EXPLORATORY DATA ANALYSIS

By S SREEJITH

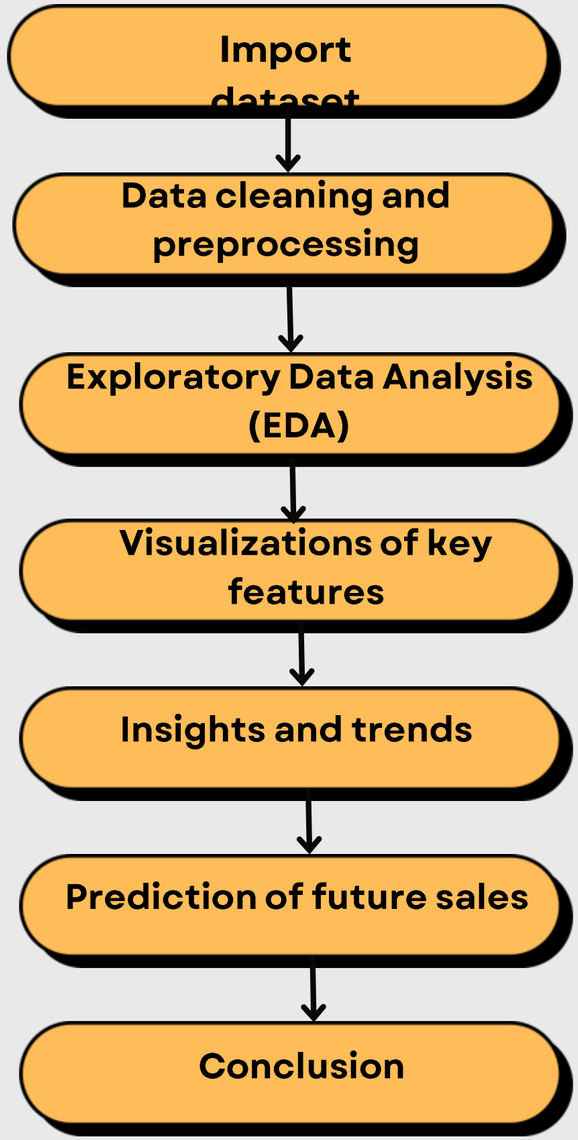
INTRODUCTION

- This presentation explains car sales data analysis and prediction.
- The dataset contains details of cars.
- It includes the company name, type of vehicle, number of sales, price, engine size and power.
- The goal is to understand sales and also predict future sales.

PROBLEM STATEMENT

- The automobile market has a wide variety of cars.
- Customers want the best balance between price, performance, and features.
- Companies need to know:
    - Which vehicles sell the most?
    - How do price and engine specifications affect sales?
    - Which manufacturers lead the market?

FLOW CHART OF WORK



TOOLS AND DATASET USED

Tools Used

- Python programming language
- Pandas library for data handling
- Matplotlib library for data visualization
- Google Colab for implementation

Dataset

- Dataset: Car Sales Dataset
- Source: Kaggle

DATASET OVERVIEW

- The dataset was taken from Kaggle (Car Sales by Gagandeep16).
- It contains information such as:

```
print(df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 157 entries, 0 to 156
Data columns (total 16 columns):
 #   Column               Non-Null Count  Dtype  
---  ------               --------------  -----  
 0   Manufacturer         157 non-null    object 
 1   Model                157 non-null    object 
 2   Sales_in_thousands   157 non-null    float64
 3   __year_resale_value  121 non-null    float64
 4   Vehicle_type         157 non-null    object 
 5   Price_in_thousands   155 non-null    float64
 6   Engine_size          156 non-null    float64
 7   Horsepower           156 non-null    float64
 8   Wheelbase            156 non-null    float64
 9   Width                156 non-null    float64
 10  Length               156 non-null    float64
 11  Curb_weight          155 non-null    float64
 12  Fuel_capacity        156 non-null    float64
 13  Fuel_efficiency      154 non-null    float64
 14  Latest_Launch        157 non-null    object 
 15  Power_perf_factor    155 non-null    float64
dtypes: float64(12), object(4)
```

ROLE OF TOOLS IN EDA

Pandas

- Loaded the dataset into a DataFrame
- Displayed the first five rows to get an overview
- Checked dataset info (column names, data types, non-null values)
- Handled missing values by filling them with zero
- Checked dataset shape (rows and columns)
- Generated statistical summary of data
- Counted category values such as vehicle type

Matplotlib

- Created visualizations such as histograms, scatter plots, and bar charts
- Added grid lines, titles, and labels for better readability
- Helped compare values clearly across categories

IMPLEMENTATION

Step 1:

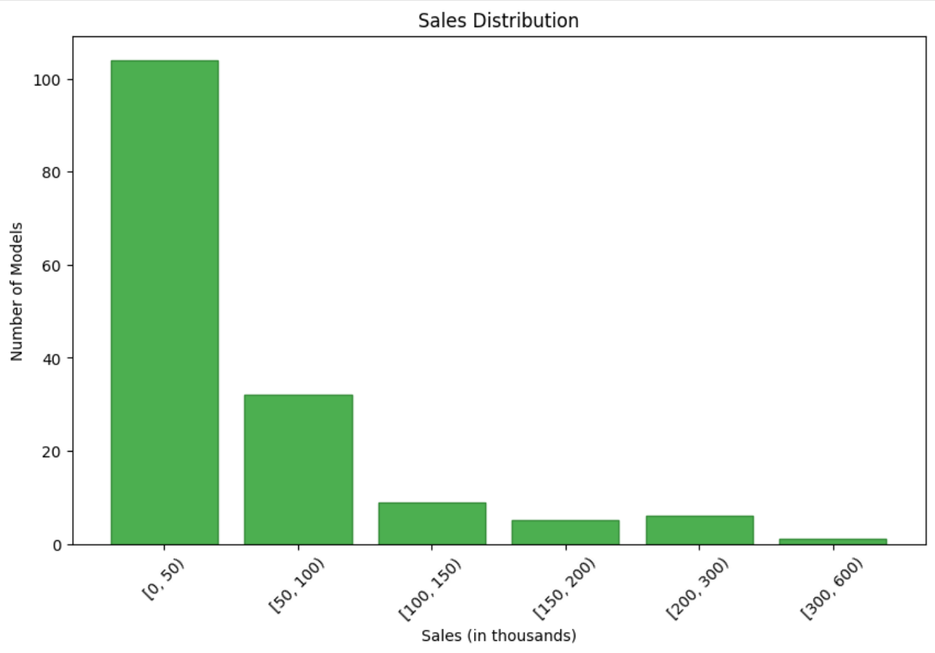
- Removed missing values where necessary.
- Checked datatypes of each column.
- Converted numeric data into usable format.
- Verified dataset structure before analysis.

Step2:

- Distribution of sales and prices.
- Count of vehicle types.
- Scatter plots for price vs horsepower, engine size vs price.
- Top manufacturers by sales.
- Average sales, price, and horsepower by vehicle type.

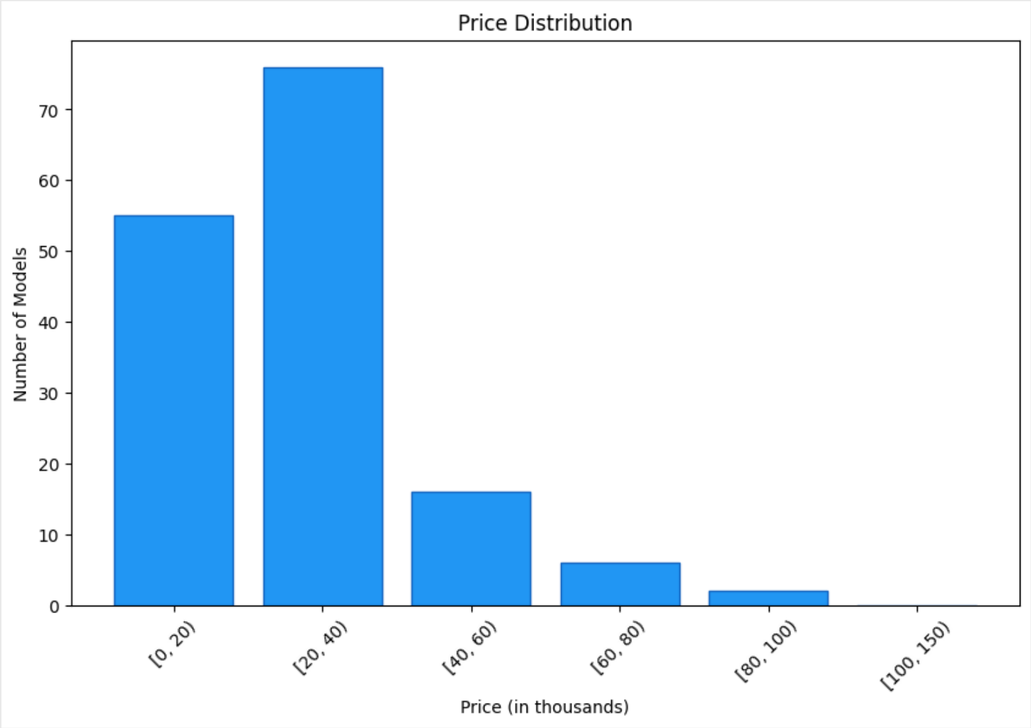
RESULTS

SALES DISTRIBUTION



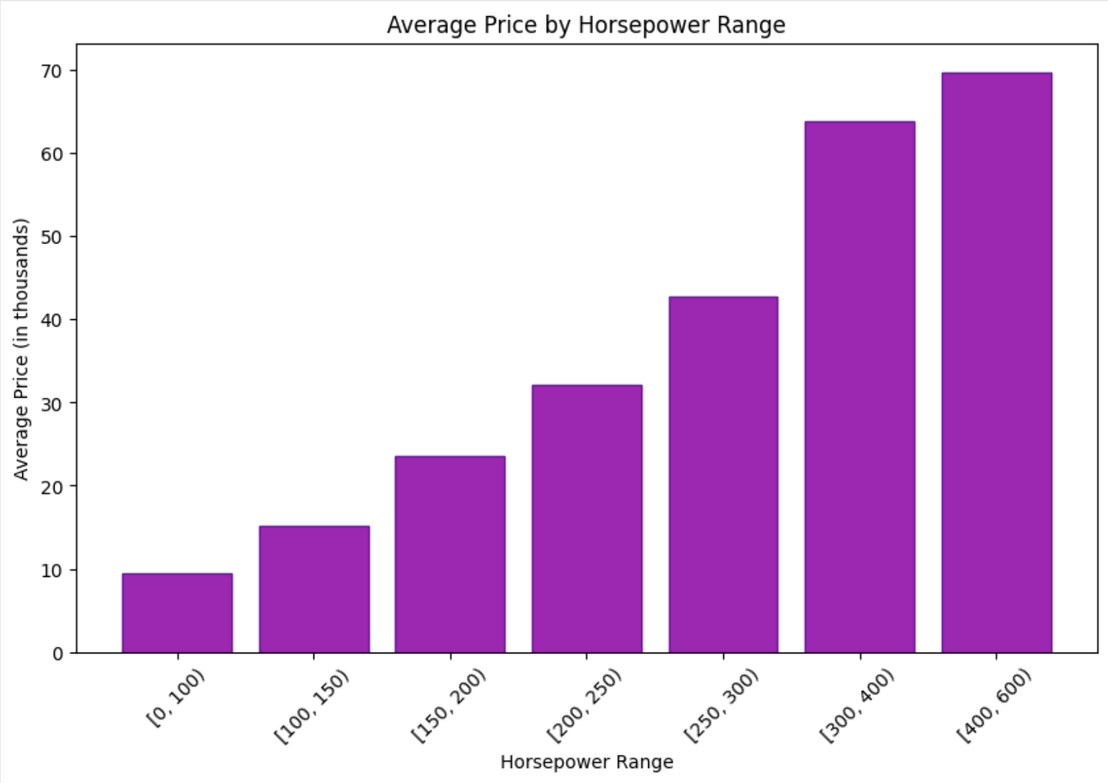
- Cars with lower sales numbers are more common.
- Most cars sold fewer than 50 thousand units.
- Only a small number of cars reached very high sales.
- This means the market is highly competitive, and very few models dominate.
- X-axis → Sales Ranges (in thousands)
- Taken from the Sales_in_thousands column and grouped into ranges (0–50, 50–100, etc.) for better readability.
- Y-axis → Number of Models
- Shows how many car models fall into each sales range (counted directly from dataset rows).
- Note: "In thousands" means sales are recorded in multiples of 1,000 (e.g., 50 → 50,000 cars).

PRICE DISTRIBUTION



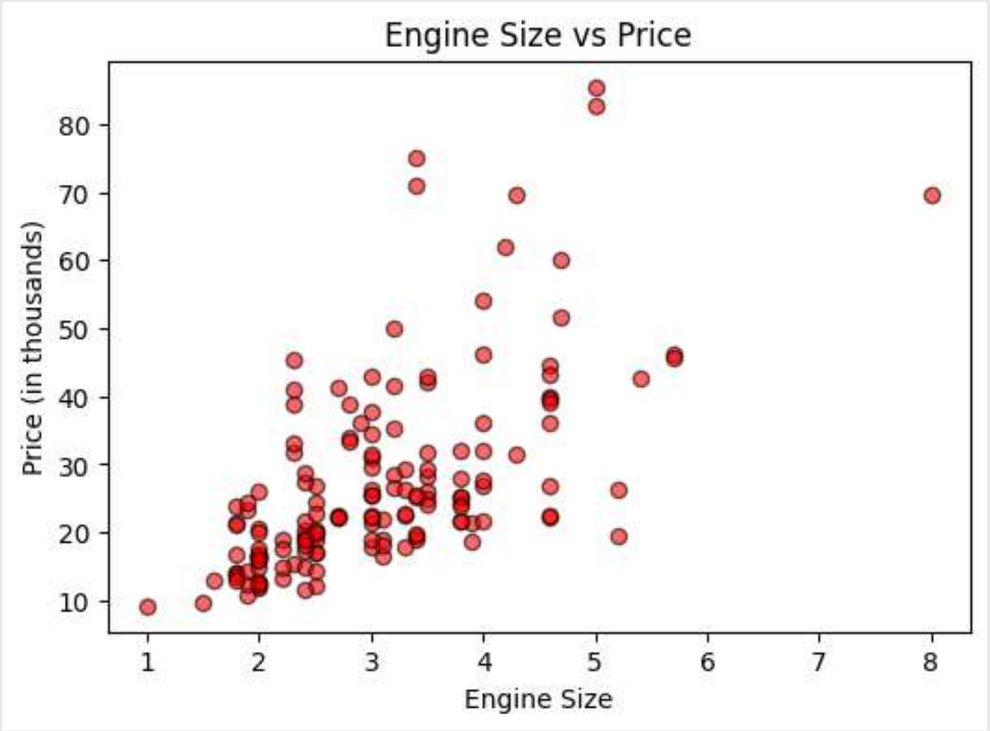
- Most cars are priced below 40 thousand.
- Very expensive cars (above 80 thousand) are rare.
- This shows the majority of models target mid-range buyers.
- X-axis → Price Ranges (in thousands)
- Taken from the Price_in_thousands column and grouped into ranges (0–20, 21–40, etc.).
- Y-axis → Number of Models
- Shows how many models fall into each price range in the dataset.
- Note: "In thousands" means price values are recorded in 1,000 units (e.g., 20 → ₹20,000).

AVERAGE PRICE BY HORSEPOWER RANGE



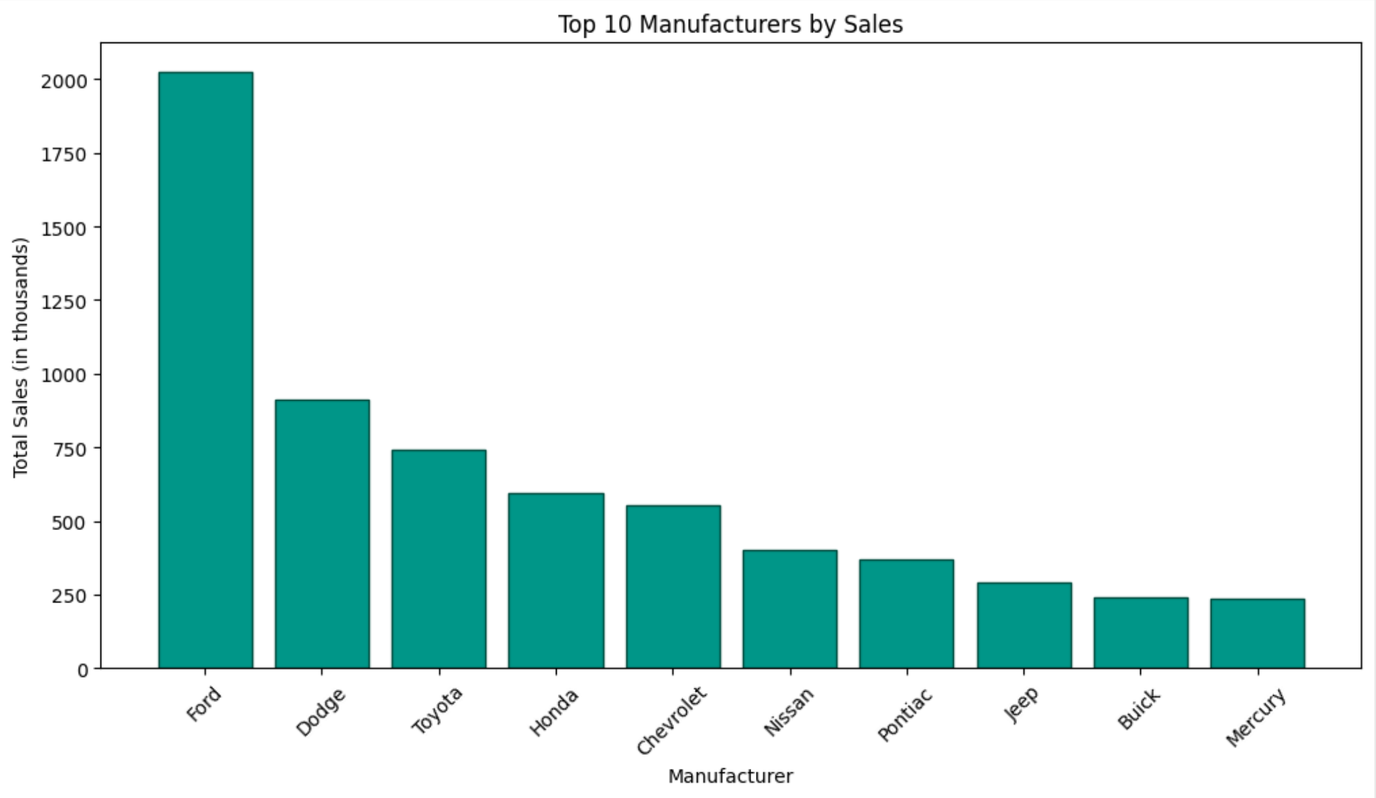
- X-axis → Horsepower Ranges
- From Horsepower, grouped into bins (100–150, 151–200, etc.).
- Y-axis → Average Price (in thousands)
- Shows the mean price for cars within each horsepower range
- This helps us understand if higher power cars are more expensive

ENGINE SIZE AND PRICE



- X-axis → Engine Size
- Each point comes directly from the Engine_size column.
- Y-axis → Price (in thousands)
- Taken directly from Price_in_thousands
- Each point in the graph shows a car's engine size compared with its price.
- This helps in checking if larger engines usually mean higher prices.

TOP COMPANIES



- X-axis → Manufacturers (Brands)
- Taken from Manufacturer column, filtered to top 10.
- Y-axis → Total Sales (in thousands)
- Sum of Sales_in_thousands for each brand Calculated by summing sales values
- This shows which companies sell the most cars overall.

FUTURE SALES PREDICTION

- Affordable cars will continue to dominate sales.
- Expensive cars will stay in small numbers.
- Passenger vehicles will grow slowly compared to cars.
- Top manufacturers are likely to maintain leadership.

CONCLUSION

- Most cars are affordable and sell in small numbers.
- Expensive cars exist but are rare.
- Sales are led by a few top manufacturers.
- Bigger and stronger vehicles usually cost more.

REFERENCES

- Dataset: Kaggle – Car Sales Dataset by gagandeep16.
- Tools: Python, Pandas, Matplotlib.

THANK YOU