Final Project

ALY 6010

Northeastern University

Professor: Dr. Amin Karimpour

Student Name: Santhi Kiran Chavali

**Appendix:**

1. **Introduction**

* Description of the dataset
* Data Frame variables
* Data cleaning

1. **Exploratory Data Analysis**

* Descriptive statistics of numeric variables
* Finding out Outliers
* Question 1 and Answer
* Question 2 and Answer
* Question 3 and Answer

1. **Univariate analysis, Histograms, and boxplots for numeric variables and dealing with outliers**
2. **Hypothesis testing**

* One sample t-test
* Question 4 and Answer
* Question 5 and Answer
* Two sample t-test
* Question 6 and Answer
* Question 7 and Answer

1. **Correlation**

* Question 8 and Answer
* Question 9 and Answer
* Question 10 and Answer

1. **Regression analysis**

* Question 11 and Answer
* Question 12 and Answer

1. **Summary and references**

**Introduction**

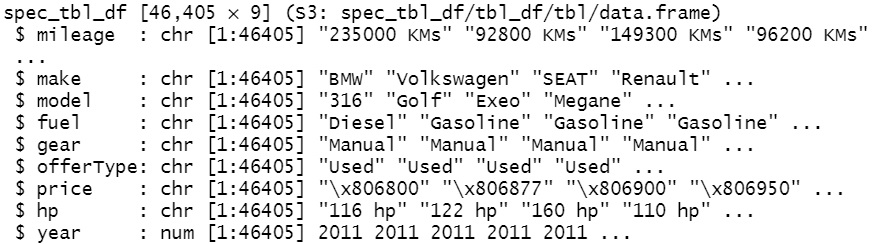
**Dataset Description:**

Autoscout24.com is the largest Pan – European Online market which offers consumers, dealers, and other finance and insurance industries a platform to buy and sell cars online. Their website contains details of different new and used cars, motorcycles, trucks, and utility vehicles. It also contains price details, ratings, and reviews that help users choose the vehicle that perfectly fulfills their requirements.

We have the dataset which contains all the details of the cars sold during the period 2011 to 2021. We analyze the dataset and help the company to improve its business. The data set is collected from Kaggle.com.

**Data Frame Variables:**

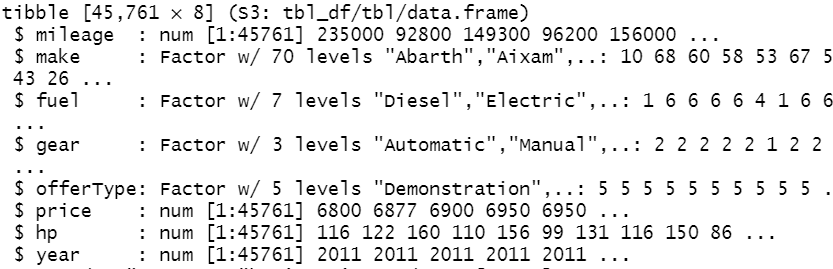
The Dataset contains 46405 records/rows and 9 variables/columns



**Data Cleaning:**

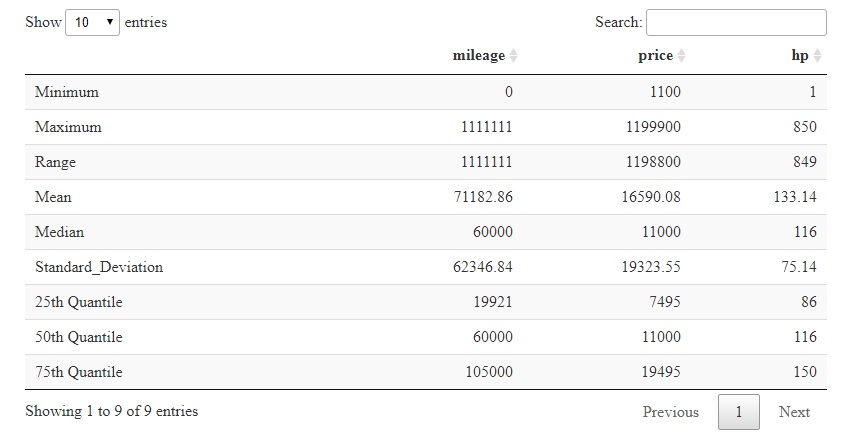
* Mileage, Price, and Horsepower columns are in character data type and also contain symbols and special characters. Hence they are converted to numeric by using gsub() and as. numeric() functions and NA values are removed from the dataset to make it useful in further analysis.
* Filtering out the blank, Others, CNG, and LPG from the fuel column as they accumulate to only 0.70% of the entire data set.

After performing the above data cleaning process we have obtained the final dataset for analysis which contains 45761 records/rows and 8 columns/variables of which four are numeric variables and four-factor variables.



**Exploratory Data Analysis**

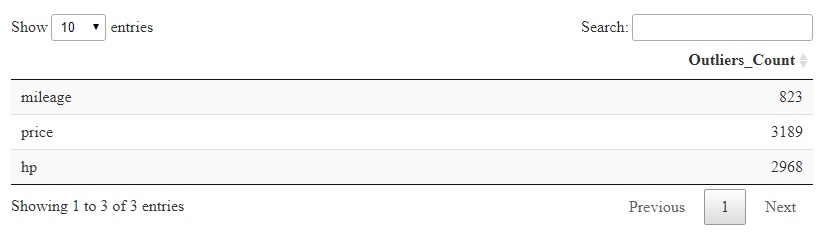
**Descriptive statistics of numeric variables**



**Observations:**

From the above Descriptive statistics table, we can observe that the mileage ranges from 0 to 1111111 with a mean of 71182.86, median of 60000, and standard deviation of 62346.84 which shows that there are new cars and as well as used cars and maximum are used car because we can see the average mileage is 71182.86. The price range of cars sold during the period 2011 to 2021 is a minimum of 1100 euros and a maximum price of 1199900 euros and the average price over the 10 years period is 16590.08 euros. The Horsepower of cars is between 1hp to 850hp with average hp of 133.14hp which shows that the average power of the engine of cars sold is 133.14

**Outliers of the numeric variables**

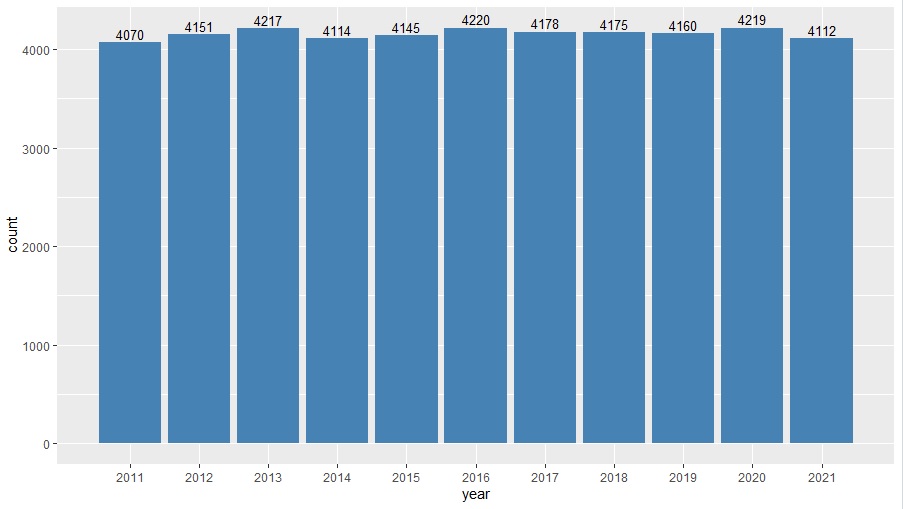


**Observations:**

We can see from the above table that all their numeric columns have outliers, The count of outliers in the mileage column is 823, the Price column is 3189 and finally, the horsepower column has 2968. These outliers will be handled in further analysis

**Question 1: How many Cars were sold year wise during the period from 2011 to 2021?**

**Answer1:** Total number of cars sold year wise

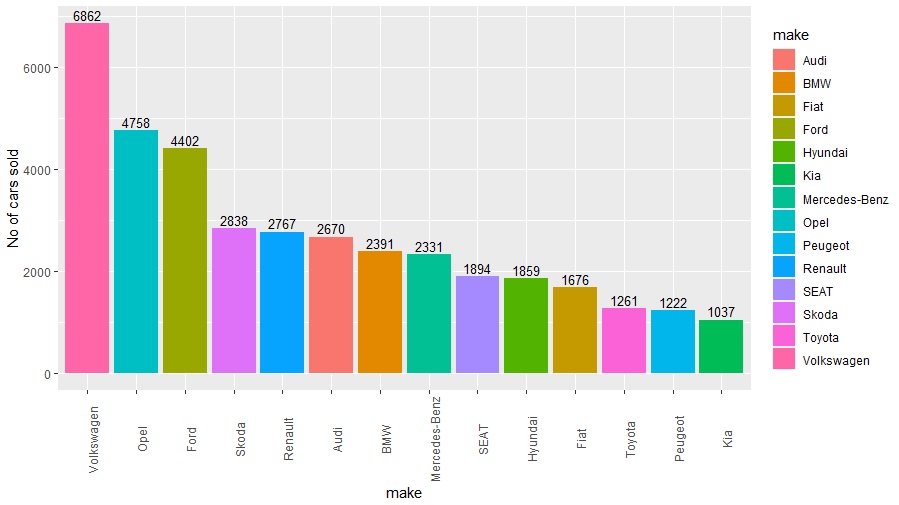


**Observations:**

From the above graph of the Total number of cars sold per year, we can observe that on average 4160 cars were sold per year through autocars24 over the period of 10 years.

**Question2: Make wise how many cars were sold during the period of 10 years**

**Answer 2:**

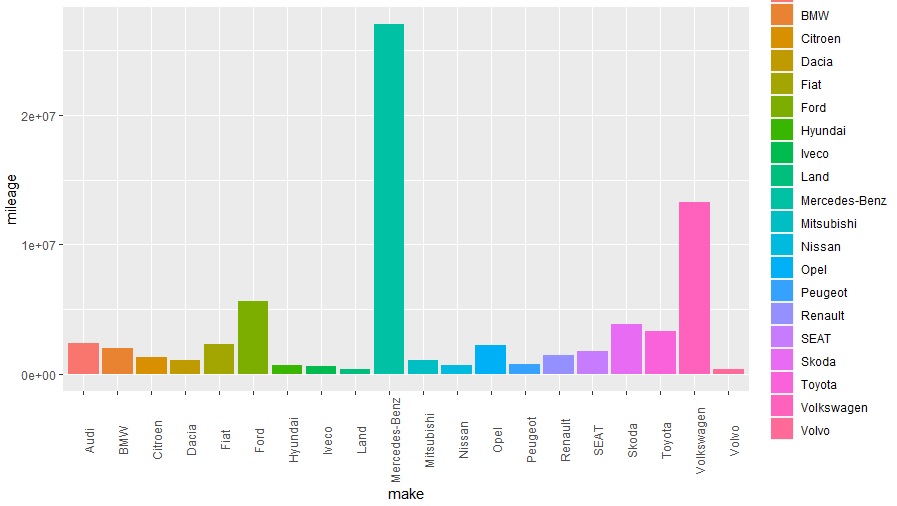


**Observations:**

From the above Bar graph pertaining to Total number of cars sold as per make we can see that Volkswagen at the top with 6862 Cars and followed by Opel make with 4758 Cars and Kia being the least number of cars sold with 1037 cars over the period from 2011 to 2021

**Question3: How many cars were sold which has driven more than 300000 Kilometres?**

**Answer**

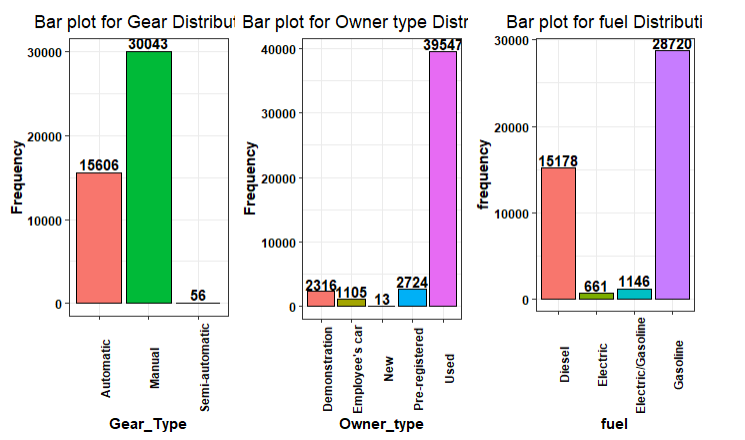


**Observations:**

The above graph pertains to the number of cars sold with mileage of more than 300000 Kilometers during the entire period. It can be observed that Mercedes-Benz make has sold the maximum number of cars which shows that even the kilometers driven is more Mercedes Benz Cars have great demand followed by Volkswagen. Volvo and Land make are the least sold cars over the period with mileage greater than 300000

**Univariate Analysis**

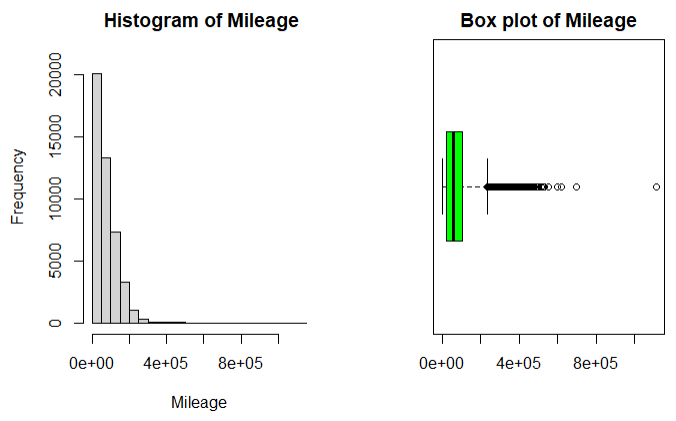
**Bar plots showing the frequency of categorical variables**

****

**Observations:**

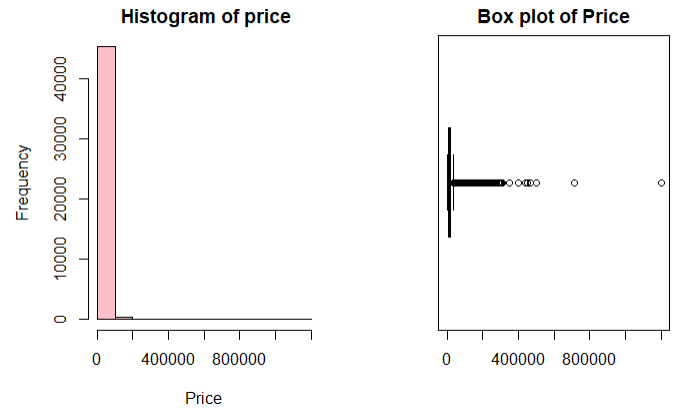
From the above bar plots, we can see that the Frequency of manual cars sold is higher when compared to automatic cars and a very less, negligible amount of Semi-Automatic Cars were sold during the period. Coming to owning type, the maximum number of cars sold are used cars with a total of 39547 out of 45705 and new cars being the least with 13 cars. Finally, Cars sold by autoscout24 during the period 2011 to 2021 maximum cars were gasoline cars with 28720 cars followed by Diesel cars with 15178 cars and electric cars being the least.

**Histograms and Boxplots of Numeric variables and dealing with the outliers**

****

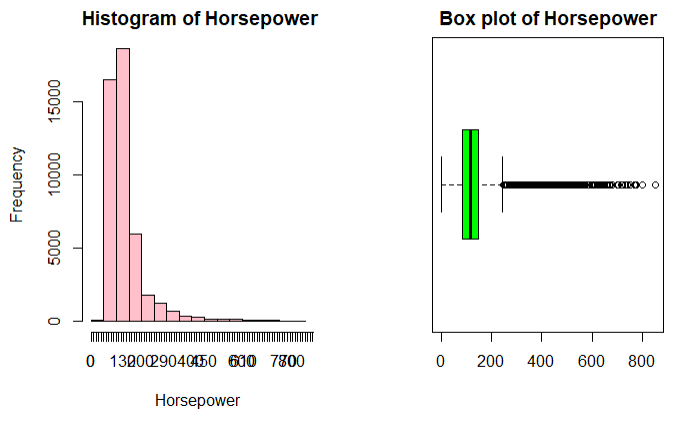
**Stats: **

**Observations:** From the above graphs we can see that the maximum car mileage is in the range between 0 to 23250. We have in total 821 outliers under the mileage variable. These outliers are the used vehicles of various make which have driven more than 232,500 Kms. So, when compared to other vehicles price of these cars will be less as they have driven more kilometers and the price also vary depending on the make. Even if the cars have driven more kilometers, based on the brand and condition of the cars prices may vary



**Stats:** 

**Observations:** From the above graphs the price of the maximum number of cars sold is in the range between 1100 Euros to 37480 Euros. We have a total of 3168 outliers under the price variable. These outliers are the costliest cars (Price > 37480 Euros) of different make that have been sold by autoscout24 during the period 2011 to 2021



**Stats: **

**Observations:** From the above graphs and stats we can see that the horsepower of the maximum cars sold is in the range of 1 to 245hp. In total, we have outliers 2956 outliers under the horsepower variable. These are the fastest cars of various make sold by autoscout24 during the period 2011 to 2021

**Hypothesis testing:**

Hypothesis testing is an act in statistics whereby an analyst tests an assumption regarding a population parameter. It is used to assess the plausibility of a hypothesis by using sample data.

A Hypothesis test involves two hypothesis

**Null Hypothesis (H0):** A null hypothesis is a type of statistical hypothesis that proposes that no statistical significance exists in a set of given observations. Hypothesis testing is used to assess the credibility of a hypothesis by using sample data.

Expressed as:

**H0:μ=μ0,**  where μ is a population mean and μ0 is some number.

**Alternative Hypothesis (HA) :** Alternative hypothesis is contradictory to the Null hypothesis. It can be expressed as

**HA:μ≠μ0.** – two-sided test, **HA:μ<μ0.** – left tailed test, **HA:μ>μ0**. – Right tailed test

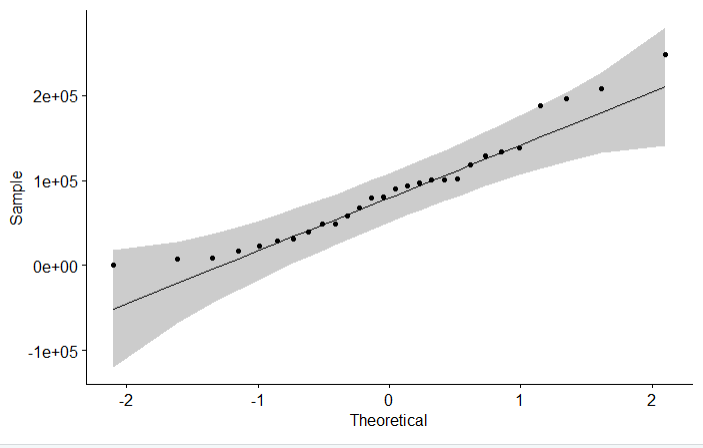
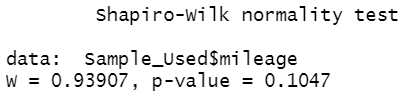
**One Sample t-test:**

One Sample t-test is used when we want to compare a sample mean with the population mean when we have no information about the variance or standard deviation of the population.

**Question 4: Whether the population means of the mileage pertaining to used cars is equal to 93750 Kms??**

**Solution:** To answer this question I have used the subset() function to separate Used cars data from the German cars dataset.

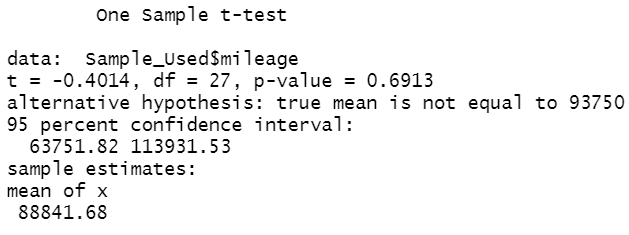
* I have taken a random sample size of 28 from the used cars and performed Shapiro-wilk normality test and Q-Q plot to know whether data is normally distributed or not



**Null Hypothesis (H0) -> μ = 93750**

**Alternative Hypothesis(HA) -> μ != 93750**

* One Sample t-test using t.test()



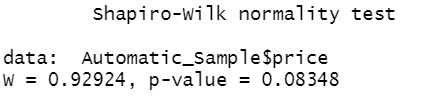
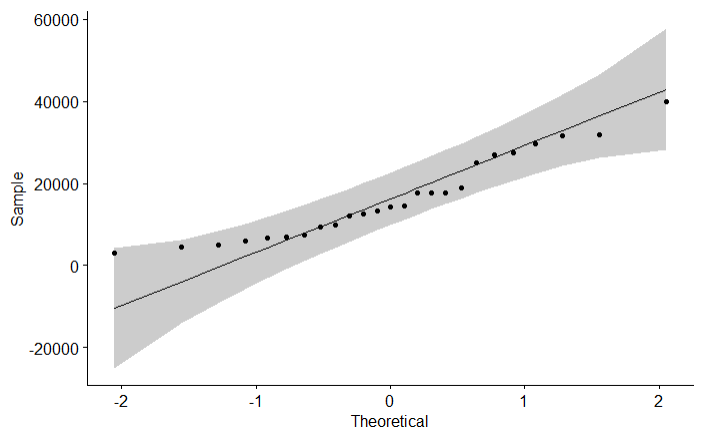
**Observations:** From the above one sample t-test we can observe that the p-value (0.6913) is greater than the significance level of 0.05. So we do not have sufficient evidence to reject the null hypothesis. Hence with a 95% confidence interval level, we can say that the population mean of the used car’s mileage can be at 93750 km

**Question 5 :** **Whether the population means of the price pertaining to automatic cars is equal to 21000 Euro at a significance level of 0.05??**

**Why 21000 Euros? – The price range of automatic cars is between 2999 Euros to 39990 Euros. So, the mean would be approximately 21000**

**Solution:**

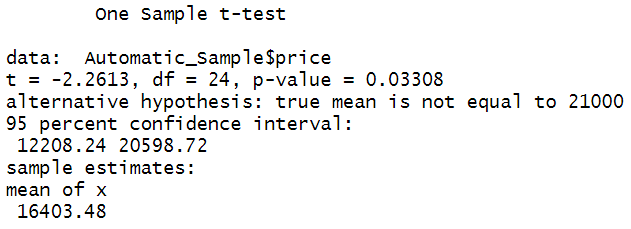
* To answer this question I have used the subset() function to separate Automatic cars data from the German cars dataset.
* I have taken a random sample size of 25 from the Automatic cars and performed Shapiro-wilk normality test and Q-Q plot to know whether data is normally distributed or not

****

**Null Hypothesis (H0) -> μ = 21000**

**Alternative Hypothesis(HA) -> μ != 21000**

* One Sample t-test using t.test()



**Observations:** From the above one sample t-test we can observe that the p-value(0.03308) is less than our significance level of 0.05. Hence we have sufficient evidence to reject the null hypothesis. So, the population mean price of automatic cars is not equal to 21000 euros. Null Hypothesis rejected**.**

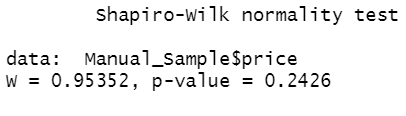
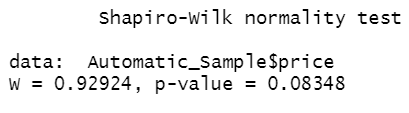
**Two Sample t-test:**

**The two-sample t-test is a method to test whether the unknown population means of two groups are equal or not. It is used to compare two process means when the data is having one nominal variable and one measurement variable.**

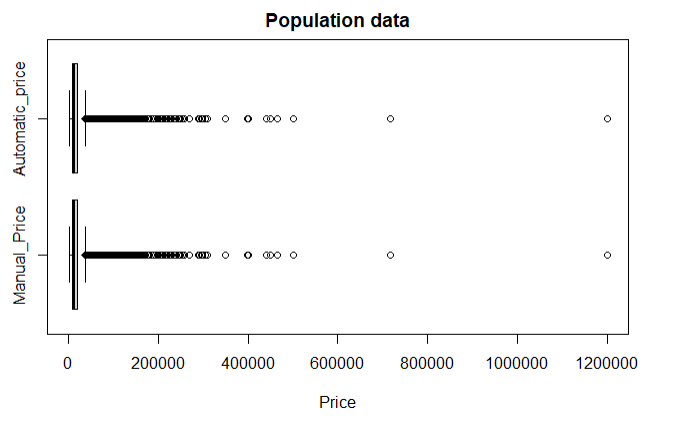
**Question 6: I want to know whether the average price of Manual cars is equal to the average price of Automatic cars?**

**Solution:**

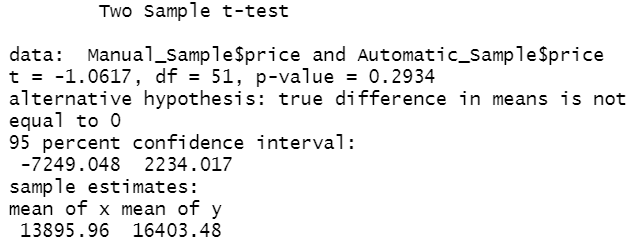
* To answer this question I have to create two separate datasets for automatic cars and manual cars using subset function
* I have collected a sample size of 25 automatic cars and 28 manual cars and checked the normality using Shapiro-wilk normality test

** **

* **Used box plot to check the standard deviation of both population**

****

* **Null Hypothesis (H0):** Average Price of Manual cars is equal to average price of the Automatic cars, **H0:μ1=μ2**
* **Alternative Hypothesis(HA):** Average Price of Manual cars is not equal to average price of the Automatic cars **HA:μ1≠μ2**
* **Significance level = 0.05**
* Conducting two sample t-test using t.test() function

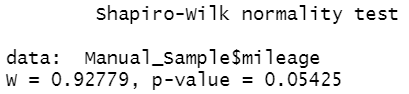
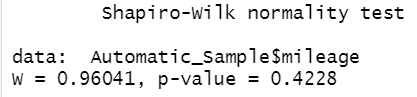


**Observations:** From the above two sample t-test we can see that the p-value (0.2934) is greater than the significance level of 0.05. Hence we fail to reject the null hypothesis. We do not have sufficient evidence to show that the average price of manual cars and automatic cars is equal.

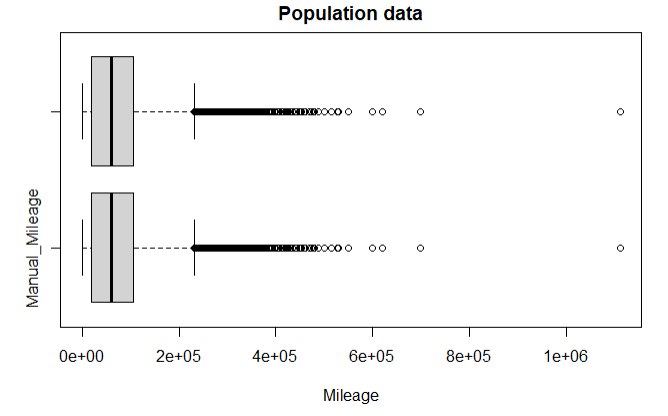
**Question 7 : I want to know whether the average mileage of Manual cars is equal to the average mileage of Automatic cars?**

**Solution:**

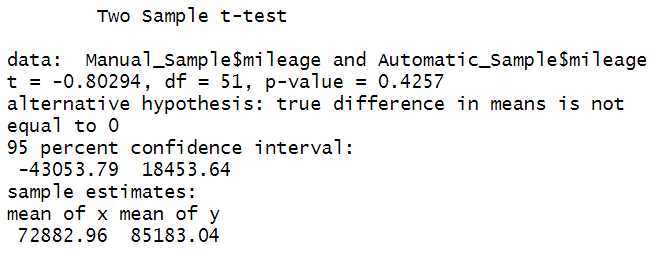
* I have collected a sample size of 25 automatic cars and 28 manual cars and checked the normality using Shapiro-wilk normality test

** **

* **Used box plot to check the standard deviation of both population**

****

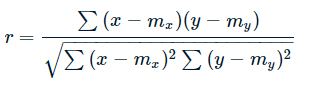
* **Null Hypothesis (H0):** Average Mileage of Manual cars is equal to average mileage of the Automatic cars, **H0:μ1=μ2**
* **Alternative Hypothesis(HA):** Average Mileage of Manual cars is not equal to average mileage of the Automatic cars **HA:μ1≠μ2**
* **Significance level = 0.05**
* Conducting two sample t-test using t.test() function

****

**Observations:** From the above two sample t-test we can see that the p-value (0.4257) is greater than the significance level of 0.05. Hence we fail to reject the null hypothesis. We do not have sufficient evidence to show that the average mileage of manual cars and automatic cars is equal.

**Correlation:**

A correlation test is used to know how strong the relationship between the variables. The correlation coefficient(r) is a way to put a value to the relationship. It ranges from -1 to +1 where -1 indicates a strong negative correlation, +1 indicates a strong positive correlation and finally, 0 indicates there is no correlation between the variables. It can be calculated y using the formula:

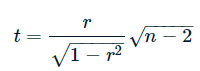


mx and my are the means of x and y variables.

The p-value (significance level) of the correlation can be determined :

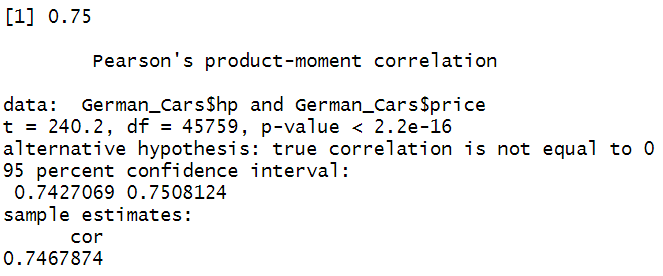
by using the correlation coefficient table for the degrees of freedom : df=n−2, where n is the number of observation in x and y variables.

or by calculating the t value as follow:



**Question 8: In General the more power of the engine the price will be more . Lets, Find out the correlation between the Horsepower and Price in the Germany cars dataset**

**Answer:**

* **Null Hypothesis:** There is no Correlation between horsepower and Price, r = 0.
* **Alternative Hypothesis:** Positive correlation between horsepower and Price, r>0.
* Conducting the correlation test using cor.test() function and visual representation.
* ****

|  |  |
| --- | --- |
|  |  |

**Observations:**

From the above correlation test, we can see that the p value is less than the significance level of 0.05 and the correlation coefficient between the variables Horsepower and Price is 0.74 by which we can say that we have strong evidence that there is a strong positive relationship between Horsepower and Price. It can also be seen visually using correlation charts and scatter plot that as the horsepower increases the price of the car increases.

**Question 9: Find out the price change over the years for both automatic cars and manual cars and the difference in prices between automatic cars and manual cars using correlation.**

**Solution:**

* To answer this question I have created two separate data sets for automatic cars and manual cars.
* To conduct a correlation test I have used Price and Year variables from both data sets.
* Using cor.test() function conducted correlation test for both automatic cars and Manual cars

|  |  |
| --- | --- |
|  |  |
|  |  |

**Observations:**

From the above correlation test conducted between the variables price and year for both Automatic cars and manual cars, we can observe that there is a positive correlation between the price and year in both cases as the p-value is less than the significance level of 0.05 in both the cases and the correlation coefficient is positive. However, we can observe that the correlation coefficient in the case of manual cars(0.56) is more than that of automatic cars(0.37). By which we can say that as the year increases there is a decent amount of rise in the price of manual cars when compared to automatic cars

**Question10: Using a correlation table or correlation chart, Show the relationship between Mileage(kilometers driven), Price and year**

**Solution:**

**I have used cor() and corrplot for displaying the correlation between the numeric variables.**

|  |  |
| --- | --- |
|  |  |

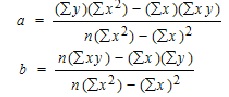
**Observations:**

By observing the above correlation chart and table pertaining to mileage, year, and price we can say that there is a moderate negative relationship between mileage and price -0.30 which is true because as the kilometers driven increase price of the vehicle decreases. We can also observe a decent positive relationship between the year and price 0.41 by which we can say that as the years increase the price of the vehicle increase. And, there is a strong negative correlation(-0.68) between the year and mileage as the vehicles sold in recent years has a chance that the vehicle is less traveled

**Regression Analysis:**

Regression analysis is a powerful statistical method that allows you to examine the relationship between two or more variables of interest. It examines the influence of one or more independent variables on a dependent variable and also provides detailed insight that can be applied to further improve products and services.

The linear regression equation can be written as Y= a + bX, Where Y is the dependent variable, X is the independent variable, b is the slope of the line and a is the y-intercept. Values of a and b can be find by using the following formulas.



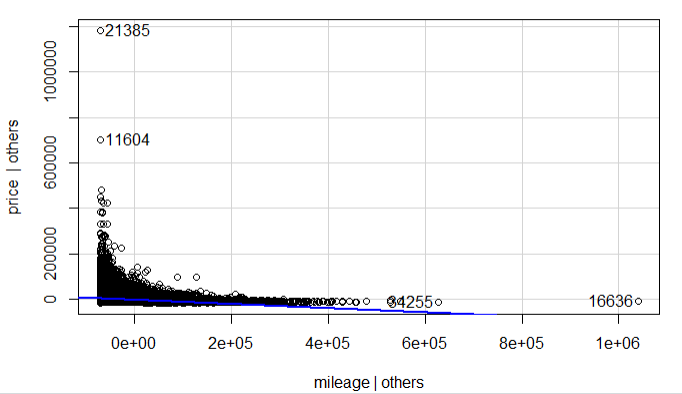
**Question 11: Find out the best regression model to predit the price from the German cars dataset.**

**Solution:**

* To Answer this question First I have tried to find the linear regression model using single variable Mileage to predict the price. In the assumption that the more kilometers driven the price will be less.
* To find the regression model to predit the price(dependant variable) using the mileage as independent variable I have used lm() function in r

|  |  |
| --- | --- |
|  |  |

**Scatter Plot:**



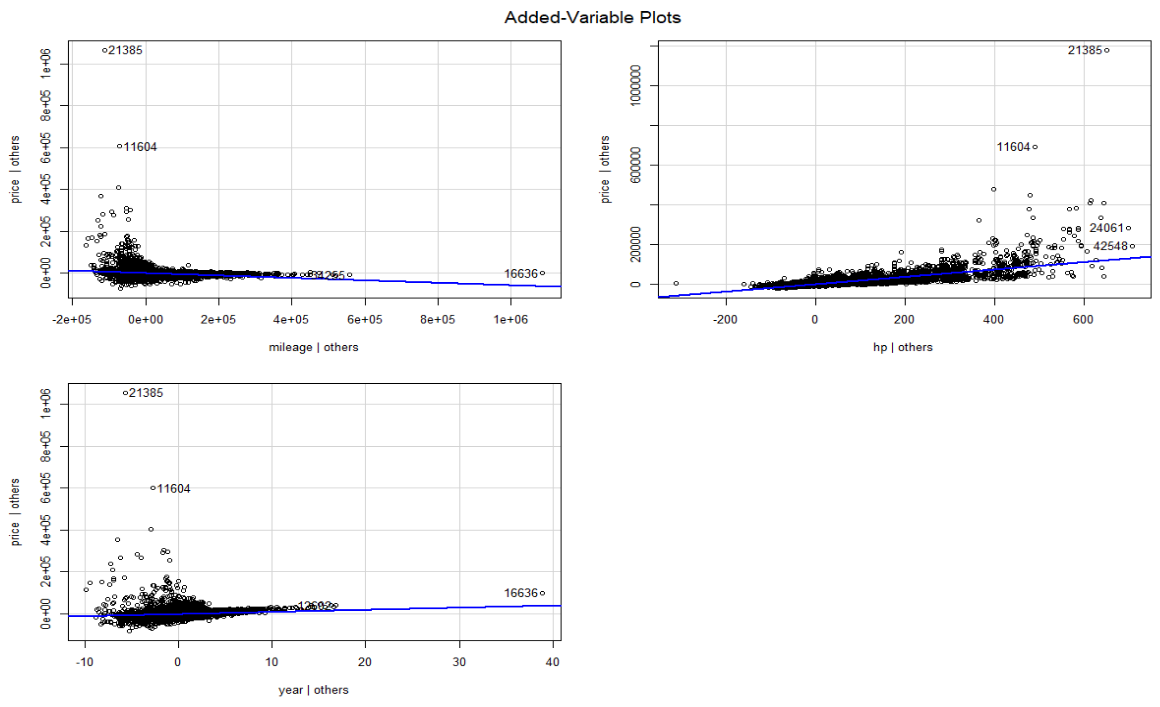
**Observation:**

From the above-conducted regression test, we can see that the intercept is 2.32e+04 and the coefficient is -9.398e-02 and the p-value is significant but the r2 is very less. In other word, very less variability is observed that this model is not the best fit. It can also be visually observed by looking at the regression table and scatter plot.

**From the above trial. We used mileage as a predictor variable, we could not find the best regression model. Hence I tried using all the numeric variables to predict the price. Hence conducted a Multiple linear regression model. The result as follows:**

|  |  |
| --- | --- |
|  |  |

**Scatter plot**

****

**Observations:**

From the multiple regression model conducted above we can observe that by using all the numeric variables mileage, year and hp as predictor variables and price as reponse variables, p value is significant in all cases and the r2 is 0.65 by which we can say that this model is a best fit to predit the price of the cars in the german cars dataset. It can also be visually observed by looking at the above regression table and scatterplots.

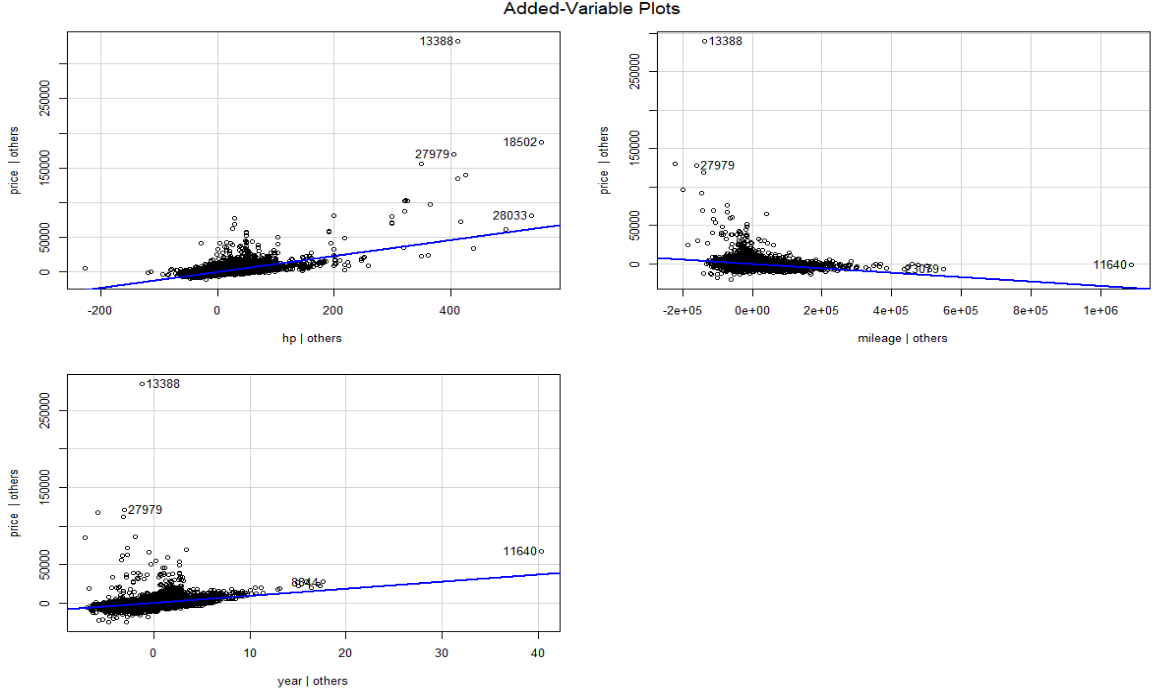
**Question 12: Build a regression model to predict the price of manual cars by taking all other variables as independent variables**

**Solution:**

Conducting a multiple linear regression model using lm() function

|  |  |
| --- | --- |
|  |  |

Scatter plot:



**Observations:**

From the multiple regression model conducted above, we can observe that by using all the numeric variables mileage, year, and hp as predictor variables and price as response variables to predict the price of the manual cars, the p-value is significant in all cases and the r2 is 0.65 by which we can say that this model is the best fit to predict the price of manual cars. It can also be visually observed by looking at the above regression table and scatterplots.

**For both regression models above it is observed that horsepower plays a major role in predicting the price of the overall dataset and manual cars. However, including the other variables like mileage and year shows the best fit to predict the price of vehicles**

**Summary:**

The dataset from the autoscout24 website contains various details related to the cars sold during the period from 2011 to 2021 like Price, engine power, Mileage, fuel type, etc. In total the dataset contains 46405 records with 9 variables. The data cleaning process is done on the dataset by changing the data types and applying the correct data types to all the variables, finding any NA values in the dataset and removing them, and filtering the data to get the required final dataset. Data analysis is done by finding the descriptive statistics of the numeric columns in the dataset and using a few graphs like bar graphs and Scatter plots to understand the data. Univariate analysis is conducted to see observe the categorical variables and used a few box plots and histograms to analyze numeric variables and deal with the outliers. In addition, hypothesis test is conducted using one sample t-test and two sample tests. Finally, correlation and regression analysis is conducted on various variables and made conclusions on the outputs. It is also represented using correlation charts and scatterplots.

References:

1. Germany Cars R. (n.d.). Kaggle.com. Retrieved October 5, 2022, from <https://www.kaggle.com/code/jcraggy/germany-cars-r/notebook>
2. Company. (n.d.). AutoScout24. Retrieved October 5, 2022, from <https://www.autoscout24.com/company/?genlnk=foot&genlnkorigin=com/all/all/home>
3. Data Carpentry contributors. (2019). Data visualization with ggplot2. Datacarpentry.org. <https://datacarpentry.org/R-ecology-lesson/04-visualization-ggplot2.html>
4. STHDA. (2019). Correlation test between two variables in R - easy guides - wiki - STHDA. Sthda.com. http://www.sthda.com/english/wiki/correlation-test-between-two-variables-in-r