Student Name: SAI MOHAN BOLLAM

Student ID: 23027511

Introduction

This report presents an exploration of the selected Mobile Price Classification dataset. This dataset consists of various characteristics relative to mobile phones such as battery power, camera specifications, memory, network connectivity, among others. Through the exploration process, various insights and relationships were identified between different the price range and other features. In this context, various tasks were carried out to achieve the assignment's objectives and they included descriptive statistics, correlation analysis, and visualizations to uncover relationships within the project data.

Data Exploration

This was the initial step and it entailed loading the project dataset and examining its structure using the pandas library as recorded below.
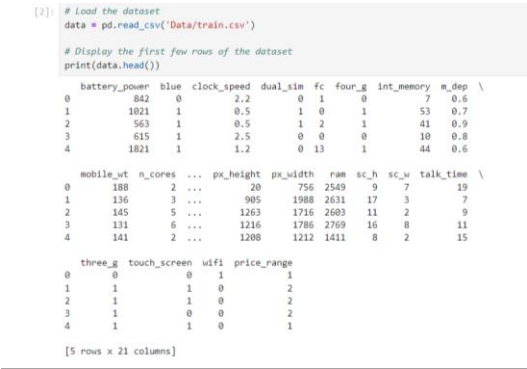

Figure 1: Dataset Loading and Exploration

The resulting output showed the first few rows of the project data thus facilitating a thorough understanding of the data and its columns.

Descriptive Statistics

Descriptive Statistics was implemented to effectively understand the distribution as well as statistics of the numeric columns in the project dataset and this was achieved by using the describe function according to (Kaur et al. 2018) and recorded below.
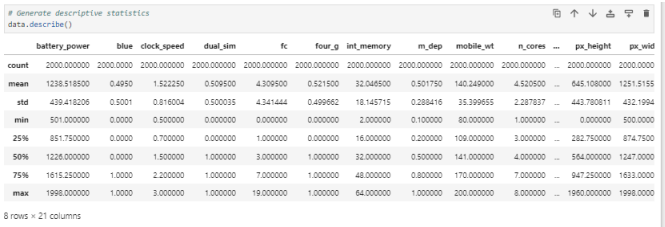

Figure 2: Descriptive Statistics

The output provided summary statistics such as count, mean, standard deviation, minimum, 25th percentile, median, 75th percentile, and maximum for each numeric column in the dataset.

Correlation Analysis

The correlation matrix showcased the pairwise correlations between all numeric columns in the dataset. According to (Vallat, 2018) the positive values indicated a positive correlation, negative values indicated a negative correlation, and values close to zero indicated a weak or no correlation.

Figure 3: Correlation Analysis

Data Visualization

The data visualization was achieved by plotting three different plots as discussed below. These visualized the relationships within the project dataset as outlined by (Sial et al. 2021).

i.   Histogram: this showed the distribution of a single variable that is "battery_power". This histogram displayed the distribution of battery power values in the dataset as shown below.

```python
# Plot a histogram of 'battery_power'
plt.hist(data['battery_power'], bins=10, edgecolor='black')
plt.xlabel('Battery Power')
plt.ylabel('Frequency')
plt.title('Distribution of Battery Power')
plt.show()
```
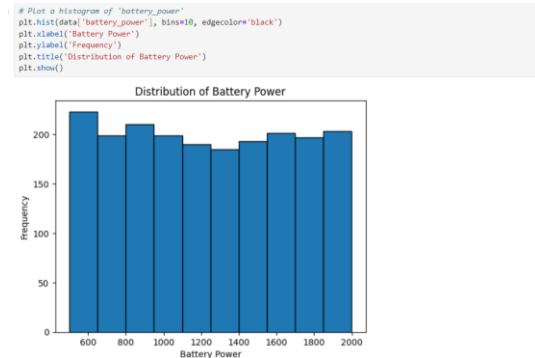
Figure 4: Battery Power Distribution

ii.   Scatter plot: The scatter plot helped to visualize the relationship between two variables. In this context, the selected variables were 'ram' (Random Access Memory) versus 'price_range'.
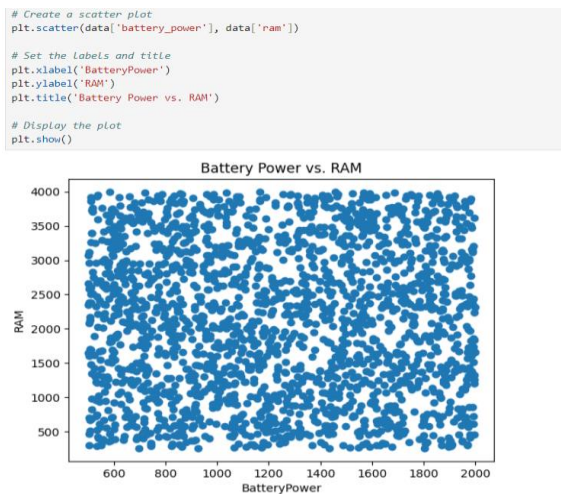
```python
# Create a scatter plot
plt.scatter(data['battery_power'], data['ram'])

# Set the labels and title
plt.xlabel('BatteryPower')
plt.ylabel('RAM')
plt.title('Battery Power vs. RAM')

# Display the plot
plt.show()
```

Figure 5: Scatter Plot

iii.   Heatmap: The heatmap visualized the correlation matrix. The heatmap will provide a visual representation of the correlation between different features in the dataset.
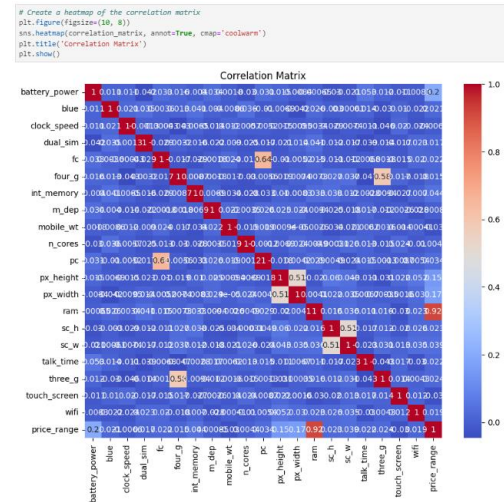
```python
# Create a heatmap of the correlation matrix
plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm')
plt.title('Correlation Matrix')
plt.show()
```

Figure 6: Correlation Heatmap

References

Dataset:
https://www.kaggle.com/datasets/iabhishekofficial/mobile-price-classification/data

Kaur, Stoltzfus, and Yellapu (2018) 'Descriptive statistics', International Journal of Academic Medicine, 4(1), pp.60-63.

Sial, Rashdi, and Khan (2021) 'Comparative analysis of data visualization libraries Matplotlib and Seaborn in Python', International Journal, 10(1).

Vallat (2018) 'Pingouin: statistics in Python', J. Open Source Softw., 3(31), p.1026.

GITHUB: https://github.com/saimohan2003/APPLIED-DATA-SCIENCE--1

CODE: APPLIED-DATA-SCIENCE--1/Mohan27.ipynb at main · saimohan2003/APPLIED-DATA-SCIENCE--1 (github.com)

DATA SET:
https://github.com/saimohan2003/APPLIED-DATA-SCIENCE--1/blob/main/train.csv