ML Applications for Business Assignment 1 Report

Name : Aditya Mishra

Roll. No. : M20MA201

- We were provided with the [Dataset](#) of the **mobile price range classifications dataset .** Which has different parts in it and the description is as :
  - It has different kinds of feature columns which shows us the property of the mobiles as eg:- RAM,Internal Memory etc.
  - There is a target variable as well as the price range of the mobile. Which we have to predict for the particular datasets. This is basically a classification problem for which we have to predict their classes of the price range.

- **So let's Do exploratory analysis and comment on the findings.**
  There are Three unique classes of the price range of the mobile as:
    - Class 0
    - Class 1
    - Class 2
    - Class 3

- And the dataset contains 2000 rows and 21 columns of the different features with the count of unique values as well as mean values :

```
battery_power : 1094
blue : 2
clock_speed : 26
dual_sim : 2
fc : 20
four_g : 2
int_memory : 63
m_dep : 10
mobile_wt : 121
n_cores : 8
pc : 21
px_height : 1137
px_width : 1109
ram : 1562
sc_h : 15
sc_w : 19
talk_time : 19
three_g : 2
touch_screen : 2
wifi : 2
price_range : 4
```

Features with the unique values

```
battery_power     1238.51850
blue                 0.49500
clock_speed          1.52225
dual_sim             0.50950
fc                   4.30950
four_g               0.52150
int_memory          32.04650
m_dep                0.50175
mobile_wt          140.24900
n_cores              4.52050
pc                   9.91650
px_height          645.10800
px_width          1251.51550
ram               2124.21300
sc_h                12.30650
sc_w                 5.76700
talk_time           11.01100
three_g              0.76150
touch_screen         0.50300
wifi                 0.50700
price_range          1.50000
```

Features with the mean of their particular column

● Let's Do the Numerical Values having dataset Exploration with the particular unique values :
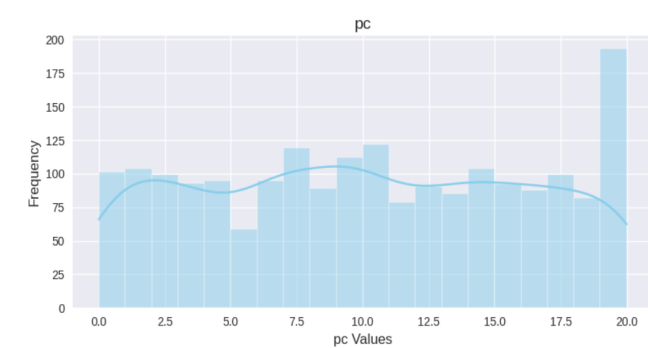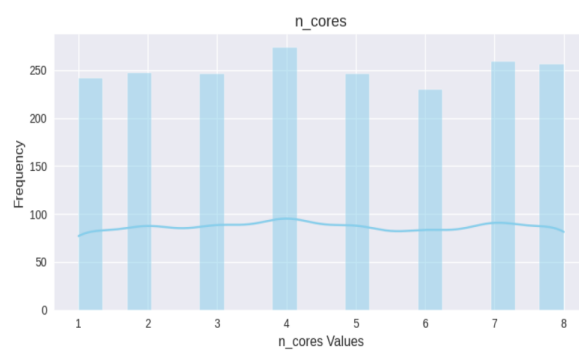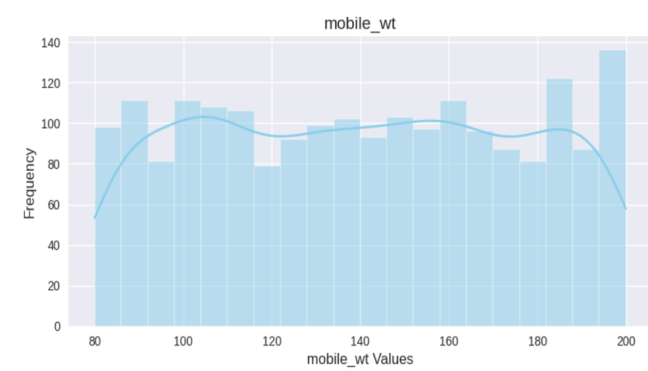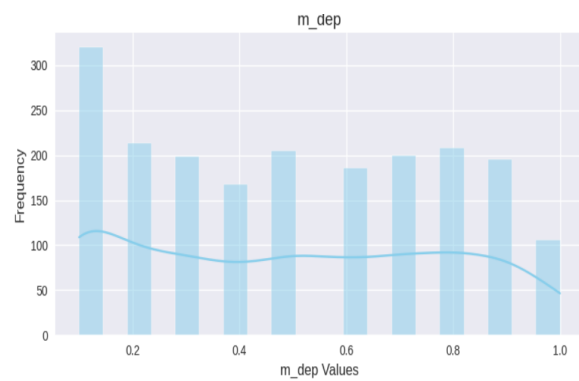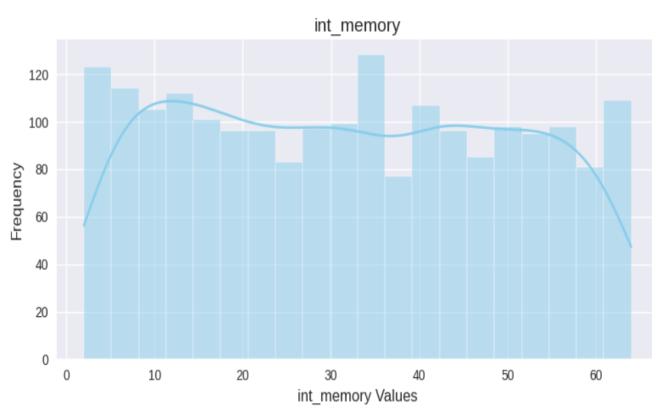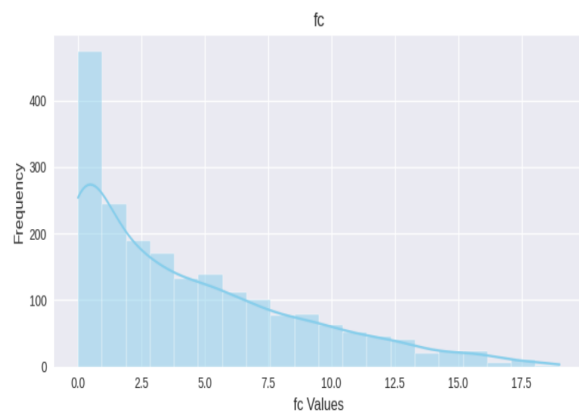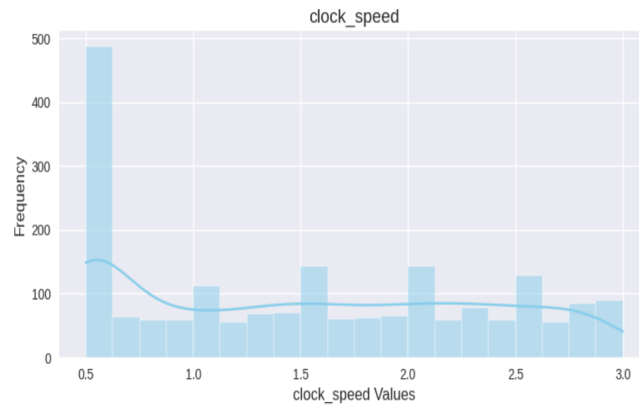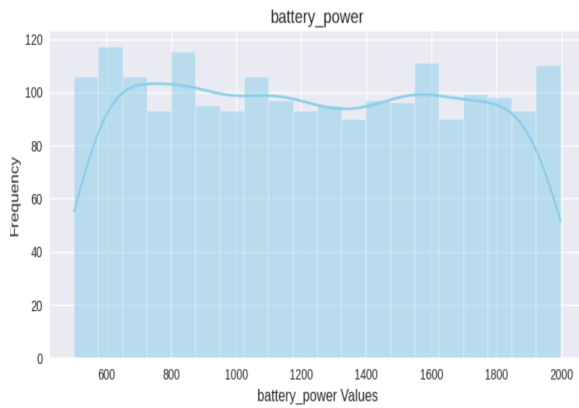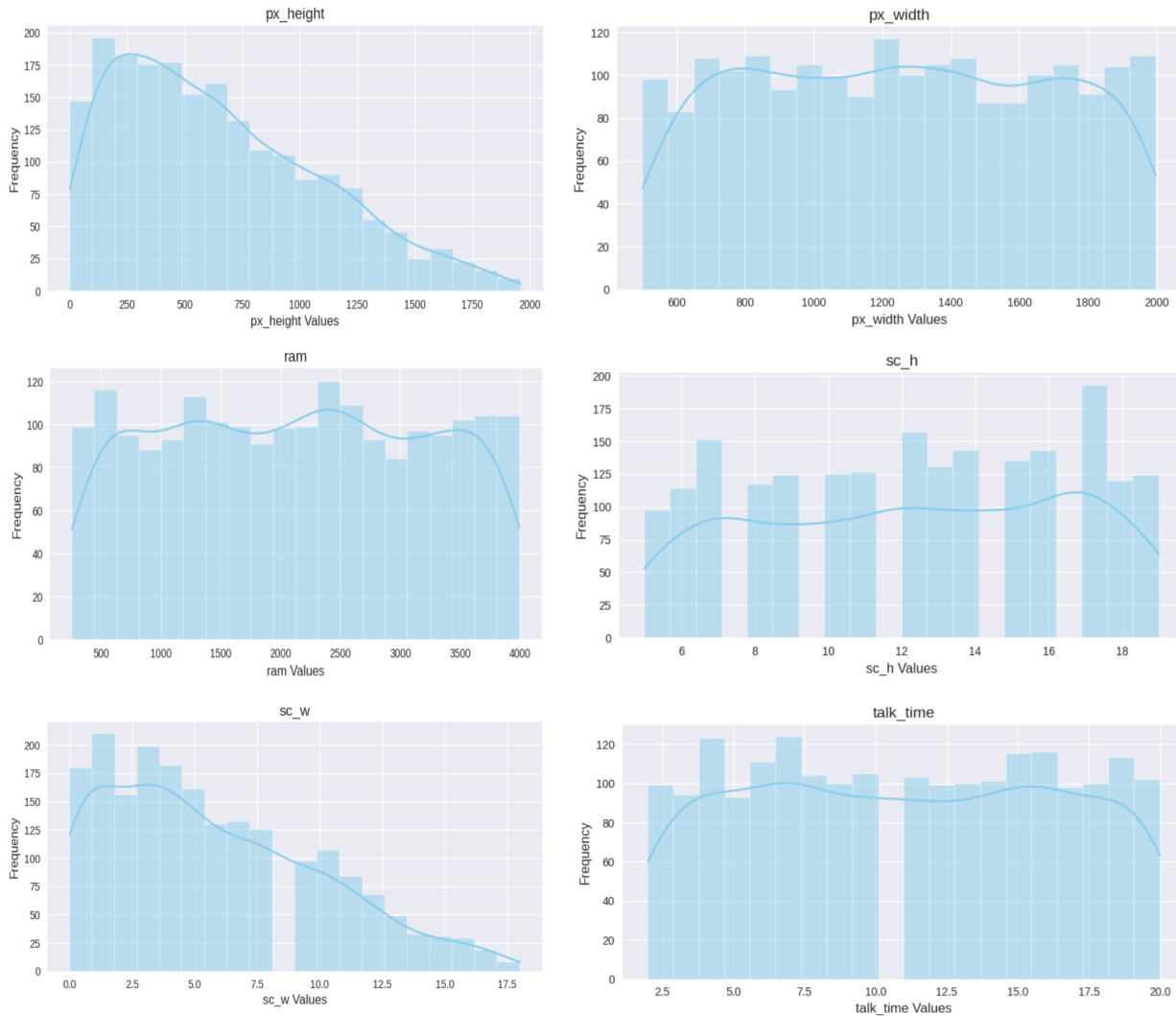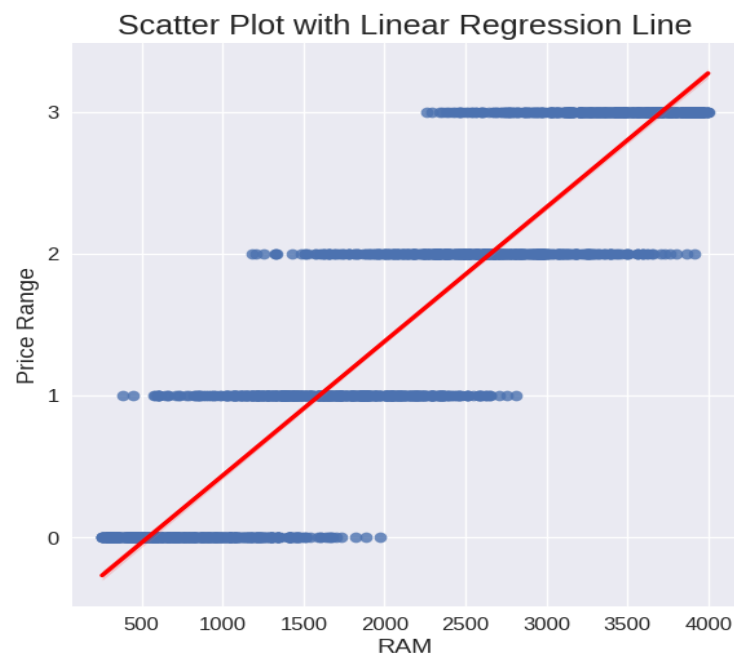
- **Now we can see that how the features are correlated with each with the help of the heatmap :**

  - The highest correlated Features are as:

    - RAM

    - Battery Power

    - Pixel Width

    - Pixel Height

- **Since the most important feature is RAM , so lets explore it with the Price Range:**

Scatter Plot with Linear Regression Line

- **We can also visualize the Battery Power with the Price range the following Bar graph :**

Battery Power's Correlation to Price Range
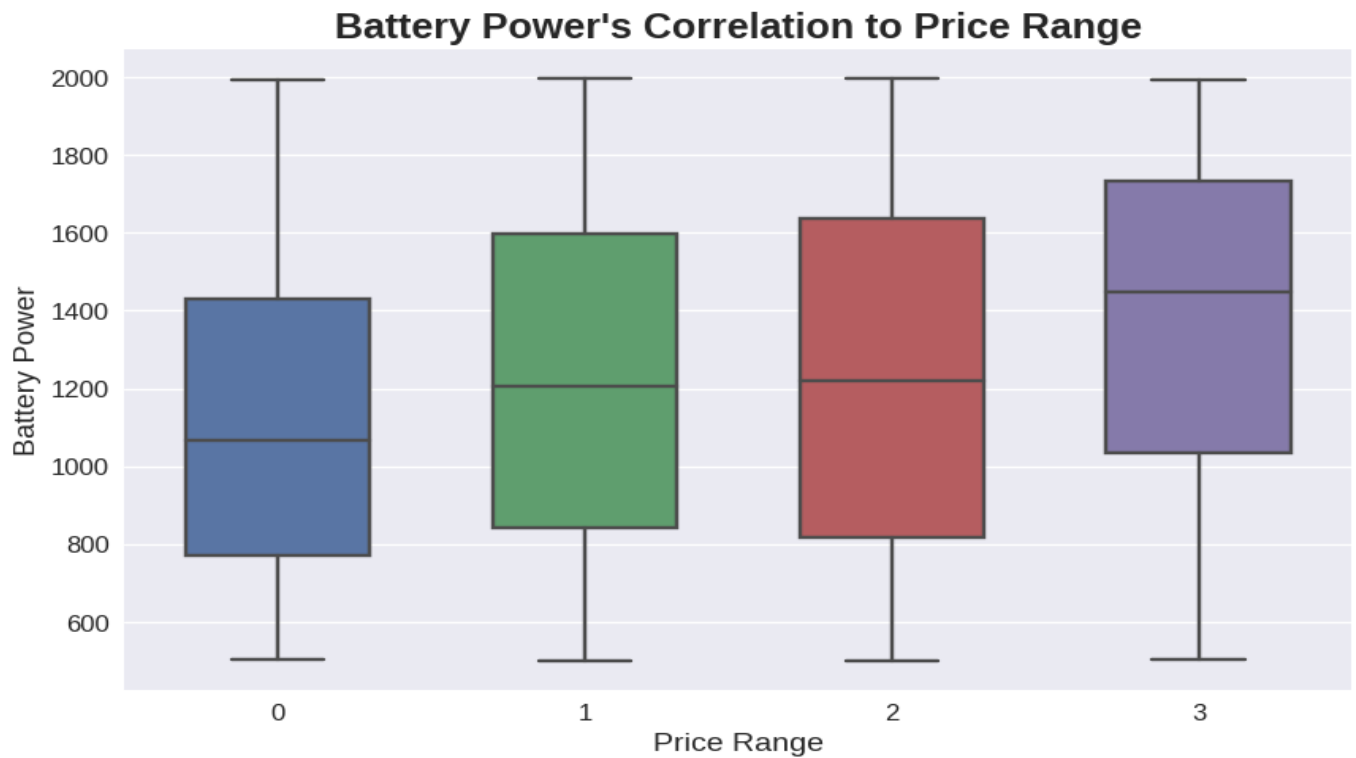
- **Lets visualize some more features with the help of the graphs :**

Percentage of Mobiles with 4G

Number of cores in mobile phones
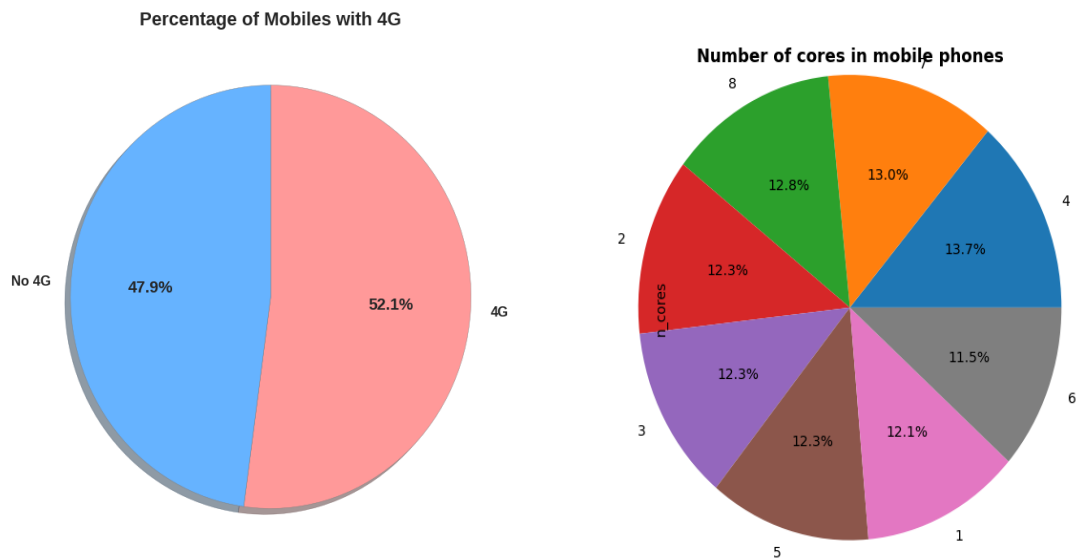
- **Outliers Removal with the help of the boxplot :**

Previously there was some outliers in the dataset which we can visualize with the help of the following Box plots :

Box Plot for sc_w   Box Plot for talk_time
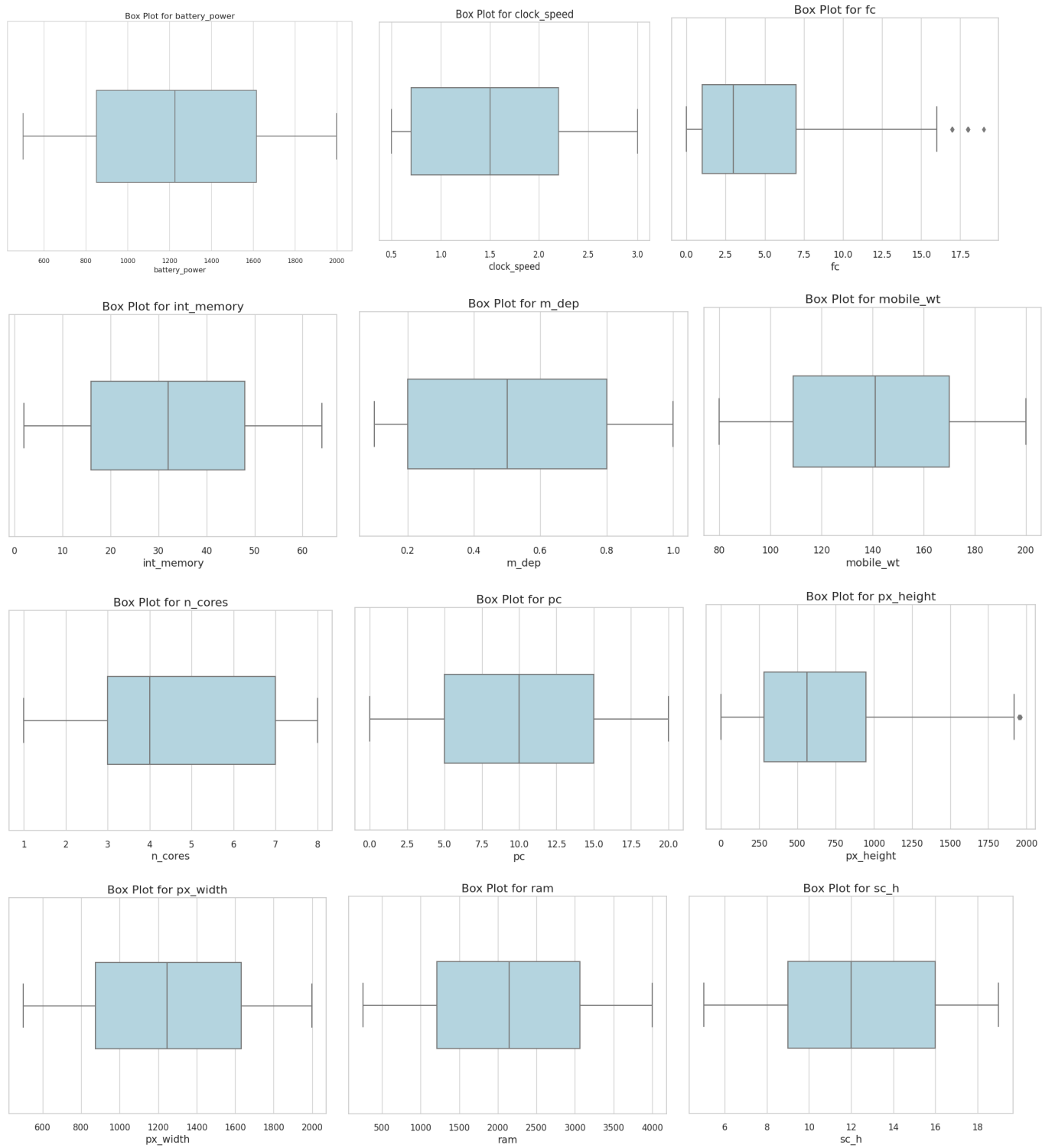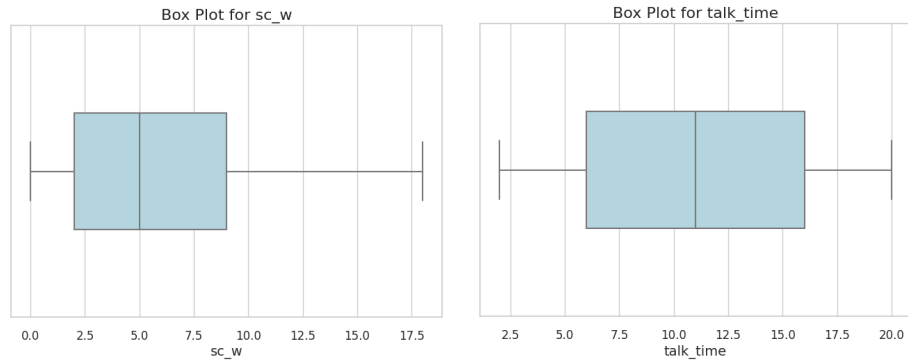
- **Appropriate ML algorithms to fit model(s) to the data :**

  - ○ **Applying Logistic Regression Model :**

These are the following reports after applying the Logistic Regression model on the dataset :

	precision	recall	f1-score	support
0	0.802817	0.754967	0.778157	151.000
1	0.517483	0.506849	0.512111	146.000
2	0.512987	0.533784	0.523179	148.000
3	0.708075	0.735484	0.721519	155.000
accuracy	0.635000	0.635000	0.635000	0.635
macro avg	0.635340	0.632771	0.633741	600.000
weighted avg	0.637419	0.635000	0.635893	600.000

Testing Accuracy we are getting as the 63.5 %

The following image shows us the Confusion matrix which summarizes the results of classification by showing the counts of: True positive, True negative, False positive, False negative.

Confusion Matrix
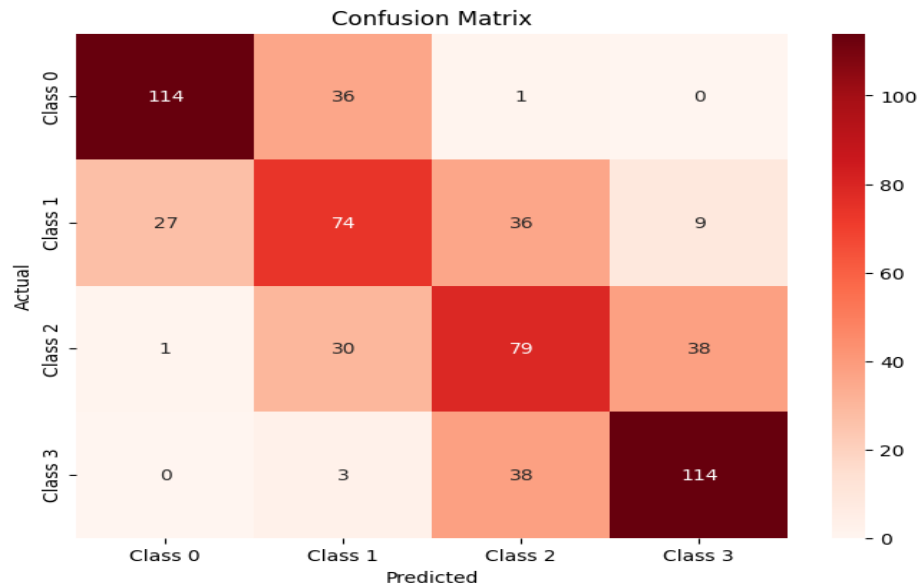
And we can also find out the top ten important variables affecting the outcome variable or variable of interest.

	Feature	Coefficient	Abs_Coefficient
8	mobile_wt	0.056074	0.056074
6	int_memory	0.018927	0.018927
14	sc_h	0.009209	0.009209
16	talk_time	0.007253	0.007253
10	pc	0.006710	0.006710
13	ram	-0.004923	0.004923
15	sc_w	0.004030	0.004030
9	n_cores	0.003598	0.003598
11	px_height	-0.002409	0.002409
4	fc	0.002093	0.002093

Top Feature Importance in Decision Tree

Important variables affecting the output of the model according to their coefficient.

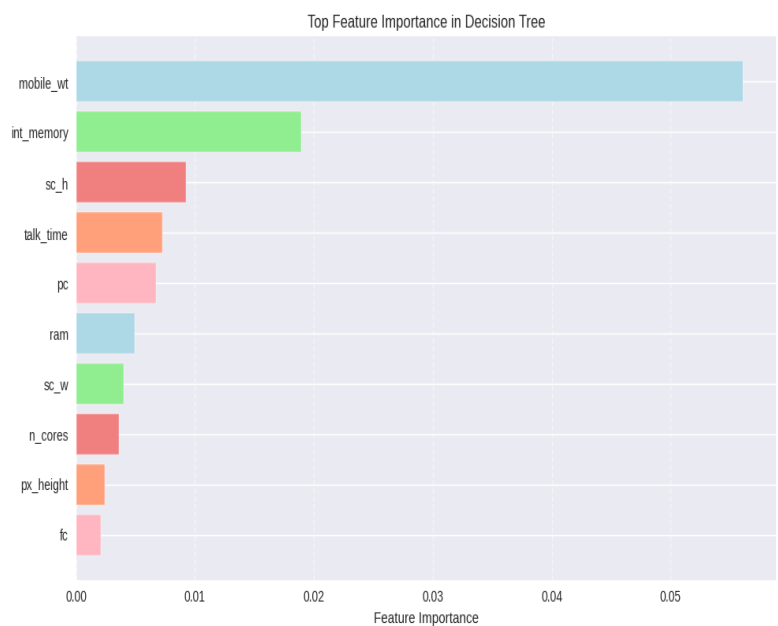○ **Applying Decision Tree Model :**

These are the following reports after applying the Decision Tree model on the dataset :

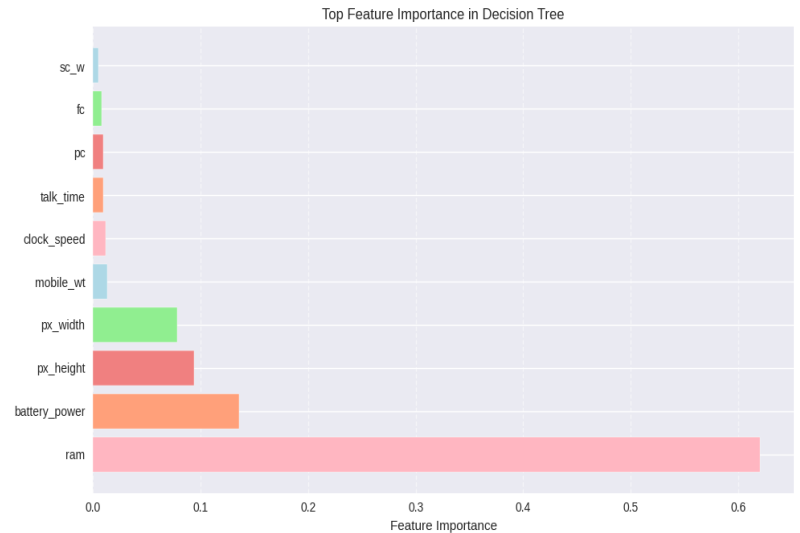|	precision	recall	f1-score	support
0	0.909722	0.867550	0.888136	151.000000
1	0.742138	0.808219	0.773770	146.000000
2	0.751724	0.736486	0.744027	148.000000
3	0.875000	0.858065	0.866450	155.000000
accuracy	0.818333	0.818333	0.818333	0.818333
macro avg	0.819646	0.817580	0.818096	600.000000
weighted avg	0.821001	0.818333	0.819158	600.000000

Testing Accuracy we are getting as the 81.83%

The following image shows us the Confusion matrix which summarizes the results of classification by showing the counts of: True positive, True negative, False positive, False negative.
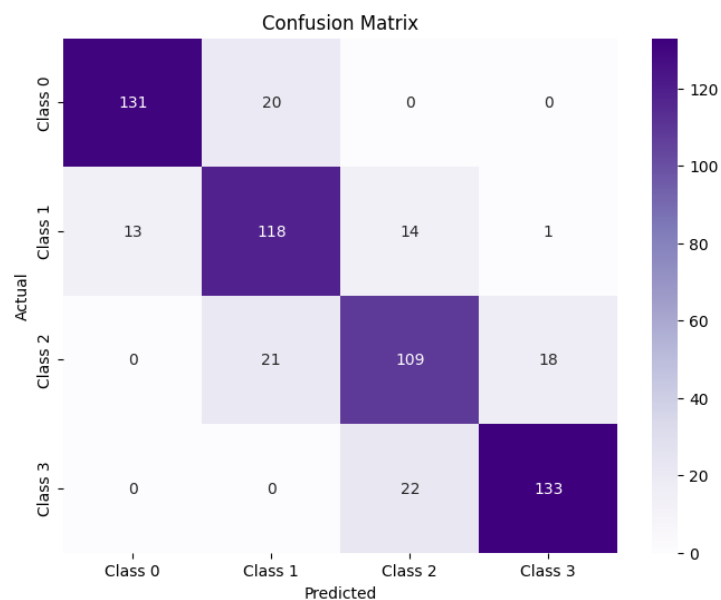
And we can also find out the top ten important variables affecting the outcome variable or variable of interest.

```
     Feature   Importance
13        ram     0.620447
0   battery_power 0.135470
11    px_height   0.093714
12    px_width    0.078303
8     mobile_wt   0.013615
2   clock_speed   0.011501
16    talk_time   0.009720
10         pc     0.009205
4          fc     0.007990
15       sc_w     0.005278
```


Top Feature Importance in Decision Tree

Important variables affecting the output of the model according to their coefficient.

○ **Applying Random Forest Model :**

These are the following reports after applying the Decision Tree model on the dataset :

	precision	recall	f1-score	support
0	0.929032	0.953642	0.941176	151.00
1	0.834483	0.828767	0.831615	146.00
2	0.805556	0.783784	0.794521	148.00
3	0.903846	0.909677	0.906752	155.00
accuracy	0.870000	0.870000	0.870000	0.87
macro avg	0.868229	0.868968	0.868516	600.00
weighted avg	0.869061	0.870000	0.869449	600.00

Testing Accuracy we are getting as the 87%

The following image shows us the Confusion matrix which summarizes the results of classification by showing the counts of: True positive, True negative, False positive, False negative.
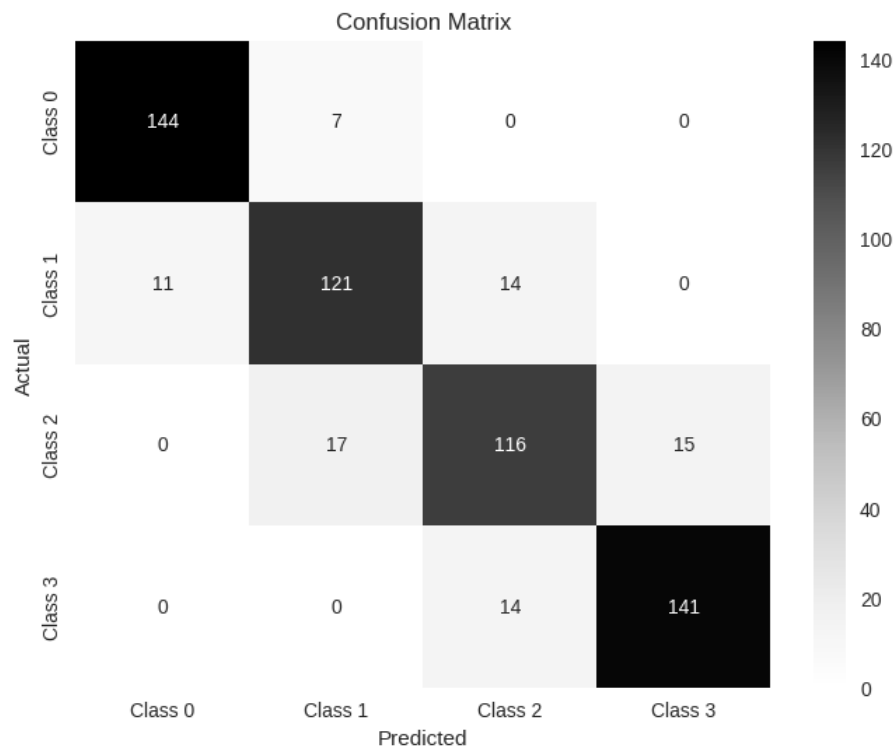
And we can also find out the top ten important variables affecting the outcome variable or variable of interest.

Important variables affecting the output of the model according to their coefficient.

The Link of the both codes are given here :

For the Part 1 : Code 1
For the Part 2 : Code 2