Q1: SUMMARY OF DATASET

The dataset, sourced from a text file named 'automobile.txt', encapsulates a comprehensive array of columns detailing automobile attributes and specifications. These columns cover a wide spectrum of automotive features, encompassing parameters such as 'fuel-type', 'aspiration', 'drive-wheels', 'wheel-base', 'length', 'width', 'height', 'curb-weight', 'engine-type', 'num-of-cylinders', 'engine-size', 'fuel-system', 'bore', 'stroke', 'compression-ratio', 'horsepower', 'peak-rpm', 'city-mpg', 'highway-mpg', and 'price'. Each attribute serves as a vital component in understanding and analysing the characteristics and performance metrics of the automobiles documented within the dataset.

Through this diverse compilation of automobile data, analysts can delve deep into the intricate details of various vehicle models, discerning patterns and trends that shed light on consumer preferences, market dynamics, and technological advancements in the automotive industry. By exploring the relationships between different attributes and visualising the data using tools like histograms and scatter plots, researchers can uncover valuable insights into pricing strategies, engine performance, and market segmentation. With a robust dataset at their disposal, stakeholders can make informed decisions, develop targeted marketing campaigns, and enhance product offerings to meet the evolving needs of customers in the automotive landscape.

Q2:DATA CLEANING

Data cleaning is a crucial step in preparing datasets for analysis, ensuring accuracy and reliability. In this process, duplicate rows are identified and removed to prevent redundancy. Moreover, rows containing missing data, often represented as NaN values, are dropped to maintain data integrity. Additionally, in the 'price' column, '?' values are replaced with NaN, followed by conversion to a numeric data type for consistency and compatibility across analyses. These steps collectively enhance the quality of the dataset, enabling more meaningful insights and informed decision-making.

Q3: MISSING DATA

In handling missing data, we began by loading the dataset and removing any duplicate rows to ensure data cleanliness. Subsequently, we identified and removed rows with missing values, employing the dropna() function. Further, we tackled '?' values in the 'price' column by replacing them with NaN, enabling consistent data handling.

The 'price' column was then converted to numeric format using pd.to_numeric(), ensuring uniformity for analysis. Following these steps, we verified the absence of NaN values in the 'price' column and inspected a sample of the cleaned dataset. Additionally, we delved into data exploration by grouping automobiles by 'make' and identifying the top 5 most expensive vehicles. Furthermore, we determined the most popular body style and common horsepower, providing insights into the dataset's characteristics.

Finally, we illustrated data distributions through captivating visualisations, showcasing the distribution of car prices by make, horsepower, and engine size, facilitating a deeper understanding of the dataset's nuances. Through these comprehensive actions, we ensured the dataset's readiness for further analysis and interpretation.

Q4: STORIES AND ASSUMPTIONS BASED ON VISUALISATIONS OF DATA:

Distribution of Car Prices by Make

The boxplot analysis provides insights into the price distribution of various car brands, showcasing their price ranges and quartile values. It highlights differences in pricing strategies and target markets among different car manufacturers, aiding in market analysis and decision-making processes. See observations below:

Mercedes-Benz: The prices for Mercedes-Benz vehicles range from a minimum of $26,000 to a maximum of $45,000. The third quartile (Q3) for Mercedes-Benz is $36,500, indicating that 75% of the Mercedes-Benz cars in the dataset have prices at or below this value.

BMW: These cars have prices ranging from a minimum of $16,000 to a maximum of $42,000. The third quartile (Q3) for BMW is $32,500, suggesting that 75% of BMW cars in the dataset have prices at or below this value.

Porsche: The price range for Porsche vehicles spans from a minimum of $30,000 to a maximum of $37,500. The third quartile (Q3) for Porsche is $36,000, indicating that 75% of Porsche cars in the dataset have prices at or below this value.

Jaguar: Jaguar cars have prices ranging from a minimum of $32,500 to a maximum of $36,000. The third quartile (Q3) for Jaguar is $36,000, suggesting that 75% of Jaguar cars in the dataset have prices at or below this value.

Volvo: Volvo cars are priced between $12,500 (minimum) and $22,500 (maximum). However, specific quartile information for Volvo is not provided in the analysis.

Other Vehicles: All other vehicles have maximum prices below $20,000, with a minimum value of $5,000.

Distribution of Car Prices by Horsepower using Histogram

The histogram depicting the distribution of car prices by horsepower provides a clear overview of the frequency distribution of prices within the dataset. By analysing the histogram, we can discern several key observations. Firstly, there appears to be a trend where higher prices correspond to lower frequencies. This suggests that cars with higher price tags are less common within the dataset, possibly indicating a smaller market segment for luxury or high-end vehicles.

Breaking down the histogram into price ranges, we observe that the price range of $20,000 to $45,000 generally exhibits a frequency averaging below 10, indicating that cars within this higher price bracket are relatively scarce in the dataset. Conversely, the price range of $10,000 to $17,500 shows a slightly higher frequency, typically ranging between 10 to 20, suggesting a broader distribution of cars within this mid-range price segment.

Moving to lower price segments, the histogram illustrates that prices ranging from $5,000 to $10,000 have a frequency between 25 to 45, indicating a higher prevalence of cars within this affordability range. Notably, the highest frequency is observed at $7,500, with a frequency of 45, indicating a significant concentration of cars at this price point.Continuing down the price spectrum, the histogram depicts that the $5,000 price point also has a substantial frequency, typically around 35, indicating another popular price point within the dataset. Finally, the lowest price range of $1,000 exhibits a frequency ranging between 25 to 27.5, indicating a moderate prevalence of cars at this lower price point.

The histogram effectively illustrates the distribution of car prices within the dataset, highlighting the prevalence of certain price ranges and revealing insights into the market segment's pricing dynamics. It provides valuable information for understanding the affordability and popularity of cars at different price points, aiding stakeholders in market analysis and decision-making processes.

Distribution of Car Prices by Engine-size using Scatter Plot

The scatter plot visualises the distribution of car prices relative to engine size. Analysing the plot, we observe a generally positive correlation between engine size and car price. As engine size increases, there is a tendency for prices to rise, indicating that larger engines are often associated with higher-priced vehicles.

Considering the specified price ranges based on engine size, we can infer that cars with engine sizes ranging from 50 to 100 cubic inches typically fall within the price range of $5,000 to $10,000. Moving to the next range of 100 to 150 cubic inches, prices span from $7,500 to $50,000, showcasing a wider variability reflecting different car models and features. Similarly, for engine sizes between 150 to 200 cubic inches, prices range from $13,500 to $37,500, indicating a higher price bracket for larger engine sizes.

As we progress to larger engine sizes, the price range further escalates. Engine sizes of 200 to 250 cubic inches are associated with prices ranging from $30,000 to $42,000, indicating premium vehicles with enhanced performance. Finally, engine sizes ranging from 250 to 350 cubic inches show a more consolidated price range, predominantly around $35,000, with some outliers reaching prices between $40,000 to $45,000, possibly representing luxury or high-performance vehicles.Overall, the scatter plot highlights the importance of engine specifications in determining the pricing dynamics of vehicles, guiding consumers and stakeholders in making informed decisions.

THIS REPORT WAS WRITTEN BY: AVRIL JACK