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.

# Generate descriptive statistics data.describe()													∓ #
	battery_power	blue	clock_speed	dual_sim	fc	four_g	int_memory	m_dep	mobile_wt	n_cores		px_height	px_wid
count	2000.000000	2000.0000	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000		2000.000000	2000.0000
mean	1238.518500	0.4950	1.522250	0.509500	4.309500	0.521500	32.046500	0.501750	140.249000	4.520500		645.108000	1251.5155
std	439.418206	0.5001	0.816004	0.500035	4.341444	0.499662	18.145715	0.288416	35.399655	2.287837		443.780811	432.1994
min	501.000000	0.0000	0.500000	0.000000	0.000000	0.000000	2.000000	0.100000	80.000000	1.000000		0.000000	500.0000
25%	851.750000	0.0000	0.700000	0.000000	1.000000	0.000000	16.000000	0.200000	109.000000	3.000000		282.750000	874.7500
50%	1226.000000	0.0000	1.500000	1.000000	3.000000	1.000000	32.000000	0.500000	141.000000	4.000000		564.000000	1247.0000
75%	1615.250000	1.0000	2.200000	1.000000	7.000000	1.000000	48.000000	0.800000	170.000000	7.000000		947.250000	1633.0000
max	1998.000000	1.0000	3.000000	1.000000	19.000000	1.000000	64.000000	1.000000	200.000000	8.000000		1960.000000	1998.0000

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.

	battery_power	blue	clock_speed	dual_sim	fc	four_g	int_memory	m_dep	mobile_wt	n_cores	-	px_height	px_width	rie
battery_power	1.000000	0.011252	0.011482	-0.041847	0.033334	0.015665	-0.004004	0.034085	0.001844	-0.029727		0.014901	-0.008402	-0.0006
blue	0.011252	1.000000	0.021419	0.035198	0.003593	0.013443	0.041177	0.004049	-0.008605	0.036161		-0.006872	-0.041533	0.0263
clock_speed	0.011482	0.021419	1.000000	-0.001315	-0.000434	-0.043073	0.006545	-0.014364	0.012350	-0.005724		-0.014523	-0.009476	0.0034
dual_sim	-0.041847	0.035198	-0.001315	1.000000	-0.029123	0.003187	-0.015679	-0.022142	-0.006979	-0.024658		-0.020875	0.014291	0.0410
fc	0.033334	0.000593	-0.000434	-0.029123	1.000000	-0.016560	-0.029133	-0.001791	0.023618	-0.013356		-0.009990	-0.005176	0.0150
four_g	0.015665	0.013443	-0.043073	0.003187	-0.016560	1.000000	0.008690	-0.001823	-0.016537	-0.029706		-0.019236	0.007448	0.0073
int_memory	-0.004004	0.041177	0.006545	-0.015679	-0.029133	0.008690	1.000000	0.006886	-0.034214	-0.028310		0.010441	-0.008335	0.0626
m_dep	0.034085	0.004049	-0.014364	-0.022142	-0.001791	-0.001823	0.006886	1.000000	0.021756	-0.003504		0.025263	0.023566	-0.0094
mobile_wt	0.001844	-0.008605	0.012350	-0.008979	0.023618	-0.016537	-0.034214	0.021756	1.000000	-0.018989		0.000939	0.000090	-0.0025
n_cores	-0.029727	0.036161	-0.005724	-0.024658	-0.013356	-0.029706	-0.028310	-0.003504	-0.018909	1.000000		-0.006872	0.024480	0.0040
pc	0.031441	-0.009952	-0.005245	-0.017143	0.644595	-0.005598	-0.033273	0.026282	0.018844	-0.001193		-0.018465	0.004196	0.0090
px_height	0.014901	-0.006872	-0.014523	-0.020875	-0.009990	-0.019236	0.010441	0.025263	0.000939	-0.006872		1.000000	0.510664	-0.0200
px_width	-0.008402	-0.041533	-0.009476	0.014291	-0.005176	0.007448	-0.008335	0.023566	0.000090	0.024480		0.510664	1.000000	0.0041
ram	-0.000653	0.026351	0.003443	0.041072	0.015099	0.007313	0.032813	-0.009434	-0.002581	0.004868		-0.020352	0.004105	1.0000
sc,h	-0.029959	-0.002952	-0.029078	-0.011949	-0.011014	0.027166	0.037771	-0.025348	-0.033855	-0.000815		0.059615	0.021599	0.0156
SC,W	-0.021421	0.000613	-0.007378	-0.016666	-0.012373	0.037005	0.011731	-0.018388	-0.020761	0.025826		0.043098	0.034699	0.0858
talk_time	0.052510	0.013934	-0.011432	-0.039404	-0.006829	-0.046628	-0.002790	0.017008	0.006209	0.013148		-0.010645	0.006720	0.0100
three_g	0.011522	-0.030236	-0.046433	-0.014008	0.001799	0.584246	-0.009366	-0.012065	0.001551	-0.014733		-0.031174	0.000350	0.0157
touch_screen	-0.010516	0.010061	0.019756	-0.017117	-0.014828	0.016758	-0.026999	-0.002638	-0.014368	0.023774		0.021891	-0.001628	-0.0004

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.

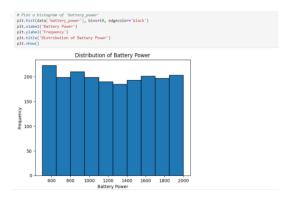


Figure 4: Battery Power Distribution

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'.

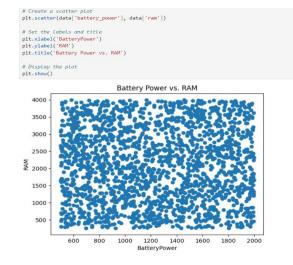


Figure 5: Scatter Plot

iii. Heatmap: The heatmap visualized the correlation matrix. The heatmap will provide a visual representation of the correlation between different features in the dataset.

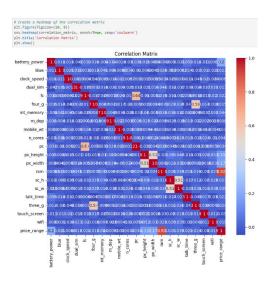


Figure 6: Correlation Heatmap

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

DATA SET:

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