FA582 Assignment 2 Report

Naveen Mathews Renji - 20016323

Problem 1

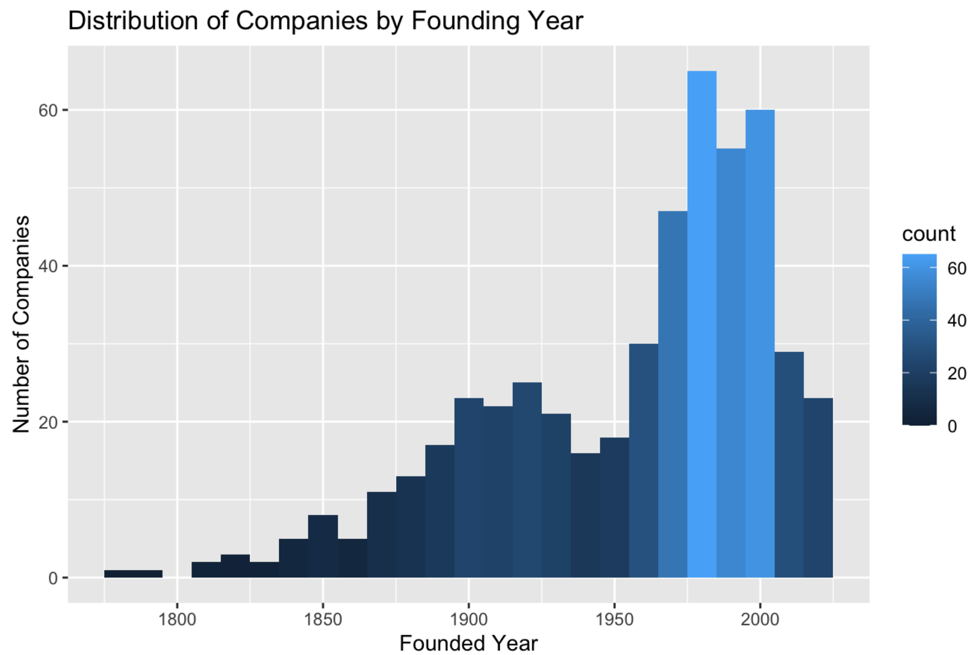
- Write an R code to scrape the website:
    - https://en.wikipedia.org/wiki/List_of_S%26P_500_companies
- Retrieve the content of the S&P 500 component stocks table (Symbol, Security, GISC Sector, GICS Sub-Industry, Headquarters Location, Date added, CIK, Founded). Create an R dataframe and perform exploratory data analysis and report summary statistics.

1. Web scraping the Data and validating it (Handling missing values, outliers and zero values)

1. I scraped the data from the link using httr and rvest, then I extracted the S&P 500 table.
2. Changed the 'Date added' and 'Founded' to Date type from character type
3. I checked the data for 0 values and found that there were none.
4. Then I checked for missing or 'NA' values and found 12 present in the 'Date added' column.
5. I am leaving it as it is as it seems very insignificant to remove or handle them.

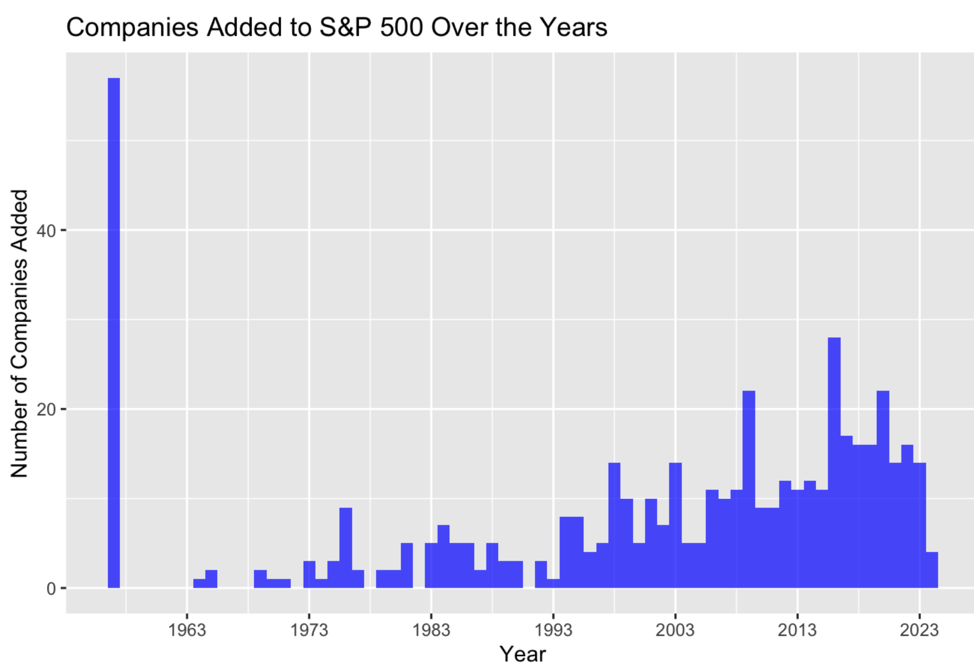
2. Performing EDA

1. Distribution of Companies by Founding Year



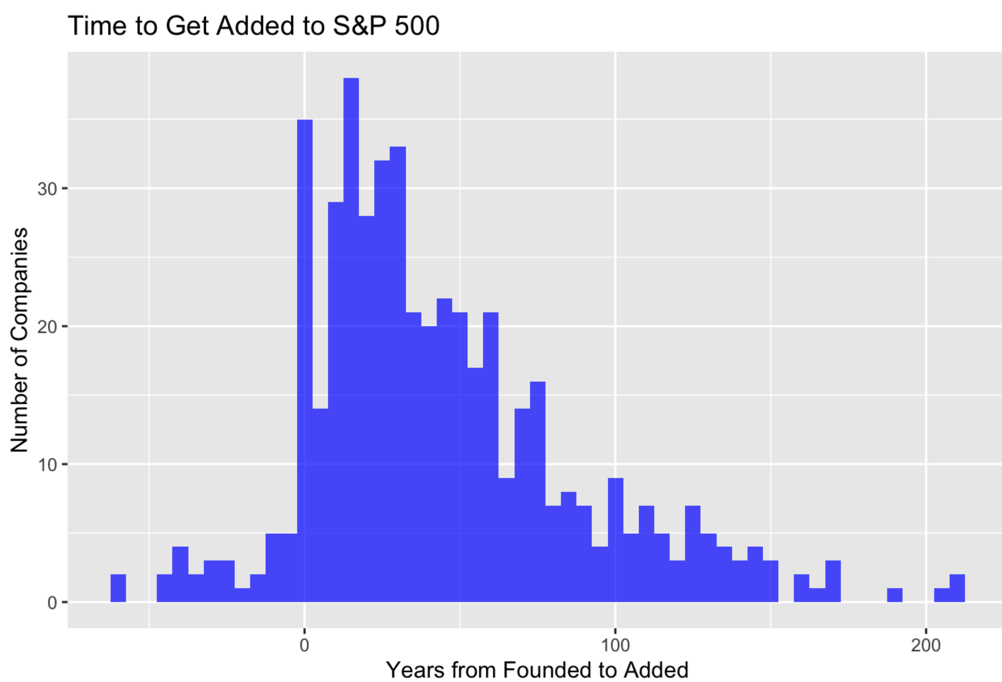
We can see that most of the companies were founded in the past century but some are even older with a few going back to even the 17th century.

2. Companies Added to S&P 500 Over the Years



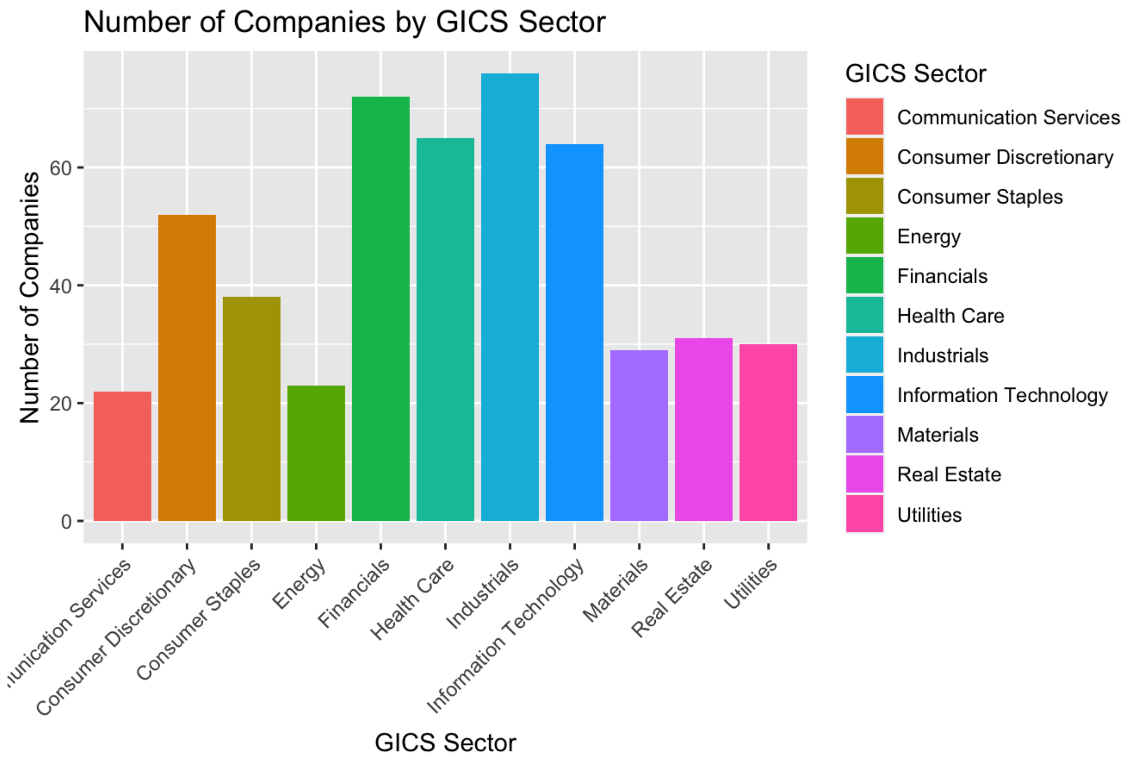
We can see that most of the companies were added in the past 30 years. 12 companies did not have the data on date added and there were 57 companies whose date of being added was on "1957-03-04" which could have been the first companies added when the index was founded.

3. Years Taken to Get Added to S&P 500 by Year Founded



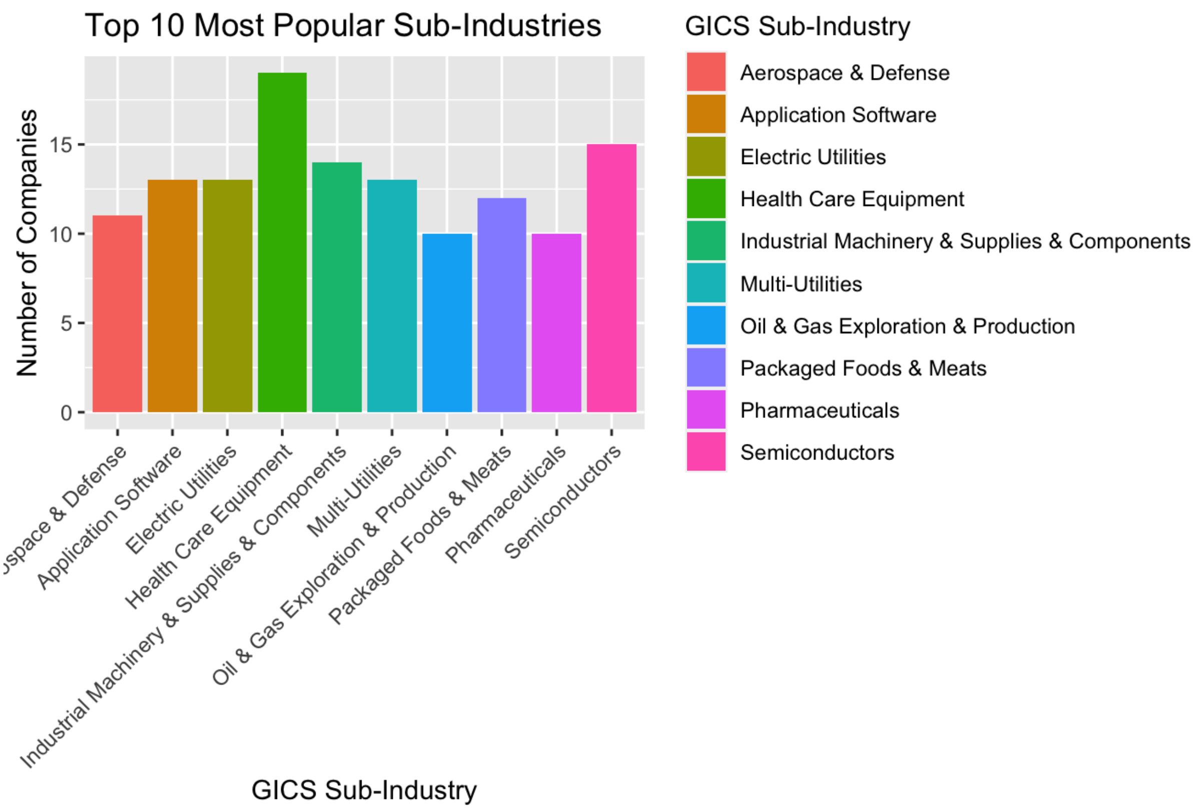
Most companies have taken less than 50 years to make it on the index though there are quite a few that have taken more than 50 years. There is also a small percentage of companies that have taken over a century and a handful that took over 2 centuries to make it which is understandable as the index was only founded around 70 years ago.

4. Number of Companies in each Sector



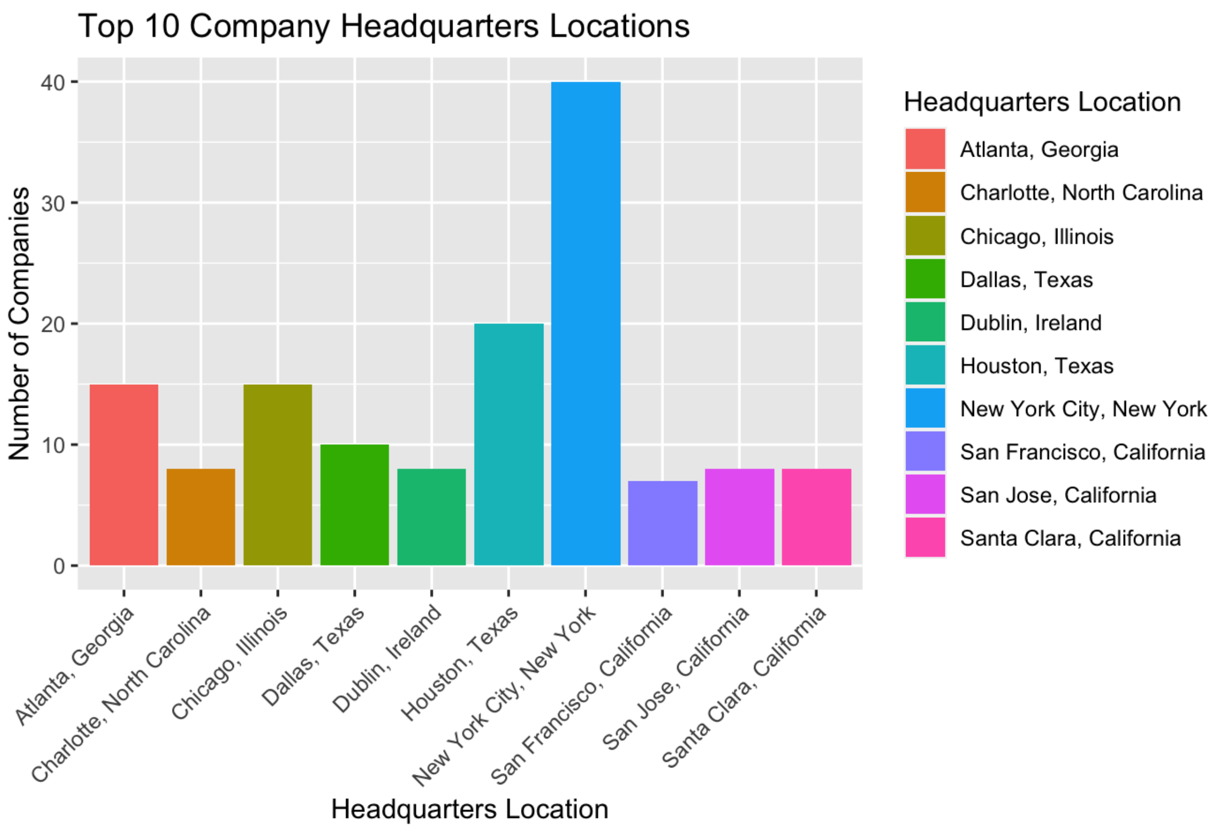
This graph allows us to see the various GICS of the companies listed. We can observe that the largest number of companies belong to 4 sectors which are Financials, Health care, industrials and IT. The Consumer Discretionary sector has the next most companies followed by an almost equal split among the rest of the sectors.

5. Top 10 Most Popular Sub-Industries



While Industrials and Financial companies are more popular by sector, healthcare equipment companies are the most popular Sub industry.

6. Top 10 Company Headquarters Locations



New York City is the clear winner when it comes to setting up office headquarters for the s&p companies but Houston Texas seems to be climbing up above the rest, this could be due to its ease of policies on setting up companies and also because they have less tax on income.

Problem 2

The data provided in the files contains several quantitative and categorical variables associate with each ticker. Please select a subset of 100 tickers from each file and use data for a specific year (ex: 2013). Use a small number of quantitative variables (10 or 12) out of ~76 columns available (example: After Tax ROE, Cash Ratio, Current Ratio, Operating Margin, Pre-Tax Margin,
Pre-Tax ROE, Profit Margin, Quick Ratio, Total Assets, Total Liabilities, Earnings Per Share, etc...).
The categorical variables available are GICS Sector, GICS Sub Industry, and possibly HQ Address
(although this is sparse data for the 100 tickers subset selected).
Next, you have to apply several distance and similarity functions to find the extreme values for distance and similarities between the subset of tickers that you chose. For each of the following cases, please define the function that allows you to calculate the quantity required, calculate the values for all ticker pairs, and rank the pairs by calculated value of distance or similarity, and report the top and bottom 10 values for each case:

a) Lp-norm for p = 1
b) Lp-norm for p = 2
c) Lp-norm for p = 3
d) Lp-norm for p = 10
e) Minkovski distance (assign different weights for the feature components in the Lp-norm
based on your assessment on the importance of the features)
f) Match-Based Similarity Computation (use a small number of equi-depth buckets, ex: 3)
g) Mahalanobis distance
h) Similarity: overlap measure
i) Similarity: inverse frequency
j) Similarity: Goodall
k) Overall similarity between tickers by using mixed type data (choose a λ value for calculation)
l) Overall normalized similarity between tickers by using mixed type data (choose a λ value for calculation)

1. Data Loading, Cleaning and Merging

1. Data Loading
    a. Two separate datasets named fundamentals.csv and securities.csv were loaded into R. Each dataset represents companies and their associated financial data. The necessary libraries like readr and dplyr were loaded to assist in data manipulation.
2. Data Structure Check
    a. We examined the structure of both data frames to understand the column names and their data types.
3. Data Filtering
    a. Year Filtering: The data was filtered to include only records from the year 2013.
4. Data Cleaning
    a. Missing or Zero Values: The code identifies rows with least missing or zero values in quantitative columns such as "After Tax ROE", "Cash Ratio", etc. It adds a new column called missing_or_zero that counts these missing or zero values per row.
    b. Top 100 Tickers: After sorting the data frame by the missing_or_zero column, the top 100 tickers were selected for further analysis.
    c. Data Subset: Data for these top 100 tickers was then extracted from both the fundamentals and securities data frames.
    d. Select Quantitative Columns: Only quantitative columns of interest were selected from the fundamentals data subset.

5. Data Merge
    a. The fundamentals and securities data frames were merged on the "Ticker Symbol", providing a comprehensive dataset.

2. Distances and Similarities

Lp norms (p=1, p=2, p=3, p=10).

- Lp-norm Function: An Lp-norm function was defined to calculate the Lp-norm distance between two vectors.
- Distance Matrix: A square matrix was created to store the Lp-norm distance values for each pair of tickers. The matrix is symmetric, meaning the distance from A to B is the same as the distance from B to A.
- Multiple Lp-norms: For each pair of tickers, distances were calculated for multiple Lp norms (p=1, p=2, p=3, p=10).

Results -

[1] "Top 10 and Bottom 10 for Lp-norm with p = 1"
[1] "Top 10:"

	Ticker_Pair	Distance
COG-EA	COG-EA	1.15e+08
ADSK-CTAS	ADSK-CTAS	1.58e+08
BCR-EA	BCR-EA	1.79e+08
BCR-COG	BCR-COG	2.36e+08
ADSK-EFX	ADSK-EFX	2.58e+08
CHD-CTAS	CHD-CTAS	2.70e+08
CBG-DRI	CBG-DRI	2.87e+08
CTAS-EFX	CTAS-EFX	2.89e+08
CF-DOV	CF-DOV	3.01e+08
ADSK-CHD	ADSK-CHD	3.54e+08

[1] "Bottom 10:"

	Ticker_Pair	Distance
ALB-CVX	ALB-CVX	3.53e+11
CERN-CVX	CERN-CVX	3.53e+11
CVX-DNB	CVX-DNB	3.54e+11
CHRW-CVX	CHRW-CVX	3.54e+11
ALXN-CVX	ALXN-CVX	3.54e+11
COO-CVX	COO-CVX	3.55e+11
CVX-DLTR	CVX-DLTR	3.55e+11
AKAM-CVX	AKAM-CVX	3.55e+11

AYI-CVX	AYI-CVX	3.56e+11
CMG-CVX	CMG-CVX	3.56e+11

[1] "Top 10 and Bottom 10 for Lp-norm with p = 2"

[1] "Top 10:"

	Ticker_Pair	Distance
COG-EA	COG-EA	9.28e+07
ADSK-CTAS	ADSK-CTAS	1.27e+08
BCR-EA	BCR-EA	1.53e+08
BCR-COG	BCR-COG	1.86e+08
CHD-CTAS	CHD-CTAS	2.03e+08
CTAS-EFX	CTAS-EFX	2.16e+08
CF-DOV	CF-DOV	2.16e+08
ADSK-EFX	ADSK-EFX	2.33e+08
CBG-DRI	CBG-DRI	2.33e+08
ALXN-COO	ALXN-COO	2.72e+08

[1] "Bottom 10:"

	Ticker_Pair	Distance
CERN-CVX	CERN-CVX	2.70e+11
ALB-CVX	ALB-CVX	2.70e+11
ALXN-CVX	ALXN-CVX	2.71e+11
CHRW-CVX	CHRW-CVX	2.71e+11
COO-CVX	COO-CVX	2.71e+11
CVX-DLTR	CVX-DLTR	2.72e+11
CVX-DNB	CVX-DNB	2.72e+11
AKAM-CVX	AKAM-CVX	2.72e+11
AYI-CVX	AYI-CVX	2.72e+11
CMG-CVX	CMG-CVX	2.72e+11

[1] "Top 10 and Bottom 10 for Lp-norm with p = 3"

[1] "Top 10:"

	Ticker_Pair	Distance
COG-EA	COG-EA	8.97e+07
ADSK-CTAS	ADSK-CTAS	1.22e+08
BCR-EA	BCR-EA	1.50e+08
BCR-COG	BCR-COG	1.79e+08
CHD-CTAS	CHD-CTAS	1.90e+08
CF-DOV	CF-DOV	1.95e+08
CTAS-EFX	CTAS-EFX	2.02e+08
CBG-DRI	CBG-DRI	2.27e+08
ADSK-EFX	ADSK-EFX	2.32e+08
ALXN-COO	ALXN-COO	2.42e+08

[1] "Bottom 10:"

	Ticker_Pair	Distance

CERN-CVX   CERN-CVX 2.55e+11
ALB-CVX   ALB-CVX 2.56e+11
ALXN-CVX   ALXN-CVX 2.56e+11
COO-CVX   COO-CVX 2.56e+11
CHRW-CVX   CHRW-CVX 2.57e+11
AKAM-CVX   AKAM-CVX 2.57e+11
CVX-DLTR   CVX-DLTR 2.57e+11
CVX-DNB   CVX-DNB 2.57e+11
CMG-CVX   CMG-CVX 2.58e+11
AYI-CVX   AYI-CVX 2.58e+11

[1] "Top 10 and Bottom 10 for Lp-norm with p = 10"
[1] "Top 10:"
Ticker_Pair Distance
COG-EA   COG-EA 8.89e+07
ADSK-CTAS   ADSK-CTAS 1.21e+08
BCR-EA   BCR-EA 1.50e+08
BCR-COG   BCR-COG 1.76e+08
CF-DOV   CF-DOV 1.78e+08
CHD-CTAS   CHD-CTAS 1.84e+08
CTAS-EFX   CTAS-EFX 1.94e+08
ALXN-COO   ALXN-COO 2.08e+08
CBG-DRI   CBG-DRI 2.25e+08
ADSK-EFX   ADSK-EFX 2.32e+08
[1] "Bottom 10:"
Ticker_Pair Distance
CERN-CVX   CERN-CVX 2.50e+11
ALB-CVX   ALB-CVX 2.50e+11
ALXN-CVX   ALXN-CVX 2.50e+11
COO-CVX   COO-CVX 2.51e+11
AKAM-CVX   AKAM-CVX 2.51e+11
CHRW-CVX   CHRW-CVX 2.51e+11
CVX-DLTR   CVX-DLTR 2.51e+11
CMG-CVX   CMG-CVX 2.52e+11
AYI-CVX   AYI-CVX 2.52e+11
CVX-DNB   CVX-DNB 2.52e+11

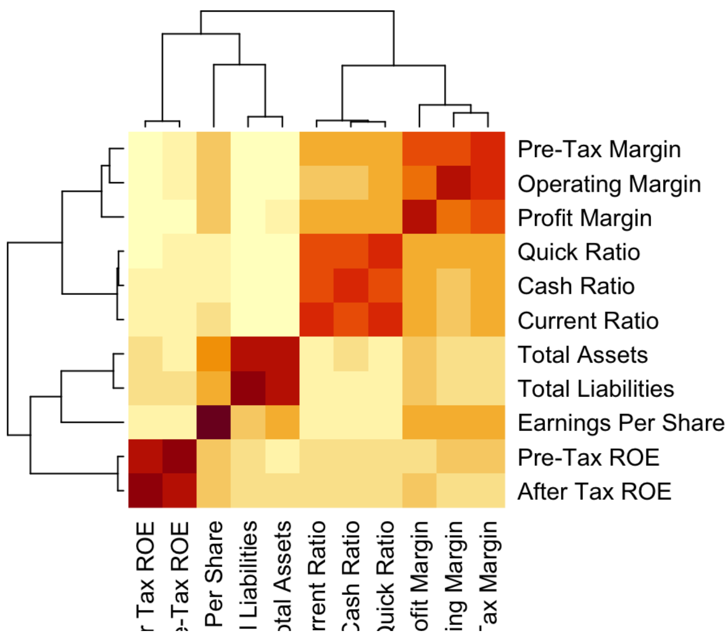
Observations -

- **Consistency Across Norms**: For all values of p (1, 2, 3, 10), the top 10 and bottom 10 ticker pairs remained the same even though the score changed. This consistency hints at a stable relationship between the financial metrics of these companies, irrespective of the distance norm used.

- **Similar Companies**: Some company pairs, like EA-COG and CTAS-ADSK, showed the smallest Lp-norm distances, making them the most financially similar based on the selected metrics.
- **Dissimilar Companies**: Pairs including CVX (Chevron) frequently appeared in the bottom 10, showing that Chevron is financially distinct from companies like CHRW, DLTR, and CMG.
- **Magnitude of Distances**: The bottom 10 pairs had notably larger Lp-norm distances compared to the top 10, indicating a substantial financial disparity probably.
- **Symmetric Distances**: As expected, the calculated distances were symmetric. For instance, the distance from EA to COG was identical to the distance from COG to EA.

Minkovski distance

We first created a correlation matrix and then generated a heatmap to help me determine the weights for the features.



- **Highly Correlated Features**: From the heatmap, we observe that some features are highly correlated (bright red). For instance, "After Tax ROE" and "Pre-Tax ROE" are very closely correlated. Such features can lead to multicollinearity if used together in some models, meaning that they carry similar information. We would typically not want to give both these features high weights as they might introduce redundancy.

- **Domain Knowledge**: Certain financial ratios are more critical than others depending on the context. For example, margins (like Operating Margin, Profit Margin, Pre-Tax Margin) are often viewed as vital indicators of a company's operational efficiency.

- **Less Correlated Features**: Features that are less correlated (light yellow) with other features bring unique information to the table and can be weighed higher.

Initial Weights (considering correlations and importance (what I am perceiving)):

Pre-Tax Margin: 0.8
Operating Margin: 0.8
Profit Margin: 0.8
Quick Ratio: 0.7
Cash Ratio: 0.6
Current Ratio: 0.6
Total Assets: 0.9
Total Liabilities: 0.9
Earnings Per Share: 0.8
Pre-Tax ROE: 0.9
After Tax ROE: 0.5

To normalize these weights, we can divide each weight by the sum of all weights:

Sum of all weights: 8.5

Normalized Weights:

Pre-Tax Margin: 0.8/8.5 = 0.0941
Operating Margin: 0.8/8.5 = 0.0941
Profit Margin: 0.8/8.5 = 0.0941
Quick Ratio: 0.7/8.5 = 0.0824
Cash Ratio: 0.6/8.5 = 0.0706
Current Ratio: 0.6/8.5 = 0.0706
Total Assets: 0.9/8.5 = 0.1059
Total Liabilities: 0.9/8.5 = 0.1059
Earnings Per Share: 0.8/8.5 = 0.0941
Pre-Tax ROE: 0.9/8.5 = 0.1059
After Tax ROE: 0.5/8.5 = 0.0588

Results

[1] "Top 10 and Bottom 10 for Weighted Minkowski Distance"
[1] "Top 10:"

	Ticker_Pair	Distance
COG-EA	COG-EA	28024768
ADSK-CTAS	ADSK-CTAS	31499020

BCR-EA	BCR-EA	37414300
BCR-COG	BCR-COG	46573807
CHD-CTAS	CHD-CTAS	51893738
CBG-DRI	CBG-DRI	57783079
CF-DOV	CF-DOV	62112837
CTAS-EFX	CTAS-EFX	63895851
ADSK-EFX	ADSK-EFX	71293615
ALXN-COO	ALXN-COO	74054585

[1] "Bottom 10:"

	Ticker_Pair	**Distance**
CERN-CVX	CERN-CVX	8.06e+10
ALB-CVX	ALB-CVX	8.07e+10
ALXN-CVX	ALXN-CVX	8.08e+10
COO-CVX	COO-CVX	8.09e+10
CHRW-CVX	CHRW-CVX	8.09e+10
AKAM-CVX	AKAM-CVX	8.10e+10
CVX-DLTR	CVX-DLTR	8.10e+10
CVX-DNB	CVX-DNB	8.11e+10
AYI-CVX	AYI-CVX	8.12e+10
CMG-CVX	CMG-CVX	8.13e+10

Observations -

Results are very similar to the Lp norms results in terms of ranking. This shows that the weights assigned didn't make a very big difference in the computation of ranking which may be due to the large value these columns hold

Match-Based Similarity Computation

Number of equi-depth buckets - 3

For each column (attribute) of the dataset, divide its values into 3 buckets of (approximately) equal size.
For each pair of rows (companies), calculate the match score, which is the number of columns where the two rows fall into the same bucket.

Results -

[1] "Top 10 most similar pairs based on match score:"

	Var1	Var2	Freq
1	ED	AEP	11
2	BCR	ALB	11
3	CMG	APH	11
4	ALB	BCR	11
5	CHK	CCL	11
6	CTL	CCL	11
7	CCL	CHK	11
8	CTL	CHK	11
9	APH	CMG	11
10	CCL	CTL	11

[1] "Bottom 10 least similar pairs based on match score:"

	Var1	Var2	Freq
1	DISCA	AAL	0
2	DOV	AAL	0
3	ATVI	AAP	0
4	CSCO	AAP	0
5	DHR	AAP	0
6	AEE	AAPL	0
7	CBG	AAPL	0
8	CRM	AAPL	0
9	CTXS	AAPL	0
10	DRI	AAPL	0

Most Similar Pairs: The top 10 pairs of companies that showcased the highest similarity based on their match scores were ED & AEP, BCR & ALB, CMG & APH, among others. All these pairs had a high match score of 11, indicating they were classified into the same bucket for all the 11 attributes.

Least Similar Pairs: On the other end of the spectrum, the 10 least similar pairs include DISCA & AAL, DOV & AAL, and ATVI & AAP, among others. These pairs had a match score of 0, suggesting they didn't fall into the same bucket for any of the attributes considered

Mahalanobis Distance

The Mahalanobis distance is a measure of the distance between a point and a distribution, which is particularly useful in multivariate data analysis. It measures distance concerning the correlations between variables.

Results -

[1] "Top 10 Mahalanobis distances:"

```
   Var1 Var2 Freq
1   AAL  AAL    0
2   AAP  AAP    0
3  AAPL AAPL    0
4  ABBV ABBV    0
5   ABT  ABT    0
6  ADBE ADBE    0
7   ADI  ADI    0
8   ADM  ADM    0
9   ADS  ADS    0
10 ADSK ADSK    0
```

[1] "Bottom 10 Mahalanobis distances:"

```
      Var1 Var2     Freq
9991   DAL  CLX 11.54225
9992   CLX  DAL 11.54225
9993   CLX AAPL 11.70798
9994  AAPL  CLX 11.70798
9995  CHTR  ADI 11.74556
9996   ADI CHTR 11.74556
9997   CLX  ADI 12.02816
9998   ADI  CLX 12.02816
9999   CLX CHTR 12.62446
10000 CHTR  CLX 12.62446
```

Top 10 Mahalanobis Distances:

The Mahalanobis distance is zero for all pairs of identical tickers (e.g., AAL to AAL, AAP to AAP, etc.). This is expected as the distance between identical points should be zero.

Bottom 10 Mahalanobis Distances:

Among different tickers, the pairs with the highest Mahalanobis distances include CLX to DAL, CLX to AAPL, and CLX to ADI, with distances of approximately 11.54, 11.71, and 12.03 respectively. The highest distance is between CLX and CHTR, at approximately 12.62.

Similarity: overlap measure

The overlap measure is a simple similarity measure which calculates the overlap between two vectors. If xi=yi, the it gets 1, else it gets 0.

1. Selected 3 categorical variables - c("GICS Sector", "GICS Sub Industry", "Address of Headquarters")
2. Selected 100 rows with least missing and zero values
3. Convert the categorical cols to character type
4. Created function to calculate the overlap similarity
5. Calculates similarity for each pair of tickers
6. Converted similarity matrices to data frames for easier sorting and viewing
7. Printed top 10 and bottom 10 similarities, excluding diagonal and duplicate pairs

Results -

```
[1] "Top 10 Overlap Similarities"
   Var1  Var2 Freq
1  GOOG GOOGL    3
2   APA   COG    3
3  ABBV   ABT    2
4     A   ABT    2
5   AME   AYI    2
6   ADP  AKAM    2
7   AAL   ALK    2
8  ABBV   AGN    2
9   AEP   LNT    2
10  AIG   ALL    2
```

```
[1] "Bottom 10 Overlap Similarities"
   Var1 Var2 Freq
1   ABT  MMM    0
2  ABBV  MMM    0
3   ACN  MMM    0
4  ATVI  MMM    0
5  ADBE  MMM    0
6   AAP  MMM    0
7   AES  MMM    0
8   AET  MMM    0
9   AMG  MMM    0
10  AFL  MMM    0
```

Top 10 Similarity Overlap:

Two ticker pairs - (GOOG GOOGL and APA COG) have a perfect overlap score of 3 indicating a perfect match in the selected variables for each ticker. The remaining in the top 10 matched 2 cols in each pair.

Bottom 10 Similarity Overlap:

Numerous ticker pairs have a similarity overlap score of 0, indicating no overlap in the selected variables. Specifically, the ticker MMM has no overlap with the other bottom tickers pair.

Similarity: inverse frequency

The similarity inverse frequency measure evaluates the similarity between pairs of tickers based on the inverse of the frequency of each value in their selected variables. Let pk(x) be the fraction of records in which the kth attribute takes on the value of ix n the data set - similarity (xi,yi) = { 1/pk(xi)^2 if xi=yi, 0 otherwise }

1. Selected 3 categorical variables - c("GICS Sector", "GICS Sub Industry", "Address of Headquarters")
2. Selected 100 rows with least missing and zero values
3. Convert the categorical cols to character type
4. Created function to calculate the similarity inverse frequency
5. Calculated p value using table()
6. Calculates similarity for each pair of tickers
7. Converted similarity matrices to data frames for easier sorting and viewing
8. Printed top 10 and bottom 10 similarities, excluding diagonal and duplicate pairs

Results -

[1] "Top 10 Inverse Frequency Similarities"

	Var1	Var2	Freq
1	CTL	T	5000
2	GOOG	GOOGL	3169
3	AMT	BXP	2900
4	AIV	BXP	2900
5	AEP	LNT	2778
6	AEE	CNP	2778
7	CHK	CVX	2778
8	AME	AYI	2600
9	AAL	ALK	2600
10	ARNC	BA	2600

[1] "Bottom 10 Inverse Frequency Similarities"

	Var1	Var2	Freq
1	ABT	MMM	0
2	ABBV	MMM	0
3	ACN	MMM	0
4	ATVI	MMM	0
5	ADBE	MMM	0
6	AAP	MMM	0

7   AES MMM   0
8   AET MMM   0
9   AMG MMM   0
10 AFL MMM   0

Top 10 Similarity Inverse Frequency:

The top ten pairs have similarity scores above and equal to 2600, with the highest being CTL T pair.

Bottom 10 Similarity Inverse Frequency:

Several ticker pairs have a similarity inverse frequency score of 0. Specifically, the ticker MMM has no overlap with the other bottom ticker pair which is consistent with the overlap similarity.

Similarity: Goodall

In goodall similarity, the similarity on the kth attribute is defined as 1-pk(xi)^2, when xi=yi and 0 otherwise - similarity (xi,yi) = { 1 - pk(xi)^2 if xi=yi, 0 otherwise }

1. Selected 3 categorical variables - c("GICS Sector", "GICS Sub Industry", "Address of Headquarters")
2. Selected 100 rows with least missing and zero values
3. Convert the categorical cols to character type
4. Created function to calculate the goodall similarity
5. Calculated p value using table()
6. Calculates similarity for each pair of tickers
7. Converted similarity matrices to data frames for easier sorting and viewing
8. Printed top 10 and bottom 10 similarities, excluding diagonal and duplicate pairs

[1] "Top 10 Goodall Similarities"

	Var1	Var2	Freq
1	APA	COG	2.99
2	GOOG	GOOGL	2.98
3	CTL	T	2.00
4	AMT	BXP	2.00
5	AIV	BXP	2.00
6	AEP	LNT	2.00
7	AEE	CNP	2.00
8	CHK	CVX	2.00
9	APA	APC	2.00
10	APC	COG	2.00

[1] "Bottom 10 Goodall Similarities"

	Var1	Var2	Freq
1	ABT	MMM	0
2	ABBV	MMM	0
3	ACN	MMM	0
4	ATVI	MMM	0
5	ADBE	MMM	0
6	AAP	MMM	0
7	AES	MMM	0
8	AET	MMM	0
9	AMG	MMM	0
10	AFL	MMM	0

Top 10 Goodall Similarity:

The highest similarity is observed between the APA COg and the GOOG GOOGL pairs which is consistent with the overlap similarity and close to 3 with the remaining pairs being equal to 2.0.

Bottom 10 Goodall Similarity:

Several ticker pairs have a similarity inverse frequency score of 0. Specifically, the ticker MMM has no overlap with the other bottom ticker pair which is consistent with the overlap similarity and inverse frequency similarity

Overall similarity between tickers by using mixed type data (choose a λ value for

calculation)

To evaluate the overall similarity between stock tickers, I used a mixed data type approach that includes both categorical and quantitative variables. The categorical variables were evaluated using the overlap similarity measure, while the quantitative ones were assessed using the Euclidean distance. I then combined these two types of similarities into an overall similarity score by using a weighted sum, parameterized by , which after several experiments I decided to use $\lambda=0.8$ as the quantitative cols are given more importance

Results -

[1] "Top 10 similarities:"

	Var1	Var2	value	rank
4020	CMG	CVX	5.94e+22	1
2508	AYI	CVX	5.94e+22	2
1182	AKAM	CVX	5.90e+22	3

```
4584  CVX  DNB 5.90e+22    4
4583  CVX DLTR 5.90e+22    5
4257  COO  CVX 5.89e+22    6
3842 CHRW  CVX 5.88e+22    7
1437 ALXN  CVX 5.88e+22    8
1268  ALB  CVX 5.85e+22    9
3648 CERN  CVX 5.85e+22   10
```

[1] "Bottom 10 similarities:"

```
     Var1 Var2    value rank
4200  COG   EA 6.89e+15 4950
913  ADSK CTAS 1.28e+16 4949
2861  BCR   EA 1.86e+16 4948
2831  BCR  COG 2.78e+16 4947
3742  CHD CTAS 3.31e+16 4946
4449 CTAS  EFX 3.74e+16 4945
3711   CF  DOV 3.74e+16 4944
940  ADSK  EFX 4.34e+16 4943
3453  CBG  DRI 4.36e+16 4942
1427 ALXN  COO 5.90e+16 4941
```

Top 10 Similarities:
The pairs of tickers with the highest overall similarity scores are primarily associated with CVX (Chevron Corporation). For instance, CMG and CVX have the highest similarity score, followed by AYI and CVX, AKAM and CVX, etc. This could indicate that these companies share a lot of similarities in terms of both their quantitative and categorical features.

Bottom 10 Similarities:
The pairs with the lowest similarity scores are more diverse and include various combinations of tickers like COG and EA, ADSK and CTAS, BCR and EA, etc. These pairs have the least similarity based on the metrics considered, implying they are quite different in their financial and categorical attributes.

Overall normalized similarity between tickers by using mixed type data (choose a λ

value for calculation)

In this analysis, we used a similar approach to the previous one but added an extra layer of normalization using the standard deviation of the variables. This makes the scale of the similarity values more interpretable. We chose λ=0.5 for this set of results, giving equal weight to both the quantitative and categorical features.

[1] "Top 10 similarities:"

	Var1	Var2	value	rank
628	ADI	CLX	171	1
250	AAPL	CLX	123	2
3977	CLX	CVX	119	3
645	ADI	CVX	117	4
201	AAPL	ADI	114	5
626	ADI	CHTR	111	6
603	ADI	AZO	107	7
604	ADI	BA	105	8
619	ADI	CCL	104	9
593	ADI	AN	103	10

[1] "Bottom 10 similarities:"

	Var1	Var2	value	rank
2339	AVY	CBG	0.0814	4950
2126	APH	CHD	0.1129	4949
3347	CAG	ECL	0.1559	4948
4898	DVA	ECL	0.1607	4947
4904	DVA	EMR	0.2055	4946
2161	APH	DOV	0.2092	4945
4928	ECL	EMR	0.2130	4944
4123	CMS	ECL	0.2264	4943
4129	CMS	EMR	0.2299	4942
2093	APD	ECL	0.2337	4941

I found similar results with lambda values of 0.1,0.7 and 0.9 in terms of ranking even though the scores changed. This indicates that euclidean distances are very high which is why it overpowers the categorical similarity weights even if it is at 90%.

Top 10 Similarities:
The tickers ADI and CLX have the highest similarity value of 171, making them the most similar pair in the dataset according to our metrics. Other notable pairs include AAPL and CLX with a similarity value of 123, and CLX and CVX with a value of 119. ADI appears frequently in the top 10 most similar pairs, which could indicate that this stock has multiple attributes that are commonly shared with other stocks.

Bottom 10 Similarities:

The least similar pairs include AVY and CBG, APH and CHD, and CAG and ECL, with extremely low similarity scores ranging from 0.0814 to 0.2337. These companies are likely very different in terms of their quantitative and categorical attributes.

Conclusion

The aim of the study was to evaluate the financial similarities and differences between various pairs of tickers, or company stocks using different similarity and distance formulas and algorithms and each providing its own unique insights.

Different Metrics for Similarity and Distance

Quantitative Similarity Measures

Lp-norm Distances
The study also utilized Lp-norm functions, focusing on norms like p=1, p=2, p=3, and p=10. Top and bottom Lp-norm distances showed remarkable consistency. Pairs like EA-COG and CTAS-ADSK often popped up as most similar, while Chevron (CVX) was frequently different from other companies like CHRW, DLTR, and CMG.

Feature Weights and Correlations
A **heatmap** revealed feature correlations, showing that "After Tax ROE" and "Pre-Tax ROE" were highly correlated, hinting at potential multicollinearity. Weights were then calculated, emphasizing key financial ratios, and these weights were **normalized**.

Minkowski Distance
When these weights were applied to the Minkowski distances at p=1, the top pairs still included EA-COG and CTAS-ADSK, confirming the findings from the Lp-norm distances.

Match-Based Similarity

Most and Least Similar: High-scoring pairs like ED & AEP and low-scoring pairs like DISCA & AAL showed that companies can be vastly different or quite similar based on multiple attributes.

Mahalanobis Distance

Closest and Furthest: Identical pairs had a distance of 0, while pairs like CLX and DAL had the highest distances, ranging around 11.54 to 11.71.

Categorical Similarity Measures

Overlap, Inverse Frequency, and Goodall: These similarities consistently showed APA COG and GOOG GOOGL to be the most similar pairs and the remaining top 10 with mostly matches with 2 of the 3 categorical columns. The bottom pairs were consistently paired with MMM showing that it is highly distinct from the rest and belongs to the GICS sector and sub-sector completely on its own.

Overall and Normalized Overall Similarity:

Overall Similarity:
In the first set of results, where we did not normalize the features, we observed that CVX (Chevron Corporation) appeared most frequently in the top 10 most similar pairs. This could indicate that CVX shares many attributes with other companies. However, the large scale of the similarity values suggested that normalization might be needed for a more interpretable analysis.

Normalized Overall Similarity:
After introducing normalization in the second set of results, the scale of similarity values became more interpretable. ADI appeared most frequently in the top 10, suggesting it shares various attributes with other stocks. Normalizing the features provided a more balanced view of the similarities and differences between the tickers.

Hence, this study offers a solid, methodological approach to understand financial and categorical similarities and differences among companies. Whether it's using match-based similarity, Mahalanobis distance, or Lp-norm functions, the results were pretty consistent. Some companies like EA-COG are almost like financial twins, while others like Chevron and MMM are the odd ones out. Overall this study was quite tedious but insightful.