S&DS363 Final Project

David Lieberman, Yavuz Ramiz Çolak, Liana Wang, Ryo Tamaki

May 8th, 2019

Introduction, Design and Primary Questions

Our dataset follows 335 S&P500 companies with 16 different variables quantifying their performance on the New York Stock Exchange as of 2017-12-31 (obtained via the Quandl API). Each row represents a company, and each column contains its performance metrics, which range from Gross Profit and Return on Equity (ROE), to Assets and Liabilities. In this report, we want to answer two questions. First, how predictive are a company's key financial metrics are of its stock price performance (our analysis follows the timeframe between 2017-12-31 to 2018-12-31). Second, we want to see whether clustering companies by industry sector reveals trends in their financial metrics – either trends within (homogeneity or substantial variation) or between sectors. In answering these questions, we will use a variety of statistical techniques:

- We will conduct Principal Component Analysis (PCA) to determine which performance metrics tend to vary greatly between stocks. Our PCA will yield a few uncorrelated principal components in the directions of greatest variance, reducing the overall dimensionality of our data. PCA will also allow us to see which performance metrics tend to correlate with one another (their weights within principal components), and effectively eliminate any correlation between variables to avoid bias that would be introduced by multicollinearity.
- We will use Discriminant Analysis (DA). This method will allow us to see which principal components (and by extension the performance metrics contained in their loading coefficients) are most important for discriminanting between companies that did or did not outperform the SP500 during that same timeframe.
- We will use Cluster Analysis and MANOVA to investigate if there are particular performance metrics indicative of a stock's membership to a given sector, and/or any significant differences between sector mean financials.

Before starting our analysis, we want to tell more about the data we are using. Our dataset has both categorical and continuous variables. Categorical Variables (3) include 'Ticker Symbol,' 'S&P500 Sector,' and 'Returns' (how much did the stock price percent change with respect to the S&P500 percent change within the same timeframe?). Continuous Variables (13) include mainly key financial ratios. The reason we preferred using ratios is that we wanted to have variables that are comparable across companies, regardless of company size. Ratios are an effective means of "standardizing" company performance. Continuous variables we used were: 'Cash&Equivalents/Liability', 'Equity/Assets', 'EBIT/EV', 'EV/EBITDA', 'EBITDA Margin', 'Gross Margin', 'Net Margin', 'ROA (Return on Asset)', 'ROE(Return on Equity)', 'ROIC (Return on Invected Capital)', 'ROS (Return on Sales)', EV (Entreprise Value) and 'YoY change in Stock Price'.

Below is a short description of each of the variables we are using:

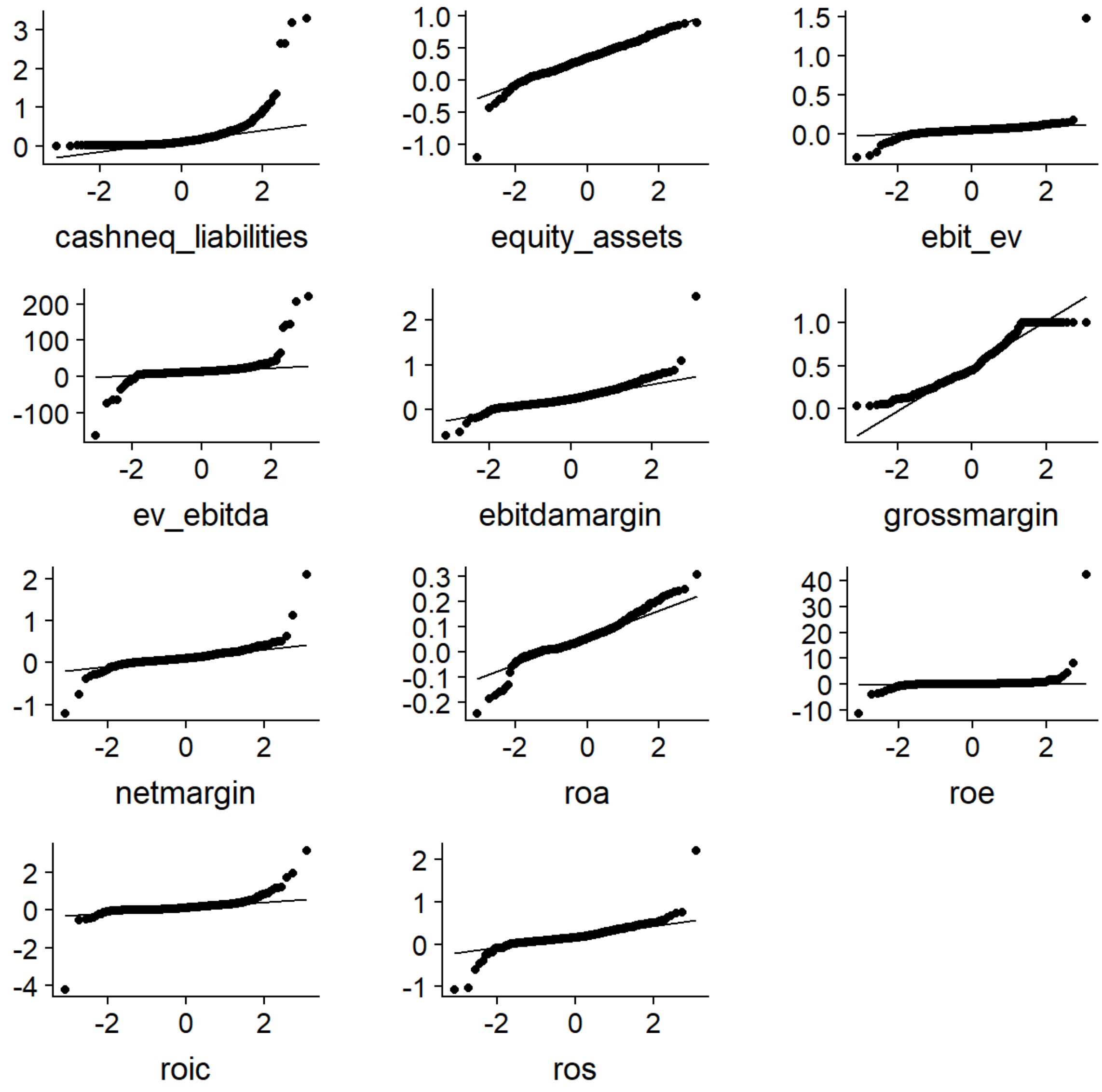
- Cash&Equivalents/Liability: Cash & Equivalents refer to a company's assets that are cash or can be converted into cash immediately. Liabilities are a company's legal financial debts or obligations.
- EBIT/EV: EBIT refers to Earnings Before Interest and Tax, and EV refers to the Enterprise Value of a company. Enterprise Value can be summarized as the measure of a company's total value. More precisely, Enterprise Value is Market Capitalization + Total Debt - Cash & Cash Equivalents of a company. So, as a ratio, EBIT/EV measures a companies pre-tax earning yield.
- EV/EBITDA: We defined EV above. EBITDA is Earnings Before Interest Tax Depreciation and Amortization. EV/EBITDA is seen as the enterprise multiple, and used to determine whether a company is overvalued, undervalued or fairly valued.
- EBITDA Margin: This ratio is calculated as EBITDA/Revenue of the company. It helps measure a company's operating profitability as a percentage of the total revenue.
- Gross Margin: This ratio is calculated as Gross Profit/Revenue. Gross profit is defined as total revenue minus the cost of goods sold (or costs of services provided).
- Net Margin: This ratio is calcualted as Net Income/Revenue. Net Income is equal to Sales minus [Cost of Goods Sold, SG&A (Selling, General and Administrative Cost), Operating Expenses, Depreciation, Other Expenses, Taxes and Interest].
- ROA: ROA is Return on Assets and it is calculated as Net Income / Total Value of a Company's Assets.

- ROE: ROE is Return on Equity and it is calculated as Net Income / Total Equity of a Company (also known as Shareholder's Equity).
- ROIC: ROIC is Return on Invested Capital and it is calculated by Net Operating Profit After Tax divided by Invested Capital.
- ROS: ROS is Return on Sales and it is calculated as Operating Profit / Net Sales. Operating Profit is also known as EBIT–Earnings Before Interest and Tax.

Studying our Data Set, Transformations

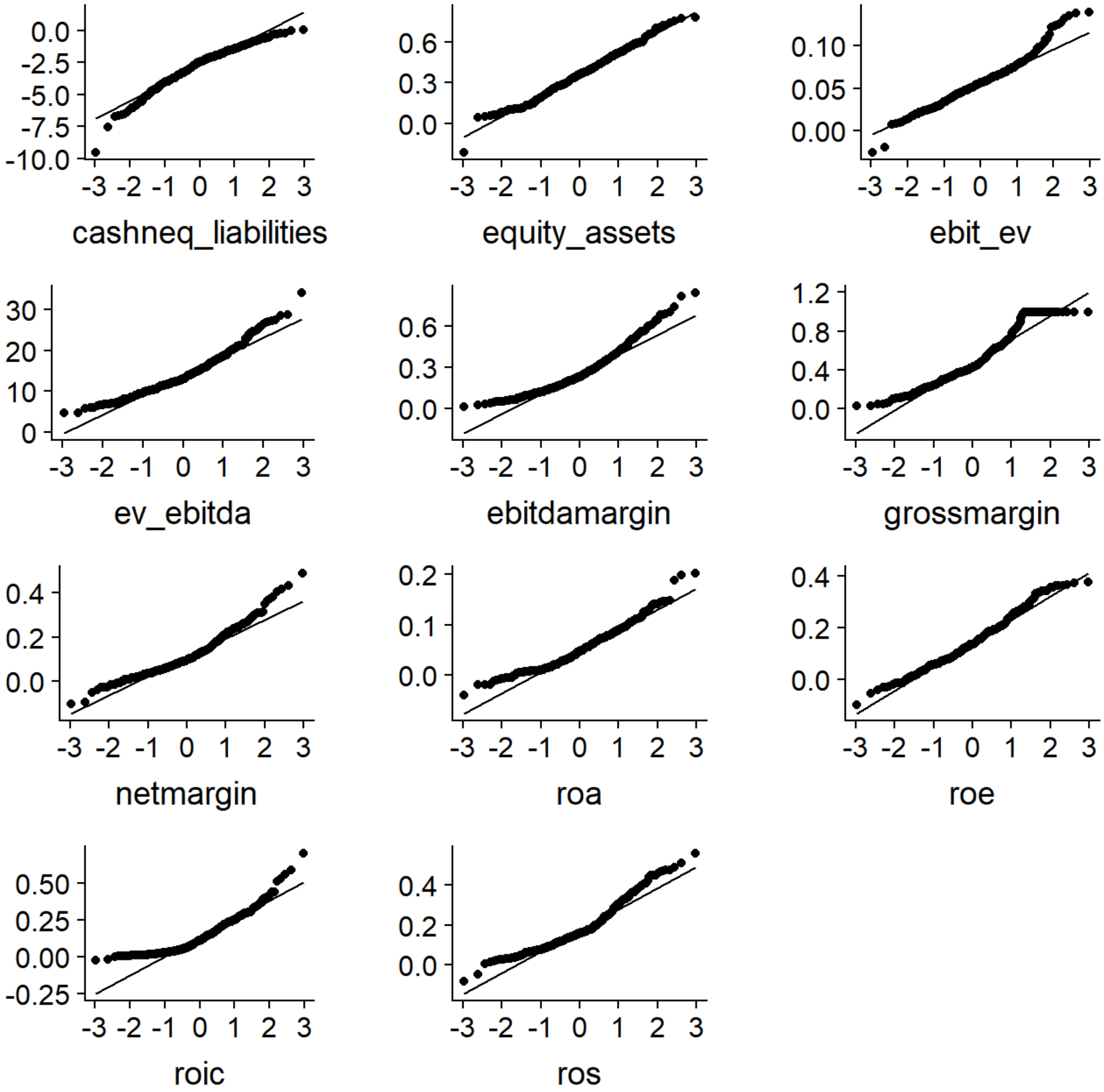
First, let's take a look at the data and make quantile-quantile plots for each performance metric to see if each variable is approximately univariate Normal.

	ticker	sector	cashneq_liabilities	equity_assets	ebit_ev	ev_ebitda	ebitdamargin	grossmargin	netmargin	roa	roe	roic	ros	ev	percent_change	returns
A	A	Health Care	0.7457533	0.5733444	0.0411919	19.572175	0.245	0.539	0.153	0.084	0.151	0.249	0.197	21411959044	-0.0500077	Mediocre
AAL	AAL	Industrials	0.0114440	-0.0147769	0.0910593	7.555655	0.152	0.586	0.030	0.024	0.496	0.081	0.104	48847306768	-0.6203675	Low
AAP	AAP	Consumer Discretionary	0.1079388	0.4026261	0.0719544	9.715571	0.088	0.436	0.051	0.056	0.149	0.145	0.062	8047601907	0.3587418	High
AAPL	AAPL	Information Technology	0.0840918	0.3571548	0.0723489	11.931053	0.324	0.385	0.211	0.140	0.363	0.188	0.280	885832939360	0.3172677	High
ABBV	ABBV	Health Care	0.1416219	0.0720058	0.0475183	17.957361	0.363	0.750	0.188	0.078	0.932	0.209	0.309	183739720155	-0.0490292	Mediocre
ABC	ABC	Health Care	0.0732321	0.0584560	0.0515906	15.545870	0.009	0.030	0.002	0.011	0.157	0.557	0.007	20605927004	0.1026892	High
ABT	ABT	Health Care	0.2128366	0.4078426	0.0290078	17.889981	0.229	0.547	0.017	0.007	0.015	0.085	0.119	112349081207	0.2109775	High
ACN	ACN	Information Technology	0.3179475	0.3944258	0.0515249	16.440430	0.143	0.300	0.094	0.163	0.424	1.163	0.121	86206756673	0.2265925	High
ADBE	ADBE	Information Technology	0.3795574	0.5820121	0.0249644	34.911900	0.348	0.862	0.232	0.124	0.212	0.485	0.303	88607797873	0.2844673	High
ADI	ADI	Information Technology	0.0954336	0.4806489	0.0264374	24.543723	0.326	0.599	0.142	0.037	0.082	0.124	0.211	40825636021	-0.0462262	Mediocre



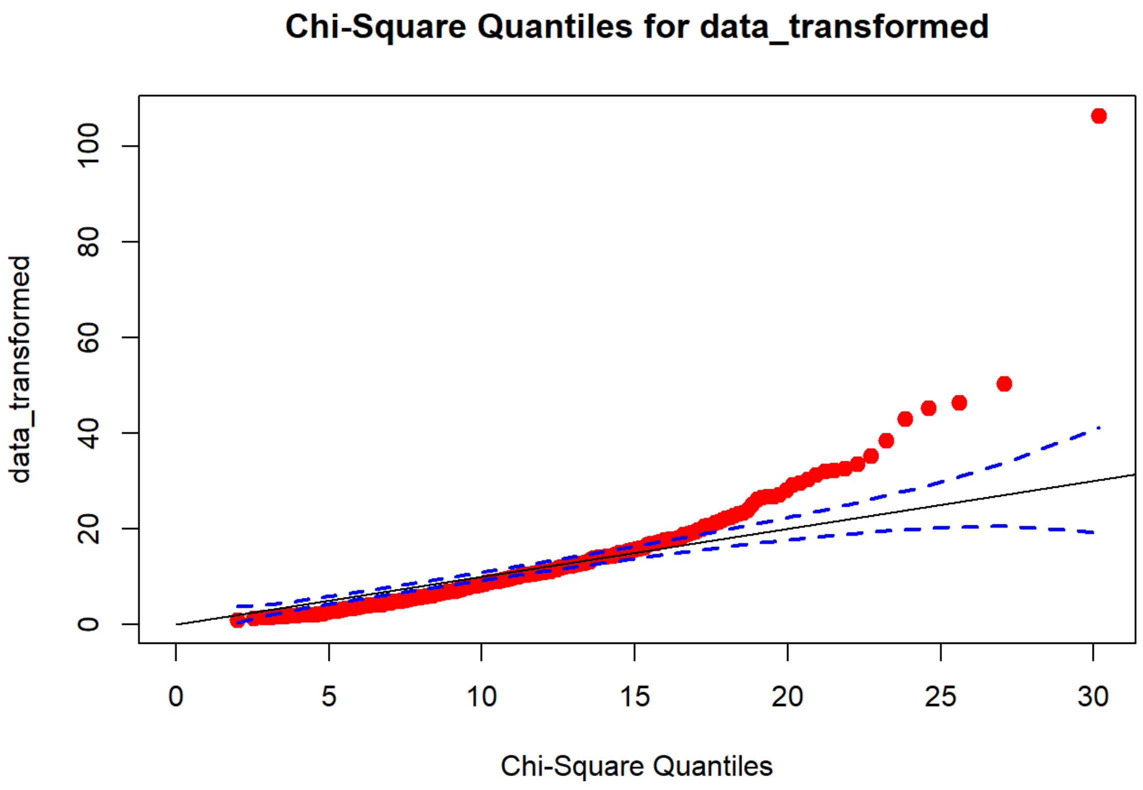
Several of these variables seem to deviate from univariate Normality. Applying a log transformation will likely help. However, there are several subtleties. Many variables take on negative values, however translating our data by a constant positive factor would disrupt our ratio quantities. Therefore, although it might be helpful to transform, by in large, our variables seem approximately Normal enough that we will leave them unchanged, except for cashneq_liabilities which is strictly positive across all our data so we will take its log. Additionally, there are many outliers that would likely disrupt our calculations later on. Therefore, we make the difficult decision to take them out early on, and proceed with our analysis without them. We are left with 335 companies.

TRANSFORM NUMERIC DATA AND REMOVE OUTLIERS



After log transforming cashneq_liabilities and removing outliers, the variables all appear to be approximately univariate Normal (save for a few remaining pesky outliers).

Now, let’s see if our data is approximately multivariate Normal using a chi-square quantile-quantile plot and seeing if the data all fall within the 95% confidence interval boundaries, denoted by the blue dotted lines on the graph below.



As we can see, our data is very far from multivariate Normal – nearly all of the data lie outside the 95% confidence limits on the chi-square quantile-quantile plot.

Section 1: Principal Components Analysis

Let’s construct a correlation matrix for our data. Hopefully, some variables will be highly correlated so our data will be a good match for Principal Component Analysis.

	cashneq_liabilities	equity_assets	ebit_ev	ev_ebitda	ebitdamargin	grossmargin	netmargin	roa	roe	roic	ros
cashneq_liabilities	1.00	0.28	-0.02	0.14	-0.23	0.03	-0.18	0.26	0.11	0.43	-0.12
equity_assets	0.28	1.00	-0.25	0.20	-0.05	-0.19	-0.02	0.49	-0.05	0.32	-0.14
ebit_ev	-0.02	-0.25	1.00	-0.71	-0.19	-0.07	0.07	0.19	0.39	0.14	0.12
ev_ebitda	0.14	0.20	-0.71	1.00	0.10	0.23	0.01	-0.05	-0.22	0.07	0.08
ebitdamargin	-0.23	-0.05	-0.19	0.10	1.00	0.55	0.62	-0.12	-0.13	-0.21	0.84
grossmargin	0.03	-0.19	-0.07	0.23	0.55	1.00	0.47	-0.14	-0.12	-0.14	0.65
netmargin	-0.18	-0.02	0.07	0.01	0.62	0.47	1.00	0.29	0.33	-0.06	0.70
roa	0.26	0.49	0.19	-0.05	-0.12	-0.14	0.29	1.00	0.78	0.62	0.02
roe	0.11	-0.05	0.39	-0.22	-0.13	-0.12	0.33	0.78	1.00	0.47	0.04
roic	0.43	0.32	0.14	0.07	-0.21	-0.14	-0.06	0.62	0.47	1.00	-0.07
ros	-0.12	-0.14	0.12	0.08	0.84	0.65	0.70	0.02	0.04	-0.07	1.00

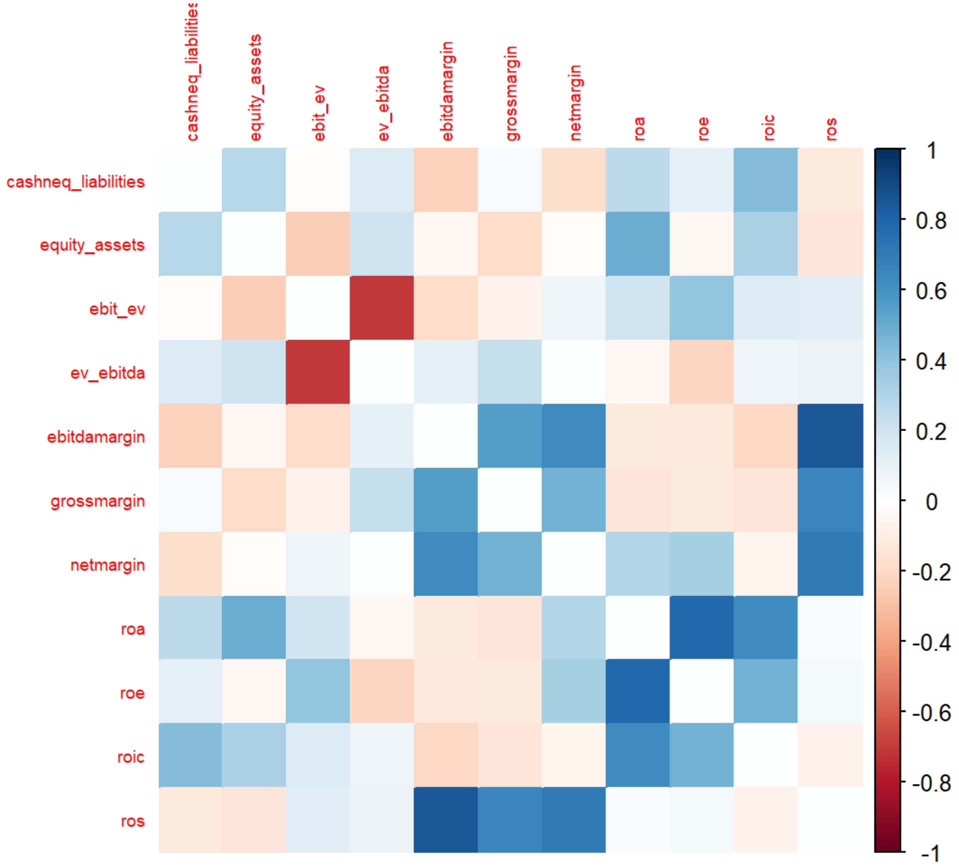
Now we sort the correlations by largest magnitude and display the first six entries.

```
correlation_vector <- as.vector(as.matrix(correlation_matrix))
correlation_vector <- correlation_vector[-as.vector(which(correlation_vector == 1))]
sorted_correlations <- correlation_vector[order(-abs(correlation_vector))][c(FALSE, TRUE)]
head(sorted_correlations)
```

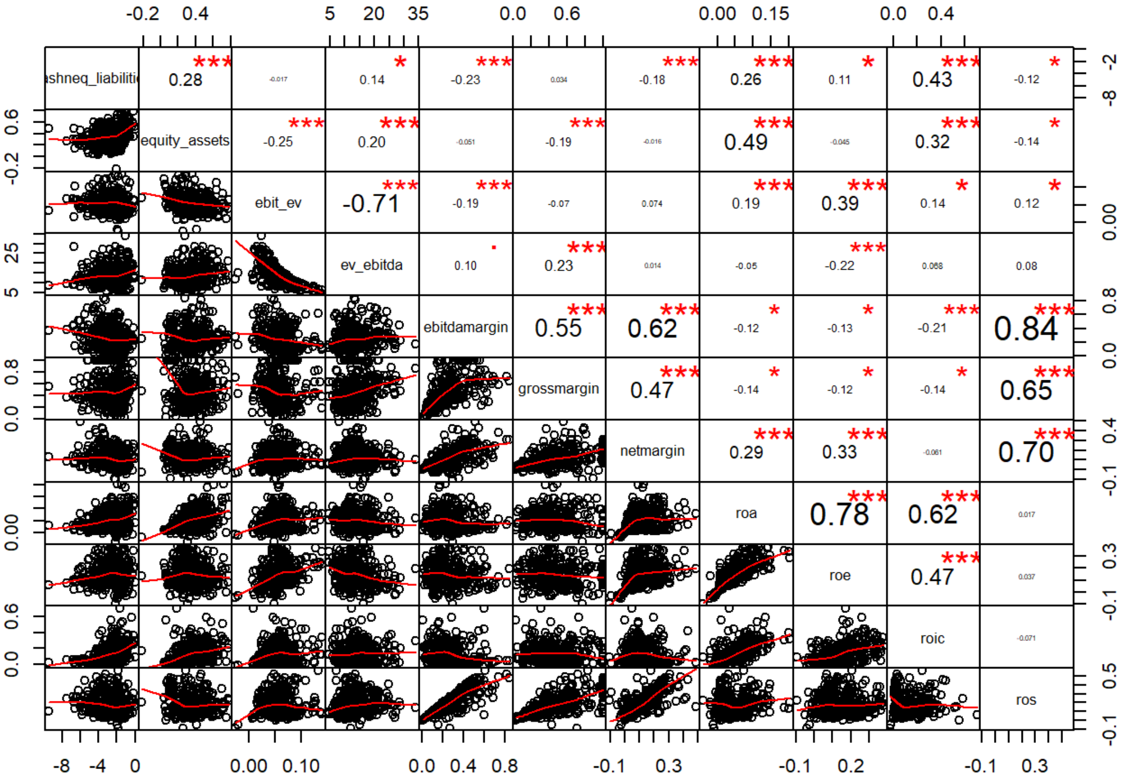
```
## [1]  0.84  0.78 -0.71  0.70  0.65  0.62
```

We observe several highly correlated variables. PCA will work well for our data.

Here’s a more graphical representation of the correlation matrix that was constructed above.



Below are scatterplots of variables we are tracking paired and projected into 2D space. We note that there is no severe clumping or drastic outliers skewing our data; in fact, several variables are noticeably linearly correlated. Therefore PCA will work well for our data and is advisable because as it stands, our data has a substantial degree of multicollinearity.



```
pc2 <- prcomp(data_transformed, scale=TRUE)

print(summary(pc2),digits=2)
```

```
## Importance of components:
##                         PC1  PC2  PC3  PC4   PC5   PC6   PC7   PC8   PC9
## Standard deviation     1.77 1.63 1.42 0.99 0.892 0.693 0.618 0.516 0.456
## Proportion of Variance 0.28 0.24 0.18 0.09 0.072 0.044 0.035 0.024 0.019
## Cumulative Proportion  0.28 0.52 0.71 0.80 0.869 0.913 0.948 0.972 0.991
##                          PC10   PC11
## Standard deviation     0.2526 0.1941
## Proportion of Variance 0.0058 0.0034
## Cumulative Proportion  0.9966 1.0000
```

According to the “eigenvalues greater than 1” rule, we will be keeping the first three principal components.

	Comp.1	Comp.2	Comp.3
cashneq_liabilities	0.2107606	0.1418846	0.2976732
equity_assets	0.1778752	0.1391838	0.4338590
ebit_ev	0.1002286	0.2435518	-0.5413428
ev_ebitda	-0.1058587	-0.0980373	0.5736262
ebitdamargin	-0.4856547	0.1481688	0.0848659
grossmargin	-0.4151067	0.1239602	0.1123826
netmargin	-0.3530608	0.3807995	-0.0132358
roa	0.2247907	0.5140928	0.1287573

	Comp.1	Comp.2	Comp.3
roe	0.1706284	0.4875412	-0.1536843
roic	0.2760203	0.3590935	0.2021214
ros	-0.4626343	0.2798275	0.0005319

Looking at the loading coefficients for the first three principal components (“major contributor”:= abs(loading coefficient) > 0.3):

- First Principal Component Major Contributors: ebitdamargin (-0.486), grossmargin (-0.415), netmargin (-0.353), ros (-0.463).

It seems that in this component we get variables that are focused on revenue and primarily costs of goods sold. All four of these, Net Margin, EBITDA Margin, Gross Margin and Return on Sales are primarly dependent on revenue, and strong revenue drives these ratios up. Similarly, these ratios are also collectively dependent on cost of goods sold–low cost of goods sold help drive these ratios up.

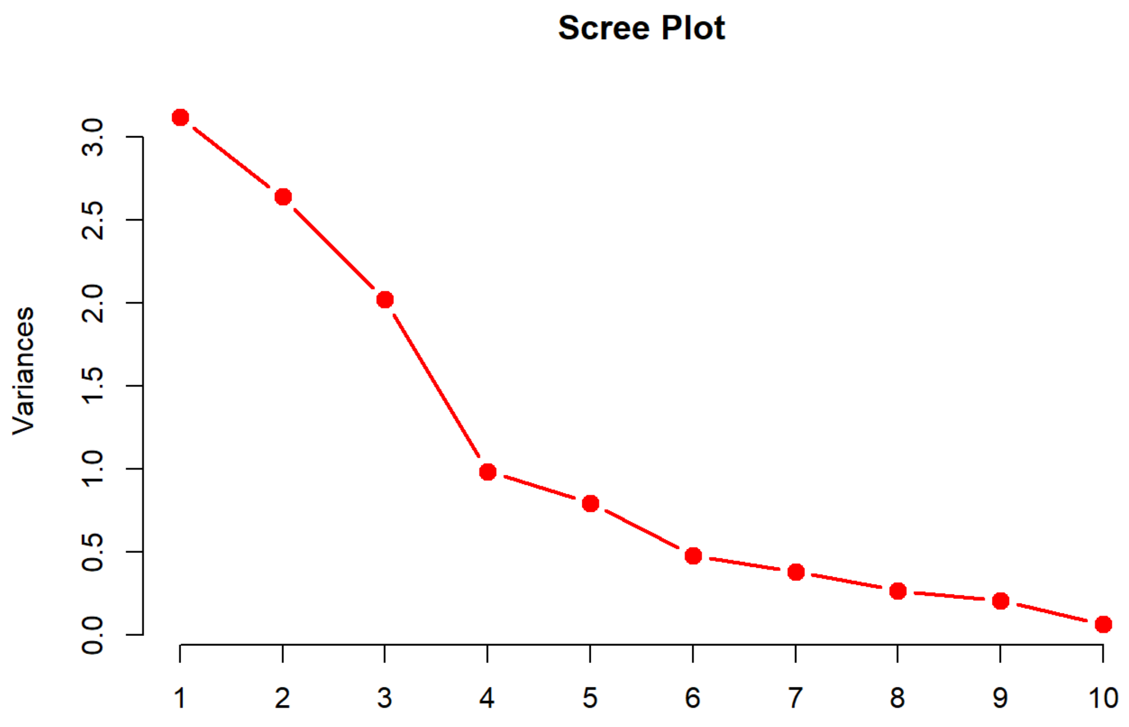
- Second Principal Component Major Contributors: netmargin (0.381), roa (0.514), roe (0.488), roic (0.359).

These four variables are trying to measure return on investments. ROA (Return on Assets), ROE (Return on Equity), ROIC (Return on Invested Capital) and Net Margin all have net income as their numerator. Therefore, these ratios are high when net income is high, and not so significant when net income is low. It is expected that these variables be grouped together in PCA.

- Third Principal Component Major Contributors: equity_assets (0.434), ebit_ev (-0.541), ev_ebitda (0.574).

These three variable all use capital structure as their numerator or denominator. Enterprise Value is Debt + Equity - Cash of the company, and these variables try to see the relation between Before Interest and Tax earnings of the company and its capital structure (Equity and Debt). So this principal component captures unique information about the Debt and Equity structure of the companies, and it makes sense that these three variables are grouped together. Also note that Equity/Assets and EV/EBITDA are positive while EBIT/EV are negative, this can be explained by the fact that in Equity/Assets and EV/EBITDA financial variables concerning the capital structure are in the numerator (i.e. Equity and Enterprise Value), and in EBIT/EV capital structure term is in the denominator (i.e. Enterprise Value).

These 3 Principal Components capture in order: Revenue & Cost Structure, Net Income and Return levels, and Capital Structure and Pre-Interes/Tax Earnings relation for the companies.



As we can see from the scree plot above, “cutting above the first elbow” leaves us with three principal components and is in good agreement with the “eigenvalues greater than one” findings above.

Here are the numerical results we obtain from performing PCA and transforming our stock data into the PCA-space. Furthermore, the summary statistics show that the PCA components appear to have approximately equal means and standard deviations and appear to be approximately univariate Normal, so we won’t restandardize the data later on.

```
##       Returns                Sectors         PC1        PC2        PC3
## A    Mediocre            Health Care -0.64761161 -1.1469320  2.4184429
## AAP      High Consumer Discretionary -1.72773597  0.3470825 -0.7126718
## AAPL     High Information Technology -0.37786940 -3.3138786 -0.5494176
## ABC      High            Health Care -3.13672078  0.9108762 -0.2483814
## ABT      High            Health Care -0.07251146  2.2091724  1.4749183
## ADI  Mediocre Information Technology  0.99397596  0.6527626  2.4782485
```

```
## [1] "Summary Statistics for PC1"
```

```
## [1] "mean: -4.67339682707185e-18"
```

```
## [1] "sd: 1.76608203692052"
```

```
## [1] "Summary Statistics for PC2"
```

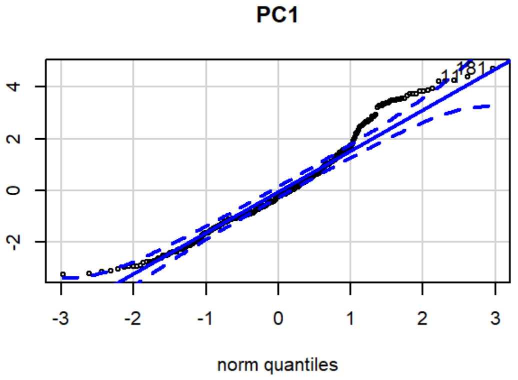
```
## [1] "mean: 8.2048860346868e-17"
```

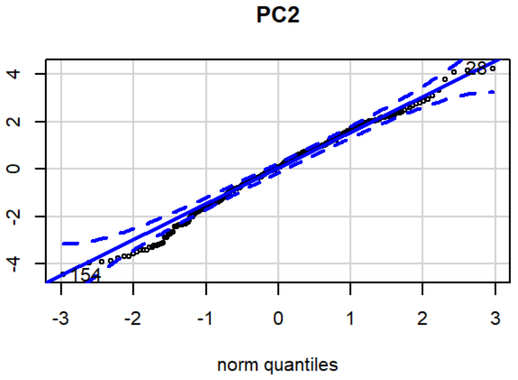
```
## [1] "sd: 1.62506724427192"
```

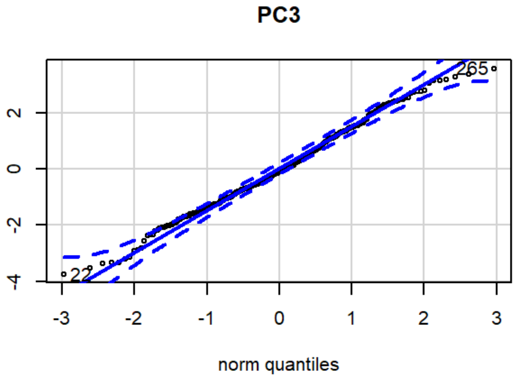
```
## [1] "Summary Statistics for PC3"
```

```
## [1] "mean: 2.50596340941948e-18"
```

```
## [1] "sd: 1.42154789131039"
```

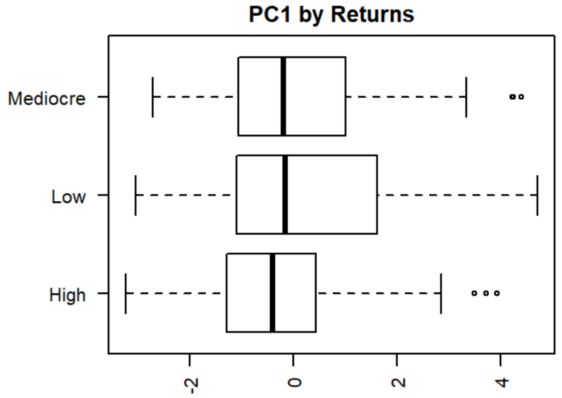


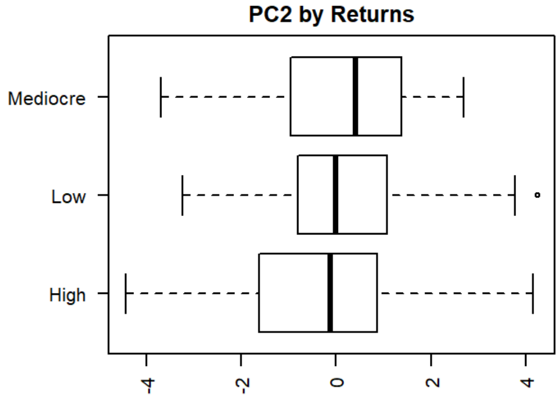


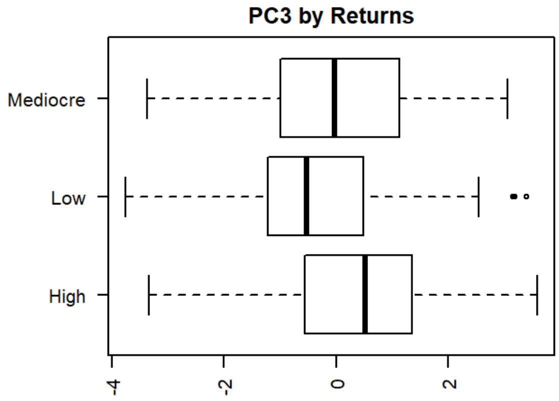


Section 2: Discriminant Analysis

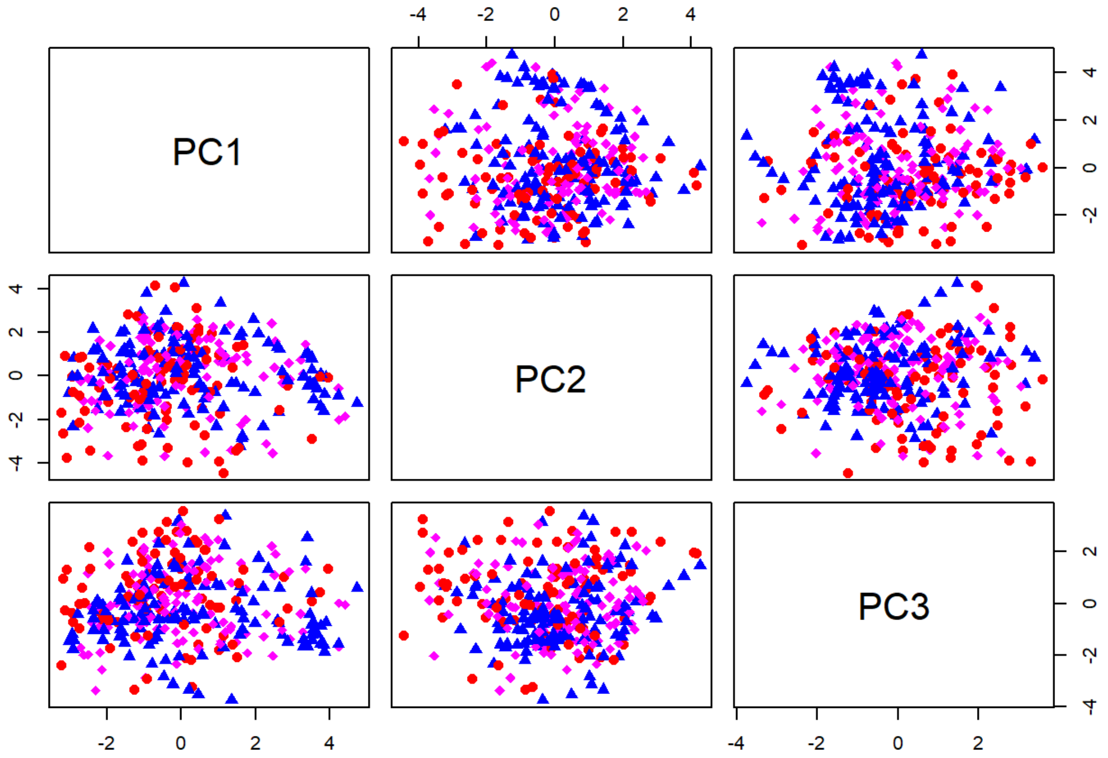
Let's take a look at boxplots of each stock's principal component values, grouped by returns. "High" means the percent change of the stock price positively outperformed by at least 150% the percent change in the SP500 between 2017-12-31 and 2018-12-31. "Low" means the stock price was at least 150% below the percent change in the SP500, while "Mediocre" means it hovered somewhere inbetween. We will be using the three principal components constructed above as our new working data, as they capture much of the variance of our original 11 continuous explanatory variables in only 3 dimensions and have the benefit of being, by construction, uncorrelated thus all but eliminating the multicollinearity once present in our data.


PC1 by Returns

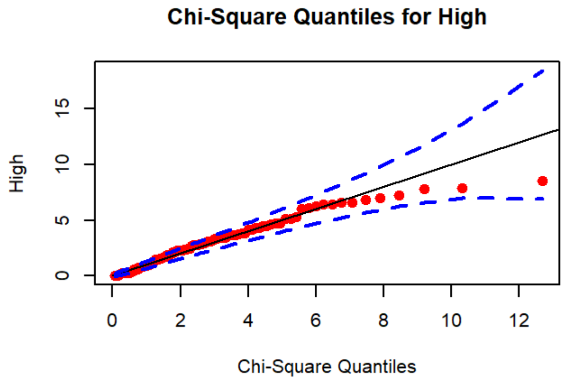

PC2 by Returns

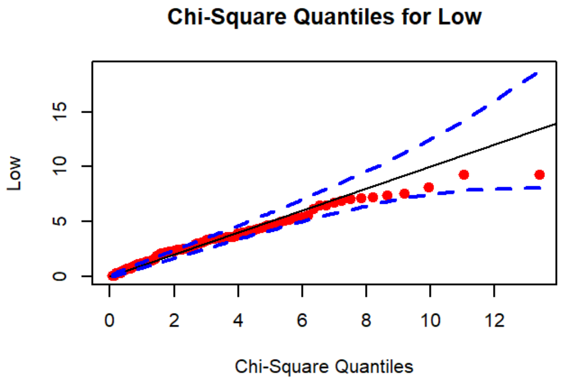

PC3 by Returns

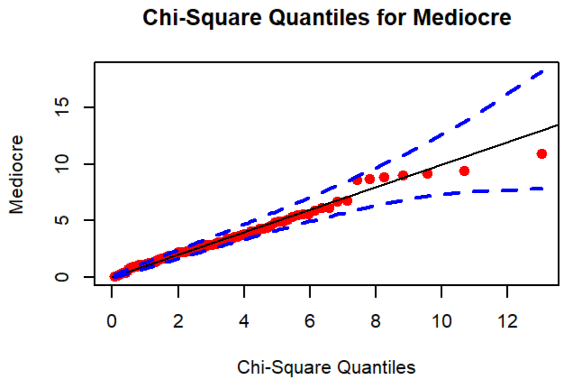
In comparing the PCA values for the High, Low, and Mediocre return stocks for the first three dimensions, we notice that the distribution of PCA values by returns appears largely symmetric. With this symmetry, we note that the our boxplots of the principle components suggest an approximately normal distribution for each group. However it should also be noted that there is some skewness for the mediocre returns grouping in the second principle component and in the high returns grouping for the third principle component.



An inspection of the principal components' projected scatters appears promising. All three groups (each has a different shape and color) seem to have a similar random scatter and "covariance footprint." In other words, it is likely that the data is multivariate Normal within groups, and the group covariance matricies are not significantly different from one another.







Indeed, when the companies are grouped by their returns (either High, Low, or Mediocre), we see that the Chi-Square Quantile plots within groups are roughly linear and within the 95% confidence boundaries. Therefore, we can conclude that within groups, the data follows an approximately multivariate Normal distribution.

```
## [1] "Covariance Matrix for High"
```

```
##           PC1        PC2        PC3
## PC1 2.4607156 0.12989332 0.19402823
## PC2 0.1298933 3.57445429 0.01980099
## PC3 0.1940282 0.01980099 2.26389279
```

```
## [1] "log-determinant"  "2.98263818904766"
```

```
## [1] "Covariance Matrix for Low"
```

```
##             PC1        PC2         PC3
## PC1  3.80220284 -0.2794256 -0.02329438
## PC2 -0.27942562  2.0677208  0.22524195
## PC3 -0.02329438  0.2252419  1.76148155
```

```
## [1] "log-determinant"  "2.60416684097937"
```

```
## [1] "Covariance Matrix for Mediocre"
```

```
##            PC1         PC2        PC3
## PC1 2.69174994  0.04665121  0.1173570
## PC2 0.04665121  2.41937204 -0.1048254
## PC3 0.11735699 -0.10482541  1.8712364
```

```
## [1] "log-determinant"  "2.49469333552727"
```

```
##  Box's M-test for Homogeneity of Covariance Matrices
## 
## data:  PCA_Data[, c("PC1", "PC2", "PC3")]
## Chi-Sq (approx.) = 20.228, df = 12, p-value = 0.0629
```

In computing the covariance matrices, by visual inspection, we note that the covariance matrices seem reasonably similar across groups. This observation is supported by the p-value (0.0629) of the Box-M test, suggesting that there is no statistically significant difference between each group's covariance matrix; therefore, our data are suitable for Linear Discriminant Analysis (LDA).

LDA, no cross-validation

```
##             High Low Mediocre
##   High       37  38       19
##   Low        23  91       17
##   Mediocre   27  63       20
```

```
## [1] 0.441791
```

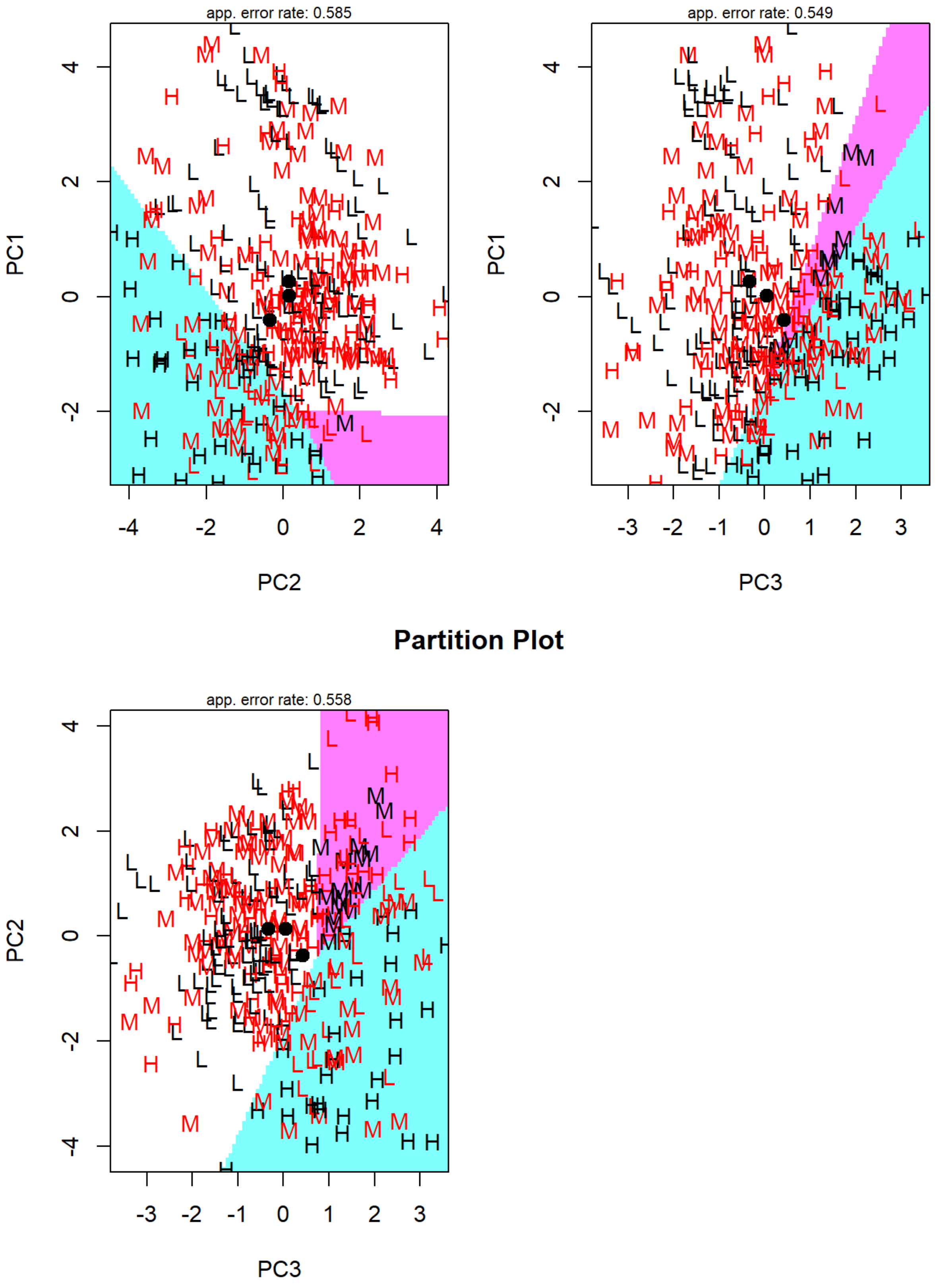
We compute our classification results in the confusion matrix above to evaluate how accurately classification via LDA (without cross-validation) worked for our data. We find that the accuracy is lacking with an accuracy rate of 44.2%, which is slightly better than guessing.

LDA, with cross-validation

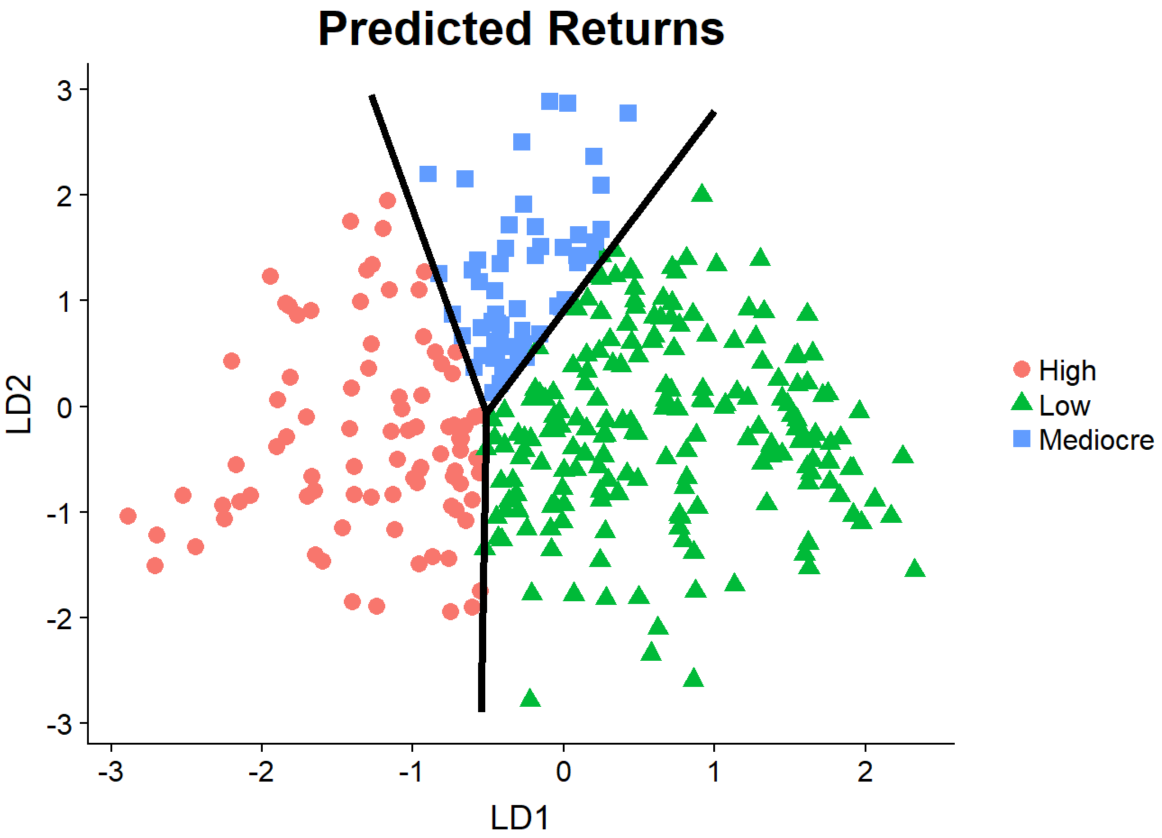
```
##             High Low Mediocre
##   High       35  39       20
##   Low        25  89       17
##   Mediocre   28  68       14
```

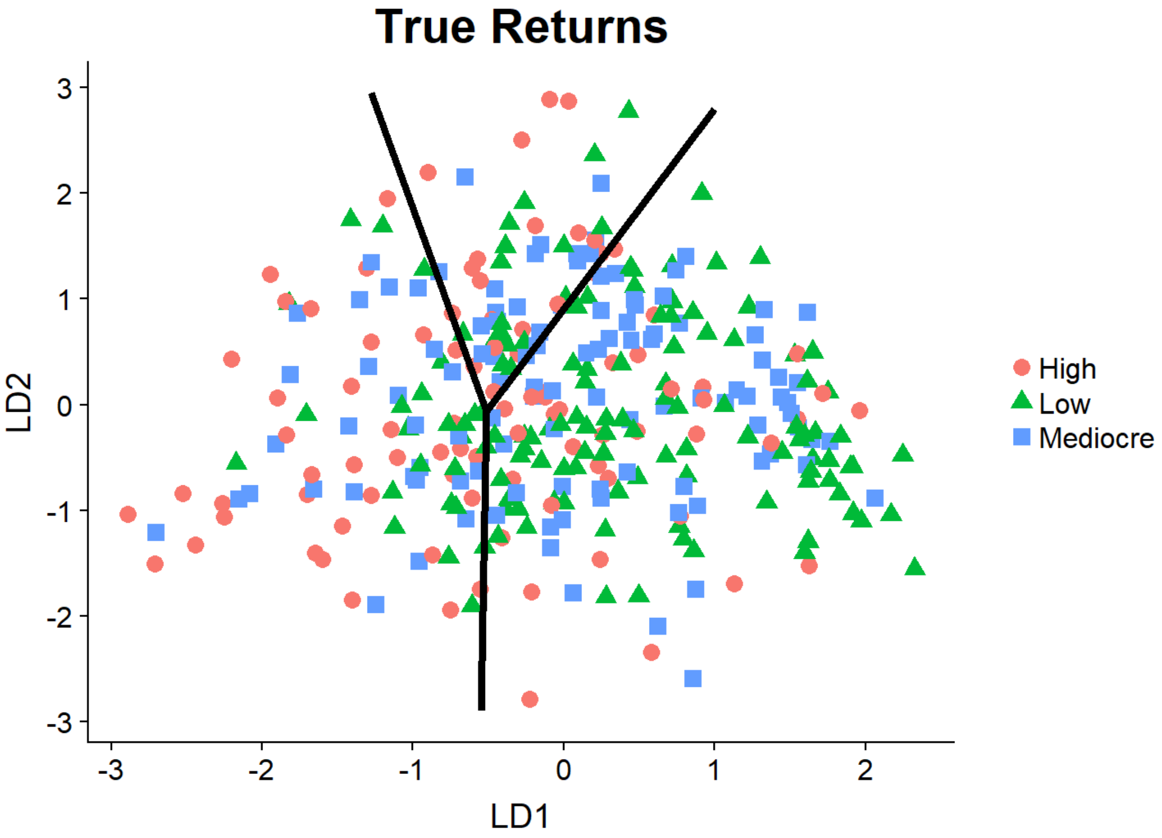
```
## [1] 0.4119403
```

We also compute LDA with cross-validation which should be more indicative of the true predictive power of our model, as we are not “double-dipping” by both constructing the model with data then using that same model to predict that same data. As expected, the prediction accuracy is even lower than before at 41.2%, a decrease of 3% and only ~8% better than guessing. The model **does not** seem to have an “Achilles heel” consistently failing to classify a particular group, rather, it seems to perform poorly across the board.

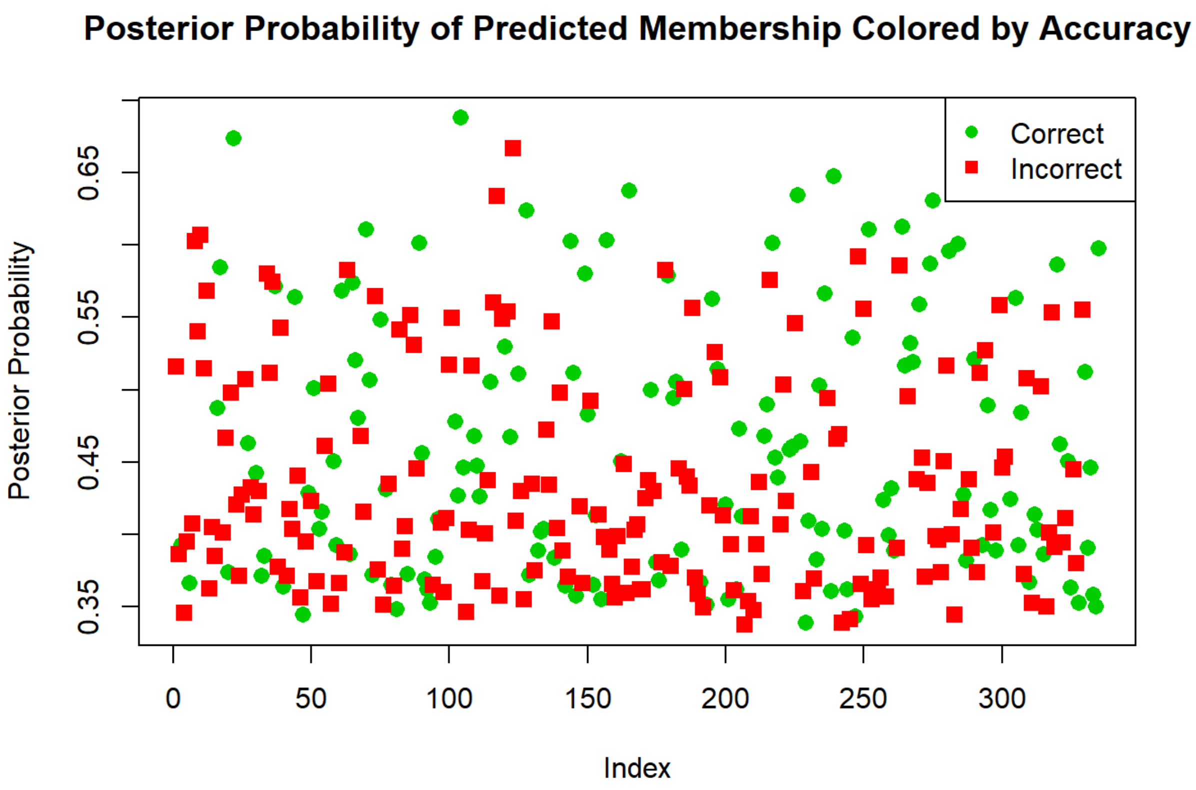


We plot the projected classification LDA boundaries into PCA-space to better visualise the accuracy of the classification method. In plotting the classifications on two dimensions between PC1, PC2, and PC3, we find that in each case, the error rate is marginally better than guessing as the error rates are between 0.549 and 0.585. In visualizing the results, we see that the model is uniformly poor.





We explictly draw the LDA classification boundaries, this time visualizing our data on the LDA space (in contrast to the projections into PCA space visualized above). We see that the predicted classifications, denoted by the different shapes and colors, when compared to the true company membership, is no where close (suprise suprise). It literally looks like our data has a random scatter with respect to the classification boundaries so it makes sense that our prediction accuracy has only been marginally better than blind guessing.



Visualizing the posterior probabilities of each stock coloured by true prediction accuracy, there appears to be an exceedingly weak relationship between our model’s predictive confidence in its classification and how well it actually did in truth. If DA were effective, we would expect that as the posterior probability increased, the accuracy would similarly increase – positive movement along the y-axis would yield increasing density of green, correct observations. As the plot above shows, however, that relation doesn’t hold.

Even though LDA was not accurate we might be inclined to try other classification techniques for comparison. Nonparametric k-nearest neighbor (leave-one-out):

```
results <- matrix(nrow = 45, ncol = 20)
for (j in 1:20) {
  for (i in 1:45) {
    test_point <- PCA_Data[i,-c(1:2)]
    train_data <- PCA_Data[-i,-c(1:2)]
    knn_prediction <- as.vector(knn(train=train_data, test=test_point, cl=PCA_Data$Returns[-i], k=j))
    truth <- as.vector(PCA_Data$Returns[i])
    results[i,j] <- truth == knn_prediction
  }
}
best_k <- which.max(colMeans(results))
best_k
```

```
## [1] 8
```

```
success_rate <- colMeans(results)[best_k]
success_rate
```

```
## [1] 0.3777778
```

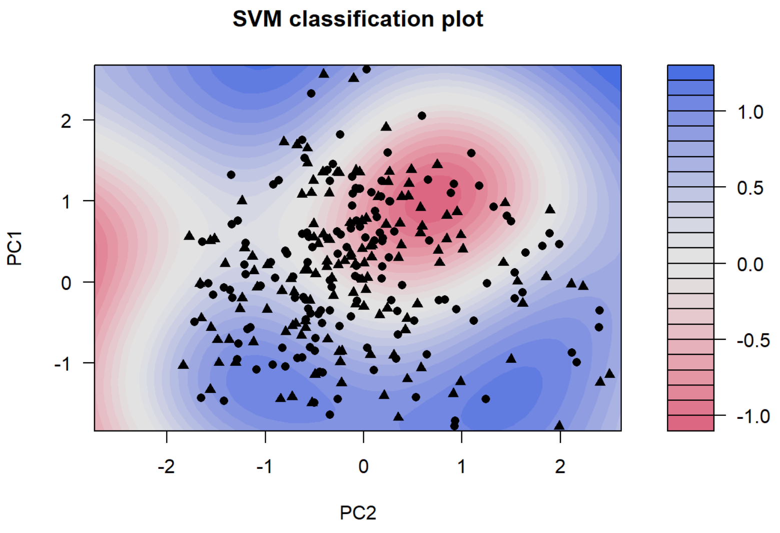
Leaving out one point, and predicting its group using the k-nearest neighbor method, and repeating this for all 335 points in our dataset, this method predicts the group with approximately 0.3777778 success rate… Even worse than LDA. We also did parameter tuning to determine the k that gave the greatest prediction accuracy, which yielded k=8

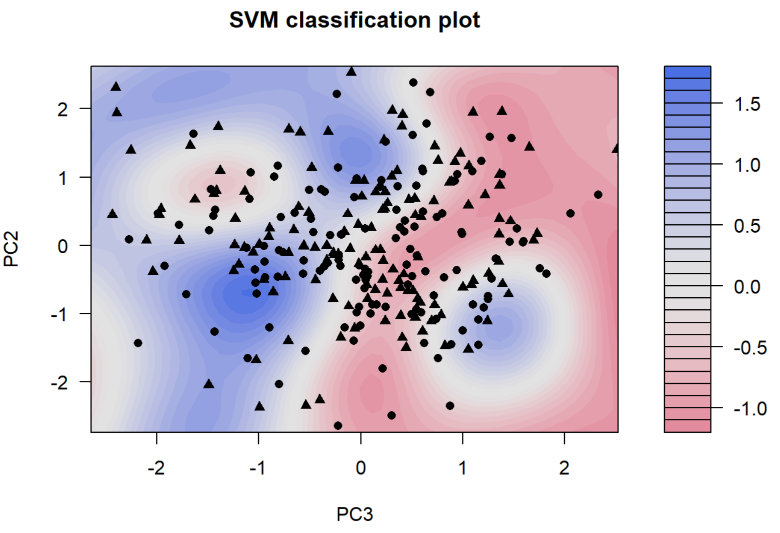
We will also try a machine learning technique: Support-Vector Machine, ML binary classification method, with cross-validation. In addition to the cross-validation of our final model’s predictive ability, we will also evaluate several potential models by tuning the SVM parameters until we select one that achieves the best results.

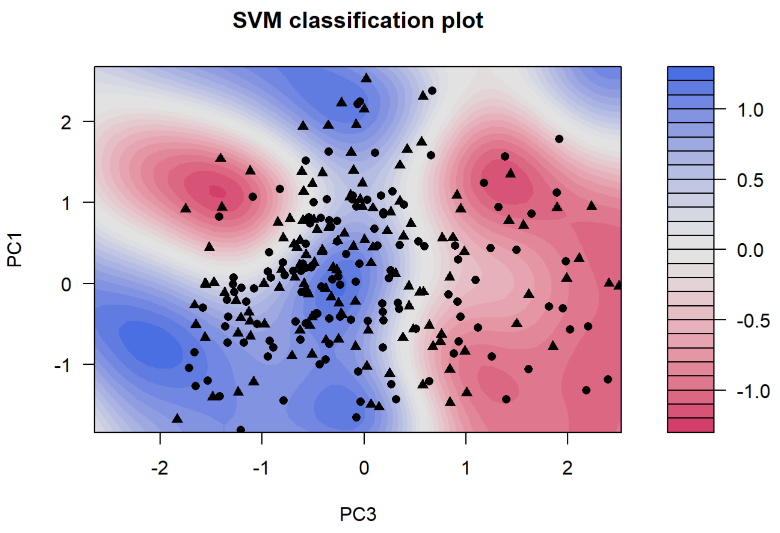
```
svm_data <- PCA_Data[,-2]
tune.out <- tune(svm, Returns~. , data=svm_data, kernel = "radial", cross = 30, scale = TRUE, ranges = list
(cost = c(0.1,1,10), gamma = c(0.5,1,2)))
bestmod <- tune.out$best.model
bestmod$tot.accuracy
```

```
## [1] 42.68657
```

We can see that the average, cross-validated accuracy of this method out-performs the cross-validated LDA, but is still roughly on-par with the not-cross-validated LDA, hardly much of an improvement.



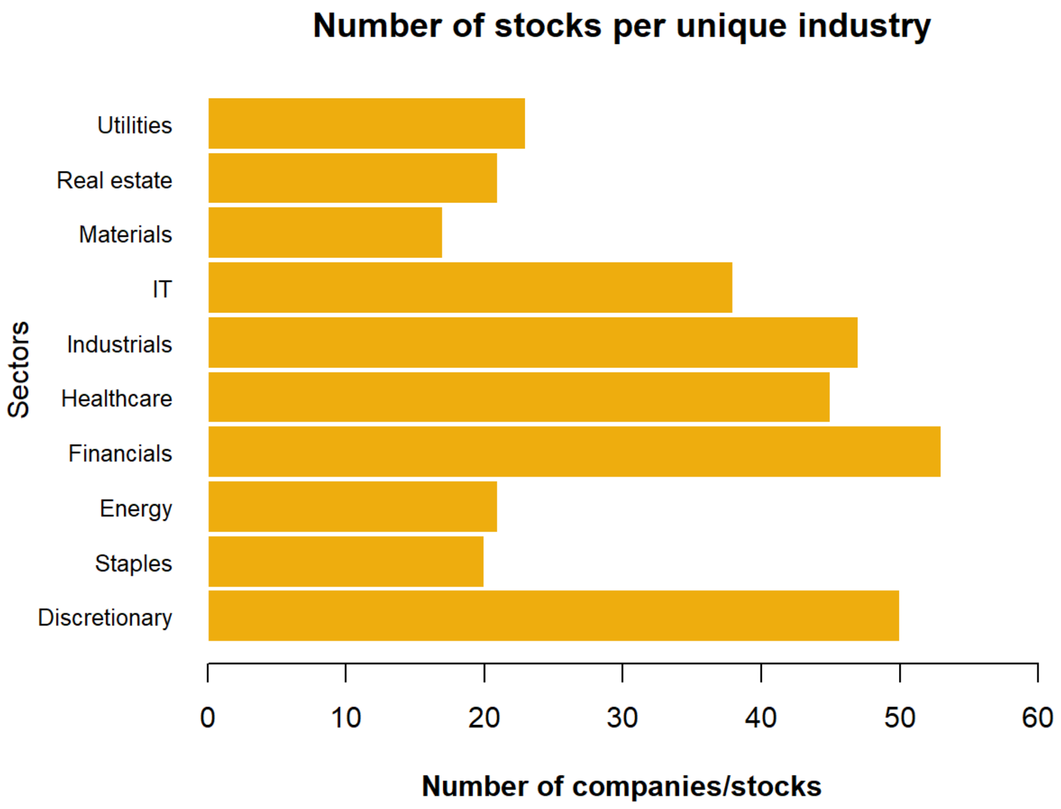


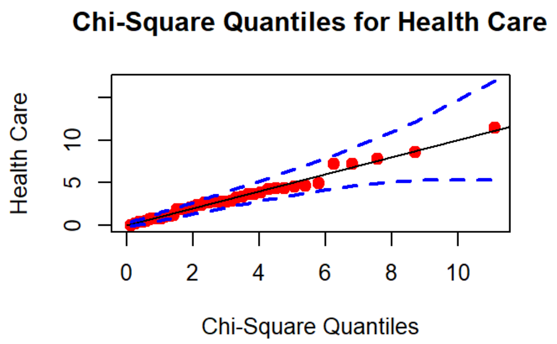


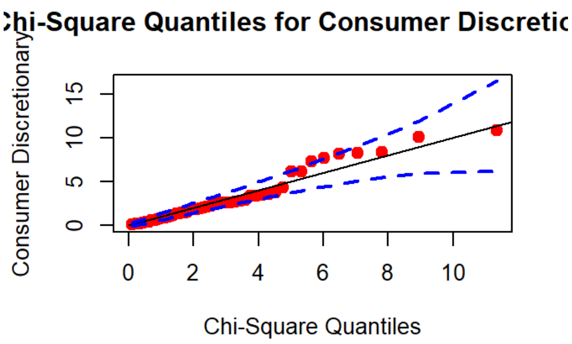
Support vector machine also has a cool “density” representation of its classification boundaries/confidence, analogous to the ggplots we constructed earlier for LDA classifications. We can see the model’s Guassian classification regions (not linear like LDA) and its classification confidence by its increasing color intensity (white-banded regions are the uncertainty of the boundary conditions). It should be noted though that, 1) this model also didn’t do a great job, so this is mostly for visual appeal rather than a true partition of space according to accurate classification, and 2) this R function can only take two arguments, so we were forced to put our stocks into strictly “Higher” or “Lower” (no Mediocre group), so this graph is somewhat different from the previous pictures (however, the SVM numerical classifications above did use all three groups).

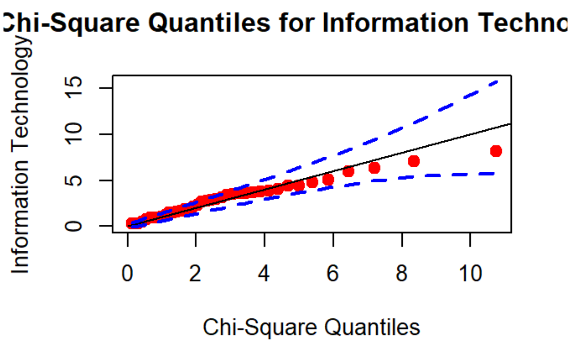
Section 3: Cluster Analysis

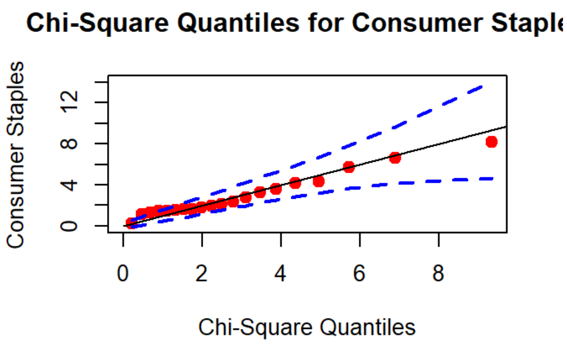
In this section, we will try to answer the following question: **“If we choose to cut our data into 10 clusters, will the companies be grouped roughly into their respective sectors?”** Answering this question will help us consider clustering and predicting within industries. One motivation behind this is that at a given year, some industries may be subject to unique industry-specific shocks that are not captured by the financial variables we used to predict stock-price performance.
There are 10 Unique Industries in our dataset!

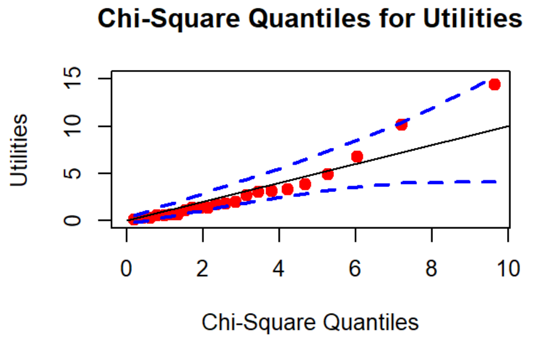


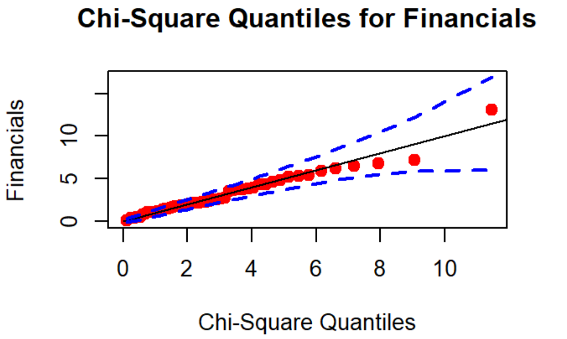

Chi-Square Quantiles for Health Care
Health Care
0
5
10
0
2
4
6
8
10
Chi-Square Quantiles

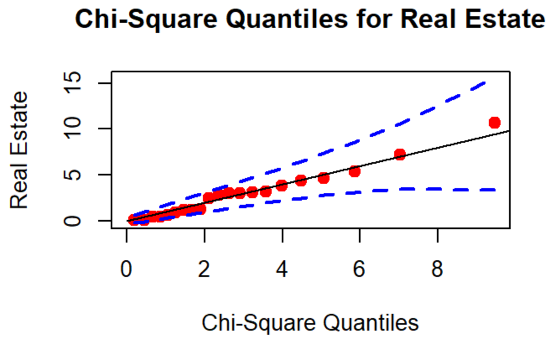

Chi-Square Quantiles for Consumer Discretic
Consumer Discretionary
0
5
10
15
0
2
4
6
8
10
Chi-Square Quantiles

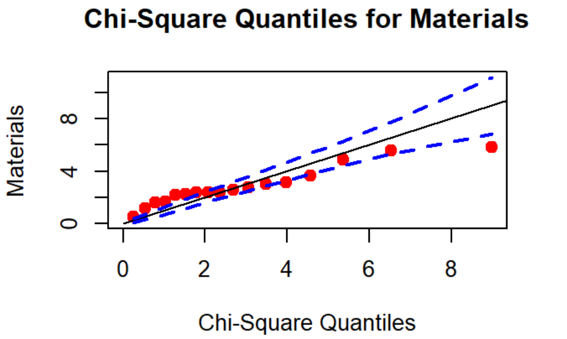

Chi-Square Quantiles for Information Technc
Information Technology
0
5
10
15
0
2
4
6
8
10
Chi-Square Quantiles

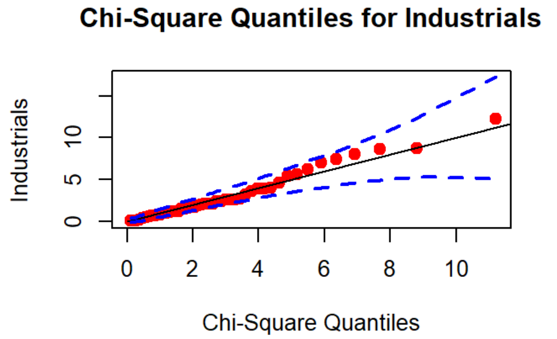

Chi-Square Quantiles for Consumer Staple
Consumer Staples
0
4
8
12
0
2
4
6
8
Chi-Square Quantiles

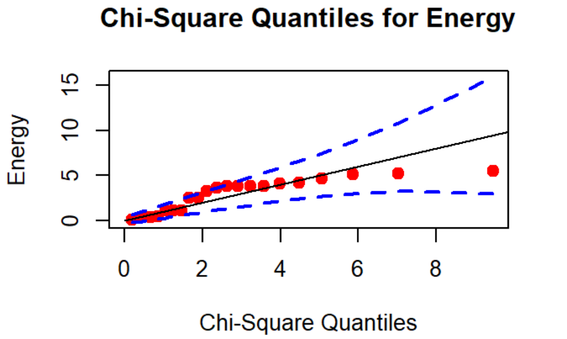

Chi-Square Quantiles for Utilities
Utilities
0
5
10
15
0
2
4
6
8
10
Chi-Square Quantiles


Chi-Square Quantiles for Financials
Financials
0
5
10
0
2
4
6
8
10
Chi-Square Quantiles


Chi-Square Quantiles for Real Estate
Real Estate
0
5
10
15
0
2
4
6
8
Chi-Square Quantiles

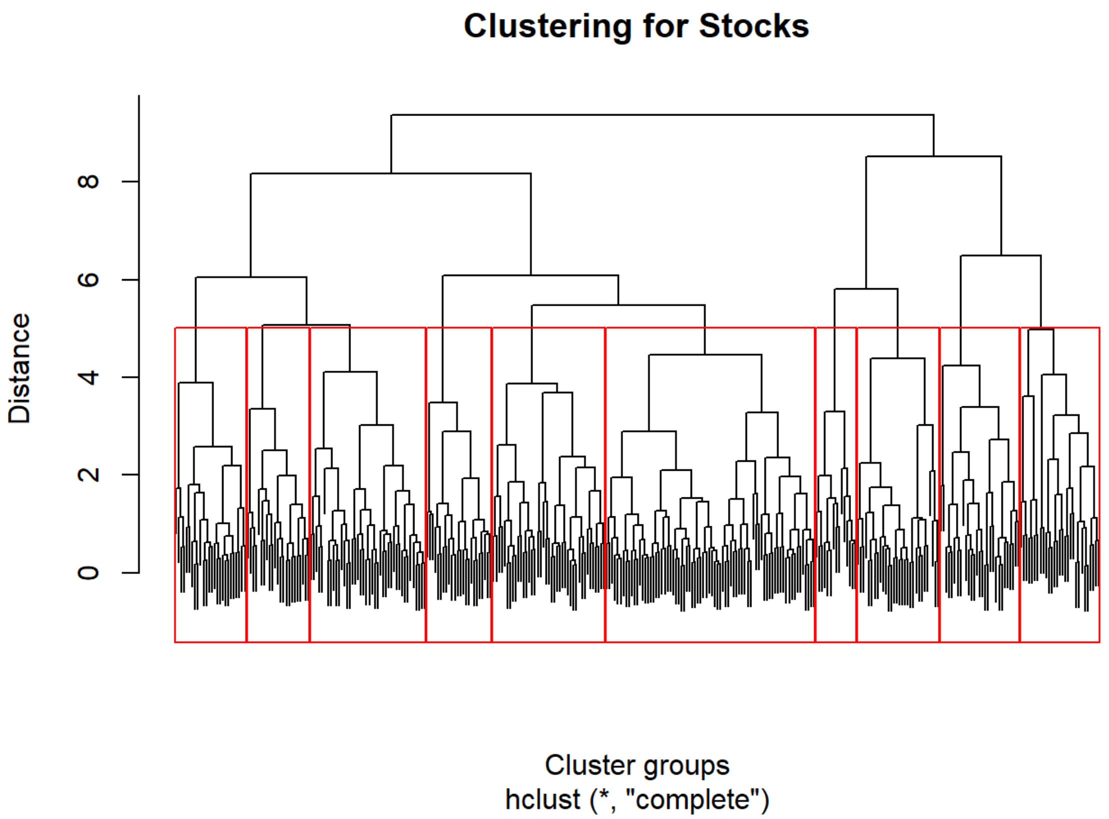

Chi-Square Quantiles for Materials
Materials
0
4
8
0
2
4
6
8
Chi-Square Quantiles


Chi-Square Quantiles for Industrials
Industrials
0
5
10
0
2
4
6
8
10
Chi-Square Quantiles


Chi-Square Quantiles for Energy
Energy
0
5
10
15
0
2
4
6
8
Chi-Square Quantiles

According to these Chi-square quantile plots, within each sector, the data appears to have an approximately multivariate Normal distribution.

We tested a couple of different ways to perform cluster analysis to try and find the distance-linkage combination that maximizes the mean membership rate across all clusters, while also giving us the most distinct clusters. We tested different distance ("euclidean" vs. "manhattan") and linkage/clustering methods ("single", "complete," and "average"), and ultimately settled on using *"Euclidean"* distance and *"complete"* linkages, which makes sense: "complete" linkages use the maximum distance between clusters and tend to make trees with distinct clusters, while also being less sensitive to noisy data. When we cut on 10 clusters, here are the results we obtain:



We can determine what proportion of each cluster is made up of stocks/companies from the highest represented industry, which we call the "membership rate." For example, if in cluster 1 the largest proportion of stocks belong to Industry A, the membership rate would be that proportion. This gives us an average membership rate of 0.3642. Listing our clusters by the industry most represented in each with the proportion of represented gives us the following results:

```
##                 industries  rates
## 1              Industrials 0.2759
## 2  Consumer Discretionary 0.2237
## 3              Health Care 0.3478
## 4              Health Care 0.3846
## 5                Utilities 0.5000
## 6               Financials 0.7667
## 7  Consumer Discretionary 0.4390
## 8               Financials 0.2069
## 9              Real Estate 0.8667
## 10              Financials 0.1667
```

To understand the threads tying each cluster together–i.e. what made the stocks in them similar–we ultimately had to go beyond simply using the membership rate of the industries. Although we cut our data into 10 clusters, they were clearly not grouped into their industry categorization: consumer discretionary and healthcare each make up the greatest proportion of two clusters; financials makes up the biggest proportion of three clusters. We decided to look at the stock tickers and the kinds of companies represented in each cluster in order to better understand the patterns.

Ultimately, we can see that there are factors which differentiate each cluster by industry–just not the broad, categorized industries that the data had been pre-sorted into. The clusters instead gave us a more fine-tuned picture of what industry each company belonged to.

Cluster 1: While “Industrials” only make up 27.59% of the stocks in this cluster, looking closely at the tickers and company descriptions shows that many “industrial” companies like AMETEK (AME), Fortive (FTV), and Rockwell (ROK) focus on industrial technologies–such as information technology or industrial automation. Meanwhile, many of the remaining non-industrial companies are also technological: we can see that Apple (AAPL), and Microsoft (MSFT) are among them, tech giants likely with industrial applications or similar supply chains. Thus, if we were to rename this cluster, we might call if industrial technology.

Clusters 2 and 7: Both of these clusters show “Consumer Discretionary” to be the highest proportion of stock tickers. Doing closer analysis of which stocks are “discretionary” and which are not within each cluster allows us to better determine the nature of the clusters. In cluster 2, most stocks–regardless of their classification–are concerned with transportation and department stores that mainly apply to the home: for example, Advance Auto Parts (AAP), LKQ corporation (LKQ) and Genuine Parts Company (GPC) all have to do with automotive parts, while Lennar Corporation (LEN) deals in home construction and upgrading. Various other companies work run cruise lines, manufacture shipping and motorcycle parts, or distribute home products. Thus, we can see that cluster 2 focuses on consumer discretionary spending which deals with **home and travel.**

For cluster 7, the stocks much more clearly represent companies which run wholesale and department stores, which generally focus on personal products and clothing. For example, Best Buy (BBY), Dollar General (DG), Dollar Tree (DLTR), Gap (GPS), and Macy’s Inc (M) are all companies which sell **everyday consumer products**–for household and personal use. While these products all fit under discretionary spending as well, they are clearly a different tier of product and an industry which caters to different consumer desires than transportation and homes. (If we had more information about consumption habits in the US, it would be interesting to look further into the difference between these two clusters, since it seems that middle- and high- income families would be much more likely to spend greater discretionary amounts on home and travel, whereas lower-income families would be more likely to shop at places like Dollar General.)

Clusters 3 and 4: Both these clusters are classified as “Healthcare” although the proportion is low in each one. Zooming closer and looking at the individual stocks again gives us a breakdown of the differences between these two clusters. In cluster 3, we can see that most of the companies–like Abbott Laboratories (ABT) and Perrigo (PRGO) are pharmaceutical or biotech companies, which manufacture drugs, OTC medication, vitamins, and supplements for patient usage. Thus, they operate in similar markets and probably also use similar supply chains, which might lead us to think of this cluster as a distinct segment of the healthcare sector, dominated by pharmaceuticals.

Cluster 4, by contrast, is mostly filled with companies that produce medical instruments and health equipment, such as Stryker (SYK)-which specializes in what is needed for trauma surgeries, as well as Medtronic (MDT) and Zimmer Biomet (ZBH)–which are both medical device companies. Some other companies that are not classified as “healthcare” may show similarities to the production of devices and parts which these med-product companies engage in: Roper Technologies (ROP), for instance, is a diversified giant which makes engineered products and parts, Analog Devices (ADI) makes semiconductors, and Akamai Tech (AKAM) provides software. We can see a difference between cluster 3 being industries that have to deal with **chemical production**, whereas cluster 4 seems more **mechanical and engineered parts manufacturing.**

Cluster 5: This cluster appears relatively accurate - half of the stocks are classified as "utilities", and the ones that are not tend to be insurance companies, such as Prudential (PRU), American Express (AXP), and Lincoln Management (LMC)–companies that also work with the everyday American consumer (we can think of them as service-oriented companies for which consumers are most likely to get monthly billed expenses, so perhaps they have similar markets and cost structures). It is also likely that many of these companies must cross-coordinate in order to respond to consumer crises–think, for example, in the aftermath of a natural disaster: utilities, insurance, and credit companies must often cooperate to respond to people's immediate needs. Thus, while other industries are captured within this cluster, it is not difficult to see the similarities between the sectors in which they operate.

Cluster 6, 8, and 10: These three clusters are all labeled as dominated by "financial" industry companies–which would appear discouraging at first. However, "finance" is a diverse industry marked by firms operating in very different kinds of financial practices. The stock tickers differentiate this as well.

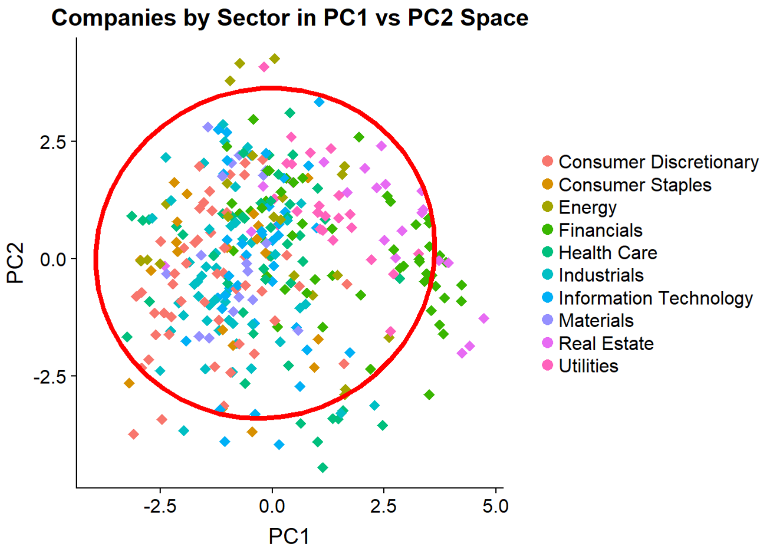
Cluster 6 has the highest percentage of stocks classified as "financial," and it is not difficult to see why: most of the stocks in this cluster are **large banks or financial institutions** that move around large blocks of money, such as BlackRock (BLK) and Bank of America (BAC). The ones which are non-financial still engage in substantial financial activity: Ventas (VTR) for example, is not classified as financial because it is a REIT (Real Estate Investment Trust)–thus it is probably labeled real estate–but its activities and the way it behaves as a company is financial in nature.

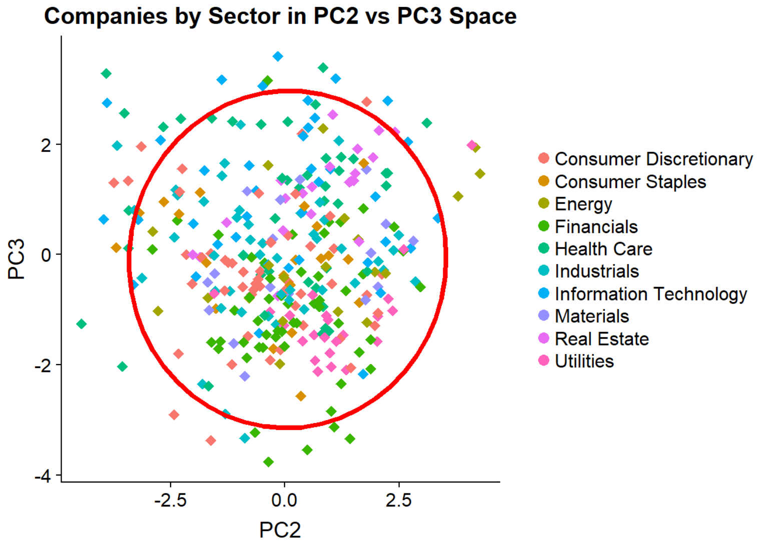
Cluster 8 has a lower membership rate: here, the financial companies are made up of **multistrat investment management** firms, which means that through diverse investments, they engage in several different activities in the market. This means that their performance is likely dependent upon and correlated with the performance of a large swath of other key players in different industries across the market, which we can identify in the remainder stocks which are primarily nonfinancial.

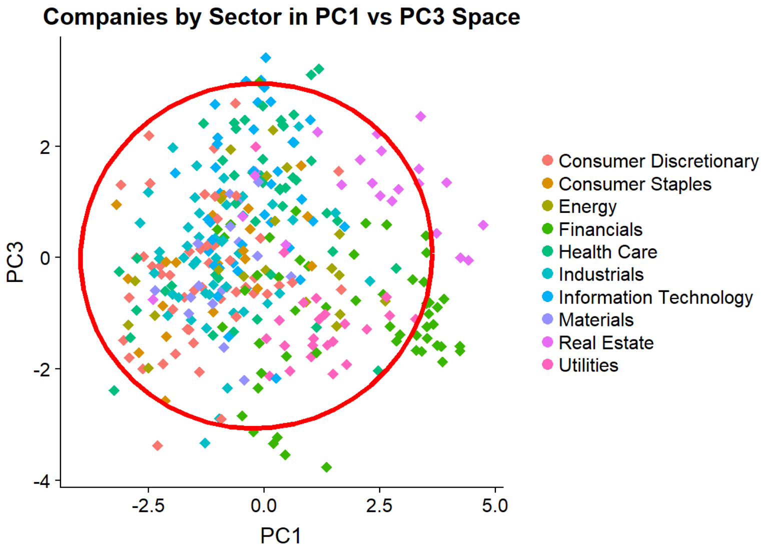
Likewise, cluster 10 has the lowest membership rate for financials, and we can see that it is because these financial institutions which primarily sell **insurance** as a financial product (a different product from the home insurance tied together with utilities in a previous cluster).For example, Progressive Corp (PGR) is a conglomerate insurance company. Note that this is also a relatively small cluster.

Cluster 9: The last cluster is real estate - which is relatively self-explanatory, and the cluster has the highest membership rate out of all the other clusters (86.67% of the stocks are directly classified as real estate). The stocks which are not classified as real estate still have activities concerned with the real estate/housing/mortgage market–for example, People's United Financial (PBCT) is not classified as "real estate", but likely fell into the cluster because of providing mortgages and other housing or real-estate related financial products which can cause these stocks' performances to be correlated. This is the most straightforward cluster which did match our original intention to have each cluster be an industry.

At the end of cluster analysis, we get 10 clusters which did not match the original preassigned industries. However, we can draw out similar characteristics among the stocks to see that the clusters roughly describe: 1) industrial technology, 2) home and travel discretionary spending, 3) pharmaceutical and chemical production, 4) health equipment and other engineered products, 5) utilities and consumer credit, 6) large banks and financial institutions, 7) personal products discretionary spending, 8) multistrategy investment management firms, 9) real estate, and 10) insurance financial firms.



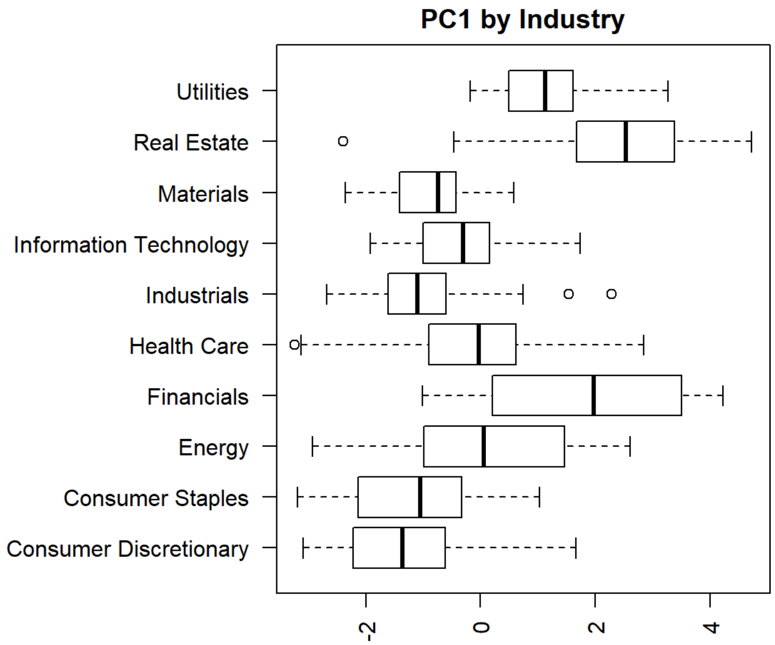


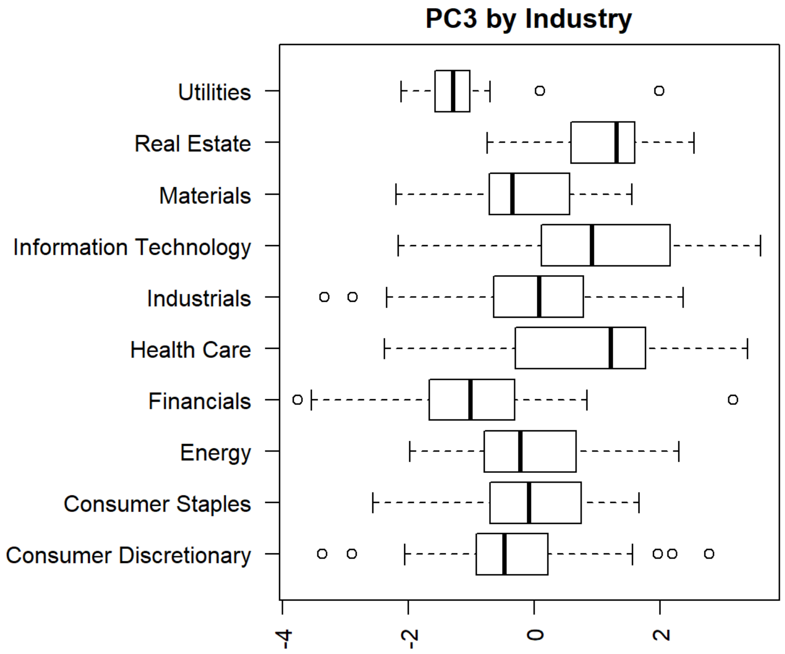


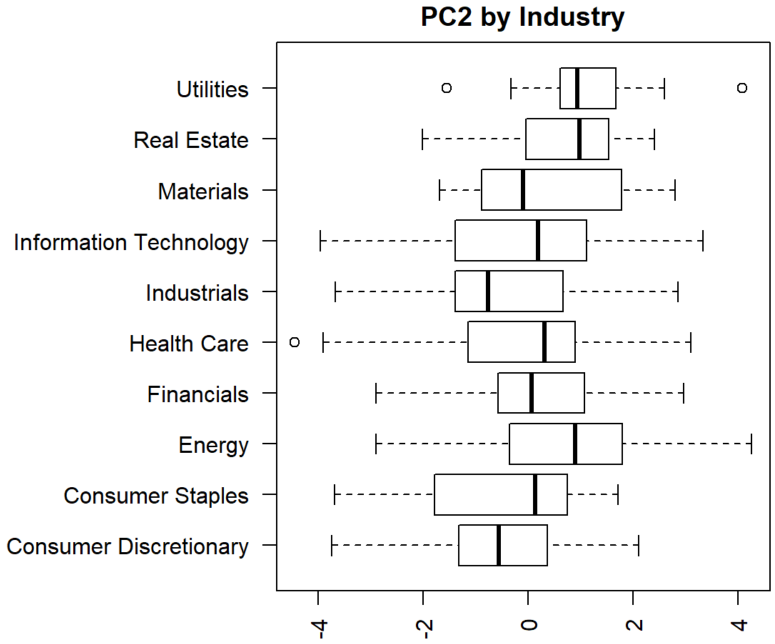
Projecting down into 2D PCA-space and coloring by industry, we can see that no single industry seems to be particularly distinct/sequestered/outlier, and that for the most part, the industries seem to be spread relatively uniformly across all three principal components.

Section 4: MANOVA

Let’s take a look at boxplots of each stock’s principal component values, grouped by industry.







These boxplots provide a bit more detailed insight into what was alluded to above. While it was somewhat difficult to discern which sectors, if any, seemed to have any substantial deviations with respect to the others, these boxplots reveal that there may be some differences after all, particularly in PC1 which concerns revenue and costs of goods sold.

Let’s run a MANOVA Wilks Lambda test to determine if there are in fact statistically significant differences in the multivariate means across all 10 sectors.

```
##                           Df  Wilks approx F num Df den Df    Pr(>F)
## factor(PCA_Data$Sectors)   9 0.3141   17.014     27 943.97 < 2.2e-16 ***
## Residuals                325
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Yes, the p-value is tiny! Thus was have reason to **reject the null hypothesis** that the differences in the multivariate means are 0 (any observed differences would have been due to chance alone – not!), and therefore there is evidence that membership to a particular sector may be indicative of more wide-reaching differences in key financial ratios and performance metrics.

Two sectors that are potentially responsible for this difference are Financials and Healthcare. Between 2017-12-31 and 2018-12-31, the financial sector stock far outperformed the healthcare sector. There are two potential real life explanations for this performance disparity. Given the overvaluation risk surrounding the various pharma and biotech companies, when global growth slowed down and uncertainty about financial outlook increased towards the end of 2018, healthcare sector was more heavily penalized. Moreover, in this risky environment, central banks became dovish for interest rates and decided to ensure financial stability–which created a relatively positive environment for the financials sector. The increased riskiness and higher debt levels of the healthcare sector are captured within our model through given financial ratios, but the external shock including global growth slowdown are not reflected in our prediction model.

If the financial ratio and performance metric explanatory variables we are tracking do indeed have predictive power of future performance we might have reason to suspect that the Financial sector multivariate mean would be significantly different than the Healthcare sector multivariate mean.

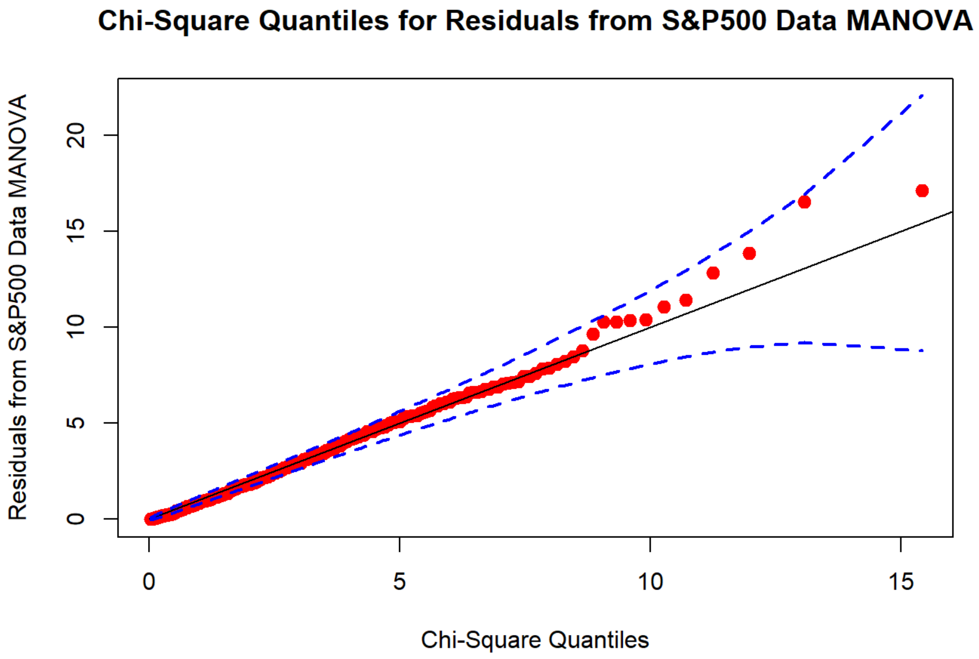
```
##  [1] Consumer Discretionary Consumer Staples       Energy
##  [4] Financials             Health Care            Industrials
##  [7] Information Technology Materials              Real Estate
## [10] Utilities
## 10 Levels: Consumer Discretionary Consumer Staples Energy ... Utilities
```

```
source("https://raw.githubusercontent.com/davidlieberman/SDS363/master/PSet5/multicontrast.R")
multicontrast(c(0,0,0,1,-1,0,0,0,0,0), PCA_Data[sorted,3:5], PCA_Data$Sectors[sorted])
```

```
##       Wilks    approx.F         df1         df2     p.value
##   0.6883486  48.7463286   3.0000000 323.0000000   0.0000000
```

Yes! The F-statistic is MASSIVE and the p-value is 0.00000. Therefore we **reject the null hypothesis** that the differences in the multivariate means is due to chance alone, and there is reason to suspect that the difference between the multivariate means on 2017-12-31 is due to the fact that ultimately one would evolve to be an extremely lucrative investment while the other would flop.

We finally want to verify if our MANOVA residuals follow a multivariate Normal distribution, as we would expect. Otherwise we might be performing tests that aren’t especially appropriate.



Yes, they do. Our MANOVA analysis is reliable!

Conclusion

To summarize our steps, we first conducted PCA and picked 3 principle components that captured high variability in our data. We then performed Linear Discriminant Analysis, k-Nearest Neighbor, and Support Vector Machine using our 3 Principal Components to see if we can predict a company's stock price performance relative to the stock price performance of S&P 500 index. However, while our prediction was more informed than blind prediction, we were not able to obtain a reliable model that would make profitable projections. We then asked if we could gain more information about financial ratios across industries. We tried to use cluster analysis to see if industry clusters would arise naturally from our data if we grouped on 10 clusters, but more successfully, in our MANOVA study, we saw that multivariate mean across all industries is not strictly conserved.

Overall, we learned that while key financial ratios are somewhat informative about a company's stock performance in a given year, they alone cannot accurately predict to the degree that one can rely on the prediction results and hope to achieve positive financial gains from investing in high performing stocks. These results are undesrstandable. Part of the reason our model may be failing to capture high accuracy is that it has limited input. A company's stock price performance depends on many factor that are hard to quantify: Management Quality, Pressures from Suppliers, Pressures from Competitors, Overvaluation Risks, Market Sentiment about the Company, etc. Yet, our project and models are still useful when combined with additional qualitative analysis. We see that our model successfully captures the differences across industries, and outliers within industries (even though we recognize we did remove extreme outliers at the outset). Using our model, we can identify the companies with significantly strong financial ratios that are above industry average and call them potential targets. We can also identify the companies with significantly poor financial ratios that are notably below industry average and call them risky companies. We can use these groups as a screen, we can avoid very risky companies, and conduct more careful analysis on the selected potential targets. With this, we can save research time, identify companies with great potential to carefully study these companies with high potential of outperforming S&P 500 and bringing significant returns to investors.