By: Alfredo Martinez
Date: 2/8/2018
Mentor: Lucas Allen

Clustering Stocks for Portfolio Diversification and Management

Introduction

Asset selection and allocation are critical challenges financial institutions and investors deal with every year, as the market changes, assets move from one place to another, continually creating new portfolios. A popular method used today when it comes to minimizing risk and maximizing return comes from building a well-diversified portfolio of stocks. An approach to cover the challenges of making a well-diversified portfolio and finding risk levels on assets is through clustering of shares based on historical risk and other metric values.

Purpose

The purpose is to group stocks base on similar data values, so these stocks can be pick from within the given groups by the clustering model, giving the investor an easy, time effective way to find a clearer picture on how a stock compares to others as far of risk levels and other key features.

Client

The clients are general market investors who want a data-driven system that can quickly analyze and give back specific details about different stocks.

Dataset

The two different data sets used came from Quandl and Nasdaq websites.

Quandl

The dataset consisted of 3,996 different codes, each representing a different company stock from different countries. There were 14 quantitative variables present for each share

Nasdaq

The dataset included 2,457 different North America Stocks consisting of 8 different quantitative and qualitative variables

Combine Dataset: Open-Filter-Combine-Historical Stock Prices

After all available company codes were open, the historical prices data set consisted of 6,553,424 data points and 24 different qualitative and quantitative variables.

Cleaning and Wrangling Data

The early first steps to working with both of the data files consisted of building each of them separately into data frames. The data frames were joined into one single filtered data frame comprising of all North America stock companies that were within both datasets. After doing this, the new data frames returned a value of 1,135 stocks companies.

For the new data frame, a for loop was used to iterate and open all the historical prices and quantitative variables available for each stock. After obtaining the new processed data frame, this was saved to the local disc as a (.csv) file and reopen to a new data frame for faster access.

For the new file, the most significant steps taking to clean our data consisted of:

1. Drop the unnecessary columns for the type of analysis.
    a. Certain features in the document had almost no values (NaN), others had irrelevant data or merely added much value to our end goal analysis.
2. Build the 'Date' column into a DateTime format.
    a. The column was originally set as object values and was changed to a time series datetime64[ns] value to set and keep track of the historical data for each company.
3. Build feature 'publicly traded years' for each company.
    a. The feature was built before any other one as we needed the historical dates for each company.
4. The model and analysis was built to represent values for the year 2017
5. Slice 'WIKI/' out of ticker values, change features with scientific values to integer values.
    a. The step was done mainly for static reasons and to have a better visual of the values at hand
6. Calculate the average daily dollars traded, daily return, daily standard deviation, daily Sharpe ratio and dividend yield for each stock.
    a. Each feature gives a stock's data value concerning market risk.
7. Check feature values and set to the right format.
    a. The data points for Market Capitalization and Average Daily Returns were converted to float values as these were a more suited format for what they represented.

The results gave a data frame table with an index and seven different columns, from which index represented the stock label for each company, and the seven columns consisted of the built features relatively to market risk.

Other Potential Datasets

As cover before, the data use for this work came directly from Quandl and Nasdaq websites. Both of the sites had different types of data which could potentially be used within the system as well. (e.g.) Nasdaq website at the moment contains other tables which can be directly downloaded and also used; some of these include files representing publicly traded companies in different exchanges, regions and industries. On the other hand, Quandl gives access several other records either at a premium price or free. (e.g.) Files with publicly traded companies financial ratios and statements.

Limitations

Before further going through the analysis, it is important to point out that there are certain limitations into the work done in these project, but can certainly be extended to a closer representation of the entire market analysis. Some of the key limitations include:

a. The Usage of only North America companies in our analysis
b. A single year of data(2017), a year where the market was up
c. No examination of the market segment/industry for these companies

Analysis of Data/Model Build Up

Data Split

Splitting our data was performed as to test whether the number of set clusters is still appropriate to unseen data. We split our data into a train (70 % of data) and test (30 % of data) objects.

Data Scale

To have our data work correctly in the clustering model we needed to scale our data since some features in the datasets had different magnitudes of numerical values.

The steps taking were performed in both our train and test data and consisted as follow:

a. Turning data frame's features columns into a matrix of values.
b. Input the new matrix with values into our preprocessing scaling function. For this dataset, we used the standard Scaler with gives each feature a mean value of 0 and a variance of 1.

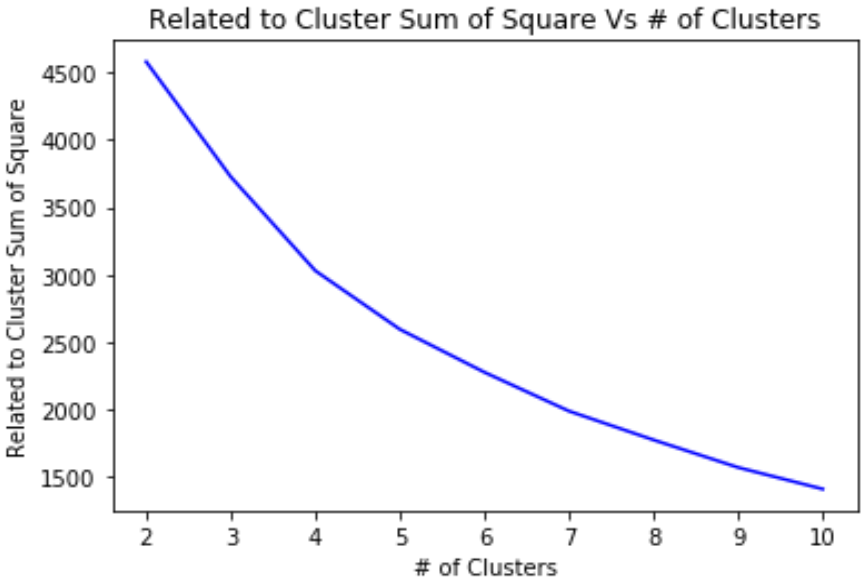
Setting Number of Clusters

Elbow Method

The elbow method looks at our data and finds a point (number of clusters) where the addition of another group doesn't offer a much better model. The process looks at the amount of variance (sum of square in our graph) that can be explained (individual group variance/total variance) within the subsequent addition of more clusters. Ideally, finding a clear breakpoint where the addition of more groups doesn't make much difference.

Through this analysis, as seen in our image (1) below we were not able to obtain a clear elbow value. A definite drop in gain by adding more cluster can easily be set to be at either four, five or any other number of groups.

(Image1)

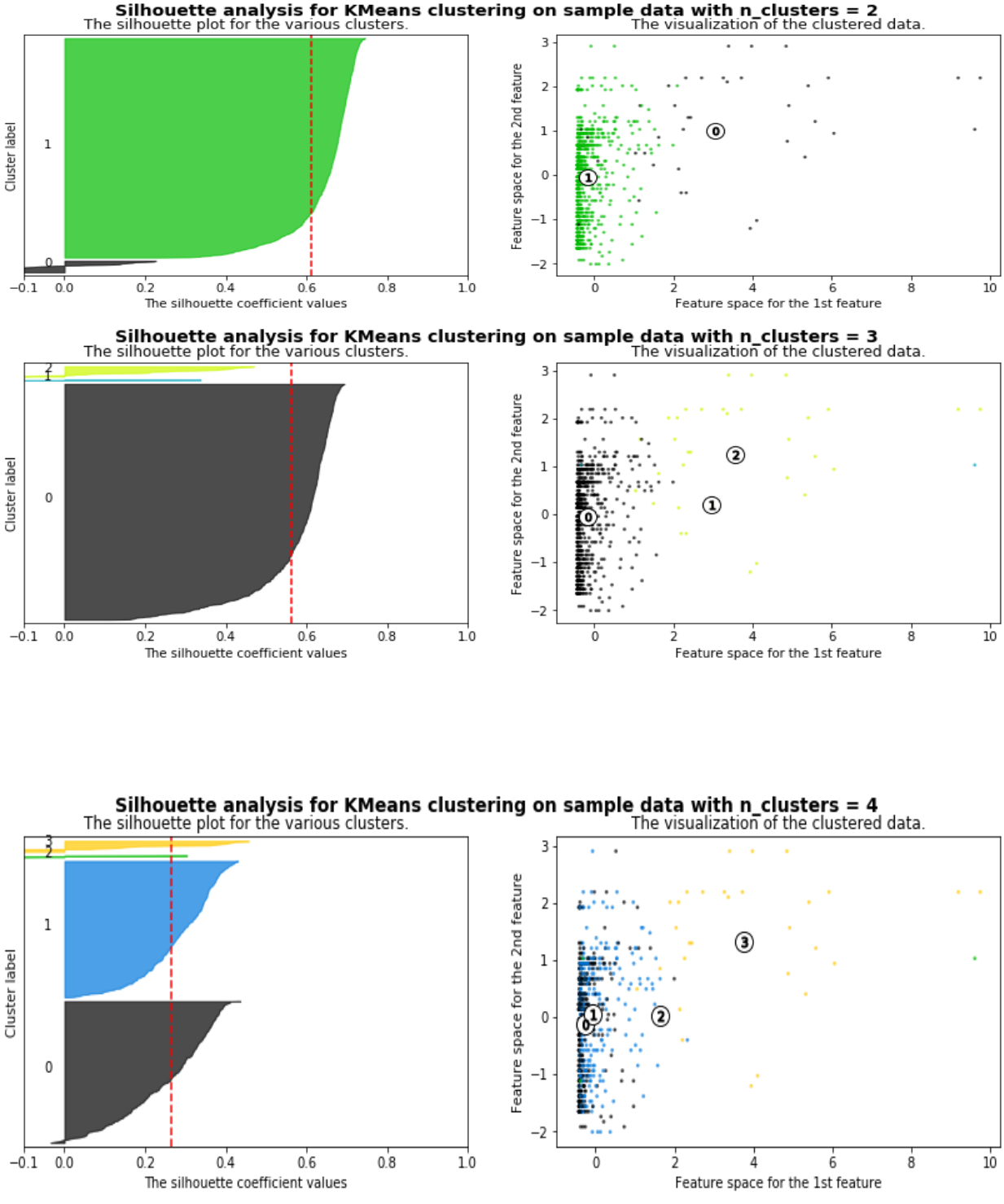


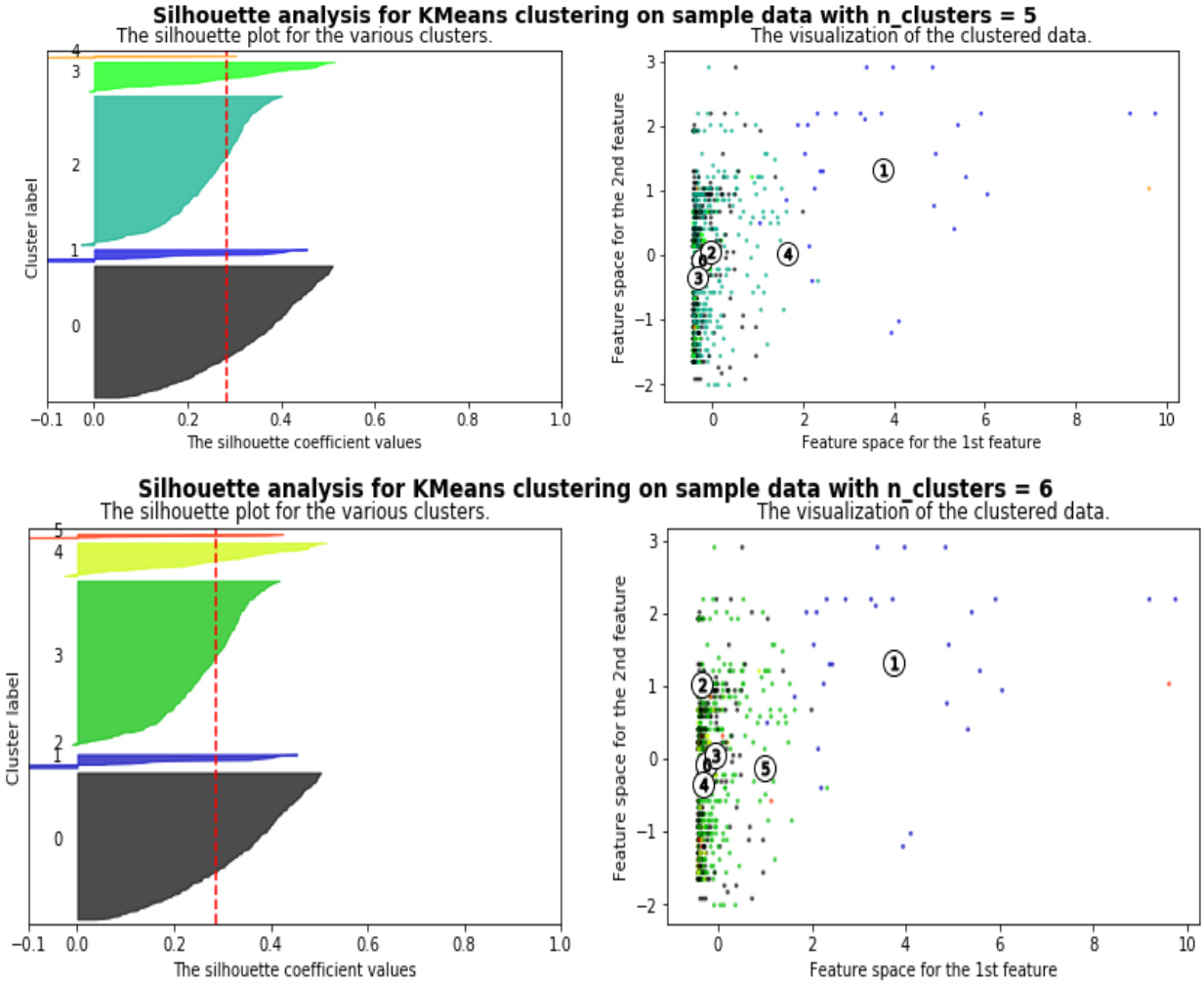
Silhouette Method:

The elbow analysis results led to further investigate and apply The Silhouette Method as a second option in finding a good optimal number of cluster for the analysis. The silhouette analysis calculates the similarity amount between a data point and its corresponding group against other clusters. The process grades the similarity from a scale from -1 to +1, where the closer the given values to one, the higher the similarity the data point has to its cluster, and the more erroneous match has to the neighbor cluster.

After running and looking at the results (image 2) from different tests with different set amounts of clusters to analyze, it gives the two-cluster group the highest silhouette coefficient (quality of clusters). However, the two clusters silhouette test captures most of the data points in only one group, which data cannot be taken as a good human perspective or being practical, and the same can be said for the three cluster analysis. Although, by looking at the rest of the groups, some tests offer a better data distribution within clusters such as a five cluster, which also show a light breaking point during the elbow method. So for the mention reasons, we conclude that for our analysis and the data in hand, a five cluster model will be the best fit.

(Image 2)


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.
Cluster label
The silhouette coefficient values
Feature space for the 2nd feature
Feature space for the 1st feature
Silhouette analysis for KMeans clustering on sample data with n_clusters = 3
The silhouette plot for the various clusters.
The visualization of the clustered data.
Cluster label
The silhouette coefficient values
Feature space for the 2nd feature
Feature space for the 1st feature
Silhouette analysis for KMeans clustering on sample data with n_clusters = 4
The silhouette plot for the various clusters.
The visualization of the clustered data.
Cluster label
The silhouette coefficient values
Feature space for the 2nd feature
Feature space for the 1st feature



Model

K-means

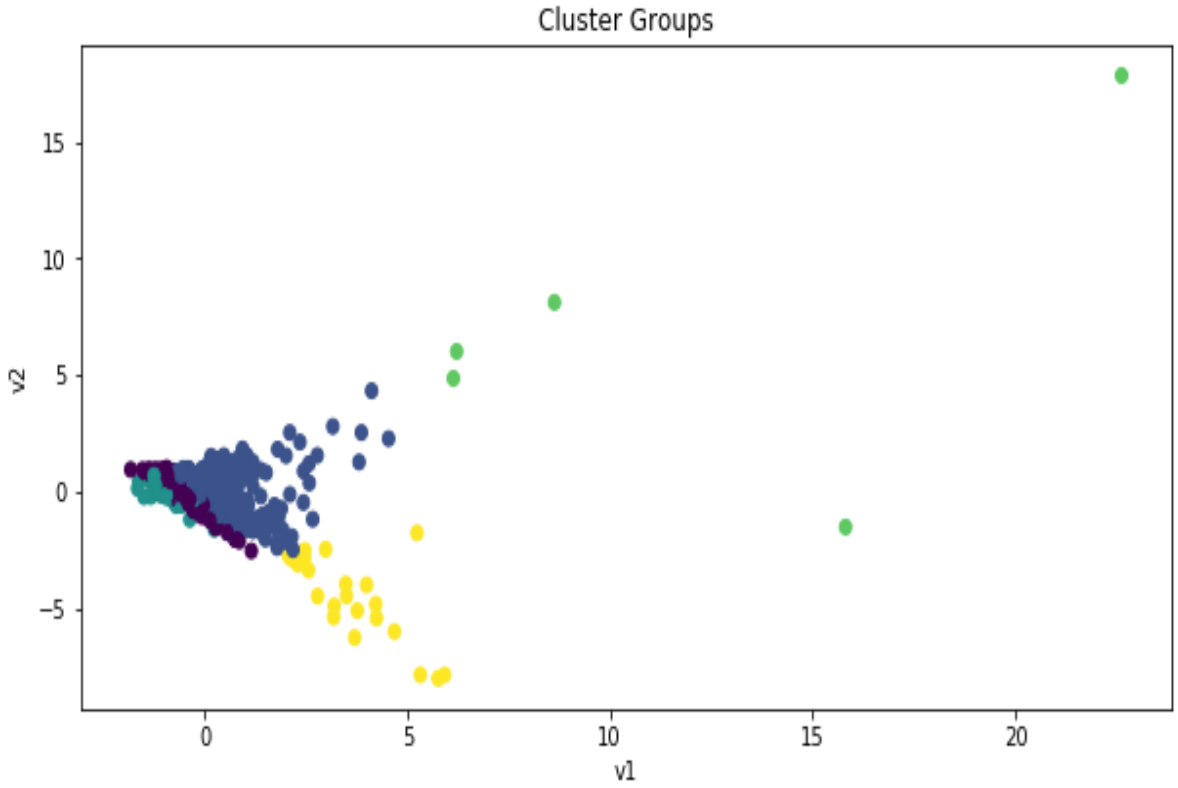
The K-means algorithm was chosen to group our data. The way its chooses to build the groupings is by randomly selecting initial points for the amount of k-means (clusters), after this the groups are build base on the points nearest to the previous randomly chosen means, than a new mean values (centroid) is calculated from the data points on each group, and the process keeps repeating until we reach convergence (same results/group values).

PCA

Principal Component Analysis is mainly use for either speed the machine learning system algorithm to be use or for data visualization. In our case, the method will be use to better visualize our data by reducing the dimensionality. The process consist of bring the several data features into a more compact manner in which the data can be better analyze. The data is bring into a two component analysis which represent the two main components of variation(providing a minimum amount of

variation in features data) and then this are assign to their corresponding target values(stock label), which can then be more easily visualize.

Groups Visualization:



The plot does show some groups with a strong differentiation, while some of the clusters seem to have an overlap within some of their group values.

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

Clusters Analysis Base on Feature Means

Group 0

This group is represented as the one with the lowest average market capitalization and dollars traded, which might not be as easy to buy and sell, meaning they have a lack of liquidity, giving investors an idea on how fast they can expect to get their money back. Also, being the group with the lowest market capitalization show for a good significant room for growth in investments overtime.

Group 1

The group is represented as the one with the highest average daily return and average daily standard deviation (volatility) value. Stocks in this group are categorized as the highest ones in return on investments, but also the riskier concerning the change in price throughout the year.

Group 2

The stocks in this group shows an overall mid-level values for most stocks features. These can be seen as the best investments for those who want a mid-risk investments in which the stock fall in a neither too risky nor safe type of investment with a decent average return on investment and dividends pay out.

Group 3

On this group, we see the lowest average of years publicly traded and dividends amount. Assets in this category have the lowest track record doing business in the market and don't have much to offer regarding dividends to customers.

Group 4

Stocks in the group had the highest averages in market capitalization, years publicly traded, Sharpe ratio and dividends pay out, also the lowest average in return and volatility. The shares in this category are big size companies which have a long history in the market, are very liquid. However, they show the lowest returns, but do offer low volatility and high performance in accordance to risk and return perspective from the Sharpe ratio and a tendency to pay a good percentage of dividends.