**Craigslist Used Cars Listing Analysis**

**IE-6200: Engineering Probability and Statistics** - **Project Report**

**Introduction :**

Thinking of selling your car ? Or want to buy a secondhand car at a cheap price ?

These questions make us think that it sure is difficult to figure out the best price for a used car. A used vehicle is priced based on its model, manufacturer, transmission type, size, condition and many more factors. To estimate price of a vehicle correctly we might look for similar condition vehicle and then fine-tune the price to best match our state. To do this you will require both domain and current market knowledge. This can be tedious when looking for a large number of vehicles.

Hence, we thought of doing some form of analysis on real-world used cars data in order to save some time and mental effort in this decision-making process. We will also try to estimate a price value for the used cars, based on historical data.

**Objective :**

The goal of this project is to perform exploratory analysis (using R) on a used cars dataset by plotting a number of graphs to give information about existing patterns, by using inferential statistical methods to perform hypothesis tests and finagling using simple linear regression to estimate the price of a car.

We will be using Craigslist Used Cars dataset available on Kaggle to achieve our purpose.

**Data Description :**

The dataset from year 1910 to 2021, contains 458260 rows (each representing a single listed used car) and 26 columns (variables describing the listed item). We initially used Excel to trim the dataset by removing the unwanted columns and removing rows nonnumeric data present in numeric columns. This reduced our dataset to 458214 rows and 16 columns. The column description is as below

1. ID — Unique ID to each listed car.
2. Price — Price in US dollar.
3. Year — The year in which the car was manufactured
4. Manufacturer —43 unique manufacture of automobiles.
5. Model — The model of the car. Like sierra classic 2500hd.
6. Condition — The condition of the car; excellent, good, fair, like new, salvage, new.
7. Cylinders — The number of cylinders in the car engine ranging from 3 to 12.
8. Fuel — There were five types of fuel, ‘diesel’, ‘gas’, ‘electric’, ‘hybrid’ and ‘other’.
9. Odometer — This is the distance that the car has traveled after it being bought.
10. Title\_Status — The cars also had 6 types of statues; ‘clean’, ‘lien’, ‘rebuilt’, ‘salvage’ , ‘parts only’ and ‘missing’.
11. Drive — There are 3 types of drive: ‘4WD, ‘FWD’ and ‘RWD’.
12. Transmission — Transmission Type like manual or automatic
13. Type — This feature identifies if a vehicle is a SUV or a mini-van. There 13 unique values in this feature.
14. Size — Size of the car
15. Paint\_Color  — Color of the car
16. State — The state is political territory and is represented in short form in the data set. Like “fl” is used for the state of Florida.

**Preprocessing :**

Further preparation and cleaning of the raw data for analysis was done in R.

Step 1 – Reading data in R

Step 2 – Removing rows with Null values

Step 3 – Identifying Outliers for Price and Odometer columns

Step 4 – Trimming Price and Odometer values to remove outliers

Step 5 – Converting Year to proper Date format

Step 6 – Merging States dataset to get the unabbreviated versions of the listed states

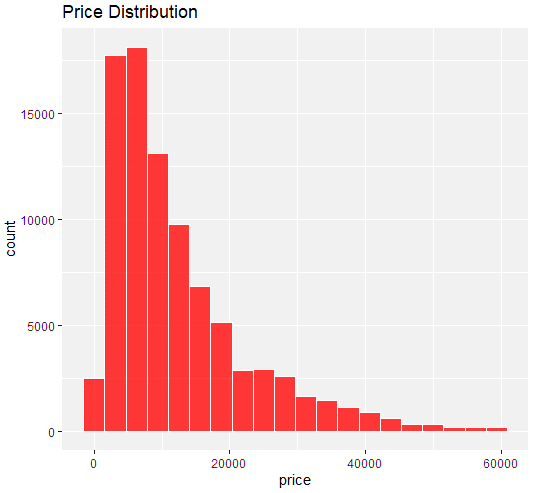
Step 7 – Breaking down price values into 6 bins for better understanding

After the above steps we were left with a clean dataset of 88631 rows and 16 columns ready to perform our next steps of analysis.

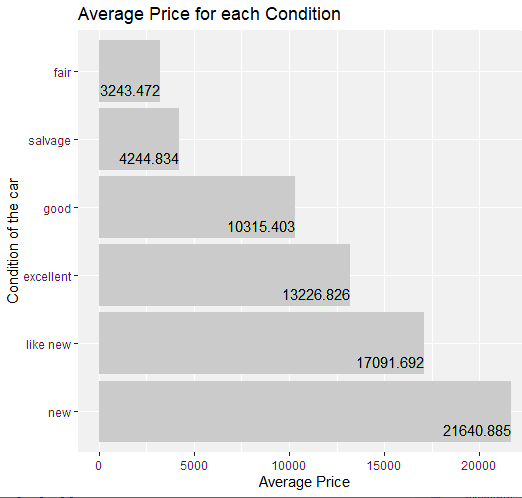
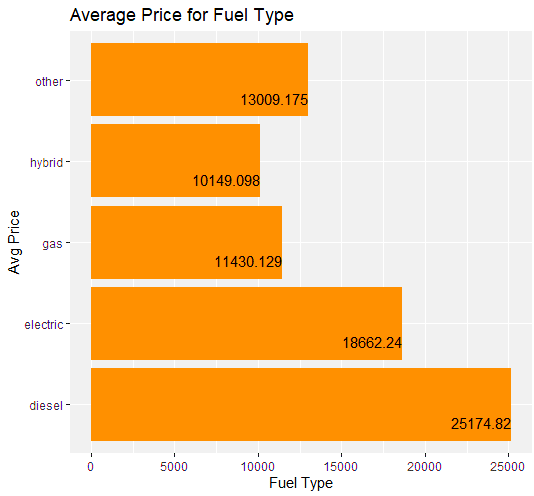
**Visualizations :**

We made a lot of visualizations for exploratory purposes and got some useful insights from each. Few of them are displayed here.

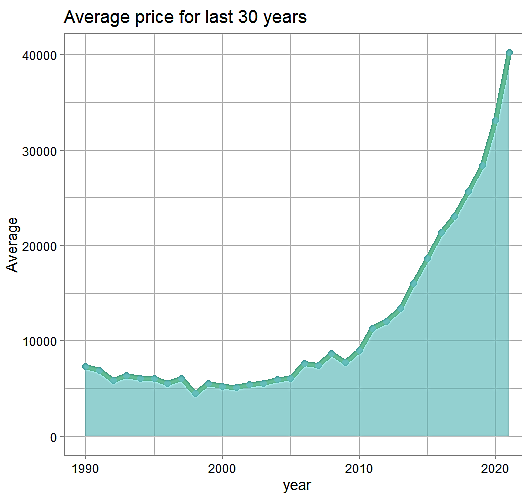
The below plot is a histogram showing the price distribution structure. We can see that the count of vehicles with lesser price is higher. One of the reasons for such a plot could be that people believe cars with lesser price are sold faster or also it could mean that smaller cars (which are cheaper) are more preferred when buying secondhand cars.



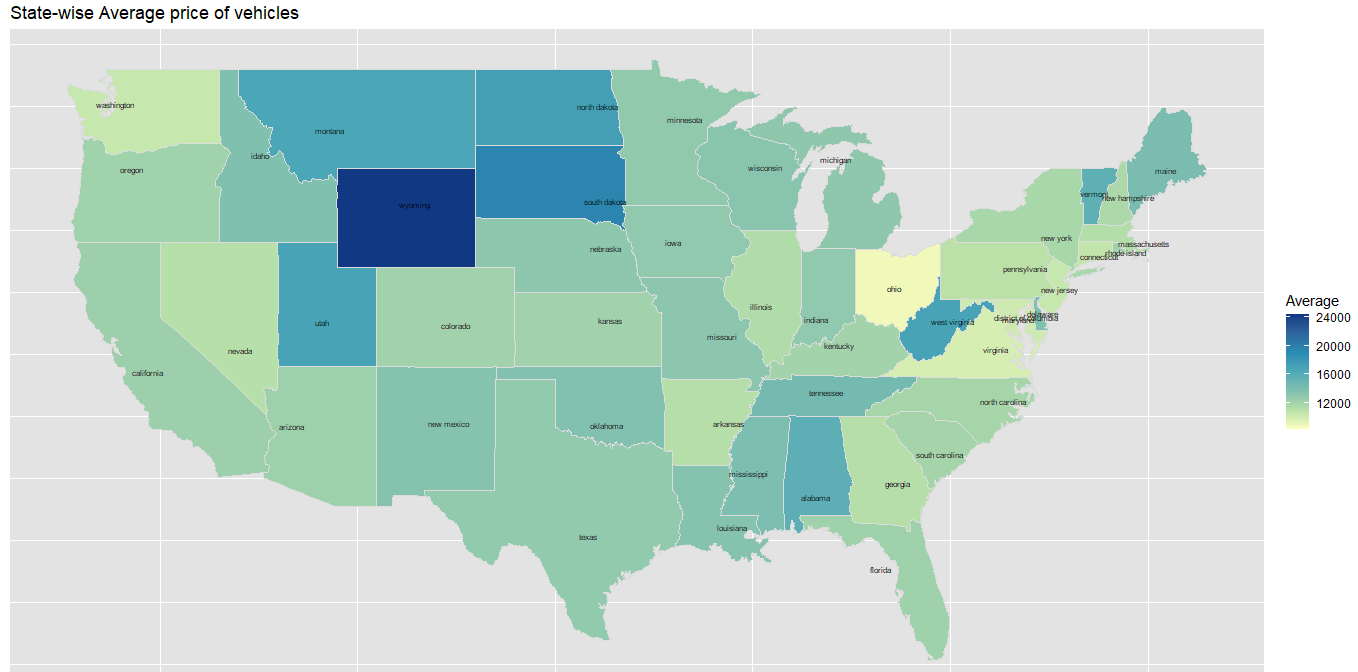
Next are the few plots show us the comparison of the average price with respect to Fuel type and Condition of the Vehicle. This shows us that used cars with fuel type diesel and condition maintained as “new” can be listed at a higher price compared to others. Similar Exploration can be done for say size, status, cylinder number etc. to get insights on those points too.



The below time series shows how the price of cars has increased as the years go by , showing increase in economy and inflation. This could mean that while using historic data of a particular model, to estimate the price, can be done by adding a few more dollars to the past value.

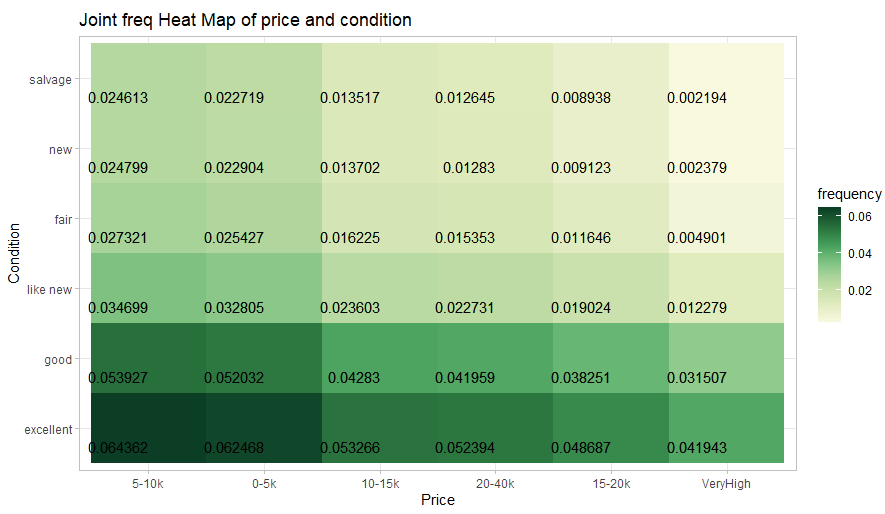


The last map plot below shows us the comparison of average price in the US based on states. We can see that Wyoming has the highest average price (good to list your car there !) and Ohio has the lowest (good to buy a used car from here !)

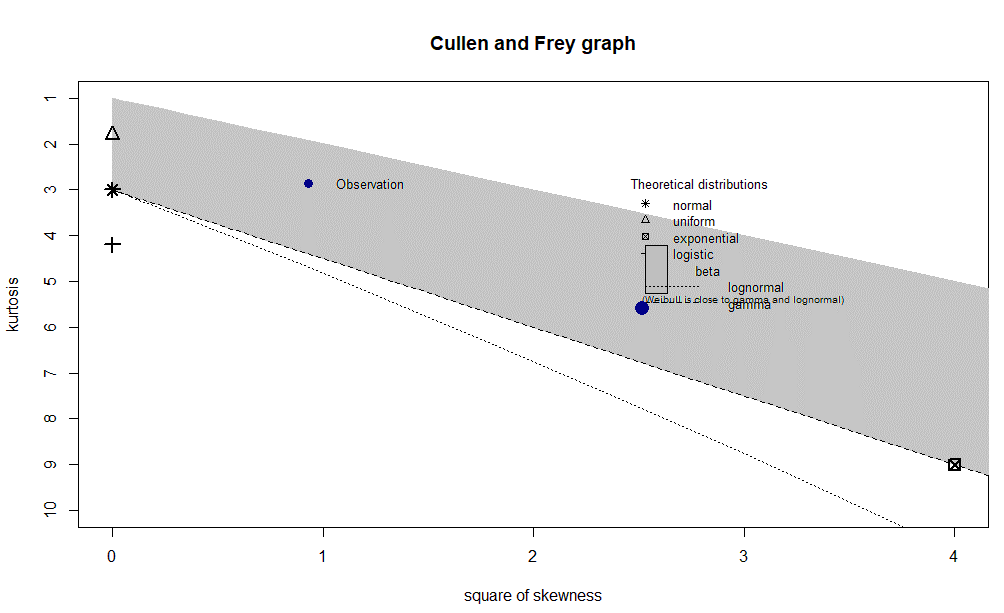


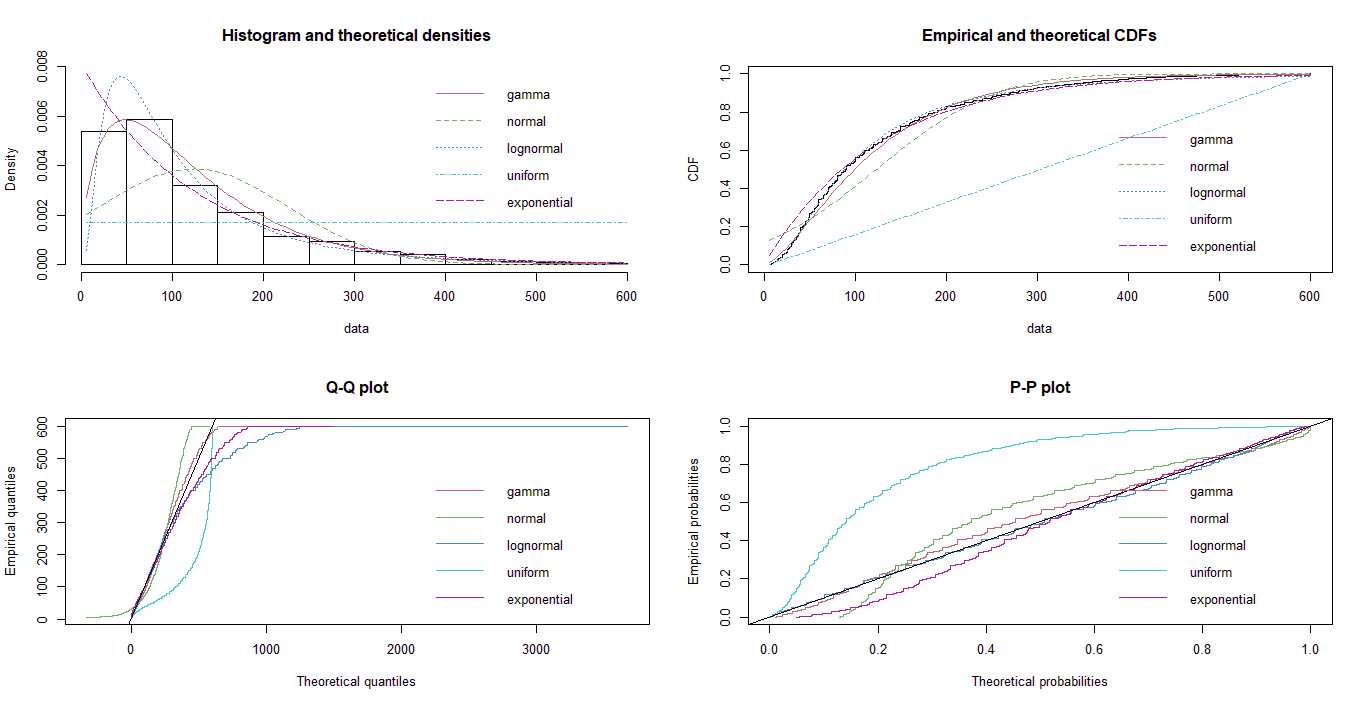
**Statistical Analysis :**

To get a more mathematical and probabilistic idea of the data we found out the statistical information of price like the mean, median, range, standard deviation, skewness, and kurtosis values. Further we utilized the PDF and CDF knowledge to find the joint probability and correlation of price with condition variable. Correlation value is as high as 0.92, which means that the condition on the used car has a huge impact on the price value of the item. Below is a heat map plot to justify the same.



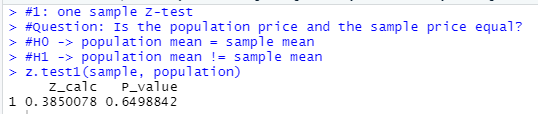
The next important step in our analysis was to know which distribution is the best fit for our data. We have already seen that the price data is negatively skewed. Now to find the goodness of fit, we first plot the Cullen and Frey graph, and then the QQ and PP plot to confirm that gamma is the best fit distribution for our dataset.



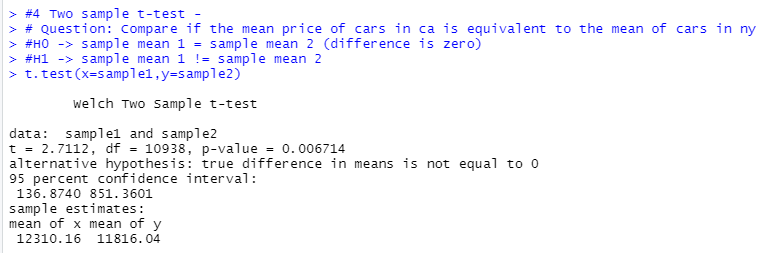


**Hypothesis Tests :**

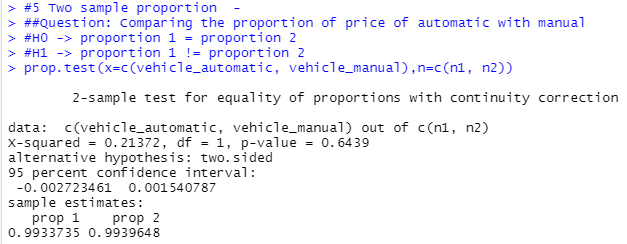
The first hypothesis tests we did was to know whether the population mean is equal to sample mean or not using z test. As it is seen in the below image, we got a p value of greater than 0.05 and hence failed to reject the null hypothesis. This forms the basis that further analysis can be done using sample instead of population, since mean values are equal.



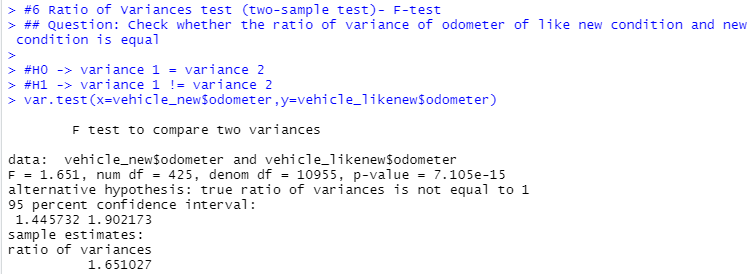
Second hypothesis test , we chose to do a two-sample t test of Compare if the mean price of cars in California is equivalent to the mean of cars in New York. We got a p value less than 0.05 and hence rejected the null hypothesis. Therefore, mean price of cars in California is not equivalent to the mean of cars in New York. This could be because of several factors like income, weather, population etc.



In order to find out whether the two transmission types have difference in proportion , we used the two-sample proportion test and got p value of greater than 0.05 and hence failed to reject the null hypothesis. This means that the proportion of the two transmission types is not equal and one of the types is preferred over the other.



Lastly, we did F-test to test the ratio of variances in odometer values of “like new” and “new” condition and got a p value less than 0.05 and hence rejected the null hypothesis. Failure of this tests shows that the condition values have different odometer variances, probably cause the “like new” condition car has travelled more miles than the “new” condition listed car.



**Advance Analytics :**

**Conclusion :**