PROJECT NAME : WEB SCRAPPING

Team Members:

- **Himanshi Agrawal**
- **Vikash Kumar**
- **Manu Ayodhi**
- **Sethu Madhav**
- **Mohammed Ahmed**
- **Syed Wadood Ahmed**
- **Mirza Zainullah Baig**
- **Pathan Shahanaz**
- **Priyanka Ramesh Burra**
- **Jyoshna Jilagam**

Title : Exploratory Data Analysis (EDA) on Cars24 Toyota

Abstract:

This project performs an **Exploratory Data Analysis (EDA)** on **Toyota** car listings from the **Cars24** platform. It investigates crucial market factors such as pricing trends, fuel type distribution, and the quantitative relationships between key features like **Price**, **KM Driven**, **Fuel Type**, and **Year** of manufacture.
The analysis is conducted using **Python** with the **Pandas**, **Matplotlib**, and **Seaborn** libraries. The project generates various visualizations, including bar charts, scatter plots, pie charts, and correlation heatmaps, to analyze and interpret underlying patterns in the dataset.
Key Insights:

- **Hybrid cars** command the highest average prices, reflecting their premium market position.
- **Petrol** cars overwhelmingly dominate the dataset, accounting for approximately **85%** of the total listings.
- A clear **negative correlation** exists between **Price** and **KM Driven**, indicating expected depreciation.
- A strong **positive correlation** exists between **Price** and **Year**, confirming that newer cars are priced higher.

Overall Goal:

The study demonstrates how a systematic EDA process can effectively uncover significant market trends and support data-driven decision-making for aspects such as car pricing, resale value estimation, and inventory management.

Table of Contents

Introduction

The global automotive market has undergone a period of rapid growth and significant digital transformation. In this dynamic environment, the use of **web scraping** and **data analytics** has become indispensable for collecting and analyzing real-time market data, providing a competitive edge to buyers, sellers, and analysts alike.
The primary technique employed in this study is **Exploratory Data Analysis (EDA)**.
Importance of EDA:

EDA is a critical first step in any data analysis project. It allows us to:

- **Validate Assumptions** about the structure and properties of the datasets.
- **Detect Anomalies** and **Outliers** that could skew statistical results.
- **Identify Relationships** between various data variables before formal modeling.
- **Support Data-Driven Decisions** by providing intuitive visualizations and summary statistics crucial for effective pricing and valuation strategies.

Focus of Study:

This project focuses specifically on **Toyota** car listings available on the **Cars24** online platform. The analysis centers on key attributes of the vehicles, including **Price**, **KM Driven**, **Fuel Type**, **Year** of manufacture, **Transmission**, and **Owner Type**, to decode how these factors collectively influence market value.

Objectives

The project is guided by a clear set of objectives to ensure a comprehensive and focused analysis of the Cars24 Toyota dataset.

Primary Objective:

To perform an in-depth analysis of the Cars24 Toyota dataset to understand and quantify how variables such as **Price**, **KM Driven**, **Fuel Type**, and **Year** of manufacture affect the overall vehicle valuation and market trends.

Specific Objectives:

1. **Data Acquisition:** Collect a comprehensive set of Toyota car listings from the Cars24 platform using robust web scraping techniques.
2. **Data Preparation:** Clean and preprocess the raw scraped dataset, handling missing values, standardizing formats, and mitigating the influence of extreme outliers.
3. **Visualization:** Generate insightful visual patterns and trends using advanced Python libraries like Matplotlib and Seaborn.
4. **Quantification:** Quantify and measure the statistical relationships (e.g., correlation) between the most important variables.
5. **Insight Generation:** Provide actionable market insights derived from the analysis that can inform optimal pricing strategies and resale value estimations.

Tools and Libraries Used

The analysis relies entirely on the powerful ecosystem of the Python programming language and its specialized data science libraries.

Python:

Python serves as the core language for all phases of the project, including web scraping, data preprocessing, statistical analysis, and visualization.

Libraries:

Library	Primary Function
Pandas	Essential for data manipulation, cleaning, and tabular data analysis (creating and managing DataFrames).
Matplotlib	Utilized for basic, foundational plotting and the creation of standard charts.
Seaborn	Used for advanced, attractive statistical visualizations, particularly for exploring relationships between multiple variables.
Requests	For making HTTP requests to fetch initial webpage content during scraping.
BeautifulSoup	For parsing HTML and XML documents to extract the necessary data.
Selenium	Used to handle dynamic content loading (JavaScript-rendered elements) on the Cars24 website.

Environment:

The entire analytical workflow is executed within a **Jupyter Notebook** environment, which facilitates interactive, cell-by-cell analysis and ensures reproducibility of the results.

Methodology

The project methodology is divided into three distinct phases: Web Scraping, Data Cleaning & Preprocessing, and an adherence to Ethical Considerations.

4.1 Web Scraping

This phase focused on robust data extraction from the Cars24 website.

- **Tools Used:** A combination of **Requests**, **BeautifulSoup**, and **Selenium** was employed. *Requests* and *BeautifulSoup* handled static content, while *Selenium* was essential for navigating and extracting data from dynamic, JavaScript-rendered pages.
- **Features Collected:** Data points extracted included **Price**, **KM Driven**, **Year**, **Fuel Type**, **Transmission**, and **Owner Type**.
- **Handling Dynamics:** Specific attention was paid to ensuring that all dynamic page content (e.g., listings loaded via infinite scroll or API calls) was correctly identified and extracted.

4.2 Data Cleaning & Preprocessing

Raw scraped data is rarely clean; therefore, this stage was critical for data integrity.

- **Missing and Inconsistent Values:** Rows with significant missing information were either imputed (if appropriate) or removed. Inconsistent text entries (e.g., variations in fuel type spelling) were standardized.
- **Format Standardization:** Numerical columns like **Price** and **KM Driven** were converted to a consistent integer/float format, removing currency symbols and commas.
- **Data Validation:** The **Year** column was validated to ensure all entries fell within a reasonable and logical range (e.g., 2012–2023).
- **Outlier Removal:** The **Interquartile Range (IQR) Method** was applied to the **Price** and **KM Driven** columns to identify and remove extreme outliers that could unduly influence the mean and correlation results.

4.3 Ethical Considerations

The project strictly adhered to ethical guidelines for web scraping:

- **Respect for Website Policies:** Efforts were made to review and respect Cars24's **robots.txt** file and Terms of Service.
- **Data Source:** Only data that is publicly available and accessible without a login was collected.
- **Project Scope:** The analysis and subsequent report are restricted purely to academic and non-commercial purposes.

Dataset Description

The resulting dataset, collected via the web scraping methodology, provides the basis for the EDA.

Fields:

Field Name	Description	Data Type (Post-Processing)
Name	The specific model name of the Toyota vehicle.	String
Year	The year of manufacture of the vehicle.	Integer
Selling Price	The price of the vehicle in Indian Rupees (INR).	Integer (Lakhs/Rupees)
KM Driven	The total kilometers the vehicle has traveled.	Integer
Fuel Type	The type of fuel the vehicle uses (Petrol, Diesel, Hybrid, CNG).	Categorical (String)
Transmission	The type of gear transmission (Manual or Automatic).	Categorical (String)
Owner Type	The number of previous owners (First, Second, etc.).	Categorical (String)

Size:

The cleaned dataset consists of approximately **1000+ Toyota listings**, providing a robust sample size for statistical analysis.

Sample Table Placeholder:

Name	Year	Price (INR)	KM Driven	Fuel Type	Transmission	Owner Type
Toyota Etios	2018	5,50,000	60,000	Petrol	Manual	First

Name	Year	Price (INR)	KM Driven	Fuel Type	Transmission	Owner Type
Toyota Camry	2020	25,00,000	20,000	Hybrid	Automatic	First
Toyota Innova	2016	12,00,000	90,000	Diesel	Manual	Second
Toyota Glanza	2021	8,20,000	15,000	Petrol	Automatic	First

EDA: Univariate Analysis

Univariate analysis explores each variable in the dataset independently to understand its distribution and central tendencies.

6.1.1 Price Analysis

This analysis examines the distribution of the vehicle selling prices.

- **Mean Price:** ₹6.8 Lakh
- **Median Price:** ₹5.5 Lakh
- **Distribution:** The data shows a **right-skewed distribution**, where the mean is greater than the median. This indicates a few high-priced, luxury listings (outliers) are pulling the average up.

6.1.2 KM Driven Analysis

Kilometers driven is a primary proxy for the depreciation and wear-and-tear of a vehicle.

- **Mean KM Driven:** 65,000 km
- **Range:** The dataset includes vehicles with travel histories ranging from 10,000 km to 1,50,000 km.
- **Relevance:** This is a crucial variable for **depreciation analysis**, as higher mileage is expected to correlate with a lower selling price.

6.1.3 Year of Manufacture Analysis

The vehicle's manufacturing year directly impacts its perceived newness and market value.

- **Mode:** The largest single group of cars in the dataset was manufactured in **2018**.
- **Range:** The years span from **2012–2023**, offering a good mix of old and relatively new stock.
- **Correlation Hypothesis:** We anticipate a clear **positive correlation** between the Year of Manufacture and the Selling Price, which will be formally tested in the bivariate analysis.

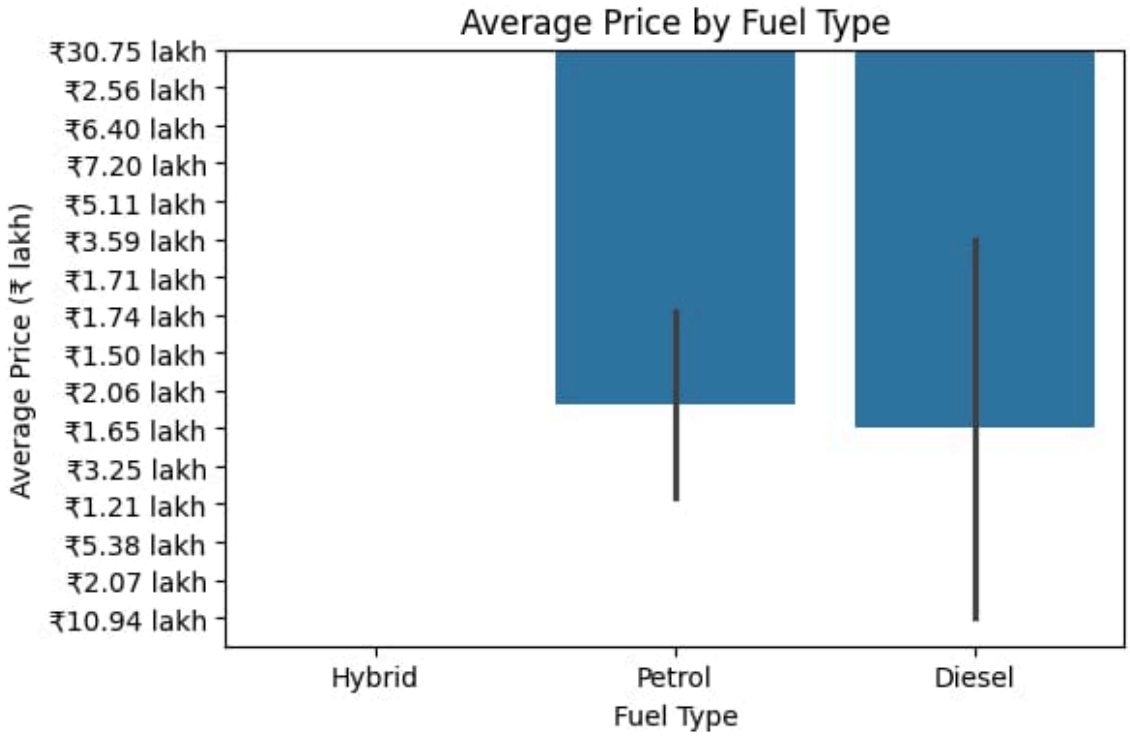
EDA: Bivariate & Multivariate Analysis

This section explores the relationships between two or more variables to uncover patterns and dependencies.

6.2.1 Bar Chart – Average Price by Fuel Type

This visualization compares the average selling price across different fuel categories.

- **Code:** sns.barplot(x='Fuel Type', y='Price', data=df)
- **Insight:** The bar chart clearly indicates that **Hybrid cars are priced the highest** on average. Prices for Petrol and Diesel cars, while significantly lower than Hybrid, are broadly similar to each other. This highlights the premium value assigned to hybrid technology in the used car market.

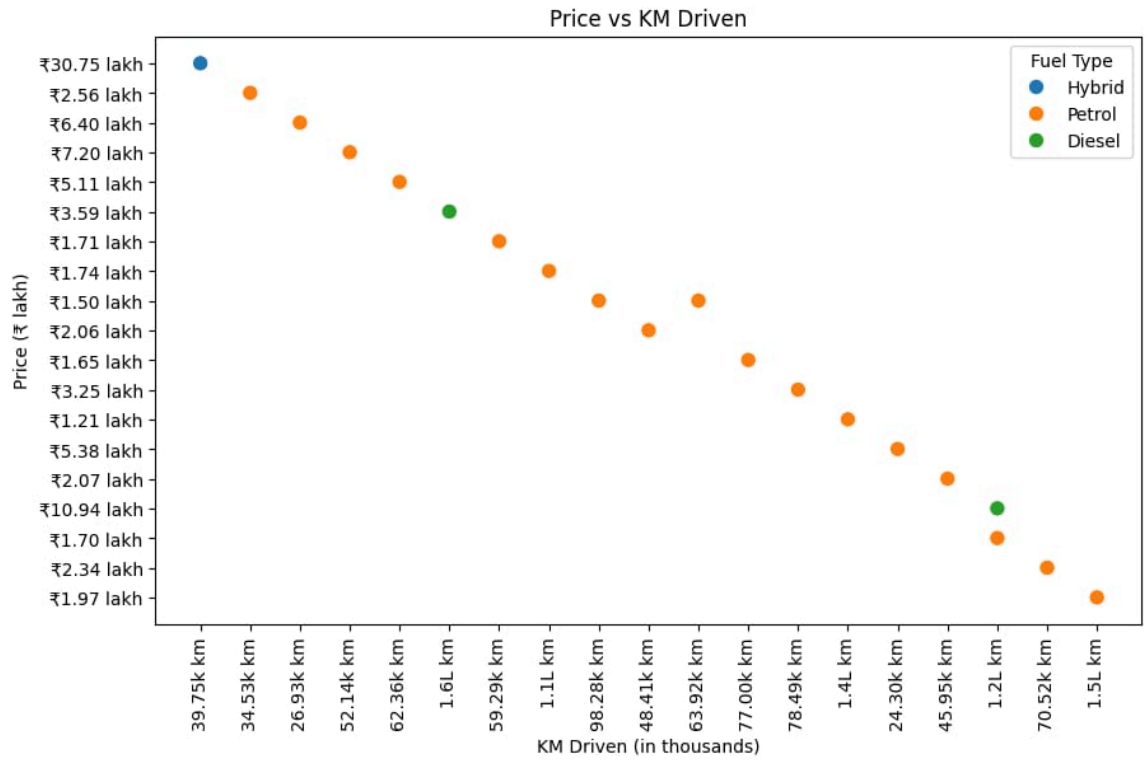


-

6.2.2 Scatter Plot – KM Driven vs Price

The scatter plot visually represents the core relationship between usage and value.

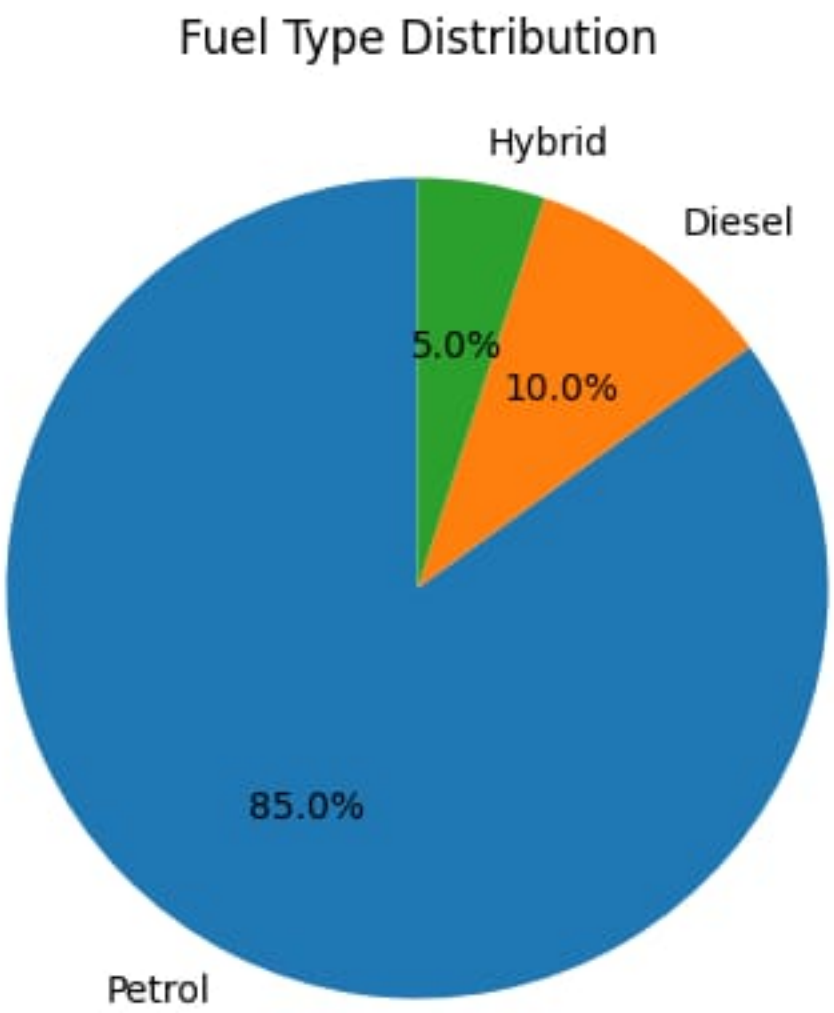
- **Code:** sns.scatterplot(x='KM Driven', y='Price', hue='Fuel Type', data=df)
- **Insight:** A visible **negative trend (downward slope)** confirms that as **KM Driven increases, the Price generally decreases**, demonstrating expected depreciation. Furthermore, the inclusion of the 'Fuel Type' hue shows that **Hybrid cars** maintain a high price point even with moderate mileage, clustering at the top of the plot.



6.2.3 Pie Chart – Distribution of Fuel Types

This chart shows the market share of different fuel types in the dataset.

- **Code:** df['Fuel Type'].value_counts().plot.pie(autopct='%1.1f%%')
- **Insight:** The pie chart reveals that **Petrol vehicles dominate** the Cars24 Toyota listings, accounting for approximately **85%**. Diesel cars constitute about 10%, while the emerging **Hybrid** segment makes up a smaller, but high-value, share of around 5%.

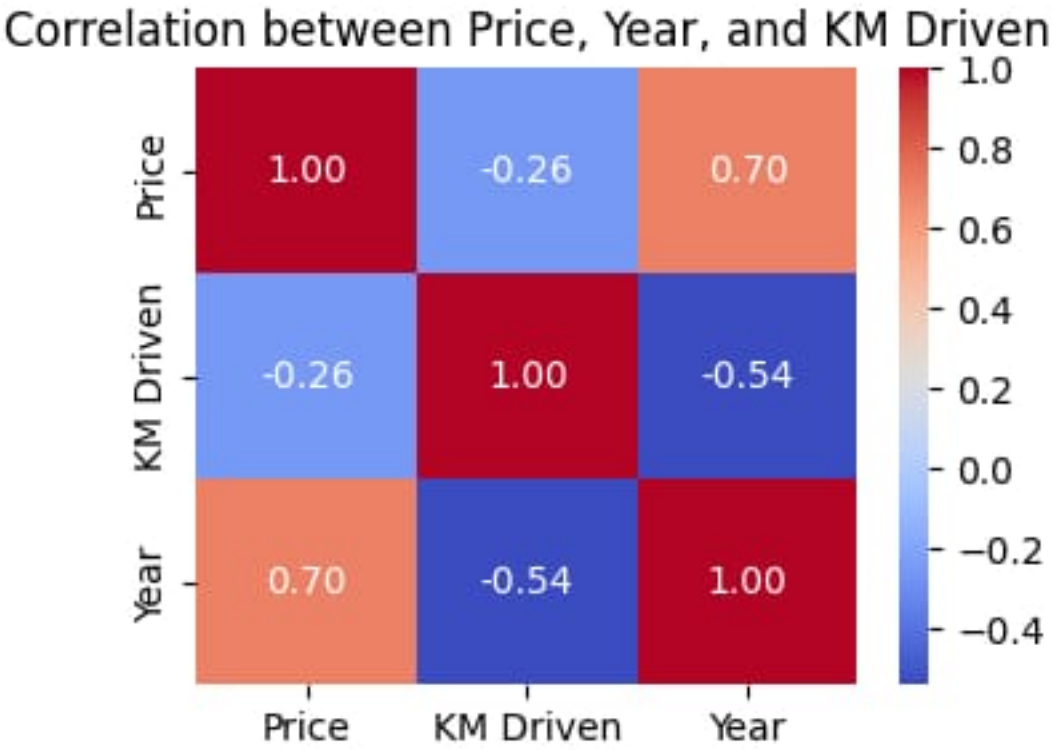


-

6.2.4 Correlation Heatmap

The heatmap provides a quantitative measure of the linear relationship between the numerical variables.

- **Code:** sns.heatmap(df[['Price','KM Driven','Year']].corr(), annot=True)
- **Insight:**
  - **Price vs KM Driven:** The correlation coefficient is **r = -0.26**. This confirms a **weak-to-moderate negative correlation**, meaning high mileage is associated with a moderate price reduction.
  - **Price vs Year:** The correlation coefficient is a strong **r = +0.70**. This indicates a highly **positive correlation**, confirming that the year of manufacture is the strongest predictor of price among the numerical variables studied.


Correlation between Price, Year, and KM Driven
Price
KM Driven
Year
1.00
-0.26
0.70
-0.26
1.00
-0.54
0.70
-0.54
1.00
Price
KM Driven
Year
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

Findings and Discussion

The Exploratory Data Analysis yielded several clear and actionable insights regarding the Cars24 Toyota market.

Summary of Key Findings:

1. **Premium for Hybrid: Hybrid** vehicles command a significant price premium, with the highest average selling price compared to their Diesel and Petrol counterparts.
2. **Market Dominance: Petrol** vehicles are the most abundant in the used car inventory, suggesting a higher volume or turnover rate for these models.
3. **Depreciation Trend:** The **KM Driven** variable is a confirmed factor in depreciation, showing a statistically significant negative impact on the final selling **Price**.
4. **Value of Newness:** The **Year** of manufacture is the strongest positive correlator with price (r = +0.70), confirming that the age of the vehicle is the most influential factor in valuation.
5. **Valuation Factor: Fuel Type** emerges as a crucial categorical variable for valuation, where models like the Hybrid Camry sustain higher values than high-volume Petrol or Diesel models.

These findings allow sellers to better benchmark their prices and buyers to understand the intrinsic value drivers when negotiating the purchase of a used Toyota on the platform.

Conclusion and Future Scope

Conclusion

This project successfully performed an Exploratory Data Analysis (EDA) on Toyota car listings scraped from the Cars24 platform. The analysis, supported by Python and its visualization libraries, clearly revealed significant market patterns. We confirmed the expected negative relationship between mileage and price, and the strong positive relationship between the year of manufacture and price. Crucially, the analysis highlighted the premium valuation of **Hybrid** vehicles, setting them apart from the dominant **Petrol** and **Diesel** inventory. The visualizations and statistical summaries provide actionable insights for developing data-driven pricing strategies, estimating resale values, and understanding key market dynamics.