**Mobile Price**

**Introduction:**

The dataset provides information about various features of mobile phones. The features include battery power, clock speed, internal memory, RAM, screen height and width, camera resolution, and several other hardware specifications. Also the dataset includes information about software features such as 3G and 4G connectivity, Wi-Fi, Bluetooth, and Dual SIM support. Here, the target column is price\_range which is a categorical column having 4 possible values.

The main goal of the dataset is to select the most significant features that can accurately predict the mobile price category.

**Preprocessing:**

**Handling Null Values:**

battery\_power 0.0

blue 0.0

clock\_speed 0.0

dual\_sim 0.0

fc 0.0

four\_g 0.0

int\_memory 0.0

m\_dep 0.0

mobile\_wt 0.0

n\_cores 0.0

pc 0.0

px\_height 0.0

px\_width 0.0

ram 0.0

sc\_h 0.0

sc\_w 0.0

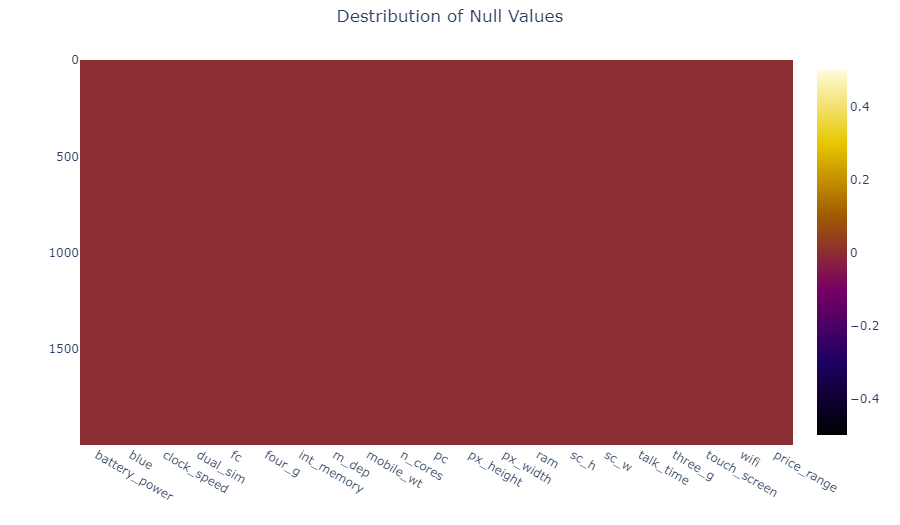
talk\_time 0.0

three\_g 0.0

touch\_screen 0.0

wifi 0.0

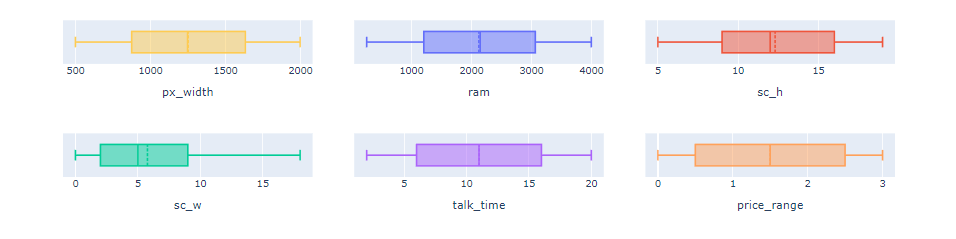
price\_range 0.0



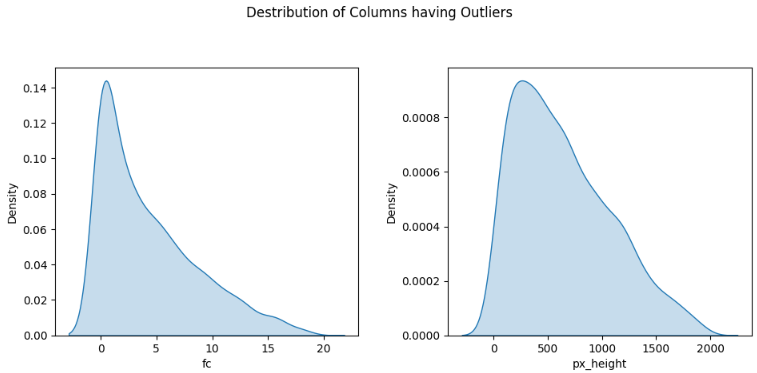
From the above heatmap we can conclude that, in this dataset there is no null values in any feature.

**Handling Outlier:**



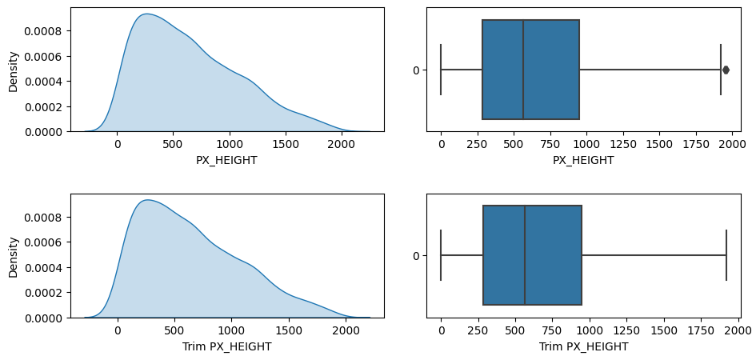


From the graph we can see only two columns has outliers. [fc, px\_height] And they both are rightly skewed. So, we have to use IQR method to handle the outliers in these two columns.



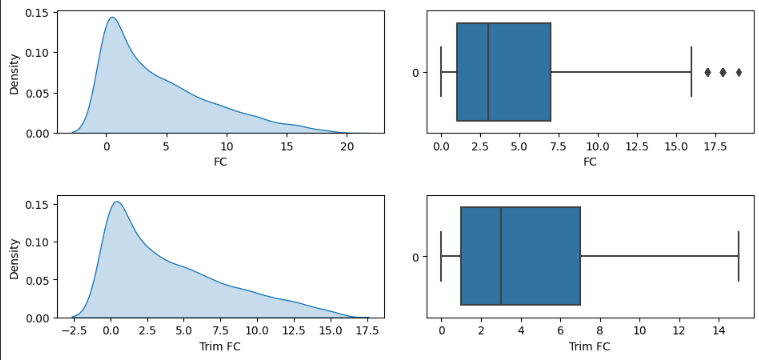
Distribution of Columns Having Outlier

**Using trimming method to handle the outlier in px\_height feature:**

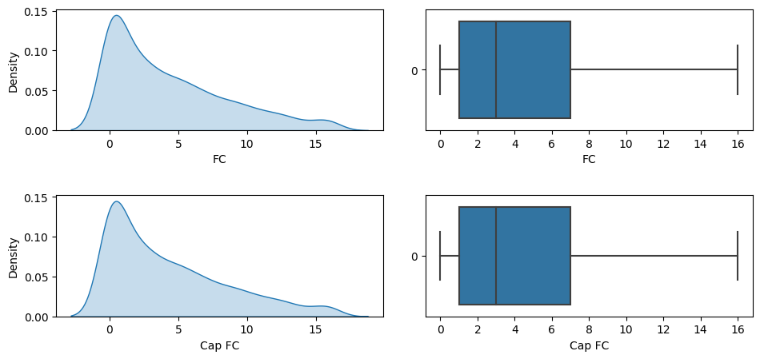


From the box plots we can see that the distribution didn’t change too much.

**Using trimming method to handle the outlier in fc feature:**



Here, there is a noticeable different the plots. So, we used capping method to compare and find out the best possible outcome.



After capping we can see that the difference is much less between the original and capped plot.

Based on the above analysis conducted, it can be concluded that the capping method is useful for the FC column, as this feature contains a greater number of outliers. Conversely, trimming may be more appropriate for the Px\_height feature, as it only contains two outliers.



**Hypothesis Test:**

On this dataset hypothesis testing can be performed to determine the statistical significance of the relationship between the target feature prize\_range. By performing Anova(for numerical columns) and Chi-square (for categorical columns) tests it is possible to test hypothesis for this dataset.

If there is no significant difference in the mean ‘prize\_range’ across different levels of categorical features in the data then it is Null Hypothesis, and if there is a significant difference then we call it alternative hypothesis.

We can do Anova test for following features:

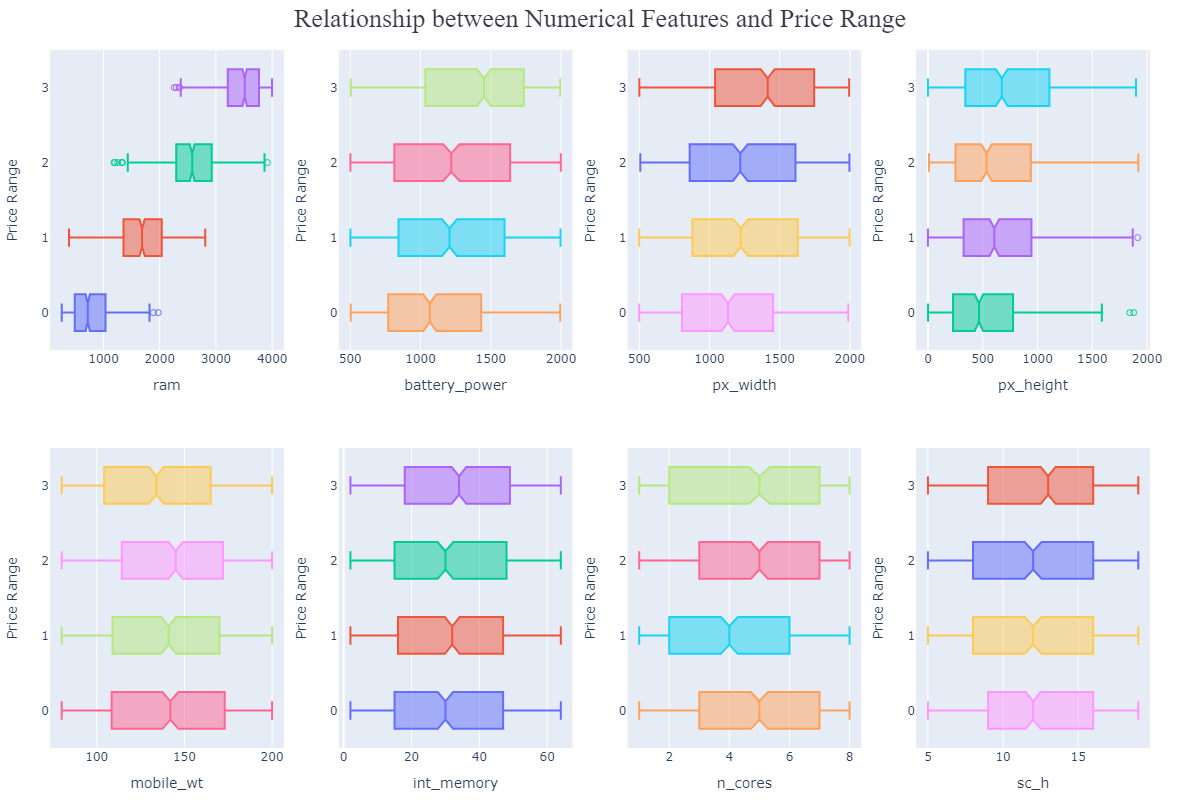
'battery\_power', 'clock\_speed', 'fc', 'int\_memory', 'm\_dep', 'mobile\_wt', 'n\_cores', 'pc', 'px\_height', 'px\_width', 'ram', 'sc\_h', 'sc\_w', 'talk\_time', 'price\_range'.

**P-value, Test of Association:**

Test of association simply means Chi-square test, which is a statistical test to determine if two categorical features are associated with each other or not. It calculates a p-value to determine the association between variables if the null hypothesis is true or false. Lower p-value means null hypothesis in rejected and alternative hypothesis is accepted.

For Chi-square test we have following features:

'blue', 'dual\_sim', 'four\_g', 'three\_g', 'touch\_screen', 'wifi'.





**Comparing Our Anova Function and Chi-square Function with Build-in Anova Function:**

**Anova:**

battery\_power

stat = 31.513495036295836, p\_value = 6.713993985856411e-20

my\_stat = 31.513495036295783, my\_p\_value = 6.713993985856411e-20

clock\_speed

stat = 0.491978773388441, p\_value = 0.6878760540629911

my\_stat = 0.49197877338861495, my\_p\_value = 0.6878760540628887

fc

stat = 0.8412347075208211, p\_value = 0.47119535676382607

my\_stat = 0.8412347075208505, my\_p\_value = 0.47119535676382607

int\_memory

stat = 2.9131423673657735, p\_value = 0.03321699701567197

my\_stat = 2.913142367365822, my\_p\_value = 0.033216997015667736

m\_dep

stat = 1.5181696567539429, p\_value = 0.20783366859699934

my\_stat = 1.5181696567539407, my\_p\_value = 0.20783366859699934

mobile\_wt

stat = 3.57796808786005, p\_value = 0.013415544766487946

my\_stat = 3.5779680878599414, my\_p\_value = 0.013415544766489661

n\_cores

stat = 2.576356312579159, p\_value = 0.052259228034057296

my\_stat = 2.5763563125790894, my\_p\_value = 0.05225922803406589

pc

stat = 0.8581096270002767, p\_value = 0.4621872664942891

my\_stat = 0.8581096270003586, my\_p\_value = 0.4621872664942498

px\_height

stat = 19.256805801002372, p\_value = 2.6162526915481838e-12

my\_stat = 19.25680580100237, my\_p\_value = 2.6162526915481838e-12

px\_width

stat = 22.405610912881148, p\_value = 2.881567837380335e-14

my\_stat = 22.405610912881055, my\_p\_value = 2.881567837380949e-14

ram

stat = 3527.774583085812, p\_value = 0.0

my\_stat = 3527.774583085812, my\_p\_value = 0.0

sc\_h

stat = 2.2300733023906982, p\_value = 0.0828023294418198

my\_stat = 2.2300733023906734, my\_p\_value = 0.08280232944182207

sc\_w

stat = 1.673070617080054, p\_value = 0.17076307782254166

my\_stat = 1.6730706170800698, my\_p\_value = 0.17076307782254166

talk\_time

stat = 1.6254402003698731, p\_value = 0.18144519465488843

my\_stat = 1.6254402003699502, my\_p\_value = 0.18144519465487718

**Chi-square:**

blue

stat = 1.379985190664331, p\_value = 0.7102329471672442

my\_stat = 1.379985190664331, my\_p\_value = 0.7102329471672442

dual\_sim

stat = 1.1650734607420992, p\_value = 0.7613927398347171

my\_stat = 1.1650734607420992, my\_p\_value = 0.7613927398347171

four\_g

stat = 3.1839157849182875, p\_value = 0.3641289090432397

my\_stat = 3.183915784918287, my\_p\_value = 0.3641289090432396

three\_g

stat = 1.3194817939999668, p\_value = 0.7245122123530843

my\_stat = 1.3194817939999668, my\_p\_value = 0.7245122123530843

touch\_screen

stat = 3.9919597018810995, p\_value = 0.26233364677575377

my\_stat = 3.991959701881099, my\_p\_value = 0.262333646775754

wifi

stat = 0.8039541146769342, p\_value = 0.8485210293812833

my\_stat = 0.8039541146769342, my\_p\_value = 0.8485210293812833

**Feature Selection Based on P-value:**

ram: 0.0

battery\_power: 6.713993985856411e-20

px\_width: 2.881567837380335e-14

px\_height: 2.6162526915481838e-12

mobile\_wt: 0.013415544766487946

int\_memory: 0.03321699701567197

n\_cores: 0.052259228034057296

sc\_h: 0.0828023294418198

sc\_w: 0.17076307782254166

talk\_time: 0.18144519465488843

m\_dep: 0.20783366859699934

touch\_screen: 0.262333646775754

four\_g: 0.3641289090432396

pc: 0.4621872664942891

fc: 0.47119535676382607

clock\_speed: 0.6878760540629911

blue: 0.7102329471672442

three\_g: 0.7245122123530843

dual\_sim: 0.7613927398347171

wifi: 0.8485210293812833

We Have to select the top features with the smallest p-values for training the model.

Ram and battery\_power have the smallest p-values, indicating a strong relationship with the target variable.

Px\_width and px\_height also have small p-values and are statistically significant, but to a slightly lesser extent.

Other features have larger p-values and may not have a strong relationship with the target variable.

As a result we select ram, battery\_power, px\_width, px\_height, mobile\_wt, int\_memory , n\_cores and sc\_h as our top features for training the model.

Relevant Exploratory Data Analysis:

1. How does the battery power vary across different price ranges?

A picture containing text, screenshot, diagram, line

Description automatically generated

2. What is the relationship between pixel height (px\_height) and pixel width (px\_width)

A picture containing text, screenshot, electric blue, design

Description automatically generated

3.Distribution of price ranges across different categorical features

From the above destribution we can say that there is a positive correlation between pixel height and pixel width. Though the relation is moderately strong. It's not highly linear.

A picture containing text, screenshot, colorfulness, parallel

Description automatically generated

From the above destribution we can conclude that majority of the phone in price range 3 have dual sim installed in them also there are not that much difference between the count of touch screen phone and button phone.

One thing also noticable is that there is more touch screen phones in price range 0 and 1 compare to price range 2 and 3. Also majority of the phone without touch screen are in price range 2.

Overall the numbers are almost similar in each category.

4. What is the frequency of 4G connectivity for different price range categories

A picture containing text, screenshot, rectangle, design

Description automatically generated

From the above destribution we can conclude that higher-priced mobile phones (Category 3) tend to have a higher frequency of 4G connectivity compared to other price range categories. On the other hand, lower-priced mobile phones (Category 2) have a higher frequency of 3G connectivity.

This information can be useful for understanding the relationship between price range and network connectivity options in the dataset. It suggests that as the price range increases, there is a higher likelihood of mobile phones having 4G connectivity

The highest frequency count of 274 in Price Range Category 3 suggests that mobile phones in this higher-priced category are more likely to offer 4G connectivity.

5. How does the weight of the mobile device ('mobile\_wt') differ across price ranges?

A picture containing text, screenshot, colorfulness, diagram

Description automatically generated

The distribution of phone weights across different price range categories suggests a potential negative correlation between price range and weight. The majority of mobile phones in the higher price range (Category 3) have relatively lower weights, typically falling within the 80-90 range. Conversely, the majority of mobile phones with weights ranging from 140 to 170 are found in the mid-price range (Category 2).

This observation indicates that as the price range increases, there is a tendency for mobile phones to have lighter weights. Similarly, within the mid-price range, there is a higher concentration of mobile phones with heavier weights.

6. Distribution of the front camera megapixels ('fc') for different price range categories

A picture containing text, screenshot, diagram, plot

Description automatically generated

The analysis indicates the following observations regarding front camera specifications and their relationship with price range categories: Majority of phones lack a front camera. Higher front camera pixel count is associated with a decrease in phone count. The highest front camera resolution (16 megapixels) is mostly found in price range 1. These findings highlight the limited prevalence of front cameras in the dataset, the inverse relationship between front camera pixel count and phone count, and the concentration of high-resolution front cameras in price range 1.

7. Visualizing the relationship between the internal memory (int\_memory) and the price range using a violin plot

A picture containing text, screenshot, plot, diagram

Description automatically generated

The box plot analysis reveals that the distribution of internal memory is similar across different price range categories. There are no outliers, indicating a consistent pattern. The internal memory ranges from 2 to 64 megapixels in each category, providing a variety of options for consumers.

8. Variation of clock speed across different price ranges

A picture containing text, diagram, screenshot, rectangle

Description automatically generated

The analysis of clock speed distribution across different price range categories indicates a consistent pattern. The distribution of clock speed appears to be similar across all price range categories, suggesting that there is no significant variation in clock speed based on the price range.

9. Distribution of talk time for mobile devicesA picture containing text, screenshot, diagram, colorfulness

Description automatically generated

Based on the distribution analysis, a notable observation is that the talk time for mobile phones in price range 0 is consistently greater compared to other price range categories, irrespective of the duration of talk time. This conclusion suggests that, within the given dataset, mobile phones falling within price range 0 tend to offer longer talk time capabilities compared to devices in other price range categories.

10. Distribution of mobile device weights across different touch screen types(0, 1)

A picture containing text, diagram, screenshot, rectangle

Description automatically generated

The analysis of Mobile Device Weights in relation to Touch Screen Types reveals that there is minimal difference observed between touch screen and non-touch screen devices. The distribution of mobile weights for both touch screen and non-touch screen devices exhibits similarities. Specifically, both touch screen and non-touch screen devices have a maximum weight of 200 and a minimum weight of 80. Additionally, the median weight for both types of devices is found to be the same.

11. Correlation between battery power and clock speed using scatter plot

A picture containing text, screenshot, font, number

Description automatically generated

The scatter plot analysis reveals a correlation coefficient of 0.012 between battery power and clock speed. This correlation coefficient indicates a very weak and nearly negligible positive correlation between the two variables.

The value of 0.012 suggests that there is almost no linear relationship between battery power and clock speed within the given dataset. It implies that changes in battery power do not consistently correspond to changes in clock speed, and vice versa. Therefore, based on this analysis, it can be concluded that battery power and clock speed have a very weak or insignificant linear relationship.

12. Correlation between the amount of internal memory and the screen size (sc\_h \* sc\_w) using scatter plot.

A picture containing text, screenshot, majorelle blue, electric blue

Description automatically generated

The analysis of the relationship between the internal memory and screen size (sc\_h \* sc\_w) of mobile devices showed a very weak correlation (r = 0.021). This means that there is almost no connection between the size of the screen and the amount of internal memory in mobile devices.

**Discussion:**

**Conclusion:**

**Reference:**