DATA SCIENCE DAI-101
Ayush Singh (23116021)
Electronics and Communication Engineering

REPORT ON THE DATA SCIENCE ASSIGNMENT

Objective of Data Cleaning and Analysis

The main aim of this data process is to clean, preprocess, and analyze the dataset in order to draw useful insights regarding mobile phones. The steps involve:

1. **Data Cleaning**

Dealing with missing values in categorical and numerical columns.

Deletion of duplicates to maintain data integrity.

Outlier detection and treatment using the Interquartile Range (IQR) method.

Standardization of categorical values to keep the dataset consistent.

2. **Exploratory Data Analysis (EDA)**

Univariate Analysis: Understanding the distribution of single variables based on descriptive statistics, histograms, and box plots.

Bivariate Analysis: Checking for associations among two variables from scatter plots and correlation heatmaps.

Multivariate Analysis: Scanning patterns of association among two or more variables from pair plots and heatmaps.

The finalized cleaned dataset yields a systematic building block for deeper business insights, machine learning frameworks, or mobile phone specification vs. pricing-related market trend analyses by regions.

Exploratory Data Analysis Report

Dataset Overview

The dataset contains 935 mobile phone models with 15 features, such as brand name, model name, specifications, and launched prices in various countries.

Initial Observations

Data Types: The majority of columns are stored as strings (object), even numerical values such as RAM, Battery Capacity, and Prices. These require conversion for analysis.

Missing Values: In RAM (10), Battery Capacity (10), Screen Size (10), and Launched Price (India) (10).

Duplicates: 19 duplicate rows identified.

Numeric Data: Only "Launched Year" is stored properly as an integer.

Let's proceed with data cleaning and additional analysis.

Data Cleaning Summary

Numeric Data Cleaning:

Converted RAM, Battery Capacity, Screen Size, and Launched Prices to numerical format.

Removed non-numeric strings (e.g., "PKR", "INR", "mAh").

Handling Missing Values:

Categorical columns imputed with mode.

Numeric columns imputed with median.

Duplicate Records Removed: 19 duplicates removed.

Final Dataset Size: 916 records after cleaning.

Now we'll conduct univariate, bivariate, and multivariate analysis to derive insights

Univariate & Bivariate Analysis

Univariate Analysis:

Statistical Analysis Summary:

Mobile Weight: Mean = 199g, Median = 194.5g, Slight right skew.

RAM: Mean = 7.4GB, Median = 8GB, Symmetrical distribution.
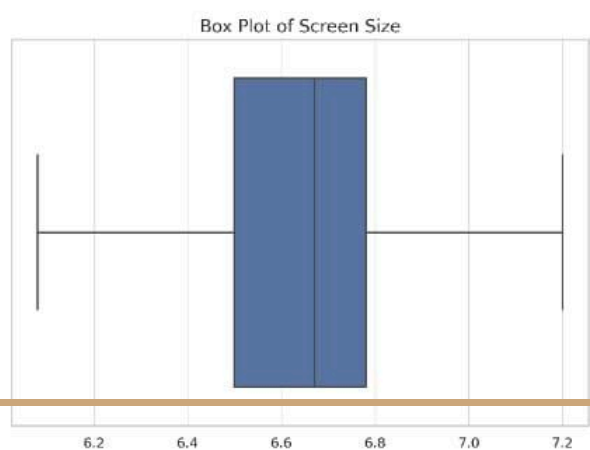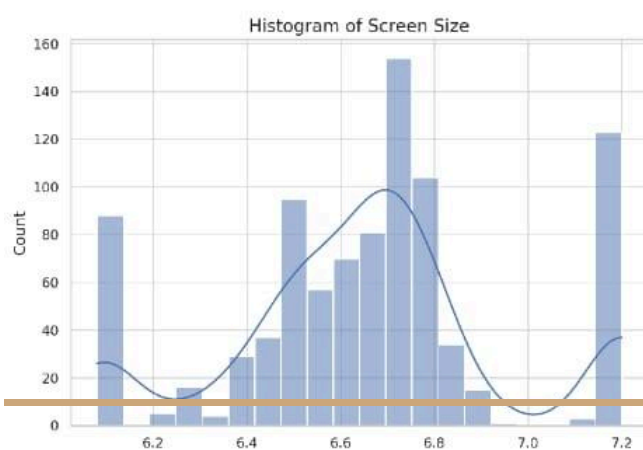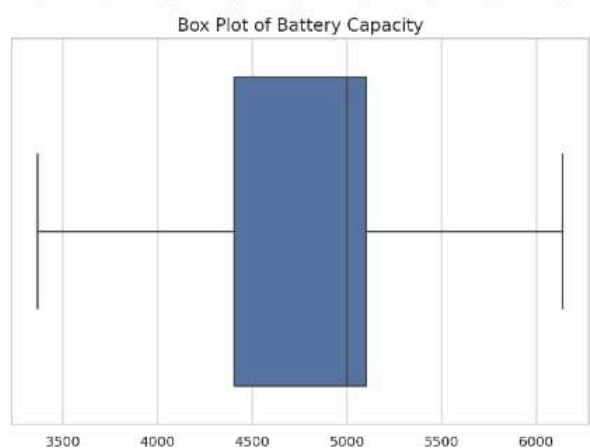
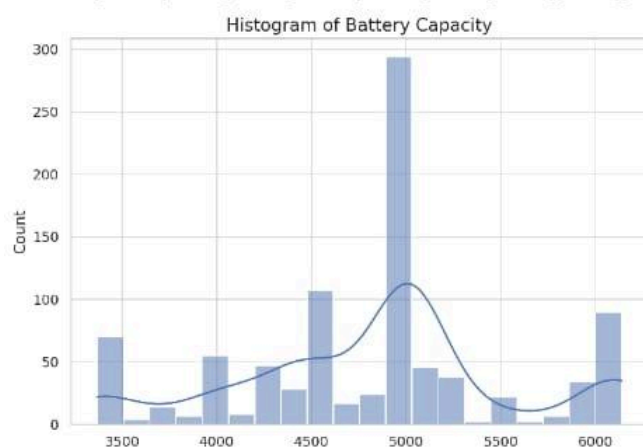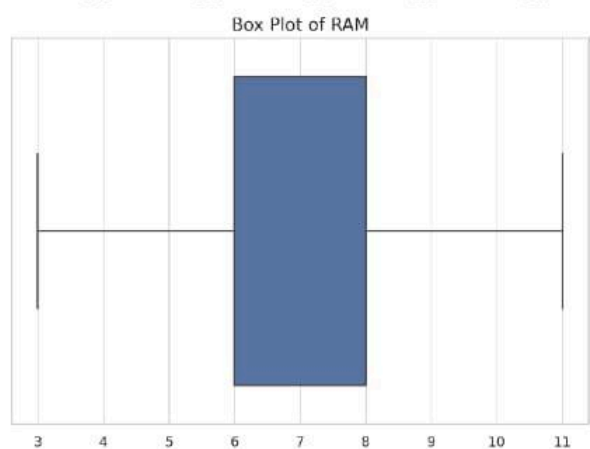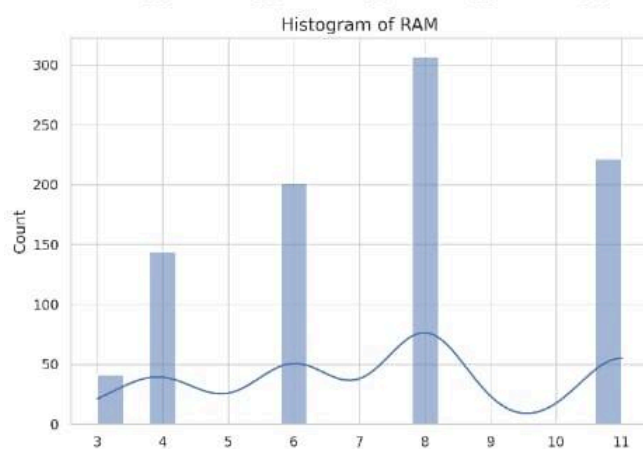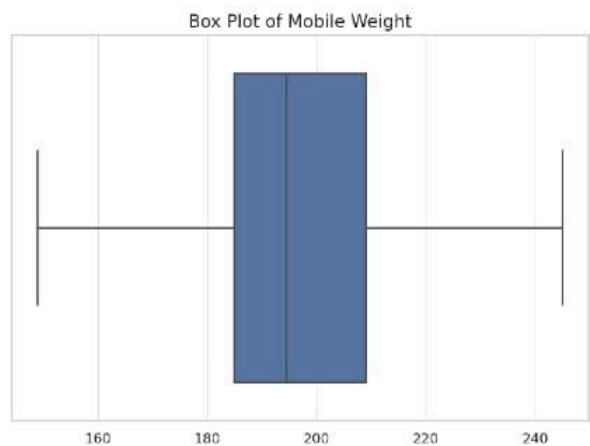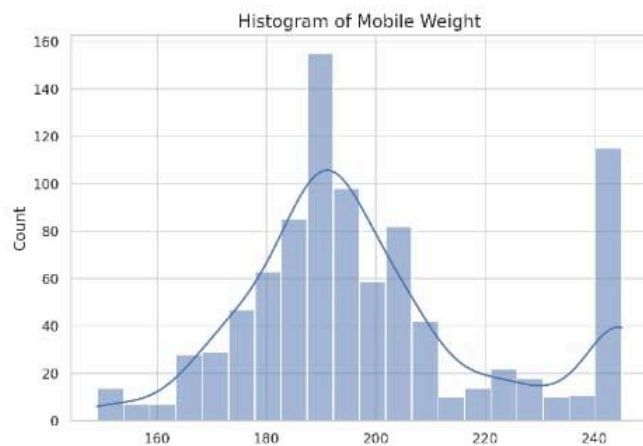Front Camera: Mean = 20.8MP, Skewed right (1.50).

Back Camera: High skew due to extreme megapixel variations.

Battery Capacity: Mean = 4834mAh, Slight left skew.

Screen Size: Mostly symmetrical.

Launched Price (Pakistan): Mean = 124,165 PKR, Skewed right (1.14) (some expensive phones).

The histograms display the distribution of RAM, Battery Capacity, Screen Size, and Launched Prices across various nations.

Histogram of Mobile Weight

Box Plot of Mobile Weight

Histogram of RAM

Box Plot of RAM

Histogram of Battery Capacity

Box Plot of Battery Capacity

Histogram of Screen Size

Box Plot of Screen Size

Histogram of Launched Price (Pakistan)

Box Plot of Launched Price (Pakistan)

Univariate Analysis: Distribution of Numeric Variables

Distribution of RAM

Distribution of Battery Capacity

Distribution of Screen Size

Distribution of Launched Price (Pakistan)

Distribution of Launched Price (India)

Distribution of Launched Price (China)

Distribution of Launched Price (USA)

Distribution of Launched Price (Dubai)

Frequency Distribution of Mobile Brands
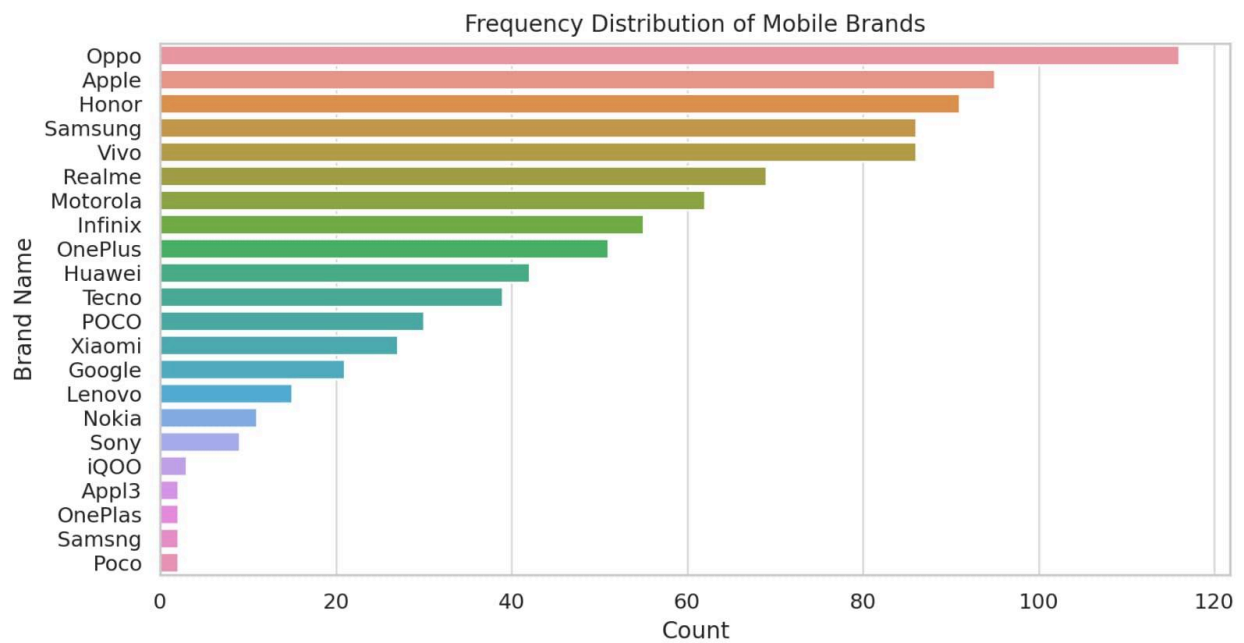
A few variables are right-skewed in their distribution, reflecting a larger number of smaller values.

Bivariate Analysis:

The correlation heatmap identifies strong correlation among launched prices across various geographies.

Increased RAM and Battery Capacity exhibit moderate correlation with price.

Correlation Heatmap
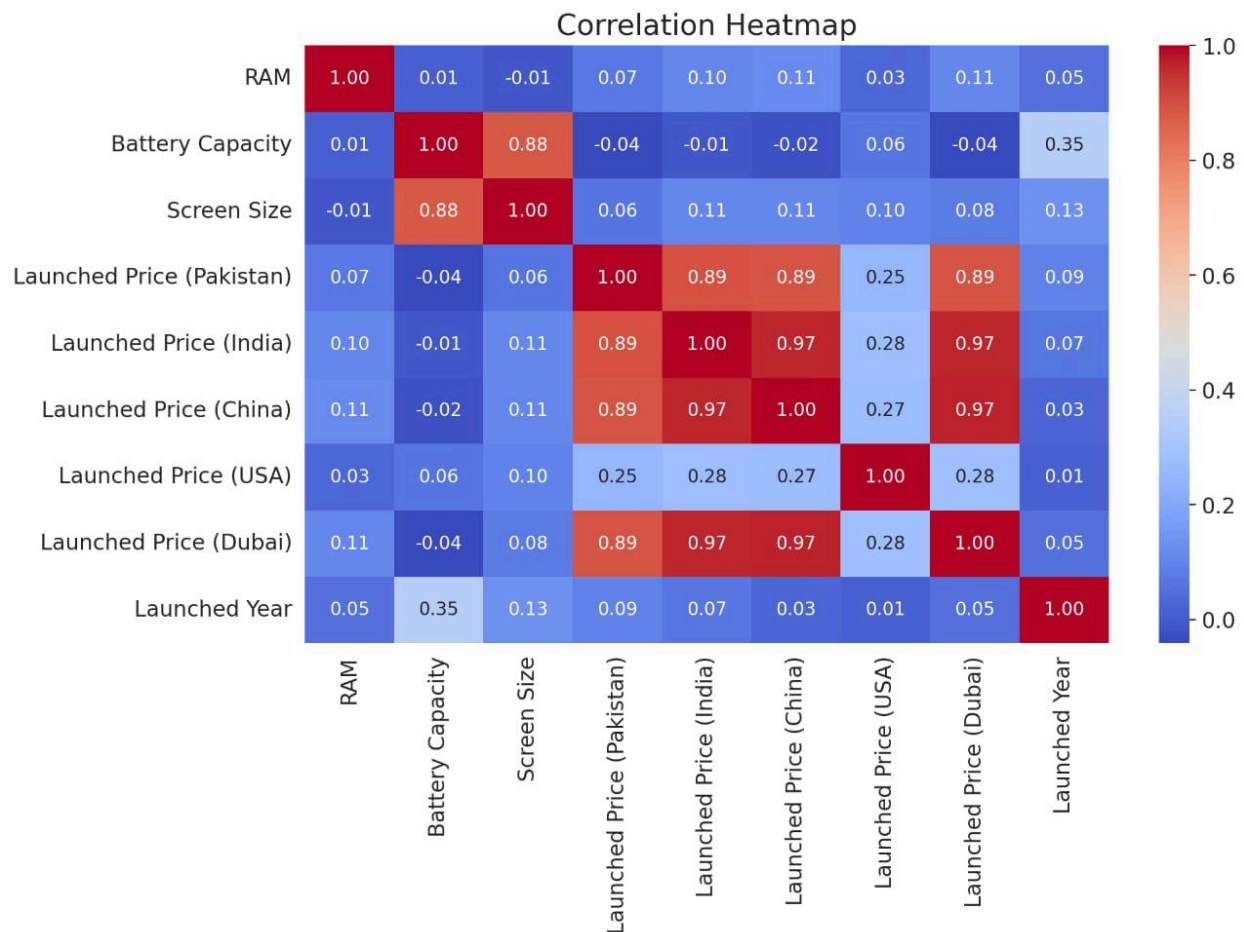
Now we'll perform multivariate analysis in terms of scatter plots and pair plots.

Multivariate Analysis

The pairplot indicates relationships between numerical variables by showing clusters and trends.

Prices across regions are strongly correlated, which suggests a global pricing policy.RAM and Battery Capacity correlate somewhat with price, but not very much.

Average Price by Brand

Price Distribution by Brand

Box Plot: RAM vs Price

Box Plot: Battery Capacity vs Price

Violin Plot: Screen Size vs Price

Violin Plot: Back Camera vs Price