**Craigslist Used Cars Listing Analysis**

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

**Introduction :**

A used vehicle is priced based on its model, manufacturer, transmission type, size, condition and many more factors. To best estimate the price of a car we look at listings of identical cars with similar condition. Getting the best price for your listing or choosing the correct car to buy requires extensive market research which can be tedious when looking at the large number of listings posted on the internet.

Through this project we aim to minimize the effort required to make the decision of selling or purchasing a vehicle by analyzing the real-world used cars data. Another focus of this project is trying to estimate the price of used car based on historical data available.

**Objective :**

The goal of this project is to perform exploratory data analysis (using R) on a used cars dataset and identifying patterns through the use of descriptive and inferential statistics and observing the corresponding graphical plots. Also, our aim is to perform hypothesis tests to draw some conclusions about the existing dataset. Finally, we apply a linear regression model to estimate the price of a car on the basis of existing information such as condition, odometer reading, number of cylinders, year of manufacturing etc. The dataset used for the project is Craigslist used cars listings available on Kaggle.

Craigslist is one the biggest used car marketplace in the world and this dataset provides a good insight into the real-world data of car listings.

**Data Description :**

The dataset car listings from year 1900 to 2021, it contains 458213 rows (each representing a single listed used car) and 26 columns (variables describing the listed item). We initially trimmed the dataset by removing the unwanted columns and this reduced our dataset to have 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 odometer reading of the listed car.
10. Title Status - The cars also had 6 types of status; ‘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. Eg. “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 unwanted columns and 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 as a new column

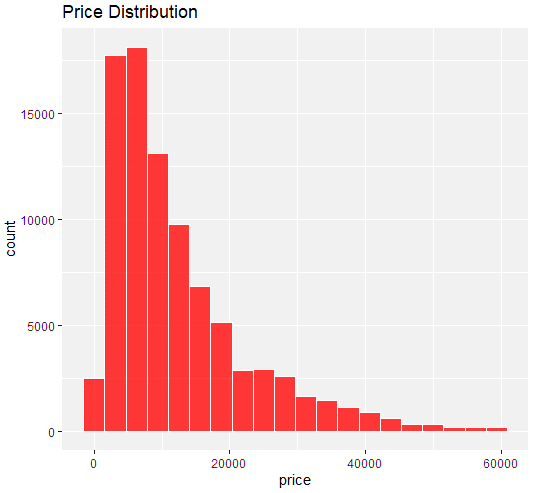
Step 7 – Creating 6 bins for price according to the values.

After preprocessing the dataset, we obtained 162,662 rows and 19 columns to perform exploratory data analysis.

**Data Visualization:**

A good way of identifying trends in the dataset is to visually inspect plots of attributes based on different statistics.

Figure 1 is a frequency histogram plot of the attribute ‘Price’. It can be observed that majority of the cars listed on Craigslist are under $20,000. One inference that can be drawn from this is that cars under $20,000 are more likely to be sold on craigslist.



Figure

Figure 2 shows the bar plot for average price of cars of each fuel type. Figure 3 depicts the average price of a listed car based on condition. It can be observed that the fuel type category ‘Diesel’ has the highest average price, similarly a car listed under ‘new’ condition label is more likely to be listed for a higher price as compared to other conditions.

Figure

Figure

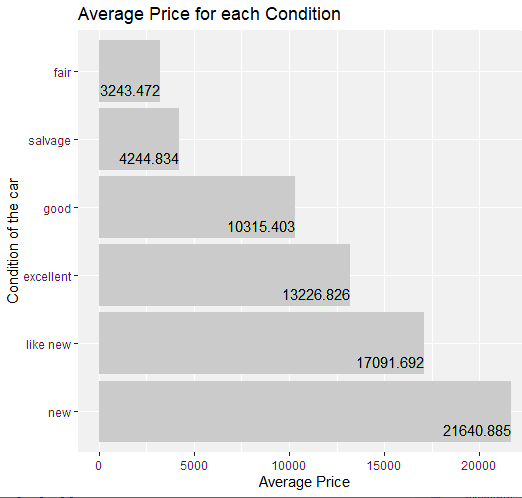
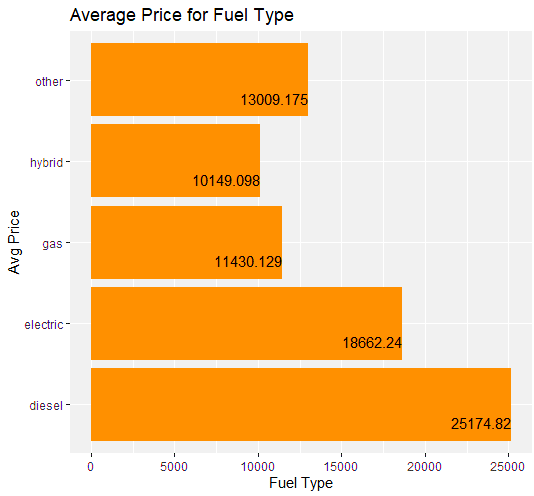
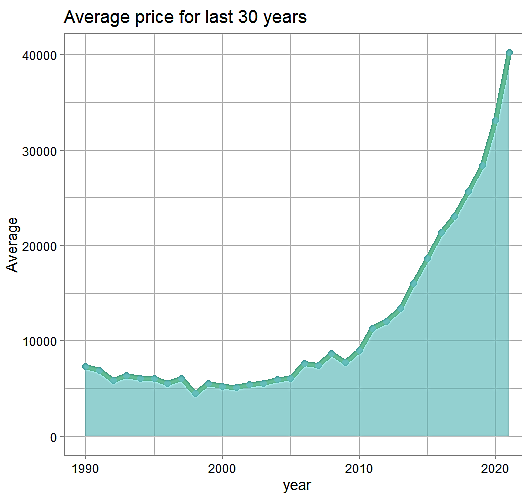
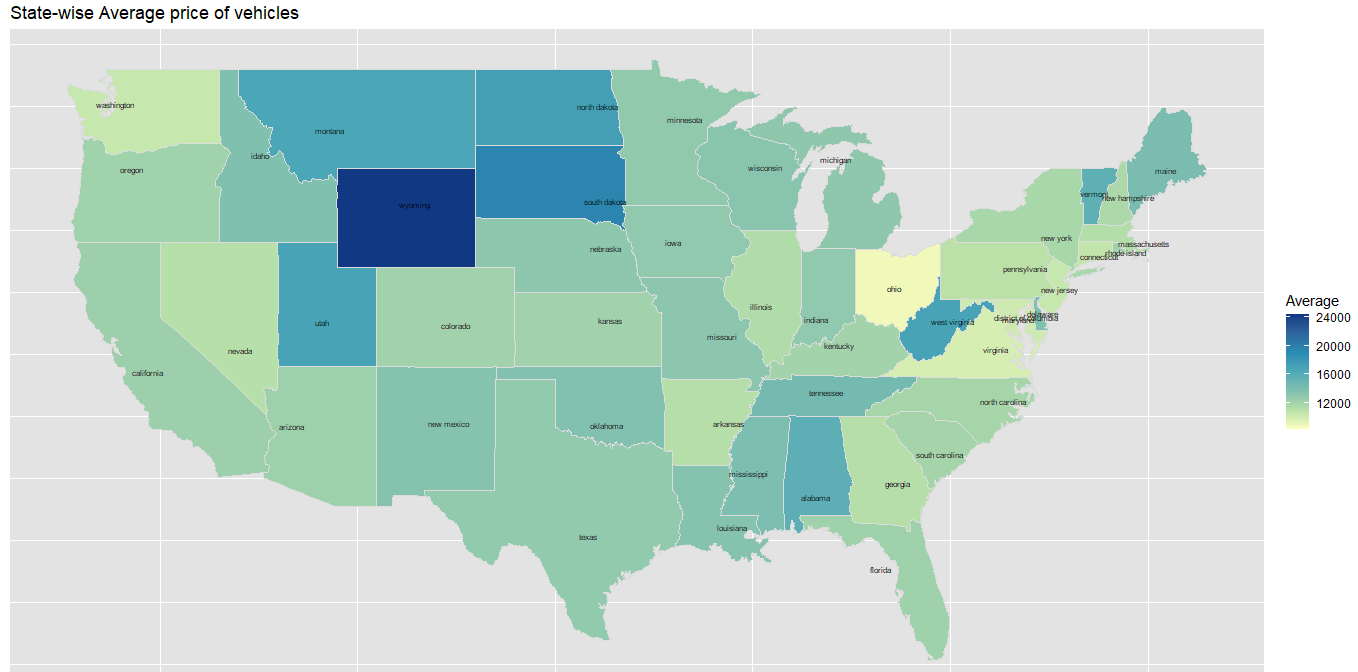


Figure 4 shows the average price of the listed cars vs the year of manufacturing. We can infer that newer cars are more likely to be listed for a higher price than older ones.



Figure

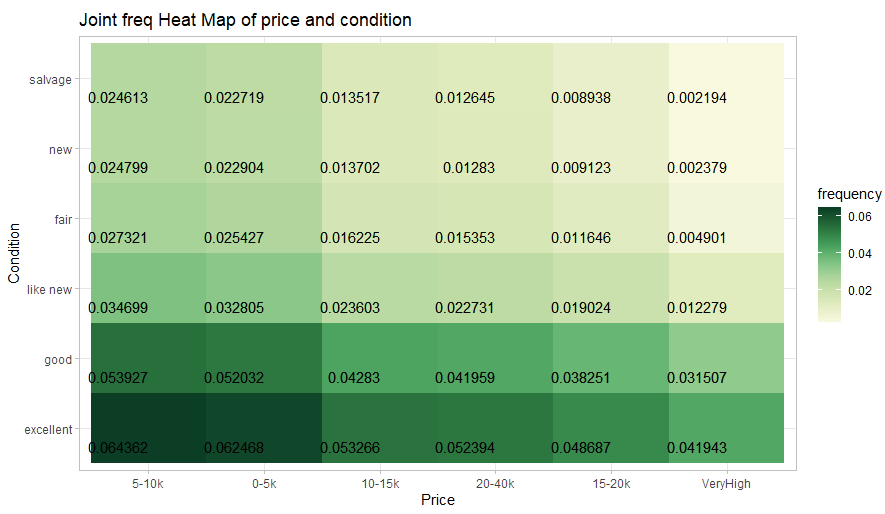
Figure 5 shows us the comparison of average price of listed cars in the US based on states. We can see that Wyoming has the highest average price (good to list your car there) and Ohio has the lowest (good to buy a used car from here).



Figure

**Statistical Analysis:**

Descriptive statistics of the numerical attributes like the mean, median, range, standard deviation, skewness, and kurtosis values were computed to give us a clear picture of the data. PDF and CDF values were calculated for car prices and a joint probability event was created for the price and condition variables. Correlation value is as high as 0.83, which means that the condition of the used car has a huge impact on the price value of the car. Figure 6 is the heat map plot of probability of joint event defined by price and condition.



Figure

Next, we try to determine the distribution that best describes the attribute Price. From data visualization we observed that the price value is a positively skewed distribution. 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.

Chart

Description automatically generated

Figure

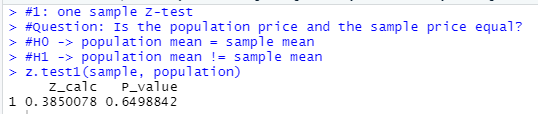
Graphical user interface, chart

Description automatically generated

Figure

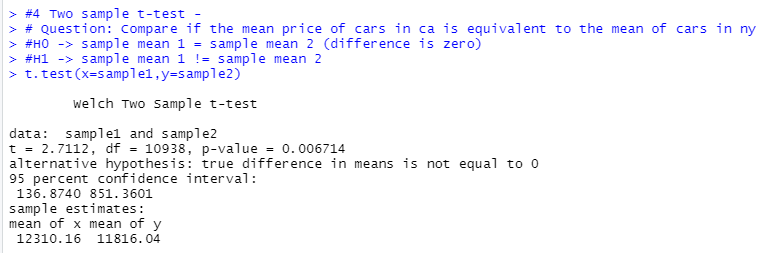
**Hypothesis Tests:**

The first hypothesis test was to determine whether the population mean is equal to sample mean or not using z test. It can be observed in Figure 9, a p value of greater than 0.05 was obtained and hence we fail to reject the null hypothesis. This forms the basis for further analysis to be done using the sample instead of population, since mean values are equal.



Figure

Second hypothesis test was a two-sample t test to determine if the mean price of cars in California is equal to the mean price of cars in New York. A p value less than 0.05 was obtained and hence we can reject the null hypothesis. Therefore, mean price of cars in California is not equal to the mean price of cars in New York. This could be because of several factors like income, weather, population etc.



Figure

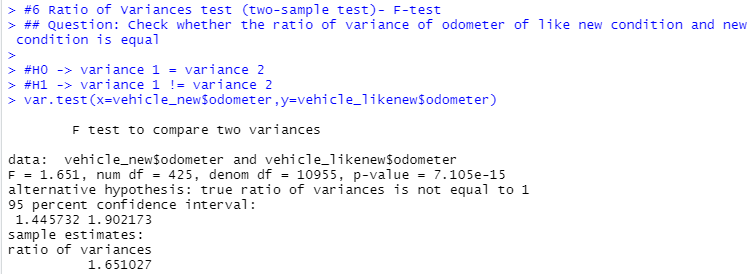
In order to find out whether the two transmission types have different proportions in the population, we employ a two-sample proportion test and obtained a p value of greater than 0.05, hence we fail to reject the null hypothesis. This means that the proportion of the two transmission types could be equal and no type is preferred over the other in the dataset.

Text

Description automatically generated

Figure

F-test was performed to test the ratio of variances in odometer values of “like new” and “new” condition and obtained p value less than 0.05 hence we reject the null hypothesis. From this we can conclude that the condition has different odometer variances, probably because the car listed under “like new” condition has travelled more miles than a car listed under “new” condition.



Figure

**Predictive Analytics (Linear Regression):**

Predicting the price of a car based on available historical data is one of the key objectives of our project. A linear regression model was trained to predict ‘price’ values based on attributes Number of Cylinders, Condition, Status, Transmission type and Fuel type.

We used a correlation matrix to identify the candidates for predictors that can be used to predict the value of attribute ‘Price’. Figure 13 represents the correlation matrix that shows the relationship between variables in the dataset with the attribute ‘Price’.

Chart, bubble chart

Description automatically generated

Figure

First, the categorical data was factored into bins and each bin was assigned a numerical value. Numerical data was already cleaned for use after importing the dataset.

The data was split into training and testing sets. We selected a 70-30 ratio of training to testing data. From the testing data another 20 percent was used for the validation set.

Model was trained on 70 percent of the dataset and it was used to predict the price of the testing and validation data.

The RMSE and R squared metrics were computed for the same, the values were 8532.2 and 0.4813 respectively.

Figure 14 shows the plot of actual values vs the values predicted by our Linear Regression model. The red line depicts the regression line which is used to predict the price variable with the help of other variables.

A picture containing graphical user interface

Description automatically generated

**Conclusion :**

Through this project we were able to apply the concepts that we learnt during the course. Using real-world data like the craigslist used car dataset gave us hands-on experience of cleaning and pre-processing data to make it usable for analysis.

After performing exploratory data analysis and visualizations we were able to observe some trends and patterns in the data. Computing the descriptive statistics confirmed that observed trends were valid and we obtained key metrics like mean, median, skewness and kurtosis values for the selected attributes.

After fitting the data to various statistical models, we were able to conclude that gamma distribution is the best fit to describe the attribute ‘Price’. The parameter values for the gamma distribution that best describes the values of Price are = 1.62, =0.01.

Test of hypothesis helped us confirm or reject important statements such as if the variances of two different listed conditions were different from each other. This led us to conclude that there was key difference in the miles driven of car ‘like new’ condition and ‘new’ condition.

Finally, the linear regression model was trained to predict the price of car using the other attributes as predictors. The model was able to predict price of car on the testing dataset with and R squared error value of 0.4813. This indicates a decent fit between predicted and actual values.