Part 4:Project Recap: Data Analysis on Used Cars

1.Introduction

This study focuses on the analysis of the used cars data to discover the patterns that helps to conclude the results mentioned in hypothesis. It uses data acquired from the Kaggle data repository [1] which is then preprocessed and visualized using RStudio [2] and Tableau [3]. The findings from the study concludes that the most popular brands in the market, such as Maruti, Hyundai, Honda, and Toyota, are in the lower price range (less than 12 lakhs), whilst the least popular brands, such as Mini, Volvo, Porsche, Jeep, and Datsun, are in the higher price range. It also explains why diesel vehicles have a larger proportion and mileage than petrol vehicles, and that mileage is inversely related to the 'number of seats' and 'power' produced. Finally, Transmission, Years Old, Power, Kilometers Driven, Engine, Fuel Type, and Owner Type are the most important factors in determining the vehicle's pricing.

2.Initial Requirements & Questionnaire

To work on the data and perform the analysis ,a dataset with key vehicle attributes such as brand name, no of seats, engine capacity, power , fuel type, mileage , kilometers driven , ownership type, transmission types are required. Steps should be taken to ensure that there are no inconsistencies or outliers in the data. If there are any it needs to be handled in the data cleaning pipeline .Finally, if the study requires in-depth analysis new attributes must be derived from the existing fields.

Some of the key questions that needs to be answered as a part of this study are as follows:

1. Which brands and type of vehicles are sold out most in the market?

2. How the mileage varies according to the fuel type and vehicle characteristics?

3. What factors determine the price of the vehicle the most?

3.Data Cleaning and Information Modeling

The dataset used for this study consists of 7251 observations and 13 features with both numerical and categorical data types such as Name, Location, Year, Kilometers_Driven, Fuel_Type, Transmission, Owner_Type, Mileage, Engine, Power, Seats, New_Price, Price. Columns which doesn't add any meaning for the modelling & analysis has been removed. Next, two new dimensions "Brand" and "years_old" has been derived from the existing fields. Summary statistics on the data revealed that fields such as seats, price, new price, engine, and power have missing numbers with "new price" accounting for the around 90% of them. This field has been removed and all the remaining fields has been filled using one of the measures of central tendencies (mean, median & mode) appropriate for that field. Features such as Mileage, Engine and Power with multiple unit metrics has been observed which in turn are transformed into single metric. Next, the numerical features has been checked for the presence of outliers using Z-score metric and has been dropped which reduced the total observations by 20%.Finally, the duplicates (4) are eliminated, and the data has been converted into appropriate data type.

4.Analysis & Results

After performing data cleaning, the obtained data is ingested into Rstudio and Tableau for modelling and visualizations. To find out which vehicles are sold out most in the market the following graphs has been produced. The result from the figures explains that brands such as Maruti, Hyundai, Honda, and Toyota are the most popular, while Mini, Volvo, Porsche, Jeep, and Datsun are the least popular. In addition to that there appears to be an inverse link between total vehicle count and price, with the most selling vehicles falling into the lower price category (below 15lakhs) and the least selling vehicles falling into the higher price category.

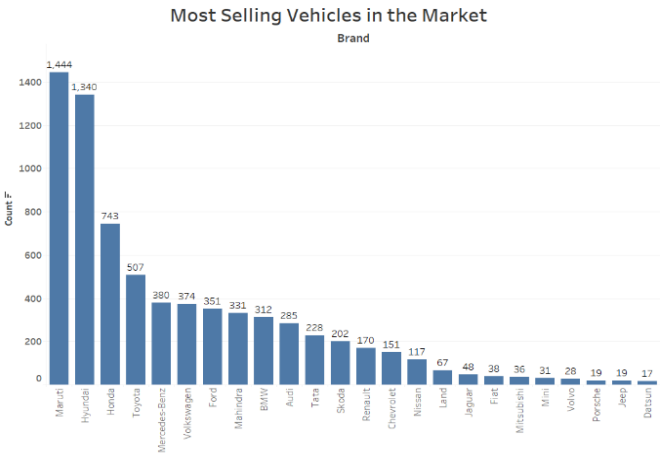

Figure 1: Frequency Distribution of Brands ***(Produced using Tableau)***

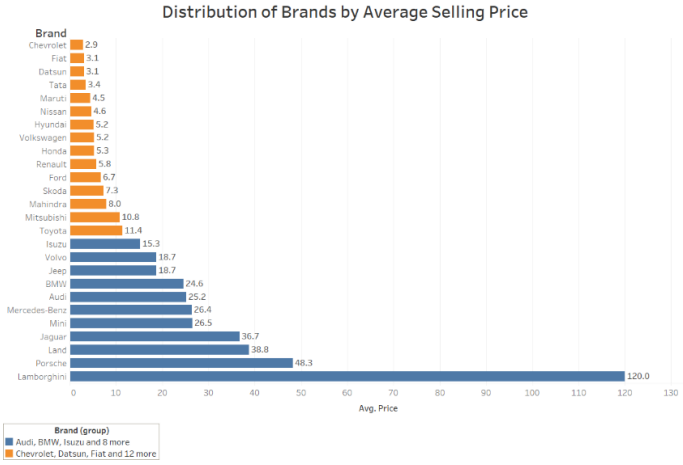

Figure 2: Vehicle Distribution by Average Price ***(Produced using Tableau)***

Next, to find out how the mileage is varying according to the fuel type and vehicle characteristics a boxplot between fuel type and Mileage has been plotted as shown in the figure 3. According to the results of the above box plot, LPG and CNG cars have the best mileage, with an average of 35 kmpl and 27 kmpl, respectively. Even though these vehicles get excellent mileage, they are not widely available. Petrol and diesel, on the other hand, provide an average mileage of around 18 kilometers per liter.

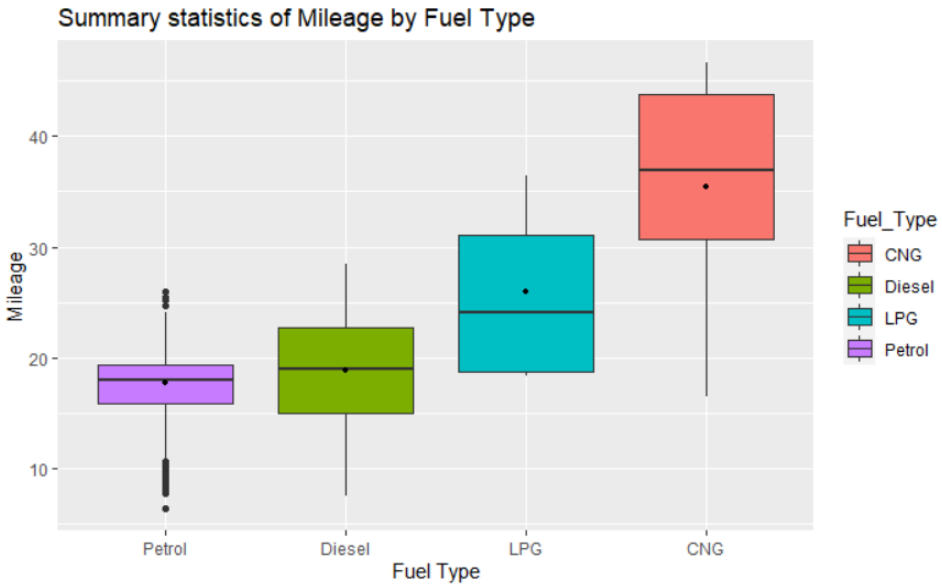

Figure 3: Box Plot of Mileage by Fuel Type ***(Produced using RStudio)***

The scatter plot explains that vehicles with a power output of more than 300 bhp are getting roughly 15 kilometers per gallon. On the other hand, vehicles with a power output of 100 to 300 horsepower are available in both petrol and diesel, with a mileage of up to 25 kilometers per gallon. Finally, vehicles with less than 100 bhp are more likely to be Petrol vehicles or CNG vehicles, with mileage of around 30kmpl and above. Based on the evidence, it can be concluded that power is inversely proportional to the amount of mileage produced.

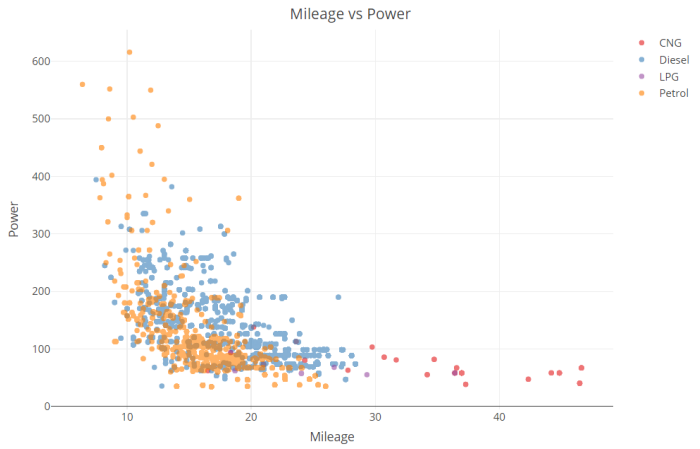

Figure 4: Relationship between Mileage and Power by Fuel Type *(Produced using RStudio)*

The average mileage of automobiles with 5 seats to 9 seats has decreased from 19kmpl to 15kmpl, as seen in the box plot. Because 2-seater vehicles are constructed with large engine cubic capacity and power, they perform similarly to 10-seater vehicles. The four-seater car has a median mileage of 12 miles per gallon and an average mileage of 14 miles per gallon.

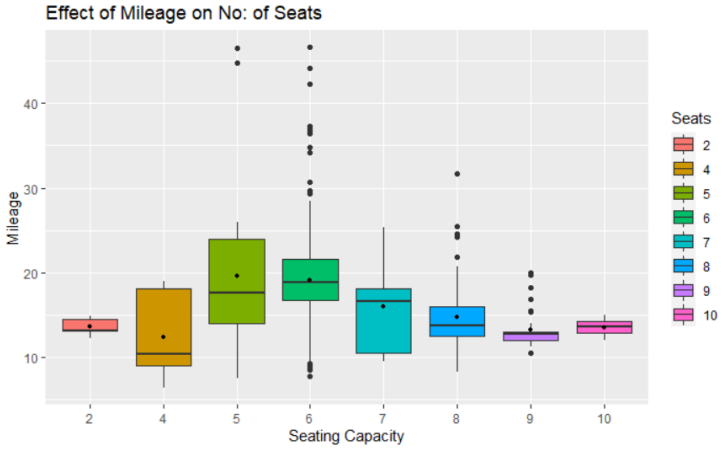

Figure 5: Box Plot of Mileage by Total number of Seats *(Produced using RStudio)*

To identify the relationships and validate the skewness in the data, a pairwise plot has been drawn as show in figure 6. The results from the figure explains that most of the features are skewed towards left side. Therefore, the data has been normalized, encoded, and scaled to improve the efficiency before fitting the models.

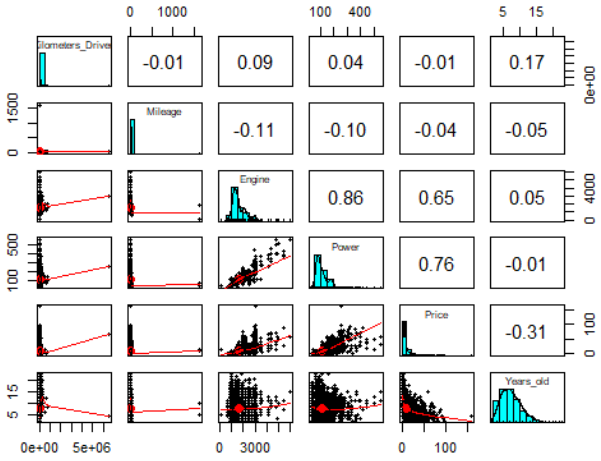

Figure 6:Pairwise plot of Vehicle Features *(Produced using R Studio)*

To identify the factors affecting the selling price as well as predict the price of unforeseen data, K nearest neighbor and linear regression model were applied on the training data using tidymodels package [4] in R. Among the two,KNN model performed better than linear regression with a root mean squared error of 4.51 and a r2 error of 0.82 as supposed to RMSE of 5.10 and an r2 of 0.77 on test data. The model also determined the most important variables based on the significance of the variables when calculating the target variable.

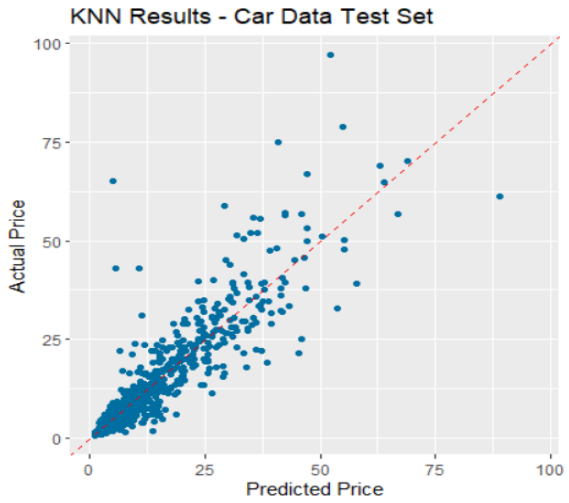

Figure 7: Actuals vs Predicted Plot
(Produced using R Studio)

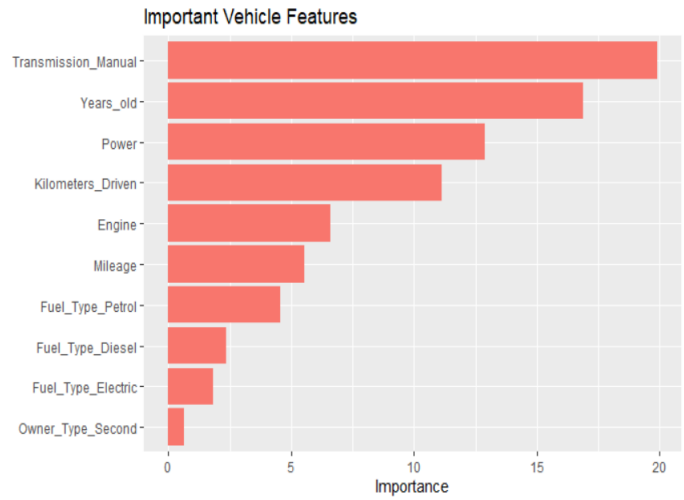

Figure 8:Variable Importance Plot
(Produced using R Studio)

5.Conclusion

1. Which brands and type of vehicles are sold out most in the market?

My hypothesis is that brands like Maruti, Honda, Tata, and Hyundai holds majority of the market because they offer automobiles with excellent mileage. Furthermore, as compared to its competitors, they are substantially less expensive. Based on the results from figure 1 and 2 , I can conclude that the hypothesis is true.

2. How the mileage varies according to the fuel type and vehicle characteristics?

When it comes to mileage, I believe that cars that run on diesel gives more mileage. In addition to that, cars with a large seating capacity and more engine power will provide fewer mileage because their size affects the speed. Based on the results it can be concluded that hypothesis is partially true because LPG and CNG are providing high mileage. On the other hand, two seaters vehicles are giving less mileage when compared to higher seating capacity vehicles.

3. What factors determine the price of the vehicle the most?

I believe that criteria such as ownership type, number of years, type of gasoline used, and mileage are the most essential aspects in calculating the final price because these are the factors that have the greatest impact on the final selling price.Based on the variable importance plot It can be concluded that the hypothesis is true, but it also explains that Transmission type, Engine and Power are important to be considered.

6.Key Challenges

One of the key challenges faced while working on the study is the cleaning the data. The data has so many missing values and some of the columns are not in proper formatting. As a result majority of the time has been spent on preprocessing to bring the data into right shape. Apart from that, after knowing that data is skewed ,applying the right techniques to normalize it has been a bit of a challenge.

7.Lessons Learnt

From this course I've learnt what to look for in the data when it is provided to the analyst before starting the analysis and right questions to ask for such that it brings value to the bussiness. It also helped me to understand how to start with basic analysis and go deeper based on the results obtained from it. Apart from that, I've learned how data with two different scales can lead to inaccurate results and role of normalization in overcoming this issue after applying in my study. Finally, I learned how to do visualizations using tableau that are dynamic and visually as well as increased my proficiency in data cleaning using RStudio.

8.References

[1] Sanam Peeyush. (2021, June 10). *Used Car Prices in India*. Kaggle. Retrieved May 1, 2022, from https://www.kaggle.com/sanamps/used-car-prices-in-india

[2] *RStudio | Open source & professional software for data science teams*. (n.d.). RStudio. Retrieved May 1, 2022, from https://www.rstudio.com/

[3] *Tableau Desktop*. (n.d.). Tableau. Retrieved May 1, 2022, from https://www.tableau.com/products/desktop

[4] *Tidymodels - Tidymodels packages*. (n.d.). Https://Www.Tidymodels.Org. Retrieved May 1, 2022, from https://www.tidymodels.org/packages/