DECISION 520Q Data Science for Business

Team C42 - Final Submission

Tracy Pham, Fengquan Sun, Devet Valecha, Yi Zhao & Tony Zitong Gao

Industry: Asset management

Dataset: The New York Stock Exchange dataset (https://www.kaggle.com/dgawlik/nyse)

Hypothesis: It is possible to select an optimal portfolio of stocks from the NYSE that outperforms the S&P 500 using purely financial information extracted from SEC-10K reports.

Business Understanding

The problem of interest is whether stock portfolios constructed based solely on firms' fundamentals could outperform the S&P 500 in a particular year (2016). Traditionally, investors seek diversification benefits by considering factors such as industry segmentation, company sizes, and asset classes. However, in this report, optimal portfolios are built purely on public information extracted from financial statements. Given the Modern Portfolio Theory (MPT) pioneered by Harry Markowitz in 1952, efficient frontiers are constructed to maximize investors' expected return at a given level of risk preference. Furthermore, given the presence of risk-free investment, tangent portfolios are also constructed to provide the best return and risk tradeoff to investors independent of preferences. All assumptions of the Portfolio Theory hold throughout this study, the U.S. 10-year treasury yields are used as a proxy for risk-free returns. This study applies unsupervised methods to facilitate clustering analysis and portfolio selections.

Data Preparation

Datasets have been cleaned and transformed with the operations outlined below:

1. Selected key datasets including:

- Prices_sa: Split-adjusted stock prices and volume data of all NYSE listed companies
- Fundamentals: Key accounting and financial metrics extracted from annual reports (mostly 10-K)
- Rf: Proxy for risk-free returns calculated on a daily basis
- SP_daily: Daily returns of the S&P 500 index from 2012 to 2016

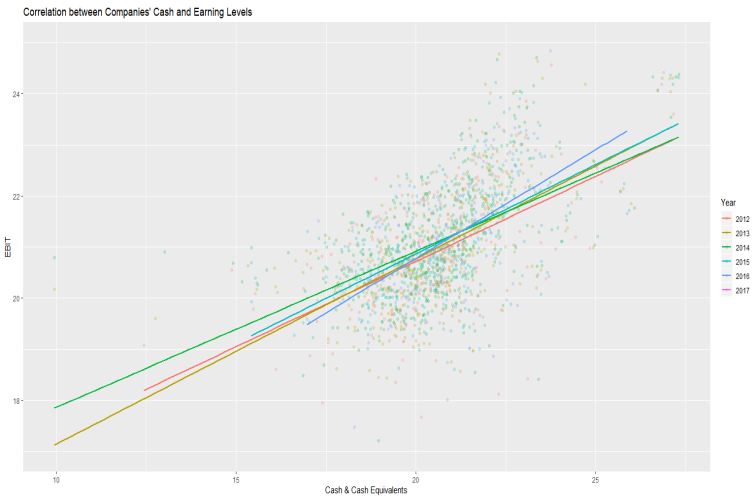
2. Examination of variables and records:

- Prices_sa: Deleted irrelevant variables, removed 34 stocks with insufficient price information, rearranged dataset formats, calculated daily log-returns
- Fundamentals: Deleted all insignificant financial metrics and confirmed the final 46 variables

The final stock returns dataset contains 467 stocks across 1762 trading days with complete information on companies' financial information and daily log-returns.

Data Understanding and Exploration

Cash and earning levels are crucial to the company's performance, both information is provided in the "fundamentals" dataset. A positive correlation is noticed in the graph below. Companies are the most efficient in 2016 in terms of generating profits from available cash flows[1].

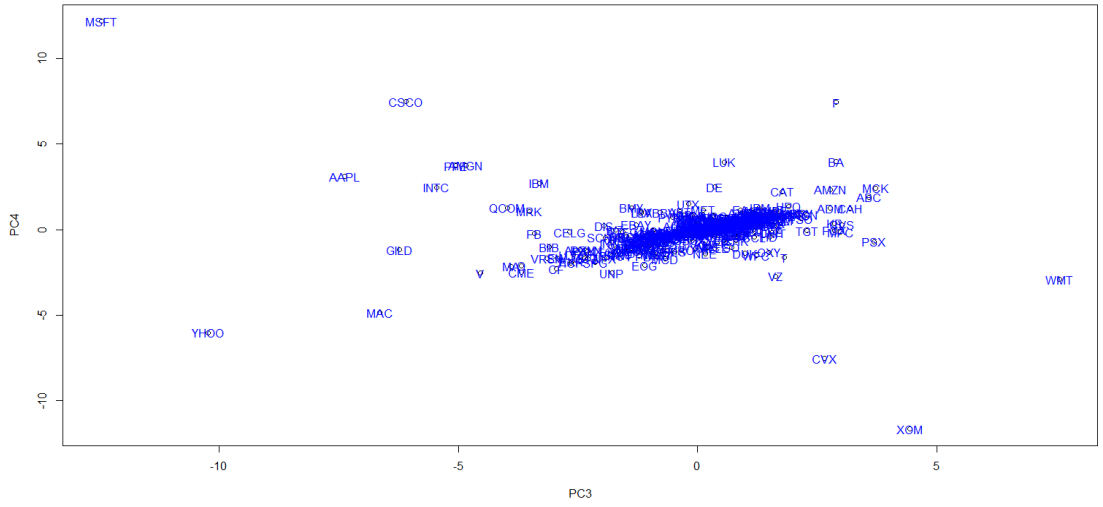


[1] High-resolution graphs are shown in the Appendix section below.

In order to reduce variable dimensions and gain a better understanding of each company's unique characteristics, the study applies the Principal Component Analysis (PCA) across each year between 2012 and 2015. The top six PCs are selected given that they explain over 75% of the total variance from the dataset. Each PC represents some latent features that a group of companies possesses and the interpretation of PCs are different across each year due to varying economic cycles and company-specific situations. As an example, a summary of latent features in 2014 are illustrated below:

Year	PCA	Interpretation	Company Examples
2014	1	Great performance, high earnings & cash flows	JPMorgan, Apple, ExxonMobile
2014	2	Low debt level, minimal long term investments	Chevron, Microsoft, Walmart
2014	3	Low operating margins, efficient production	Walmart, Phillips 66
2014	4	Heavily invested in R&D	Microsoft, Cisco, Ford
2014	5	Low score: High ROE, efficient operations	Wynn Resorts, Allegion
2014	6	Low AR, high Interest payments & debt levels	Verizon, HCA Healthcare

Above interpretations are derived from the calculation of PC loadings as well as the dimensionality reduction plots. In particular, PCs are graphed in the pairwise order and below is an example that outlines the relationship between PC3 and PC4 in year 2014:

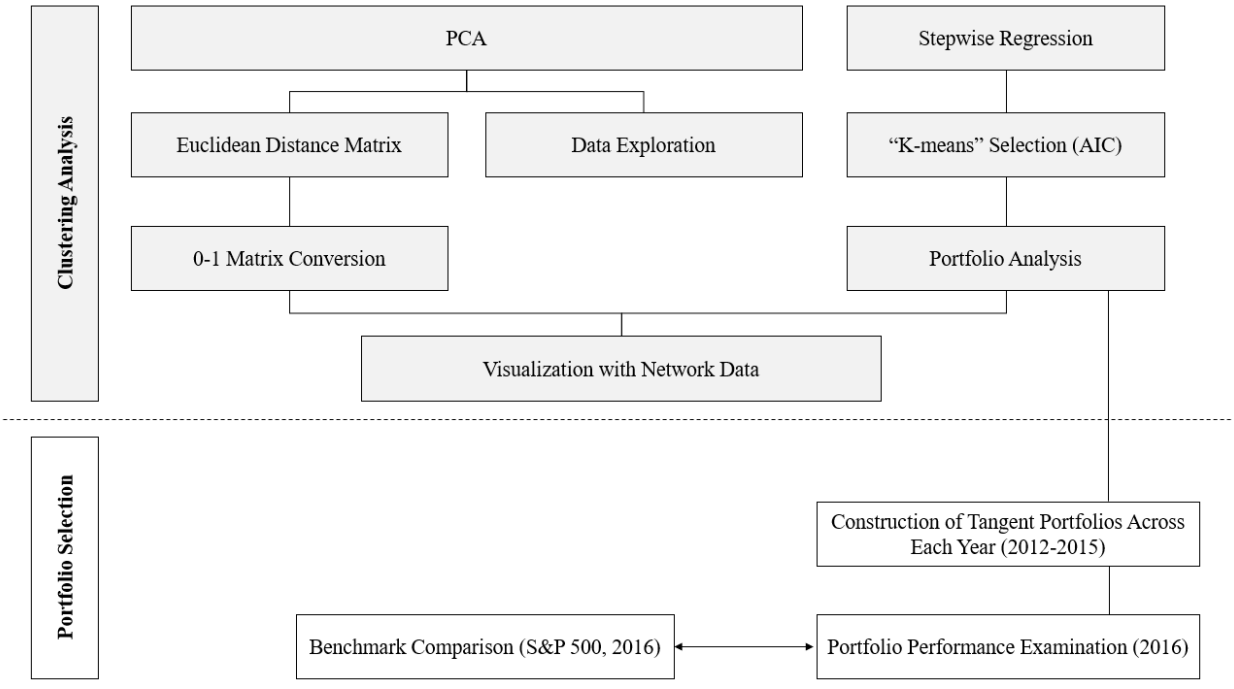


The PCA assesses the company's operational performance using a combination of financial metrics. The outcome is further analyzed in the Modelling section of the report where the Euclidean distance is calculated based on each company's six-dimensional coordinates.

Modelling of Portfolio Selection

The key methodology used in this study is the K-means Clustering Analysis from a family of unsupervised machine learning techniques. The model construction follows a set structure that is outlined below. Key steps include variable selection, "k-means" selection, visualization of multidimensional observations, integrated insights with network data and portfolio derivation using average returns and standard deviations. Key assumptions include no transaction and agency costs are involved with portfolio investment.

Structure of Model Construction



Phase 1: Clustering Analysis

a. Explanatory Variable Selection

Stepwise regression from both directions is used for automatic selection of regressors which are correlated with the response variable - *Net Cash Flow for Operating.* Taking outcomes from the regression, the significance of variables is further assessed based on accounting equations and consideration of multicollinearity. With an AIC of 50604.13, there are 11 variables remained including: *Accounts Payable*, *COGS*, *EBIT*, *Gross Profit*, *Long-term Debt*, *Long-term Investment*, *Net Cash Flow for Financing*, *Net Cash Flow for Investing*, S*hort-term Debt to Long-term Debt Ratio*, *Retained Earnings and Short-term Investments.*

b. "K-means" Selection via Information Criterion (IC)

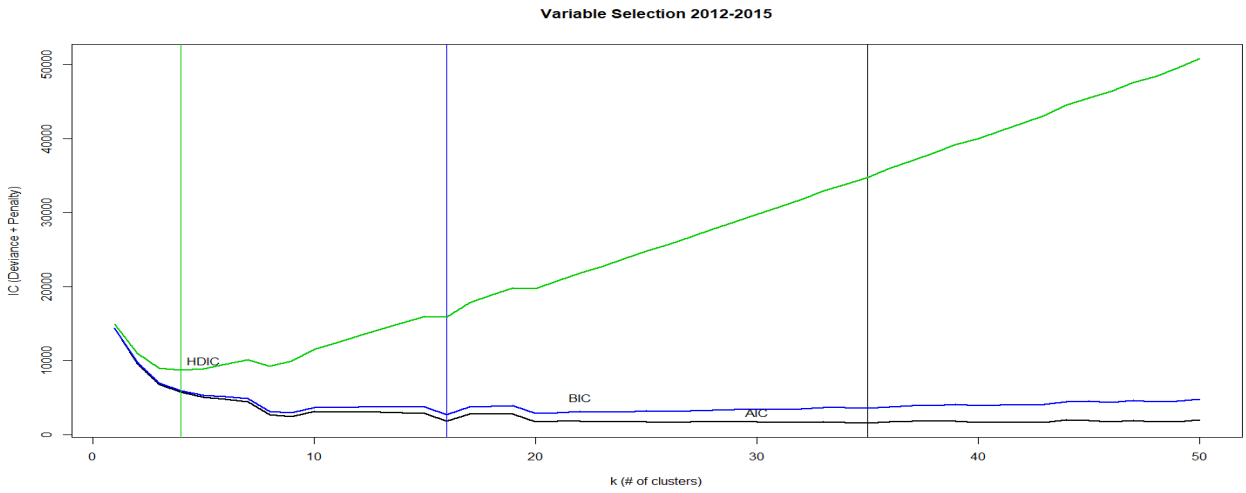
The selection of "k" centres matter greatness to the overall fitness of the model due to concerns of overfitting and model bias, A single cluster produces high bias, where as "k=n" clusters lack of flexibility in out-of-sample usage. In order to regularize "k-means", Information Criterion (IC) is applied by minimising the deviance function and results are shown below:

$$\min_{k} Deviance\ of\ using\ k\ centres + \lambda\{\#\ of\ clusters\} * \{\#\ of\ features\}$$

AIC = 35, bic = 16, HDIC = 4

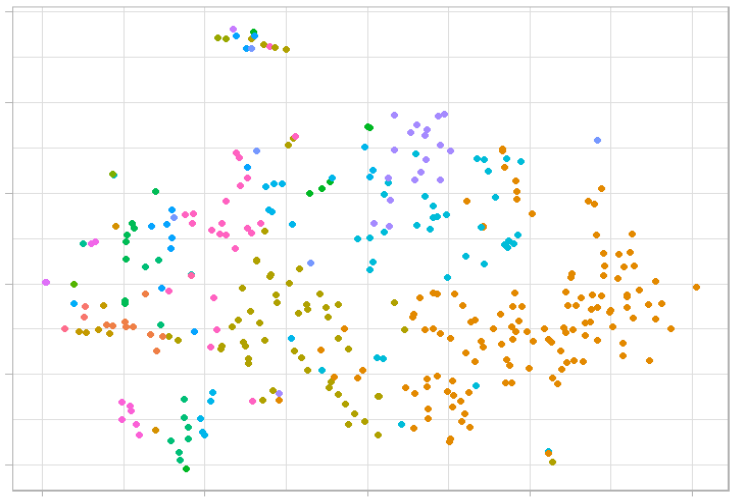
In general, R provides three options of IC and for the purpose of this study, AIC = 35 is used for the selection of "k" centres[2].

[2] Given that the initial k cluster centers locations are assigned randomly, the selection process is deemed to cause inconsistency in IC application. Consequently, results for portfolio selection may differ depends on the selection of k centres.



c. Clustering Analysis via Dimensionality Reduction Techniques

Clustering visualization is done by t-SNE[3] plots given the 11-dimensional coordinates assigned to each company. This method allows the 11-dimensional objects to be projected on a 2D graph. While it is hard to interpret the axes and distance among data points, the structure of the data is preserved. In other words, companies in the same cluster remain close proximity with each other.. The 11-dimensional clusters could therefore be visualized as shown below (2013):



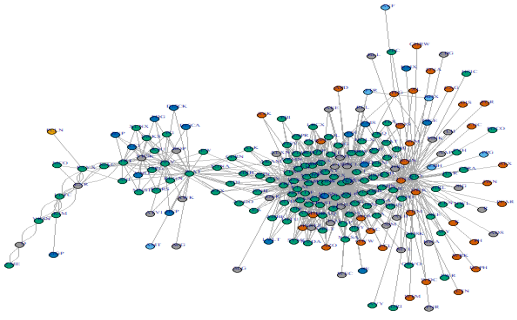
[3] t-SNE is a machine learning algorithm for the visualization of multi-dimensional datasets.

d. Network Data Visualization

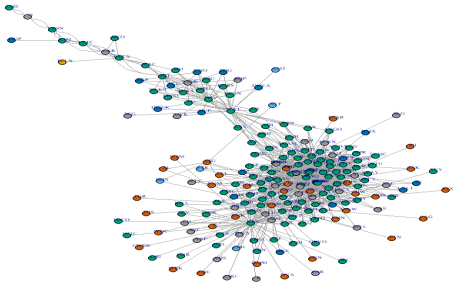
PCA converts a set of correlated variables into several principal components. Six principal components are utilized to define the fundamentals of each company from 2012 to 2015. The distance matrix is built based on the Euclidean distance[4] between two companies, each with a six dimensional coordinate. Threshold[5] is set to transform the distance matrix into a 0-1 matrix.

The next step is to combine network data with k-means clustering by applying different colors to the nodes on the graph, with each color corresponding to one cluster. The nodes represent companies selected by the distance threshold and edges represent the connections between each company. Specifically, if there is an edge between two nodes, the distance between the two companies is less than or equal to 1. The companies that have no connection with others are left out. This is also another way of visualizing the 11-dimensional k-means clustering: If the nodes from the same cluster are connected to each other, it reinforces the clustering analysis because companies within one cluster are also close to each other according to the six PCs.

Network - 2012



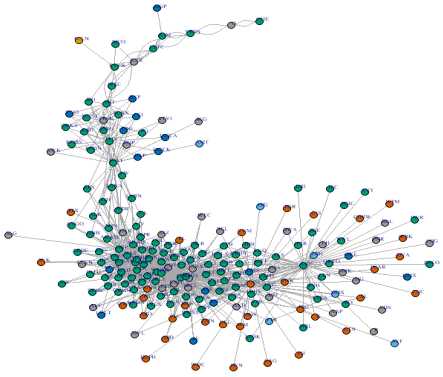
Network - 2013



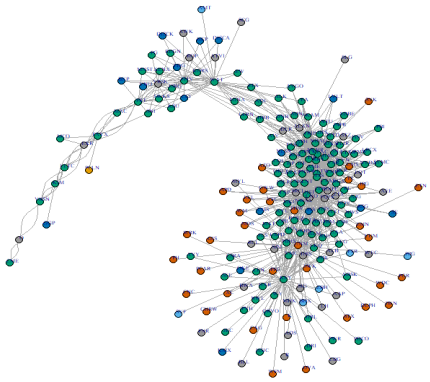
[4] Euclidean Distance: euclidean distance $=\sqrt[2]{(a-b)^T(a-b)}$, where a and b are vectors

[5] Distance = 1: when the distance is less than or equals 1, then it is assigned 1 in the 0-1 matrix to show their closer relationship

Network - 2014



Network - 2015



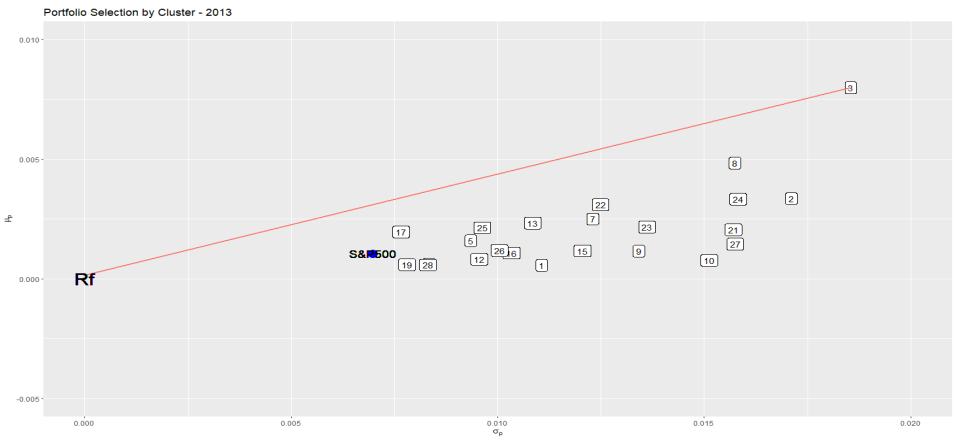
Phase 2: Portfolio Selection

a. Construction of Tangent Portfolios (2012-2015)

Every cluster represents a portfolio, hence for each year there are 35 portfolios differing in size. In the presence of the risk-free asset, company weights are calculated by deriving the tangency portfolio using matrix manipulation. Based on the optimal weights, portfolio performances are evaluated by mean-variance analysis.

b. Selection of Optimal Portfolio by Sharpe Ratio

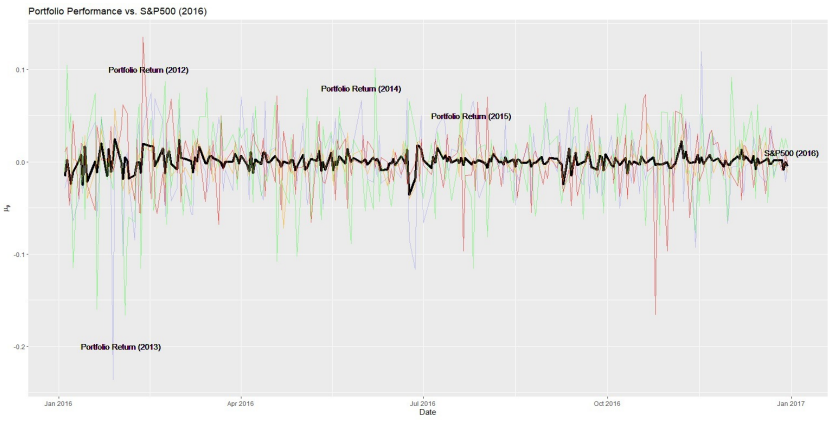
For each year, taking the average return and standard deviation into consideration, the optimal portfolio in that year among all tangent portfolios in each cluster is selected based on highest Sharpe Ratio . A list of companies that belong to each year's optimal portfolio is attached in the Appendix section. Example from 2013 is shown below:



Model Performance Evaluation

Four historical optimal portfolios, based on PCA analysis, k-means clustering analysis and mean-variance analysis, are chosen to simulate their daily performance in 2016, to see how true it holds in 2016. Basically the daily clustering portfolio returns are calculated through $R_p = w_1x_1 + w_2x_2 + w_3x_3 + \ldots + w_tx_t$, where weight w_i is got from mean-variance analysis and x_i denotes the daily returns of the company.

Daily returns of S&P 500 index in 2016, which is the leading stock index due to its broader scope, is utilized as a benchmark for validation with the four optimal portfolios.



The graph above validates the fact that S&P 500 has a short-term mean return of approximately 0 and a smaller volatility compared to chosen portfolios, given its wider coverage of companies. Mean of each portfolio's daily return and daily variance are shown as below:

	Portfolio (2012)	Portfolio (2013)	Portfolio (2014)	Portfolio (2015)	S&P500 (2016)
Return	-0.0005498826	0.0004406561	-0.001679459	0.0005518815	0.000397619
St. Dev	0.0314922747	0.0362790837	0.043535788	0.0142937545	0.008252354

Both portfolio 2013 and portfolio 2015 outperform the S&P in 2016. Four portfolios all have a higher standard deviation compared to S&P. Since there is only one company in the tangent portfolio in portfolio 2015, portfolio 2013 is the best after the simulation.

Solution Deployment

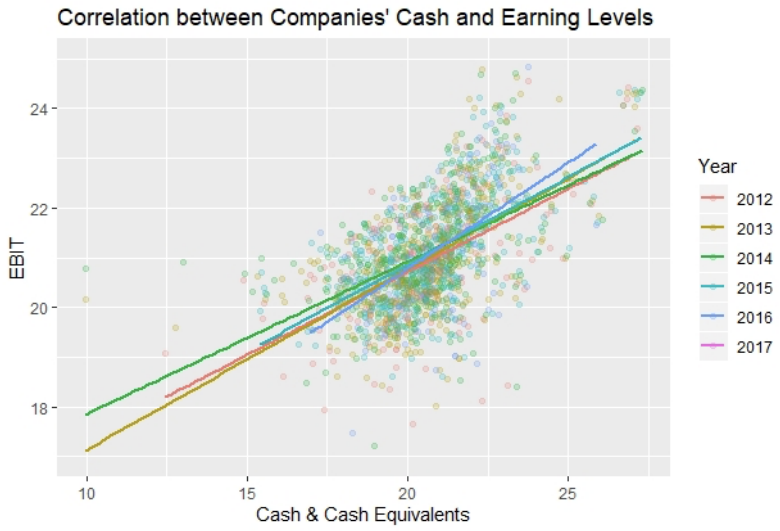
This project illustrates that judgement based on companies' historical fundamentals after clustering actually can provide some insights on investment. Though with higher variance, the four portfolios constructed based on historical data can achieve more perk returns than S&P, and the daily return of two portfolios even outperform the S&P.

The project proposes a way to selecting optimal portfolio based on historical optimal portfolios. Using clustering method and financial analysis, the portfolio creation process requires less effort than traditional portfolio management methods. Furthermore, machine learning eliminates the influence of biases by asymmetric information.

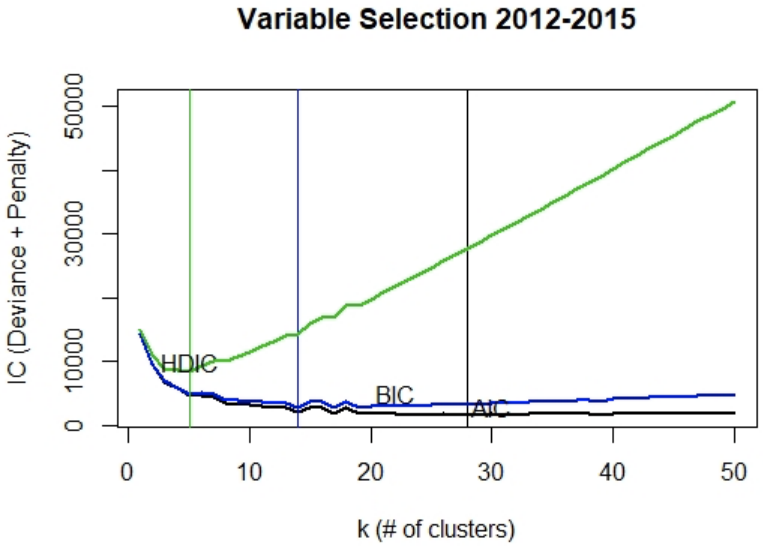
However, there exist several potential risks associated with optimal method and need further improvement. In this project, transaction fees are excluded from modeling for simplicity. However, transaction in risky asset incurs proportional transaction costs, which would in turn affect the total return. Second, while this analysis sounds sufficient to assist with diversification, there is no strong evidence that financial profile is a good indicator to define the similarity among stocks. Also, further analysis could be conducted to test the algorithm within each industry segmentation of stocks. The time period for the proposal from this project is one year. Multi-period portfolio selection ought to be considered in further research. Last but not least, the method of k-means in clustering since k-means may well generate some problems because it basically generates a set of random clusters somehow each time. It is possible that one set of cluster are not representative and would no doubt cast some concern to the model.

Appendix

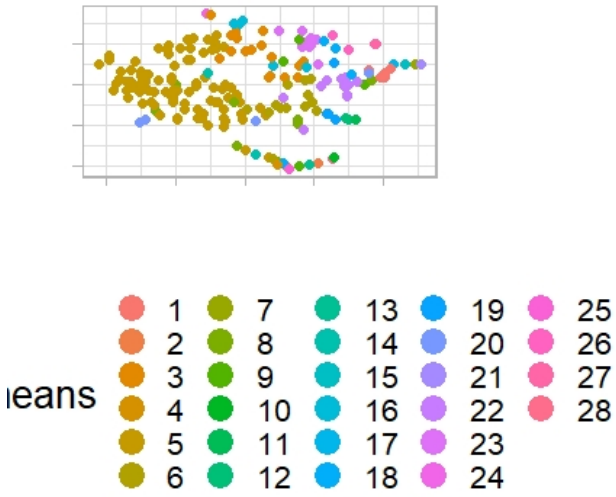
Correlation between companies' cash and earning levels



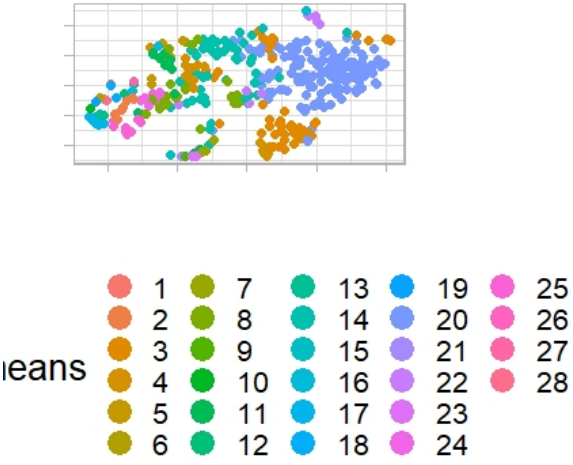
K selection via Information Criterion (IC)



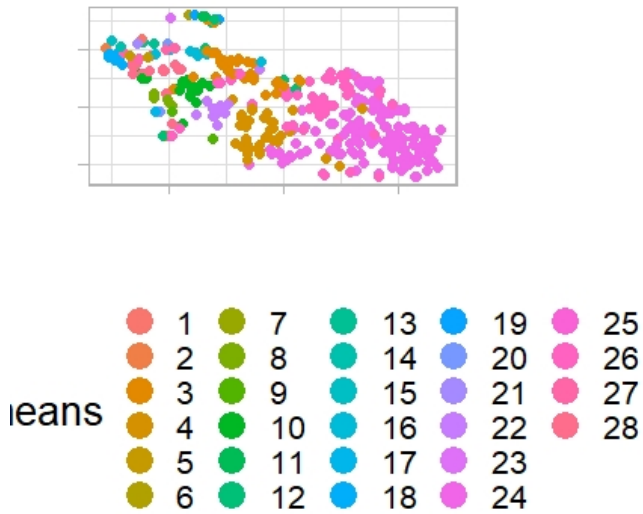
Dreduction Plot - 2012



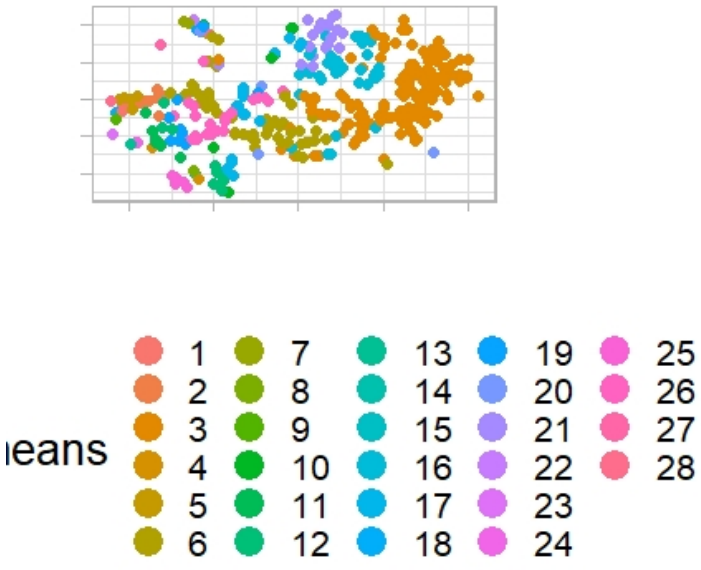
Dreduction Plot - 2013



Dreduction Plot - 2014



Dreduction Plot - 2015



Optimal Portfolio Company List - From 2012 to 2015

```
> opt12
 [1] CME  ZBH  UHS  CF   GWW  ATVI DGX  HSY  MAT  VFC  ECL  SHW  ORLY GPC  YUM  TXT
448 Levels: AAL AAP AAPL ABBV ABC ABT ADBE ADI ADM ADS ADSK AEE AEP AFL AIG AIV AIZ AJG AKAM ALB ALK ALL ALLE ALXN AMAT AME AMG ... ZTS
> opt13
 [1] O    HST  TDG  XL   SLG  HCP  EQIX CXO  AWK  VNO  WEC  ADS  KMX  IRM  WYNN ES   SYMC DLR  FIS  SCG  VRSN FTR
[23] BXP  AEE  R    WY   ATVI LVLT CMS  RCL  DVA  NLSN NFX  WYN  GGP  DISCA DISCK RSG  CPB  NBL  ZTS  SEE  LB   BSX
[45] CNP  MAS  MU   AJG
448 Levels: AAL AAP AAPL ABBV ABC ABT ADBE ADI ADM ADS ADSK AEE AEP AFL AIG AIV AIZ AJG AKAM ALB ALK ALL ALLE ALXN AMAT AME AMG ... ZTS
> opt14
  [1] WLTW TSS  KLAC QRVO WAT  FFIV TRIP O    LLTC MCHP RHT  EXR  VRSK COO  SNI  VRSN FRT  EFX  PAYX IDXX KIM  CMG  SRCL TDC  IFF  DNB
 [27] FLIR ULTA SWKS AKAM ISRG FAST HST  DPS  VMC  GPN  MLM  MNST KORS AIV  URBN FBHS USB  TDG  HOLX UA   UAA  XLNX TIF  SLG  ADSK ALLE
 [53] CTXS EQIX LEN  ILMN DRI  KSU  ROP  EW   LNT  ALB  ADI  ALXN EQT  NDAQ WEC  SJM  HP   AYI  MTD  PG   REGN XL   COG  CHD  JBHT BCR
 [79] XRAY PDCO PNW  TGNA PKI  AWK  IRM  RHI  SEE  RRC  LH   CTAS SYMC AME  LKQ  LEG  SIG  DLR  VAR  CXO  XEC  FIS  HRB  FSLR XYL  HAS
[105] LRCX GRMN MKC  PNR  CHRW EA   ATVI WY   CLX  ALK  HRL  EXPD MAC  WRK  APH  HRS  ZTS  R    PWR  MAT  NFX  MHK  WYN  MJN  FMC  MAS
[131] COL  FLS  HAR  TAP  NWL  BWA  HSIC BLL  PBI  CBG  JEC  EXPE
448 Levels: AAL AAP AAPL ABBV ABC ABT ADBE ADI ADM ADS ADSK AEE AEP AFL AIG AIV AIZ AJG AKAM ALB ALK ALL ALLE ALXN AMAT AME AMG ... ZTS
> opt15
[1] MSFT
448 Levels: AAL AAP AAPL ABBV ABC ABT ADBE ADI ADM ADS ADSK AEE AEP AFL AIG AIV AIZ AJG AKAM ALB ALK ALL ALLE ALXN AMAT AME AMG ... ZTS
```

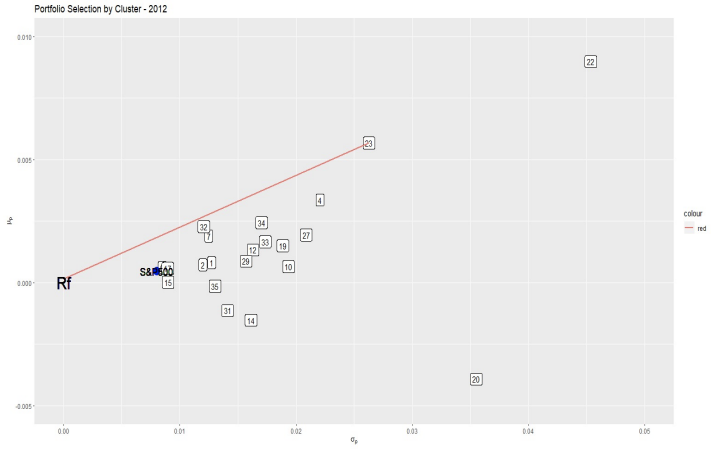
Optimal Portfolio Company List - 2016

	Portfolio (2012)	Portfolio (2013)	Portfolio (2014)	Portfolio (2015)	S&P500 (2016)
Return	-0.0005498826	0.0004406561	-0.001679459	0.0005518815	0.000397619
St. Dev	0.0314922747	0.0362790837	0.043535788	0.0142937545	0.008252354

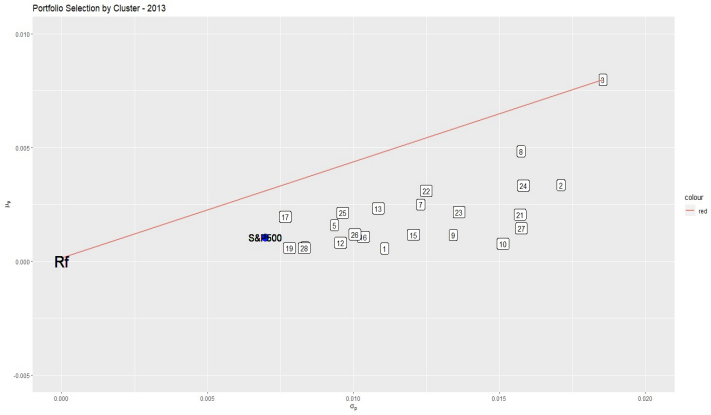
Min IC for K selection

```
> which.min(kaic)
[1] 28
> which.min(kbic)
[1] 14
> which.min(kHDic)
[1] 5
```

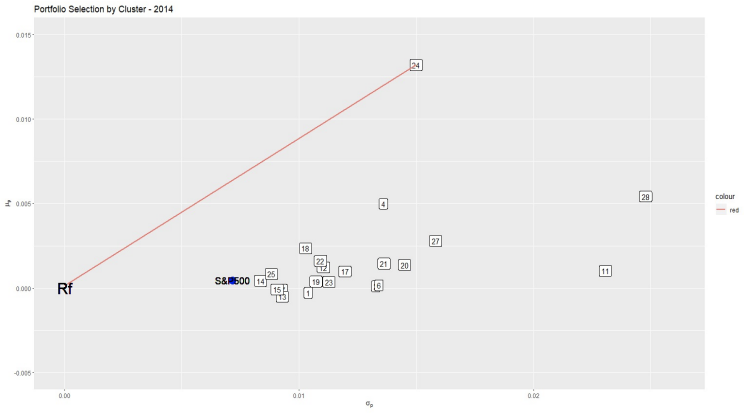
Optimal Portfolio Selection - 2012



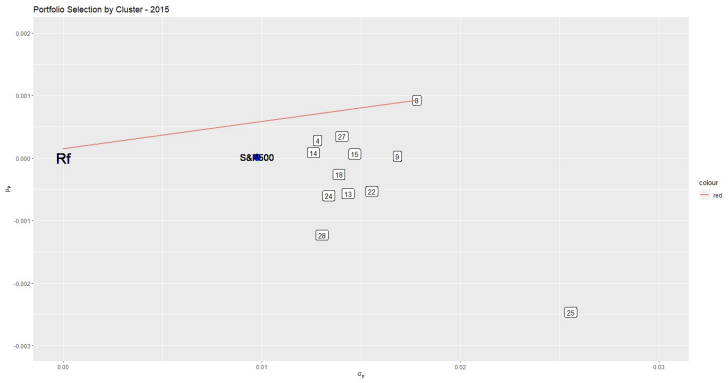
Optimal Portfolio Selection - 2013



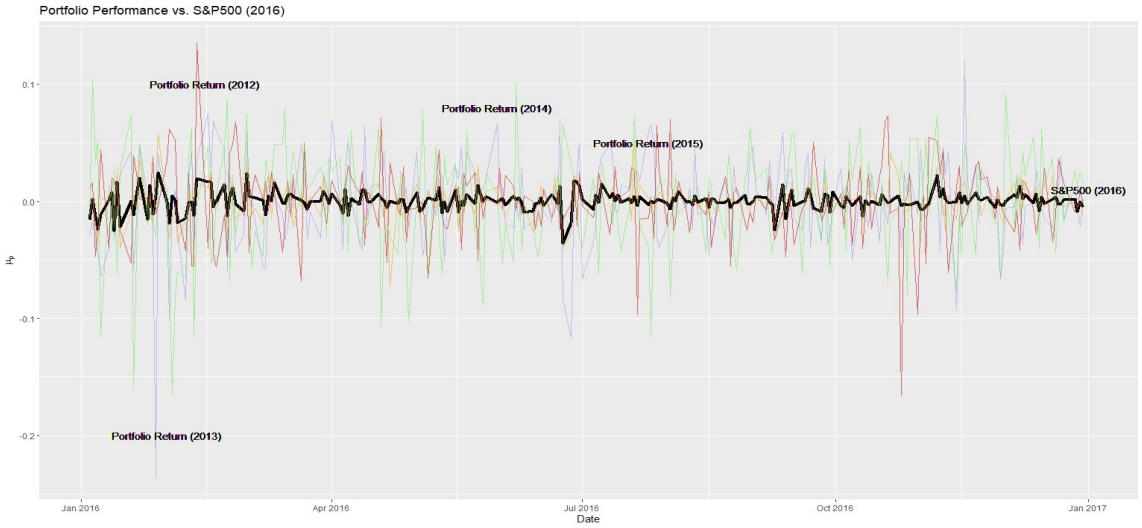
Optimal Portfolio Selection - 2014



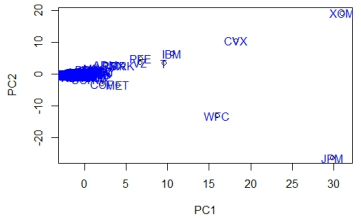
Optimal Portfolio Selection - 2015



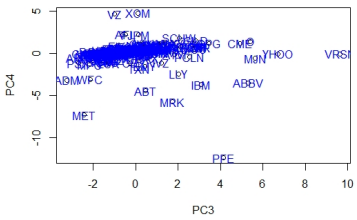
Portfolio Performance Simulation in 2016



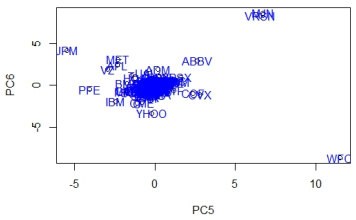
Distribution of Principal Components across years



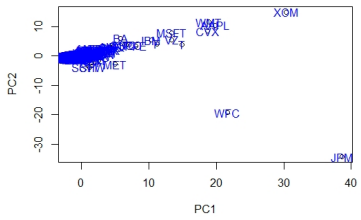
PCA 2012, PC1 vs PC2



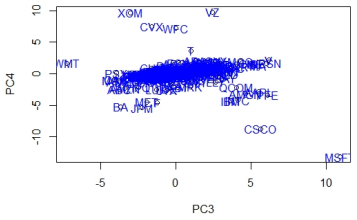
PCA 2012, PC3 vs PC4



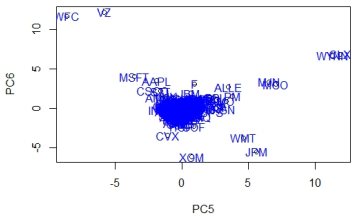
PCA 2012, PC5 vs PC6



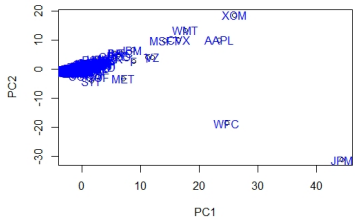
PCA 2013, PC1 vs PC2



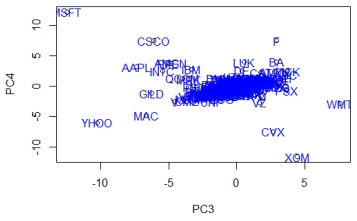
PCA 2013, PC3 vs PC4



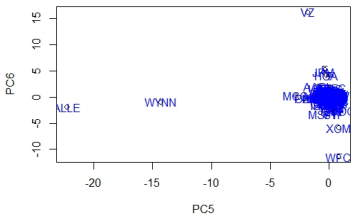
PCA 2013, PC5 vs PC6



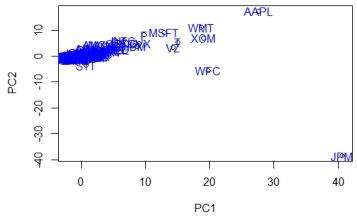
PCA 2014, PC1 vs PC2



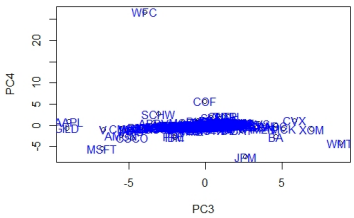
PCA 2014, PC3 vs PC4



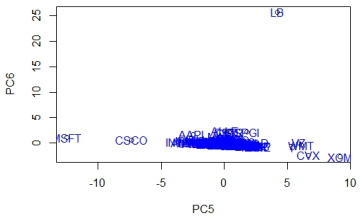
PCA 2014, PC5 vs PC6



PCA 2015, PC1 vs PC2



PCA 2015, PC3 vs PC4



PCA 2015, PC5 vs PC6

Interpretation of Principal Components Across years

Year	PCA	Interpretation	Examples (Scores High)
2012	1	High Revenue and Profit, High Payables	Exxon Mobil, JP Morgan
2012	2	Low on cash, few assets, low on long term debt, high on short term debt	Exxon Mobil, Chevron Corporation
2012	3	Running losses, very low profits	Bank of America, Archer Daniels Midland
2012	4	High on borrowing, low on investing	Wells Fargo, JP Morgan
2012	5	Low on liabilities, low on returns	Banking firms(Eg:Citi Bank,Wells Fargo)
2012	6	very low return, low cash flow	Bank of America, Exxon Mobil, Realty Income Corp
2013	1	High profit, high earnings	Exxon Mobil, JP Morgan, Bank of America
2013	2	high short term debt, low cash, low investment	Apple, Walmart, Chevron Corporation
2013	3	High profits and margins, low costs	Microsoft, Visa, Verisign
2013	4	very low on financing, high on investing (low dividend)	Bank Of America, Microsoft, Cisco
2013	5	Very high Return on Equity (high dividend)	Wynn hotels, Clorox
2013	6	High on Assets, low expenditure for investments (Capital and R&D)	Bank of America, Exxon Mobil, Chevron
2014	1	High earnings, High operating cash flows	Unicorn Companies
2014	2	low debt level, low long term investments	Big Tech Companies
2014	3	low profit and operating margins, efficient production	Walmart, Ford, Energy Companies
2014	4	high short term invesments, high on R&D, negligible fixed assets	Research based Tech Companies
2014	5	Low Return, Ineffecient utilization of Assets	Wynn hotels, Allegian
2014	6	Low Receivables, High Interest, High on debt	Verizon, JP Morgan, Ford
2015	1	High earnings, High operating cash flows	Unicorn Companies
2015	2	High on Borrowing, low on investing	Walmart, Apple, Microsoft, Ford
2015	3	low profit and operating margins, efficient production	Energy Companies, Walmart, Ford, Big Tech Firms
2015	4	High debt level, low investment	Microsoft, Cisco, Ford
2015	5	Low short term investment, low common stock, highbfixed assets	Exxon, Chevron, Walmart, Verizon
2015	6	Very high return on equity	L Brands