**Name:** Syed Muhammad Faizan Abbas Kazmi

**Roll no :** 21L-5675

**Section:** BDS 3B

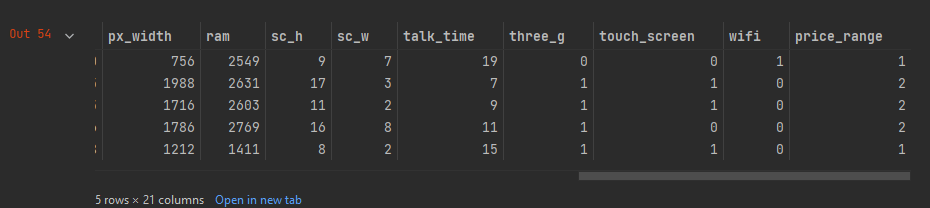
**Detailed Report**

Bob has started his own mobile company. He wants to give tough fight to big companies like Apple, Samsung etc. He does not know how to estimate price of mobiles his company creates. In this competitive mobile phone market, you cannot simply assume things. To solve this problem, he collects sales data of mobile phones of various companies. Bob wants to find out some relation between features of a mobile phone (eg:- RAM, Internal Memory etc.) and its selling price. But he is not so good at Machine Learning. So, he needs your help to solve this problem. In this problem you do not have to predict actual price but a price range indicating how high the price is. Dataset is related to the mobile price prediction based on different features. Data set consist of the 2000 rows and 21 features with the total size of the 42000. Datasets consist of the following properties.

**Dataset Dimension**: contains 2000 rows with the 21 features

Based on these 21 features, we are predicting the price of phone. By using the pandas iloc function we are selecting all these features for prediction.

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

****

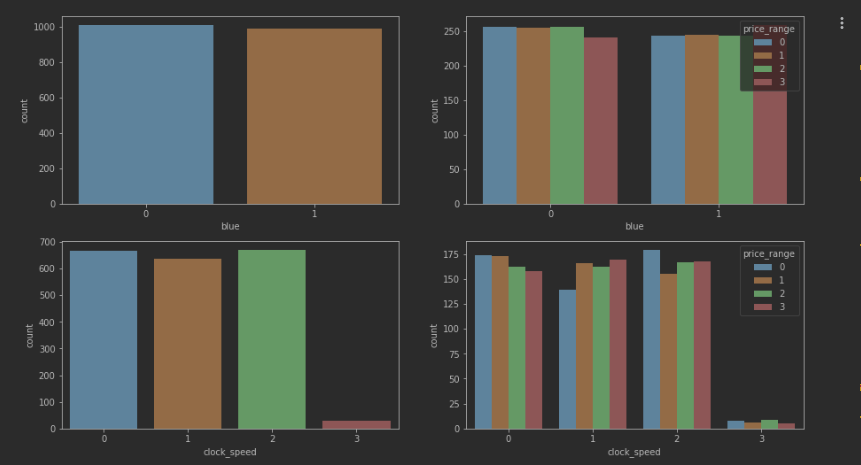
**Data Pre-processing:**

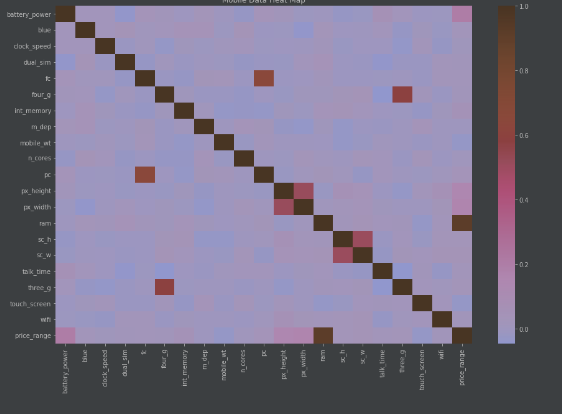
On the following Dataset we have performed the Exploratory Data Analysis technique on our dataset to clean it make suitable to use it for the prediction.

* Conversion of datatype into integer 16 for less memory usage.
* Calculated the mean, standard deviation, max of our data
* Checking the null values in our data
* Replacing the null values from its mean values
* Normalization of data and also print the z-score

For plotting of the data, we have used the matplotlib and seaborn

We have potted the categorical features by using the histogram because to clearly differentiate between the price range based on categorical features some of the features are color, ram, speed…

****

****

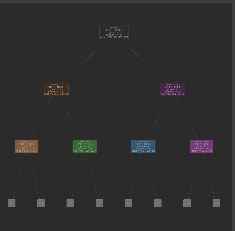
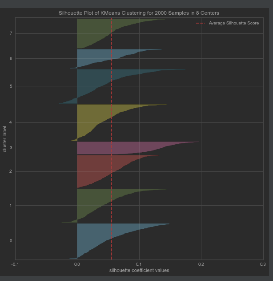
**Algorithms**

Algorithms we have used in our projects are K-means which is unsupervised learning technique and second algorithm we have used is Random Forest which is supervised learning technique. K-mean is a clustering algorithm which clustered the data based on centroids on the basis of randomly initialized data points, assign these data points to the nearest centroids no of clusters we are creating for our dataset is 7.

Random forest is an algorithm which uses the supervised learning technique and Random Forest is better and reliable than the decision tree. Random forest generates the multiples trees on the data genome values instead of Decision tree. Decision tree generates a single tree for a data. Random forest performs better than the Decision tree and its accuracy rate is higher than that.

**Results and accuracy**

**Accuracy rate of our model is: 82%**

****