STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

Predicting Used Car Prices

A team project report submitted for

MA 541 - Statistical Methods

By
SATHVIK VADLAPATLA
ANWAR BASHA DUDEKULA
AFRAS
SUDHAM SAI RITHEESH

Under the guidance of

Prof. HONG DO

Stevens Institute of Technology

I pledge my honor that I have abided by the Stevens Honor System.

Table of Contents

1.0 Introduction

1.1 Background on the Used Car Market

The used car market has been experiencing immense amount of hike over the period of last 10 years, making it a pivotal sector in the automotive industry. This upswing is driven by many reasons which includes economic fluctuations which brings out consumers to look out for low cost deputy to new vehicles and improvements in the durability of vehicle's lifspan. As a result, used cars have become an tempting atlernate for a broad demographic, ranging from first-time buyers to seasoned enthusiasts looking for vintage or specific models.

The used car market dynamics have been influenced by many different factors like the economic condition, technological advancement, and chioce of the customer. Such factors not only affect the price of the vehicle but also effects the demand and supply in different parts of the region. Furthermore, the trend with respect to digital media platforms has completely evolved the manner in which dealers and client interact in terms of buying and selling preowned cars, initiating a level of transparency and accessibility which was not experienced ever before.

1.2 Importance of Data Analysis in the Automobile Industry

In today's data-centric generation , the automobile industry makes use of data analysis to not only increase operational efficiencies but it also understands the way consumer reacts with reapect to the trends in the market. For preowned car dealerships and marketplaceson social media, data analysis gives critical observations that helps in taking strategy based decisions like managing inventory, strategical pricing, and client targeting.

Advanced analytics helps in recognising patterns and predicting the price of a car based on the future trends , which are important in order optimize sales of the preowned cars and buyer satisfaction. By analyzing data from the past, dealers will be able to predict the car models which are going to be in high demand and make changes their inventories periodically. Moreover, data analytics gives rise to personalized marketing, in which the client is targeted based on the behaivour of buying and their peferences thereby, increasing the probability of sale.

1.3 Overview of the Dataset Used

The dataset that we have used in this project gives us detailed insights on different preowned cars that have been sold, that captures a detailed set of features which plays an important role in identifying the trends and patterns in the used car market. The dataset includes columns that gives us a brief description of specifications in each car, sales details, and information of the customer.

Following is a brief description on the columns that exisrts in the dataset:

- car_name: The make and model of the car.

- registration_year: The year in which the car was initially registered.
- insurance_validity: The status of the car's insurance coverage.
- fuel_type: The type of fuel the car uses (e.g., Petrol, Diesel).
- seats: The number of seats in the car.
- kms_driven: The total kilometers the car has been driven.
- ownership: Indicates how many previous owners the car has had.
- transmission: Specifies whether the car has an automatic or manual transmission.
- manufacturing_year: The year the car was manufactured.
- Mileage (kmpl): The car's fuel efficiency in kilometers per liter.
- engine(cc): The engine capacity in cubic centimeters.
- max_power (bhp): The maximum power output of the car in brake horsepower.
- torque (Nm): The peak torque produced by the engine in Newton meters.
- price (in lakhs): The selling price of the car in lakhs.

This dataset is a very valuable accomodity in conducting statistical analysis and building predictive models to understand various important factors which influences price of the car and preferences of a client in the used car market. By looking into the variables such as mileage, engine size, and price, our project aims to identify trends and give valuable insights which could be the decisive factor for the dealers to optimize their inventory and strategize their pricing.

2.0 Objectives

2.1 Main Goals of the Project

Our main goal of doing this project is to utilize the resources of data analysis to pull out valuable insights from the used car dataset, which can help in understanding the dynamics of the used car market. By applying various statistical methods and data visualization techniques, the project aims to:

- Identify Key Trends: Understanding important trends within the used car sales data, like the impact of features like mileage, engine size, and age with respect to selling price of the car.

- Optimize Pricing Strategy: Imploy a model that helps the dealers to predict a good market value of the pre owned cars on the basis of their attributes, assisting dealerships in setting competitive and profitable pricing.

- Enhance Inventory Management: Helping the dealership to optimize their inventory by identifying the market trends, as in the type of cars which are demanded highly by the buyers based on the historical data.

- Improve Customer Targeting: Analyzing customer behaivour and buying patterns to alter marketing strategies that helps in effectively targeting potential buyers, thus enhancing sales conversions.

- Predict Future Sales: Use time-series analysis to forecast future sales trends, which helps the dealres to stategize their sale effectively.

2.2 Specific Questions the Project Aims to Answer

Through indepth analysis of the dataset, our project amis to answer various questions that are necessary in making strategic decisions in the used car market:

1. What are the most important factors that impacts the price of a used car?

- Determining features (e.g., mileage, engine capacity, age) most importantly effects the selling price of a car.

2. What is the corelation of different attributes with customer buying behaivour?

- Exploring relationships among specifications of the car (like fuel type and transmission) and their popularity among different customer segments.

3. Does the historical sales data indicate any particular preferences for car types based on geographical regions?

- Analyze if certain models or types of cars sell better in specific regions or cities and why that might be the case.

4. What is the impact of the car's age and kilometers driven on its resale value?

- Assess how depreciation and usage affect the car's market value to aid dealers in assessing trade-ins and resales.

5. Can we predict future trends in car sales based on past data?

- Develop predictive models to forecast sales, which can guide inventory and marketing strategies.

6. What pricing strategy should be adopted to maximize profitability while remaining competitive?

- Use the data to model optimal pricing strategies that attract customers and maximize margins based on market conditions and car features.

Although analysis of issues, the project will provide comprehensive insights that will further fuel improvements in customer satisfaction, sales strategies, and performance in the used vehicle market. That approach guarantees sustainability and growth in a cutthroat market by advancing long-term planning and forecasting, in addition to assisting with current business decisions.

3.0 Methods

3.1 Description of the Data Preprocessing Steps

To ensure the accuracy and effectiveness of the analysis, the data underwent several preprocessing steps. These steps were critical to preparing the dataset for a thorough exploration and ensuring that the subsequent statistical analyses and model predictions were based on clean and relevant data.

- Cleaning data: The stage included in the stage was very vital was dealing with all the missing values found in our data, removing duplicate data, and correcting any inconsistencies present in it. For example, we identified significant missing values in columns such as 'mileage(kmpl)', 'engine(cc)', and so on which had to be dealt with either by imputation or deletion depending on their nature as well as extent.

- Data Transformation: We altered certain attributes such as 'price(in lakhs)' to depict car prices variations accurately. This transformation aims to improve analytical model performance by normalizing skewed data distribution through logarithmic transformations.

- Feature Encoding: Categorical variables such as ‘fuel_type’ and ‘transmission’ were encoded to numerical formats using methods like one-hot encoding or label encoding. This step was essential for preparing the dataset for machine learning algorithms, which require numerical input.

- Feature Engineering: In order to improve model performance and offer deeper insights, new features were extracted from the available data. For instance, to ascertain the influence of the vehicles' age on demand and price, the age of the automobiles was computed using the 'manufacturing_year'.

- Data Scaling: Standardisation or normalisation procedures were employed to scale the features so that variables with higher values would not adversely affect the models. For distance-based algorithms that are sensitive to the range of data points, this is especially crucial.

3.2 Overview of the Analytical Techniques Used

The project employed a variety of analytical techniques ranging from descriptive statistics to complex predictive modeling, tailored to extract the maximum insights from the dataset:

- Descriptive Statistics: An initial inquiry requires computing elementary statistical quantities, such as

mean, median, mode, variance, dispersion, and standard deviation among others, on various attributes with a view to grasping the data's central tendencies and spread.

- Data Visualization: This involved a wide range of graphical tools employments. For instance, histograms showed how different numbers are distributed; bar charts compared categorical data while scatter plots helped in identifying relationships between pairs of values. Also, box plots were used to find outliers among others. Thus, they made it possible to detect any patterns, trends or anomalies in the information displayed on them that might not have been seen otherwise. Essentially, these visual methods unearthed what could not be seen through mere inspection of the numbers.

- Inferential Statistics: This stage utilized techniques like t-tests as well as ANOVA (Analysis of Variance). Their aim was to establish if there were any significant differences between group means thereby enabling one to determine how much influence each variable had on car prices and customer tastes.

- Correlation Analysis: The relationship between various factors was examined utilizing correlation coefficients to identify those that are most strongly associated with price and other dependent variables.

- Predictive Modeling: Predictive modeling involved creating several machine learning models such as linear regression, decision trees and random forests among others, for purposes such as predicting car prices or sales volume. These models were tested on the cleaned dataset which had also been used in their training as well as validation processes before being evaluated based on their performances according to RMSE (root mean square error) values or R-squared ones too alternatively.

- Time Series Analysis: For predicting future sales trends, time series analysis was conducted, taking into account seasonal variations and other temporal factors that could impact car sales.

By leveraging these methods, the project aims to provide comprehensive insights into the used car market, aiding stakeholders in making informed decisions based on data-driven evidence.

4.0 Prediction Model Building

4.1 Summary of the Dataset

The data set used in this study is a collection of records from the sales of second-hand cars, with each transaction described by some related features. It covers information about 1,553 pre-owned vehicles, featuring different makes, models and conditions. Every entry represents a unique sale event accompanied by categorical and numerical points necessary for holistic understanding.

In order to ensure correctness, the dataset was prepared carefully through various preprocessing stages including handling missing values; removing duplicates; converting categorical data into appropriate forms for analysis etcetera. The wealthiness of this set allows one to look at many things about what affects the market for used cars.

4.2 Description of Variables Used in Analysis

The dataset contains several variables, each providing insights into different aspects of the used car

transactions. Here is a detailed description of these variables:

- car_name: The make and model of the car, which is critical for identifying the car's brand and model-specific market trends.
- registration_year: Year in which the car was first registered; this helps in calculating the car's age, which is a vital factor in pricing used cars.
- insurance_validity: Indicates whether the car's insurance is still valid, affecting the car's readiness for transfer to a new owner.
- fuel_type: Type of fuel the car uses (e.g., Petrol, Diesel), influencing the car's operating cost and environmental impact.
- seats: The number of seats in the car, affecting the car's utility and target market segment (e.g., families might prefer cars with more seats).
- kms_driven: Total kilometers the car has been driven, which is a direct indicator of wear and tear and affects the car's value.
- ownership: Number of previous owners, with fewer previous owners often seen as preferable in the used car market.
- transmission: Specifies whether the car has manual or automatic transmission, impacting the driving experience and maintenance costs.
- manufacturing_year: The year the car was manufactured, used to calculate the car's age.
- mileage(kmpl): Fuel efficiency in kilometers per liter, important for cost-conscious consumers.
- engine(cc): The engine capacity in cubic centimeters, indicating the power and potential fuel consumption of the car.
- max_power(bhp): The maximum power output in brake horsepower, providing a measure of the car's performance capabilities.
- torque(Nm): The peak torque output in Newton meters, which affects the car's acceleration and hauling ability.
- price(in lakhs): The selling price of the car, expressed in lakhs, which is the primary dependent variable for analyses related to pricing strategies.

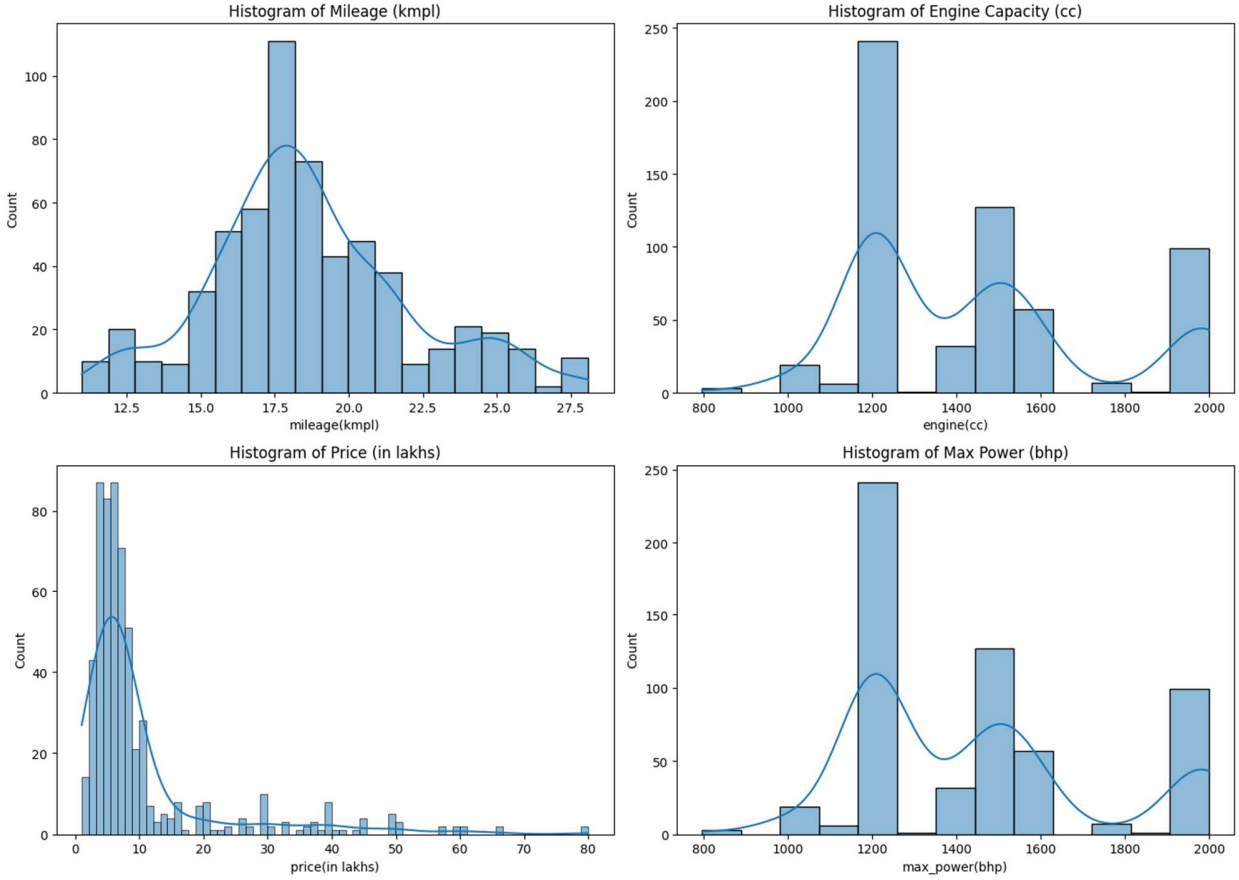
Collectively, these variables present an all-inclusive system to evaluate the market of used cars. They allow us to look into different things that may have an effect on car value or consumer choice such as how age relates with mileage and price also, what fuels are preferred over others in terms of transmission type etc.

The aim of this project is to come up with insights which can be acted upon so as to enable dealerships make their sales strategies better and ensure that clients are satisfied. Another thing that will happen during this investigation is getting ready for prediction modelling whereby prices as well as sales patterns shall be forecasted based on these very same variables.

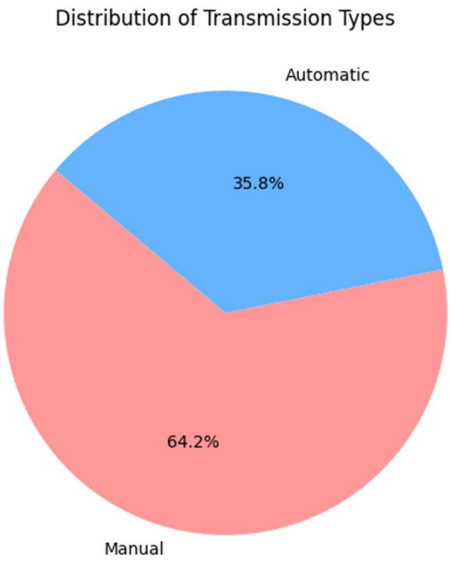
5.0 Data Visualization

Data visualization plays an important role in our project by providing visual insights of the dataset, disclosing underlying trends and insights which could not be visible using the raw data. We have used various different types of graphical representations to determine the price distribution, analysing the corelation among various different variables and in understanding the demographic characteristics of the vehicles sold.

Histograms:

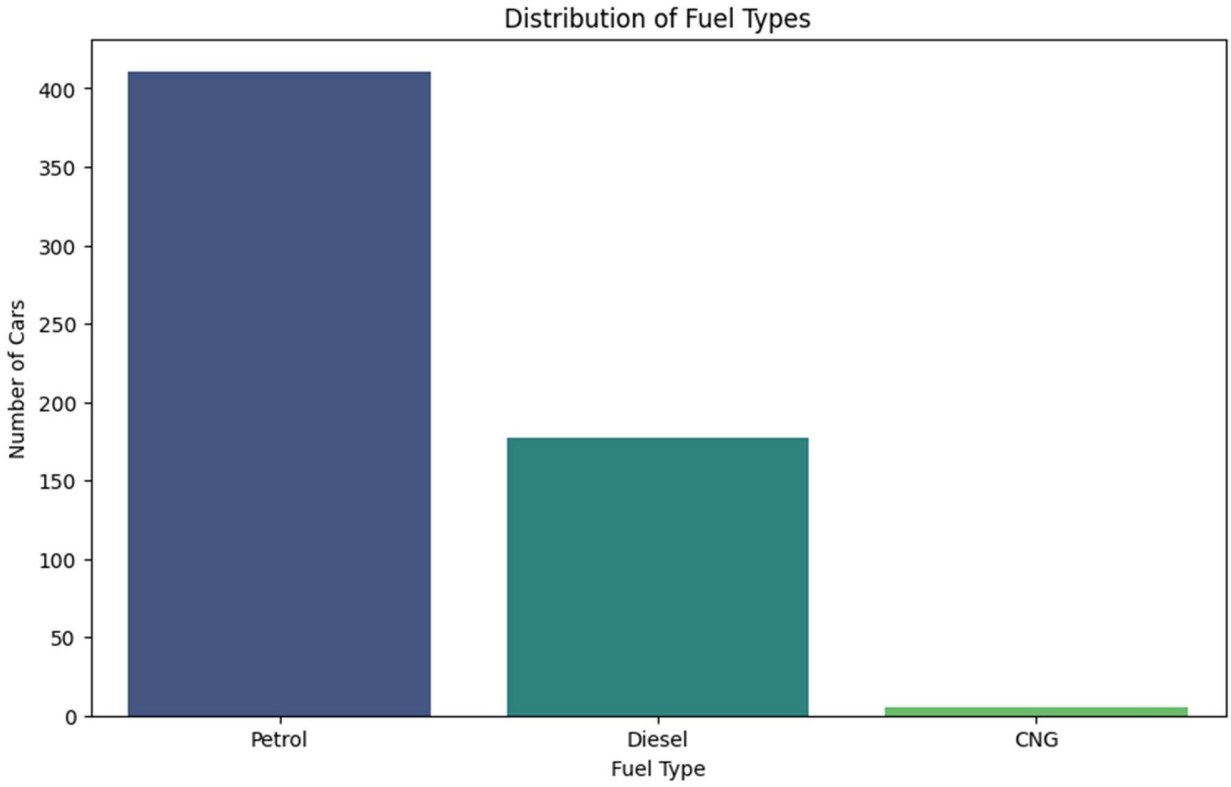


Pie Chart for Transmission Types:



This pie chat shows the distribution among the types of transmission in the cars that have been used previously. The percentages for automatic vs. manual transmissions gives us a clear pictorial representation of the type preferred. Manual vehicles have an upper hand among the customers with 64.2% cars sold, whereas automatic transmission is less preferred with 35.8%

Bar Graph for Fuel Types:



This bar graph shows the distribution of fuel types among the used cars in dataset. This visual helps us in identifying the fuel type most common in the used car market. From this graph, we can come to an understanding that the petrol cars are most common in the dataset with over 400 cars, followed by diesel cars with just less than 200 and CNG being the most unpopular with 10 cars.

Summary Statistics

We have worked out the summary statistics for relevant numerical columns in order to understand the data set properly and identify any trends or unusual values.

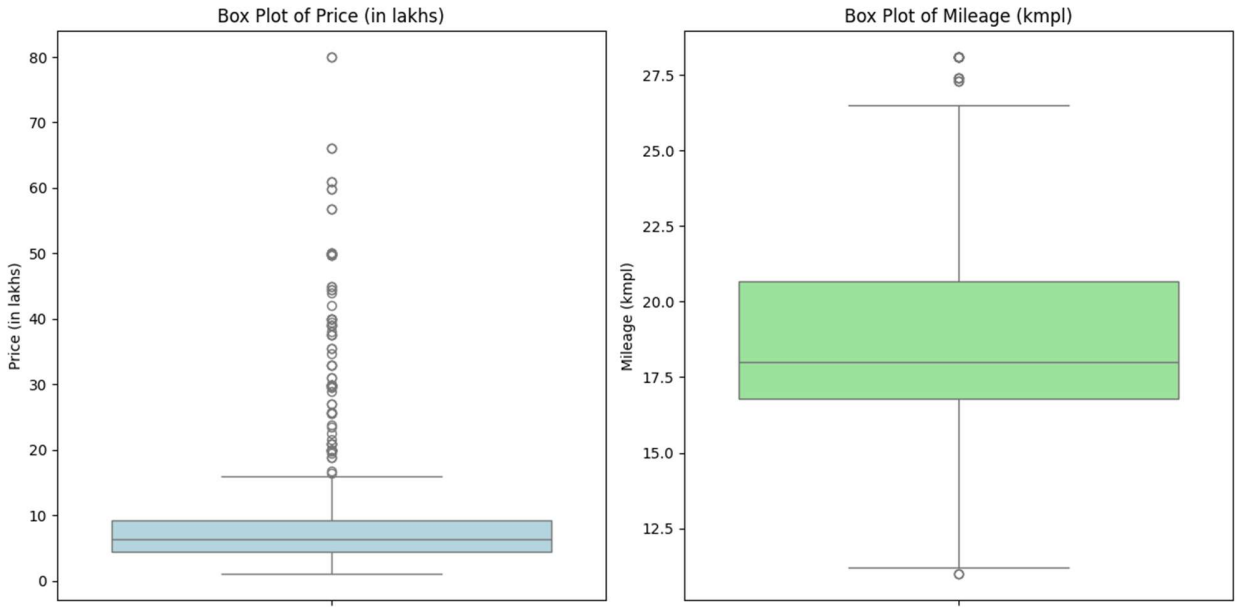
Measures of Center (Mean, Median, Mode)
Measures of Dispersion (Range, Variance, Standard Deviation)
Measures of Location (Percentiles, 5-number summary)

We have calculated statistics for price(in lakhs) and mileage(kmpl) as they are decisive variables that hepls us in understanding the value and efficiency of used cars.

Summary Statistics for Price(in Lakhs) and Mileage(kmpl) Variables:

Price in Lakhs:
Measures of Center:
Mean: ₹10.35 lakhs
Median: ₹6.34 lakhs
Mode: ₹6.25 lakhs (most common price)
Measures of Dispersion:
Range: ₹1 lakh to ₹80 lakhs
Variance: ₹142.85
Standard Deviation: ₹11.95
Measures of Location:
25th, 50th, 75th Percentiles: ₹4.5 lakhs, ₹6.34 lakhs, ₹9.22 lakhs
5-number Summary: [₹1, ₹4.5, ₹6.34, ₹9.22, ₹80]
Mileage (kmpl):
Measures of Center:
Mean: 18.69 kmpl
Median: 18.00 kmpl
Mode: 21.4 kmpl (most common mileage)
Measures of Dispersion:
Range: 10.98 kmpl to 28.09 kmpl
Variance: 11.61
Standard Deviation: 3.41
Measures of Location:
25th, 50th, 75th Percentiles: 16.8 kmpl, 18.0 kmpl, 20.68 kmpl
5-number Summary: [10.98, 16.8, 18.0, 20.68, 28.09]

Visualizing 5-number Summary with Boxplots:



The box plots provide a visual summary of the distribution of prices and mileage.

Box Plot Analysis:

Price (in lakhs): The box plot indicates that the range lies predominantly between ₹1 lakh and ₹10 lakhs, with the median being at around ₹6.34 lakhs. Several outliers can also be obvserved above ₹10 lakhs, with highest being 80 lakhs indicating that there are expensive cars in the dataset as well.

Mileage (kmpl): The distribution of mileage lies mostly in between 17 kmpl and 21 kmpl indicating that it is fairly compact with the median at around 18 kmpl. Few outliers can also we observed in the graph with the lowest mileage of 10.5 kmpl and highest of over 27.5 kmpl.

These graphical visualizations and statistic summaries provides us with a strong understanding of the dataset's key characteristics.

Statistical Inference

- In order to explore the distribution of key variables, We have conducted univariate analysis
- For exploring the relationships between different variables like the impact of engine size on mileage we have performed Performed multivariate analysis.

For performing Univariate Analysis, in order to go in depth regarding the distribution and charcteristics of individual variables, we have focused on two main variables: Price in lakhs and Mileage(kmpl) because

these are considered to be the main factors in the pre owned car market.

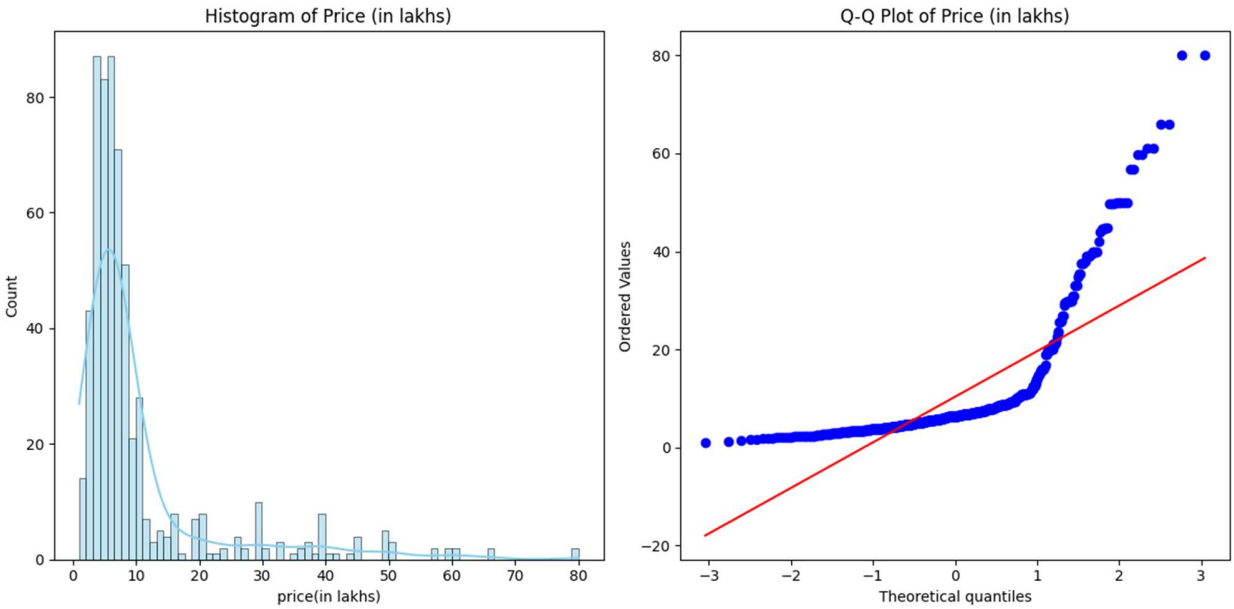
Univariate Analysis:

- Distribution Analysis: We are going to examine the distributions visually by using the histograms and Q-Q plots in order to determine the normality of the distributions.

- Hypothesis Testing: Testing the hypothesis with respect to the population drawn. (testing if the average price of cars is above a certain value) in our case we have taken the certain value of 5 Lakh Rupees.

Distribution Analysis for 'Price (in lakhs)':

Histogram and Q-Q Plot:



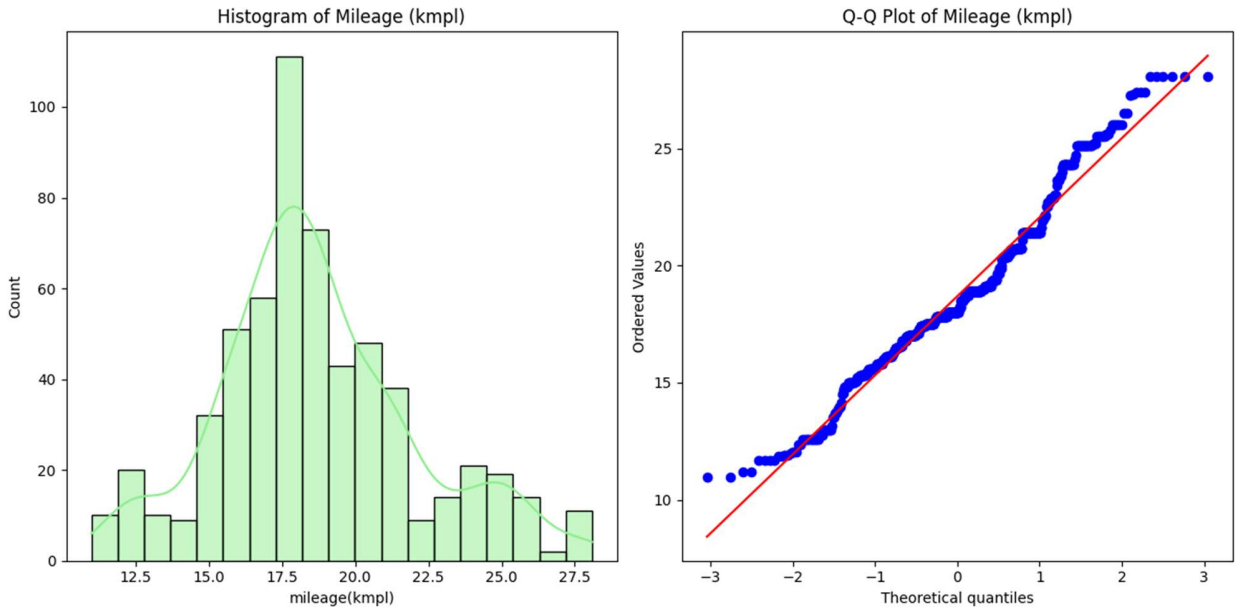
Analysis of Price (in lakhs)

Histogram: We can observe that the distribution of car prices is Right Skewed with most of the cars being priced low and there is a long tail when it comes to cars with higher prices.The histogram suggests us that most of the cars are cheap and affordable, while there are cars which are expensive as well, offering a wide price range to the customer.

Q-Q Plot: The Q-Q plot suggests us that the plot tends to deviate significantly from the straight line, especially when it comes to the upper quantiles, which confirms that there is a lack of normality in the distribution of prices. It is evident that there is skewness and also confirms the presence of outliers.

Distribution Analysis for 'Mileage (kmpl)':

Histogram and Q-Q Plot for Mileage (kmpl):



Analysis of Mileage (kmpl)

Histogram: It can be observed that the histogram for mileage is skewed to the right middle part which suggests that most of the cars have moderate mileage but there are also some cars with unusually high mileage, while there are a few cars with less mileage.

Q-Q Plot: Similar to the price distribution, the Q-Q plot for mileage deviates slightly from the line, particularly at the higher quantiles, which indicates us that the distribution is not completely normal and has some skewness.

Hypothesis Testing

We are going to Perform a simple hypothesis test to determine if the average price of cars is significantly different when compared to market-assumed average price. We are going to determine if the average price is higher than 5 Lakh Rupees.

Hypothesis:

Null Hypothesis H_0: $\mu = 5$ (The average price of cars is ₹5 lakhs)

Alternate Hypothesis H_a: $\mu > 5$ (The average price of cars is greater than ₹5 lakhs)

We are going to use one samle t-test in order to test our hypothesis.

Upon performing the one sample t-test for Price(in Lakhs), we have observed the following results,

T-statistic: 10.89
P-value: 1.30×10^{-25}

As the P-Value obtained is significantly small (way less when compared to conventional alpha level, like

0.05), we reject the Null Hypothesis. This suggests us that the average price for cars in out dataset is significantly greater that ₹5 lakhs.

Multivariable Analysis

In order to explore relationships between the variables, We have performed multivariate analysis . We are going to analyze the relationship between engine size (cc) and mileage (kmpl) using correlation and simple linear regression.

Correlation Coefficient

Upon analyzing the correlational coefficient in between engine(cc) and mileage(kmpl)we have found that it is approximately −0.431, which indicates us that there is a miderately negative relationship which suggests that, as size of engine increases, the mileage decreases, which is expected because larger engines generally tends to consume more amount of fuel.

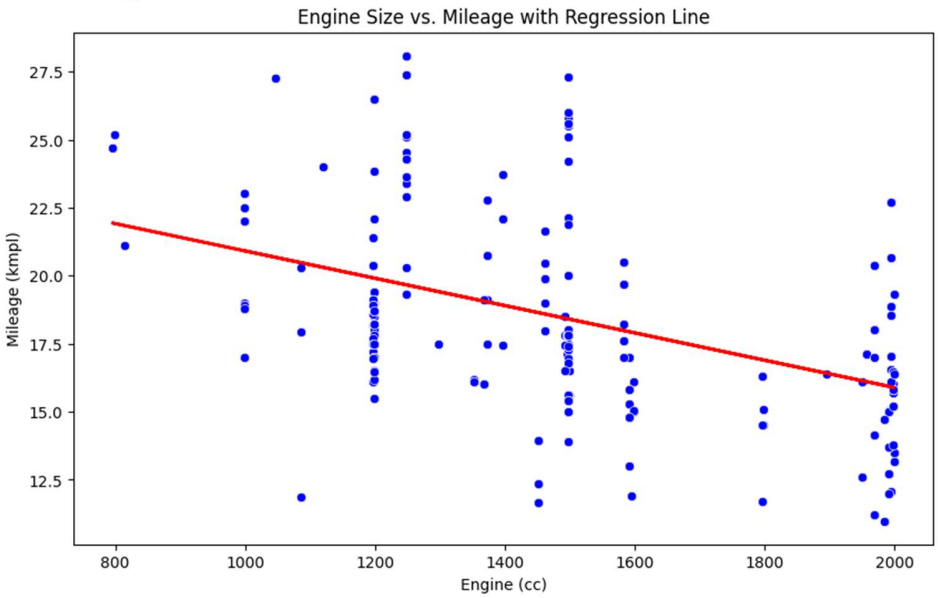
Simple Linear Regression

We have performed a simple linear regression analysis, where we have taken engine(cc) as the independent variable and mileage(kmpl) as the dependent variable.

Upon performing Linear Regression we got the Results as follows :
Slope: −0.0050
Intercept: 25.93
The linear regression equation derived from the analysis is:
Mileage (kmpl)=25.93−0.0050×Engine (cc)

This equation indicates that for every 1 cc increase in engine size, the mileage decreases by approximately 0.005 kmpl. The intercept(25.93) suggests us that if the engine size was hypothetically zero, the mileage would be around 25.93 kmpl, which serves as a theoretical baseline.

Visualization of the Regression Line:



This scatter plot with the regression line gives us a visual confirmation of the negative relationship between

engine size (‘engine(cc)’) and mileage (‘mileage(kmpl)’). The red slanting line represents the linear regression model, which comfirms the trend that as engine size increases, the mileage tends to decrease.

ANOVA Test for Price Differences by Fuel Type:

In order to statistically test if the differences in prices across fuel types are significant, we have performed an Analysis of Variance (ANOVA) Test.

Upon performing the ANOVA test, the results indicate that there are significant differences in the prices of cars based on their fuel type:

F-statistic: 22.31
P-value: 4.58×10^{-10}

The extremely low p-value (far less than 0.05) allows us to reject the null hypothesis, indicating that the average prices of used cars vary significantly among different fuel types.

Relationship between Transmission Type and Mileage (kmpl):

To analyze if there's a significant difference in mileage between cars with manual and automatic transmissions. We have performed a t-test to compare the two groups.

Upon performing the T-test, Results for Difference in Mileage by Transmission Type are as follows,
T-statistic: 11.66
P-value: 1.10×10^{-27}

The results of the t-test indicated that there is a significant difference in mileage between the manual and automatic transmissons. The positive T-statistic gives us an indication that, on an average, manual transmission cars have higher mileage when compared to those with automatic transmission. On the basis of very small p-value, we reject the null hypothesis and come to a conclusion that the difference in mileage is statistically significant.

Correlation between kms_driven and price_in_lakhs:

To determine if the distance a car travels affects the price, we will calculate the correlation coefficient between kms_driven and price_in_lakhs.

The correlation coefficient between kms_driven and price_in_lakhs is approximately −0.358, which indicates us a moderate negative relationship. This suggests us that as the kilometers range of the car increases, the price of the car decreses, which is consistent with the expectation that more usage of the car generally tends to lower the resale value.

In our project, We've conducted several analysis which explores different relationships within the dataset:

- Fuel type significantly affects car prices, as evidenced by the ANOVA Test results.
- Transmission type significantly influences mileage—with manual transmissions generally offering a better mileage when compared to automatic cars.

- A moderate negative correlation exists between kilometers driven and car prices—concluding a fact that, as the kilometers of a car increases, the price significantly decreases in most cases.

These analysis plays an important role in providing us with comprehensive understanding of the factors that influence the used car market.

6.0 Conclusion

The detailed analysis of the used car dataset helped us in disclosing several crucial insights into factors that influence prices of the car and performance metrics like mileage. Our investigation spanned from cleaning the initial data cleaning to indpth statistical inference, which helped us in providing a strong foundation for understanding the trends of the market in the used car zone.

Firstly, we identified various significant variances of cars with respect to the fuel type. The ANOVA test results showed that there is a strong statistical significance (p-value = 4.58×10^{-10}), which ondicated us that the type of fuel used plays an important role indicating the prices of a used car. Sprcifically, cars with certain fuel types demand prices on the higher side, lwhich reflects the market preferences significantly.

Secondly, we founf that the transmission type significantly affects the mileage of a car, with manual transmission vehicles having higher mileage than the counterpart. We came to this conclusion using the t-test which yeilded a p-value of 1.10×10^{-27}, which suggests us a noticable difference in fuel efficiency in between transmission types. This finding is quite crutial for prople who have fuel economy as their top priority.

Furthermore, we have explored the relationship between the kilometers driven and the car's selling price which came out with moderate negative correlation (correlation coefficient = -0.358). This suggested us that higher usage,with respect to kilometers on the car, generally results in lower resale values.

These findings are instrumental for stakeholders in the used car market, including buyers, sellers, and market analysts. For buyers, understanding these factors can aid in making informed purchasing decisions that align with their preferences for fuel type, transmission, and usage level. Sellers can better position their vehicles in the market, adjusting pricing based on these key influencing factors to optimize sales strategies. Market analysts can leverage these insights to forecast trends and advise on inventory management based on prevalent market conditions.

In conclusion, the analysis not only underscores the complexities inherent in the used car market but also highlights specific factors that significantly influence car valuation and operational performance. By continuing to delve deeper into these relationships, stakeholders can enhance their strategic decisions and operational efficiency in this dynamic market.

References

Dataset- https://www.kaggle.com/datasets/rishabhkarn/used-car-dataset

1. Pandas Development Team. pandas Documentation, version 1.3.3. Available at https://pandas.pydata.org/pandas-docs/version/1.3.3/
2. Matplotlib Development Team. Matplotlib Documentation, version 3.4.3. Available at https://matplotlib.org/3.4.3/contents.html
3. Seaborn Documentation. Version 0.11.2. Available at https://seaborn.pydata.org/

4. SciPy Community. SciPy Reference Guide, version 1.7.1. Available at https://docs.scipy.org/doc/scipy/reference/
5. Statsmodels Developers. Statsmodels Documentation. Available at https://www.statsmodels.org/stable/index.html
6. Wu, J.-D.; Hsu, C.-C.; Chen, H.-C. An expert system of price forecasting for used cars using adaptive neuro-fuzzy inference. *Expert Syst. Appl.* **2009**, *36*, 7809–7817.
7. Arawomo, D.F.; Osigwe, A.C. Nexus of fuel consumption, car features and car prices: Evidence from major institutions in Ibadan. *Renew. Sustain. Energ. Rev.* **2016**, *59*, 1220–1228.