Milestone 3 New Additions Group 4

PCA

```
Standard deviations (1, .., p=6):
[1] 2.326420e+11 6.536499e+09 1.275279e+05 3.536727e+02
[5] 2.434643e+00 7.125505e-06

Rotation (n x k) = (6 x 6):
                             PC1           PC2           PC3
Currentprice        6.071158e-10 -5.458383e-09  2.356367e-04
Marketcap           9.990311e-01 -4.400922e-02 -3.908048e-09
Ebitda              4.400922e-02  9.990311e-01 -4.179373e-06
Revenuegrowth      -1.957633e-13 -1.683860e-11 -4.488302e-07
Fulltimeemployees   1.878351e-07  4.175154e-06  1.000000e+00
Weight              2.339998e-14 -1.093757e-15 -1.032925e-11
                             PC4           PC5           PC6
Currentprice        1.000000e+00  1.744690e-04  2.847096e-09
Marketcap          -8.458258e-10 -6.948045e-13 -2.342791e-14
Ebitda              6.411188e-09  1.607363e-11  3.834060e-17
Revenuegrowth      -1.744689e-04  1.000000e+00  2.503437e-08
Fulltimeemployees  -2.356368e-04  4.077189e-07  9.669605e-12
Weight             -2.842726e-09 -2.503486e-08  1.000000e+00
Importance of components:
                            PC1       PC2    PC3   PC4
Standard deviation     2.326e+11 6.536e+09 127528 353.7
Proportion of Variance 9.992e-01 7.900e-04      0   0.0
Cumulative Proportion  9.992e-01 1.000e+00      1   1.0
                          PC5       PC6
Standard deviation     2.435 7.126e-06
Proportion of Variance 0.000 0.000e+00
Cumulative Proportion  1.000 1.000e+00
```

	Currentprice	Marketcap	Ebitda
Currentprice	1.00000000	0.36823710	0.26036144
Marketcap	0.36823710	1.00000000	0.84244189
Ebitda	0.26036144	0.84244189	1.00000000
Revenuegrowth	-0.02783496	-0.01861679	-0.04000994
Fulltimeemployees	0.17112352	0.31746401	0.37456243
Weight	0.36805457	0.99999909	0.84240041

	Revenuegrowth	Fulltimeemployees
Currentprice	-0.02783496	0.17112352
Marketcap	-0.01861679	0.31746401
Ebitda	-0.04000994	0.37456243
Revenuegrowth	1.00000000	-0.03664587
Fulltimeemployees	-0.03664587	1.00000000
Weight	-0.01861419	0.31722433

	Weight
Currentprice	0.36805457
Marketcap	0.99999909
Ebitda	0.84240041
Revenuegrowth	-0.01861419
Fulltimeemployees	0.31722433
Weight	1.00000000

The first tecchnique used is PCA on the company.csv. Due to limited time to remake this milestone we are going to briefly cover what we found and how we can address it for the final. The general idea for PCA on the company's data set was utilizing the current price, marketcap, ebitda, revenuegain, and full time employees. Utilizing the current price as the independent variable. After running our PCA initial run we found that in the Cumulative Proportion PC1 has a value way above 1. That would mean scaling is an issue and we would address that later for the final. We also provided correlations to the Comapny and numeric values. We might look at a regular regression and see if there has been a change after scaling is applied to this dataset.

Looking at the highest correlation value overall is Weight and Marketcap 0.99999909. The lowest overall value was with Revenuegrowth and Fulltimeemployees at -0.03664587.

Per each variable Currentprice the highest correlation is to Marketcap at 0.36823710 and the lowest is Revenuegrowth at -0.02783496. Marketcap the highest is Weight at 0.99999909 and the lowest is Revenuegrowth at 00.01861679. Having a variable at 99% correlation means that Weight will influence the Marketcap and if per day goes up because of each and vice versa. Ebitda the highest is Marketcap at 0.84244189 and lowest is Revenuegrowth at -0.04000994. Revenuegrowth the highest is Weight -0.01861419 and lowest Ebitda -0.040000994. Revenuegrowth has the most negative influencing variables in correlation. It also has no positive variables correlated. For Fulltimeemployees the highest is Ebitda at 0.37456243 and lowest is Currentprice 0.17112352.

We can also see that we get about 99% of our variance in PC1. By PC3 we already have about 100% therefore we only probably need PC1, PC2, and possibly PC3; however, it is unlikely 3 is necessary. This could be an issue with scaling that only PC1 and PC2 had significance to our results that PC3 and PC4 had no value.

Principal Component Formulas:
PC1 = 6.071158e-10(Currentprice) + 9.990311e-01(Marketcap) + 4.400922e-02(Ebitda) - 1.957633e-13(Revenuegrowth) + 1.878351e-07(Fulltimeemployees) + 2.339998e-14(Weight)

PC2 = -5.458383e-09(Currentprice) - 4.400922e-02(Marketcap) + 9.990311e-01(Ebitda) - 1.683860e-11(Revenuegrowth) + 4.175154e-06(Fulltimeemployees) - 1.093757e-15(Weight)

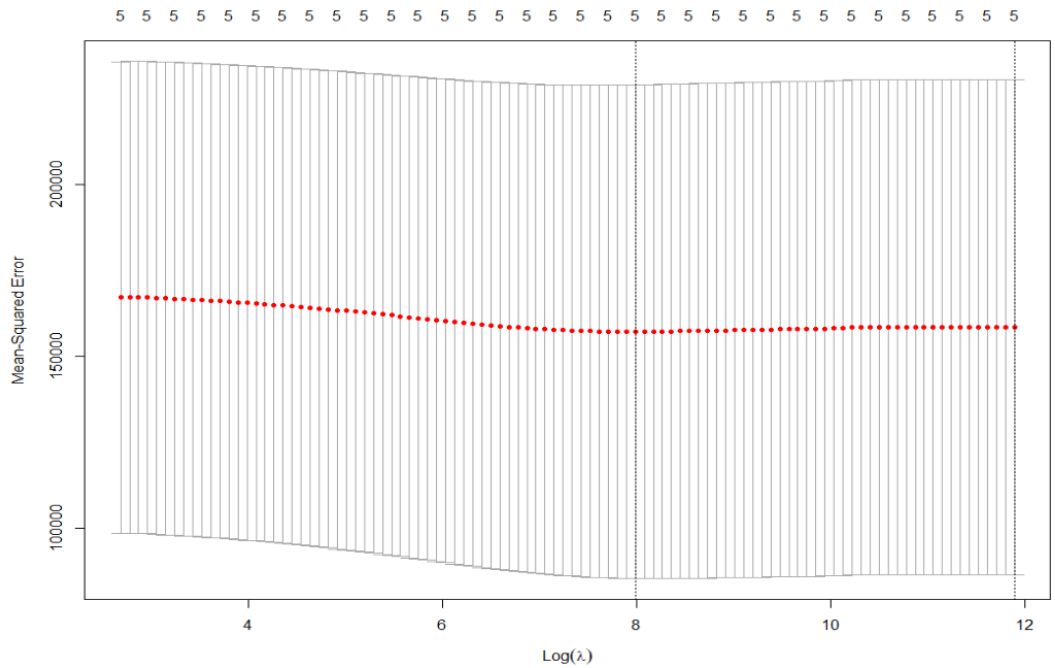
Method 2 Lasso and Ridge Regression on Companies for CurrentPrice

OLS:

```
  Exchange            Symbol           Shortname
Length:402         Length:402         Length:402
Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character


  Longname            Sector           Industry
Length:402         Length:402         Length:402
Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character


 Currentprice       Marketcap             Ebitda
Min.    :  12.13   Min.   :6.602e+09   Min.   :-4.127e+09
1st Qu.:  64.63   1st Qu.:1.807e+10   1st Qu.: 1.020e+09
Median : 117.45   Median :3.238e+10   Median : 2.193e+09
Mean   : 204.49   Mean   :7.969e+10   Mean   : 5.175e+09
3rd Qu.: 221.34   3rd Qu.:6.325e+10   3rd Qu.: 4.633e+09
Max.   :5493.75   Max.   :2.350e+12   Max.   : 1.220e+11
Revenuegrowth         City              State
Min.   :-0.2450   Length:402         Length:402
1st Qu.: 0.0760   Class :character   Class :character
Median : 0.1370   Mode  :character   Mode  :character
Mean   : 0.2887
3rd Qu.: 0.2577
Max.   :22.4860
  Country           Fulltimeemployees Longbusinesssummary
Length:402         Min.   :      0   Length:402
Class :character   1st Qu.:   8725   Class :character
Mode  :character   Median :  19136   Mode  :character
                   Mean   :  52382
                   3rd Qu.:  50950
                   Max.   :2200000
    Weight
Min.   :0.0001546
1st Qu.:0.0004232
Median :0.0007583
Mean   :0.0018665
3rd Qu.:0.0014814
Max.   :0.0551192
```

Ridge:

```
[1] "Training set RMSE:"
[1] 364.6091
[1] "Testing set RMSE:"
[1] 312.7322
[1] "Min Lambda:"
[1] 2940.453
[1] "Lambda.1se:"
[1] 146347

Call:  cv.glmnet(x = xTrain, y = yTrain, nfolds = 7, alpha = 0)

Measure: Mean-Squared Error

    Lambda Index Measure    SE Nonzero
min   2940    43  157112 71840       5
1se 146347     1  158415 72085       5
[1] "Ridge RMSE:"
[1] 300.8339
[1] "Testing set RMSE:"
[1] 312.7322

Call:
lm(formula = Currentprice ~ Marketcap + Ebitda + Revenuegrowth +
    Fulltimeemployees + Weight, data = df_companies_train)

Residuals:
    Min      1Q  Median      3Q     Max
-1286.7  -115.4   -62.4    16.1  5323.9

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        1.600e+02  2.130e+01   7.509 3.99e-13 ***
Marketcap          1.452e-07  5.896e-08   2.463   0.0142 *
Ebitda            -4.049e-09  2.680e-09  -1.511   0.1316
Revenuegrowth     -6.447e+00  1.580e+01  -0.408   0.6834
Fulltimeemployees -2.009e-04  1.481e-04  -1.356   0.1757
Weight            -6.158e+06  2.516e+06  -2.448   0.0148 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 367.4 on 396 degrees of freedom
Multiple R-squared:  0.1581,    Adjusted R-squared:  0.1475
F-statistic: 14.88 on 5 and 396 DF,  p-value: 2.185e-13


Call:  glmnet(x = xTrain, y = yTrain, alpha = 0, lambda = 2940.453)

  Df %Dev Lambda
1  5 5.69   2940
[1] "------------------------------------------------------------"
```



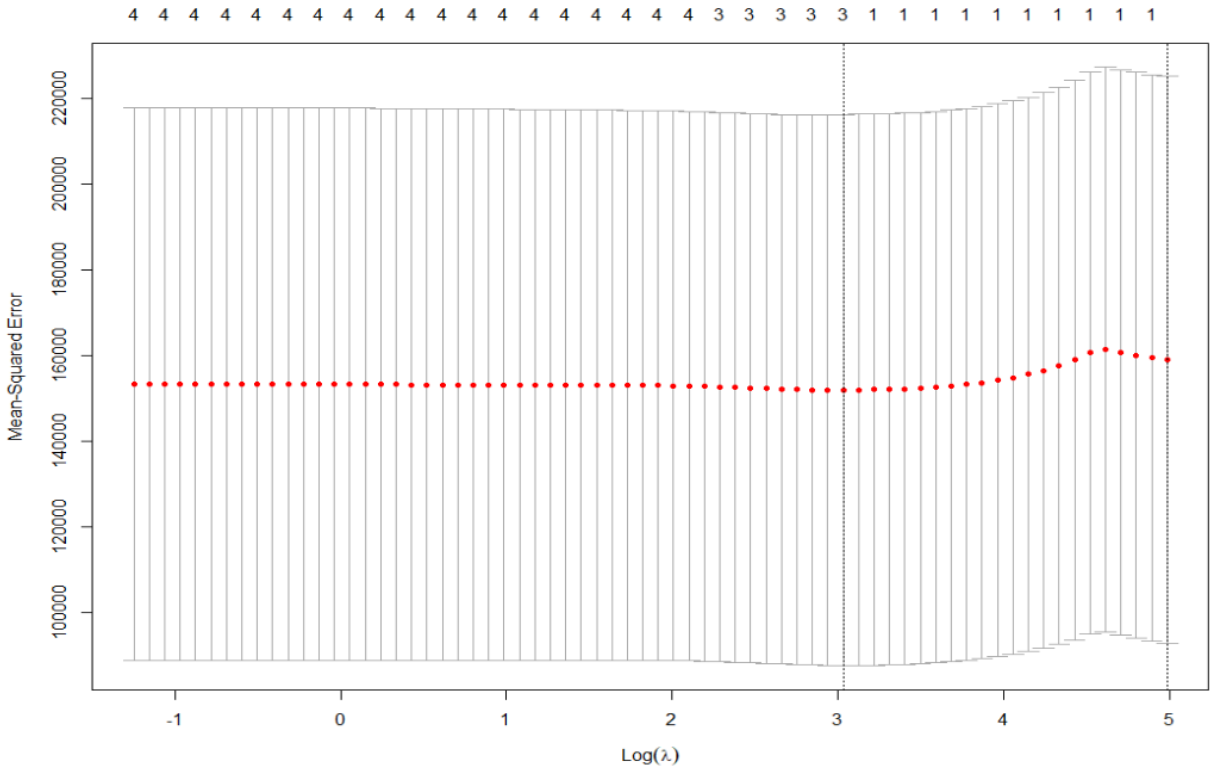
The second technique wasdoing Ridge Regression on the companies table for companies and their currentprice. We first had to take our initial dataset of companies and make that our training set. From there we had to make a test set from it with the help of seed able to split the data. Thus making our OLS test set. We had 402 entries in our training and about 102 in our test set. From there we were able to run the RMSE of both training and test set being 364.6091 and 312.7322 respectively. Those values are about ~52 off from each other that might cause overfitting based on RMSE values.From there mad min lambdaas 2940.453. Our lambda lse was 146347.

After running residuals it can be found on the signfiicnace that Marketcap and Weight had significant stars of 1 at 0.05. The Adj-R^2 is 0.1475 which is rather small. This might contradict that overfitting might not be an issue because of the low R^2 nowhere near 1.0. Another reason why there might not be overfitting is because the test set was not even higher than the training set. It was nowhere near proportionally higher which is a requirement for overfitting.

When looking at the ridge glmnet plot the value of the currentprice of companies did go down after log 6. It decreased from ~17,000 to ~16,500. After the rest of the dataset stayed the saem from log 6 on.

Lasso

```
[1] "Min Lambda:"
[1] 20.74432
[1] "Lambda.1se:"
[1] 146.347
[1] "Lasso RMSE:"
[1] 297.7286
[1] "Testing set RMSE:"
[1] 312.7322

Call:
lm(formula = Currentprice ~ Marketcap + Ebitda + Revenuegrowth +
    Fulltimeemployees + Weight, data = df_companies_train)

Residuals:
    Min      1Q  Median      3Q     Max
-1286.7  -115.4   -62.4    16.1  5323.9

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        1.600e+02  2.130e+01   7.509 3.99e-13 ***
Marketcap          1.452e-07  5.896e-08   2.463   0.0142 *
Ebitda            -4.049e-09  2.680e-09  -1.511   0.1316
Revenuegrowth     -6.447e+00  1.580e+01  -0.408   0.6834
Fulltimeemployees -2.009e-04  1.481e-04  -1.356   0.1757
Weight            -6.158e+06  2.516e+06  -2.448   0.0148 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 367.4 on 396 degrees of freedom
Multiple R-squared:  0.1581,    Adjusted R-squared:  0.1475
F-statistic: 14.88 on 5 and 396 DF,  p-value: 2.185e-13


Call:  glmnet(x = xTrain, y = yTrain, alpha = 1, lambda = 20.74432)

  Df %Dev Lambda
1  3 13.4  20.74
```

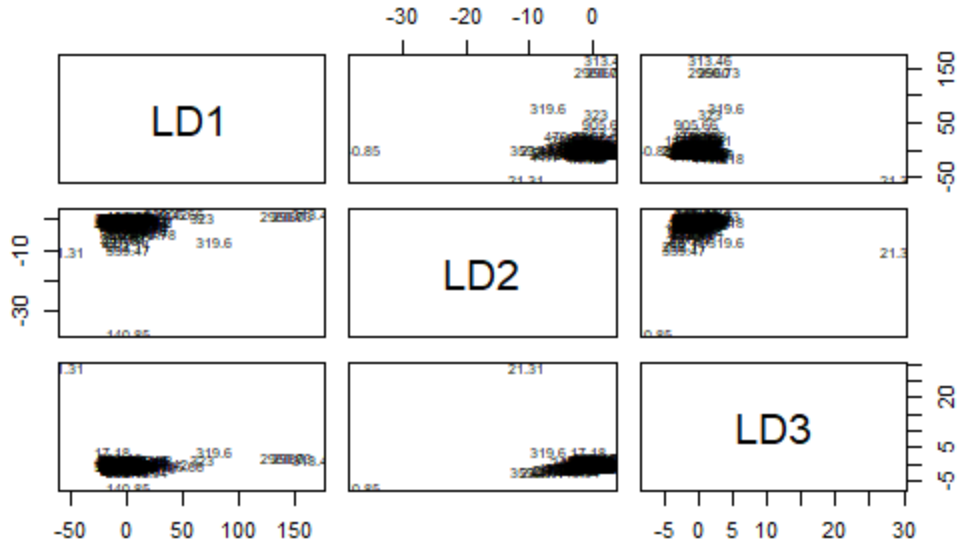


When looking at the second half of the technique this time on lasso regression of the same dataset. After runningLasso Min Lambda was 20.74432 and Lambda.1se wsas 146.347. The same ideaology was applited to the creation of the test set. We made a seed and created a test set based on the same number of observations which was 402 in the training and test was 102. From there the training RMSE was 297.7286 andtesting rmse was 312.7322. If looking at overfitting based on the RMSE of training and test it can be assumed that it does not exist. There is not much of a high difference where it can be felt as overfitting. Again the Weight Marketcap were significant at 1 star equaling 0.05. The Adj-R^2 is the same from the Ridge regression. That again is another reason for no overfitting since R^2 value is nowhere near 1.0. It can be assumed by doing a Lasso regression multicolinearity issues have been removed too.

When looking at the lasso glmnet plot there is an increase of about 1,000 after log 4 on. Which in per company terms if the current price is increasing over 1,000 that means their value of the company is going up and price per raises from the stock market. This is opposite from the ridge regression where it started and went down. The Lasso went up as the log went further on.

Method 3 LDA on Companies

```
counts    399   -none- numeric
means    1596   -none- numeric
scaling    12   -none- numeric
lev       399   -none- character
svd         3   -none- numeric
N           1   -none- numeric
call        3   -none- call
terms       3   terms  call
xlevels     0   -none- list
Call:
lda(Currentprice ~ Marketcap + Ebitda + Revenuegrowth +
Fulltimeemployees,
    data = df_companies_train)
```



While still working on this we started on LDA on the company's dataset. We are looking at the Currentprice on Marketcap, Ebitda, Revenuegrowth, and Fulltimeemployees. From there creation of a confusion plot was made. Plans after we want to understand performance more with a test set and running predictions on the test set. We may discriminate histograms for a further look.

Method Lasso/Ridge Regression not used

```
Call:
lm(formula = TradingVariance ~ ., data = df3)

Residuals:
       Min         1Q     Median         3Q        Max
-5.357e-10  0.000e+00  0.000e+00  0.000e+00  2.102e-08

Coefficients:
             Estimate Std. Error    t value Pr(>|t|)
(Intercept)  1.084e-11  3.089e-14  3.508e+02   <2e-16 ***
AdjClose    -1.056e-13  2.953e-15 -3.577e+01   <2e-16 ***
Close       -1.000e+00  1.652e-14 -6.054e+13   <2e-16 ***
High         1.000e+00  1.698e-14  5.890e+13   <2e-16 ***
Low         -4.148e-15  1.630e-14 -2.550e-01    0.799
Open         3.868e-15  1.585e-14  2.440e-01    0.807
Volume       6.935e-24  9.438e-22  7.000e-03    0.994
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.103e-11 on 999356 degrees of freedom
  (49210 observations deleted due to missingness)
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 2.8e+27 on 6 and 999356 DF,  p-value: < 2.2e-16

[1] 2.103089e-11
```

A second technique is based on calculating a new column called trading variance which takes the high variable minus close variable. From there we can run a lasso or ridge regression to run against mainly the close and volume variables but also looking at the rest too. We noticed throughout the entire Milestone 2 that we had mutliple instances of multicollineraity but no observed overfitting thus doing a regression on lasso feels it would work best. Granted we have not gotten this fully developed yet but if ridge ends up working better we will use that. This is currently our only image we can produce the regression with a calcualted RMSE from the training set. We may end up with a test set for this to fully work but will get done for the final booklet. We plan to make predictions with the use of a test set and cross validate the two matrices. From there we cna use the matrices produced and compare.