Research Topics

PhD in Economics | Finance – 2019/2020

Lecture notes - Financial Database Management

Pedro Pires
Course contents:

➢ **Section I**

- Overview of WRDS
- Accessing WRDS - Web query in detail
- Data management in STATA
  - Aggregating/Summarizing data
  - Compounding returns
  - Merging datasets
  - Fuzzy matching

➢ **Section II**

- Brief introduction to SAS and SQL
- Data management in SAS Studio
  - I/O (Input-Output) files
  - Querying data
  - Aggregating/Summarizing data
  - Merging datasets
  - Fuzzy matching
## Overview of WRDS

### Financial Information by Source

#### Company Financials
- Balance Sheet, Income Statement, and Cash Flows
- Annual and Quarterly Frequencies
- Industrial and Financial Services (Banks) Format

#### Financial Markets
- Stock Prices, Returns, Shares Outstanding and Volume
- Daily and Monthly Frequencies
- Index, CDS, Bonds, Interest Rates, Futures, Options, Swaps

#### Ownership
- Institutional and Mutual Fund Holdings
- Ownership of securities at the institution level (e.g. Fidelity)
- Ownership by mutual funds (e.g. Fidelity Magellan)

#### Other
- Bond Ratings
- Executive Compensation
- Analyst Forecasts (e.g. future EPS, recommendations)
- IPOs, SEOs, M&A
- Syndicated Loans

#### US: Compustat North America
##### Worldwide: WorldScope, Factset Fundamentals, Compustat Global

#### Banks: Compustat Bank, BvD BankFocus

#### Private: BvD Orbis, BvD Amadeus

#### US: CRSP, Compustat North America
##### Worldwide: DataStream, Bloomberg

#### US: CRSP Mutual Funds, Thomson Reuters
##### Worldwide: FactSet, Lipper

#### US: Compustat Execucomp, Boardex
##### Worldwide: Bloomberg, IBES, Thomson One, Eikon, SDC, DealScan
Contents coverage

➢ **Company fundamentals**
  - Compustat North America
  - Compustat Bank
  - FactSet Fundamentals
  - Bureau Van Dijk
  - Thomson Reuters Eikon - DataStream/WorldScope

➢ **Segments: Geographical breakdown of sales**
  - FactSet Revere
  - Compustat Historical Segments

➢ **Issuer credit ratings**
  - Compustat S&P Ratings

➢ **Analyst estimates**
  - Thomson Reuters IBES

➢ **Executive compensation**
  - Compustat Execucomp – Executive Compensation
  - Boardex

➢ **M&A and Equity/Debt underwriting**
  - Thomson Reuters Eikon
  - Thomson Reuters SDC
Contents coverage

➢ Stocks
   • CRSP
   • Compustat Supplemental Short Interest File

➢ Database management
   • CRSP-Compustat Merged (CCM)
   • Merging CRSP and Compustat by CUSIP
   • Fuzzy Matching

➢ Institutional ownership and mutual fund holdings
   • CRSP Mutual Funds
   • Thomson Reuters - Mutual Fund Holdings
   • Lipper
   • Thomson Reuters Institutional Holdings
   • FactSet LionShares Ownership

➢ Syndicated loans
   • Thomson Reuters LPC DealScan

➢ Others:
   • Eventus
   • WorldBank WDI
   • IMF Macroeconomic and financial data
Accessing WRDS

Register for a WRDS Account

You must have an account to use WRDS. Please use the form below to request one. If you already have one and want to transfer it, click here.

Institution: [University of Lisbon]  
Institution not listed?  

Affiliation With Institution: Ph.D.  
What affiliation type should I use?  

First Name:  

Last Name:  

Email (assigned by your institution):  

Department (optional):  

Desired Username:  

[Register button]
Three ways of using WRDS:

- 3 complementary ways of using WRDS
  - Web Query
  - UNIX
    - Secure SHell (SSH) Client (please install the WinSCP client)
  - SAS Studio
    - PCSAS (SAS)

- You can download the same exact data using any of the alternative methods. Different strengths.

- In this course we will focus on (1) managing data obtained through Web Queries (using STATA); and (2) SAS Studio.
(1) Web Query

- Just need an **Internet Browser**
- Simple **Point & Click query interface**
- Access to search and help tools
- Choice of **different output formats**: txt, csv,xlsx, sas, stata, spss
- Instantaneously **save output to your PC**
Accessing WRDS

(2.1) Install WinSCP for Windows or an alternative for Mac

[Image of WinSCP website]

WinSCP 5.15 is a major application update. New features and enhancements include:

- Files can be optionally encrypted when storing them on SFTP server.
- Local UNC paths can be browsed.
- Compare Files extensions.
- Dark theme.
- Files can be copied to the clipboard.
- Coloring files in file panels based on a file mask.
- Improved incremental search in file panels.
- Support OpenSSH AES-256-CTR-encrypted keys.
- Improvements to directory synchronization.
- List of all changes.

[Image of WinSCP license agreement]

License Agreement
Please read the following important information before continuing.

Please read the following License Agreement. You must accept the terms of this agreement before continuing with the installation.

You can also review this license and further details online at:
https://winscp.net/eng/docs/license

A. GNU General Public License
B. License of WinSCP Icon Set
C. Privacy Policy

A. GNU GENERAL PUBLIC LICENSE
Version 3, 29 June 2007

Copyright (C) 2007 Free Software Foundation, Inc. <https://www.fsf.org/>
Everyone is permitted to copy and distribute verbatim copies of this license

[Image of WinSCP setup window]

Help
Accept > Cancel
(2.1) UNIX SAS - SSH Client

- Needs a SSH Client (e.g. WinSCP)
  → Host: wrds-cloud.wharton.upenn.edu
- Programming in SAS (No license needed) using a text editor (e.g. Notepad)
- Up to 16 GB RAM per process (Max 5 jobs)
- 10 GB individual storage space in directory /home/unlpt/username
- 4 TB temporary disk space located in /sastemp (deleted in 1 week or full)
Accessing WRDS

(2.1) UNIX SAS - SSH Client

Advanced Settings

- Remote directory: /home/unlpt/username
- Uncheck “Remember last used directory”
- Connection: Timeouts – 60 seconds
(2.1) UNIX SAS - SSH Client (Programming in SAS)

- Write a SAS program using a text editor
- Upload the file to /home/unlpt/username (make sure the file extension is .sas)
- Run the program in the command window (qsas filename)
- Download the file stored in the WRDS server to your pc
(2.2) SAS Studio

- SAS studio is a **friendly visual interface** to the SAS server running on WRDS
- All standard browsers are supported (Firefox, Google Chrome, Internet Explorer, Safari)
(3) PCSAS

- Needs a SAS License (software installed on your PC)
- Remote access to the WRDS server
  \[ \text{wrds-cloud.wharton.upenn.edu} \]
- Easier to edit and debug programs
Web Query Characteristics

➢ Select Data Source
  • Select Database (e.g. COMPUSTAT)
  • Select Dataset (e.g. COMPUSTAT/North America/Fundamentals Annual)

➢ Uniform Query Interface (whatever the database you use in WRDS)
  • Step 1: Select Date Range
  • Step 2: Select Companies/Entities (e.g. IBM, MSFT, XOM)
  • Step 3: Select Variables (e.g. return, price, total assets, net income)
  • Step 4: Select output format (e.g. txt, xlsx, csv, stata, sas)

➢ Download file to your PC
Web Query in detail

COMPUSTAT (Web Query)

➢ Select Dataset

[Image of COMPUSTAT dataset selection]

Compustat - Capital IQ from Standard & Poor's

For more about this dataset, see the Dataset List, Manuals and Overviews or FAQs.

Compustat

Databases in this section are updated every day unless otherwise noted. Update schedules should not be confused with end-of-day or end-of-month data such as stock prices.

North America - Annual Updates

Point & Click

[Image of Point & Click feature]

As a reminder, subscribers to Compustat Annual Updates also have access to Compustat Daily Updates, which can be accessed at this link. The data structure between the two databases is identical, with the only difference being how often each database is updated with new data.

For more about this dataset, see the Dataset List, Manuals and Overviews or FAQs.
Web Query in detail

**COMPUSTAT (Web Query)**

- **Step 1: Select Date Range**

  Set Begin and End Dates (in Fiscal or Calendar Year)

  **Help pages:**
  - Variable descriptions
  - Dataset list
  - Manuals and FAQs
Web Query in detail

COMPUSTAT (Web Query)

- Step 2: Select Companies / Entities

**Step 2: Apply your company codes.**

- TIC
- GVKEY
- CUSIP
- SIC
- NAICS
- CIK

Select an option for entering company codes:

- **Option 1:** Insert company identifier (code lookup tool)
- **Option 2:** Browse and upload file with list of identifiers
- **Option 3:** Entire dataset

Search by: Ticker, GVKEY, CUSIP, ...

Example: IBM MSFT DELI

Save code list

Search the entire database

This method allows you to search the entire database of records. Please be aware that this method can take a very long time to run because it is dependent upon the size of the database.
Web Query in detail

COMPUSTAT (Web Query)

➢ Step 3: Select Variables

Screening Variables

Several screening variables are pre-selected to produce one record per GVKEY-DATADATE pair, while keeping the vast majority of records. Examples of excluded rows include those with restated data, different views of the same data (pro forma, pre-FASB). Click on each variable for a more detailed explanation.

Click on the question mark button to get detailed information on each item
Web Query in detail

COMPUSTAT (Web Query)

➢ Step 2b: Conditional Statements

Conditional Statements (Optional)

How does this work?

Example 1

- Assets $\geq$ 10000 millions
- Employees $\geq$ 500

Example 2

- R&D expense $> 0$
- Standard Industry Classification Code (SIC) = 2834 OR 3674
Web Query in detail

COMPUSTAT (Web Query)

➢ Step 4: Select Output Format

**Step 4: Select query output.**
Select the desired format of the output file. For large data requests, select a compression type to expedite downloads. If you enter your email address, you will receive an email that contains a URL to the output file when the data request is finished processing.

- **Output Format**
  - fixed-width text (*.txt)
  - comma-delimited text (*.csv)
  - Excel spreadsheet (*.xlsx)
  - tab-delimited text (*.txt)
  - HTML table (*.html)
  - SAS Windows_32 dataset (*.sas7bdat)
  - SAS Windows_64 dataset (*.sas7bdat)
  - SAS Solaris_64 dataset (*.sas7bdat)
  - dBase file (*.dbf)
  - STATA file (*.dta)
  - SPSS file (*.sav)

- **Compression Type**
  - None
  - zip (*.zip)
  - gzip (*.gz)

- **Date Format**
  - YYMDDn8. (e.g. 19840725)
  - DATE9. (e.g. 25JUL1884)
  - DDMYY8. (e.g. 250784)
  - MDDYY10. (e.g. 07/25/1984)
  - DDMYY10. (e.g. 25/07/1984)
  - YYMDDs10. (e.g. 1984/07/25)

E-Mail Address *(Optional)*

E-mail

Edit Preferences

Custom Field *(Optional)*

Save this query to myWRDS

Query Name

Submit Query
Web Query in detail

COMPUSTAT (Web Query)

➢ Save File to your PC

Data Request Summary

Your output is complete. Click on the link below to open the output file.

ebde6dfa316fcf54.dta (4 KB, 9 observations 17 variables)

Download instructions
Internet Explorer and Firefox users… Right-click and select "Save Target As…"

Citation instructions
To cite this data use the following format:


<table>
<thead>
<tr>
<th>Data Request ID</th>
<th>ebde6dfa316fcf54</th>
</tr>
</thead>
<tbody>
<tr>
<td>Libraries/Data Sets</td>
<td>compa/funda /</td>
</tr>
<tr>
<td>Frequency/Date Range</td>
<td>ann / 01Jan2010 - 31Jul2019</td>
</tr>
<tr>
<td>Search Variable</td>
<td>TIC</td>
</tr>
<tr>
<td>Input Codes</td>
<td>IBM</td>
</tr>
<tr>
<td>Conditional Statements</td>
<td>n/a</td>
</tr>
<tr>
<td>Output format/Compression</td>
<td>dta /</td>
</tr>
<tr>
<td>Variables Selected</td>
<td>CONM AT NI SALE XRD CAPX EMP</td>
</tr>
<tr>
<td>Extra Variables and Parameters Selected</td>
<td>C INDL STD</td>
</tr>
</tbody>
</table>

• Save File when output is complete
• Right-Click and Select “Save Target (Link) As”
• Don’t Left-Click the URL for large files
### COMPUSTAT (Web Query)

**Example: IBM**

<table>
<thead>
<tr>
<th>gvkey</th>
<th>endDate</th>
<th>fycast</th>
<th>tic</th>
<th>at</th>
<th>capx</th>
<th>EMP</th>
<th>nl</th>
<th>sale</th>
<th>xrd</th>
</tr>
</thead>
<tbody>
<tr>
<td>00666</td>
<td>31dec2010</td>
<td>2010</td>
<td>IBM</td>
<td>113452</td>
<td>4185</td>
<td>426.751</td>
<td>14833</td>
<td>99871</td>
<td>5720</td>
</tr>
<tr>
<td>00666</td>
<td>31dec2011</td>
<td>2011</td>
<td>IBM</td>
<td>116433</td>
<td>4108</td>
<td>433.362</td>
<td>15855</td>
<td>106916</td>
<td>5990</td>
</tr>
<tr>
<td>00666</td>
<td>31dec2012</td>
<td>2012</td>
<td>IBM</td>
<td>119213</td>
<td>4081</td>
<td>434.246</td>
<td>16604</td>
<td>104507</td>
<td>6034</td>
</tr>
<tr>
<td>00666</td>
<td>31dec2013</td>
<td>2013</td>
<td>IBM</td>
<td>126223</td>
<td>5623</td>
<td>431.212</td>
<td>16483</td>
<td>89751</td>
<td>5859</td>
</tr>
<tr>
<td>00666</td>
<td>31dec2014</td>
<td>2014</td>
<td>IBM</td>
<td>117552</td>
<td>3740</td>
<td>379.592</td>
<td>12022</td>
<td>92793</td>
<td>5437</td>
</tr>
<tr>
<td>00666</td>
<td>31dec2015</td>
<td>2015</td>
<td>IBM</td>
<td>110495</td>
<td>3579</td>
<td>377.757</td>
<td>15190</td>
<td>81741</td>
<td>5247</td>
</tr>
<tr>
<td>00666</td>
<td>31dec2016</td>
<td>2016</td>
<td>IBM</td>
<td>117470</td>
<td>3567</td>
<td>380.3</td>
<td>11872</td>
<td>79920</td>
<td>5751</td>
</tr>
<tr>
<td>00666</td>
<td>31dec2017</td>
<td>2017</td>
<td>IBM</td>
<td>125356</td>
<td>3229</td>
<td>366.6</td>
<td>5753</td>
<td>79139</td>
<td>5787</td>
</tr>
<tr>
<td>00666</td>
<td>31dec2018</td>
<td>2018</td>
<td>IBM</td>
<td>123382</td>
<td>3396</td>
<td>350.6</td>
<td>5728</td>
<td>79591</td>
<td>5379</td>
</tr>
</tbody>
</table>

**Consolidated Statement of Financial Position**

International Business Machines Corporation and Subsidiary Companies

($) in millions except per share amounts

<table>
<thead>
<tr>
<th>At December 31:</th>
<th>Notes</th>
<th>2018</th>
<th>2017</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total assets</td>
<td></td>
<td>$123,382</td>
<td>$125,356</td>
</tr>
</tbody>
</table>

**Employees**

**Variable Name = EMP**

This item represents the number of people employed by the company and its consolidated subsidiaries in thousands.
Company fundamentals

• **Compustat North America** [Coverage: North America | Public]

• **Compustat Global [Not Subscribed]** [Coverage: Global | Public]

• **Compustat Bank** [Coverage: US Banking Institutions]

• **FactSet Fundamentals** [Coverage: Global | Public + some Private]

• **Bureau Van Dijk – Orbis** [Coverage: Global | Public + Private]

• **Bureau Van Dijk – Other [Not Subscribed]**
  - **Amadeus** [Coverage: Europe | Public + Private]
  - **BankFocus** [Coverage: Global | Banks | Public + Private]
  - **Isis** [Coverage: Global | Insurance companies | Public + Private]
  - **Osiris** [Coverage: Global | Major companies | Public + Private]

• **Outside WRDS...**
  - **DataStream/WorldScope** [Coverage: Global | Public]
  - **Bloomberg** [Coverage: Global | Public] *

* Not covered in this course
Financial reports for worldwide firms – Comparative Analysis


https://wrds-www.wharton.upenn.edu/documents/692/March_16_2012_international_0HjpzsR.pdf

1) **Compustat Global** features greater coverage of large companies in more developed countries and provides a wider range of accounting data items than any other databases;

2) **BvD Osiris** offers lesser variety of accounting data items, but it also contains a higher number of small firms from developing countries;

3) **FactSet Fundamentals** Database provides a balance in the firm size and quantity of accounting items with a reasonable geographical coverage.
✓ **Compustat North America Fundamentals**

- **Understanding the Data:**
  - Coverage: North America (US, Canada)
  - Over 31,000 public companies (14,650 active)
  - Includes ADR companies
  - Company fundamental annual data beginning in 1950
  - Monthly market data beginning in 1962
  - Fiscal and Calendar presentations of the data
Compustat North America Fundamentals

- **GVKEY:** Global Company Key (GVKEY) is a unique identifier that represents each company in COMPUSTAT.

- **DATADATE:** Data Date represents the reporting date for a data record.

- **{GVKEY; DATADATE}** DATADATE is often combined with GVKEY to make company records unique.

- **COMMON IDENTIFIERS:** CUSIP, SEDOL, ISIN, TICKER

- **CURRENCY:** Data expressed in Local Currency.
Web Query in detail (Compustat)

Compustat North America Fundamentals

➢ Compustat has both active and inactive companies. Field DLRSN indicates the reason a company became inactive.
Compustat North America Fundamentals

➢ Compustat is mainly used for its Fundamentals data...

➢ But Compustat also has other useful datasets:
  ➢ Index Constituents
  ➢ Ratings
  ➢ Security Daily/Monthly
  ➢ Bank Fundamentals

➢ Compustat is also the vendor of other databases: EXECUCOMP and CAPITAL IQ
Compustat Bank Fundamentals contains financial data of the biggest and most important banks in the US since 1950.

In accounting there is a difference between balance sheets and income statements for industrial companies and Banks.

Compustat Industrial Annual includes both industrial companies and banks, with banks financial reports represented as Industrial statements.

Compustat Bank annual features bank related items on the balance sheet and income statement.
✓ **FactSet Fundamentals** *(currently, not subscribed)*

- FactSet purchased a copy of the Thomson Worldscope Database in April 2008, with historical information for over 43,000 companies dating back to 1980.
- FactSet covers more than 73,000 public and private companies in the Americas, Europe/Africa and Asia-Pacific regions.
- There are over 5,000 private companies within the Fundamentals database.
**FactSet Fundamentals – Global coverage**

**Example:** Find the sales and total assets for firms incorporated in Portugal with financial reports during 2017.

**Dataset: Fundamentals - International (v2)**

<table>
<thead>
<tr>
<th>factset_em-d</th>
<th>entity_prop_name</th>
<th>isin</th>
<th>iso_country-p</th>
<th>date</th>
<th>ff_sales</th>
<th>ff_assets</th>
</tr>
</thead>
<tbody>
<tr>
<td>0C88HS-K</td>
<td>Banco Comercial Português SA</td>
<td>PTBCP0AM0015</td>
<td>PT</td>
<td>31Dec2017</td>
<td>2972.641959559555</td>
<td>72041.625</td>
</tr>
<tr>
<td>0C88JS-K</td>
<td>Twerol SGPS SA</td>
<td>PTTPC0AM0009</td>
<td>PT</td>
<td>31Dec2017</td>
<td>0</td>
<td>269.1045390559559597</td>
</tr>
<tr>
<td>0C60PQ-E</td>
<td>Litho Formas SA</td>
<td>PTLIT0UE0005</td>
<td>PT</td>
<td>31Dec2017</td>
<td>2.4729746</td>
<td>2.897246909559555999</td>
</tr>
<tr>
<td>0C97FR-K</td>
<td>Sport Lisboa e Benfica-Futebol SA</td>
<td>PTSLB0AM0010</td>
<td>PT</td>
<td>30Jun2017</td>
<td>128.2350000000000001</td>
<td>507.7120000000000002</td>
</tr>
<tr>
<td>0C831E-K</td>
<td>Altri SGPS SA</td>
<td>PTALTOAE0002</td>
<td>PT</td>
<td>31Dec2017</td>
<td>456.03516</td>
<td>1210.0855065559555999</td>
</tr>
<tr>
<td>0566G7-K</td>
<td>Banco BPI SA</td>
<td>PTBPB0AM0004</td>
<td>PT</td>
<td>31Dec2017</td>
<td>834.3809999999999997</td>
<td>29640.20899999999999</td>
</tr>
<tr>
<td>0567DP-E</td>
<td>Caixa Geral de DepsCitos SA</td>
<td>PTCDP0AM0004</td>
<td>PT</td>
<td>31Dec2017</td>
<td>3419.9084119559559990</td>
<td>93247.91395999999999</td>
</tr>
<tr>
<td>056867-K</td>
<td>BRISA - Auto-estradas de Portugal SA</td>
<td>PTBRI0AM0004</td>
<td>PT</td>
<td>31Dec2017</td>
<td>670.9180000000000001</td>
<td>3383.703</td>
</tr>
<tr>
<td>05992S-K</td>
<td>EDP-Energias de Portugal SA</td>
<td>PTCED0AM0009</td>
<td>PT</td>
<td>31Dec2017</td>
<td>15745.94699999999999</td>
<td>42075.04899999999999</td>
</tr>
<tr>
<td>0S8J0N-E</td>
<td>The Navigator Co. SA</td>
<td>PTPTN0AM0006</td>
<td>PT</td>
<td>31Dec2017</td>
<td>1636.8341600000000001</td>
<td>2439.13629595599999</td>
</tr>
<tr>
<td>0S8J2L-K</td>
<td>Jerônimo Martins SGPS SA</td>
<td>PTCJN0AM0001</td>
<td>PT</td>
<td>31Dec2017</td>
<td>14276.48</td>
<td>48411.01499999999999</td>
</tr>
<tr>
<td>0S8J2M-K</td>
<td>Semaço Sociedade de Investimento e Gestao SGPS SA</td>
<td>PTSEM0AM0004</td>
<td>PT</td>
<td>31Dec2017</td>
<td>2164.6532775955959980</td>
<td>3556.887071</td>
</tr>
<tr>
<td>0S8J2Y-K</td>
<td>SonaeCM SGPS SA</td>
<td>PTSCN0AM0006</td>
<td>PT</td>
<td>31Dec2017</td>
<td>139.55982</td>
<td>1105.608969</td>
</tr>
<tr>
<td>0S8J6E-K</td>
<td>Mota-Engil SGPS SA</td>
<td>PTPME0AE0005</td>
<td>PT</td>
<td>31Dec2017</td>
<td>2597.2935999999999999</td>
<td>4614.08099999999999</td>
</tr>
<tr>
<td>0S8J0O-K</td>
<td>SAG GEST - Solucoes Automovel Globais SA</td>
<td>PTSGE0AM0004</td>
<td>PT</td>
<td>31Dec2017</td>
<td>619.7074999999999998</td>
<td>550.1285000000000003</td>
</tr>
<tr>
<td>0S8J1L-K</td>
<td>Novaface SGPS SA</td>
<td>PTNFA0AM0006</td>
<td>PT</td>
<td>31Dec2017</td>
<td>139.7210000000000001</td>
<td>184.1226</td>
</tr>
<tr>
<td>0S8J1R-E</td>
<td>Imprese SGPS SA</td>
<td>PTIMP0AM0000</td>
<td>PT</td>
<td>31Dec2017</td>
<td>290.6781259999999999</td>
<td>390.0922930000000001</td>
</tr>
<tr>
<td>0S8J2Y-K</td>
<td>Inapa-Investimentos, Participações e Gestões SA</td>
<td>PTINH0AE0008</td>
<td>PT</td>
<td>31Dec2017</td>
<td>894.4999999999999996</td>
<td>670.7100000000000004</td>
</tr>
<tr>
<td>0S8J30-E</td>
<td>COFINA SGPS SA</td>
<td>PTCO0AM0003</td>
<td>PT</td>
<td>31Dec2017</td>
<td>91.035724</td>
<td>113.99072</td>
</tr>
<tr>
<td>0S8J3M-E</td>
<td>Corticeira Amorim SGPS SA</td>
<td>PTCOR0AE0006</td>
<td>PT</td>
<td>31Dec2017</td>
<td>701.1690000000000004</td>
<td>869.4070000000000004</td>
</tr>
<tr>
<td>0S8J69-E</td>
<td>Toyota Coeteno Portugal SA</td>
<td>PTSTC0AP0018</td>
<td>PT</td>
<td>31Dec2017</td>
<td>390.0335299999999996</td>
<td>298.4806700000000001</td>
</tr>
<tr>
<td>0S8J6H-E</td>
<td>Iberoil SGPS SA</td>
<td>PTISI0AM0008</td>
<td>PT</td>
<td>31Dec2017</td>
<td>466.329084</td>
<td>436.9633020000000001</td>
</tr>
<tr>
<td>0S8J6N-E</td>
<td>Estoril Sol SGPS SA</td>
<td>PTESO0AM0005</td>
<td>PT</td>
<td>31Dec2017</td>
<td>107.487377</td>
<td>180.0428300000000001</td>
</tr>
<tr>
<td>0S8J6W-E</td>
<td>Pardachica-Participações FClimax SGPS SA</td>
<td>PTFCO0AM0000</td>
<td>PT</td>
<td>31Dec2017</td>
<td>932.8999999999999995</td>
<td>13902.038</td>
</tr>
<tr>
<td>0S8J72-E</td>
<td>Sonae SGPS SA</td>
<td>PTSON0AM0001</td>
<td>PT</td>
<td>31Dec2017</td>
<td>5710.12193000000002</td>
<td>5604.6022359999999999</td>
</tr>
</tbody>
</table>
FactSet Fundamentals – Global coverage

**Example:** Find the sales and total assets for firms incorporated in Portugal with financial reports during 2017.

**Dataset:** Fundamentals - International (v2)

<table>
<thead>
<tr>
<th>factset_en-d</th>
<th>entity_proper_name</th>
<th>isin</th>
<th>iso_country</th>
<th>date</th>
<th>ff_sales</th>
<th>ff_assets</th>
</tr>
</thead>
<tbody>
<tr>
<td>059B85-E</td>
<td>Sonae Capital, SGPS, SA</td>
<td>PTSP0AEB0008</td>
<td>PT</td>
<td>31Dec2017</td>
<td>177,156,765,000,000</td>
<td>516,126,605,000,000</td>
</tr>
<tr>
<td>059B86-E</td>
<td>Futebol Clube do Porto SAD</td>
<td>PTFC0AEB0005</td>
<td>PT</td>
<td>30Jun2017</td>
<td>97,670,327,999,9999</td>
<td>375,424,937,999,9999</td>
</tr>
<tr>
<td>059B87-E</td>
<td>Glintt-Global Intelligent Technologies SGPS</td>
<td>PTTF0AEB0007</td>
<td>PT</td>
<td>31Dec2017</td>
<td>71,066,651,000,000</td>
<td>166,022,024,000,000</td>
</tr>
<tr>
<td>059B88-E</td>
<td>Gia Portuguesa de Almocas SA</td>
<td>PTGPA0AEB0006</td>
<td>PT</td>
<td>30Jun2017</td>
<td>32,012,622,000,000</td>
<td>19,026,263,000,000</td>
</tr>
<tr>
<td>059B89-E</td>
<td>Sporting Clube de Portugal Futebol SAD</td>
<td>PTCFC0AEB0010</td>
<td>PT</td>
<td>30Jun2017</td>
<td>61,402,999,999,9999</td>
<td>345,497,000,000,000</td>
</tr>
<tr>
<td>059B90-E</td>
<td>Compta Equipamentos e Servicos de Informatica SA</td>
<td>PTETE0AEB0007</td>
<td>PT</td>
<td>31Dec2017</td>
<td>29,128,764,000,000</td>
<td>26,762,660,000,000</td>
</tr>
<tr>
<td>059B91-E</td>
<td>Reditux SA</td>
<td>PTRED0AEB0010</td>
<td>PT</td>
<td>31Dec2017</td>
<td>41,136,9999</td>
<td>171,125,670,000,000</td>
</tr>
<tr>
<td>059B92-E</td>
<td>Imobiliaria Construtores GuDo-Parol SA</td>
<td>PTGU0AEB0007</td>
<td>PT</td>
<td>31Dec2017</td>
<td>0,029,999,999,999,999</td>
<td>62,410,301,000,000</td>
</tr>
<tr>
<td>059B93-E</td>
<td>VNL Vista Alegre Atlantis SGPS SA</td>
<td>PTVVA0AEB0001</td>
<td>PT</td>
<td>31Dec2017</td>
<td>64,930,000,000,000</td>
<td>277,816</td>
</tr>
<tr>
<td>059B94-E</td>
<td>CIFAN Companhia Industrial Produtora de Antibioticos SA</td>
<td>PTCP0AEB0002</td>
<td>PT</td>
<td>31Dec2017</td>
<td>19,262,819,279,9999</td>
<td>27,820,152,570,000</td>
</tr>
<tr>
<td>059B95-E</td>
<td>Lisergtica - Impressao e Artes Graficas SA</td>
<td>PTLIG0AEB0002</td>
<td>PT</td>
<td>31Dec2017</td>
<td>16,969,491,999,9999</td>
<td>16,603,259,000,000</td>
</tr>
<tr>
<td>059B96-E</td>
<td>Galp Energia SGPS SA</td>
<td>PTGAL0AEB0005</td>
<td>PT</td>
<td>31Dec2017</td>
<td>1,151,530</td>
<td>12,350</td>
</tr>
<tr>
<td>059B97-E</td>
<td>CTT - Correios de Portugal, SA</td>
<td>PTCTC0AEB0001</td>
<td>PT</td>
<td>31Dec2017</td>
<td>703,582,163,000,000</td>
<td>1,606,765,930</td>
</tr>
<tr>
<td>059B98-E</td>
<td>Luiz Sodre SA</td>
<td>PTETE0AEB0005</td>
<td>PT</td>
<td>31Dec2017</td>
<td>401,924,599</td>
<td>659,213,531,999,999</td>
</tr>
<tr>
<td>059B99-E</td>
<td>Grupo Media Capital SGPS SA</td>
<td>PTGMC0AEB0003</td>
<td>PT</td>
<td>31Dec2017</td>
<td>126,910,591,999,999</td>
<td>301,244,316,999,999</td>
</tr>
<tr>
<td>059C01-E</td>
<td>Constru Engenharia SA</td>
<td>PTCEU0AEB0003</td>
<td>PT</td>
<td>31Dec2017</td>
<td>140,878,156,999,999</td>
<td>421,540,036,999,999</td>
</tr>
<tr>
<td>059C02-E</td>
<td>Empresa de Desenvolvimento e Infra-estruturas do Alqueva SA</td>
<td>PT5180AEB0005</td>
<td>PT</td>
<td>31Dec2017</td>
<td>28,180,444,999,9999</td>
<td>812,591,16</td>
</tr>
<tr>
<td>059C03-E</td>
<td>Sonae Sierra SGPS SA</td>
<td>PTSG0AEB0008</td>
<td>PT</td>
<td>31Dec2017</td>
<td>181,411,11</td>
<td>2,822,083,999,9999</td>
</tr>
<tr>
<td>059C04-E</td>
<td>Martifer SGPS</td>
<td>PTMFR0AEB0003</td>
<td>PT</td>
<td>31Dec2017</td>
<td>116,871,771</td>
<td>376,100,489,999,999</td>
</tr>
<tr>
<td>059C05-E</td>
<td>Sonagi SGPS</td>
<td>PTSON0AEB0007</td>
<td>PT</td>
<td>31Dec2017</td>
<td>7,218,749,999,9999</td>
<td>117,508,84</td>
</tr>
<tr>
<td>059C06-E</td>
<td>Sonae Industria SGPS SA</td>
<td>PT38P0AEB0025</td>
<td>PT</td>
<td>31Dec2017</td>
<td>230,978,239,999,999</td>
<td>401,032,709,999,999</td>
</tr>
<tr>
<td>059C07-E</td>
<td>Redes Energeticas Nacionais SGPS SA</td>
<td>PTREN0AEB0005</td>
<td>PT</td>
<td>31Dec2017</td>
<td>746,147,000,000,000</td>
<td>928,963,000,000,000</td>
</tr>
<tr>
<td>059C08-E</td>
<td>NOV SGPS</td>
<td>PTNOV0AEB0006</td>
<td>PT</td>
<td>31Dec2017</td>
<td>1,546,939,999,9999</td>
<td>2567,067</td>
</tr>
<tr>
<td>059C09-E</td>
<td>Ramada Investimentos e Industria SA</td>
<td>PTTRV0AEB0004</td>
<td>PT</td>
<td>31Dec2017</td>
<td>156,646,600,000,000</td>
<td>314,116,080,000,000</td>
</tr>
<tr>
<td>059C10-E</td>
<td>Sunol + Compad SA</td>
<td>PTSL0AEB0009</td>
<td>PT</td>
<td>31Dec2017</td>
<td>356,130,160,000,000</td>
<td>646,219,395,000,000</td>
</tr>
<tr>
<td>059C11-E</td>
<td>Teixeira Duarte SA</td>
<td>PTTD0AEB0000</td>
<td>PT</td>
<td>31Dec2017</td>
<td>1,035,637,999,999,999</td>
<td>2,294,353,999,999,999</td>
</tr>
<tr>
<td>059C12-E</td>
<td>Brise Concessao Rodoviaria SA</td>
<td>PTBT0AEB0000</td>
<td>PT</td>
<td>31Dec2017</td>
<td>576,590,221,999,9999</td>
<td>2974,026,269</td>
</tr>
</tbody>
</table>
✓ **BvD – Bureau Van Dijk**

**Orbis**
- Flagship product, information on over 220 million public and private companies worldwide

**Amadeus (not subscribed)**
- European public and private company information (21 million companies)

**BankFocus (not subscribed)**
- Detailed information on 38,000 banks (28,000 US and 10,000 Non-US)

**Isis (not subscribed)**
- Detailed information on public and private insurance companies around the world. ISIS contains information on 11,700 companies in 190 different countries.

**Osiris (not subscribed)**
- Listed and major unlisted/delisted companies globally (80,000 companies)
Web Query in detail (BvD Orbis)

✓ BvD Orbis

➢ Company Financials
  ➢ Financials for Industrial Companies
  ➢ Financials for Banks
  ➢ Financials for Insurance Companies

➢ Company size categories
  ➢ Very large companies (VL)
  ➢ Large companies (L)
  ➢ Medium companies (M)
  ➢ Small companies (S)

➢ Shareholders
  ➢ All current shareholders First Level
BvD Orbis

- **Company size categories**

  - **Very large companies (VL)**
    - Companies that match at least one of the following conditions:
      - Operating Revenue $\geq 100$ million EUR ($130$ million USD)
      - Total assets $\geq 200$ million EUR ($260$ million USD)
      - Employees $\geq 1,000$
      - Listed

  - **Large companies (L)**
    - Operating Revenue $\geq 10$ million EUR ($13$ million USD)
    - Total assets $\geq 20$ million EUR ($26$ million USD)
    - Employees $\geq 150$
    - Not Very Large

  - **Medium companies (M)**
    - Operating Revenue $\geq 1$ million EUR ($1.3$ million USD)
    - Total assets $\geq 2$ million EUR ($2.6$ million USD)
    - Employees $\geq 15$
    - Not Very Large or Large

  - **Small companies (S)**
    - Companies not included in another category
BvD Orbis

Consolidation codes

- For a given company, ORBIS provides financial statements with different consolidation codes:
  - C1: consolidated account aggregating all companies belonging to the group (affiliates, subsidiaries, etc.), where the company has no unconsolidated account
  - C2: consolidated account aggregating all companies belonging to the group (affiliates, subsidiaries, etc.) where the company also presents an unconsolidated account
  - U1: unconsolidated account of a company with no consolidated account
  - U2: unconsolidated account of a company which also has a consolidated account
BvD Orbis

Survivorship bias (data availability up to 5/6 years)


Known issues:

- Download speed (the BvD is in general slow)
- Survivorship bias: Both Orbis (up to 5 years) and Amadeus (at most 10 years) contain a number of years of historic financial data
- Reporting lag (about 2 years)

EDP – Energias de Portugal:
Web Query in detail (BvD Orbis)

BvD Orbis – Web Interface

Orbis Web Interface @ Teresa and Alexandre Soares dos Santos Library
Access to resources

All the resources are available to Nova SBE registered students, faculty, and staff.
If you are not affiliated with Nova SBE and you wish to access Social Sciences DataLab resources, please fill in this form.

Microdata sets

Please read the access conditions here.
Contents coverage

➢ **Segments: Geographical breakdown of sales**
  
  - FactSet Revere
  
  - Compustat - Historical Segments

➢ **Segments: breakdown of sales by customers**
  
  - Compustat - Customer Segments
 ✓ Factset Revere

➢ Product designed to answer 3 questions: What do companies do? Who do they work with? Where do they sell?

➢ GeoRev focus on geographic revenue (Where do companies sell?)

➢ GeoRev currently covers more than 14,000 publicly traded companies around the world with historical data going back as far as 2005.

![Figure 2: Intel’s % Revenues by Geography as disclosed in FY 2011 10-K](image)

- Taiwan: 31.6%
- United States: 15.6%
- China: 15.0%
- Europe: 12.9%
- Asia/Pacific: 10.2%
- Japan: 9.3%
- Americas (ex. US): 5.4%

4 out of the 7 disclosed segments are country specific and require no further estimation.

Any inquiry regarding revenues exposure to countries beside Taiwan, U.S., China and Japan will require algorithmic estimations.
Web Query in detail (Revere)

Factset Revere

<table>
<thead>
<tr>
<th>Features</th>
<th>FactSet GeoRev™</th>
<th>MSCI Economic Exposure Security Data Modules</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Sources</td>
<td>Public filings</td>
<td>Public filings</td>
</tr>
<tr>
<td>Estimation Factor</td>
<td>GDP weight</td>
<td>GDP weight</td>
</tr>
<tr>
<td>History</td>
<td>Up to 9 years</td>
<td>Up to 9 years (only for = 2,500 large and mid companies in MSCI ACWI)</td>
</tr>
<tr>
<td>Taxonomy for Data Mapping and Normalization</td>
<td>Yes</td>
<td>Undisclosed</td>
</tr>
<tr>
<td>Calculation Engine</td>
<td>Estimates calculation 100% automated; no manual calculation to minimize errors</td>
<td>Undisclosed</td>
</tr>
<tr>
<td>Confidence Factor</td>
<td>Yes. Incorporates information from each company’s unique disclosure</td>
<td>Yes, but rudimentary</td>
</tr>
<tr>
<td>Data Granularity</td>
<td>280+ super-regions, regions, areas, and countries. Ideally suited for user customization</td>
<td>244 regions and countries</td>
</tr>
<tr>
<td>Coverage*</td>
<td>14,000 large, mid and small cap companies in developed and emerging markets (Russell 3000 and MSCI ACWI included)</td>
<td>8,700 companies in MSCI ACWI FM IMI</td>
</tr>
<tr>
<td>Customizable Estimation On Demand**</td>
<td>Yes</td>
<td>Undisclosed</td>
</tr>
</tbody>
</table>

* Coverage as of Feb 2014. FactSet’s coverage will expand to 15,000 companies by 2015 with a target goal of about 48,000 companies.
** Customers may have the option to request coverage of securities as we expand current coverage to entire global equity market.
Compustat – Historical Segments

- Similar to FactSet GeoRev for US companies

- Find the geographical exposure of Apple sales (AAPL) in 2018
✓ **Compustat – Historical Segments**

➢ Find the geographical exposure of Apple sales (AAPL) in 2018
Web Query in detail (Compustat Historical Segments)

Exercise COMPSEG-1: Find the geographical exposure of Apple sales (AAPL) in 2018

```
use "D:\Documents\FDM - PhD 2020\data\compsegment_aapl.dta", clear
order gvkey comm tic datadate srcrevdate stype sales curcds srcs snms
keep if stype=="GEOSEG" & sales!=0
egen total_sales = sum(sales)
gen percent = sales/total_sales
gsort -percent
format comm %--16s
format snms %--24s
br gvkey comm tic datadate srcrevdate stype sales curcds srcs snms total_sales percent
```

<table>
<thead>
<tr>
<th>gvkey</th>
<th>comm</th>
<th>tic</th>
<th>datadate</th>
<th>srcrevdate</th>
<th>stype</th>
<th>sales</th>
<th>curcds</th>
<th>srcs</th>
<th>snms</th>
<th>total_sales</th>
<th>percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>001690</td>
<td>APPLE INC</td>
<td>AAPL</td>
<td>30sep2018</td>
<td>30sep2018</td>
<td>GEOSEG</td>
<td>112093</td>
<td>USD</td>
<td>5</td>
<td>Americas</td>
<td>265595</td>
<td>.4220445</td>
</tr>
<tr>
<td>001690</td>
<td>APPLE INC</td>
<td>AAPL</td>
<td>30sep2018</td>
<td>30sep2018</td>
<td>GEOSEG</td>
<td>62420</td>
<td>USD</td>
<td>5</td>
<td>Europe</td>
<td>265595</td>
<td>.2350195</td>
</tr>
<tr>
<td>001690</td>
<td>APPLE INC</td>
<td>AAPL</td>
<td>30sep2018</td>
<td>30sep2018</td>
<td>GEOSEG</td>
<td>81942</td>
<td>USD</td>
<td>5</td>
<td>Greater China</td>
<td>265595</td>
<td>.1955654</td>
</tr>
<tr>
<td>001690</td>
<td>APPLE INC</td>
<td>AAPL</td>
<td>30sep2018</td>
<td>30sep2018</td>
<td>GEOSEG</td>
<td>21733</td>
<td>USD</td>
<td>5</td>
<td>Japan</td>
<td>265595</td>
<td>.0818276</td>
</tr>
<tr>
<td>001690</td>
<td>APPLE INC</td>
<td>AAPL</td>
<td>30sep2018</td>
<td>30sep2018</td>
<td>GEOSEG</td>
<td>17407</td>
<td>USD</td>
<td>5</td>
<td>Rest of Asia Pacific</td>
<td>265595</td>
<td>.0653396</td>
</tr>
</tbody>
</table>

Geog. Revenue Breakdown:

- Americas: 42.2%
- Asia/Pacific: 34.3%
- Europe: 23.5%
Compustat – Customer Segments

- Compustat Segments Data provides **business and geographic detail, product information** and **customer data** for over 70% of the companies in the North American (NA) database.

Coverage

- **Current Segments**: Past 8 years of information for over 8,000 North American companies.
- **Historical Segments**: Since 1976, information for over 23,000 North American companies.
Web Query in detail (Compustat Customer Segments)

Compustat Segments – Customer

**Example:** Find the % sales of Colgate-Palmolive to reported customers in fiscal year 2017.

**Dataset:** Customer Segments

**Dataset:** Fundamentals Annual
Web Query in detail (Compustat Customer Segments)

Compustat Segments – Customer

**Example:** Find the \% sales of Colgate-Palmolive to reported customers in fiscal year 2017.

```
use "C:\Users\Pedro Fires\Downloads\compsegment_customer.dta", clear
format comm %24s
format cnms %24s
merge m:1 gvkey using "C:\Users\Pedro Fires\Downloads\comp_colgatepalmolive.dta"
gen double salescs_pct = salecs/sale*100
format salescs_pct %4.2f
b r gvkey cid cnms ctype salescs sid stype srcdate comm tic cusip sale salescs_pct if gvkey=="003170" & year(srcdate)==2017 & ctype=="COMPANY"
```

<table>
<thead>
<tr>
<th>gvkey</th>
<th>cid</th>
<th>cnms</th>
<th>ctype</th>
<th>salescs</th>
<th>sid</th>
<th>stype</th>
<th>srcdate</th>
<th>comm</th>
<th>tic</th>
<th>cusip</th>
<th>sale</th>
<th>salescs_pct</th>
</tr>
</thead>
<tbody>
<tr>
<td>003170</td>
<td>9</td>
<td>Wal-Mart Stores Inc</td>
<td>COMPANY</td>
<td>1689.94</td>
<td>0</td>
<td>31Dec2017</td>
<td>COLGATE-PALMOLIVE CO</td>
<td>CL</td>
<td>194162103</td>
<td>15454</td>
<td>11.00</td>
<td></td>
</tr>
</tbody>
</table>

**Distribution; Raw Materials; Competition; Trademarks and Patents**

The Company’s products are marketed by a direct sales force at individual operating subsidiaries or business units, and by distributors or brokers. The Oral, Personal and Home Care products are sold to a variety of retail and wholesale customers and distributors. Pet Nutrition products are sold by authorized pet supply retailers and veterinarians. Many of the Company’s products are also sold online through various e-commerce platforms and retailers. The Company’s sales to Wal-Mart Stores, Inc. and its affiliates represent approximately 11\% of the Company’s Net sales in 2017. No other customer represents more than 10\% of the Company’s Net sales.
Contents coverage

- Issuer credit ratings
  - Compustat S&P Ratings (North America)
  - Alternative: Bloomberg terminal - RATC function (Global)
Web Query in detail (Compustat S&P Ratings)

✓ **S&P Ratings (data until FEB2017)**

➢ Compustat S&P Ratings from 1973
  
  ➢ **S&P Domestic Long-Term Issuer Credit Rating**

  ➢ **S&P Domestic Short-Term Issuer Credit Rating**

---

**Credit Rating Scales by Agency, Long-Term**

<table>
<thead>
<tr>
<th>Moody's</th>
<th>S&amp;P</th>
<th>Fitch</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aaa</td>
<td>AAA</td>
<td>AAA</td>
</tr>
<tr>
<td>Aa1</td>
<td>AA+</td>
<td>AA+</td>
</tr>
<tr>
<td>Aa2</td>
<td>AA</td>
<td>AA</td>
</tr>
<tr>
<td>Aa3</td>
<td>AA-</td>
<td>AA-</td>
</tr>
<tr>
<td>A1</td>
<td>A+</td>
<td>A+</td>
</tr>
<tr>
<td>A2</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>A3</td>
<td>A-</td>
<td>A-</td>
</tr>
<tr>
<td>Baa1</td>
<td>BBB+</td>
<td>BBB+</td>
</tr>
<tr>
<td>Baa2</td>
<td>BBB</td>
<td>BBB</td>
</tr>
<tr>
<td>Baa3</td>
<td>BBB-</td>
<td>BBB-</td>
</tr>
<tr>
<td>Ba1</td>
<td>BB+</td>
<td>BB+</td>
</tr>
<tr>
<td>Ba2</td>
<td>BB</td>
<td>BB</td>
</tr>
<tr>
<td>Ba3</td>
<td>BB-</td>
<td>BB-</td>
</tr>
<tr>
<td>B1</td>
<td>B+</td>
<td>B+</td>
</tr>
<tr>
<td>B2</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>B3</td>
<td>B-</td>
<td>B-</td>
</tr>
<tr>
<td>Caa1</td>
<td>CCC+</td>
<td>CCC+</td>
</tr>
<tr>
<td>Caa2</td>
<td>CCC</td>
<td>CCC</td>
</tr>
<tr>
<td>Caa3</td>
<td>CCC-</td>
<td>CCC-</td>
</tr>
<tr>
<td>Ca</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>C</td>
<td>D</td>
<td>D</td>
</tr>
<tr>
<td>/</td>
<td>D</td>
<td>D</td>
</tr>
<tr>
<td>/</td>
<td>D</td>
<td>D</td>
</tr>
</tbody>
</table>

Superscript: "Junk"

---

WOLFSTREET.com
Contents coverage

➢ Analyst estimates

  • Thomson Reuters IBES
Thomson Reuters IBES (not subscribed – temporarily)

- Institutional Brokers’ Estimate System (I/B/E/S)
- Analysts forecast data (eps, sales, etc.), consensus estimates and trade recommendations
- Global coverage starting in 1980
- More than 22,000 thousand companies and 930 brokers

For additional insights about IBES watch e-learning video from WRDS:
[http://wrds-web.wharton.upenn.edu/wrds/E-Learning/_000Video/IBES_at_WRDS/index.cfm](http://wrds-web.wharton.upenn.edu/wrds/E-Learning/_000Video/IBES_at_WRDS/index.cfm)
Summary Statistics contains one record for each forecast period for each Thomson Reuters statistical period:

- Statistical period is the date when the set of summary statistics was calculated.
- Forecast period represents the period end for which the forecasts were made for.
### I/B/E/S EPS Forecast for Portugal Telecom

<table>
<thead>
<tr>
<th>ticker</th>
<th>oftic</th>
<th>cfname</th>
<th>statpers</th>
<th>measure</th>
<th>fpi</th>
<th>curcode</th>
<th>numest</th>
<th>numup</th>
<th>numdown</th>
<th>meanest</th>
<th>stdev</th>
<th>highest</th>
<th>lowest</th>
<th>epedats</th>
</tr>
</thead>
<tbody>
<tr>
<td>@FT2</td>
<td>FTC</td>
<td>PORTUGAL TELECOM</td>
<td>15jan2012</td>
<td>EPS</td>
<td>1</td>
<td>EUR</td>
<td>28</td>
<td>2</td>
<td>3</td>
<td>.55</td>
<td>.1</td>
<td>.8</td>
<td>.41</td>
<td>31dec2011</td>
</tr>
<tr>
<td>2</td>
<td>@FT2</td>
<td>FTC</td>
<td>PORTUGAL TELECOM</td>
<td>16feb2012</td>
<td>EPS</td>
<td>1</td>
<td>EUR</td>
<td>28</td>
<td>0</td>
<td>4</td>
<td>.54</td>
<td>.1</td>
<td>.8</td>
<td>.41</td>
</tr>
<tr>
<td>3</td>
<td>@FT2</td>
<td>FTC</td>
<td>PORTUGAL TELECOM</td>
<td>18mar2012</td>
<td>EPS</td>
<td>1</td>
<td>EUR</td>
<td>28</td>
<td>2</td>
<td>5</td>
<td>.92</td>
<td>.09</td>
<td>.79</td>
<td>.4</td>
</tr>
<tr>
<td>4</td>
<td>@FT2</td>
<td>FTC</td>
<td>PORTUGAL TELECOM</td>
<td>15apr2012</td>
<td>EPS</td>
<td>1</td>
<td>EUR</td>
<td>29</td>
<td>1</td>
<td>13</td>
<td>.47</td>
<td>.11</td>
<td>.81</td>
<td>.31</td>
</tr>
<tr>
<td>5</td>
<td>@FT2</td>
<td>FTC</td>
<td>PORTUGAL TELECOM</td>
<td>17may2012</td>
<td>EPS</td>
<td>1</td>
<td>EUR</td>
<td>29</td>
<td>0</td>
<td>4</td>
<td>.40</td>
<td>.12</td>
<td>.81</td>
<td>.29</td>
</tr>
<tr>
<td>6</td>
<td>@FT2</td>
<td>FTC</td>
<td>PORTUGAL TELECOM</td>
<td>14jun2012</td>
<td>EPS</td>
<td>1</td>
<td>EUR</td>
<td>29</td>
<td>3</td>
<td>4</td>
<td>.45</td>
<td>.12</td>
<td>.81</td>
<td>.27</td>
</tr>
<tr>
<td>7</td>
<td>@FT2</td>
<td>FTC</td>
<td>PORTUGAL TELECOM</td>
<td>19jul2012</td>
<td>EPS</td>
<td>1</td>
<td>EUR</td>
<td>29</td>
<td>1</td>
<td>13</td>
<td>.4</td>
<td>.13</td>
<td>.81</td>
<td>.34</td>
</tr>
<tr>
<td>8</td>
<td>@FT2</td>
<td>FTC</td>
<td>PORTUGAL TELECOM</td>
<td>16aug2012</td>
<td>EPS</td>
<td>1</td>
<td>EUR</td>
<td>29</td>
<td>2</td>
<td>5</td>
<td>.90</td>
<td>.12</td>
<td>.81</td>
<td>.26</td>
</tr>
<tr>
<td>9</td>
<td>@FT2</td>
<td>FTC</td>
<td>PORTUGAL TELECOM</td>
<td>20sep2012</td>
<td>EPS</td>
<td>1</td>
<td>EUR</td>
<td>29</td>
<td>1</td>
<td>3</td>
<td>.36</td>
<td>.12</td>
<td>.81</td>
<td>.25</td>
</tr>
<tr>
<td>10</td>
<td>@FT3</td>
<td>FTC</td>
<td>PORTUGAL TELECOM</td>
<td>13oct2012</td>
<td>EPS</td>
<td>1</td>
<td>EUR</td>
<td>28</td>
<td>1</td>
<td>3</td>
<td>.36</td>
<td>.09</td>
<td>.6</td>
<td>.28</td>
</tr>
<tr>
<td>11</td>
<td>@FT2</td>
<td>FTC</td>
<td>PORTUGAL TELECOM</td>
<td>13nov2012</td>
<td>EPS</td>
<td>1</td>
<td>EUR</td>
<td>27</td>
<td>3</td>
<td>5</td>
<td>.34</td>
<td>.07</td>
<td>.52</td>
<td>.25</td>
</tr>
<tr>
<td>12</td>
<td>@FT2</td>
<td>FTC</td>
<td>PORTUGAL TELECOM</td>
<td>20dec2012</td>
<td>EPS</td>
<td>1</td>
<td>EUR</td>
<td>27</td>
<td>2</td>
<td>4</td>
<td>.34</td>
<td>.06</td>
<td>.52</td>
<td>.27</td>
</tr>
<tr>
<td>13</td>
<td>@FT3</td>
<td>FTC</td>
<td>PORTUGAL TELECOM</td>
<td>17jan2013</td>
<td>EPS</td>
<td>1</td>
<td>EUR</td>
<td>26</td>
<td>0</td>
<td>1</td>
<td>.63</td>
<td>.06</td>
<td>.52</td>
<td>.26</td>
</tr>
<tr>
<td>14</td>
<td>@FT2</td>
<td>FTC</td>
<td>PORTUGAL TELECOM</td>
<td>14feb2013</td>
<td>EPS</td>
<td>1</td>
<td>EUR</td>
<td>27</td>
<td>1</td>
<td>7</td>
<td>.32</td>
<td>.06</td>
<td>.5</td>
<td>.24</td>
</tr>
<tr>
<td>15</td>
<td>@FT2</td>
<td>FTC</td>
<td>PORTUGAL TELECOM</td>
<td>14mar2013</td>
<td>EPS</td>
<td>1</td>
<td>EUR</td>
<td>29</td>
<td>2</td>
<td>3</td>
<td>.34</td>
<td>.1</td>
<td>.62</td>
<td>.16</td>
</tr>
<tr>
<td>16</td>
<td>@FT2</td>
<td>FTC</td>
<td>PORTUGAL TELECOM</td>
<td>13apr2013</td>
<td>EPS</td>
<td>1</td>
<td>EUR</td>
<td>38</td>
<td>4</td>
<td>4</td>
<td>.34</td>
<td>.09</td>
<td>.55</td>
<td>.19</td>
</tr>
<tr>
<td>17</td>
<td>@FT2</td>
<td>FTC</td>
<td>PORTUGAL TELECOM</td>
<td>16may2013</td>
<td>EPS</td>
<td>1</td>
<td>EUR</td>
<td>28</td>
<td>2</td>
<td>2</td>
<td>.34</td>
<td>.05</td>
<td>.55</td>
<td>.18</td>
</tr>
<tr>
<td>18</td>
<td>@FT2</td>
<td>FTC</td>
<td>PORTUGAL TELECOM</td>
<td>20jun2013</td>
<td>EPS</td>
<td>1</td>
<td>EUR</td>
<td>27</td>
<td>1</td>
<td>10</td>
<td>.3</td>
<td>.09</td>
<td>.55</td>
<td>.10</td>
</tr>
<tr>
<td>19</td>
<td>@FT2</td>
<td>FTC</td>
<td>PORTUGAL TELECOM</td>
<td>18jul2013</td>
<td>EPS</td>
<td>1</td>
<td>EUR</td>
<td>19</td>
<td>4</td>
<td>5</td>
<td>.29</td>
<td>.1</td>
<td>.55</td>
<td>.15</td>
</tr>
<tr>
<td>20</td>
<td>@FT3</td>
<td>FTC</td>
<td>PORTUGAL TELECOM</td>
<td>15aug2013</td>
<td>EPS</td>
<td>1</td>
<td>EUR</td>
<td>19</td>
<td>2</td>
<td>5</td>
<td>.28</td>
<td>.1</td>
<td>.49</td>
<td>.01</td>
</tr>
<tr>
<td>21</td>
<td>@FT2</td>
<td>FTC</td>
<td>PORTUGAL TELECOM</td>
<td>19sep2013</td>
<td>EPS</td>
<td>1</td>
<td>EUR</td>
<td>28</td>
<td>6</td>
<td>4</td>
<td>.3</td>
<td>.12</td>
<td>.49</td>
<td>.01</td>
</tr>
<tr>
<td>22</td>
<td>@FT2</td>
<td>FTC</td>
<td>PORTUGAL TELECOM</td>
<td>17oct2013</td>
<td>EPS</td>
<td>1</td>
<td>EUR</td>
<td>23</td>
<td>4</td>
<td>6</td>
<td>.31</td>
<td>.14</td>
<td>.49</td>
<td>.02</td>
</tr>
<tr>
<td>23</td>
<td>@FT3</td>
<td>FTC</td>
<td>PORTUGAL TELECOM</td>
<td>14nov2013</td>
<td>EPS</td>
<td>1</td>
<td>EUR</td>
<td>33</td>
<td>2</td>
<td>4</td>
<td>.31</td>
<td>.14</td>
<td>.53</td>
<td>.05</td>
</tr>
<tr>
<td>24</td>
<td>@FT2</td>
<td>FTC</td>
<td>PORTUGAL TELECOM</td>
<td>18dec2013</td>
<td>EPS</td>
<td>1</td>
<td>EUR</td>
<td>22</td>
<td>2</td>
<td>2</td>
<td>.3</td>
<td>.14</td>
<td>.53</td>
<td>.05</td>
</tr>
</tbody>
</table>
Contents coverage

➢ Executive compensation

• Compustat Execucomp – Executive Compensation

• Boardex [Not subscribed]
✓ **Compustat Execucomp – Executive Compensation**

- Executive compensation data (DEF14A SEC form)
- Contains over 2872 companies, both active and inactive
- Covers the S&P 1500 since 1994, top 5 earners (e.g., CEO, CFO) in each company
- Date field is based on Fiscal Year
- Units: Thousands of Dollars
Compustat Execucomp – Executive Compensation

Executive Compensation of Facebook in 2018

Step 1: Choose your date range.
Date range
2018 to 2018

Step 2: Apply your company codes.
- Ticker: FB

Selected
- Company Name
- EXEC_FULLNAME
- YEAR - Fiscal Year
- TOTAL_SEC - Total Compensation - As Reported in SEC Filings
- TDC2 - Total Compensation (Salary + Bonus + Other An...
Web Query in detail (Execucomp)

Compustat Execucomp – Executive Compensation

➢ Executive Compensation of Facebook in 2018

https://www.sec.gov/Archives/edgar/data/1326801/000132680119000025/facebook2019definitiveprox.htm

<table>
<thead>
<tr>
<th>execu_fullname</th>
<th>co_per_role</th>
<th>coname</th>
<th>total_sec</th>
<th>tds2</th>
<th>gvkey</th>
<th>year</th>
<th>ticker</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mark Elliot Zuckerberg</td>
<td>46275</td>
<td>FACEBOOK INC</td>
<td>22554,543</td>
<td>22554,543</td>
<td>170617</td>
<td>2018</td>
<td>FB</td>
</tr>
<tr>
<td>Michael T. Schreepfer</td>
<td>49276</td>
<td>FACEBOOK INC</td>
<td>19757,363</td>
<td>158767,482</td>
<td>170617</td>
<td>2018</td>
<td>FB</td>
</tr>
<tr>
<td>Sheryl Kara Sandberg</td>
<td>49278</td>
<td>FACEBOOK INC</td>
<td>29728,418</td>
<td>196746,541</td>
<td>170617</td>
<td>2018</td>
<td>FB</td>
</tr>
<tr>
<td>David M. Wexner</td>
<td>50979</td>
<td>FACEBOOK INC</td>
<td>19686,113</td>
<td>19686,113</td>
<td>170617</td>
<td>2018</td>
<td>FB</td>
</tr>
<tr>
<td>Christopher K. Cox</td>
<td>50981</td>
<td>FACEBOOK INC</td>
<td>19686,113</td>
<td>19686,113</td>
<td>170617</td>
<td>2018</td>
<td>FB</td>
</tr>
</tbody>
</table>

2018 Summary Compensation Table

The following table presents summary information regarding the total compensation awarded to, earned by, or paid to each of the named executive officers for services rendered to us for the years ended December 31, 2018, 2017, and 2016.

<table>
<thead>
<tr>
<th>Name and Principal Position</th>
<th>Fiscal Year</th>
<th>Salary ($)</th>
<th>Bonus ($)</th>
<th>Stock Awards ($)</th>
<th>All Other Compensation ($)</th>
<th>Total ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mark Zuckerberg</td>
<td>2018</td>
<td>1</td>
<td>—</td>
<td>—</td>
<td>22,554,542</td>
<td>22,554,543</td>
</tr>
<tr>
<td></td>
<td>2017</td>
<td>1</td>
<td>—</td>
<td>—</td>
<td>9,101,965</td>
<td>9,101,966</td>
</tr>
<tr>
<td></td>
<td>2016</td>
<td>1</td>
<td>—</td>
<td>—</td>
<td>6,014,212</td>
<td>6,014,212</td>
</tr>
<tr>
<td>Sheryl K. Sandberg</td>
<td>2018</td>
<td>843,077</td>
<td>638,310</td>
<td>—</td>
<td>18,423,523</td>
<td>23,728,418</td>
</tr>
<tr>
<td></td>
<td>2017</td>
<td>795,769</td>
<td>640,578</td>
<td>—</td>
<td>21,072,431</td>
<td>25,196,211</td>
</tr>
<tr>
<td></td>
<td>2016</td>
<td>738,077</td>
<td>1,209,451</td>
<td>—</td>
<td>16,008,428</td>
<td>19,556,357</td>
</tr>
<tr>
<td>David M. Wexner</td>
<td>2018</td>
<td>753,846</td>
<td>499,494</td>
<td>18,423,523</td>
<td>9,250</td>
<td>19,688,113</td>
</tr>
<tr>
<td></td>
<td>2017</td>
<td>711,539</td>
<td>633,917</td>
<td>21,072,431</td>
<td>9,000</td>
<td>22,426,287</td>
</tr>
<tr>
<td></td>
<td>2016</td>
<td>665,846</td>
<td>613,309</td>
<td>14,031,304</td>
<td>9,576</td>
<td>18,235,375</td>
</tr>
<tr>
<td>Christopher K. Cox</td>
<td>2018</td>
<td>753,846</td>
<td>499,494</td>
<td>18,423,523</td>
<td>9,250</td>
<td>19,688,113</td>
</tr>
<tr>
<td>Former CFO</td>
<td>2017</td>
<td>711,539</td>
<td>567,404</td>
<td>21,072,431</td>
<td>9,000</td>
<td>22,360,374</td>
</tr>
<tr>
<td></td>
<td>2016</td>
<td>658,846</td>
<td>613,309</td>
<td>14,031,304</td>
<td>9,518</td>
<td>16,531,180</td>
</tr>
<tr>
<td>Mike Schreepfer</td>
<td>2018</td>
<td>753,846</td>
<td>570,744</td>
<td>18,423,523</td>
<td>9,250</td>
<td>19,757,363</td>
</tr>
<tr>
<td>CTO</td>
<td>2017</td>
<td>711,539</td>
<td>633,917</td>
<td>21,072,431</td>
<td>9,000</td>
<td>22,426,287</td>
</tr>
<tr>
<td></td>
<td>2016</td>
<td>658,846</td>
<td>859,356</td>
<td>14,931,596</td>
<td>9,377</td>
<td>16,459,175</td>
</tr>
</tbody>
</table>
Compensation Data (BoardEx)

Are US CEOs Paid More? New International Evidence*

Nuno Fernandes, IMD International
Miguel A. Ferreira, Nova School of Business and Economics
Pedro Matos, University of Virginia, Darden School of Business
Kevin J. Murphy, University of Southern California, Marshall School of Business

(forthcoming in The Review of Financial Studies)

Abstract

This paper challenges the widely accepted stylized fact that CEOs in the United States are paid significantly more than their foreign counterparts. Using CEO pay data across 14 countries with mandated pay disclosures, we show that the US pay premium is economically modest and primarily reflects the performance-based pay demanded by institutional shareholders and independent boards. Indeed, we find no significant difference in either level of CEO pay or the use of equity-based pay between US and non-US firms exposed to international and US capital, product, and labor markets. We also show that US and non-US CEO pay has largely converged in the 2000s.

Appendix A: Variables definition and data sources

<table>
<thead>
<tr>
<th>Variable</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Pay</td>
<td>Total CEO compensation in US$ (US firms: ExecComp; non-US firms: BoardEx, corporate filings)</td>
</tr>
<tr>
<td>Salary</td>
<td>Salary in US$ (ExecComp: salary; BoardEx: salary)</td>
</tr>
<tr>
<td>Other pay</td>
<td>Other compensation in US$ (ExecComp: other compensation; BoardEx: other pay and pensions)</td>
</tr>
<tr>
<td>Bonuses</td>
<td>Non-equity incentive-plan compensation in US$ (ExecComp: bonus plus target value of non-equity incentive compensation; BoardEx: bonus)</td>
</tr>
<tr>
<td>Equity pay (stock and options)</td>
<td>Stock and options awards in US$ (ExecComp: grant-date fair value of stock awards plus grant-date fair value of option awards; BoardEx: market value of shares plus long-term incentive plans plus Black-Scholes option value)</td>
</tr>
</tbody>
</table>

The disclosure situation has improved over the past decade. Regulations mandating disclosure of executive pay were introduced in Ireland and South Africa in 2000 and in Australia in 2004. In May 2003, the European Union (EU) Commission issued an “Action Plan” recommending that all listed companies in the EU report details on individual compensation packages, and that EU member countries pass rules requiring such disclosure. By 2006, six EU members (in addition to the United Kingdom and Ireland) had mandated CEO-level disclosure: Belgium, France, Germany, Italy, Netherlands, and Sweden. In addition, although not in the EU, Norway also adopted EU-style disclosure rules, and Switzerland demanded similar disclosure for the “highest-paid” executive.

September 2012

Our primary data source on compensation for US CEOs is Standard and Poor’s (S&P’s) ExecComp database, while our primary source for CEOs of firms based outside the United States is BoardEx, compiled by the UK-based firm Management Diagnostics Limited. Together, these two sources (identified as “BoardEx & Exec” in Table 1) account for 2,899 of the 3,263 firms in our sample. BoardEx is the leading database on board composition of publicly listed firms, and includes detailed biographic information on individual executives and board members in nearly 50 countries, including countries that do not have mandatory disclosure requirements for executive compensation. In addition to providing biographic information, BoardEx also includes detailed compensation data for top executives – including salaries, other pay, bonuses, payouts under long-term plans, option grants, and share grants.
Contents coverage

➢ M&A and Equity/Debt underwriting

  • Thomson Reuters SDC
  • Thomson Reuters Eikon
Outside WRDS... (SDC)

✓ **SDC Global Issues – Thomson Reuters Eikon**

- **Mergers and Acquisitions**
  - more than 500,000 transactions globally since 1980

- **Equity**
  - more than 100,000 transactions (IPO/SEO) globally since 1970

- **Bonds**
  - more than 300,000 transactions globally since 1970

- **Syndicated Loans**
  - more than 120,000 transactions globally since 1982
Thomson Reuters SDC – WRDS

- Mergers and Acquisitions
- New Issues

Duplicated records: One Deal_ID with 2 target companies

Duplicated records: One Deal_ID with several records (legal - legal advisor name, datefin – date of financial statement, ...)

---

Outside WRDS... (SDC)
SDC Global Issues – Thomson Reuters Eikon

How to use TR Eikon’s Deal Analytics (User Guide)

1 – Open Excel and Select the “Thomson Reuters” Add-in Tab. Click on offline/login and insert login credentials

2 – Click on “Deal Screener”
SDC Global Issues – Thomson Reuters Eikon

How to use TR Eikon’s Deal Analytics

3 – The “Set Up Screen” window will pop up. Select the deal type in the “Universe” ListBox (M&A Deals, Equity Deals, Bond Deals, Loan Deals).

4 – Pick “Equity Deals” as an example. The extraction process of data from TR Eikon Deal Analytics is a bit cumbersome as there is a download limit. So the data needs to be subdivided before exporting to excel. Subset the data either by setting up “Popular Filters” which are fields already available in the screener or by inserting other available fields in the screener.
SDC Global Issues – Thomson Reuters Eikon

How to use TR Eikon’s Deal Analytics

4.1 – “Popular filters”
Outside WRDS... (SDC)

SDC Global Issues – Thomson Reuters Eikon

How to use TR Eikon’s Deal Analytics

4.2 – Insert other available Fields
How to use TR Eikon’s Deal Analytics

5 – Example: Set-Up Screen for Convertibles.

5.1 – Select “Equity Deals” in the “Universe” ListBox

5.2 – Click on the magnifier next to “Issue Type”

5.3 – Select “Convertible (CVT)” from the “Issue Type” ListBox. Click OK.
How to use TR Eikon’s Deal Analytics

5.4 – Select “United States” from the “Issuer Nation/Region” TextBox

5.5 – Check the record count (number of records in the query to the Deal Analytics dataset)

5.6 – Make sure the number of records doesn’t exceed the 3’000 download limit
SDC Global Issues – Thomson Reuters Eikon

How to use TR Eikon’s Deal Analytics

5.7 – A way to comply with the download limit is to subset data using date intervals using the field “Issue Date”

5.8 – Check the record count until it’s below 3’000

5.9 – Run Screen

5.10 – Click “Build Report” and “Add report items” to add more fields to the screener
How to use TR Eikon’s Deal Analytics

5.11 – Pick the fields you wish to include in the output. Select e.g. “Offer Currency”; “ISIN”; “Security Type”; “Use of Proceeds”; “Gross Spread”; “Principal Amount”; “Maturity Date”; “Equity Deal Type”; “Issuer SDC CUSIP”; “Issuer Name”; “Convertible Issue Flag”; “Conversion Shares per Bond/Share”; “Conversion Price”; “Conversion Premium (%)”; “Conversion Notes”; “First Conversion Date”; “Last Conversion Date”.

5.12 – Click “Insert” when the query is ready to download
SDC Global Issues – Thomson Reuters Eikon

How to use TR Eikon’s Deal Analytics

5.13 – Data exported to Excel:

<table>
<thead>
<tr>
<th>Equity Deal Number</th>
<th>Issuer Name</th>
<th>Gross Proceeds Inc.</th>
<th>Issuer Status</th>
<th>Issuer Type</th>
<th>Issuer Nation</th>
<th>Issue Date</th>
<th>Convertible Issue Flag (Y/N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>24087757013</td>
<td>Punta Gorda Isles Inc</td>
<td>7.5</td>
<td>Convertible</td>
<td>United States of America</td>
<td>03/31/1987 Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1601826033</td>
<td>HealthVest</td>
<td>33.113</td>
<td>Convertible</td>
<td>United States of America</td>
<td>10/09/1987 Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1241482007</td>
<td>Dennisson Manufacturing Co.</td>
<td>37.099</td>
<td>Convertible</td>
<td>United States of America</td>
<td>02/01/1985 Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1196634097</td>
<td>Club Med Inc.</td>
<td>0.13</td>
<td>Convertible</td>
<td>United States of America</td>
<td>06/17/1992 Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1103498007</td>
<td>Community Psychiatric Centers</td>
<td>15</td>
<td>Convertible</td>
<td>United States of America</td>
<td>02/01/1980 Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1103473007</td>
<td>Oak Industries Inc</td>
<td>35</td>
<td>Convertible</td>
<td>United States of America</td>
<td>09/01/1980 Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1103465007</td>
<td>Digicon Finance NV</td>
<td>18</td>
<td>Convertible</td>
<td>United States of America</td>
<td>09/01/1980 Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1103419007</td>
<td>Community Psychiatric Centers</td>
<td>15</td>
<td>Convertible</td>
<td>United States of America</td>
<td>04/01/1981 Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1103401007</td>
<td>Massmutual Mortgage Reality</td>
<td>21</td>
<td>Convertible</td>
<td>United States of America</td>
<td>07/01/1981 Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1103397007</td>
<td>Blue Ridge Petroleum Corp.</td>
<td>25</td>
<td>Convertible</td>
<td>United States of America</td>
<td>07/01/1981 Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1103344007</td>
<td>AMI American Medical Intl Inc</td>
<td>25</td>
<td>Convertible</td>
<td>United States of America</td>
<td>04/30/1982 Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1103339007</td>
<td>Southern California Edison Co.</td>
<td>50</td>
<td>Convertible</td>
<td>United States of America</td>
<td>07/01/1982 Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1103314007</td>
<td>Duncan Larrie Security Holding</td>
<td>15</td>
<td>Convertible</td>
<td>United States of America</td>
<td>04/29/1983 Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1103252007</td>
<td>Viacom International Inc</td>
<td>50</td>
<td>Convertible</td>
<td>United States of America</td>
<td>10/01/1984 Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1103235007</td>
<td>National Patent Development Corp.</td>
<td>50</td>
<td>Convertible</td>
<td>United States of America</td>
<td>01/01/1985 Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1103225007</td>
<td>Wang Laboratories Inc</td>
<td>74.198</td>
<td>Convertible</td>
<td>United States of America</td>
<td>02/01/1985 Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1103195007</td>
<td>St Paul Cos Inc.</td>
<td>100</td>
<td>Convertible</td>
<td>United States of America</td>
<td>03/01/1985 Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1103191007</td>
<td>Newmont Mining Corp.</td>
<td>60</td>
<td>Convertible</td>
<td>United States of America</td>
<td>03/01/1985 Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1103190007</td>
<td>National Patent Development Corp.</td>
<td>17.495</td>
<td>Convertible</td>
<td>United States of America</td>
<td>03/01/1985 Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1103188007</td>
<td>Louisiana Land &amp; Exploration Co.</td>
<td>50</td>
<td>Convertible</td>
<td>United States of America</td>
<td>03/01/1985 Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1103187007</td>
<td>Louisiana Land &amp; Exploration Co.</td>
<td>50</td>
<td>Convertible</td>
<td>United States of America</td>
<td>03/01/1985 Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1103154007</td>
<td>American General Corp.</td>
<td>300</td>
<td>Convertible</td>
<td>United States of America</td>
<td>05/01/1985 Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1103153007</td>
<td>Viacom International Inc</td>
<td>50</td>
<td>Convertible</td>
<td>United States of America</td>
<td>05/11/1985 Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1103149007</td>
<td>Pan American Corporation</td>
<td>0.767</td>
<td>Convertible</td>
<td>United States of America</td>
<td>05/31/1985 Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1103141007</td>
<td>ICN Pharmaceuticals Inc</td>
<td>19.227</td>
<td>Convertible</td>
<td>United States of America</td>
<td>03/31/1985 Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1103136007</td>
<td>Coleco Inc</td>
<td>0.614</td>
<td>Convertible</td>
<td>United States of America</td>
<td>05/31/1985 Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1103135007</td>
<td>Continental Health Affiliates Inc.</td>
<td>13.425</td>
<td>Convertible</td>
<td>United States of America</td>
<td>05/31/1985 Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1103131007</td>
<td>Limited Inc</td>
<td>50</td>
<td>Convertible</td>
<td>United States of America</td>
<td>07/01/1985 Y</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5.14 – Use the Excel Formula to change date ranges thus avoiding using the set up screener window

```
=IFS(TF.Screen(Deal,DealExpFmt(Deal,Active,1,IssuerStatus=Equity),curr=USD,(IN(TF.EQ.IssueType(),"Convertible")=1),(IN(TF.EQ.IssuerNation(),"United States of America")=1),(TF.EQ.IssueDate()>="01/01/1980" And TF.EQ.IssueDate()<="12/31/1992"))
```

'IFS(TF.Screen(Deal,DealExpFmt(Deal,Active,1,IssuerStatus=Equity),curr=USD,(IN(TF.EQ.IssueType(),"Convertible")=1),(IN(TF.EQ.IssuerNation(),"United States of America")=1),(TF.EQ.IssueDate()>="01/01/1980" And TF.EQ.IssueDate()<="12/31/1992"))
Contents coverage

➢ **Stocks**

  • CRSP

  • Compustat Supplemental Short Interest File

➢ **Database Management (Merging Datasets)**

  • CRSP-Compustat Merged (CCM)

  • Merging CRSP and Compustat by CUSIP

  • Fuzzy Matching
CRSP – The Center for Research in Security Prices

- Comprehensive collection of daily and monthly security price, return, and volume data for the NYSE, AMEX and NASDAQ stock markets.

- CRSP U.S. Stock databases do not include securities for international companies unless they are ADRs, cross-listed, or traded on the major stock exchanges mentioned above.

- Daily and Monthly data for roughly 28,000 securities traded on major exchanges from 1925-present.

- Complete historical information (bias-free).

- Accurate total returns calculation.
Information in CRSP:

- Price and quote data.
- Holding period returns with and without dividends.
- Shares outstanding.
- Trading volume.
- Security delisting information.
CRSP - Identifiers

- PERMNO: Unique and permanent issue identification number
- PERMCO: Permanent company identification number
- While PERMCO applies to only one company more than one PERMNO (securities) can be associated with one PERMCO (company)

Common Identifiers:
- CUSIP (latest eight-character CUSIP identifier)
- Ticker
CRSP (Web Query)

Web Query in detail (CRSP)

CRSP (Web Query)

Step 3: Query Variables.
How does this work?

Selected

Selected

Search All

Identifying Information

Time Series Information

Share Information

Select

Search All

Share Class

Nasdaq Issue Number

Exchange Code

Header Exchange Code

SIC Code

Cusip

Company Name

CRSP Permanent Company Number

Share Code

Ticker

Ncusip
CRSP (Web Query)

**Price (PRC)** is the closing price or the negative bid/ask average for a trading day. If the number in the price field has a negative sign it is a bid/ask average and not an actual closing price.

**Share Volume (VOL)** is the sum of the trading volumes during that period. It is expressed in units of one share, for daily data, and on hundred shares for monthly data. Volume is set to -99 if the value is missing.

**Holding Period Return (RET)** is the change in the total value of an investment in a common stock over some period of time per dollar of initial investment. It’s already adjusted to split events and includes dividends.

**Nbr. of Shares Outstanding (SHROUT)** is the number of publicly held shares recorded in thousands.
CRSP (Web Query)

- Factors to adjust price and shares outstanding for split events (stock splits, stock dividends, spin-offs, stock distributions, and rights)
  - Adjusted Price = \( \frac{PRC}{CFACPR} \)
  - Adjusted Shares Outstanding = \( SHROUT \times CFACSHR \)
- Holding period returns are already adjusted!

For most distribution cases (excl. e.g. DISTCD 6225):

\[
r(t) = \left( \frac{|PRC(t)| + d(t)}{CFACPR(t)} \right) / \left( \frac{|PRC(t')|}{CFACPR(t')} \right) - 1
\]

### Table: RET(t) Reason For Missing Return

<table>
<thead>
<tr>
<th>RET(t)</th>
<th>Reason For Missing Return</th>
</tr>
</thead>
<tbody>
<tr>
<td>-66.0</td>
<td>more than 10 periods between time ( t ) and the time of the preceding price ( t )?</td>
</tr>
<tr>
<td>-77.0</td>
<td>not trading on the current exchange at time ( t )</td>
</tr>
<tr>
<td>-88.0</td>
<td>no return, array index ( t ) not within range of ( \text{BEGRET and ENDRET} )</td>
</tr>
<tr>
<td>-99.0</td>
<td>missing return due to missing price at time ( t )</td>
</tr>
</tbody>
</table>

\[
r(KO, 1990-03-09) = \frac{(72+0.4)/16}{72.375/16} - 1 = 0.000345
\]
**CRSP – Share Code**

<table>
<thead>
<tr>
<th>First Digit</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Ordinary Common Shares</td>
</tr>
<tr>
<td>2</td>
<td>Certificates</td>
</tr>
<tr>
<td>3</td>
<td>ADRs (American Depository Receipts)</td>
</tr>
<tr>
<td>4</td>
<td>SBIs (Shares of Beneficial Interest)</td>
</tr>
<tr>
<td>7</td>
<td>Units (Depository Units, Units of Beneficial Interest, Units of Limited Partnership Interest, Depository Receipts, etc.)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Second Digit</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Securities which have not been further defined.</td>
</tr>
<tr>
<td>1</td>
<td>Securities which need not be further defined.</td>
</tr>
<tr>
<td>2</td>
<td>Companies incorporated outside the US</td>
</tr>
<tr>
<td>3</td>
<td>Americus Trust Components (Primes and Scores).</td>
</tr>
<tr>
<td>4</td>
<td>Closed-end funds.</td>
</tr>
<tr>
<td>5</td>
<td>Closed-end fund companies incorporated outside the US</td>
</tr>
<tr>
<td>8</td>
<td>REIT's (Real Estate Investment Trusts).</td>
</tr>
</tbody>
</table>

Usually you want to include share codes 10 and 11 (Ordinary common shares incorporated in the US)
Web Query in detail (CRSP)

✓ Merging datasets in STATA

Stata:
- **edit**
- **copy paste**
- save the 3 files (e.g. ds1, ds2, ds3)

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Ticker</td>
<td>Year</td>
<td>Industry</td>
<td>Ticker</td>
<td>Year</td>
<td>Assets</td>
<td>Industry</td>
<td>IndustryName</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>AAPL</td>
<td>2015</td>
<td>A</td>
<td>AAPL</td>
<td>2015</td>
<td>80</td>
<td>A</td>
<td>Technology</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>AAPL</td>
<td>2016</td>
<td>A</td>
<td>AAPL</td>
<td>2016</td>
<td>90</td>
<td>B</td>
<td>Banking</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>JPM</td>
<td>2015</td>
<td>B</td>
<td>JPM</td>
<td>2015</td>
<td>200</td>
<td>C</td>
<td>Beverage</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>JPM</td>
<td>2016</td>
<td>B</td>
<td>JPM</td>
<td>2016</td>
<td>210</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>KO</td>
<td>2015</td>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>KO</td>
<td>2016</td>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

```
^#* STATA MERGE - EXAMPLE
^#* STATA MERGE - EXAMPLE
^cd "D:\FDIM - PhD 2020\data"

* example - merge 1:1
use ds1, clear
merge 1:1 ticker year using ds2
br

* example - merge m:1
use ds1, clear
merge m:1 industry using ds3
br

* example - merge 1:m
use ds3, clear
merge 1:m industry using ds1
sort ticker year
br
```
Merging datasets in STATA

merge 1:1 ticker year using "path\filename.dta"

<table>
<thead>
<tr>
<th>ticker</th>
<th>year</th>
<th>industry</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAPL</td>
<td>2015</td>
<td>A</td>
</tr>
<tr>
<td>AAPL</td>
<td>2016</td>
<td>A</td>
</tr>
<tr>
<td>JPM</td>
<td>2015</td>
<td>B</td>
</tr>
<tr>
<td>JPM</td>
<td>2016</td>
<td>B</td>
</tr>
<tr>
<td>KO</td>
<td>2015</td>
<td>C</td>
</tr>
<tr>
<td>KO</td>
<td>2016</td>
<td>C</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ticker</th>
<th>year</th>
<th>assets</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAPL</td>
<td>2015</td>
<td>80</td>
</tr>
<tr>
<td>AAPL</td>
<td>2016</td>
<td>90</td>
</tr>
<tr>
<td>JPM</td>
<td>2015</td>
<td>200</td>
</tr>
<tr>
<td>JPM</td>
<td>2016</td>
<td>210</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ticker</th>
<th>year</th>
<th>industry</th>
<th>assets</th>
<th>_merge</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAPL</td>
<td>2015</td>
<td>A</td>
<td>80</td>
<td>matched (3)</td>
</tr>
<tr>
<td>AAPL</td>
<td>2016</td>
<td>A</td>
<td>90</td>
<td>matched (3)</td>
</tr>
<tr>
<td>JPM</td>
<td>2015</td>
<td>B</td>
<td>200</td>
<td>matched (3)</td>
</tr>
<tr>
<td>JPM</td>
<td>2016</td>
<td>B</td>
<td>210</td>
<td>matched (3)</td>
</tr>
<tr>
<td>KO</td>
<td>2015</td>
<td>C</td>
<td>.</td>
<td>master only (1)</td>
</tr>
<tr>
<td>KO</td>
<td>2016</td>
<td>C</td>
<td>.</td>
<td>master only (1)</td>
</tr>
</tbody>
</table>
Merging datasets in STATA

merge m:1 industry using “path\filename.dta”

<table>
<thead>
<tr>
<th>ticker</th>
<th>year</th>
<th>industry</th>
<th>industryname</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAPL</td>
<td>2015</td>
<td>A</td>
<td>Technology</td>
</tr>
<tr>
<td>AAPL</td>
<td>2016</td>
<td>A</td>
<td>Technology</td>
</tr>
<tr>
<td>JPM</td>
<td>2015</td>
<td>B</td>
<td>Banking</td>
</tr>
<tr>
<td>JPM</td>
<td>2016</td>
<td>B</td>
<td>Banking</td>
</tr>
<tr>
<td>KO</td>
<td>2015</td>
<td>C</td>
<td>Beverage</td>
</tr>
<tr>
<td>KO</td>
<td>2016</td>
<td>C</td>
<td>Beverage</td>
</tr>
</tbody>
</table>

Web Query in detail (CRSP)
Merging datasets in STATA

```
merge 1:m industry using "path\filename.dta"
```

**ds3**

<table>
<thead>
<tr>
<th>industry</th>
<th>industryname</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A Technology</td>
</tr>
<tr>
<td>2</td>
<td>B Banking</td>
</tr>
<tr>
<td>3</td>
<td>C Beverage</td>
</tr>
</tbody>
</table>

**ds1**

<table>
<thead>
<tr>
<th>ticker</th>
<th>year</th>
<th>industry</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAPL</td>
<td>2015</td>
<td>A</td>
</tr>
<tr>
<td>AAPL</td>
<td>2016</td>
<td>A</td>
</tr>
<tr>
<td>JPM</td>
<td>2015</td>
<td>B</td>
</tr>
<tr>
<td>JPM</td>
<td>2016</td>
<td>B</td>
</tr>
<tr>
<td>KO</td>
<td>2015</td>
<td>C</td>
</tr>
<tr>
<td>KO</td>
<td>2016</td>
<td>C</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>industry</th>
<th>industryname</th>
<th>ticker</th>
<th>year</th>
<th>_merge</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Technology</td>
<td>AAPL</td>
<td>2015</td>
<td>matched (3)</td>
</tr>
<tr>
<td>A</td>
<td>Technology</td>
<td>AAPL</td>
<td>2016</td>
<td>matched (3)</td>
</tr>
<tr>
<td>B</td>
<td>Banking</td>
<td>JPM</td>
<td>2015</td>
<td>matched (3)</td>
</tr>
<tr>
<td>B</td>
<td>Banking</td>
<td>JPM</td>
<td>2016</td>
<td>matched (3)</td>
</tr>
<tr>
<td>C</td>
<td>Beverage</td>
<td>KO</td>
<td>2015</td>
<td>matched (3)</td>
</tr>
<tr>
<td>C</td>
<td>Beverage</td>
<td>KO</td>
<td>2016</td>
<td>matched (3)</td>
</tr>
</tbody>
</table>
Web Query in detail (CRSP)

✓ Merging datasets in STATA

### Syntax

- **One-to-one merge on specified key variables**
  
  ```stata
  merge 1:1 varlist using filename [, options]
  ```

- **Many-to-one merge on specified key variables**
  
  ```stata
  merge m:1 varlist using filename [, options]
  ```

- **One-to-many merge on specified key variables**
  
  ```stata
  merge 1:m varlist using filename [, options]
  ```

- **Many-to-many merge on specified key variables**
  
  ```stata
  merge m:m varlist using filename [, options]
  ```

### Description

`merge` joins corresponding observations from the dataset currently in memory (called the master dataset) with those from `filename.dta` (called the using dataset), matching on one or more key variables. `merge` can perform match merges (one-to-one, one-to-many, many-to-one, and many-to-many), which are often called `joins` by database people. `merge` can also perform sequential merges, which have no equivalent in the relational database world.

`merge` is for adding new variables from a second dataset to existing observations. You use `merge`, for instance, when combining hospital patient and discharge datasets. If you wish to add new observations to existing variables, then see `[D] append`. You use `append`, for instance, when adding current discharges to past discharges.

By default, `merge` creates a new variable, `_merge`, containing numeric codes concerning the source and the contents of each observation in the merged dataset. These codes are explained below in the match results table.

- **Key variables cannot be strls.**
- **If filename is specified without an extension, then .dta is assumed.**
Web Query in detail (CRSP)

The original data in memory are called the master data. The data in `filename.dta` are called the using data. After `merge`, the merged result is left in memory. The `id` variable is called the key variable. Stata jargon is that the datasets were merged on `id`.

Observations for `id==1` existed in both the master and using datasets and so were combined in the merged result. The same occurred for `id==2`. For `id==5` and `id==4`, however, no matches were found and thus each became a separate observation in the merged result. Thus each observation in the merged result came from one of three possible sources:

### m:1 merges

In an m:1 merge, the key variable or variables uniquely identify the observations in the using data, but not necessarily in the master data. Suppose you had person-level data within regions and you wished to bring in regional data. Here is an example:

```
   . merge m:1 region using filename

in memory + in filename.dta = merged result

  id  region  a    region  x    id  region  a  x  _merge
  1   2  26    1   15    1   2  26  13   3
  2   1  29    2   13    2   1  29  15   3
  3   2  22    3   12    3   2  22  13   3
  4   3  21    4   11    4   3  21  12   3
  5   1  24    5   14    5   1  24  15   3
  6   5  20    6   10    6   5  20  3    3

To bring in the regional information, we need to merge on `region`. The values of `region` identify individual observations in the using data, but it is not an identifier in the master data.

We show the merged dataset sorted by `id` because this makes it easier to see how the merged dataset was constructed. For each observation in the master data, `merge` finds the corresponding observation in the using data. `merge` combines the values of the variables in the using dataset to the observations in the master dataset.
Basic data management in STATA (Merge CRSP - Compustat)

✓ **CRSP-Compustat Merged (CCM)**

➢ **Easiest way to merge CRSP and Compustat**

➢ **Use either CRSP or Compustat identifiers, (i.e. PERMNO|PERMCO; GVKEY)**

➢ **The match between CRSP and COMPUSTAT data is highly accurate, but it is not complete or entirely unambiguous**

   ➢ CRSP database covers stock prices on public stock exchanges while COMPUSTAT does not require a company to have a traded stock. There are cases where a company that is covered by COMPUSTAT has valid stock price data for only part of its history.

   ➢ Mergers: sometimes there is disagreement between CRSP and COMPUSTAT over which is the surviving company.

   ➢ The match between COMPUSTAT’s GVKEY and CRSP’s PERMNO is not one-to-one. For example, a company might have multiple equity issues.

   ➢ Different calendar/fiscal regime.

   ➢ Different frequencies: Compustat data is either annual or quarterly while CRSP data is either monthly or daily.
CRSP-Compustat Merged (CCM)

- **Primary link types - 41%**
  - LC: Link research complete.
  - LU: Link is unresearched by CRSP and is established by comparing CUSIPs
  - LS: these links mainly relate to ETFs

- **Secondary link types – 2%**
  - LX, LD, LN

- **Non-matching link types – 57%**
  - NR: No link available; confirmed by research
  - NU: No link available; not yet confirmed
Basic data management in STATA (Merge CRSP - Compustat)

CRSP-Compustat Merged (CCM)

- CCM Link structure

<table>
<thead>
<tr>
<th>MNEMONIC</th>
<th>FIELD NAME</th>
</tr>
</thead>
<tbody>
<tr>
<td>LPERMNO</td>
<td>CRSP PERMNO link during link period. It is set to zero if there is no CRSP link during the range.</td>
</tr>
<tr>
<td>LPERMCO</td>
<td>CRSP PERMCO link during link period. It is set to zero if there is no CRSP link during the range.</td>
</tr>
<tr>
<td>LIID</td>
<td>Security identifier</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LINKPRIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>P = Primary, identified by Compustat in monthly security data.</td>
</tr>
<tr>
<td>J = Joiner secondary issue of a company, identified by Compustat in monthly security data.</td>
</tr>
<tr>
<td>C = Primary, assigned by CRSP to resolve ranges of overlapping or missing primary markers from Compustat in order to produce one primary security throughout the company history.</td>
</tr>
<tr>
<td>N = Secondary, assigned by CRSP to override Compustat. Compustat allows a US and Canadian security to both be marked as Primary at the same time. For purposes of the link, CRSP allows only one primary at a time and marks the others as N.</td>
</tr>
</tbody>
</table>
Basic data management in STATA (Merge CRSP - Compustat)

CCM – CRSP/Compustat Merged

Example: Get the annual sales and assets for Boeing (BA), Dell (DELL), Ebay (EBAY), and General Dynamics (GD) for the fiscal years 2010:2017.

Dataset: CCM

CRSP/Compustat Merged - Fundamentals Annual

Changes to CCM

As of the February 2014 release, USEDFLAG is no longer used in the WRDS CCM queries. Please select LINKTYPES LC, LU, and LS for the same results. These represent the vast majority of the links between CRSP securities and Compustat companies, without introducing duplicate data.

The WRDS-created linking dataset (ccmxpf_linktable) has been deprecated. It will continue to be created for a transition period of 1 year. SAS programmers should use the Link History dataset (ccmxpf_linkhist) from CRSP.

Please refer to the WRDS knowledgebase and our new overview of CCM for more information.

Step 1: Choose your date range.

Date Variable: Fiscal Year

Date range

2010-01 to 2017-12

Step 2: Apply your company codes.

Select an option for entering company codes

BA GD EBAY DELL

Code List Name
Basic data management in STATA (Merge CRSP - Compustat)

CCM – CRSP/Compustat Merged

**Example:** Get the annual sales and assets for Boeing (BA), Dell (DELL), Ebay (EBAY), and General Dynamics (GD) for the fiscal years 2010:2017.

**Dataset:** CCM

**Step 3: Linking Options**

```
use "C:\Users\Pedro Pires\Downloads\ccm_crsp_comp_example1.dta", clear
format comm %24s
b r g v key l i d l i n kprim l i l d l i n k t y p e 1 p e r m n o 1 p e r m c o d a t a d a t e f y e a r t i c c o m m a t s a l e
```
**Basic data management in STATA (Merge CRSP - Compustat)**

**CCM – CRSP/Compustat Merged**

**Example:** Get the annual sales and assets for Boeing (BA), Dell (DELL), Ebay (EBAY), and General Dynamics (GD) for the fiscal years 2010:2017.

**Dataset: CCM**

<table>
<thead>
<tr>
<th>qvkey</th>
<th>linkprim</th>
<th>lid</th>
<th>linktype</th>
<th>lpermno</th>
<th>lpermno</th>
<th>datekey</th>
<th>fyear</th>
<th>tic</th>
<th>omm</th>
<th>at</th>
<th>sale</th>
</tr>
</thead>
<tbody>
<tr>
<td>002285</td>
<td>P</td>
<td>01</td>
<td>LU</td>
<td>19561</td>
<td>20515</td>
<td>31dec2010</td>
<td>2010</td>
<td>BA</td>
<td>BOEING CO</td>
<td>68565</td>
<td>64306</td>
</tr>
<tr>
<td>002285</td>
<td>P</td>
<td>01</td>
<td>LU</td>
<td>19561</td>
<td>20315</td>
<td>31dec2011</td>
<td>2011</td>
<td>BA</td>
<td>BOEING CO</td>
<td>79996</td>
<td>66735</td>
</tr>
<tr>
<td>002285</td>
<td>P</td>
<td>01</td>
<td>LU</td>
<td>19561</td>
<td>20315</td>
<td>31dec2012</td>
<td>2012</td>
<td>BA</td>
<td>BOEING CO</td>
<td>88886</td>
<td>81688</td>
</tr>
<tr>
<td>002285</td>
<td>P</td>
<td>01</td>
<td>LU</td>
<td>19561</td>
<td>20315</td>
<td>31dec2013</td>
<td>2013</td>
<td>BA</td>
<td>BOEING CO</td>
<td>92663</td>
<td>86623</td>
</tr>
<tr>
<td>002285</td>
<td>P</td>
<td>01</td>
<td>LU</td>
<td>19561</td>
<td>20315</td>
<td>31dec2014</td>
<td>2014</td>
<td>BA</td>
<td>BOEING CO</td>
<td>99199</td>
<td>90762</td>
</tr>
<tr>
<td>002285</td>
<td>P</td>
<td>01</td>
<td>LU</td>
<td>19561</td>
<td>20315</td>
<td>31dec2015</td>
<td>2015</td>
<td>BA</td>
<td>BOEING CO</td>
<td>94408</td>
<td>96114</td>
</tr>
<tr>
<td>002285</td>
<td>P</td>
<td>01</td>
<td>LU</td>
<td>19561</td>
<td>20315</td>
<td>31dec2016</td>
<td>2016</td>
<td>BA</td>
<td>BOEING CO</td>
<td>89997</td>
<td>94571</td>
</tr>
<tr>
<td>002285</td>
<td>P</td>
<td>01</td>
<td>LU</td>
<td>19561</td>
<td>20315</td>
<td>31dec2017</td>
<td>2017</td>
<td>BA</td>
<td>BOEING CO</td>
<td>92338</td>
<td>93392</td>
</tr>
<tr>
<td>005046</td>
<td>P</td>
<td>01</td>
<td>LC</td>
<td>12052</td>
<td>20791</td>
<td>31dec2010</td>
<td>2010</td>
<td>GD</td>
<td>GENERAL DYNAMICS CORP</td>
<td>32545</td>
<td>32466</td>
</tr>
<tr>
<td>005046</td>
<td>P</td>
<td>01</td>
<td>LC</td>
<td>12052</td>
<td>20791</td>
<td>31dec2011</td>
<td>2011</td>
<td>GD</td>
<td>GENERAL DYNAMICS CORP</td>
<td>34882</td>
<td>32677</td>
</tr>
<tr>
<td>005046</td>
<td>P</td>
<td>01</td>
<td>LC</td>
<td>12052</td>
<td>20791</td>
<td>31dec2012</td>
<td>2012</td>
<td>GD</td>
<td>GENERAL DYNAMICS CORP</td>
<td>34309</td>
<td>31682</td>
</tr>
<tr>
<td>005046</td>
<td>P</td>
<td>01</td>
<td>LC</td>
<td>12052</td>
<td>20791</td>
<td>31dec2013</td>
<td>2013</td>
<td>GD</td>
<td>GENERAL DYNAMICS CORP</td>
<td>35448</td>
<td>31218</td>
</tr>
<tr>
<td>005046</td>
<td>P</td>
<td>01</td>
<td>LC</td>
<td>12052</td>
<td>20791</td>
<td>31dec2014</td>
<td>2014</td>
<td>GD</td>
<td>GENERAL DYNAMICS CORP</td>
<td>35355</td>
<td>30852</td>
</tr>
<tr>
<td>005046</td>
<td>P</td>
<td>01</td>
<td>LC</td>
<td>12052</td>
<td>20791</td>
<td>31dec2015</td>
<td>2015</td>
<td>GD</td>
<td>GENERAL DYNAMICS CORP</td>
<td>31997</td>
<td>31649</td>
</tr>
<tr>
<td>005046</td>
<td>P</td>
<td>01</td>
<td>LC</td>
<td>12052</td>
<td>20791</td>
<td>31dec2016</td>
<td>2016</td>
<td>GD</td>
<td>GENERAL DYNAMICS CORP</td>
<td>32872</td>
<td>31353</td>
</tr>
<tr>
<td>005046</td>
<td>P</td>
<td>01</td>
<td>LC</td>
<td>12052</td>
<td>20791</td>
<td>31dec2017</td>
<td>2017</td>
<td>GD</td>
<td>GENERAL DYNAMICS CORP</td>
<td>35046</td>
<td>30973</td>
</tr>
<tr>
<td>014489</td>
<td>P</td>
<td>01</td>
<td>LC</td>
<td>11081</td>
<td>9533</td>
<td>31jan2011</td>
<td>2010</td>
<td>DELL</td>
<td>DELL TECHNOLOGIES INC</td>
<td>30599</td>
<td>61494</td>
</tr>
<tr>
<td>014489</td>
<td>P</td>
<td>01</td>
<td>LC</td>
<td>11081</td>
<td>9533</td>
<td>31jan2012</td>
<td>2011</td>
<td>DELL</td>
<td>DELL TECHNOLOGIES INC</td>
<td>44553</td>
<td>62071</td>
</tr>
<tr>
<td>014489</td>
<td>P</td>
<td>01</td>
<td>LC</td>
<td>11081</td>
<td>9533</td>
<td>31jan2013</td>
<td>2012</td>
<td>DELL</td>
<td>DELL TECHNOLOGIES INC</td>
<td>47540</td>
<td>56940</td>
</tr>
<tr>
<td>114524</td>
<td>P</td>
<td>01</td>
<td>LU</td>
<td>86356</td>
<td>16285</td>
<td>31dec2010</td>
<td>2010</td>
<td>EBAY</td>
<td>EBAY INC</td>
<td>22003.762</td>
<td>9156.273999999999999</td>
</tr>
<tr>
<td>114524</td>
<td>P</td>
<td>01</td>
<td>LU</td>
<td>86356</td>
<td>16285</td>
<td>31dec2011</td>
<td>2011</td>
<td>EBAY</td>
<td>EBAY INC</td>
<td>27320.218</td>
<td>11651.654</td>
</tr>
<tr>
<td>114524</td>
<td>P</td>
<td>01</td>
<td>LU</td>
<td>86356</td>
<td>16285</td>
<td>31dec2012</td>
<td>2012</td>
<td>EBAY</td>
<td>EBAY INC</td>
<td>37074</td>
<td>14072</td>
</tr>
<tr>
<td>114524</td>
<td>P</td>
<td>01</td>
<td>LU</td>
<td>86356</td>
<td>16285</td>
<td>31dec2013</td>
<td>2013</td>
<td>EBAY</td>
<td>EBAY INC</td>
<td>41480</td>
<td>12047</td>
</tr>
<tr>
<td>114524</td>
<td>P</td>
<td>01</td>
<td>LU</td>
<td>86356</td>
<td>16285</td>
<td>31dec2014</td>
<td>2014</td>
<td>EBAY</td>
<td>EBAY INC</td>
<td>45132</td>
<td>17902</td>
</tr>
<tr>
<td>114524</td>
<td>P</td>
<td>01</td>
<td>LU</td>
<td>86356</td>
<td>16285</td>
<td>31dec2015</td>
<td>2015</td>
<td>EBAY</td>
<td>EBAY INC</td>
<td>17785</td>
<td>8532</td>
</tr>
<tr>
<td>114524</td>
<td>P</td>
<td>01</td>
<td>LU</td>
<td>86356</td>
<td>16285</td>
<td>31dec2016</td>
<td>2016</td>
<td>EBAY</td>
<td>EBAY INC</td>
<td>23847</td>
<td>6897</td>
</tr>
<tr>
<td>114524</td>
<td>P</td>
<td>01</td>
<td>LU</td>
<td>86356</td>
<td>16285</td>
<td>31dec2017</td>
<td>2017</td>
<td>EBAY</td>
<td>EBAY INC</td>
<td>25991</td>
<td>9567</td>
</tr>
</tbody>
</table>
✓ **Compustat Supplemental Short Interest File**

- Data on shorted stocks for listed companies at the New York Stock Exchange, American Stock Exchange, and NASDAQ.
- Short interest positions are compiled by the exchanges twice each month, at mid-month and end-month.

---

**Abstract**

Stocks are short-sale constrained when there is a strong demand to sell short and a limited supply of shares to borrow. Using data on both short interest (a proxy for demand) and institutional ownership (a proxy for supply) we find that constrained stocks underperform during the period 1988–2002 by a significant 215 basis points per month on an equally weighted basis, although by only an insignificant 39 basis points per month on a value-weighted basis. For the overwhelming majority of stocks, short interest and institutional ownership levels make short selling constraints unlikely.
✓ **Merging without CCM**

➢ Compustat and CRSP have different universes with slightly different coverage

<table>
<thead>
<tr>
<th></th>
<th>CRSP</th>
<th>COMPUSTAT</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Coverage</strong></td>
<td>Stock market data from major stock exchanges (NYSE, AMEX, NASDAQ)</td>
<td>Accounting data for public, OTC, and private companies</td>
</tr>
<tr>
<td><strong>Primary Identifiers</strong></td>
<td>PERMNO, PERMCO</td>
<td>GVKEY</td>
</tr>
<tr>
<td><strong>Secondary Identifiers</strong></td>
<td>Header &amp; Historical 8-digit CUSIP</td>
<td>Header 9-digit CUSIP</td>
</tr>
<tr>
<td><strong>Additional Identifiers</strong></td>
<td>Ticker Symbol, Company Name (Header &amp; Historical)</td>
<td>Ticker Symbol, Company Name (Header only)</td>
</tr>
<tr>
<td><strong>Inclusion Policy</strong></td>
<td>All data after IPO</td>
<td>A firm must have some number of years of history as public company before inclusion</td>
</tr>
<tr>
<td><strong>Historical Coverage</strong></td>
<td>1925–present</td>
<td>1950–present</td>
</tr>
</tbody>
</table>

➢ WRDS recommends starting with the SAS sample program `merge_funda_crsp_bycusip.sas` to merge CRSP and Compustat Xpressfeed databases by CUSIP.

https://wrds-www.wharton.upenn.edu/pages/support/research-wrds/macros/wrds-macro-merge_funda_crsp_bycusipsas/
Merging without CCM

➢ CCM vs. CUSIP-Matched CRSP and Compustat companies

SAS code provided in:

Figure I: CCM v.s. CUSIP Matched CRSP and Compustat Companies

Source: SAS code provided by WRDS
Merging without CCM

- **Compustat data is annual or quarterly, while CRSP data is monthly or daily.**

- **Different calendar/fiscal regime:** Compustat data needs to be converted from fiscal to calendar regime.

- **When merging both sets, it is important to ask the question, “When did the market know about the accounting information?”**
Understanding CUSIP’s

Example using CUSIP for Pfizer (PFE)

6-Digit Issuer Number

Issue Number

Check Digit

8-Digit CUSIP

7 1 7 0 8 1 10 3
Basic data management in STATA (Merge CRSP - Compustat)

Merge CRSP/Compustat by CUSIP

**Example:** Merge CRSP common stocks (Share Codes 10,11) of US pharmaceutical preparation companies (SIC 2834) with Compustat for the period 2012:2013.

**STEP 1 - Dataset: CRSP**

- **CRSP Monthly Stock**
  - **Step 1:** Choose your date range.
    - Date range: 2012-01 to 2013-12
  - **Step 2:** Apply your company codes.
    - Select an option for entering company codes:
      - Ticker
      - PERMNO
      - PERMCO
      - CUSIP
      - NCUSIP
      - CRSP Permanent Company Number
      - Share Code
      - Price
      - CRSP
      - CRSP Permanent Company Number
      - Number of Shares Outstanding

- Extract CRSP Monthly data for pharmaceutical preparation stocks (sic 2834, share code 10/11) for 2 yrs. ending in 2013.
Basic data management in STATA (Merge CRSP - Compustat)

Merge CRSP/Compustat by CUSIP

**Example:** Merge CRSP common stocks (Share Codes 10,11) of US pharmaceutical preparation companies (SIC 2834) with Compustat for the period 2012:2013.

**STEP2 - Dataset: Compustat Fundamentals Annual**

- **Query Form**
- **Variable Descriptions**
- **Map**

**Compustat Annual Updates - Further**

**Step 1: Choose your date range.**

- **Date Variable:** Fiscal Year
- **Date range:** From 2010-01 to 2013-12

**Screening Variables**

Several screening variables are pre-selected to produce one record per GVKEY-DATADATE pair, while keeping the vast majority of records. Examples of excluded rows include those with restated data, different views of the same data (pro forma, pre-FASB). Click on each variable for a more detailed explanation.

**Conditional Statements (Optional)**

**How does this work?**

- **Condition:** Foreign Incorporation Code
- **Value:** USA

**Data Request Summary**

Your output is complete. Click on the link below to open the output: 6574c9be81a0d47.dta (5.6 MB, 33068 observations 19 variables)
## Merge CRSP/Compustat by CUSIP

**Example:** Merge CRSP common stocks (Share Codes 10,11) of US pharmaceutical preparation companies (SIC 2834) with Compustat for the period 2012:2013.

### STEP3 – Merge CRSP and COMPUSTAT Datasets by CUSIP

```plaintext
clear all
set more off

cd "D:\Documents\FDM - PhD 2020\data"

* We are going to begin with the Compustat file
use ex1_crspcomp_merge_cusip_comp.dta, clear

* drop records with missing cusip - common identifier in CRSP & COMPSTAT
drop if cusip=="

* CRSP CUSIP only has 8 digits - Remove last digit from Compustat CUSIP
gen cusip8 = substr(cusip,1,8)

* Check for duplicates on the merging keys (cusip8-fyear)
duplicates report cusip8 fyear

* create new variable comp_fyear with a copy of fyear from Compustat
gen comp_fyear = fyear

* Save file
sa ex1_crspcomp_aux1.dta, replace
```
Basic data management in STATA (Merge CRSP - Compustat)

**Merge CRSP/Compustat by CUSIP**

**Example:** Merge CRSP common stocks (Share Codes 10,11) of US pharmaceutical preparation companies (SIC 2834) with Compustat for the period 2012:2013.

**STEP3 – Merge CRSP and COMPUSTAT Datasets by CUSIP**

To make sure accounting information is publicly available when a new portfolio is formatted, Fama and French (1993) chose the last trading date in June as the portfolio construction date, while using the Market Capitalization for the previous year end, to ensure both data availability and consistency of their empirical analysis.

The following figure demonstrates the logic of time dimension alignment in Fama and French (1993). A more detailed description of this strategy, as well as a reproduction of the methodology, is documented in Fama-French Factors.
Basic data management in STATA (Merge CRSP - Compustat)

**Merge CRSP/Compustat by CUSIP**

**Example:** Merge CRSP common stocks (Share Codes 10,11) of US pharmaceutical preparation companies (SIC 2834) with Compustat for the period 2012:2013.

**STEP3 – Merge CRSP and COMPUSTAT Datasets by CUSIP**

**Compustat fiscal years vs. calendar years**

Compustat annual files include a fiscal year-end month variable (fyr) that takes values form 1 (January) to 12 (December) depending on the month in which a company’s fiscal year ends.

- For firms with a fiscal year end between January and May, the fiscal year lags one year behind the calendar year
- For firms with a fiscal year end between June and December, the fiscal year is the same as the calendar year

<table>
<thead>
<tr>
<th>FYR</th>
<th>Fiscal year-end</th>
<th>Calendar year</th>
<th>Fiscal year</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>January 31</td>
<td>t</td>
<td>t-1</td>
</tr>
<tr>
<td>2</td>
<td>February 28</td>
<td>t</td>
<td>t-1</td>
</tr>
<tr>
<td>3</td>
<td>March 31</td>
<td>t</td>
<td>t-1</td>
</tr>
<tr>
<td>4</td>
<td>April 30</td>
<td>t</td>
<td>t-1</td>
</tr>
<tr>
<td>5</td>
<td>May 31</td>
<td>t</td>
<td>t-1</td>
</tr>
<tr>
<td>6</td>
<td>June 30</td>
<td>t</td>
<td>t</td>
</tr>
<tr>
<td>7</td>
<td>July 31</td>
<td>t</td>
<td>t</td>
</tr>
<tr>
<td>8</td>
<td>August 31</td>
<td>t</td>
<td>t</td>
</tr>
<tr>
<td>9</td>
<td>September 30</td>
<td>t</td>
<td>t</td>
</tr>
<tr>
<td>10</td>
<td>October 31</td>
<td>t</td>
<td>t</td>
</tr>
<tr>
<td>11</td>
<td>November 30</td>
<td>t</td>
<td>t</td>
</tr>
<tr>
<td>12</td>
<td>December 31</td>
<td>t</td>
<td>t</td>
</tr>
</tbody>
</table>
Basic data management in STATA (Merge CRSP - Compustat)

**Example (1):** Merge CRSP common stocks (Share Codes 10,11) of US pharmaceutical preparation companies (SIC 2834) with Compustat for the period 2012:2013.

**STEP3 – Merge CRSP and COMPUSTAT Datasets by CUSIP**

* Now, We are going to open the CRSP file
  use exl_crspcomp_merge_cusip_crsp.dta, clear

* drop recors with missing cusip - common identifier in CRSP & COMPUSTAT
  drop if cusip=="

* create a new variable cusip8 (copy of cusip) to facilitate the merge command
  gen cusip8 = cusip

* Check for duplicates
  duplicates report cusip8 date

* You need to think on how you want to merge the CRSP data with the Compustat data
* In CRSP you'll get daily or monthly data in a calendar regime
* In Compustat you'll get annual or quarterly data in a fiscal regime
* In compustat datadate identifies the calendar date of the financial report
* For illustrative purposes only, in this example we will merge CRSP market data
  on calendar months from July of year(t) to June of year(t+1)
* with Compustat accounting data in fiscal year (t-1)
* Accordingly, create a key variable with the time identifier needed to merge with Compustat
  gen fyear = year(date)-1
  replace fyear = year(date)-2 if month(date) <= 6
Basic data management in STATA (Merge CRSP - Compustat)

Example (1): Merge CRSP common stocks (Share Codes 10,11) of US pharmaceutical preparation companies (SIC 2834) with Compustat for the period 2012:2013.

STEP3 – Merge CRSP and COMPUSTAT Datasets by CUSIP

```
* Now, we can merge CRSP with Compustat on cusip8-fyear
* issue the merge command
merge m:1 cusip8 fyear using ex1_crspcomp_aux1.dta, keep(match master)
* drop observations without a match
drop if _merge!=3
format comnam %32s
format comm %32s
sort permno date
gen yearfrac = (date - datadate)/365
su yearfrac, d
br permno date cusip ticker comnam ret gvkey comp fyear datadate comm at
```

Syntax:

- One-to-one merge on specified key variables
  ```
  merge 1:1 varlist using filename [, options]
  ```
- Many-to-one merge on specified key variables
  ```
  merge m:1 varlist using filename [, options]
  ```
Basic data management in STATA (Merge CRSP - Compustat)

**Example (1):** Merge CRSP common stocks (Share Codes 10,11) of US pharmaceutical preparation companies (SIC 2834) with Compustat for the period 2012:2013.

**STEP3 – Merge CRSP and COMPUSSTAT Datasets by CUSIP**

<table>
<thead>
<tr>
<th>permno</th>
<th>date</th>
<th>cusip</th>
<th>company</th>
<th>ret</th>
<th>qvkey</th>
<th>comp_fyear</th>
<th>datasetc</th>
<th>at</th>
</tr>
</thead>
<tbody>
<tr>
<td>10116</td>
<td>31jan2012</td>
<td>00509L80</td>
<td>ACURA PHARMACEUTICALS INC</td>
<td>-.0515759561</td>
<td>011929</td>
<td>2010</td>
<td>31dec2010</td>
<td>25.493</td>
</tr>
<tr>
<td>10116</td>
<td>29feb2012</td>
<td>00509L80</td>
<td>ACURA PHARMACEUTICALS INC</td>
<td>-.021148017</td>
<td>011929</td>
<td>2010</td>
<td>31dec2010</td>
<td>25.493</td>
</tr>
<tr>
<td>10116</td>
<td>30mar2012</td>
<td>00509L80</td>
<td>ACURA PHARMACEUTICALS INC</td>
<td>.067901246</td>
<td>011929</td>
<td>2010</td>
<td>31dec2010</td>
<td>25.493</td>
</tr>
<tr>
<td>10116</td>
<td>30apr2012</td>
<td>00509L80</td>
<td>ACURA PHARMACEUTICALS INC</td>
<td>-.0669735991</td>
<td>011929</td>
<td>2010</td>
<td>31dec2010</td>
<td>25.493</td>
</tr>
<tr>
<td>10116</td>
<td>31may2012</td>
<td>00509L80</td>
<td>ACURA PHARMACEUTICALS INC</td>
<td>-.126935005</td>
<td>011929</td>
<td>2010</td>
<td>31dec2010</td>
<td>25.493</td>
</tr>
<tr>
<td>10116</td>
<td>29jun2012</td>
<td>00509L80</td>
<td>ACURA PHARMACEUTICALS INC</td>
<td>.113475241</td>
<td>011929</td>
<td>2010</td>
<td>31dec2010</td>
<td>25.493</td>
</tr>
<tr>
<td>10116</td>
<td>31jul2012</td>
<td>00509L80</td>
<td>ACURA PHARMACEUTICALS INC</td>
<td>-.407643318</td>
<td>011929</td>
<td>2011</td>
<td>31dec2011</td>
<td>37.173</td>
</tr>
<tr>
<td>10116</td>
<td>31aug2012</td>
<td>00509L80</td>
<td>ACURA PHARMACEUTICALS INC</td>
<td>-.123665923</td>
<td>011929</td>
<td>2011</td>
<td>31dec2011</td>
<td>37.173</td>
</tr>
<tr>
<td>10116</td>
<td>28sep2012</td>
<td>00509L80</td>
<td>ACURA PHARMACEUTICALS INC</td>
<td>.067894669</td>
<td>011929</td>
<td>2011</td>
<td>31dec2011</td>
<td>37.173</td>
</tr>
<tr>
<td>10116</td>
<td>30oct2012</td>
<td>00509L80</td>
<td>ACURA PHARMACEUTICALS INC</td>
<td>-.132183924</td>
<td>011929</td>
<td>2011</td>
<td>31dec2011</td>
<td>37.173</td>
</tr>
<tr>
<td>10116</td>
<td>30nov2012</td>
<td>00509L80</td>
<td>ACURA PHARMACEUTICALS INC</td>
<td>.211920559</td>
<td>011929</td>
<td>2011</td>
<td>31dec2011</td>
<td>37.173</td>
</tr>
<tr>
<td>10116</td>
<td>31dec2012</td>
<td>00509L80</td>
<td>ACURA PHARMACEUTICALS INC</td>
<td>.213114738</td>
<td>011929</td>
<td>2011</td>
<td>31dec2011</td>
<td>37.173</td>
</tr>
<tr>
<td>10116</td>
<td>31jan2013</td>
<td>00509L80</td>
<td>ACURA PHARMACEUTICALS INC</td>
<td>-.197687688</td>
<td>011929</td>
<td>2011</td>
<td>31dec2011</td>
<td>37.173</td>
</tr>
<tr>
<td>10116</td>
<td>28feb2013</td>
<td>00509L80</td>
<td>ACURA PHARMACEUTICALS INC</td>
<td>.101604246</td>
<td>011929</td>
<td>2011</td>
<td>31dec2011</td>
<td>37.173</td>
</tr>
<tr>
<td>10116</td>
<td>20mar2013</td>
<td>00509L80</td>
<td>ACURA PHARMACEUTICALS INC</td>
<td>.033300666</td>
<td>011929</td>
<td>2011</td>
<td>31dec2011</td>
<td>37.173</td>
</tr>
<tr>
<td>10116</td>
<td>30apr2013</td>
<td>00509L80</td>
<td>ACURA PHARMACEUTICALS INC</td>
<td>.150234699</td>
<td>011929</td>
<td>2011</td>
<td>31dec2011</td>
<td>37.173</td>
</tr>
<tr>
<td>10116</td>
<td>31may2013</td>
<td>00509L80</td>
<td>ACURA PHARMACEUTICALS INC</td>
<td>-.077551045</td>
<td>011929</td>
<td>2011</td>
<td>31dec2011</td>
<td>37.173</td>
</tr>
<tr>
<td>10116</td>
<td>28jun2013</td>
<td>00509L80</td>
<td>ACURA PHARMACEUTICALS INC</td>
<td>-.189141889</td>
<td>011929</td>
<td>2011</td>
<td>31dec2011</td>
<td>37.173</td>
</tr>
<tr>
<td>10116</td>
<td>31jul2013</td>
<td>00509L80</td>
<td>ACURA PHARMACEUTICALS INC</td>
<td>.088106365</td>
<td>011929</td>
<td>2012</td>
<td>31dec2012</td>
<td>29.054</td>
</tr>
</tbody>
</table>

\[ t = 2012 / Jul2012 – Jun2013 / fyear t-1 = 2011 \]
Basic data management in STATA (Merge CRSP - Compustat)

Example (1): Merge CRSP common stocks (Share Codes 10,11) of US pharmaceutical preparation companies (SIC 2834) with Compustat for the period 2012:2013.

STEP3 – Merge CRSP and COMPUSTAT Datasets by CUSIP

```
. su yearfrac, d

    -----.  Percentiles    Smallest
          |            | Largest
    1%    5808219          .1671233
    5%    5835617          .2520548
    10%   .6630137         .3287671  Obs      2,851
    25%   .8328767         .3342466  Sum of Wgt. 2,851

    50%    1.084931   Mean     1.079154

          |            | Std. Dev.
    Largest  1.331507    2  Variance .1055117
    90%    1.493151    2  Skewness .2090855
    95%    1.506849    2  Kurtosis 2.500372
    99%    1.920548    2

. br permno date cusip ticker comnam ret gvkey comp_fyear datadate comm at
```
Example (2): Merge CRSP common stocks (Share Codes 10,11) of US pharmaceutical preparation companies (SIC 2834) with Compustat for the period 2012:2013. Most recent report as long as it’s at least 3 months old. – JOINBY --

STEP3 – Merge CRSP and COMPUSTAT Datasets by CUSIP

```
clear all
set more off

cd "D:\Documents\FDM - PhD 2020\data"

* open CRSP file
use ex1_crspcomp_merge_cusip_crsp.dta, clear

* drop records with missing cusip - common identifier in CRSP & COMPUSTAT
drop if cusip=="

* create a new variable cusip8 (copy of cusip) to facilitate the merge command
gen cusip8 = cusip

* Check for duplicates
duplicates report cusip8 date

* check distinct # obs. before joinby
distinct cusip8 date, joint

* For illustrative purposes only, in this example we will merge CRSP market data,
* for each CRSP date, with the most recent financial report provided that
* the report is at least 3 months old (consider a year fraction of 0.25)

* issue the joinby command by cusip8 to create a cartesian product
* that includes every compustat observations with the same cusip8 for each cusip8-date of CRSP
joinby cusip8 using ex1_crspcomp_aux1.dta, unmatched(master)

* check distinct # obs. after joinby
distinct cusip8 date, joint
```
Basic data management in STATA (Merge CRSP - Compustat)

Example (2): Merge CRSP common stocks (Share Codes 10,11) of US pharmaceutical preparation companies (SIC 2834) with Compustat for the period 2012:2013.

STEP3 – Merge CRSP and COMPUSTAT Datasets by CUSIP

* axp. year fraction difference in days / 365 days
  gen yearfrac = (date - datadate)/365

* drop if compustat datadate is missing
  drop if datadate == .

* drop if financial data is less than 3 months old
  * Prior to 2003, most public companies were allowed to file the annual meeting report (10K) within 90 days
  * of fiscal year-end. After 2005 certain large companies were required to file their 10K within 60 days.
  drop if yearfrac < 0.25

* sort by permno date and descendent compustat report date
  qsort permno date -datadate

* count distinct permno-date before dropping duplicates
  distinct permno date, joint

duplicates drop permno date, force

* count distinct permno-date after dropping duplicates
  distinct permno date, joint

* Check the nr. ob observations following the merge command
  su

format commnam %32s
format comm %32s

sort permno date
br permno date cusip ticker commnam ret gvykey comp_fyear datadate comm at
su yearfrac, d

Instead of year fraction (yearfrac) you can use something more precise:

gen ddays = date - datadate

drop if ddays < 90
Basic data management in STATA (Merge CRSP - Compustat)

Example (2): Merge CRSP common stocks (Share Codes 10,11) of US pharmaceutical preparation companies (SIC 2834) with Compustat for the period 2012:2013.

**STEP3 – Merge CRSP and COMPUSTAT Datasets by CUSIP**

<table>
<thead>
<tr>
<th>permno</th>
<th>date</th>
<th>cusip</th>
<th>ticker</th>
<th>commnam</th>
<th>ret</th>
<th>gvkey</th>
<th>comp_fyear</th>
<th>datadate</th>
<th>at</th>
</tr>
</thead>
<tbody>
<tr>
<td>10116</td>
<td>31jan2012</td>
<td>00509L80</td>
<td>ACUR</td>
<td>ACURA PHARMACEUTICALS INC</td>
<td>-.051875991</td>
<td>011929</td>
<td>2010</td>
<td>31dec2010</td>
<td>25.493</td>
</tr>
<tr>
<td>10116</td>
<td>29feb2012</td>
<td>00509L80</td>
<td>ACUR</td>
<td>ACURA PHARMACEUTICALS INC</td>
<td>-.021148017</td>
<td>011929</td>
<td>2010</td>
<td>31dec2010</td>
<td>25.493</td>
</tr>
<tr>
<td>10116</td>
<td>30mar2012</td>
<td>00509L80</td>
<td>ACUR</td>
<td>ACURA PHARMACEUTICALS INC</td>
<td>.067901246</td>
<td>011929</td>
<td>2010</td>
<td>31dec2010</td>
<td>25.493</td>
</tr>
<tr>
<td>10116</td>
<td>30apr2012</td>
<td>00509L80</td>
<td>ACUR</td>
<td>ACURA PHARMACEUTICALS INC</td>
<td>-.066473991</td>
<td>011929</td>
<td>2011</td>
<td>31dec2011</td>
<td>37.173</td>
</tr>
<tr>
<td>10116</td>
<td>31may2012</td>
<td>00509L80</td>
<td>ACUR</td>
<td>ACURA PHARMACEUTICALS INC</td>
<td>-.126935005</td>
<td>011929</td>
<td>2011</td>
<td>31dec2011</td>
<td>37.173</td>
</tr>
<tr>
<td>10116</td>
<td>29jun2012</td>
<td>00509L80</td>
<td>ACUR</td>
<td>ACURA PHARMACEUTICALS INC</td>
<td>.113475241</td>
<td>011929</td>
<td>2011</td>
<td>31dec2011</td>
<td>37.173</td>
</tr>
<tr>
<td>10116</td>
<td>31jul2012</td>
<td>00509L80</td>
<td>ACUR</td>
<td>ACURA PHARMACEUTICALS INC</td>
<td>-.407643318</td>
<td>011929</td>
<td>2011</td>
<td>31dec2011</td>
<td>37.173</td>
</tr>
<tr>
<td>10116</td>
<td>31aug2012</td>
<td>00509L80</td>
<td>ACUR</td>
<td>ACURA PHARMACEUTICALS INC</td>
<td>-.123658923</td>
<td>011929</td>
<td>2011</td>
<td>31dec2011</td>
<td>37.173</td>
</tr>
<tr>
<td>10116</td>
<td>28sep2012</td>
<td>00509L80</td>
<td>ACUR</td>
<td>ACURA PHARMACEUTICALS INC</td>
<td>.067846669</td>
<td>011929</td>
<td>2011</td>
<td>31dec2011</td>
<td>37.173</td>
</tr>
<tr>
<td>10116</td>
<td>30oct2012</td>
<td>00509L80</td>
<td>ACUR</td>
<td>ACURA PHARMACEUTICALS INC</td>
<td>-.132183924</td>
<td>011929</td>
<td>2011</td>
<td>31dec2011</td>
<td>37.173</td>
</tr>
<tr>
<td>10116</td>
<td>30nov2012</td>
<td>00509L80</td>
<td>ACUR</td>
<td>ACURA PHARMACEUTICALS INC</td>
<td>.211920569</td>
<td>011929</td>
<td>2011</td>
<td>31dec2011</td>
<td>37.173</td>
</tr>
<tr>
<td>10116</td>
<td>31dec2012</td>
<td>00509L80</td>
<td>ACUR</td>
<td>ACURA PHARMACEUTICALS INC</td>
<td>.213114738</td>
<td>011929</td>
<td>2011</td>
<td>31dec2011</td>
<td>37.173</td>
</tr>
<tr>
<td>10116</td>
<td>31jan2013</td>
<td>00509L80</td>
<td>ACUR</td>
<td>ACURA PHARMACEUTICALS INC</td>
<td>-.157657668</td>
<td>011929</td>
<td>2011</td>
<td>31dec2011</td>
<td>37.173</td>
</tr>
<tr>
<td>10116</td>
<td>28feb2013</td>
<td>00509L80</td>
<td>ACUR</td>
<td>ACURA PHARMACEUTICALS INC</td>
<td>.101604246</td>
<td>011929</td>
<td>2011</td>
<td>31dec2011</td>
<td>37.173</td>
</tr>
<tr>
<td>10116</td>
<td>28mar2013</td>
<td>00509L80</td>
<td>ACUR</td>
<td>ACURA PHARMACEUTICALS INC</td>
<td>.033980683</td>
<td>011929</td>
<td>2011</td>
<td>31dec2011</td>
<td>37.173</td>
</tr>
<tr>
<td>10116</td>
<td>30apr2013</td>
<td>00509L80</td>
<td>ACUR</td>
<td>ACURA PHARMACEUTICALS INC</td>
<td>.180234699</td>
<td>011929</td>
<td>2012</td>
<td>31dec2012</td>
<td>29.084</td>
</tr>
<tr>
<td>10116</td>
<td>31may2013</td>
<td>00509L80</td>
<td>ACUR</td>
<td>ACURA PHARMACEUTICALS INC</td>
<td>-.077651048</td>
<td>011929</td>
<td>2012</td>
<td>31dec2012</td>
<td>29.084</td>
</tr>
</tbody>
</table>

31jan2012 / most recent report date 31dec2011 / 1 month difference only
30apr2012 / most recent report date 31dec2011 / 4 months difference
Basic data management in STATA (Merge CRSP - Compustat)

**Example (2):** Merge CRSP common stocks (Share Codes 10,11) of US pharmaceutical preparation companies (SIC 2834) with Compustat for the period 2012:2013.

**STEP3 – Merge CRSP and COMPUSTAT Datasets by CUSIP**

```
. su yearfrac, d
```

<table>
<thead>
<tr>
<th>Percentiles</th>
<th>Smallest</th>
<th>Largest</th>
<th>Std. Dev.</th>
<th>Variance</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>.3287671</td>
<td>1.00274</td>
<td>.2908622</td>
<td>.0846124</td>
<td>.0142461</td>
<td>1.835334</td>
</tr>
<tr>
<td>5%</td>
<td>.3315068</td>
<td>1.334247</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10%</td>
<td>.4136986</td>
<td>1.589041</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25%</td>
<td>.4986301</td>
<td>1.668493</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50%</td>
<td>.7506849</td>
<td>1.750685</td>
<td>.7851856</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Basic data management in STATA (Merging datasets)

Fuzzy matching in STATA – Database management functions

CROSS

Cartesian Product $\rightarrow$ Generate all possible combinations $\rightarrow$ CROSS

[D] cross — Form every pairwise combination of two datasets

Syntax

cross using filename

Menu

Data > Combine datasets > Form every pairwise combination of two datasets

Description

cross forms every pairwise combination of the data in memory with the data in filename.

Crossing refers to merging two datasets in every way possible. That is, the first observation of the data in memory is merged with every observation of filename, followed by the second, and so on. Thus the result will have $N_1N_2$ observations, where $N_1$ and $N_2$ are the number of observations in memory and in filename, respectively.
Basic data management in STATA (Merging datasets)

Fuzzy matching in STATA – Database management functions

JOINBY

Cartesian Product → Generate all possible combinations within groups → JOINBY

Title

[D] joinby — Form all pairwise combinations within groups

Syntax

joinby [varlist] using filename [, options]

options      Description

Options
When observations match:
update replace missing data in memory with values from filename
replace replace all data in memory with values from filename

When observations do not match:
unmatched (none) ignore all; the default
unmatched (both) include from both datasets
unmatched (master) include from data in memory
unmatched (using) include from data in filename

_merge (varname) varname marks source of resulting observation; default is _merge
nolabel do not copy value-label definitions from filename

varlist may not contain strs.
Basic data management in STATA (Merging datasets)

STATA – edit – copy/paste

<table>
<thead>
<tr>
<th>A</th>
<th>ticker</th>
<th>B</th>
<th>comm</th>
<th>C</th>
<th>state</th>
<th>D</th>
<th>tic</th>
<th>E</th>
<th>name</th>
<th>F</th>
<th>state</th>
<th>G</th>
<th>cstate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>ticker</td>
<td></td>
<td>comm</td>
<td>state</td>
<td>tic</td>
<td>name</td>
<td>state</td>
<td>cstate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>NVDA</td>
<td>NVIDIA</td>
<td>CA</td>
<td>NVDA</td>
<td>NVIDIA</td>
<td>CA</td>
<td>CA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>WFC</td>
<td>WELLS FARGO</td>
<td>CA</td>
<td>WFC</td>
<td>WELLS FARGO</td>
<td>CA</td>
<td>CA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>INTC</td>
<td>INTEL</td>
<td>CA</td>
<td>INTC</td>
<td>INTEL</td>
<td>CA</td>
<td>CA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>AAPL</td>
<td>APPLE</td>
<td>CA</td>
<td>AAPL</td>
<td>APPLE</td>
<td>CA</td>
<td>CA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>BA</td>
<td>BOEING</td>
<td>IL</td>
<td>BA</td>
<td>BOEING</td>
<td>IL</td>
<td>IL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>CAT</td>
<td>CATERPILLAR</td>
<td>IL</td>
<td>CAT</td>
<td>CATERPILLAR</td>
<td>IL</td>
<td>IL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>HON</td>
<td>HONEYWELL</td>
<td>NJ</td>
<td>HON</td>
<td>HONEYWELL</td>
<td>NJ</td>
<td>NJ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>JNJ</td>
<td>J&amp;J</td>
<td>NJ</td>
<td>JNJ</td>
<td>J&amp;J</td>
<td>NJ</td>
<td>NJ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>JPM</td>
<td>JPMORGAN</td>
<td>NY</td>
<td>JPM</td>
<td>JPMORGAN</td>
<td>NY</td>
<td>NY</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>XOM</td>
<td>EXXON MOBIL</td>
<td>TX</td>
<td>KO</td>
<td>COCA-COLA</td>
<td>GA</td>
<td>GA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MSFT</td>
<td>MICROSOFT</td>
<td>WA</td>
<td>WA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ticker</th>
<th>comm</th>
<th>state</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>NVDA</td>
<td>NVIDIA</td>
</tr>
<tr>
<td>2</td>
<td>WFC</td>
<td>WELLS FARGO</td>
</tr>
<tr>
<td>3</td>
<td>INTC</td>
<td>INTEL</td>
</tr>
<tr>
<td>4</td>
<td>AAPL</td>
<td>APPLE</td>
</tr>
<tr>
<td>5</td>
<td>BA</td>
<td>BOEING</td>
</tr>
<tr>
<td>6</td>
<td>CAT</td>
<td>CATERPILLAR</td>
</tr>
<tr>
<td>7</td>
<td>HON</td>
<td>HONEYWELL</td>
</tr>
<tr>
<td>8</td>
<td>JNJ</td>
<td>J&amp;J</td>
</tr>
<tr>
<td>9</td>
<td>JPM</td>
<td>JPMORGAN</td>
</tr>
<tr>
<td>10</td>
<td>XOM</td>
<td>EXXON MOBIL</td>
</tr>
</tbody>
</table>
Basic data management in STATA (Merging datasets)

```stata
use "C:\Users\Pedro Pires\Desktop\ds1.dta", clear
cross using "C:\Users\Pedro Pires\Desktop\ds2.dta"
sort ticker name
br
```
Basic data management in STATA (Merging datasets)

* JOBY - example

use "C:\Users\Pedro Pires\Desktop\ds1.dta", clear
joinby state using "C:\Users\Pedro Pires\Desktop\ds2.dta", unmatched(master)

sort ticker name

br
✓ **Fuzzy matching in STATA**

➢ Using common identifiers (such as ISIN, CUSIP, and SEDOL) is the best way to merge datasets from different providers.

➢ In practical applications, however, sometimes one has to merge datasets which only contain a string (e.g. company name, fund name). In these cases, fuzzy matching algorithms provide similarity scores that help identifying correct matches.

---

**Syntax**

Data in two columns in the same dataset

```
matchit varname1 varname2 [, options]
```

Data in two different datasets (with indexation)

```
matchit idmaster txtmaster using filename.dta, idusing(varname) txtusing(varname) [options]
```

<table>
<thead>
<tr>
<th>Options</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>simimethod</strong></td>
<td>String matching method. Default is bigram. Other built-in simimethod are: ngram, ngram_circ, token, soundex and token_soundex.</td>
</tr>
<tr>
<td><strong>score</strong></td>
<td>Specifies similarity score. Default is jaccard. Other built-in options are simple and minsimple.</td>
</tr>
<tr>
<td><strong>weights</strong></td>
<td>Weighting transformation. Default is noweights. Built-in options are simple, log and root.</td>
</tr>
<tr>
<td><strong>generate</strong></td>
<td>Specifies the name for the similarity score variable. Default is simiscore.</td>
</tr>
</tbody>
</table>
Fuzzy matching in STATA

Decomposition of text into *grams* of 2 moving chars (*bigrams*)

Weights: based on *grams* frequency (default: no weights – all *grams* eq. 1)

**Description**

`matchit` provides a similarity score between two different text strings by performing many different string-based matching techniques. It returns a new numeric variable (`similscore`) containing the similarity score, which ranges from 0 to 1. A `similscore` of 1 implies a perfect similarity according to the string matching technique chosen and decreases when the match is less similar. `similscore` is a relative measure which can (and often do) change depending on the technique chosen. For more information on these techniques refer to Raffo & Lhuillery (2009).

**Examples for "John Smith":**

<table>
<thead>
<tr>
<th>#</th>
<th>token_grams</th>
<th>bigram</th>
<th>token</th>
<th>soundex</th>
<th>soundex</th>
<th>ngram,1</th>
<th>ngram,3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Jo</td>
<td>John</td>
<td>J525</td>
<td>J500</td>
<td>J</td>
<td>Joh</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>oh</td>
<td>Smith</td>
<td>S530</td>
<td>o</td>
<td>ohn</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>hn</td>
<td></td>
<td>h</td>
<td>hn_</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>n_</td>
<td></td>
<td>n</td>
<td>n_S</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>_S</td>
<td></td>
<td>_</td>
<td>_Sm</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Sm</td>
<td></td>
<td>S</td>
<td>Smi</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>mi</td>
<td></td>
<td>m</td>
<td>mit</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>it</td>
<td></td>
<td>i</td>
<td>ith</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>th</td>
<td></td>
<td>t</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td>h</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Notes:** "_" = a blank space.

ngram, 2 is equivalent to bigram.
Basic data management in STATA (Merging datasets)

**Fuzzy matching in STATA - GED (Generalized Edit Distance)**

<table>
<thead>
<tr>
<th>Obs</th>
<th>String1</th>
<th>String2</th>
<th>Distance</th>
<th>Operation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>baboon</td>
<td>baboon</td>
<td>0</td>
<td>match</td>
</tr>
<tr>
<td>2</td>
<td>baXboon</td>
<td>baboon</td>
<td>100</td>
<td>insert</td>
</tr>
<tr>
<td>3</td>
<td>baoon</td>
<td>baboon</td>
<td>100</td>
<td>delete</td>
</tr>
<tr>
<td>4</td>
<td>baXoon</td>
<td>baboon</td>
<td>100</td>
<td>replace</td>
</tr>
<tr>
<td>5</td>
<td>baboonX</td>
<td>baboon</td>
<td>50</td>
<td>append</td>
</tr>
<tr>
<td>6</td>
<td>baboo</td>
<td>baboon</td>
<td>10</td>
<td>truncate</td>
</tr>
<tr>
<td>7</td>
<td>babboon</td>
<td>baboon</td>
<td>20</td>
<td>double</td>
</tr>
<tr>
<td>8</td>
<td>babon</td>
<td>baboon</td>
<td>20</td>
<td>single</td>
</tr>
<tr>
<td>9</td>
<td>baobon</td>
<td>baboon</td>
<td>20</td>
<td>swap</td>
</tr>
<tr>
<td>10</td>
<td>bab oon</td>
<td>baboon</td>
<td>10</td>
<td>blank</td>
</tr>
<tr>
<td>11</td>
<td>bab,oon</td>
<td>baboon</td>
<td>30</td>
<td>punctuation</td>
</tr>
<tr>
<td>12</td>
<td>bXaoon</td>
<td>baboon</td>
<td>200</td>
<td>insert+delete</td>
</tr>
<tr>
<td>13</td>
<td>bXaYoon</td>
<td>baboon</td>
<td>200</td>
<td>insert+replace</td>
</tr>
<tr>
<td>14</td>
<td>bXoon</td>
<td>baboon</td>
<td>200</td>
<td>delete+replace</td>
</tr>
<tr>
<td>15</td>
<td>Xbaboon</td>
<td>baboon</td>
<td>200</td>
<td>finsert</td>
</tr>
<tr>
<td>16</td>
<td>aboon</td>
<td>baboon</td>
<td>200</td>
<td>trick question: swap+delete</td>
</tr>
<tr>
<td>17</td>
<td>Xaboon</td>
<td>baboon</td>
<td>200</td>
<td>freplace</td>
</tr>
<tr>
<td>18</td>
<td>axoon</td>
<td>baboon</td>
<td>300</td>
<td>fdelete+replace</td>
</tr>
<tr>
<td>19</td>
<td>axoo</td>
<td>baboon</td>
<td>310</td>
<td>fdelete+replace+truncate</td>
</tr>
<tr>
<td>20</td>
<td>axon</td>
<td>baboon</td>
<td>320</td>
<td>fdelete+replace+single</td>
</tr>
<tr>
<td>21</td>
<td>baby</td>
<td>baboon</td>
<td>120</td>
<td>replace+truncate*2</td>
</tr>
<tr>
<td>22</td>
<td>balloon</td>
<td>baboon</td>
<td>200</td>
<td>replace+insert</td>
</tr>
</tbody>
</table>
Basic data management in STATA (Merging datasets)

Fuzzy matching in STATA - Levenshtein distance

Calculates the distance between strings using the Levenshtein distance metric. Levenshtein distance, or edit distance, is the smallest number of edits required to make one string match a second string. An edit may be an insertion, deletion, or substitution of any single letter.

```
help strdist
```

**Title**

```
strdist — Calculate the Levenshtein distance, or edit distance, between strings.
```

**Syntax**

```
strdist {varname1"string1"} {varname2"string2"} [if] [in] [, generate(newvar) maxdist(integer)]
```
Basic data management in STATA (Merging datasets)

Fuzzy matching in STATA

COMPGED Function

The COMPGED function returns the generalized edit distance between two strings. Specifically, the COMPGED function returns a generalization of the Levenshtein edit distance, which is a measure of dissimilarity between two strings. The Levenshtein edit distance is the number of operations (deletions, insertions, or replacement) of a single characters that are required to transform string-1 into string-2.

Each operation basically 'costs' a certain value. For example, if string-1 is the result of inserting a space into string-2, this has a cost of 10. The more dramatic the operation, the greater the cost. The COMPGED will return the total cost for all operations that occur. The costs returned by COMPGED can be altered by using CALL COMPDCOST so that the cost are specific to your needs. A common use I have seen for using the COMPGED function is using it to compare email addresses.

```plaintext
email1='JohnDoe@abc.com';
email2='John_Doe@abc.com';
cost=compged(email1,email2);
```

The value of COST will be 30 which is the cost of adding punctuation to a string.

COMPLEV Function

The COMPLEV function is very similar to the COMPGED function. The difference is that the Levenshtein edit distance that is computed by COMPLEV is a special case of the generalized edit distance that is computed by COMPGED. The result is the COMPLEV executes much more quickly than COMPGED. However, the COMPLEV function is not as powerful or versatile as the COMPGED function. The COMPLEV function is generally most useful when comparing simple strings and when speed of comparison is important.
Contents coverage

➢ Institutional ownership and mutual fund holdings

- CRSP Mutual Funds
- Thomson Reuters - Mutual Fund Holdings
- Lipper
- Thomson Reuters Institutional Holdings
- FactSet LionShares Ownership
Web Query in detail (CRSP Mutual Funds)

CRSP Mutual Funds

- **Mutual Fund data (1962-2018)**
  - Survivorship-bias-free data
  - Historical performance of open-end mutual funds
  - History of each mutual fund’s name, investment style, fee structure, and asset allocation.
  - Monthly total returns, monthly total net assets, monthly/daily net asset values, and dividends.
CRSP Mutual Funds

NOTE ABOUT RETURNS:

Daily and monthly returns values are calculated as a change in NAV including reinvested dividends from one period to the next. Naves are net of all management expenses and 12b-fees. Front and rear load fees are excluded.

Returns are calculated as follows:

\[ R_t = \left[ \frac{\text{Nav}_t \times \text{cumfact}}{\text{Nav}_{t-1}} \right] - 1 \]

DATABASE TABLES OVERVIEW

<table>
<thead>
<tr>
<th>TABLE</th>
<th>NAME</th>
<th>DEFINITION</th>
</tr>
</thead>
<tbody>
<tr>
<td>contact_info</td>
<td>Contact Information</td>
<td>Current and historical contact information</td>
</tr>
<tr>
<td>daily_nav</td>
<td>Daily Net Asset Value</td>
<td>Net Asset Value for each trading day</td>
</tr>
<tr>
<td>daily_returns</td>
<td>Daily Returns</td>
<td>Returns for each trading day</td>
</tr>
<tr>
<td>dividends</td>
<td>Dividends</td>
<td>Fund dividends</td>
</tr>
<tr>
<td>front_load_det</td>
<td>Front Load Detail</td>
<td>Details of front load fees</td>
</tr>
<tr>
<td>front_load_GRP</td>
<td>Front Load Group</td>
<td>Effective dates for front load fees</td>
</tr>
<tr>
<td>fund_fees</td>
<td>Fund Fees</td>
<td>Fees associated with each fund</td>
</tr>
<tr>
<td>fund_hdr</td>
<td>Fund Header</td>
<td>Most recent identification information for each fund</td>
</tr>
<tr>
<td>fund_hdr_hist</td>
<td>Historical Fund Header</td>
<td>Historical identification information for each fund</td>
</tr>
<tr>
<td>fund_style</td>
<td>Fund Style</td>
<td>Style attributes for each fund</td>
</tr>
<tr>
<td>fund_summary</td>
<td>Fund Summary</td>
<td>Summary data for each fund</td>
</tr>
<tr>
<td>holdings</td>
<td>Holdings</td>
<td>Portfolio holding information</td>
</tr>
<tr>
<td>holdings_CC_info</td>
<td>Holdings Company Information</td>
<td>Information about companies held in portfolios</td>
</tr>
<tr>
<td>crsp_portno_map</td>
<td>CRSP PORTNO Map</td>
<td>Map to portfolio for security holdings info</td>
</tr>
<tr>
<td>monthly_nav</td>
<td>Monthly Net Asset Value</td>
<td>Net Asset Values as of the last trading day of each month</td>
</tr>
<tr>
<td>monthly RETURNS</td>
<td>Monthly Returns</td>
<td>Monthly holding period returns</td>
</tr>
<tr>
<td>monthly_tna</td>
<td>Monthly Total Net Assets</td>
<td>Total Net Assets as of the last trading day of each month</td>
</tr>
<tr>
<td>rear_load_det</td>
<td>Rear Load Detail</td>
<td>Details of rear load fees</td>
</tr>
<tr>
<td>rear_load_GRP</td>
<td>Rear Load Group</td>
<td>Effective dates for rear load fees</td>
</tr>
</tbody>
</table>
### CRSP Mutual Funds

#### Fund Fees (FUND_FEES)

<table>
<thead>
<tr>
<th>NAME</th>
<th>DATATYPE</th>
<th>DEFINITION</th>
</tr>
</thead>
<tbody>
<tr>
<td>*crsp_fundno</td>
<td>INTEGER</td>
<td>Unique identifier for fund</td>
</tr>
<tr>
<td>*begdt</td>
<td>DATE</td>
<td>Beginning of date range for fee information. For annual data, information is valid for the year portion of the begin date.</td>
</tr>
<tr>
<td>enddt</td>
<td>DATE</td>
<td>End of date range for fee information</td>
</tr>
</tbody>
</table>
| actual_12b1 | FLOAT    | Reported as the ratio of the total assets attributed to marketing and distribution costs. Represents the actual fee paid in the most recently completed fiscal year as reported in the Annual Report Statement of Operations. Represented in decimal format.  
  - -99 & 0 (prior to 1998) are used when no 12b-1 fee is reported.                                                                 |
| max_12b1    | FLOAT    | Maximum contractual 12b-1 fee. Represented in decimal format.  
  - A fund with an actual 12b-1 value of -99 may have a max_12b1 fee indicating the maximum even though none is currently being reported.  
  - Null values are used when no maximum value is reported.                                                                 |
| exp_ratio   | FLOAT    | Expense Ratio as of the most recently completed fiscal year. Represented in decimal format.  
  - Ratio of total investment that shareholders pay for the fund's operating expenses, which include 12b-1 fees  
  - exp_ratio may include waivers and reimbursements, causing it to appear to be less than the fund management fee. |
| mngt_fee    | FLOAT    | Management Fee ($)/ Average Net Assets ($)  
  - The fee is calculated using ratios based on the line items reported in the Statement of Operations. The management fee can be offset by fee waivers and/or reimbursements which will make this value differ from the contractual fees found in the prospectus.  
  - Reimbursements can lead to negative Management Fees.                                                                 |
| turn_ratio  | FLOAT    | Fund Turnover Ratio. Minimum (of aggregated sales or aggregated purchases of securities), divided by the average 12-month Total Net Assets of the fund. If fiscal_yearend is present, turn_ratio is for the twelve months ending on fiscal_yearend. If fiscal_yearend is missing, then turn_ratio is for the twelve months ending on the begdt. |
CRSP Mutual Funds

<table>
<thead>
<tr>
<th>NAME</th>
<th>DATATYPE</th>
<th>DEFINITION</th>
<th>DATA AVAILABILITY</th>
</tr>
</thead>
<tbody>
<tr>
<td>fund_name</td>
<td>VARCHAR(140)</td>
<td>The full name of the fund</td>
<td></td>
</tr>
<tr>
<td>nasdaq</td>
<td>VARCHAR(5)</td>
<td>NASDAQ ticker symbol. Available for NASDAQ listed funds only.</td>
<td></td>
</tr>
<tr>
<td>ncusip</td>
<td>VARCHAR(9)</td>
<td>Fund CUSIP. Available for internal use by client only. Go to <a href="http://www.crsp.ChicagoBooth.edu/crsp/policies/index.html">www.crsp.ChicagoBooth.edu/crsp/policies/index.html</a> for CUSIP licensing information.</td>
<td></td>
</tr>
<tr>
<td>first_offer_dt</td>
<td>DATE</td>
<td>Date when the fund was first offered</td>
<td></td>
</tr>
<tr>
<td>mgmt_name</td>
<td>VARCHAR(80)</td>
<td>Management Company Name</td>
<td></td>
</tr>
<tr>
<td>mgr_name</td>
<td>VARCHAR(30)</td>
<td>Portfolio Manager Name</td>
<td></td>
</tr>
<tr>
<td>mgr_dt</td>
<td>DATE</td>
<td>Date that the current portfolio manager assumed responsibility for the portfolio</td>
<td></td>
</tr>
<tr>
<td>adv_name</td>
<td>VARCHAR(40)</td>
<td>Fund Advisor Name</td>
<td>December 1999</td>
</tr>
<tr>
<td>open_to_inv</td>
<td>VARCHAR(1)</td>
<td>Identifies if the fund is open to investors: Y = Yes, N = No</td>
<td>December 1999</td>
</tr>
<tr>
<td>retail_fund</td>
<td>VARCHAR(1)</td>
<td>Identifies if a fund is a retail fund or not: Y = Yes, N = No</td>
<td>December 1999</td>
</tr>
<tr>
<td>Inst_fund</td>
<td>VARCHAR(1)</td>
<td>Identifies if a fund is an institutional fund: Y = Yes, N = No</td>
<td>December 1999</td>
</tr>
</tbody>
</table>
## CRSP Mutual Funds

<table>
<thead>
<tr>
<th>NAME</th>
<th>DATATYPE</th>
<th>DEFINITION</th>
<th>DATA AVAILABILITY</th>
</tr>
</thead>
</table>
| m_fund       | VARCHAR(1)| Identifies if a fund was originally a “M” fund:  
Y = Yes  
N = No  
Note: “M” funds are “dead” funds that were introduced to the database when originally built by Mark Carhart in order to eliminate survivor bias.                                                                                       |                   |
| index_fund_flag | VARCHAR(3) | Identifies if a fund is an index fund:  
B = Index-based fund - utilizes indexes as its primary filter for the purchase and sale of securities. This is accomplished by investing in the components of one or more indexes, or by investing in a small percentage of securities within the index in an attempt to capture the best performers. In each case, the option to invest a portion of assets outside the securities held by the index is left open.  
D = Pure Index fund - objective is to match the total investment performance of a publicly recognized securities market index. The fund will hold virtually all securities in the noted index with weightings equal to those in the index.  
E = Index fund enhanced - objective is to exceed the total investment performance of a publicly recognized securities market. This is accomplished by investing primarily in derivatives based on the index itself and/or the securities within the index, or by utilizing different weightings for the securities held by the index.  
| June 2008       |
| vau_fund      |           | Y = Fund is a variable annuity underlying fund  
N = Fund is not a variable annuity underlying fund                                                                                                                                                                                                                                                                                      |                   |
| et_flag       | VARCHAR(1)| Identifies if a fund is an ETF or ETN:  
F = ETF  
N = ETN                                              |                   |
| end_dt        | DATE      | Date of latest NAV data                                                                                                                                                                                                                                                                                                                  |                   |
| dead_flag     | VARCHAR(1)| Identifies if a fund is dead or still active:  
Y = Yes  
N = No                                              |                   |
CRSP Mutual Funds

CRSP Style Code

The CRSP US Survivor-Bias-Free Mutual Funds database includes style and objective codes from three different sources over the life of the database. No single source exists for its full-time range.

- Wiesenberger Objective codes are populated between 1962 – 1993.
- Strategic Insight Objective codes are populated between 1993 – 1998.
- Lipper Objective codes begin 1998.

The CRSP Style Code builds continuity within the database by using the three afore mentioned codes as its base and provides consistency with those codes provided by our different sources.

The CRSP Style Code consists of up to four characters, with each position defined. Reading Left to Right, the four codes represent an increasing level of granularity. For example, a code for a particular mutual fund is EDYG, where:

- \( E = \text{Equity}, \ D = \text{Domestic}, \ Y = \text{Style}, \ G = \text{Growth} \)

Codes with less than four characters exist, and it simply means that they are defined to a less granular level.
## Web Query in detail (CRSP Mutual Funds)

<table>
<thead>
<tr>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
<th>Level 4</th>
<th>Strategic Insights</th>
<th>Wiesenberger</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equity</td>
<td>Domestic</td>
<td>Sector</td>
<td>Gold (G)</td>
<td>GLD Equity Gold</td>
<td>GPM Gold and Precious Metals</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Health (H)</td>
<td>HLT Equity USA Health</td>
<td>HLT Health Sector</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Financial (F)</td>
<td>FIN Equity USA Financial Sector</td>
<td>FIN Financial Sector</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Natural Resources (N)</td>
<td>NTR Equity Natural Resources &amp; Energy</td>
<td>ENR Energy/Natural Resources</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Real Estate (R)</td>
<td>RLE Equity USA Real Estate</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Technology (T)</td>
<td>TEC Equity USA Technology</td>
<td>TCH Technology Sector</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Utilities (U)</td>
<td>UTI Equity USA Utilities</td>
<td>UTL Utilities</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Consumer Goods (G)</td>
<td>CG CONSUMER GOODS FUNDS</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Commodities (C)</td>
<td>CMD COMMODITIES FUNDS</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Consumer Services (S)</td>
<td>CS CONSUMER SERVICES FUNDS</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Industrials (I)</td>
<td>ID INDUSTRIALS FUNDS</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Materials (M)</td>
<td>BM BASIC MATERIALS FUNDS</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Telecom (A)</td>
<td>TL Telecommunication Funds</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Cap-based (C)</td>
<td>SEC Equity USA Misc Sectors</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Large Cap (L)</td>
<td>SP S&amp;P 500 Index Objective Funds</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mid Cap (M)</td>
<td>GMC Equity USA Midcaps</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Small Cap (S)</td>
<td>SCG Equity USA Small Companies</td>
<td>SCG Small Capitalization Growth</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Micro Cap (I)</td>
<td>MR Micro-Cap Funds</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Style (Y)</td>
<td>AGG Equity USA Aggressive Growth</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Growth (G)</td>
<td>GRO Equity USA Growth</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Growth &amp; Income (B)</td>
<td>GRI Equity USA Growth &amp; Income</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Hedged (H)</td>
<td>LSE Long/Short Equity Funds</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Short (S)</td>
<td>OPI Option Income</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Income (I)</td>
<td>IEQ Equity Income</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ENV Equity USA Environmental</td>
<td></td>
</tr>
</tbody>
</table>
Web Query in detail (TR s12)

**Thomson Reuters - Mutual Funds Holdings**

- Holding information for mutual funds that report with the SEC
  - **Type 1: Fund characteristics** (fund name, country, ...)
  - **Type 2: Stock characteristics** (cusip, stock name, shares outstanding, price)
  - **Type 3: Stock holdings** (cusip, # of shares held)
  - **Type 4: Change in holdings** (net change in shares held)
  - **s12 master file (Types 1/2/3/4)**
  - **Link to the CRSP Mutual Funds database using WRDS MFLinks (Not subscribed)**

- Details on the holdings of Fidelity Magellan and other Fidelity funds are available in the S12 sets, while the aggregate for all Fidelity funds is in the S34 set.
Web Query in detail (TR s12)

Thomson Reuters – Mutual Fund Holdings

Example: Get the security holdings for Fidelity Magellan and show the top 20 holdings for June 2018.

Dataset: s12 master file

Thomson Reuters Mutual Fund Holdings - Type 3: Stock Holdings Code Lookup

fidelity magellan

The Company Code Lookup Tool is designed to find all identifiers associated with a specific company. This tool can retrieve identifiers for multiple companies, then add the codes directly to your query. You can also use this tool to save and download codes for future queries. Downloaded codes are saved in the text (.txt) file format.

1. Start by entering a company name (or part of the name) into the search box above.
2. After the results are displayed, make your selections and choose an identifier.
3. Next, refine your list of codes or add more.
4. Last, choose whether to insert your selections into your web query or to download them as a text file for later use.

1 results found that start with "fidelity magellan"

<table>
<thead>
<tr>
<th>ENTITY_NAME</th>
<th>INCORPORATION_COUNTRY</th>
<th>TFN_FUND_NUM</th>
<th>FIRST_DATE</th>
<th>LAST_DATE</th>
</tr>
</thead>
<tbody>
<tr>
<td>FIDELITY MAGELLAN FUND</td>
<td>UNITED STATES</td>
<td>21858</td>
<td>30JUN1981</td>
<td>30JUN2018</td>
</tr>
</tbody>
</table>

Need more help?

Take a moment to read the full step-by-step instructions to using this tool, including a complete list of identifiers available.
**Thomson Reuters – Mutual Fund Holdings**

**Example:** Get the security holdings for Fidelity Magellan and show the top 20 holdings for June 2018.

**Dataset:** sec filing form N-Q [https://www.sec.gov/Archives/edgar/data/61397/000137949118004145/filing924.htm](https://www.sec.gov/Archives/edgar/data/61397/000137949118004145/filing924.htm)

<table>
<thead>
<tr>
<th>stock</th>
<th>shares</th>
<th>value</th>
<th>stkname</th>
<th>shares</th>
<th>ppr</th>
<th>holding</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microsoft Corp.</td>
<td>9,704,700</td>
<td>956,889</td>
<td>MICROSOFT CORP</td>
<td>9,704,700</td>
<td>98.81</td>
<td>$956,880</td>
</tr>
<tr>
<td>Amazon.com, Inc. (a)</td>
<td>362,200</td>
<td>615,668</td>
<td>AMazon.com INC</td>
<td>362,200</td>
<td>1698.17</td>
<td>$615,077</td>
</tr>
<tr>
<td>UnitedHealth Group, Inc.</td>
<td>1,947,584</td>
<td>477,820</td>
<td>UNITEDHEALTH GROUP INC</td>
<td>1,947,584</td>
<td>245.34</td>
<td>$477,820</td>
</tr>
<tr>
<td>Alphabet Class A (a)</td>
<td>393,316</td>
<td>444,128</td>
<td>ALPHABET INC</td>
<td>393,316</td>
<td>1129.19</td>
<td>$444,128</td>
</tr>
<tr>
<td>Alphabet Class C (a)</td>
<td>306,554</td>
<td>442,415</td>
<td>ALPHABET INC</td>
<td>306,554</td>
<td>1115.66</td>
<td>$442,415</td>
</tr>
<tr>
<td>Apple, Inc.</td>
<td>2,355,513</td>
<td>436,029</td>
<td>APPLE INC</td>
<td>2,355,513</td>
<td>182.91</td>
<td>$430,847</td>
</tr>
<tr>
<td>Berkshire Hathaway, Inc. Class B (a)</td>
<td>1,929,673</td>
<td>360,173</td>
<td>BERKSHIRE HATHAWAY INC</td>
<td>1,929,673</td>
<td>188.65</td>
<td>$360,173</td>
</tr>
<tr>
<td>JPMorgan Chase &amp; Co.</td>
<td>3,450,533</td>
<td>360,171</td>
<td>JPMORGAN CHASE &amp; CO</td>
<td>3,450,533</td>
<td>104.2</td>
<td>$360,171</td>
</tr>
<tr>
<td>Facebook, Inc. Class A (a)</td>
<td>1,716,200</td>
<td>333,402</td>
<td>FACEBOOK INC</td>
<td>1,716,200</td>
<td>194.32</td>
<td>$333,402</td>
</tr>
<tr>
<td>Home Depot, Inc.</td>
<td>1,643,163</td>
<td>320,581</td>
<td>HOME DEPOT INC</td>
<td>1,643,163</td>
<td>195.1</td>
<td>$320,581</td>
</tr>
<tr>
<td>Northrop Grumman Corp.</td>
<td>1,025,342</td>
<td>315,498</td>
<td>NORTHROP GRUMMAN CORP</td>
<td>1,025,342</td>
<td>307.7</td>
<td>$315,498</td>
</tr>
<tr>
<td>Boston Scientific Corp. (a)</td>
<td>7,789,172</td>
<td>254,706</td>
<td>BOSTON SCIENTIFIC CORP</td>
<td>7,789,172</td>
<td>32.7</td>
<td>$254,706</td>
</tr>
<tr>
<td>DowDuPont, Inc.</td>
<td>3,829,470</td>
<td>252,439</td>
<td>DOWDUPONT INC</td>
<td>3,829,470</td>
<td>65.92</td>
<td>$252,439</td>
</tr>
<tr>
<td>United Technologies Corp.</td>
<td>2,017,321</td>
<td>252,226</td>
<td>UNITED TECHNOLOGIES CORP</td>
<td>2,017,321</td>
<td>125.03</td>
<td>$252,226</td>
</tr>
<tr>
<td>ConocoPhillips Co.</td>
<td>3,601,083</td>
<td>250,707</td>
<td>CONOCOPHILLIPS</td>
<td>3,601,083</td>
<td>69.62</td>
<td>$250,707</td>
</tr>
<tr>
<td>American Tower Corp.</td>
<td>1,730,050</td>
<td>249,421</td>
<td>AMERICAN TOWER CORP</td>
<td>1,730,050</td>
<td>144.17</td>
<td>$249,421</td>
</tr>
<tr>
<td>Monster Beverage Corp. (a)</td>
<td>4,310,100</td>
<td>246,969</td>
<td>MONSTER BEVERAGE CORP</td>
<td>4,310,100</td>
<td>57.3</td>
<td>$246,969</td>
</tr>
<tr>
<td>Intuit, Inc.</td>
<td>1,188,200</td>
<td>242,755</td>
<td>INTUIT INC</td>
<td>1,188,200</td>
<td>204.33</td>
<td>$242,755</td>
</tr>
<tr>
<td>EOG Resources, Inc.</td>
<td>1,946,292</td>
<td>242,177</td>
<td>EOG RESOURCES INC</td>
<td>1,946,292</td>
<td>124.43</td>
<td>$242,177</td>
</tr>
</tbody>
</table>
Lipper for Investment Management Not subscribed

- Provides mutual fund data (returns, TNA, benchmark, fund family, total expense ratio (TER), loads, fund domicile, countries notified for sale, etc.)

- Coverage of over 213,000 share classes of more than 117,000 funds located in over 60 countries

- Ferreira, Keswani, Miguel and Ramos (2012) find that Lipper covers 87% of the total TNA of worldwide equity mutual funds (as of Dec/2007)
<table>
<thead>
<tr>
<th>Lipper_ID</th>
<th>Name</th>
<th>CUSIP</th>
<th>ISIN_Code</th>
<th>Asset_ Universe</th>
<th>Asset_Type</th>
<th>Asset_Status</th>
<th>Geographical_Focus</th>
<th>Lipper_Global</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Lipper Growth Fund Index</td>
<td>761724202</td>
<td>US7617242026</td>
<td>Mutual Funds</td>
<td>Equity</td>
<td>Liquidated</td>
<td>United States of America</td>
<td>Equity Global</td>
</tr>
<tr>
<td>2</td>
<td>Reynolds Opportunity Fund</td>
<td>757591042</td>
<td>US7575910426</td>
<td>Mutual Funds</td>
<td>Equity</td>
<td>Liquidated</td>
<td>United States of America</td>
<td>Equity North America</td>
</tr>
<tr>
<td>3</td>
<td>PNC Large Cp Growth A</td>
<td>603814293</td>
<td>US6038142930</td>
<td>Mutual Funds</td>
<td>Equity</td>
<td>Active</td>
<td>United States of America</td>
<td>Equity North America</td>
</tr>
<tr>
<td>4</td>
<td>T Rowe Price Japan Fund</td>
<td>773594708</td>
<td>US7735947085</td>
<td>Mutual Funds</td>
<td>Equity</td>
<td>Active</td>
<td>Japan</td>
<td>Equity Japan</td>
</tr>
<tr>
<td>5</td>
<td>Dreyfus International Growth Fund I</td>
<td>2619691030</td>
<td>US2619691030</td>
<td>Mutual Funds</td>
<td>Equity</td>
<td>Liquidated</td>
<td>Global Ex US</td>
<td>Equity Global &amp; U.S.</td>
</tr>
<tr>
<td>6</td>
<td>Touchstone HAM Large Cp Growth Fund II</td>
<td>691594702</td>
<td>US6915947026</td>
<td>Mutual Funds</td>
<td>Equity</td>
<td>Liquidated</td>
<td>United States of America</td>
<td>Equity North America</td>
</tr>
<tr>
<td>7</td>
<td>Scudder Dynamic Growth Fund (Class A)</td>
<td>81114R1032</td>
<td>US81114R1032</td>
<td>Mutual Funds</td>
<td>Equity</td>
<td>Merged</td>
<td>United States of America</td>
<td>Equity Mid Cap</td>
</tr>
<tr>
<td>8</td>
<td>BlueSource New Dimensions Fund A</td>
<td>000345108</td>
<td>US0003451081</td>
<td>Mutual Funds</td>
<td>Equity</td>
<td>Merged</td>
<td>United States of America</td>
<td>Equity North America</td>
</tr>
<tr>
<td>9</td>
<td>Lead Abbot Affiliated Fund A</td>
<td>544001102</td>
<td>US5440011001</td>
<td>Mutual Funds</td>
<td>Equity</td>
<td>Active</td>
<td>United States of America</td>
<td>Equity North America</td>
</tr>
<tr>
<td>10</td>
<td>Seigman Capital Fund A</td>
<td>815326012</td>
<td>US8153260122</td>
<td>Mutual Funds</td>
<td>Equity</td>
<td>Merged</td>
<td>United States of America</td>
<td>Equity Mid Cap</td>
</tr>
<tr>
<td>11</td>
<td>American Funds AMCEF Fund A</td>
<td>0023975063</td>
<td>US0023975056</td>
<td>Mutual Funds</td>
<td>Equity</td>
<td>Active</td>
<td>United States of America</td>
<td>Equity North America</td>
</tr>
<tr>
<td>12</td>
<td>Midas Investors Ltd</td>
<td>0053630100</td>
<td>US0053630100</td>
<td>Mutual Funds</td>
<td>Equity</td>
<td>Merged</td>
<td>United States of America</td>
<td>Equity Sector Gold/Proc. Metals</td>
</tr>
<tr>
<td>13</td>
<td>American Growth Fund D</td>
<td>0053330100</td>
<td>US0053330100</td>
<td>Mutual Funds</td>
<td>Equity</td>
<td>Active</td>
<td>United States of America</td>
<td>Equity North America</td>
</tr>
<tr>
<td>14</td>
<td>Invesco Charlem Fund A</td>
<td>001413033</td>
<td>US001413033</td>
<td>Mutual Funds</td>
<td>Equity</td>
<td>Liquidated</td>
<td>United States of America</td>
<td>Equity North America</td>
</tr>
<tr>
<td>15</td>
<td>Smith Barney Equities Funds, Inc.; Class A Shares</td>
<td>001413033</td>
<td>US001413033</td>
<td>Mutual Funds</td>
<td>Equity</td>
<td>Liquidated</td>
<td>United States of America</td>
<td>Equity North America</td>
</tr>
</tbody>
</table>

**Fund Manager**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>S&amp;P 500 TR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Russell 1000 Growth TR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSCI Japan NR USD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Russell 1000 Growth TR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Russell 1000 Growth TR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Russell MidCap Growth TR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S&amp;P 500 TR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Russell 1000 Value TR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Russell MidCap Growth TR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S&amp;P 500 TR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Russell 1000 Value TR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Russell MidCap Growth TR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S&amp;P 500 TR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Russell 1000 Value TR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Thomson Reuters Institutional Holdings

➢ Formerly known as CDA/Spectrum

➢ Section 13(f) of the Securities Exchange Act (1975)

➢ Institutional Common Stock Holdings as reported on Form 13F filed with the SEC

➢ Ownership information by institutional managers with $100 million or more in Assets Under Management
Thomson Reuters 13F Ownership Tool

➢ Ownership of Pfizer in 2014-06-30

Step 1: Choose your date range.

Date Variable: File Date

Date range

2014-06 to 2014-06

Step 2: Apply your company codes.

Select an option for entering company codes

FPE

Data Request Summary

Your output is complete. Click on the link below to open the file:
a305fc46e1616662.dta (3 KB, 1 observations 18 variables)
Thomson Reuters 13F Ownership Tool

- Ownership of Pfizer in 2014-06-30
Web Query in detail (Thomson Reuters - Ownership)

**Thomson Reuters 13F Ownership Tool**

- Ownership of Pfizer in 2014-06-30

<table>
<thead>
<tr>
<th>rdate</th>
<th>stkname</th>
<th>ticker</th>
<th>instown</th>
<th>instown_hhi</th>
<th>instown_perc</th>
</tr>
</thead>
<tbody>
<tr>
<td>30Jun2014</td>
<td>PFIZER INC</td>
<td>PFE</td>
<td>4475315092</td>
<td>.024950791</td>
<td>.6309282995</td>
</tr>
</tbody>
</table>

You can access more detailed data (e.g. by institution or fund) using the SAS code used in the Tool provided by WRDS (Moussawi and Palacios, 2009) or by using the Thomson Reuters raw files (Web Query)

[http://wrds-web.wharton.upenn.edu/wrds/research/applications/index.cfm](http://wrds-web.wharton.upenn.edu/wrds/research/applications/index.cfm)
FactSet/LionShares Fundamentals and Ownership Not Subscribed

➢ Contains global equity ownership data for approximately 13,000 institutions and 33,000 mutual funds

➢ History dating back to 1999

➢ Unadjusted Package (survivorship bias free)

➢ Fundamentals Data (same as Worldscope)

➢ Stock Ownership Summary Tool (Ferreira and Matos, JFE 2008) programmed by Pedro Pires
Web Query in detail (FactSet/LionShares)

Factset - Stock Ownership Summary

This research application provides institutional ownership statistics by firm and was developed by Miguel Ferreira (Nova School of Business and Economics, miguel.ferreira@novasbe.pt) and Pedro Matos (University of Virginia - Darden School of Business, matosp@virginia.edu).

Rules for usage - Please respect the following three rules when using the Stock Ownership Summary file:
2. Please do not share this file as it is for academic use only. Please refer others to this web page.
3. If you find any errors please notify us at support ownership@novasbe.pt so that we can update the file and notify others. We collect your e-mail address for notification of problems and updates.

For further details, please find here the Stock Ownership Overview developed by Miguel Ferreira and Pedro Matos.

For the SAS code that generates the data below at the firm/quarter level, please see the sample programs provided by the authors.

For more about this dataset, see the data below at the firm/quarter level, please see the Sample programs provided by the authors.

Step 1: What date range do you want to use?
Date Variable: RKQUARTER

I would like data from start date: 2010-01, to end date: 2012-01 (YYYY-mm).

Step 2: How would you like to search this dataset?
What format are your company codes?

FACTSET_ENTITY_ID
• TIC
• CUSIP
• ISIN

<table>
<thead>
<tr>
<th>Firms</th>
<th>Stock</th>
<th>Quarter (Report Date)</th>
<th>Firms name</th>
<th>ISIN (primary)</th>
<th>TIC (primary)</th>
<th>Market capitalization in USD1000000</th>
<th>Total institutional ownership in USD1000000</th>
<th>Number of institutional owners</th>
<th>Total institutional ownership ratio in market capitalization</th>
<th>Domestic institutional ownership ratio in percentage of market</th>
<th>Foreign institutional ownership ratio in percentage of market</th>
</tr>
</thead>
<tbody>
<tr>
<td>05HH95-E</td>
<td>201001</td>
<td>ENERGIAS DE PORTUGAL SA</td>
<td>PTEDDOAM0009</td>
<td>EDP</td>
<td>14223.14</td>
<td>13504.49</td>
<td>367</td>
<td>12.8%</td>
<td>0.3%</td>
<td>12.5%</td>
<td></td>
</tr>
<tr>
<td>05HH95-E</td>
<td>201001</td>
<td>ENERGIAS DE PORTUGAL SA</td>
<td>PTEDDOAM0009</td>
<td>EDP</td>
<td>10828.32</td>
<td>1320.95</td>
<td>354</td>
<td>12.2%</td>
<td>0.2%</td>
<td>12.0%</td>
<td></td>
</tr>
<tr>
<td>05HH95-E</td>
<td>201001</td>
<td>ENERGIAS DE PORTUGAL SA</td>
<td>PTEDDOAM0009</td>
<td>EDP</td>
<td>12434.57</td>
<td>1372.07</td>
<td>341</td>
<td>12.0%</td>
<td>0.2%</td>
<td>12.4%</td>
<td></td>
</tr>
<tr>
<td>05HH95-E</td>
<td>201001</td>
<td>ENERGIAS DE PORTUGAL SA</td>
<td>PTEDDOAM0009</td>
<td>EDP</td>
<td>12106.05</td>
<td>1783.99</td>
<td>330</td>
<td>13.7%</td>
<td>0.2%</td>
<td>13.5%</td>
<td></td>
</tr>
<tr>
<td>05HH95-E</td>
<td>201001</td>
<td>ENERGIAS DE PORTUGAL SA</td>
<td>PTEDDOAM0009</td>
<td>EDP</td>
<td>13635.41</td>
<td>2079.03</td>
<td>317</td>
<td>15.2%</td>
<td>0.2%</td>
<td>15.1%</td>
<td></td>
</tr>
<tr>
<td>05HH95-E</td>
<td>201001</td>
<td>ENERGIAS DE PORTUGAL SA</td>
<td>PTEDDOAM0009</td>
<td>EDP</td>
<td>11257.14</td>
<td>1603.34</td>
<td>204</td>
<td>14.2%</td>
<td>0.2%</td>
<td>14.3%</td>
<td></td>
</tr>
<tr>
<td>05HH95-E</td>
<td>201001</td>
<td>ENERGIAS DE PORTUGAL SA</td>
<td>PTEDDOAM0009</td>
<td>EDP</td>
<td>11549.01</td>
<td>1647.42</td>
<td>807</td>
<td>14.6%</td>
<td>0.2%</td>
<td>14.4%</td>
<td></td>
</tr>
<tr>
<td>05HH95-E</td>
<td>201201</td>
<td>ENERGIAS DE PORTUGAL SA</td>
<td>PTEDDOAM0009</td>
<td>EDP</td>
<td>10524.96</td>
<td>1536.23</td>
<td>305</td>
<td>14.6%</td>
<td>0.1%</td>
<td>14.5%</td>
<td></td>
</tr>
</tbody>
</table>
Contents coverage

➢ Syndicated loans

  • Thomson Reuters LPC DealScan
Thomson Reuters LPC DealScan

Also known as Loan Pricing Corporation Deal Scan

Global syndicated bank loan market

Detailed historical (origination) information on loan pricing, contracts details, terms and conditions on over 240,000 transactions since 1988

Information includes borrower, lender, purpose, amount, fees, and covenants.

Carey and Hrycray (1999) estimated that DealScan loans covered between half and three-quarters of the volume for outstanding commercial and industrial loans in the US

For additional insights about DealScan watch e-learning video from WRDS:

http://wrds-web.wharton.upenn.edu/wrds/E-Learning/_000Video/Overview_of_Dealscan/index.cfm
Thomson Reuters LPC DealScan

➢ **Company**
  - Company information on all borrowers & lenders
  - Main identifier: companyID
  - No historical information
  - Contains the ultimate parent of each company

➢ **Package**
  - Details on each loan contract or deal (named package)
  - Uniquely identified by packageID
  - One package may include several facilities (groups facilities into each deal)
  - Information on deal status: (completed 95%, noinfo 1.5%, in process, closed, cancelled, suspended, rumour, etc.)

➢ **Facility**
  - Details the information provided in the package file
  - Uniquely identified by facilityID
  - Contains packageID, BorrowerCompanyID
Thomson Reuters LPC DealScan

- **Lenders**
  - Company information on the lenders & participants in each facilityID
  - Contains the LenderRole

- **Current Facility Pricing**
  - Details on loan pricing: fees, base rates and spreads

- **DealScan-Compustat linking table** (mostly for US firms)
  - Prof. Michael Roberts
  - gvkey
Contents coverage

➢ Others:

• Eventus

• Datastream

• WorldBank WDI

• IMF Macroeconomic and financial data
Eventus

Performs event studies using CRSP market data

Text file with 2 columns:
- Col (1) PERMNO or CUSIP
- Col (2) Date
Web Query in detail (Eventus)

**Step 2: Choose Market Indices and Benchmark Options.**

**Market Index**
- CRSP Equity Weighted
- CRSP Value Weighted
- CRSP Equity Weighted + Value Weighted
- CRSP Equity Weighted + S&P500
- Exclude Dividends (NDC/DIVX)

**Additional Benchmark Options**
(MM - market model is selected by default)
- No Market Model (NOMM)
- Market-Adjusted Returns (MAR)
- No Market Adjustment (RAW)
- Comparison Period Mean-Adjusted Returns (CP)

**Step 3: Choose Estimation Options.**

**Estimation Period:**
- End Before Event Date (EST)
- Minimum Estimation Length (MINESTN)
- Maximum Estimation Length (EStLEN)
- POOL

**Autodate**
- None

**Estimate Method**
- OLS
- EGARCH
- GARCH
- ScholesWilliams

---

**Market Model:**
- Using CRSP Value Weighted
- Parameter Estimation ($\alpha$, $\beta$) using pre-defined window (-246,-46)
- Cumulative abnormal return window (-30,+30)

**Alternative Benchmark: Market-Adjusted Returns (MAR, $\beta=1$)**

**Step 4: Choose windows to search.**

**Event Period:**
- PRE 30
- POST 30

- Allow OVERLAP with estimation period

**Alternative Windows**

<table>
<thead>
<tr>
<th>#</th>
<th>Begin</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-30</td>
<td>-2</td>
</tr>
<tr>
<td>2</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>30</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Step 5: Choose tests to search.**

*How does this work?*
### Market Adjusted Returns, Value Weighted Index

<table>
<thead>
<tr>
<th>Days</th>
<th>N</th>
<th>Mean Cumulative Abnormal Return</th>
<th>Precision Weighted CAAR</th>
<th>Positive: Uncorrected Patell Z</th>
<th>Portfolio Time-Series (CDA) t</th>
<th>Generalized Sign Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>(-30,-2)</td>
<td>18</td>
<td>-1.09%</td>
<td>-1.17%</td>
<td>6:12</td>
<td>-0.488</td>
<td>-0.169</td>
</tr>
<tr>
<td>(-1,+1)</td>
<td>18</td>
<td>2.00%</td>
<td>1.13%</td>
<td>10:8</td>
<td>1.437$</td>
<td>0.966</td>
</tr>
<tr>
<td>(+1,+10)</td>
<td>18</td>
<td>5.93%</td>
<td>2.98%</td>
<td>10:8</td>
<td>2.100*</td>
<td>1.567$</td>
</tr>
<tr>
<td>(+2,+30)</td>
<td>18</td>
<td>5.18%</td>
<td>2.33%</td>
<td>11:7</td>
<td>0.954</td>
<td>0.804</td>
</tr>
</tbody>
</table>

The symbols $, *, **, and *** denote statistical significance at the 0.10, 0.05, 0.01 and 0.001 levels, respectively, using a generic one-tail test. The symbols (,< or,), etc. correspond to $,* and show the direction and significance of a generic one-tail generalized sign test.

### Market Model Abnormal Returns, Value Weighted Index

<table>
<thead>
<tr>
<th>Days</th>
<th>N</th>
<th>Mean Cumulative Abnormal Return</th>
<th>Precision Weighted CAAR</th>
<th>Positive: Uncorrected Patell Z</th>
<th>Portfolio Time-Series (CDA) t</th>
<th>Generalized Sign Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>(-30,-2)</td>
<td>18</td>
<td>-2.26%</td>
<td>-1.64%</td>
<td>6:12</td>
<td>-0.677</td>
<td>-0.352</td>
</tr>
<tr>
<td>(-1,+1)</td>
<td>18</td>
<td>1.61%</td>
<td>1.04%</td>
<td>9:9</td>
<td>1.322$</td>
<td>0.778</td>
</tr>
<tr>
<td>(+1,+10)</td>
<td>18</td>
<td>5.39%</td>
<td>2.69%</td>
<td>10:8</td>
<td>1.889*</td>
<td>1.427$</td>
</tr>
<tr>
<td>(+2,+30)</td>
<td>18</td>
<td>4.65%</td>
<td>2.57%</td>
<td>10:8</td>
<td>1.046</td>
<td>0.723</td>
</tr>
</tbody>
</table>

The symbols $, *, **, and *** denote statistical significance at the 0.10, 0.05, 0.01 and 0.001 levels, respectively, using a generic one-tail test. The symbols (,< or,), etc. correspond to $,* and show the direction and significance of a generic one-tail generalized sign test.
Outside WRDS... (Eikon)

✓ Thomson Reuters Eikon - DataStream/WorldScope

Tools to request Static Data and Time-Series Data

Opens excel layout useful to download huge datasets (several lists and data fields)

Datastream Navigator to explore data

Tool to create a single list of securities using identifiers (preferred key: datastream codes)
Outside WRDS... (Eikon)

Static Request

<table>
<thead>
<tr>
<th>Series/Lists</th>
<th>P-PTC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data/Type</td>
<td>ENAME.ISIN.DSCD</td>
</tr>
</tbody>
</table>

Find Series

Refine Search

Category: Portugal Telecom

Results for Portugal Telecom (Silicon)

<table>
<thead>
<tr>
<th>Code</th>
<th>Name</th>
<th>ISIN</th>
<th>Datostream Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>PT</td>
<td>PORTUGAL TELECOM SGPS</td>
<td>PTPTGAM0009</td>
<td>152311</td>
</tr>
</tbody>
</table>
Outside WRDS... (Eikon)

Time-Series Request

[Image of Time Series Request dialog box with selected fields highlighted]
Create List of Securities

1. Click on "Lists (Create from Range)" in the top menu.
2. Select "Datastream List" for the list type.
3. Choose "Mnemonic (L#)" as the mnemonic.
4. Click "OK" to confirm.
5. In the Time Series Request, select the series list and set the start and end dates.
6. Choose the desired options and click "Submit".
7. A message will appear indicating that the list has been successfully uploaded.
Constituent Lists of Securities

Market Indices and Research Lists

Constituent Lists (323,683)

Credit Default Swaps (82,257)
Economics (5,841,175)
Equities (240,780)
Equity Indices (344,915)
Exchange Rates (10,997)
Futures (268,661)
Interest Rates (19,950)
Investment Trusts (2,024)
Options (10,228,993)
Unit Trusts (422,856)
User Portfolios & Lists (47)
Warrants (2,359,552)
Converting data to US$ or US$ Cents.

Outside WRDS... (Eikon)
Request Table Tool

![Datastream Interface]

- **Process Table**: YES
- **Format**: TSL
- **Series Lookup**: RCTM
- **Datastream Code**: WC08001
- **Update**: June 30, 2020
- **Start Date**: 12/31/2010
- **End Date**: 12/31/2017
- **Freq**: Monthly

### Data Destination

#### Name

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Apple Inc - Market Capitalization</td>
<td>WC08001</td>
<td>992816</td>
<td>296455299</td>
<td>376535168</td>
<td>375351685</td>
<td>376535168</td>
<td>376535168</td>
<td>376535168</td>
</tr>
<tr>
<td>Microsoft Corp - Market Capitalization</td>
<td>WC08001</td>
<td>713463</td>
<td>241923680</td>
<td>217440860</td>
<td>217440860</td>
<td>217440860</td>
<td>217440860</td>
<td>217440860</td>
</tr>
<tr>
<td>Exxon Mobil Corp - Market Capitalization</td>
<td>WC08001</td>
<td>900338</td>
<td>364054640</td>
<td>401263040</td>
<td>401263040</td>
<td>401263040</td>
<td>401263040</td>
<td>401263040</td>
</tr>
<tr>
<td>Johnson &amp; Johnson - Market Capitalization</td>
<td>WC08001</td>
<td>912212</td>
<td>169351299</td>
<td>179637265</td>
<td>179637265</td>
<td>179637265</td>
<td>179637265</td>
<td>179637265</td>
</tr>
<tr>
<td>Wal-Mart Stores Inc - Market Capitalization</td>
<td>WC08001</td>
<td>915648</td>
<td>189617680</td>
<td>204259960</td>
<td>204259960</td>
<td>204259960</td>
<td>204259960</td>
<td>204259960</td>
</tr>
<tr>
<td>Wells Fargo &amp; Co - Market Capitalization</td>
<td>WC08001</td>
<td>906195</td>
<td>163073157</td>
<td>145037577</td>
<td>145037577</td>
<td>145037577</td>
<td>145037577</td>
<td>145037577</td>
</tr>
</tbody>
</table>
Outside WRDS... (Eikon)

Other Data Categories...

Example: Republic of Portugal 5 Yrs. CDS Spread
Outside WRDS... (WDI)

WDI

- World Bank collection of development indicators
IMF

- International Monetary Fund (Macroeconomic and financial data)
  - www.data.imf.org
### IMF DATA

#### ACCESS TO MACROECONOMIC & FINANCIAL DATA

**Historical Public Debt Database (HPDD)**

To display data please select the desired Time, Country, and Indicator on the left. For detailed help documents and video tutorials please use “How to Use Query” in the IMF knowledge repository.

#### Historical Public Debt (HPDD)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Canada</td>
<td>76.18</td>
<td>72.14</td>
<td>70.87</td>
<td>70.13</td>
<td>66.84</td>
<td>67.83</td>
<td>79.28</td>
<td>81.10</td>
<td>81.51</td>
<td>84.84</td>
<td>86.0</td>
</tr>
<tr>
<td>China</td>
<td>26.57</td>
<td>26.17</td>
<td>26.10</td>
<td>25.38</td>
<td>29.04</td>
<td>27.00</td>
<td>32.56</td>
<td>33.09</td>
<td>33.09</td>
<td>34.02</td>
<td>36.8</td>
</tr>
<tr>
<td>France</td>
<td>64.21</td>
<td>65.74</td>
<td>67.21</td>
<td>64.45</td>
<td>64.40</td>
<td>68.06</td>
<td>79.00</td>
<td>81.70</td>
<td>85.21</td>
<td>89.58</td>
<td>92.4</td>
</tr>
<tr>
<td>Germany</td>
<td>62.92</td>
<td>64.69</td>
<td>66.91</td>
<td>66.33</td>
<td>63.54</td>
<td>64.93</td>
<td>72.43</td>
<td>81.00</td>
<td>78.31</td>
<td>79.52</td>
<td>77.0</td>
</tr>
<tr>
<td>Greece</td>
<td>101.46</td>
<td>102.87</td>
<td>107.39</td>
<td>103.57</td>
<td>103.10</td>
<td>109.42</td>
<td>126.74</td>
<td>146.25</td>
<td>172.10</td>
<td>159.56</td>
<td>177.6</td>
</tr>
<tr>
<td>Ireland</td>
<td>30.94</td>
<td>29.63</td>
<td>27.35</td>
<td>24.81</td>
<td>25.01</td>
<td>44.34</td>
<td>61.68</td>
<td>86.30</td>
<td>109.61</td>
<td>119.48</td>
<td>119.4</td>
</tr>
<tr>
<td>Italy</td>
<td>100.48</td>
<td>100.09</td>
<td>101.94</td>
<td>102.56</td>
<td>99.78</td>
<td>102.39</td>
<td>112.52</td>
<td>115.38</td>
<td>116.50</td>
<td>123.34</td>
<td>129.0</td>
</tr>
<tr>
<td>Japan</td>
<td>169.57</td>
<td>180.66</td>
<td>186.44</td>
<td>186.00</td>
<td>183.01</td>
<td>191.81</td>
<td>210.25</td>
<td>215.82</td>
<td>231.63</td>
<td>238.01</td>
<td>244.4</td>
</tr>
<tr>
<td>Portugal</td>
<td>55.70</td>
<td>57.46</td>
<td>62.53</td>
<td>61.62</td>
<td>68.44</td>
<td>71.67</td>
<td>83.61</td>
<td>96.18</td>
<td>111.39</td>
<td>126.21</td>
<td>129.0</td>
</tr>
<tr>
<td>Spain</td>
<td>47.64</td>
<td>45.26</td>
<td>42.28</td>
<td>38.91</td>
<td>35.51</td>
<td>39.40</td>
<td>52.70</td>
<td>60.07</td>
<td>69.46</td>
<td>85.41</td>
<td>93.6</td>
</tr>
<tr>
<td>Switzerland</td>
<td>59.45</td>
<td>60.77</td>
<td>58.26</td>
<td>51.67</td>
<td>49.53</td>
<td>49.40</td>
<td>47.34</td>
<td>46.13</td>
<td>46.03</td>
<td>46.63</td>
<td>46.2</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>35.76</td>
<td>38.66</td>
<td>40.01</td>
<td>41.02</td>
<td>42.21</td>
<td>50.27</td>
<td>64.20</td>
<td>75.74</td>
<td>81.32</td>
<td>84.82</td>
<td>86.0</td>
</tr>
<tr>
<td>United States</td>
<td>58.52</td>
<td>65.49</td>
<td>64.89</td>
<td>63.64</td>
<td>64.01</td>
<td>72.85</td>
<td>86.03</td>
<td>94.73</td>
<td>98.99</td>
<td>102.50</td>
<td>104.4</td>
</tr>
</tbody>
</table>
IMF

**IMF DATA** ACCESS TO MACROECONOMIC & FINANCIAL DATA

Monetary and Financial Statistics (MFS)

The MFS database contains the aggregated surveys covering i) the Central Bank, ii) Depository Corporations, and iii) Other Financial Corporations. The key macroeconomic aggregates in this dataset include i) monetary base and broad money, ii) credit aggregates (including credit to the private sector), and iii) foreign assets and liabilities. The statistics are based on the Monetary and Financial Statistics Manual and Compilation Guide (Manual).

Detailed monetary statistics based on the standardized report forms reflecting the conceptual framework of the above Manual and its predecessors.

Reporters Map

- SRF Countries
- Non-SRF Countries
- Non Reporters

Course contents:

➢ Section II

• Brief introduction to SAS and SQL
• Data management in SAS Studio
  ✓ I/O (Input-Output) files
  ✓ Querying data
  ✓ Aggregating/Summarizing data
  ✓ Merging datasets
  ✓ Fuzzy matching
Accessing WRDS

UNIX – SAS Studio

- SAS studio is a **friendly visual interface** to the SAS server running on WRDS
- It looks like the SAS interface on the PC but has remote access to WRDS data like on UNIX
- All standard browsers are supported (Firefox, Google Chrome, Internet Explorer, Safari)
Brief introduction to SAS

SAS Studio on WRDS

SAS Studio is a web application that lets you access SAS through your browser.

How It Works

SAS Studio is an interface to the SAS server running on the WRDS environment. It complements SAS on Unix, SAS Connect on a PC, and the SAS Share interface to R, Matlab, and other languages. Since a new connection is established, you will be prompted to enter your password.

It looks like the SAS interface found on the PC, but has direct access to WRDS data like on Unix.

All standard browsers are supported, including Firefox, Chrome, Internet Explorer, and Safari.

Who should use it?

At this time, SAS Share is available to Faculty, staff, PhD students, and research assistants.

People using a Mac and those who happen to be at a computer that does not have PC SAS installed will find SAS Share very convenient. Unix users looking for a more friendly interface may also find it useful.

Programming With WRDS

- SAS Programming
- MATLAB Programming
- R Programming
- Python Programming
- Other Languages: C, Fortran
Why SAS?

➢ **WRDS platform** is built under a **SAS/SQL backend**.

➢ **SAS** doesn’t load the entire dataset to your PC RAM. It works sequentially in the dataset which is stored in the disk. Good at **handling large datasets**.

➢ **SAS** allows you to work remotely in the **WRDS Server**.

  ▪ You don’t need to have a super laptop even when working with big datasets.

  ▪ You don’t need to download all datasets to your PC. You can work in the server and then just download the final output file.

  ▪ **WRDS datasets** and **sample programs/research applications** are in **SAS**.

➢ Other Options: **Stata**, **R**, **Matlab**, **Python**, etc.. **R** is the Open source counterpart of Matlab or SAS. The number of researchers using it is expanding fast.
User-friendly interface

Create new folder “bootcamp” in your home directory
Hands-on example:

Hands-on example:


Filtering in the Table View

- Select **CRSP** in the left navigation menu under **SAS libraries**
- Inside **CRSP** click on **MSF**
- Select the columns **permno, date, prc, ret**
- Filter the table to get the results needed
(1) Filtering in the Table View

Filtering in the Table View

SAS Studio

Filtering in the Table View

Filter Table Rows

Clear Filter

permno in (14541, 11850, 13928) and year(date) eq 2014;

Total rows: 4152033 Total columns: 21 Filtered rows: 36

<table>
<thead>
<tr>
<th>PERMNO</th>
<th>DATE</th>
<th>PRC</th>
<th>RET</th>
</tr>
</thead>
<tbody>
<tr>
<td>11850</td>
<td>20140131</td>
<td>92.1600</td>
<td>-0.03226</td>
</tr>
<tr>
<td>11850</td>
<td>20140228</td>
<td>95.2700</td>
<td>0.051432</td>
</tr>
<tr>
<td>11850</td>
<td>20140331</td>
<td>97.6800</td>
<td>0.011646</td>
</tr>
<tr>
<td>11850</td>
<td>20140430</td>
<td>102.4100</td>
<td>0.048423</td>
</tr>
<tr>
<td>11850</td>
<td>20140530</td>
<td>101.5300</td>
<td>-0.011620</td>
</tr>
<tr>
<td>11850</td>
<td>20140630</td>
<td>105.6900</td>
<td>0.001492</td>
</tr>
<tr>
<td>11850</td>
<td>20140731</td>
<td>99.8400</td>
<td>-0.127820</td>
</tr>
<tr>
<td>11850</td>
<td>20140830</td>
<td>98.4500</td>
<td>0.012230</td>
</tr>
<tr>
<td>11850</td>
<td>20140930</td>
<td>94.0600</td>
<td>-0.054394</td>
</tr>
<tr>
<td>11850</td>
<td>20141031</td>
<td>96.7100</td>
<td>0.022833</td>
</tr>
<tr>
<td>11850</td>
<td>20141120</td>
<td>99.5400</td>
<td>-0.055664</td>
</tr>
<tr>
<td>11850</td>
<td>20141231</td>
<td>92.4500</td>
<td>0.021036</td>
</tr>
<tr>
<td>13928</td>
<td>20140131</td>
<td>64.8500</td>
<td>-0.096760</td>
</tr>
<tr>
<td>13928</td>
<td>20140228</td>
<td>65.5000</td>
<td>0.031486</td>
</tr>
<tr>
<td>13928</td>
<td>20140331</td>
<td>70.2600</td>
<td>0.057638</td>
</tr>
<tr>
<td>13928</td>
<td>20140430</td>
<td>74.2100</td>
<td>0.055290</td>
</tr>
</tbody>
</table>
(2) Program a Query

```sas
libname bootcamp '/home/unlpt/pmpires/bootcamp';
proc sql;
create table bootcamp.myexample1 as
  select permno, date, prc, ret
  from crsp.msf
where permno in (14541, 11850, 13928) and year(date) eq 2014;
quit;
```
(2) Program a Query (cont.)

```
*Program 1*

**Results**

NOTE: Library BOOTCAMP was successfully assigned as follows:
Engine:   WS
Physical Name: /home/unlpt/pmires/bootcamp

59    proc sql;
60    create table bootcamp.myexample1 as
61    select permno, data, ppc, rct
62    from crsp.msf
63    where permno in (11850, 11850, 13928) and year(data) eq 2014;
NOTE: Table BOOTCAMP.MYEXAMPLE1 created, with 36 rows and 4 columns.

65    quit;
NOTE: PROCEDURE SQL used (Total process time):
      real time   0.08 seconds
      cpu time   0.01 seconds

67    proc print datat=bootcamp.myexample1 nodots label;
68    run;
NOTE: There were 36 observations read from the data set BOOTCAMP.MYEXAMPLE1.
NOTE: PROCEDURE PRINT used (Total process time):
      real time   0.11 seconds
      cpu time   0.10 seconds
```

<table>
<thead>
<tr>
<th>PERMNO</th>
<th>Date of Observation</th>
<th>Price or Bid/Ask Average</th>
<th>Returns</th>
</tr>
</thead>
<tbody>
<tr>
<td>11850</td>
<td>20140131</td>
<td>92.16000</td>
<td>-0.089328</td>
</tr>
<tr>
<td>11850</td>
<td>20140228</td>
<td>96.27000</td>
<td>0.051432</td>
</tr>
<tr>
<td>11850</td>
<td>20140331</td>
<td>97.65000</td>
<td>0.014846</td>
</tr>
<tr>
<td>11850</td>
<td>20140430</td>
<td>102.41000</td>
<td>0.048423</td>
</tr>
<tr>
<td>11850</td>
<td>20140530</td>
<td>100.53000</td>
<td>-0.011620</td>
</tr>
<tr>
<td>11850</td>
<td>20140630</td>
<td>100.60000</td>
<td>0.001462</td>
</tr>
<tr>
<td>11850</td>
<td>20140731</td>
<td>98.89000</td>
<td>-0.017282</td>
</tr>
<tr>
<td>11850</td>
<td>20140829</td>
<td>99.44000</td>
<td>0.012230</td>
</tr>
<tr>
<td>11850</td>
<td>20140930</td>
<td>94.03000</td>
<td>-0.054394</td>
</tr>
<tr>
<td>11850</td>
<td>20141031</td>
<td>98.71000</td>
<td>0.028283</td>
</tr>
<tr>
<td>11850</td>
<td>20141128</td>
<td>90.54000</td>
<td>-0.056664</td>
</tr>
<tr>
<td>11850</td>
<td>20141231</td>
<td>92.45000</td>
<td>0.021096</td>
</tr>
<tr>
<td>13928</td>
<td>20140131</td>
<td>64.06000</td>
<td>-0.080670</td>
</tr>
<tr>
<td>13928</td>
<td>20140228</td>
<td>66.50000</td>
<td>0.034348</td>
</tr>
<tr>
<td>13928</td>
<td>20140331</td>
<td>70.35000</td>
<td>0.057965</td>
</tr>
<tr>
<td>13928</td>
<td>20140430</td>
<td>74.31000</td>
<td>0.050290</td>
</tr>
<tr>
<td>13928</td>
<td>20140530</td>
<td>79.94000</td>
<td>0.085049</td>
</tr>
<tr>
<td>13928</td>
<td>20140630</td>
<td>85.73000</td>
<td>0.072429</td>
</tr>
<tr>
<td>13928</td>
<td>20140731</td>
<td>82.50000</td>
<td>-0.029161</td>
</tr>
<tr>
<td>13928</td>
<td>20140829</td>
<td>81.22000</td>
<td>-0.015515</td>
</tr>
<tr>
<td>13928</td>
<td>20140930</td>
<td>76.52000</td>
<td>-0.057868</td>
</tr>
<tr>
<td>13928</td>
<td>20141031</td>
<td>72.16000</td>
<td>-0.047660</td>
</tr>
<tr>
<td>13928</td>
<td>20141129</td>
<td>69.07000</td>
<td>-0.084260</td>
</tr>
<tr>
<td>13928</td>
<td>20141231</td>
<td>69.06000</td>
<td>0.045265</td>
</tr>
<tr>
<td>14541</td>
<td>20140131</td>
<td>110.30000</td>
<td>-0.100317</td>
</tr>
<tr>
<td>14541</td>
<td>20140228</td>
<td>112.03000</td>
<td>0.04103</td>
</tr>
<tr>
<td>14541</td>
<td>20140331</td>
<td>118.91000</td>
<td>0.031041</td>
</tr>
<tr>
<td>14541</td>
<td>20140430</td>
<td>125.52000</td>
<td>0.055588</td>
</tr>
<tr>
<td>14541</td>
<td>20140530</td>
<td>122.79000</td>
<td>-0.013225</td>
</tr>
<tr>
<td>14541</td>
<td>20140630</td>
<td>130.54000</td>
<td>0.063197</td>
</tr>
</tbody>
</table>
(2) Program a Query (cont.) - Download a file
Upload a file

➢ Upload a file from your PC to WRDS

1. Select the folder in SAS Studio.
2. Click on "Upload" to select files.
3. Navigate to the file on your PC.
4. Enter the file path in the upload field.
5. Select the file you want to upload and click "Choose Files".
More on using SAS Studio...

➢ Defining a SAS Library

**New Library**

To create a library for this session, specify these values:

- **Name:** ppiex
- **Path:** /home/unlpt/ppire/footcamp
- **Options:** LIBNAME options (separated by spaces)

- [ ] Re-create this library at start-up (adds the library to the SAS autoexec file)

**SAS Studio**

- **Folders**
- **Common Tasks**
- **Code Snippets**
- **SAS Libraries**
  - [ ] MRKTSAMP
  - [ ] MSFANLY
  - [ ] MSFWST
  - [ ] NASTRAQ
  - [ ] OHTRIAL
  - [ ] OPTIONM
  - [ ] OTC
  - [ ] PACAP
  - [ ] PEX
  - [ ] PPIEX
  - [ ] PPIEXM
  - [ ] PPIEXSAMP
  - [ ] PUBLIC
  - [ ] PWI
- **File Shortcuts**
Brief introduction to SAS

Introduction

➢ A Table is a two-dimensional representation of data consisting of columns and rows

<table>
<thead>
<tr>
<th></th>
<th>name</th>
<th>year</th>
<th>score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>John</td>
<td>2012</td>
<td>55</td>
</tr>
<tr>
<td>2</td>
<td>John</td>
<td>2013</td>
<td>65</td>
</tr>
<tr>
<td>3</td>
<td>John</td>
<td>2014</td>
<td>70</td>
</tr>
<tr>
<td>4</td>
<td>Pedro</td>
<td>2012</td>
<td>75</td>
</tr>
<tr>
<td>5</td>
<td>Pedro</td>
<td>2013</td>
<td>80</td>
</tr>
<tr>
<td>6</td>
<td>Pedro</td>
<td>2014</td>
<td>90</td>
</tr>
<tr>
<td>7</td>
<td>Marta</td>
<td>2012</td>
<td>75</td>
</tr>
<tr>
<td>8</td>
<td>Marta</td>
<td>2013</td>
<td>65</td>
</tr>
<tr>
<td>9</td>
<td>Marta</td>
<td>2014</td>
<td>80</td>
</tr>
</tbody>
</table>

Terminology

<table>
<thead>
<tr>
<th>Data Processing</th>
<th>SAS</th>
<th>SQL</th>
</tr>
</thead>
<tbody>
<tr>
<td>File</td>
<td>Dataset</td>
<td>Table</td>
</tr>
<tr>
<td>Record</td>
<td>Observation</td>
<td>Row</td>
</tr>
<tr>
<td>Field</td>
<td>Variable</td>
<td>Column</td>
</tr>
</tbody>
</table>
PROC SQL

➢ Use the “proc sql” instruction to turn SAS into an efficient SQL environment

```sql
proc sql;
[SQL Instruction]
quit;
```

Structure of a SQL Query

• SELECT columns (variables)
• FROM tables (datasets)
• WHERE row (observation) conditions that must be met
• GROUP BY summarize by these variables
• HAVING summary conditions that must be met
• ORDER BY sort by these columns
Brief introduction to SAS

**Querying data and storing results**

- By default the results of a query are displayed in the SAS output window

```sas
proc sql;
  select * from saslibrary.table;
quit;
```

- You can create a dataset (table) to store the output of your query

```sas
proc sql;
  create table table as
  select * from saslibrary.table;
quit;
```
Brief introduction to SAS

Sorting

```sql
select name, year, score
from work.scores
order by year, score desc;
```

Creating new variables

```sql
select name, year, score, 90 - score as distance
from work.scores
```

Duplicates

```sql
select distinct name
from work.scores
```
Conditional logic (select statement)

```sql
select name, year, score,
    case
        when score < 50 then 'Fail'
        else 'Pass'
    end as mark
from work.scores
```

Subsetting data

```sql
select name, year, score
from work.scores
where year = 2014

(...)

where year = 2014 and name = 'Pedro'
```
**Brief introduction to SAS**

**BETWEEN operator**

```sql
select name, year, score
from work.scores
where score between 80 and 90
```

**LIKE operator**

```sql
select name, year, score
from work.scores
where name like '%Ped%'
```

- Contains: ‘%Ped%’
- Starts with: ‘Ped%’
- Ends with: ‘%dro’

**IN operator**

```sql
select name, year, score
from work.scores
where name in ('Pedro', 'Marta')
```

- Strings in SQL are case sensitive
- Use lower and upper functions to convert strings to lower- or upper-case
### Standard comparison operators

<table>
<thead>
<tr>
<th>Operator</th>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>EQ</td>
<td>=</td>
<td>Equal to</td>
</tr>
<tr>
<td>NE</td>
<td>^= or &lt;&gt;</td>
<td>Not equal to</td>
</tr>
<tr>
<td>GE</td>
<td>&gt;=</td>
<td>Greater than or equal to</td>
</tr>
<tr>
<td>GT</td>
<td>&gt;</td>
<td>Greater than</td>
</tr>
<tr>
<td>LE</td>
<td>&lt;=</td>
<td>Less than or equal to</td>
</tr>
<tr>
<td>LT</td>
<td>&lt;</td>
<td>Less than</td>
</tr>
</tbody>
</table>
Summarizing or aggregating data

```sql
select year, avg(score) as avg_score
from work.scores
group by year
```

Subsetting with group by clauses

```sql
select name, avg(score) as avg_score
from work.scores
group by name
having avg(score) > 70
```
Subqueries

➢ A query expression nested as part of another query expression

```sql
select *
from address
where name in
    (select distinct name from scores
     where score ge 75 and year eq 2014);
```

Creating a table and sorting with SAS Data Step

```sas
data table1;
set scores (keep=name score);
where year=2014;
run;

proc sort data=table1;
by descending score;
run;
```
Relational database

- Data in tables can be logically related according to common keys (variables)

**Table: Scores**

<table>
<thead>
<tr>
<th>name</th>
<th>year</th>
<th>score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>John</td>
<td>55</td>
</tr>
<tr>
<td>2</td>
<td>John</td>
<td>65</td>
</tr>
<tr>
<td>3</td>
<td>John</td>
<td>70</td>
</tr>
<tr>
<td>4</td>
<td>Pedro</td>
<td>75</td>
</tr>
<tr>
<td>5</td>
<td>Pedro</td>
<td>80</td>
</tr>
<tr>
<td>6</td>
<td>Pedro</td>
<td>90</td>
</tr>
<tr>
<td>7</td>
<td>Marta</td>
<td>75</td>
</tr>
<tr>
<td>8</td>
<td>Marta</td>
<td>65</td>
</tr>
<tr>
<td>9</td>
<td>Marta</td>
<td>80</td>
</tr>
</tbody>
</table>

9 records

**Table: Address**

<table>
<thead>
<tr>
<th>name</th>
<th>city</th>
<th>zip</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Almada</td>
<td>5220</td>
</tr>
<tr>
<td>2</td>
<td>Lisbon</td>
<td>1500</td>
</tr>
<tr>
<td>3</td>
<td>Cascais</td>
<td>2850</td>
</tr>
</tbody>
</table>

3 records
Brief introduction to SAS

Relational database

➢ Data in tables can be logically related according to common keys (variables)

```sas
data scores;
input name $ year score;
datalines;
John 2012 55
John 2013 65
John 2014 70
Pedro 2012 75
Pedro 2013 80
Pedro 2014 90
Marta 2012 75
Marta 2013 65
Marta 2014 80
;
run;
```

```sas
data address;
input name $ city $ zip;
datalines;
John Almada 5220
Pedro Lisbon 1500
Marta Cascais 2850
;
run;
```
# Brief introduction to SAS

## Output Data

Table: WORK.SCORES

<table>
<thead>
<tr>
<th>Columns</th>
<th>Select all</th>
<th>name</th>
<th>year</th>
<th>score</th>
</tr>
</thead>
<tbody>
<tr>
<td>✔️</td>
<td>✔️</td>
<td>John</td>
<td>2012</td>
<td>55</td>
</tr>
<tr>
<td>✔️</td>
<td>✔️</td>
<td>John</td>
<td>2013</td>
<td>65</td>
</tr>
<tr>
<td>✔️</td>
<td>✔️</td>
<td>John</td>
<td>2014</td>
<td>70</td>
</tr>
<tr>
<td>✔️</td>
<td>✔️</td>
<td>Pedro</td>
<td>2012</td>
<td>75</td>
</tr>
<tr>
<td>✔️</td>
<td>✔️</td>
<td>Pedro</td>
<td>2013</td>
<td>80</td>
</tr>
<tr>
<td>✔️</td>
<td>✔️</td>
<td>Pedro</td>
<td>2014</td>
<td>80</td>
</tr>
<tr>
<td>✔️</td>
<td>✔️</td>
<td>Marta</td>
<td>2012</td>
<td>75</td>
</tr>
<tr>
<td>✔️</td>
<td>✔️</td>
<td>Marta</td>
<td>2013</td>
<td>65</td>
</tr>
<tr>
<td>✔️</td>
<td>✔️</td>
<td>Marta</td>
<td>2014</td>
<td>80</td>
</tr>
</tbody>
</table>

Table: WORK.ADDRESS

<table>
<thead>
<tr>
<th>Columns</th>
<th>Select all</th>
<th>name</th>
<th>city</th>
<th>zip</th>
</tr>
</thead>
<tbody>
<tr>
<td>✔️</td>
<td>✔️</td>
<td>John</td>
<td>Almada</td>
<td>5220</td>
</tr>
<tr>
<td>✔️</td>
<td>✔️</td>
<td>Pedro</td>
<td>Lisbon</td>
<td>1500</td>
</tr>
<tr>
<td>✔️</td>
<td>✔️</td>
<td>Marta</td>
<td>Cascais</td>
<td>2850</td>
</tr>
</tbody>
</table>
Joining Tables

➢ Cartesian Product: Each row of the first table is combined with every row from the second table (number of rows table 1 * number of rows table 2; can be huge!)

```
create table cartesian as
select *
from work.scores, work.address
order by name, year;
```

- 27 records (=9*3)
- Now, we have three duplicate records for each pair (name, year)

NOTE: The execution of this query involves performing one or more Cartesian product joins that can not be optimized.
Inner Join

- Returns the subset of rows from the first table that matches rows from the second table

```
create table inner as
select *
from scores as t1, address as t2
where t1.name = t2.name
order by t1.name, t1.year;

(...) from scores as t1
inner join address as t2
  on t1.name = t2.name
```

Use table alias to abbreviate table names
Brief introduction to SAS

**Left Outer Join**

- Returns all the rows from the first table and the fields matched with the second table

```sql
proc sql;
delete from address
where name eq 'John';
create table left as
select *
from scores as t1
Left join address as t2
  on t1.name = t2.name
order by t1.name, t1.year;
```

**Table: Address**

<table>
<thead>
<tr>
<th>name</th>
<th>city</th>
<th>zip</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pedro</td>
<td>Lisbon</td>
<td>1500</td>
</tr>
<tr>
<td>Marta</td>
<td>Cascais</td>
<td>2850</td>
</tr>
</tbody>
</table>

**Table: Scores**

<table>
<thead>
<tr>
<th>name</th>
<th>year</th>
<th>score</th>
<th>city</th>
<th>zip</th>
</tr>
</thead>
<tbody>
<tr>
<td>John</td>
<td>2012</td>
<td>55</td>
<td></td>
<td></td>
</tr>
<tr>
<td>John</td>
<td>2013</td>
<td>65</td>
<td></td>
<td></td>
</tr>
<tr>
<td>John</td>
<td>2014</td>
<td>70</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marta</td>
<td>2012</td>
<td>75</td>
<td>Cascais</td>
<td>2850</td>
</tr>
<tr>
<td>Marta</td>
<td>2013</td>
<td>80</td>
<td>Cascais</td>
<td>2850</td>
</tr>
<tr>
<td>Pedro</td>
<td>2012</td>
<td>75</td>
<td>Lisbon</td>
<td>1500</td>
</tr>
<tr>
<td>Pedro</td>
<td>2013</td>
<td>80</td>
<td>Lisbon</td>
<td>1500</td>
</tr>
<tr>
<td>Pedro</td>
<td>2014</td>
<td>90</td>
<td>Lisbon</td>
<td>1500</td>
</tr>
</tbody>
</table>
Full Outer Join

➢ Returns all matching and non-matching rows

create table full as

select coalesce (t1.name, t2.name) as name, t1.year, t1.score, t2.city, t2.zip
from scores as t1
full join address as t2
on t1.name = t2.name
order by calculated name, t1.year;
Brief introduction to SAS

data scores;
input name $ year $ score;
datalines;
John 2012 55
John 2013 65
John 2014 70
Pedro 2012 75
Pedro 2013 80
Pedro 2014 90
Marta 2012 75
Marta 2013 65
Marta 2014 80
;
run;

data address;
input name $ city $ zip;
datalines;
John Almada 5220
Pedro Lisbon 1500
Marta Cascais 2850
;
run;

proc sql;
create table inner as
select *
from scores as t1, address as t2
where t1.name = t2.name
order by t1.name, t1.year;

proc sql; delete from address where name eq 'John';

proc sql;
create table left as
select *
from scores as t1
left join address as t2
  on t1.name = t2.name
order by t1.name, t1.year;

proc sql;
create table full as
select coalesce (t1.name, t2.name) as name,
t1.year, t1.score, t2.city, t2.zip
from scores as t1
full join address as t2
  on t1.name = t2.name
order by calculated name, t1.year;