Web Scraping Project Report:

Hyundai Cars from Cars24.com

Team B Members Vinayak Ramavath, Bhukya Raju, Nikhil Bhoi, K .Bhargav Sai, Diya Singh, Muhammed Sahal, Racha Navya Sree, Hemanth, Akash Kumar, R Sireesha, Tanvi Desai, Shubham

1. Objective

To automatically extract and clean data for used Hyundai vehicles from Cars24.com, collecting key attributes—including model name, kilometers driven, year, fuel type, transmission, price, and location—for market analysis.

2. Tools & Technologies

- Python with Selenium WebDriver to automate a browser and handle dynamic JavaScript content.
- BeautifulSoup for parsing the fully loaded HTML.
- Pandas for data structuring, cleaning, and exporting to CSV.

3. Methodology

- Dynamic Scraping: Selenium was used to simulate a real user, navigating to search results, automatically scrolling to load all listings, and iterating through pagination.
- Data Cleaning: Raw text was processed to remove units and symbols. Key conversions included:
  - "82,657 km" → Integer 82657
  - "₹ 6.5 Lakh" → Float 650000.0
- CSV Export: Cleaned data was compiled into a structured Pandas DataFrame and saved for analysis.

4. Key Challenges & Solutions

- Challenge: Dynamic Content Loading
  - Solution: Used Selenium to automate a Chrome browser, programmatically scrolling and waiting for all elements to load before extraction.
- Challenge: Inconsistent Data Formats
  - Problem: Raw text contained units and was inconsistently formatted (e.g., "3.55 Lakh", "1,03,048 km").
  - Solution: Built a robust parsing pipeline using string operations and regular expressions to isolate and standardize numerical values.

5.Visualization :

i)Manual vs. Automatic Transmission Count

This bar chart illustrates the number of cars available for each type of transmission.

Explanation & Insights:

- **The plot clearly shows that manual transmission cars are far more common in this dataset than automatic transmission cars.**
- **This suggests that the used Hyundai car market in this region is dominated by manual models, which could be due to factors like lower initial cost, better fuel efficiency perceptions, or consumer preference.**

ii)Fuel Type Distribution

This bar chart displays the distribution of cars based on their fuel type.

Explanation & Insights:

- **Petrol is the most prevalent fuel type, with a significantly higher count than both CNG and Diesel.**
- **This indicates a strong market preference or availability for petrol-powered Hyundai cars over other fuel options in the used car segment.**

iii)Distribution of Kilometers Driven by Transmission Type

This violin plot compares the distribution of kilometers driven for both manual and automatic transmission cars. The width of each "violin" represents the frequency of cars at that mileage.

Explanation & Insights

- **Both manual and automatic cars show a wide range of kilometers driven, but the distribution for manual cars is much wider, indicating a greater variety in their usage history.**
- **The densest part of the distribution for both types appears to be below 100,000 km, which is typical for the used car market.**
- **The plot helps visualize the density and spread of mileage for each category.**

iv)Average Kilometers Driven by Location

This bar chart shows the average kilometers driven for cars listed in different locations.

Explanation & Insights:

- **The chart highlights significant variations in the average mileage of cars based on their listed location.**
- **Locations like "Vasai West Tarkhad" and "Kandivali West Mumbai Maharashtra" have cars with higher average kilometers driven, suggesting that cars from these areas might have been used more for longer commutes.**
- **This visualization is useful for potential buyers to understand typical usage patterns in different parts of the city.**

- ***Note: The non-English labels are rendered as they appear in the source data.***

v)Scatter Plot: Kilometers Driven vs. Transmission Type

This scatter plot visualizes the relationship between the type of transmission and the kilometers driven for each car in the dataset. Jitter (a small random value) was added to the transmission axis to prevent points from overlapping and make the distribution clearer.

Explanation & Insights:

- **This plot confirms that there are many more data points for manual cars than for automatic ones.**
- **It also shows that cars with very high mileage (over 150,000 km) are predominantly manual transmission vehicles.**
- **This might suggest that manual cars are either driven for longer distances or are kept in service for a longer period.**

vi)Heatmap of Transmission vs. Fuel Type

This heatmap shows the joint distribution of car counts based on both their transmission and fuel type. Each cell's color and number represent the quantity of cars with that specific combination.

Explanation & Insights:

- **The most common combination of vehicle in the dataset is a petrol-powered manual transmission car, as indicated by the high number in that cell.**
- **Diesel cars are also predominantly manual.**
- **There are very few CNG cars with automatic transmission, which could point to their rarity in the market.**
- **This chart is excellent for understanding the interplay between these two categorical features.**

vii)Number of Cars by Year of Manufacture

This line plot shows the trend in the number of available used cars according to their year of manufacture.

Explanation & Insights:

- **The plot shows a clear trend: the number of available used cars is higher for more recent manufacturing years.**
- **There's a noticeable peak around 2017-2019, suggesting that cars from this period are currently very common in the resale market, possibly as they come off their initial ownership or lease periods.**

The decline in later years is expected as newer cars are less likely to be sold

6.Conclusion:

The Team Hyundai car data scraping project successfully achieved its objectives of extracting comprehensive automotive data from Cars24.com. Despite encountering various technical challenges, the team implemented effective solutions that demonstrated adaptability and problem-solving skills. The project provided valuable hands-on experience in web scraping, data cleaning, and analysis while maintaining ethical standards and technical excellence. The extracted data provides a solid foundation for further analysis of the used car market, specifically for Hyundai vehicles, and the methodologies developed can be applied to similar data extraction projects. The experience gained will be invaluable for future data science endeavors.