**USED GM CAR PRICE PREDICTIONS**

**ADVANCED DATA ANALYTICS**

**MEET KHOLIA**

**INFORMATION TECHNOLOGY MANAGEMENT**

**SCHOOL OF APPLIED TECHNOLOGY**

**ILLINOIS INSTITUTE OF TECHNOLOGY**

Table of Contents

[Business Objective 2](#_Toc446679425)

[Selected Data Overview 2](#_Toc446679426)

[Data Flow Diagram 3](#_Toc446679427)

[Exploratory Data Analysis 4](#_Toc446679428)

[Cleaning the data: 4](#_Toc446679429)

[Exploring PRICE variable: 4](#_Toc446679430)

[Exploring MAKE variable: 5](#_Toc446679431)

[Exploring TYPE variable: 5](#_Toc446679432)

[Outliers Treatment: 6](#_Toc446679433)

[Finding Outliers: 7](#_Toc446679434)

[Another way to find outliers: 8](#_Toc446679435)

[Detailed Analysis of the Data: 9](#_Toc446679436)

[Correlation: 9](#_Toc446679437)

[Regression Analysis 10](#_Toc446679438)

[Linear Regression: 10](#_Toc446679439)

[Scatter Plot with fitted line and 95% confidence interval: 11](#_Toc446679440)

[Linear Regression Summary: 11](#_Toc446679441)

[ANOVA (Analysis of Variance) 12](#_Toc446679442)

[ANOVA with the factors: 14](#_Toc446679443)

[ANVOA with factors summary: 18](#_Toc446679444)

[Interaction effects in ANOVA: 19](#_Toc446679445)

[Tukey – Kramer Test 20](#_Toc446679446)

[Conclusion 23](#_Toc446679447)

# 

# Business Objective

* To predict car price for given sample data based on different characteristics of the car such as MAKE of the car, MODEL of the car, CRUISE control, etc.
* Analyze which characteristic will affect car’s selling price.

# Selected Data Overview

* Data is collected from Kelly Blue Book for several hundred 2005 used General Motors (GM) cars.
* This dataset is a sample of over eight hundred 2005 GM cars were selected, then retail price was calculated from the tables provided in the 2005 Central Edition of the Kelly Blue Book.
* This dataset will help predicting the price of a used car depending on the data of more than 800 cars we have. So, on Y-axis the dependent variable will always be the Price of the car. While, all the other variables will be independent variables and will be used on X-axis.

|  |  |  |
| --- | --- | --- |
|  | **NUMERICAL** | **CATEGORICAL** |
| **Y - AXIS** | PRICE | - |
| **X – AXIS** | MILEAGE | - |
| - | MAKE |
| - | MODEL |
| - | TRIM |
| - | TYPE |
| - | CYLINDER |
| - | LITER |
| - | DOORS |
| - | CRUISE |
| - | SOUND |
| - | LEATHER |

# 

# Data Flow Diagram

* Following Data Flow Diagram explains data extraction, integration and presentation :

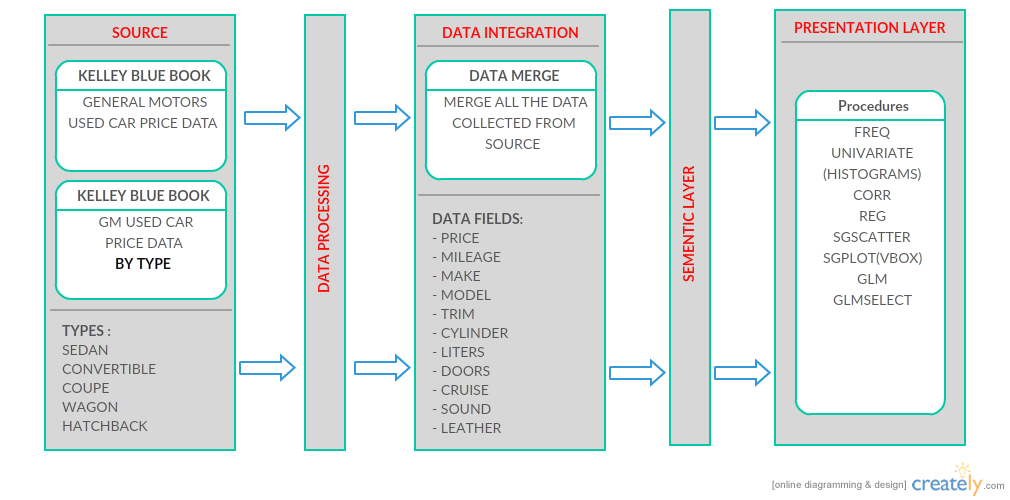


Figure 3.1

# 

# Exploratory Data Analysis

## Cleaning the data:

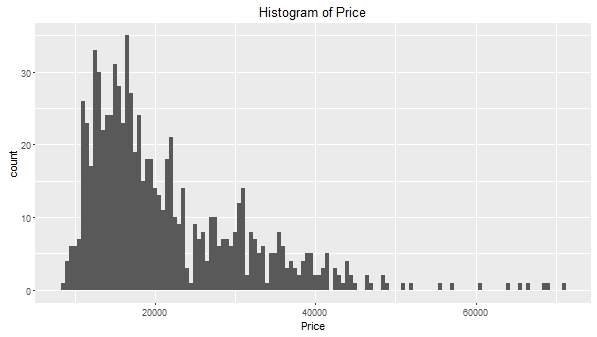
* As data was collected from a reliable and good source, it did not have any missing or NULL values.
* But, when reading the data, R interpreted some categorical data as numeric as their levels was defined as a number.
* E.g. for Cruise =1/0 ( 1= Car has Cruise control, 0= Car doesn’t have Cruise control). R read Cruise variable as numeric.
* Code to convert : GM$Cruise <- as.factor(GM$Cruise)
* So, converting this and variable like this in factors was the first step in Exploratory Data Analysis.

## Exploring PRICE variable:

* All the Price value are in USD
* Exploring minimum, maximum, mean and quartiles for PRICE variables.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Minimum** | **1st Quartile** | **Median** | **Mean** | **3rd Quartile** | **Maximum** |
| $8638.93 | $14273.08 | $18024.99 | $21343.14 | $26717.32 | $70755.47 |

* From Figure 4.1, we can say that Price variable histogram is right skewed and Price not normally distributed.
* We can also say that, most of the cars has Price within the range of $10000 and $50000. Only some of them are above $50000 and below $10000.

Figure 4.1

## Exploring MAKE variable:

* In the following table, we can see MAKE-wise mean price of the car:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Buick** | **Cadillac** | **Chevrolet** | **Pontiac** | **SAAB** | **Saturn** |
| **AVG Price =** | $20815.11 | $40936.34 | $16427.60 | $18412.10 | $29494.70 | $13978.81 |

* In Figure 4.2, we can find a boxplot of the Make/Price.
* We can easily see in the figure that Cadillac made cars are more expensive as compared to any other make. Also, Saturn made cars are the least expensive.
* Here, Chevrolet and Cadillac has many values lying above 95% range. These are outliers, we will see how to treat them in a few steps.

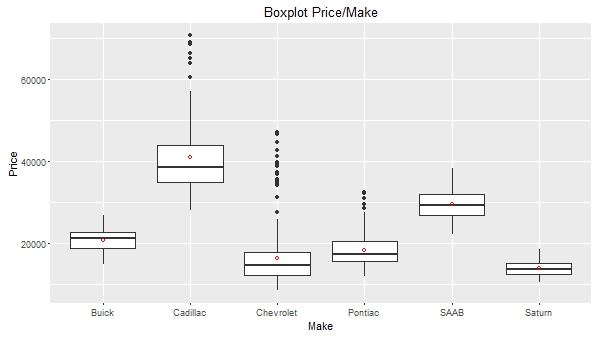


Figure 4.2

## Exploring TYPE variable:

* In the following table, we can see TYPE-wise mean price of the car:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Convertible** | **Coupe** | **Hatchback** | **Sedan** | **Wagon** |
| **AVG Price =** | $40831.71 | $17726.93 | $14170.93 | $21067.93 | 22859.25 |

* Figure 4.3, we can find boxplot for Price/Type.
* We ca see that Convertible cars are significantly expensive than all the other types of cars.
* Though Convertible and Sedan have some outliers, we cannot say that Sedan is less expensive because number of outliers is a lot.

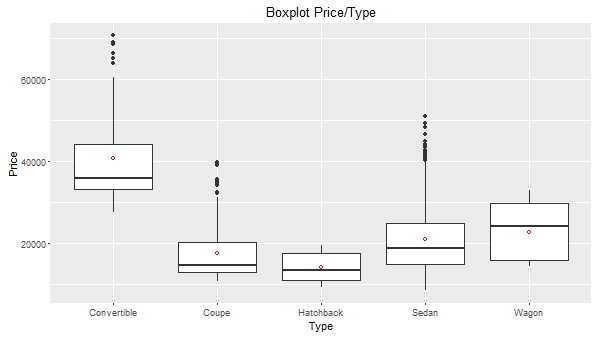


Figure 4.3

## Outliers Treatment:

* We saw that Price is not normally distributed. So, we can obviously plot a boxplot to see where each quantiles lies.
* Figure 4.4 shows a boxplot for Price. We can see that 95% quantile is around $45000. We will consider Prices more than $50000 as outliers.
* Yet we do not know what Make or Type or Model of the cars has price more than $50000.



Figure 4.4

## Finding Outliers:

* To find which Make of the car has Price > 50000, following code can be used:

> which.max(table(GM[GM$Price >50000,"Make"]))

Cadillac

* So, we can say that all the cars with Price >50000, most of them are Cadillac Made. We can also plot a bar graph for the same which is shown in Figure 4.5

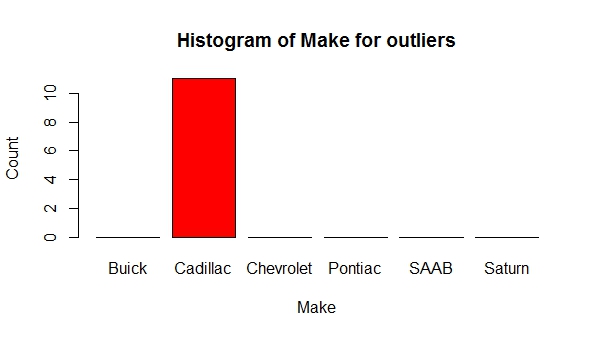


Figure 4.5

* Now, to find which Type of cars which are Cadillac made has Price>50000, following code can be used:

> which.max(table(GM[GM$Price >50000 & GM$Make == "Cadillac","Type"]))

Convertible

* Now we know that our outliers are Cadillac Made and 10 them are Convertibles and only 1 is

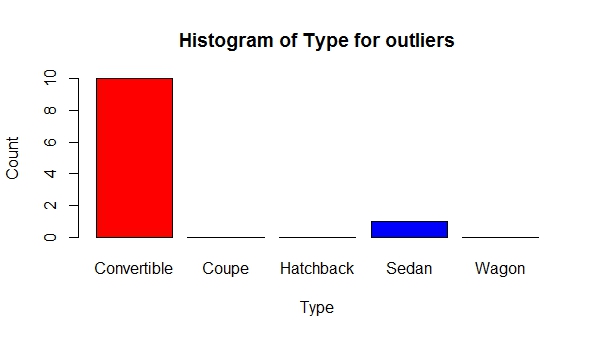
Sedan. We can see that in Figure 4.6 also.

Figure 4.6

## Another way to find outliers:

* Now all of the process done above can be summarized from one graph which is shown in Figure 4.7
* We can say that only Convertibles and Sedan goes over $50000 and also we can make out that

all of them are Cadillac made because of the (Gold) color.

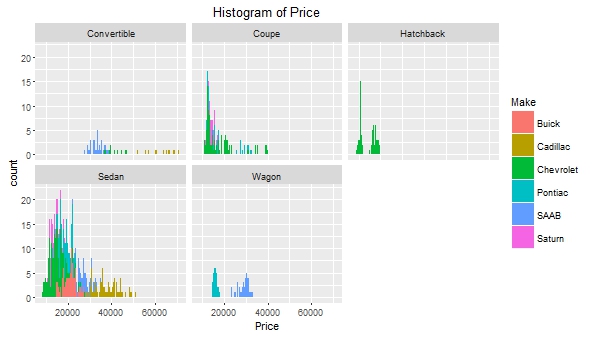


Figure 4.7

## Detailed Analysis of the Data:

* Figure 4.8 shows where and how price is distributed among Make and Type both variables.
* Excluding outliers Chevrolet Convertibles and Cadillac Sedan has the highest selling price.
* Also, Pontiac Wagon, Saturn Coupe and Sedan and Chevrolet Hatchback has the lower selling price.

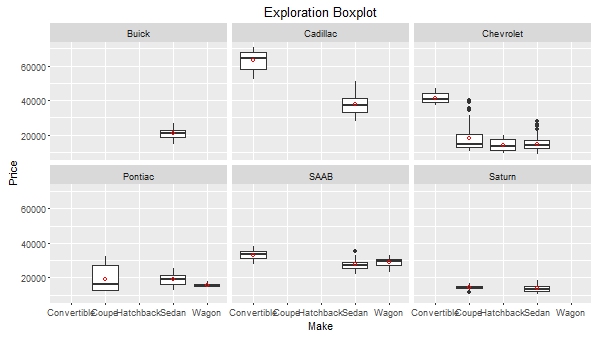


Figure 4.8

## Correlation:

* Correlation between the only numeric variable - Mileage and Price **r =** **- 0.1430505**
* From the Pearson’s coefficient above we can say that there is a weak negative relationship between Price and Mileage.
* As Mileage increases, Price will decrease very slowly.
* How Price will decrease will Mileage, we can find out in Linear regression step ahead.

# Regression Analysis

## Linear Regression:

* Since, we have only one Numeric variable (Mileage) we will perform linear regression model on Mileage over Price.
* Figure 5.1 and 5.2 show output of linear regression performed.

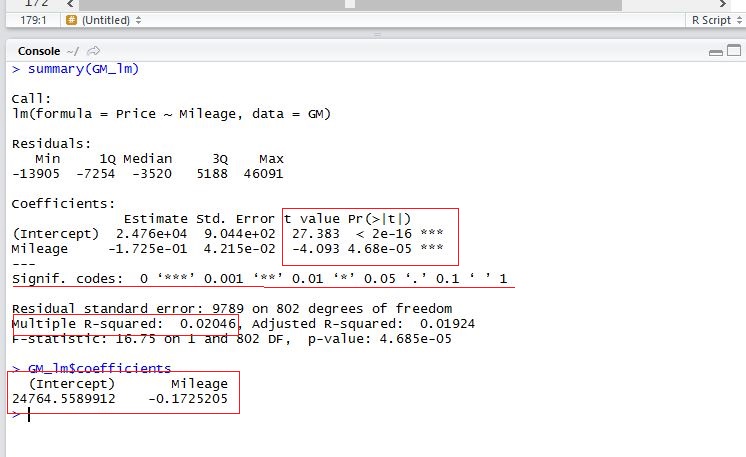


Figure 5.1

* The Pr value is significantly low for Mileage as shown above.

**Pr(>|t|) = 0.0000468**

So, wecan reject the NULL hypothesis that Price does not depend on Mileage.

* But, when we look at the r-squared value it is significantly low.

**Multiple R – Squared = 0.02046**

This says that Mileage explains only 2% variability in Price which is significantly low.

* Equation for the fitted line:

**MPG = - 0.1725 \* (Mileage) + 24764.55899**

## Scatter Plot with fitted line and 95% confidence interval:

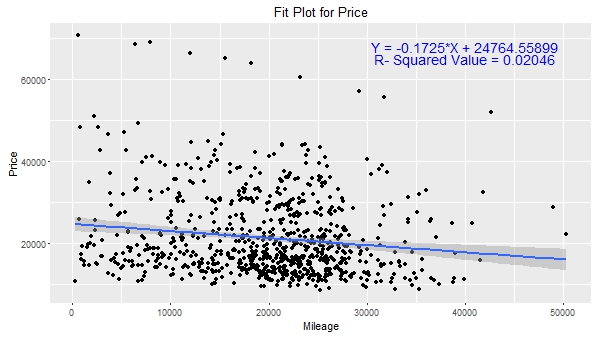


Figure 5.2

## Linear Regression Summary:

* Regression analysis on Mileage shows that it is not very significant to derive the car price.
* Low R squared value (2%) tells us that Mileage only shows 2% variability on Price which is not what I assumed. I assume that Mileage must have a high significance on Price but it doesn’t. That means other variables must have large significance on it. To find out which we have perform ANOVA on our data.

# ANOVA (Analysis of Variance)

* We already figured out that Mileage does not affect Price much. So, there must be some variables out of other 10 categorical variables which makes high influence on Price variable.
* Now, Price variable is influence by which categorical variable and how, can be analyzed using ANOVA.
* Figure 6.1 shows output of ANOVA procedure in r.

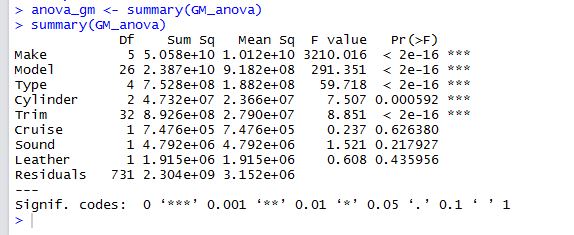


Figure 6.1

* To find out which one has significant influence on Price, r has given us Significance codes are shown beside Pr value. S
* Significance codes can be interpreted as follows:
  + \*\*\* = High significance
  + \*\* = Medium significance
  + \* = Low significance
  + . = Very low significance
  + blank = No significance
* Here, we have \*\*\* for Make, Model, Type, Cylinder and Trim which means they have high significance among all 10 categorical variables.
* I am not sure why Liter variable is not there, I tried various things on it but could not figure out. But, I ran anova on Liter variable separately but it has no significance (blank). So, that we are good to go further.
* Now, we will run anova for only those variables who makes significant influence on Price.
* Figure 6.2 is the output of the anova for only significant variables.
* We can see in the output that Cylinder is not as significant as Make, Model and Type when we consider only significant variables.
* We can say that Cylinder is less significant that Model, Make and Type but more significant than all other variables like Doors, Leather, Sound, Trim, Cruise and Liter.

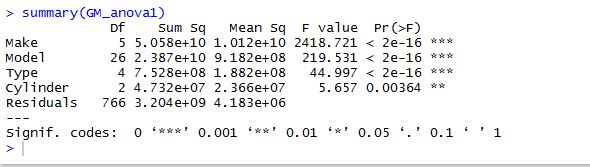


Figure 6.2

## ANOVA with the factors:

* Till now we know that Make is the most significant influencer on Price but we need to figure out which make out of all makes affects the most.
* Now, here we will consider all the variables again to see which level has how much influence on Price of the car.
* Figure 6.3 shows that considering all variables will exclude some levels because of the singularity. So, that is why we have to summarize in a way that we have minimum singularity so that we can analyze mostly every levels of each variables.

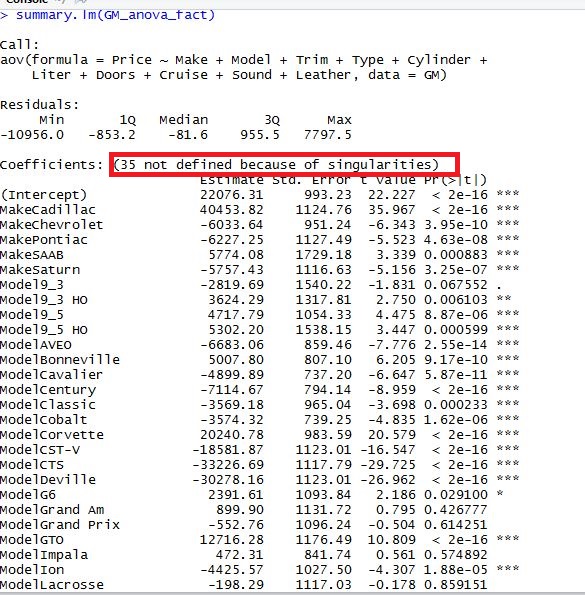


Figure 6.3

* Figure 6.4 shows how each Make, Type, Cylinder, Cruise, Leather and Sound affects Price.
* As seen in the figure below, we can say that all Makes are significant except Saturn. Also, Saturn has the higher Pr value than all the other Make which makes sure that Saturn is the least significant Make that can affect Price variable.
* Figure 6.4 also shows that Leather(1) and Cruise(1) has no significance on Price variable. Which means If you have Leather seat cars or if you have Cruise control in the car, it will not significantly give you higher or lesser Price for the car.
* Coming to Liter variable, we can see how Liter1.8 is very significant while Liter2 is not at all significant.
* Overall **R-squared value** for the model is **92.33%.**

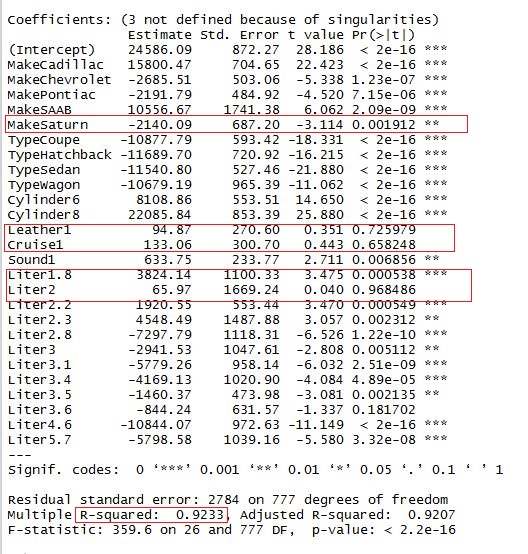


Figure 6.4

* Figure 6.5 shows significance of Model over Price.
* Figure 6.5 shows how ModelGTO,Model9\_3,Model9\_3 HO and Model9\_5 have significance in ascending order. ModelGTO is the one of the least significant Model while Model9\_5 is one of many highly significant models.

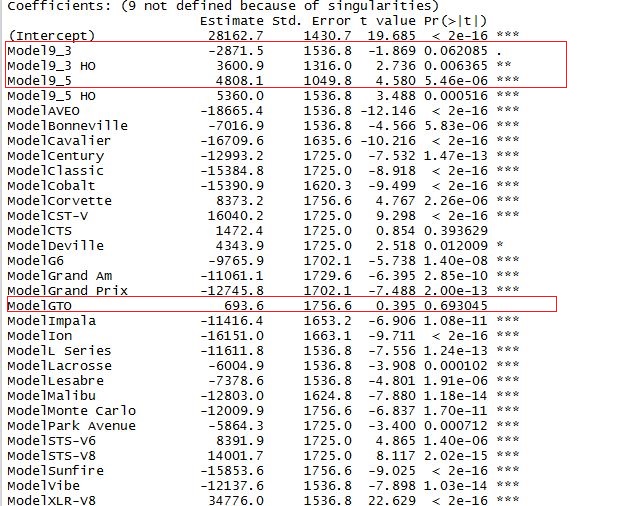


Figure 6.5

* Figure 6.6 shows significance of Trim over Price.
* We can see the highlighted box shows some of the most significant Trims as well as the least significant Trims of the car which will affect car Price.
* Overall **R-squared value** for Model is **97%.**

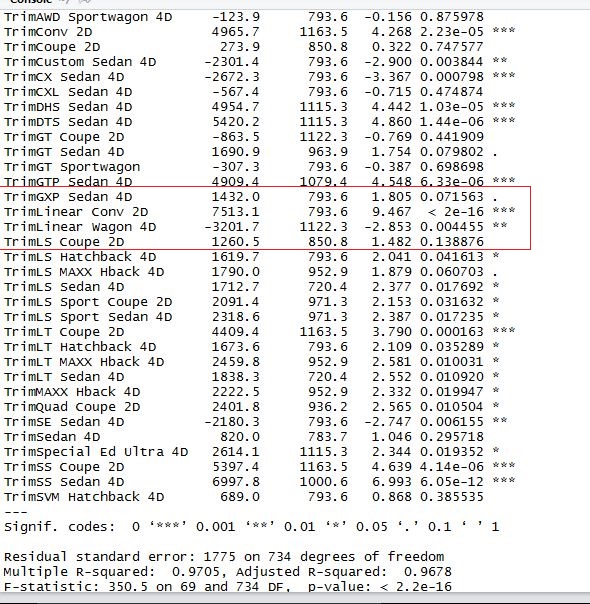


Figure 6.6

* Figure 6.7 show how number of doors affects Price of the car.
* We can see here that Doors4 (4 Door car) has a higher significance. From which we can say that if number of doors of the car is 4 than it will affect car Price highly.
* But, if you notice we have **R-squared value** of just ~**2%** which is not that significant. So, we can not actually the statement above stated.
* Finally, because of the lower R-squared value, I think Doors does not have much significance on Price.

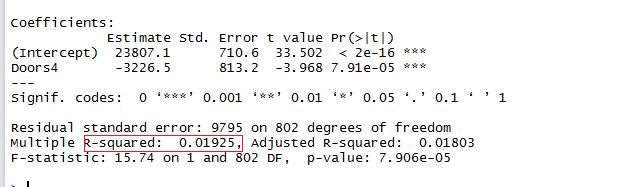


Figure 6.7

## ANVOA with factors summary:

* In summary, if we three model mentioned above, we overall get a nice R-squared value because third model (Doors) has only one variable while other two has more than two variables which very high R-squared value.
* From analysis done above, we can say that Make has the highest influence while Leather, Cruise, Sound and Doors has very less influence on the Price variable.

## Interaction effects in ANOVA:

* Interaction effect can be found between different variables of a data set.
* Null hypothesis for interaction would state that the difference between columns are the same for each row and the difference between rows are the same for each column.
* Statistically saying, the regression line of those two variables will be the same if we get a Pr value as 1.
* Now, trying interactions for Model, Make and Type, we can get output as shown in figure 6.7
* From the output we can see that Model:Type is not at all significant and Make:Type is somewhat significant to determine the Price value.
* But here we have all three variables highly significant on its own. So, that it would be better that we do not include interactions in our model further.

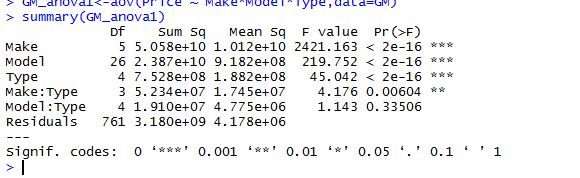


Figure 6.7

# Tukey – Kramer Test

* Tukey-Kramer test will be give us a 95% confidence interval values(lower and upper) and a

“p-adj” value.

* If p-adj value is 0 than it means that there is a lot of difference between those two particular levels of the same variable. While when it is 1, it says that both levels are statistically the same.
* So, if both are statistically the same, they do not make a high influence on response variable.
* Thus, if p-adj is 0, both levels are high influencers and vice-versa.
* We will be running Tukey-Kramer test for only Make, Model, Type and Cylinder variables as those are the most significant variables among every independent variable.
* Figure 7.1 shows the output of Tukey-Kramer test for Make.
* All the p-adj values are 0 here.
* We can say that none of the Make is statistically same and all of them (every type) will have different effect on Price with respect to different make type.

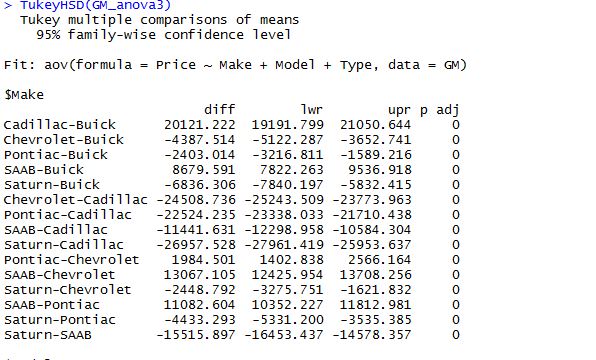


Figure 7.1

* Figure 7.2 shows output of Tukey-Kramer’s test for Type:

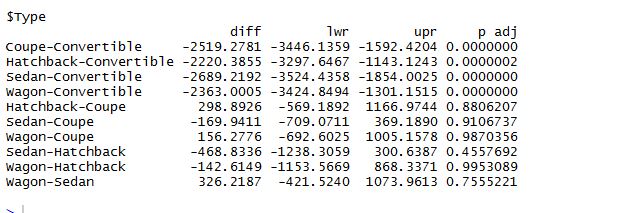


Figure 7.2

* We can see in the output that all the Convertible Type of cars as p-adj value as 0.0000, which means that if convertible car Price will be so much different(Higher) than all the other types.
* We can also see from here that Hatchback has p-adj value near to 1 for Sedan, Wagon and Coupe which means hatchback will not much significance on price because it is the same as the other three.
* Between Sedan, Wagon and Coupe, all of them has p-adj value near 1 except Sedan-Hatchback. Which means that Sedan make an influence on Price as given the price of hatchback (base level).
* Figure 7.3 shows output of Tukey-Kramer’s test for Cylinder:

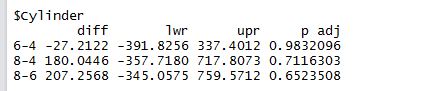
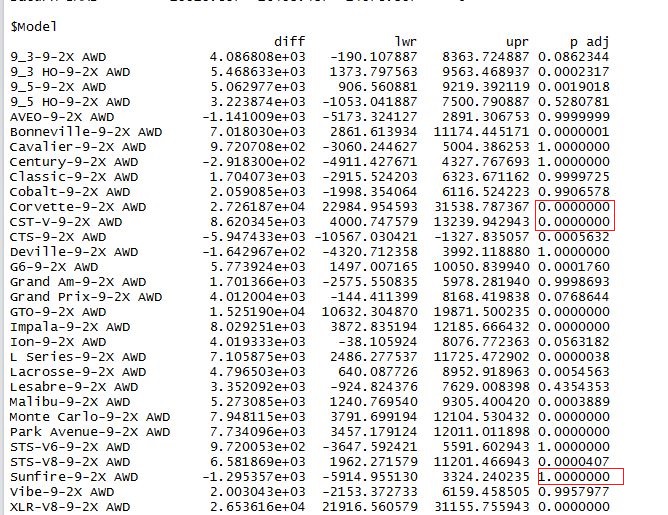


Figure 7.3

* We can see here that, Cylinder 8 has less p-adj value than 6 and 4.
* This means that if we have 8 Cylinders in the car it will make some(not high because it is still not close to 0) difference to price value.
* Figure 7.4 shows output of Tukey-Kramer’s test for Model:
* We have so many combinations here but for the sake of the explanation I have cropped the output and highlighted the one I am going to explain.
* So, Corveetee-9 -2x AWD and CST-V-9 and 2x AWD has p-adj value of 0 which means there is no difference between their means and they make no significant influence on Price variable.
* But, 2x-AWD will make a difference over Sunfire-9 Model. So, means of those two varies a lot which makes those combination of Models significant.
* Again, this figure is just a snapshot of sample. We have 32 levels of Models which will calculate almost 496 combinations of levels (32C 2 Combinations).
* We can go through each Model and find out if it is significant of not.
* But analyzing overall p-adj value, I can say that XLR-V8-9 , Corvette and Cavalier has the highest 0 in p-adj value which also makes sense because XLR-V8-9 is for Cadillac convertible. And of course it has higher prices than others.

Figure 7.4

# Conclusion

After performing all the regression analysis we can conclude the following results:

* Mileage is not as significant as we expected to be to predict Price of the car.
* Most significant variable is Make of the car. All Makes are significant except Saturn. Saturn is the least significant Make that can affect Price variable.
* Type of the car is also very significant. Among all Types, Cadillacs are the most significant but that falls under outliers. But, as a whole all Types have the equal significance over Price.
* Some of the Model, Trim and Liter values are very significant while some are not significant at all. So, I will say that Model, Trim and Liter have influence on Price but not as significant as Make and Type.
* Leather and Cruise has almost the same influence on Price. Not as significant as Make, Type, Model and Trim but cars with Leather seat (Leather=1) and cars with Cruise control (Cruise=1), it will not significantly give you higher or lesser Price for the car.
* Figure 8.1 shows significance of the variable with F value and Pr value in descending order:

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Significance** | **F-Value** | **Pr -Value** |
| MAKE | VERY HIGH | 3210.016 | <2e – 16 |
| MODEL | HIGH | 291.351 | <2e – 16 |
| TYPE | VERY HIGH | 59.718 | <2e – 16 |
| CYLINDER | MEDIUM | 1.64 | 0.0006 |
| TRIM | MEDIUM | 8.8 | <0.0001 |
| LITERS | MEDIUM | 5.51 | 0.0192 |
| MILEAGE | LOW | 16.75 | 4.685e -05 |
| DOORS | LOW | 7.507 | 0.000592 |
| SOUND | LOW | 1.521 | 0.2192 |
| LEATHER | VERY LOW | 0.608 | 0.4359 |
| CRUISE | VERY LOW | 0.237 | 0.6263 |

Figure 8.1