# Used Cars Price Analysis – Final Project Report

## 🔍 Objective

The primary goal of this project is to extract and analyze data from online used car marketplace such as Car Dekho. By leveraging web scraping and data analysis techniques, the project aims to:  
Understand pricing patterns based on car attributes

* Understand pricing patterns based on car attributes
* Identify value-for-money brands and models
* Assist potential buyers in making informed decisions
* Explore which factors (e.g., mileage, fuel type, brand) have a significant impact on used car prices

This data-driven approach helps overcome the uncertainty faced by buyers in the second-hand automobile market.

## 🛠 Tools & Technologies Used

To execute the various stages of the project, the following tools and technologies were used:

* **Python**: Programming language used for automating web scraping and preprocessing tasks
* **BeautifulSoup**: Python library used for parsing and navigating HTML content efficiently
* **requests**: Used to send HTTP requests and fetch web pages
* **re (Regular Expressions)**: Utilized for pattern matching, particularly to extract the year from car names
* **openpyxl**: Used to create and manipulate Excel files programmatically
* **Excel**: For initial inspection, manual cleaning, and exploratory analysis
* **Power BI**: A business intelligence tool for creating interactive dashboards and visualizations to derive insights

## ✅ Step-by-Step Project Workflow

## 1. Data Collection (Web Scraping)

The first step in the project was collecting real-world data from the CarDekho used car listings page. This was achieved by writing a Python script to scrape essential details of each listed vehicle.  
  
**Key Features Extracted:**

* **Year**: Manufacturing year of the vehicle, extracted using a regular expression
* **Name**: Car make and model name
* **Price**: Listed price (e.g., ₹4.5 Lakhs)
* **Mileage**: Fuel efficiency in kmpl or km/kg
* **Fuel Type**: Petrol, Diesel, CNG, etc.
* **Transmission**: Manual or Automatic

**Script Explanation:**

1. **Library Import**: The script begins by importing required libraries: BeautifulSoup, requests, openpyxl, and re.
2. **Workbook Creation**: An Excel workbook is created using openpyxl, and headers are added.
3. **User-Agent Header**: To mimic a browser and avoid bot detection, a User-Agent string is included in the HTTP request.
4. **Request and Parsing**: The HTML content of the used car page is fetched using requests and parsed using BeautifulSoup.
5. **Car Listing Extraction**: Specific HTML classes and IDs are used to navigate through the DOM and locate each car listing.
6. **Data Extraction and Cleaning**:
   1. Car year is extracted using regex.
   2. Name is cleaned by removing the year substring.
   3. Mileage, fuel type, and transmission are split from a combined string.
7. **Data Storage**: Cleaned data is appended row-by-row to the Excel sheet.
8. **File Output**: The Excel file is saved locally as Hema\_project.xlsx.

This script ensures the collection of structured, readable, and analysable data from an otherwise unstructured web page.

.

## 2. Data Cleaning & Transformation

Once the data is scraped, it undergoes a thorough cleaning process:

* **Null and Duplicate Removal**: All missing values and duplicate records are dropped.
* **Format Standardization**:
  + Price values are converted from string format (e.g., ₹4.5 Lakhs) to numeric.
  + Mileage values are unified with consistent units (e.g., converting km/kg to kmpl where necessary).
* **String Cleanup**: Special characters and unnecessary white spaces are removed.
* **Brand Normalization**: Brand names are standardized to avoid variations (e.g., "Hyundai" vs "HYUNDAI").

This step ensures that the dataset is clean, consistent, and ready for meaningful analysis.

## 3. Data Analysis in Power BI

After cleaning, the data is imported into **Power BI** where an interactive dashboard is built. The dashboard helps visualize patterns and trends.

**Key Visuals:**

* **Brand vs. Average Price**: A bar chart showing the average price of cars by brand. It helps identify expensive vs. budget-friendly brands.
* **Fuel Type Distribution**: A pie chart or bar graph to analyze the proportion of petrol, diesel, and other fuel types in the market.
* **Number of Cars per Brand**: Identifies the most listed car brands, indicating popularity and availability.
* **EMI vs. Price Comparison**: Illustrates the correlation between a car's price and its potential monthly instalment, helping assess affordability.

The dashboard is dynamic, allowing filters by brand, year, transmission, and fuel type.

## 4. Insights & Observations

* **Value for Money**: Brands like **Maruti** and **Hyundai** consistently offer high mileage at lower prices, making them suitable for budget-conscious buyers.
* **Fuel Trends**: Despite increasing interest in electric vehicles, **diesel** and **petrol** cars still dominate the used market.
* **Transmission Preference**: Manual cars are far more common, reflecting consumer preferences and price affordability.
* **Price-EMI Relationship**: A direct correlation exists, where higher car prices naturally result in higher monthly EMIs.

These insights can guide dealerships and buyers in evaluating used car options more strategically.

## 5. Challenges Faced

|  |  |  |
| --- | --- | --- |
| Challenge | Description | Solution |
| Inconsistent HTML tags | Nested and varying classes  in the website's DOM | Used targeted tag and class navigation, and try-except  blocks |
| Unavailable or Missing Data | Some listings lacked fields  like mileage or transmission | Defaulted to "N/A" using  validation |
| Extracting Year | Year embedded in the car  name | Used regex to isolate year |
| Dynamic Elements | Some content loaded dynamically | Focused on static HTML  sections |

## 6. Conclusion

The project effectively demonstrates how a combination of web scraping, data cleaning, and visualization can provide actionable insights into the used car market. Key takeaways include:

* Identifying affordable and reliable car brands
* Understanding fuel and transmission trends
* Evaluating pricing structures using real data

This data-centric approach can empower buyers, dealers, and financial institutions alike to make smarter, more confident decisions.

## ✨ Future Scope

To enhance the project further, the following improvements are proposed:

* **City-Level Analysis**: Introduce geographic filters to compare pricing trends across locations.
* **Predictive Modelling**: Use machine learning algorithms to predict car prices based on features.
* **Automated Data Refresh**: Schedule scraping scripts to run weekly and update the dashboard automatically.
* **Integration with APIs**: Incorporate real-time data from APIs for more accurate and updated listings.

## 📅 Project Duration

* **Start Date:** 17th March 2025
* **End Date:** 16th April 2025

This timeline includes data collection, cleaning, dashboard building, and documentation preparation.

## 📄 Submitted Deliverables

- Hema\_project.xlsx  
- Hema Project Cleaned\_data.xlsx  
- Hema Project.pbix  
- Hema\_project.pptx

## 💼 Submitted To

Company: Analytics Space  
Mentor: Chethan V S, Director

## ✨ Thank You!

This project reflects the integration of coding, analytical thinking, and business storytelling. It provides a comprehensive solution for understanding and navigating the used car market effectively.