Capstone Project - 3

Mobile Price Range Prediction

Team Members

Puroshotam Kumar Singh

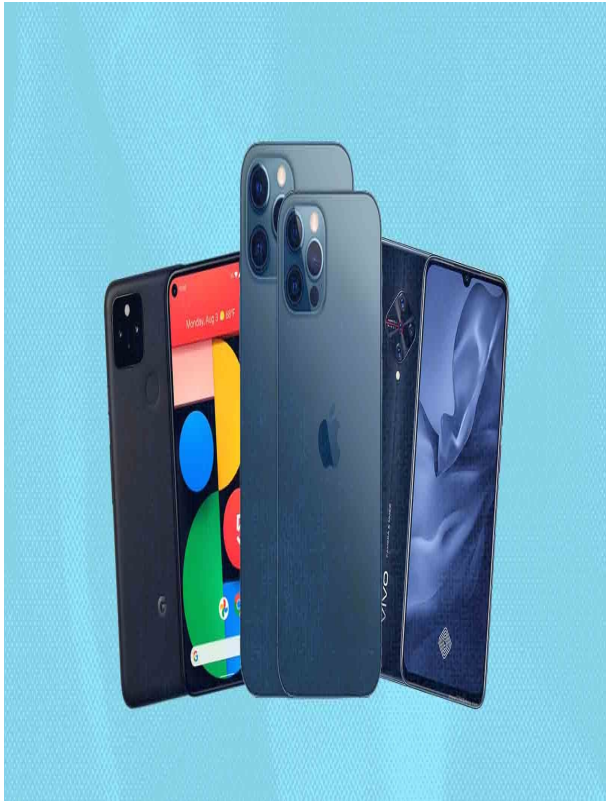
Vikram Pandey

Content

Problem Statement

- Price of a mobile phone is influenced by various factors. Brand name, newness of the model, specifications such as internal memory, camera, ram, sizes, connectivity etc., are some of the important factors in determining the price. As a business point of view, it becomes an utmost priority to analyse these factors from time to time and come up with best set of specifications and price ranges so that people buy their mobile phones.

- Hence, through this exercise and our predictions we will try to help companies estimate price of mobiles to give tough competition to other mobile manufacturer and also it will be useful for customers to verify that they are paying best price for a mobile.

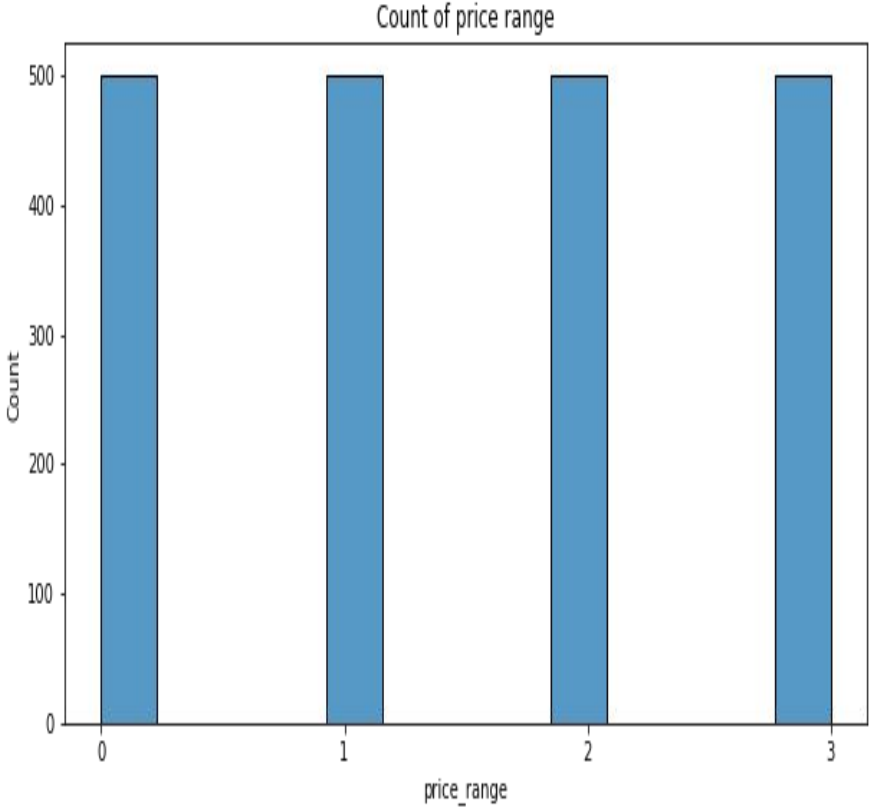


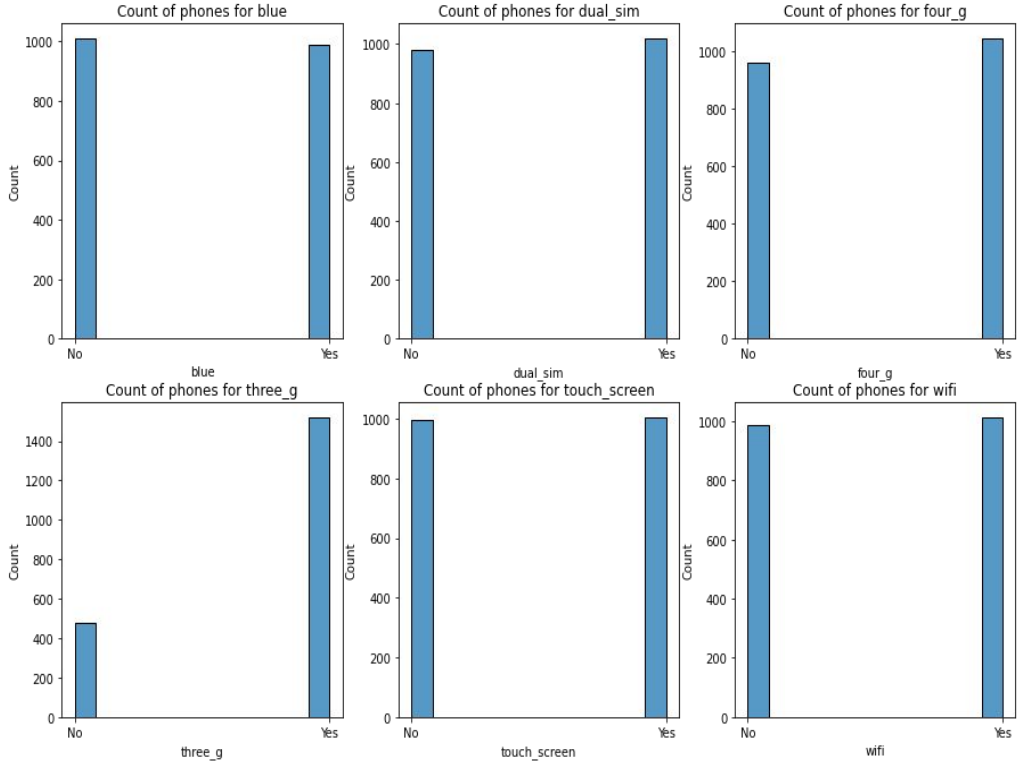
Data Summary

- **The dataset contains 2000 rows and 21 columns.**
- **The contents of the data had these features:**
  - **Battery_power -** Total energy a battery can store in one time measured in mAh
  - **Clock_speed -** speed at which microprocessor executes instructions
  - **Fc , Pc -** Front and Primary Camera megapixels
  - **Int_memory -** Internal Memory in Gigabytes
  - **M_dep -** Mobile Depth in cm
  - **Mobile_wt -** Weight of mobile phone
  - **N_cores -** Number of cores of processor
  - **Px_height, Px_width -** Pixel Resolution Height and Width
  - **Ram -** Random Access Memory in Megabytes
  - **Sc_h, Sc_w -** Screen Height and width of mobile in cm
  - **Talk_time -** longest time that a single battery charge will last when you are on call
  - **Blue, 4g, 3g, dual_sim, touchscreen, wifi -** Some supported and unsupported categories
  - **Price_range -** This is the target variable with value of 0(low cost), 1(medium cost), 2(high cost) and 3(very high cost).

Exploratory Data Analysis

EDA - Univariate analysis

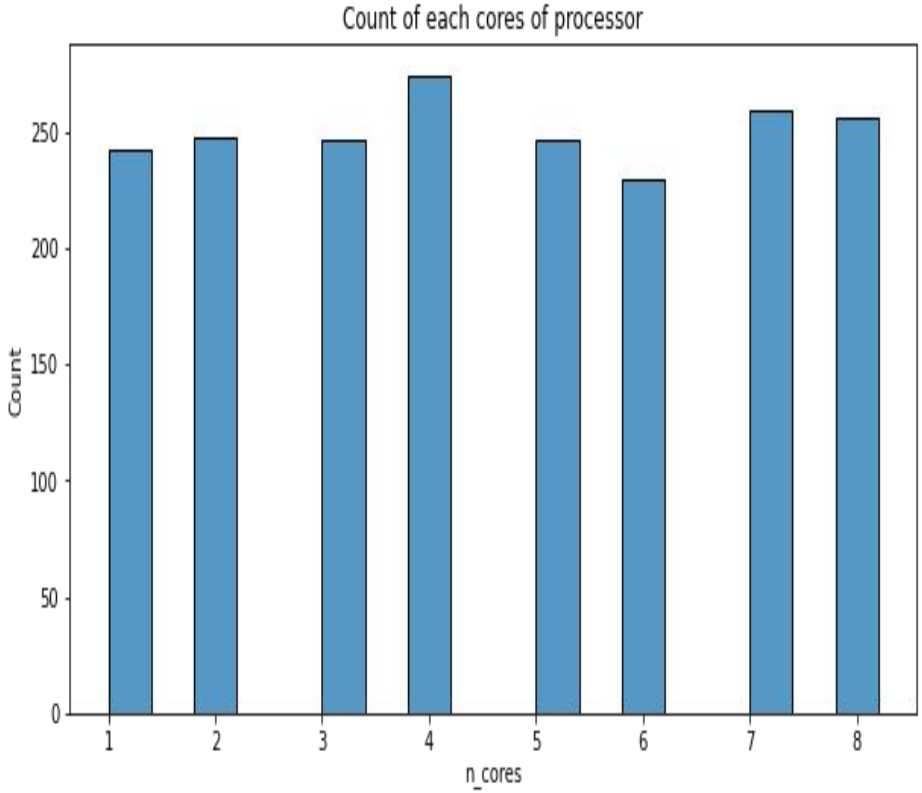




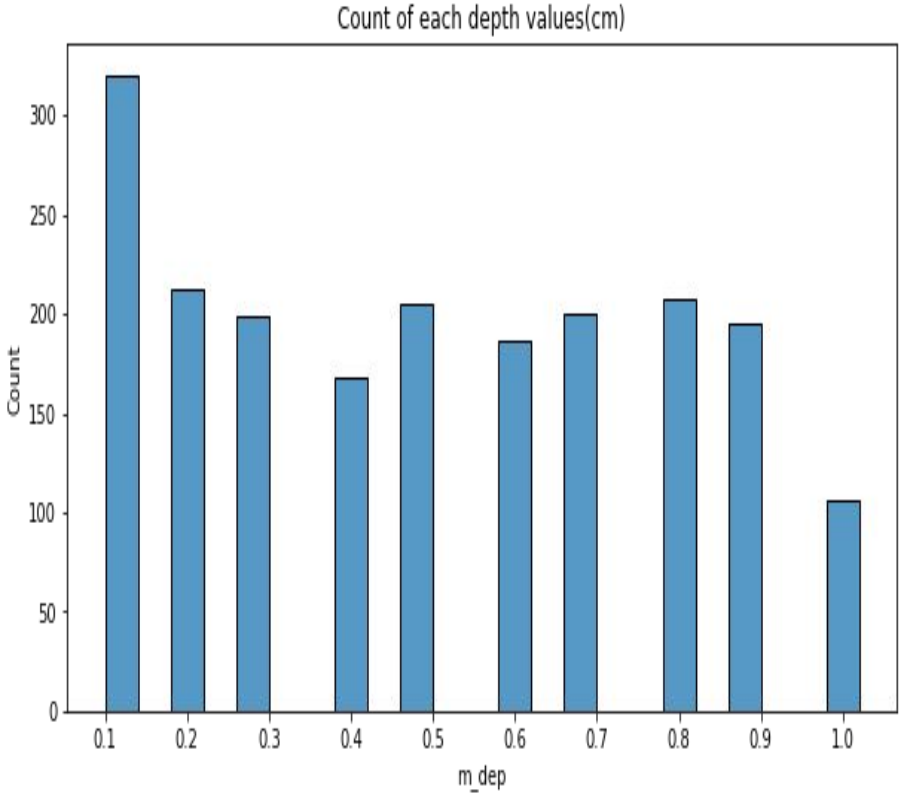
Price_range has equal no of observations

Except 3g, other dichotomous types have equal no of observations

EDA - Univariate analysis



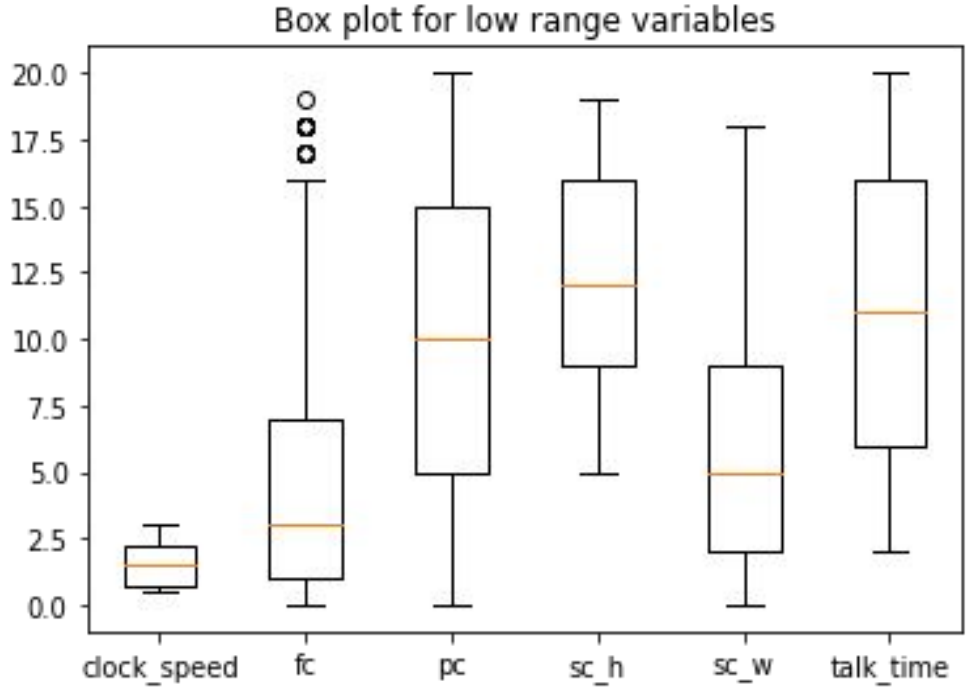
Equal no. of observations, highest being 4 cores

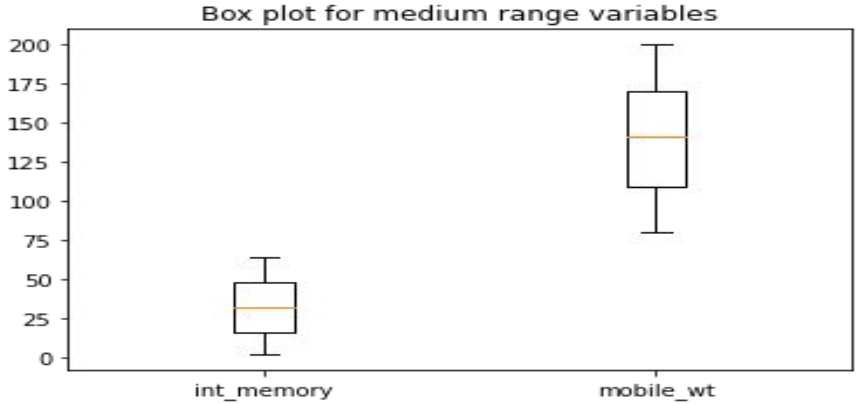


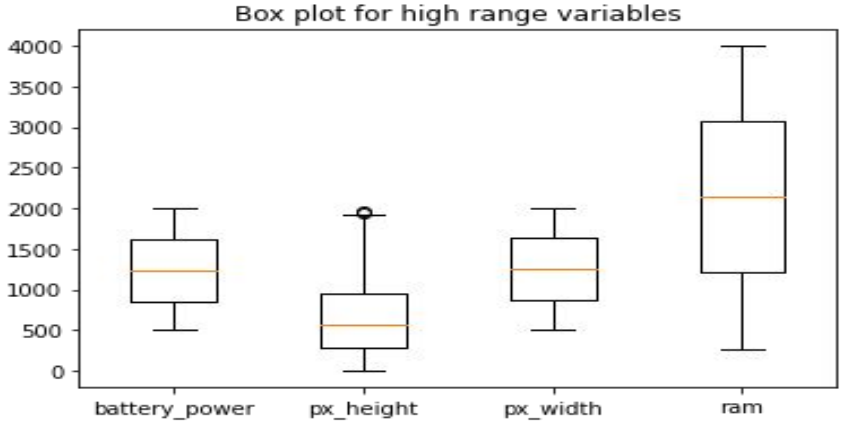
0.1cm has highest and 1cm has least no. of observations.

Univariate analysis - Numerical Variable

Descriptive stats using boxplots

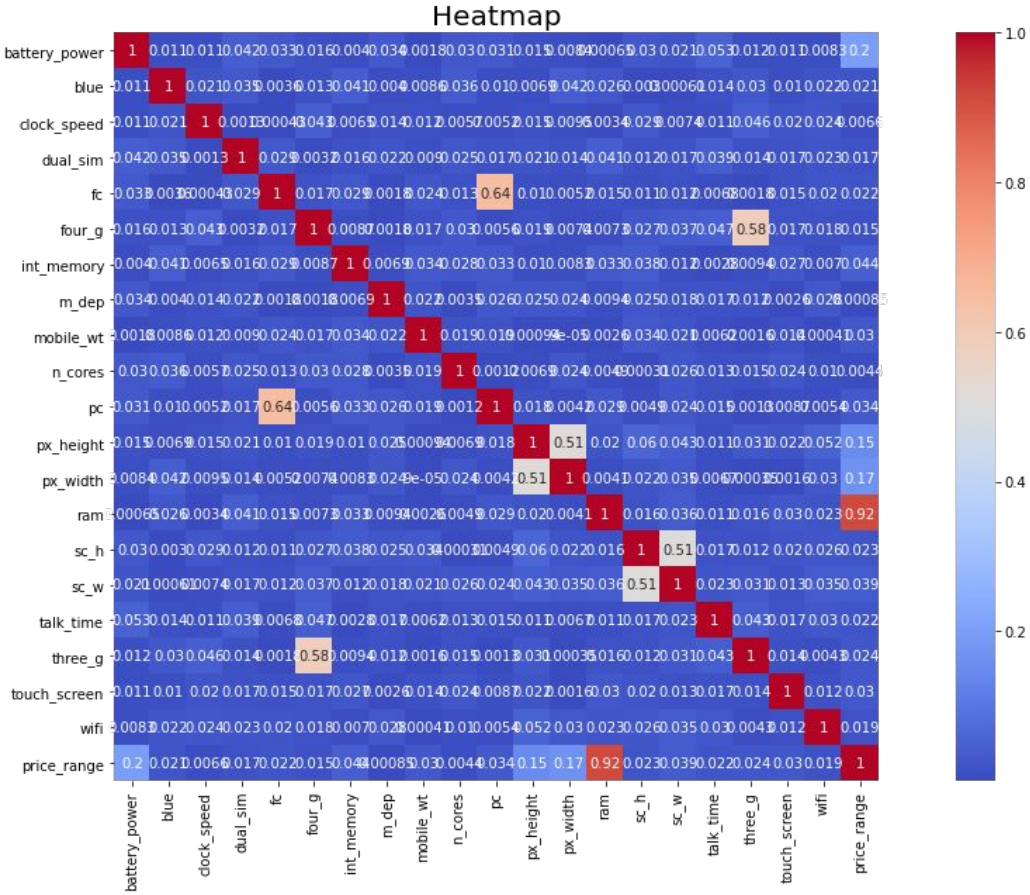




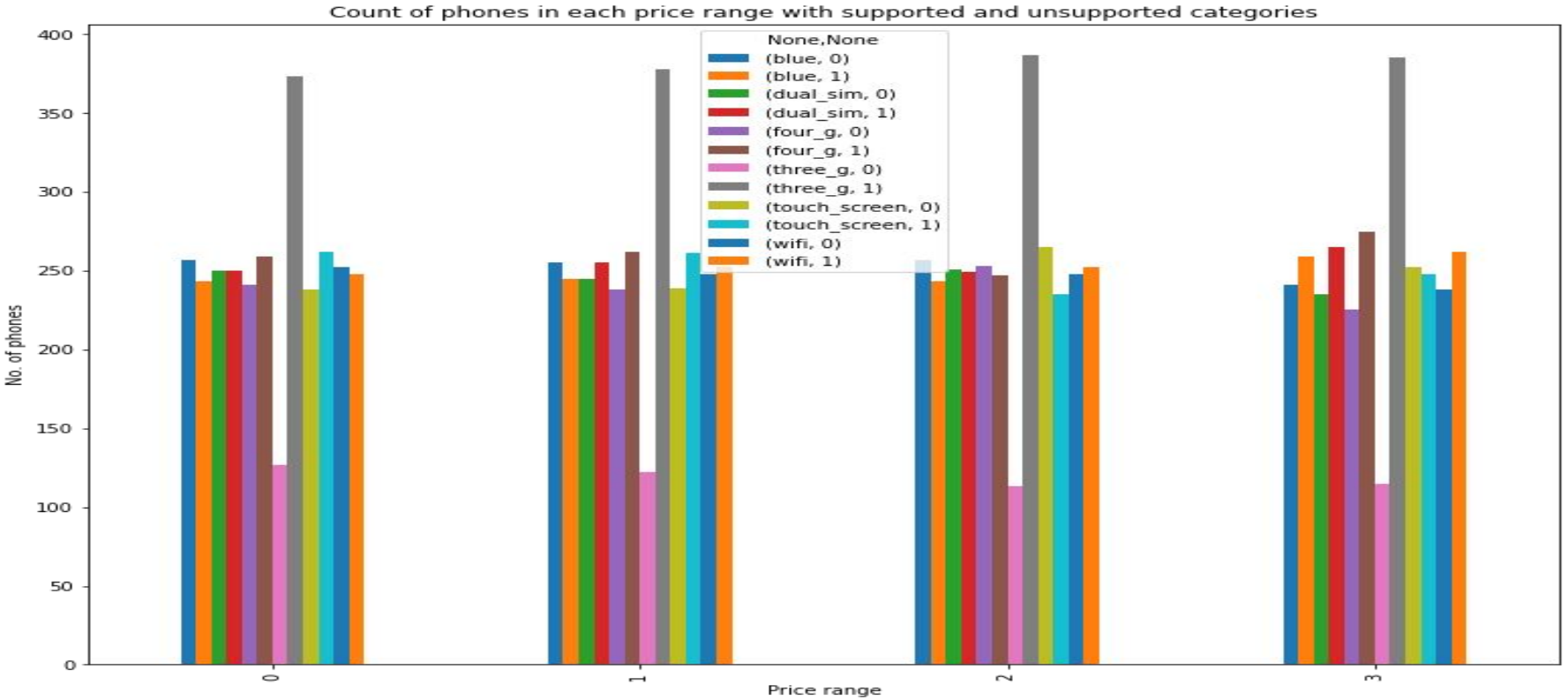


Multivariate analysis

- Pc is correlated with Fc.
- px_height and px_width are moderately correlated.
- Sc_h and sc_w are moderately correlated.
- Ram is highly correlated with price_range.

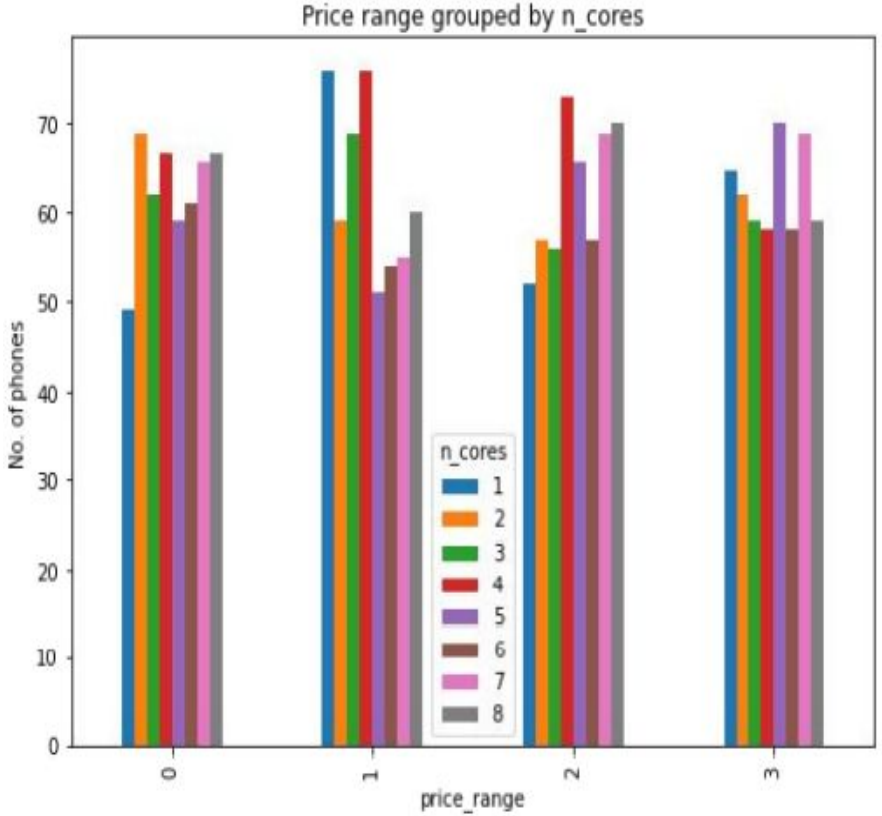


Multivariate analysis - Categorical variables

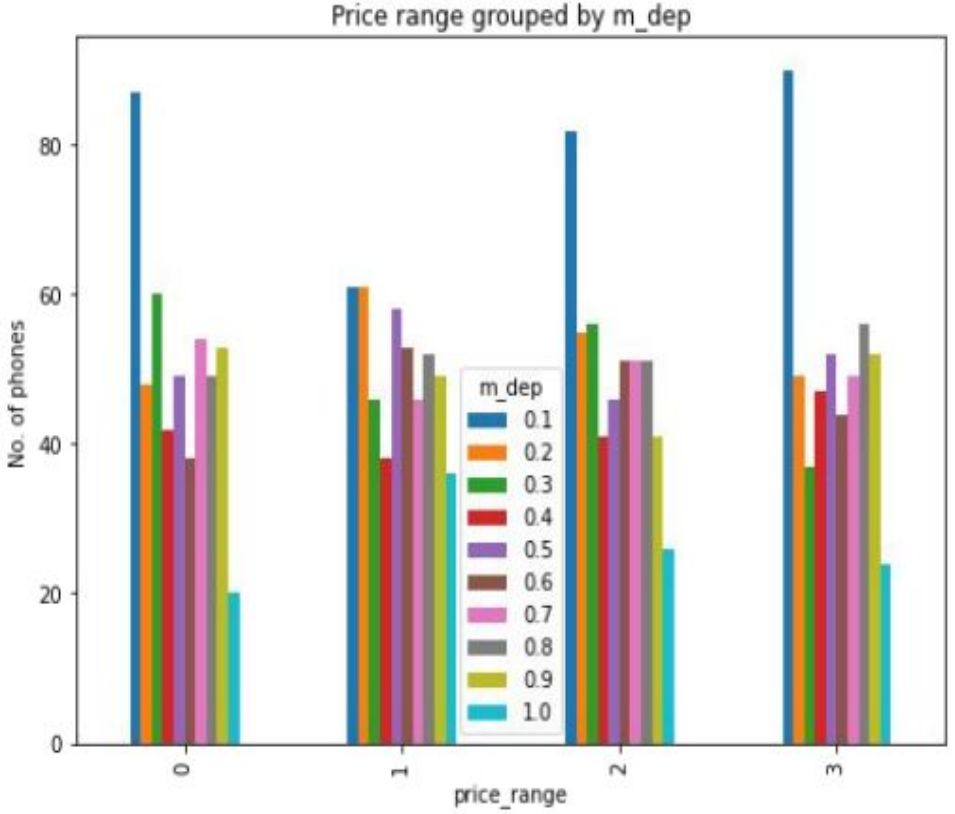


Almost equal no. of observations for each price range for each category.

Multivariate analysis - n_cores & m_dep

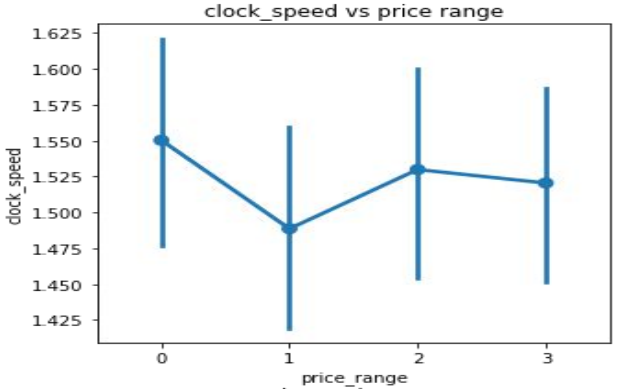


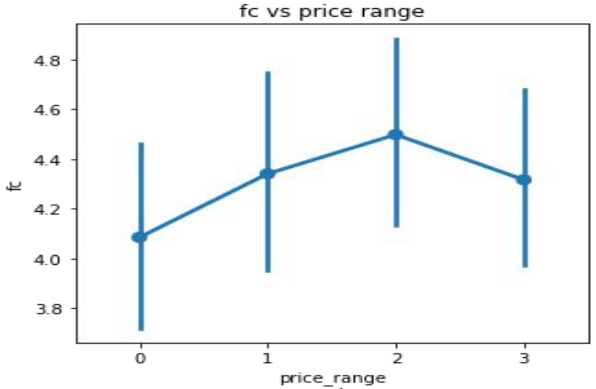
Count of phones with n_cores for each price range

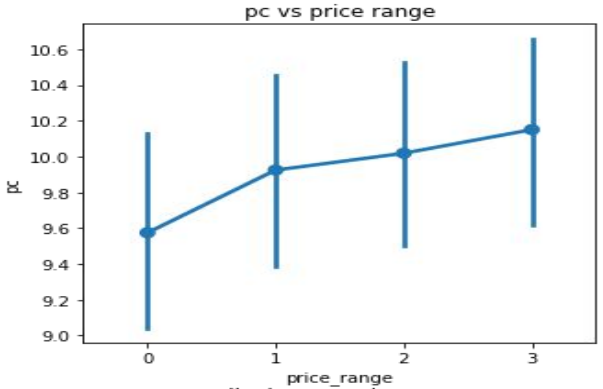


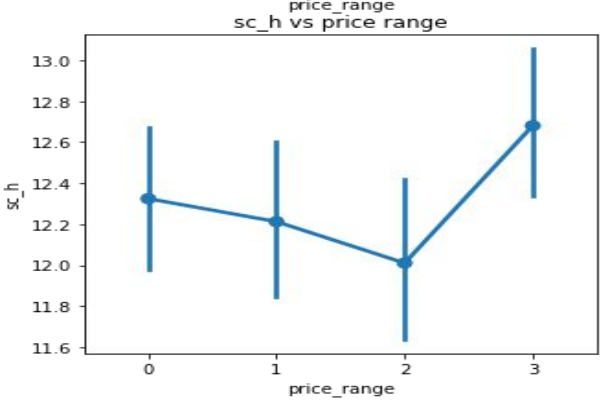
Count of phones with m_dep for each price range

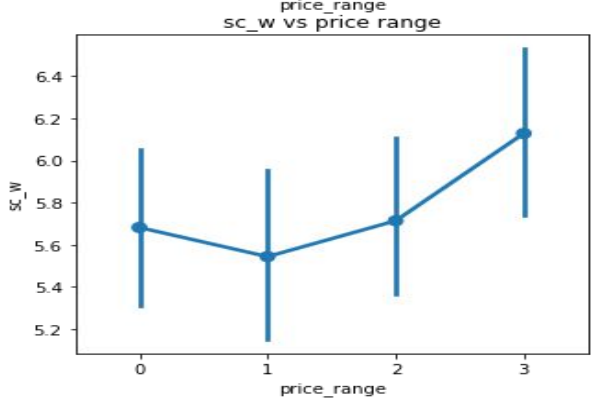
Multivariate analysis - Numerical variables

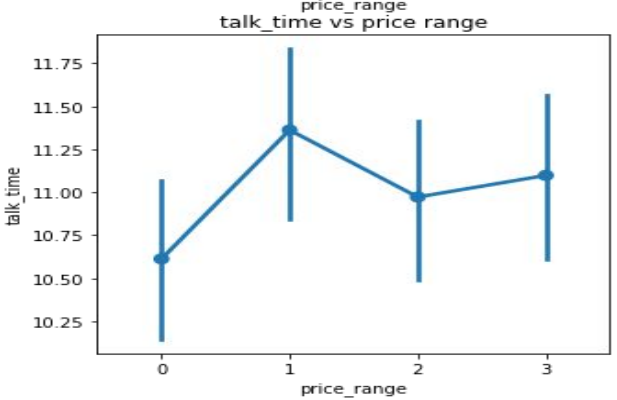






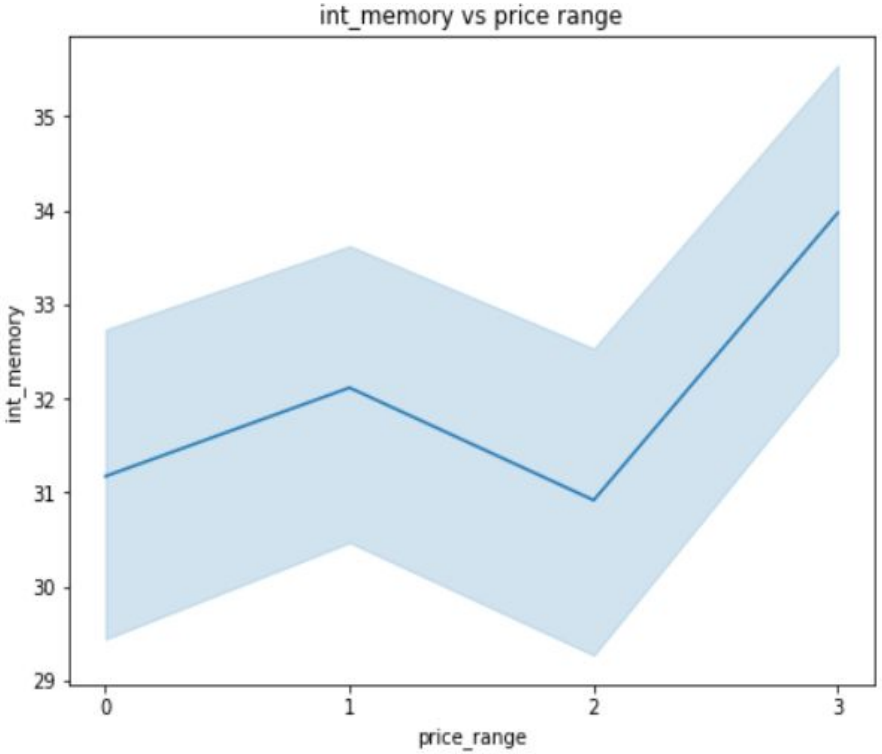


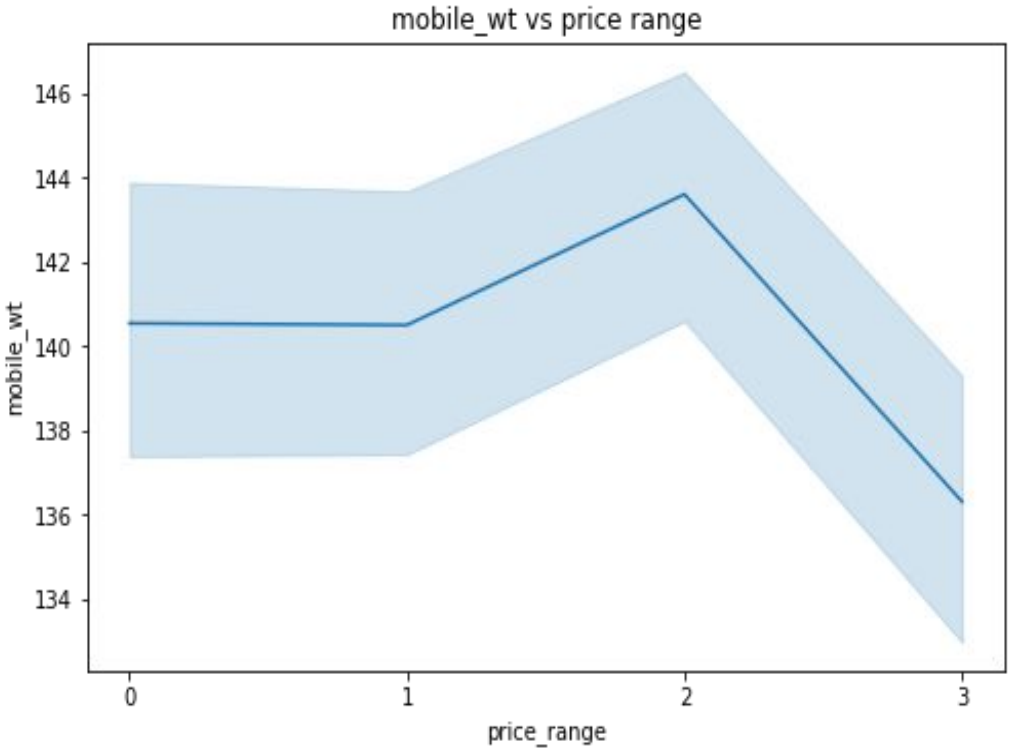




- **Clock_speed is high for low price range phones, talk_time is also less.**
- **Pc, fc, sc_w are in increasing trend.**

Multivariate analysis - int_memory, mobile_wt

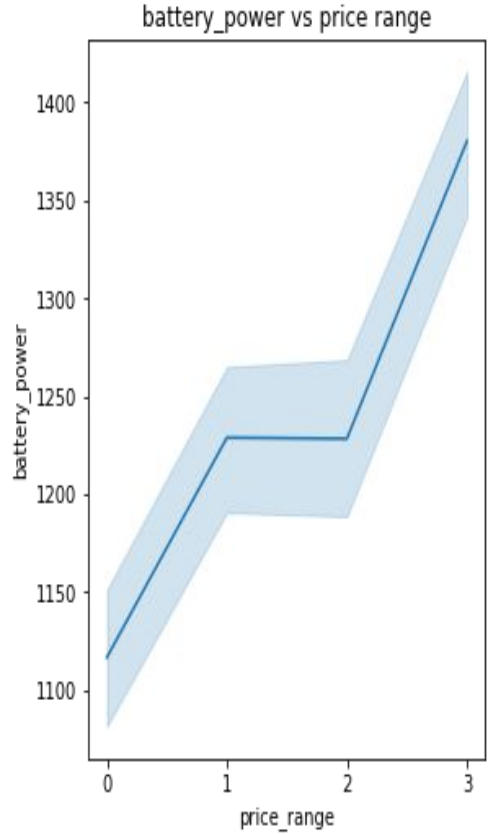


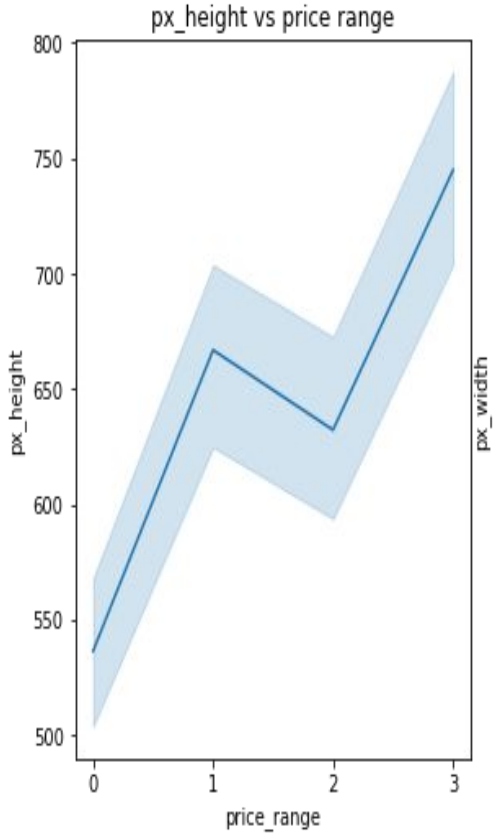


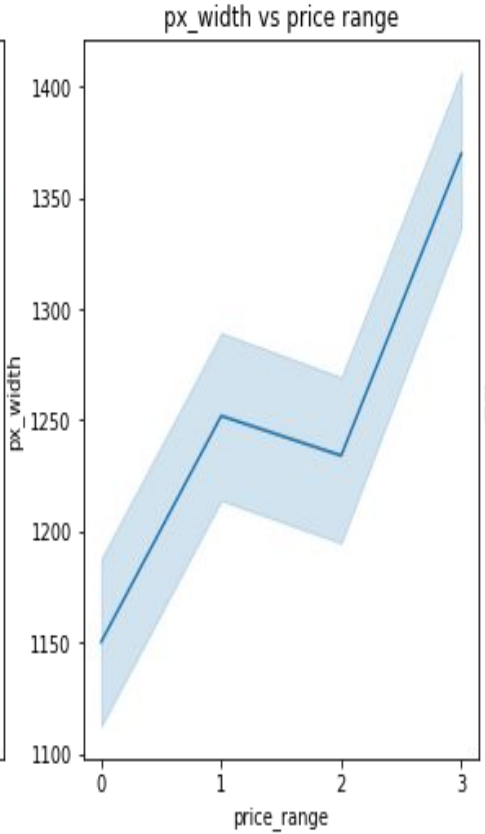
Increase in internal memory for very high prices

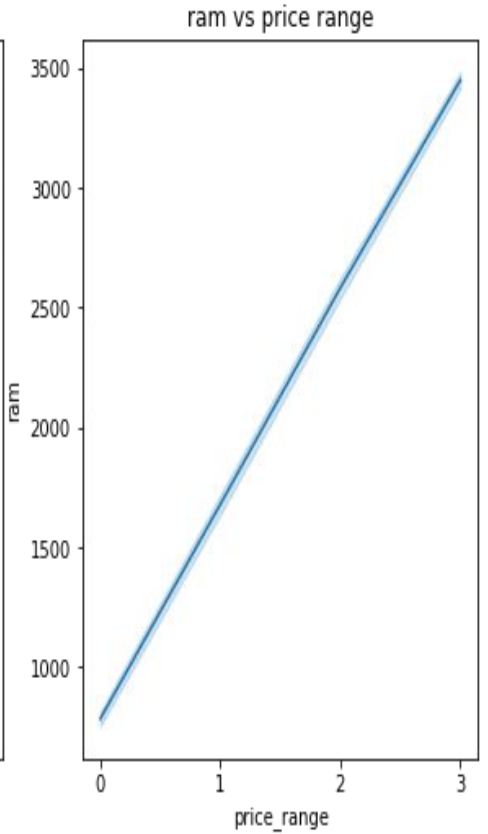
Decrease in mobile weight for very high prices

Multivariate analysis - Continued



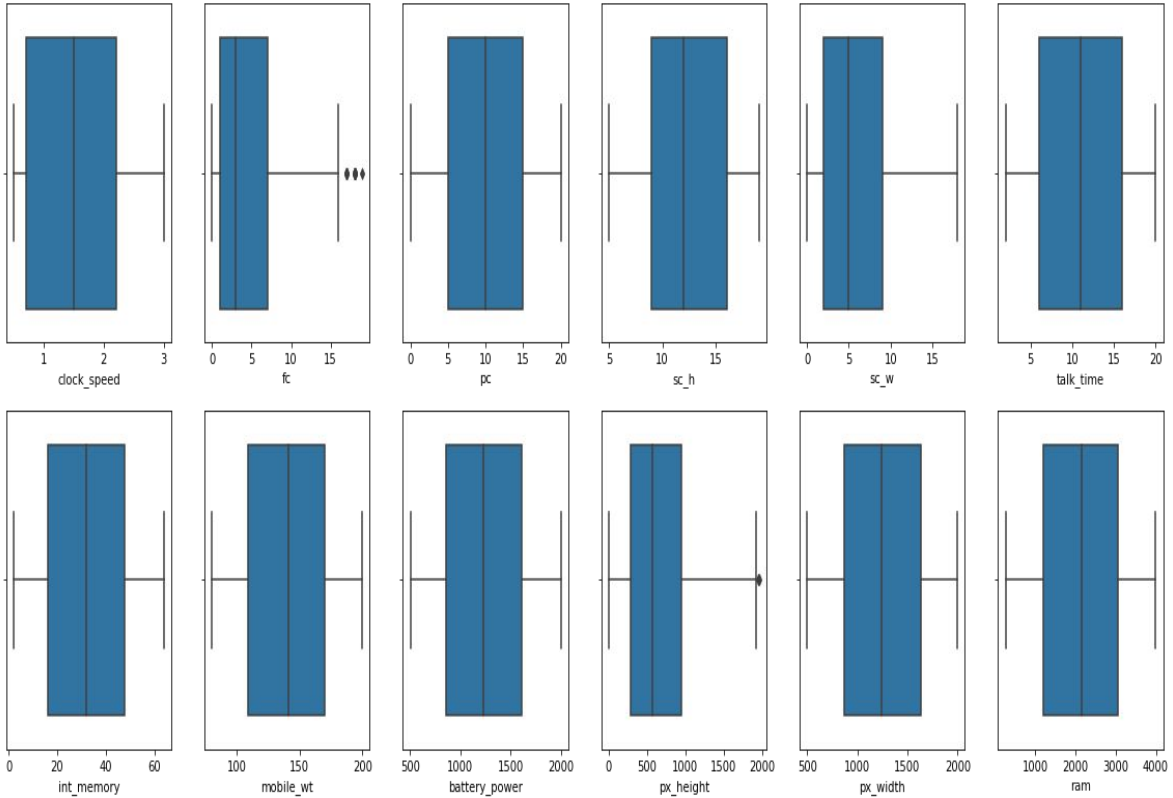






Data Wrangling

- No null values.
- Outliers were close to maximum value, so they can be ignored.
- px_height, px_width and sc_h, sc_w were modified into single columns.



Machine Learning Models

- **Two models experimented: Random Forest Classifier and XGBoost Classifier. The evaluation results are:**

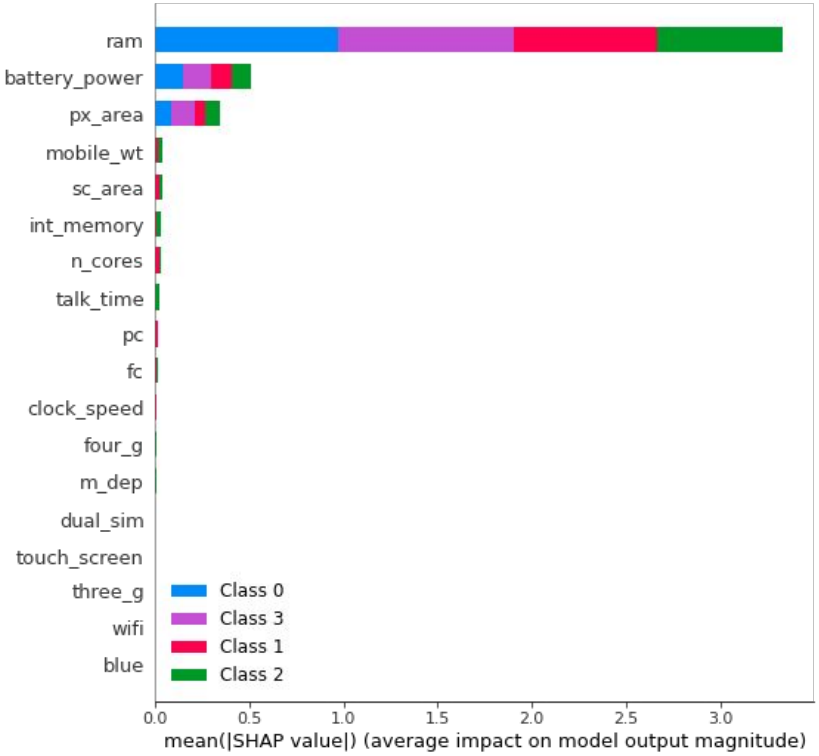
		Model	Accuracy	Precision	Recall	F1_score
Training set	0	Random Forest - Before hyperparameter tuning	1.00	[1.0, 1.0, 1.0, 1.0]	[1.0, 1.0, 1.0, 1.0]	[1.0, 1.0, 1.0, 1.0]
	1	Random Forest - After hyperparameter tuning	0.86	[[0.88, 0.78, 0.83, 0.94]]	[[0.97, 0.77, 0.74, 0.95]]	[[0.93, 0.78, 0.79, 0.95]]
	2	XGBoost - Before hyperparameter tuning	0.98	[[1.0, 0.96, 0.97, 0.99]]	[[0.98, 0.98, 0.97, 0.99]]	[[0.99, 0.97, 0.97, 0.99]]
	3	XGBoost - After hyperparameter tuning	0.89	[[0.92, 0.86, 0.83, 0.93]]	[[0.96, 0.83, 0.85, 0.92]]	[[0.94, 0.85, 0.84, 0.93]]
Test set	0	Random Forest - Before hyperparameter tuning	0.89	[0.91, 0.79, 0.86, 1.0]	[0.99, 0.88, 0.79, 0.9]	[0.95, 0.83, 0.82, 0.95]
	1	Random Forest - After hyperparameter tuning	0.82	[[0.86, 0.67, 0.77, 0.94]]	[[0.99, 0.77, 0.6, 0.89]]	[[0.92, 0.72, 0.67, 0.92]]
	2	XGBoost - Before hyperparameter tuning	0.87	[[0.95, 0.79, 0.78, 0.99]]	[[0.98, 0.92, 0.78, 0.82]]	[[0.96, 0.85, 0.78, 0.9]]
	3	XGBoost - After hyperparameter tuning	0.85	[[0.89, 0.78, 0.78, 0.93]]	[[0.97, 0.85, 0.74, 0.84]]	[[0.93, 0.81, 0.76, 0.88]]

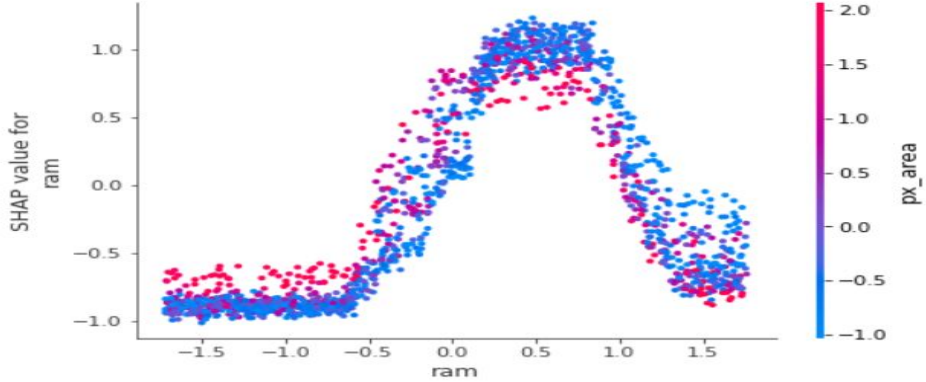
Model selection and validation

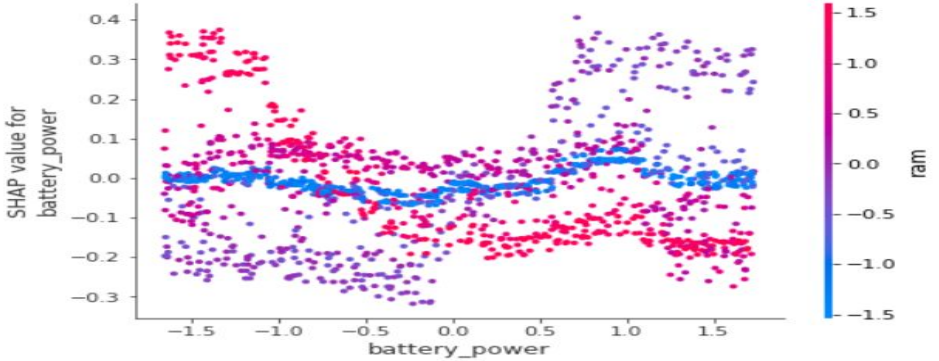
- **Random Forest and XGBoost initially overfitted with default hyperparameters.**
- **Overfitting was tackled with the help of hyperparameter tuning using Random Search.**
- **The best performance was given by XGBoost model, with accuracy of 0.89 and 0.85 for training and test set respectively.**
- **The best hyperparameter values were:**
  - learning rate = 0.13
  - n_estimators = 13
  - max_depth = 13
  - min_child_weight = 10
  - gamma = 1
  - subsample = 0.5

Model Explanation

- Shap technique were implemented to understand the working of the best model.
- The most important features were ram, battery_power, px_height and px_width.







Challenges

- The most challenging part in this exercise was to find an optimal set of hyperparameters that could give us the best performance.
- It took hours to try every combinations and finally selecting the best values.
- The model could even perform better with finer tunings and more number of instances.

Conclusions

- We build a predictive model, which could help companies to estimate price of mobiles in much effective way.
- To predict the cost of various different types of products, same procedure can be performed.
- By specifying economic range, a good product can be suggested to a customer.