Clustering Stocks in The Bases of Risk Factors

By: Alfredo Martinez
Mentor: Lucas Allen
February 2018

Introduction

Purpose:

1. Build a diversification system which groups stocks base on similar data and risk values.
2. Build a K-Means model which best fits each stock to others in similarity.
3. Find a strong amount of cluster which can best represent the data on hand.

Data Sets

- Quandl
  - Consisting of 3,996 different company ticker codes
  - 14 Quantitate variables for each ticker
- Nasdaq
  - Contains 2,457 different North America Stocks
  - 8 Quantitative and qualitative variables
- Combine Data File
  - 6,553,424 historical prices data points
  - 24 different qualitative and quantitative variables

Feature Selection and Engineering

1. Market Capitalization
2. Years Publicly Traded
3. Dollars Traded
4. Average Daily Return
5. Average Volatility
6. Average Sharpe Ratio
7. Average Dividend Yield

Other Potential Datasets

Quandl

- Premium members can get access to several different financial statements and ratios from companies

Nasdaq

- Data with companies in different exchanges, regions, industries

Limitations

- Usage of North America Companies only
- Single year data (2017)
- Lack of different industry for companies

Preparing Data for Modeling

Split data for train and test methods

- Check to see if clusters amount is still appropriate for test/unseen/new data.

Data Scale

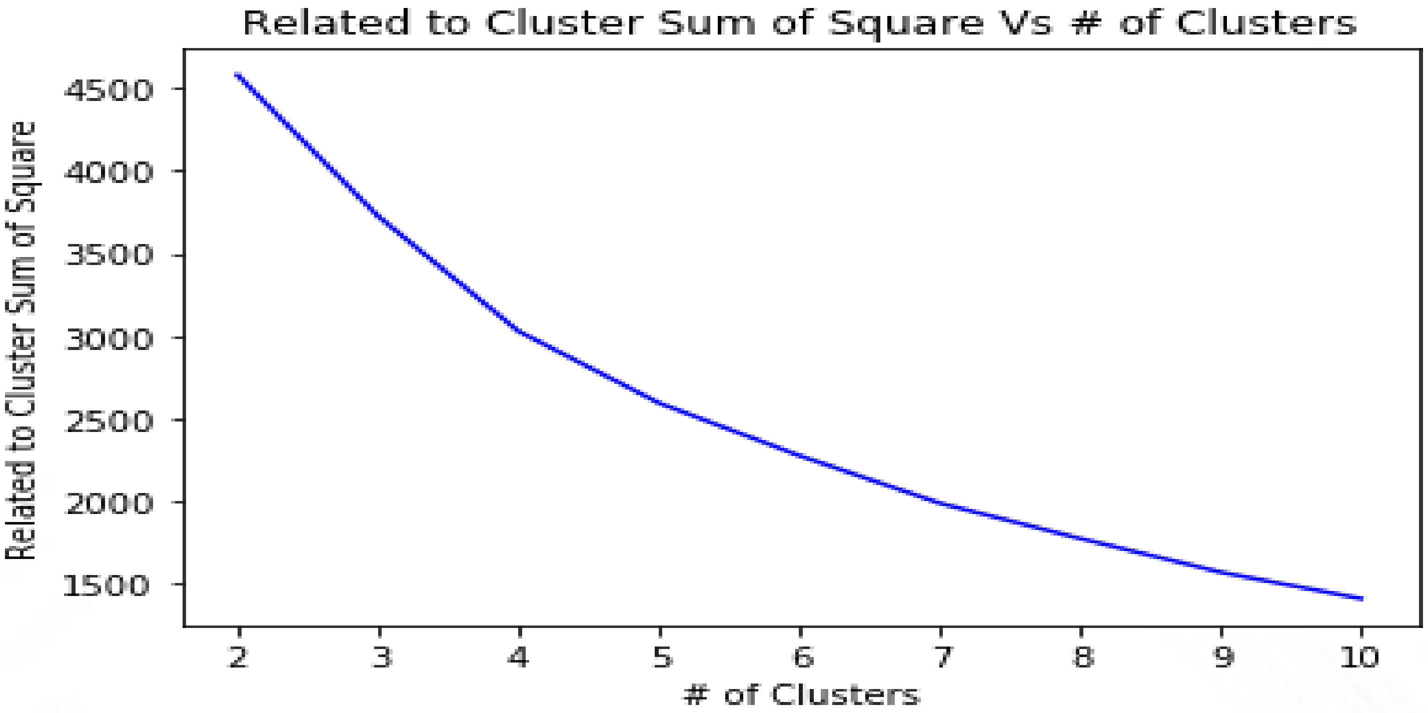
- Levels up the different magnitudes of numerical values in our features for better analysis

Setting Optimal Number of Clusters

Different Methods

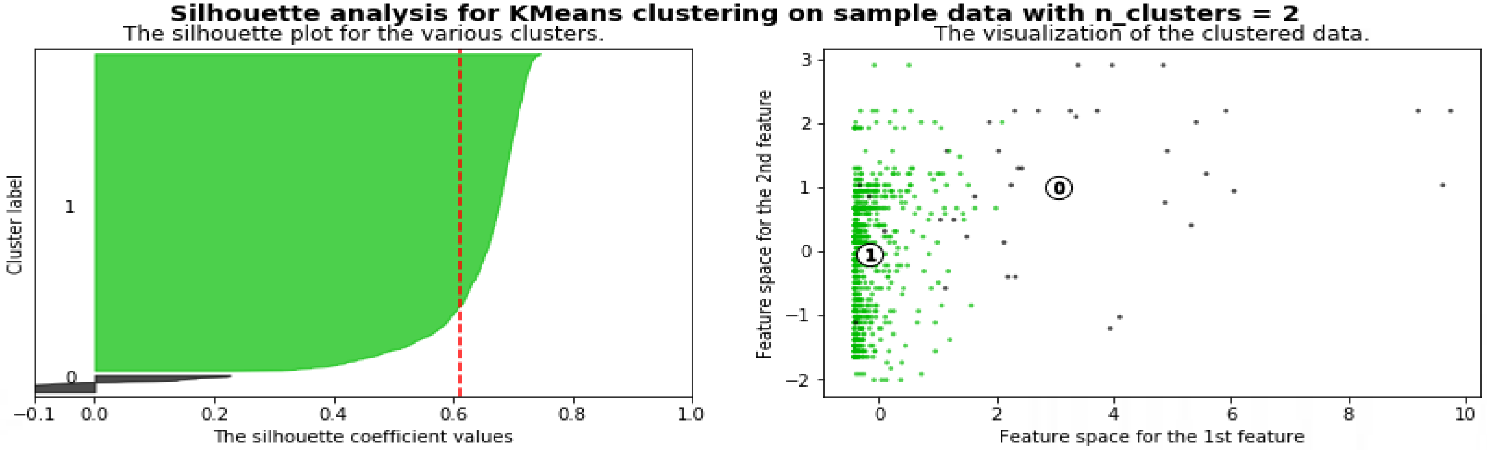
1. Elbow Method
   - Find a point where the addition of another cluster doesn't offer a much better model
2. Silhouette Method
   - Grades quality in similarity from data points to their given cluster

Elbow Method



The analysis was not very successful at identifying a good elbow

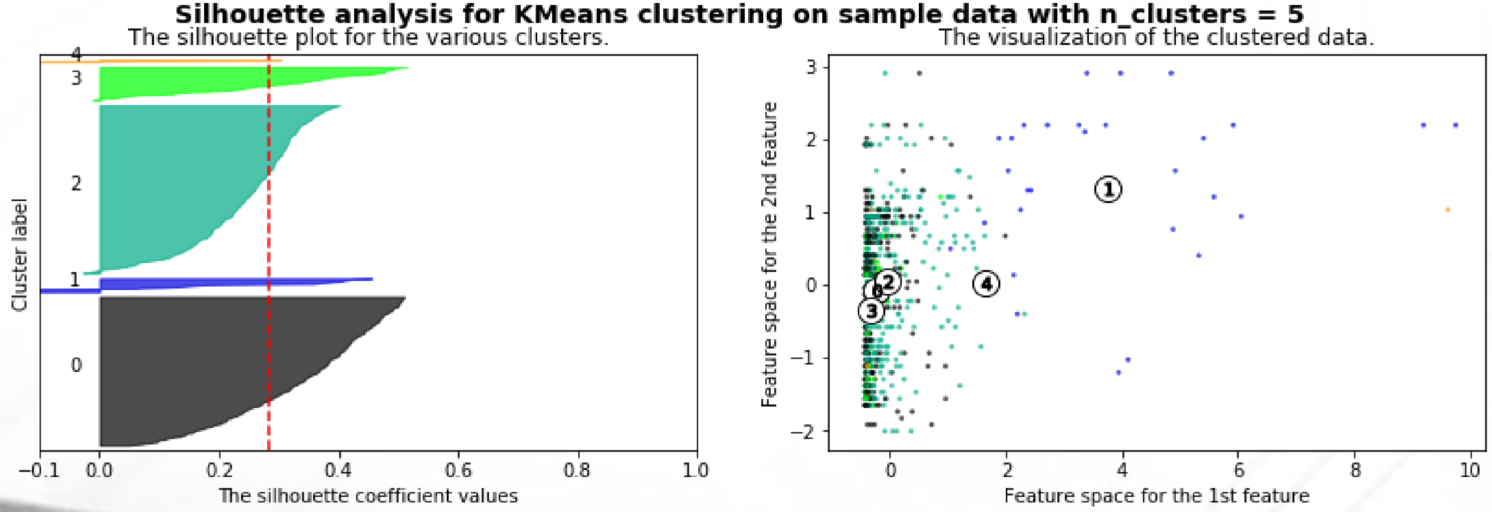
Silhouette Method



Silhouette analysis for KMeans clustering on sample data with n_clusters = 2

The silhouette plot for the various clusters.

The visualization of the clustered data.



Silhouette analysis for KMeans clustering on sample data with n_clusters = 5

The silhouette plot for the various clusters.

The visualization of the clustered data.

Silhouette Method Continues..

2- cluster

Analysis gives the highest coefficient grading values. However, grouping all of our data in two groups does not become very useful and practical.

5- cluster

From all other options, the five cluster grouping gives a better distribution of out data and a fair amount of quality in similarity from data points to its clusters

K-Means

The algorithm use for the analysis was base on K-means:

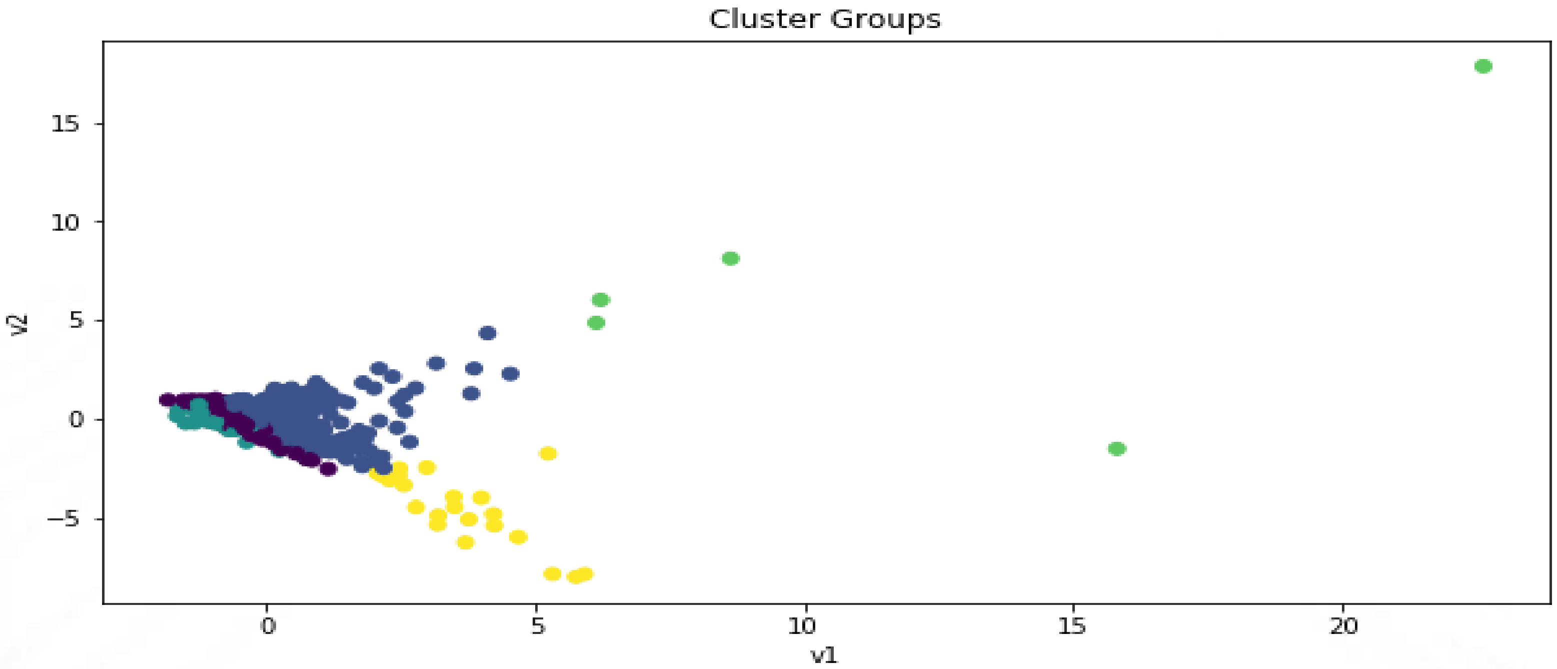
- It chooses a random point as the initial mean value for the number amount of k(clusters)
- Cluster are build by relating each point with its nearest mean.
- Finds new mean values within new clusters group values and becomes the new centroid.
- The cycle continues until convergence(same result) shows

PCA for Data Visualization

Principal Component Analysis

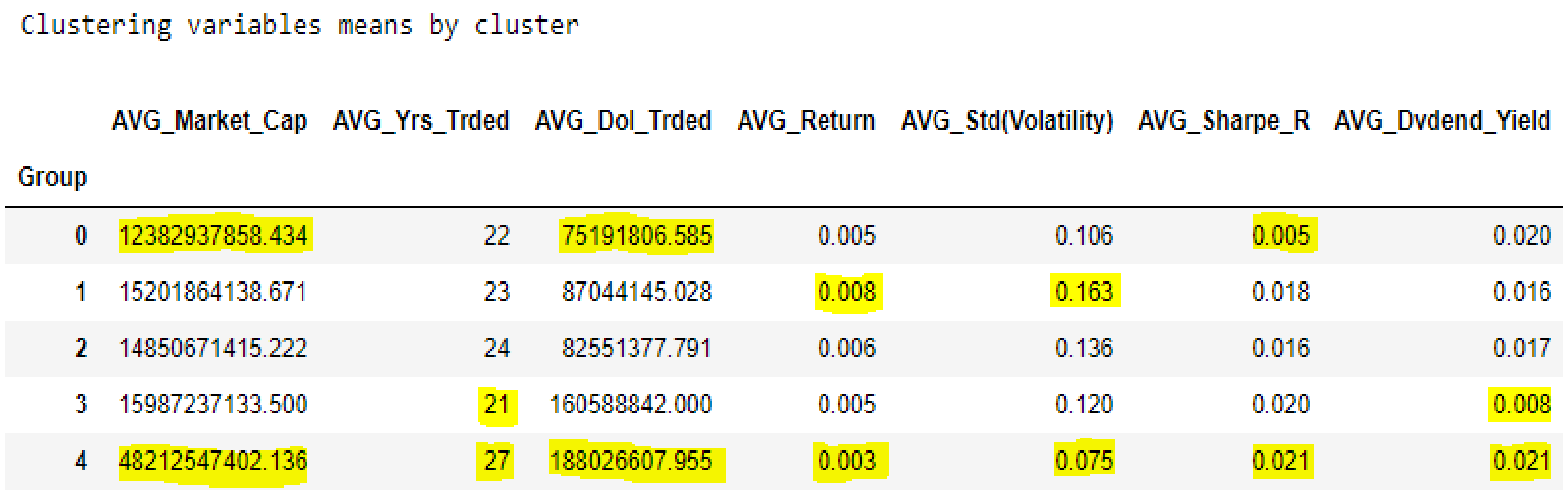
- Process consist of bringing all the features into a more compact manner in which the data can be better analyze.
- It provides a minimum amount of variation in features data and than assign to their corresponding target values(stock labels)

PCA for Data Visualization Continues..



The plot show some groups with a strong difference, while some are more similar within their neighboring groups.

Features’ Averages by Group



```
Clustering variables means by cluster
```

Group	AVG_Market_Cap	AVG_Yrs_Trded	AVG_Dol_Trded	AVG_Return	AVG_Std(Volatility)	AVG_Sharpe_R	AVG_Dvdend_Yield
0	12382937858.434	22	75191806.585	0.005	0.106	0.005	0.020
1	15201864138.671	23	87044145.028	0.008	0.163	0.018	0.016
2	14850671415.222	24	82551377.791	0.006	0.136	0.016	0.017
3	15987237133.500	21	160588842.000	0.005	0.120	0.020	0.008
4	48212547402.136	27	188026607.955	0.003	0.075	0.021	0.021

- The highlighted results in the table above show some of the main difference between the clustering groups build within our data
- The groups can be used by different investors as way to classify and better analyze a particular investment in compare to others investments

Thank you!