****

**American University of Sharjah**

**College of Arts and Sciences**

-------------------------------------------------------------------------------------------------------------------------------

**STA 301: Foundations of Statistics for Data Science**

**Spring 2021**

-------------------------------------------------------------------------------------------------------------------------------

**Submitted to**

**Dr. Hana Sulieman**

**Department of Mathematics and Statistics**

-------------------------------------------------------------------------------------------------------------------------------

**Project Report**

-------------------------------------------------------------------------------------------------------------------------------

**Team Number: 10**

-------------------------------------------------------------------------------------------------------------------------------

**Date of Submission: 22nd May 2021**

-------------------------------------------------------------------------------------------------------------------------------

**Name & ID:**

|  |  |
| --- | --- |
| **Mahira Pathan** | **g00085003** |
| **Prem Rajendran** | **b00084833** |

**Section 1: Introduction**

**OBJECTIVE:**

The objective of this project is to analyze the UK Used Car Dataset and devise a model to predict the prices of used cars based on the various predictors.

**CONTENT:**

This data set contains information about the various used cars in the UK automotive market. Each brand was separated into different datasets, with the same parameters. This data set was combined and cleaned via spreadsheet softwares like Excel and Minitab. The data set was split into the training and test sets using Python at a ratio of 70% and 30%.

The cleaned data set contains information of price, transmission, mileage, fuel type, road tax, miles per gallon (mpg), and engine size.

Description of the dependent variable (target variable):

The price of the car is the independent variable. It represents the resale values of a car based on its different features. The price values range from £5 to £147500.

Description of the independent variables (features):

The features of the model are transmission, mileage, fuel type, road tax, miles per gallon (mpg), and engine size.

Mileage refers to the number of miles that a used car has traversed across its life span. Road tax refers to the Vehicle Excise Duty (VED) paid for vehicles in the UK. Miles per gallon is a measure of the engine's efficiency, which gives us a measure of how many miles a vehicle can travel per gallon of fuel.

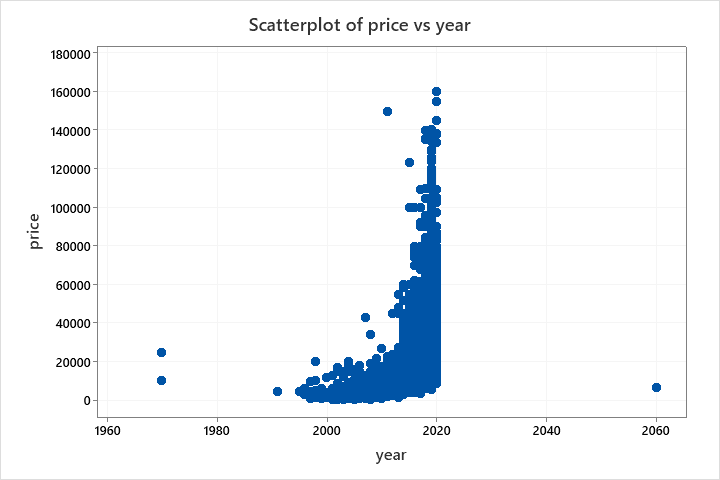
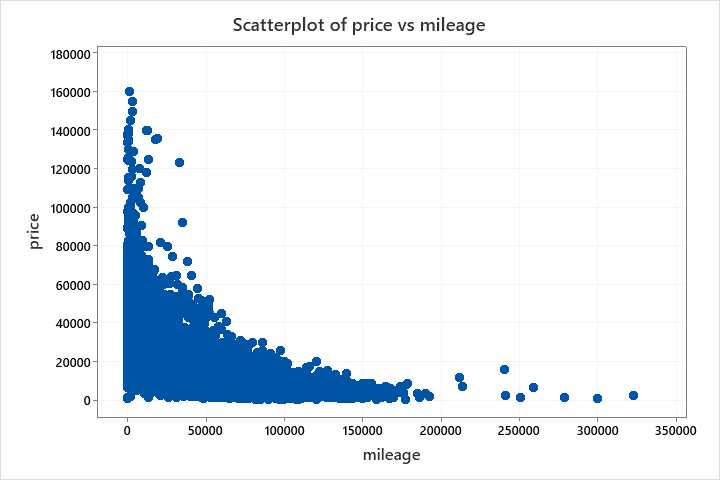
Fuel type refers to the type of fuel that the engine uses. It can be of 5 types, namely: Diesel, Petrol, Hybrid, Electric and Other.

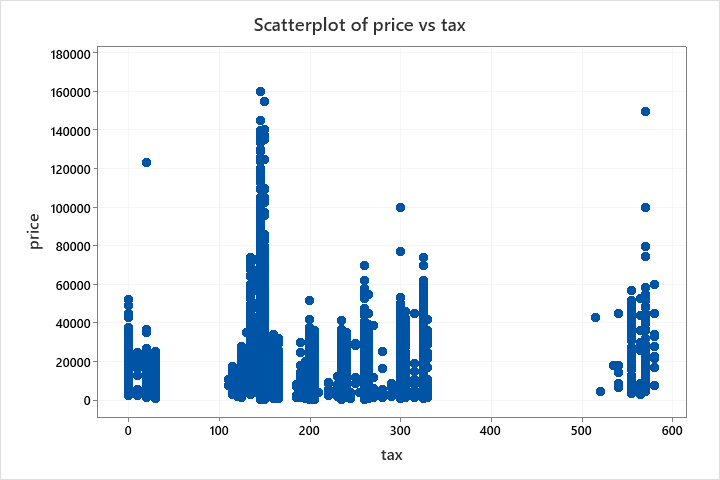
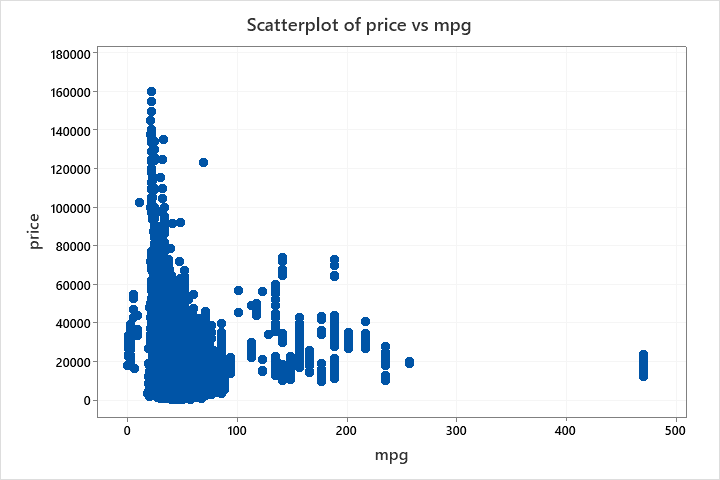
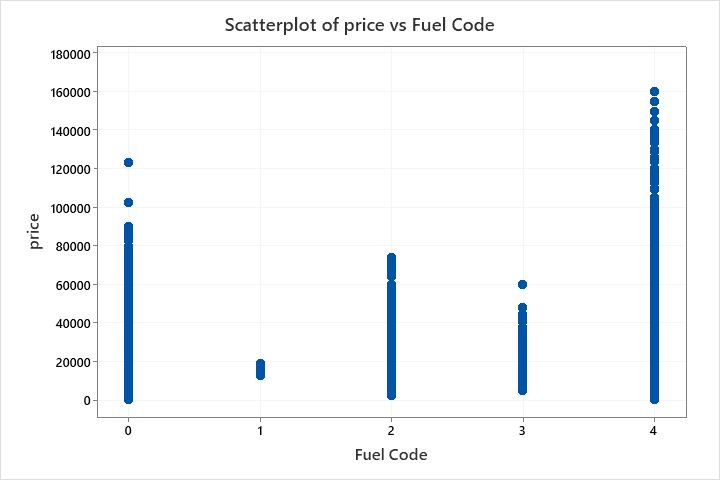
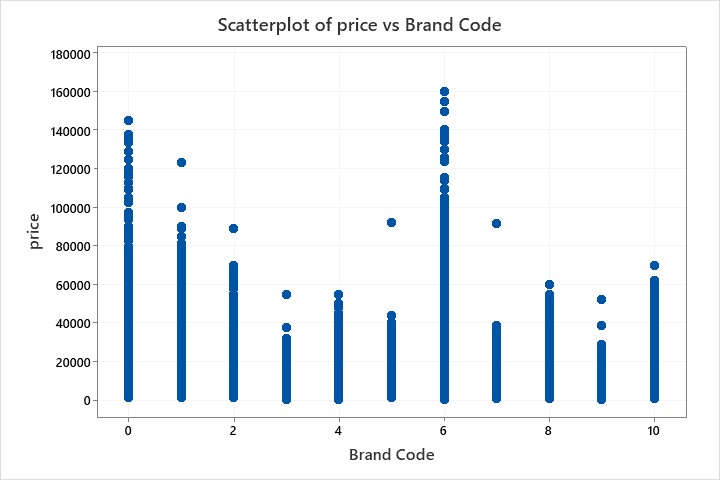
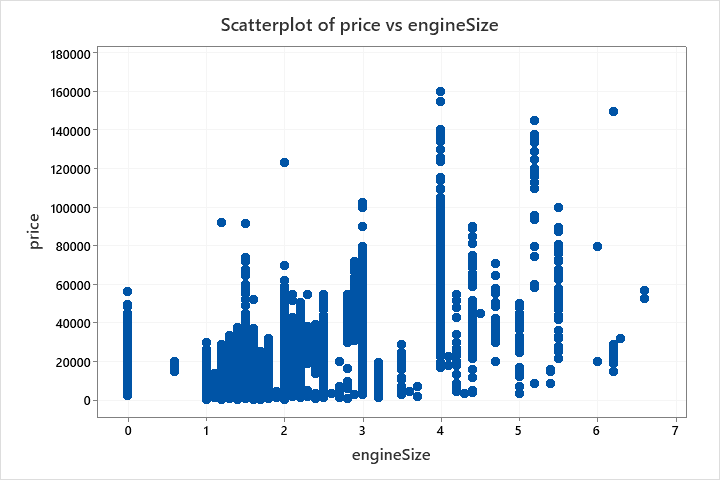
Transmission refers to the type of gearbox that the vehicle uses. There are 4 types of transmissions: Manual, Automatic, Semi-Automatic and Other.

The prices and features are divided into 11 car brands: Audi, BMW, CClass, Focus, Ford, Hyundai, Mercedes, Skoda, Toyota, Vauxhall and Volkswagen.

**Section 2: Exploratory Analysis and Descriptive Statistics**

Analysis of Scatter Plots:

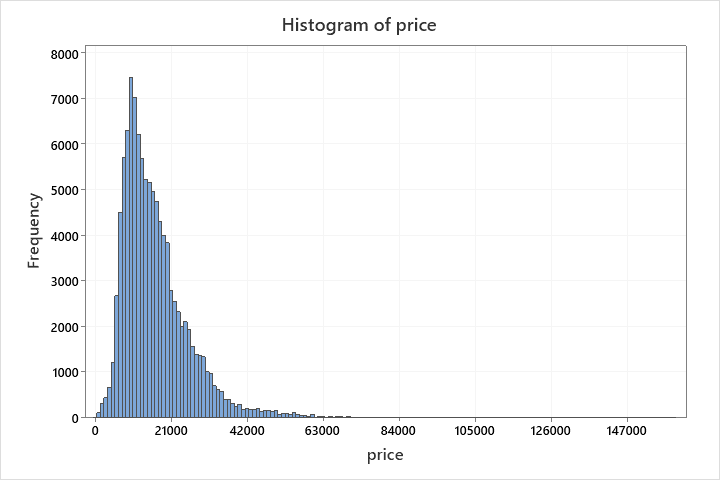
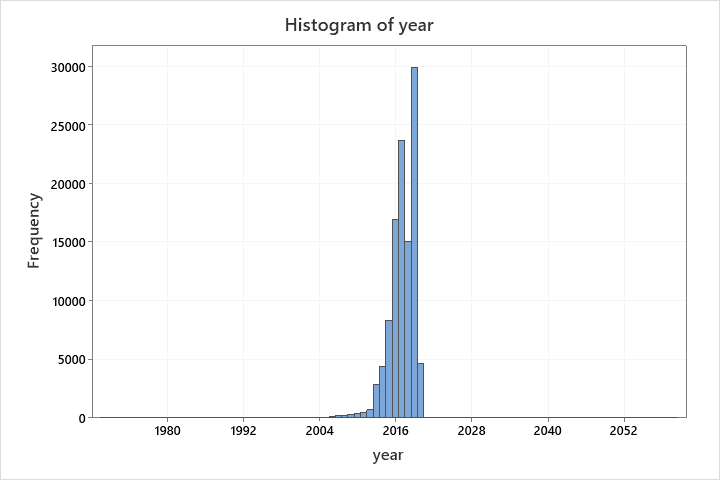
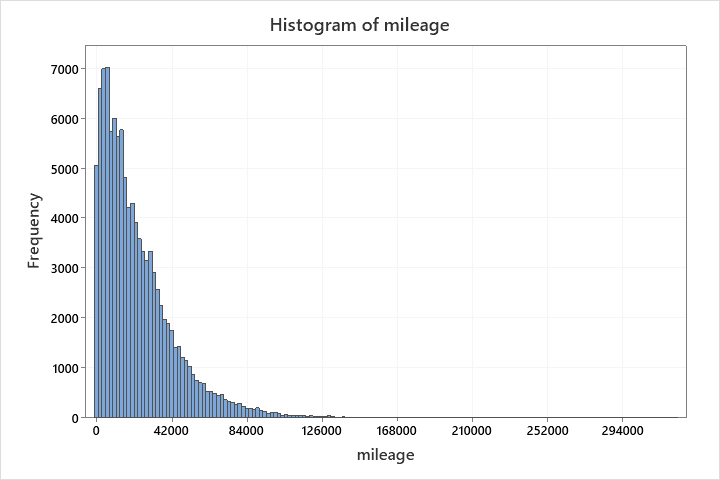
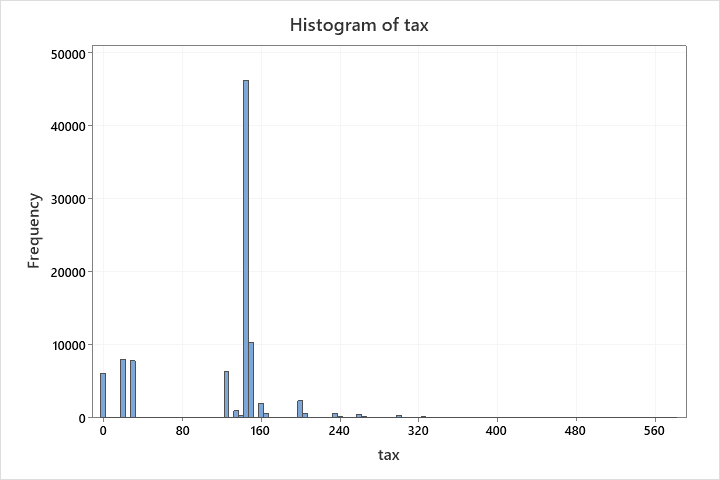
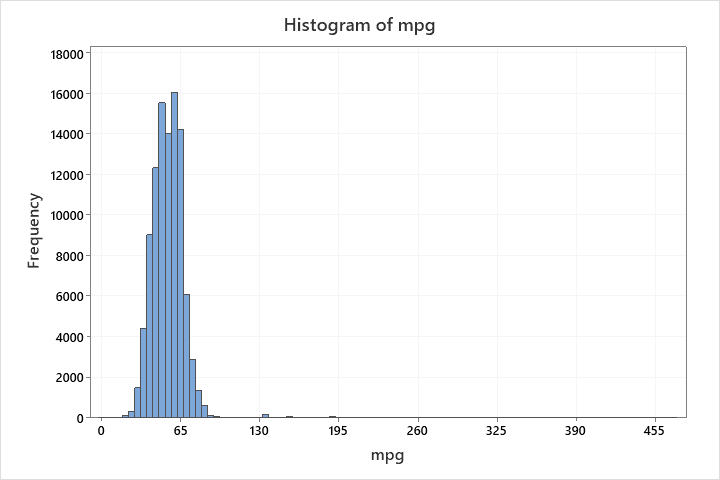
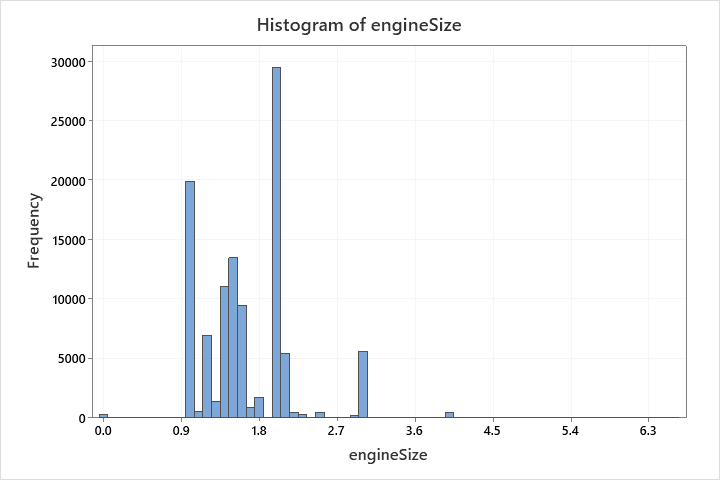
**Scatterplot of price vs year, mileage, tax, mpg, engineSize, Brand Code, Transmission Code, Fuel Code**

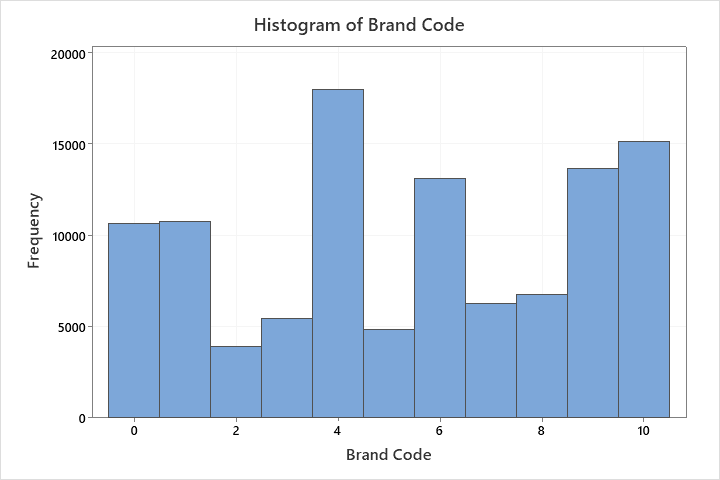
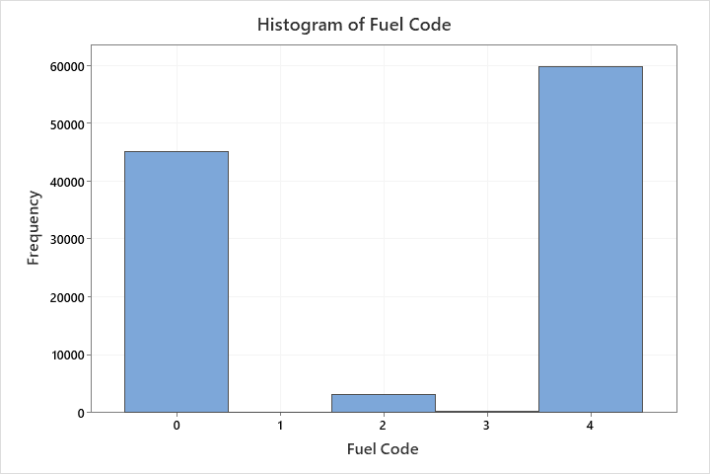
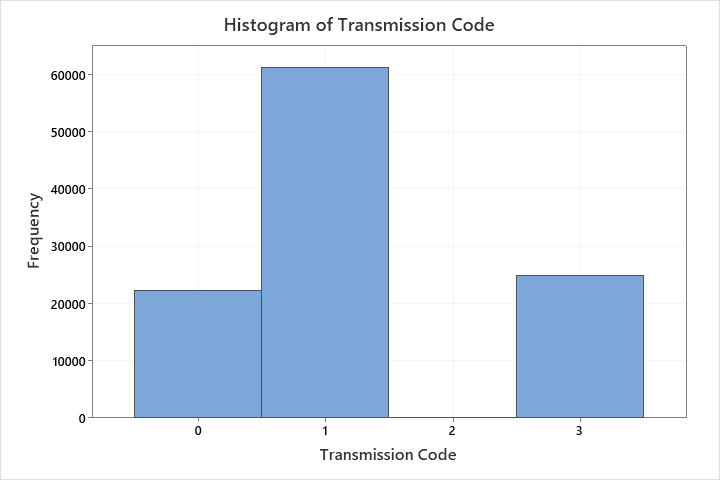


The following conclusions can be drawn from the scatter plots about the relationship between the predictors and the response variable:

1. Price and mileage: negative non-linear correlation
2. Price and mpg: negative correlation
3. Price and tax: inconclusive
4. Price and engineSize: positive correlation
5. Price and year: positive correlation (newer cars have higher resale price)

Skewness of the data:

**Histogram of price, year, mileage, tax, mpg, engineSize, Brand Code, Fuel Code, Transmission Code**

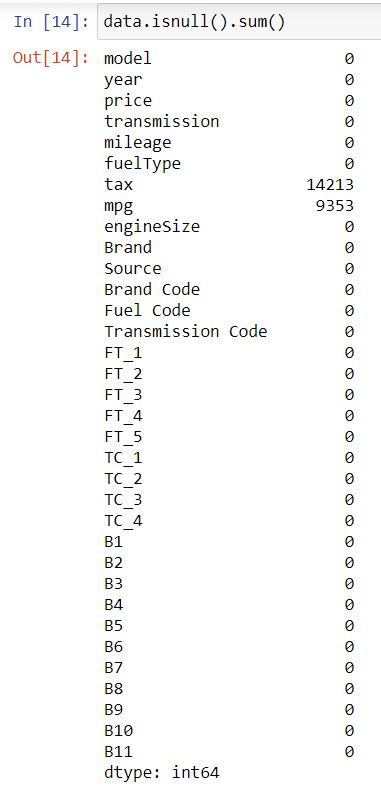
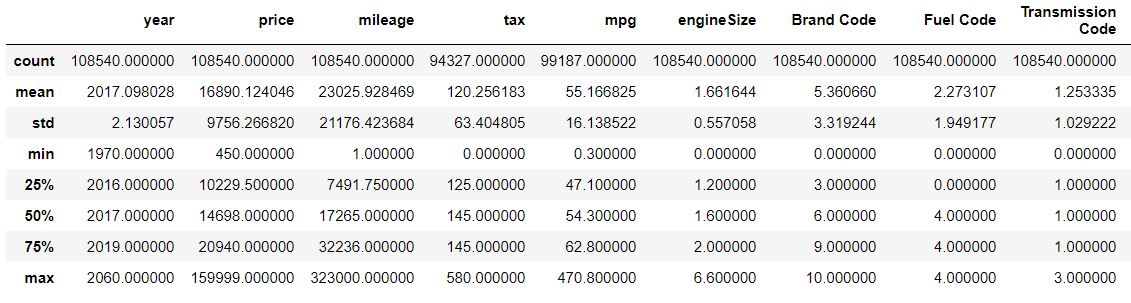
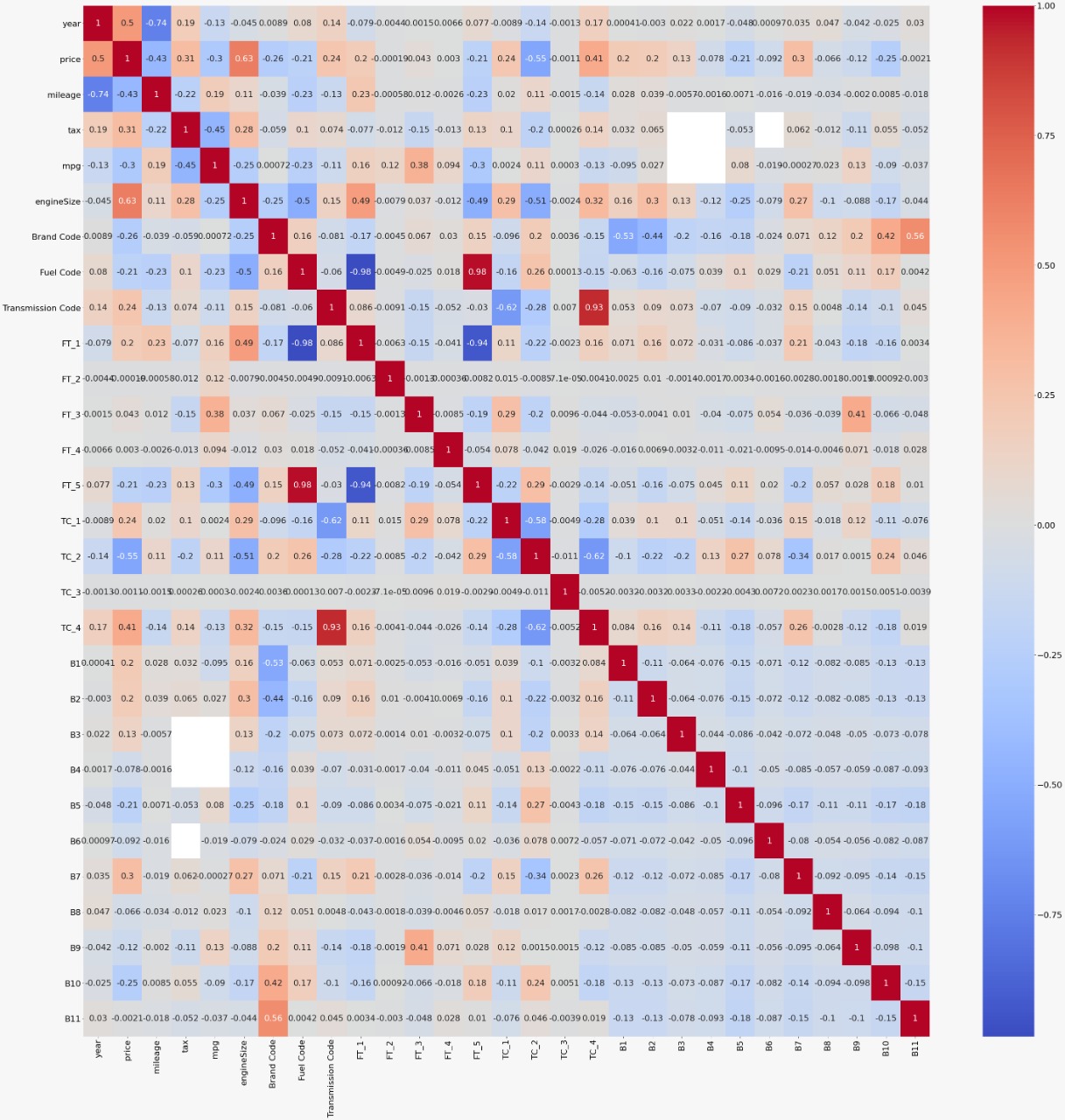
The year values are left skewed which is expected since recent data on the cars is more widely accessible as compared to older data. The prices and mileage values are right skewed. The skewness of the prices concurs with the fact that fewer people will be willing to pay a high price for a second-hand car. We can also see that there are comparatively very few observations for the transmission type “Other”. The prices and mileage values appear to be right skewed, whereas the values for year are left skewed. The mog values seem to have a normal distribution.

Outliers:

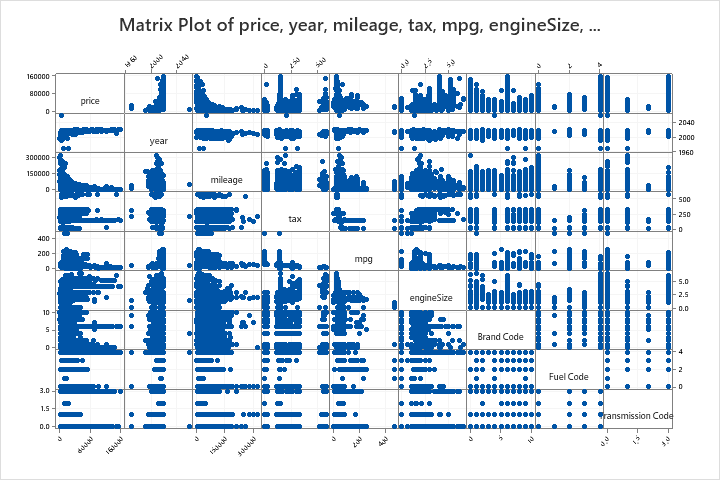
The following outliers were detected:

1. Cars with year = 2060 (Brand: Ford, Model:Fiesta), year = 1970 (Brand: Toyota, Model: Yaris), year=1971 (Brand: Mercedes, Model: MClass)
2. Cars with mpg = 470.8 (Brand: BMQ, Model: i3)

Heat Map:



**Matrix Plot of price, year, mileage, tax, mpg, engineSize, ...**



The correlation matrix and the heat map verify our conclusions about the relationships between the predictors and the price and price vs tax shows a positive correlation. The highest correlation is between price and engine size because cars with larger engines have better acceleration and hence better performance.

Based on the initial analysis of the data, it is assumed that we would require a multiple linear regression model with categorical predictors (for the transmission, fuel type and brand).

**Section 3: Model Building**

**Model 1:**

Upon fitting a multiple regression model with categorical predictors as follows:

**Regression Analysis: price versus year, mileage, tax, mpg, engineSize, Brand, transmission, fuelType**

**Method**

|  |  |
| --- | --- |
| Categorical predictor coding | (1, 0) |
| Rows unused | 9992 |

**Regression Equation**

|  |  |  |
| --- | --- | --- |
| price | = | -3121679 + 1552.5 year - 0.09225 mileage - 3.401 tax - 54.20 mpg + 9619.5 engineSize + 0.0 Brand\_Audi - 2313.9 Brand\_BMW - 3765.1 Brand\_Ford - 264.9 Brand\_Mercedes - 4480.2 Brand\_Skoda - 6345 Brand\_Toyota - 6775.7 Brand\_Vauxhall - 3010.2 Brand\_Volkswagen + 0.0 transmission\_Automatic - 1931.6 transmission\_Manual + 2236 transmission\_Other + 388.8 transmission\_Semi-Auto + 0.0 fuelType\_Diesel + 20954 fuelType\_Electric + 5018 fuelType\_Hybrid + 4872 fuelType\_Other + 1321.1 fuelType\_Petrol |

**Coefficients**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Term** | **Coef** | **SE Coef** | **T-Value** | **P-Value** | **VIF** |
| Constant | -3121679 | 27430 | -113.80 | 0.000 |  |
| year | 1552.5 | 13.6 | 114.31 | 0.000 | 2.33 |
| mileage | -0.09225 | 0.00139 | -66.54 | 0.000 | 2.46 |
| tax | -3.401 | 0.356 | -9.56 | 0.000 | 1.42 |
| mpg | -54.20 | 1.65 | -32.94 | 0.000 | 2.01 |
| engineSize | 9619.5 | 53.7 | 179.17 | 0.000 | 2.55 |
| Brand |  |  |  |  |  |
| BMW | -2313.9 | 80.8 | -28.64 | 0.000 | 1.85 |
| Ford | -3765.1 | 76.6 | -49.14 | 0.000 | 2.53 |
| Mercedes | -264.9 | 78.1 | -3.39 | 0.001 | 2.02 |
| Skoda | -4480.2 | 95.1 | -47.09 | 0.000 | 1.55 |
| Toyota | -6345 | 102 | -61.93 | 0.000 | 1.93 |
| Vauxhall | -6775.7 | 80.7 | -84.00 | 0.000 | 2.24 |
| Volkswagen | -3010.2 | 75.0 | -40.14 | 0.000 | 2.13 |
| transmission |  |  |  |  |  |
| Manual | -1931.6 | 60.5 | -31.93 | 0.000 | 2.51 |
| Other | 2236 | 2178 | 1.03 | 0.305 | 1.00 |
| Semi-Auto | 388.8 | 59.8 | 6.50 | 0.000 | 1.79 |
| fuelType |  |  |  |  |  |
| Electric | 20954 | 2465 | 8.50 | 0.000 | 1.03 |
| Hybrid | 5018 | 148 | 33.85 | 0.000 | 1.71 |
| Other | 4872 | 383 | 12.73 | 0.000 | 1.05 |
| Petrol | 1321.1 | 52.8 | 25.02 | 0.000 | 1.92 |

**Model Summary**

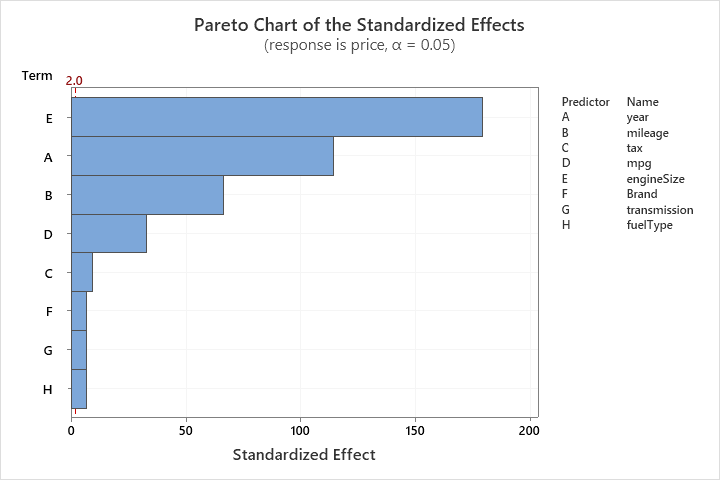
|  |  |  |  |
| --- | --- | --- | --- |
| **S** | **R-sq** | **R-sq(adj)** | **R-sq(pred)** |
| 4865.11 | 76.58% | 76.57% | 76.52% |

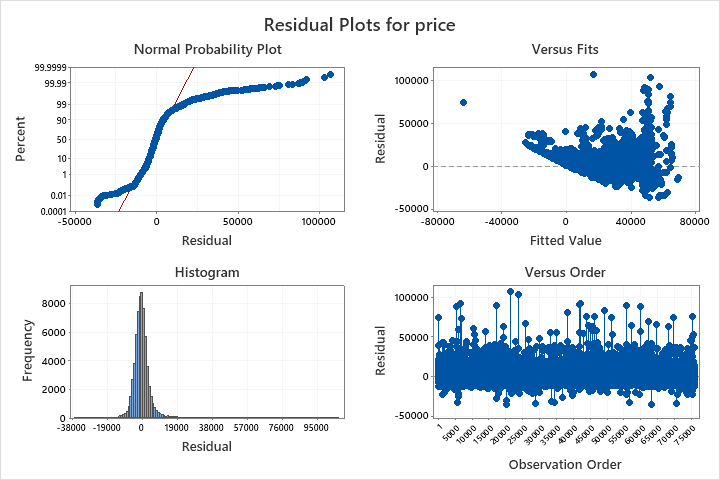
**Analysis of Variance**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Adj SS** | **Adj MS** | **F-Value** | **P-Value** |
| Regression | 19 | 5.10554E+12 | 2.68712E+11 | 11352.78 | 0.000 |
| year | 1 | 3.09272E+11 | 3.09272E+11 | 13066.39 | 0.000 |
| mileage | 1 | 1.04811E+11 | 1.04811E+11 | 4428.13 | 0.000 |
| tax | 1 | 2160972269 | 2160972269 | 91.30 | 0.000 |
| mpg | 1 | 25680578316 | 25680578316 | 1084.97 | 0.000 |
| engineSize | 1 | 7.59799E+11 | 7.59799E+11 | 32100.60 | 0.000 |
| Brand | 7 | 2.31748E+11 | 33106786343 | 1398.72 | 0.000 |
| transmission | 3 | 42028448893 | 14009482964 | 591.88 | 0.000 |
| fuelType | 4 | 44578332679 | 11144583170 | 470.85 | 0.000 |
| Error | 65966 | 1.56137E+12 | 23669304 |  |  |
| Lack-of-Fit | 64088 | 1.55625E+12 | 24283043 | 8.91 | 0.000 |
| Pure Error | 1878 | 5117649733 | 2725053 |  |  |
| Total | 65985 | 6.66691E+12 |  |  |  |

**Durbin-Watson Statistic**

|  |  |
| --- | --- |
| Durbin-Watson Statistic = | 2.00395 |





Our model has a clear lack of fit due to the p-value being less than the significance level (α=0.05). All the coefficients (except the indicator variable for the Mercedes brand) are significant as their p-value is less than the significance level (α=0.05). However, our model has a fairly high R2 value of 76.52% which suggests a decent predictive capability of our model.

However, our model assumptions are violated as our residuals are not normally distributed. Hence, we will perform the Brown-Forsythe test to check for equal variances.

**Test for Equal Variances: price versus year, transmission, mileage, fuelType, tax, mpg, engineSize, Brand**

**Method**

|  |  |
| --- | --- |
| Null hypothesis | All variances are equal |
| Alternative hypothesis | At least one variance is different |
| Significance level | α = 0.05 |

**Tests**

|  |  |  |
| --- | --- | --- |
| **Method** | **Test Statistic** | **P-Value** |
| Multiple comparisons | — | 0.000 |
| Levene | 3.59 | 0.000 |

*Samples are omitted from the tests if their standard deviations are 0 or missing.*

\* NOTE \* The graphical summary cannot be displayed because the multiple comparison intervals  
cannot be calculated.

As we can clearly see, we have a p-value less than the significance level (α=0.05). Hence, we reject the null hypothesis. Therefore, our residuals do not have constant variance.

Using best subset regression procedure for variable selection:

**Best Subsets Regression: price versus year, mileage, tax, mpg, engineSize, Brand Code, Fuel Code, Transmission Code**

**Response is price**

65986 cases used, 9992 cases contain missing values

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Vars** | **R-Sq** | **R-Sq (adj)** | **R-Sq (pred)** | **Mallows Cp** | **S** | **y e a r** | **m i l e a g e** | **t a x** | **m p g** | **e n g i n e S i z e** | **B r a n d  C o d e** | **F u e l  C o d e** | **T r a n s m i s s i o n  C o d e** |
| 1 | 40.9 | 40.9 | 40.9 | 73315.7 | 7729.6 |  |  |  |  | X |  |  |  |
| 1 | 24.4 | 24.4 | 24.4 | 112130.3 | 8740.5 | X |  |  |  |  |  |  |  |
| 2 | 67.8 | 67.8 | 67.8 | 9949.5 | 5706.8 | X |  |  |  | X |  |  |  |
| 2 | 65.0 | 65.0 | 65.0 | 16484.5 | 5947.4 |  | X |  |  | X |  |  |  |
| 3 | 70.2 | 70.2 | 70.2 | 4106.8 | 5482.8 | X | X |  |  | X |  |  |  |
| 3 | 68.7 | 68.7 | 68.7 | 7732.4 | 5622.9 | X |  |  |  | X | X |  |  |
| 4 | 71.2 | 71.2 | 71.2 | 1749.6 | 5389.8 | X | X |  |  | X | X |  |  |
| 4 | 70.6 | 70.6 | 70.6 | 3231.0 | 5448.4 | X | X |  |  | X |  |  | X |
| 5 | 71.6 | 71.6 | 71.6 | 1004.8 | 5360.1 | X | X |  |  | X | X |  | X |
| 5 | 71.5 | 71.5 | 71.5 | 1070.8 | 5362.7 | X | X |  |  | X | X | X |  |
| 6 | 71.9 | 71.9 | 71.8 | 323.5 | 5332.7 | X | X |  |  | X | X | X | X |
| 6 | 71.8 | 71.8 | 71.8 | 460.1 | 5338.2 | X | X |  | X | X | X |  | X |
| 7 | 71.9 | 71.9 | 71.9 | 127.5 | 5324.8 | X | X |  | X | X | X | X | X |
| 7 | 71.9 | 71.9 | 71.9 | 288.2 | 5331.2 | X | X | X |  | X | X | X | X |
| 8 | 72.0 | 72.0 | 72.0 | 9.0 | 5319.9 | X | X | X | X | X | X | X | X |

We can see that the last model with all the 8 predictors is the best because it has the highest R-sq, R-sq (pred) and R-sq (adj) value of 72.0 and the lowest Mallows Cp of 9.0.

We will try to improve this model by applying mathematical transformations where we see fit.

**Model 2:**

Now, we refit the model with transforming mileage to be 1/mileage based on the scatter plot.

**Regression Analysis: price versus year, 1/mileage, tax, mpg, engineSize, Brand, transmission, fuelType**

**Method**

|  |  |
| --- | --- |
| Categorical predictor coding | (1, 0) |
| Rows unused | 9992 |

**Regression Equation**

|  |  |  |
| --- | --- | --- |
| price | = | -4422711 + 2196.47 year + 8633 1/mileage - 1.757 tax - 60.37 mpg + 9410.5 engineSize + 0.0 Brand\_Audi - 2285.4 Brand\_BMW - 3539.9 Brand\_Ford - 49.0 Brand\_Mercedes - 4437.0 Brand\_Skoda - 6023 Brand\_Toyota - 6765.4 Brand\_Vauxhall - 2916.7 Brand\_Volkswagen + 0.0 transmission\_Automatic - 2106.1 transmission\_Manual + 2627 transmission\_Other + 511.7 transmission\_Semi-Auto + 0.0 fuelType\_Diesel + 23719 fuelType\_Electric + 5330 fuelType\_Hybrid + 5283 fuelType\_Other + 1894.9 fuelType\_Petrol |

**Coefficients**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Term** | **Coef** | **SE Coef** | **T-Value** | **P-Value** | **VIF** |
| Constant | -4422711 | 19637 | -225.23 | 0.000 |  |
| year | 2196.47 | 9.73 | 225.85 | 0.000 | 1.12 |
| 1/mileage | 8633 | 824 | 10.48 | 0.000 | 1.01 |
| tax | -1.757 | 0.367 | -4.79 | 0.000 | 1.41 |
| mpg | -60.37 | 1.70 | -35.61 | 0.000 | 2.00 |
| engineSize | 9410.5 | 55.3 | 170.11 | 0.000 | 2.54 |
| Brand |  |  |  |  |  |
| BMW | -2285.4 | 83.4 | -27.39 | 0.000 | 1.85 |
| Ford | -3539.9 | 79.0 | -44.80 | 0.000 | 2.53 |
| Mercedes | -49.0 | 80.5 | -0.61 | 0.543 | 2.01 |
| Skoda | -4437.0 | 98.2 | -45.18 | 0.000 | 1.55 |
| Toyota | -6023 | 106 | -57.02 | 0.000 | 1.93 |
| Vauxhall | -6765.4 | 83.3 | -81.22 | 0.000 | 2.24 |
| Volkswagen | -2916.7 | 77.4 | -37.68 | 0.000 | 2.13 |
| transmission |  |  |  |  |  |
| Manual | -2106.1 | 62.4 | -33.76 | 0.000 | 2.51 |
| Other | 2627 | 2248 | 1.17 | 0.242 | 1.00 |
| Semi-Auto | 511.7 | 61.7 | 8.29 | 0.000 | 1.78 |
| fuelType |  |  |  |  |  |
| Electric | 23719 | 2544 | 9.32 | 0.000 | 1.03 |
| Hybrid | 5330 | 153 | 34.85 | 0.000 | 1.71 |
| Other | 5283 | 395 | 13.37 | 0.000 | 1.05 |
| Petrol | 1894.9 | 53.8 | 35.23 | 0.000 | 1.87 |

**Model Summary**

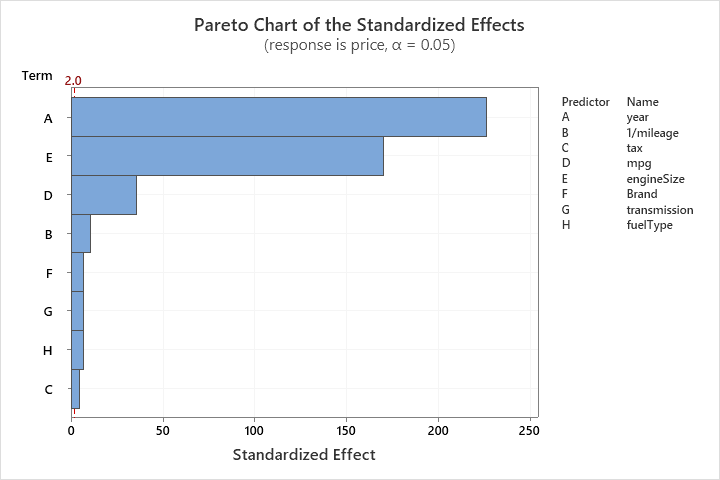
|  |  |  |  |
| --- | --- | --- | --- |
| **S** | **R-sq** | **R-sq(adj)** | **R-sq(pred)** |
| 5021.58 | 75.05% | 75.04% | 74.99% |

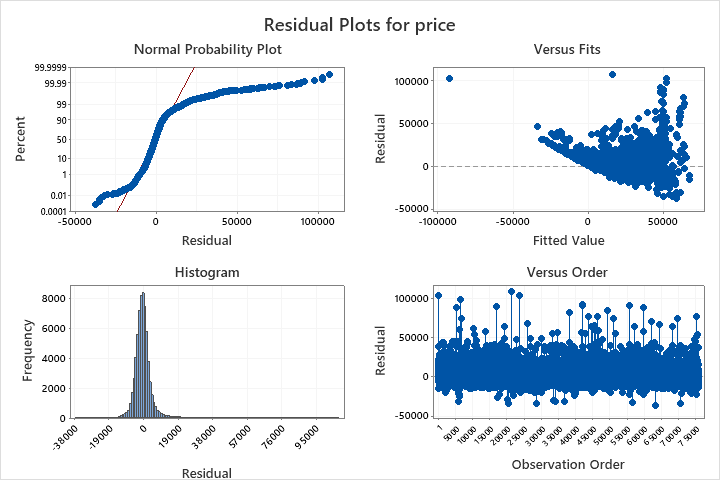
**Analysis of Variance**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Adj SS** | **Adj MS** | **F-Value** | **P-Value** |
| Regression | 19 | 5.00349E+12 | 2.63342E+11 | 10443.35 | 0.000 |
| year | 1 | 1.28619E+12 | 1.28619E+12 | 51006.49 | 0.000 |
| 1/mileage | 1 | 2767026015 | 2767026015 | 109.73 | 0.000 |
| tax | 1 | 579474041 | 579474041 | 22.98 | 0.000 |
| mpg | 1 | 31971440566 | 31971440566 | 1267.89 | 0.000 |
| engineSize | 1 | 7.29696E+11 | 7.29696E+11 | 28937.56 | 0.000 |
| Brand | 7 | 2.33727E+11 | 33389527405 | 1324.13 | 0.000 |
| transmission | 3 | 52778900815 | 17592966938 | 697.68 | 0.000 |
| fuelType | 4 | 65769344761 | 16442336190 | 652.05 | 0.000 |
| Error | 65966 | 1.66341E+12 | 25216216 |  |  |
| Lack-of-Fit | 64088 | 1.65830E+12 | 25875285 | 9.50 | 0.000 |
| Pure Error | 1878 | 5117649733 | 2725053 |  |  |
| Total | 65985 | 6.66691E+12 |  |  |  |

**Durbin-Watson Statistic**

|  |  |
| --- | --- |
| Durbin-Watson Statistic = | 2.00186 |





From the above results we can see an improvement in the f-value for the lack-of-fit test, indicating that our model is better fitted now. However, we still have a p-value less than the significance level (α=0.05). Hence, our model is still not well fitted.

We see a very slight change in our R2 value, as it decremented by a very small amount. This is however cancelled out by our better fitting; hence we can claim a good predictive capability for our new model.

The Mercedes indicator variable and the “other” transmission indicator variable shows us a high p-value, which suggests that these predictors are not significant to the model.

Our residuals are yet again not normally distributed which violates our model assumptions. We also notice various outliers in our plots, which affect the model adversely.

Hence, we will try to improve the model by removing these outliers and applying a Box-Cox transformation with an optimal value of λ.

**Model 3:**

The outliers removed include a data value with a year of 2060 (which does not make sense in our context), two data values with year 1970 (which were far behind most of our other values) and an unusually large mpg value of 470.8 (which is physically impossible as the maximum possible mpg value recorded in human history is 14.573 mpg). We have removed these outliers to improve our model. We have also applied a Box-Cox transformation with an optimal value of λ, generated by Minitab.

After removing the outliers and applying the Box-Cox transformation, the following model was obtained.

**Regression Analysis: price versus year, 1/mileage, tax, mpg, engineSize, Brand, transmission, fuelType**

**Method**

|  |  |
| --- | --- |
| Categorical predictor coding | (1, 0) |
| Rows unused | 9992 |
|  |  |
| Box-Cox transformation |  |
| Rounded λ | 0.0411146 |
| Estimated λ | 0.0411146 |
| 95% CI for λ | (0.0356146, 0.0466146) |

**Regression Equation**

|  |  |  |
| --- | --- | --- |
| price^0.0411146 | = | -16.5840 + 0.008959 year + 0.01325 1/mileage + 0.000012 tax - 0.000351 mpg + 0.019624 engineSize + 0.0 Brand\_Audi - 0.006508 Brand\_BMW - 0.015301 Brand\_Ford - 0.000656 Brand\_Mercedes - 0.017713 Brand\_Skoda - 0.025501 Brand\_Toyota - 0.028615 Brand\_Vauxhall - 0.010362 Brand\_Volkswagen + 0.0 transmission\_Automatic - 0.008073 transmission\_Manual + 0.00958 transmission\_Other + 0.000861 transmission\_Semi-Auto + 0.0 fuelType\_Diesel + 0.05141 fuelType\_Electric + 0.024471 fuelType\_Hybrid + 0.012152 fuelType\_Other - 0.003001 fuelType\_Petrol |

**Coefficients for Transformed Response**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Term** | **Coef** | **SE Coef** | **T-Value** | **P-Value** | **VIF** |
| Constant | -16.5840 | 0.0462 | -358.80 | 0.000 |  |
| year | 0.008959 | 0.000023 | 391.49 | 0.000 | 1.13 |
| 1/mileage | 0.01325 | 0.00193 | 6.87 | 0.000 | 1.02 |
| tax | 0.000012 | 0.000001 | 14.06 | 0.000 | 1.49 |
| mpg | -0.000351 | 0.000005 | -67.68 | 0.000 | 2.40 |
| engineSize | 0.019624 | 0.000132 | 148.65 | 0.000 | 2.63 |
| Brand |  |  |  |  |  |
| BMW | -0.006508 | 0.000195 | -33.30 | 0.000 | 1.85 |
| Ford | -0.015301 | 0.000185 | -82.71 | 0.000 | 2.53 |
| Mercedes | -0.000656 | 0.000189 | -3.47 | 0.001 | 2.03 |
| Skoda | -0.017713 | 0.000230 | -77.03 | 0.000 | 1.55 |
| Toyota | -0.025501 | 0.000247 | -103.15 | 0.000 | 1.93 |
| Vauxhall | -0.028615 | 0.000195 | -146.81 | 0.000 | 2.24 |
| Volkswagen | -0.010362 | 0.000181 | -57.22 | 0.000 | 2.13 |
| transmission |  |  |  |  |  |
| Manual | -0.008073 | 0.000147 | -54.92 | 0.000 | 2.54 |
| Other | 0.00958 | 0.00526 | 1.82 | 0.068 | 1.00 |
| Semi-Auto | 0.000861 | 0.000144 | 5.96 | 0.000 | 1.79 |
| fuelType |  |  |  |  |  |
| Electric | 0.05141 | 0.00832 | 6.18 | 0.000 | 1.00 |
| Hybrid | 0.024471 | 0.000365 | 67.05 | 0.000 | 1.76 |
| Other | 0.012152 | 0.000938 | 12.95 | 0.000 | 1.04 |
| Petrol | -0.003001 | 0.000131 | -22.84 | 0.000 | 2.04 |

**Model Summary for Transformed Response**

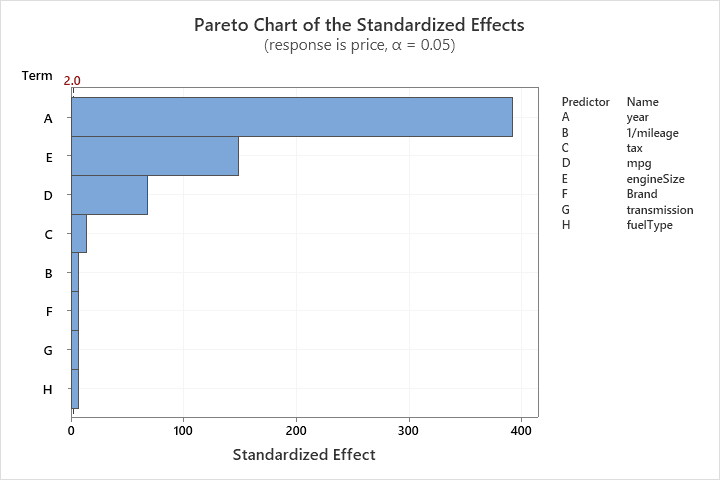
|  |  |  |  |
| --- | --- | --- | --- |
| **S** | **R-sq** | **R-sq(adj)** | **R-sq(pred)** |
| 0.0117473 | 87.09% | 87.08% | 87.06% |

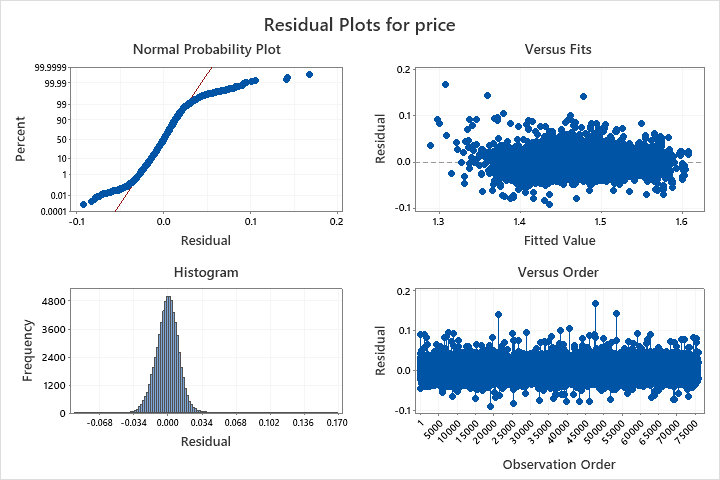
**Analysis of Variance for Transformed Response**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Adj SS** | **Adj MS** | **F-Value** | **P-Value** |
| Regression | 19 | 61.3584 | 3.2294 | 23401.38 | 0.000 |
| year | 1 | 21.1503 | 21.1503 | 153263.43 | 0.000 |
| 1/mileage | 1 | 0.0065 | 0.0065 | 47.24 | 0.000 |
| tax | 1 | 0.0273 | 0.0273 | 197.65 | 0.000 |
| mpg | 1 | 0.6322 | 0.6322 | 4581.20 | 0.000 |
| engineSize | 1 | 3.0495 | 3.0495 | 22098.12 | 0.000 |
| Brand | 7 | 4.1465 | 0.5924 | 4292.47 | 0.000 |
| transmission | 3 | 0.6530 | 0.2177 | 1577.24 | 0.000 |
| fuelType | 4 | 0.6758 | 0.1690 | 1224.34 | 0.000 |
| Error | 65935 | 9.0990 | 0.0001 |  |  |
| Lack-of-Fit | 64057 | 9.0659 | 0.0001 | 8.02 | 0.000 |
| Pure Error | 1878 | 0.0332 | 0.0000 |  |  |
| Total | 65954 | 70.4574 |  |  |  |

**Durbin-Watson Statistic for Transformed Response**

|  |  |
| --- | --- |
| Durbin-Watson Statistic = | 2.00490 |





From the above results, we can conclude that our newly fitted model has a lower f-value, indicating a worse lack of fit. However, our R2 value spiked by a large amount of nearly 10%. This could indicate a better predictive capability in our new model.

Our newly fitted model shows us that all predictors are significant except the indicator variable for “other” transmissions, as it has a p-value larger than our significance level (α=0.05). The residuals show better normal distribution, but this is still not adequate.

We will try to improve our model by adding interaction terms without applying a Box-Cox transformation, since the Box-Cox transformation affected our lack-of-fit test. We will also remove all the rows that correspond to “other” transmission modes as it shows us a large level of insignificance.

**Model 4:**

We removed all the rows in our data set corresponding to “other” transmission modes (around 20 rows) as it does not affect our overall sample size.

Also, we selected our interaction term considering the fact that the engine size of a car affects the mpg, and hence the interaction term engineSize\*mpg is added to the model and the results are as follows:

**Regression Analysis: price versus year, 1/mileage, tax, mpg, engineSize, mpg\*engineSize, Brand, transmission, fuelType**

**Method**

|  |  |
| --- | --- |
| Categorical predictor coding | (1, 0) |
| Rows unused | 9989 |

**Regression Equation**

|  |  |  |
| --- | --- | --- |
| price | = | -4374513 + 2168.72 year + 8117 1/mileage - 6.521 tax + 221.41 mpg + 17792 engineSize - 220.63 mpg\*engineSize + 0.0 Brand\_Audi - 1554.3 Brand\_BMW - 4217.8 Brand\_Ford + 596.2 Brand\_Mercedes - 4711.5 Brand\_Skoda - 6929.4 Brand\_Toyota - 6674.2 Brand\_Vauxhall - 3234.7 Brand\_Volkswagen + 0.0 transmission\_Automatic - 1911.0 transmission\_Manual + 724.3 transmission\_Semi-Auto + 0.0 fuelType\_Diesel + 10843 fuelType\_Electric + 8137 fuelType\_Hybrid + 4869 fuelType\_Other - 784.3 fuelType\_Petrol |

**Coefficients**

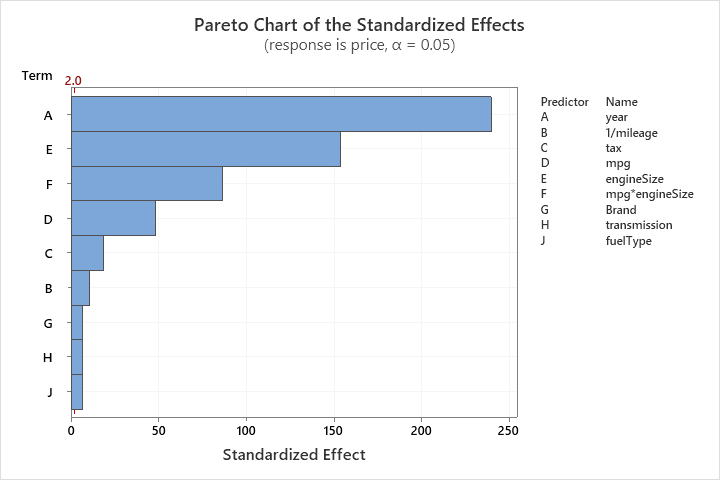
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Term** | **Coef** | **SE Coef** | **T-Value** | **P-Value** | **VIF** |
| Constant | -4374513 | 18288 | -239.20 | 0.000 |  |
| year | 2168.72 | 9.06 | 239.47 | 0.000 | 1.13 |
| 1/mileage | 8117 | 763 | 10.64 | 0.000 | 1.02 |
| tax | -6.521 | 0.349 | -18.69 | 0.000 | 1.49 |
| mpg | 221.41 | 4.58 | 48.31 | 0.000 | 12.02 |
| engineSize | 17792 | 116 | 153.86 | 0.000 | 12.89 |
| mpg\*engineSize | -220.63 | 2.55 | -86.49 | 0.000 | 20.47 |
| Brand |  |  |  |  |  |
| BMW | -1554.3 | 77.6 | -20.04 | 0.000 | 1.86 |
| Ford | -4217.8 | 73.8 | -57.12 | 0.000 | 2.58 |
| Mercedes | 596.2 | 74.9 | 7.96 | 0.000 | 2.03 |
| Skoda | -4711.5 | 91.2 | -51.69 | 0.000 | 1.56 |
| Toyota | -6929.4 | 98.6 | -70.31 | 0.000 | 1.96 |
| Vauxhall | -6674.2 | 77.1 | -86.51 | 0.000 | 2.24 |
| Volkswagen | -3234.7 | 71.7 | -45.10 | 0.000 | 2.13 |
| transmission |  |  |  |  |  |
| Manual | -1911.0 | 58.2 | -32.84 | 0.000 | 2.55 |
| Semi-Auto | 724.3 | 57.2 | 12.67 | 0.000 | 1.79 |
| fuelType |  |  |  |  |  |
| Electric | 10843 | 3292 | 3.29 | 0.001 | 1.00 |
| Hybrid | 8137 | 145 | 56.04 | 0.000 | 1.78 |
| Other | 4869 | 372 | 13.08 | 0.000 | 1.04 |
| Petrol | -784.3 | 56.2 | -13.95 | 0.000 | 2.39 |

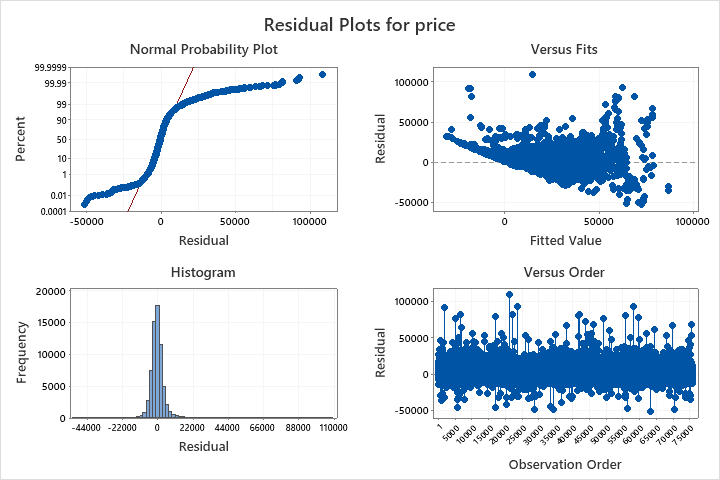
**Model Summary**

|  |  |  |  |
| --- | --- | --- | --- |
| **S** | **R-sq** | **R-sq(adj)** | **R-sq(pred)** |
| 4647.38 | 78.64% | 78.63% | 78.58% |

**Analysis of Variance**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Adj SS** | **Adj MS** | **F-Value** | **P-Value** |
| Regression | 19 | 5.24255E+12 | 2.75924E+11 | 12775.33 | 0.000 |
| year | 1 | 1.23859E+12 | 1.23859E+12 | 57346.96 | 0.000 |
| 1/mileage | 1 | 2445565712 | 2445565712 | 113.23 | 0.000 |
| tax | 1 | 7542568366 | 7542568366 | 349.22 | 0.000 |
| mpg | 1 | 50398734511 | 50398734511 | 2333.47 | 0.000 |
| engineSize | 1 | 5.11282E+11 | 5.11282E+11 | 23672.46 | 0.000 |
| mpg\*engineSize | 1 | 1.61550E+11 | 1.61550E+11 | 7479.82 | 0.000 |
| Brand | 7 | 2.67020E+11 | 38145656630 | 1766.15 | 0.000 |
| transmission | 2 | 50328848892 | 25164424446 | 1165.12 | 0.000 |
| fuelType | 4 | 70508283428 | 17627070857 | 816.14 | 0.000 |
| Error | 65930 | 1.42397E+12 | 21598170 |  |  |
| Lack-of-Fit | 36075 | 1.36621E+12 | 37871302 | 19.57 | 0.000 |
| Pure Error | 29855 | 57760111961 | 1934688 |  |  |
| Total | 65949 | 6.66652E+12 |  |  |  |





The above results confirm that although our new model has lack of fit, the f-value has increased. However, it is still not enough to attain a proper fitting of the data. The R2 (pred) value is at a level of 78.58%, which shows us that our model has a good predictive capability.

This time, we see that all our predictors are significant to the model as they all have a low p-value less than our significance level (α=0.05). However, the residuals are skewed and not randomly scattered in our model.

We will apply a Box-Cox transformation to our model to see its impact as a final measure to test for improvements.

**Model 5:**

Applying the Box-Cox transformation to Model 4 we get:

**Regression Analysis: price versus year, 1/mileage, tax, mpg, engineSize, mpg\*engineSize, Brand, transmission, fuelType**

**Method**

|  |  |
| --- | --- |
| Categorical predictor coding | (1, 0) |
| Rows unused | 9992 |
|  |  |
| Box-Cox transformation |  |
| Rounded λ | 0.0528665 |
| Estimated λ | 0.0528665 |
| 95% CI for λ | (0.0463665, 0.0593665) |

**Regression Equation**

|  |  |  |
| --- | --- | --- |
| price^0.0528665 | = | -24.2686 + 0.012853 year + 0.01958 1/mileage + 0.000018 tax - 0.000281 mpg + 0.034114 engineSize - 0.000141 mpg\*engineSize + 0.0 Brand\_Audi - 0.009039 Brand\_BMW - 0.022545 Brand\_Ford - 0.000715 Brand\_Mercedes - 0.025788 Brand\_Skoda - 0.037329 Brand\_Toyota - 0.041034 Brand\_Vauxhall - 0.015116 Brand\_Volkswagen + 0.0 transmission\_Automatic - 0.011726 transmission\_Manual + 0.01280 transmission\_Other + 0.001338 transmission\_Semi-Auto + 0.0 fuelType\_Diesel + 0.0716 fuelType\_Electric + 0.036116 fuelType\_Hybrid + 0.01737 fuelType\_Other - 0.005435 fuelType\_Petrol |

**Coefficients for Transformed Response**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Term** | **Coef** | **SE Coef** | **T-Value** | **P-Value** | **VIF** |
| Constant | -24.2686 | 0.0664 | -365.35 | 0.000 |  |
| year | 0.012853 | 0.000033 | 390.74 | 0.000 | 1.13 |
| 1/mileage | 0.01958 | 0.00277 | 7.07 | 0.000 | 1.02 |
| tax | 0.000018 | 0.000001 | 13.81 | 0.000 | 1.49 |
| mpg | -0.000281 | 0.000017 | -16.85 | 0.000 | 12.02 |
| engineSize | 0.034114 | 0.000420 | 81.22 | 0.000 | 12.89 |
| mpg\*engineSize | -0.000141 | 0.000009 | -15.27 | 0.000 | 20.47 |
| Brand |  |  |  |  |  |
| BMW | -0.009039 | 0.000282 | -32.09 | 0.000 | 1.86 |
| Ford | -0.022545 | 0.000268 | -84.05 | 0.000 | 2.58 |
| Mercedes | -0.000715 | 0.000272 | -2.63 | 0.009 | 2.03 |
| Skoda | -0.025788 | 0.000331 | -77.89 | 0.000 | 1.56 |
| Toyota | -0.037329 | 0.000358 | -104.29 | 0.000 | 1.96 |
| Vauxhall | -0.041034 | 0.000280 | -146.44 | 0.000 | 2.24 |
| Volkswagen | -0.015116 | 0.000260 | -58.03 | 0.000 | 2.13 |
| transmission |  |  |  |  |  |
| Manual | -0.011726 | 0.000211 | -55.47 | 0.000 | 2.55 |
| Semi-Auto | 0.001338 | 0.000208 | 6.44 | 0.000 | 1.79 |
| fuelType |  |  |  |  |  |
| Electric | 0.0716 | 0.0120 | 5.98 | 0.000 | 1.00 |
| Hybrid | 0.036116 | 0.000527 | 68.48 | 0.000 | 1.78 |
| Other | 0.01737 | 0.00135 | 12.88 | 0.000 | 1.04 |
| Petrol | -0.005435 | 0.000204 | -26.61 | 0.000 | 2.39 |

**Model Summary for Transformed Response**

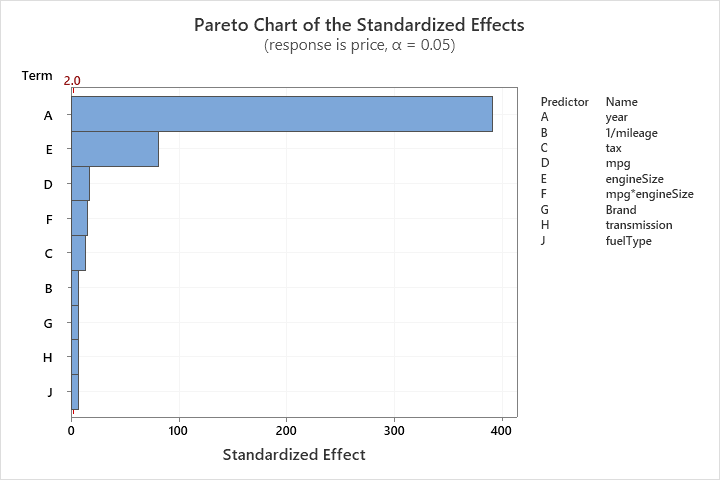
|  |  |  |  |
| --- | --- | --- | --- |
| **S** | **R-sq** | **R-sq(adj)** | **R-sq(pred)** |
| 0.0168811 | 87.13% | 87.13% | 87.10% |

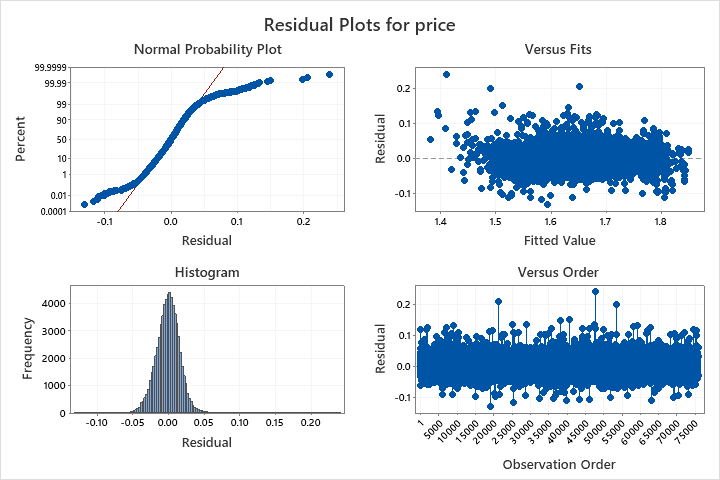
**Analysis of Variance for Transformed Response**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Adj SS** | **Adj MS** | **F-Value** | **P-Value** |
| Regression | 20 | 127.214 | 6.3607 | 22320.46 | 0.000 |
| year | 1 | 43.509 | 43.5086 | 152677.47 | 0.000 |
| 1/mileage | 1 | 0.014 | 0.0142 | 49.93 | 0.000 |
| tax | 1 | 0.054 | 0.0543 | 190.64 | 0.000 |
| mpg | 1 | 0.081 | 0.0809 | 283.90 | 0.000 |
| engineSize | 1 | 1.880 | 1.8797 | 6596.20 | 0.000 |
| mpg\*engineSize | 1 | 0.066 | 0.0664 | 233.14 | 0.000 |
| Brand | 7 | 8.628 | 1.2325 | 4325.02 | 0.000 |
| transmission | 3 | 1.385 | 0.4615 | 1619.53 | 0.000 |
| fuelType | 4 | 1.466 | 0.3665 | 1286.26 | 0.000 |
| Error | 65934 | 18.789 | 0.0003 |  |  |
| Lack-of-Fit | 64056 | 18.720 | 0.0003 | 7.94 | 0.000 |
| Pure Error | 1878 | 0.069 | 0.0000 |  |  |
| Total | 65954 | 146.003 |  |  |  |

**Durbin-Watson Statistic for Transformed Response**

|  |  |
| --- | --- |
| Durbin-Watson Statistic = | 2.00461 |





In our final model, we obtain a significantly worse f-value for our lack-of-fit measure. This would imply that our model does not fit well. However, we obtain a large R2 value of 87.10%.

The predictors are all significant yet again in our model, as per our previous model. The normal probability plot of residuals has improved with heavy tails and the residuals show random scattering unlike the previous model.

**One-way ANOVA:**

1. **Fuel Type and Price**

Using Linear Regression Model:

**Regression Equation**

|  |  |  |
| --- | --- | --- |
| price | = | 19206.1 FT\_1 + 14487 FT\_2 + 19207 FT\_3 + 17501 FT\_4 + 15020.5 FT\_5 |

The mean values for FT\_1, FT\_2, FT\_3, FT\_4, and FT\_5 are 19206.1, 14487, 19207, 17501 and 15020.5 respectively.

**Regression Equation**

|  |  |  |
| --- | --- | --- |
| price | = | 17084 + 2122 FT\_1\_1 - 2597 FT\_2\_1 + 2123 FT\_3\_1 + 417 FT\_4\_1 |

The following recoding was applied: FT\_1 = FT\_1\_1; FT\_2 = FT\_2\_1; FT\_3 = FT\_3\_1; FT\_4 = FT\_4\_1; FT\_5 = FT\_5\_1.

FT\_1, FT\_3, and FT\_4 have a positive effect on the price, i.e., they are effective in increasing the price of a car by 2122, 2123 and 417 times respectively, over the grand population mean. The level effect of FT\_5 is calculated as: effect = - (2122 – 2597 + 2123 + 417) = -2065.

FT\_2 and FT\_5 are effective in decreasing the price of a car by 2597 and 2065 times respectively, over the grand mean = 17084.

Comparisons:

**Method**

|  |  |
| --- | --- |
| Null hypothesis | All means are equal |
| Alternative hypothesis | Not all means are equal |
| Significance level | α = 0.05 |

*Equal variances were assumed for the analysis.*

**Means**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **fuelCode** | **N** | **Mean** | **StDev** | **95% CI** |
| 0 | 31649 | 19206.1 | 9918.4 | (19100.4, 19311.8) |
| 1 | 2 | 14487 | 2104 | (1190, 27784) |
| 2 | 2217 | 19207 | 7543 | (18808, 19607) |
| 3 | 166 | 17501 | 7774 | (16042, 18961) |
| 4 | 41905 | 15020.5 | 9448.0 | (14928.6, 15112.3) |

*Pooled StDev = 9594.44*

**Tukey Pairwise Comparisons**

**Grouping Information Using the Tukey Method and 95% Confidence**

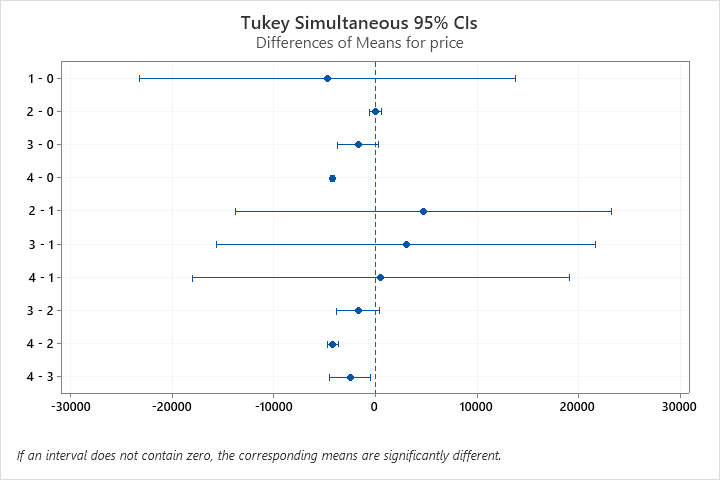
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **fuelCode** | **N** | **Mean** | **Grouping** | |
| 2 | 2217 | 19207 | A |  |
| 0 | 31649 | 19206.1 | A |  |
| 3 | 166 | 17501 | A |  |
| 4 | 41905 | 15020.5 |  | B |
| 1 | 2 | 14487 | A | B |

*Means that do not share a letter are significantly different.*

**Tukey Simultaneous Tests for Differences of Means**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Difference of Levels** | **Difference of Means** | **SE of Difference** | **95% CI** | **T-Value** | **Adjusted P-Value** |
| 1 - 0 | -4719 | 6785 | (-23237, 13799) | -0.70 | 0.958 |
| 2 - 0 | 1 | 211 | (-574, 577) | 0.01 | 1.000 |
| 3 - 0 | -1705 | 747 | (-3743, 333) | -2.28 | 0.150 |
| 4 - 0 | -4185.6 | 71.5 | (-4380.6, -3990.6) | -58.58 | 0.000 |
| 2 - 1 | 4720 | 6787 | (-13805, 23246) | 0.70 | 0.958 |
| 3 - 1 | 3014 | 6825 | (-15614, 21643) | 0.44 | 0.992 |
| 4 - 1 | 533 | 6784 | (-17984, 19051) | 0.08 | 1.000 |
| 3 - 2 | -1706 | 772 | (-3813, 401) | -2.21 | 0.176 |
| 4 - 2 | -4187 | 209 | (-4758, -3616) | -20.02 | 0.000 |
| 4 - 3 | -2481 | 746 | (-4517, -444) | -3.32 | 0.008 |

*Individual confidence level = 99.37%*



Based on the Tukey’s method, there are 5\*4/2= 10 interactions. From the table and interval plots we can see that only fuel codes 4 – 0, 4 – 2, and 4 – 3 have significantly different means.

**Fisher Pairwise Comparisons**

**Grouping Information Using the Fisher LSD Method and 95% Confidence**

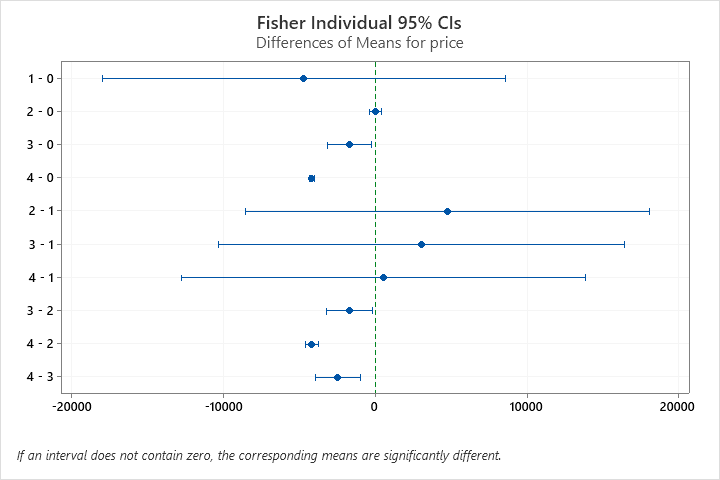
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **fuelCode** | **N** | **Mean** | **Grouping** | | |
| 2 | 2217 | 19207 | A |  |  |
| 0 | 31649 | 19206.1 | A |  |  |
| 3 | 166 | 17501 |  | B |  |
| 4 | 41905 | 15020.5 |  |  | C |
| 1 | 2 | 14487 | A | B | C |

*Means that do not share a letter are significantly different.*

**Fisher Individual Tests for Differences of Means**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Difference of Levels** | **Difference of Means** | **SE of Difference** | **95% CI** | **T-Value** | **Adjusted P-Value** |
| 1 - 0 | -4719 | 6785 | (-18017, 8579) | -0.70 | 0.487 |
| 2 - 0 | 1 | 211 | (-412, 415) | 0.01 | 0.995 |
| 3 - 0 | -1705 | 747 | (-3168, -241) | -2.28 | 0.022 |
| 4 - 0 | -4185.6 | 71.5 | (-4325.6, -4045.5) | -58.58 | 0.000 |
| 2 - 1 | 4720 | 6787 | (-8583, 18024) | 0.70 | 0.487 |
| 3 - 1 | 3014 | 6825 | (-10363, 16391) | 0.44 | 0.659 |
| 4 - 1 | 533 | 6784 | (-12764, 13831) | 0.08 | 0.937 |
| 3 - 2 | -1706 | 772 | (-3219, -193) | -2.21 | 0.027 |
| 4 - 2 | -4187 | 209 | (-4597, -3777) | -20.02 | 0.000 |
| 4 - 3 | -2481 | 746 | (-3943, -1018) | -3.32 | 0.001 |

*Simultaneous confidence level = 71.42%*



Bonferroni adjustment:

**Fisher Pairwise Comparisons**

**Grouping Information Using the Fisher LSD Method and 50% Confidence**

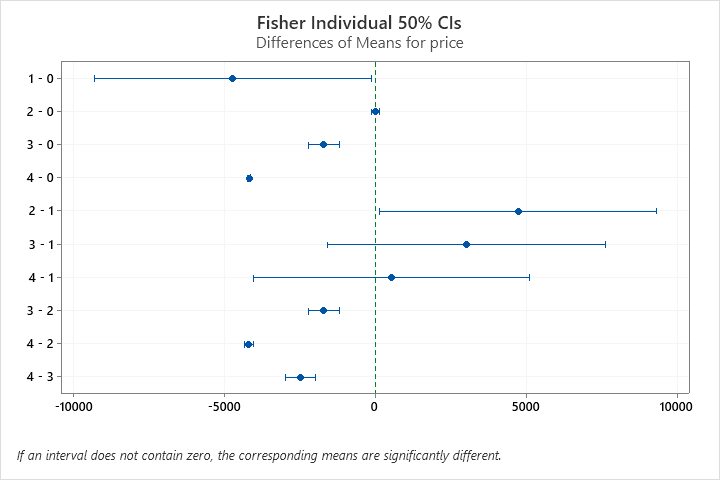
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **fuelCode** | **N** | **Mean** | **Grouping** | | |
| 2 | 2217 | 19207 | A |  |  |
| 0 | 31649 | 19206.1 | A |  |  |
| 3 | 166 | 17501 |  | B |  |
| 4 | 41905 | 15020.5 |  |  | C |
| 1 | 2 | 14487 |  | B | C |

*Means that do not share a letter are significantly different.*

**Fisher Individual Tests for Differences of Means**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Difference of Levels** | **Difference of Means** | **SE of Difference** | **50% CI** | **T-Value** | **Adjusted P-Value** |
| 1 - 0 | -4719 | 6785 | (-9295, -143) | -0.70 | 0.487 |
| 2 - 0 | 1 | 211 | (-141, 144) | 0.01 | 0.995 |
| 3 - 0 | -1705 | 747 | (-2208, -1201) | -2.28 | 0.022 |
| 4 - 0 | -4185.6 | 71.5 | (-4233.8, -4137.4) | -58.58 | 0.000 |
| 2 - 1 | 4720 | 6787 | (142, 9298) | 0.70 | 0.487 |
| 3 - 1 | 3014 | 6825 | (-1589, 7618) | 0.44 | 0.659 |
| 4 - 1 | 533 | 6784 | (-4043, 5110) | 0.08 | 0.937 |
| 3 - 2 | -1706 | 772 | (-2227, -1185) | -2.21 | 0.027 |
| 4 - 2 | -4187 | 209 | (-4328, -4046) | -20.02 | 0.000 |
| 4 - 3 | -2481 | 746 | (-2984, -1978) | -3.32 | 0.001 |

*Simultaneous confidence level = 3.81%*



The analysis of the difference of means can be done based on the plots and their description.

1. **Brand and Price**

Using Linear Regression Model:

**Regression Equation**

|  |  |  |
| --- | --- | --- |
| price | = | 22949.1 B1 + 22717.7 B2 + 23624 B3 + 13492 B4 + 12255.5 B5 + 12701 B6 + 24843.2 B7 + 14205 B8 + 12513 B9 + 10425.3 B10 + 16879.0 B11 |

The coefficients of the brands gives the mean values of the prices for those brands.

**Regression Equation**

|  |  |  |
| --- | --- | --- |
| price | = | 16964.0 + 5985.1 B1\_1 + 5753.7 B2\_1 + 6660 B3\_1 - 3472 B4\_1 - 4708.5 B5\_1 - 4263 B6\_1 + 7879.2 B7\_1 - 2759 B8\_1 - 4451 B9\_1 - 6538.7 B10\_1 |

The following recoding was applied: B1\_1 = B1\_1; B2 = B2\_1; B3 = B3\_1; B4 = B4\_1; B5 = B5\_1; B6 = B6\_1; B7 = B7\_1; B8= B8\_1; B9= B9\_1; B10 = B10\_1; B11 = B11\_1.

B1\_1 has a positive effect on the price, i.e., it is effective in increasing the price of a car by 5985.1 over the grand population mean = 16964.0. The same can be interpreted for the other brands with positive or negative effect based on the positive or negative sign respectively. The level effect of B11\_1 is calculated as: effect = - (5985.1 + 5735.7 + 6660 – 3472 – 4708.5 - 4263 +7879.2 – 2759 – 4451 - 6538) = - 85.8.

Comparisons:

**Tukey Pairwise Comparisons**

**Grouping Information Using the Tukey Method and 95% Confidence**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Brand Code** | **N** | **Mean** | **Grouping** | | | | | | | |
| 6 | 9074 | 24843 | A |  |  |  |  |  |  |  |
| 2 | 2721 | 23624 |  | B |  |  |  |  |  |  |
| 0 | 7472 | 22949 |  |  | C |  |  |  |  |  |
| 1 | 7556 | 22718 |  |  | C |  |  |  |  |  |
| 10 | 10692 | 16879.0 |  |  |  | D |  |  |  |  |
| 7 | 4343 | 14204.9 |  |  |  |  | E |  |  |  |
| 3 | 3889 | 13491.9 |  |  |  |  |  | F |  |  |
| 5 | 3379 | 12701 |  |  |  |  |  |  | G |  |
| 8 | 4695 | 12512.8 |  |  |  |  |  |  | G |  |
| 4 | 12612 | 12255.5 |  |  |  |  |  |  | G |  |
| 9 | 9506 | 10425.3 |  |  |  |  |  |  |  | H |

*Means that do not share a letter are significantly different.*

**Fisher Pairwise Comparisons**

**Grouping Information Using the Fisher LSD Method and 95% Confidence**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Brand Code** | **N** | **Mean** | **Grouping** | | | | | | | | |
| 6 | 9074 | 24843 | A |  |  |  |  |  |  |  |  |
| 2 | 2721 | 23624 |  | B |  |  |  |  |  |  |  |
| 0 | 7472 | 22949 |  |  | C |  |  |  |  |  |  |
| 1 | 7556 | 22718 |  |  | C |  |  |  |  |  |  |
| 10 | 10692 | 16879.0 |  |  |  | D |  |  |  |  |  |
| 7 | 4343 | 14204.9 |  |  |  |  | E |  |  |  |  |
| 3 | 3889 | 13491.9 |  |  |  |  |  | F |  |  |  |
| 5 | 3379 | 12701 |  |  |  |  |  |  | G |  |  |
| 8 | 4695 | 12512.8 |  |  |  |  |  |  | G | H |  |
| 4 | 12612 | 12255.5 |  |  |  |  |  |  |  | H |  |
| 9 | 9506 | 10425.3 |  |  |  |  |  |  |  |  | I |

*Means that do not share a letter are significantly different.*

Bonferroni adjustment:

**Fisher Pairwise Comparisons**

**Grouping Information Using the Fisher LSD Method and 90.91% Confidence**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Brand Code** | **N** | **Mean** | **Grouping** | | | | | | | | | |
| 6 | 9074 | 24843 | A |  |  |  |  |  |  |  |  |  |
| 2 | 2721 | 23624 |  | B |  |  |  |  |  |  |  |  |
| 0 | 7472 | 22949 |  |  | C |  |  |  |  |  |  |  |
| 1 | 7556 | 22718 |  |  |  | D |  |  |  |  |  |  |
| 10 | 10692 | 16879.0 |  |  |  |  | E |  |  |  |  |  |
| 7 | 4343 | 14204.9 |  |  |  |  |  | F |  |  |  |  |
| 3 | 3889 | 13491.9 |  |  |  |  |  |  | G |  |  |  |
| 5 | 3379 | 12701 |  |  |  |  |  |  |  | H |  |  |
| 8 | 4695 | 12512.8 |  |  |  |  |  |  |  | H |  |  |
| 4 | 12612 | 12255.5 |  |  |  |  |  |  |  |  | I |  |
| 9 | 9506 | 10425.3 |  |  |  |  |  |  |  |  |  | J |

*Means that do not share a letter are significantly different.*

The one-way ANOVA comparisons of the brands and the price gives information about the differences in means.

**Two-way ANOVA:**

**General Linear Model: price versus FT\_1, FT\_2, FT\_3, FT\_4, FT\_5, B1, B2, B3, B4, B5, B6, B7, B8, B9, B10, B11**

**Comparisons for price**

**Tukey Pairwise Comparisons: FT\_1**

**Grouping Information Using the Tukey Method and 95% Confidence**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **FT\_1** | **N** | **Mean** | **Grouping** | |
| 0 | 63363 | 6249.40 | A |  |
| 1 | 45177 | -7138.28 |  | B |

*Means that do not share a letter are significantly different.*

**Tukey Pairwise Comparisons: FT\_2**

**Grouping Information Using the Tukey Method and 95% Confidence**

|  |  |  |  |
| --- | --- | --- | --- |
| **FT\_2** | **N** | **Mean** | **Grouping** |
| 0 | 108534 | -3.107 | A |
| 1 | 6 | -885.775 | A |

*Means that do not share a letter are significantly different.*

**Tukey Pairwise Comparisons: FT\_3**

**Grouping Information Using the Tukey Method and 95% Confidence**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **FT\_3** | **N** | **Mean** | **Grouping** | |
| 0 | 105311 | 5361.16 | A |  |
| 1 | 3229 | -6250.04 |  | B |

*Means that do not share a letter are significantly different.*

**Tukey Pairwise Comparisons: FT\_4**

**Grouping Information Using the Tukey Method and 95% Confidence**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **FT\_4** | **N** | **Mean** | **Grouping** | |
| 0 | 108287 | 12363.3 | A |  |
| 1 | 253 | -13252.2 |  | B |

*Means that do not share a letter are significantly different.*

**Tukey Pairwise Comparisons: B1**

**Grouping Information Using the Tukey Method and 95% Confidence**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **B1** | **N** | **Mean** | **Grouping** | |
| 1 | 10668 | 2092.85 | A |  |
| 0 | 97872 | -2981.73 |  | B |

*Means that do not share a letter are significantly different.*

**Tukey Pairwise Comparisons: B2**

**Grouping Information Using the Tukey Method and 95% Confidence**

|  |  |  |  |
| --- | --- | --- | --- |
| **B2** | **N** | **Mean** | **Grouping** |
| 0 | 97759 | 3368.47 | A |
| 1 | 10781 | -4257.35 | A |

*Means that do not share a letter are significantly different.*

**Tukey Simultaneous Tests for Differences of Means**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Difference of B2 Levels** | **Difference of Means** | **SE of Difference** | **Simultaneous 95% CI** | **T-Value** | **Adjusted P-Value** |
| 1 - 0 | -7626 | 4725 | (-16887, 1635) | -1.61 | 0.107 |

*Individual confidence level = 95.00%*

**Tukey Pairwise Comparisons: B3**

**Grouping Information Using the Tukey Method and 95% Confidence**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **B3** | **N** | **Mean** | **Grouping** | |
| 0 | 104641 | 3631.47 | A |  |
| 1 | 3899 | -4520.36 |  | B |

*Means that do not share a letter are significantly different.*

**Tukey Pairwise Comparisons: B4**

**Grouping Information Using the Tukey Method and 95% Confidence**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **B4** | **N** | **Mean** | **Grouping** | |
| 0 | 103086 | 1338.65 | A |  |
| 1 | 5454 | -2227.53 |  | B |

*Means that do not share a letter are significantly different.*

**Tukey Pairwise Comparisons: B5**

**Grouping Information Using the Tukey Method and 95% Confidence**

|  |  |  |  |
| --- | --- | --- | --- |
| **B5** | **N** | **Mean** | **Grouping** |
| 0 | 90575 | 2058.39 | A |
| 1 | 17965 | -2947.27 | A |

*Means that do not share a letter are significantly different.*

**Tukey Pairwise Comparisons: B6**

**Grouping Information Using the Tukey Method and 95% Confidence**

|  |  |  |  |
| --- | --- | --- | --- |
| **B6** | **N** | **Mean** | **Grouping** |
| 0 | 103680 | 2980.82 | A |
| 1 | 4860 | -3869.70 | A |

*Means that do not share a letter are significantly different.*

**Tukey Pairwise Comparisons: B7**

**Grouping Information Using the Tukey Method and 95% Confidence**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **B7** | **N** | **Mean** | **Grouping** | |
| 0 | 95421 | 1872.37 | A |  |
| 1 | 13119 | -2761.25 |  | B |

*Means that do not share a letter are significantly different.*

**Tukey Pairwise Comparisons: B8**

**Grouping Information Using the Tukey Method and 95% Confidence**

|  |  |  |  |
| --- | --- | --- | --- |
| **B8** | **N** | **Mean** | **Grouping** |
| 0 | 102273 | -418.557 | A |
| 1 | 6267 | -470.326 | A |

*Means that do not share a letter are significantly different.*

**Tukey Pairwise Comparisons: B9**

**Grouping Information Using the Tukey Method and 95% Confidence**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **B9** | **N** | **Mean** | **Grouping** | |
| 0 | 101802 | 2584.83 | A |  |
| 1 | 6738 | -3473.72 |  | B |

*Means that do not share a letter are significantly different.*

**Tukey Pairwise Comparisons: B10**

**Grouping Information Using the Tukey Method and 95% Confidence**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **B10** | **N** | **Mean** | **Grouping** | |
| 0 | 94908 | 7635.29 | A |  |
| 1 | 13632 | -8524.18 |  | B |

*Means that do not share a letter are significantly different.*

**Tukey Pairwise Comparisons: FT\_1\*B1**

**Grouping Information Using the Tukey Method and 95% Confidence**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **FT\_1\*B1** | **N** | **Mean** | **Grouping** | | | |
| 0 1 | 5091 | 9506.05 | A |  |  |  |
| 0 0 | 58272 | 2992.75 |  | B |  |  |
| 1 1 | 5577 | -5320.35 |  |  | C |  |
| 1 0 | 39600 | -8956.21 |  |  |  | D |

*Means that do not share a letter are significantly different.*

**Tukey Pairwise Comparisons: FT\_1\*B2**

**Grouping Information Using the Tukey Method and 95% Confidence**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **FT\_1\*B2** | **N** | **Mean** | **Grouping** | | |
| 0 0 | 59609 | 8508.5 | A |  |  |
| 0 1 | 3754 | 3990.3 | A | B |  |
| 1 0 | 38150 | -1771.6 |  | B | C |
| 1 1 | 7027 | -12505.0 |  |  | C |

*Means that do not share a letter are significantly different.*

**Tukey Pairwise Comparisons: FT\_1\*B3**

**Grouping Information Using the Tukey Method and 95% Confidence**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **FT\_1\*B3** | **N** | **Mean** | **Grouping** | | |
| 0 0 | 61804 | 8034.2 | A |  |  |
| 0 1 | 1559 | 4464.6 | A |  |  |
| 1 0 | 42837 | -771.2 |  | B |  |
| 1 1 | 2340 | -13505.3 |  |  | C |

*Means that do not share a letter are significantly different.*

**Tukey Pairwise Comparisons: FT\_1\*B4**

**Grouping Information Using the Tukey Method and 95% Confidence**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **FT\_1\*B4** | **N** | **Mean** | **Grouping** | | | |
| 0 0 | 59818 | 6884.8 | A |  |  |  |
| 0 1 | 3545 | 5614.0 |  | B |  |  |
| 1 0 | 43268 | -4207.5 |  |  | C |  |
| 1 1 | 1909 | -10069.1 |  |  |  | D |

*Means that do not share a letter are significantly different.*

**Tukey Pairwise Comparisons: FT\_1\*B5**

**Grouping Information Using the Tukey Method and 95% Confidence**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **FT\_1\*B5** | **N** | **Mean** | **Grouping** | | |
| 0 0 | 51160 | 8356.4 | A |  |  |
| 0 1 | 12203 | 4142.4 | A | B |  |
| 1 0 | 39415 | -4239.6 |  | B | C |
| 1 1 | 5762 | -10037.0 |  |  | C |

*Means that do not share a letter are significantly different.*

**Tukey Pairwise Comparisons: FT\_1\*B6**

**Grouping Information Using the Tukey Method and 95% Confidence**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **FT\_1\*B6** | **N** | **Mean** | **Grouping** | | |
| 0 0 | 60111 | 9675.6 | A |  |  |
| 0 1 | 3252 | 2823.2 | A | B |  |
| 1 0 | 43569 | -3713.9 |  | B | C |
| 1 1 | 1608 | -10562.6 |  |  | C |

*Means that do not share a letter are significantly different.*

**Tukey Pairwise Comparisons: FT\_1\*B7**

**Grouping Information Using the Tukey Method and 95% Confidence**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **FT\_1\*B7** | **N** | **Mean** | **Grouping** | | |
| 0 0 | 59431 | 6410.7 | A |  |  |
| 0 1 | 3932 | 6088.1 | A |  |  |
| 1 0 | 35990 | -2665.9 |  | B |  |
| 1 1 | 9187 | -11610.6 |  |  | C |

*Means that do not share a letter are significantly different.*

**Tukey Pairwise Comparisons: FT\_1\*B8**

**Grouping Information Using the Tukey Method and 95% Confidence**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **FT\_1\*B8** | **N** | **Mean** | **Grouping** | |
| 0 1 | 4198 | 6363.34 | A |  |
| 0 0 | 59165 | 6135.45 | A |  |
| 1 0 | 43108 | -6972.57 |  | B |
| 1 1 | 2069 | -7304.00 |  | B |

*Means that do not share a letter are significantly different.*

**Tukey Pairwise Comparisons: FT\_1\*B9**

**Grouping Information Using the Tukey Method and 95% Confidence**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **FT\_1\*B9** | **N** | **Mean** | **Grouping** | | | |
| 0 0 | 57128 | 9854.73 | A |  |  |  |
| 0 1 | 6235 | 2644.07 |  | B |  |  |
| 1 0 | 44674 | -4685.06 |  |  | C |  |
| 1 1 | 503 | -9591.50 |  |  |  | D |

*Means that do not share a letter are significantly different.*

**Tukey Pairwise Comparisons: FT\_1\*B10**

**Grouping Information Using the Tukey Method and 95% Confidence**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **FT\_1\*B10** | **N** | **Mean** | **Grouping** | | |
| 0 0 | 52554 | 13644.5 | A |  |  |
| 1 0 | 42354 | 1626.1 |  | B |  |
| 0 1 | 10809 | -1145.7 |  | B |  |
| 1 1 | 2823 | -15902.6 |  |  | C |

*Means that do not share a letter are significantly different.*

**Tukey Pairwise Comparisons: FT\_2\*B2**

**Grouping Information Using the Tukey Method and 95% Confidence**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **FT\_2\*B2** | **N** | **Mean** | **Grouping** | |
| 1 0 | 3 | 5095.19 | A | B |
| 0 0 | 97756 | 1641.75 | A |  |
| 0 1 | 10778 | -1647.97 |  | B |
| 1 1 | 3 | -6866.74 | A | B |

*Means that do not share a letter are significantly different.*

**Tukey Pairwise Comparisons: FT\_2\*B5**

**Grouping Information Using the Tukey Method and 95% Confidence**

|  |  |  |  |
| --- | --- | --- | --- |
| **FT\_2\*B5** | **N** | **Mean** | **Grouping** |
| 0 0 | 90571 | 2837.47 | A |
| 1 0 | 4 | 1279.30 | A |
| 0 1 | 17963 | -2843.68 | A |
| 1 1 | 2 | -3050.85 | A |

*Means that do not share a letter are significantly different.*

**Tukey Pairwise Comparisons: FT\_3\*B1**

**Grouping Information Using the Tukey Method and 95% Confidence**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **FT\_3\*B1** | **N** | **Mean** | **Grouping** | | |
| 0 1 | 10640 | 8288.45 | A |  |  |
| 0 0 | 94671 | 2433.87 |  | B |  |
| 1 1 | 28 | -4102.75 |  | B |  |
| 1 0 | 3201 | -8397.33 |  |  | C |

*Means that do not share a letter are significantly different.*

**Tukey Pairwise Comparisons: FT\_3\*B2**

**Grouping Information Using the Tukey Method and 95% Confidence**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **FT\_3\*B2** | **N** | **Mean** | **Grouping** | |
| 0 0 | 94828 | 7558.7 | A |  |
| 0 1 | 10483 | 3163.6 | A |  |
| 1 0 | 2931 | -821.8 | A | B |
| 1 1 | 298 | -11678.3 |  | B |

*Means that do not share a letter are significantly different.*

**Tukey Pairwise Comparisons: FT\_3\*B3**

**Grouping Information Using the Tukey Method and 95% Confidence**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **FT\_3\*B3** | **N** | **Mean** | **Grouping** | |
| 0 1 | 3748 | 5767.7 | A |  |
| 0 0 | 101563 | 4954.6 | A |  |
| 1 0 | 3078 | 2308.3 | A |  |
| 1 1 | 151 | -14808.4 |  | B |

*Means that do not share a letter are significantly different.*

**Tukey Pairwise Comparisons: FT\_3\*B5**

**Grouping Information Using the Tukey Method and 95% Confidence**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **FT\_3\*B5** | **N** | **Mean** | **Grouping** | |
| 0 0 | 87368 | 8181.41 | A |  |
| 0 1 | 17943 | 2540.91 | A | B |
| 1 0 | 3207 | -4064.63 |  | B |
| 1 1 | 22 | -8435.45 | A | B |

*Means that do not share a letter are significantly different.*

**Tukey Pairwise Comparisons: FT\_3\*B6**

**Grouping Information Using the Tukey Method and 95% Confidence**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **FT\_3\*B6** | **N** | **Mean** | **Grouping** | | |
| 0 0 | 100800 | 8536.47 | A |  |  |
| 0 1 | 4511 | 2185.85 | A | B |  |
| 1 0 | 2880 | -2574.83 |  | B | C |
| 1 1 | 349 | -9925.25 |  |  | C |

*Means that do not share a letter are significantly different.*

**Tukey Pairwise Comparisons: FT\_3\*B7**

**Grouping Information Using the Tukey Method and 95% Confidence**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **FT\_3\*B7** | **N** | **Mean** | **Grouping** | |
| 0 1 | 12946 | 6655.0 | A |  |
| 0 0 | 92365 | 4067.3 | A |  |
| 1 0 | 3056 | -322.6 | A |  |
| 1 1 | 173 | -12177.5 |  | B |

*Means that do not share a letter are significantly different.*

**Tukey Pairwise Comparisons: FT\_3\*B8**

**Grouping Information Using the Tukey Method and 95% Confidence**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **FT\_3\*B8** | **N** | **Mean** | **Grouping** | | |
| 0 0 | 99062 | 7574.30 | A |  |  |
| 0 1 | 6249 | 3148.02 |  | B |  |
| 1 1 | 18 | -4088.67 | A | B | C |
| 1 0 | 3211 | -8411.41 |  |  | C |

*Means that do not share a letter are significantly different.*

**Tukey Pairwise Comparisons: FT\_3\*B9**

**Grouping Information Using the Tukey Method and 95% Confidence**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **FT\_3\*B9** | **N** | **Mean** | **Grouping** | | |
| 0 0 | 100616 | 7929.20 | A |  |  |
| 0 1 | 4695 | 2793.12 |  | B |  |
| 1 0 | 1186 | -2759.54 |  | B |  |
| 1 1 | 2043 | -9740.55 |  |  | C |

*Means that do not share a letter are significantly different.*

**Tukey Pairwise Comparisons: FT\_3\*B10**

**Grouping Information Using the Tukey Method and 95% Confidence**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **FT\_3\*B10** | **N** | **Mean** | **Grouping** | | |
| 0 0 | 91681 | 11304.8 | A |  |  |
| 1 0 | 3227 | 3965.8 |  | B |  |
| 0 1 | 13630 | -582.5 |  | B | C |
| 1 1 | 2 | -16465.9 |  |  | C |

*Means that do not share a letter are significantly different.*

**Tukey Pairwise Comparisons: FT\_4\*B2**

**Grouping Information Using the Tukey Method and 95% Confidence**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **FT\_4\*B2** | **N** | **Mean** | **Grouping** | |
| 0 0 | 97542 | 13114.1 | A |  |
| 0 1 | 10745 | 11612.5 | A |  |
| 1 0 | 217 | -6377.2 | A |  |
| 1 1 | 36 | -20127.2 |  | B |

*Means that do not share a letter are significantly different.*

**Tukey Pairwise Comparisons: FT\_4\*B3**

**Grouping Information Using the Tukey Method and 95% Confidence**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **FT\_4\*B3** | **N** | **Mean** | **Grouping** | | |
| 0 0 | 104394 | 13060.6 | A |  |  |
| 0 1 | 3893 | 11666.0 |  | B |  |
| 1 0 | 247 | -5797.6 | A | B |  |
| 1 1 | 6 | -20706.7 |  |  | C |

*Means that do not share a letter are significantly different.*

**Tukey Pairwise Comparisons: FT\_4\*B5**

**Grouping Information Using the Tukey Method and 95% Confidence**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **FT\_4\*B5** | **N** | **Mean** | **Grouping** | | |
| 0 0 | 90323 | 13893.5 | A |  |  |
| 0 1 | 17964 | 10833.1 | A | B |  |
| 1 0 | 252 | -9776.7 |  | B | C |
| 1 1 | 1 | -16727.6 |  |  | C |

*Means that do not share a letter are significantly different.*

**Tukey Pairwise Comparisons: FT\_4\*B6**

**Grouping Information Using the Tukey Method and 95% Confidence**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **FT\_4\*B6** | **N** | **Mean** | **Grouping** | | |
| 0 0 | 103428 | 14730.0 | A |  |  |
| 0 1 | 4859 | 9996.6 |  | B |  |
| 1 0 | 252 | -8768.4 |  |  | C |
| 1 1 | 1 | -17736.0 |  |  | C |

*Means that do not share a letter are significantly different.*

**Tukey Pairwise Comparisons: FT\_4\*B7**

**Grouping Information Using the Tukey Method and 95% Confidence**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **FT\_4\*B7** | **N** | **Mean** | **Grouping** | | |
| 0 1 | 13112 | 13134.9 | A |  |  |
| 0 0 | 95175 | 11591.6 |  | B |  |
| 1 0 | 246 | -7846.9 |  | B |  |
| 1 1 | 7 | -18657.4 |  |  | C |

*Means that do not share a letter are significantly different.*

**Tukey Pairwise Comparisons: FT\_4\*B8**

**Grouping Information Using the Tukey Method and 95% Confidence**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **FT\_4\*B8** | **N** | **Mean** | **Grouping** | |
| 0 1 | 6258 | 13420.0 | A |  |
| 0 0 | 102029 | 11306.6 | A |  |
| 1 0 | 244 | -12143.7 |  | B |
| 1 1 | 9 | -14360.6 |  | B |

*Means that do not share a letter are significantly different.*

**Tukey Pairwise Comparisons: FT\_4\*B9**

**Grouping Information Using the Tukey Method and 95% Confidence**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **FT\_4\*B9** | **N** | **Mean** | **Grouping** | | |
| 0 0 | 101654 | 14962.5 | A |  |  |
| 0 1 | 6633 | 9764.1 |  | B |  |
| 1 0 | 148 | -9792.8 |  | B |  |
| 1 1 | 105 | -16711.5 |  |  | C |

*Means that do not share a letter are significantly different.*

**Tukey Pairwise Comparisons: FT\_4\*B10**

**Grouping Information Using the Tukey Method and 95% Confidence**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **FT\_4\*B10** | **N** | **Mean** | **Grouping** | | |
| 0 0 | 94656 | 17667.3 | A |  |  |
| 0 1 | 13631 | 7059.3 |  | B |  |
| 1 0 | 252 | -2396.7 |  | B | C |
| 1 1 | 1 | -24107.6 |  |  | C |

*Means that do not share a letter are significantly different.*

**Analysis of Variance**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Adj SS** | **Adj MS** | **F-Value** | **P-Value** |
| FT\_1 | 1 | 28353401319 | 28353401319 | 439.58 | 0.000 |
| FT\_2 | 1 | 3738017 | 3738017 | 0.06 | 0.810 |
| FT\_3 | 1 | 538218980 | 538218980 | 8.34 | 0.004 |
| FT\_4 | 1 | 676756716 | 676756716 | 10.49 | 0.001 |
| B1 | 1 | 2398429147 | 2398429147 | 37.18 | 0.000 |
| B2 | 1 | 168007623 | 168007623 | 2.60 | 0.107 |
| B3 | 1 | 1376577078 | 1376577078 | 21.34 | 0.000 |
| B4 | 1 | 47111926428 | 47111926428 | 730.41 | 0.000 |
| B5 | 1 | 39078951 | 39078951 | 0.61 | 0.436 |
| B6 | 1 | 183588540 | 183588540 | 2.85 | 0.092 |
| B7 | 1 | 512182787 | 512182787 | 7.94 | 0.005 |
| B8 | 1 | 57619 | 57619 | 0.00 | 0.976 |
| B9 | 1 | 4749943040 | 4749943040 | 73.64 | 0.000 |
| B10 | 1 | 687582153 | 687582153 | 10.66 | 0.001 |
| FT\_1\*B1 | 1 | 12724842552 | 12724842552 | 197.28 | 0.000 |
| FT\_1\*B2 | 1 | 54495089837 | 54495089837 | 844.87 | 0.000 |
| FT\_1\*B3 | 1 | 59377657412 | 59377657412 | 920.57 | 0.000 |
| FT\_1\*B4 | 1 | 19519075042 | 19519075042 | 302.62 | 0.000 |
| FT\_1\*B5 | 1 | 4735064085 | 4735064085 | 73.41 | 0.000 |
| FT\_1\*B6 | 1 | 10596 | 10596 | 0.00 | 0.990 |
| FT\_1\*B7 | 1 | 1.14507E+11 | 1.14507E+11 | 1775.28 | 0.000 |
| FT\_1\*B8 | 1 | 313799798 | 313799798 | 4.87 | 0.027 |
| FT\_1\*B9 | 1 | 2118159222 | 2118159222 | 32.84 | 0.000 |
| FT\_1\*B10 | 1 | 10406579736 | 10406579736 | 161.34 | 0.000 |
| FT\_2\*B2 | 1 | 56389181 | 56389181 | 0.87 | 0.350 |
| FT\_2\*B5 | 1 | 1216645 | 1216645 | 0.02 | 0.891 |
| FT\_3\*B1 | 1 | 56694710 | 56694710 | 0.88 | 0.348 |
| FT\_3\*B2 | 1 | 3915821391 | 3915821391 | 60.71 | 0.000 |
| FT\_3\*B3 | 1 | 22403683684 | 22403683684 | 347.34 | 0.000 |
| FT\_3\*B5 | 1 | 30676766 | 30676766 | 0.48 | 0.490 |
| FT\_3\*B6 | 1 | 97776270 | 97776270 | 1.52 | 0.218 |
| FT\_3\*B7 | 1 | 15970908294 | 15970908294 | 247.61 | 0.000 |
| FT\_3\*B8 | 1 | 1218703358 | 1218703358 | 18.89 | 0.000 |
| FT\_3\*B9 | 1 | 439324671 | 439324671 | 6.81 | 0.009 |
| FT\_3\*B10 | 1 | 143967829 | 143967829 | 2.23 | 0.135 |
| FT\_4\*B2 | 1 | 3780620587 | 3780620587 | 58.61 | 0.000 |
| FT\_4\*B3 | 1 | 1020407462 | 1020407462 | 15.82 | 0.000 |
| FT\_4\*B5 | 1 | 14960773 | 14960773 | 0.23 | 0.630 |
| FT\_4\*B6 | 1 | 17715969 | 17715969 | 0.27 | 0.600 |
| FT\_4\*B7 | 1 | 986312625 | 986312625 | 15.29 | 0.000 |
| FT\_4\*B8 | 1 | 152491056 | 152491056 | 2.36 | 0.124 |
| FT\_4\*B9 | 1 | 138418758 | 138418758 | 2.15 | 0.143 |
| FT\_4\*B10 | 1 | 121849552 | 121849552 | 1.89 | 0.169 |
| Error | 108496 | 6.99810E+12 | 64500999 |  |  |
| Total | 108539 | 1.03313E+13 |  |  |  |

**Regression Equation**

|  |  |  |
| --- | --- | --- |
| price | = | -444 + 6694 FT\_1\_0 - 6694 FT\_1\_1 + 441 FT\_2\_0 - 441 FT\_2\_1 + 5806 FT\_3\_0 - 5806 FT\_3\_1 + 12808 FT\_4\_0 - 12808 FT\_4\_1 - 2537 B1\_0 + 2537 B1\_1 + 3813 B2\_0 - 3813 B2\_1 + 4076 B3\_0 - 4076 B3\_1 + 1783.1 B4\_0 - 1783.1 B4\_1 + 2503 B5\_0 - 2503 B5\_1 + 3425 B6\_0 - 3425 B6\_1 + 2317 B7\_0 - 2317 B7\_1 + 26 B8\_0 - 26 B8\_1 + 3029 B9\_0 - 3029 B9\_1 + 8080 B10\_0 - 8080 B10\_1 - 719.4 FT\_1\*B1\_0 0 + 719.4 FT\_1\*B1\_0 1 + 719.4 FT\_1\*B1\_1 0 - 719.4 FT\_1\*B1\_1 1 - 1553.8 FT\_1\*B2\_0 0 + 1553.8 FT\_1\*B2\_0 1 + 1553.8 FT\_1\*B2\_1 0 - 1553.8 FT\_1\*B2\_1 1 - 2291.1 FT\_1\*B3\_0 0 + 2291.1 FT\_1\*B3\_0 1 + 2291.1 FT\_1\*B3\_1 0 - 2291.1 FT\_1\*B3\_1 1 - 1147.7 FT\_1\*B4\_0 0 + 1147.7 FT\_1\*B4\_0 1 + 1147.7 FT\_1\*B4\_1 0 - 1147.7 FT\_1\*B4\_1 1 - 395.9 FT\_1\*B5\_0 0 + 395.9 FT\_1\*B5\_0 1 + 395.9 FT\_1\*B5\_1 0 - 395.9 FT\_1\*B5\_1 1 + 0.9 FT\_1\*B6\_0 0 - 0.9 FT\_1\*B6\_0 1 - 0.9 FT\_1\*B6\_1 0 + 0.9 FT\_1\*B6\_1 1 - 2155.5 FT\_1\*B7\_0 0 + 2155.5 FT\_1\*B7\_0 1 + 2155.5 FT\_1\*B7\_1 0 - 2155.5 FT\_1\*B7\_1 1 - 139.8 FT\_1\*B8\_0 0 + 139.8 FT\_1\*B8\_0 1 + 139.8 FT\_1\*B8\_1 0 - 139.8 FT\_1\*B8\_1 1 + 576 FT\_1\*B9\_0 0 - 576 FT\_1\*B9\_0 1 - 576 FT\_1\*B9\_1 0 + 576 FT\_1\*B9\_1 1 - 684.6 FT\_1\*B10\_0 0 + 684.6 FT\_1\*B10\_0 1 + 684.6 FT\_1\*B10\_1 0 - 684.6 FT\_1\*B10\_1 1 - 2168 FT\_2\*B2\_0 0 + 2168 FT\_2\*B2\_0 1 + 2168 FT\_2\*B2\_1 0 - 2168 FT\_2\*B2\_1 1 + 338 FT\_2\*B5\_0 0 - 338 FT\_2\*B5\_0 1 - 338 FT\_2\*B5\_1 0 + 338 FT\_2\*B5\_1 1 - 390 FT\_3\*B1\_0 0 + 390 FT\_3\*B1\_0 1 + 390 FT\_3\*B1\_1 0 - 390 FT\_3\*B1\_1 1 - 1615 FT\_3\*B2\_0 0 + 1615 FT\_3\*B2\_0 1 + 1615 FT\_3\*B2\_1 0 - 1615 FT\_3\*B2\_1 1 - 4482 FT\_3\*B3\_0 0 + 4482 FT\_3\*B3\_0 1 + 4482 FT\_3\*B3\_1 0 - 4482 FT\_3\*B3\_1 1 + 317 FT\_3\*B5\_0 0 - 317 FT\_3\*B5\_0 1 - 317 FT\_3\*B5\_1 0 + 317 FT\_3\*B5\_1 1 - 250 FT\_3\*B6\_0 0 + 250 FT\_3\*B6\_0 1 + 250 FT\_3\*B6\_1 0 - 250 FT\_3\*B6\_1 1 - 3611 FT\_3\*B7\_0 0 + 3611 FT\_3\*B7\_0 1 + 3611 FT\_3\*B7\_1 0 - 3611 FT\_3\*B7\_1 1 + 2187 FT\_3\*B8\_0 0 - 2187 FT\_3\*B8\_0 1 - 2187 FT\_3\*B8\_1 0 + 2187 FT\_3\*B8\_1 1 - 461 FT\_3\*B9\_0 0 + 461 FT\_3\*B9\_0 1 + 461 FT\_3\*B9\_1 0 - 461 FT\_3\*B9\_1 1 - 2136 FT\_3\*B10\_0 0 + 2136 FT\_3\*B10\_0 1 + 2136 FT\_3\*B10\_1 0 - 2136 FT\_3\*B10\_1 1 - 3062 FT\_4\*B2\_0 0 + 3062 FT\_4\*B2\_0 1 + 3062 FT\_4\*B2\_1 0 - 3062 FT\_4\*B2\_1 1 - 3379 FT\_4\*B3\_0 0 + 3379 FT\_4\*B3\_0 1 + 3379 FT\_4\*B3\_1 0 - 3379 FT\_4\*B3\_1 1 - 973 FT\_4\*B5\_0 0 + 973 FT\_4\*B5\_0 1 + 973 FT\_4\*B5\_1 0 - 973 FT\_4\*B5\_1 1 - 1059 FT\_4\*B6\_0 0 + 1059 FT\_4\*B6\_0 1 + 1059 FT\_4\*B6\_1 0 - 1059 FT\_4\*B6\_1 1 - 3088 FT\_4\*B7\_0 0 + 3088 FT\_4\*B7\_0 1 + 3088 FT\_4\*B7\_1 0 - 3088 FT\_4\*B7\_1 1 - 1083 FT\_4\*B8\_0 0 + 1083 FT\_4\*B8\_0 1 + 1083 FT\_4\*B8\_1 0 - 1083 FT\_4\*B8\_1 1 - 430 FT\_4\*B9\_0 0 + 430 FT\_4\*B9\_0 1 + 430 FT\_4\*B9\_1 0 - 430 FT\_4\*B9\_1 1 - 2776 FT\_4\*B10\_0 0 + 2776 FT\_4\*B10\_0 1 + 2776 FT\_4\*B10\_1 0 - 2776 FT\_4\*B10\_1 1 |

The two-way ANOVA gives information about the significance of means and the effects (Alpha’s and Beta’s) of the different levels of predictors on the price and their interaction. FT\_2, B2, B5, B6, B8, FT1\*B6, FT\_2\*B2, FT\_2\*B5, FT\_3\*B1, FT\_3\*B5, FT\_3\*B6, FT\_3\*B10, FT\_4\*B5, FT\_4\*B6, FT\_4\*B8, FT\_4\*B9 and FT\_4\*B10 are insignificant based on the two-way ANOVA. The coefficients of the different levels of fuel types and brands and their interactions give information about their effect on the price over the population grand mean.

**Section 4: Checking the Accuracy of the Model Predictors**

We will now test our selected model (model 5) with our 30% test data set. The results obtained are as follows:

**Regression Analysis: price versus year, 1/mileage, tax, mpg, engineSize, mpg\*engineSize, Brand, transmission, fuelType**

**Method**

|  |  |
| --- | --- |
| Categorical predictor coding | (1, 0) |
| Rows unused | 14210 |
|  |  |
| Box-Cox transformation |  |
| Rounded λ | 0.0807457 |
| Estimated λ | 0.0807457 |
| 95% CI for λ | (0.0712457, 0.0902457) |
| Test set ID | Index |
| Test set fraction | 69.9% |

**Regression Equation**

|  |  |  |
| --- | --- | --- |
| price^0.0807457 | = | -47.667 + 0.024692 year + 0.02710 1/mileage + 0.000053 tax - 0.000034 mpg + 0.078801 engineSize - 0.000548 mpg\*engineSize + 0.0 Brand\_Audi - 0.017357 Brand\_BMW - 0.047162 Brand\_Ford - 0.002325 Brand\_Mercedes - 0.05167 Brand\_Skoda - 0.07760 Brand\_Toyota - 0.083169 Brand\_Vauxhall - 0.030556 Brand\_Volkswagen + 0.0 transmission\_Automatic - 0.024391 transmission\_Manual + 0.0004 transmission\_Other + 0.003362 transmission\_Semi-Auto + 0.0 fuelType\_Diesel + 0.0840 fuelType\_Electric + 0.07387 fuelType\_Hybrid + 0.01887 fuelType\_Other - 0.011469 fuelType\_Petrol |

**Coefficients for Transformed Response**

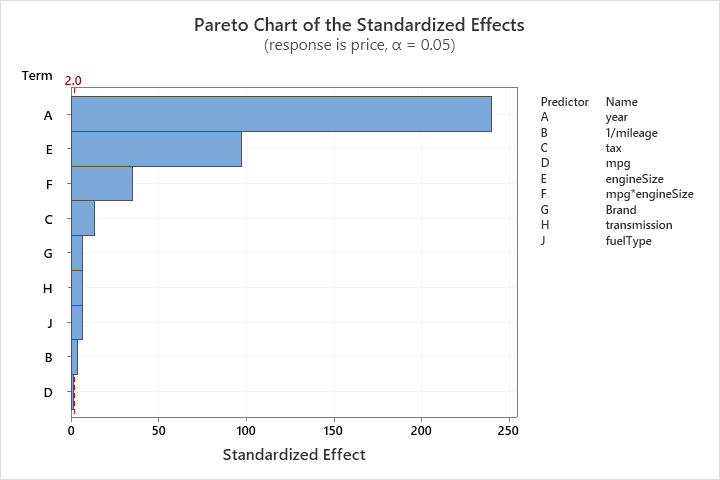
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Term** | **Coef** | **SE Coef** | **T-Value** | **P-Value** | **VIF** |
| Constant | -47.667 | 0.208 | -229.01 | 0.000 |  |
| year | 0.024692 | 0.000103 | 239.51 | 0.000 | 1.12 |
| 1/mileage | 0.02710 | 0.00761 | 3.56 | 0.000 | 1.01 |
| tax | 0.000053 | 0.000004 | 13.32 | 0.000 | 1.46 |
| mpg | -0.000034 | 0.000022 | -1.51 | 0.130 | 2.97 |
| engineSize | 0.078801 | 0.000810 | 97.25 | 0.000 | 4.75 |
| mpg\*engineSize | -0.000548 | 0.000016 | -34.96 | 0.000 | 5.92 |
| Brand |  |  |  |  |  |
| BMW | -0.017357 | 0.000901 | -19.27 | 0.000 | 1.86 |
| Ford | -0.047162 | 0.000847 | -55.65 | 0.000 | 2.52 |
| Mercedes | -0.002325 | 0.000857 | -2.71 | 0.007 | 2.06 |
| Skoda | -0.05167 | 0.00104 | -49.62 | 0.000 | 1.57 |
| Toyota | -0.07760 | 0.00113 | -68.75 | 0.000 | 1.95 |
| Vauxhall | -0.083169 | 0.000890 | -93.42 | 0.000 | 2.26 |
| Volkswagen | -0.030556 | 0.000834 | -36.63 | 0.000 | 2.11 |
| transmission |  |  |  |  |  |
| Manual | -0.024391 | 0.000669 | -36.47 | 0.000 | 2.53 |
| Other | 0.0004 | 0.0249 | 0.01 | 0.989 | 1.00 |
| Semi-Auto | 0.003362 | 0.000658 | 5.11 | 0.000 | 1.78 |
| fuelType |  |  |  |  |  |
| Electric | 0.0840 | 0.0253 | 3.32 | 0.001 | 1.03 |
| Hybrid | 0.07387 | 0.00166 | 44.57 | 0.000 | 1.80 |
| Other | 0.01887 | 0.00410 | 4.60 | 0.000 | 1.03 |
| Petrol | -0.011469 | 0.000637 | -18.01 | 0.000 | 2.29 |

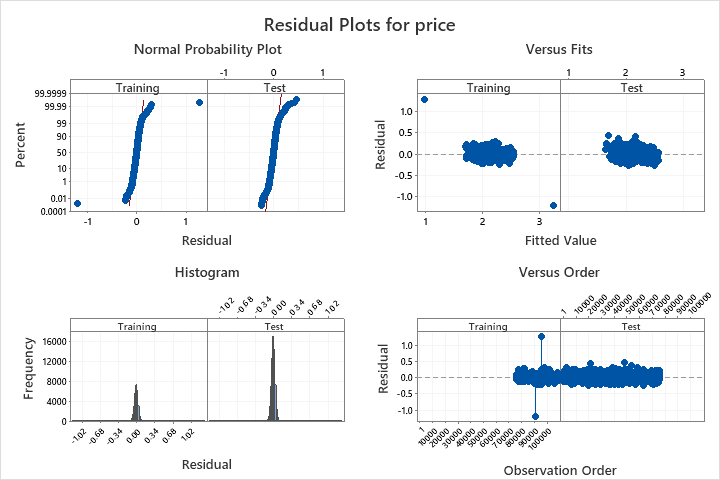
**Model Summary for Transformed Response**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **S** | **R-sq** | **R-sq(adj)** | **R-sq(pred)** | **Test S** | **Test R-sq** |
| 0.0351848 | 85.85% | 85.84% | 85.75% | 0.0338300 | 87.04% |

**Analysis of Variance for Transformed Response**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Adj SS** | **Adj MS** | **F-Value** | **P-Value** |
| Regression | 20 | 212.738 | 10.6369 | 8592.22 | 0.000 |
| year | 1 | 71.017 | 71.0166 | 57365.45 | 0.000 |
| 1/mileage | 1 | 0.016 | 0.0157 | 12.69 | 0.000 |
| tax | 1 | 0.220 | 0.2197 | 177.47 | 0.000 |
| mpg | 1 | 0.003 | 0.0028 | 2.29 | 0.130 |
| engineSize | 1 | 11.707 | 11.7072 | 9456.77 | 0.000 |
| mpg\*engineSize | 1 | 1.513 | 1.5132 | 1222.29 | 0.000 |
| Brand | 7 | 15.390 | 2.1985 | 1775.90 | 0.000 |
| transmission | 3 | 2.678 | 0.8928 | 721.15 | 0.000 |
| fuelType | 4 | 2.719 | 0.6798 | 549.15 | 0.000 |
| Error | 28320 | 35.059 | 0.0012 |  |  |
| Lack-of-Fit | 18881 | 32.906 | 0.0017 | 7.64 | 0.000 |
| Pure Error | 9439 | 2.153 | 0.0002 |  |  |
| Total | 28340 | 247.797 |  |  |  |





We have an accuracy rate of 87.04%, giving us a misclassification rate of 12.96%.

**Section 5: Discussion of Overall Results**

An experimental study was conducted on this dataset. The researchers influenced the response variable by making a transformation on the mileage predictor, removed a particular level for the transmission in the final model and added interaction terms to better fit the model.

First, we started analyzing the dataset with all of the predictors by fitting a multiple regression model with categorical predictors. We found that the model is adequate and all the predictors are significant except the “Other” level of transmission, but it violates the assumption of normality and independence of the residuals. Assuming not normal residuals, we conducted the Brown-Forsythe test and found that the variance of the residuals is not constant.

Then we performed subset regression for the variable selection of the model which indicated that the model with all of the predictors is the best. Upon analyzing the scatter plot of mileage, we transformed the mileage predictor as 1/mileage and found that the model was still adequate but the Mercedes brand and Other transmission level is insignificant. However, the F-value for the lack of fit increased from 8.91 to 9.50. Then, we removed the outliers and refitted the previous model with Box-Cox transformation which improved the P-value for the “Other” transmission level and the R-sq value as compared to the other two models. Considering the relationship between engine size and mpg we refit the model with the interaction term and removed the “Other” level of transmission (the histogram of transmission levels also showcased very few observations for “Other”) and found that the interaction term was significant. However, it has high VIF for the interaction terms because of structural multicollinearity. We compared this model with another model that has Box-Cox transformation and found that Model 5 was the best. It is an adequate model with all predictors found to be significant and has an R-sq value of 87.13%, R-sq (pred) of 87.10% and R-sq (adj) of 87.13%. Despite there being a lack of fit, the residuals show random scattering and the normal probability plot of residuals has improved with heavy tails.

Upon making predictions for the test set, the model has a misclassification rate of 12.96% which is not very large.

*Verdict:* From our various model building procedures, we found that model 5 was the best one in terms of its R2 value, significance of predictors and normality and independence of residuals. It is not perfect i.e., it still has a lack of fit and does not satisfy constant variance assumption of residuals; however, it gives us the best model from all our procedures. Ways to improve the model would include using a different type of machine learning algorithm such as Random Forest Regression or Multi-Layer Perceptron Regressor, which are beyond the scope of this course. We can also consider adding other predictors that affect the response variable.