Executive Summary

Objective

The main object of this project is to group similar companies based on a clustering analysis of the publicly trades stocks. The result is intended to facilitate an insight into the structure of the stock market.

The various elements of the project are summarized as follows:

a. To describe the analytic workflow

b. To describe the rationale of the methods and techniques used in the analysis

c. A summary of the results – **five clusters of listed companies**

Workflow

Dataset Exploration

- shape of the data
- summary statistics
- idenity missing values
- Distributions of some features(e.g. price)
- Correlations among features

Data Cleaning

- Reduce non-numeric features(e.g. Quarter end)
- Reduce feature with just one unique value(e.g. 'Split factor')
- Set the index as ticker
- Drop missing values

Feature Engineering

- create new features: *Debts/Asset*, *Cash from investing activities per share*, *Cash from operating activities per share*
- Features Selection

dimensionality reduction

- PCA
- Plot cumulative explained variance ratio

Clustering Analysis

- elbow plot
- silhouette score and plot
- Kmeans analysis
- results: 5 clusters

Methods used for cleaning data

Feature selection and drop NAs to deal with missing values

The summary statistics and the distribution plots for several of the features reveal the fact that the companies vary significantly in almost all the features, such as price. The distributions of the feature dataset are highly skewed. Besides, the features are obviously on different scales. Some features, such as Assets, are measured in U.S dollars, and the maximum number can reach 3.401105e+12, while some features are presented as ratios. In addition, the features in the datasets are financial

metrics that present the performance of a given company. Thus, imputing the missing values with mean/median might be inappropriate because it fails to represent the real financial performance of a given company and might cause problems. For example, the value of current assets might turn out to be large than the value of the assets.

Based on the above reasons, I decided to perform feature selection and remove the rows with missing values. I chose to keep the variables that demonstrate a decent value of correlations among all the important financial metrics. Also, I keep the variables that allow us to compare the performance of companies more easily, such as EPS basic and Book value of equity per share. This method allows me to set aside the less important variables and avoid those too similar variables, which will help us focus on the signals rather than the noises.

The benefits of creating new features

To better compare the similarities and differences between companies, I created three new features. They are:

- *Debts/Assets*: A high debt-to-asset ratio indicates higher risk and a less healthy financial structure.
- *Cash from investing activities per share*: It enables us to compare cash from investing activities across different companies with vastly different number of shares
- *Cash from operating activities per share*: It enables us to compare cash from operating activities across different companies with vastly different numbers of shares.

Techniques used for clustering analysis

PCA:

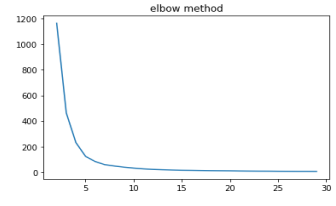
The dataset contains 12 variables after feature selections. As our goal is to group the companies and visualize the corresponding clusters on a 2-dimensional scatterplot, the principal component analysis is therefore considered an ideal method. In our analysis, two principal components can explain 80% of the variance of the processed data.

K-means:

Combining the k-means algorithm and PCA enables us to apply the data in the new feature space and then distinguish and label the different clusters.

K-means is a preferred clustering technique in our analysis based on three main reasons. First, we can specify the number of clusters. Given that the dataset demonstrates a significant amount of variance, we can determine the cluster number with elbow method and silhouette scores, without creating too many subgroups. Second, in the K-means method, the convergence is guaranteed, and it is specialized to clusters of different sizes and shapes. Third, k-means allows a better visualization in this case. Given that the pre-processed dataset still contains 559 observations that exhibit high variances, a scatterplot will be more intuitive and better visualization than a dendrogram in this scenario.

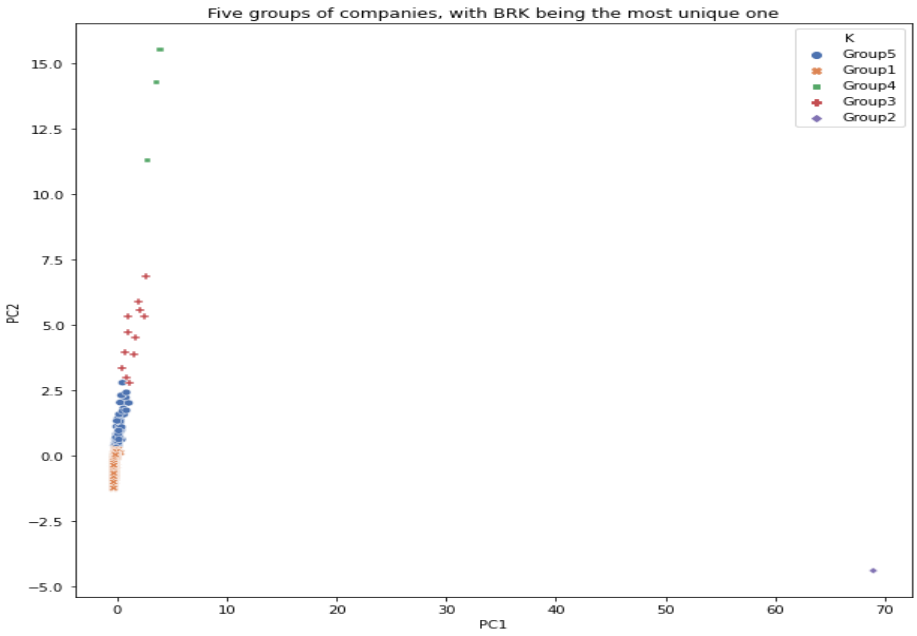
According to the elbow method, we consider the optimal number of clusters as 5.



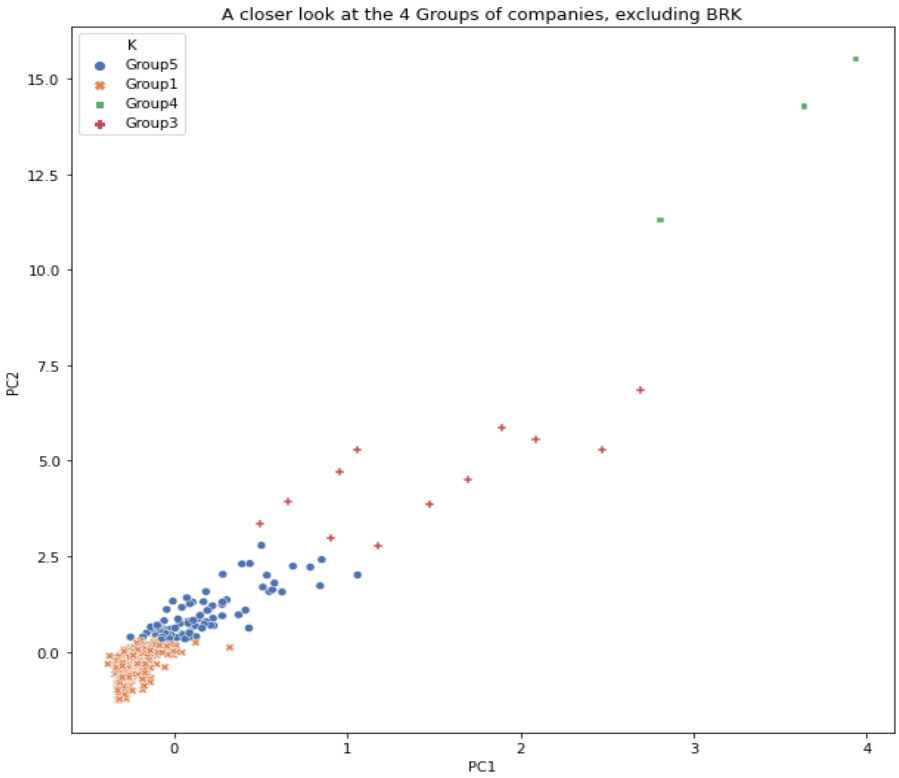
A summary of project results

The clustering analysis indicates that the 559 companies in our pre-processed dataset can be segmented into **5 sets of like-group.**

- Group1: 460 companies, including Paychex, Costco, and Kroger
- Group2: 1 company- Berkshire Hathaway
- Group3: 12 companies, including Morgan Stanley, Goldman Sachs, and Verizon
- Group4: 3 companies, including JPMorgan Chase, Bank of America, and Wells Fargo
- Group5: 83 companies, including ORCL, Walgreens, and Comcast



The scatterplot on the left visualizes the 5 sets of like groups among 559 companies. The group2 only contains 1 company ---Berkshire Hathaway Inc. And it is noticeably much farther away from other companies, indicating a huge difference between BRK and other companies. The stock price of Berkshire Hathaway Inc is $308,080, much higher than the average stock price of $525 in the dataset.



A closer look at the other 4 groups of companies, excluding BRK, tells us that **Group1 and Group5** are more similar to each other, given that these two clusters are much closer. Most of the companies fall into these two groups. The two groups have similar level of "Book value of equity per share" (around 35) but shows different levels of average Debt/Assets ratios.

In comparison, group 3 and group 4 are quite different than other clusters and have more variability within the group. For example, Group5 demonstrate a higher average Debt/Assets ratio than other groups.