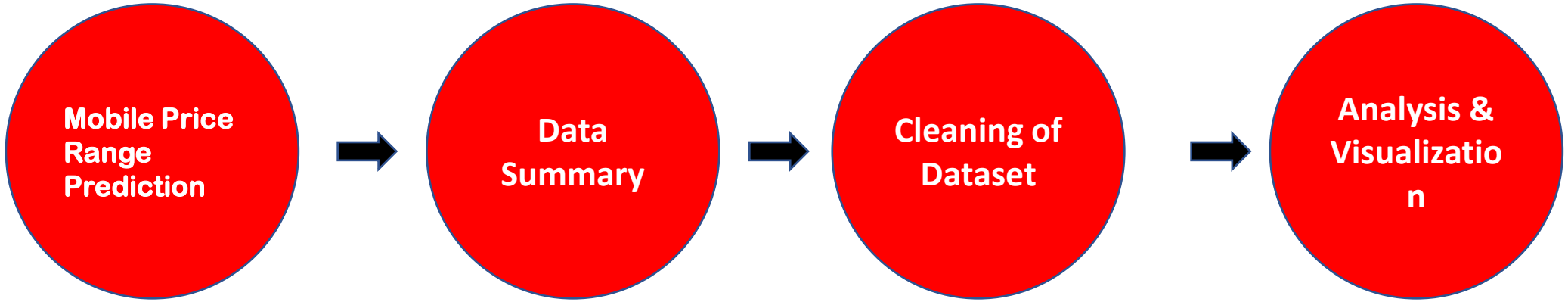
# Capstone Project

# Mobile Price Range Prediction

**Rushikesh s Shete (rushishete101@gamil.com)**

# Problem statement :

In the competitive mobile phone market companies wantto understand sales data of mobile phones and factors which drive the prices. The objective is to find out some relation between features of a mobile phone(eg:- RAM,Internal Memory, etc) and its selling price. In this problem, we do not have to predict theactual price but a price range indicating how high the price is.



# Points to discuss :

- Data description and summary
- Exploratory data analysis
- Heat map
- Machine learning algorithms
  1. Logistic regression
  2. Decision tree
  3. Random forest classifier
  4. Xgboost  classifier
- conclusion

# Data description :

The data contains information regarding mobile phone features, specifications etc and their price range. The various features and information can be used to predict the price range of a mobile phone.

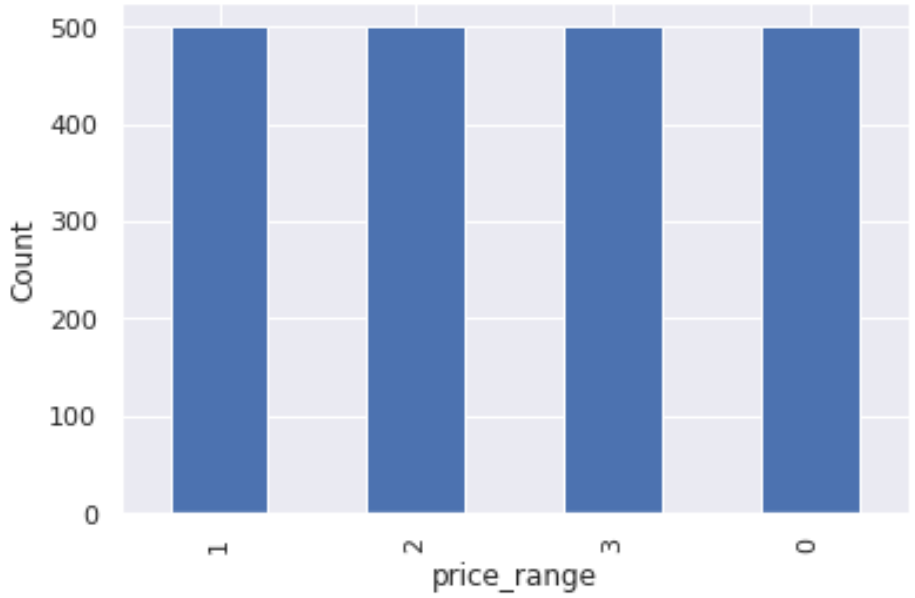
- Battery_power - Total energy a battery can store in one time measured inmAh
- Blue - Has bluetooth or not
- Clock_speed - speed at which microprocessor executes instructions
- Dual_sim - Has dual sim support or not
- Fc - Front Camera mega pixels
- Four_g - Has 4G or not
- Int_memory - Internal Memory in Gigabytes
- M_dep - Mobile Depth in cm
- Mobile_wt - Weight of mobile phone

# Data description :

- N_cores - Number of cores of processor
- Pc - Primary Camera mega pixels
- Px_height - Pixel Resolution Height
- Px_width - Pixel Resolution Width
- Ram - Random Access Memory in Mega Bytes
- Sc_h - Screen Height of mobile in cm
- Sc_w - Screen Width of mobile in cm
- Talk_time - longest time that a single battery charge will last when you are
- Three_g - Has 3G or not
- Touch_screen - Has touch screen or not
- Wifi - Has wifi or not
- Price_range - This is the target variable with value of 0(low cost), 1(medium cost),
- 2(high cost) and 3(very high cost).

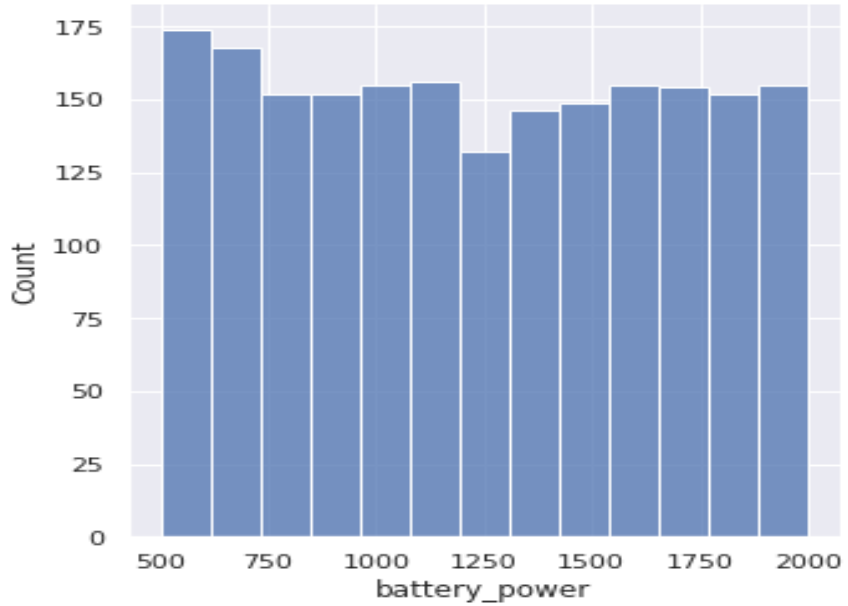
# Data Visualization

## price




● there are mobile phones in 4 price ranges. the number of elements is almost similar

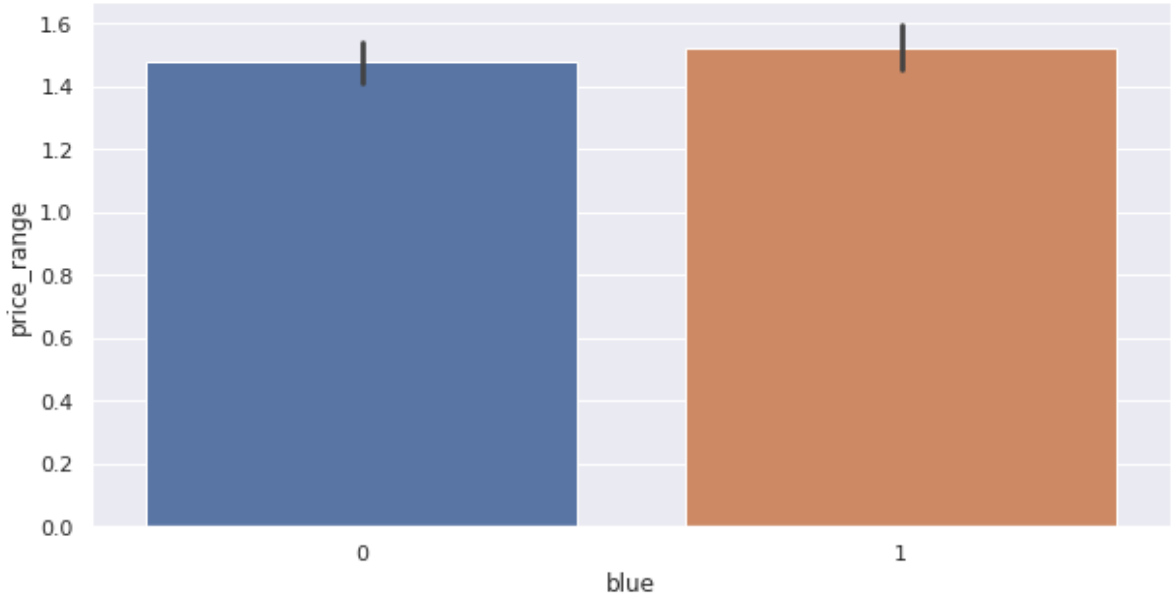
## Battery




● this plot shows how the battery mAh is spread. there is a gradual increase as the price range increases

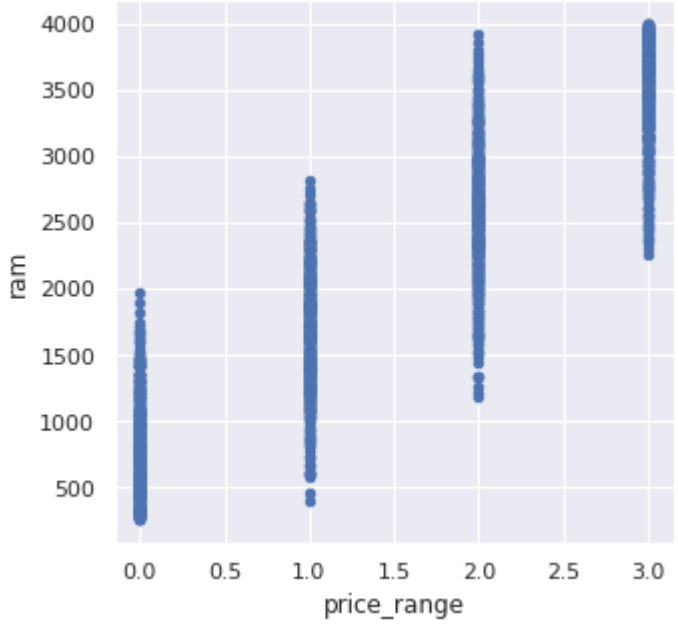
# Data Visualization

## Bluetooth




- half the devices have Bluetooth, and half don’t

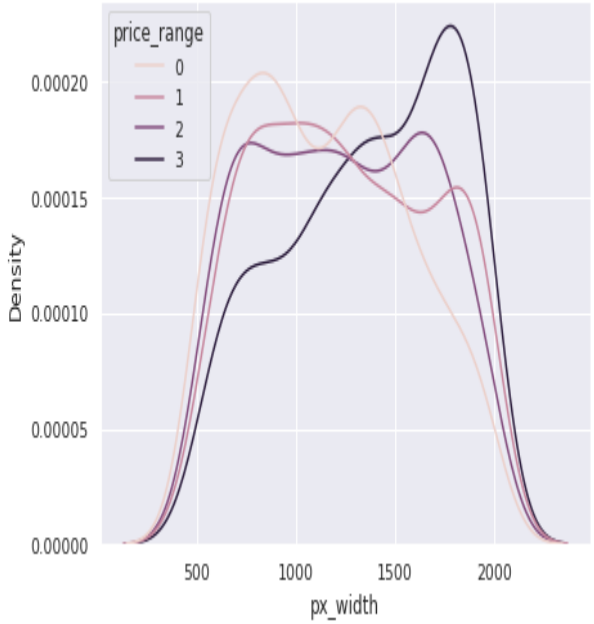
## RAM

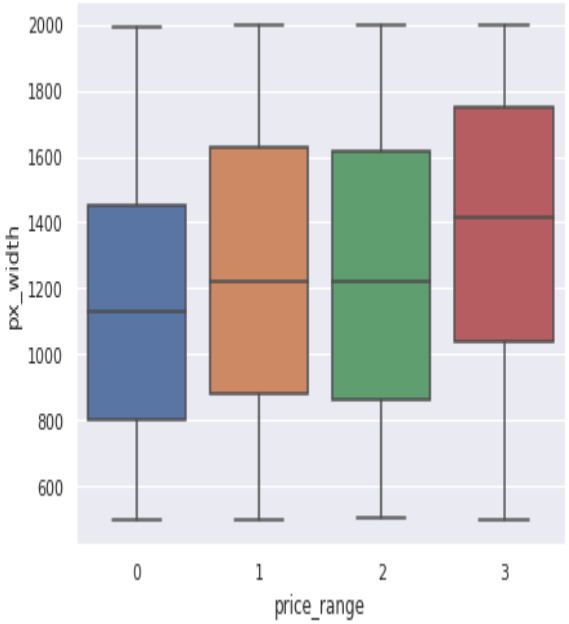



- Ram has continuous increase with price range while moving from Low cost to Very high cost

# Data Visualization

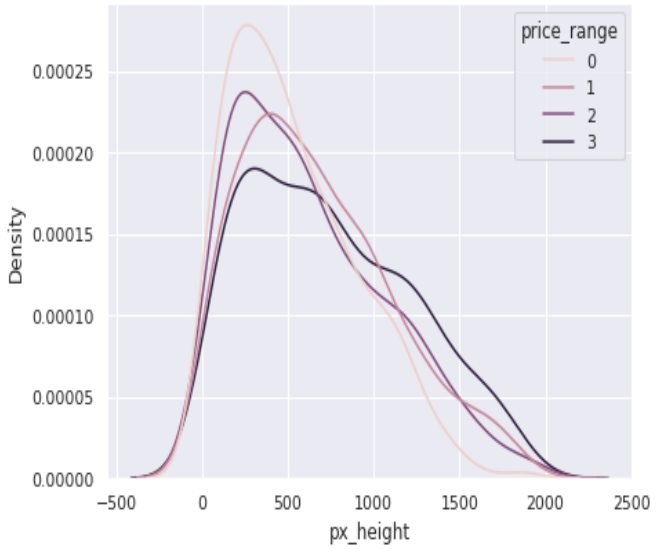
## Px_width

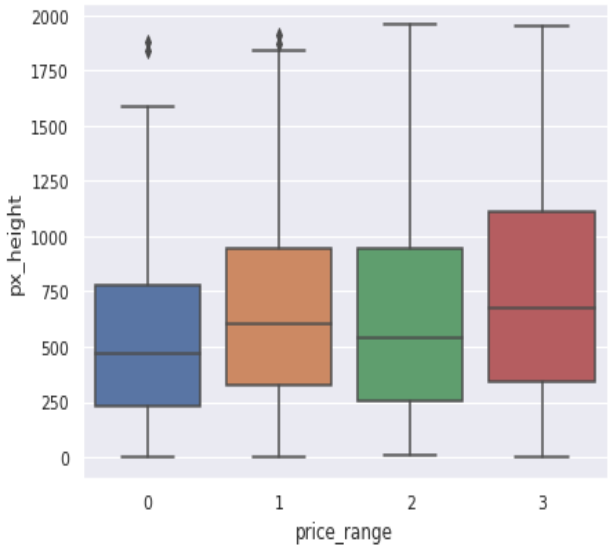



- There is not a continuous increase in pixel width as we move from Low cost to Very high cost. Mobiles with 'Medium cost' and 'High cost' has almost equal pixel width. so we can say that it would be a driving factor in deciding price_range.

## px_height

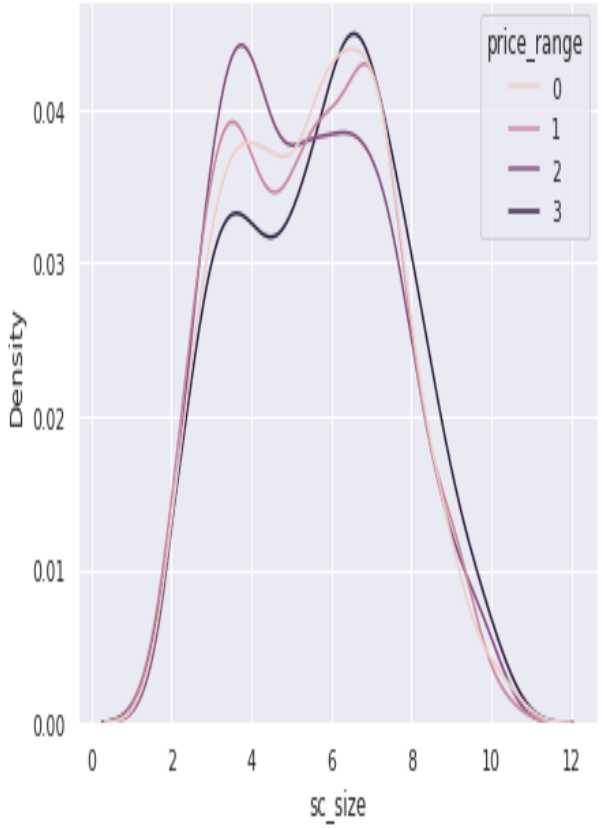


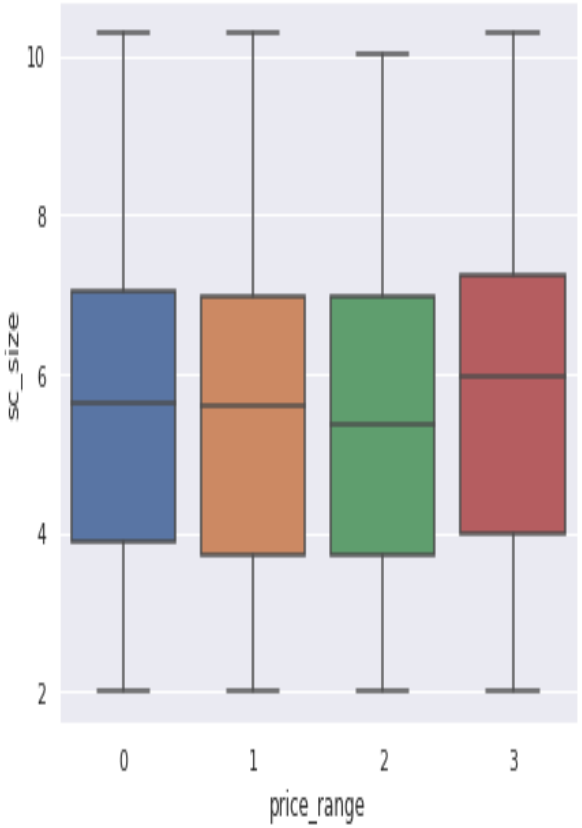




- Pixel height is almost similar as we move from Low cost to Very high cost.little variation in pixel_height

# Data Visualization

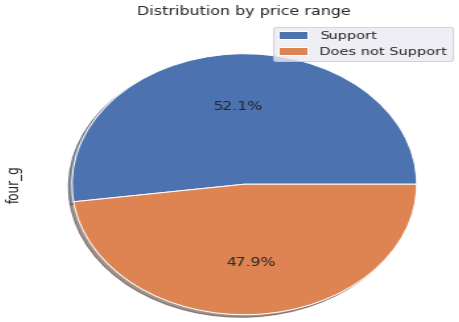
## screen_size

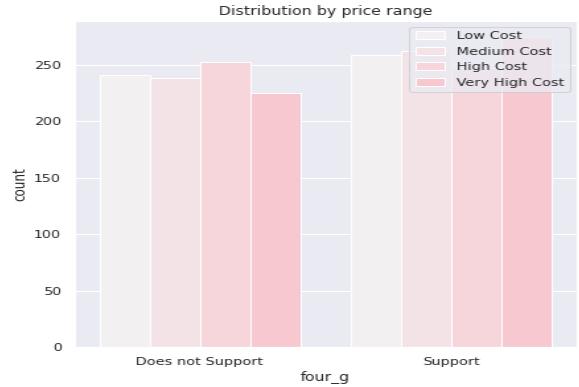


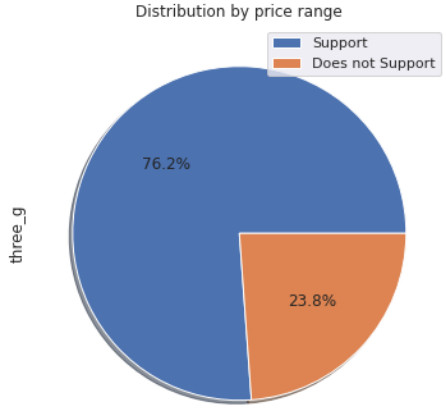



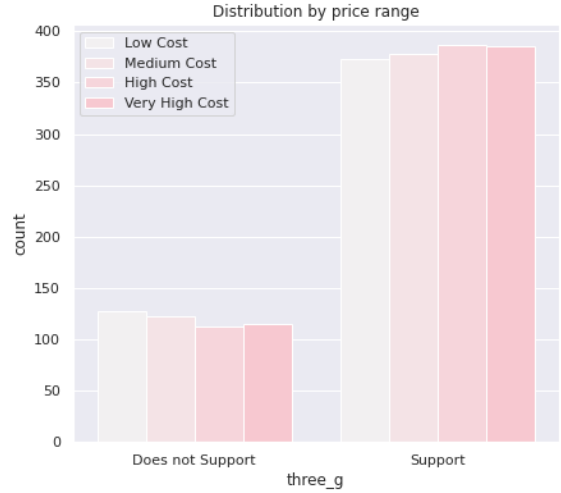

- Combining the sc_height and sc_width into one column that is sc_size, Screen Size shows little variation along the target variables. This can be helpful in predicting the target categories

## 4G and 3G



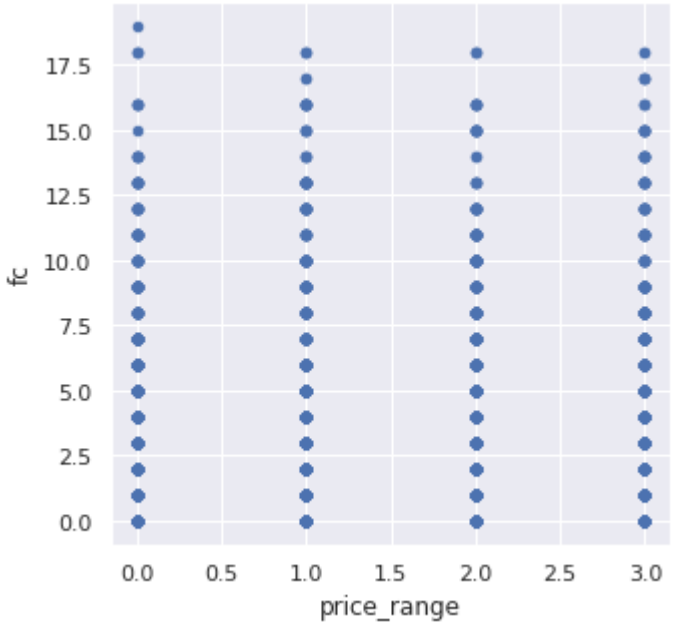




- 50% of the phones support 4_g and 76% of phones support 3_g,feature 'three_g' play an important feature in prediction

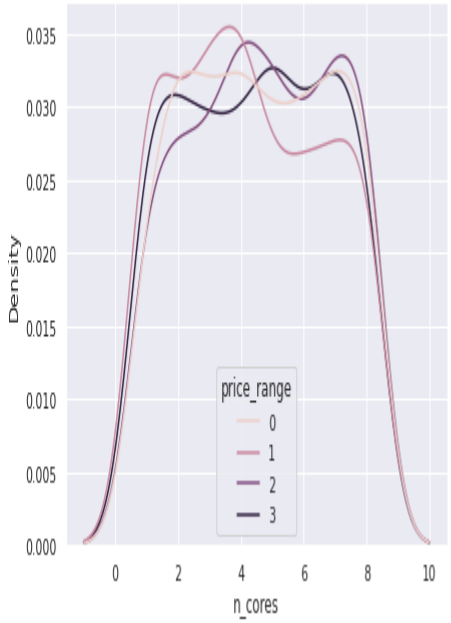
# Data Visualization

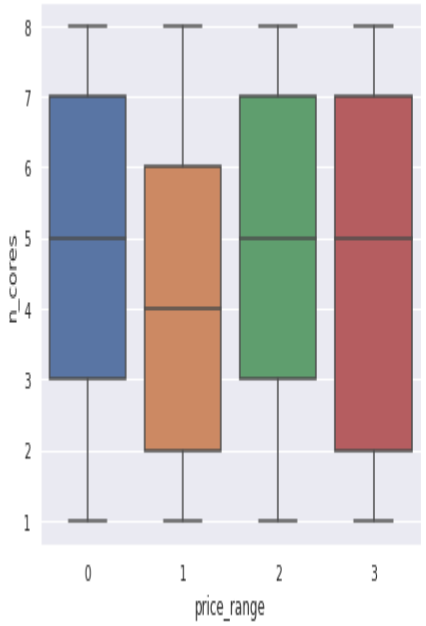
## FC (front camera megapixels)




● This features distribution is almost similar along all the price ranges variable, it may not be helpful in making predictions

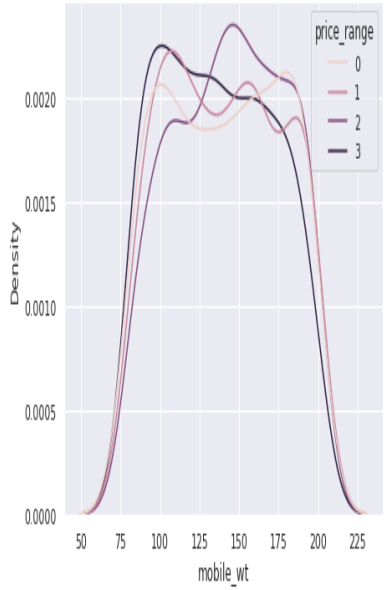
## PC (Primary camera Megapixels )

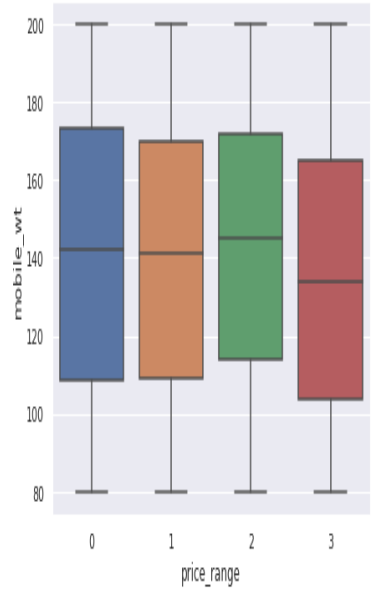



● Primary camera megapixels are showing a little variation along the target categories, which is a good sign for prediction.

## mobile weight



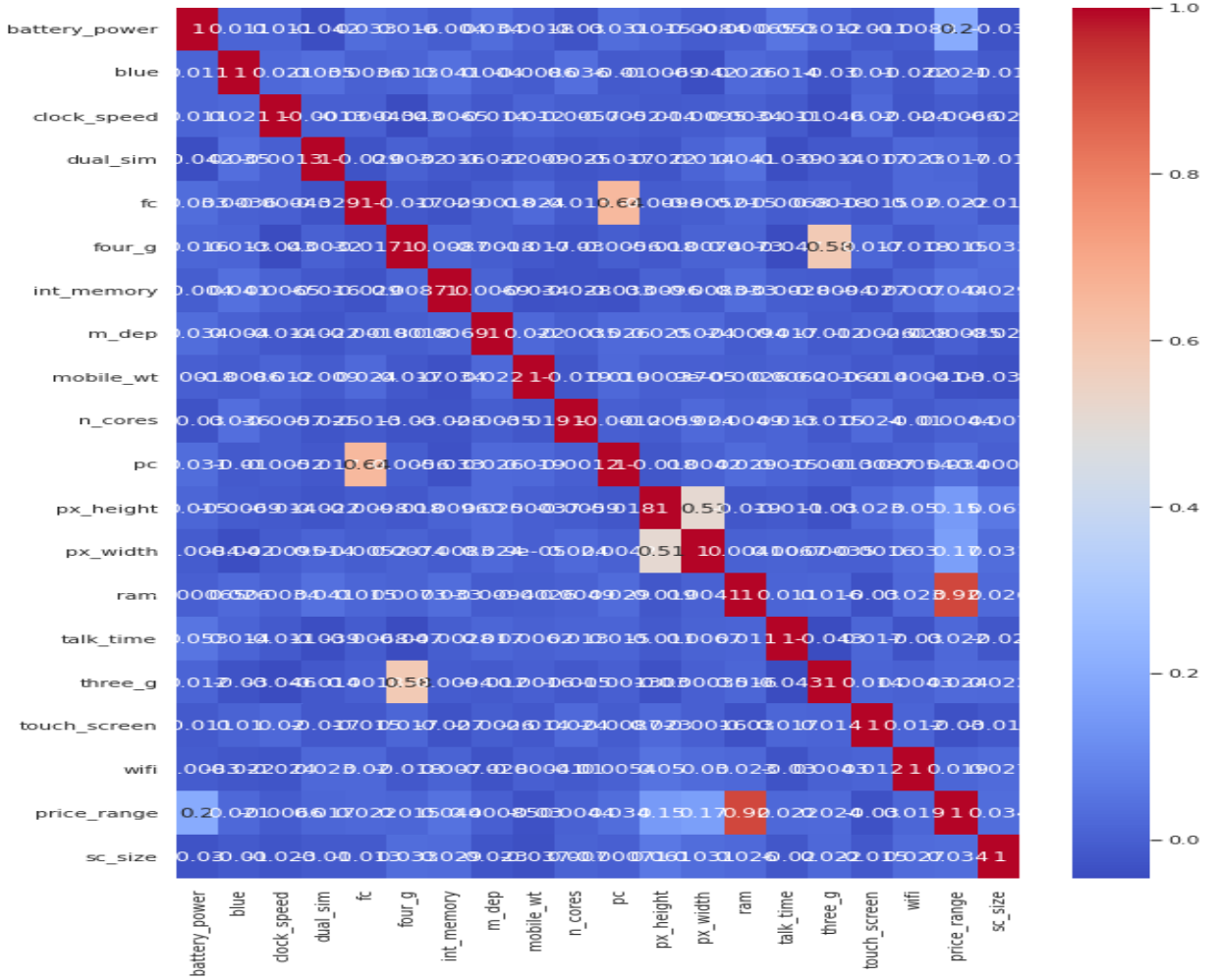




● costly phones are lighter

# Data Visualization

## Heat map

- RAM and price_range shows high correlation which is a good sign, it signifies that RAM will play major deciding factor in estimating the price range.
- There is some collinearity in feature pairs ('pc', 'fc') and ('px_width', 'px_height'). Both correlations are justified since there are good chances that if front camera of a phone is good, the back camera would also be good.
- Also, if px_height increases, pixel width also increases, that means the overall pixels in the screen. We can replace these two features with one feature. Front Camera megapixels and Primary camera megapixels are different entities despite of showing colinearity. So we'll be keeping them as they are.



# Data Visualization

## ML algorithms :

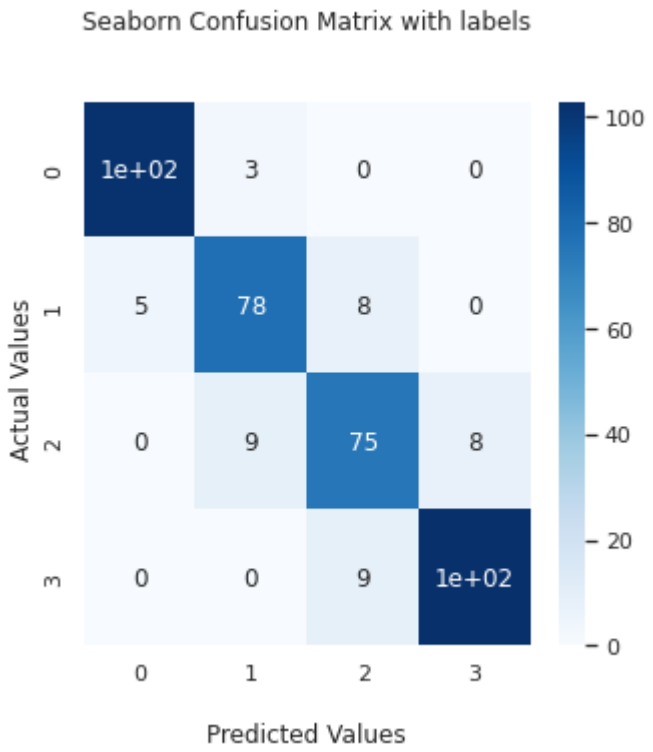
1. Logistic regression
2. Decision tree
3. Random Forest classification
4. XGboost

# *Feature Engineering & Data Pre-processing*

- **Logistic regression**

```
from sklearn.metrics import classification_report
print('Classification report for Logistic Regression (Test set)= ')
print(classification_report(y_pred_test, y_test))
```

```
Classification report for Logistic Regression (Test set)=
              precision    recall  f1-score   support

           0       0.97      0.95      0.96       107
           1       0.86      0.87      0.86        90
           2       0.82      0.82      0.82        92
           3       0.92      0.93      0.92       111

    accuracy                           0.90       400
   macro avg       0.89      0.89      0.89       400
weighted avg       0.90      0.90      0.90       400
```




Train_accuracy : 92% Test_accuracy: 90%

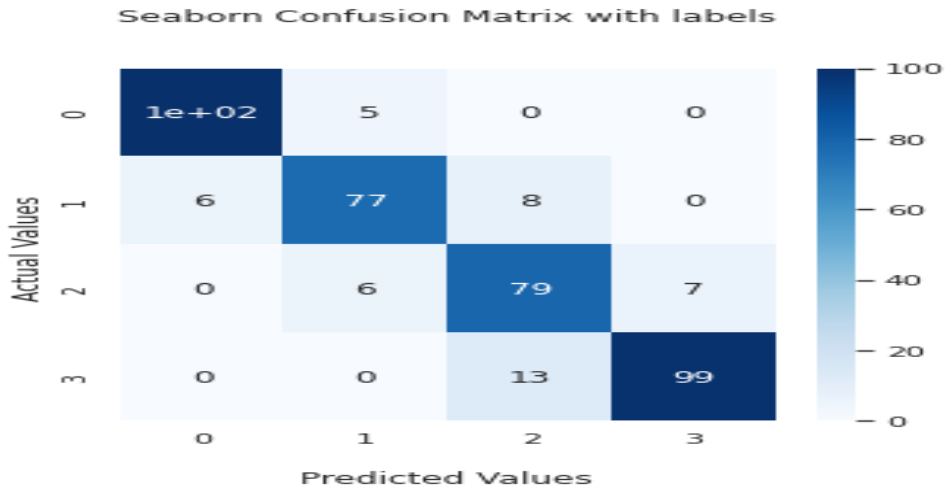
Logistic Regression performed better than the other two model

# *Feature Engineering & Data Pre-processing*

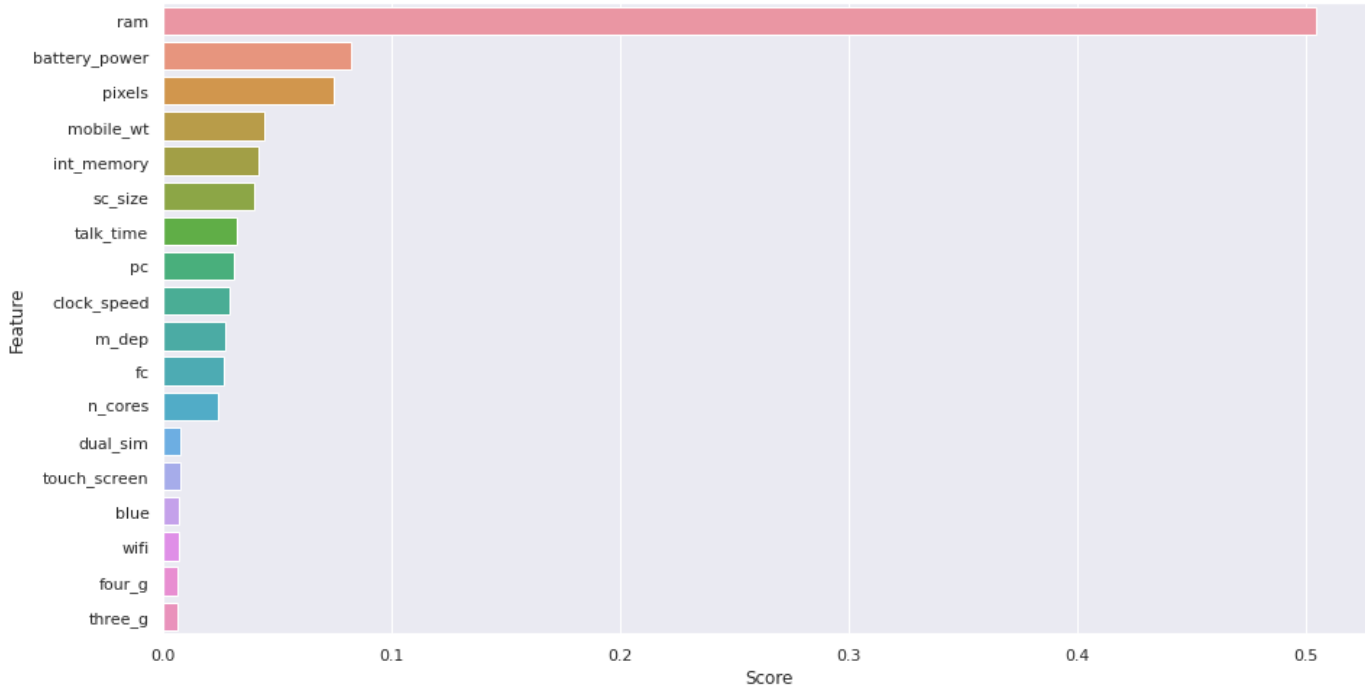
- **Random forest classifier with hyper parameter tuning :**

```
In [47]: print(classification_report(y_test, y_pred))
```

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.94 | 0.95 | 0.95 | 105 |
| 1 | 0.88 | 0.85 | 0.86 | 91 |
| 2 | 0.79 | 0.86 | 0.82 | 92 |
| 3 | 0.93 | 0.88 | 0.91 | 112 |
| accuracy | | | 0.89 | 400 |
| macro avg | 0.89 | 0.89 | 0.88 | 400 |
| weighted avg | 0.89 | 0.89 | 0.89 | 400 |




Train_accuracy : 86.5%




As we can see the top 3 important features of our dataset are: RAM, battery_power ,pixels

•RAM, battery power, pixels played more significant role in deciding the price range of mobile phone.

# *Feature Engineering & Data Pre-processing*

## Decision tree

Test  accuracy : 84%

In [66]:
```
# Evaluation metrics for test

print('Classification report for Decision Tree (Test set)= ')
print(classification_report(y_pred_test, y_test))
```

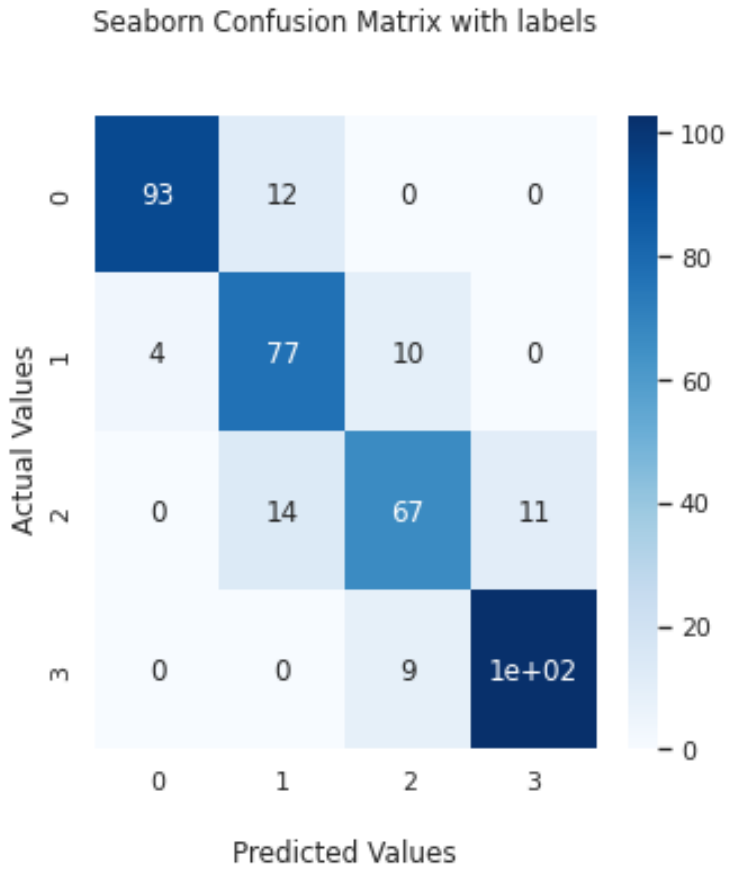
```
Classification report for Decision Tree (Test set)=
              precision    recall  f1-score   support

           0       0.87      0.98      0.92        93
           1       0.81      0.73      0.77       101
           2       0.78      0.67      0.72       108
           3       0.81      0.93      0.87        98

    accuracy                           0.82       400
   macro avg       0.82      0.83      0.82       400
weighted avg       0.82      0.82      0.82       400
```

## Decision tree with hyperparameter tuning

Test_accuracy : 82%

In [68]:
```
# Prediction

y_pred_test = grid.predict(X_test)
y_pres_train = grid.predict(X_train)
# Evaluation metrics for test

print('Classification Report for Decision Tree (Test set)= ')
print(classification_report(y_test, y_pred_test))
```

```
Classification Report for Decision Tree (Test set)=
              precision    recall  f1-score   support

           0       0.96      0.89      0.92       105
           1       0.75      0.85      0.79        91
           2       0.78      0.73      0.75        92
           3       0.90      0.92      0.91       112

    accuracy                           0.85       400
   macro avg       0.85      0.84      0.84       400
weighted avg       0.85      0.85      0.85       400
```



# *Feature Engineering & Data Pre-processing*

## XGboost

Test_accuracy : 89%

In [71]:

```
# Applying XGBoost

from xgboost import XGBClassifier

xgb = XGBClassifier(max_depth = 5, learning_rate = 0.1)
xgb.fit(X_train, y_train)
XGBClassifier(max_depth=5, objective='multi:softprob')
# Prediction

y_pred_train = xgb.predict(X_train)
y_pred_test = xgb.predict(X_test)
# Evaluation metrics for test

score = classification_report(y_test, y_pred_test)
print('Classification Report for XGBoost(Test set)= ')
print(score)
```

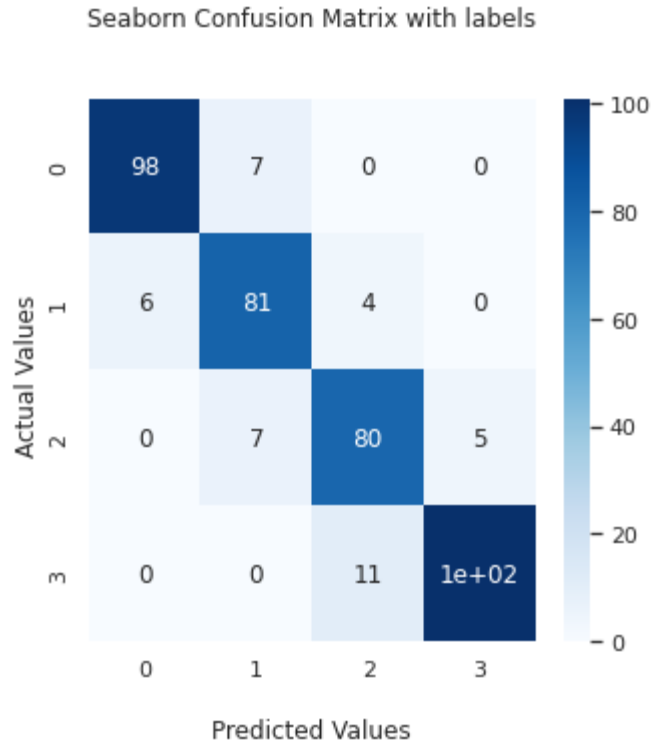
```
Classification Report for XGBoost(Test set)=
              precision    recall  f1-score   support

           0       0.95      0.93      0.94       105
           1       0.83      0.88      0.86        91
           2       0.81      0.84      0.82        92
           3       0.94      0.89      0.92       112

    accuracy                           0.89       400
   macro avg       0.88      0.89      0.88       400
weighted avg       0.89      0.89      0.89       400
```

## XGboost with hyperparameter tuning

Test_accuracy : 90%

```
y_pred_train = grid.predict(X_train)
y_pred_test = grid.predict(X_test)
# Evaluation metrics for test

score = classification_report(y_test, y_pred_test)
print('Classification Report for tuned XGBoost(Test set)= ')
print(score)
```

```
Classification Report for tuned XGBoost(Test set)=
              precision    recall  f1-score   support

           0       0.94      0.93      0.94       105
           1       0.85      0.89      0.87        91
           2       0.84      0.87      0.86        92
           3       0.95      0.90      0.93       112

    accuracy                           0.90       400
   macro avg       0.90      0.90      0.90       400
weighted avg       0.90      0.90      0.90       400
```



# Conclusions :

- From EDA we can see that here are mobile phones in 4 price ranges. The number of elements is almost similar.
- half the devices have Bluetooth, and half don’t
- there is a gradual increase in battery as the price range increases
- Ram has continuous increase with price range while moving from Low cost to Very high cost
- costly phones are lighter
- RAM, battery power, pixels played more significant role in deciding the price range of mobile phone.
- form all the above experiments we can conclude that logistic regression and, XGboosting with using hyperparameters we got the best results
- Logistic regression shows approx accuracy of 90% and the other two shows approx 80% accuracy, which shows that the data were properly classified.
- The accuracy and performance of the model is evaluated by using confusion matrix


***You have successfully completed your Machine Learning Capstone Project !!!***


***You have successfully completed your Machine Learning Capstone Project !!!***