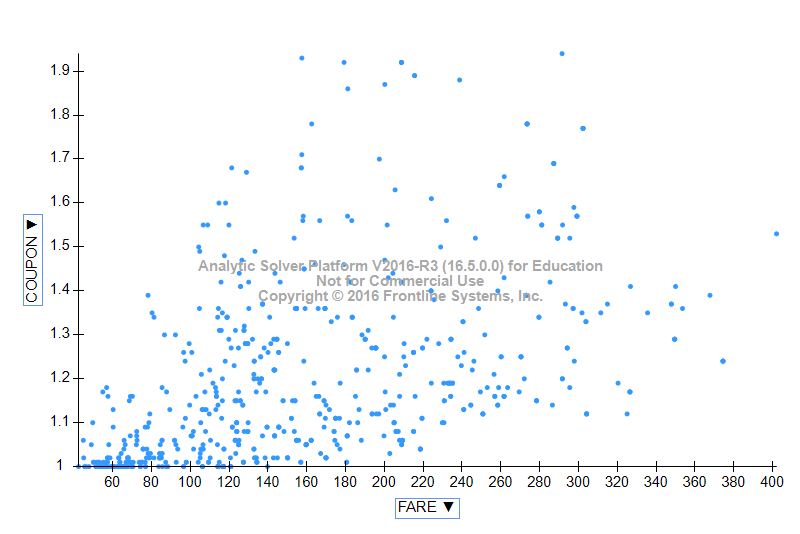
*Big Data Homework 4*

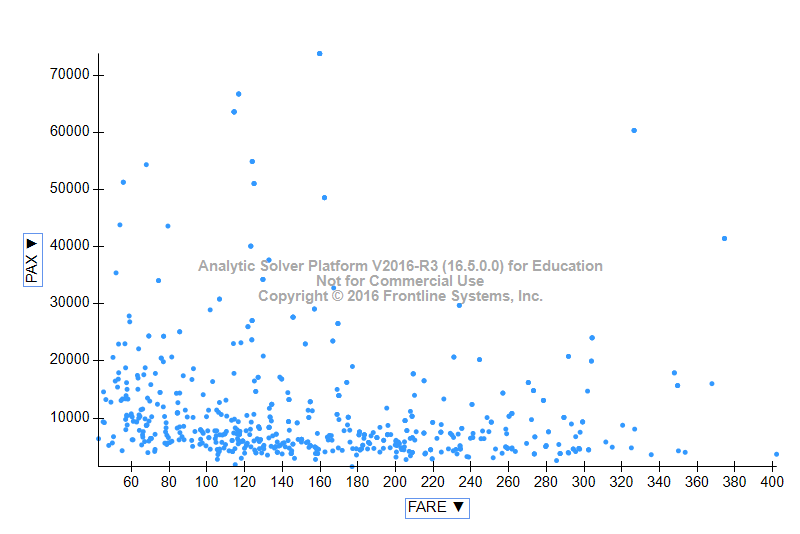
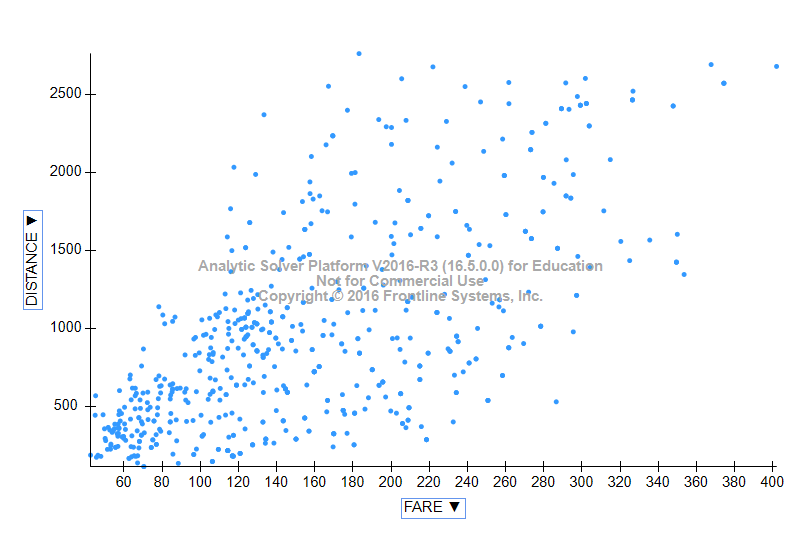
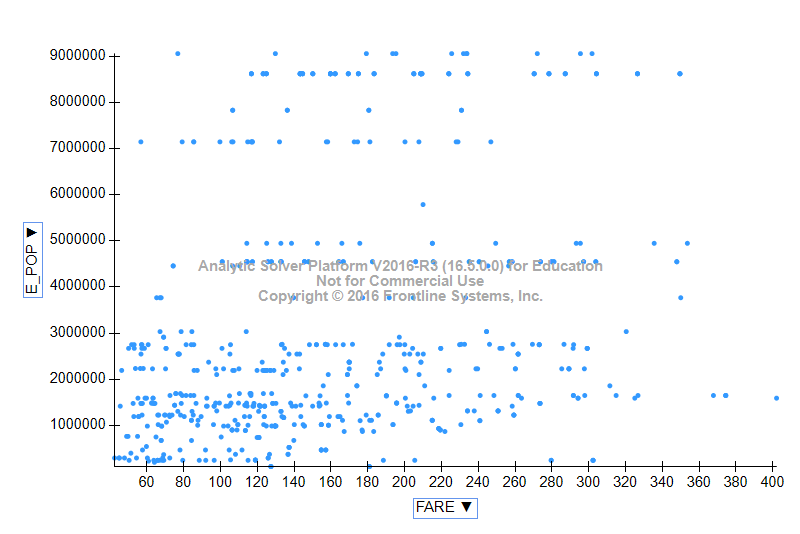
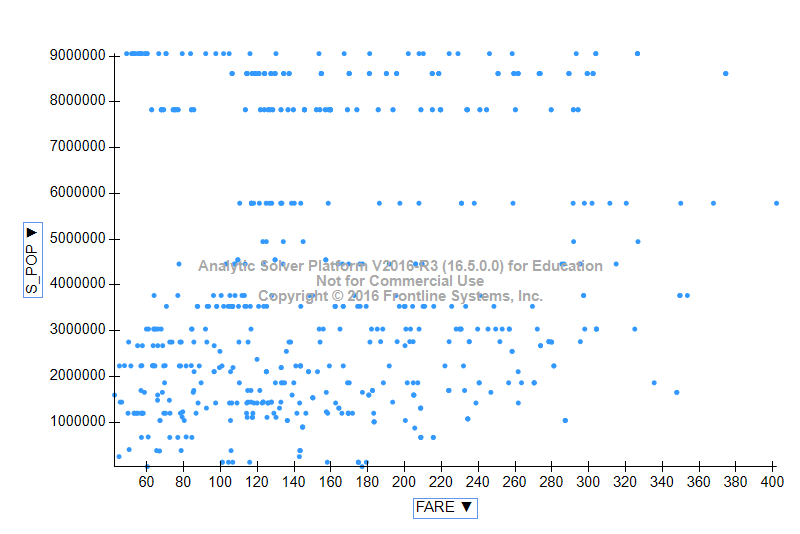
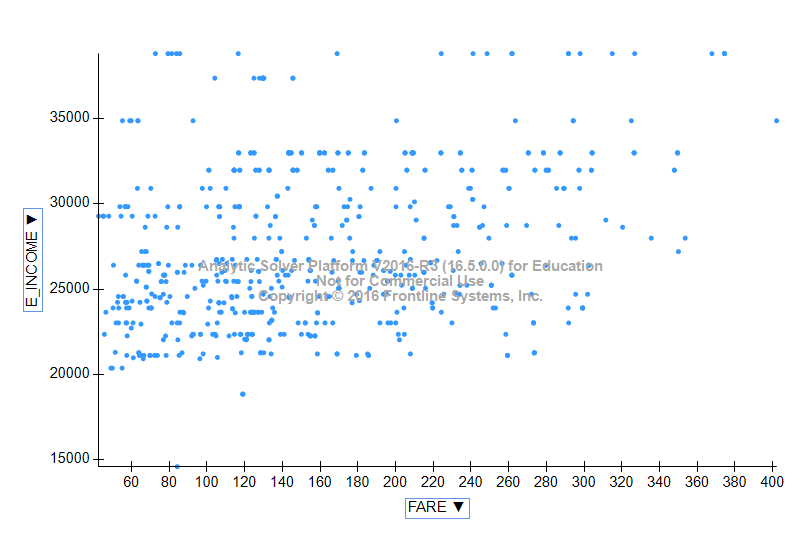
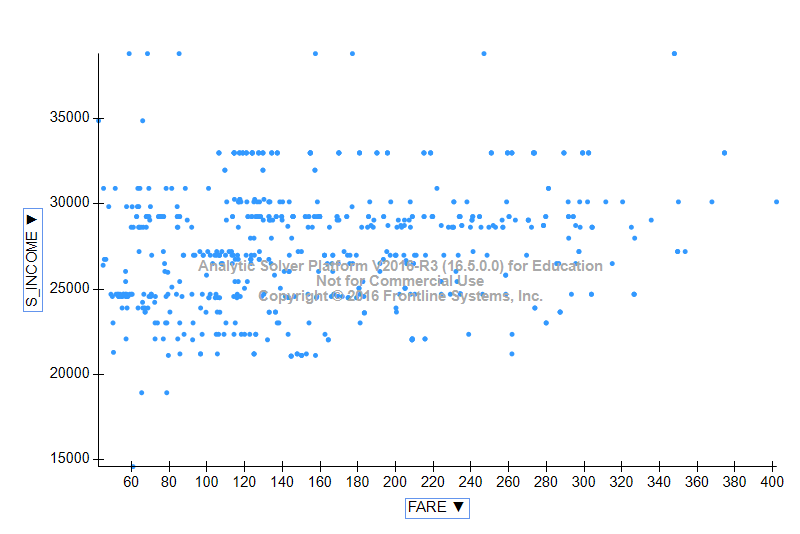
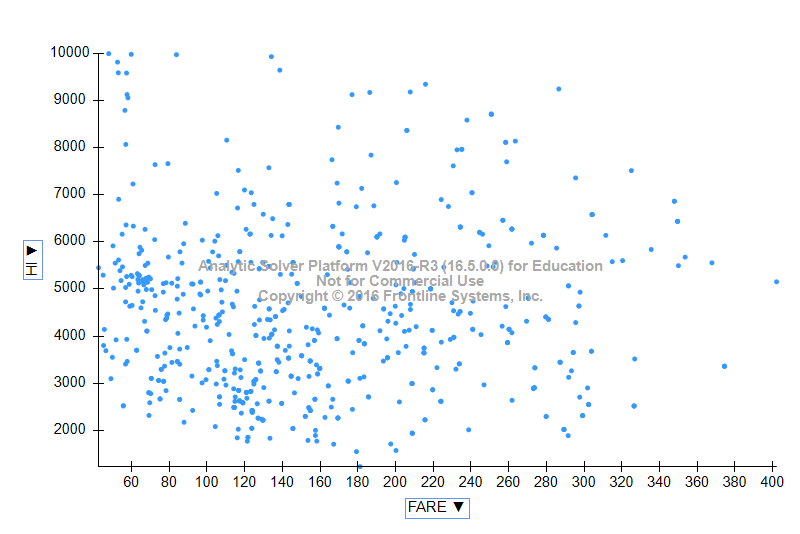
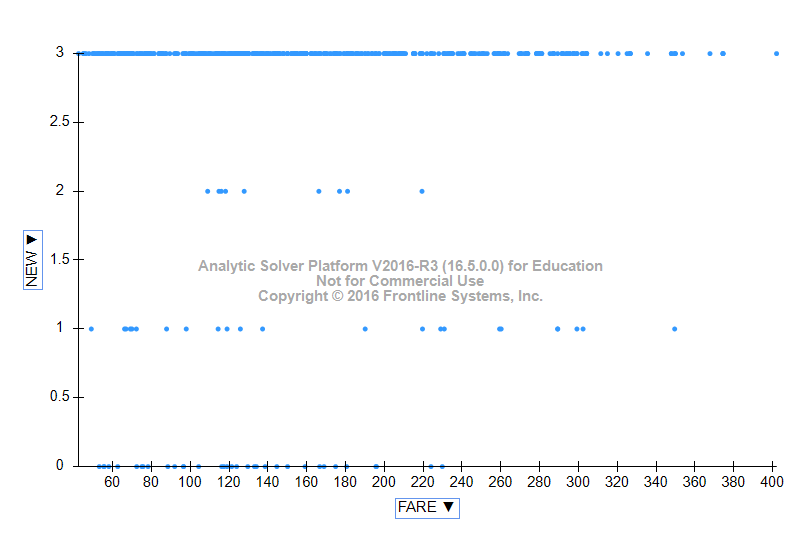
*a) Explore the numerical predictors and response (FARE) by creating a correlation*

*table and examining some scatterplots between FARE and those predictors.*

*What seems to be the best single predictor of FARE?*

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *COUPON* | *NEW* | *HI* | *S\_INCOME* | *E\_INCOME* | *S\_POP* | *E\_POP* | *DISTANCE* | *PAX* | *FARE* |
| COUPON | 1 |  |  |  |  |  |  |  |  |  |
| NEW | 0.020223 | 1 |  |  |  |  |  |  |  |  |
| HI | -0.34725 | 0.054147 | 1 |  |  |  |  |  |  |  |
| S\_INCOME | -0.0884 | 0.026597 | -0.02738 | 1 |  |  |  |  |  |  |
| E\_INCOME | 0.046889 | 0.113377 | 0.082393 | -0.13886 | 1 |  |  |  |  |  |
| S\_POP | -0.10776 | -0.01667 | -0.1725 | 0.517187 | -0.14406 | 1 |  |  |  |  |
| E\_POP | 0.09497 | 0.058568 | -0.06246 | -0.27228 | 0.458418 | -0.28014 | 1 |  |  |  |
| DISTANCE | 0.746805 | 0.080965 | -0.31237 | 0.028153 | 0.176531 | 0.018437 | 0.11564 | 1 |  |  |
| PAX | -0.33697 | 0.010495 | -0.16896 | 0.138197 | 0.259961 | 0.284611 | 0.314698 | -0.10248 | 1 |  |
| FARE | 0.496537 | 0.09173 | 0.025195 | 0.209135 | 0.326092 | 0.145097 | 0.285043 | 0.670016 | -0.09071 | 1 |





**I**n the scatter plot, fare is constantly increasing with increase in distance. Hence, it can be concluded that distance is the best single predictor. n the correlation table, we found a higher correlation between distance and fare.

*b) Explore the categorical predictors (excluding the first four) by computing the*

*percentage of flights in each category. Create a pivot table with the average fare*

*in each category. Which categorical predictor seems best for predicting FARE?*

In the data, Categorical Predictors are VACATION, SW, SLOT, GATE.

Right side tables refer to the number of flights and their percentage in each category.

Left side tables refer to the average fare in each category.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Vacation |  |  |  | Vacation |  |
|  |  |  |  |  |  |
|  | **Average of FARE** | |  |  | **Count of FARE** |
| No | 173.5525 |  |  | No | 73.35% |
| Yes | 125.9808824 |  |  | Yes | 26.65% |
| **Grand Total** | **160.8766771** |  |  | **Grand Total** | **100.00%** |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
| SW |  |  |  | SW |  |
|  |  |  |  |  |  |
|  | **Average of FARE** | |  |  | **Count of FARE** |
| No | 188.1827928 |  |  | No | 69.59% |
| Yes | 98.38226804 |  |  | Yes | 30.41% |
| **Grand Total** | **160.8766771** |  |  | **Grand Total** | **100.00%** |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
| Slot |  |  |  | Slot |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
| **Row Labels** | **Average of FARE** | |  | **Row Labels** | **Count of FARE** |
| Controlled | 186.0593956 |  |  | Controlled | 28.53% |
| Free | 150.8256798 |  |  | Free | 71.47% |
| **Grand Total** | **160.8766771** |  |  | **Grand Total** | **100.00%** |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
| Gate |  |  |  | Gate |  |
| **Row Labels** | **Average of FARE** | |  | **Row Labels** | **Count of FARE** |
| Constrained | 193.1290323 |  |  | Constrained | 19.44% |
| Free | 153.0959533 |  |  | Free | 80.56% |
| **Grand Total** | **160.8766771** |  |  | **Grand Total** | **100.00%** |
|  |  |  |  |  |  |

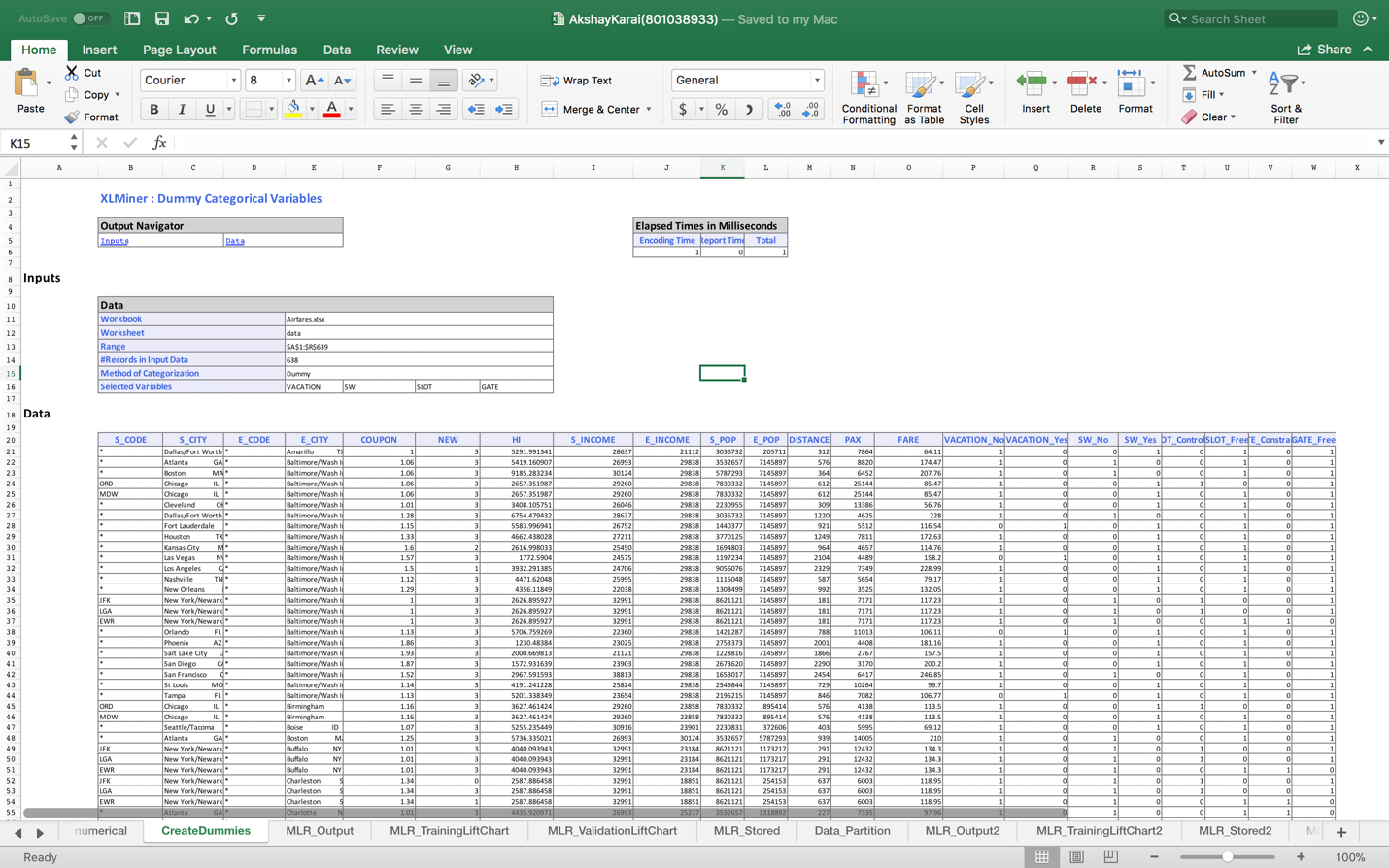
*c)Find a model for predicting the average fare on a new route:*

*i) Convert categorical variables (e.g., SW) into dummy variables. Then*

*partition the data into training and validation sets. The model will be fit to*

*the training data and evaluated on the validation set.*

Create dummies: convert categorical into numerical using dummies

**

*ii) Why should the data be partitioned into training, and validation? What will the training set be used for? What will the validation set be used for?*

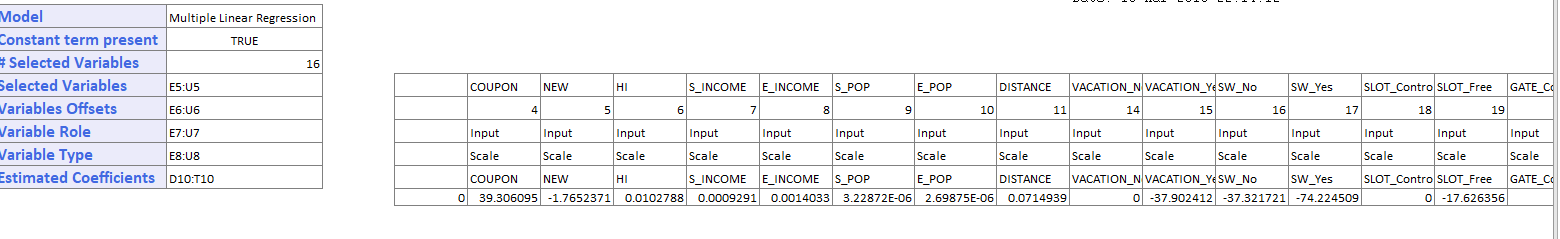
Ans: -So as to rectify the errors in the model selected, the data is to be partitioned into training and validation sets. If there is overfitting, the model will work perfectly on training data and fails on test data. Hence, to predict its accuracy, we need to check the model on validation set. Accuracy on training set is not the right metric to evaluate a model because the model will already be trained on it. Considering accuracy of other data would be the right metric.

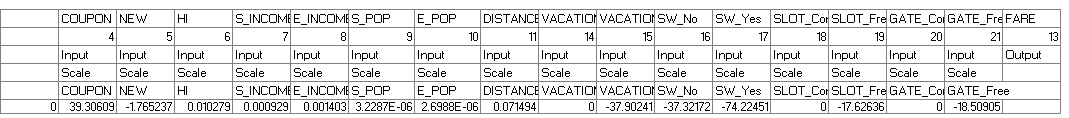
Training set is to fit the parameters in the model whereas Validation set is to tune the parameters.

*iii) Use stepwise regression to reduce the number of predictors. You can*

*ignore the first four predictors (S CODE, S CITY, E CODE, E CITY).*

*Report the estimated mode l selected.*





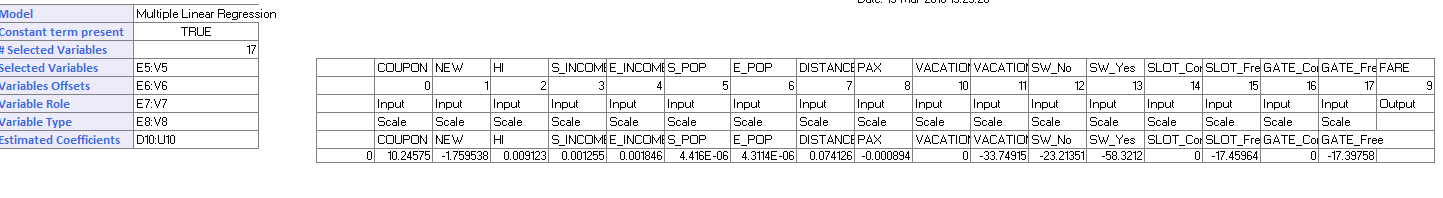
*Iv) Repeat (iii) using exhaustive search instead of stepwise regression.*

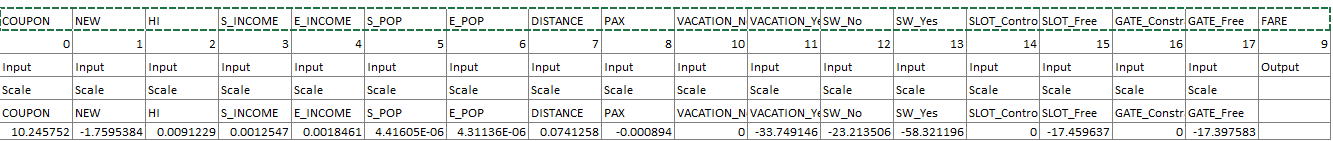
*Compare the resulting best model to the one you obtained in (iii) in terms*

*of the predictors that are in the model.*

*Ans:*

The exhaustive search model and their corresponding values are shown in the table. The description of this are present in excel sheet MLR\_Output1, MLR\_Stored2.





Stepwise regression model contains 16 predictors which includes dummy variables as well and exhaustive model contains 17 predictors which includes dummy variables .When Pax is excluded in the exhaustive model then we can see that there is very slight difference in RMS Error value. So, there is no uniform increase or decrease from stepwise to exhaustive search.

*v) Compare the predictive accuracy of both models (iii) and (iv) using measures such as RMSE and average error and lift charts.*

Stepwise Regression model:

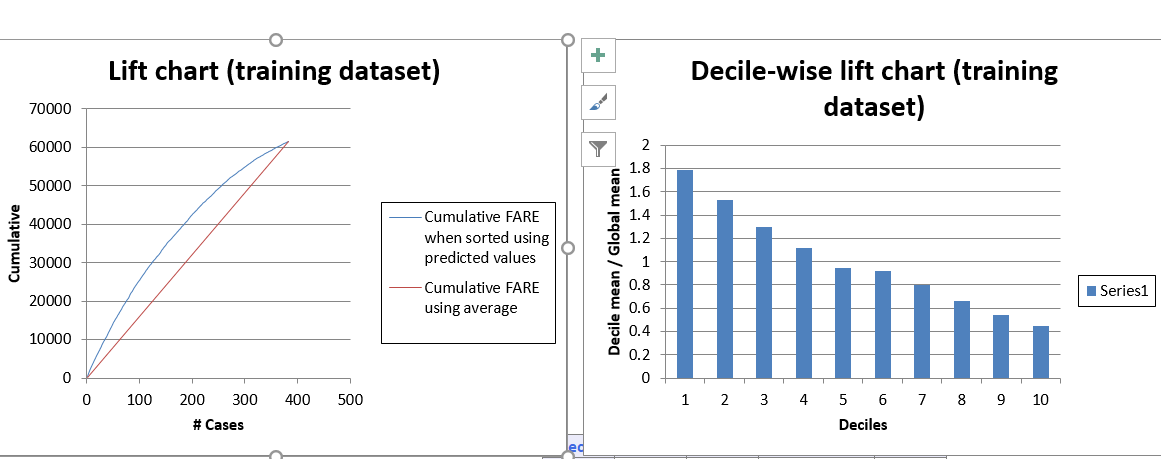
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Training Data Scoring** | | | | | |
|  |  |  |  |  |  |
|  | **Total sum of squared errors** | **RMS Error** | **Average Error** |  |  |
|  | 502743.57 | 36.230462 | 3.36905E-14 |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
| **Validation Data Scoring** | | | | | |
|  |  |  |  |  |  |
|  | **Total sum of squared errors** | **RMS Error** | **Average Error** |  |  |
|  | 511292.22 | 44.77798 | -24.1207313 |  |  |

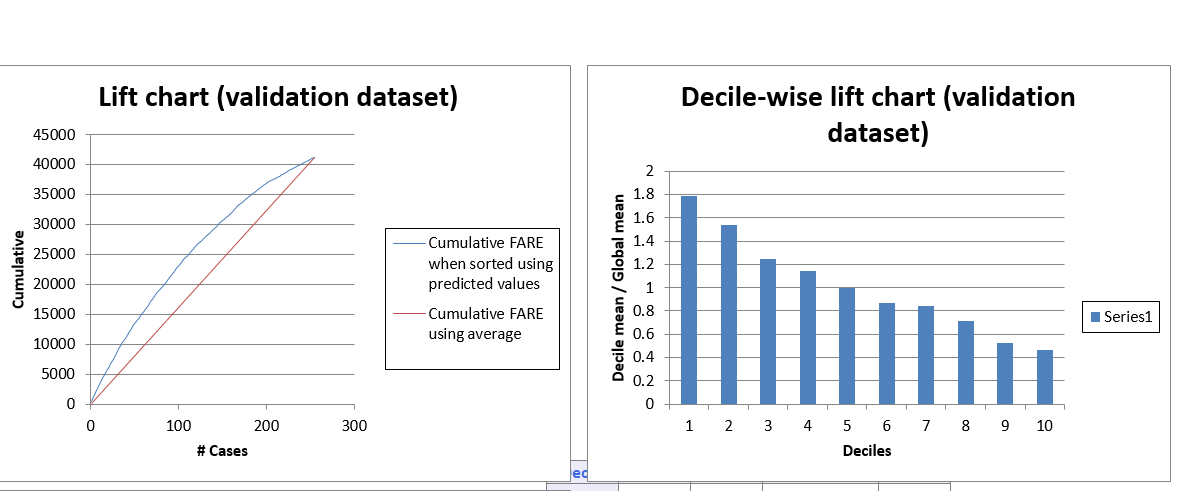
Exhaustive Search:

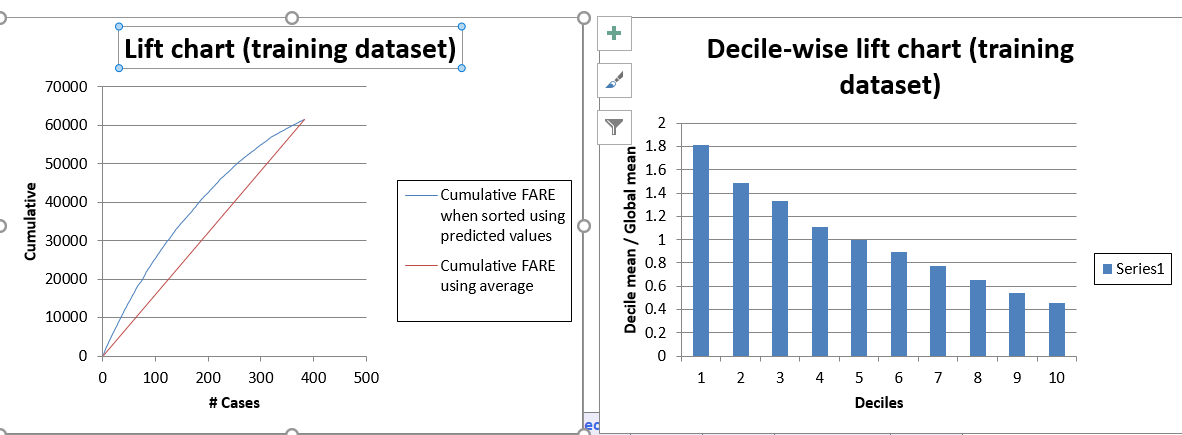
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
|  | **Training Data Scoring** | | | | |
|  |  |  |  |  |  |
|  |  | **Total sum of squared errors** | **RMS Error** | **Average Error** |  |
|  |  | 476204.638 | 35.26122724 | 6.86054E-14 |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  | **Validation Data Scoring** | | | | |
|  |  |  |  |  |  |
|  |  | **Total sum of squared errors** | **RMS Error** | **Average Error** |  |
|  |  | 473821.1311 | 43.10594023 | -22.4059602 |  |
|  |  |  |  |  |  |

Exhaustive Search has less magnitude of errors i.e., higher predictive accuracy

Decile chart for stepwise Regression:





Decile chart for Exhaustive Search:

Decile-wise lift charts for validation dataset and training dataset looks similar in both the models.

*vi) Using model (iv), predict the average fare on a route with the following*

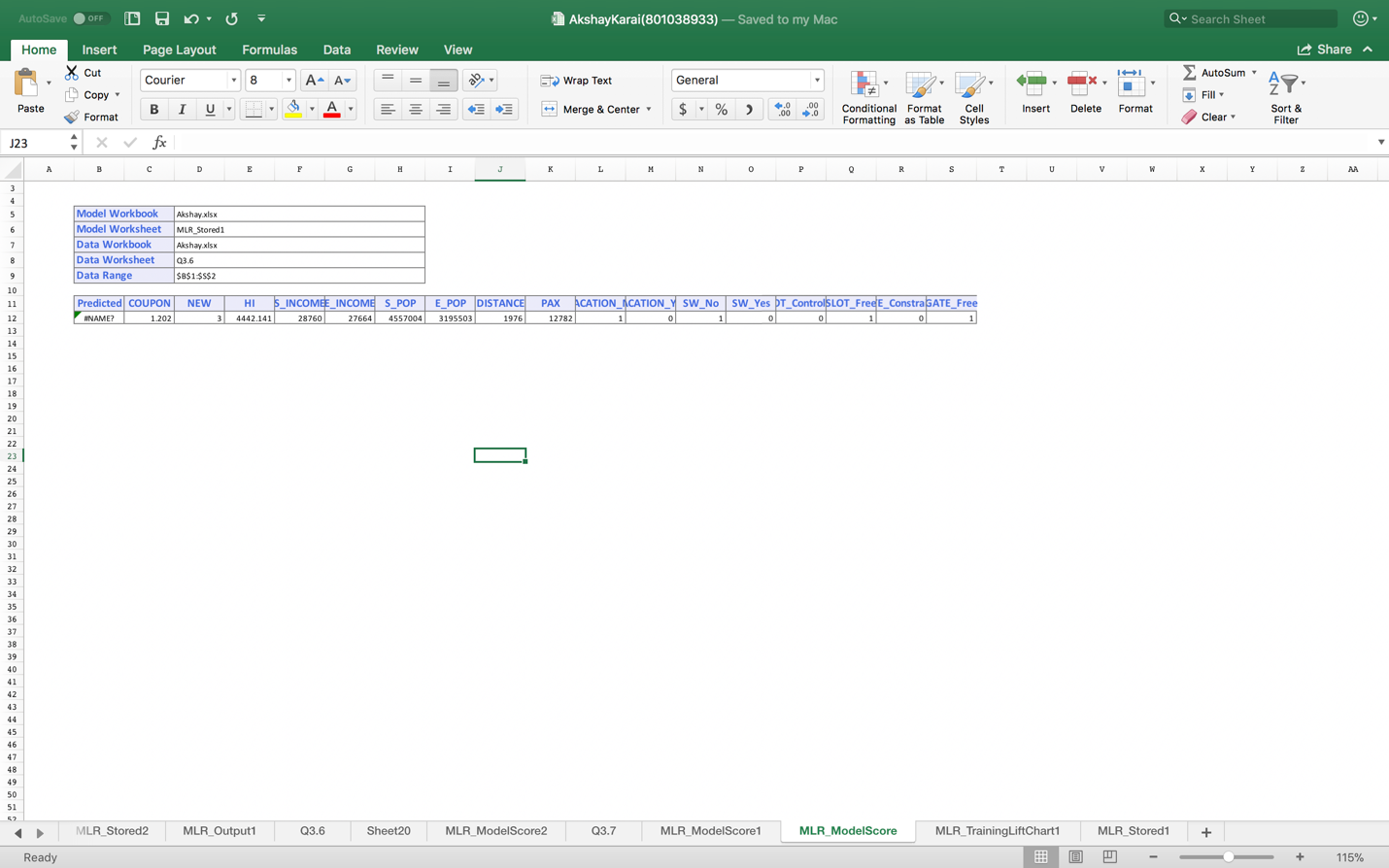
*characteristics: COUPON = 1.202, NEW = 3, VACATION = No, SW = No,*

*HI = 4442.141, S INCOME = $ 28,760, E INCOME = $27,664, S POP =*

*4,557,004, E POP = 3,195,503, SLOT = Free, GATE = Free, PAX = 12,782,*

*DISTANCE = 1976 miles.*

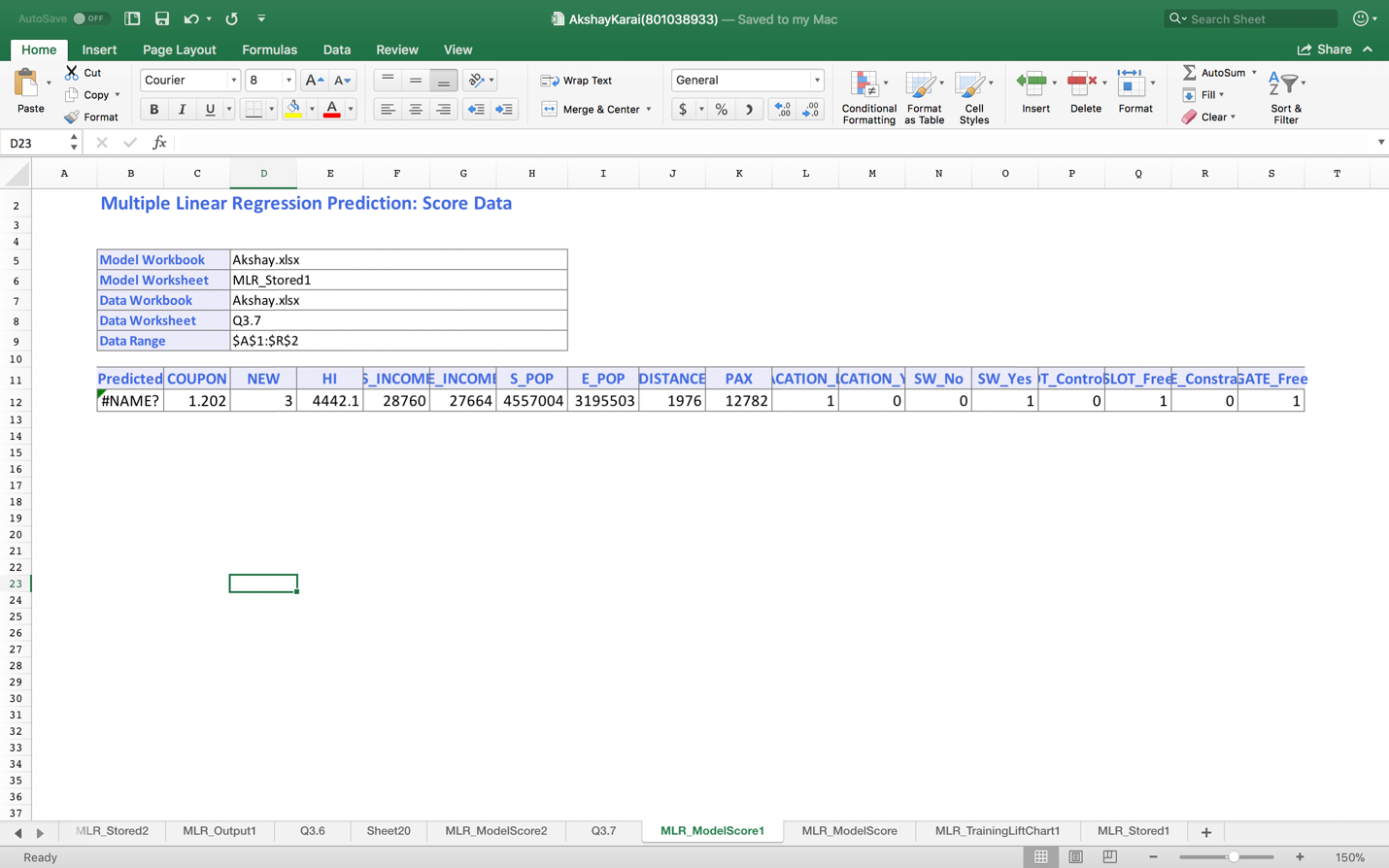
*Ans:* MLR\_ModelScore

**

*vii) Using model (iv), predict the reduction in average fare on the route in (vi)*

*if Southwest decides to cover this route.*

*Ans:* MLR\_ModelScore1

**

*viii) In reality, which of the factors will not be available for predicting the*

*average fare from a new airport (i.e., before flights start operating on those*

*routes)? Which ones can be estimated? How?*

*Ans:*

**PAX**: Number of passengers that would be travelling in that route will not be available before flights start operating.

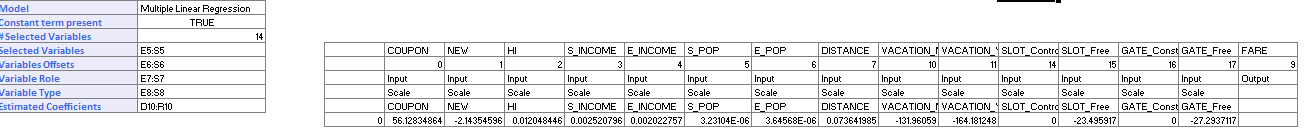
**SW:** One more factor that will not be available before flights start operating is presence of Southwest Airlines in that route. Southwest may or may not announce service on that route.

*ix) Select a model that includes only factors that are available before flights*

*begin to*

*operate on the new route. Use an exhaustive search to find such a*

*model.*



*x) Use the model in (ix) to predict the average fare on a route with*

*characteristics COUPON = 1.202, NEW = 3, VACATION = No, SW = No,*

*HI = 4442.141, S INCOME = $ 28,760, E INCOME = $ 27,664, S POP = 4,557,004, E POP = 3,195,503, SLOT = Free, GATE = Free, PAX = 12,782, DISTANCE = 1976 miles.*

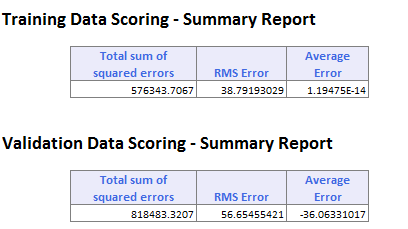
For the features available before flights begin, the fare prediction using exhaustive search model is 245.59. Sheet-MLR\_ModelScore

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | **Model Workbook** | | Pujitha.xlsb.xlsx | | | | |  |  |  |  |  |  |  |  |  |  |  |
|  | **Model Worksheet** | | MLR\_Stored1 | | | | |  |  |  |  |  |  |  |  |  |  |  |
|  | **Data Workbook** | | Pujithaa.xlsb.xlsx | | | | |  |  |  |  |  |  |  |  |  |  |  |
|  | **Data Worksheet** | | Q3.6 | | | | |  |  |  |  |  |  |  |  |  |  |  |
|  | **Data Range** | | $B$1:$S$2 | | | | |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | **Predicted** | **COUPON** | **NEW** | **HI** | **S\_INCOME** | **E\_INCOME** | **S\_POP** | **E\_POP** | **DISTANCE** | **PAX** | **VACATION\_No** | **VACATION\_Yes** | **SW\_No** | **SW\_Yes** | **SLOT\_Controlled** | **SLOT\_Free** | **GATE\_Constrained** | **GATE\_Free** |
|  | 245.5965013 | 1.202 | 3 | 4442.141 | 28760 | 27664 | 4557004 | 3195503 | 1976 | 12782 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

*xi) Compare the predictive accuracy of this model with model (iv). Is this*

*model good enough, or is it worthwhile re-evaluating the model once flights begin on the new route?*

Exhaustive search



Comparing flights