

# Scraping Data Using Python

Harness the power of Python to extract valuable data from the web, transforming it into actionable insights. Embark on a journey of data-driven discovery and uncover hidden opportunities.



**Evoastra Ventutre Mini Project**

**Internship Presentation**

**By Team A**



# Team Leads

**Sachin Sharma**

Team Lead

**Nidhi Gummaraju**

Team Lead

**Muhammad Umar**

Co - Team Lead

**Santosh Shinde**

Co - Team Lead

# Team Members

**Mohit**

**Syed Musa**

**Solanki**

**Akansha**

**Srinivas**

**Dixit**

**Akshaya**

**Sahil**

**Hannah**

**Pranay**

**Rikesh**

**Mahesh**

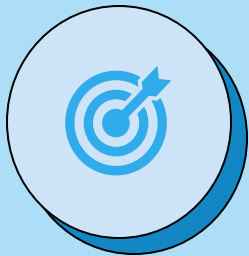
**Vignesh**

**Sandeep**

**Raval**



# Content



INTRODUCTION AND  
OBJECTIVE



DATA COLLECTION  
AND WEB SCRAPING



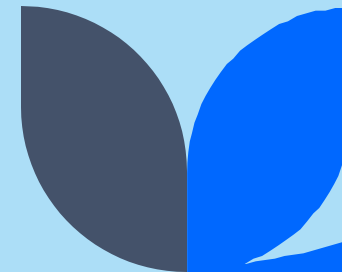
DATA PROCESSING  
AND CLEANING



DATA ANALYSIS AND  
INSIGHTS

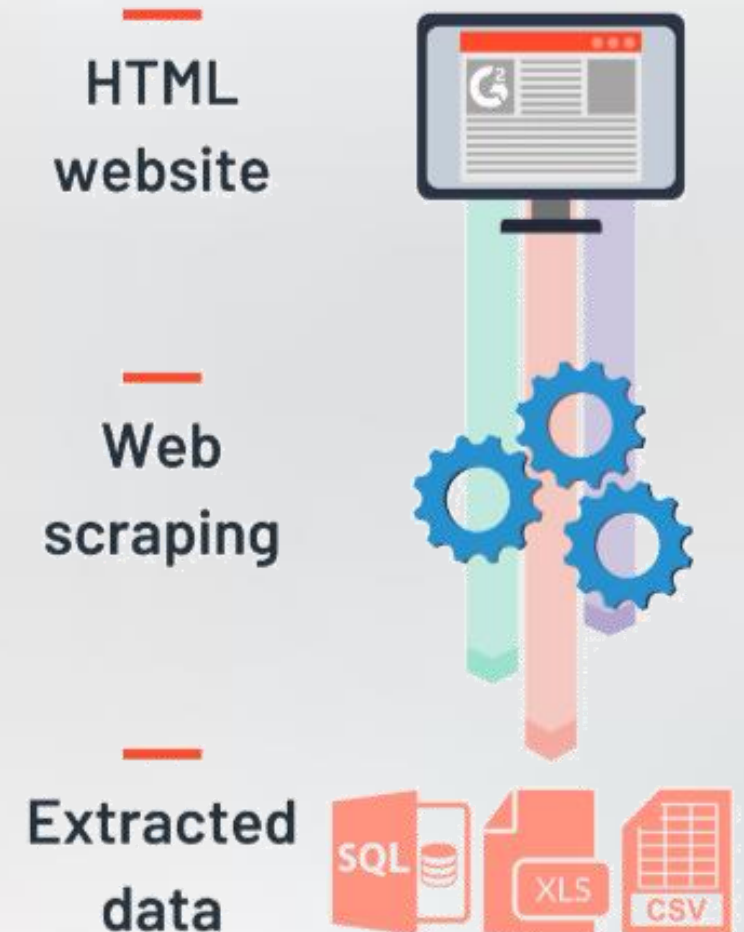


CHALLENGES AND  
CONCLUSION



# Objectives

Our objectives are to collect and scrap the data on cars listed for sale in Mumbai from Cars24.com, ensuring data accuracy through cleaning and processing and providing valuable information for potential buyers and sellers.



# Research and Planning

1

## Identify Data Sources

Determine the websites and web pages that contain the data essential to the project's objectives.

2

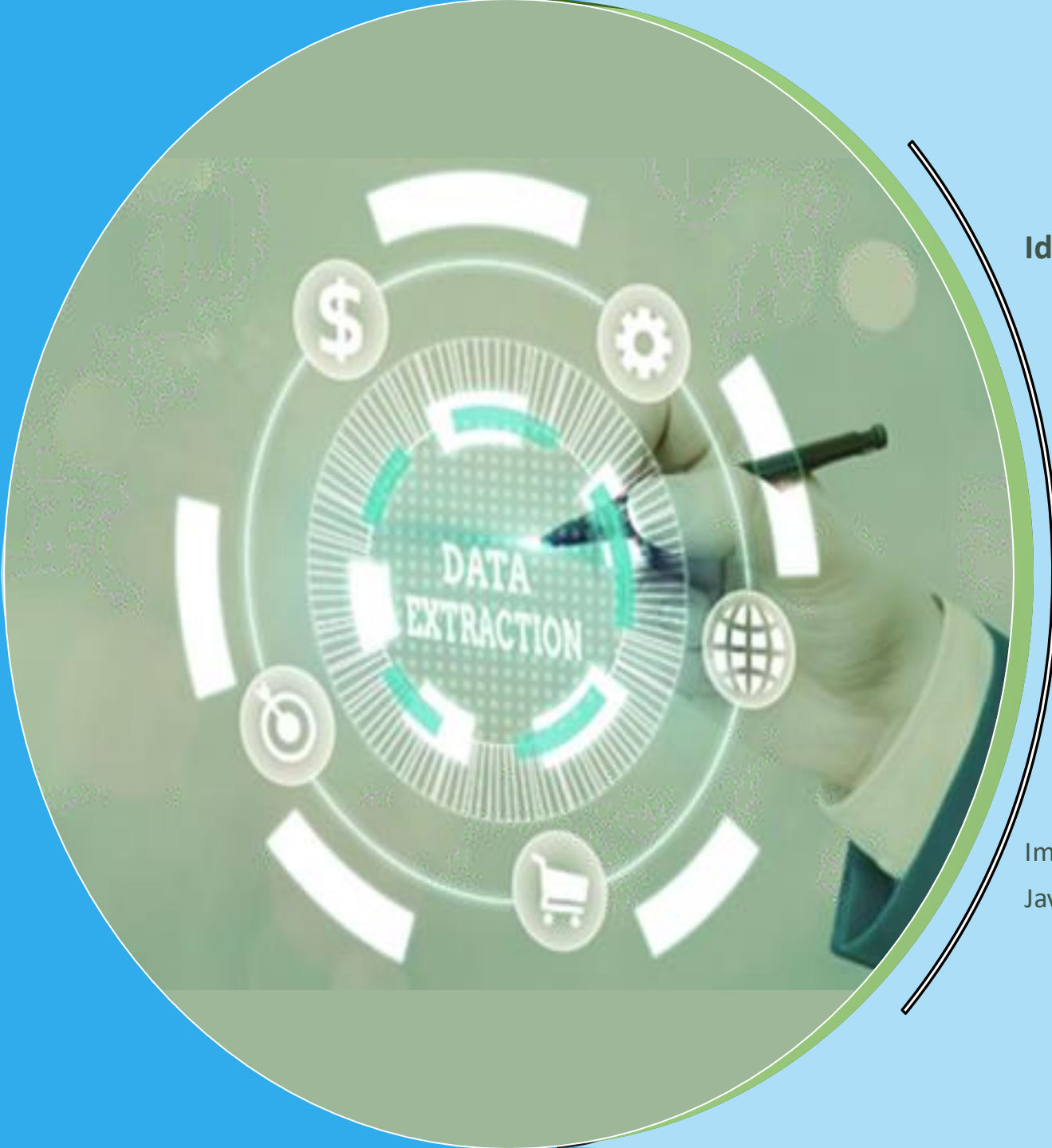
## Outline Scraping Strategies

Develop a comprehensive plan for navigating the target websites and extracting the required data.

3

## Anticipate Challenges

Proactively consider potential obstacles, such as website changes or changes or rate limiting, and devise contingency plans.



## Identify HTML Structure

Analyze the target web pages to understand their HTML structure and locate the data elements to be extracted.

## Develop Scraping Scripts

Write Python code to automate the process of navigating the websites and extracting the desired data.

## Handle Dynamic Content

Implement techniques to capture data from websites with JavaScript-driven content or pagination.

# Data Cleaning and Organization

## DATA CLEANING CHECKLIST

**Up-to-date data**

Data should be up-to-date. In order to obtain maximum value from the data analysis.



**Missing values**

Count missing values and analyze where in the data they are missing. Missing values can disrupt some analyses and skew the results.




**Duplicates**

Duplicate IDs indicate multiple records for one person, e.g. someone holds multiple functions at the same time.




**Numerical outliers**

Numerical outliers are fairly easy to detect and remove. Define minimum and maximum to spot outliers easily.



**Check IDs**

Check data labels of all the fields to see whether some categorical values are mislabeled.



**Define valid output**

Define valid data labels for categorical data. Define data ranges for numerical variables. Non-matching data is presumably wrong.



### Eliminate Duplicates

Implement deduplication algorithms to ensure the dataset is free from redundant information.

### Handle Missing Data

Fill in gaps or replace missing values using appropriate techniques, such as imputation or interpolation.

### Standardize Data Format

Normalize the data structure and ensure consistent formatting across different data sources.

### Organize for Analysis

Store the cleaned and structured data in a format suitable for subsequent data analysis and analysis and visualization.

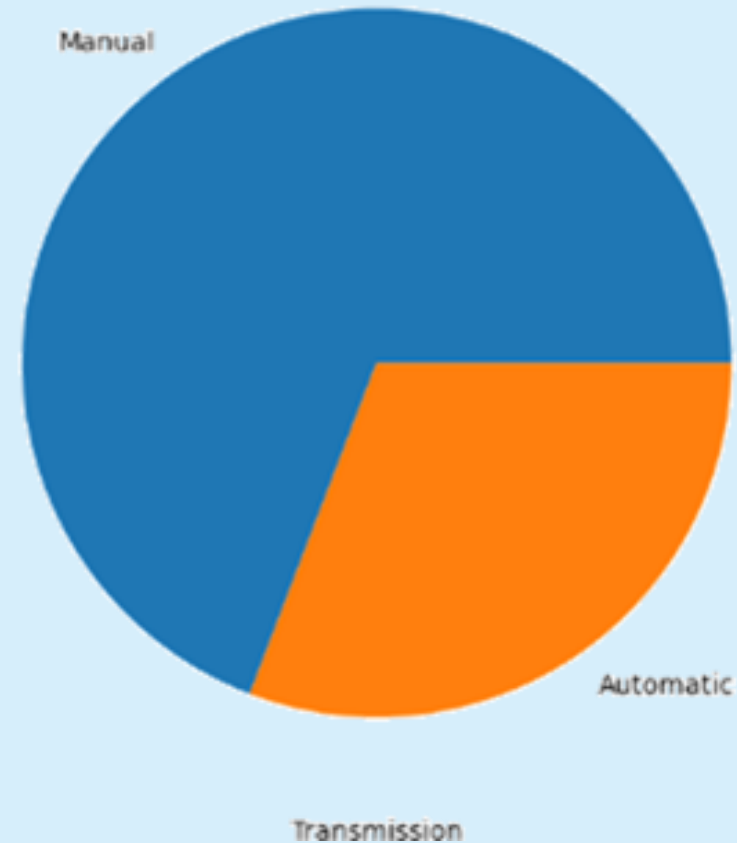


# Data Analysis 01

## Bar Chart of Car Brands:

- Data analysis involves examining data to uncover patterns, draw conclusions, and make decisions. It uses statistical and computational methods to extract meaningful insights from raw data
- **Description:** This bar chart illustrates the distribution of different car brands within the dataset. Maruti and Hyundai are the most prominent brands, indicating their dominance in the Mumbai automotive market. Other brands like Honda, Tata, and Renault have a significantly lower presence.

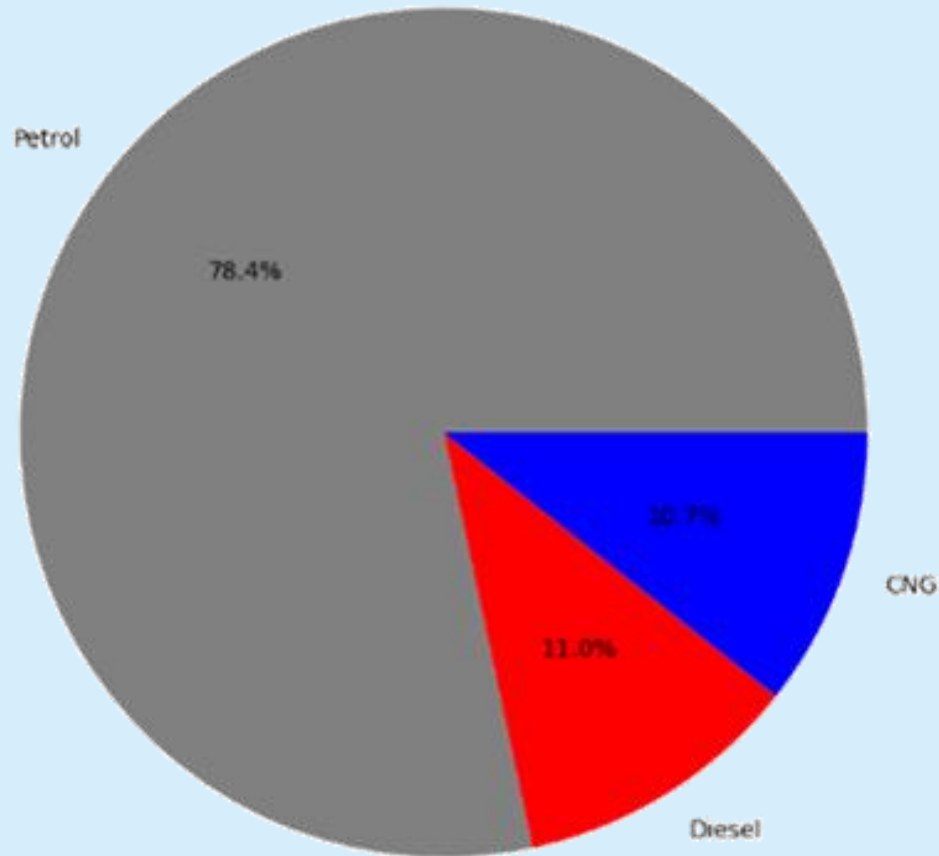
Bar Chart of Transmission Types



# Data Analysis 03

## Pie Chart of Fuel Types

Pie Chart of Fuel Types



### Dominance of Petrol Cars:

- **Observation:** Petrol-powered vehicles account for a significant 78.4% of the total car listings.
- **Interpretation:** This dominance suggests that petrol is the preferred fuel type among consumers in Mumbai, possibly due to its widespread availability and infrastructure.

### Presence of Diesel Cars:

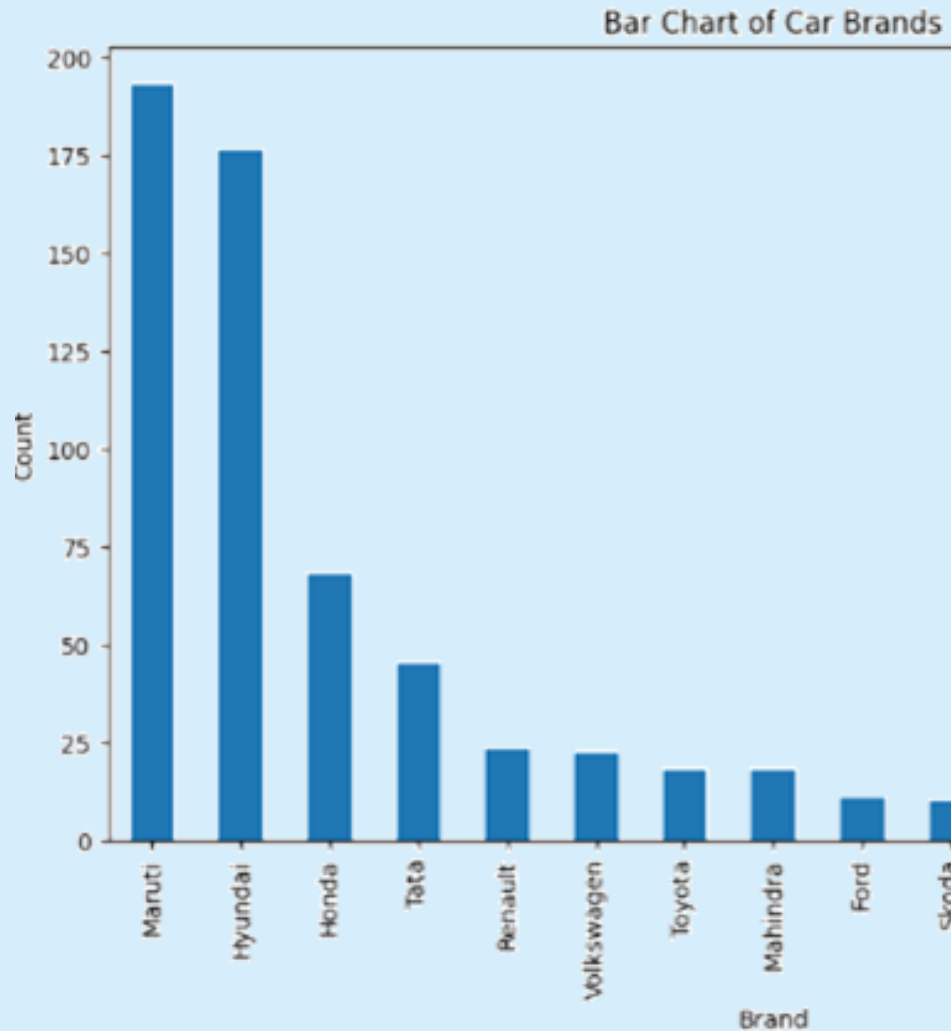
- **Observation:** Diesel vehicles make up 11.0% of the listings.
- **Interpretation:** While significantly less than petrol, diesel cars still hold a noticeable portion of the market. This could be due to diesel's fuel efficiency and lower running costs for long-distance travel.

### CNG Vehicles:

- **Observation:** CNG (Compressed Natural Gas) cars constitute 10.7% of the market.
- **Interpretation:** The presence of CNG vehicles reflects a growing trend toward more eco-friendly and cost-effective alternatives to petrol and diesel. The adoption of CNG may be driven by governmental incentives and the increasing number of CNG refueling stations in Mumbai.

# Data Analysis 02

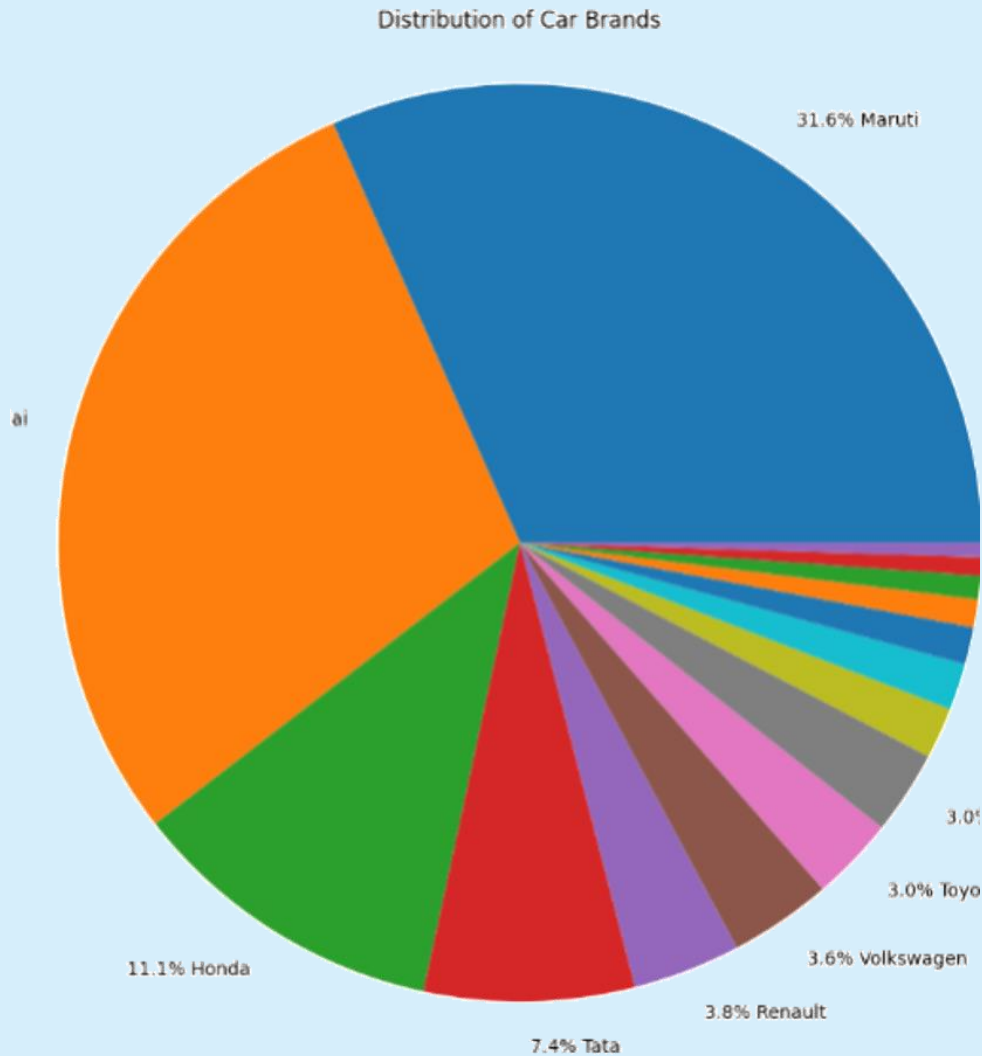
Bar Chart of Transmission Types:



**Description:** The bar chart shows the proportion of cars with manual and automatic transmission types. The majority of cars in the dataset have manual transmissions, suggesting that manual vehicles are more commonly listed or preferred in Mumbai.

# Data Analysis 04

## Pie Chart of Car Brand Distribution



### Maruti's Market Share:

- **Observation:** Maruti leads the market with a 31.6% share.
- **Interpretation:** Maruti's dominance highlights its strong brand presence and consumer trust in Mumbai. Known for its affordability and extensive service network, Maruti has become a go-to choice for many car buyers.

### Significant Players (Honda, Tata, Hyundai):

- **Observation:** Honda, Tata, and Hyundai follow with shares of 11.1%, 7.4%, and 11.1%, respectively.
- **Interpretation:** These brands represent significant alternatives to Maruti, each catering to different segments of the market. Honda and Hyundai offer a mix of premium and economy vehicles, while Tata is known for its durability and value-for-money offerings.

### Smaller Brand Segments:

- **Observation:** The remaining brands, such as Renault, Volkswagen, and Ford, each occupy smaller segments of the market.
- **Interpretation:** These smaller shares indicate that while these brands are present, they have not captured a large portion of the market, possibly due to higher pricing, limited service networks, or lower consumer awareness.

# Challenges and Solutions

## Dynamic Content:

Websites that load content dynamically using JavaScript may require Selenium to simulate user interactions and access the data.

## Website Structure Changes:

Websites often undergo changes in HTML structure or class names, requiring updates to your scraping code.

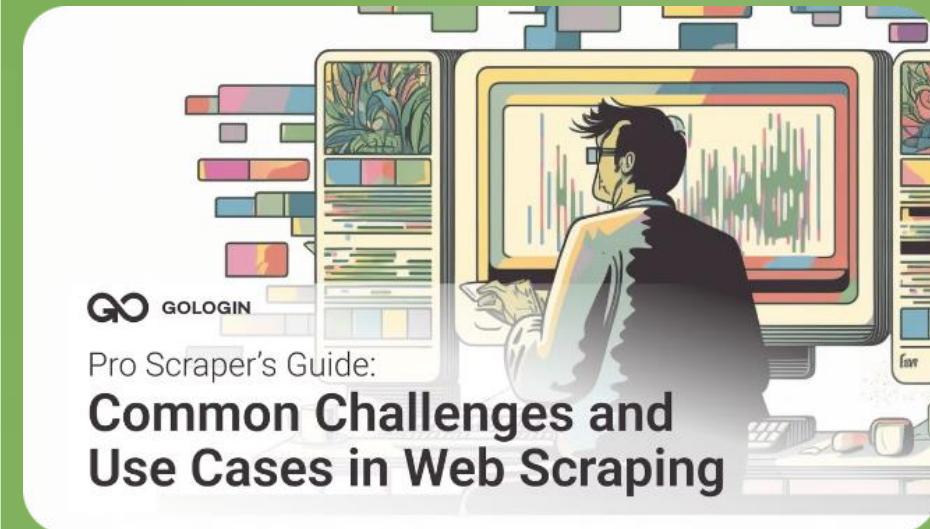
## Handling Cookies and Sessions:

Some websites use cookies or sessions to manage user interactions, requiring Selenium to maintain state across requests.

**Data Quality:** Ensuring scraped data is accurate and reliable requires thorough validation and cleaning to handle errors, missing data, or inconsistencies.

## Performance:

Scraping large amounts of data can be slow, especially when using Selenium for complex interactions or when handling large datasets.



# Conclusion

- **Successful Data Collection:** Scraped detailed information on any cars in Mumbai From Cars24.com Using Python libraries like BeautifulSoup and Selenium
- **Key Insights:**
  - Most Popular models identified.
  - Average pricing trend analyzed.
  - Mileage ranges , fuel type and transmission preferences determined.
- **Challenges Overcome:** Managed dynamic web content and ensured data accuracy.
- **Value :** Provides valuable insights for buyers and sellers in the Mumbai any car market

# THANK YOU



**Evoastra Ventutre Mini Project**

**Internship Presentation**

**By Team A**

