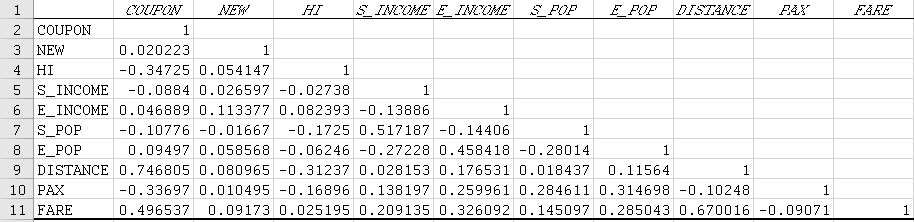
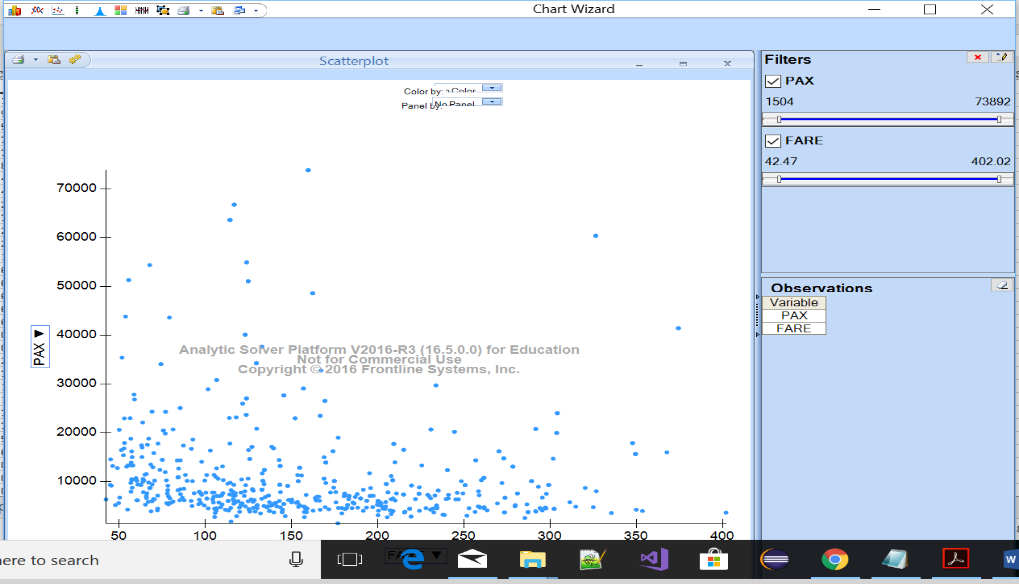
1. Below are the numerical variables in the given dataset.
2. COUPON
3. NEW
4. HI
5. S\_INCOME
6. E\_INCOME
7. S\_POP
8. E\_POP
9. DISTANCE
10. PAX

The Correlation matrix between the Numerical predictors and FARE is given below:

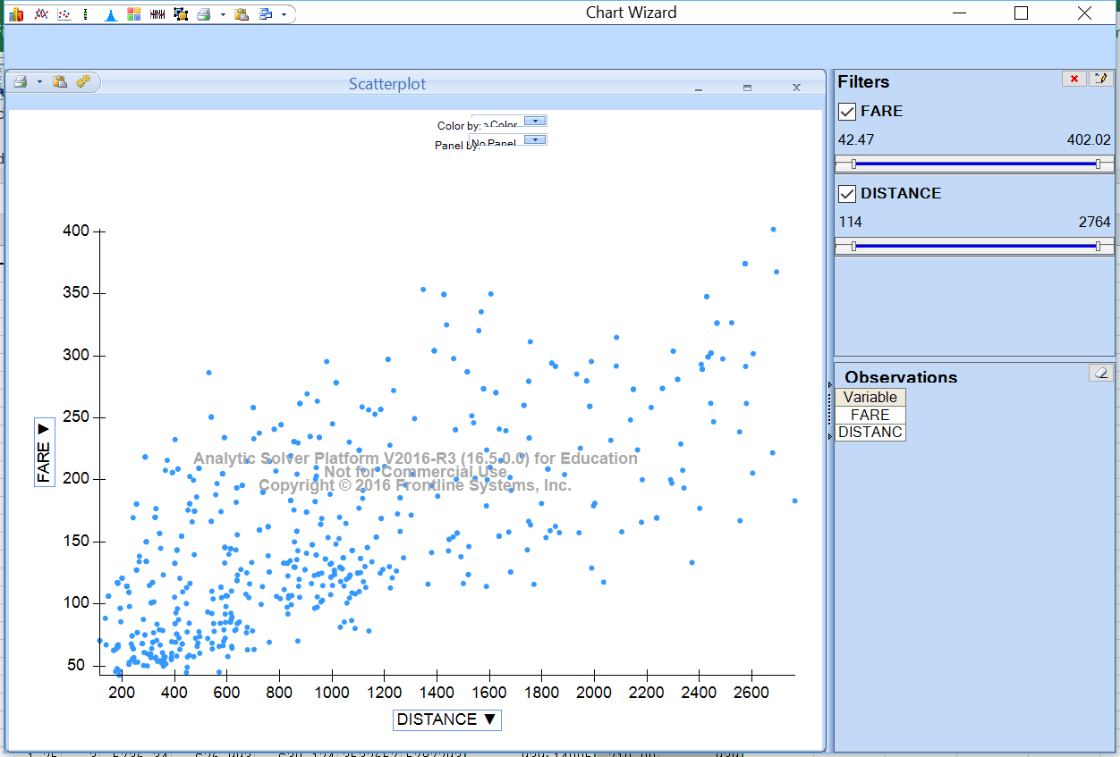


The Scatter plots between different numerical predictors and FARE is given below:

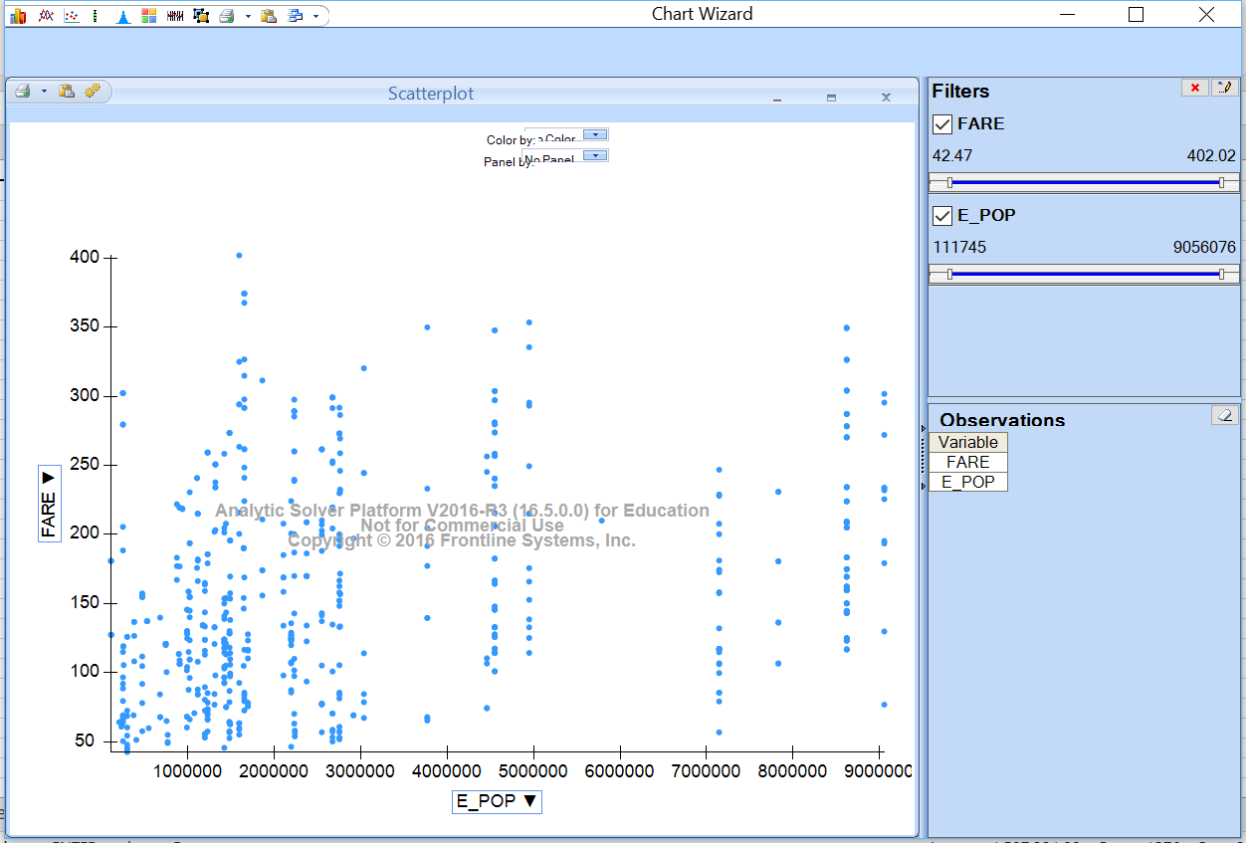
1. PAX vs FARE



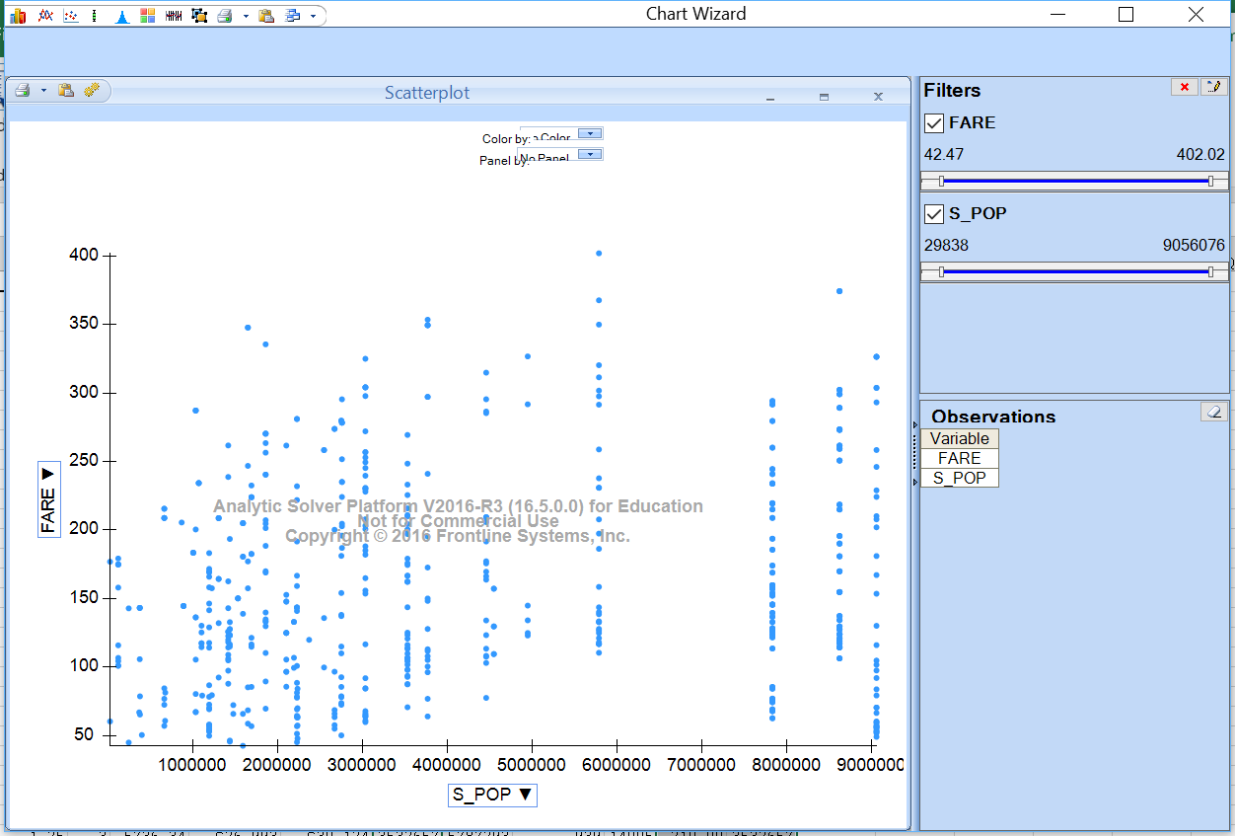
1. DISTANCE vs FARE

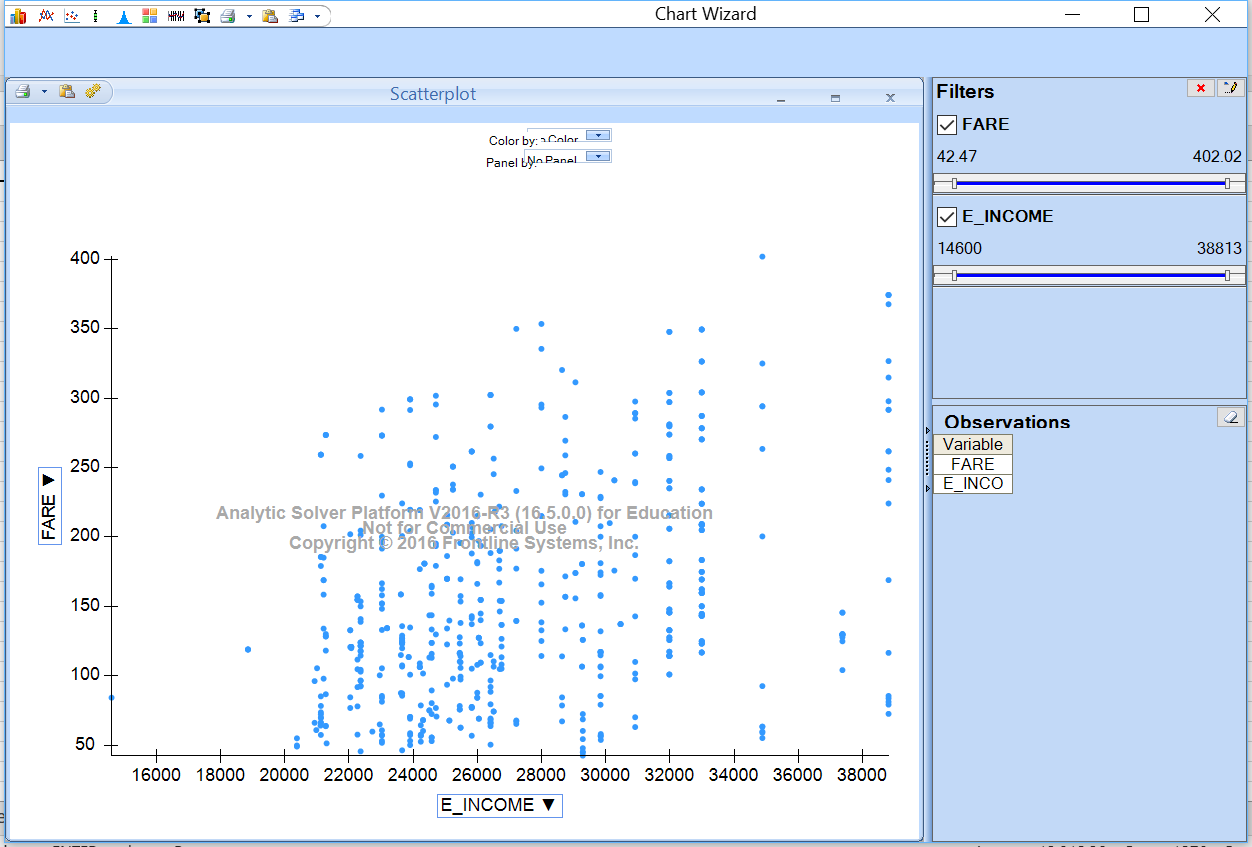
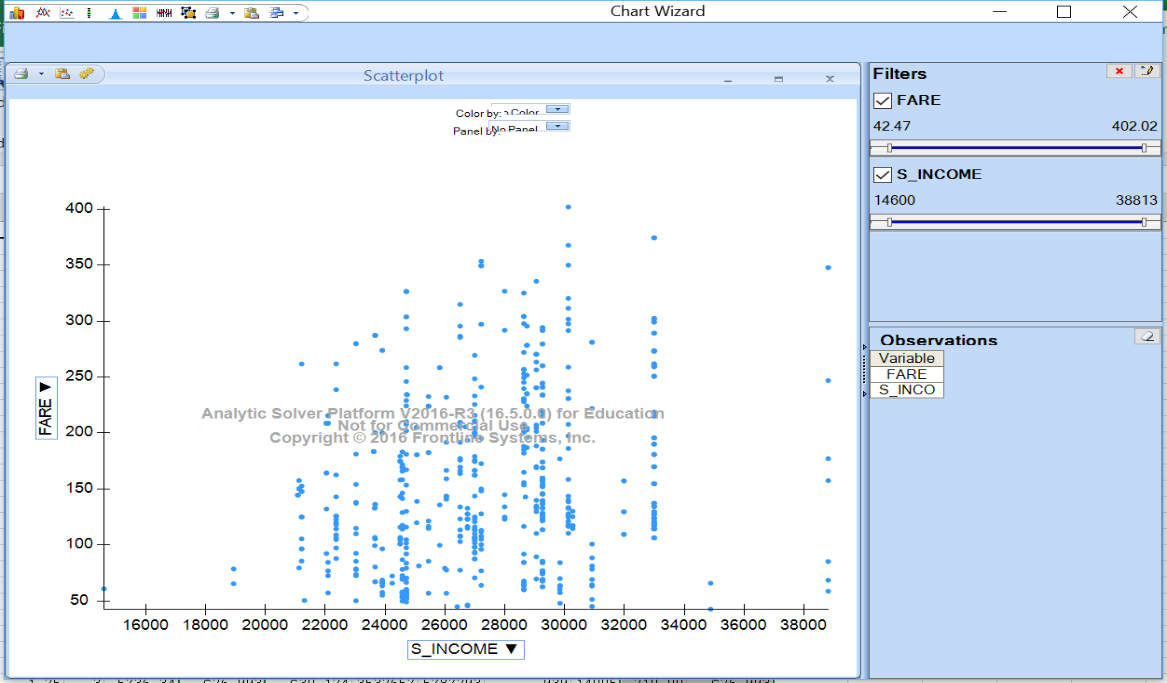


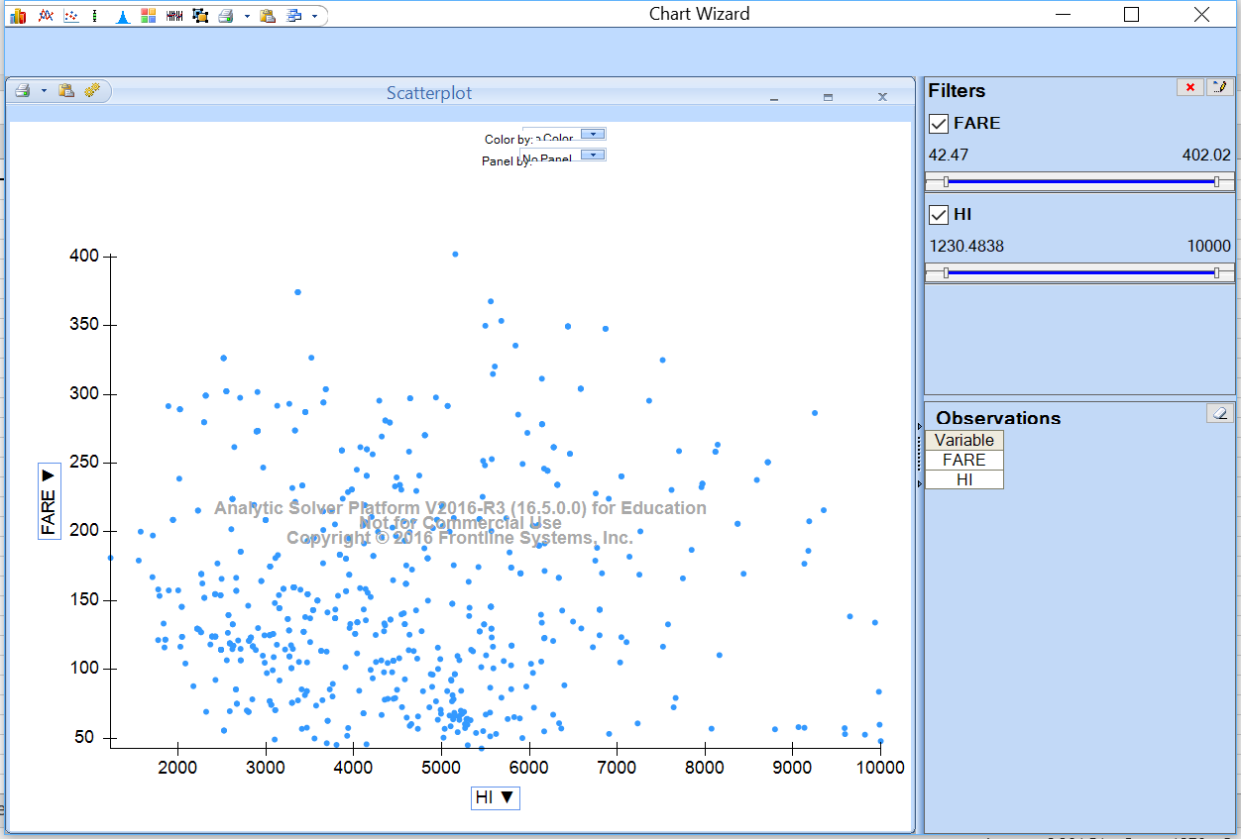
1. E\_POP vs FARE



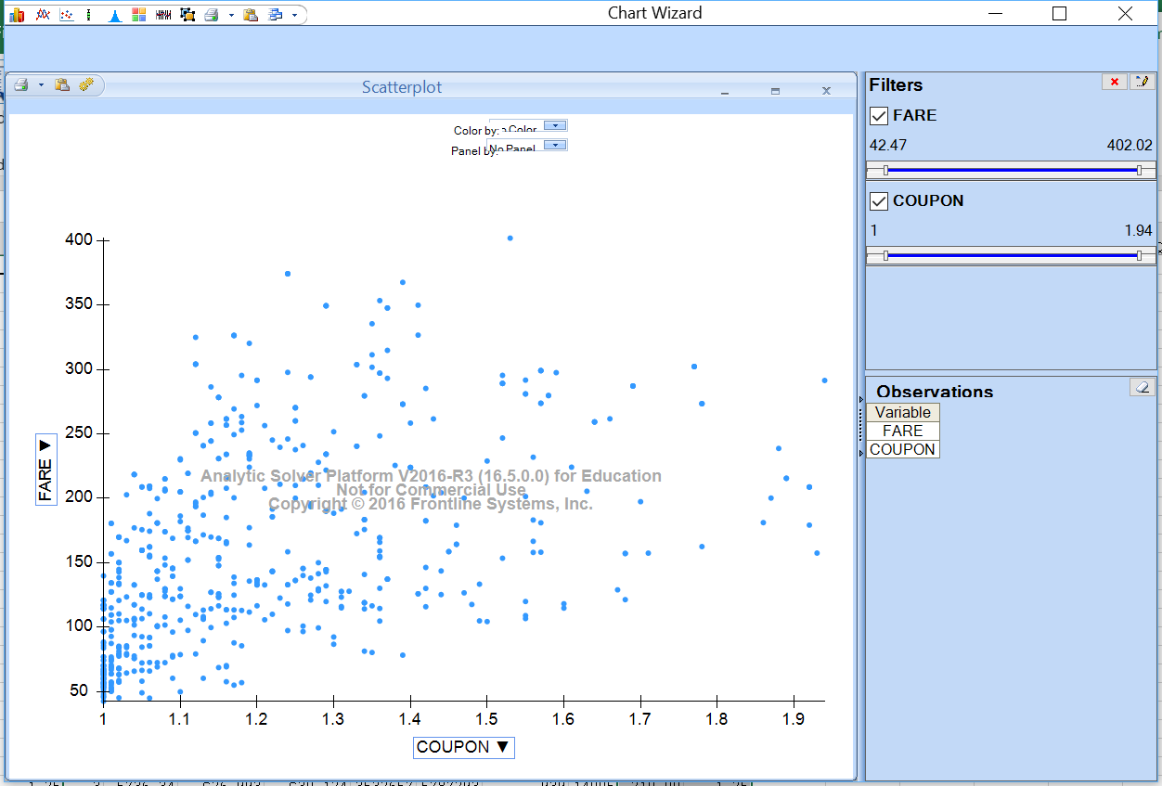
1. S\_POP vs FARE



1. E\_INCOME vs FARE
2. S\_INCOME vs FARE
3. HI vs FARE



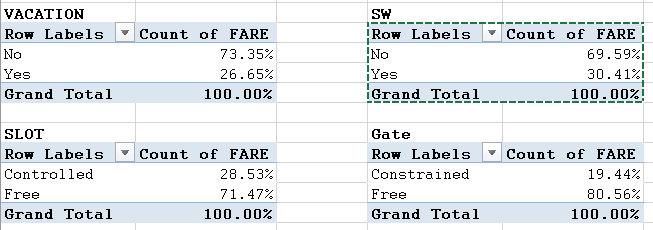
1. COUPON vs FARE



By looking at the correlation matrix and the scatter plots, we can conclude that **DISTANCE** variable is highly correlated to FARE**.** Hence, **DISTANCE** can be used as a numerical predictor.

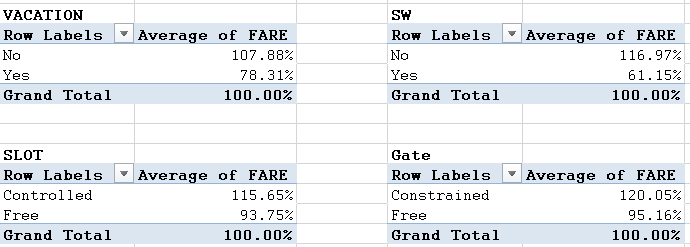
1. Below are the categorical variables in the given dataset.
2. VACATION
3. SW
4. SLOT
5. GATE

The percentage of the categories of each categorical variable is below:



We can use pivot table for finding the best categorical variable that can be used for predicting. This table

combine the information of multiple variables and compute the sizes of different categories of the variables. Hence, the best categorical variable is the one which has the maximum difference between the categories.



According to the tables, **SW** variable has the maximum difference between the categories and hence SW can be the best predictor.

1. The Category variables taken into consideration are

VACATION

SW

SLOT

GATE

This can be done by the following procedure

Go to XLMiner Platform 🡪 Transfor 🡪 Transform Categorical Data 🡪 Create Dummies

Automatically, all the categorical variables will be selected and if we press enter, a new worksheet will be created with the new attributes for the categorical variables.

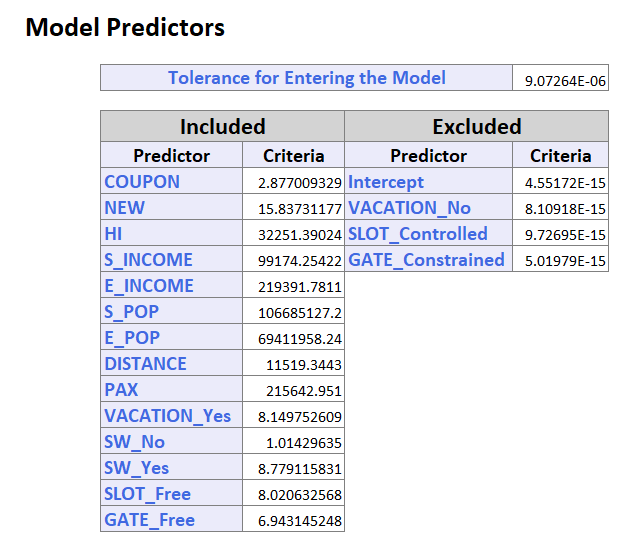
The partition of this data can be done by the following procedure

XLMiner Platform 🡪 Partition 🡪 Standard Partition 🡪 select all the dataset to be partitioned 🡪 Pick up rows randomly 🡪 Set the percentage as 60% for training and 40% for Test/Validation 🡪 ok

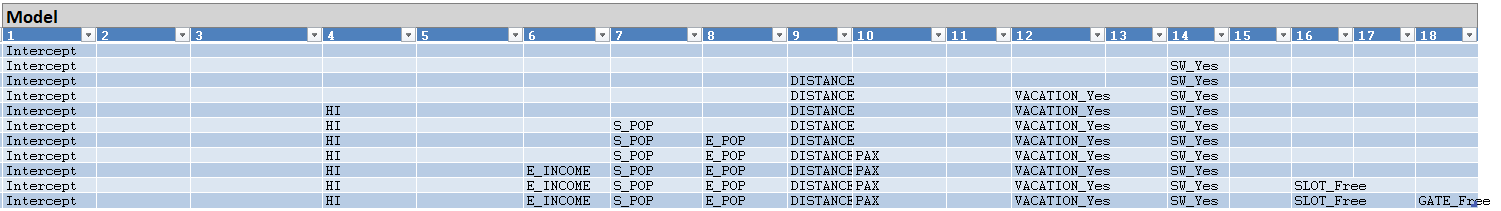
1. The data will be partitioned into training and validation because, the training data will be used to train the model, make understand the model for a set of inputs what exactly be the output. After the model is trained on the trained dataset, in order to test the performance or accuracy of the trained model, it is tested on the validation dataset. It predicts the output values and the deviate

on can be verified. This determines the accuracy and performance of the model.

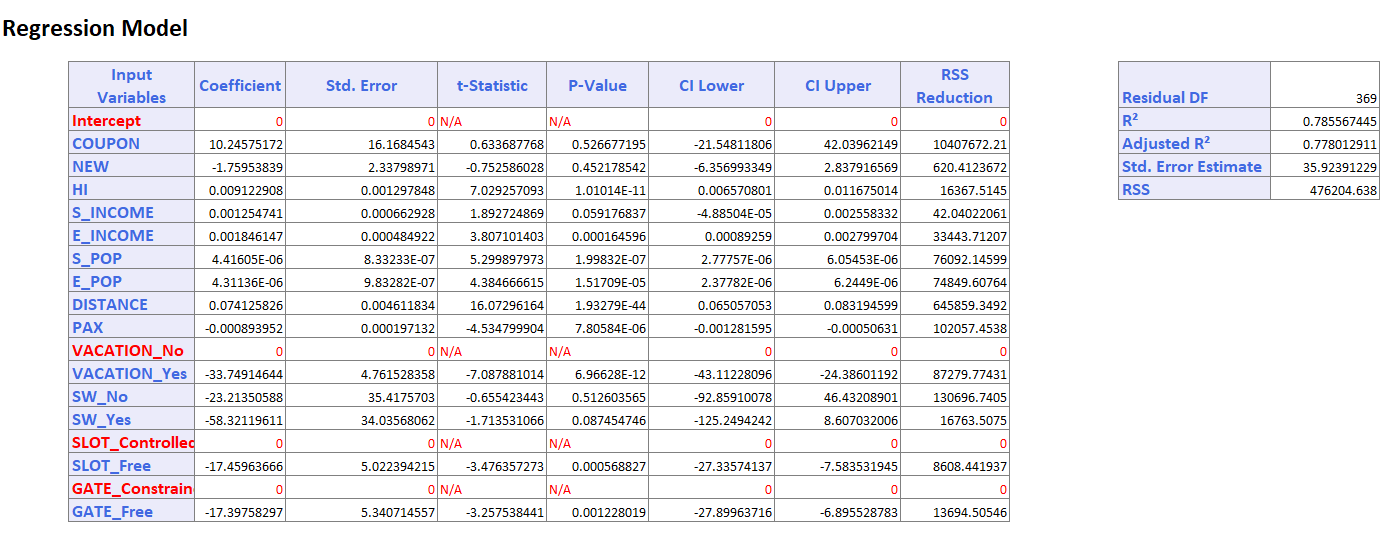
1. When we apply step wise regression on the above dataset, we get the following predictors excluded.



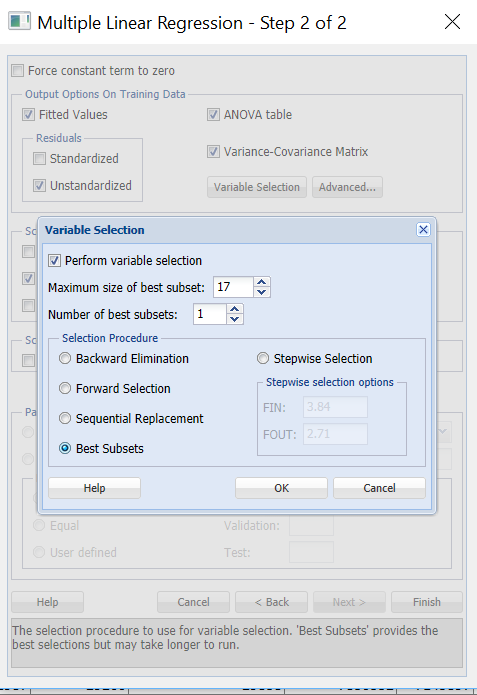
**Model selected:**



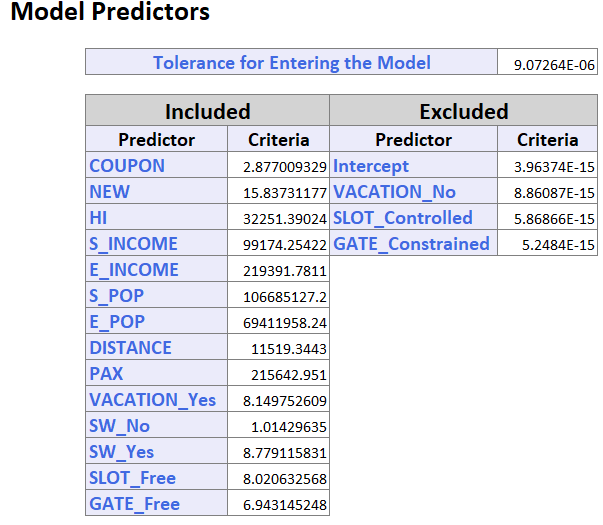
The Regression Model created is as follows:



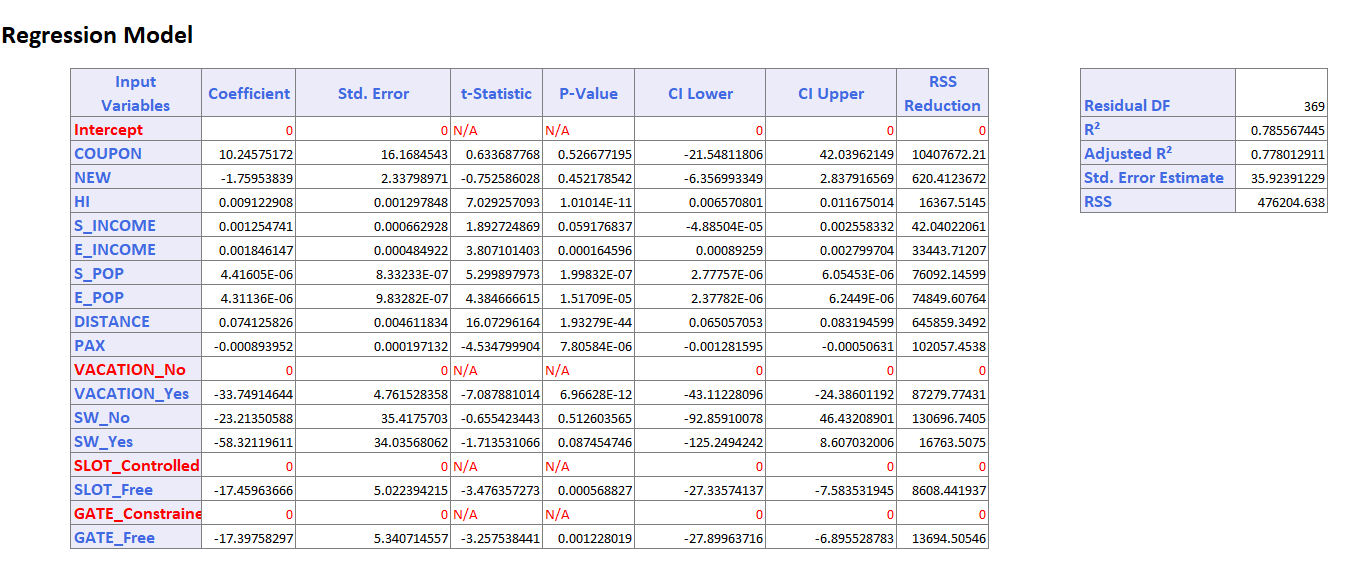
**iv.** After applying the Exhaustive search on the same dataset the following are the predictors used.

Below is the selection for the Exhaustive search model.

Below are the predictors used in this model.



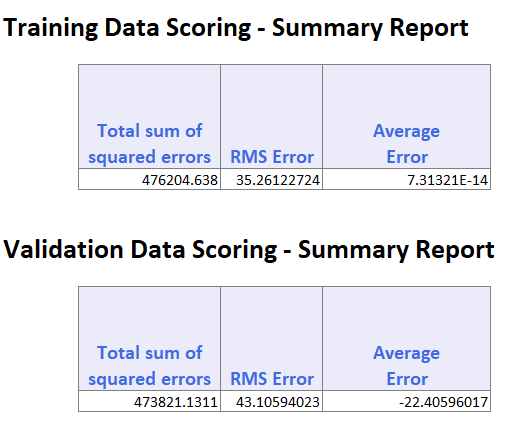
The model is as follows:

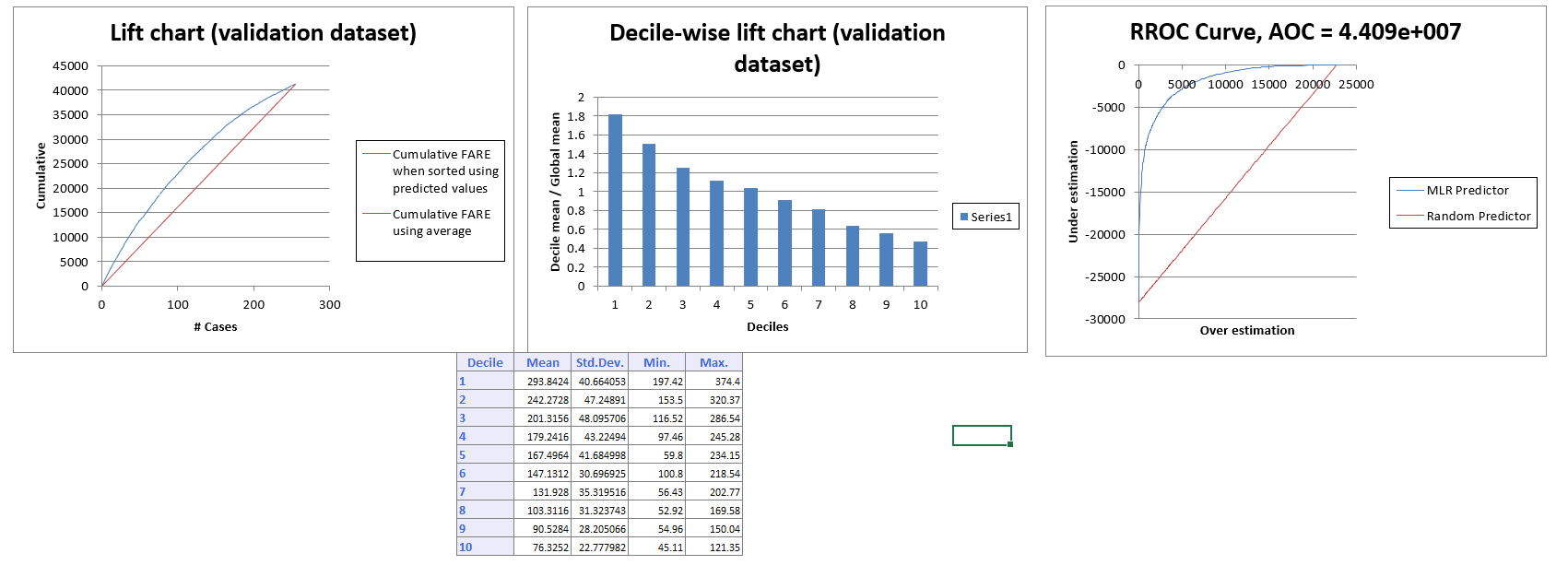


There is no change in the number of predictors when we compare with the models. There are same predictors which were taken into consideration in both the model.

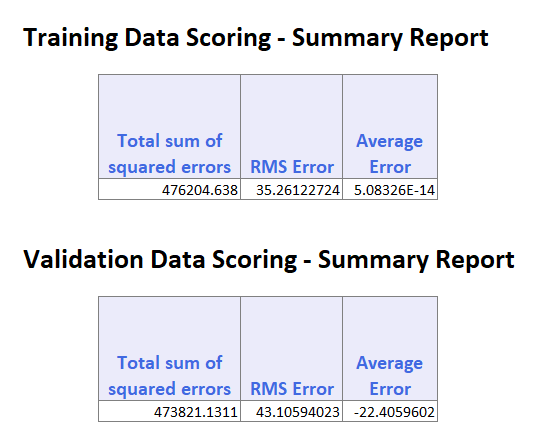
**v.** The Predictive accuracy of the both the models are mentioned below:

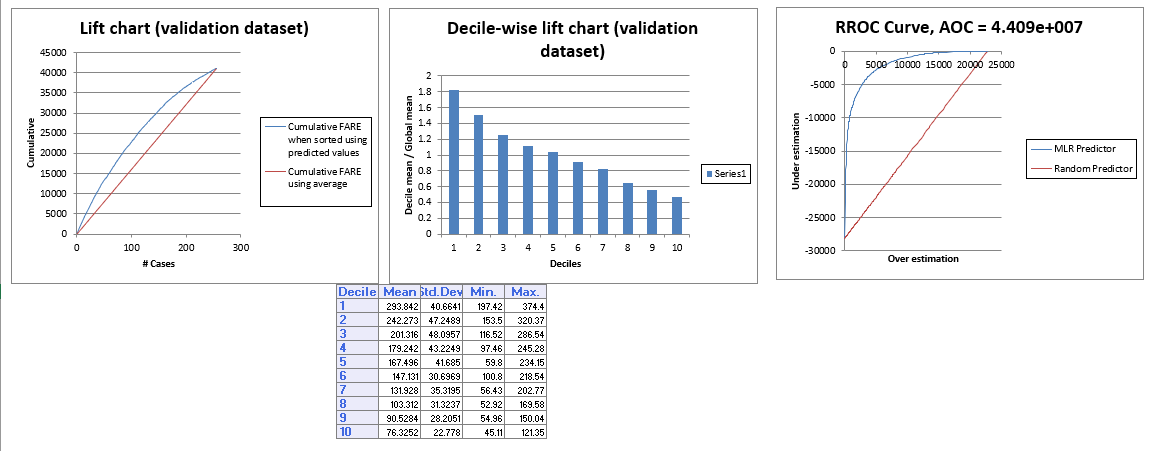
**Model (iii):** The RMSE, Average error and lift chart of the Validation data is below





**Model (iv):** The RMSE, Average Error and Lift chart of the validation data for this model are below.





**vi.**  The formula to be used in calculating the average fare is as follows:

Y = http://www.stat.yale.edu/Courses/1997-98/101/beta.gif0 + http://www.stat.yale.edu/Courses/1997-98/101/beta.gif1*x*1 + http://www.stat.yale.edu/Courses/1997-98/101/beta.gif2*x*2 + ... + http://www.stat.yale.edu/Courses/1997-98/101/beta.gifp*x*p

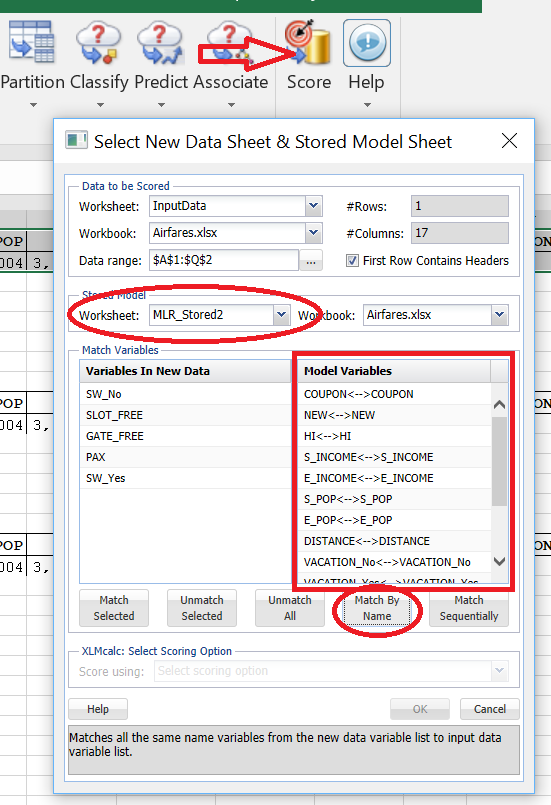
This can be calculated using the Regression model that is generated and all the beta values are the coefficients in the generated regression table.

There is also another procedure to calculate the predicted Fare with given input values.

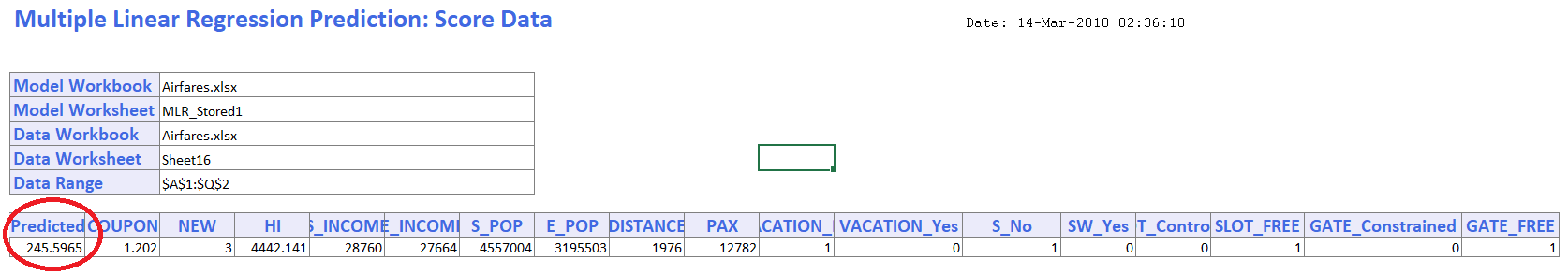
Please find the below procedure in predicting the output value from the input variables.



* Initially, construct a table with all given input values as shown below.
* Next, select this table and click XLMiner Platform 🡪 Score. Here we select the model stored worksheet (named as MLR\_Stored).
* Match all the variables by name and then click ok. A new worksheet will be created with the predicted Fare value with the given inputs.



* The above procedure will use the same formula of predicting the output value. The output worksheet is as follows

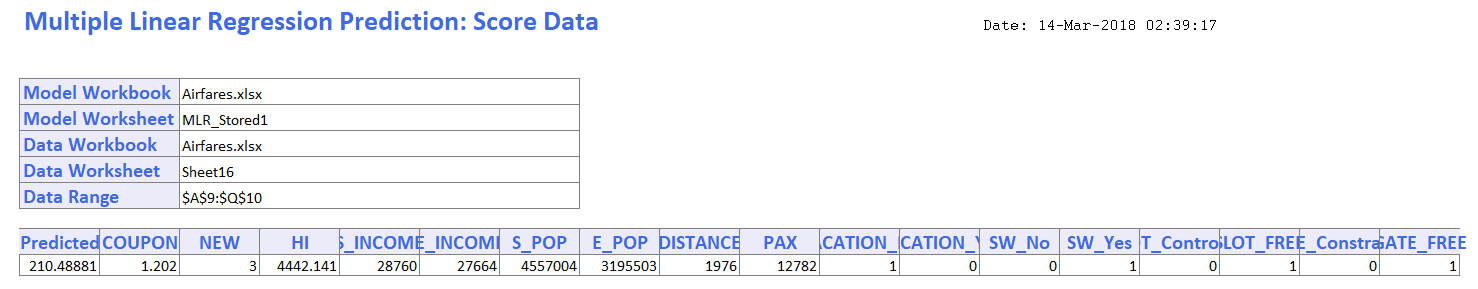


As we can see in the above screenshot, the predicted Fare value with the provided inputs is **245.5965$**.

**vii.** Similar to the procedure we followed in the previous bit, we need to create the table with the input values, but this time SW value should be **Yes.**



The output generated will be as follows:



As we can see in the above predicted value of the Fare is **210.48881 $**.

Reduction in the average fare after covering the route will be

245.5965 - 210.48881 = **35.1$**

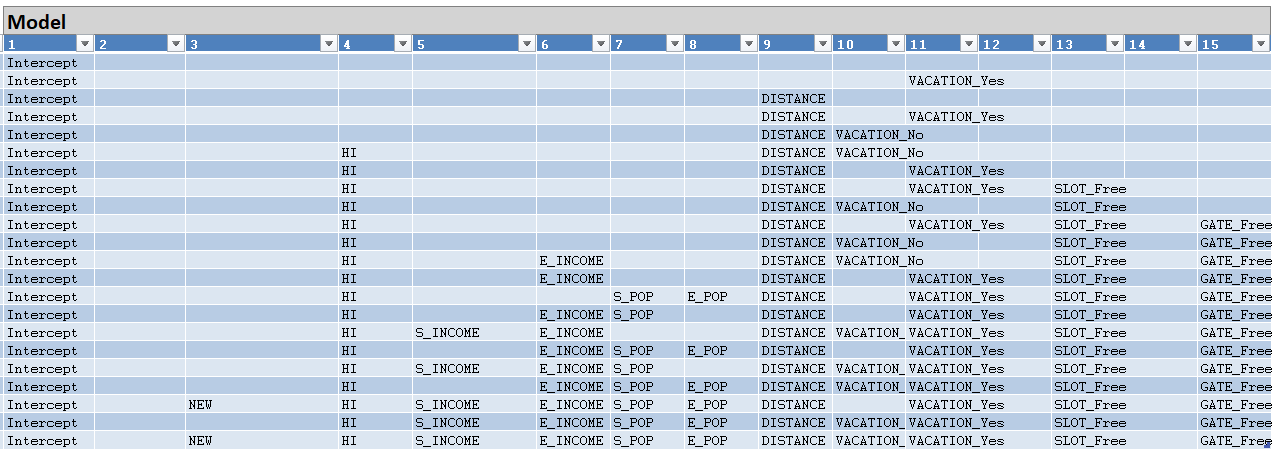
**viii.** The factors are known about the new airport are: the distance travelled, demographics of the city where the new airport is located, and whether this city is a vacation destination. Hence, we can consider all the variables into consideration as predictors except two variables.

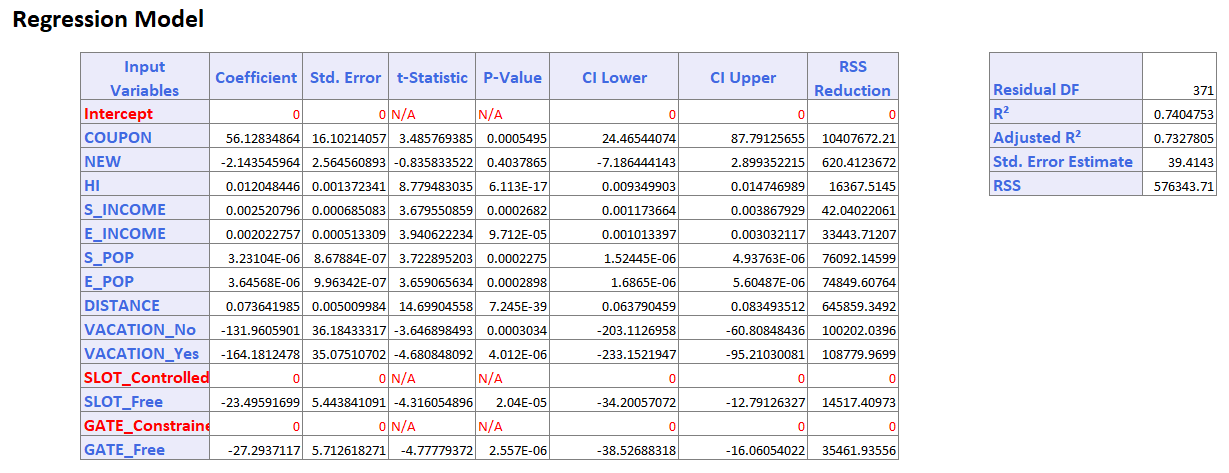
COUPON can be estimated as it depends on the number of stops in the new route. NEW can be determined i.e. the number of new careers in that route. VACATION can be estimated depending upon the destination place. HI can be estimated as it depends upon the market concentration and we are already given with the geographic data. S\_INCOME, E\_INCOME, S\_POP and E\_POP can be estimated as the starting and ending cities are known and their attributes can be estimated. Similarly, SLOT, GATE and DISTANCE can also be estimated.

The factors unknown is the number of passengers who will travel this route (PAX) and the major unknown factor is whether Southwest or another discount airline will travel on these new routes. Southwest's strategy.

**ix.** Now, we need to create a model with the attributes that can be estimated for the new route.

Below is the screen shot for the new model. This is using the exhaustive search.

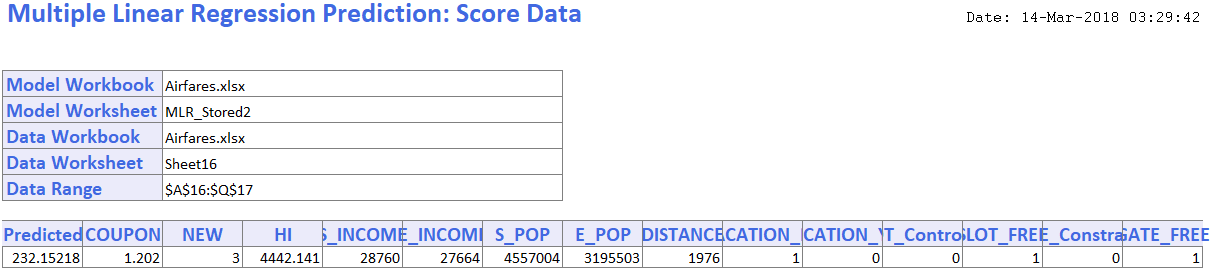


 Regression Model is as below:

**x.** Given, different input values and we need to determine the average output Fare. For this, we follow the same procedure which we used in (vi) to determine the output variable. Below is the table taken into consideration.

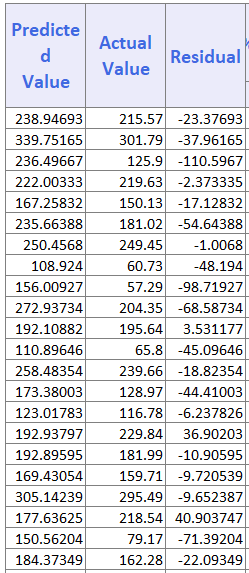
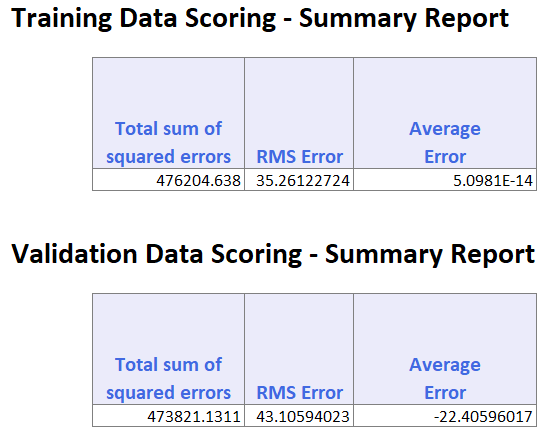


The output generated is as follows:

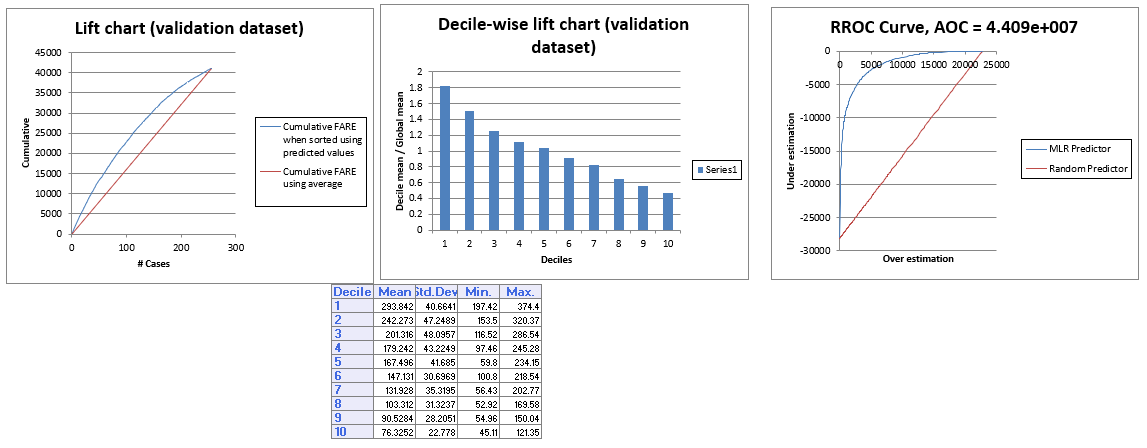


The predicted average fare is **232.15218 $.**

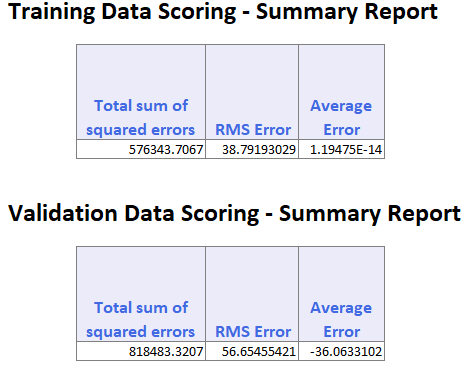
**xi.** Model accuracy of **model (iv)** is as follows:

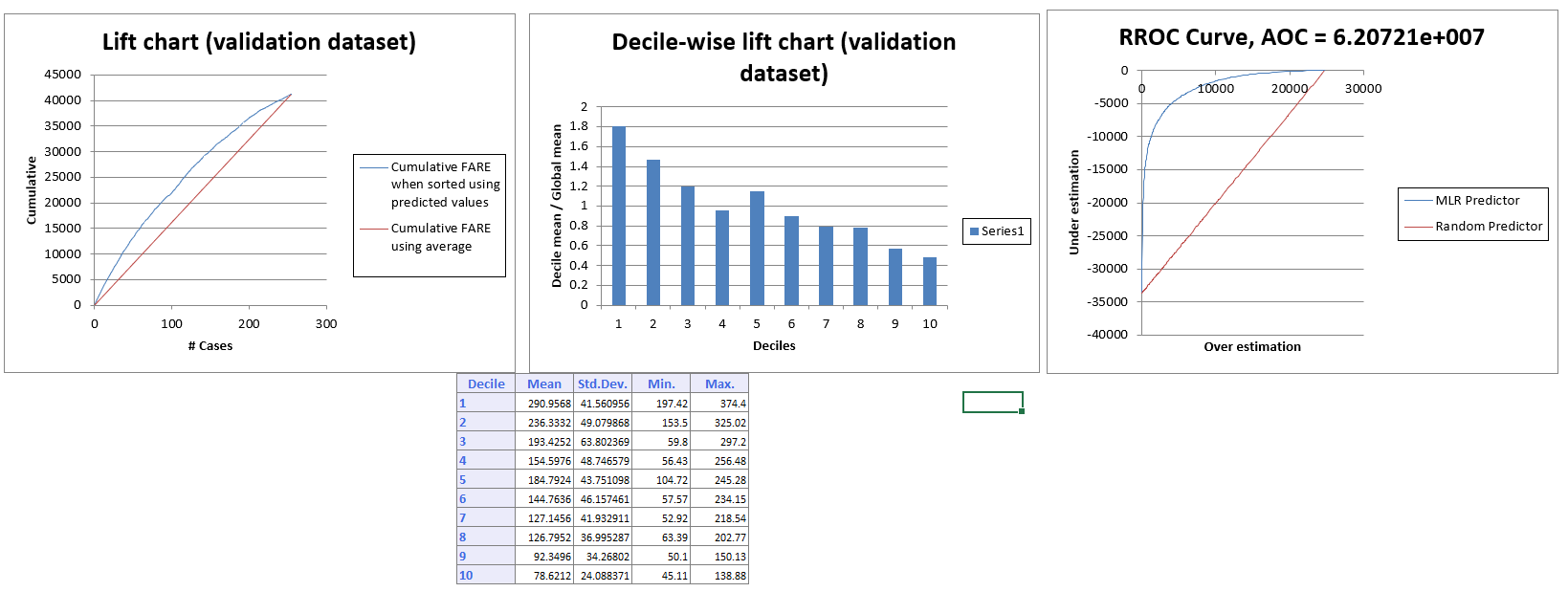
 

The above screenshot denotes the predicted vales on the validation data.

The lift charts of this model are as follows:

Model and accuracy of the new model is as follows:



The Lift chart of the new model on the validation data is as follows:

By considering the average error and the RSME values, we can understand that the new model is less accurate than the model (iv). Also, by observing the lift chart, we could observe that the lift in the new model is slightly less compared to the lift in model (iv) which concludes that model (iv) is more accurate the newly predicted model.

**Find the excel attachment which includes all the detailed reports and outputs of the above discussed questions.**

