**Data Preparation Project: Step1, Proposal**

**Dataset:** https://www.kaggle.com/datasets/iabhishekofficial/mobile-price-classification/data?select=train.csv

1. **Variables:**

battery\_power: Total energy battery can store from a charge (mAh)

blue: Has Bluetooth or not (binary)

clock\_speed: Speed at which microprocessor

executes instructions (seconds)

dual\_sim: Has dual sim support or not (binary)

fc: Front camera mega pixels

four\_g: Has 4G or not (binary)

int\_memory: Internal Memory in Gigabytes

m\_dep: Mobile phone depth (cm)

mobile\_wt: Weight of mobile phone

n\_cores: Number of cores of processor

pc: Primary camera mega pixels

px\_height: Pixel resolution height

px\_width: Pixel resolution width

ram: RAM in mega bytes

sc\_h: Screen height of mobile phone (cm)

sc\_w: Screen width of mobile phone (cm)

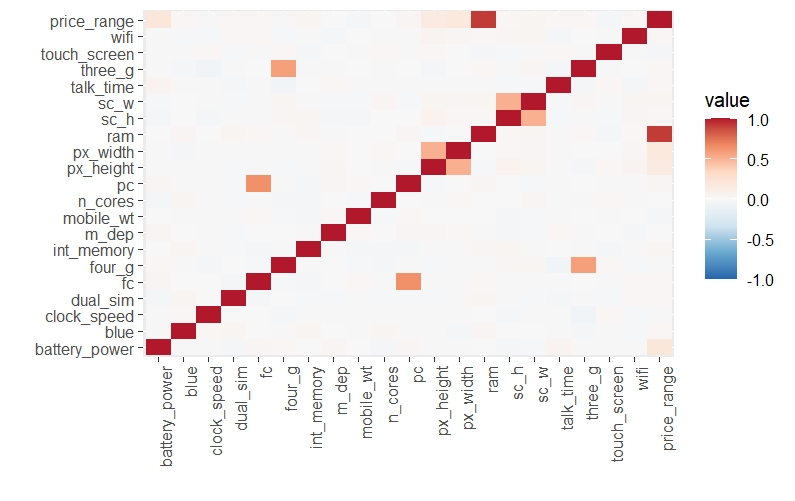
talk\_time: Time battery lasts when talking on phone (hrs)

three\_g: Has 3G or not (binary)

wifi: Has wifi or not (binary)

price\_range: Target variable, ranging from 0-3, corresponding from low to high cost

**Plots and Tables:**

A graph with a number of rectangular objects

Description automatically generated with medium confidence

A graph with a row of squares

Description automatically generatedA graph showing a number of cores

Description automatically generated

A graph with a bar graph

Description automatically generated with medium confidenceA graph of different sizes and sizes

Description automatically generated

The above plots detail some important findings about the nature of our data. The first and most crucial of these is the fact that there are hardly any correlations between any of our variables, save for a few between our response variable, price\_range, and some predictor variables that are intrinsically intertwined. Some of these include dimensional properties, such as length and width of the phone, or of the camera/pixels of the phone. To those unfamiliar with how computers and phones work, increasing the number of cores may also be believed to increase the RAM, so the boxplot above shows how this is not the case. The only boxplot truly worth mentioning is the one between price range and RAM, which shows large evidence of having different statistics for RAM for different levels of the price. We also notice that there are very few outliers across this data, save for some at the price level of 2. We can say something somewhat similar about the battery\_power variable, however these values are still far too close to one another to consider saying they have significant differences in statistics. We can also see that the counts for the different levels of our response are perfectly distributed, making analysis easier.

1. It would be best to take an approach using classification to predict the response ‘price\_range’ for this data. This is because the response is a categorical variable with 4 levels, and we must classify our data to predict which level it will go in to based on our predictors. A regression model would be much better suited if the variable was a continuous set of values corresponding to exact prices, as opposed to just binning them. A method like random forest or SVM might be the best method to employ as both can handle the mixed bag of variable types that this data presents. Random forest might be best as it can narrow down the plethora of non-important variables that seem to be present based on the correlation matrix.