Assignment #6

Syamala Srinivasan

Introduction:

Principle Components Analysis is a method of combining variables to ensure that the least important variables are excluded from model creation while the most important remain. This assignment explores utilizing principle component analysis to ensure each variable is free of multicollinearity and is independent of each other.

Results:

Task 1:

The first step to conduct before the data can be utilized in a model is to view the data for inconsistencies and complexities. Because of the nature of our data being daily stock prices, these data points will be highly correlated each day. This presents the requirement to remove this correlation to ensure they are independent. To do so, the log of today's stock price minus the log of yesterday's stock price is taken.

Now we have a data frame with the log returns of the prices without the correlation:

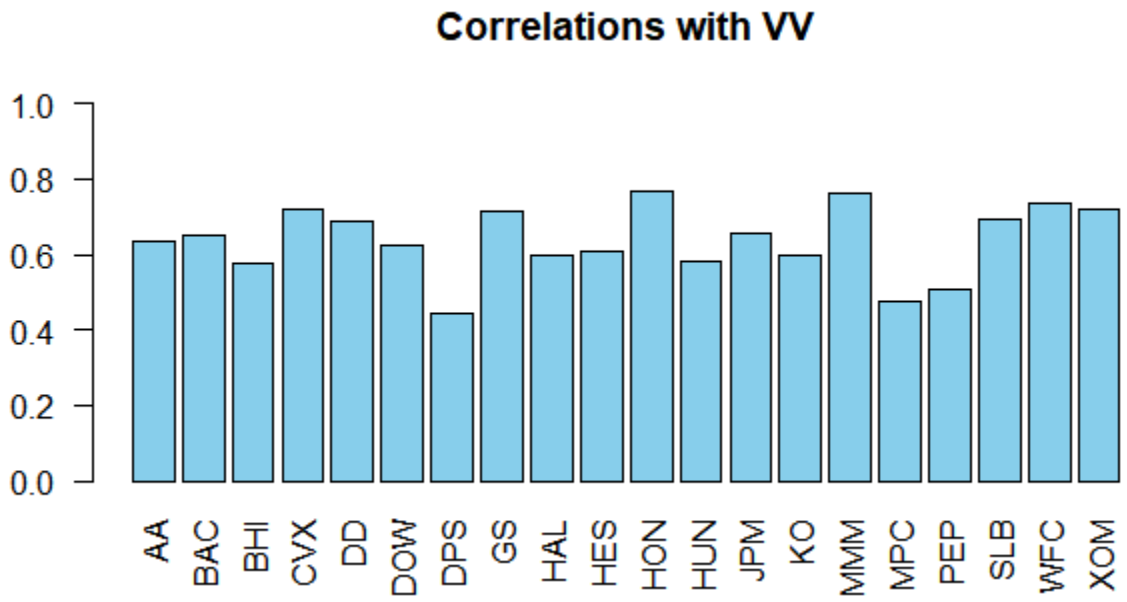
```
'data.frame':   501 obs. of  21 variables:
 $ AA : num  0.02356 -0.00957 -0.0216 0.02799 0.00212 ...
 $ BAC: num  0.00172 0.08256 -0.02082 0.01446 0.05583 ...
 $ BHI: num  0.00995 -0.01387 0.00862 0.00622 0.00715 ...
 $ CVX: num  -0.00172 -0.00985 -0.00727 0.01084 -0.00394 ...
 $ DD : num  0.01091 -0.00683 -0.01423 0.00844 0.01518 ...
 $ DOW: num  0.00536 0.00632 0.00595 -0.00033 0.02186 ...
 $ DPS: num  0.00546 0.00621 -0.00698 0 0.00259 ...
 $ GS : num  -0.00652 -0.00169 -0.01234 0.0135 0.03772 ...
 $ HAL: num  0.028 -0.0161 0.0121 0.0114 0.0265 ...
 $ HES: num  0.01022 -0.02401 -0.0207 0.00847 0.02876 ...
 $ HON: num  -0.0009 0.00108 -0.0074 0.0083 0.01675 ...
 $ HUN: num  -0.00807 -0.00508 0.00811 -0.00608 0.03593 ...
 $ JPM: num  -0.000858 0.020672 -0.009009 -0.001698 0.021024 ...
 $ KO : num  -0.00629 -0.00489 -0.00636 0 0.00608 ...
 $ MMM: num  0.00823 -0.00452 -0.00514 0.00598 0.00511 ...
 $ MPC: num  0.01042 -0.05604 -0.00818 -0.02236 0.02771 ...
 $ PEP: num  0.00511 -0.00782 -0.01261 0.00519 -0.00107 ...
 $ SLB: num  -0.00759 -0.02165 -0.00427 0.01523 0.02766 ...
 $ WFC: num  0.00456 0.01598 -0.00276 0.01236 0.00375 ...
 $ XOM: num  0.000233 -0.003027 -0.007491 0.004454 0.00257 ...
 $ VV : num  0.0012 0.00326 -0.00206 0.00223 0.0092 ...
```

Task 2:

Once our data is correctly defined, we can attempt to understand the correlations in the data.

The below table indicates the correlation of each stock price to the index fund. It is immediately clear that many of the stocks have a very similar correlation to the index fund. There are three stocks that

have less of a correlation; DPS, MPC and PEP. The stocks with the highest correlation are HON and MMM. This would lead me to become most interested in.



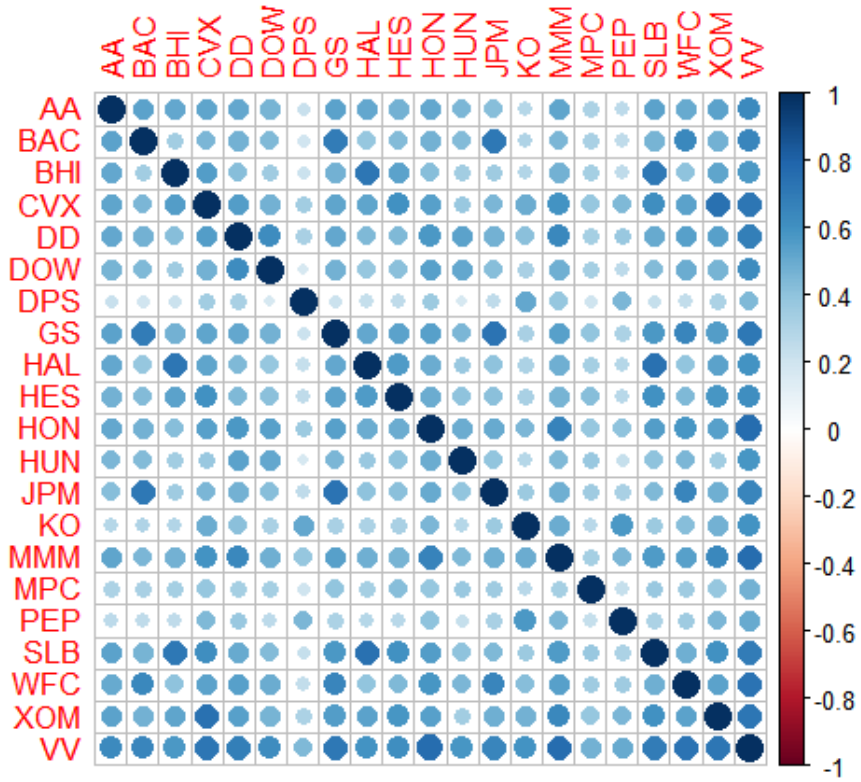

```
       AA       BAC       BHI       CVX       DD       DOW       DPS       GS
0.6324106 0.6501877 0.5774988 0.7209041 0.6895190 0.6264550 0.4435005 0.7121620
      HAL       HES       HON       HUN       JPM       KO       MMM       MPC
0.5974989 0.6107960 0.7683784 0.5819449 0.6578478 0.5997988 0.7608489 0.4731198
      PEP       SLB       WFC       XOM       VV
0.5075264 0.6928534 0.7335731 0.7211079 1.0000000
> |
```

Task 3:

Of course, the stocks can have a correlation to each other as well as the index fund. In the below graph, we are able to indicate which stocks have a higher correlation by the size and color of the circle. We can see that for the majority of the stocks they are most highly correlated with the index fund than any other stock. However, there are a few that stand out. For example, JPM and BAC, HOU and MMM, and SLB and BHI and three examples of stocks that have a relatively high correlation with each other as well as with the index fund.

There are also stocks that are consistently showing a weak correlation amongst all stocks. DPS for example does not have a strong correlation with any stocks.

It is also worth noting that none of the stocks have a negative correlation with each other.



The correlation plot provides an additional level of detail that the bar plot does not. With the use of the correlation plot we are able to easily identify the correlations between individual stocks and their significance. We are also able to easily identify if there is a positive or negative correlation.

Data visualizations and statistical graphs essentially have the same goals, to assist the user in more easily understanding the data. Thus, the two overlap frequently as statistical graphs can be a function of data visualization. Statistical graphs would include the bar graph above while data visualization would include the correlation plot. Fundamentally, whichever is used depends on the ultimate goal and story the analyst is attempting to tell through the data.

Utilizing the correlation plot we can attempt to determine which stocks would have a high or low VIF value.

High: XOM, SLB and MMM

Low: DPS, MPC and PEP

The model created with these stocks has a moderate R-Squared value of .7581 and a very small error of .0038.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.0001357  0.0001699   0.799  0.42480
SLB         0.1352313  0.0151259   8.940  < 2e-16 ***
DPS         0.0912938  0.0196642   4.643 4.41e-06 ***
MPC         0.0502081  0.0093745   5.356 1.31e-07 ***
PEP         0.0814806  0.0268636   3.033  0.00255 **
MMM         0.2976615  0.0269811  11.032  < 2e-16 ***
XOM         0.1823520  0.0276403   6.597 1.08e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.003758 on 494 degrees of freedom
Multiple R-squared:  0.7581,    Adjusted R-squared:  0.7551
F-statistic:   258 on 6 and 494 DF,  p-value: < 2.2e-16
```

Task 4:

To verify the assumption that the VIF values for each stock are able to be predicted via the correlation plot, they are each calculated:

```
     SLB      DPS      MPC      PEP      MMM      XOM
1.777348 1.330570 1.233408 1.510014 2.028488 2.140522
```

XOM and MMM are indeed high as expected, however, SLB is not extremely significant. DPS, MPC and PEP are higher than anticipated but still extremely close to 1 which indicates they have a very low multicollinearity.

To verify if the selected values were indeed able to be estimated based on the correlation graph, the VIF values of the full model will be calculated.

```
     BAC       GS      JPM      WFC      BHI      CVX       DD      DOW      DPS      HAL      HES
2.558097 3.190808 2.844537 2.528808 2.603510 2.909686 2.432674 1.961953 1.524399 2.902240 2.095666
     HON      HUN       KO      MMM      MPC      PEP      SLB      XOM
2.447013 1.721319 1.967512 2.670404 1.376185 1.719788 3.257595 2.924084
```

It is easily noticeable that in fact, the largest VIF values were not for the stocks previously selected. In fact, GS, SLB and XOM are the three largest. This excludes MMM from the earlier estimate which still has a large VIF value but not the largest. In regard to the smallest MPC, DPS, and PEP are the smallest which were the same as the ones selected earlier.

There are no major multicollinearity concerns for either model. GS and SLB should be looked into with more detail to ensure that the correlation is not an indicator of a larger problem, however none of the VIF values are larger than or near 5 which would indicate a pressing concern.

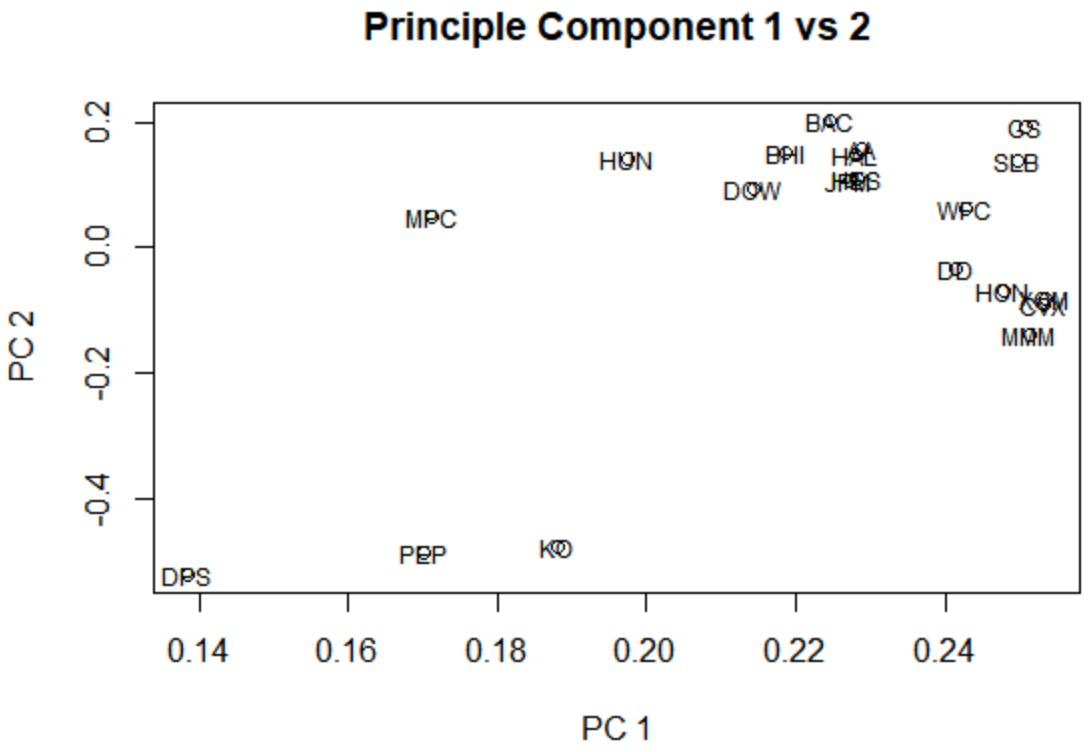
The full model (summary below) has an R-Squared of .88 which indicates that this is likely an accurate and good model to utilize.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.846e-05  1.213e-04   0.730 0.466020
BAC          3.037e-02  9.474e-03   3.205 0.001440 **
GS           3.528e-02  1.358e-02   2.598 0.009675 **
JPM          2.019e-02  1.323e-02   1.526 0.127769
WFC          7.829e-02  1.581e-02   4.952 1.02e-06 ***
BHI          1.834e-02  1.152e-02   1.593 0.111884
CVX          5.925e-02  2.067e-02   2.866 0.004337 **
DD           1.148e-02  1.624e-02   0.707 0.480021
DOW          3.671e-02  1.070e-02   3.431 0.000652 ***
DPS          5.722e-02  1.495e-02   3.828 0.000146 ***
HAL         -5.837e-04  1.208e-02  -0.048 0.961476
HES          4.589e-03  9.699e-03   0.473 0.636297
HON          1.085e-01  1.607e-02   6.751 4.25e-11 ***
HUN          2.988e-02  7.184e-03   4.160 3.78e-05 ***
KO           9.194e-02  1.843e-02   4.990 8.45e-07 ***
MMM          1.117e-01  2.198e-02   5.080 5.41e-07 ***
MPC          1.059e-02  7.032e-03   1.506 0.132741
PEP          2.024e-02  2.036e-02   0.994 0.320703
SLB          4.807e-02  1.454e-02   3.306 0.001019 **
XOM          6.115e-02  2.294e-02   2.665 0.007947 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.002669 on 481 degrees of freedom
Multiple R-squared:  0.8812,    Adjusted R-squared:  0.8765
F-statistic: 187.8 on 19 and 481 DF,  p-value: < 2.2e-16
```

Task 5:

The below plot indicates which stocks have a higher correlation with principle components.

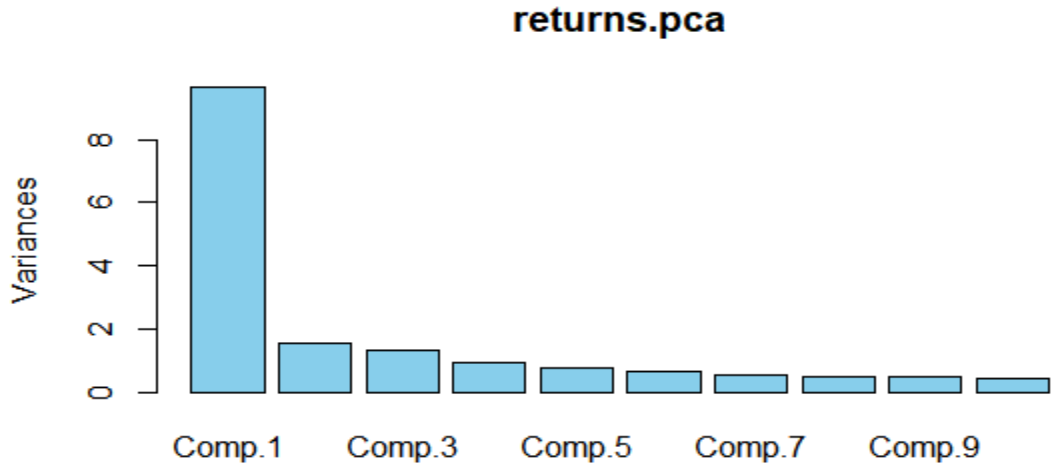
There is a bit of a grouping towards the top right where PC1 is greater than .22 and PC2 is greater than -.02.

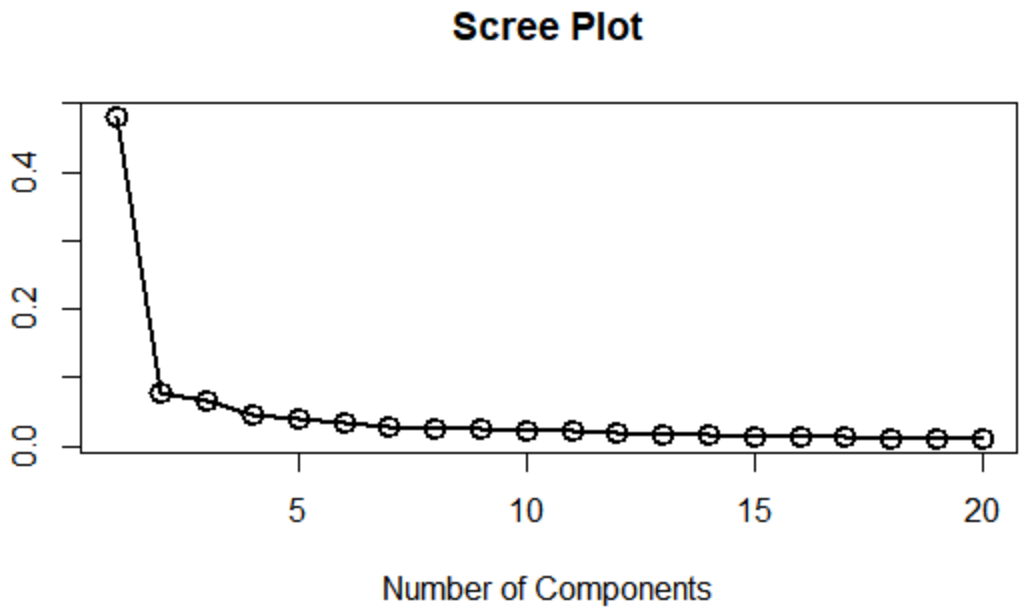


Task 6:

The next step is to plot the variances of each principle component.

In the graphs below it is evident that the first principle component has significantly more variance than the other components. Each of the other principle components have much smaller variances in a decreasing order.

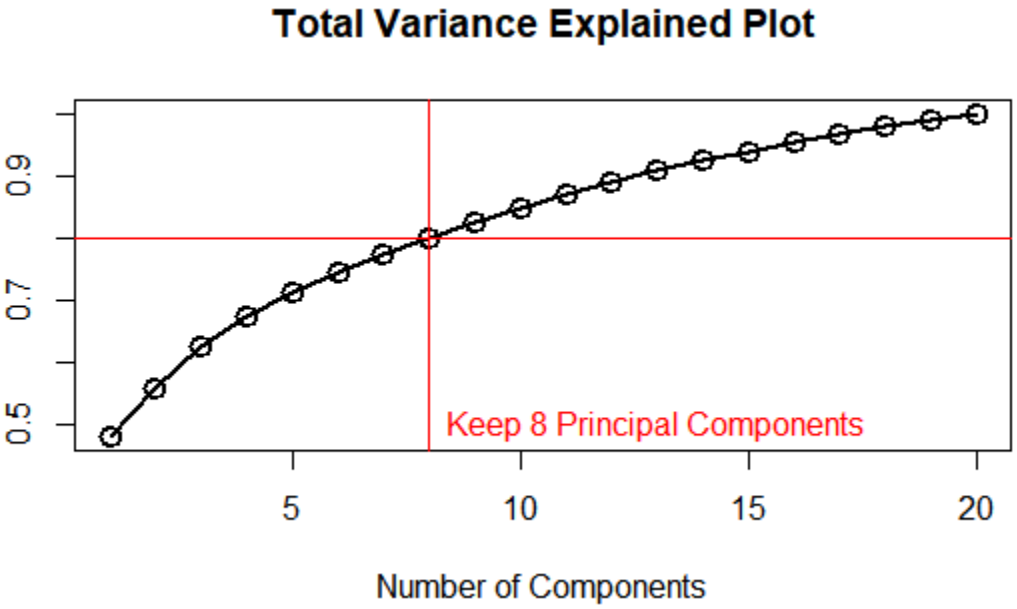




The scree plot indicates further how much more significant the first component is compared to the others.

This is understandable as typically the first five principle components typically explain 86% of variance. To look deeper into this hypothesis, we will see exactly how much each component explains the variance.

Typically we want to include the components that explain between 70 and 90% of the variance. In the data for this assignment this would include 8 components as the 8th component is when we reach exactly 80% of the variance explained.



```
Importance of components:
                          Comp.1     Comp.2     Comp.3     Comp.4     Comp.5     Comp.6
Standard deviation     3.1042633 1.24037473 1.16079159 0.97348817 0.89191173 0.8163381
Proportion of Variance 0.4818225 0.07692647 0.06737186 0.04738396 0.03977533 0.0333204
Cumulative Proportion  0.4818225 0.55874901 0.62612087 0.67350483 0.71328015 0.7466005
                           Comp.7     Comp.8     Comp.9    Comp.10    Comp.11   Comp.12
Standard deviation     0.74727540 0.71606462 0.70486968 0.68141987 0.65836107 0.6385577
Proportion of Variance 0.02792103 0.02563743 0.02484206 0.02321665 0.02167196 0.0203878
Cumulative Proportion  0.77452158 0.80015900 0.82500107 0.84821772 0.86988968 0.8902775
                          Comp.13    Comp.14    Comp.15    Comp.16    Comp.17   Comp.18
Standard deviation     0.5925250 0.57888423 0.54494939 0.52563057 0.50927436 0.4919940
Proportion of Variance 0.0175543 0.01675535 0.01484849 0.01381437 0.01296802 0.0121029
Cumulative Proportion  0.9078318 0.92458713 0.93943562 0.95324999 0.96621801 0.9783209
                          Comp.19    Comp.20
Standard deviation     0.4710181 0.46013436
Proportion of Variance 0.0110929 0.01058618
Cumulative Proportion  0.9894138 1.00000000
```

Task 7:

In order to begin creating a predictive model we must first create the PCA scores. These scores are created for each of the twenty stocks. These stocks are mapped into the principle components to create individual scores.

These scores are then utilized to create two data sets, the training and the test data sets. 70% of the data will be put into the training dataset while the remaining 30% will be put into the test data set.

As determined earlier, utilizing 8 of the principle components will explain 80% of our data, therefore only the first 8 of the 20 principle components will be utilized in the predictive model.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.422e-04  1.447e-04   5.820 1.33e-08 ***
Comp.1       2.221e-03  4.777e-05  46.500  < 2e-16 ***
Comp.2      -4.528e-04  1.138e-04  -3.979 8.42e-05 ***
Comp.3      -2.146e-04  1.321e-04  -1.624    0.105
Comp.4       1.373e-04  1.465e-04   0.937    0.349
Comp.5       1.325e-04  1.643e-04   0.807    0.420
Comp.6       7.483e-05  1.735e-04   0.431    0.667
Comp.7       3.971e-05  1.862e-04   0.213    0.831
Comp.8      -2.456e-04  2.049e-04  -1.199    0.231
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.002737 on 350 degrees of freedom
Multiple R-squared:  0.8623,    Adjusted R-squared:  0.8592
F-statistic:   274 on 8 and 350 DF,  p-value: < 2.2e-16
```

The model for the 8 principle components, pca1, has a large R-Squared value of .86 as well as a very small standard error of only .003.

Additionally, it is beneficial to calculate the mean absolute error. Mean Absolute Error (MAE) indicates the difference between the forecasted value and the actual observed value. This allows an estimate of how large or small the forecasted error to be on average. The MEA for this model is calculated to be

.002. This is an extremely small MAE which leads us to believe that there is a very small risk for error with this model. We also want to calculate the MAE out of sample meaning we want to know what the error is like predicting new values, not predicting already experienced values. The calculated out of sample MAE is also .002.

When calculating the VIF for each of the principle components, they all are calculated to be very close to 1.

```
  Comp.1   Comp.2   Comp.3   Comp.4   Comp.5   Comp.6   Comp.7   Comp.8
1.005105 1.005991 1.010478 1.019104 1.004348 1.011619 1.005242 1.005808
```

Task 8:

Next two models will be created to compare the in-sample and out-of-sample models for the raw data.

The first model created, Model 1, has a slightly higher adjusted R-Squared than previously as the original R-Squared was roughly .75 and is now .78. This incidates that the model may be slightly better with the raw data than the adjusted data.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.0000574  0.0001979   0.290    0.772
SLB         0.1631890  0.0185895   8.779  < 2e-16 ***
DPS         0.1070140  0.0231383   4.625 5.31e-06 ***
MPC         0.0649215  0.0121212   5.356 1.56e-07 ***
PEP         0.0386827  0.0325123   1.190    0.235
MMM         0.3166109  0.0293679  10.781  < 2e-16 ***
XOM         0.1615433  0.0308447   5.237 2.84e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.003693 on 345 degrees of freedom
Multiple R-squared:  0.7925,    Adjusted R-squared:  0.7889
F-statistic: 219.6 on 6 and 345 DF,  p-value: < 2.2e-16
```

The MAE for the training set is .003, this is slightly higher than the previous calculated MAE. When the test set is utilized the MAE is still .003.

Next the full model will be produced.

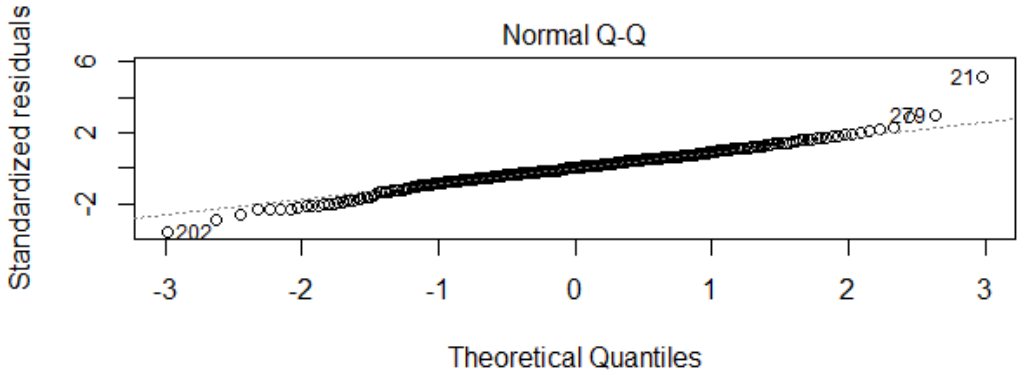
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.926e-05  1.456e-04   0.201  0.84087
BAC          1.655e-02  1.172e-02   1.413  0.15863
GS           3.013e-02  1.591e-02   1.893  0.05922 .
JPM          2.513e-02  1.724e-02   1.457  0.14598
WFC          7.727e-02  1.889e-02   4.090 5.41e-05 ***
BHI          3.174e-02  1.415e-02   2.243  0.02557 *
CVX          6.647e-02  2.478e-02   2.682  0.00769 **
DD          -6.916e-03  1.875e-02  -0.369  0.71240
DOW          3.389e-02  1.308e-02   2.591  0.00998 **
DPS          7.368e-02  1.811e-02   4.069 5.90e-05 ***
HAL          7.149e-03  1.549e-02   0.461  0.64481
HES          1.371e-02  1.220e-02   1.125  0.26156
HON          1.239e-01  2.030e-02   6.102 2.92e-09 ***
HUN          2.570e-02  8.850e-03   2.904  0.00393 **
KO           9.571e-02  2.291e-02   4.177 3.78e-05 ***
MMM          1.318e-01  2.509e-02   5.252 2.70e-07 ***
MPC          1.419e-02  9.342e-03   1.519  0.12963
PEP         -1.524e-02  2.506e-02  -0.608  0.54365
SLB          4.693e-02  1.947e-02   2.411  0.01646 *
XOM          3.858e-02  2.657e-02   1.452  0.14746
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.002683 on 332 degrees of freedom
Multiple R-squared:  0.8946,    Adjusted R-squared:  0.8885
F-statistic: 148.2 on 19 and 332 DF,  p-value: < 2.2e-16
```

The second model produces an adjusted R-Squared of .89. This R-Squared is .02 higher than it originally was. This is not much of a significant increase but is an increase, nevertheless. When the MAE for the training set of the second model is calculated it is .002 which is the same as it was previously. The MAE for the test set also remains the same. The error is slightly less for the full model than the adjusted model.

Model	**Data-Set**	**MAE**
Model 1	Training	.003
Model 1	Test	.003
Pca1	Training	.002
Pca1	Test	.002
Model 2	Training	.002
Model 2	Test	.002

It could be argued that Model 2 is the “best” model due to the low MAE and the higher R-Squared.

In fact, looking at the residuals, it is easy to tell that Model 2 is also very normalized and has very few outliers which may influence the high R-Squared value.



Task 9:

Our decision to only include the initial 8 components limited the scope of the analysis but may have excluded some significant values. For example, when a model of the full data-set is created it is indicated that components 9 through 11 are significant. Additionally the R-Squared value of this model is higher and the error is lower indicating that this could be a more accurate model.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.939e-04  1.453e-04   4.776 2.69e-06 ***
Comp.1       2.281e-03  4.392e-05  51.926  < 2e-16 ***
Comp.2      -4.497e-04  1.167e-04  -3.853 0.000140 ***
Comp.3      -2.108e-04  1.318e-04  -1.599 0.110833
Comp.4       1.737e-04  1.498e-04   1.160 0.246928
Comp.5       7.074e-05  1.766e-04   0.401 0.688924
Comp.6       2.682e-04  1.728e-04   1.552 0.121576
Comp.7       1.128e-04  2.024e-04   0.557 0.577710
Comp.8      -6.294e-04  2.067e-04  -3.044 0.002519 **
Comp.9       5.916e-04  2.147e-04   2.756 0.006174 **
Comp.10     -6.490e-04  2.192e-04  -2.962 0.003282 **
Comp.11      7.859e-04  2.233e-04   3.520 0.000492 ***
Comp.12      3.015e-04  2.244e-04   1.343 0.180033
Comp.13     -8.847e-05  2.443e-04  -0.362 0.717500
Comp.14     -4.694e-04  2.448e-04  -1.918 0.056013 .
Comp.15      1.586e-04  2.661e-04   0.596 0.551505
Comp.16     -3.905e-04  2.794e-04  -1.398 0.163066
Comp.17     -1.951e-04  2.898e-04  -0.673 0.501253
Comp.18      2.758e-04  3.074e-04   0.897 0.370347
Comp.19      7.766e-05  2.984e-04   0.260 0.794810
Comp.20      3.977e-05  3.291e-04   0.121 0.903885
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.002682 on 331 degrees of freedom
Multiple R-squared:  0.895,     Adjusted R-squared:  0.8886
F-statistic:   141 on 20 and 331 DF,  p-value: < 2.2e-16
```

Another method to determine which components to select would have been automated variable selection.

The backwards automated variable selection method suggests keeping 10 components. These components are 1,2,3,6,8,9,10,11,14 and 16. These are very different selections than previously chosen. This new model has a slightly (.03) higher R-Squared than the original and has the same standard error.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0006840  0.0001427   4.792 2.47e-06 ***
Comp.1       0.0022790  0.0000434  52.511  < 2e-16 ***
Comp.2      -0.0004538  0.0001152  -3.940 9.90e-05 ***
Comp.3      -0.0002211  0.0001301  -1.699 0.090271 .
Comp.6       0.0002595  0.0001705   1.522 0.129011
Comp.8      -0.0006383  0.0002039  -3.130 0.001901 **
Comp.9       0.0005893  0.0002114   2.788 0.005595 **
Comp.10     -0.0006701  0.0002164  -3.097 0.002120 **
Comp.11      0.0007891  0.0002193   3.598 0.000368 ***
Comp.14     -0.0004831  0.0002420  -1.996 0.046687 *
Comp.16     -0.0004282  0.0002754  -1.555 0.120973
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.002663 on 341 degrees of freedom
Multiple R-squared:  0.8933,    Adjusted R-squared:  0.8902
F-statistic: 285.5 on 10 and 341 DF,  p-value: < 2.2e-16
```

Model	**Data-Set**	**MAE**
Model 1	Training	.003
Model 1	Test	.003
Pca1	Training	.002
Pca1	Test	.002
Model 2	Training	.002
Model 2	Test	.002
Backwards Model	Test	.002

The model created utilizing the backwards variable selection produced what would be considered the best model as this model had the largest R-Squared and had a similar MAE to the other models.

Conclusions:

There are multiple ways to select the most efficient variables for a model. Two methods were explored in this assignment, principle component analysis and backwards automated variable selection. Neither selected is typically more accurate than the other as they both produce results that would need to be reviewed and evaluated before accepting.