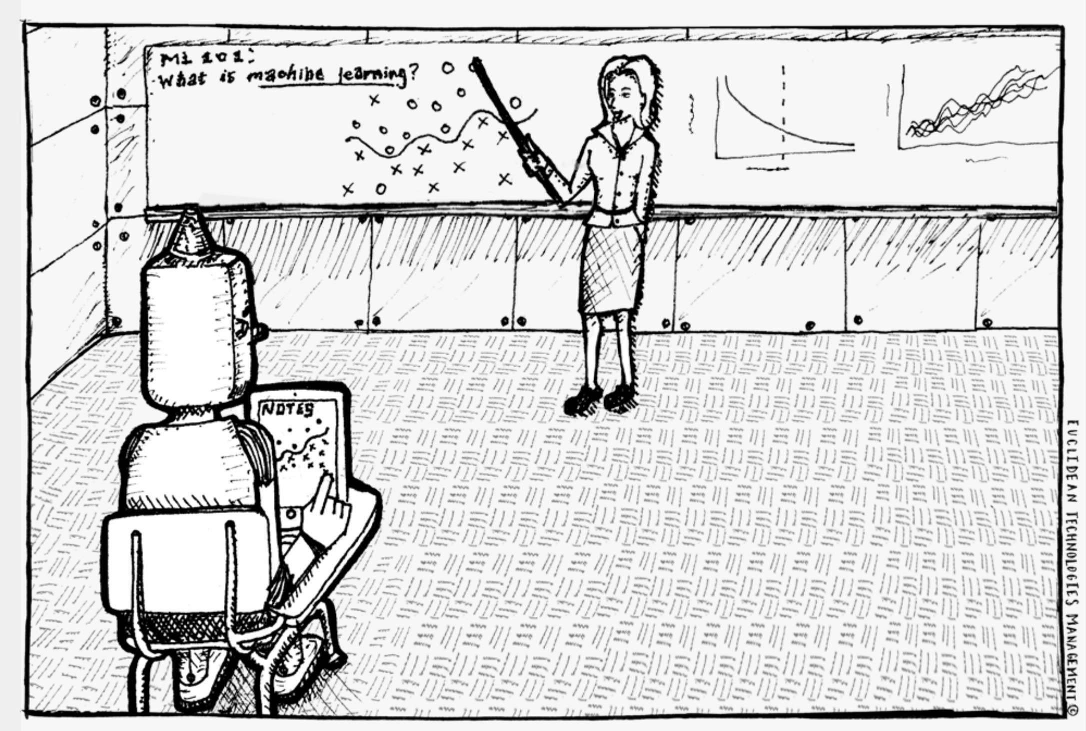
QF632: Financial Data Science

Course Project



Data Frame | Table

- Rows = number of observations
- Columns = number of features
- This is a **cross sectional** data frame (a single point in time)

	make	price	mpg	rep78	headroom	trunk	weight	length	turn	displacement	gear_ratio	foreign
1	AMC Concord	4,099	22	3	2.5	11	2,930	186	40	121	3.58	Domestic
2	AMC Pacer	4,749	17	3	3.0	11	3,350	173	40	258	2.53	Domestic
3	AMC Spirit	3,799	22	.	3.0	12	2,640	168	35	121	3.08	Domestic
4	Buick Century	4,816	20	3	4.5	16	3,250	196	40	196	2.93	Domestic
5	Buick Electra	7,827	15	4	4.0	20	4,080	222	43	350	2.41	Domestic
6	Buick LeSabre	5,788	18	3	4.0	21	3,670	218	43	231	2.73	Domestic
7	Buick Opel	4,453	26	.	3.0	10	2,230	170	34	304	2.87	Domestic
8	Buick Regal	5,189	20	3	2.0	16	3,280	200	42	196	2.93	Domestic
9	Buick Riviera	10,372	16	3	3.5	17	3,880	207	43	231	2.93	Domestic
10	Buick Skylark	4,082	19	3	3.5	13	3,400	200	42	231	3.08	Domestic
11	Cad. Deville	11,385	14	3	4.0	20	4,330	221	44	425	2.28	Domestic
12	Cad. Eldorado	14,500	14	2	3.5	16	3,900	204	43	350	2.19	Domestic
13	Cad. Seville	15,906	21	3	3.0	13	4,290	204	45	350	2.24	Domestic
14	Chev. Chevette	3,299	29	3	2.5	9	2,110	163	34	231	2.93	Domestic
15	Chev. Impala	5,705	16	4	4.0	20	3,690	212	43	250	2.56	Domestic
16	Chev. Malibu	4,504	22	3	3.5	17	3,180	193	31	200	2.73	Domestic
17	Chev. Monte Carlo	5,104	22	2	2.0	16	3,220	200	41	200	2.73	Domestic
18	Chev. Monza	3,667	24	2	2.0	7	2,750	179	40	151	2.73	Domestic
19	Chev. Nova	3,955	19	3	3.5	13	3,430	197	43	250	2.56	Domestic
20	Dodge Colt	3,984	30	5	2.0	8	2,120	163	35	98	3.54	Domestic
21	Dodge Diplomat	4,010	18	2	4.0	17	3,600	206	46	318	2.47	Domestic
22	Dodge Magnum	5,886	16	2	4.0	17	3,600	206	46	318	2.47	Domestic
23	Dodge St. Regis	6,342	17	2	4.5	21	3,740	220	46	225	2.94	Domestic
24	Ford Fiesta	4,389	28	4	1.5	9	1,800	147	33	98	3.15	Domestic
25	Ford Mustang	4,187	21	3	2.0	10	2,650	179	43	140	3.08	Domestic
26	Linc. Continental	11,497	12	3	3.5	22	4,840	233	51	400	2.47	Domestic
27	Linc. Mark V	13,594	12	3	2.5	18	4,720	230	48	400	2.47	Domestic

Variables

Enter filter text here

Name	Label
make	Make and M...
price	Price
mpg	Mileage (mpg)
rep78	Repair Recor...
headroom	Headroom (in.)
trunk	Trunk space...
weight	Weight (lbs.)
length	Length (in.)
turn	Turn Circle (...
displacement	Displacemen...
gear_ratio	Gear Ratio
foreign	Car type

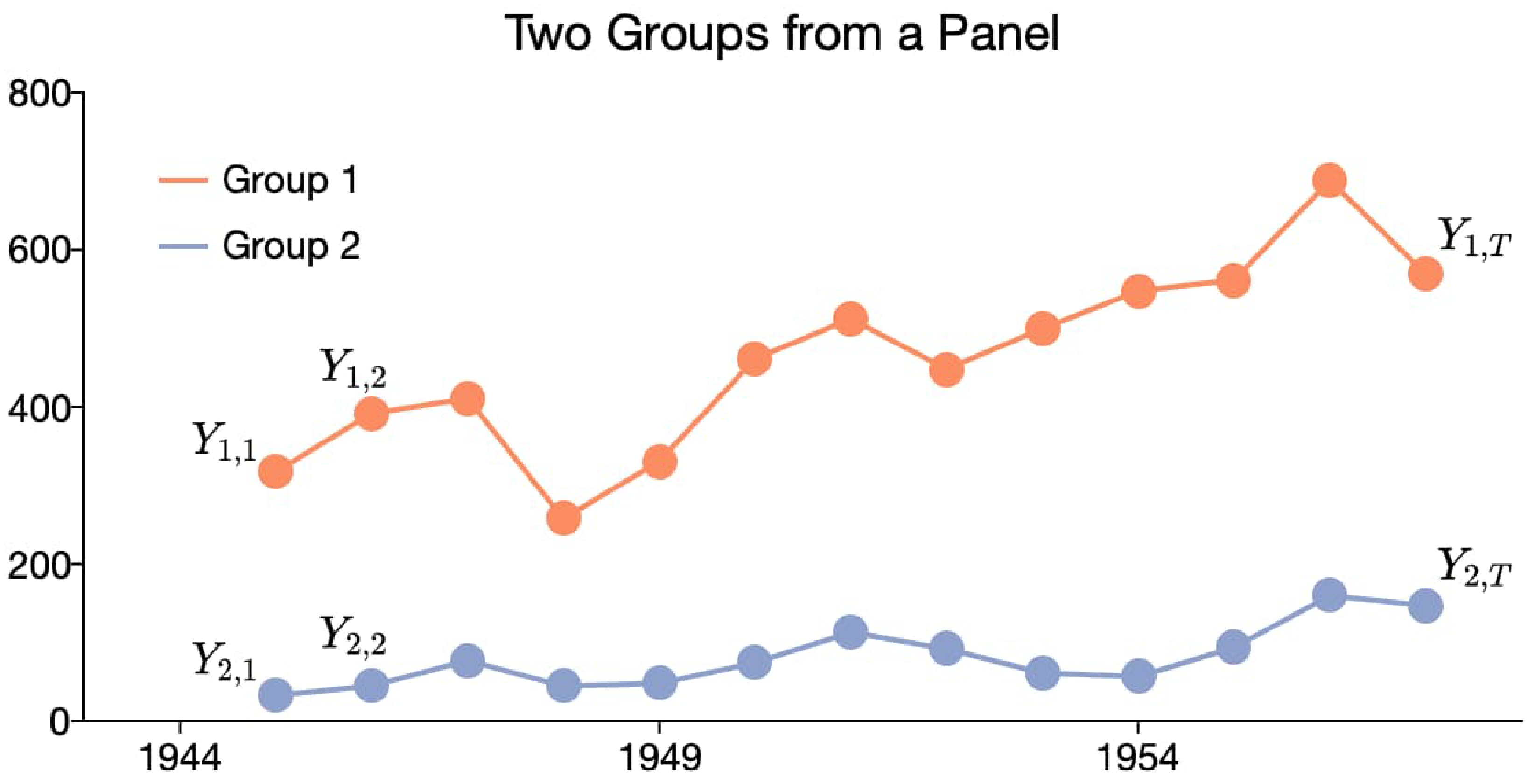
Properties

Variables	
Name	mpg
Label	Mileage (mpg)
Type	int
Format	%8.0g
Value Label	
Notes	
Data	
Filename	auto.dta
Label	1978 Automobile Data
Notes	1 note
Variables	12
Observations	74
Size	3.11K
Memory	64M

Categorical **Numerical** **Ordinal** **Numerical** **Categorical**

Time Series Data

- 2 time series data in a single data frame forms a panel data frame



Date	Y1	Y2
1944	300	50
1945	325	60
1946	330	75
1947	322	55
1948	323	55
1949	325	60
1950	328	65
1951	329	70
1952	335	68
1953	334	69
1954	332	71
1955	337	72
1956	339	73
1957	341	75
1958	342	76

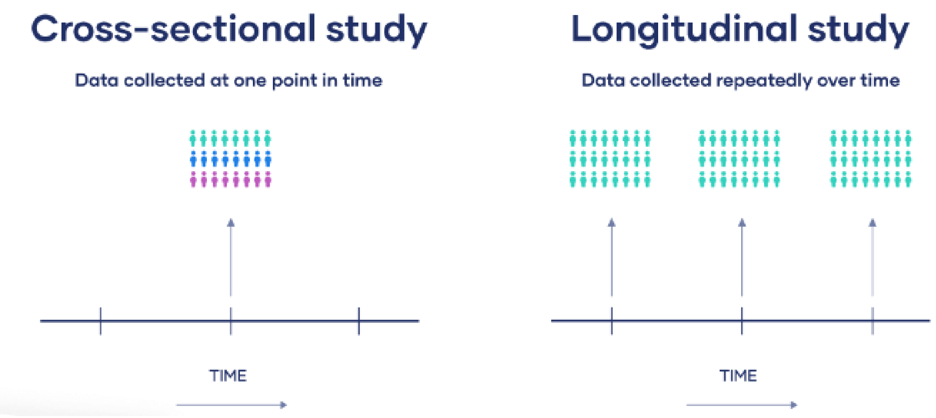
Panel Data

- Multiple cross sectional data frames ➔ Panel/Longitudinal data

```
> dt[date=="2015-01-03"]
     brandparent   freq              metric value         MEAN       date
   1:    24 Sevres weekly            comments     0   273.714894 2015-01-03
   2:    24 Sevres weekly      comments_pp_pxfw     0     3.750638 2015-01-03
   3:    24 Sevres weekly        interactions     0 19606.714894 2015-01-03
   4:    24 Sevres weekly interactions_pp_pxfw     0   285.982128 2015-01-03
   5:    24 Sevres weekly               likes     0 14308.795745 2015-01-03
  ---
2328:         Zegna weekly            pictures    54    28.493617 2015-01-03
2329:         Zegna weekly               posts    56    34.531915 2015-01-03
2330:         Zegna weekly              videos     2     6.038298 2015-01-03
2331:         Zegna weekly          videoviews     0 59898.974468 2015-01-03
2332:         Zegna weekly    videoviews_pp_pxfw     0  2195.305957 2015-01-03
```

+

```
> dt[date=="2015-01-10"]
     brandparent   freq              metric value         MEAN       date
   1:    24 Sevres weekly            comments     0   273.714894 2015-01-10
   2:    24 Sevres weekly      comments_pp_pxfw     0     3.750638 2015-01-10
   3:    24 Sevres weekly        interactions     0 19606.714894 2015-01-10
   4:    24 Sevres weekly interactions_pp_pxfw     0   285.982128 2015-01-10
   5:    24 Sevres weekly               likes     0 14308.795745 2015-01-10
  ---
2328:         Zegna weekly            pictures    38    28.493617 2015-01-10
2329:         Zegna weekly               posts    39    34.531915 2015-01-10
2330:         Zegna weekly              videos     1     6.038298 2015-01-10
2331:         Zegna weekly          videoviews     0 59898.974468 2015-01-10
2332:         Zegna weekly    videoviews_pp_pxfw     0  2195.305957 2015-01-10
```



Folding Wide Data Frames into Long Data Frames

- This is panel data
- **Wide** data frame ➔ a subject's repeated responses will be in a single row, and each response is in a separate column.

  65084 rows (observations) X 24 columns (features)

	brandparent	dt	freq	comments	comments_pf_pxfw	comments_pfp_pxfw	comments_pp_pxfw	followers	followgrowth	interactions	interactions_pf_pxfw	interactions_pfp_pxfw
1:	24 Sevres	1/10/2015	weekly	0	NA	NA	0.0	NA	NA	0	NA	NA
2:	24 Sevres	1/12/2019	weekly	504	0.00502	7.27e-05	7.3	NA	NA	61693	0.61400	0.00889857
3:	24 Sevres	1/13/2018	weekly	176	NA	NA	3.1	NA	NA	45665	NA	NA
4:	24 Sevres	1/14/2017	weekly	0	NA	NA	0.0	NA	NA	0	NA	NA
5:	24 Sevres	1/16/2016	weekly	0	NA	NA	0.0	NA	NA	0	NA	NA

65080:	Zegna	9/30/2018	monthly	9857	0.02033	3.99e-04	193.3	484944	0.03818833	703002	1.44966	0.02842463
65081:	Zegna	9/30/2018	quarterly	10741	0.02215	2.33e-04	113.1	484944	0.08117797	1086522	2.24051	0.02358432
65082:	Zegna	9/5/2015	weekly	567	NA	NA	18.3	NA	NA	46972	NA	NA
65083:	Zegna	9/8/2018	weekly	534	0.00114	7.11e-05	33.4	NA	NA	292791	0.62388	0.03899230
65084:	Zegna	9/9/2017	weekly	377	NA	NA	14.0	NA	NA	205965	NA	NA

	interactions_pp_pxfw	likes	likes_pf_pxfw	likes_pfp_pxfw	likes_pp_pxfw	pictures	posts	videos	videoviews	videoviews_pf_pxfw	videoviews_pfp_pxfw	videoviews_pp_pxfw
1:	0.0	0	NA	NA	0.0	0	0	0	0	NA	NA	0.0
2:	894.1	49368	0.49134	0.00712082	715.5	67	69	2	11821	0.11765	0.00170506	171.3
3:	815.4	18108	NA	NA	323.4	44	56	12	27381	NA	NA	488.9
4:	0.0	0	NA	NA	0.0	0	0	0	0	NA	NA	0.0
5:	0.0	0	NA	NA	0.0	0	0	0	0	NA	NA	0.0

65080:	13784.4	449481	0.92687	0.01817396	8813.4	39	51	12	243664	0.50246	0.00985212	4777.7
65081:	11437.1	537254	1.10787	0.01166177	5655.3	67	95	28	538527	1.11049	0.01168940	5668.7
65082:	1515.2	46405	NA	NA	1496.9	29	31	2	0	NA	NA	0.0
65083:	18299.4	51673	0.11010	0.00688153	3229.6	8	16	8	240584	0.51263	0.03203966	15036.5
65084:	7628.3	72999	NA	NA	2703.7	16	27	11	132589	NA	NA	4910.7

- **Long** data frame ➔ each row is one time point per subject. So each subject will have data in multiple rows. Any variables that don't change across time will have the same value in all the rows.

  802126 rows (observations) X 5 columns (features)

	brandparent	dt	freq	metric	value
1:	24 Sevres	1/3/2015	weekly	comments	0.00000000
2:	24 Sevres	1/3/2015	weekly	comments_pp_pxfw	0.00000000
3:	24 Sevres	1/3/2015	weekly	interactions	0.00000000
4:	24 Sevres	1/3/2015	weekly	interactions_pp_pxfw	0.00000000
5:	24 Sevres	1/3/2015	weekly	likes	0.00000000

802122:	Wrangler	6/30/2019	quarterly	followgrowth	NA
802123:	Yoox	6/30/2019	quarterly	followgrowth	0.23428595
802124:	Zalando	6/30/2019	quarterly	followgrowth	0.15409778
802125:	Zara	6/30/2019	quarterly	followgrowth	0.05843839
802126:	Zegna	6/30/2019	quarterly	followgrowth	0.03052609

Long Data Frames

- Long format allows data to be stored more densely and operations applied/scales more easily (group by), while the wide format has more explanatory power if tabular formats are required in a report (like and Excel spreadsheet) – but only one variable can be "displayed"

```
> dt[,MEAN:=mean(value,na.rm=TRUE),by=.(brandparent,freq,metric)]
> dt
         brandparent        dt      freq               metric      value         MEAN
      1:   24 Sevres  1/3/2015    weekly             comments 0.00000000 2.737149e+02
      2:   24 Sevres  1/3/2015    weekly     comments_pp_pxfw 0.00000000 3.750638e+00
      3:   24 Sevres  1/3/2015    weekly         interactions 0.00000000 1.960671e+04
      4:   24 Sevres  1/3/2015    weekly interactions_pp_pxfw 0.00000000 2.859821e+02
      5:   24 Sevres  1/3/2015    weekly                likes 0.00000000 1.430880e+04
     ---
802122:     Wrangler 6/30/2019 quarterly         followgrowth         NA          NaN
802123:         Yoox 6/30/2019 quarterly         followgrowth 0.23428595 2.242676e-01
802124:      Zalando 6/30/2019 quarterly         followgrowth 0.15409778 1.069618e-01
802125:         Zara 6/30/2019 quarterly         followgrowth 0.05843839 6.621934e-02
802126:        Zegna 6/30/2019 quarterly         followgrowth 0.03052609 6.263951e-02
```

- One reason for setting up the data in one format or the other is simply that different analyses require different set ups. If we filter the data frame to just one *brandparent*, and *freq* – a wide format is more easily visualizable

```
> dcast.data.table(dt[brandparent=="Zara" & freq=="quarterly"],dt~metric,value.var="value")
            dt> dcast.data.table(dt[brandparent=="Zara" & freq=="quarterly"],dt~metric,value.var="value")
            dt comments comments_pf_pxfw comments_pfp_pxfw comments_pp_pxfw followers followgrowth interactions interactions_pf_pxfw interactions_pfp_pxfw interactions_pp_pxfw    likes
 1: 12/31/2015    46829               NA                NA            384.7        NA           NA      7598354                   NA                    NA              63782.6  7485119
 2: 12/31/2016    37210               NA                NA            295.4        NA           NA     22307941                   NA                    NA             181232.3  8952192
 3: 12/31/2017    45369               NA                NA            269.3        NA           NA     33739332                   NA                    NA             203753.1 12653450
 4: 12/31/2018    64374          0.00187          1.23e-05            328.1  34434455   0.07546270     25503228              0.74266            0.00485673             131731.9 12936888
 5:  3/31/2015    45757               NA                NA            335.0        NA           NA      4936807                   NA                    NA              37774.5  4891050
 6:  3/31/2016    42036               NA                NA            407.0        NA           NA      9137324                   NA                    NA              90805.6  6472361
 7:  3/31/2017    29856               NA                NA            210.5        NA           NA     16251019                   NA                    NA             118186.2  9004335
 8:  3/31/2018    45215               NA                NA            299.5        NA           NA     25022939                   NA                    NA             173107.2 10342749
 9:  3/31/2019    52874          0.00149          1.00e-05            294.2  36664325   0.06475694     23521692              0.66266            0.00446447             130828.2 13214658
10:  6/30/2015    61174               NA                NA            428.5        NA           NA      7714268                   NA                    NA              55398.6  7653094
11:  6/30/2016    40467               NA                NA            353.3        NA           NA     16808882                   NA                    NA             149671.9  7033269
12:  6/30/2017    30775               NA                NA            240.2        NA           NA     19887031                   NA                    NA             158235.0 10527189
13:  6/30/2018    36080               NA                NA            234.2        NA           NA     25891947                   NA                    NA             167752.0  9134181
14:  6/30/2019    40771          0.00108          7.49e-06            230.3  38806929   0.05843839     16586502              0.44122            0.00304524              94337.1 11726792
15:  9/30/2015    51740               NA                NA            410.1        NA           NA      7232754                   NA                    NA              58847.2  7181014
16:  9/30/2016    67524               NA                NA            663.2        NA           NA     16397684                   NA                    NA             159065.6  7845239
17:  9/30/2017    31754               NA                NA            196.2        NA           NA     27870841                   NA                    NA             176114.2 10964878
18:  9/30/2018    44499          0.00141          1.07e-05            270.5  32018270           NA     20837124              0.65655            0.00501630             124650.1 10441578
```

Types of Data

- **Nominal/Categorical**
  Labels with no natural order
  Example: nationality, colour, gender, etc.
- **Ordinal**
  Labels where there is a natural order, but cannot perform any arithmetical operation
  Example: small/medium/large, primary/secondary/undergraduate/postgraduate, etc.
- **Discrete**
  Finite, whole numbers, cannot be fractionalized or decimalized
  Example: number of students in a class, days in a month, mobile phone number, etc.
- **Continuous**
  Measurably expressed in the form of a fractional/decimal number
  Example: height and weight, frequency spectrum, car speed, a period of time, etc.

- *Encoding*
  Special treatment for discrete, categorical data – use one hot encoding.
  This transforms nominal/categorical data into numerical features that allow further arithmetic manipulation, but still preserves the lack of an ordinal relationship between the variables.

Original Data

Team	Points
A	25
A	12
B	15
B	14
B	19
B	23
C	25
C	29

One-Hot Encoded Data

Team_A	Team_B	Team_C	Points
1	0	0	25
1	0	0	12
0	1	0	15
0	1	0	14
0	1	0	19
0	1	0	23
0	0	1	25
0	0	1	29

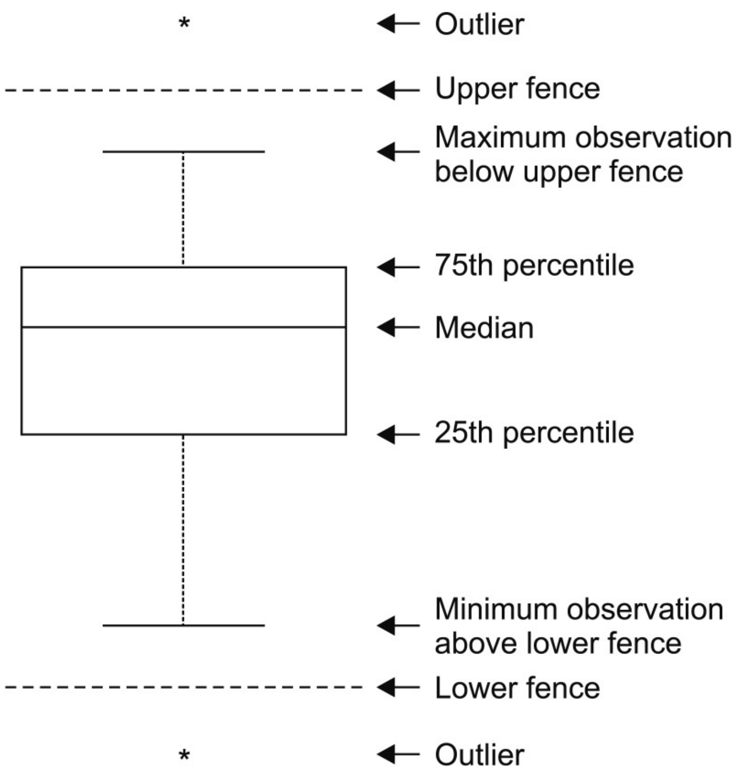
Data Cleansing: Missing Data

Types of missing values	Description	Possible causes
Missing completely at random	Missing data occur completely at random without being influenced by other data.	Consent withdrawal, omission of major exams, death, discontinued follow-up and serious adverse reactions.
Missing at random	Missing data occur at a specific time point in conjunction with participant dissatisfaction with study outcomes and ongoing participation	Refusal to continue measurements.
Not missing at random	Missing data occur when a patient who is not satisfied with study outcomes performs the required measurements on his own, before the scheduled measurement.	If a patient finds the results of self-measurement dissatisfactory in addition to dissatisfaction related to the study, the patient may refuse further measurements.

- Solution 1: Simply remove the offending observation
  - But this could introduce biases as there could be a systematical problem with the data gathering process that renders those data missing. Need to understand why a particular data point/field is missing
  - Not enough observations – every data point is precious. Simply discard missing data is not a reasonable practice, as valuable information may be lost and inferential power compromised
- Solution 2: Imputation
  - A better approach. Simple approach is to replace with mean/median by group.
  - A statistically more reliable method is to build a model to predict/impute that the missing value should have been.
  - Challenging as the dimensionality of the problem increases

Data Cleansing: Outliers

- **Identifying outliers**
  - Different methods can be used to identify outliers. For a normal distribution assumption, one of these methods measures the distance between a data point and the center of all data points to determine an outlier. The data points that do not fall within three SD of the mean are identified as outliers. However, this method is not considered appropriate because the mean and SD are statistically sensitive to the presence of outliers.
  - Alternatively, the median and quartile range are more useful because these statistics are less sensitive to outliers. In addition, box plots can be used to identify the outliers. In this box plot, any data that lies outside the upper or lower fence lines is considered outliers.
- **Treating outliers**
  - **Trimming:** Simple removal of offending observation. A data set that excludes outliers is first analyzed. The trimmed estimators such as mean decrease the variance in the data and cause a bias based on under- or overestimation. Given that the outliers are also observed values, excluding them from the analysis makes this approach inadequate for the treatment of outliers.
  - **Winsorization:** Threshold the offending observation (capping). This approach involves modifying the weights of outliers or replacing the values being tested for outliers with expected values. The weight modification method allows weight modification without discarding or replacing the values of outliers, thus limiting the influence of the outliers. The value modification method allows the replacement of the values of outliers with an appropriate value excluding outliers.
  - **Robust estimation:** When the nature of the population distributions is known, this approach is considered appropriate because it produces estimators robust to outliers, and estimators are consistent.



Statistical data preparation: management of missing values and outliers by S.K. Kwak and J.H. Kim (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5548942/)

Harvesting Data from the Web

- The Internet is a trove of not just unstructured data, but also structured data. Most of the time, APIs and data feeds aren't available.
- Web harvesting can be a one-off event (to get data for pilot studies) or once the usefulness of the harvested data has been established, programmatically scraped on a regular basis – essentially automating what can be a repetitive copy-and-paste.
- Challenging when dealing with dynamic content and dynamic rendering (e.g. infinite scrolling, JS, Ajax).
- Packages for web scraping:
    - R: Rcurl, Rvest, urltools, jsonlite, XML, RSelenium
    - Python: BeautifulSoup, Selenium, Scrapy, lxml
- Please harvest responsibly – always make sure of the following:
    - No robots.txt (see footnote below)
    - Throttle your scrapes (sleep/pause in between calls) – to reduce chances of your efforts being misclassified as DDOS attempts
- Simple workflow
    - Understand (1) the hierarchy of the website, and (2) the page structure of the webpages you want to systematically harvest – very investigative in nature
    - If table tags exist, this is even easier.
    - Remove duplicates (if harvesting on a regular basis) or keep track of change logs

Robots.txt is a text file webmasters create to instruct web robots (typically search engine robots) how to crawl pages on their website. The robots.txt file is part of the the robots exclusion protocol (REP), a group of web standards that regulate how robots crawl the web, access and index content, and serve that content up to users. The REP also includes directives like meta robots, as well as page-, subdirectory-, or site-wide instructions for how search engines should treat links (such as "follow" or "nofollow"). In practice, robots.txt files indicate whether certain user agents (web-crawling software) can or cannot crawl parts of a website. These crawl instructions are specified by "disallowing" or "allowing" the behavior of certain (or all) user agents.

Harvesting Data from the Web

- Don't be a burden
- Don't violate copyright
- Don't breach GDPR
- Beware of login and website terms and conditions
- The first rule of scraping the web is: do not harm the website. The second rule of web crawling is: do NOT harm the website.
  This means that the volume and frequency of queries you make should not burden the website's servers or interfere with the website's normal operations.
  You can accomplish this in a number of ways: Limit the number of concurrent requests to the same website from a single IP.
  Respect the delay that crawlers should wait between requests by following the crawl-delay directive outlined in the robots.txt file.
  If possible it is more respectful if you can schedule your crawls to take place at the website's off-peak hours.

Course Project

Course Project

- Series of mini-projects to underscore the different and important phases in financial data science – from acquiring a dataset, cleaning and analyzing it as well as iteratively building impactful insights out of it.
- Course project 1
  - Data cleaning and exploration
  - Simple statistical data analysis, outlier/missing data handling
- Course project 2
  - Data acquisition and investigative analysis and exploration
  - Web harvesting and building analytical insights in tracking trends, storytelling/narration coupled with visualization
- Course project 3
  - Building an unsupervised model to cluster stocks, modeling covariance/distance matrix structures
  - Understanding how companies/stocks are grouped together and investigate better ways of recategorizing peer groups
- *Course project 4: … < will retain optionality on this >*

Course Project 1: Data Cleaning and Exploration

- Data Scientists and Analysts spend almost 80% of their time cleaning and analyzing datasets.
- We are working with an external research firm who specializes in the application of machine learning to forecasting prices of financial instruments. This firm has developed a proprietary system, that we would like to investigate. To demonstrate the effectiveness of their forecasting system, the vendor has sent us the attached sample dataset.
- The dataset includes signal values generated by the proposed system as well as historical prices for a well-known broad market ETF.
- Before using the data in our production systems, we need to run through a few things:
    1. Review the quality of the data, list any potential errors, and propose corrected values. Please list each quality check error and correction applied.
    2. Please analyze the signal's effectiveness or lack thereof in forecasting ETF price, using whatever metrics you think are most relevant.
    3. Run any exploratory data analysis you think is important and highlight any interesting insights you come across.
    4. Write a summary for the team addressing your observations about the efficacy and believability of the product, and recommendation for next steps.
    5. Please include all the intermediate steps, and lay out your thinking as well.

Course Project 2: Data Acquisition and Analysis

- Salary data is an important component of a company's cost structure. For obvious reasons, this is not a data point that many will disclose readily. However, there are sources where such data needs to be filed and reported, usually driven by policy requirements.
- One example is H-1B visa data. The H-1B is a visa in the United States under the Immigration and Nationality Act, that allows U.S. employers to temporarily employ foreign workers in specialty occupations. A specialty occupation requires the application of specialized knowledge and a bachelor's degree or the equivalent of work experience. Essentially, this is to say that outside talent can be imported into the USA if a particular individual possesses a unique skillset that is not otherwise available within the local population.
- An example website that openly makes all H1B visa application data available is https://h1bdata.info/
- The schema of the dataset is straightforward. EMPLOYER | JOB TITLE | BASE SALARY | LOCATION | SUBMIT DATE | START DATE

EMPLOYER	JOB TITLE	BASE SALARY	LOCATION	SUBMIT DATE	START DATE
FACEBOOK INC	08 06 2024	160,025	MENLO PARK, CA	03/12/2021	08/07/2021
FACEBOOK INC	1101 DEXTER AVE N	161,233	SEATTLE, WA	09/28/2021	11/01/2021
FACEBOOK INC	ACADEMIC COLLABORATOR	111,000	MENLO PARK, CA	03/03/2021	09/01/2021
FACEBOOK INC	ACCESSIBILITY SPECIALIST	199,693	SAN FRANCISCO, CA	02/03/2021	05/31/2021
FACEBOOK INC	ACCOUNTING SYSTEMS MANAGER	227,000	MENLO PARK, CA	04/22/2021	05/03/2021
FACEBOOK INC	ADS RESEARCH LEAD, MARKETING SCIENCE RESEARCH	203,840	NEW YORK, NY	04/01/2021	09/06/2021
FACEBOOK MIAMI INC	AGENCY DIRECTOR, LATAM	331,672	MIAMI, FL	02/08/2021	08/05/2021
FACEBOOK MIAMI INC	AGENCY DIRECTOR, LATAM	342,882	MIAMI, FL	08/26/2021	10/01/2021
FACEBOOK INC	AI RESEARCH SCIENTIST	160,000	MENLO PARK, CA	02/17/2021	08/05/2021
FACEBOOK INC	AI RESEARCH SCIENTIST	160,000	NEW YORK, NY	03/02/2021	09/01/2021

Source: https://h1bdata.info/index.php?em=facebook&job=&city=&year=2021

Course Project 2: Data Acquisition and Analysis

As the lead scientist assigned to this project, there are a few tasks that needs to be performed:

1. Harvest the data from https://h1bdata.info
2. This is a fairly "easy" website given that most of the data we need for our analysis is structured and hence stored in tables within the page. Pay special attention to how to "grab" data by looking at the way in which the website (search options) were built.
   The modifiable dimensions are Location and Year. Ignore the free text search. You should retrieve all companies across all years. It seems like a lot but it's not.
   Hint: this is how I store my files:

Name	Date modified	Type	Size
1_XENIA_2021.csv	4/16/2022 5:50 PM	Microsoft Excel C...	1 KB
2_YORK_2021.csv	4/16/2022 5:50 PM	Microsoft Excel C...	86 KB
3_YORKTOWN%20HEIGHTS_2021.csv	4/16/2022 5:50 PM	Microsoft Excel C...	89 KB
4_YONKERS_2021.csv	4/16/2022 5:50 PM	Microsoft Excel C...	6 KB
5_YPSILANTI_2021.csv	4/16/2022 5:50 PM	Microsoft Excel C...	6 KB
6_YUMA_2021.csv	4/16/2022 5:50 PM	Microsoft Excel C...	4 KB
7_YAKIMA_2021.csv	4/16/2022 5:50 PM	Microsoft Excel C...	3 KB
8_YARDLEY_2021.csv	4/16/2022 5:50 PM	Microsoft Excel C...	2 KB
9_YOUNGSTOWN_2021.csv	4/16/2022 5:50 PM	Microsoft Excel C...	2 KB
10_YORBA%20LINDA_2021.csv	4/16/2022 5:50 PM	Microsoft Excel C...	2 KB

3. Once you have the data, explore and play with it. Run through some exploratory data analysis. (This is a fairly clean, high quality dataset so very little to do in terms of pre-processing for missing data/outliers). At best need to disambiguate companies e.g. Amazon, Amazon.com, Amazon Services LLC. etc.
4. What interesting trends do you see within the dataset? You can start by focusing on the larger companies and ignore the smaller companies with fewer employees that are applying for H1B visas.
5. Many ways to tease out insights from this dataset. For this kind of job datasets, essentially we are keen on exploring salary trends by company, as well as determining how expensive is it to hire people by roles (software engineers, data scientists, quantitative researchers, portfolio managers etc). For example (this is not an exhaustive list of questions), (i) Which is the most expensive city in the USA to build a startup? Who is the biggest hirer of H1B visa applicants? (ii) Is it better to be an analytical employee (e.g. data scientist/engineer/specialist) in a technology or investment management company? (iii) How has the trend of salary for data scientists been over the years? (iv) More qualitative: how much more useful will this dataset be if joined with other datasets/metadata? What types of data would that be (give examples)?

Course Project 2: Data Acquisition and Analysis

Appendix on H1B Visa (just some background, for context – not essential to the modelling)

- The Immigration Act of 1990 limits to 65,000 the number of foreign nationals who may be issued a visa or otherwise provided H-1B status each fiscal year (FY). An additional 20,000 H-1Bs are available to foreign nationals holding a master's or higher degree from U.S. universities. In addition, excluded from the ceiling are all H-1B non-immigrants who work at (but not necessarily for) universities, non-profit research facilities associated with universities, and government research facilities.
- Universities can employ an unlimited number of foreign workers otherwise qualifying for the H-1B as cap-exempt. This also means that contractors working at but not directly employed by the institutions may be exempt from the cap as well. However, employers must show 1) the majority of the worker's duties will be performed at the qualifying institution, organization or entity and 2) the job duties directly and predominantly further the essential purpose, mission objectives or functions of the qualifying institution, organization or entity. Free Trade Agreements carve out 1,400 H-1B1 visas for Chilean nationals and 5,400 H-1B1 visas for Singapore nationals. However, if these reserved visas are not used, then they are made available in the next fiscal year to applicants from other countries. Due to these unlimited exemptions and roll-overs, the number of H-1B visas issued each year is significantly more than the 65,000 cap, with 117,828 having been issued in FY2010, 129,552 in FY2011, and 135,991 in FY2012.
- In past years, the cap was not always reached. For example, in FY1996, the INS (now known as USCIS) announced on August 20, 1996 that a preliminary report indicated that the cap had been exceeded, and processing of H-1B applications was temporarily halted. However, when more accurate numbers became available on September 6, it became apparent the cap had not been reached after all, and processing resumed for the remainder of the fiscal year.
- The United States Citizenship and Immigration Services starts accepting applications on the first business day of April for visas that count against the fiscal year starting in October. For instance, H-1B visa applications that count against the FY 2013 cap were submitted starting Monday, 2012 April 2. USCIS accepts H-1B visa applications no more than 6 months in advance of the requested start date. Beneficiaries not subject to the annual cap are those who currently hold cap-subject H-1B status or have held cap-subject H-1B status at some point in the past six years.

Application Process: The process of getting a H-1B visa has three stages:

1. The employer files with the United States Department of Labor a Labor Condition Application (LCA) for the employee, making relevant attestations, including attestations about wages (showing that the wage is at least equal to the prevailing wage and wages paid to others in the company in similar positions) and working conditions.
2. With an approved LCA, the employer files a Form I-129 (Petition for a Nonimmigrant Worker) requesting H-1B classification for the worker. This must be accompanied by necessary supporting documents and fees.
3. Once the Form I-129 is approved, the worker may begin working with the H-1B classification on or after the indicated start date of the job, if already physically present in the United States in valid status at the time. If the employee is outside the United States, he/she may use the approved Form I-129 and supporting documents to apply for the H-1B visa. With a H-1B visa, the worker may present himself or herself at a United States port of entry seeking admission to the United States, and get a Form I-94 to enter the United States. (Employees who started a job on H-1B status without a H-1B visa because they were already in the United States still need to get a H-1B visa if they ever leave and wish to reenter the United States while on H-1B status.)

Course Project 3: Building a Better Company Classification Scheme

- Many institutional investors rely on traditional index providers to categorize stocks into different sectors and industries. The reason for doing so is to allow risk taking to be done as efficiently as possible. For example, and speaking from a stock-selection perspective, if you are overweight Netflix, ideally you are also underweight a stock that are driven by the same economic drivers so that you are best able to isolate idiosyncratic stock risk, i.e. maybe Disney/Hulu?.
- One such index classification scheme is the Global Industry Classification Standard (GICS) is an industry taxonomy developed in 1999 by MSCI and Standard & Poor's (S&P) for use by the global financial community. The GICS structure consists of 11 sectors, 24 industry groups, 69 industries and 158 sub-industries into which S&P has categorized all major public companies.
- But because almost every institution uses this same set of classification rules, it can lead to very crowded ways of thinking about segmenting company risk. Sometimes it might be preferable to develop your own stock classification scheme such that you would be better able to segment companies along the right business and economic drivers.
- One such approach is to use how the sell-side analysts organize themselves to cover companies. Often they try to cover as many names within their investment sphere that are as similar as much as possible in terms of business, industry and macro drivers. This is done so that there are minimal overlaps between analysts, thus maximizing coverage of all investable stocks while minimizing the use of human labour.
- As it is with any new dataset, get comfortable with it and explore. In our actual problem, we enriched this dataset with other datasets as well as metadata to form even more efficient clusters to address our needs.

Source: https://en.wikipedia.org/wiki/Global_Industry_Classification_Standard

Course Project 3: Building a Better Company Classification Scheme

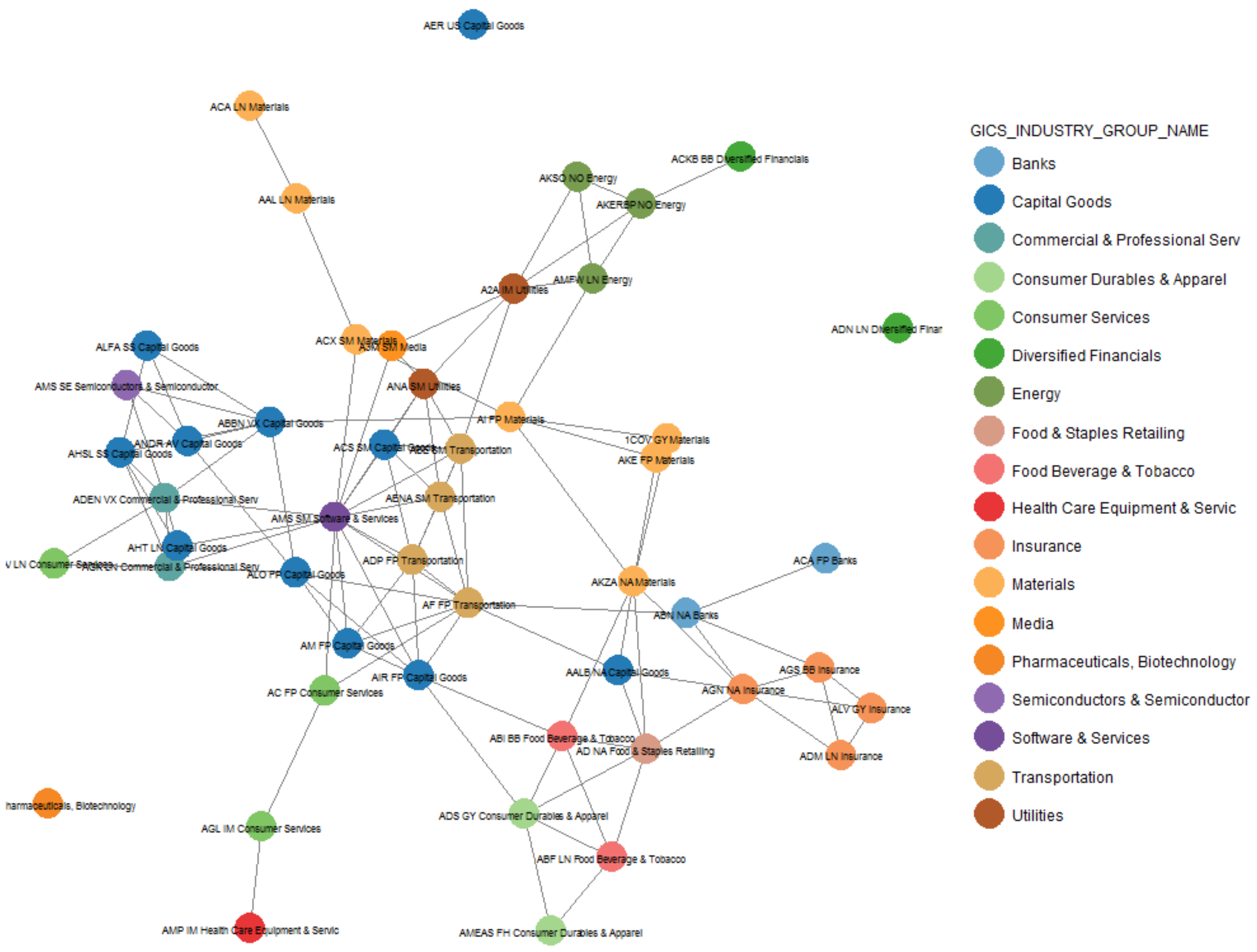
```
> df[order(RATING)]
        ANALYST       DATE BROKER RATING     RECOMMENDATION TARGET_PRICE       BBTICKER GICS_SECTOR_NAME       GICS_INDUSTRY_GROUP_NAME
   1: JamJones  2/17/2020  RBets      1        underperform    9.100e+01 NESN SW Equity Consumer Staples       Food, Beverage & Tobacco
   2: Ioatikis  2/13/2020  MoInc      1                sell    9.100e+01 NESN SW Equity Consumer Staples       Food, Beverage & Tobacco
   3: JefStent  2/13/2020  Exbas      1        underperform    1.030e+02 NESN SW Equity Consumer Staples       Food, Beverage & Tobacco
   4: Luiector 10/21/2019  Exbas      1        underperform    2.600e+02  ROG SW Equity      Health Care Pharmaceuticals, Biotechnology
   5: Amit Roy  3/27/2018  FoLLP      1         underweight    2.350e+02  ROG SW Equity      Health Care Pharmaceuticals, Biotechnology
  ---
8672: KazAndac   3/6/2020  Deank     NA           not rated   -2.420e-14 BIRG ID Equity       Financials                          Banks
8673: Sylarker  2/26/2020  J.gan     NA    Rating Suspended   -2.420e-14  GFS LN Equity      Industrials Commercial & Professional Serv
8674: Phihards  9/19/2018  Gochs     NA suspended coverage   -2.420e-14  GFS LN Equity      Industrials Commercial & Professional Serv
8675: Kriiksen   6/4/2018  SEies     NA suspended coverage   -2.420e-14  GFS LN Equity      Industrials Commercial & Professional Serv
8676: Micchill 10/26/2017  Chrch     NA   no rating system    3.466e+01  LHA GR Equity      Industrials                 Transportation
```

- We always compare against peers in the same sector/industry – allows more calibrated risk taking to isolate idiosyncratic risk, because we are then able to net out common business upside drivers/downside risk.
- GICS industry code (8 digits) consists of:
  11 sectors, 24 industry groups, 69 industries and 158 sub-industries
  `(sector | industry group | industry | sub-industry)`
- Key in this project is the modelling of the covariance/distance matrix

45	Information Technology	4510	Software & Services	451020	IT Services	45102010	IT Consulting & Other Services
						45102020	Data Processing & Outsourced Services
						45102030	Internet Services & Infrastructure
				451030	Software	45103010	Application Software
						45103020	Systems Software
		4520	Technology Hardware & Equipment	452010	Communications Equipment	45201020	Communications Equipment
				452020	Technology Hardware, Storage & Peripherals	45202030	Technology Hardware, Storage & Peripherals
				452030	Electronic Equipment, Instruments & Components	45203010	Electronic Equipment & Instruments
						45203015	Electronic Components
						45203020	Electronic Manufacturing Services
						45203030	Technology Distributors
		4530	Semiconductors & Semiconductor Equipment	453010	Semiconductors & Semiconductor Equipment	45301010	Semiconductor Equipment
						45301020	Semiconductors

Source: https://en.wikipedia.org/wiki/Global_Industry_Classification_Standard

Course Project 3: Building a Better Company Classification Scheme



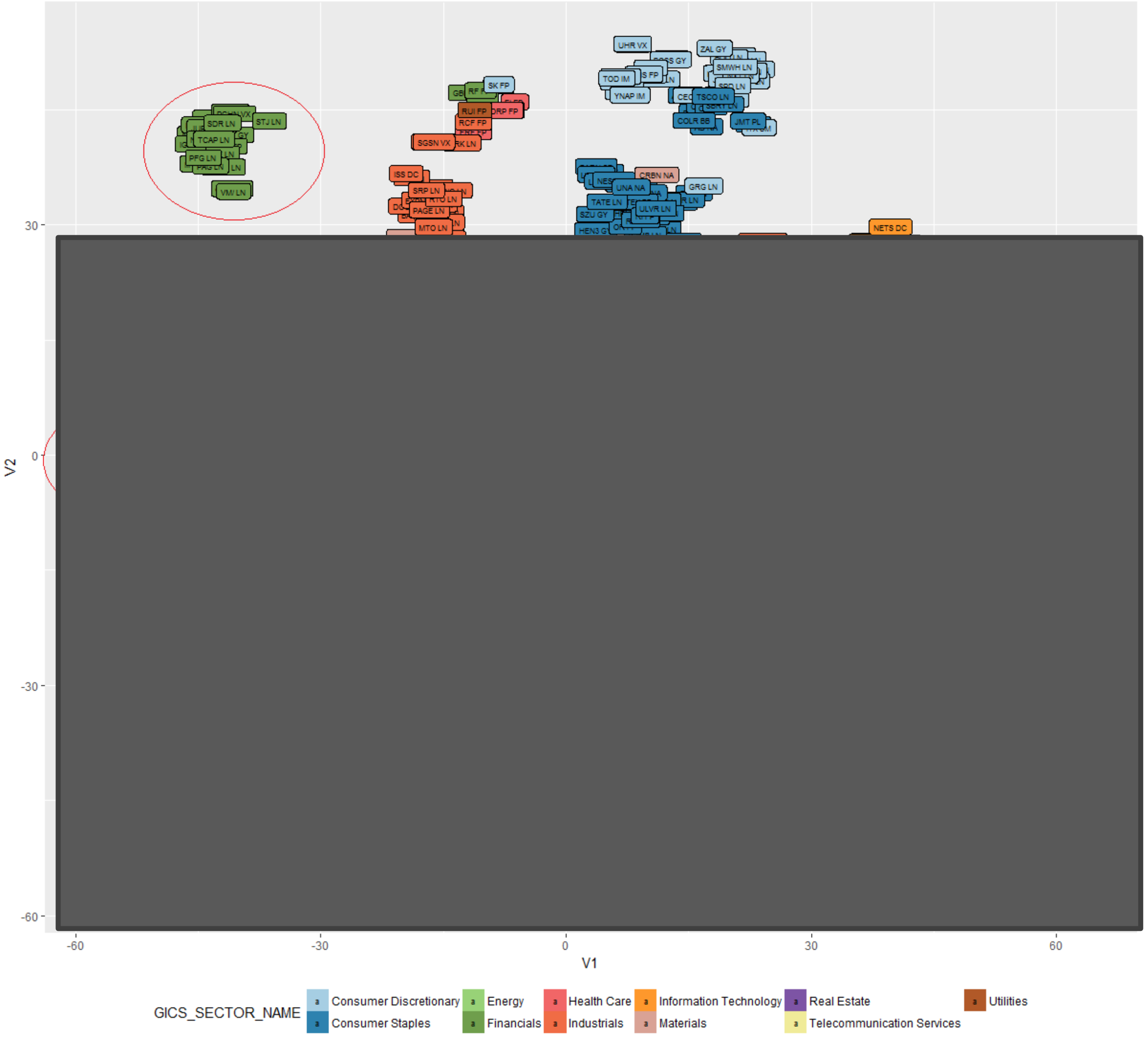
We don't usually like to depend on standard company classification methodologies – based on the following data frame of analyst coverage of companies, we would like to pursue a recategorization of companies based on exactly this - analyst co-coverage. Note some columns are redundant/not needed.

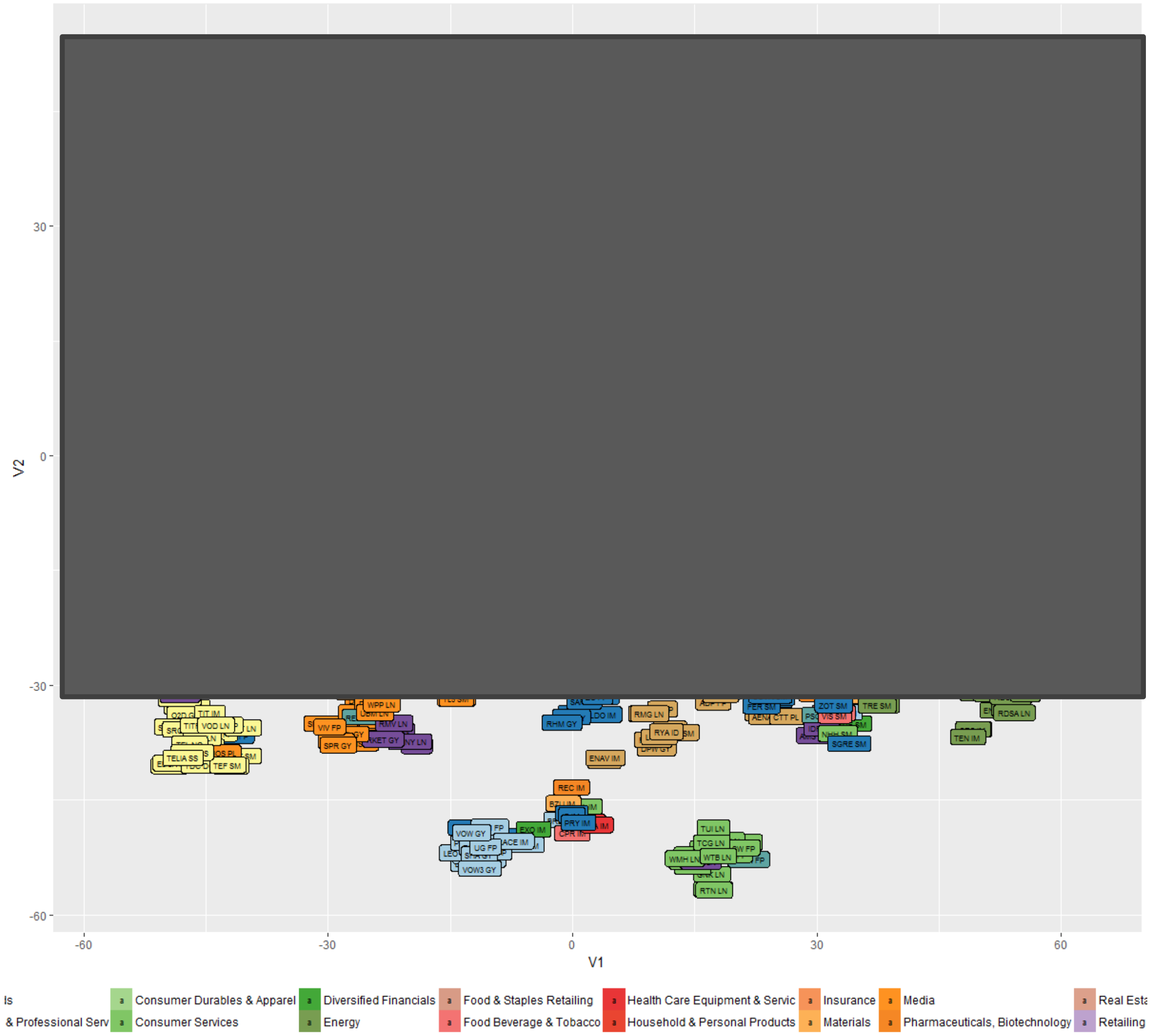
1. Which company has the higher analyst coverage? (Look at histogram)
2. Which analyst covers the most companies? (Look at histogram)
3. Based on how analysts organize themselves into covering companies,
   a. Could you model the similarity or conversely, the distance matrix between the companies based on this analyst co-coverage
   b. How would the results change if you were to restrict the dataset to only analysts having companies covered within 1s.d. of the distribution found in Qn. 2?
   c. If further restricted to a smaller subset?
4. Which sectors are the most heterogenous? (Look at the clusters formed by industry groups per sector – use t-SNE to visualize)
5. Similarly, which sectors are the most homogenous?
6. What type of companies tend to be outliers in terms of the clusters?
7. Feel free to explore and provide deeper insights in the structure of the clusters/network as part of the outputs.

Source: https://en.wikipedia.org/wiki/Global_Industry_Classification_Standard

Course Project 3: Building a Better Company Classification Scheme

Example of what the clusters should look like.





Colour-coded by original GICS Sectors (Left) and Industry Groups (Right)