Question 1: Read the dataset and display first 10 rows, shape and columns

Shape: (1061, 7)
Columns: ['name', 'selling_price', 'year', 'seller_type', 'owner', 'km_driven', 'ex_showroom_price']

Question 2: Missing values and handling approach

Checked missing values and calculated percentage of missing data.

Approach for handling missing values:

- For small percentages, impute numeric columns with median and categorical with mode.

- For large percentages (>40%), consider dropping the column.

- Drop rows where 'selling_price' is missing.

- Keep an unmodified copy of raw data for reference.

Question 3: Distribution of selling prices (Histogram)

A histogram was plotted for 'selling_price'.

Summary statistics:

Count: 1061, Mean: 59638, Std: 56304, Min: 5000, Max: 760000.

Observation: The distribution is right-skewed with most bikes priced under ■1,00,000.

Question 4: Average selling price by seller_type

Bar plot was created for average selling price by seller_type.

Observation: 'Individual' sellers have the highest average selling price.

Question 5: Average km_driven for each ownership type

Calculated and plotted average km_driven grouped by owner type.

Observation: Higher ownership counts generally correspond to higher kilometers driven.

Question 6: Outlier detection and removal using IQR method

Before IQR removal: count = 1061, min = 350, max = 880000

After IQR removal: count = 1022, min = 350, max = 86000

IQR bounds: lower = -30750, upper = 87250

Observation: Extreme outliers (km_driven > 87250) were removed, improving data reliability.

Question 7: Scatter plot of year vs selling_price

Scatter plot shows a positive correlation between year and selling_price.

Newer bikes (recent years) have higher selling prices.

Question 8: One-hot encoding of seller_type

Applied one-hot encoding using pandas get_dummies().

Created new columns for each seller type (e.g., seller_Individual, seller_Dealer, etc.) and displayed first 5 rows.

Question 9: Correlation matrix heatmap

Generated heatmap of numeric columns.

Top correlations:

- ex_showroom_price & selling_price: 0.9186

- year & selling_price: 0.4022

- km_driven & year: 0.2887

Observation: Selling price is strongly correlated with ex_showroom_price and moderately with year.

Question 10: Summary Findings

Important factors affecting selling_price:

- Year (bike age): newer bikes are priced higher.

- km_driven: higher distance reduces price.

- Ownership and seller_type also impact value.

Data cleaning and feature engineering:

- Converted km_driven and year to numeric.

- Used IQR method to remove outliers.

- One-hot encoded seller_type.

- Suggested median/mode imputations for missing data.

Final note: Include all plots and code outputs with this report for full credit.