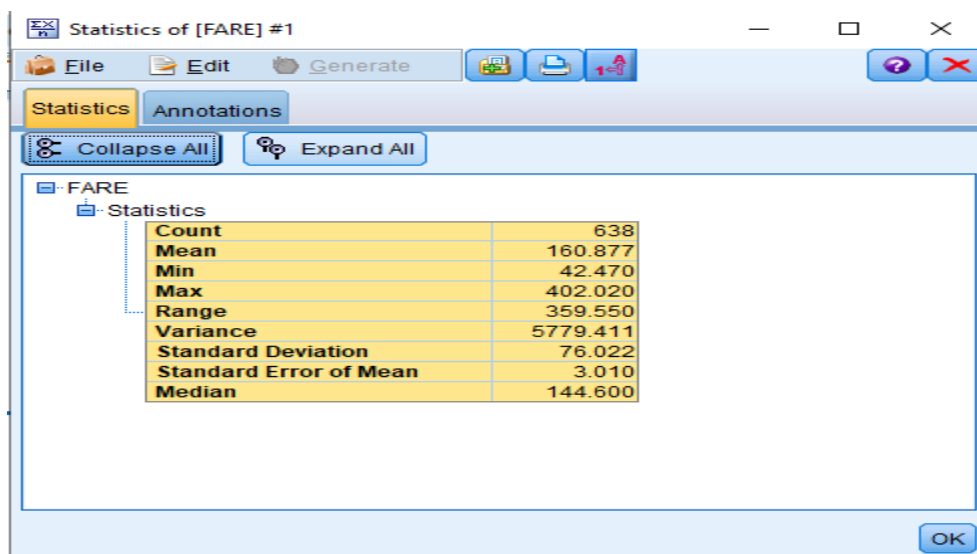


Find the MLR model for predicting the average fare on a new route:

For this assignment, we have given a data set of airlines, with 638 data fields and we have to perform the MLR model for predicting the average fare on a new route while answering the followings:

a) Convert categorical variables (e.g. SW) into indicator variables.

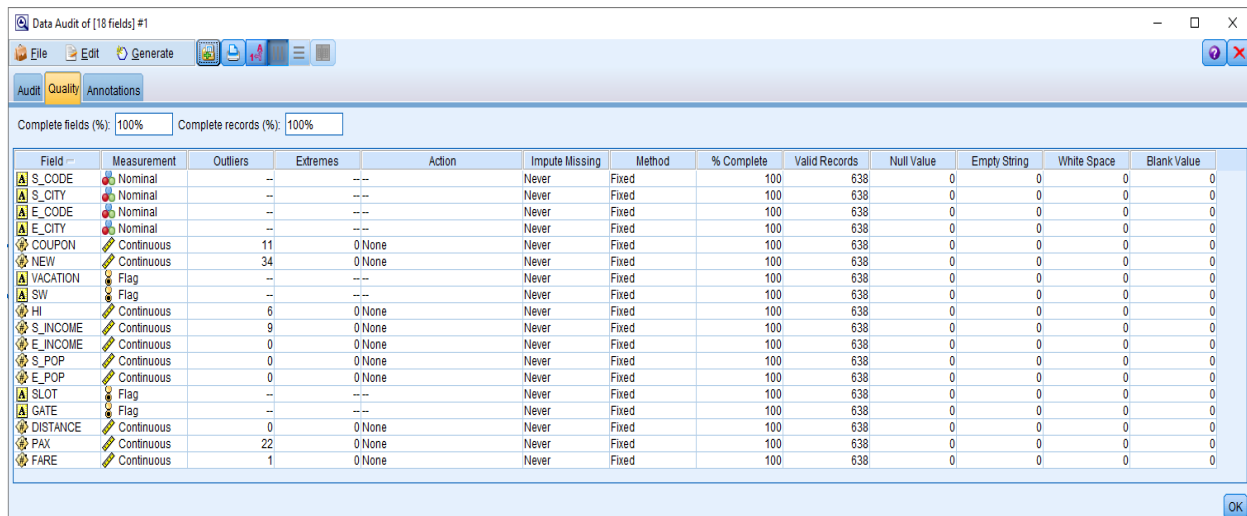
Before we convert the categorical variables into indicator variables, we have first checked the statistics of the given dataset to see if the provided dataset makes sense or not. We have only performed the statistics on fare prices in Airlines dataset as we have to predict the MLR model for predicting the average fare on a new route. The provided data makes sense, below is the snapshot of the statistics.



Statistics of [FARE] #1

| Statistic | Value |
|------------------------|----------|
| Count | 638 |
| Mean | 160.877 |
| Min | 42.470 |
| Max | 402.020 |
| Range | 359.550 |
| Variance | 5779.411 |
| Standard Deviation | 76.022 |
| Standard Error of Mean | 3.010 |
| Median | 144.600 |

The next step, I did is the check the quality of data, to check if we have any missing values in the provided dataset, the provided dataset is of good quality without any missing values.



Data Audit of [18 fields] #1

| Field | Measurement | Outliers | Extremes | Action | Impute Missing | Method | % Complete | Valid Records | Null Value | Empty String | White Space | Blank Value |
|----------|-------------|----------|----------|--------|----------------|--------|------------|---------------|------------|--------------|-------------|-------------|
| S_CODE | Nominal | -- | -- | | Never | Fixed | 100 | 638 | 0 | 0 | 0 | 0 |
| S_CITY | Nominal | -- | -- | | Never | Fixed | 100 | 638 | 0 | 0 | 0 | 0 |
| E_CODE | Nominal | -- | -- | | Never | Fixed | 100 | 638 | 0 | 0 | 0 | 0 |
| E_CITY | Nominal | -- | -- | | Never | Fixed | 100 | 638 | 0 | 0 | 0 | 0 |
| COUPON | Continuous | 11 | 0 None | | Never | Fixed | 100 | 638 | 0 | 0 | 0 | 0 |
| NEW | Continuous | 34 | 0 None | | Never | Fixed | 100 | 638 | 0 | 0 | 0 | 0 |
| VACATION | Flag | -- | -- | | Never | Fixed | 100 | 638 | 0 | 0 | 0 | 0 |
| SW | Flag | -- | -- | | Never | Fixed | 100 | 638 | 0 | 0 | 0 | 0 |
| HI | Continuous | 6 | 0 None | | Never | Fixed | 100 | 638 | 0 | 0 | 0 | 0 |
| S_INCOME | Continuous | 9 | 0 None | | Never | Fixed | 100 | 638 | 0 | 0 | 0 | 0 |
| E_INCOME | Continuous | 0 | 0 None | | Never | Fixed | 100 | 638 | 0 | 0 | 0 | 0 |
| S_POP | Continuous | 0 | 0 None | | Never | Fixed | 100 | 638 | 0 | 0 | 0 | 0 |
| E_POP | Continuous | 0 | 0 None | | Never | Fixed | 100 | 638 | 0 | 0 | 0 | 0 |
| SLOT | Flag | -- | -- | | Never | Fixed | 100 | 638 | 0 | 0 | 0 | 0 |
| GATE | Flag | -- | -- | | Never | Fixed | 100 | 638 | 0 | 0 | 0 | 0 |
| DISTANCE | Continuous | 0 | 0 None | | Never | Fixed | 100 | 638 | 0 | 0 | 0 | 0 |
| PAX | Continuous | 22 | 0 None | | Never | Fixed | 100 | 638 | 0 | 0 | 0 | 0 |
| FARE | Continuous | 1 | 0 None | | Never | Fixed | 100 | 638 | 0 | 0 | 0 | 0 |

Restructure

Settings Annotations

Available fields:

GATE

Available values:

Create restructured fields:

SW_No
SW_Yes
VACATION_No
VACATION_Yes
SLOT_Controlled
SLOT_Free
GATE_Constrained
GATE_Free

☒ Include field names

☒ Use values from other field(s) ☐ Create numeric value flags

Value field(s):

OK Cancel Apply Reset

| Table (63 fields, 638 records) #1 | | | | | | | | | | | | | | | | | | | | | | | |
|-----------------------------------|-------------|------|----------|----------|------------|------|------|------|--|--|--|--|--|--|--|-------|--------|-------------|--------------|-----------------|-----------|------------------|-----------|
| | File | Edit | Generate | | | | | | | | | | | | | | | | | | | | |
| Tabl | Annotations | | | | | | | | | | | | | | | | | | | | | | |
| | | | | VACATION | SW | | SLOT | GATE | | | | | | | | SW_No | SW_Yes | VACATION_No | VACATION_Yes | SLOT_Controlled | SLOT_Free | GATE_Constrained | GATE_Free |
| 23 | * | * | No | Yes | Free | Free | | | | | | | | | | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 |
| 24 | * | * | Yes | Yes | Free | Free | | | | | | | | | | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| 25 | * | * | No | Yes | Controlled | Free | | | | | | | | | | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 |
| 26 | * | * | No | Yes | Free | Free | | | | | | | | | | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 |
| 27 | * | * | No | Yes | Free | Free | | | | | | | | | | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 1 |
| 28 | * | * | No | No | Free | Free | | | | | | | | | | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 |
| 29 | * | * | No | No | Controlled | Free | | | | | | | | | | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 1 |
| 30 | * | * | No | No | Controlled | Free | | | | | | | | | | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 |
| 31 | * | * | No | No | Free | Con. | | | | | | | | | | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 |
| 32 | * | * | No | No | Controlled | Free | | | | | | | | | | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 |
| 33 | * | * | No | No | Controlled | Free | | | | | | | | | | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 1 |
| 34 | * | * | No | No | Free | Con. | | | | | | | | | | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 |
| 35 | * | * | No | No | Free | Con. | | | | | | | | | | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 |
| 36 | * | * | No | No | Free | Con. | | | | | | | | | | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 |
| 37 | * | * | No | No | Controlled | Con. | | | | | | | | | | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 |
| 38 | * | * | No | No | Free | Con. | | | | | | | | | | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 |
| 39 | * | * | No | No | Free | Con. | | | | | | | | | | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 |
| 40 | * | * | No | No | Controlled | Con. | | | | | | | | | | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 |
| 41 | * | * | No | No | Controlled | Con. | | | | | | | | | | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 |
| 42 | * | * | No | No | Free | Con. | | | | | | | | | | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 |
| 43 | * | * | No | No | Controlled | Free | | | | | | | | | | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 |
| 44 | * | * | No | No | Free | Free | | | | | | | | | | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 |
| 45 | * | * | No | Yes | Controlled | Free | | | | | | | | | | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 |

Marist College | Fall 2020

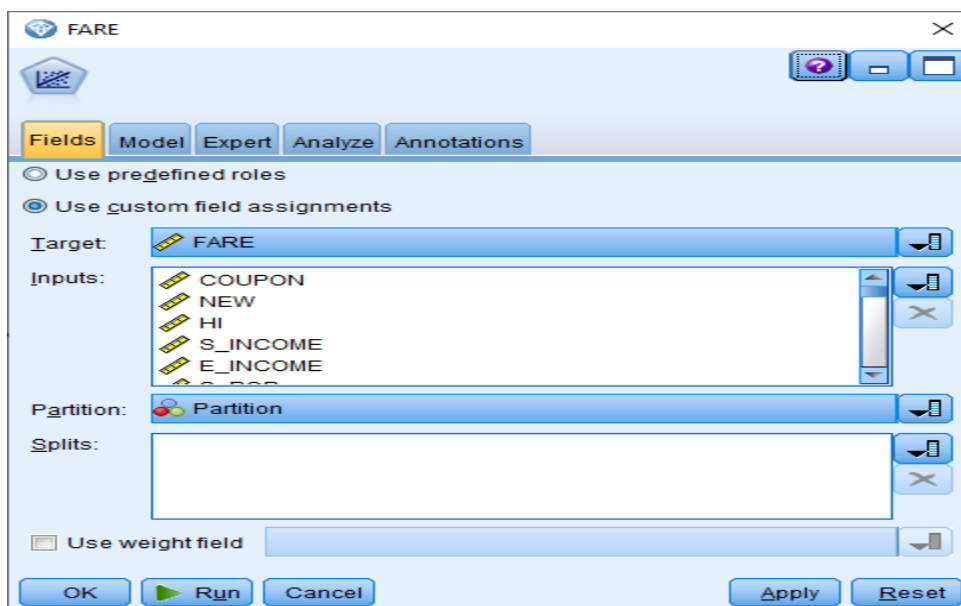
As required, we have partitioned the data into training and test subset (70% - 30% ratio). For this we have used the partition node, below is the snapshot for it.

[illegible]

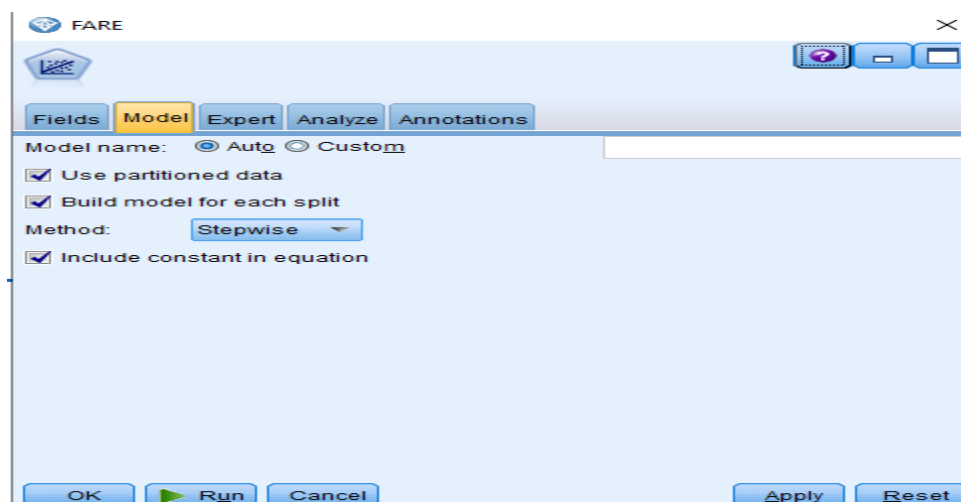
From the above snapshot, we can see that the data has been divided for training and testing purposes (chosen at random).

c) Use stepwise regression to reduce the number of predictors using the training data

For this step of the assignment, we have used the regression model node from the SPSS and selected all the fields variable (features) as input expect the SW 0 field and FARE fields (which implies that we are performing stepwise regression all the all the data which are operated by South West Airlines and not operated SW Airlines fields will act as our reference indicator) and selected the FARE field as a target field, and most importantly chosen the partition data to perform the required regression, which can be seen below.

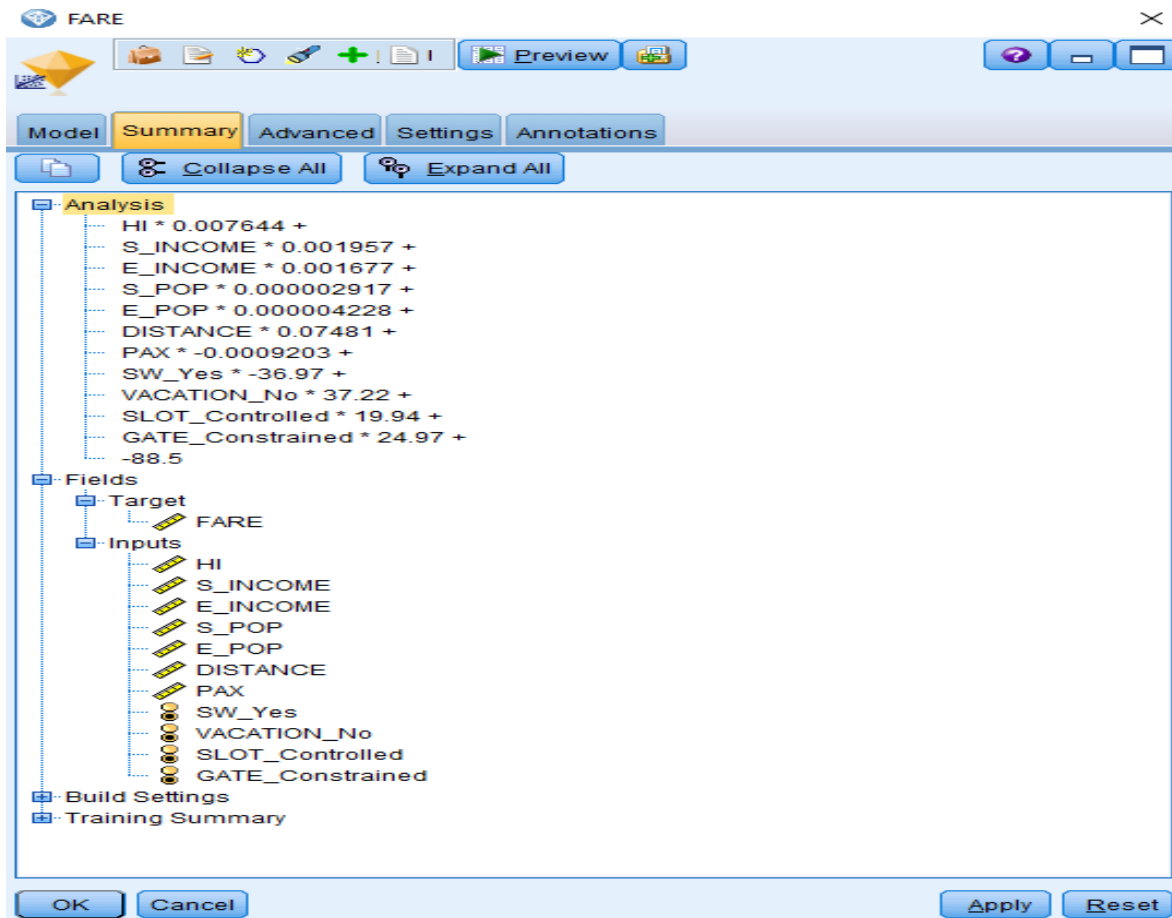


As required we have selected the stepwise regression method from our partitioned data, which can be seen below:



d) Report the model parameters: regression coefficient estimates with their standard errors, goodness of fit metrics (R-squared, adjusted R-squared), standard error of the estimate (s), t-test values (scores and p-values), F-test values (F-score and p-value)

Below is the snapshot for the summary for MLR model which is the representation of the equation, here MLR is represented here through a linear combination of the selected predictors.



Below is the snapshot of details representation of the predictors and all the features explaining the fit of the model. If we look into here, we can see the R square and adjusted R square value which represents the fit for the model. Here the 11th model has the highest R square and adjusted R square value which means that 11th model is the 78.4% and the highest fit for this given data, If we are to select a model we would select the 11th model.

The screenshot shows the FARE software interface. The 'Model Summary' tab is selected, displaying a table with 5 columns: Model, R, R Square, Adjusted R Square, and Std. Error of the Estimate. The table lists 11 models. Model 11 has the highest R Square (.784) and Adjusted R Square (.778). Below the table, a list of predictors is provided for each model, labeled from 'a' to 'k'. A red arrow points to the list of predictors for Model 11 (labeled 'k').

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|-------|-------------------|----------|-------------------|----------------------------|
| 1 | .675 ^a | .456 | .454 | 57.128918 |
| 2 | .775 ^b | .600 | .598 | 49.009547 |
| 3 | .837 ^c | .700 | .698 | 42.491159 |
| 4 | .849 ^d | .722 | .719 | 40.997202 |
| 5 | .859 ^e | .738 | .735 | 39.779505 |
| 6 | .871 ^f | .758 | .755 | 38.300048 |
| 7 | .873 ^g | .762 | .758 | 38.067835 |
| 8 | .876 ^h | .767 | .763 | 37.652728 |
| 9 | .878 ⁱ | .772 | .767 | 37.341233 |
| 10 | .883 ^j | .779 | .774 | 36.742794 |
| 11 | .885 ^k | .784 | .778 | 36.410904 |

a. Predictors: (Constant), DISTANCE
b. Predictors: (Constant), DISTANCE, SW_Yes
c. Predictors: (Constant), DISTANCE, SW_Yes, VACATION_No
d. Predictors: (Constant), DISTANCE, SW_Yes, VACATION_No, HI
e. Predictors: (Constant), DISTANCE, SW_Yes, VACATION_No, HI, SLOT_Controlled
f. Predictors: (Constant), DISTANCE, SW_Yes, VACATION_No, HI, SLOT_Controlled, GATE_Constrained
g. Predictors: (Constant), DISTANCE, SW_Yes, VACATION_No, HI, SLOT_Controlled, GATE_Constrained, E_INCOME
h. Predictors: (Constant), DISTANCE, SW_Yes, VACATION_No, HI, SLOT_Controlled, GATE_Constrained, E_INCOME, PAX
i. Predictors: (Constant), DISTANCE, SW_Yes, VACATION_No, HI, SLOT_Controlled, GATE_Constrained, E_INCOME, PAX, S_POP
j. Predictors: (Constant), DISTANCE, SW_Yes, VACATION_No, HI, SLOT_Controlled, GATE_Constrained, E_INCOME, PAX, S_POP, E_POP
k. Predictors: (Constant), DISTANCE, SW_Yes, VACATION_No, HI, SLOT_Controlled, GATE_Constrained, E_INCOME, PAX, S_POP, E_POP, S_INCOME

ANOVA

Here, we have selected stepwise model to perform the MLR which means the it will select predictors in stepwise squence in different collections and monitor the fit of the data, and adding its evalution to our final above summary. Below are the details of coffieicients from pour selected model, in this case 11th model the complete details of the cofficient can be seen from the provided the .str file.

In our selected model, the standard error of estimation is also lowest (36.41) unit as compare the the total FARE mean for the South west Airlines opertions from our given dataset.

| Coefficients | | | | | |
|--------------|------------------|-----------------------------|------------|---------------------------|------|
| Model | | Unstandardized Coefficients | | Standardized Coefficients | Sig. |
| | | B | Std. Error | Beta | |
| 11 | (Constant) | -88.499 | 23.459 | | .000 |
| | DISTANCE | .075 | .003 | .645 | .000 |
| | SW_Yes | -36.971 | 4.519 | -.223 | .000 |
| | VACATION_No | 37.224 | 4.461 | .209 | .000 |
| | HI | .008 | .001 | .173 | .000 |
| | SLOT_Controlled | 19.939 | 4.610 | .116 | .000 |
| | GATE_Constrained | 24.966 | 5.042 | .124 | .000 |
| | E_INCOME | .002 | .000 | .100 | .000 |
| | PAX | -.001 | .000 | -.156 | .000 |
| | S_POP | 2.917E-6 | .000 | .114 | .000 |
| | E_POP | 4.228E-6 | .000 | .152 | .000 |
| | S_INCOME | .002 | .001 | .088 | .003 |

The important thing to notice from the above table is that not all the predictors are selected and some of the predictors values are also very small which can be further neglected. Here we can notice that some of the coefficients of our predictors are negative which implies that increase in the predictor will decrease in the response variables. We will discuss more on these value in our regression equation.

The other thing to consider is the sign for the t test coeffieicients, t-statistic, indicates whether or not an independent variable is correlated to the dependent variable; that is, it tells you whether or not the independent variable helps to explain the dependent variable and, therefore, whether or we should leave the independent variable in the model, in our case the sign of t-statistics is of same sign as of the other coeffieicients which makes sense.

One more important factor to notice is that all the values of sig. if less than 0.001 which makes which means our model is significant and it make sense to the data. Below is the anova test results.

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|-------|------------|----------------|-----|-------------|---------|-------------------|
| 1 | Regression | 1204529.350 | 1 | 1204529.350 | 369.067 | .000 ^b |
| | Residual | 1439297.539 | 441 | 3263.713 | | |
| | Total | 2643826.889 | 442 | | | |
| 2 | Regression | 1586975.199 | 2 | 793487.600 | 330.353 | .000 ^c |
| | Residual | 1056851.690 | 440 | 2401.936 | | |
| | Total | 2643826.889 | 442 | | | |
| 3 | Regression | 1851213.017 | 3 | 617071.006 | 341.773 | .000 ^d |
| | Residual | 792613.872 | 439 | 1805.499 | | |
| | Total | 2643826.889 | 442 | | | |
| 4 | Regression | 1907649.383 | 4 | 476912.346 | 283.746 | .000 ^e |
| | Residual | 736177.506 | 438 | 1680.771 | | |
| | Total | 2643826.889 | 442 | | | |
| 5 | Regression | 1952314.141 | 5 | 390462.828 | 246.752 | .000 ^f |
| | Residual | 691512.748 | 437 | 1582.409 | | |
| | Total | 2643826.889 | 442 | | | |
| 6 | Regression | 2004261.243 | 6 | 334043.540 | 227.722 | .000 ^g |
| | Residual | 639565.646 | 436 | 1466.894 | | |
| | Total | 2643826.889 | 442 | | | |
| 7 | Regression | 2013442.261 | 7 | 287634.609 | 198.484 | .000 ^h |
| | Residual | 630384.628 | 435 | 1449.160 | | |
| | Total | 2643826.889 | 442 | | | |
| 8 | Regression | 2028532.981 | 8 | 253566.623 | 178.854 | .000 ⁱ |
| | Residual | 615293.907 | 434 | 1417.728 | | |
| | Total | 2643826.889 | 442 | | | |
| 9 | Regression | 2040065.679 | 9 | 226673.964 | 162.564 | .000 ^j |
| | Residual | 603761.210 | 433 | 1394.368 | | |
| | Total | 2643826.889 | 442 | | | |
| 10 | Regression | 2060612.662 | 10 | 206061.266 | 152.834 | .000 ^k |
| | Residual | 583214.227 | 432 | 1350.033 | | |
| | Total | 2643826.889 | 442 | | | |
| 11 | Regression | 2072426.947 | 11 | 188402.450 | 142.110 | .000 ^l |
| | Residual | 571399.941 | 431 | 1325.754 | | |
| | Total | 2643826.889 | 442 | | | |

From the above annova test we can see the F-statistics which is also a measure of fit, The "F" column provides a statistic for testing the hypothesis that it is not equal to zero. Here from the above we can see that in our 11th model the F statistics is equal to 142.110 which is less the 0.001, providing strong evidence against the null hypothesis.

e) Write and explain the regression equation

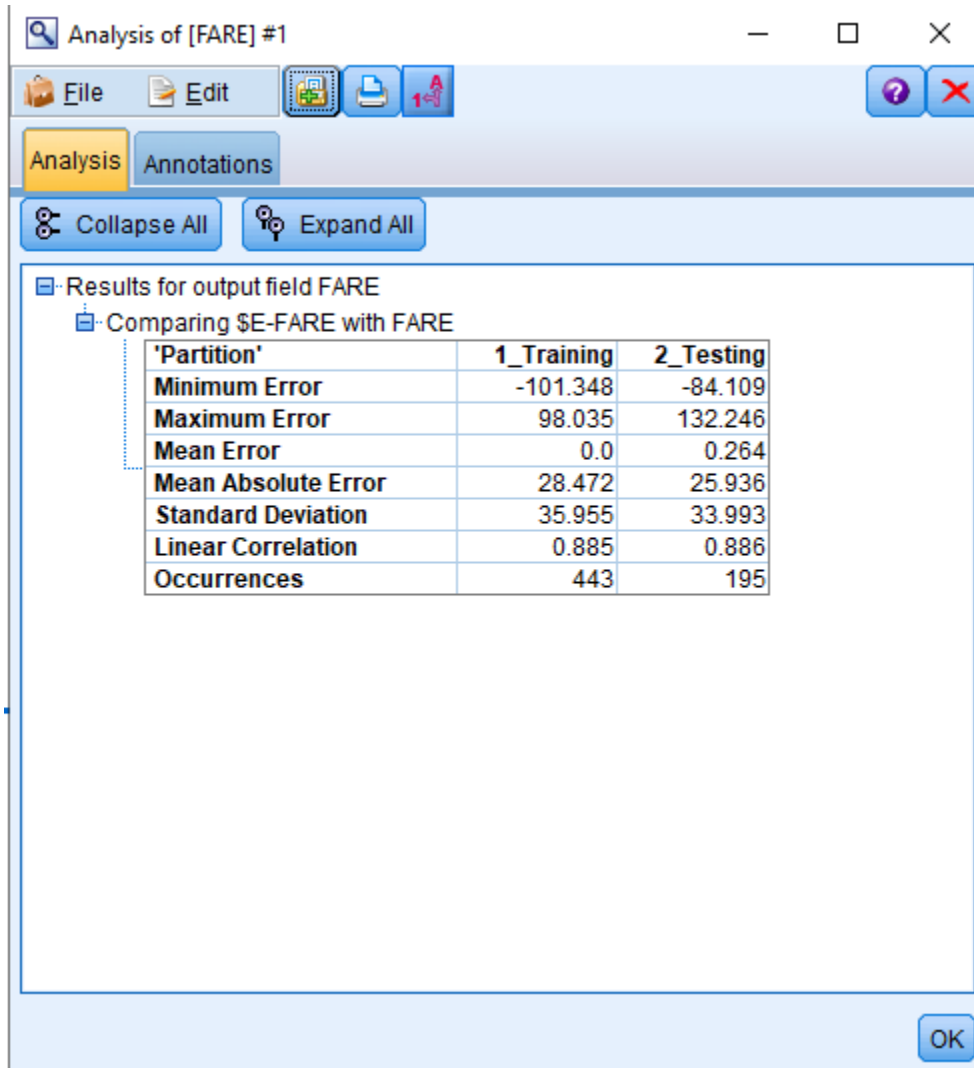
In the Airline Data set FARE for the new route

$$= (.075) * \text{Distance} + (-36.971) * \text{SW_Yes} + (37.224) * \text{VACATION_No_} + (19.939) * \text{SLOT_Controlled} + (24.966) * \text{GATE_Constrained.} - 88.499$$

From the above equation, we can see that Southwest Airlines provided will lower the fare for the new route and non-vacation destination route and busy airports can increase the price for the fares for new

route. The other predictors have very little effect on the equation therefore, they are been removed eliminated from the equation as they have little effects.

f) Using the test subset, compute the predictive accuracy metrics (MAE, max, min errors, std dev of the predictive error)

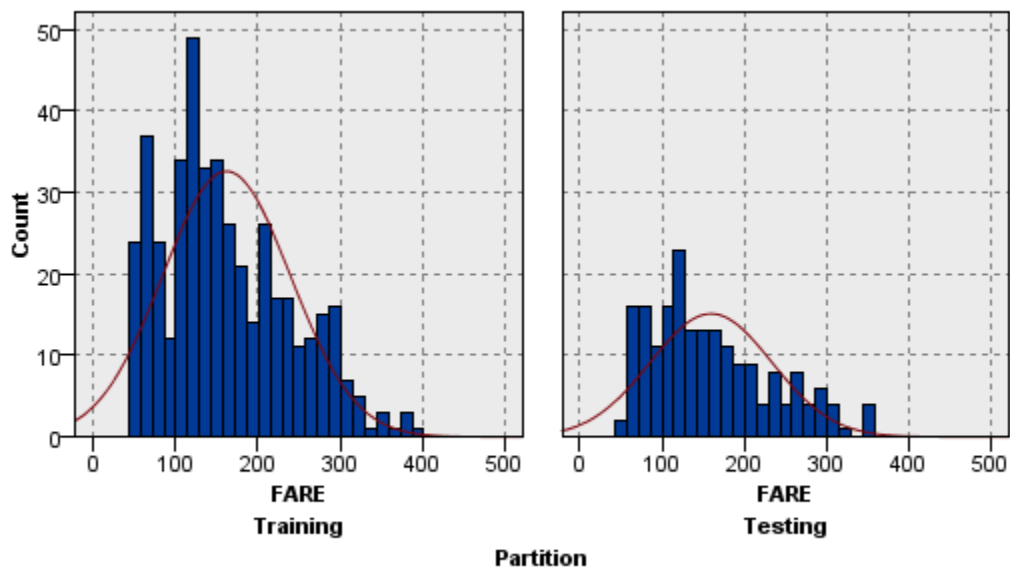


| 'Partition' | 1_Training | 2_Testing |
|---------------------|------------|-----------|
| Minimum Error | -101.348 | -84.109 |
| Maximum Error | 98.035 | 132.246 |
| Mean Error | 0.0 | 0.264 |
| Mean Absolute Error | 28.472 | 25.936 |
| Standard Deviation | 35.955 | 33.993 |
| Linear Correlation | 0.885 | 0.886 |
| Occurrences | 443 | 195 |

The above the predictive accuracy metrics, which is performed using the analysis node.

g) What is the typical predictive error that you can expect with this model?

The typical predictive error for model estimated to be 29/144.6 which is estimated to be 20%.



The model residuals appear slightly skewed and not normal. We cannot assume the error values are normally distributed, so model predictive performance may decrease.