Submitted by: Rajan Ghimire

Studet ID: C0924991

About the dataset :

(Note: The dataset you provided was text based data so i choosed another data as per our discussion. This is the old data.)

The dataset comprises detailed attributes of mobile phones, capturing various technical specifications and features, with a focus on assessing their price range. It includes 21 attributes, both categorical and continuous.

This dataset is downloaded form this link : https://www.kaggle.com/datasets/iabhishekofficial/mobile-price-classification/data?select=train.csv

Now, let's look at each column, its type, and a short description of the column.

	Column	Type	Description
0	battery_power	Continuous	Total energy a battery can store in one time measured in (mAh)
1	blue	Categorical	Has Bluetooth or not (0: No, 1: Yes)
2	clock_speed	Continuous	Speed at which microprocessor executes instructions (GHz)
3	dual_sim	Categorical	Has dual SIM support or not (0: No, 1: Yes)
4	fc	Continuous	Front camera (Megapixels)
5	four_g	Categorical	Has 4G or not (0: No, 1: Yes)
6	int_memory	Continuous	Internal memory in (Gigabytes)
7	m_dep	Categorical	Mobile depth in (Cm)
8	mobile_wt	Continuous	Weight of mobile phone (grams)
9	n_cores	Categorical	Number of cores of processor
10	pc	Continuous	Primary camera (Megapixels)
11	px_height	Continuous	Pixel resolution height
12	px_width	Continuous	Pixel resolution width
13	ram	Continuous	Random access memory in (Megabytes)
14	sc_h	Continuous	Screen height of mobile in (Cm)
15	sc_w	Continuous	Screen width of mobile in (Cm)
16	talk_time	Continuous	Longest time that a single battery charge will last when you are constantly talking on the phone (hours)
17	three_g	Categorical	Has 3G or not (0: No, 1: Yes)
18	touch_screen	Categorical	Has touch screen or not (0: No, 1: Yes)
19	wifi	Categorical	Has WiFi or not (0: No, 1: Yes)
20	price_range	Categorical	Target variable with value of 0: (Low Cost), 1: (Medium Cost), 2: (High Cost), and 3: (Very High Cost)

Statisticts of Numerical Columns:

The table below shows the statistical information such as mean, mode, median, Q1, and Q3 for each of the numerical features. This helps us to find:

- Central Tendency and Distribution
- Variability and Spread
- Identifying Outliers
- Feature Comparison

	Column	Mean	Median	Mode	Q1	Q3
0	battery_power	1238.52	1226	618	851.75	1615.25
1	clock_speed	1.52225	1.5	0.5	0.7	2.2
2	fc	4.3095	3	0	1	7
3	int_memory	32.0465	32	27	16	48
4	mobile_wt	140.249	141	182	109	170
5	pc	9.9165	10	10	5	15
6	px_height	645.108	564	347	282.75	947.25
7	px_width	1251.52	1247	874	874.75	1633

	Column	Mean	Median	Mode	Q1	Q3
8	ram	2124.21	2146.5	1229	1207.5	3064.5
9	sc_h	12.3065	12	17	9	16
10	sc_w	5.767	5	1	2	9
11	talk_time	11.011	11	7	6	16

Identifying Missing values in each Column :

Analyzing missing values, it can be stated that there are no missing values present in any of the dataset's columns, as the percentage was 0% for all features. This means that there is no need to do imputation or data cleaning to handle missing value, which simplifies analysis and modeling.

	Column	Missing_Percentage
0	battery_power	0
1	blue	0
2	clock_speed	0
3	dual_sim	0
4	fc	0
5	four_g	0
6	int_memory	0
7	m_dep	0
8	mobile_wt	0
9	n_cores	0
10	pc	0
11	px_height	0
12	px_width	0
13	ram	0
14	sc_h	0
15	sc_w	0
16	talk_time	0
17	three_g	0
18	touch_screen	0
19	wifi	0
20	price_range	0

Identifying which of the categorical features are important to us:

Chi-square tests to determine the importance of categorical features relative to this target.

	Column	Chi2	p-value
0	m_dep	28.7291	0.374155
1	n_cores	20.5245	0.488294
2	touch_screen	3.88014	0.274701
3	four_g	3.17988	0.364714
4	blue	1.43214	0.698018

Based on the chi-square analysis, the top 3 most important categorical columns with respect to the "price_range" target variable are:

1. **m_dep**
2. **n_cores**
3. **touch_screen**

These columns have the highest chi-square statistics, indicating a stronger association with the target variable "price_range".

Now, lets have a look at count of each of the categorical features:

m_dep count

	m_dep	Count
0	0.1	320
1	0.2	213
2	0.8	208
3	0.5	205
4	0.7	200
5	0.3	199
6	0.9	195
7	0.6	186
8	0.4	168
9	1	106

n_cores count

	n_cores	Count
0	4	274
1	7	259
2	8	256
3	2	247
4	3	246
5	5	246
6	1	242
7	6	230

touch_screen count

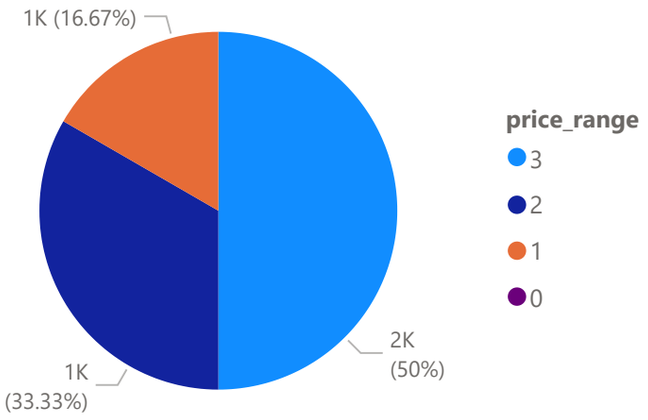
	touch_screen	Count
0	1	1006
1	0	994

Analysis of Feature Correlations in Mobile Phone Specifications Dataset

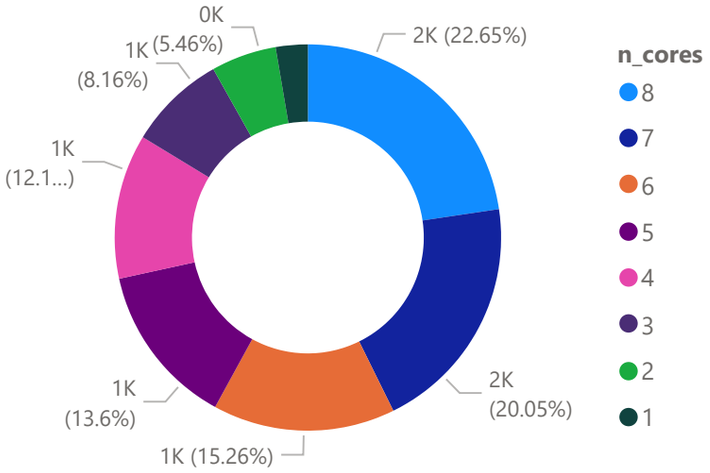
Correlation matrix displays the nature of relationship between two numerical variables in the data set, where it varies from -1 to +1. If the value is positive, it shows a positive relation; conversely, if the value is negative, it reveals a negative relation. It is noteworthy that, despite the generally low levels of correlation between most pairs of variables, pc (primary camera megapixels) and fc (front camera megapixels) have a positive linear correlation of 0. 6446, which indicates that higher primary camera resolution is likely to be accompanied by higher front camera resolution as well.

	battery_power	clock_speed	fc	int_memory	mobile_wt	pc	px_height	px_width	ram	
battery_power	1	0.0114816	0.0333344	-0.00400368	0.00184438	0.0314407	0.0149008	-0.00840183	-0.000652926	-0.
clock_speed	0.0114816	1	-0.000433898	0.00654515	0.0123497	-0.00524504	-0.0145229	-0.00947565	0.00344303	-0.
fc	0.0333344	-0.000433898	1	-0.0291328	0.023618	0.644595	-0.00998991	-0.00517563	0.015099	-(
int_memory	-0.00400368	0.00654515	-0.0291328	1	-0.0342142	-0.0332734	0.0104413	-0.00833485	0.0328132	0.
mobile_wt	0.00184438	0.0123497	0.023618	-0.0342142	1	0.0188439	0.000939324	8.97616e-05	-0.00258054	-0.
pc	0.0314407	-0.00524504	0.644595	-0.0332734	0.0188439	1	-0.0184655	0.00419594	0.0289835	0.0
px_height	0.0149008	-0.0145229	-0.00998991	0.0104413	0.000939324	-0.0184655	1	0.510664	-0.0203519	0.
px_width	-0.00840183	-0.00947565	-0.00517563	-0.00833485	8.97616e-05	0.00419594	0.510664	1	0.00410522	0.
ram	-0.000652926	0.00344303	0.015099	0.0328132	-0.00258054	0.0289835	-0.0203519	0.00410522	1	0.
sc_h	-0.0299586	-0.0290776	-0.011014	0.0377711	-0.0338547	0.00493752	0.0596153	0.0215986	0.0159961	
sc_w	-0.0214209	-0.00737836	-0.0123726	0.0117305	-0.0207605	-0.0238192	0.0430383	0.0346992	0.0355757	(
talk_time	0.0525104	-0.0114319	-0.00682865	-0.00279029	0.0062085	0.014657	-0.0106453	0.00671994	0.01082	-0.

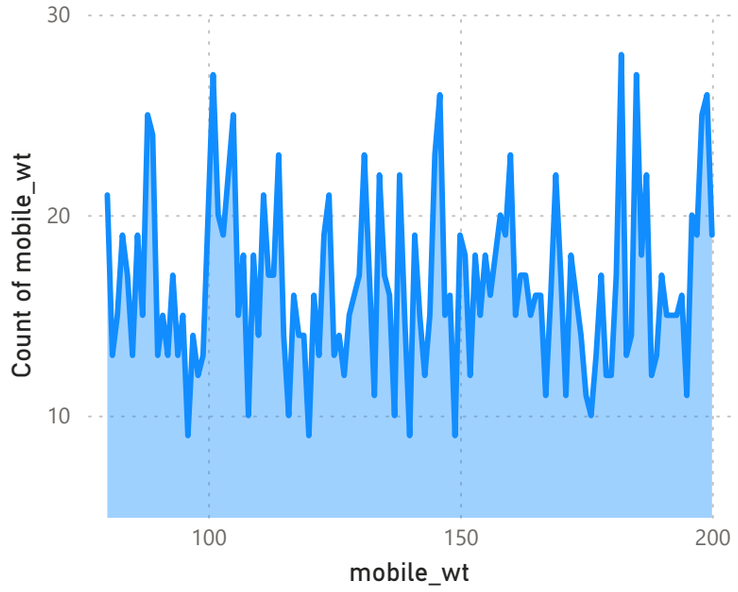
Sum of price_range by price_range



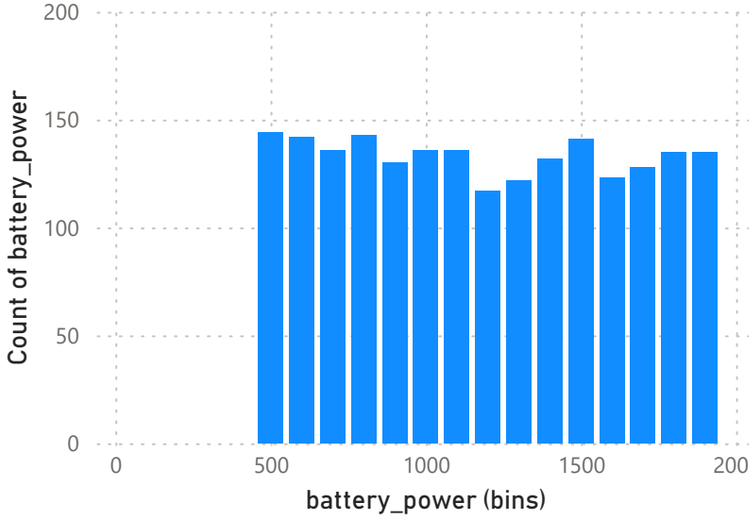
Sum of n_cores by n_cores



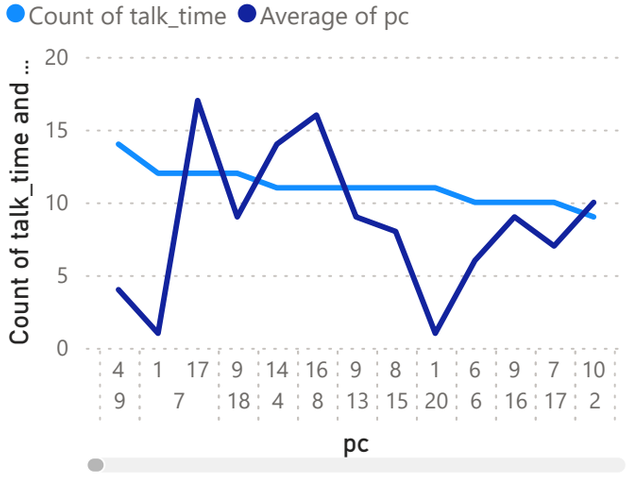
Count of mobile_wt by mobile_wt



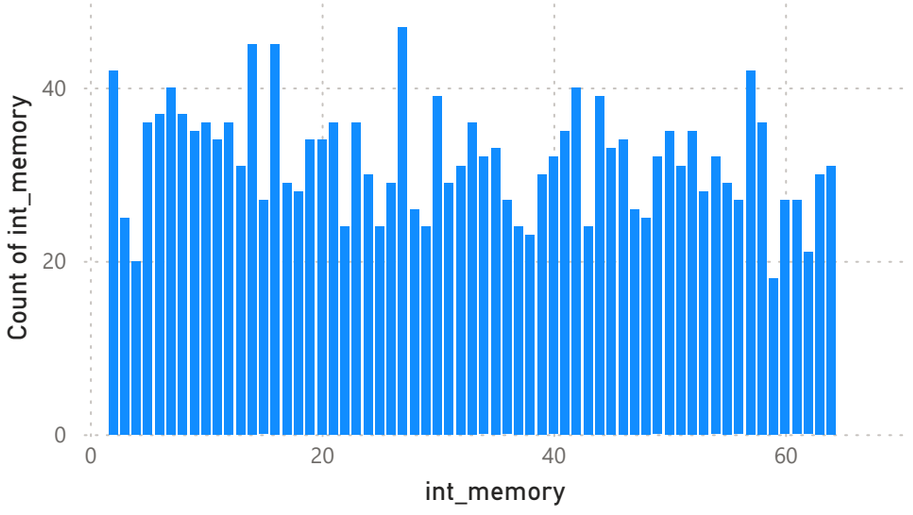
Count of battery_power by battery_power (bins)



Count of talk_time and Average of pc by talk_time and pc



Count of int_memory by int_memory



Quick measure