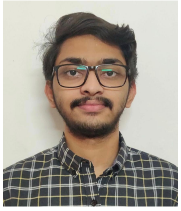
Exploratory Data Analysis

PROJECT

by

Group 4



Aadesh Minz
ID: 202103002
Course: B.Tech in Maths And Computing

Asish Joel Batha
ID: 202103015
Course: B.Tech in Maths And Computing

Raj Kariya
ID: 202103048
Course: B.Tech in Maths And Computing

Course Code: IT 462
Semester: Winter 2024

Under the guidance of

Dr. Gopinath Panda

Dhirubhai Ambani Institute of Information and Communication Technology

April 29,2024

Acknowledgment

I wanted to take a moment to express my sincere appreciation for the exceptional guidance and support you provided me during my project ”Vehicle Sales” Your mentorship has been invaluable and played a pivotal role in ensuring the success of this endeavor.

I consider myself very lucky to have had the opportunity to benefit from your expertise and mentorship. Your insightful advice, coupled with extensive knowledge in the field, has significantly influenced the quality and scope of the project. Your constructive feedbacks and suggestions not only helped me navigate challenges but deepened my understanding of the subject.

I would also like to show my gratitude to the entire team at DAIICT for fostering a culture of collaboration and innovation. The resources and facilities provided by the institution have been instrumental in facilitating comprehensive research and analysis, thereby enriching the outcome of the project.

Moreover, I am thankful to my peers and colleagues for their unwavering support and camaraderie throughout this journey. Their contributions have undeniably enhanced the development of the ”Used Vehicle sales” project.

Entering into the ”Vehicle sales” project has been an immensely fulfilling experience for me. I am confident that the knowledge and skills acquired during this endeavor will serve as a solid foundation for my future pursuits.

Once again, I want to express my deepest gratitude for your invaluable guidance and support. Your mentorship was indispensable, and I am genuinely appreciative of the opportunity to learn from you.

Sincerely,

Aadesh Minz, 202103002
Asish Joel Batha, 202103015
Raj Kariya, 202103048

DECLARATION

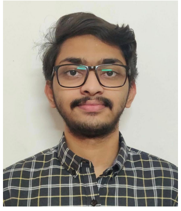
We, [202103002,202103015,202103048] now declare that the EDA project work presented in this report is our original work and has not been submitted for any other academic degree. All the sources cited in this report have been appropriately referenced.

We acknowledge that the data utilized in this project has been sourced from https://www.cars24.com/. We affirm that we have complied with the terms and conditions specified on the website for accessing and using the dataset. We hereby confirm that the dataset employed in this project is accurate and authentic to the best of our knowledge.

We acknowledge that we have received no external help or assistance in conducting this project except for the guidance provided by our mentor, Prof. Gopinath Panda. We declare no conflict of interest in conducting this EDA project.

We have now signed the declaration statement and confirmed the submission of this report on 29 April 2024.

Aadesh Minz
ID: 202103002
Course: B.Tech in Maths And Computing



Asish Joel Batha
ID: 202103015
Course: B.Tech in Maths And Computing

Raj Kariya
ID: 202103048
Course: B.Tech in Maths And Computing

CERTIFICATE

This is to certify that Group 4 comprising Aadesh Minz, Asish Joel Batha and Raj Kariya has completed an exploratory data analysis (EDA) project on the PROJECT, which was obtained from Cars24.com.

The EDA project presented by Group 4 is their original work. It was completed under the guidance of the course instructor, Prof. Gopinath Panda, who provided support and guidance throughout the project. The project is based on a thorough analysis of the Used Vehicle sales dataset, and the results presented in the report are based on the data obtained from the dataset.

This certificate is issued to recognize the successful completion of the EDA project on the Vehicle Sales, which demonstrates the analytical skills and knowledge of the students of Group 4 in the field of data analysis.

Signed,
Dr. Gopinath Panda,
IT 462 Course Instructor
Dhirubhai Ambani Institute of Information and Communication Technology
Gandhinagar, Gujarat, INDIA.

September 6, 2024

Contents

List of Figures

Abstract

Major consumer hassles with used car buying and selling. How does a customer decide on a fair price for a used car? How does a customer sort out fuel-efficient vehicles from factors affected by age and mileage?. This project is to conduct Exploratory Data Analysis (EDA) on the dataset from one of the top sources of second-hand car deals online, i.e., cars24.com. The cleaning of the dataset had been carried out for this project with handling of the missing values, standardization of formats, and encoding of categorical variables. Few of the features were engineered if necessary which helped in understanding the pattern and preference, which in turn selected the features using visualizations like scatter plots and histograms. The regression models include Random Forest, Linear Regression, which predicts the vehicle pricing and fuel efficiency results from one another. Evaluation of the models was based on MAE, RMSE, and R^2 to ensure accuracy, reliability, and deliverance of valuable insights for its customers and the business strategies in this used automobile sector.

Chapter 1. Introduction

Buying and selling used cars is is quite challenging.Customers rather more than often face challenges especially in figuring out what might be a fair price for selling or buying a used car or even figuring out what might be the best fuel efficient vehicles as there can be many factors such as age of the car, Km-driven etc.,

As a part of our course project for exploratory data analysis we have used the concepts taught in the course to solve the above problems. We have performed EDA on a dataset based on used car selling and buying which we have acquired by web-scraping a well known website Cars24.com a platform well known for buying and selling used cars.In this project, we developed prediction models for vehicle pricing and fuel economy in respect to exhaustive Exploratory Data Analysis (EDA) of a used vehicle sale dataset. In order to prepare the dataset for analysis, first, we took care of cleaning the data, handling missing values, standardizing formats, and transforming the categorical variables using encoding techniques. This was necessary to make sure that the analysis which follows is accurate and reliable.

Once the data had been cleaned, we visualized it in order to draw out underlying patterns and linkages. We used several types of charts, from scatter plots, where we could observe relationships for continuous variables, to histograms, in order to understand distributions for both continuous and categorical data frequencies. These graphics made it easy to spot some important patterns, such as the predominance of some particular auto brands or the preferences for fuel type.

We applied selection approaches to our dataset, ensured that only the important features were incorporated in our models, and derived new variables for feature engineering. This was a very important step in streamlining the model with only the main important variables, and thus enhancing the performance of the model.

After which techniques such as Random Forest, Linear Regression, among others were used to build regression models for both the pricing and fuel efficiency prediction. To confirm that these models provided accurate predictions according to our EDA insights, the evaluation of the models has been carefully made based on MAE, RMSE, and R^2. Such in-depth research, showing our ability to treat and model data, together with yielding pertinent insight, may be of help in the choice of customers and business tactics in the used automobile sector.

1.1 Project idea

This project will look at and analyze automobile sales data of used cars to find hidden patterns, trends, and insights such as what might be the price of a used car based on various parameters that may be derived from the data to help in making decisions. It will go through a number of car sales-related subjects, like mileage, pricing, and other aspects of the vehicles being sold, in order to help readers understand the dynamics at play in the used automotive market.

Objective: The main objective is to develop a predictive model in order to predict the selling price of used cars for customers using different car attributes, such as engine capacity, mileage, brand of the car, model of the car, and make year. This model should allow both the dealerships and independent sellers to place competitive prices that meet the conditions of the market and the characteristics of the vehicles.

Model type: These are the regression analysis methods by which one can apply, for predicting the prices of used cars accurately. This model would be able to quantify and explain how each of the different features influences the price of a car based on historical sales data.

Goal: Develop a model that calculates a car's fuel economy (mileage) based on the engine specs, year of manufacture, and other pertinent information. This analysis is required in order to make available to a customer who wants purchase cars who seeks to shortlist cars where fuel economy predominates, with the most mileage.

Model Type: Reliable models of forecasting automobile fuel efficiency are developed using regression analysis. Mileage forecasting from the available car data will be done with very high accuracy using modern techniques such as XGBoost and other advanced regression methods. This predictive capacity is meant to provide information to future buyers, persuading them to take up more environment-friendly automobiles.

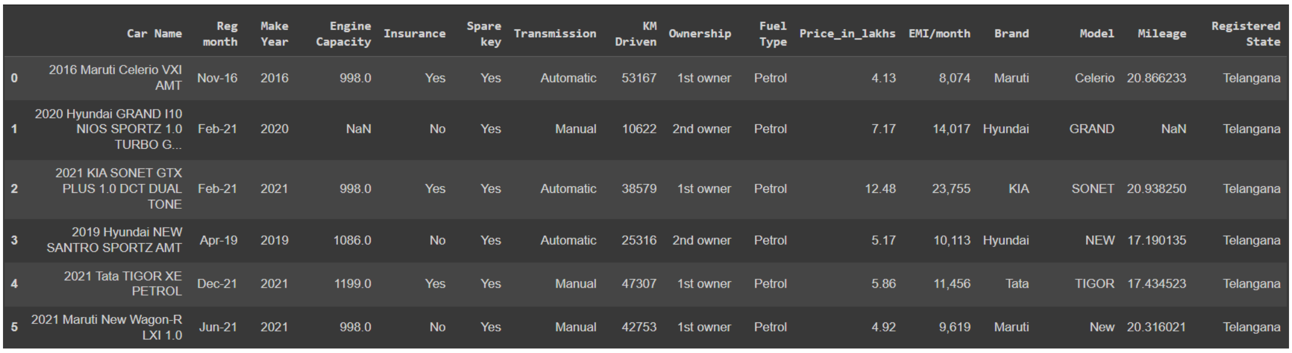
1.2 Data Collection

We have collected data from a website named cars24.com which is a website is famous selling used cars all over india. We have used a popular library which is used for webscraping named as Beautiful soup. It gives developers an easy way to parse documents in HTML and XML so they can search and navigate the web page's content's tree-like structure. Beautiful Soup is a powerful tool for extracting data from a variety of websites because it can handle malformed markup.

We have extracted data on used cars such as its name, model , manufactured year, Price of the car, Mileage, engine capacity, KM-Driven etc all the data related to a used car has been extracted from the website in this step. In the end we were able to extract a dataset of 16 columns and 700 rows consisting of various data types of numerical and categorical data of used cars of various brands dating from 2010 - 2023 and prices.

Data Sources

- We have performed web-scrapping technique by using Beautiful soup the data from a well known website known as Cars24.com which is a platform used to sell and buy used cars. You can find the link uploaded in the references section.



	Car Name	Reg month	Make Year	Engine Capacity	Insurance	Spare key	Transmission	KM Driven	Ownership	Fuel Type	Price_in_lakhs	EMI/month	Brand	Model	Mileage	Registered State
0	2016 Maruti Celerio VXI AMT	Nov-16	2016	998.0	Yes	Yes	Automatic	53167	1st owner	Petrol	4.13	8,074	Maruti	Celerio	20.866233	Telangana
1	2020 Hyundai GRAND I10 NIOS SPORTZ 1.0 TURBO G...	Feb-21	2020	NaN	No	Yes	Manual	10622	2nd owner	Petrol	7.17	14,017	Hyundai	GRAND	NaN	Telangana
2	2021 KIA SONET GTX PLUS 1.0 DCT DUAL TONE	Feb-21	2021	998.0	Yes	Yes	Automatic	38579	1st owner	Petrol	12.48	23,755	KIA	SONET	20.938250	Telangana
3	2019 Hyundai NEW SANTRO SPORTZ AMT	Apr-19	2019	1086.0	No	Yes	Automatic	25316	2nd owner	Petrol	5.17	10,113	Hyundai	NEW	17.190135	Telangana
4	2021 Tata TIGOR XE PETROL	Dec-21	2021	1199.0	Yes	Yes	Manual	47307	1st owner	Petrol	5.86	11,456	Tata	TIGOR	17.434523	Telangana
5	2021 Maruti New Wagon-R LXI 1.0	Jun-21	2021	998.0	No	Yes	Manual	42753	1st owner	Petrol	4.92	9,619	Maruti	New	20.316021	Telangana

Figure 1.1: First 5 rows of the dataset

1.3 Dataset Description

1. **Car Name:** This column contains the full name and model of the car. It is crucial for identifying the specific vehicle and can be used to analyze preferences for certain car brands and models in the used car market.

2. **Reg month:** The month and year when the car was initially registered. This information helps determine the age of the car, which is a significant factor in pricing and market valuation.

3. **Make Year:** The year the car was manufactured. The age of the car (derived from this year) can affect its price, performance expectations, and desirability in the used car market.

4. **Engine Capacity:** Measured in cubic centimeters (cc), this indicates the volume of the car's engine chamber. It is an important metric that often correlates with the power output and performance of the vehicle, influencing buyer choice and pricing.

5. **Insurance:** Indicates whether the car comes with insurance. This can be a selling point, as it adds value and provides assurance to potential buyers regarding potential future costs.

6. **Spare key:** Specifies whether an additional key is available with the vehicle. This is a minor detail but can be important for convenience and security considerations for the buyer.

7. **Transmission**: This describes whether the car has an automatic or manual transmission system. This attribute is significant as it affects the driving experience and can influence buyer preference based on ease of driving or fuel efficiency.

8. **KM Driven:** The total kilometers the car has been driven, which is a direct indicator of wear and tear. This is a critical factor in assessing the condition and value of a used car.

9. **Ownership:** Indicates whether the car was owned by a first owner, second owner, etc. The number of previous owners can affect the car's resale value, as fewer owners often correlate with better maintenance and condition.

10. **Fuel Type:** This column identifies the type of fuel the car uses (e.g., Petrol, Diesel, CNG). Fuel type impacts operating costs and environmental considerations, influencing buyer choice based on usage patterns and fuel price volatility.

11. *Price_in_Lakhs* : The price of the car listed in lakhs. It is the dependent variable for any predictive modeling aimed at determining car prices and is essential for understanding market pricing trends.

12. **EMI/month:** Estimated monthly installment if the car is purchased on loan. This can help buyers understand the financial commitment involved in purchasing the vehicle.

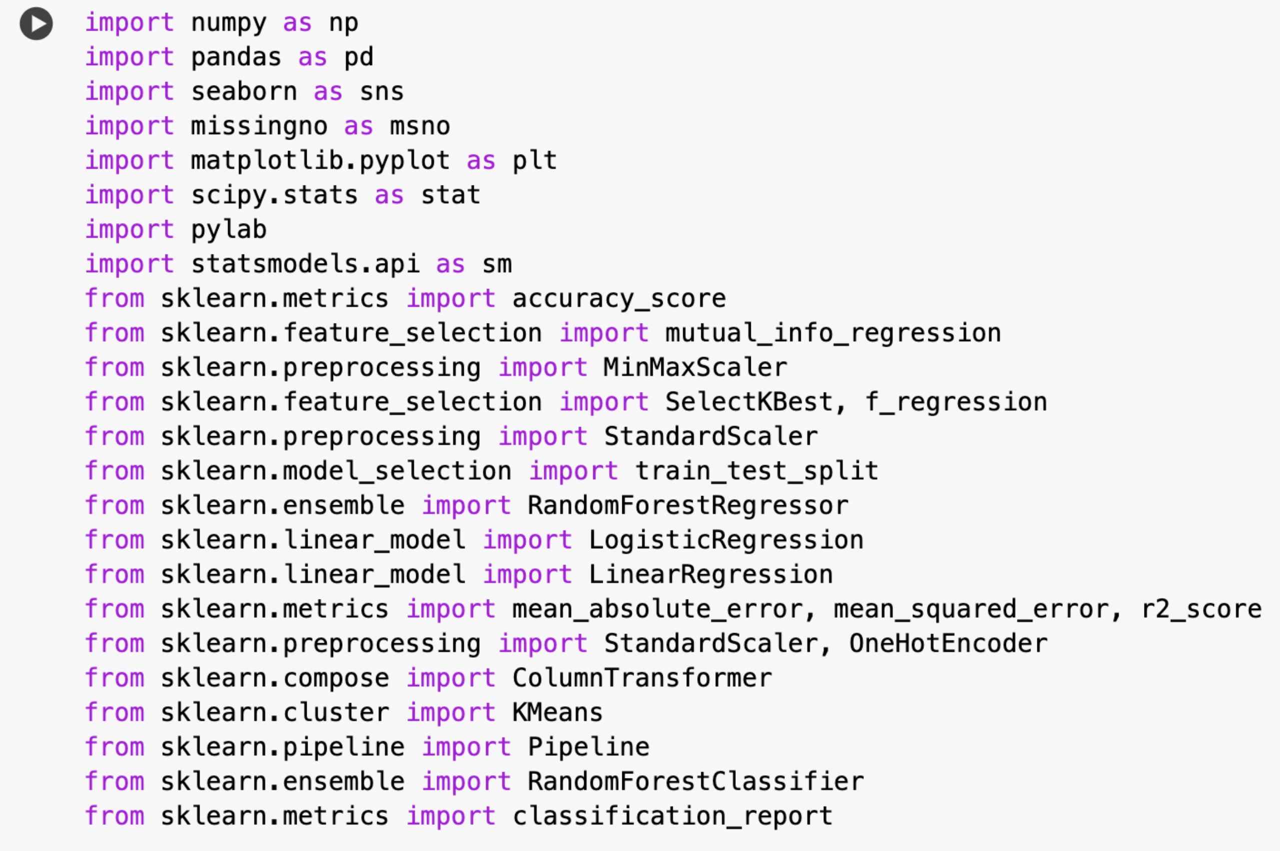
13. **Brand:** The brand of the car, which is crucial for brand-based pricing and popularity analysis. Some brands might command a premium due to perceived reliability, status, or performance.

14. **Model:** The specific model of the car under a brand. This allows for detailed analysis at a more granular level, comparing different models within or across brands.

15. **Mileage:** Fuel efficiency in km/l. Higher mileage cars are often more desirable as they imply lower fuel costs, making this a key factor in the economic decision-making of buyers.

16. **Registered State:** The state in which the car was originally registered. This can affect the resale value due to regional variations in market demand, taxation, and vehicle registration laws.

1.4 Packages required



```
import numpy as np
import pandas as pd
import seaborn as sns
import missingno as msno
import matplotlib.pyplot as plt
import scipy.stats as stat
import pylab
import statsmodels.api as sm
from sklearn.metrics import accuracy_score
from sklearn.feature_selection import mutual_info_regression
from sklearn.preprocessing import MinMaxScaler
from sklearn.feature_selection import SelectKBest, f_regression
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor
from sklearn.linear_model import LogisticRegression
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.compose import ColumnTransformer
from sklearn.cluster import KMeans
from sklearn.pipeline import Pipeline
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report
```

Figure 1.2: Required Library

Pandas (pd)

- **NumPy (np):** NumPy is a fundamental package for scientific computing with Python. It provides support for arrays, matrices, and a collection of mathematical functions to operate on these data structures efficiently. Pandas (pd): Pandas is a powerful data manipulation and analysis library. It provides data structures like DataFrame and Series, along with functions to manipulate and analyze structured data.

- **Seaborn (sns):** Seaborn is a Python data visualization library based on Matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

- **Missingno (msno):** Missingno is a Python library that provides tools to visualize missing data in datasets. It helps to understand the completeness of the dataset by displaying the distribution of missing values.

- **Matplotlib.pyplot (plt):** Matplotlib is a plotting library for Python. The pyplot module provides a MATLAB-like interface for creating static, interactive, and animated visualizations.
- **Scipy.stats (stat):** Scipy is a scientific computing library for Python. The stats module provides a wide range of statistical functions and distributions, including hypothesis tests, probability distributions, and statistical functions.
- **PyLab:** PyLab is a module that bundles several libraries, including NumPy, SciPy, and Matplotlib, into a single namespace for easy access to their functions. It is often used for interactive scientific computing and plotting.
- **Statsmodels.api (sm):** Statsmodels is a Python library for estimating and interpreting statistical models. It provides classes and functions for various statistical models, including linear regression, generalized linear models, and time series analysis.
- **Scikit-learn (sklearn):** Scikit-learn is a machine learning library for Python. It provides simple and efficient tools for data mining and data analysis, including supervised and unsupervised learning algorithms, cross-validation, and feature selection.
- **OneHotEncoder:** OneHotEncoder is a preprocessing technique in scikit-learn used to convert categorical variables into a binary matrix representation.
- **ColumnTransformer:** ColumnTransformer is a utility in scikit-learn for applying different transformations to different columns of a dataset, allowing for flexible preprocessing pipelines.
- **KMeans:** KMeans is an unsupervised learning algorithm in scikit-learn used for clustering data into groups based on similarity.
- **Pipeline:** Pipeline is a utility in scikit-learn for chaining together multiple preprocessing and modeling steps into a single object, enabling streamlined workflows.
- A Python package called Beautiful Soup is used for online scraping, specifically for obtaining data from XML and HTML documents. It gives users an easy-to-use

interface for searching and manipulating the contents of web pages by browsing the parse tree created by various markup languages. Beautiful Soup is an effective tool for collecting data from web pages for a variety of uses, including automation, research, and data analysis. It tackles basic parsing issues including encoding detection and tag balance.

Chapter 2. Data Cleaning

Data cleaning is a very fundamental part of the process of data analysis. It ensures the quality of data that forms the basis for analysis accurate and reliable. This section comprises the steps of data cleaning that took place, including how issues such as missing values, inconsistent formatting, and outliers in the vehicle sales dataset were handled.

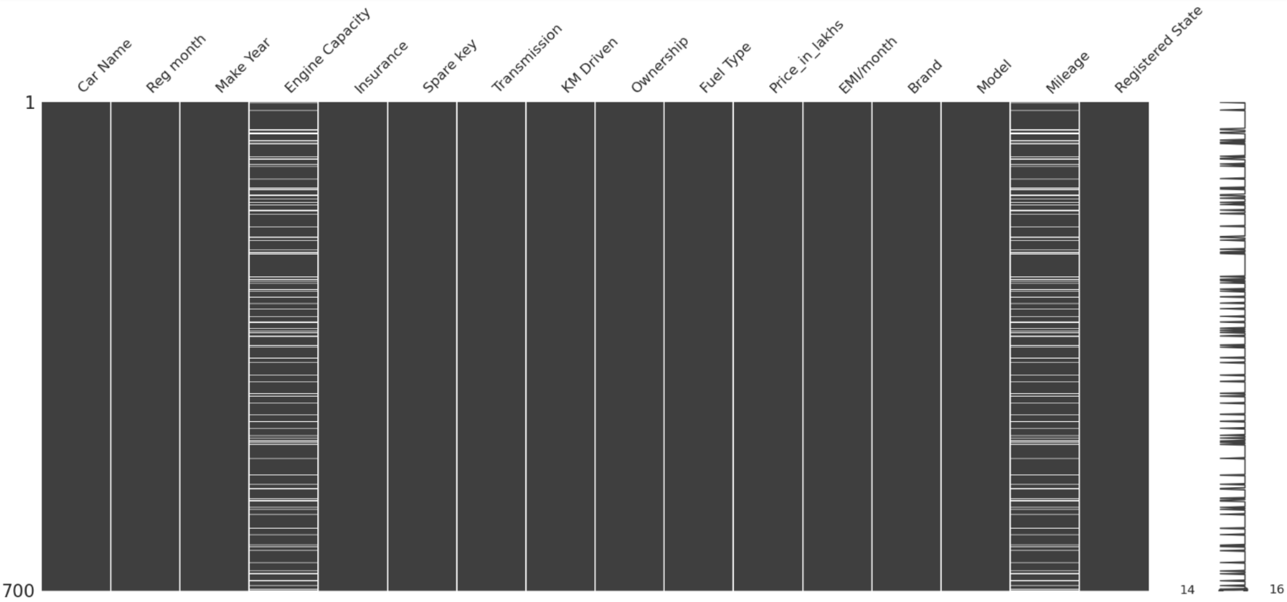


Figure 2.1: Required Library

2.0.1 Missing Data Analysis

Here we have performed missing data before the doing all of the analysis such as visualizing, encoding, selection etc., as missing values in a dataset create ambiguity and bias in prediction and model fitting in the data, which then results in biased or misleading results. From the dataset which we have acquired the missing values occurred in two columns :

- Engine Capacity: There were around 86 entries missing, which represented the volume of the car's engine chamber. Missing values here can influence any analysis related to the performance of the vehicle such as Mileage, Fuel efficiency etc.,
- Mileage: In this column around 86 of the entries were missing, which are key points to make decisions affecting fuel efficiency and cost-effectiveness, determining purchase considerations by buyers.

The presence of missing values in these columns could therefore alter any conclusions drawn for factors influencing car prices and preferences.

2.0.2 Imputation

In order to handle the missing values effectively, the following imputation strategies were employed:

- Engine Capacity: We have used mean valued imputation method for engine capacity, because it is a common practice to use the mean when the data distribution is not heavily skewed, in as much it provides a centrality measure less likely to introduce bias when compared with the median or mode in this context.
- Mileage: Here we have use the median Imputation methods The missing values in the mileage column were imputed with the median. This approach is used mostly when handling outliers or skewed data, as the median gives a robust central value with little influence from the extremes.

These imputation methods are used such that there are no bias while filling the missing values and hence will provide us more accurate and close to real world data. Which is good for performing analysis.

2.0.3 Data Formatting and Cleanup

Further steps in the data cleaning process were standardizing the formats of numerical and categorical data.

- Categorical Data: The features Insurance and Spare key have been converted from 'Yes'/'No' to a boolean format (True/False). Such a standard eases analyses involving logical operations or comparisons.
- Numerical Data: The field 'EMI/month' had comma separators between the numbers. Hence the data in the field did not read as numerical. Conversion of this

data type into float and elimination of such separators signified allowing smooth progression of subsequent numerical operations. Addressing these data cleaning tasks, the dataset was prepared for robust and reliable exploratory analysis, visualization, and modeling.

By addressing these data cleaning tasks, the dataset was prepared for robust and reliable exploratory analysis, visualization, and modeling.

Chapter 3. Visualization

As we know in EDA visualization plays an integral part as it is the graphical representation of data to communicate insights, patterns, and relationships effectively. It involves using visual elements such as charts, graphs, and maps to convey complex information in a clear and intuitive manner.

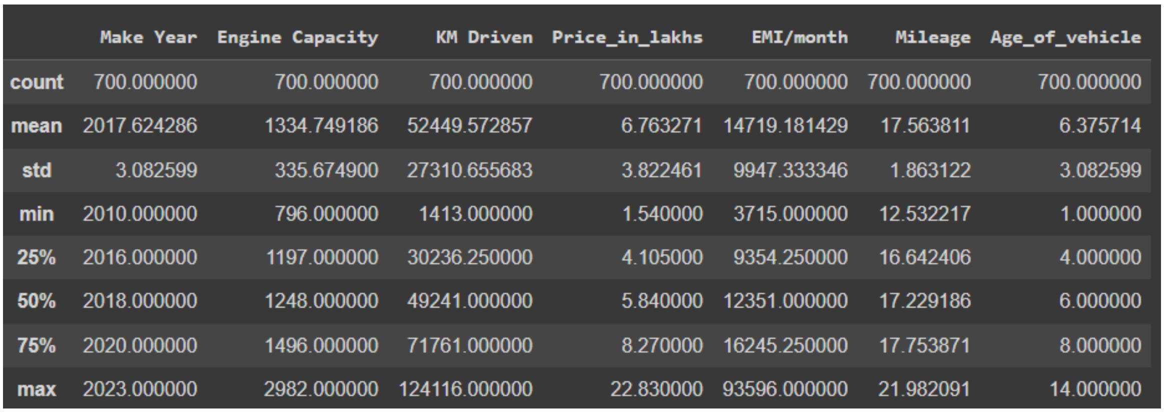
In data analysis and exploration, visualization plays a crucial role in understanding the characteristics and distributions of the data, identifying trends and outliers, and making informed decisions. So here in the this project we have several plots such as

Histograms - which is used to display the frequency distribution of numerical data.
Bar plots - which is used to compare the values of categorical variables.
Pie charts - used to show the proportions of different categories in a dataset.
box plots - used to summarize the distribution of data and identify outliers.
Violin plots - which combine aspects of box plots and kernel density estimation to visualize the distribution of data
Scatter plots - which is use to show the relationship between two numerical variables by plotting individual data points on a Cartesian plane, revealing patterns, correlations, and outliers in the data through their distribution and clustering.

Along with these basic plots we have shown some line graphs to show the trends of the variables with respect to time and KDE plots to show the density and distribution of the data.

3.1 Univariate analysis

Here in Univariate analysis which is basically analysis pertaining and conducted on only one column such as distributions and shapes of the data set and various statistical analysis.



	Make Year	Engine Capacity	KM Driven	Price_in_lakhs	EMI/month	Mileage	Age_of_vehicle
count	700.000000	700.000000	700.000000	700.000000	700.000000	700.000000	700.000000
mean	2017.624286	1334.749186	52449.572857	6.763271	14719.181429	17.563811	6.375714
std	3.082599	335.674900	27310.655683	3.822461	9947.333346	1.863122	3.082599
min	2010.000000	796.000000	1413.000000	1.540000	3715.000000	12.532217	1.000000
25%	2016.000000	1197.000000	30236.250000	4.105000	9354.250000	16.642406	4.000000
50%	2018.000000	1248.000000	49241.000000	5.840000	12351.000000	17.229186	6.000000
75%	2020.000000	1496.000000	71761.000000	8.270000	16245.250000	17.753871	8.000000
max	2023.000000	2982.000000	124116.000000	22.830000	93596.000000	21.982091	14.000000

Figure 3.1: Summary Statics of numerical datatype

Graph Description(Fig 3.2): Displays the frequency distribution of engine capacities among vehicles in the dataset. Observations: The distribution is skewed to the right from the histogram. A greater proportion of the cars, in that case, lie with engine capacities on the lower side, while very few lie with much bigger capacities. This show that among the range of engine capacities there is an inclination of consumers

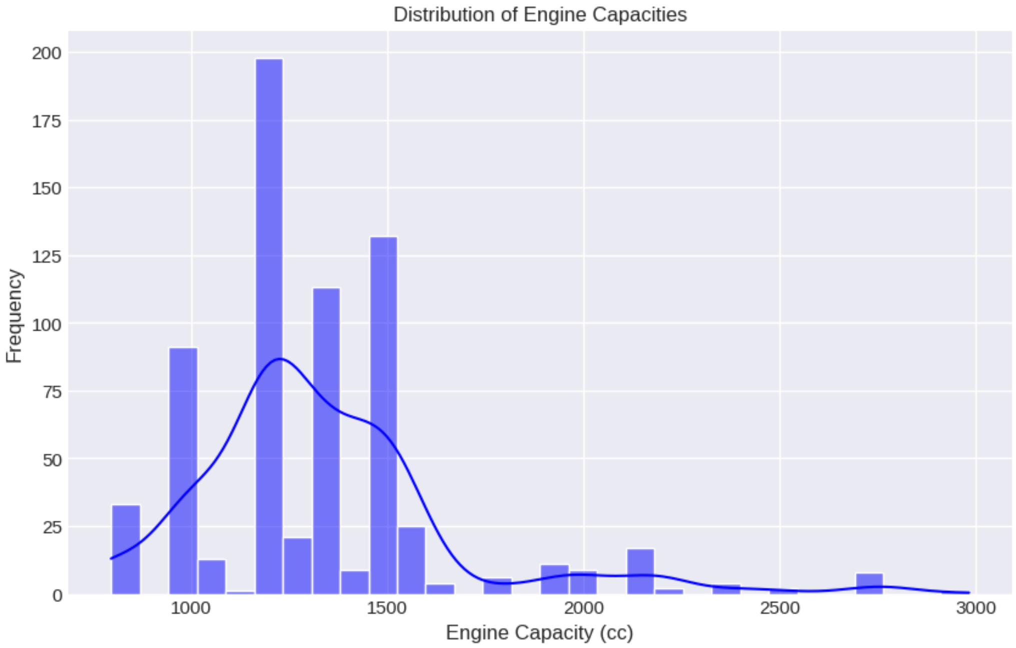


Figure 3.2: Distribution Of Engine Capabilities

towards anything more compact or cheap which might be due to cheaper prices and more mileage.

Graph Description (Fig 3.3): Displays the frequency distribution of mileage among vehicles in the dataset.

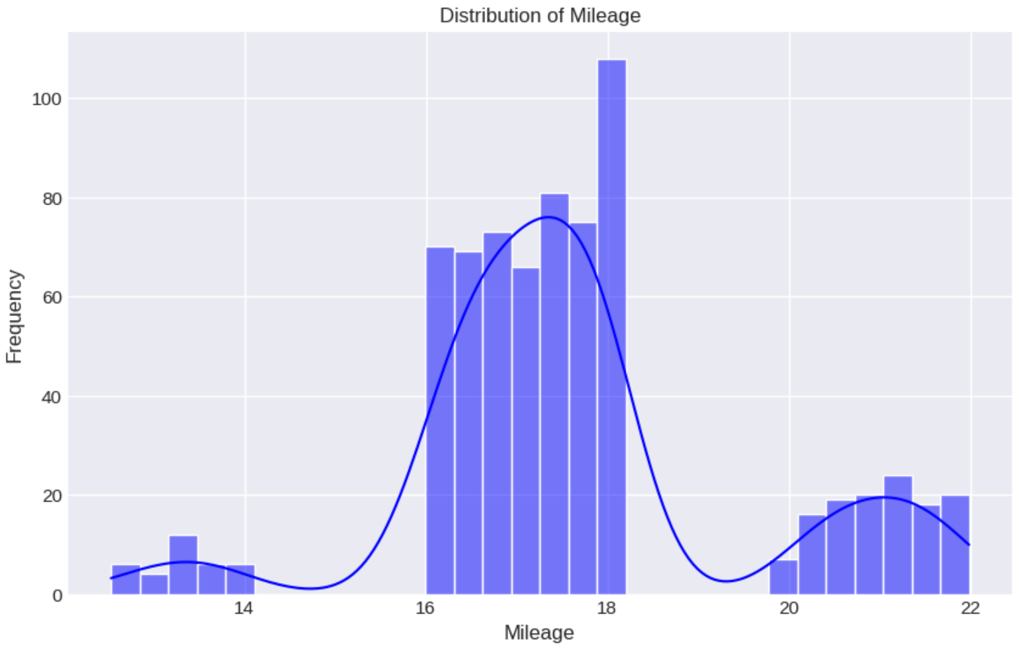


Figure 3.3: Distribution Of Mileage

Observations : The histogram for the distribution of mileage exhibits a bimodal distribution, where there are two peaks, suggesting two different groupings or types of vehicles within the dataset. The primary peak occurs around the $17 - 18km/l$ range, while a secondary, smaller peak is present around $21km/l$. The above distribution is line with engine capacity as greater the engine capacity the lower the mileage we will get and since we can see in the above graph of engine capacity shows that the maximum engines have the capacity have the range around 1000-2000cc which is being reflected in the mileage by having primary peak at 17-18 km range and the second one at 21-22km.

The mean and variance of mileage is $mean = 17.61km$, $variance = 3.94km$ respectively.

Graph Description (Fig 3.4): The histogram displays the distribution of kilometers driven for vehicles in the dataset, with the x-axis representing kilometers driven and the y-axis showing the frequency of vehicles.

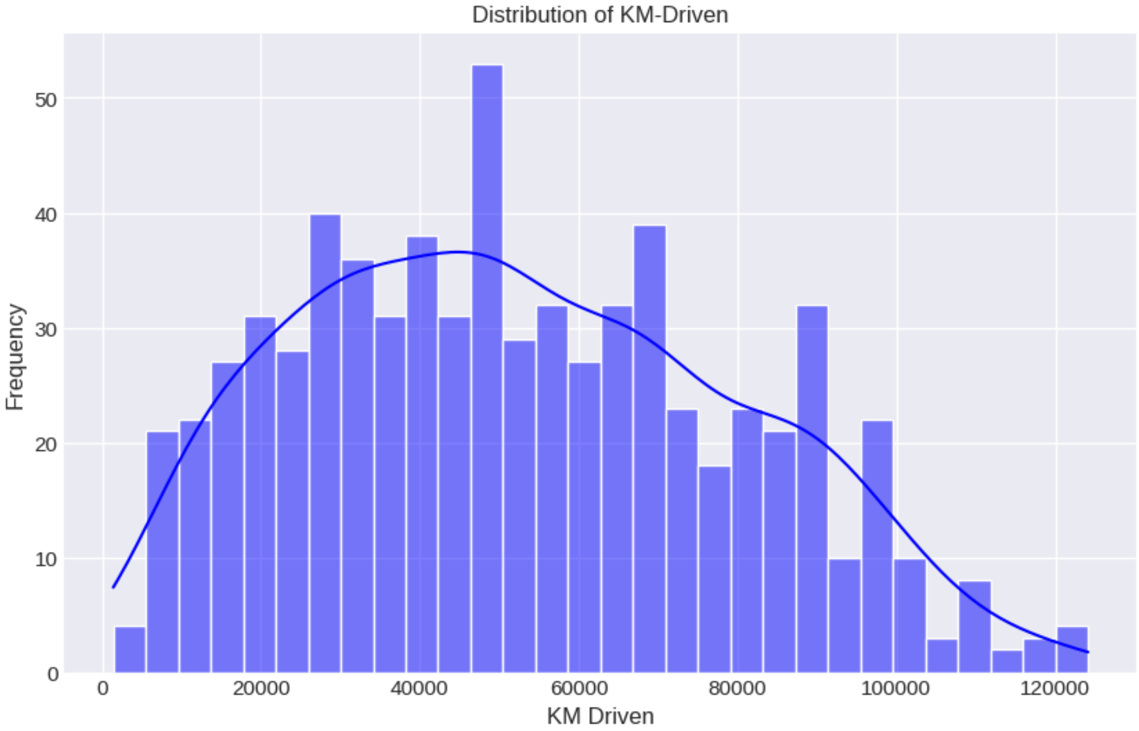


Figure 3.4: Distribution of KM-Driven

Observations : Overlay of the line plot with a histogram brings out the trend of the distribution clear, depicting how similar the plots are across each kilometer range. Still, many vehicles have significantly higher kilometers on them, but the distribution is right-skewed, peaking at lower kilometers. This suggests most vehicles are driven relatively less. This shows that so many used cars with relatively less driven , are preferred more. The observed skewness of the data distribution shows the possible consequence for car valuation, as greater mileage values are often synonymous with a corresponding lesser resale value because of more use and a shorter expected life.

Graph Description (Fig 3.5): The distribution of vehicle prices (in lakhs) within the dataset is shown on the graph, which is a histogram with an overlaying curve. The x-axis represents price ranges, and the y-axis shows the frequency of automobiles falling into each price range.

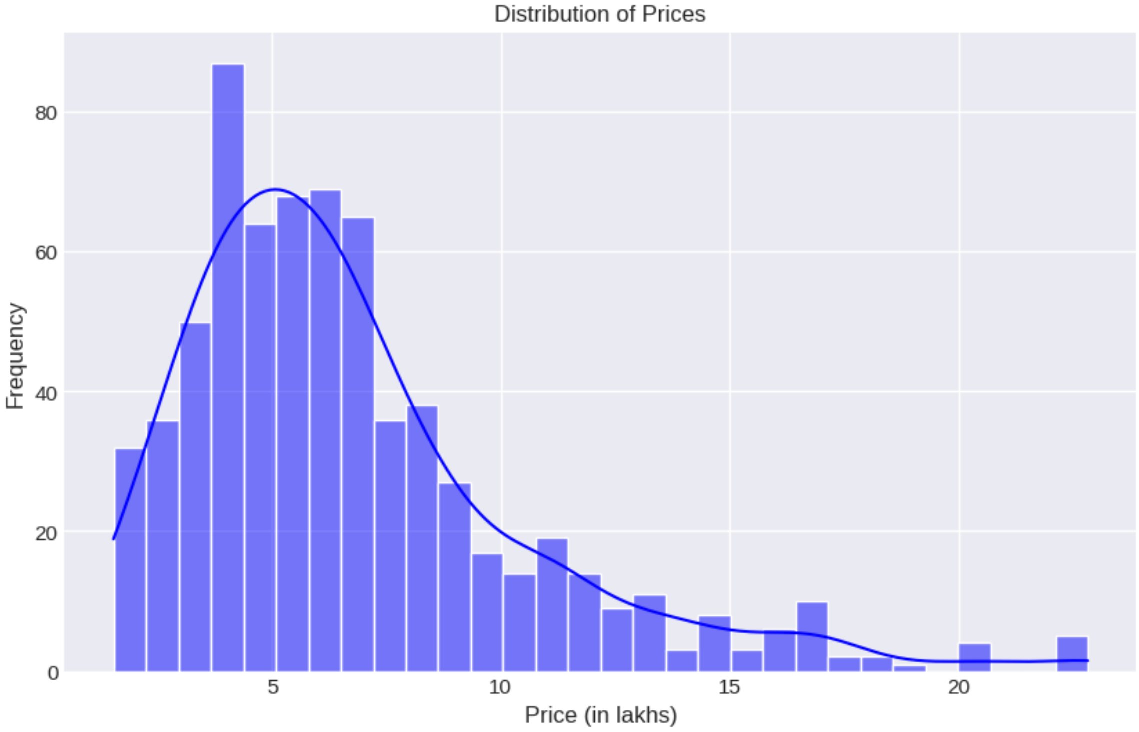


Figure 3.5: Distribution of Prices

Observations : The analysis of the histogram shows that the car prices have a right-skewed distribution, with the mode at the lower end of the price and the tail pointing higher. This indicates that many of the cars are falling in the lower price range and not so many of them are falling in the higher price range. Even though they are outliers with respect to the bulk of the data, the existence of outliers at abnormally high prices is pointed out. While an absence of luxury or high-cost vehicles would point towards the possibility that either the used-car market has less supply or less demand for luxury or high-cost vehicles, the frequency's concentration at the lower end, on the other hand, signifies a market where either cheap vehicles are common or are highly in demand.

Graph Description (Fig 3.6): The histogram shows the distribution of Equated Monthly Installment (EMI) amounts for the automobiles in the dataset. With the x-axis representing the monthly EMI amount and the y-axis representing the frequency of occurrences within the dataset.

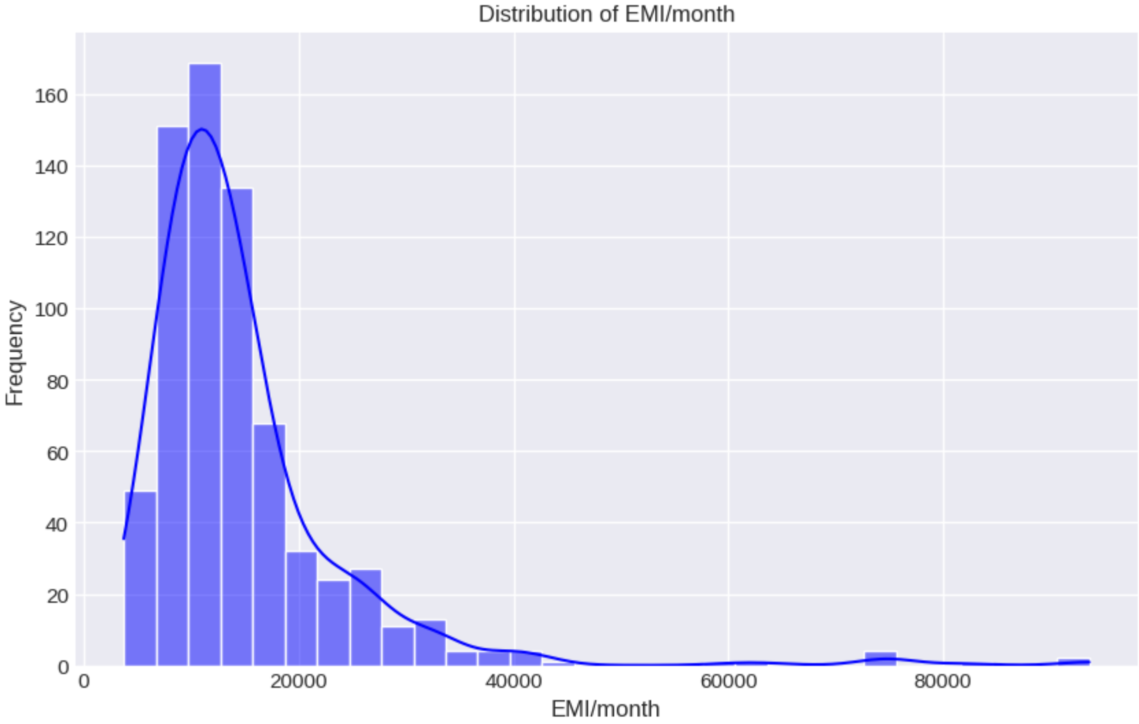


Figure 3.6: Distribution of EMI/Month

Observations : The histogram shows the distribution of Equated Monthly Installments (EMI) paid for cars; it is a right-skewed pattern with a very pronounced peak at lower EMI values, and the tail extends far from it to higher EMIs. This distribution may reflect the case that a greater number of lesser-priced vehicles are being financed, since it shows most of the vehicles have lower monthly payments. And even if it's much higher EMI, some nature has very few, which shows that there exist luxury cars with huge financing commitments. The tails of the distribution would also be affected by the longer-term loans or by higher rates associated with certain classes of auto loans. Further, the right skewness observed in the price and EMI distributions indicates a probable relationship in the sense that cheaper cars usually have smaller EMIs. Such relationships between the EMI amounts, vehicle prices, and kilometers driven further enforce how strong the relation between several numbers of influencing factors gets distributed with the EMIs, such as loan terms and interest rates.

Graph Description (Fig 3.7): The x-axis shows the months of the year, while the y-axis shows the number of automobile registrations in each month. The bar chart shows the counts of car registrations by month. Every bar on the graph represents a separate month, and the height of the bar indicates the total number of cars registered in that month.

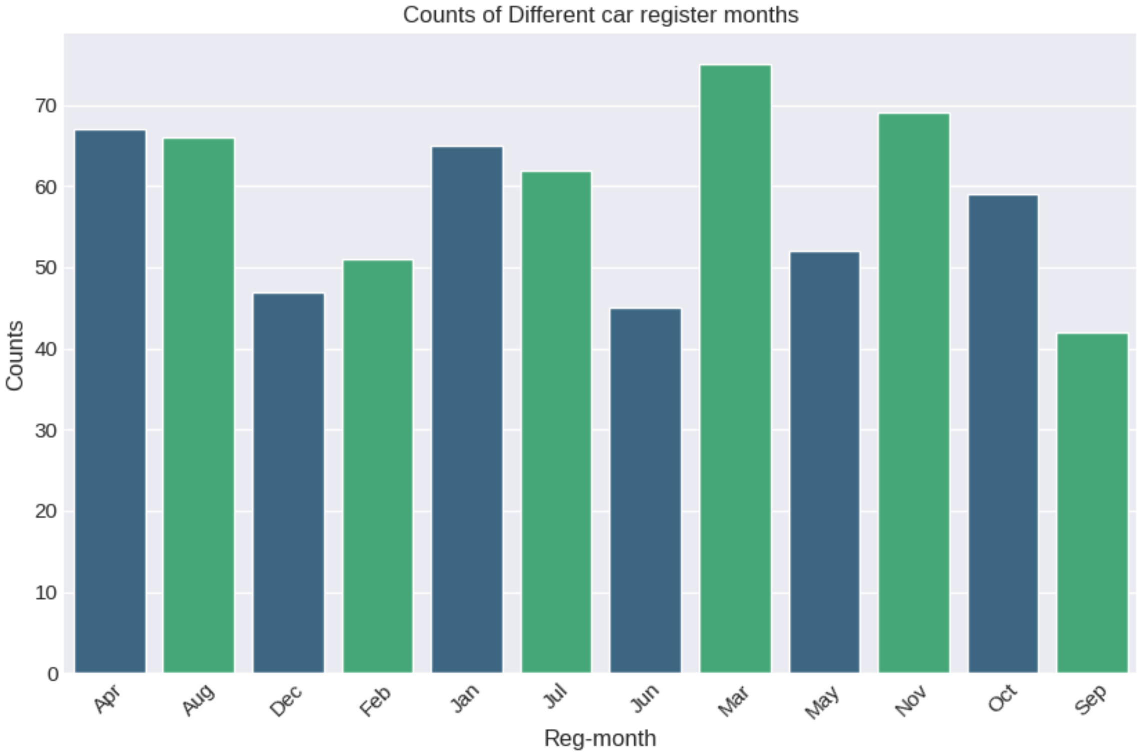


Figure 3.7: Counts of different car register months

Observations : The car registrations by month bar chart shows a distribution that lacks a clear trend, fluctuating from month to month without a regular pattern of rising or falling numbers. If some kind of seasonality is, in fact, apparent, then there might be no possibility to identify any recurrent peaks or troughs. Besides, the data is not skinned, thereby indicating that month-on-month variations in registration numbers are highly random. Some observation points to higher registration numbers in some months due to such events as festivals, financial year-ends, or new car model launches, while in some other months, lower figures may be occasioned by adverse weather, a slowdown in economic activities, or fewer promotional offers on the part of car sellers.

Graph Description (Fig 3.8): The counts of cars grouped by model year are shown in the bar chart. Vehicle model years are listed on the x-axis, while the number of vehicles for each year is indicated on the y-axis. Every bar is a distinct model year, and the height of the bar indicates the quantity of cars from that year in the collection.

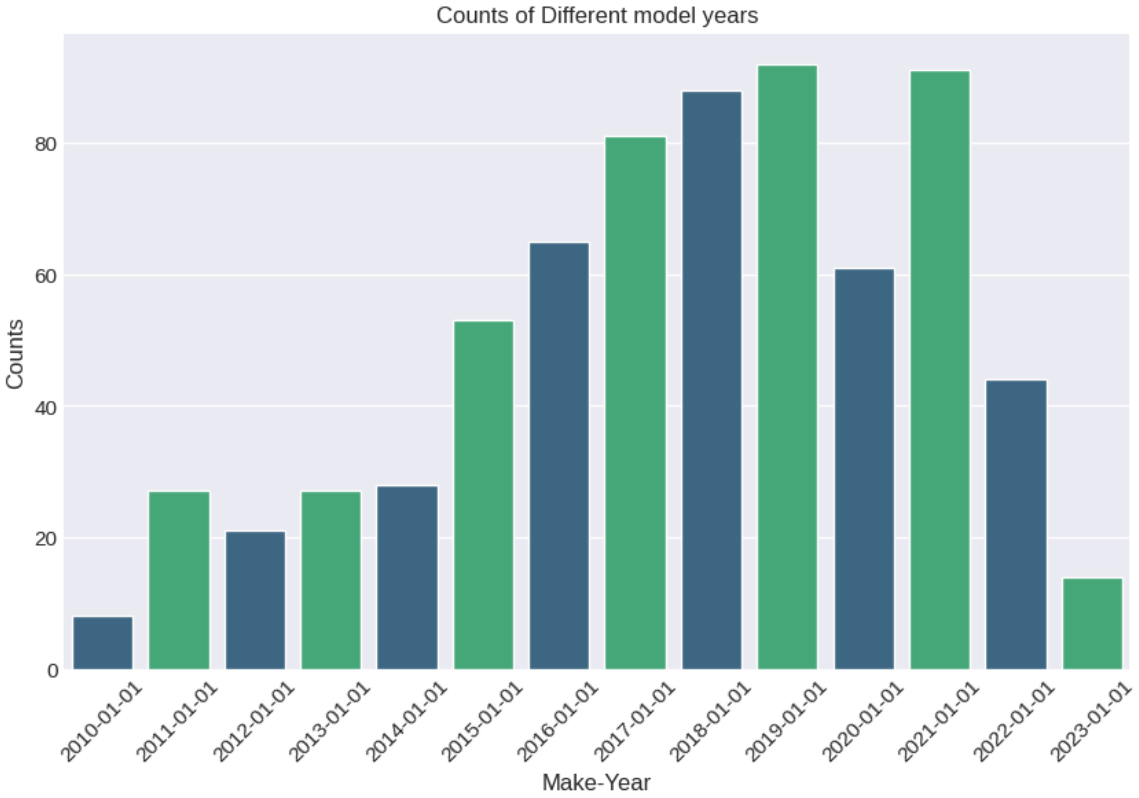


Figure 3.8: Counts of Different model years

Observations : It can be seen from the bar chart below, representing vehicle counts by model year, that a rising trend in the counts of vehicles as the model years approach the present suggests that newer models are more common in the dataset. This clearly gives the pattern of increasing gradually the production and sales of more recent models or a preference by the market for newer models.

Graph Description (Fig 3.9): The bar chart displays vehicle counts categorized by their insurance status. The x-axis separated the vehicles into two groups based on whether they are insured ('True') or uninsured ('False'). The y-axis indicates the number of vehicles within each category. Two bars are present, each representing the count of vehicles for the corresponding insurance status.

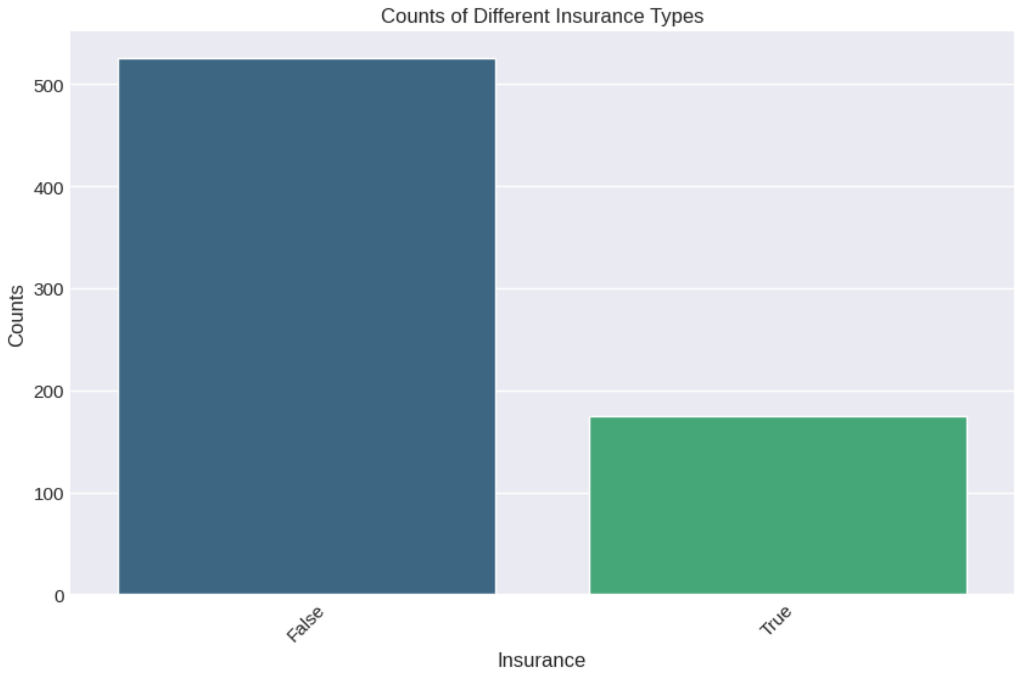


Figure 3.9: Counts of Different Insurance Types

Observations : This plot shows the the number of cars between the insured status (displaying "True") and not insured status (displaying "False"). The graph highlights how most cars in the dataset are assured, which might be attributed to the choice of the owner for safeguard, regulation from lenders, or regulatory requirements. This could have an impact on the resale value or act as a determinant for a potential buyer of whether to purchase the vehicle or not, even though there seems no visibility of any link with the period of registration, model years, or EMI values. This may affect the attractiveness of insured automobiles to buyers and may affect the distribution of prices and then EMI variations.

Graph Description (Fig 3.10): The bar chart presents vehicle counts categorized by the availability of spare keys. The x-axis divides vehicles into two groups: those possessing a spare key ('True') and those without one ('False'). Meanwhile, the y-axis denotes the number of vehicles within each category.

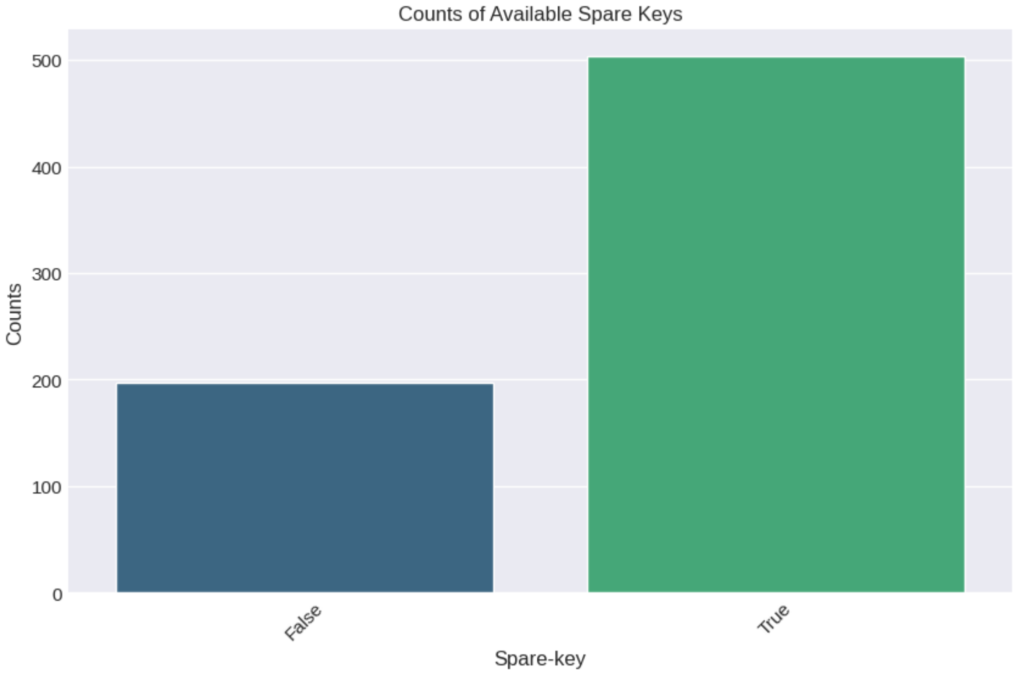


Figure 3.10: Counts Of Available Spare Keys

Observations : On spare key availability, the bar chart also displays a high variation between vehicles that are with or without spare keys. Skewness analysis cannot apply, as the data is purely categorical and presents only two possible outcomes. Still, it is apparent that most of the respondents really prefer their cars coming with a set of spare keys. When buying a used car, the availability of spare keys is very crucial since it tells more about how well the car was taken care of by the previous owner. Although, might have some small effects, such as a consumer might be more interested and assured to buy that car. Over can have a little increase in price.

Graph Description (Fig 3.11): Vehicle counts are shown in the bar chart according to transmission types. Vehicles are divided into two transmission types using the x-axis: automatic and manual. In the meantime, the number of cars for each type of transmission is indicated on the y-axis. The total number of cars with either an automatic or manual transmission is shown by each bar.

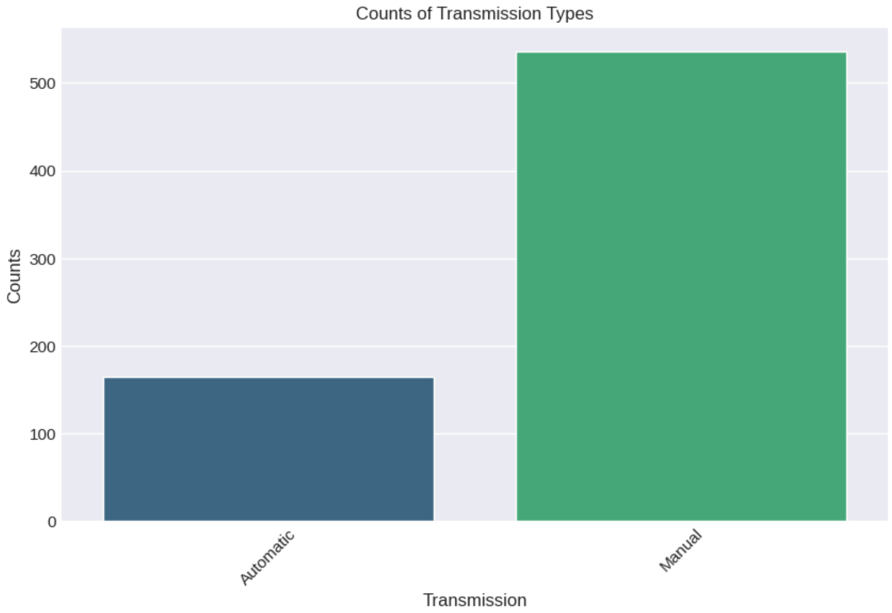


Figure 3.11: Counts Of Transmission Types

Observations : It shows a bar graph of the number of cars in manual and automatic transmissions; apparently, the number of vehicles transmitted manually far outweighs those transmitted automatically. This, in reality, means the car market in the data set apparently leans to have more of the manual transmissions or it is more prevalent in manuals. The higher figures for manual transmissions would likely be due to cheaper prices, control, or the involvement in driving perception of ease of access in the areas where data was collected. While no clear correlation of registration months or even model years is seen, the increase in numbers of the automatic counts might be representing growth in automatic transmission adoption in newer model years. Further, in an indirect way, the type of transmission has an influence on price and EMI distribution, as automatic vehicles carry relatively higher prices and may affect the overall price distribution and subsequently EMIs.

Graph Description (Fig 3.12): The dataset's distribution of car ownership kinds is shown in the bar chart. Cars are sorted by ownership status on the x-axis and are marked as ”first owner,” ”second owner,” and ”third owner.” The number of cars for each ownership type is shown on the y-axis. The number of cars that fall under each of the aforementioned ownership categories is shown by each bar.

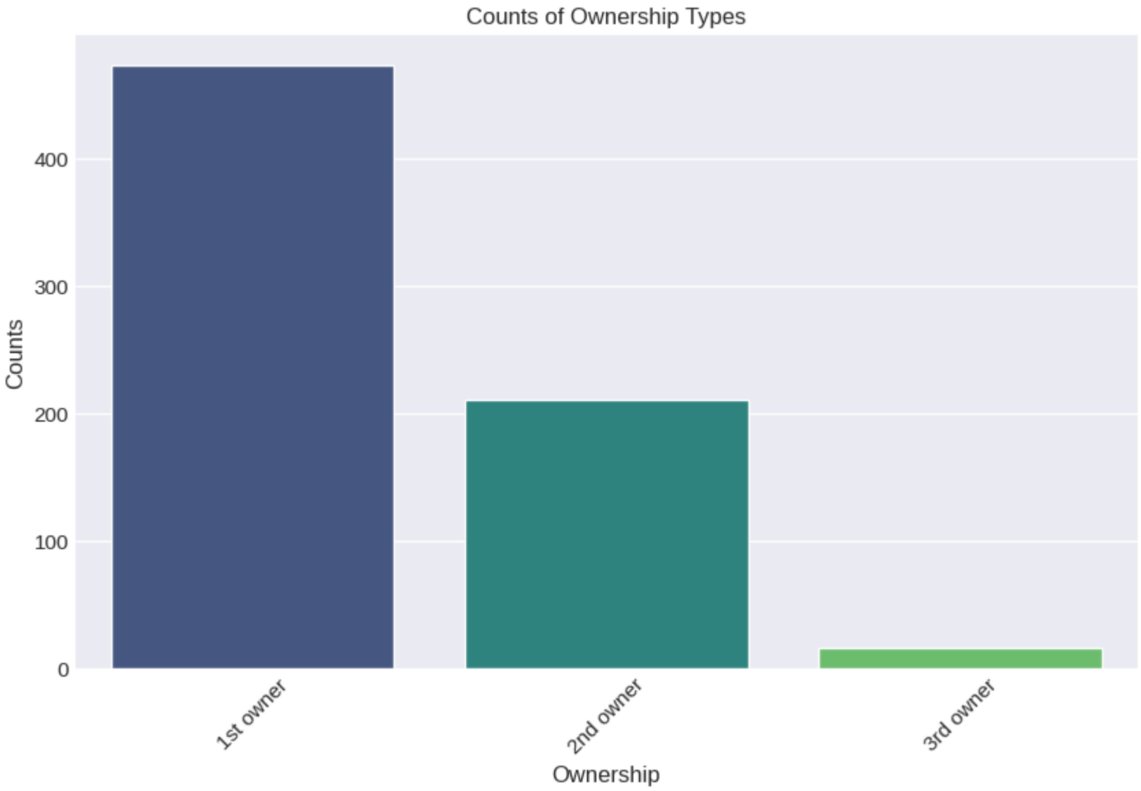


Figure 3.12: Counts of Ownership Types

Observation : The categories of car ownership from this dataset are distributed as shown in the bar chart: most of the vehicles belong to the first owner, fewer to the second owner, and even fewer to the third owner. On the other hand, it points toward the much higher prevalence of newer or less traded cars in this market, usually with only one previous owner. Cars of newer model years might be less exposed to getting resold, consequently affecting the registration month distribution in an indirect manner.

Graph Description (Fig 3.13): The bar chart illustrates the frequency of vehicles categorized by fuel type within the dataset. On the x-axis, different fuel types are listed: CNG, Diesel, and Petrol. The y-axis denotes the count of vehicles associated with each fuel type. Each bar on the chart represents the total number of vehicles for the corresponding fuel type.

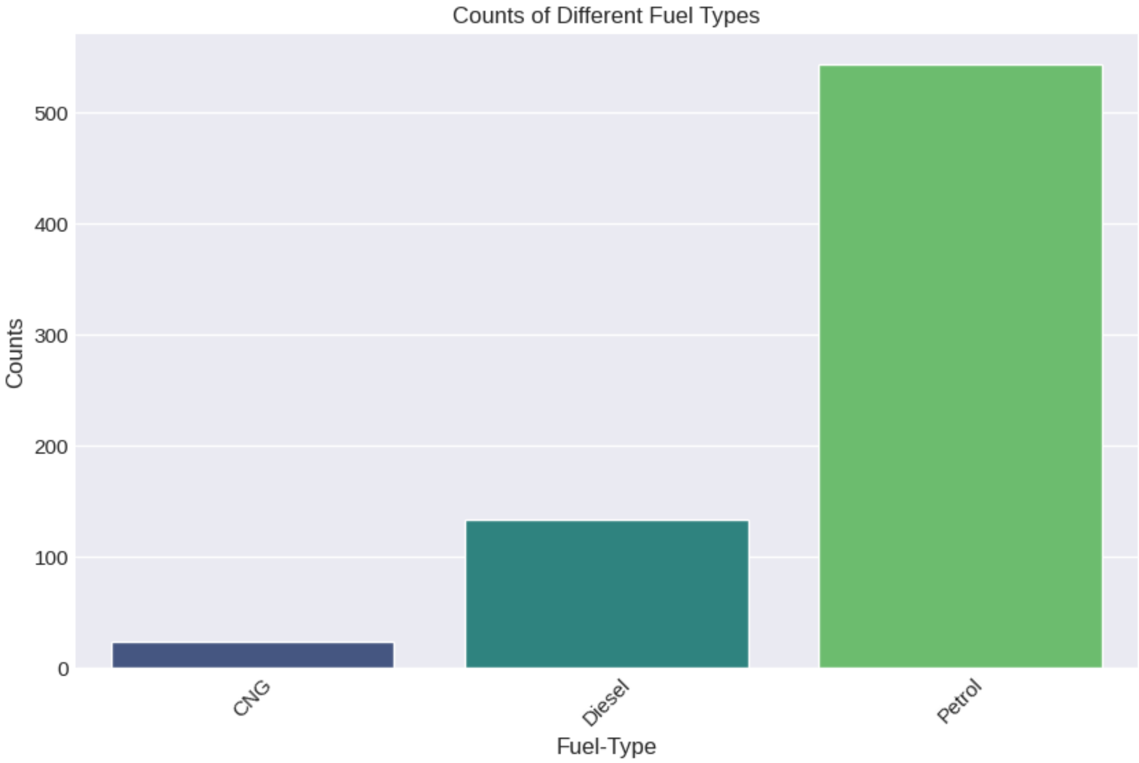


Figure 3.13: Counts of Different Fuel Types

Observations : The fuel type distribution of vehicles indicates that gasoline is the most preferred fuel, followed by diesel, and CNG vehicles are the least prevalent. This represents customer preferences or what's available on the market, and it might also be related to price distributions.

Graph description (Fig 3.14): The bar chart visually presents the distribution of vehicles categorized by their brand within the dataset. On the x-axis, various car brands are listed, while the y-axis indicates the count of vehicles associated with each brand. Each bar on the chart corresponds to the number of vehicles from a specific brand, providing an overview of the dataset's composition in terms of vehicle brands.

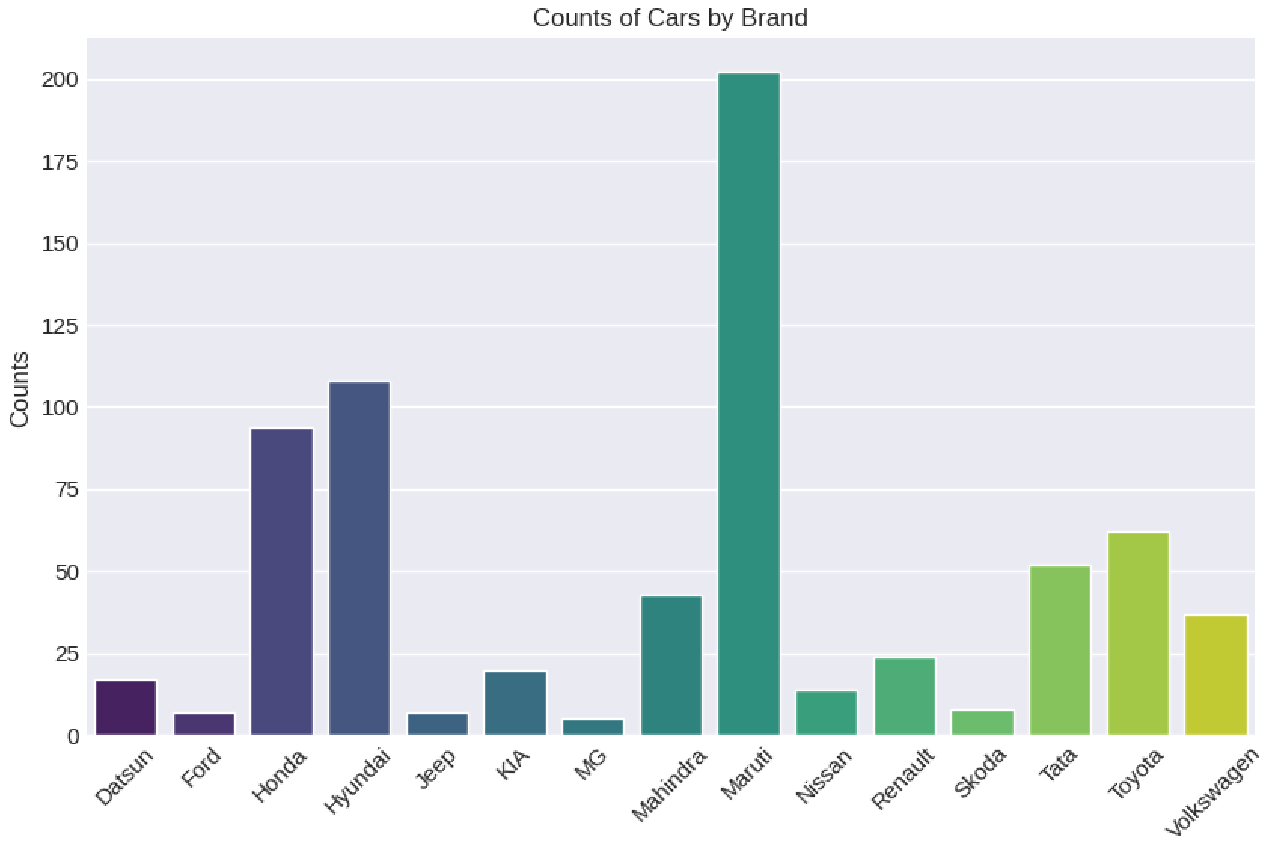


Figure 3.14: Counts Of Cars By Brand

Observations : A bar chart is produced for the distribution of vehicles by brand, where one brand has the largest count displayed dominantly, meaning there is widespread use of that particular brand or appeal in automobiles. There will be differences in the distribution of brands, and a few less-strong brands compared to others. There may be relationships between the most common brands and the most common fuel type, transmission type, and ownership percentage of plot linkages associated with other plots. The distribution of brands may also affect the distribution of prices, as some brands may relatively be more expensive than others, where the interest rate may also be higher. This is a graph to show the distribution of registered cars per state in India for the given data, which indicates that the state with the highest number of registered cars has just over double the state with the lowest number of registered cars.

Graph Description (Fig 3.15): The bar chart visualizes the distribution of registered vehicles categorized by Indian states within the dataset. On the x-axis, the names of Indian states are listed, while the y-axis indicates the count of vehicles registered in each state. Each bar on the chart represents the total number of vehicles registered in a specific state, providing an overview of vehicle registration distribution across different regions of India.

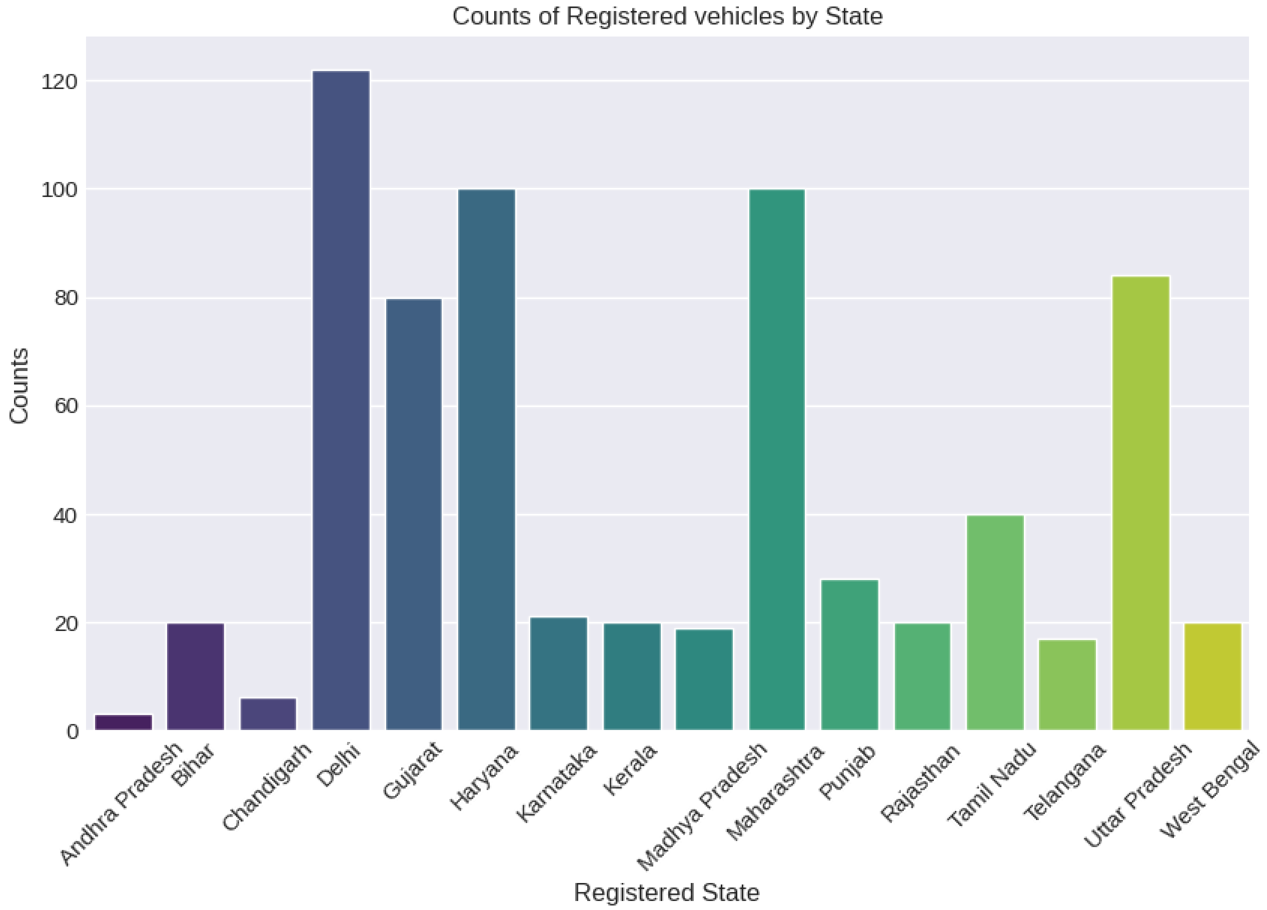


Figure 3.15: Counts Of Registered Vehicles By State

Observations : This is because the number of states having too many numbers could be more affluent, have large urban populations, or have higher car sales rates, which lead to a higher number of people having a car. Perhaps certain brands are more prevalent in a particular state than the other; thus, it can be said that this method of interpreting the brand distribution chart is rather elusive. But in the distribution of car registrations by state, this may not show a direct relation to the fuel types, brands, or kinds of ownership. States with higher car registration rates may also have an impact on price and EMI distributions because higher registration areas may see an increase in demand, and hence, higher car costs.

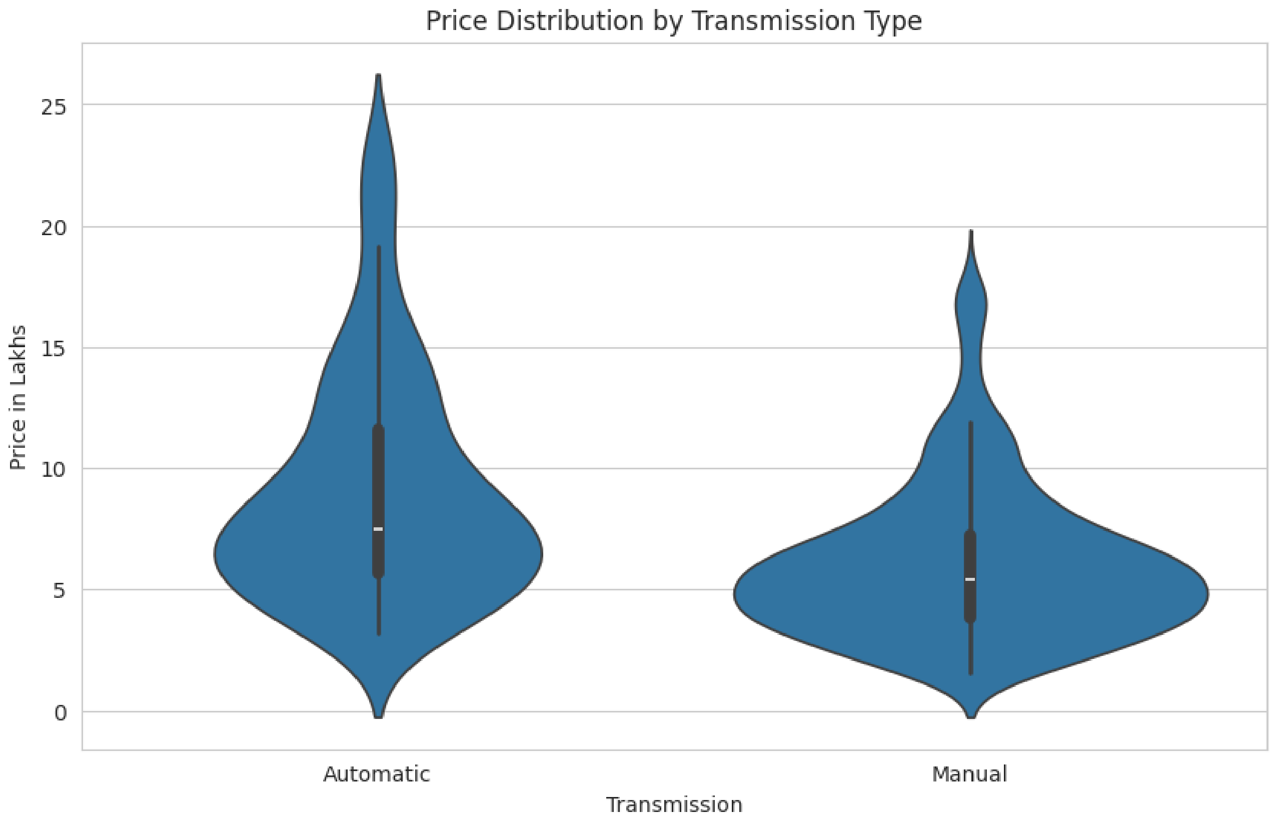


Figure 3.16: Price Distribution by Transmission Type

Observation (fig : 3.16): The price distributions for cars with automatic and manual transmissions are shown in the violin plot. We can observe that there is wider density for Manual Cars in price range of 5 - 8 lakhs which suggests that in budget price range manual transmission dominates whereas automatic tend to bit more costlier.

3.2 Multivariate Analysis :

Graph Description (Fig 3.17): The scatter plot visualizes the relationship between the manufacturing year of vehicles and their prices. On the x-axis, the manufacturing years of the vehicles are listed, ranging from 2010 to 2022, while the y-axis denotes the price of the vehicles in lakhs. Each dot on the plot represents an individual vehicle, positioned based on its manufacturing year and corresponding price.

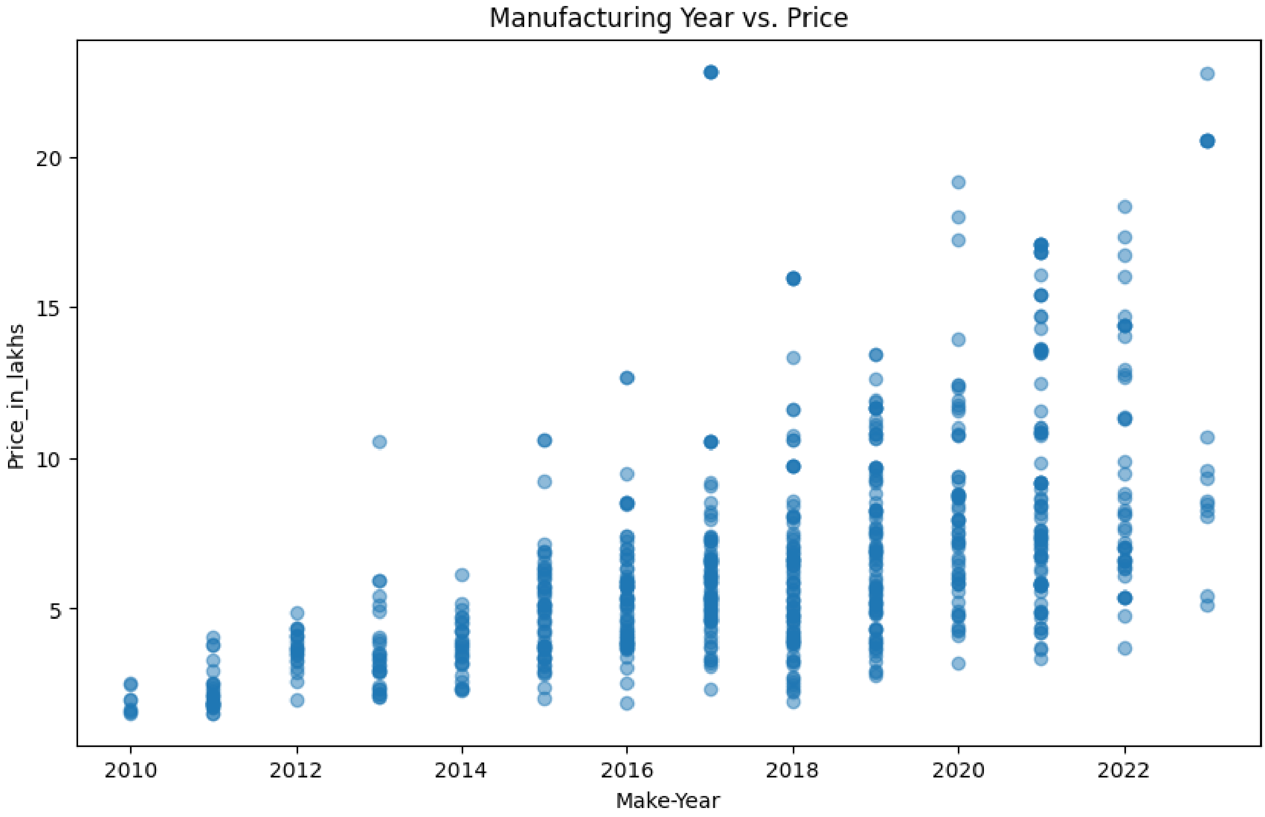


Figure 3.17: Manufacturing Year vs. Price

Observation : The scatter plot describes the relationship between the price of a car and the year of the car. It generally shows an increase in prices with an increase in manufacturing year, indicative of the usual depreciating effect on automobiles. Even within the same year of manufacture, there is a great difference in price, which means other factors, such as features, brand, model, mileage, and condition, are influencing the cost. The plot shows the spread and dispersion of vehicle prices in many years of manufacture. Relationship with other plots: The scatter plot might have an indirect relationship to the count of cars by brand or by fuel type, because some brands or fuel types are more or less used in certain manufacturing years and because they may affect the average price. Other than that, the year of make is likely to relate to the distribution of ownership types, since a new car would have had fewer previous owners compared to an old one.

Graph Description (Fig 3.18): The scatter plot illustrates the relationship between engine capacity and the price of vehicles. Engine capacity, presumably measured in cubic centimeters (cc), is depicted on the x-axis, while the price of vehicles in lakhs is represented on the y-axis. Each point on the plot represents a vehicle, positioned according to its engine capacity and corresponding price.

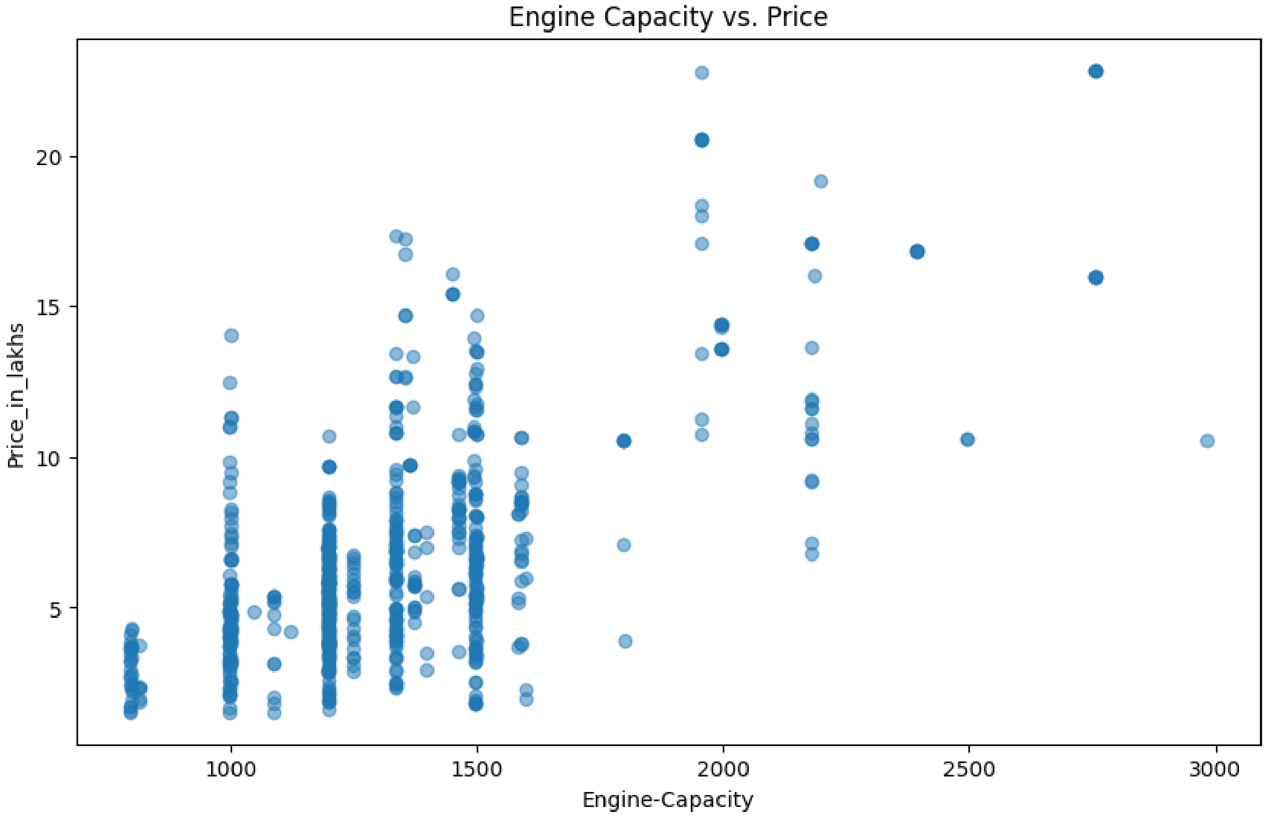


Figure 3.18: Engine Capacity vs price

Observation(Fig 3.18): The scatter diagram, which depicts a general trend of rising costs with bigger engine capacities, shows the association between engine capacity and car pricing. Nevertheless, there is a sizable price variance at every capacity level, indicating that additional factors affect the cost of an automobile. Engine displacement and brand-specific automobile distributions may be related; certain brands prefer larger engines, which may have an effect on pricing.

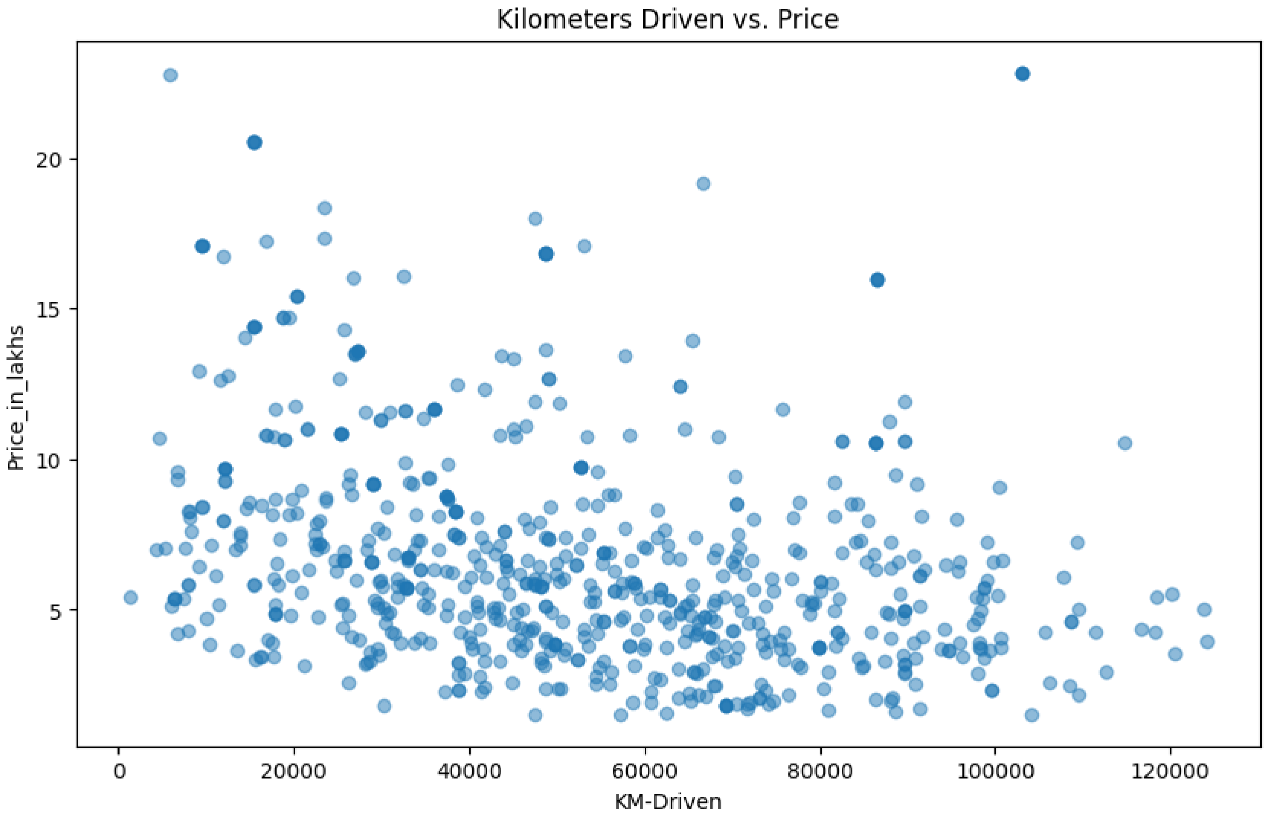


Figure 3.19: KM Driven vs. Price

Observation (Fig 3.19): From the scatter plot of price and kilometers, key findings about the relationship of the two variables with the price of the vehicle are revealed by the analysis. Most of the data points are clustered in the bottom left of this graph, indicative of the fact that cars usually cost less with less mileage. This illustrates a clear negative correlation with kilometers driven and pricing, where pricing gradually decreases with an increase in mileage. The plot shows denser distributions for lower kilometers and wider spreads for the larger ones. Outliers, like expensive cars with high mileage, may make an inference of deviations from the norm, which will typically mean luxury or collector type vehicles.

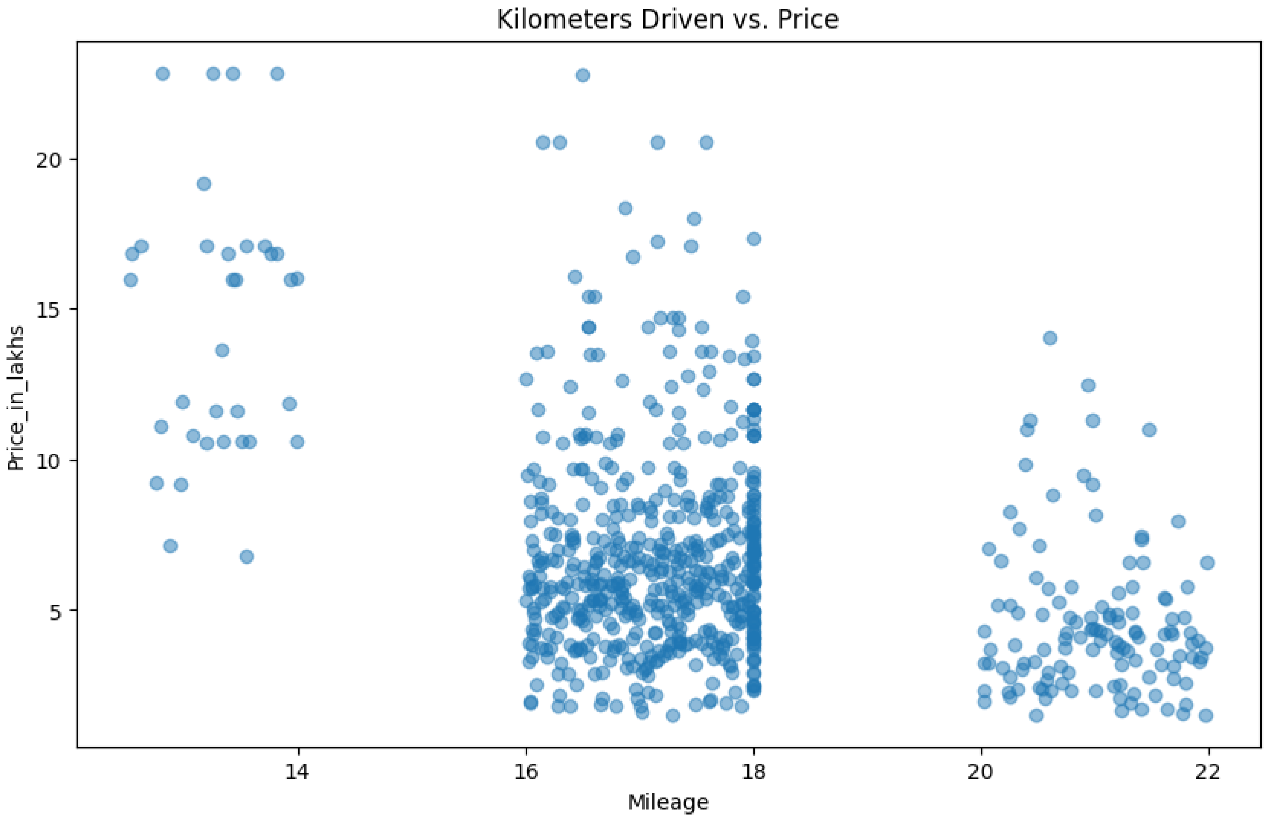


Figure 3.20: Mileage vs Price

Observation (Fig 3.20):This variable relationship of more mileage to cheaper or higher prices basically means that there must be some other important variables in consideration. Outliers with a higher mileage and price could be a signal that they are premium or efficient cars, while outliers with lower mileage but a higher price could denote performance cars or vehicles with desirable features beyond fuel efficiency. Overall tendencies show that the new car is much more expensive, and generally, the larger capacity of the engine links with a higher cost.

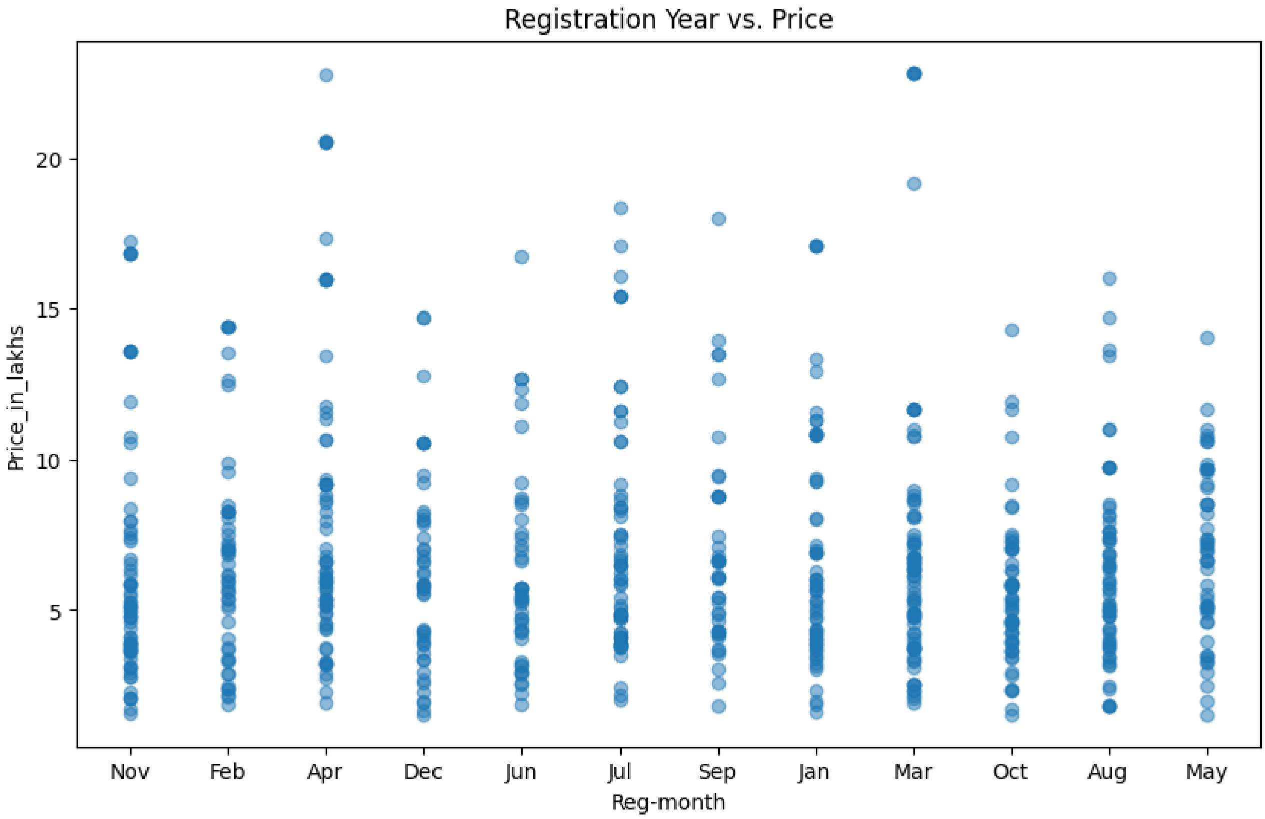


Figure 3.21: Registration Year vs Price

Observation (Fig 3.21): The price and production year are positively related, meaning newer cars have better conditions and improved amenities. A higher engine displacement will mean more than power; it may signal more costly vehicles, i.e., higher-end vehicles, for which indeed the cost is affected. Here's where this trend goes in time: more and more powerful cars. Unlike earlier studies, the ones that have depicted the price in context with the mileage and kilometers driven, this emphasizes how important the engine capacity and production year are while deciding on the price for a car.

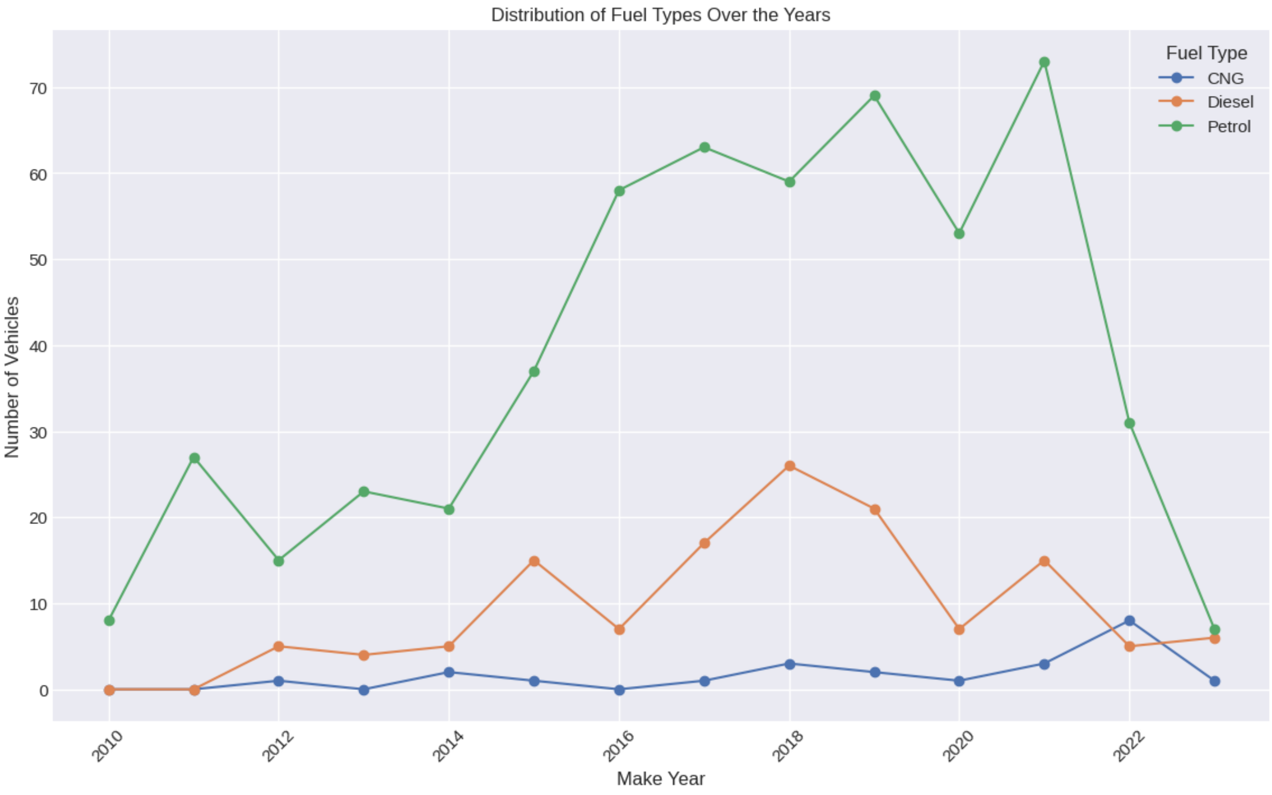


Figure 3.22: Distribution Of Fuel Types Over the Years

Observation (Fig 3.22): The line plot above indicates an upward trend for car prices over the production years, all things being considered concerning the fuel type. The reason behind the observed increase might be changes in consumer preferences, inflation, or technological changes. Diesel cars, on the other hand, normally have a higher selling price, more so in the last years, probably due to better fuel consumption and technology advancement. During 2018, though, the years before have shown an increase in the cost of diesel cars, maybe as a result of growing demand or advancements in technology. In total, CNG vehicles usually have the same price, while with new, the price hike is in most cases the highest. Most probably, a higher number and model varieties are available in this category: from the entry type to the luxurious type with additional features, given the price differential for the diesel cars.

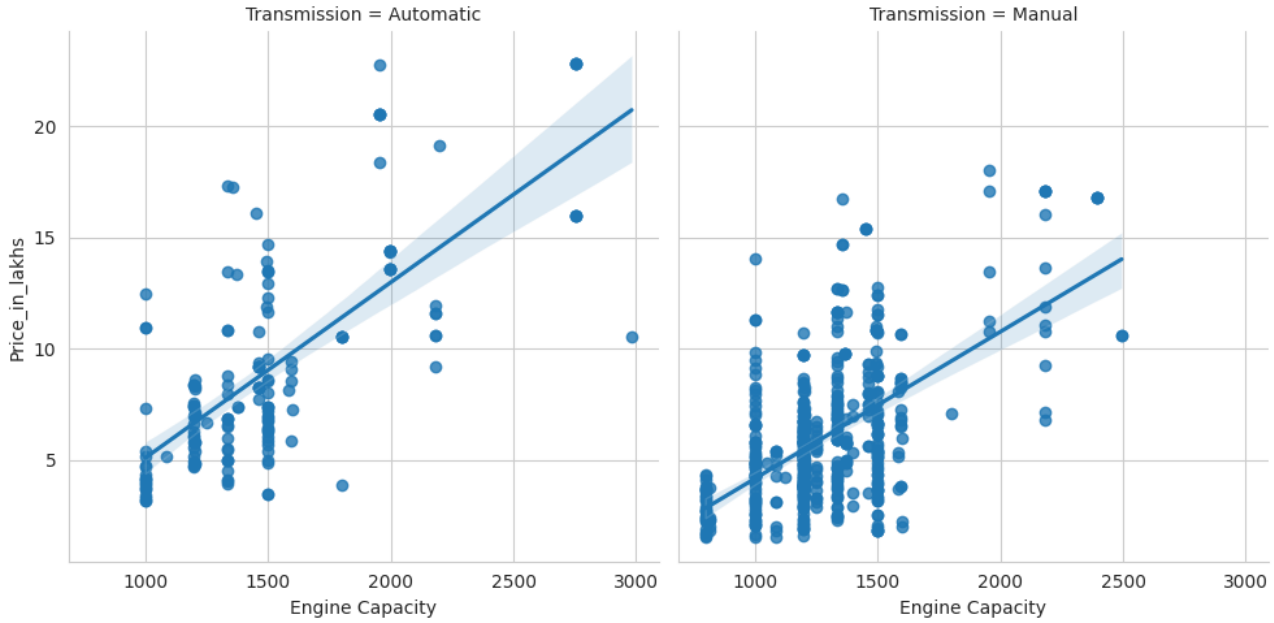


Figure 3.23: Engine Capacity Transmission (Type = Automatic and Manual) vs price

Observation (Fig 3.23): The scatter plots show that, for both manual and automatic transmissions, there is a positive link between engine size and vehicle price. All engine sizes show that automatic cars are always more expensive; a larger range of costs suggests a more varied selection. Price distributions for cars with manual transmissions are denser, especially for mid-range engine capacities. The trend line for automatic automobiles is steeper, indicating a higher link between engine capacity and cost. This association may be attributed to changing consumer preferences and the dynamics of the market for automatic transmissions.

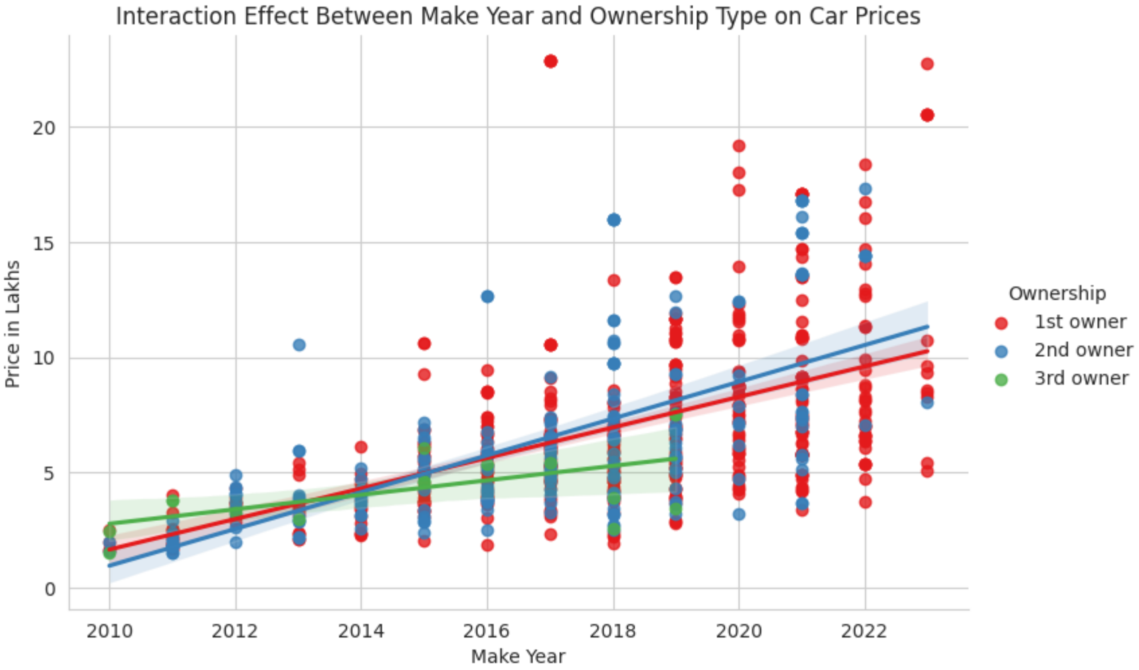


Figure 3.24: Interaction Effect Between Make Year And Ownership Type on Car Prices

Observation (Fig 3.24): As can be observed in the scatter plots above, mostly cars that have a single owner fetch more compared to those which have changed hands several times. Prices show depreciation since they are on the decline with every subsequent owner. Within a make year, the price varies widely, especially for the newer models. For the cars that bore only one owner, the regression lines aim higher to point at a positive relationship between make year and price, showing better value retention. In the case of several owners, this difference becomes less marked due to different depreciation rates and buyers' opinions regarding the history of the vehicle. It shows the line-plot of an overall average price for vehicles as time increases; much stable from 2010 to 2014 and a visibly great increase thereafter.

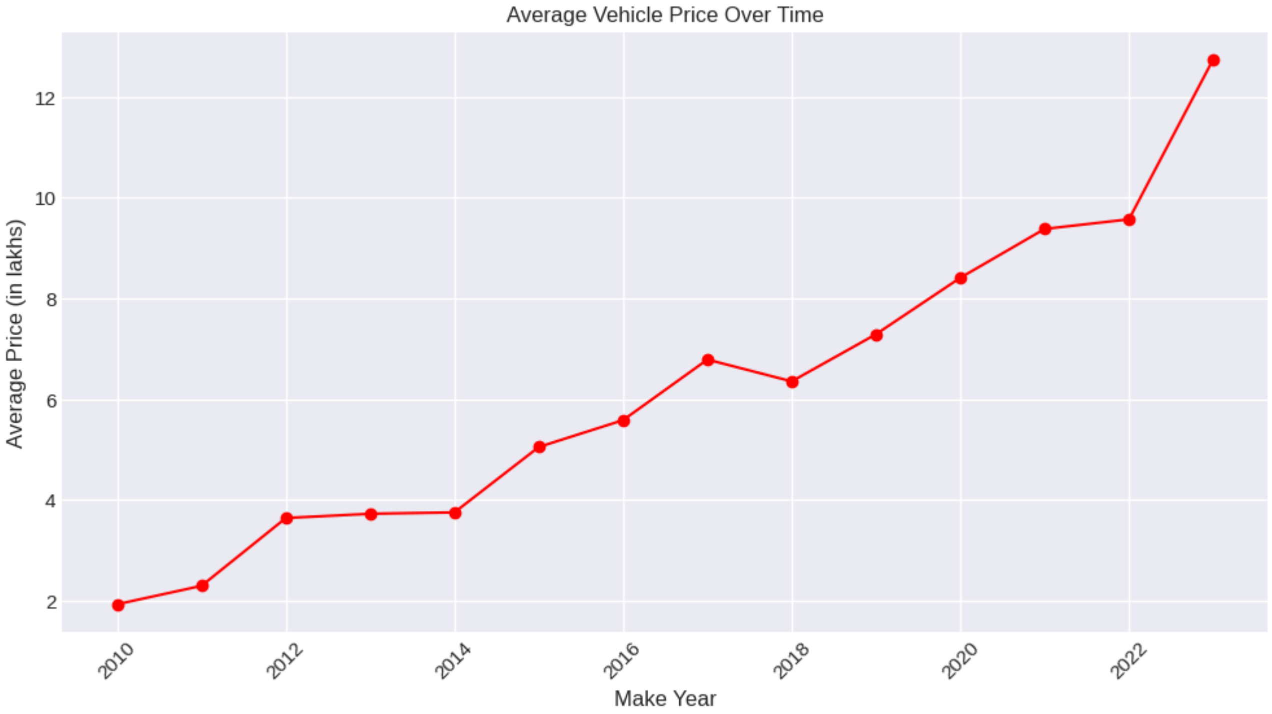


Figure 3.25: Average Vehicle Price Over Time

Observation(Fig 3.25) - The most significant increase is in the final year, a suggestion of potential market movements or changes in consumers' preferences. However, there's a plot with variance data every year, making it hard to know the price distribution. The line plot clearly shows that there is a significant difference between the average price of vehicles in the states and that some seem to increase linearly with time, while others seem to fluctuate.

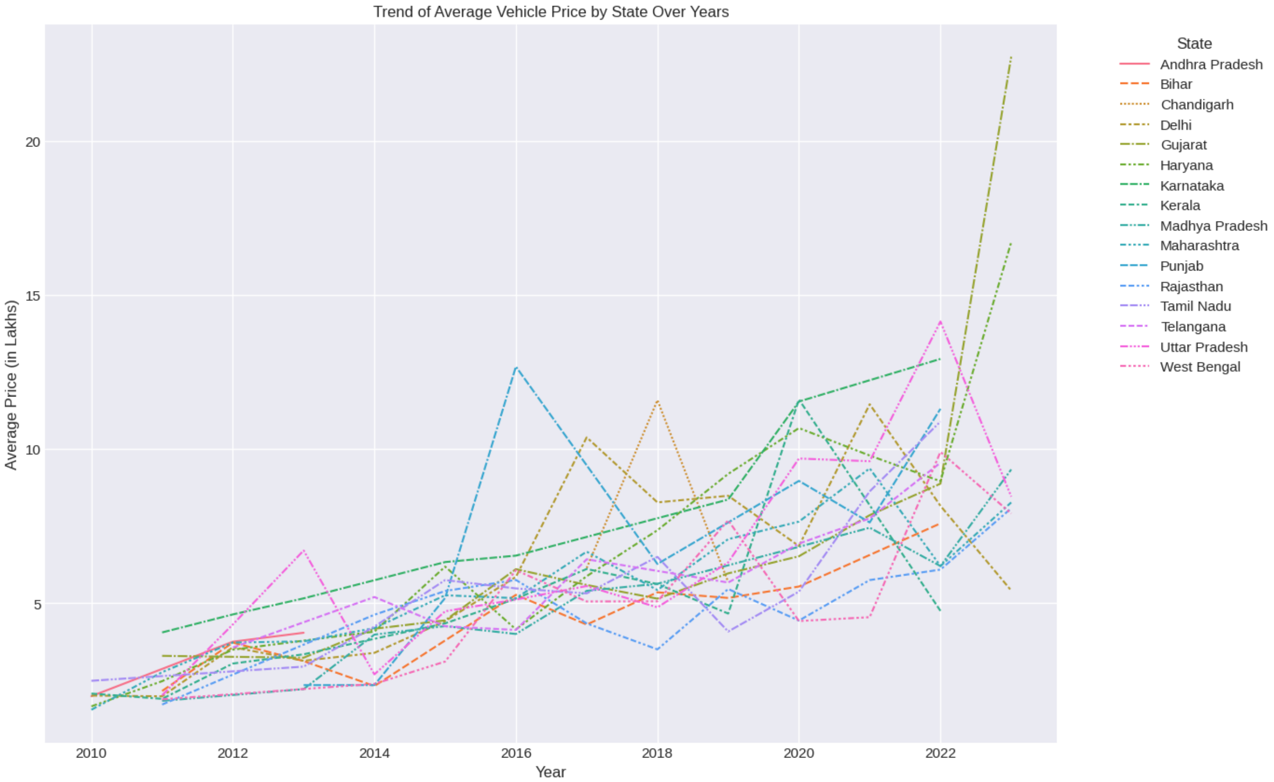


Figure 3.26: Trend Of Average Vehicle Price by State Over Years

Observation(Fig 3.26) : These spikes in recent years could result from such factors such as, shifting consumer demand, or reduced supply. This line crossover would suggest state-specific influence on pricing trends, such as regional economic conditions and taxation policies.

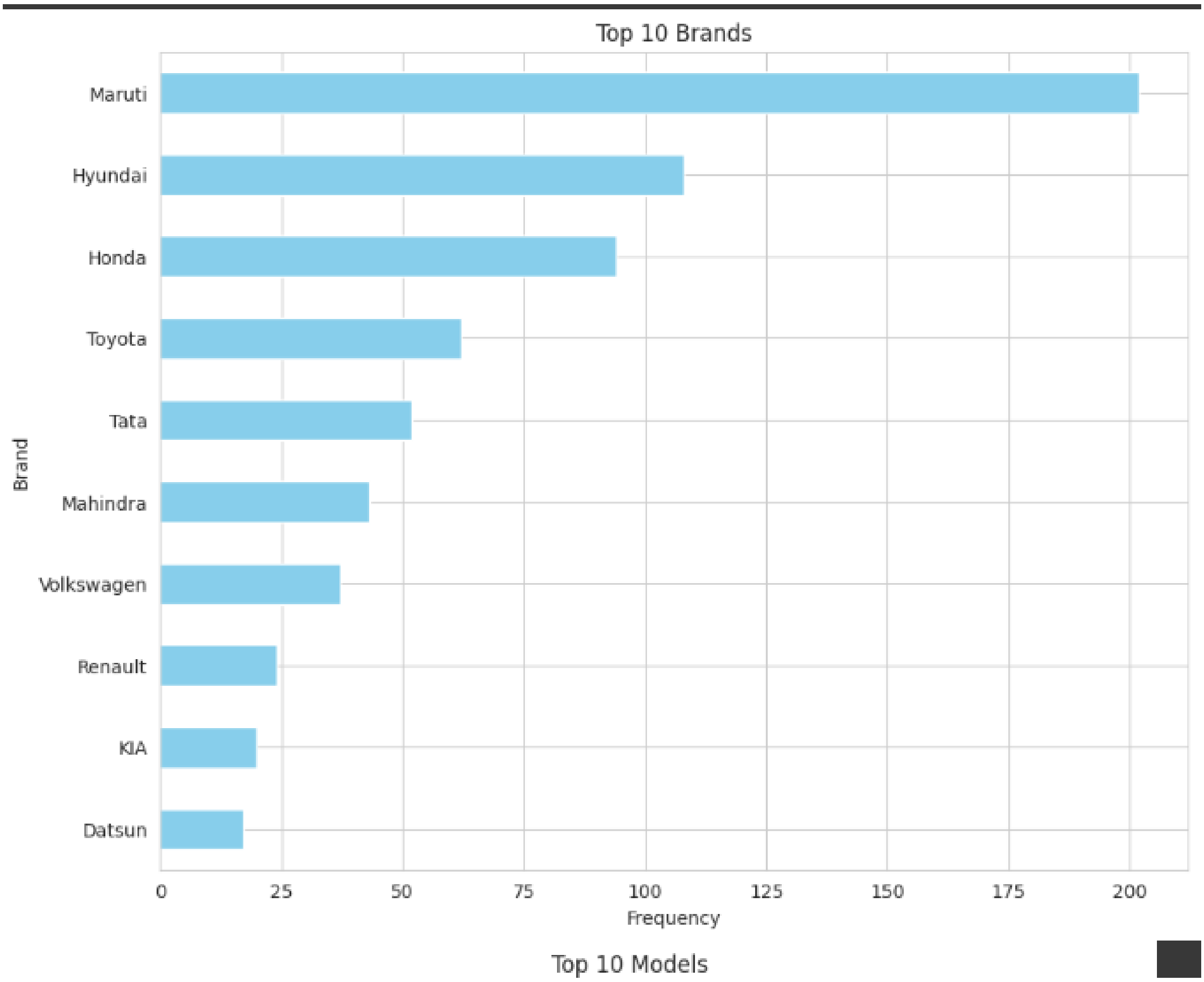


Figure 3.27: Top 10 selling brands

Observation (Fig 3.27) : The frequency of Maruti stands out significantly compared to other well-known brands like Hyundai, Honda, Toyota, and Tata, indicating Maruti's dominance in the Indian automobile market.

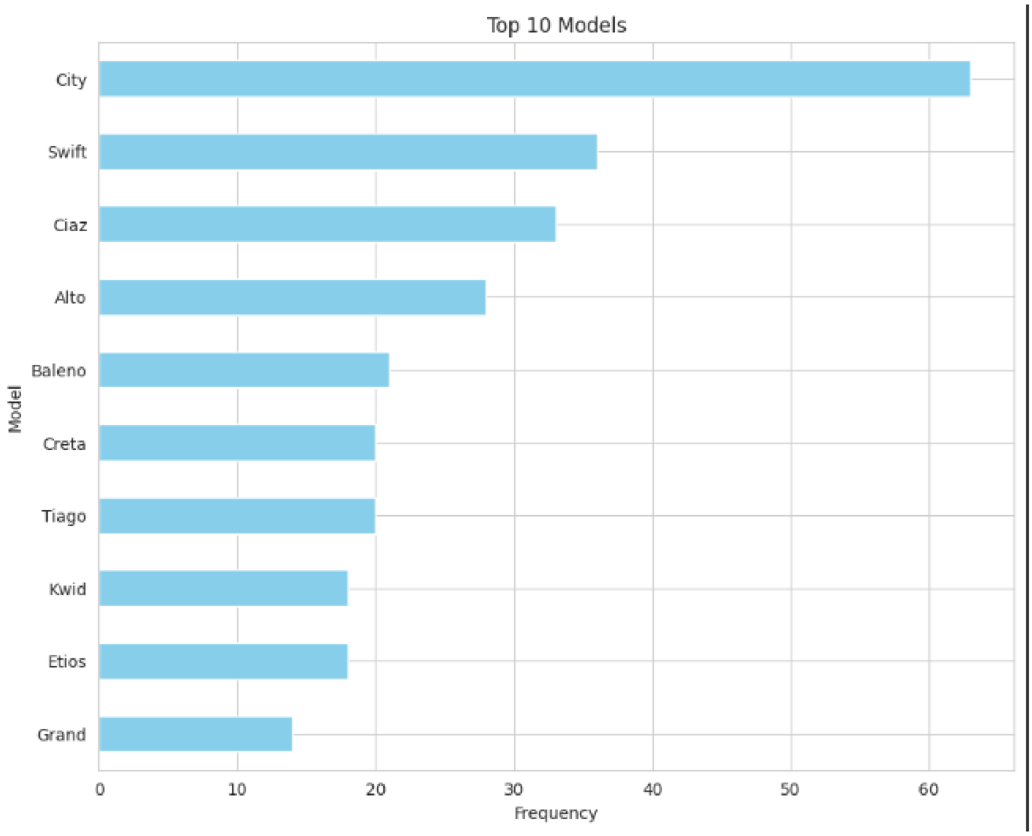


Figure 3.28: Top 10 selling models

Observation (Fig 3.28) : The frequency of Honda city stands out significantly compared to other well-known car models like swift, Honda city, ciaz, and alto by having more than 60 car of the same model which is then followed by Maruti swift.However we can observe that among the top 10 model most them belong to only maruti brand showing that maruti is more popular brand than honda despite having the best selling model of them all.

3.2.1 Inferences

- **Engine Capacity:**Inference: The histogram displays a right-skewed distribution, suggesting that consumers have a preference for smaller, more economical engine capacity. This desire is probably brought about by lower costs and improved mileage.
- **Mileage: Inference:** Based on the bimodal distribution, two different vehicle groups are suggested, with the major peak located approximately 17–18 km/l and the secondary peak located approximately 21–22 km/l. This is consistent with the distribution of engine capacity, since larger engines often have lesser mileage.
- **KM-Driven:** Deduction: Given that the distribution is right-skewed, the majority of cars have comparatively few miles on them. This implies that even though there are a lot of used cars on the market, people tend to favor cars with lesser mileage.
- **Price_in_lakhs**: The bulk of automobiles fall into the lower price bracket, with fewer cars overall, according to the right-skewed distribution of car pricing.
- **Reg month:**The month-by-month car registration bar chart lacks a discernible structure and displays erratic monthly fluctuations. Certain months may see higher registration numbers due to festivals or the introduction of new car models.
- **Insurance:** The majority of the automobiles in the sample had insurance, which could affect their buyer appeal and resale value. It's unclear how this affects other characteristics like the model year or registration month.
- **Transmission:**There is a noticeable difference between the number of cars with manual and automatic transmissions, with manual transmission cars being more common. This is demonstrated by the bar graph. This might be a reflection of the market or customer preferences.
- **Ownership:** The bulk of cars are listed as being owned by their original owner, indicating a market where newer or less traded autos are prevalent.
- **Fuel Type:** The most common type of vehicle is gasoline, followed by diesel, and the least common type is CNG. This is an indication of the market's preferences or availability for cars, maybe correlated with price distributions.

- **Brand:** The most prominently displayed brand with the highest count is a single brand, suggesting that it is popular or widely used in cars. Differences in ownership percentages, fuel kinds, and transmission types may all be related to variations in brand distribution.

Chapter 4. Feature Engineering

4.1 Feature extraction

Feature extraction is the process of creating new variables from existing data in order to improve the prediction capacity of models and extract deeper insights.
Here, we'll explore some key features extracted from a vehicle sales dataset:

1. **Reg Month**:This is derived from the registration year and may be another relevant feature. The registration month can have a say in things like price or demand, for instance.

2. Convert both **"Insurance"** and **"Spare Key"** from 'Yes'/'No' to boolean values. This way, we have a consistent format where more models, or rather more machine learning models, will recognize these features as binary (True/False) values that they should be viewed at and thus will make the processing and analysis a lot easier.

3. **Missing Value Imputation**: Filling the missing values in "Engine Capacity" and "Mileage" fields is very important. The methods of imputations to be used (mean and median imputation) are able to impose assumptions over the data but allow models to train over completed data.

4. **Age** : we have created a new column named 'Age' which is made by subtracted the 'Make-Year of the car with current year. This can be used in predicting the price of the car, as the price of the car is heavily dependent on the year they made. generally the older they are more cheaper they get.

5. Also for the Km-Driven column the dataset consisted of distance travelled along with an abbreviation of 'KM' at the end each value so we extracted the numeric value value from it.

4.2 Feature selection

Feature selection is the process of reducing the number of input variables when developing a predictive model. It is desirable to reduce the complexity of the model to increase interpretability and reduce overfitting.

1. **Variance Threshold**: This technique was used to remove features with low variance, based on the premise that variables with lower variance do not contribute significant information. This method helps streamline the data input without sacrificing essential details.

2. **Mutual Information Scores** : This method ranked features based on their relationship with the target variable 'Price_in_Lakhs'. Features such as 'KM Driven', 'Engine Capacity', and 'Make Year' scored highly, suggesting they are strong predictors of vehicle price. High mutual information scores indicate that these variables share a significant amount of information with the price, making them valuable for the model.

3. **Pearson Correlation**: This technique identifies and eliminates features that are highly correlated with each other, a phenomenon known as multicollinearity. For example, if two features, like "Engine Capacity" and "Horsepower" (if it were included), exhibit a strong correlation, meaning one can be predicted from the other, having both might be redundant. Removing one of these highly correlated features reduces the model's feature space without losing significant information, thereby streamlining the analysis and improving model performance by addressing potential multi-collinearity issues.

   These methods collectively enhance the predictive model by focusing on the most informative and independent features, supporting more robust and interpretable outcomes.

4.3 Data Tranformation

Transformation techniques like logarithmic, square root, or Box-Cox transformations, etc., can be applied in order to bring the skewness to a minimum and make the data distribution more symmetric. We've tried to implement and check all transformation through our dataset and out of it, we choose the best possible transformation for the feature that best fits the Q-Q plots.

- **Logarithmic Transformation:**

- **Definition:** Logarithmic transformation is the taking of the log of the values in a given data set. The natural logarithm (base e) and the base 10 logarithm are typical logarithmic transformations.
- **Use Case:** Logarithmic transformation is useful when the data is highly skewed, especially when there are large differences between the magnitudes of values. It can help in stabilizing the variance and making the distribution more symmetrical.
- **Application:** It is often applied to data with exponential growth patterns or data with long tails.
- **In our use case of problem statement of Price prediction, we have implemented Log transformation on top feature of dataset using feature selection which gives us 'KM Driven', 'Mileage'.**

- **Box-Cox Transformation:**
  - Definition: The Box-Cox transformation is a family of power transformations that includes both logarithmic and square root transformations as special cases. It involves a parameter, lambda (λ), which determines the type of transformation applied.
  - Use Case: The Box-Cox transformation is a versatile technique that can handle different types of data distributions. By estimating the optimal value of lambda, it aims to achieve the best possible transformation to make the data distribution closer to normal.
  - Application: It is widely used in regression analysis and other statistical modeling techniques where the assumption of normality is important.
  - **In our data transformation process, we considered Engine Capacity as a feature transformation variable for distribution of dataset.**

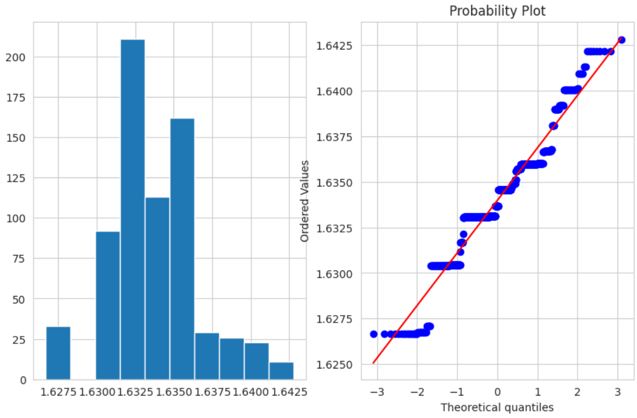


Figure 4.1: Feature Transformation for Engine Capacity

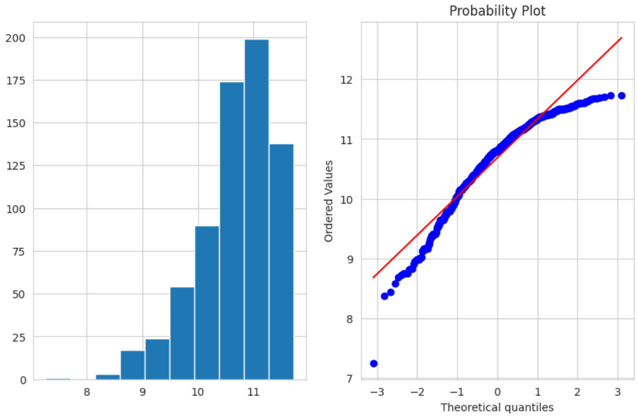


Figure 4.2: Feature Transformation for KM Driven

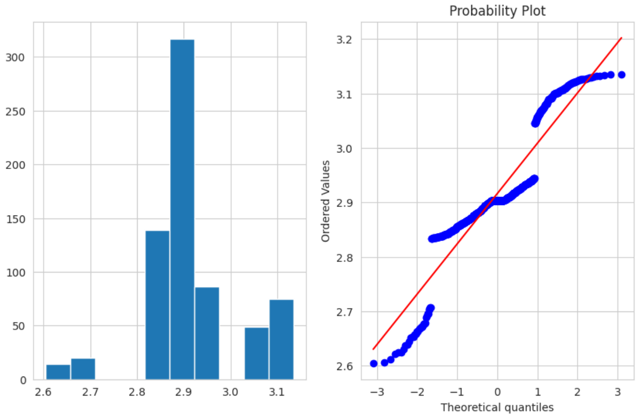


Figure 4.3: Feature Transformation for Mileage

Chapter 5. Model fitting

Model tuning is a fundamental step in the machine learning process, where algorithms are trained with data for predictions or classifying data points. Here, we will focus on the development of prediction models for two of the main issues in the Vehicle sale dataset: the problem of car price prediction and the problem of fuel economy. Each problem requires certain target variables, features, and modeling methods respectively. .

5.1 Model Fitting and Prediction Scores for Price Prediction

5.1.1 Random Forest Regressor

Model Description:

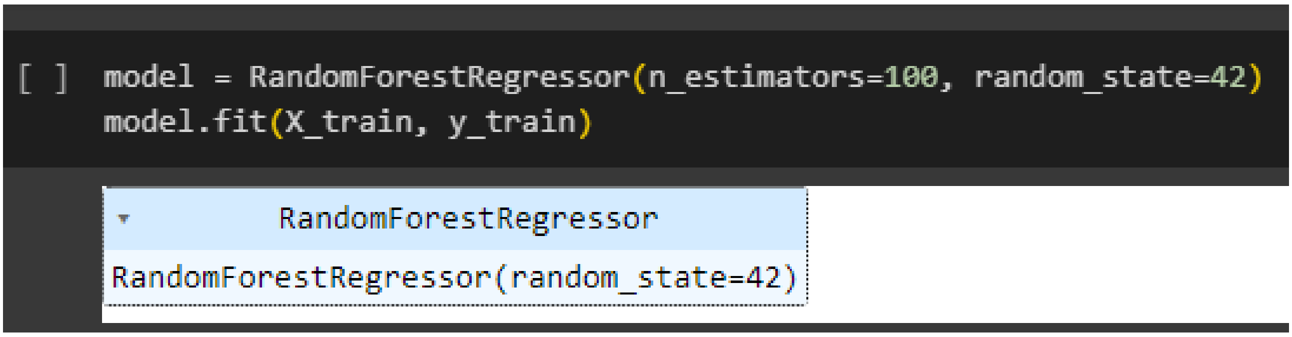
RandomForest Regressor is an ensemble learning method (which is considered for both classifcition and regresson task) that operates by constructing a multiple decision trees at training instances and outputting the average prediction of the individual trees. This method helps improve predictive accuracy and control over-fitting.

Encoding Techniques

we have taken Categorical variables such as Transmission and Fuel Type and were encoded using one-hot encoding to convert them into a format suitable for model input like 'Transmission_Automatic','Transmission_Manual', and 'Fuel Type_CNG', 'Fuel Type_Diesel', 'Fuel Type_Petrol' which is necessary for handling categorical data for better model Traning.

Model Parameters

1. **n_estimators**: Set to 100, which means 100 trees were used to build the forest. A higher number of trees generally improves the performance but also makes the computation slower.



```
[ ]  model = RandomForestRegressor(n_estimators=100, random_state=42)
     model.fit(X_train, y_train)
```

```
RandomForestRegressor
RandomForestRegressor(random_state=42)
```

Figure 5.1: model Fit

2. **max_depth:** Limiting the depth of each tree helps prevent overfitting. Typically set based on the complexity of the data and the relationship between features.
3. **random_state:** Ensures reproducibility of results by controlling the randomness of the forest's bootstrap aggregation.

Model fitting using feature selection

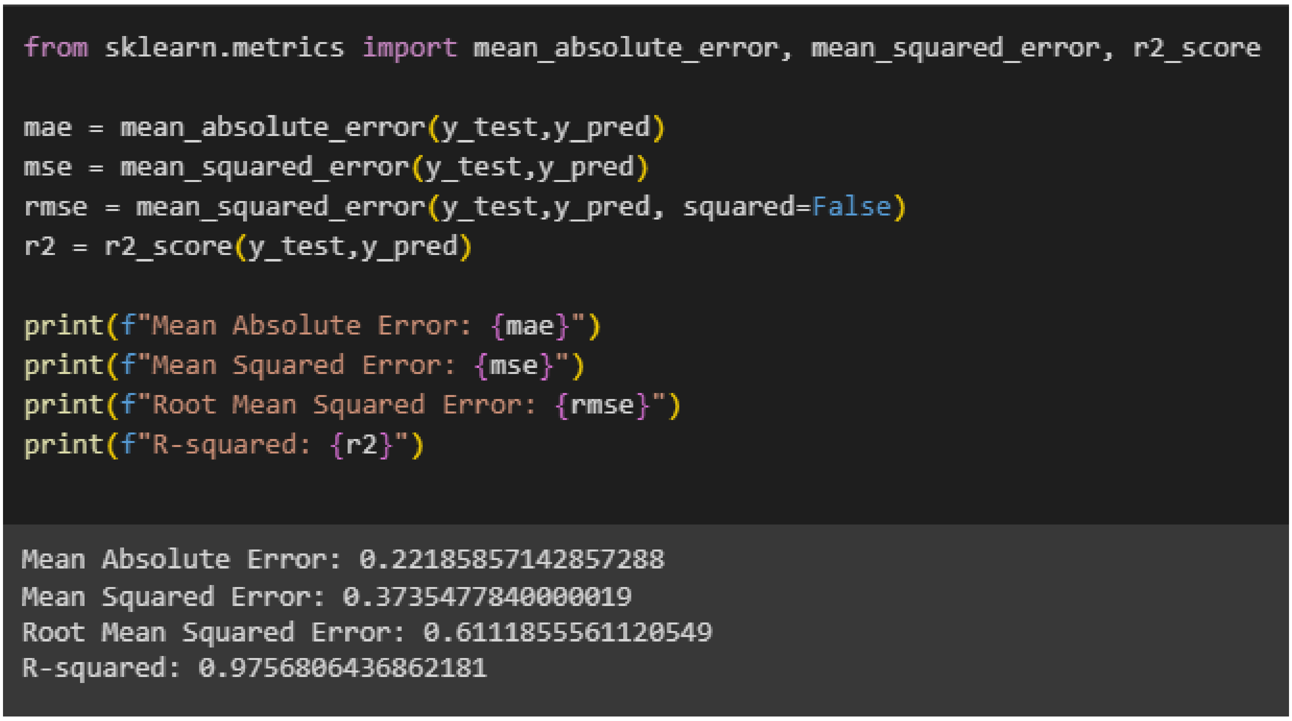
For the price prediction model,we took important features based on mutual information scores and correlation with the target variable (**'Price_in_Lakhs'** and Features like 'BoxCox_Engine Capacity', 'Log_KM Driven', 'Log_Mileage', 'Insurance','Spare key', 'Ownership', 'Make Year','Transmission_Automatic','Transmission_Manual' were included because they directly influence a vehicle's market price.

Model fitting without feature selection

In this model fitting, we manually choosen feature variable based on practical intuition and assumption which involves 'Price_in_lakhs' as our target variable and other than these columns 'Brand', 'Price_in_lakhs','Model','Registered State' , we have taken rest column in dataset.

Model Training and Results

The model was trained using a split of 80% of the data for training and 20% for testing to evaluate its performance. This split helps in validating the model against unseen data, providing an estimate of its performance in real-world scenarios.



```
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score

mae = mean_absolute_error(y_test,y_pred)
mse = mean_squared_error(y_test,y_pred)
rmse = mean_squared_error(y_test,y_pred, squared=False)
r2 = r2_score(y_test,y_pred)

print(f"Mean Absolute Error: {mae}")
print(f"Mean Squared Error: {mse}")
print(f"Root Mean Squared Error: {rmse}")
print(f"R-squared: {r2}")
```

```
Mean Absolute Error: 0.22185857142857288
Mean Squared Error: 0.3735477840000019
Root Mean Squared Error: 0.6111855561120549
R-squared: 0.9756806436862181
```

Figure 5.2: model Fit

Model Prediction Scores and Significance

1. Mean Absolute Error is 0.221 Meaning, on average, the model goes wrong in its prediction of the actual sale price by about 0.22. This is a very small error, meaning the model has high accuracy in predicting vehicle prices.

2. The mean squared error is 0.373 which statistic reveals that there must be some predictions in which deviation from the actual value is way too huge, although this should not be that frequent, since MAE is relatively low.

3. Root mean Squared error came to be 0.6111855561120549 this value demonstrates a moderate average error per prediction. It highlights that while the model is generally accurate, errors in particular expectations (possibly exceptions values) can exist.

4. The R^2 value is very high, at approximately 0.976, which means that about 97.6 percent of the variance in the car prices is explained by the predictors included in the model

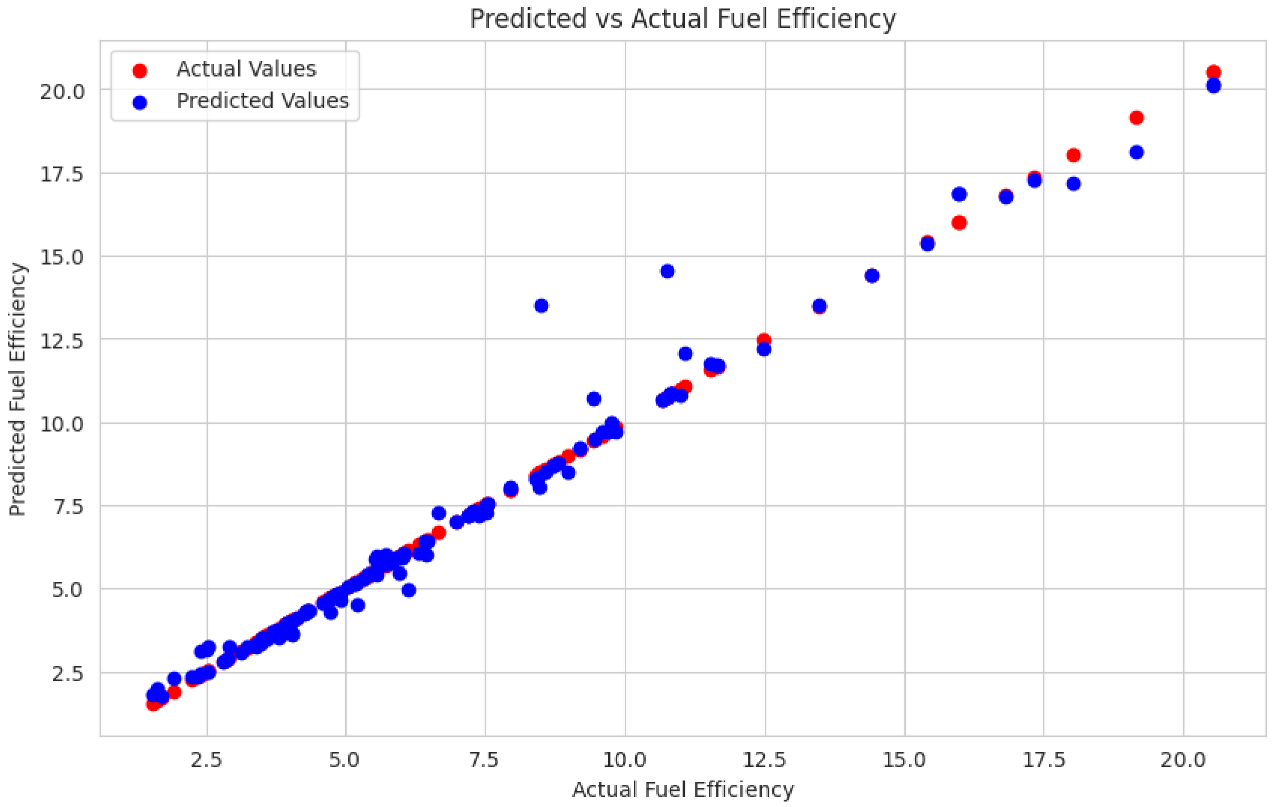


Figure 5.3: Regressor Fit using RandomForest

Model Evulation:

- Mean Absolute Error (MAE): Represents the average absolute difference between observed actual outcomes and predictions by the model. A lower MAE suggests a model with better accuracy in predicting continuous outcomes.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

- Root Mean Squared Error (RMSE): Measures the square root of the average of the squares of the errors. RMSE is sensitive to outliers and provides a higher weight to large errors, making it a robust metric against outlier influences.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- R-squared (R^2): Represents the proportion of variance in the dependent variable that is predictable from the independent variables. An R^2 score close to 1 indicates that the model explains a large portion of the variance in the target variable.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

5.1.2 Linear Regression

Model Overview

Linear Regression is a fundamental statistical method used in predictive modeling, particularly well-suited for understanding the relationship between multiple explanatory variables and a continuous dependent variable.

Features and Target Variable

- Target Variable: **Price_in_lakhs** – the price of the vehicle.
- Features: Included variables such as Engine Capacity, KM Driven, Make Year, and encoded categorical variables such as **Transmission_Automatic**, **Transmission_Manual** etc..

Model Fitting and Parameters

- Model Parameters:Linear Regression does not come with complex hyperparameter tuning, hence making it very plain and transparent for interpretation. The key parameter in the Ordinary Least Squares (OLS) model is the coefficient assigned to each feature, quantifying the impact of each feature on the target variable.
- Training Process: The fitting and training of the model were undertaken by fitting to a designated training dataset, involving the calculation of the best-fit line that minimizes the sum of the difference of squares between observed targets in the dataset and those predicted by the linear approximation.

Prediction and Evaluation

Evaluation Metrics: The model was evaluated using:

- **Mean Absolute Error (MAE)**
- **Root Mean Squared Error (RMSE)**
- **R-squared** (R^2)

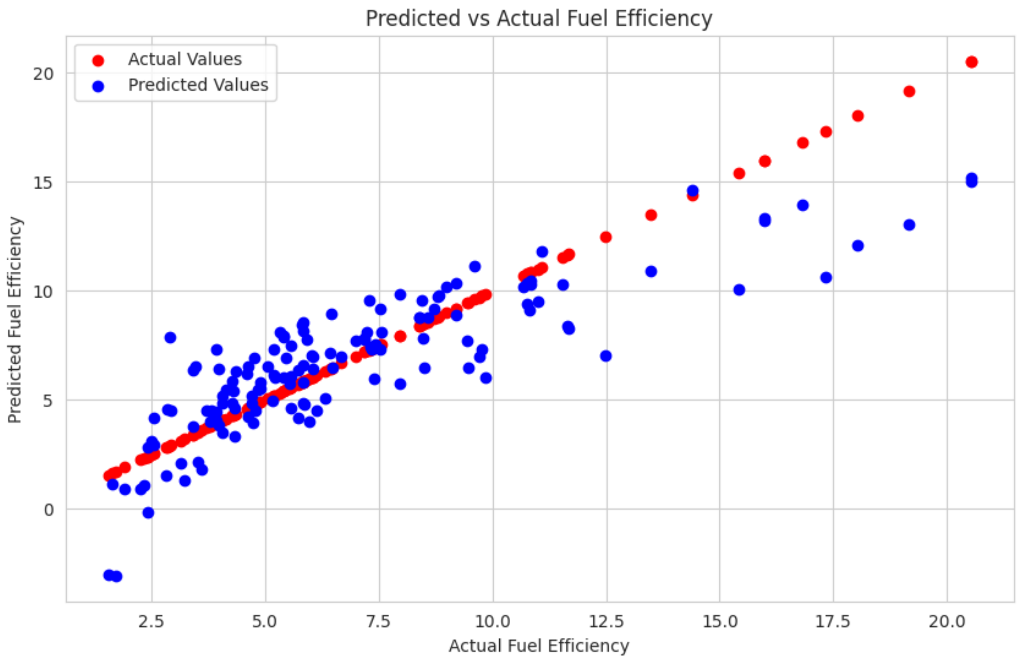
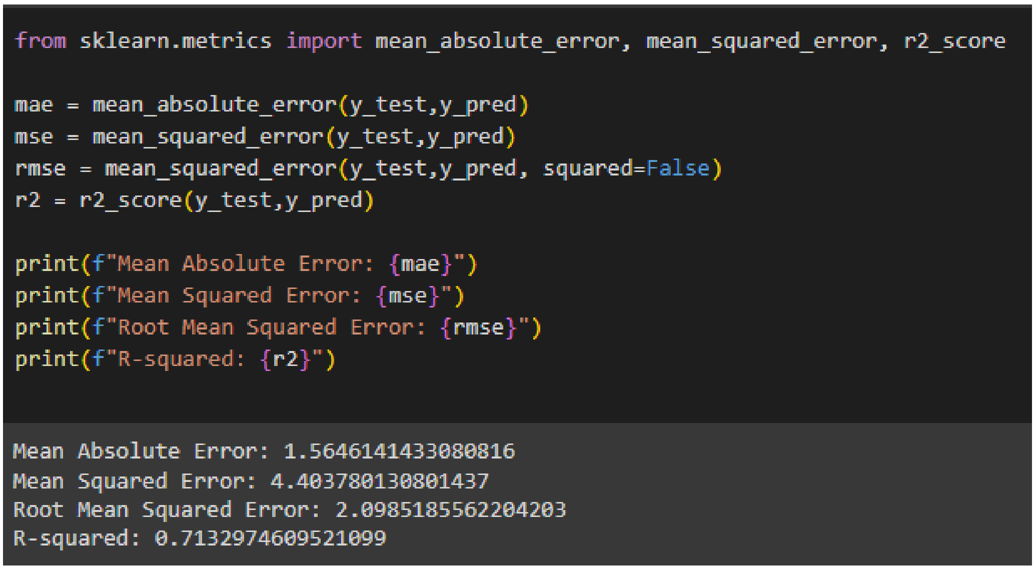


Figure 5.4: Linear regression without feature selection



```
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score

mae = mean_absolute_error(y_test,y_pred)
mse = mean_squared_error(y_test,y_pred)
rmse = mean_squared_error(y_test,y_pred, squared=False)
r2 = r2_score(y_test,y_pred)

print(f"Mean Absolute Error: {mae}")
print(f"Mean Squared Error: {mse}")
print(f"Root Mean Squared Error: {rmse}")
print(f"R-squared: {r2}")
```

```
Mean Absolute Error: 1.5646141433080816
Mean Squared Error: 4.403780130801437
Root Mean Squared Error: 2.0985185562204203
R-squared: 0.7132974609521099
```

Figure 5.5: Linear regression without feature selection

5.1.3 XGBoost Regressor

Model Overview

XGBoost is an advanced implementation of gradient boosting that uses decision trees and gradient boosting frameworks. It is highly efficient, flexible, and portable, often providing better performance than other types of models.

Features and Target Variable

Same as used in Linear Regression to ensure consistency and comparability between the models.

Model Fitting and Parameters

1. **n_estimators**: Number of gradient boosted trees. Higher numbers typically give better performance.
2. **max_depth**: The maximum depth of each tree. Limited depth helps to prevent overfitting.
3. **learning rate**: Determines the impact of each tree on the final outcome. Lower rates are typically used in conjunction with higher **n_estimators**.
4. **Training Process**:XGBoost builds upon the base gradient boosting framework but uses more accurate approximations for finding the best tree model. Like linear regression, XGBoost was trained on the same data set but with an iterative refined type of prediction based on the computed gradients.

Fitting and Evaluation

Fig(5.6) and fig(5.7) demonstrates the model Fitting and evaluation score of XGBoost regressor.

Evaluation Metrics: Utilized the same metrics as linear regression to evaluate performance. Due to the nature of XGBoost, expectations were for a higher accuracy and a better fit as measured by these metrics.

Comparative Analysis of Model Outputs

Interpretation of Scores:

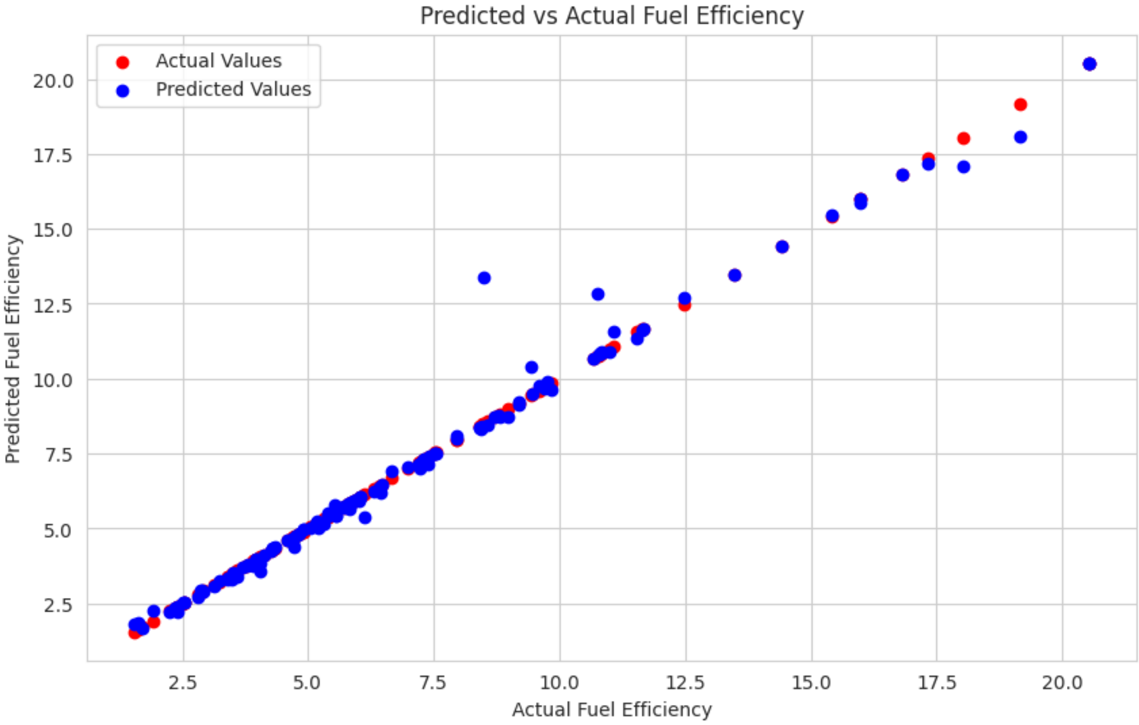
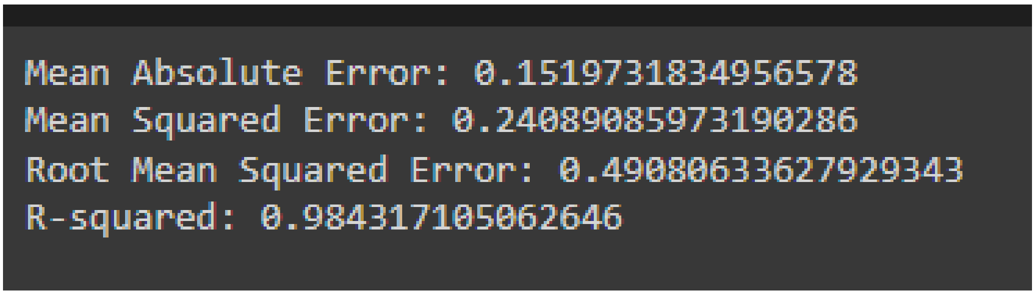


Figure 5.6: XGBoost regressor plot



```
Mean Absolute Error: 0.1519731834956578
Mean Squared Error: 0.24089085973190286
Root Mean Squared Error: 0.49080633627929343
R-squared: 0.984317105062646
```

Figure 5.7: XGBoost metrics

- **Linear Regression** showa lower R^2 and higher errors if the relationships between features and target are non-linear or if significant interactions between features are present, which linear regression cannot capture effectively.
- **XGBoost**, with its ability to handle non-linearities and complex interactions, is expected to show a higher R^2 and lower errors, indicating 1st best performance.
- **RandomForest Regressor**, with its ability to use ensumble technique multiple regression are trained with complex interactions, is expected to also show a higher R^2 and lower errors, indicating 2nd best performance.
- **Linear Regression with Feature selection** : also failed to show improved R^2 and leads higher errors , thought transformation, one reason might be inability of caputre relationships between features and target are non-linear or usage of inappropriate transformation thus linear regression with feature selection cannot capture effectively.

Significance and Implication

- Pricing Strategy Optimization :

  This ability to accurately forecast used car prices enables the dealer, or even an individual vendor, to hone their pricing strategies. Understanding the value drivers—make, model, year, mileage, and engine capacity—helps the sellers put up competitive yet profitable prices. This reduces the chance of underpricing or overpricing vehicles and therefore maximizes the income, but guarantees that there is a turnover in the stock.

- Consumer Trust and Market Transparency :

  Transparent and accurate price predictions contribute to the transparency of the used car market and give consumers confidence. This makes buyers know that the pricing given is data-backed, and buyers will be more confident that the price given to them is fair. This is vital in order to maintain a customer loyalty base while improving their buying experience.

5.2 Model Fitting and Prediction Scores for Fuel Efficiency Prediction

Problem Statement : Fuel efficiency, often measured as mileage (km per liter), is a crucial factor influencing vehicle purchasing decisions. Accurately predicting this attribute can help manufacturers and dealerships position their vehicles more effectively in the market. In this section, we explore the methodologies used to model and predict fuel efficiency using advanced machine learning techniques, focusing particularly on Linear Regression and XGBoost for their comparative analysis.

5.2.1 Linear Regression for Fuel Efficiency

Model Overview:

Linear Regression is used here to understand the linear relationships between multiple explanatory variables and the continuous dependent variable, fuel efficiency (Mileage).

Features and Target Variable

- Target Variable: Mileage
- Features: Includes Engine Capacity, Make Year, KM Driven, and categorical variables like Transmission and Fuel Type encoded using one-hot encoding.

Model Fitting and Parameters

- Model Parameters: No complex hyperparameters; the model focuses on coefficients which provide a direct understanding of how each feature impacts fuel efficiency.
- Training Process: Fit the model to the training data, ensuring the minimization of the residual sum of squares between observed targets and predictions.

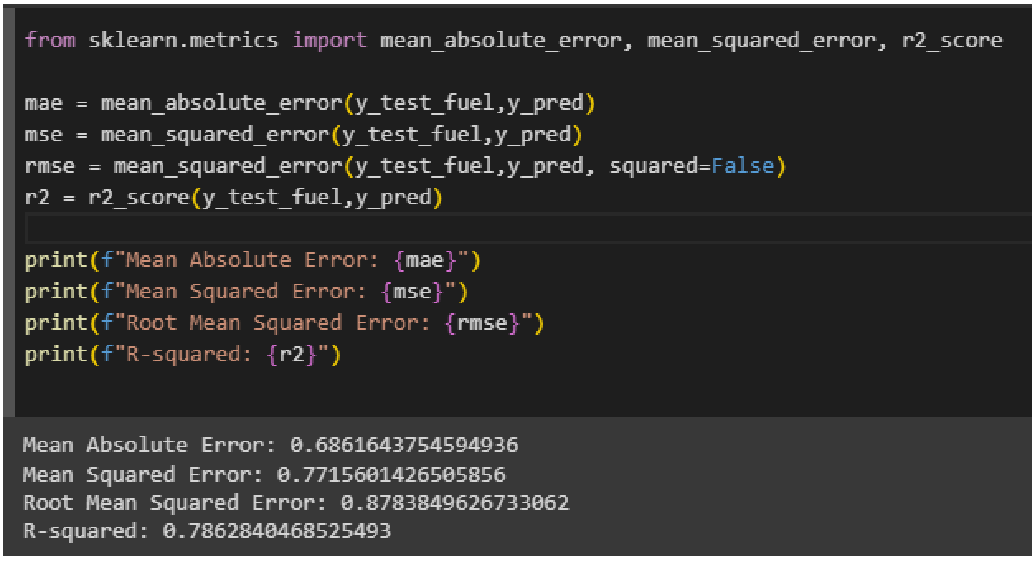
Evaluation Metrics: for fuel Prediction

Fig(5.8) shows the performance of model for Fuel Efficiency prediction.

5.2.2 XGBoost for Fuel Efficiency

Features and Target Variable

Same features as in Linear Regression to maintain consistency in comparison.



```
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score

mae = mean_absolute_error(y_test_fuel,y_pred)
mse = mean_squared_error(y_test_fuel,y_pred)
rmse = mean_squared_error(y_test_fuel,y_pred, squared=False)
r2 = r2_score(y_test_fuel,y_pred)

print(f"Mean Absolute Error: {mae}")
print(f"Mean Squared Error: {mse}")
print(f"Root Mean Squared Error: {rmse}")
print(f"R-squared: {r2}")
```

```
Mean Absolute Error: 0.6861643754594936
Mean Squared Error: 0.7715601426505856
Root Mean Squared Error: 0.8783849626733062
R-squared: 0.7862840468525493
```

Figure 5.8: Fuel effeciency prediction through Linear regression

Model Fitting and Parameters

- **Model Parameters:**
  - **n_estimators:** Adjusts based on the complexity of the data and overfitting risk.
  - **learning_rate:** Typically set low to allow more robust convergence.
  - **max_depth:** Controlled to prevent the model from becoming overly complex and overfitting.
- **Training Process:** Utilizes gradient boosting framework to build trees sequentially, where each new tree helps to correct errors made by previously trained trees.

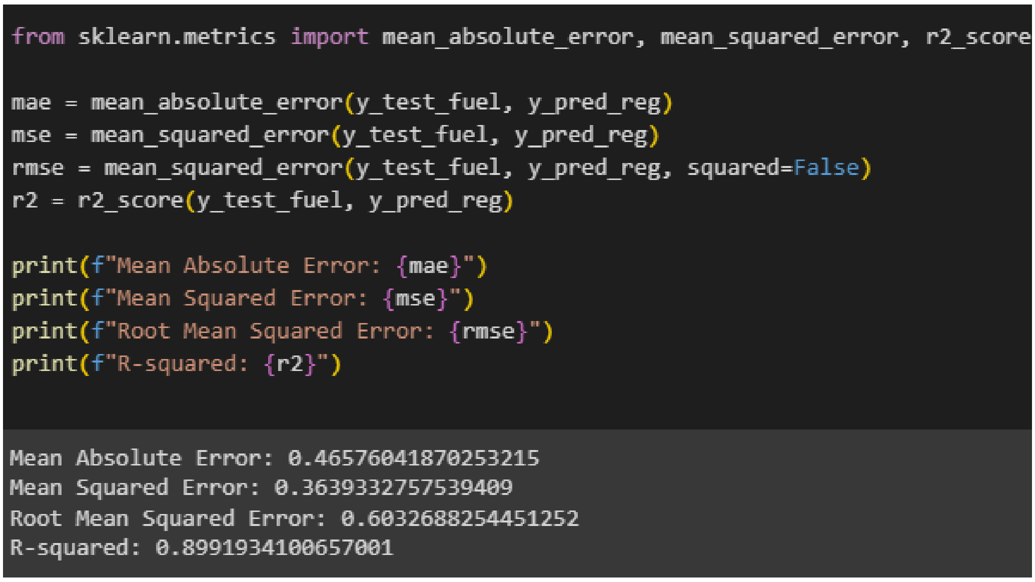
Prediction and Evaluation

Fig(5.9) demonstrate the performance of XGBoost regressor. Evaluation Metrics: Same metrics as used for Linear Regression are applied to measure the precision and effectiveness of the XGBoost model.

5.2.3 RandomForest Regressor

Overview

he RandomForest Regressor is an ensemble learning method that builds multiple decision trees during training and outputs the average expectation of the individual trees



```
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score

mae = mean_absolute_error(y_test_fuel, y_pred_reg)
mse = mean_squared_error(y_test_fuel, y_pred_reg)
rmse = mean_squared_error(y_test_fuel, y_pred_reg, squared=False)
r2 = r2_score(y_test_fuel, y_pred_reg)

print(f"Mean Absolute Error: {mae}")
print(f"Mean Squared Error: {mse}")
print(f"Root Mean Squared Error: {rmse}")
print(f"R-squared: {r2}")
```

```
Mean Absolute Error: 0.46576041870253215
Mean Squared Error: 0.3639332757539409
Root Mean Squared Error: 0.6032688254451252
R-squared: 0.8991934100657001
```

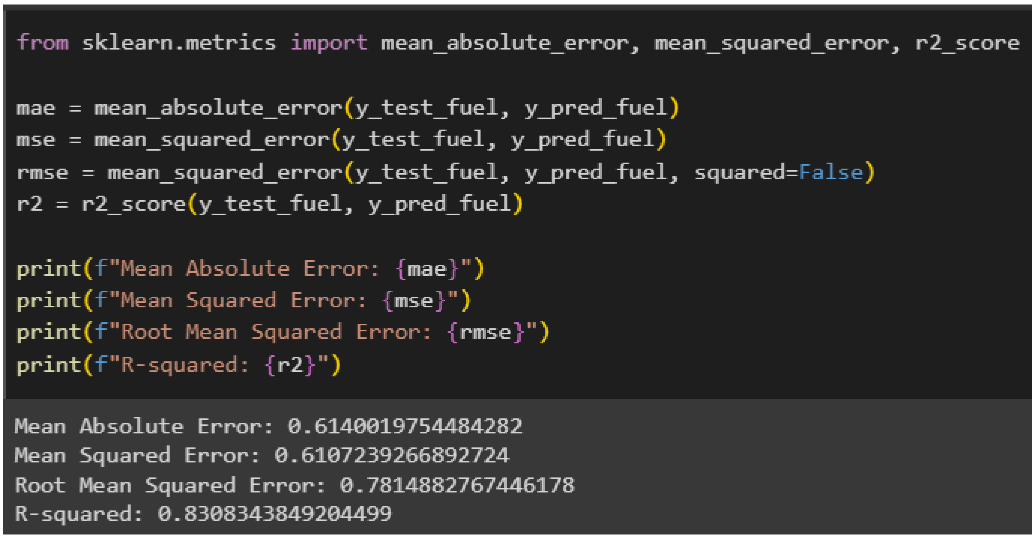
Figure 5.9: XGBoost model for Fuel Effeciency Prediction

to improve the predictive accuracy and control over-fitting. It is particularly effective for regression tasks due to its capacity to handle complex datasets with mixed types of features and its strength against overfitting.

Model Configuration and Training

The configuration of the RandomForest model for predicting fuel efficiency involves several key parameters:

- n_estimators=100: The model uses 100 trees, offering a good balance between computational efficiency and model performance. More trees can improve accuracy but at the cost of increased computational resources.
- max_depth=10: Restricting the depth of the trees to 10 helps prevent the model from learning overly complex patterns that do not generalize well, reducing the risk of overfitting.
- max_features='sqrt': This setting controls the number of features to consider when looking for the best split; using the square root of the number of features helps ensure that each tree in the forest is different and makes the model more robust.
- random_state=42: This ensures that the results are reproducible and consistent across different runs of the model.



```
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score

mae = mean_absolute_error(y_test_fuel, y_pred_fuel)
mse = mean_squared_error(y_test_fuel, y_pred_fuel)
rmse = mean_squared_error(y_test_fuel, y_pred_fuel, squared=False)
r2 = r2_score(y_test_fuel, y_pred_fuel)

print(f"Mean Absolute Error: {mae}")
print(f"Mean Squared Error: {mse}")
print(f"Root Mean Squared Error: {rmse}")
print(f"R-squared: {r2}")
```

```
Mean Absolute Error: 0.6140019754484282
Mean Squared Error: 0.6107239266892724
Root Mean Squared Error: 0.7814882767446178
R-squared: 0.8308343849204499
```

Figure 5.10: RandomForest Metrcis

Model Training:

The model is trained on a training dataset (X_train_fuel, y_train_fuel), where it learns to predict the fuel efficiency based on input features like engine specifications, make year, and other relevant attributes.

Prediction and Model Evaluation

After training, the model uses the test set (X_test_fuel) to predict fuel efficiency, creating predictions (y_pred_fuel) that can be compared against the actual values to evaluate the model's performance.

Key Evaluation Metrics:

- Mean Absolute Error.
- Mean Squared Error .
- R-squared.

Interpretation of Prediction Scores

- MAE and RMSE: These two statistics give measures of the average magnitude of error and standard deviation of the prediction errors, respectively. The metrics of MAE and RMSE help to know the amount of error that can be expected from the model's predictions and give a clear measure of model accuracy.

- R^2 : A high value of R^2 would, therefore, indicate more of the model's explanation of the variance in fuel efficiency, and that simply means effective learning and prediction capabilities. It informs the stakeholders on the degree to which the unseen data is likely going to be well predicted based on the model developed.

Comparative Analysis of Model Outputs

Interpretation of Scores:

- Linear Regression has low score than other model, possibilly non-linear relationships or interactions that are significant in determining fuel efficiency, which linear regression cannot capture.
- XGBoost is expected to perform better due to its ability to handle complex and non-linear interactions between variables, typically reflected in higher R^2 and lower MAE and RMSE scores.

Significance and Implications

- Operational Efficiency: It can be used for potential to correctly predict fuel efficiency will be helpful to automotive companies so that they can design and manufacture fuel-efficient motor vehicles and may even sell their vehicles with efficiency numbers..
- Consumer Information: Provide the consumer with accurate fuel efficiency estimates that will help him or her in buying a vehicle.
- Environmental Impact: Helps to access the impact on the environment, considering that it is based on the predicted fuel efficiency of a vehicle, hence contributing to the realization of a larger effort in conserving the environment.

Chapter 6. Conclusion & future scope

We successfully constructed predictive models addressing two important components of the automotive market: car pricing and fuel efficiency, thanks to our thorough research of the used vehicle dataset. Regression models such as XGBoost for mileage estimation and Linear Regression and Random Forest for pricing prediction showed great ability to interpret the intricate correlations between different vehicle attributes and the desired results. Our findings show that while engine specs and make year are critical for fuel efficiency, factors like brand, make year, and engine capacity have a substantial impact on the price and fuel efficiency of the used car costs. These models support sellers' pricing strategies as well as customers' decision-making by providing information on anticipated fuel economy. In the end, this analysis is an essential resource for those involved in the automobile industry.

6.1 Future plan

If we can have a more detailed dataset consisting more detail like the wear and tear of the car, any damage or accidents in the past, color etc., with more data in general would greatly enhance the accuracy of the prediction and analysis of the data.

Using more sophisticated learning algorithms and approaches we can find more patterns and interactions between variables that are currently unrecognized by current models, which will in turn help increase the accuracy of the prediction.

Group Contribution

Member 1

Aadesh Minz has contributed in collecting the data by web scraping and fitting and prediction of the data.

Member 2

Asish Joel has contributed in cleaning the data and generated all the visual plots and analysis of the data.

Member 3

Raj Kariya contributed in the section of Feature Engineering and feature selection.

Short Bio

1. **Asish Joel Batha** is a third year undergraduate student pursuing BTech in Maths And Computing from Dhirubhai Ambani Institute of Information and Communication Technology. He is a passionate student with skills in development and data science. Outside of work, Asish enjoys , playing badminton and cricket,and listening to some insightful podcasts.
2. **Raj Kariya** is a third year undergraduate student pursuing BTech in Maths And Computing from Dhirubhai Ambani Institute of Information and CommunicationTechnology. Raj is a passionate student with Machine-Learning and Data Science expertise. He enjoys working with friends and finding solution to the problems.Outside of work, Raj enjoys playing badminton, and watching movies.

3. **Aadesh Minz** is a third-year undergraduate student at Dhirubhai Ambani Institute of Information and Communication Technology, pursuing BTech in Math and Computing. His native Village is Chhattisgarh(Bilaspur) and has done his schooling at Space Central School, Sriharikota .He is an enthusiast in the field of Machine Learning and Data Science; hence, he has been able to do a lot of projects together. Aadesh is equally good as a leader, with great communication skills and the ability to crack hard problems with effective solutions while working in a team.

He is working on experimenting with new technologies and programming languages, like Python, JavaScript, and Node.js. He is a good reader, player, and movie watcher when he is free. He is an adept problem solver in software development. 32

References

[1] Cars24 Services Pvt. Ltd., *Vehicle Sales Data*, obtained from Cars24 website, https://www.cars24.com

[2] Group-4 Exploratory data Analysis of Vehicle sales https://colab.research.google.com/drive/1TQ1zcIkL2vH5Z--69-D_GJya4RoPlaa_?usp=sharing

[3] Group- 4 Web- Scrapping from Cars24 https://colab.research.google.com/drive/1qE87qLJ9hngMW0Zc0aWpb89-jyy0hOYX?usp=sharing