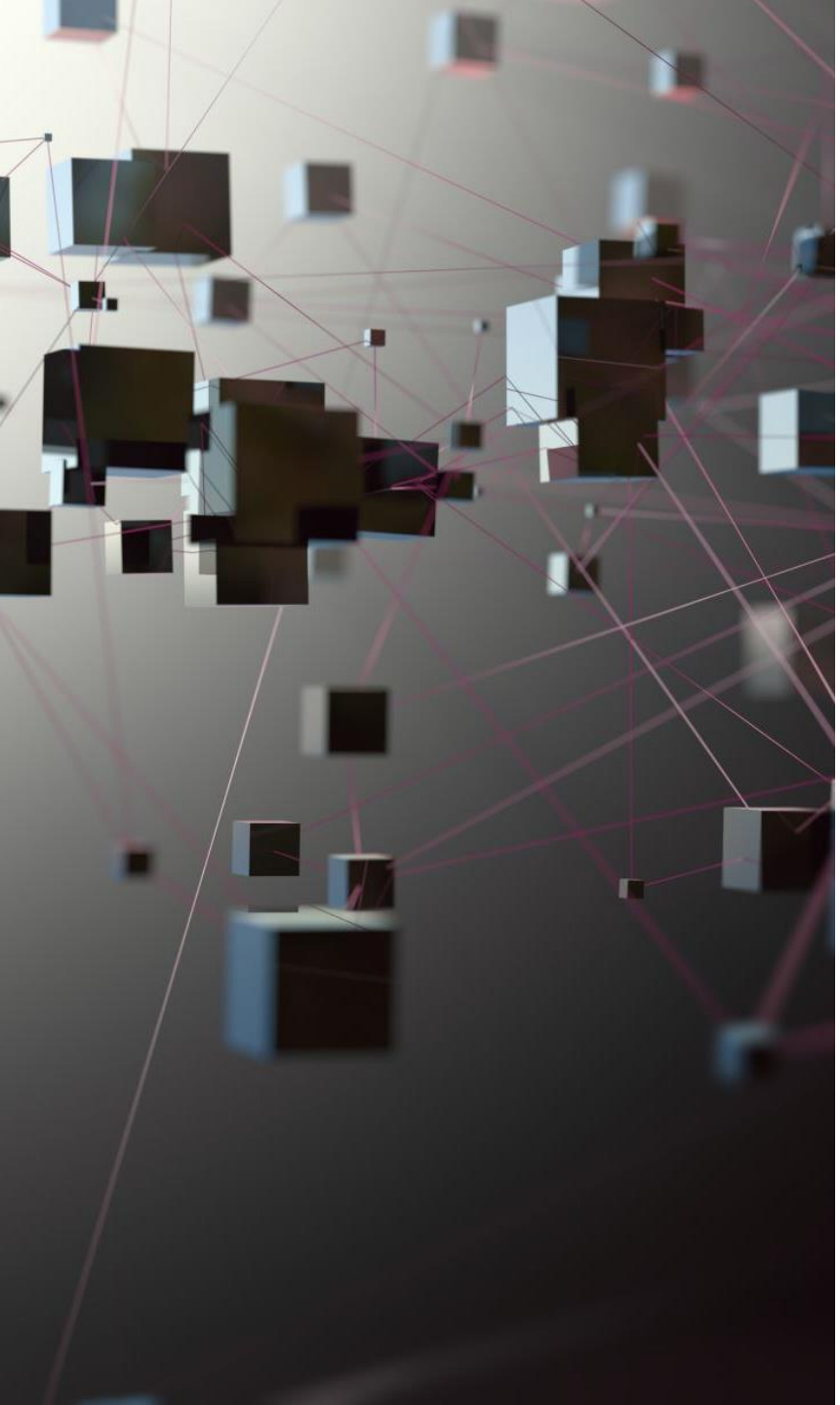
Scraping Project

- EVOASTRA VENTURES MINI PROJECT
- Mentor-Aniket Manwatkar
- Team D



Topics

- Introduction
- Tools and Technologies
- Project Workflow
- Challenges and Solutions
- Analysis and Insights
- Data Visualisation
- Conclusion

Project Introduction

Objective

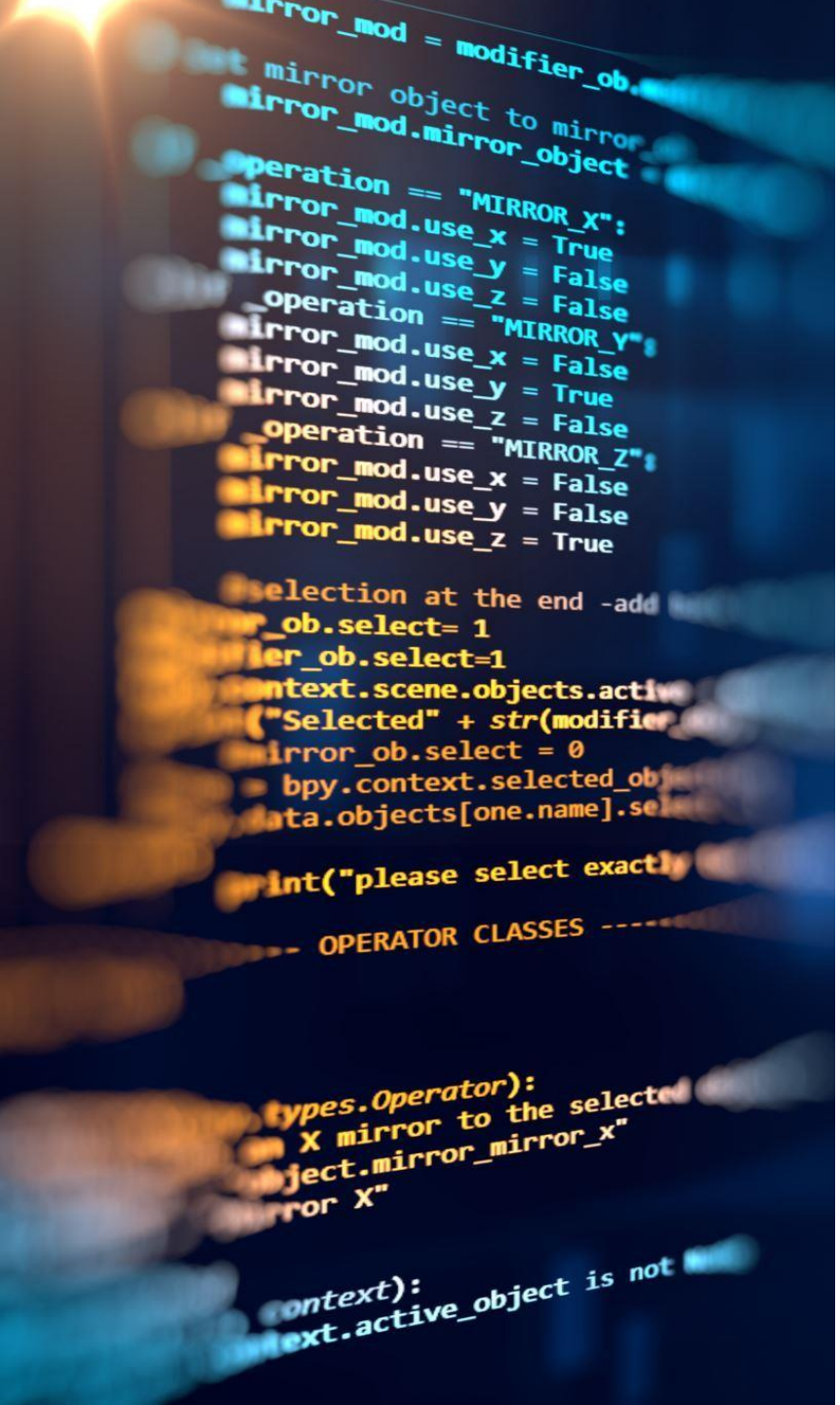
To scrape detailed car listings from Cars24.com to build a comprehensive dataset for analysis.

Purpose and Goals of the Scraping Project

The goal of this project was to gather detailed information about used cars

We focused on extracting key data points such as:

- **Company Name**
- **Model**
- **Year of Manufacture**
- **Price (in lakhs)**
- **Kilometers Driven**
- **Fuel Type**
- **Transmission**
- **Location**



Tools & Technologies

Programming Language: Python

Libraries Used:

- BeautifulSoup & Selenium for parsing HTML and XML.
- Requests for handling HTTP requests.
- Pandas &Numpy for data manipulation.
- Matplotlib & Seaborn for data visualization.
- Warnings to ignore warnings

Environment: Jupyter Notebook

Project Workflow

- **Step 1:** Target website is https://www.cars24.com/.
- **Step 2:** Sent HTTP requests to fetch the cars24.com website.
- **Step 3:** Parse the HTML content.
- **Step 4:** Extract relevant data e.g Company name,model,KM driven,fuel &transmission type,price and location using BeautifulSoup & Selenium library.
- **Step 5:** Store the data in a structured format i.e in CSV.
- **Step 6:** Perform EDA
- **Step 7:** Data Visualisation

Challenges and Solutions

- **Scattered Data Extraction**
- **Issue:** Data was sometimes spread across multiple elements with inconsistent class names and attributes.
- **Solution:** Identified unique and consistent patterns in the HTML structure, allowing for precise data extraction.

Analysis and insights

Total Entries-3247

Columns in dataset – 8

Data Types- Float(1),Int(2),Object(5)

The dataset has no null values present and also No duplicate entries are found,it is a clean dataset.

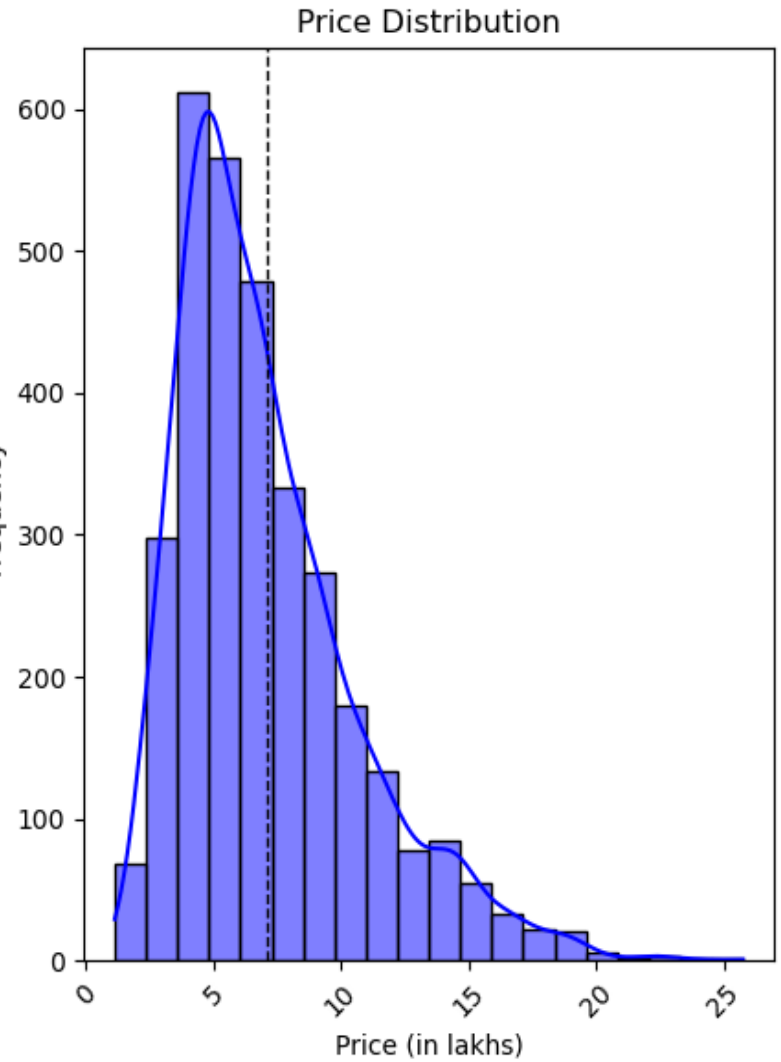
Car Prices Overview:

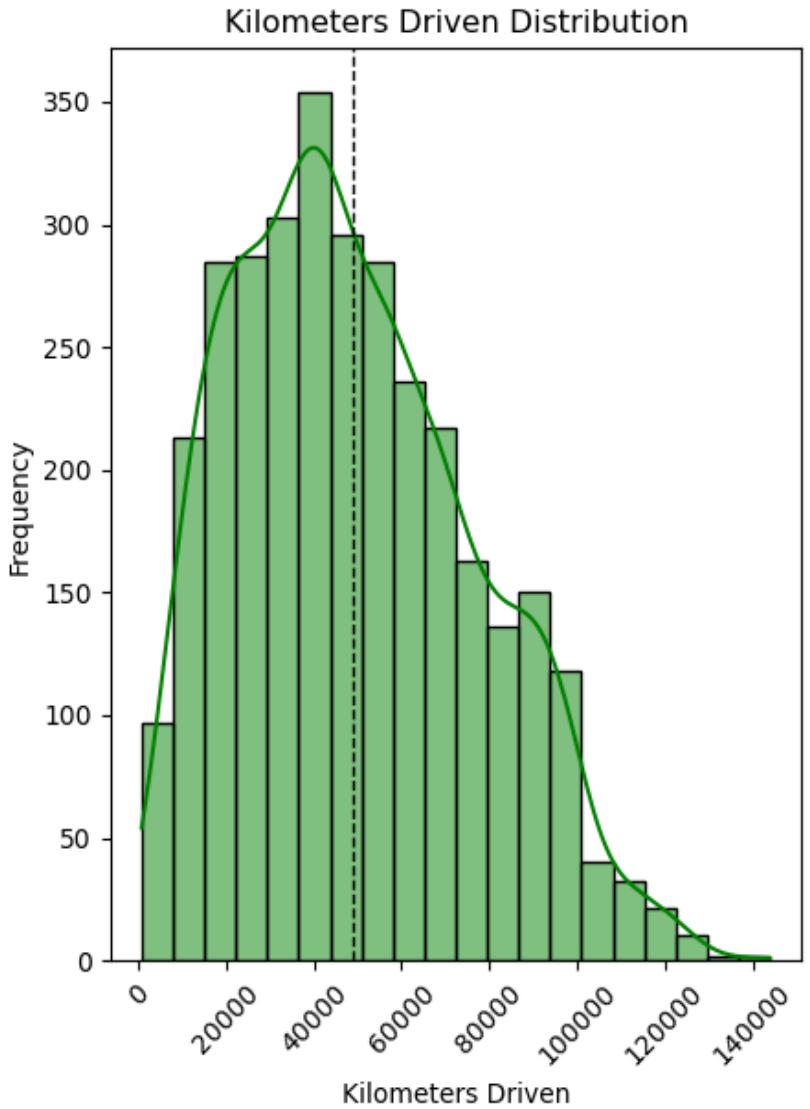
Average Price: ₹7.09 lakhs

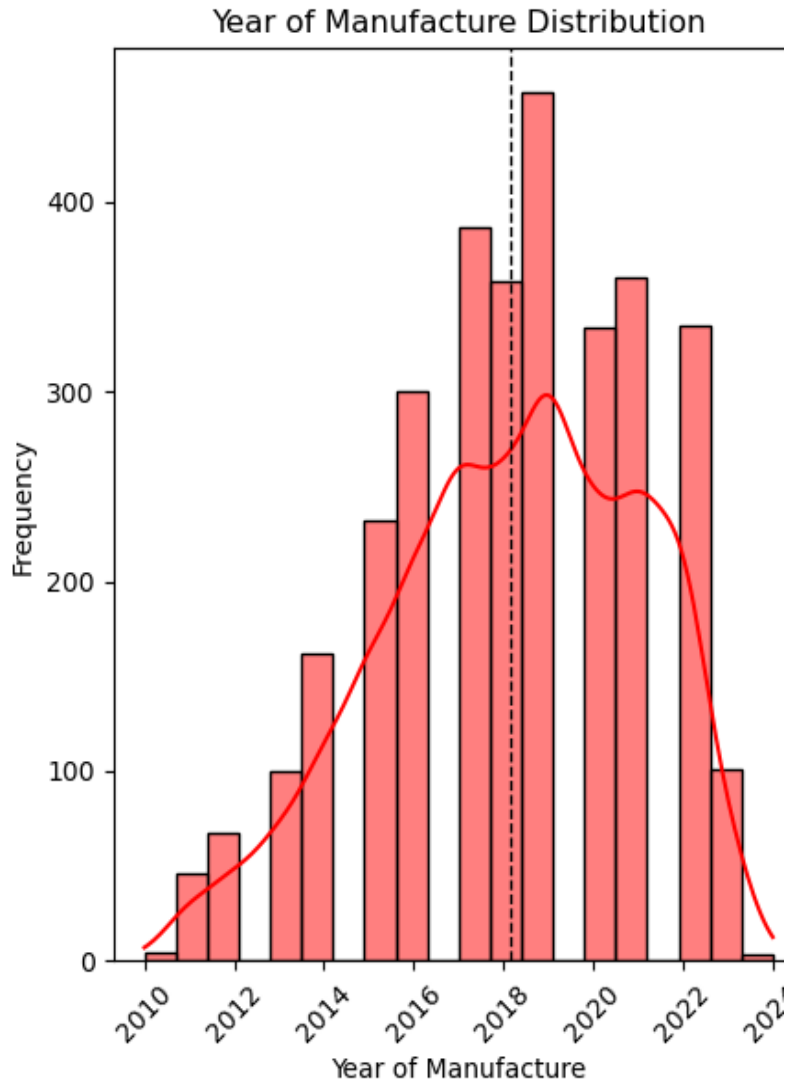
Standard Deviation: ₹3 lakhs

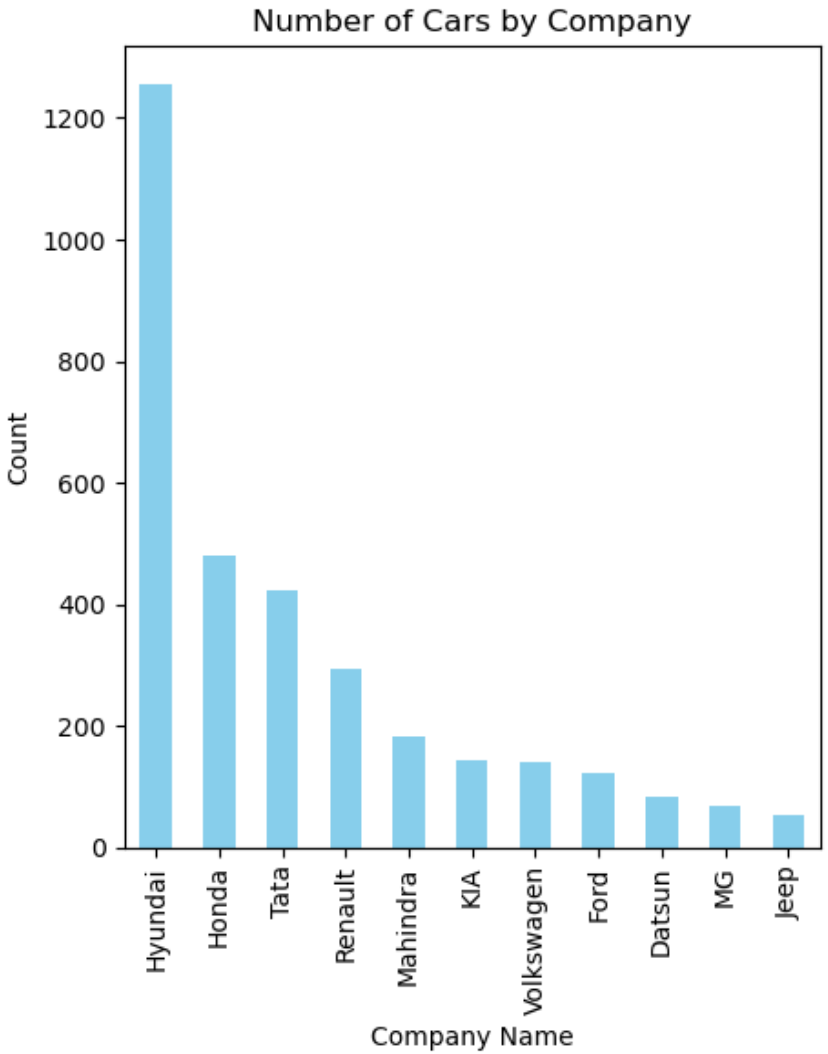
Insight:

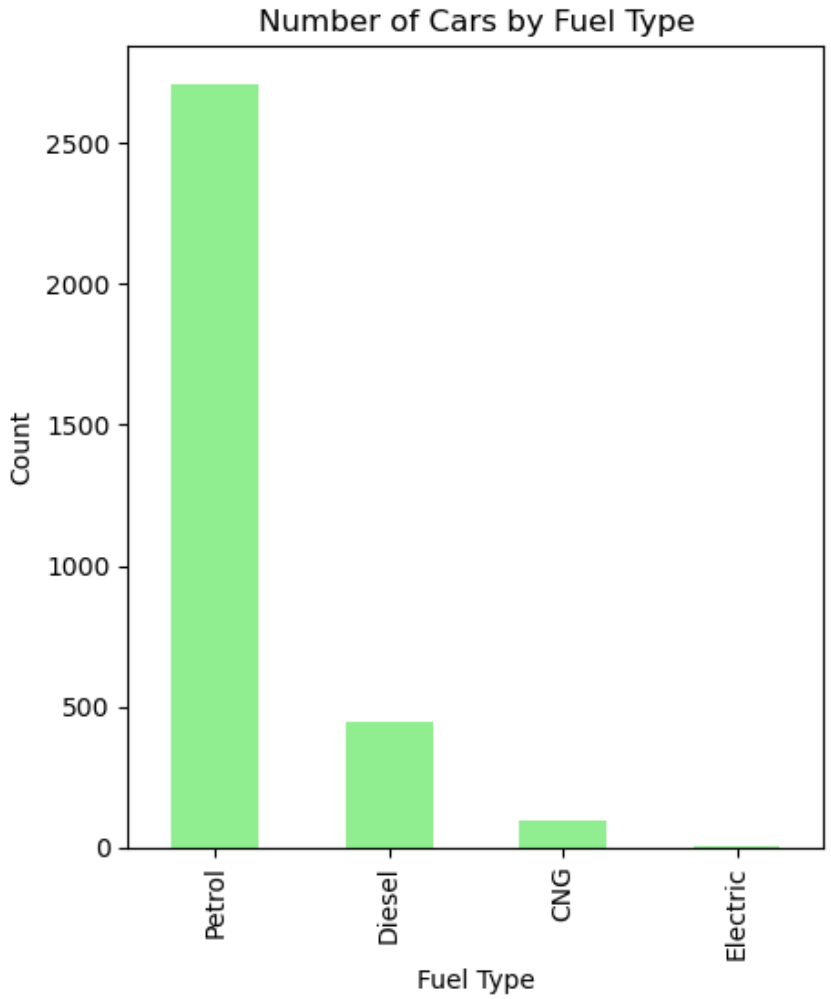
The dataset reveals a diverse range of car prices, a strong preference for Hyundai and petrol vehicles, a prevalence of manual transmissions, and Bangalore as a key market

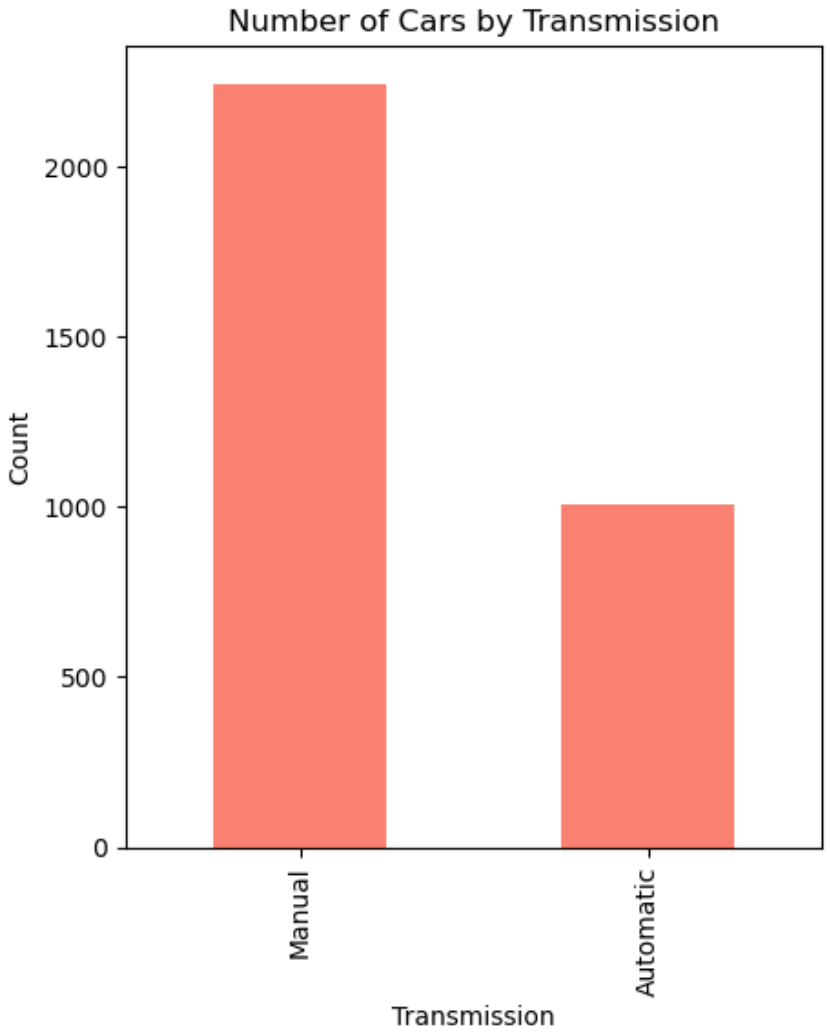

Price Distribution
Frequency
Price (in lakhs)
0
100
200
300
400
500
600
0
5
10
15
20
25

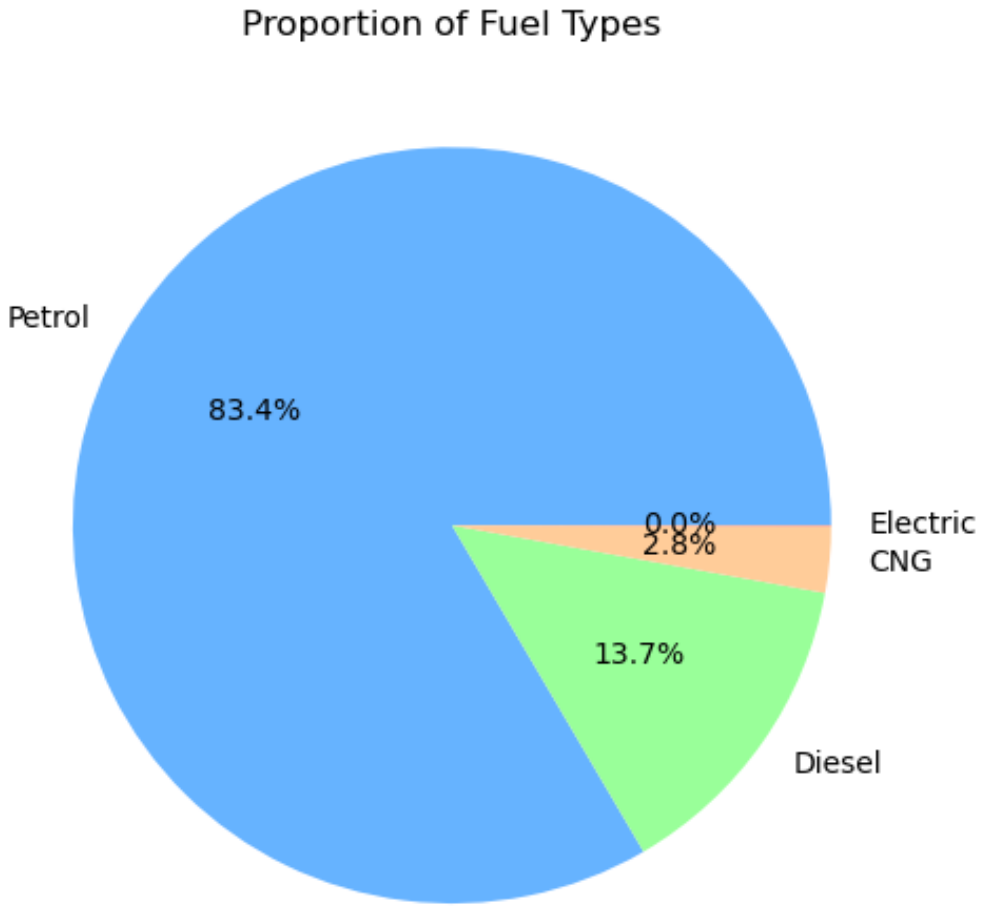

Kilometers Driven Distribution
Frequency
Kilometers Driven
0
50
100
150
200
250
300
350
0
20000
40000
60000
80000
100000
120000
140000

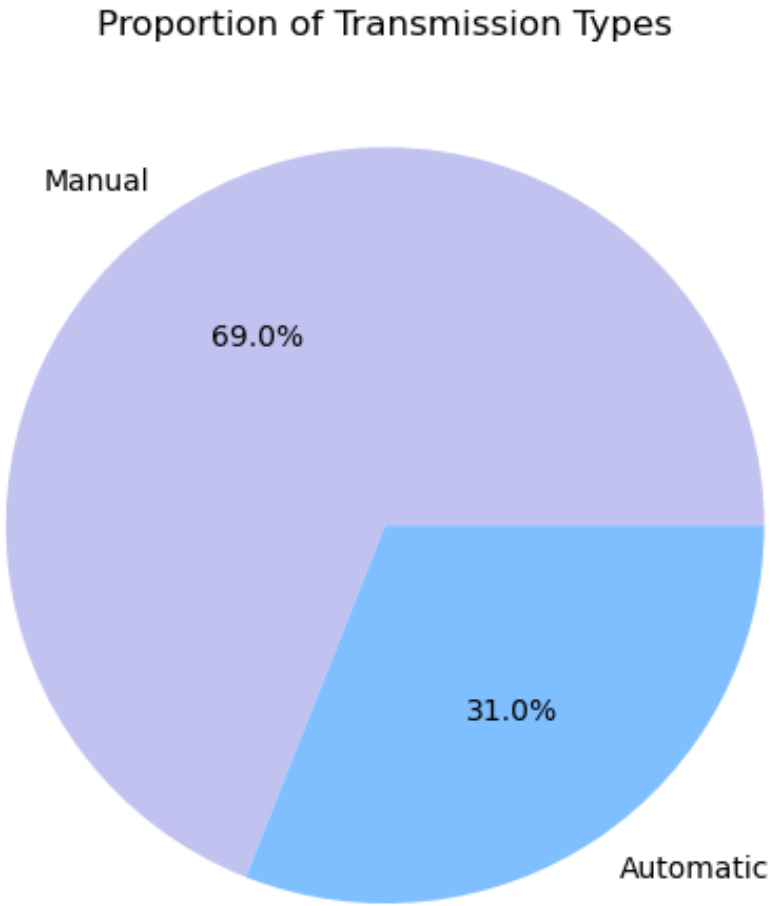

Year of Manufacture Distribution
Frequency
Year of Manufacture
0
100
200
300
400
2010
2012
2014
2016
2018
2020
2022

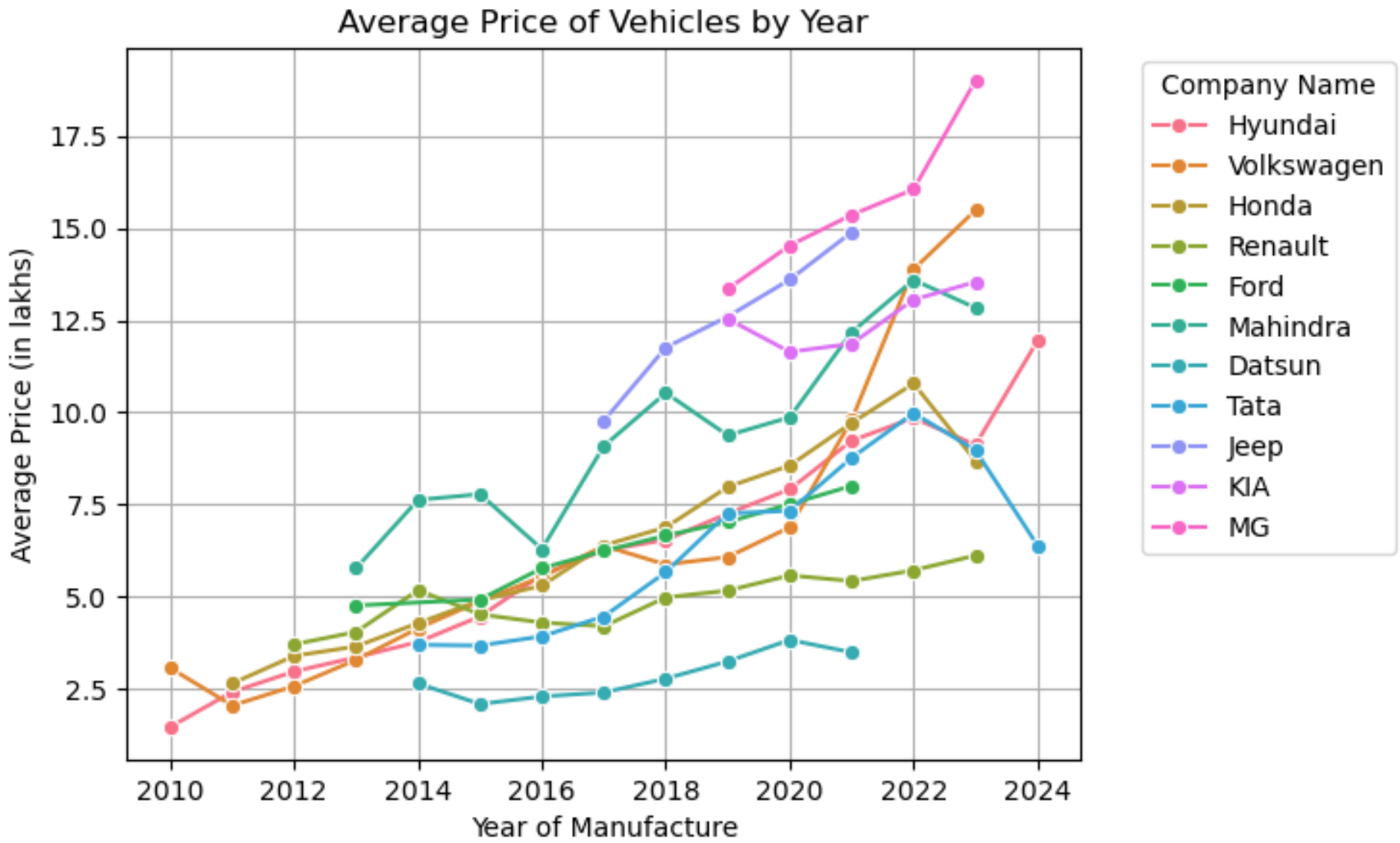

Number of Cars by Company
Count
0
200
400
600
800
1000
1200
Hyundai
Honda
Tata
Renault
Mahindra
KIA
Volkswagen
Ford
Datsun
MG
Jeep
Company Name


Number of Cars by Fuel Type
Count
0
500
1000
1500
2000
2500
Petrol
Diesel
CNG
Electric
Fuel Type


Number of Cars by Transmission
Count
0
500
1000
1500
2000
Manual
Automatic
Transmission


Proportion of Fuel Types
Petrol
83.4%
0.0%
Electric
2.8%
CNG
13.7%
Diesel


Proportion of Transmission Types
Manual
69.0%
31.0%
Automatic


Average Price of Vehicles by Year
Average Price (in lakhs)
17.5
15.0
12.5
10.0
7.5
5.0
2.5
2010
2012
2014
2016
2018
2020
2022
2024
Year of Manufacture
Company Name
Hyundai
Volkswagen
Honda
Renault
Ford
Mahindra
Datsun
Tata
Jeep
KIA
MG

Predicting Prices with Machine Learning Models

- **To accurately predict prices, we will leverage machine learning (ML) models that can identify complex patterns and relationships within the data.**
- **some models that can be used for predicting prices:**
- **Linear Regression:** Simple and interpretable, useful for understanding linear relationships and predicting.
- **Decision Trees and Random Forests:** These models can handle non-linearities and interactions between features effectively.
- **Gradient Boosting Machines:** Improve prediction accuracy by combining multiple weak learning models.
- **Neural Networks:** Capture complex patterns through deep learning techniques and highly customizable.

Conclusion

- **Successful Data Extraction:**
- Efficiently scraped and collected relevant car data, including key attributes like price, brand, and specifications.
- **Overcoming Challenges:**
- Addressed and resolved issues such as handling dynamic content and managing scattered data, ensuring accurate and comprehensive data collection.