Z

A USED CAR PREDICTION PROJECT

Overview

The used car market in Kenya has undergone significant changes since the country's independence. In the early post-independence years of the 1960s through the 1980s, Kenya's automotive landscape was dominated by new vehicles; however, in the 1990s the market had a turning point as economic liberalization policies relaxed car importation restrictions, opening the door to an increasing number of used car imports, primarily from Japan. This trend accelerated into the 2000s, with used cars becoming increasingly popular due to their affordability compared to new vehicles. Due to these relaxed policies used cars have become quite popular and affordable making up the majority of vehicle sales in Kenya, with about 80% of all vehicles sold being pre-owned these is due to increase in their demand among may car buyers due to their affordability The market used is made up of foreign or locally used vehicle

Navigating this market can be challenging, as both buyers and sellers often face the difficulty of determining fair prices due to various influencing factors such as age, mileage, and condition of the vehicles. Accurate price estimation is crucial for making informed decisions in this dynamic market. These project aims to offer car buyers and sellers an efficient, reliable and accurate solution in pricing of different vehicle models based on their various features by leveraging data driven insights analysis and machine learning tools and techniques

z

This project focuses on creating a machine learning model tailored specifically to the Kenyan market, aiming to predict the value of used cars based on various features.

Problem Statement

The Kenyan used car market lacks a reliable, data-driven mechanism for accurately predicting vehicle prices. This absence leads to inefficiencies and uncertainties for buyers, sellers, and other stakeholders in the automotive sector. Currently, pricing is often based on subjective assessments or incomplete information, resulting in potential overpricing or undervaluing of vehicles hence this situation often leads to prolonged negotiation processes, unfair deals, and a lack of trust in the market.

OBJECTIVES

Main Objective

1. To develop a machine learning model that predicts prices of used cars accurately in Kenya based on the vehicle various features

Specific Objectives

2. To determine which vehicle features that have the greatest impact on the price of the vehicle

3. To determine which car brands is popular in the Kenya used car market

4. To determine vehicle features are the most common in our dataset

z

SUCCESS CRITERIA

An R squared of 0.8 which shows the model has predicted 80% of the variance on the target variable

A Low mean absolute error suggesting the model has the small average absolute difference between predicted and actual values.

CONSTRAINTS

Lack of enough Kenya data to adequately train our models

z

DATA UNDERSTANDING

The dataset for this project was sourced from several Kenya car bazaars and show rooms in Nairobi County and its environments. This dataset contains car listings up to 2021, capturing key information such as the car's Vehicle name, year of manufacture, mileage, fuel type, transmission, and the listed price. This data forms the foundation for building the predictive model and offers valuable insights into the Kenyan used car market.

Dataset Overview:

With a substantial size of 6,019 entries, this dataset offers valuable insights for our analysis. It encompasses a range of features pertinent to the Kenyan automotive market. Below is a description of the dataset's columns:

- Total Rows: 6,019

- Total Columns: 11

COLUMNS	**DESCRIPTION**	**DATA TYPE**	**VARIABLE TYPE**
CAR PRICE	Unique identifier for each vehicle listing.	FLOAT	TARGET
NAME	Model name of the vehicle.	STRING	PREDICTOR
YEAR	Year the vehicle was manufactured.	INTEGER	PREDICTOR

z

KILOMETERS DRIVEN	Total distance covered by the vehicle in kilometers	INTEGER	PREDICTOR
FUEL TYPE	Type of fuel used by the vehicle (e.g., petrol, diesel).	STRING	PREDICTOR
TRANSMISSION	Transmission type of the vehicle (e.g., manual, automatic).	STRING	PREDICTOR
USE	where the vehicle was previously used (e.g. foreign, local)	STRING	PREDICTOR
ENGINE	Engine capacity or specification of the vehicle	INTEGER	PREDICTOR
POWER	Power output of the vehicle's engine, typically in brake horsepower	INTEGER	PREDICTOR
SEATS	Number of seats available in the vehicle.	INTEGER	PREDICTOR

z

DATA PREPARATION

1: Data Cleaning

At these stage we will clean our data using the following steps

Completeness

The column engine and power have missing values. To assess the extent of the missing data, we calculate the percentage of null values and found engine at 0.6% and power at 2.38%which were not significantly large and could not afford to lose any data due to our limited dataset we filled the null values with median to represents the middle value of the dataset and it is less affected by outliers compared to the mean. This ensures that the imputation reflects a central, robust estimate of the typical value

Consistency

No duplicates values

Uniformity

First , we strip the 'cc' in engine and 'bhp' in power ,then change their data types from object to floats and seats from float to integer to reflects the real-world scenario accurately by first filling its 42 null values with the mode because most the car seats range between 4 to 7 seats hence representing the real world phenomena

Let’s also change the columns naming of name to model, kilometers driven to mileage for easy understandability and also standardize the text in the columns with strings for uniformity

Z

Validity

We dropped the 'No' column it does not have any relevance in our analysis

We examined our data for outliers using interquartile method

The dataset contains outliers such as Mileage (39,000), and Power (60.03), among others. These statistics describe specific automotive attributes such as the year of manufacturing, mileage, engine capacity, power, number of seats, and price. While these outliers might occasionally mislead analysis, we are not removing them in this case because each value represents a genuine and realistic phenomena in the real-world.

2: Feature engineering

By creating new columns like age and brand we are able to perform deeper analysis to gain more insights, the age of the vehicle was calculated by subtracting the vehicle manufacture year by our current year 2024 to get the latest age as per now

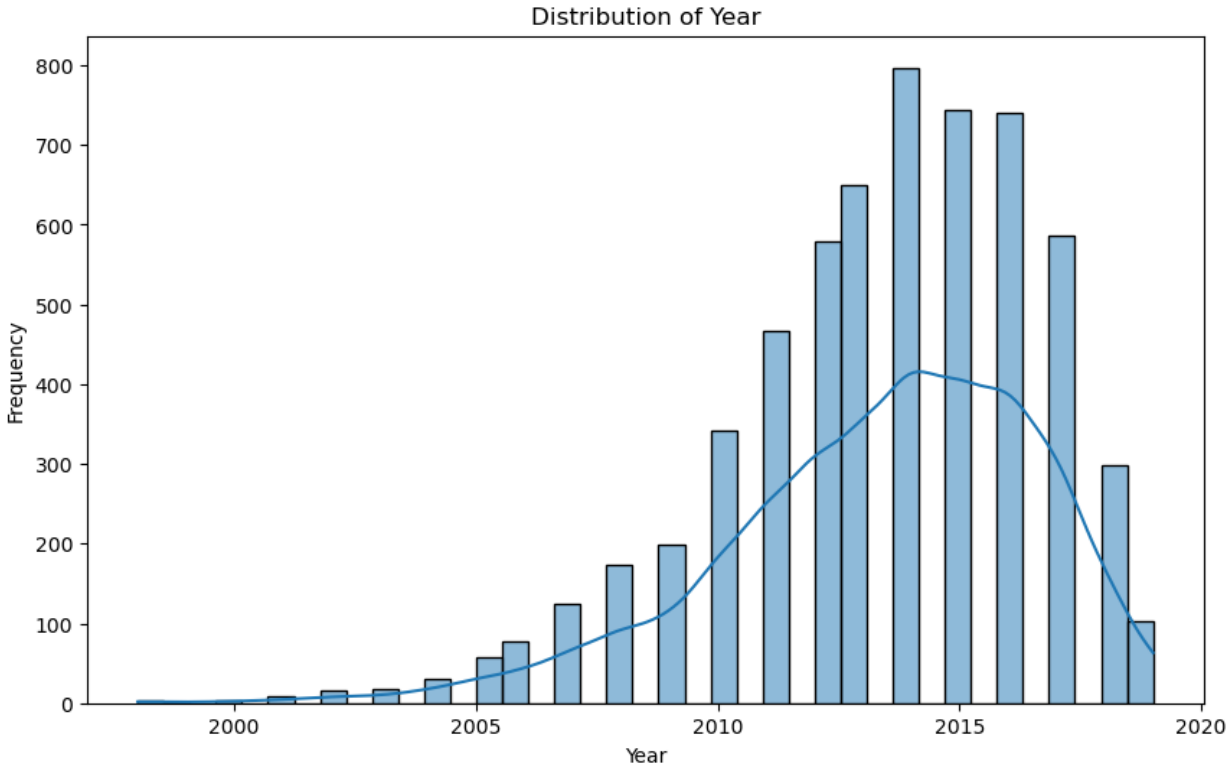
Z

3: EDA (Exploratory Data Analysis)

We will analysis our data in the following steps

a) **Univalent analysis**

Distribution of year



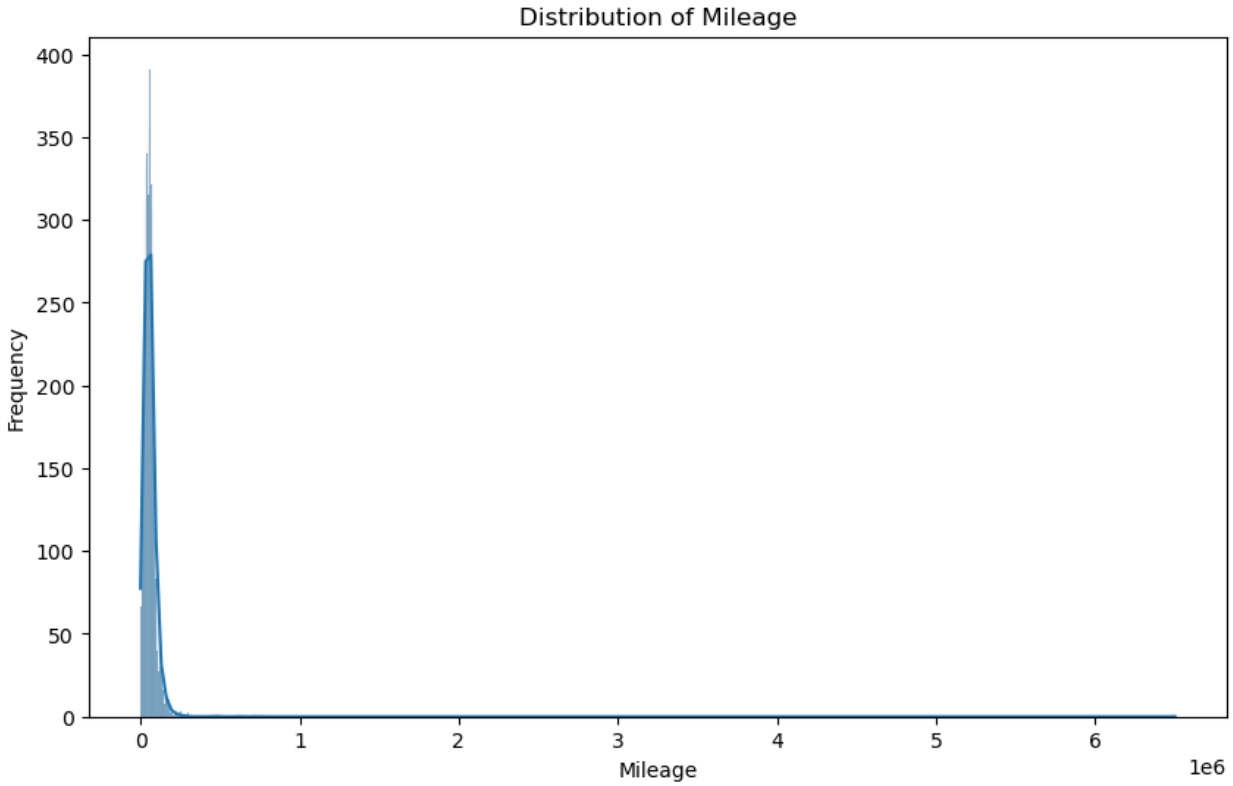
Observations

Left skewedness implying the distribution of most vehicles were manufactured around 2010, indicating a higher prevalence of cars that are about 10 years old in the listings.

This trend suggests that cars from this period are more commonly available in the second-hand market, which can influence pricing strategies.

Z

Distribution of mileage



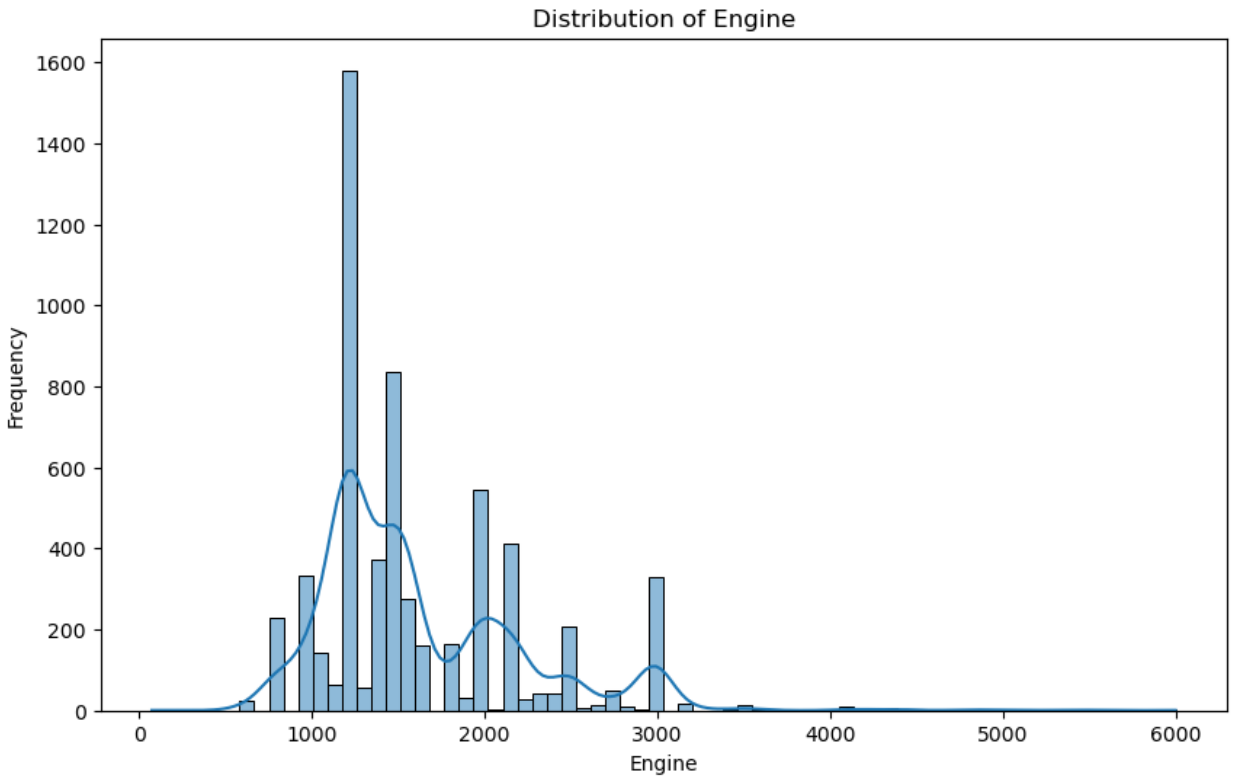
Observations

Right skewedness mean that majority of vehicles have lower kilometers driven, with a sharp decline in frequency as the kilometers increase.

Vehicles with fewer kilometers are more prevalent and likely more desirable due to less wear and tear, affecting their market value.

Z

Distribution of engine size



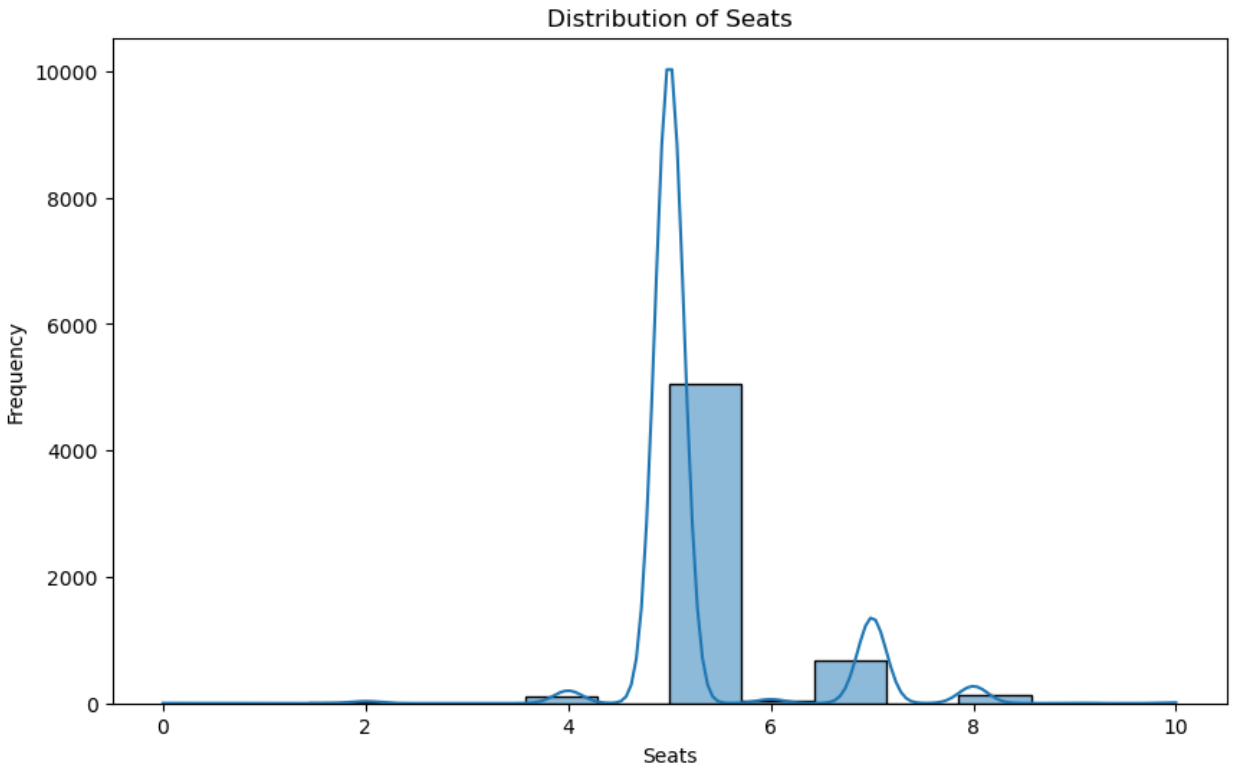
Observations

Right-skewed suggests that smaller engine sizes are more common in the dataset. The long tail to the right indicates that there are relatively few instances with very large engine sizes compared to smaller ones.

Mid-sized engines are the most common, suggesting a balance of performance and fuel efficiency that appeals to buyers.

Z

Distribution of Seats



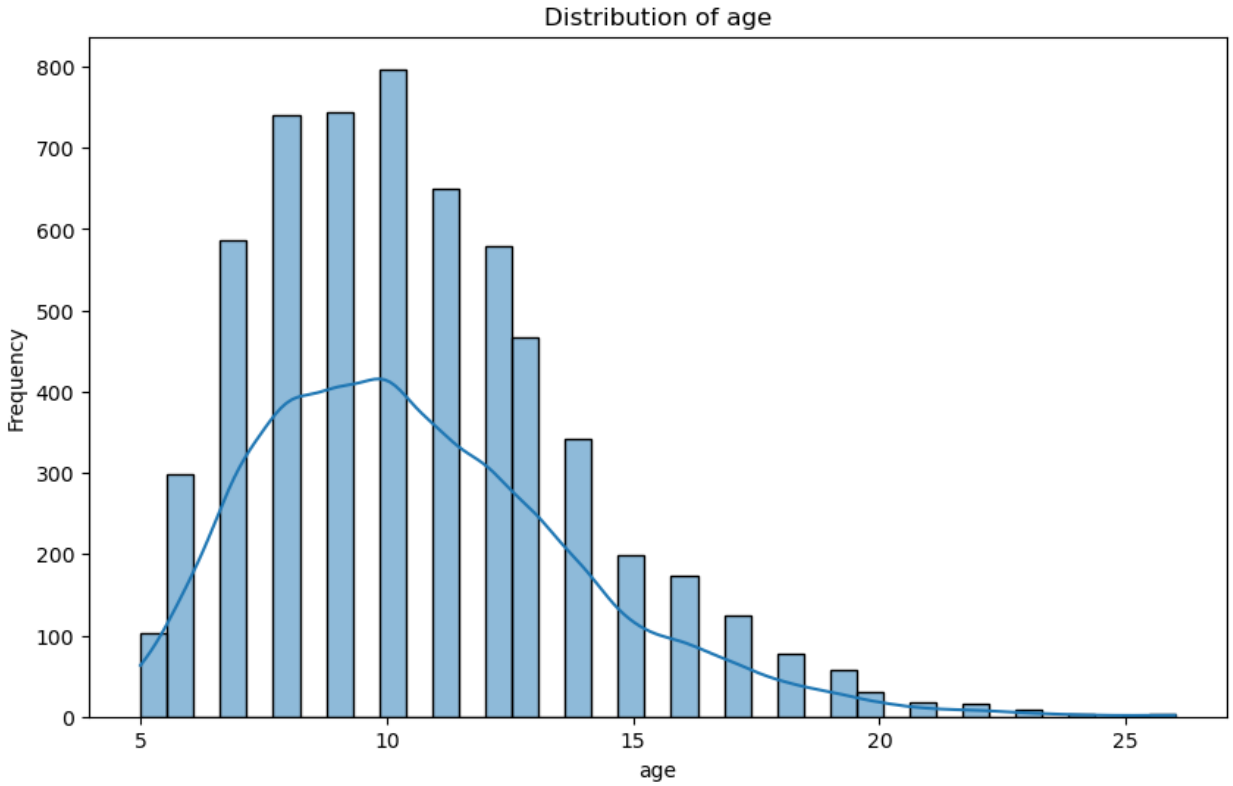
Observations

The most common number of seats is at 4 seats, indicating that most vehicles are designed to accommodate four passengers.

This reflects market preferences for family oriented cars, which can guide inventory and pricing decisions.

Z

Distribution of Age



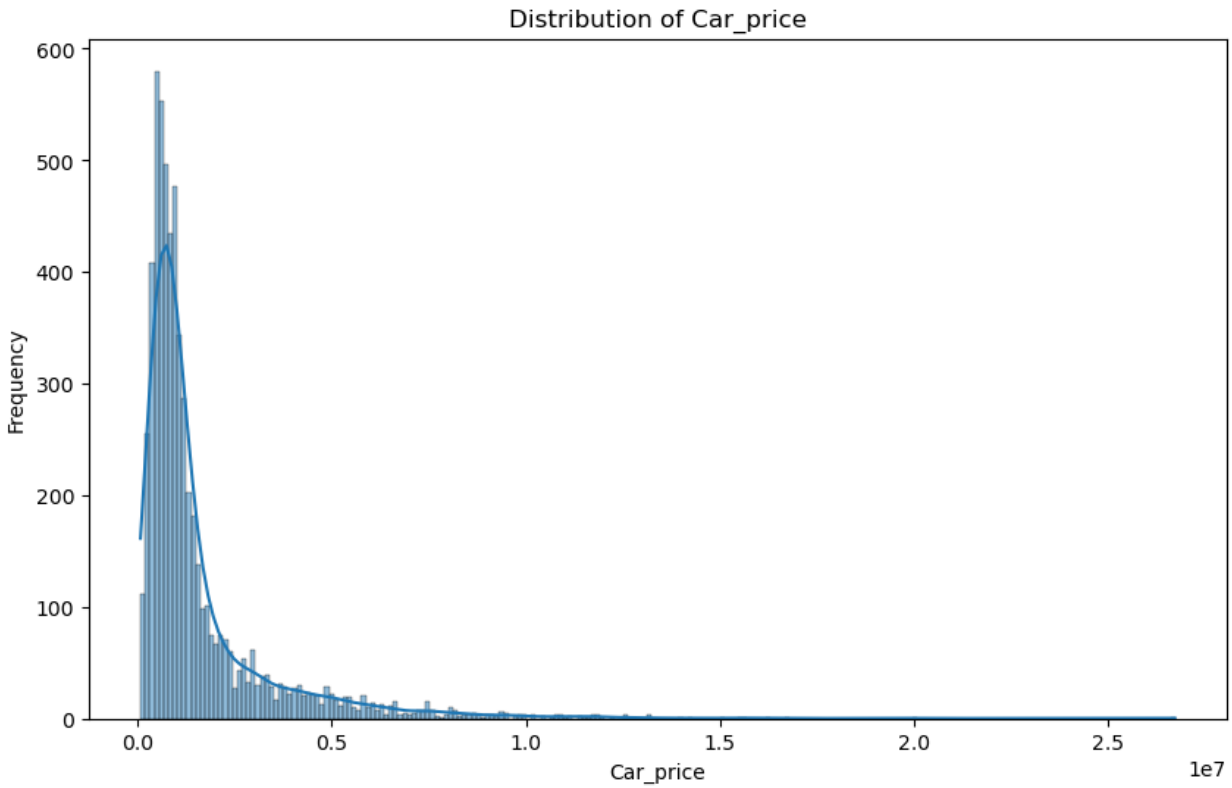
Observations

Right skewed shows newer vehicles 5yrs old, with a decline as the age increases.

Newer vehicles are likely command higher prices due to better their condition and lower mileage.

Z

Distribution of Car price



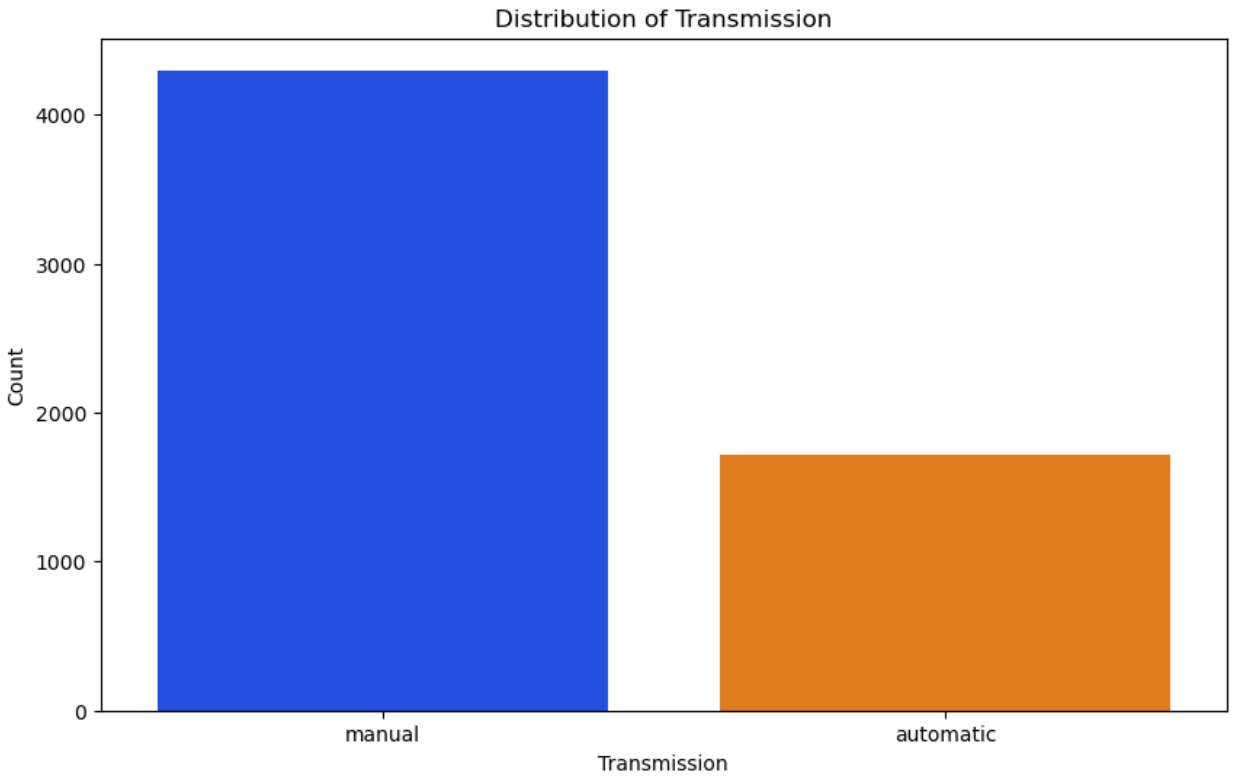
Observations

Right skewed showing that most used cars fall into the lower price range

Higher priced used cars are less common, possibly due to affordability constraints.

Z

Distribution of transmission



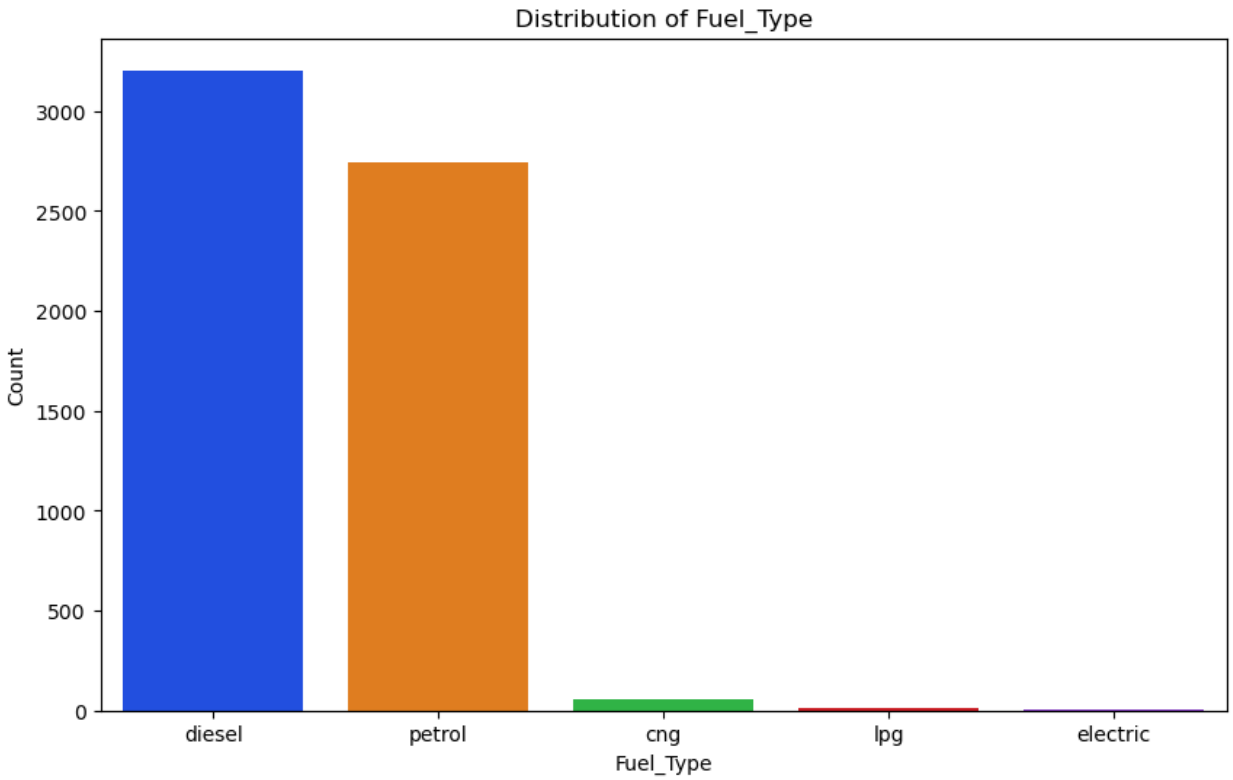
Observations

The manual category has a significantly higher count (taller bar) compared to the automatic category

The preference for manual transmissions may vary based on factors like driving habits, fuel efficiency, and personal preferences

Z

Distribution of Fuel type



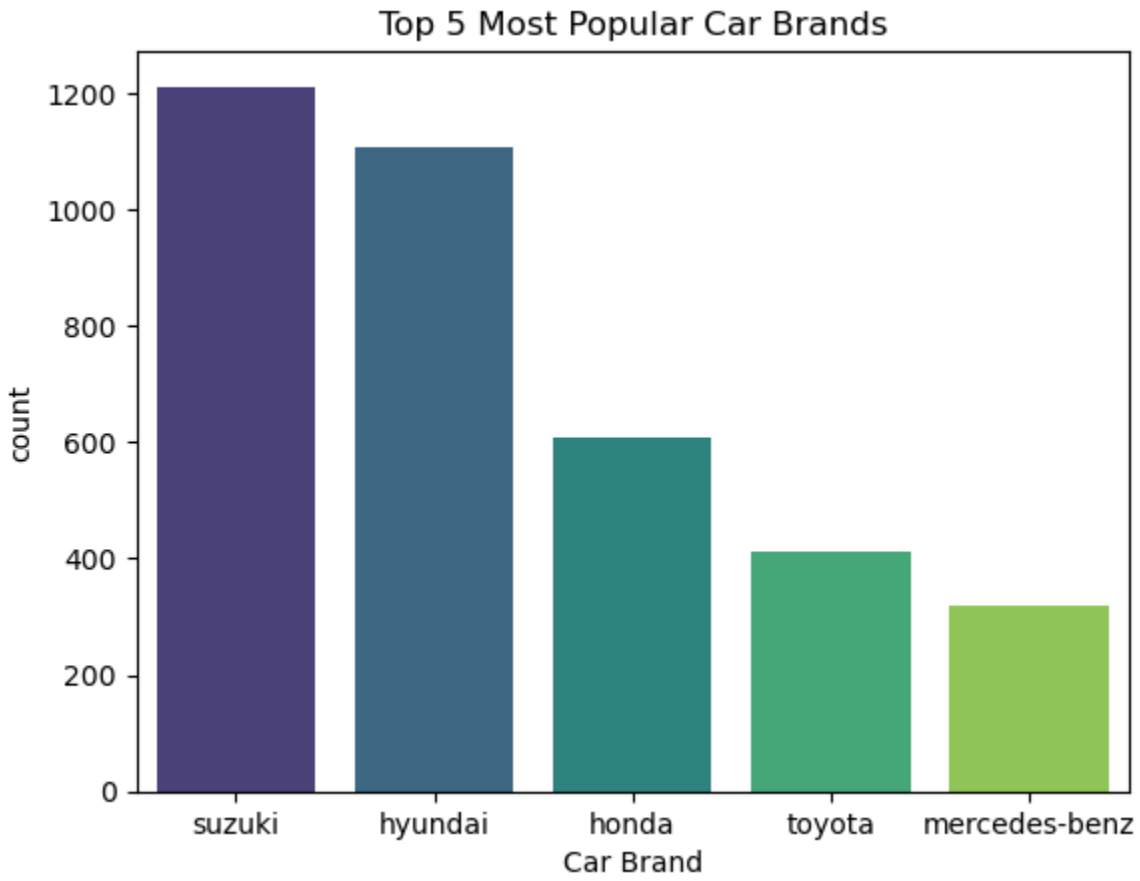
Observations

Diesel and petrol are the dominant fuel types in this dataset

This distribution reflects the current market affordability or availability of different fuel options.

z

Distribution of 5 Top brands



Observations

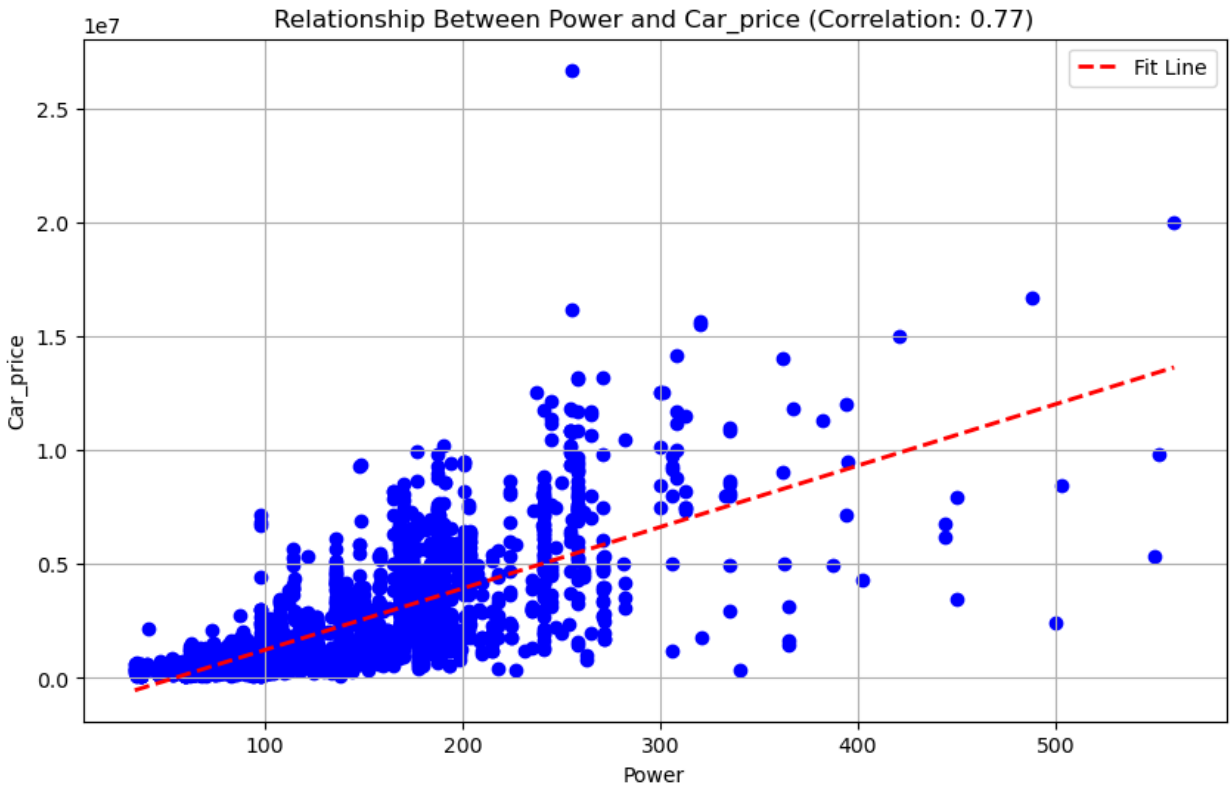
Suzuki and Hyundai are the most popular brands, followed by Honda, Toyota, and Mercedes-Benz although more data is need to conclude these

Understanding brand popularity helps in targeting marketing efforts and managing inventory to meet consumer demand.

Z

Bivalent analysis

Scatter plot of power and car price



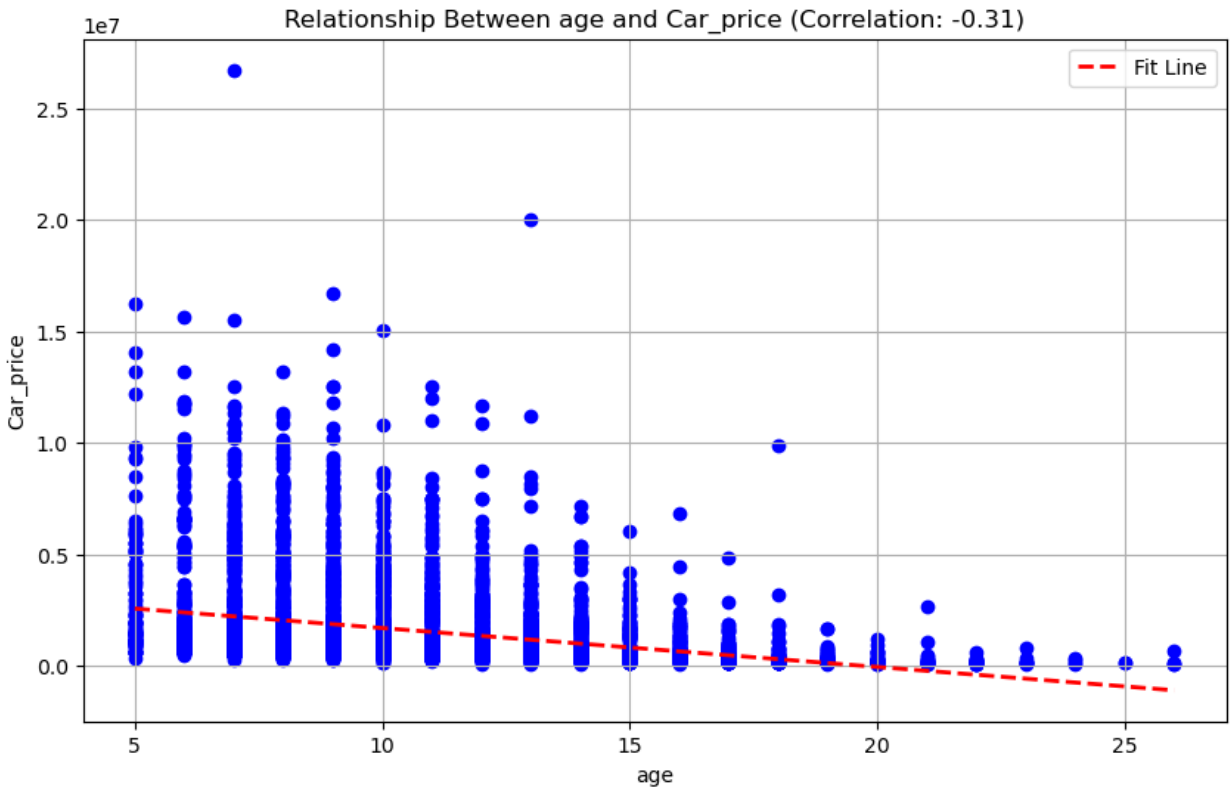
Observations

This indicates a strong positive correlation between car power and price; as the power increases, so does the price

Cars with more power sell for much higher prices

Z

Scatter plot of age and car price

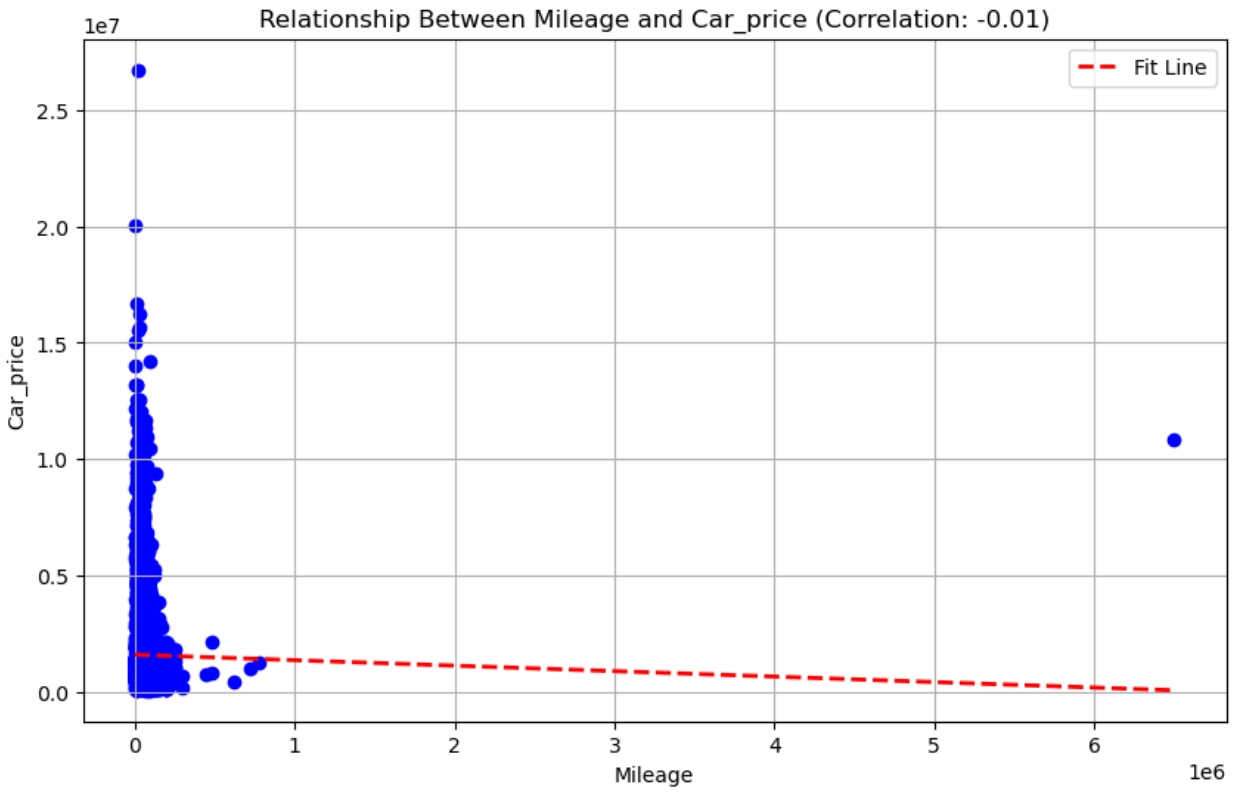


Observations

A weak negative correlation between the two variables, the more the car ages the lower the price indicating cars more than 10 years old depreciate more in terms of the value

Z

Scatter plot mileage and car price



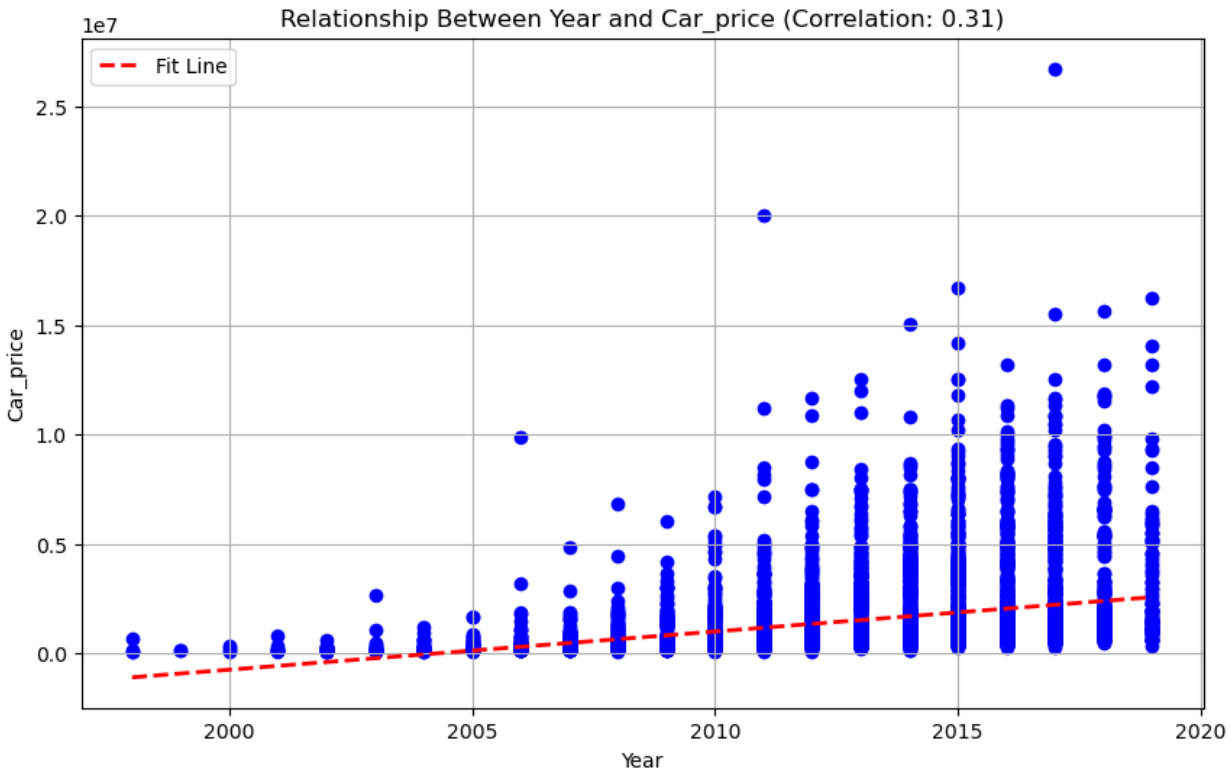
Observations

Weak negative correlation: In this dataset, there is no relationship between a car's mileage and its price

Mileage alone cannot be used alone to determine the price of the vehicle other features must be included

Z

Scatter plot between year and caprice



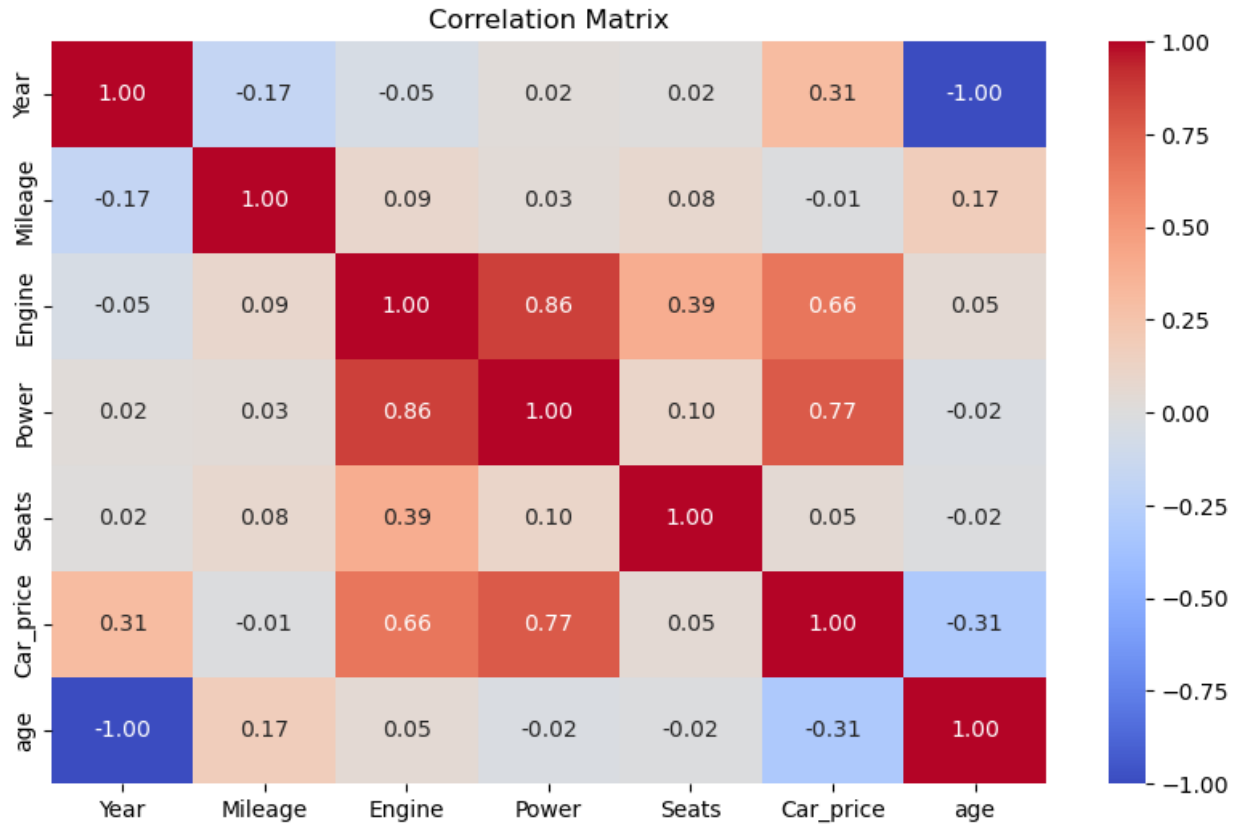
Observations

This indicates a positive but relatively weak correlation between the year of manufacturing and car prices meaning as cars age, their value typically depreciates, resulting in lower prices

Z

Multicolinearity analysis

Correlation matrix



Observations

The matrix demonstrates that several elements have significant positive relationships, such as the correlation between engine and power (0.86) and the correlation between seats and car pricing (0.77). This implies that as engine power increases, so does overall vehicle power, and as the number of seats in a vehicle increases, so will the car's price.

On the other side, the matrix shows negative correlations, such as the correlation between year and mileage (-0.17), indicating an inverse relationship between these two characteristics. This may imply that as the year of the vehicle grows, the mileage tends to decrease, or vice versa.

The presence of both positive and negative correlations indicates that the dataset may contain multicollinearity, which might have repercussions when modelling

VIF Analysis

The VIF calculates the variance inflation factor to determine multicolinearity among the features

FEATURE	**VIF CONSTANT**
Year	infinite
Mileage	1.044489
Engine	6.350604
Power	6.610625
Seats	1.629408
Car price	3.120482
Age	infinite

The high VIF values for the "Engine" and "Power" features suggest that they may be closely related, and including both of them in a regression model could lead to issues with multicollinearity

z

Modelling and Evaluation

Data preprocessing steps followed:

Feature selection

We will use two methods:

Variance Inflation Factor (VIF) analysis

According to the VIF we will drop the power column because it has very high vif value compared to the rest and it's very highly correlatable to engine

Domain Knowledge

The features like brand, year, mileage, transmission, use, seats, and fuel type are significant because these features are relatable to the real world

Age of the car does vary overtime hence its unstable to be used for modelling

We will use 8 features for modelling; brand, year, mileage, transmission, use, seats, fuel type, engine

Train test split

We will used the 80:20 split ensuring 80% of our data is for training and 20% for testing

One hot Encoding

Why encoded our categorical variables to represent them as numerical for our machine learning models to understand our data

z

Why one hot encoding it ensures that the categorical variables are represented in a way that satisfies the linearity assumption, as each binary column can have a separate coefficient in the model while preserving the information about the categorical variables

Scaling

Why scale the numerical data only:

Numerical variables can have vastly different scales and distributions, resulting in some variables dominating the model's learning process.

Scaling the numerical variables ensures that all variables are on the same scale, preventing specific variables from having a disproportionate impact on the model.

Scaling categorical variables would distort the inherent information they provide.

Why standard scaler:

The Standard Scaler is less sensitive to outliers in the data compared to Min-Max Scaling the Standard Scaler performs better with skewed distributions than Min-Max Scaling, which can compress the data range for severely skewed features.

z

Evaluation

Why the evaluation metrics?

R-squared

It measures general the proportion of variance in the dependent variable explained by the independent variables

It's easy to interpret (higher values indicate better model performance) and also it’s easy to use it to compare different models

Mean_absolute_error

It calculates the average of absolute differences between predicted and actual values

It's less sensitive to outliers compared to RMSE and is easy to interpreted

Root_mean_square_error

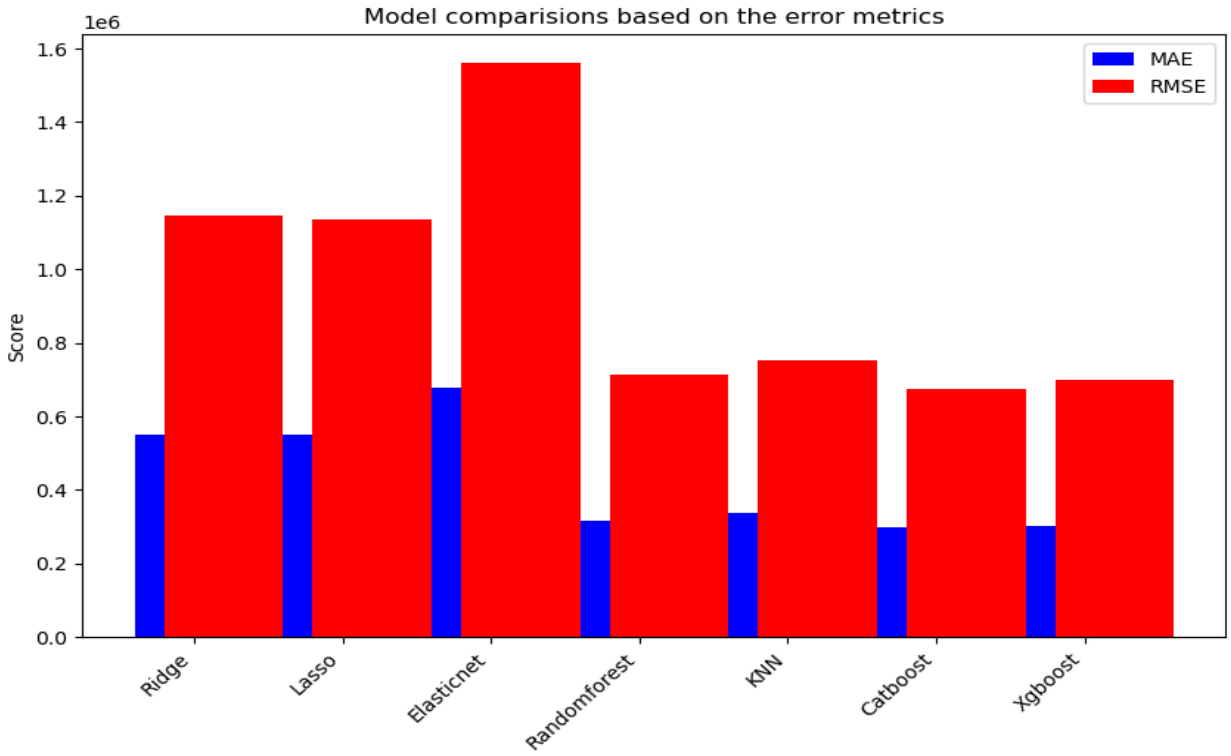
Its square the root of the average of mean_square_error to determine the differences between predicted and actual values

It penalizes large undesirable errors more heavily

We used Nine Models; Linear regression, Ridge regression, Lasso regression, Elastic Net, Random forest repressor, KNN repressor, Cat boost regressor,XGBoost repressor and Recurrent Neural Network

Z

MODEL	R-SQUARED	MAE (Mean Absolute Error)
Cat Boost Regression	0.87	298328.91
XGBoost Regression	0.86	302261.48
Random Forest Regressor	0.85	315234.6
KNN Regressor	0.84	337761.66
Lasso Regression	0.63	5492235.34
Ridge Regression	0..62	550384.84
Elastic Net Regression	0.29	678576.8
Linear Regression	-2.37	3632345302271830.06



Based on the error metrics displayed in the graph, the models can be compared as follows: XGBoost and Cat Boost appear to be the best-performing models, with the lowest MAE and RMSE scores, indicating the smallest average prediction errors. The Random Forest model follows closely behind these two, indicating that ensemble approaches are quite effective on this dataset. KNN has slightly higher error rates yet works pretty well. The Lasso and Ridge regression models have similar performance, with moderate error rates that are greater than the ensemble approaches but lower than Elastic Net. The Elastic Net model has by far the highest error rates, with both its MAE and RMSE being much bigger than any other models, showing that it may not be suitable for predicting the prices.

Our best performing model was the **Cat boost regressor** model why?

It had the highest Test Set R^2 Score:

On the test set, Cat boost achieved the highest R^2 score of 0.86, showing strong predictive performance on unseen data.

Lowest MAE (Mean Absolute Error):

Cat boost has the lowest MAE of 298,328.91, suggesting it has the smallest average absolute difference between predicted and actual values.

Lowest RMSE (Root Mean Squared Error):

With an RMSE of 674,239.75, Cat boost has the lowest average root squared difference between predicted and actual values.

z

Conclusions

Cat boost is the best-performing model.

Engine and power are the best qualities that positively affect the vehicle's price.

Suzuki was the most popular automobile brand.

Vehicles produced in 2015 attracted greater prices.

Most of the cars featured four to seven seats.

Most automobile costs were between 500,000 and 1.5 million.

Recommendation

Deployment and constant evaluation of the Cat boost model

Collection and analysis of quality Kenyan data to conclude on the best performing used car brand in kenya as well us improving on the model performance through providing more data for training and testing the models

Dealers should import newer vehicles in regard to year of manufacturing, cars models manufactured from 2015 and above command higher prices

Importation of family oriented vehicles which have 4 to 7 seats

Importation of affordable vehicles prices ranging between 500,000 to 1.5million

Preference of smaller engine capacities offer buyers value for money in terms of fuel savings and maintenance cost

Tuning of the power output in vehicles could proof to be an investment as it raises the car value bring and when sold it has a high return on investment