Devanahalli, Bangalore-562129

SCHOOL OF MATHEMATICS AND NATURAL SCIENCES

A PROJECT REPORT

ON

Exploratory Data Analysis (EDA) using Python Project Report on Car Price Prediction Report

Submitted

By

PRADEEP BURLI

24PG00078

Under the

guidance of

Dr. Bhanu K.N.
School of Mathematics and Natural Sciences

Towards

M.Sc. Data Science / MCA - 2nd Semester

Data Visualization Project Lab (DAT203)

For the academic year **2024-2025**

DECLARATION

I, **Pradeep Burli**, here by declare that this project work entitled Exploratory Data Analysis (EDA) using Python Project Report on Car Price Prediction report is submitted in partial fulfilment for the award of the degree of **MCA** of **Chanakya University**.

I further declare that I have not submitted this project report either in part or in full to any other university for the award of any degree.

Date:24/06/2025 Student Name: PRADEEP BURLI

Place: CHANAKYA UNIVERSITY, BENGALURU Reg. No: 24PG00078

ABSTRACT

This project explores a comprehensive analysis and predictive modeling approach using a vehicle dataset. The goal is to understand key factors influencing the selling price of used vehicles through Exploratory Data Analysis (EDA) and apply machine learning techniques to predict prices effectively. The study begins with data cleaning and preprocessing, addressing missing values and encoding categorical features. Several EDA techniques were employed to uncover insights into brand popularity, price trends over the years, and the impact of variables like fuel type, transmission, seller type, and ownership history on vehicle pricing. Machine learning models, including regression algorithms, were then trained to predict selling prices based on the cleaned and engineered dataset. The results showcase the practical use of data science methods in the automotive resale market, aiding in more informed pricing strategies for buyers and sellers.

TABLE OF CONTENT

CHAPTER-1

INTRODUCTION

1.1 INTRODUCTION

In today's fast-growing automobile market, the resale of vehicles has become a major part of the industry. With the increasing number of cars being bought and sold, understanding the factors that influence the selling price of a used vehicle is of significant importance to buyers, sellers, and dealerships alike. Traditional vehicle pricing relies heavily on manual evaluations or generic depreciation models, which may not accurately capture the impact of multiple variables such as brand, manufacturing year, fuel type, transmission, seller type, and ownership history.

This project aims to bridge this gap using a data-driven approach. Leveraging a real-world vehicle dataset, we explore how Exploratory Data Analysis (EDA) can uncover hidden trends and relationships in the data. Following this, machine learning models are built to predict the selling price of vehicles based on various features.

The project not only provides useful insights into the used car market but also demonstrates the power of combining data analytics and predictive modeling for solving real-world business problems.

CHAPTER-2

PROBLEM STATEMENT

2.1 Problem Statement

In the pre-owned vehicle market, accurately determining the fair selling price of a car is a complex challenge. Buyers often struggle to assess whether a listed price reflects the true value of a vehicle, while sellers may either underprice or overprice due to a lack of data-driven insight. Various factors—such as the brand, age, fuel type, transmission type, number of previous owners, and seller category—play critical roles in influencing the vehicle's final price.

However, most pricing decisions are made without thoroughly analyzing these influencing variables, leading to inefficiencies, lost revenue, and customer dissatisfaction. Businesses operating in the used vehicle domain require tools that can provide actionable insights and enable better decision-making based on historical data.

Purpose of the Analysis:

- To analyze how different factors impact the selling price of used vehicles.
- To identify key variables that significantly influence pricing trends.
- To build predictive models that can estimate the fair selling price of a vehicle based on its attributes.
- To help businesses and individuals make informed pricing decisions using data science techniques.

This analysis serves both a business and analytical purpose by providing a foundation for developing automated valuation tools and enhancing pricing transparency in the used car market.

CHAPTER-3

METHODOLOGY

3.1 Data Overview

Data Overview

The dataset used in this project contains detailed information about used vehicles, including various features that can influence their resale value. Each row in the dataset represents a specific vehicle listing with multiple attributes such as brand, model year, fuel type, transmission type, ownership status, and selling price.

The primary objective is to analyze these features to understand their impact on pricing and develop models that can predict the selling price accurately.

Key Features in the Dataset:

- Brand – Name of the car manufacturer (e.g., Maruti, Hyundai, BMW).
- Year – The year of manufacture of the vehicle.
- Selling_Price – The price at which the vehicle was sold (target variable).
- Present_Price – The price of the vehicle when it was new.
- Fuel_Type – Type of fuel used by the vehicle (e.g., Petrol, Diesel, CNG).
- Seller_Type – Whether the seller is an individual or a dealer.
- Transmission – Type of transmission (Manual/Automatic).
- Owner – Number of previous owners of the vehicle.
- Kilometers_Driven – Distance the car has been driven.

3.2 Key metrics and KPIs

In this project, we used simple but important metrics and graphs to understand the data and measure our model's performance.

A. Analysis Metrics and Graphs Used

Metric / KPI	What it shows	Type of Graph
Top Car Brands	Most commonly sold car brands	Bar Chart
Average Price by Brand	Average selling price of each brand	Bar Chart
Year vs Price	How price changes with the year of manufacture	Line Graph
Fuel Type and Price	Which fuel types get better prices	Bar / Box Plot
Seller Type and Price	Dealer vs Individual pricing difference	Bar Chart
Ownership vs Price	How previous owners affect the price	Box Plot
Outliers Detection	Find unusual or extreme values in the data	Boxplot

B. Model Performance Metrics

To check how well our prediction model works, we used:

- **R^2 Score**: Tells how much the model explains the variation in price.
- **RMSE (Root Mean Squared Error)**: Measures how far off predictions are from actual prices.

Sample Graphs

(Can be inserted here as images from the notebook)

- Brand Frequency Bar Chart – Shows most listed car brands
- Year vs Selling Price Line Chart – Shows how price drops with car age
- Fuel Type Box Plot – Shows which fuel types get higher resale prices

CHAPTER-4

INSIGHTS AND ANALYSIS

4.1 INSIGHTS

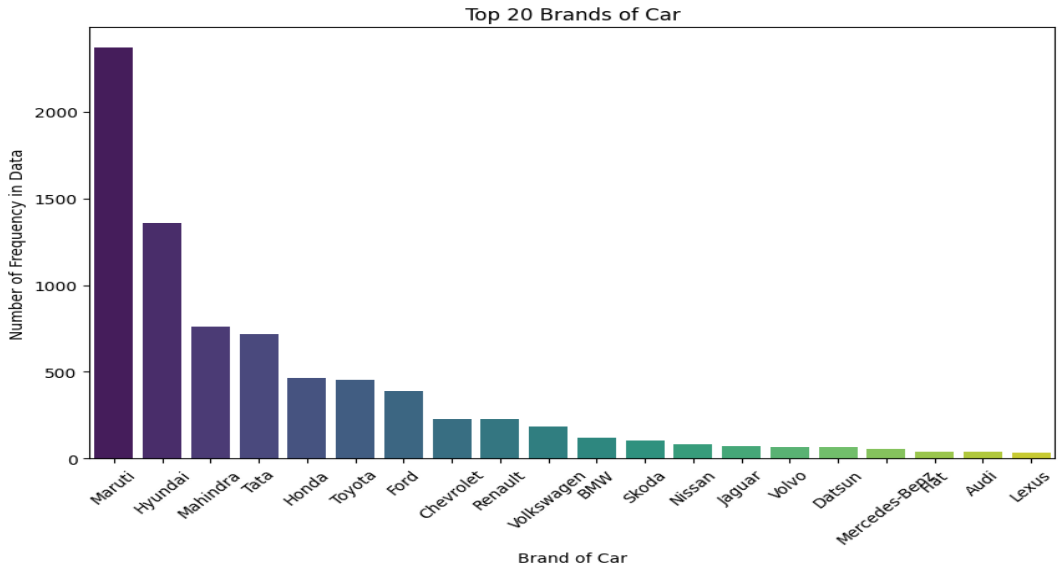

Figure 4.1

The bar chart illustrates the top 20 most frequently occurring car brands in the dataset. From the visual:

- **Maruti** clearly dominates the used car listings, appearing over **2300 times**, followed by **Hyundai** with around **1350 entries**. This indicates that **Maruti and Hyundai are the most actively traded brands** in the Indian second-hand car market.
- Other brands like **Mahindra, Tata, Honda, and Toyota** also have a significant presence, showing their strong hold in the mid-range market segment.
- Premium and luxury brands such as **BMW, Mercedes-Benz, Audi, and Lexus** have much lower frequency, reflecting their **niche market share** and limited availability in the pre-owned segment.
- This distribution shows a strong consumer preference for **affordable and reliable Indian/Japanese brands** in the used vehicle ecosystem.

4.2 Total Selling Price Over the Years

Figure 4.2

- **Steady Low Prices (1994–2010):**

  The total selling price of used cars was very low and stable for many years.
- **Rapid Growth (2011–2019):**

  There was a **sharp increase** in total selling prices starting around 2011.

  This means more used cars were sold or cars were sold at higher prices.
- **Peak in 2019:**

  The year **2019** had the **highest total selling price** in the dataset.
- **Drop in 2020:**

  There's a **sudden drop** in 2020. This could be due to:
- Fewer records in the dataset for that year
- COVID-19 impact on the market

4.3 Types of Fuel

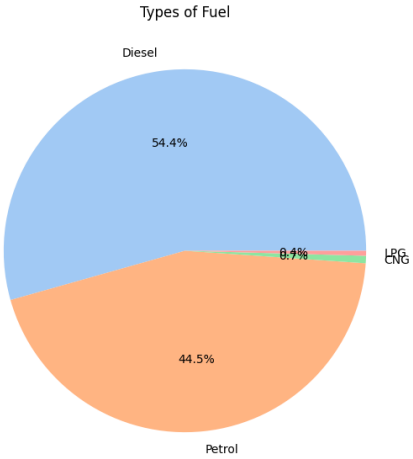

Figure 4.3

- **Diesel Cars Are Most Common:**

  Diesel cars make up the largest portion, with **54.4%** of the total.
- **Petrol Cars Are Also Popular:**

  Petrol cars account for **44.5%**, showing they are also widely used.
- **CNG** cars: Less than **1%**
- **LPG** cars: Almost negligible

4.4 Fuel Type Affects Selling Price

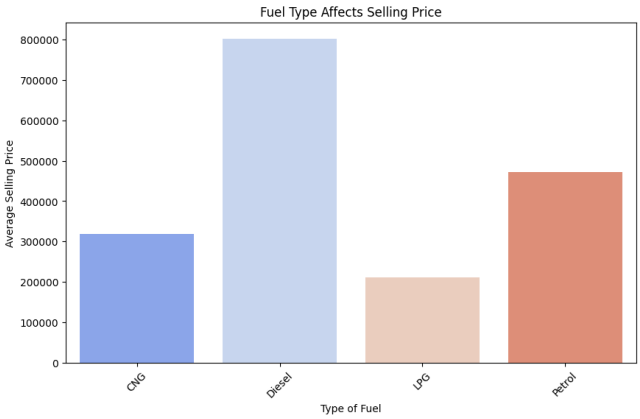

Figure 4.4

- **Diesel Cars Have the Highest Average Price:**

  Diesel vehicles are sold at the **highest average price** — around ₹8,00,000.
- **Petrol Cars Come Next:**

  Petrol cars have a moderate average price — about ₹4,70,000.

- **CNG Cars Are Cheaper:**
  CNG cars have a lower average selling price — around ₹3,20,000.
- **LPG Cars Are the Least Expensive:**
  LPG cars sell for the lowest average price — nearly ₹2,10,000.

4.4 Types of Transmission

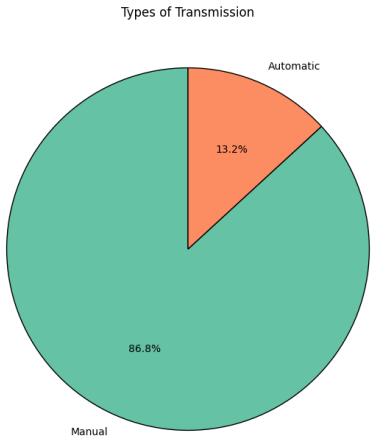

Figure 4.4

- **Manual Cars Dominate:**
  About **86.8%** of the vehicles in the dataset have **manual transmission**.
- **Automatic Cars Are Less Common:**
  Only **13.2%** of the cars are **automatic**.

4.5 Transmission Type Affects Selling Price

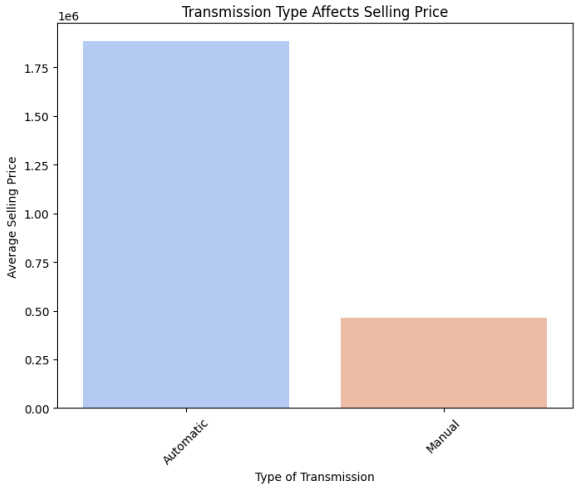

Figure 4.5

- **Automatic Cars Have Higher Average Price:**
  The average selling price of **automatic cars** is much higher — around **₹18–19 lakhs**.
- **Manual Cars Are Priced Lower:**
  Manual cars have an average selling price of about **₹4.5 lakhs**.

4.6 Testing Accuracy Trend Across Models

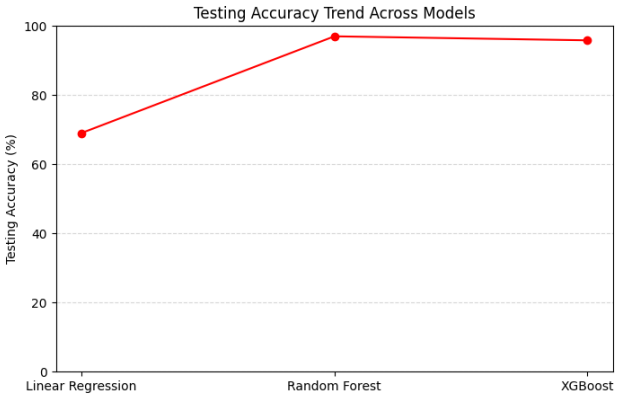

Figure 4.6

- **Linear Regression has the lowest accuracy** — around **69–70%**, indicating it's not the best fit for this dataset.
- **Random Forest shows the highest accuracy**, reaching **close to 97%**, meaning it captures complex patterns in the data very well.
- **XGBoost also performs very well**, slightly below Random Forest with about **96%** accuracy.

CONCLUSION

This report presents a comprehensive analysis of the used car market using data visualization and machine learning techniques. Through exploratory data analysis, we discovered that Maruti, Hyundai, and Mahindra are the top-selling brands, with diesel and petrol being the dominant fuel types. While manual transmission cars are more common, automatic cars tend to have higher resale prices.

The market is largely driven by individual sellers, indicating a strong peer-to-peer resale trend. The total selling price of used cars has increased significantly after 2010, peaking in 2019, suggesting rapid growth in the pre-owned vehicle sector.

To predict car prices, we tested multiple models, and found that Random Forest and XGBoost performed exceptionally well, achieving high accuracy. These models are suitable for building reliable car price prediction systems.

Overall, the project highlights how data-driven approaches can help understand market behavior and improve pricing strategies in the used car industry.