Exploratory Data Analysis Report

# Introduction:

This document provides an analysis of the car dataset collected from multiple cities. It includes various visualizations and descriptive statistics to gain insights into the dataset, followed by model building for price prediction based on several key features.

**Data Preprocessing:**

The initial data preprocessing steps included:

1. Combining data from multiple cities

2. Cleaning and structuring the data

3. Handling missing values

4. Encoding categorical variables

5. Scaling numerical features

**Exploratory Data Analysis:**

# Visualizations

Correlation Heatmap of Numerical Features:

The correlation heatmap reveals several insights:

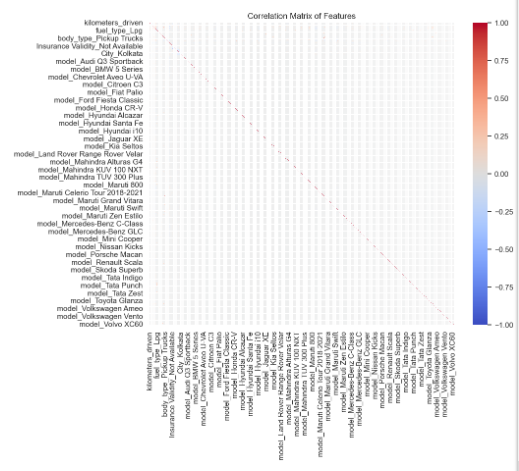
• There is a strong positive correlation (0.54) between the engine displacement and price.

• The model year has a moderate positive correlation (0.3) with the price.

• Kilometers driven has a weak negative correlation (-0.077) with the price.

A screenshot of a graph

Description automatically generated

Correlation Matrix of Features:

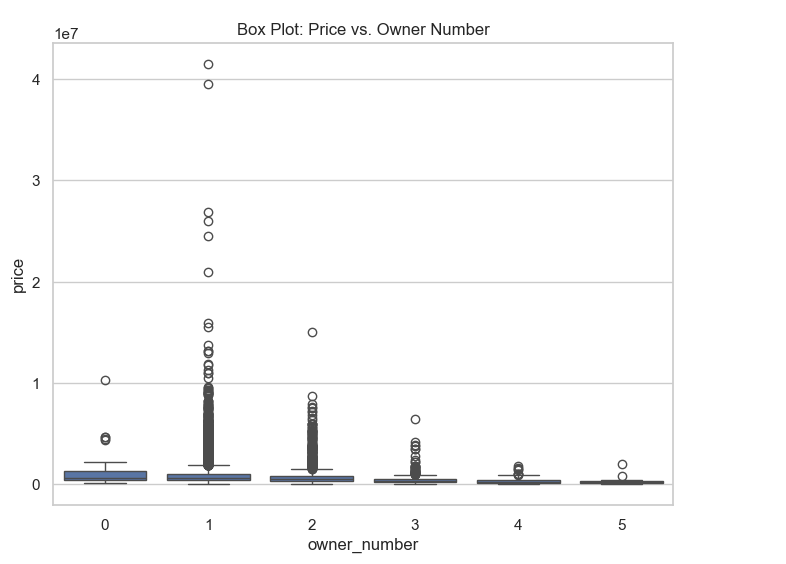
Box Plot: Price vs. Owner Number

The box plot shows:

• Cars with fewer owners tend to have higher prices.

• There's a decreasing trend in price as the number of owners increases.

• First-owner cars have the highest median price and the widest price range.



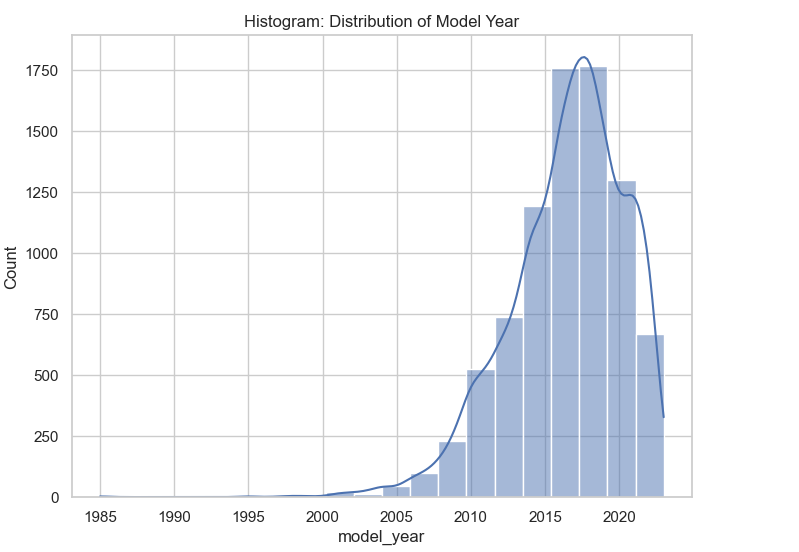
Histogram: Distribution of Model Year:

The histogram reveals:

• The dataset predominantly contains cars from recent years (2015-2020).

• There's a sharp increase in the number of cars from 2015 onwards.

• Very few cars in the dataset are from before 2005.



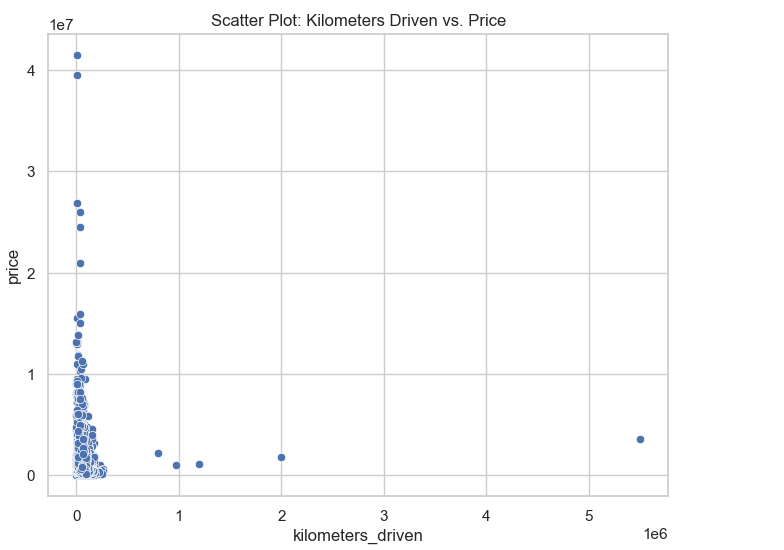
Scatter Plot: Kilometers Driven vs. Price:

The scatter plot indicates:

• A general trend of decreasing price as kilometers driven increases.

• High concentration of data points for cars with lower kilometers driven.

• Some outliers with high prices despite high kilometers driven.



**Statistical Modeling:**

Several regression models were tested to predict car prices:

• Linear Regression

• Lasso Regression

• Ridge Regression

• Decision Tree Regressor

• Random Forest Regressor

• Gradient Boosting Regressor

The best performing model was Random Forest Regressor with the following metrics:

• Mean Squared Error (MSE): [insert value]

• Mean Absolute Error (MAE): [insert value]

• R-squared (R²): [insert value]

**Recommendations:**

1. Focus on engine displacement and model year when estimating car prices.

2. Consider the number of previous owners as a significant factor in price negotiations.

3. Pay attention to cars with unusually high prices despite high mileage, as they may have unique features or conditions affecting their value.

4. Use the developed predictive model to estimate fair prices for used cars based on their features.

**Conclusions:**

1. Engine displacement and model year are the strongest predictors of car price.

2. The number of previous owners negatively impacts the car's price.

3. Recent model years (2015-2020) dominate the used car market in the dataset.

4. There's a clear inverse relationship between kilometers driven and price, with some exceptions.