College Data: Private and Public University Acceptance Rate

By: Jackson Stroup, Jack Michalowski, and Liam O'Connor

MA 362 10:00 AM

10/23/2020

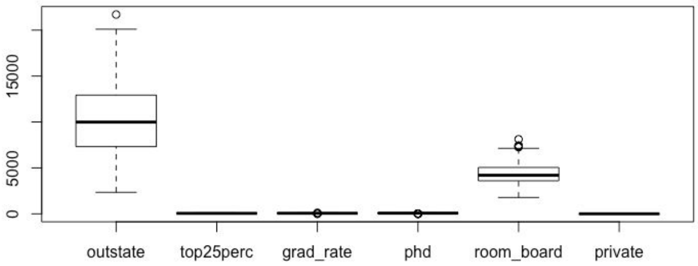
Table of Contents:

Dataset Description:

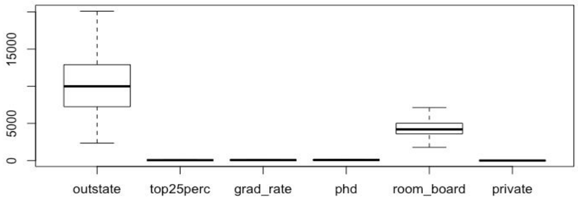
The “college data” dataset, created by Fares Sayah, is a dataset found on kaggle.com that looks at different admission statistics from 777 colleges and universities. The dataset includes variables such as the out of state tuition, the room and board fees, the graduation rate, the percentage of faculty that hold PHD’s, the percentage of students who were in the top 25% of their high school class, and whether it is a private or public school. For the linear regression model, Out of state tuition is the response variable, while room and board, graduation rate, PHD percentage, private vs public, and top 25 percentage are all explanatory variables.

Outlier Removal

First, let’s check the boxplots of each variable to see if there are outliers.



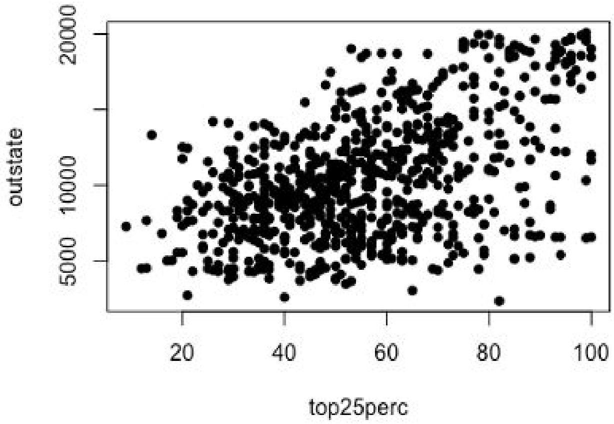
Looking at the boxplot, it is clear that outliers exist in the outstate, grad_rate, phd, and room_board variables. Upon further investigation, there is 1 outlier in the outstate variable, 4 in the grad_rate variable, 8 in the phd variable, and 7 in the room_board variable. There are 20 outliers total in the dataset. Let’s remove the outliers. After identifying the outliers in R, and removing the 20 data points, a new boxplot was created.



After creating the new boxplot, it is clear that no outliers exist in the dataset anymore. We can now continue with generating scatter plots for the model.

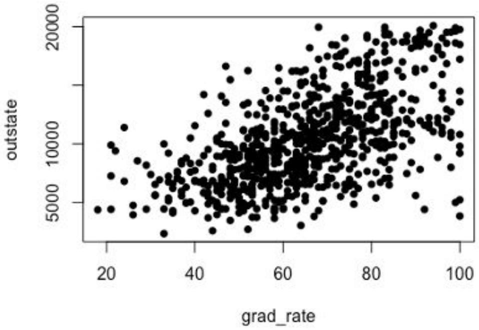
Scatter Plots:

Let’s now create scatter plots for each variable to take a look at the linear relation between each x-y pair. Let’s start with the top25 percent variable.



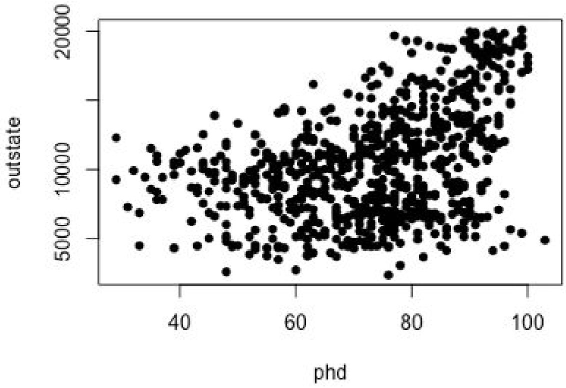
Looking at the scatter plot, there is an approximate linear and positive relationship between the 2 variables. Upon further analysis, we can see that the correlation between the 2 variables is equal to .49008. This shows that there is a moderate positive linear relationship between the top 25 percent and outstate variable.

Let’s look at the grad_rate variable next.



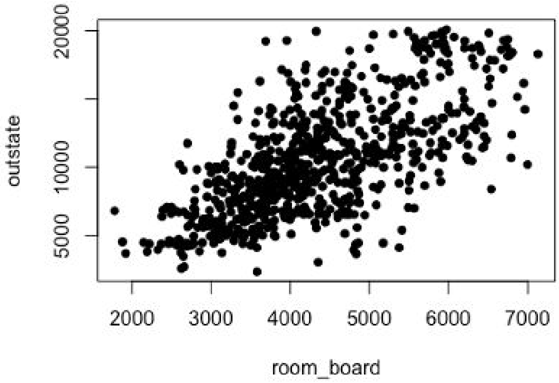
Looking at the output, there is a clear positive linear relationship between grad_rate and outstate tuition. We can confirm this by looking at the correlation between the 2 variables. The correlation is equal to .5853, which shows that there is a moderately strong positive linear relationship between the grad_rate and outstate variables.

Next let's look at the relationship between the phd and outstate variables.



Looking at the output, there is a somewhat positive linear relationship between the phd and outstate variables. Looking at the correlation, it is equal to .409936. This means that there is a somewhat weak, positive relationship between the phd and outstate variable.

Finally, let's look at the relationship between room_board and outstate tuition variables.



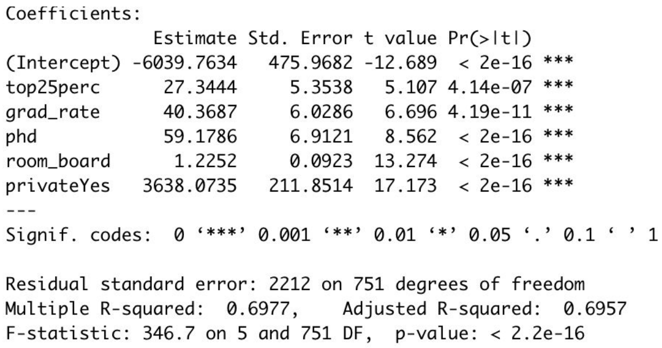
Looking at the output, it is clear that there is a positive linear relationship between room_board and outstate variables. We can confirm this by looking at the correlation, which is equal to .66147. This shows that there is a somewhat strong, positive relationship between room_board and outstate tuition.

After looking at each explanatory variable, it is clear that there is a positive linear relationship between each x-y pair. This provides evidence that a multiple regression model is appropriate for this dataset.

Multiple Linear Regression Model:

In order to create the multiple linear regression model, we must create a test and train dataset. After removing outliers, there are now 757 points in the dataset. For the train dataset, we will need to take 80% of the data, which is 606 points. We will then use the other 151 data points for the test dataset. Using the randomize feature in R, 606 random data points were selected, and we can now create the MLR model based off of the train dataset.

model1=lm(outstate~top25perc+grad_rate+phd+room_board+private)



```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -6039.7634   475.9682 -12.689  < 2e-16 ***
top25perc      27.3444     5.3538   5.107 4.14e-07 ***
grad_rate      40.3687     6.0286   6.696 4.19e-11 ***
phd            59.1786     6.9121   8.562  < 2e-16 ***
room_board      1.2252     0.0923  13.274  < 2e-16 ***
privateYes   3638.0735   211.8514  17.173  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2212 on 751 degrees of freedom
Multiple R-squared:  0.6977,	Adjusted R-squared:  0.6957 
F-statistic: 346.7 on 5 and 751 DF,  p-value: < 2.2e-16
```

Estimated regression model:

outstatehat=-6,039.7634+27.3444top25perc+40.3687gradrate+59.1786phd+1.2252roomboard+3638.0735privateyes

Interpret the Parameters :

beta0: Is not useful in the model, since you cannot have a negative tuition

beta1=27.344: For each 1 percent increase in the percentage of students from the top 25% of their high school graduating class, the out of state tuition will increase by $27.34, keeping all other variables fixed.

beta2=40.37: For each 1 percent increase in the percentage of students that graduate from the school, the out of state tuition will increase by $40.37, keeping all other variables constant

beta3=59.1786: For each 1 percent increase in the percentage of instructors that obtain their PHD, the out of state tuition will increase by approximately $59.18, keeping all other variables fixed.

beta4=1.23: For each 1 dollar increase in the room and board fees, the out of state tuition will increase by approximately $1.23, keeping all other variables fixed.

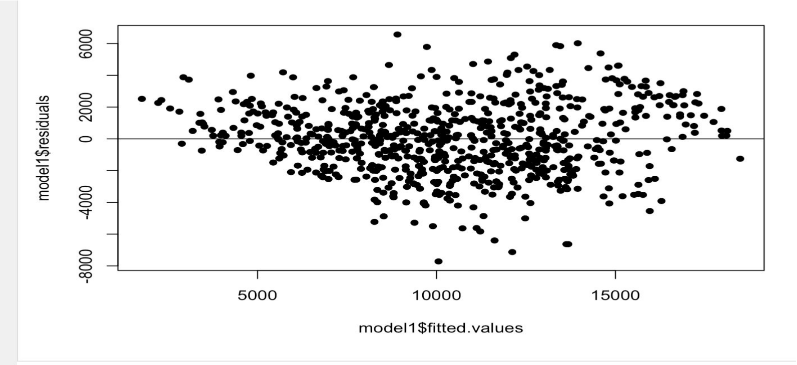
beta5= 3638.07: If a school identifies as private, the out of state tuition will increase by $3638.07, keeping all other variables fixed.

Looking at the multiple coefficient of determination, we can see that r-squared is equal to .6977. This means that 69.77% of the observed sample variation in the response variable is explained by the regression model. The adjusted r squared is .6957 which means that 69.57% of the observed tuition can be explained by the model which includes top25perc, gradrate, phd, roomboard and privateyes.

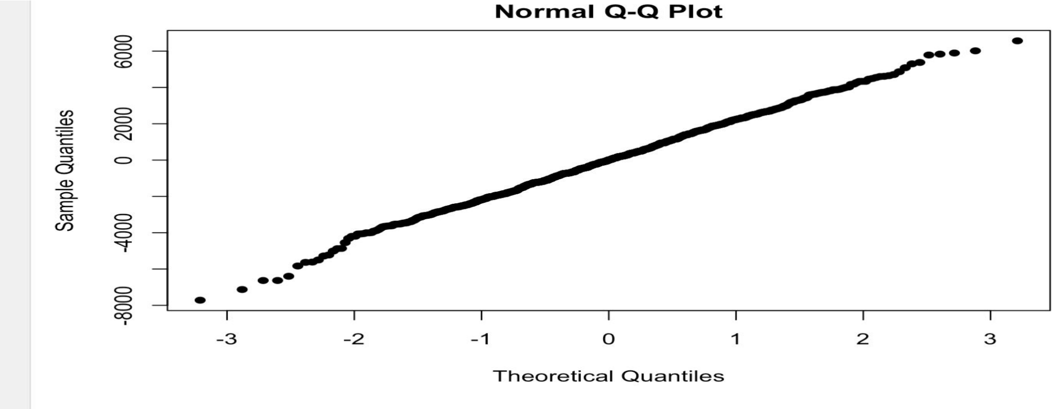
Looking at the standard error, we see that standard error is equal to 2,212. This means that the average difference between the predicted tuition and the actual tuition is equal to $2,212.

Residual Analysis:

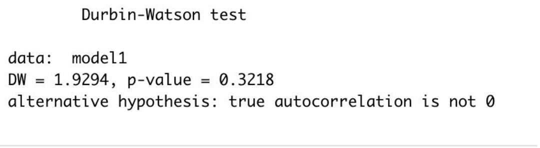
First, let’s check the regression assumptions. Let's take a look at the residual plot below.



Assumption 1 states that the mean of the probability distribution of errors is 0. Based on the model above, the assumption of E(e)=0 is satisfied because the points are symmetric around x-axis, random, and no pattern. The assumption of V(e)= constant for all the value combinations of the predictor variables is also satisfied because points form an approximately consistant horizontal band on the residual plot seen above. Next let’s look at the normal probability plot for the next assumption.



Based on the above normal probability plot of residuals, we can say the errors are normally distributed becasue the plot is positive and linear.

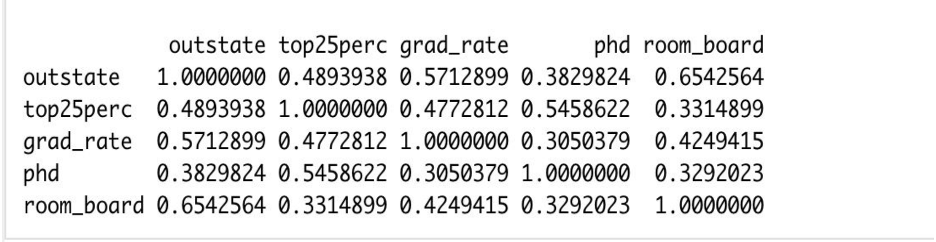


```
        Durbin-Watson test

data:  model1
DW = 1.9294, p-value = 0.3218
alternative hypothesis: true autocorrelation is not 0
```

For our last assumption, we turn to the Durbin-Watson Test. As shown by the R output, we have a DW test statistic of approximately 2. This result shows that there is no correlation in the errors, and that they are independent. Thus, assumption 4 is satisfied. In conclusion, all assumptions are satisfied.

Checking for Multicollinearity and Interpreting Outcomes:

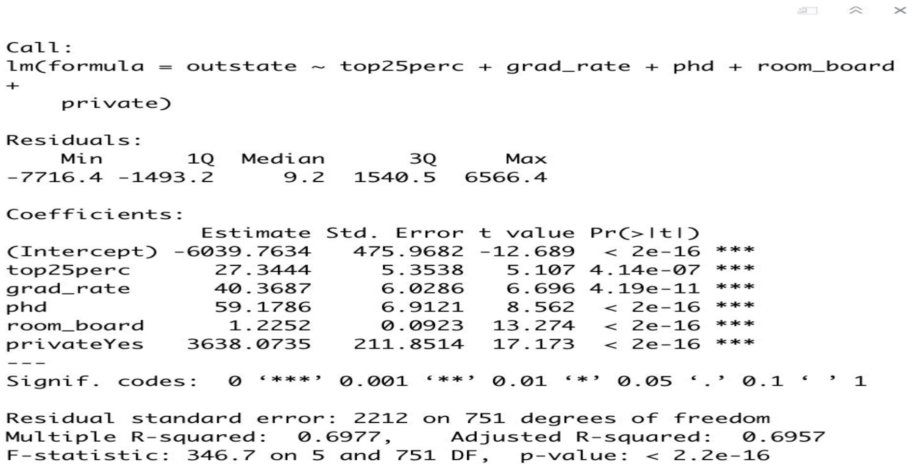
1) **Significant correlation between pairs of independent variables in the model:**



```
           outstate top25perc grad_rate       phd room_board
outstate  1.0000000 0.4893938 0.5712899 0.3829824  0.6542564
top25perc 0.4893938 1.0000000 0.4772812 0.5458622  0.3314899
grad_rate 0.5712899 0.4772812 1.0000000 0.3050379  0.4249415
phd       0.3829824 0.5458622 0.3050379 1.0000000  0.3292023
room_board 0.6542564 0.3314899 0.4249415 0.3292023  1.0000000
```

The above correlation coefficient matrix from R shows that room_board and outstate (r=.6542564) have a strong linear correlation. In addition, grad_rate and outstate(r=.5712899) also have a strong correlation. As a result, it can be concluded that there is multicollinearity in the model.

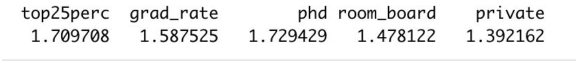
2) **Opposite signs in estimated parameters:**



```
Call:
lm(formula = outstate ~ top25perc + grad_rate + phd + room_board
+
    private)

Residuals:
    Min      1Q  Median      3Q     Max 
-7716.4 -1493.2     9.2  1540.5  6566.4 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -6039.7634   475.9682 -12.689  < 2e-16 ***
top25perc      27.3444     5.3538   5.107 4.14e-07 ***
grad_rate      40.3687     6.0286   6.696 4.19e-11 ***
phd            59.1786     6.9121   8.562  < 2e-16 ***
room_board      1.2252     0.0923  13.274  < 2e-16 ***
privateYes   3638.0735   211.8514  17.173  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2212 on 751 degrees of freedom
Multiple R-squared:  0.6977,    Adjusted R-squared:  0.6957 
F-statistic: 346.7 on 5 and 751 DF,  p-value: < 2.2e-16
```

All signs are positive except intercept, which is not useful in the model. This indicates multicollinearity.

3) **VIF > 10**



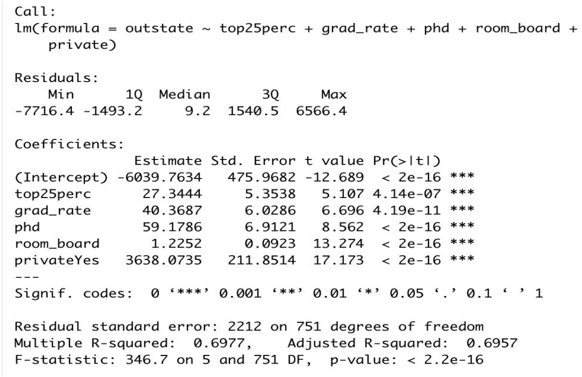
```
 top25perc  grad_rate        phd room_board    private 
  1.709708   1.587525   1.729429   1.478122   1.392162 
```

Unfortunately, none of the estimated parameters have a VIF>10, so we cannot say there's multicollinearity.

4) **Non significant t-tests for ALL of the individual parameters:**
If we refer back to the R output in part 2 above, all of the pvalues are <0.05 which means all of the parameters are significant, thus no multicollinearity.

In conclusion, our model does in fact show multicollinearity because of the correlation matrix showing strong positive correlation, as well as the opposite sign (intercept) in parameters.

Testing the Utility of the Model:



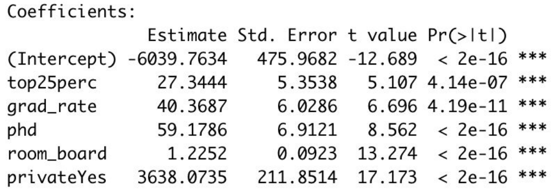
```
Call:
lm(formula = outstate ~ top25perc + grad_rate + phd + room_board +
    private)

Residuals:
    Min      1Q  Median      3Q     Max
-7716.4 -1493.2     9.2  1540.5  6566.4

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -6039.7634   475.9682 -12.689  < 2e-16 ***
top25perc      27.3444     5.3538   5.107 4.14e-07 ***
grad_rate      40.3687     6.0286   6.696 4.19e-11 ***
phd            59.1786     6.9121   8.562  < 2e-16 ***
room_board      1.2252     0.0923  13.274  < 2e-16 ***
privateYes   3638.0735   211.8514  17.173  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2212 on 751 degrees of freedom
Multiple R-squared:  0.6977,    Adjusted R-squared:  0.6957
F-statistic: 346.7 on 5 and 751 DF,  p-value: < 2.2e-16
```

In order to test the utility of the model, we need to use the Global F Test in order to test the usefulness of the model.

H0: b1=b2=b3=b4=b5=0   All predictor variables are not useful in the model.
Ha: bi ≠ 0 i=1,2,3,4,5   At least one predictor variable is useful in the model.

The pvalue of the model is 2.2x10^-16. This pvalue is < α which is 0.05. Thus, we reject H0….data support Ha. At the 5% significance level, data provide evidence to say that at least one of the predictor variables is useful in the model. Therefore, the model passes the Global F test.

Predictor Variable Usefulness:



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -6039.7634   475.9682 -12.689  < 2e-16 ***
top25perc      27.3444     5.3538   5.107 4.14e-07 ***
grad_rate      40.3687     6.0286   6.696 4.19e-11 ***
phd            59.1786     6.9121   8.562  < 2e-16 ***
room_board      1.2252     0.0923  13.274  < 2e-16 ***
privateYes   3638.0735   211.8514  17.173  < 2e-16 ***
```

P values for the estimated parameters

We must check the individual t tests/pvalues and compare them to our alpha 0.05.
top25perc: pval(4.14x10^-7) < 0.05 : reject H0, useful in the model.
grad_rate: pval4.19x10^-11()<0.05 : reject H0, useful in the model.
Phd: pval(2x10^-16) < 0.05 : reject H0, useful in the model.
room_board: pval(2x10^-16) : reject H0, useful in the model.
privateYes: pval(2x10^-16) : reject H0, useful in the model.

Overall, because all of the estimated parameters have a pvalue < 0.05, all of the variables are useful in the model. When we tested the utility of the model with the Global test, we concluded that at least one of the predictor variables is useful. The individual t tests confirmed this, so we can say that the result of the Global F Test is consistent with the individual tests.

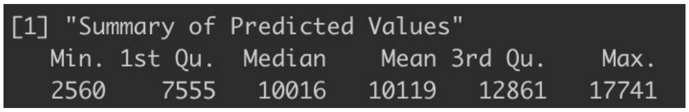
Removing any not useful predictor variables:

Based on the previous section, we concluded that all of the predictor variables are useful. Therefore, we do not need to remove any variables from our original model1.

Comparing Models:

Since all of the predictor variables in model1 are useful, there was no need to remove any. With this being said, there is no need to create a model2 since it would essentially be the same as model1 for the reason that was just stated. Therefore, model1 is the better model because there is no need for a second model. We will now move forward with using the first model.

Predicting the Test Dataset:

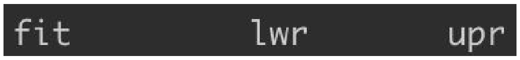


```
[1] "Summary of Predicted Values"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   2560    7555   10016   10119   12861   17741 
```

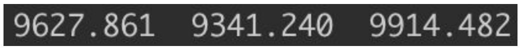
Shown above are the predicted values for the Test dataset.

Prediction & Confidence Intervals for the Predictor Variable Values in the Test Dataset:

Confidence Interval:



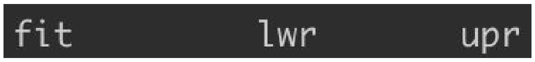
```
fit            lwr        upr
```



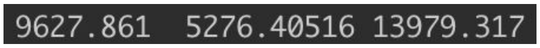
```
9627.861  9341.240  9914.482
```

We are 95% confident that the average out of state tuition cost is between $9,342.24 and $9914.48.

Prediction Interval:



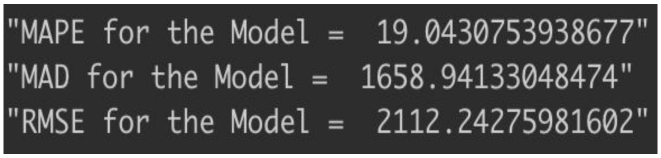
```
fit            lwr        upr
```



```
9627.861  5276.40516 13979.317
```

We are 95% confident that the predicted average out of state tuition is between $5,276.41 and 13,979.32.

Prediction Accuracy with MAPE, MAD, and RMSE:



```
"MAPE for the Model =  19.0430753938677"
"MAD for the Model =  1658.94133048474"
"RMSE for the Model =  2112.24275981602"
```

Shown above are the values for prediction accuracy using MAPE, MAD, and RMSE.