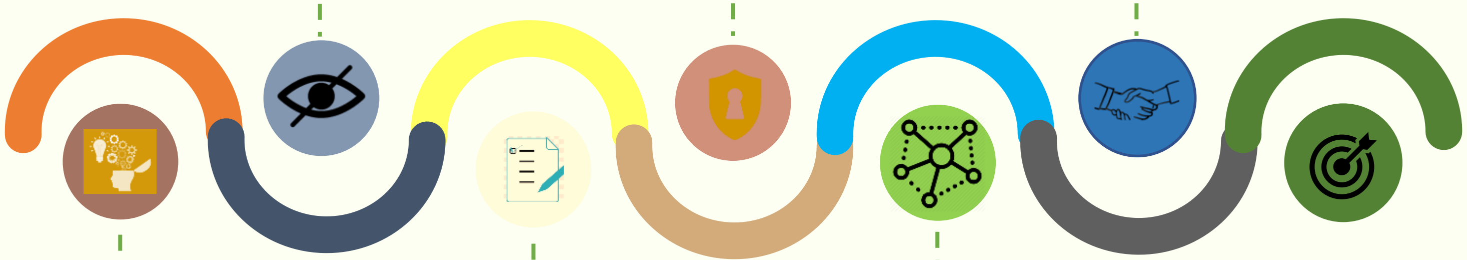
Capstone Project -3

Supervised Machine Learning – Classification

Jouher Lais Khan

Approach



Business Problem

Data Collection and Preprocessing

Importing important libraries and modules - Data Cleaning - Missing Data Handling -

Exploratory Data Analysis

Categorical Features - Continuous Features - EDA Conclusion

Feature Selection and Outlier Detection

Feature Engineering - Outlier Detection and Treatment

Modelling Train-Test Split

Model Performance and Evaluation

Visualizing Model Performances

Conclusion

Problem Description

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

Data Description

Battery_power - Total energy a battery can store in one time measured in mAh

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

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

•**Touch_screen -** Has touch screen or not

Wifi - Has wifi or not

•**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

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 Processing

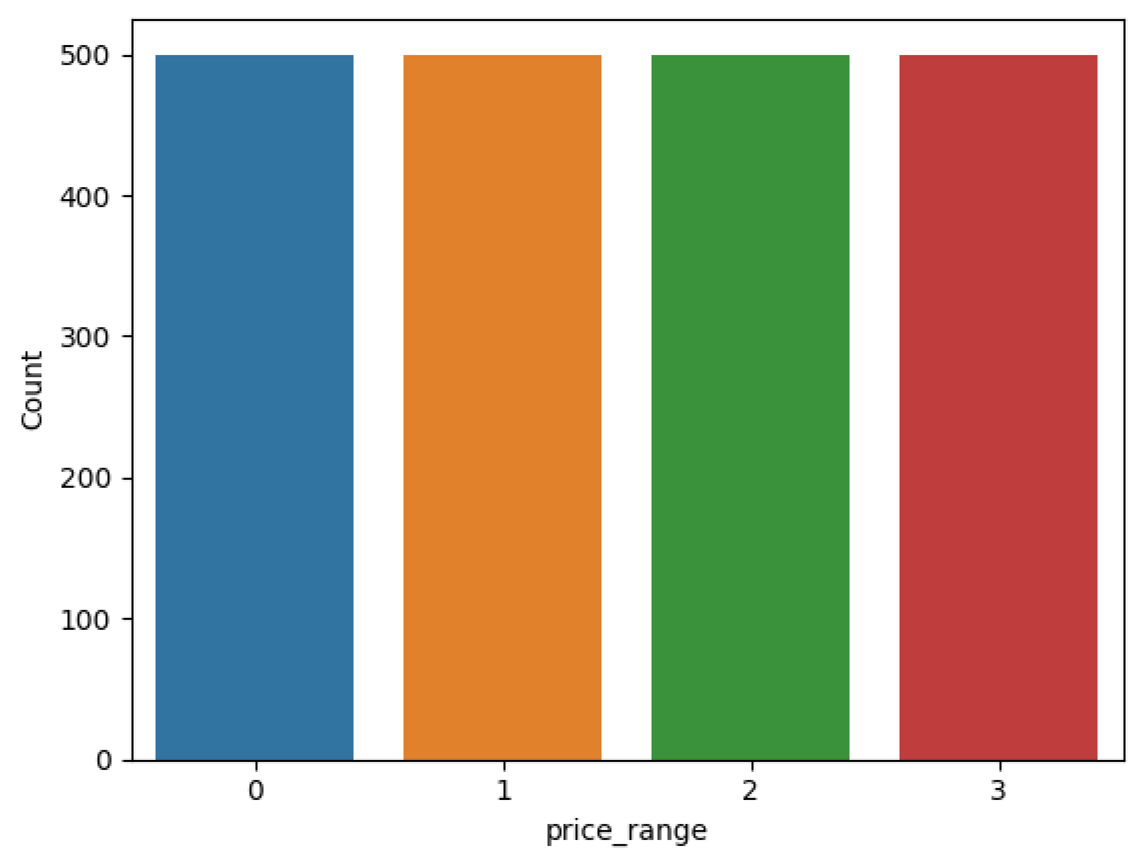
1. Minimum value of px_height and sc_w cannot be zero so we have convert the zero values to the mean value of column.

2. Data Contains no null value

3. Data Contains no duplicates value

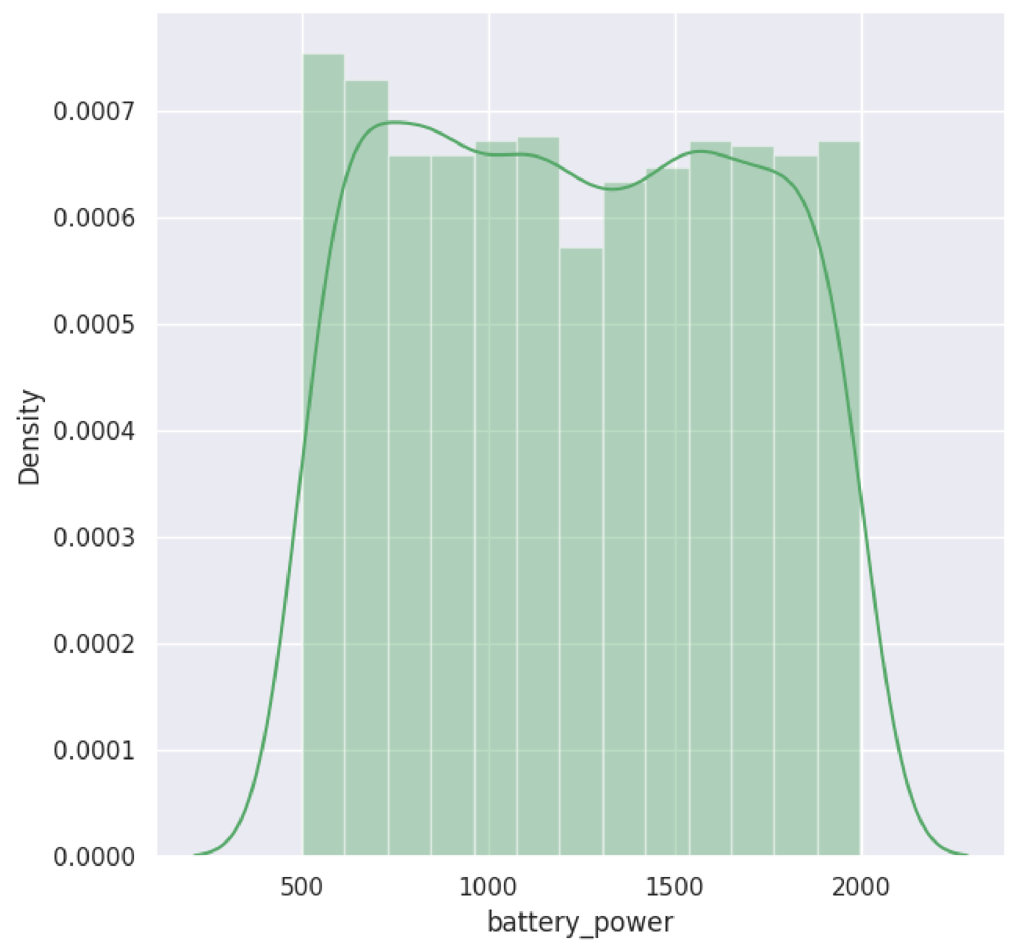
Exploratory Data Analysis

Price



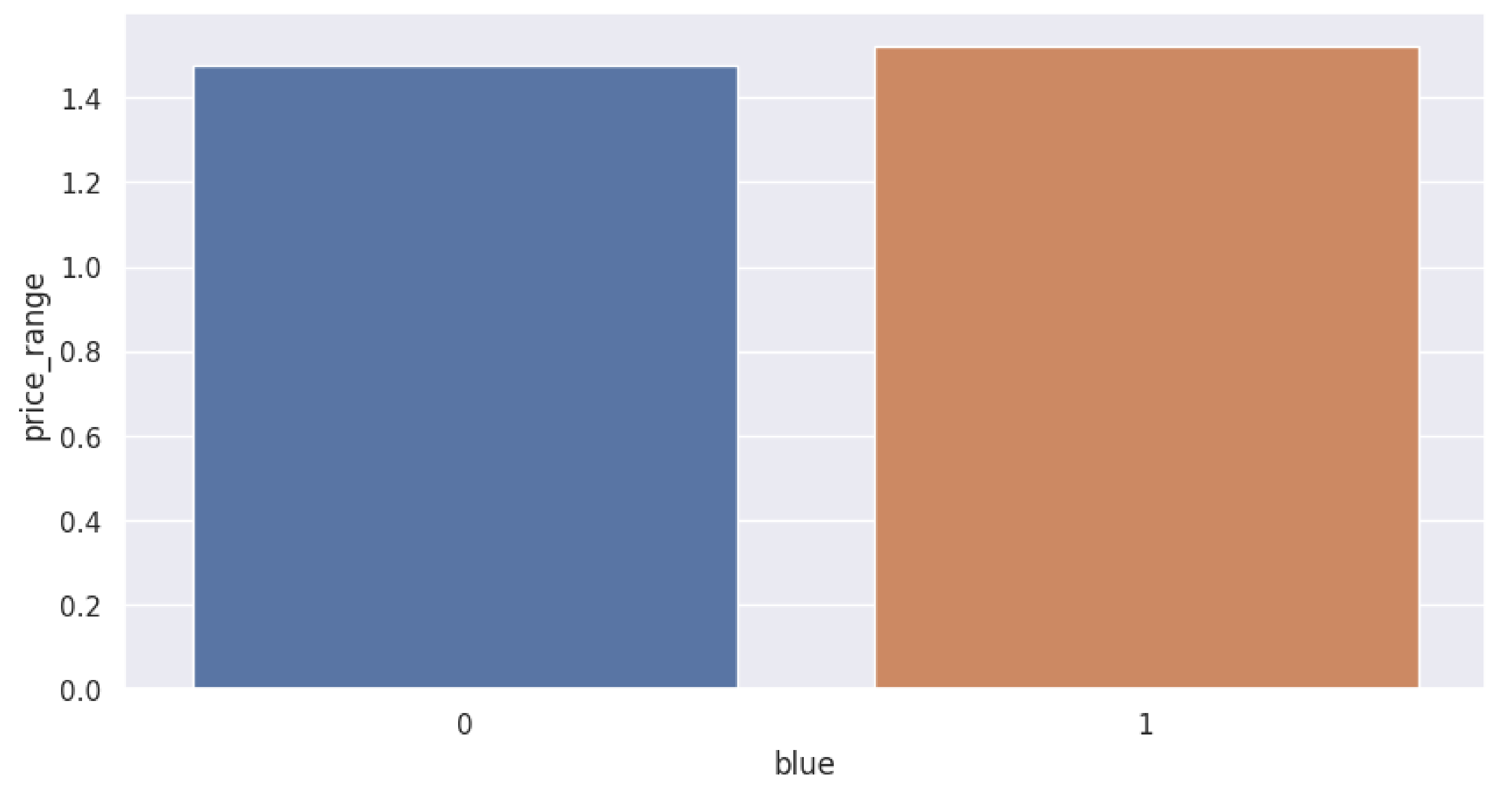
There are mobile phones in 4 price ranges. The number of elements is similar.

Battery Power



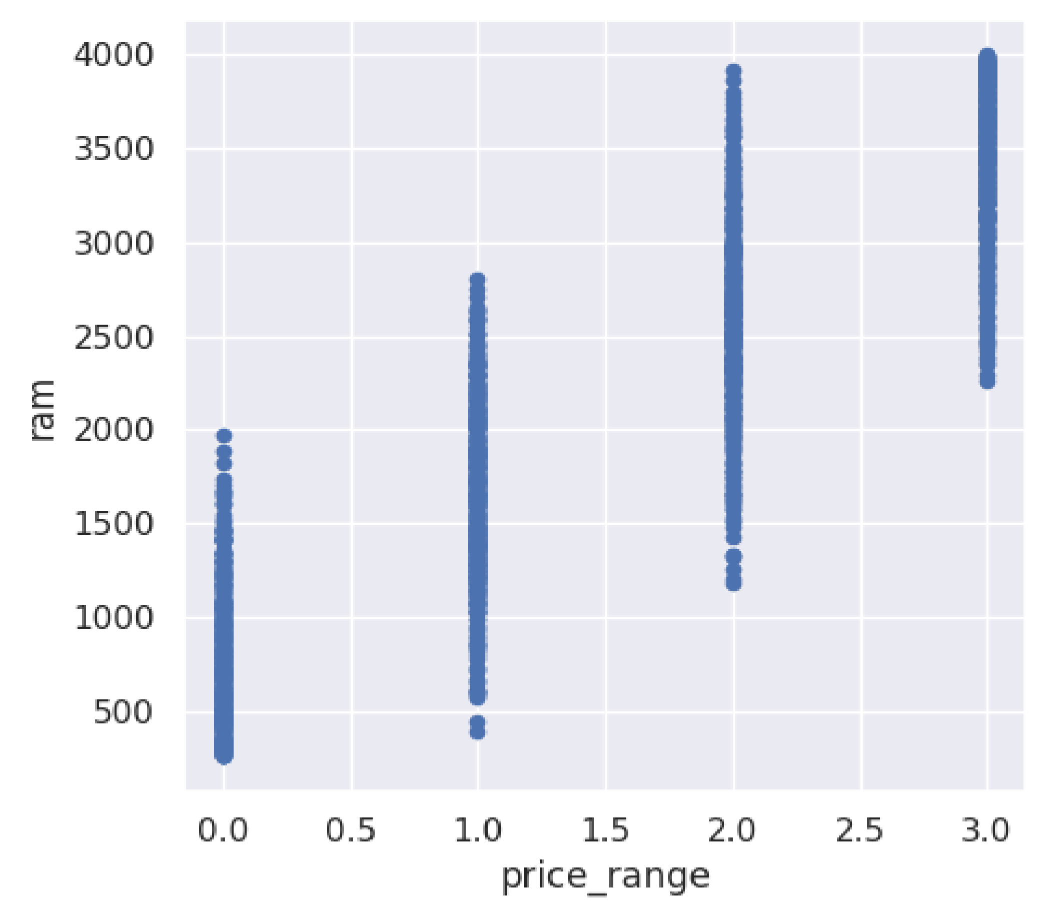
This plot shows how the battery mAh is spread. There is a gradual increase as the price range increases

Bluetooth



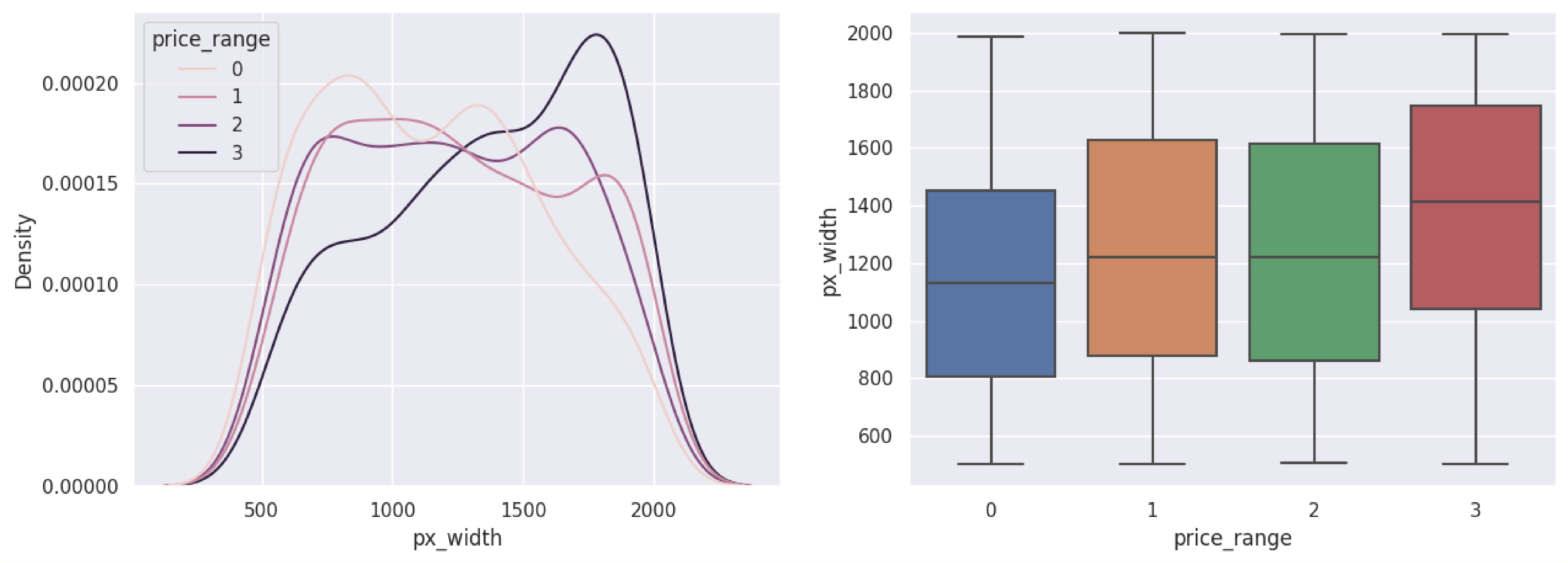
Half the devices have Bluetooth, and half don't.

Ram



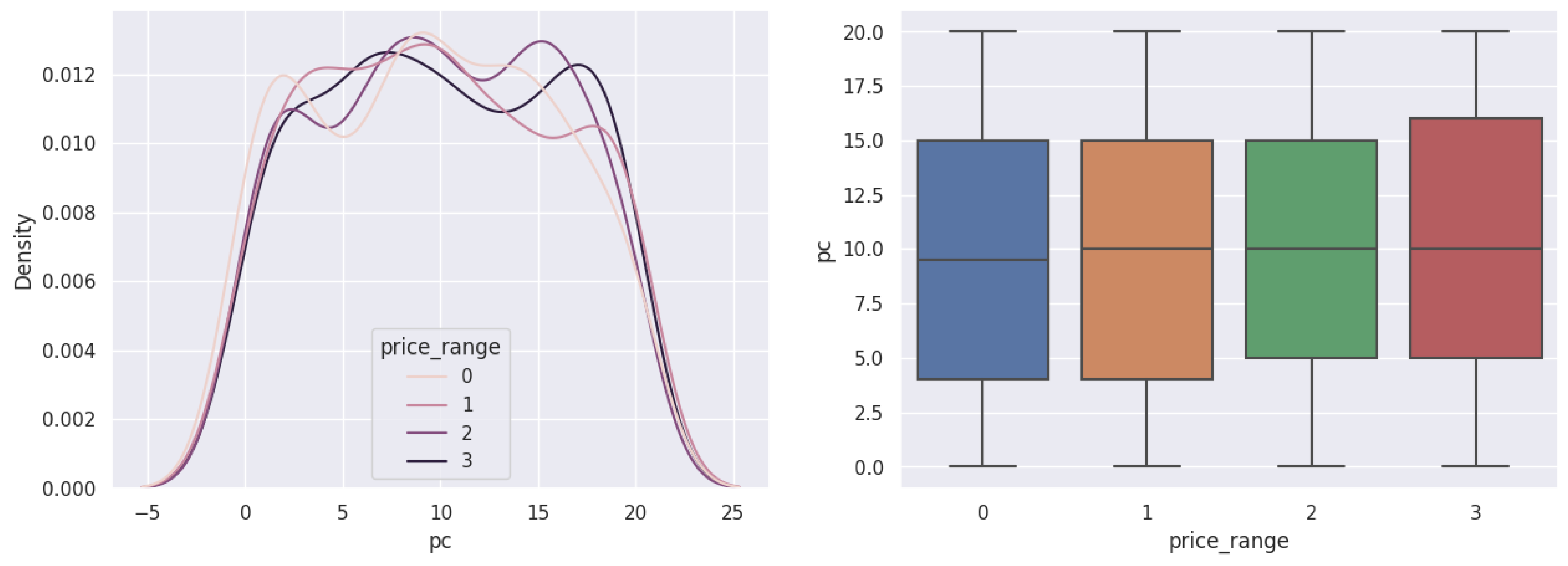
Ram has continuous increase with price range while moving from Low cost to Very high cost

Pixel Width



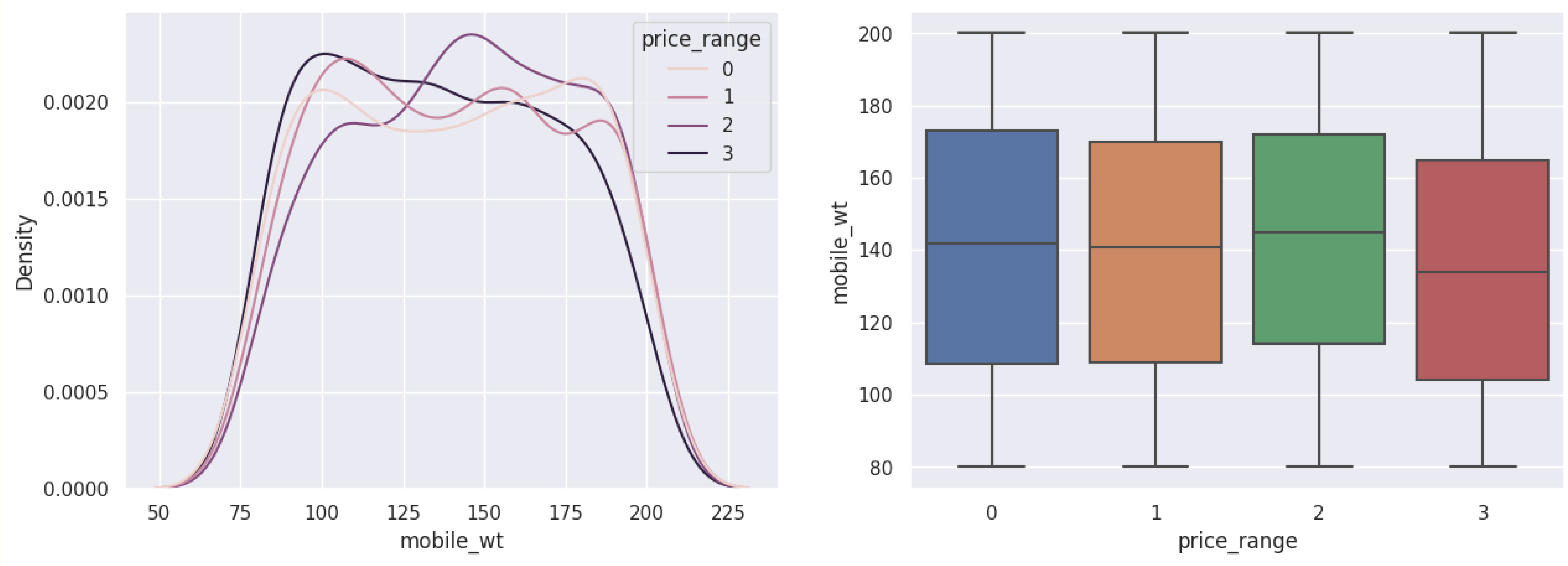
There is not a continuous increase in pixel width as we move from Low cost to Very high cost. Mobiles with 'Medium cost' and 'High cost' has almost equal pixel width. so we can say that it would be a driving factor in deciding price_range.

Primary Camera(Mega Pixels)



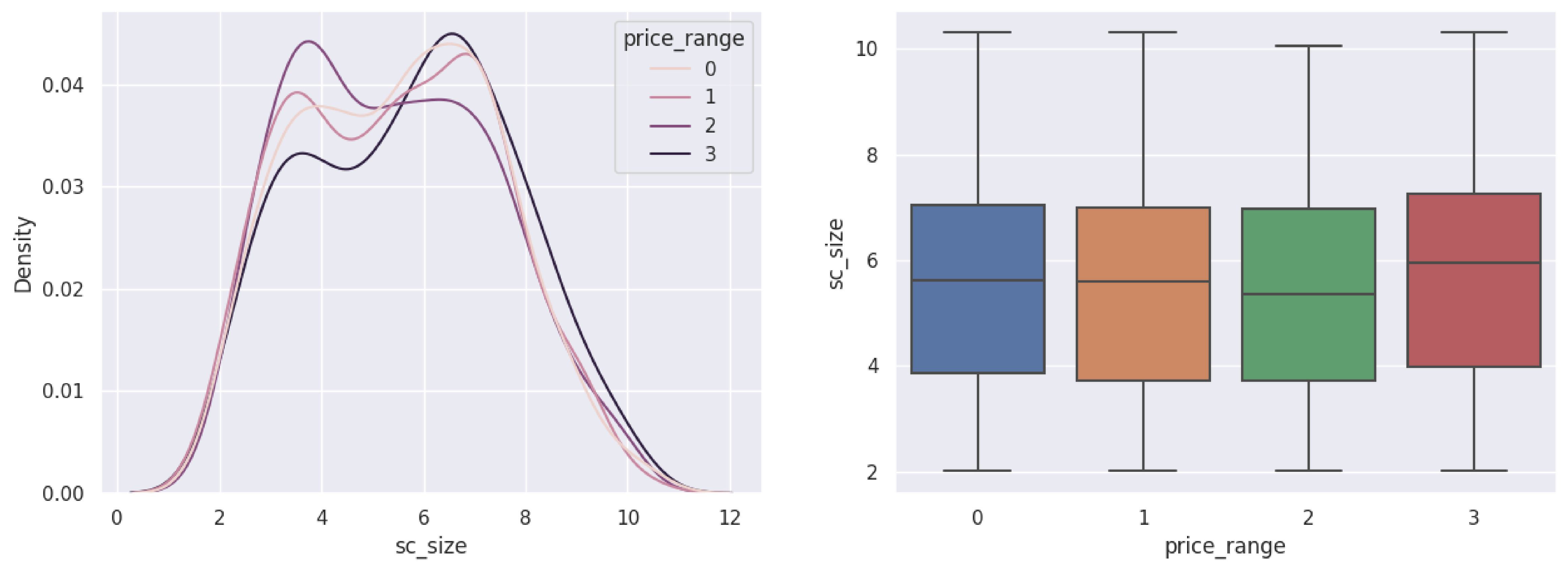
Primary camera megapixels are showing a little variation along the target categories, which is a good sign for prediction.

Mobile weight



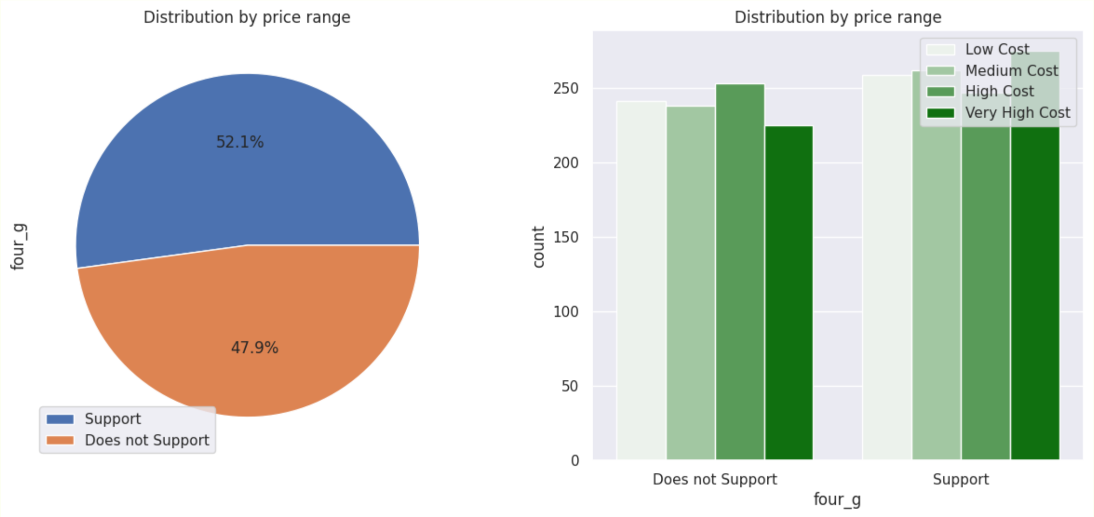
Costly phones are lighter

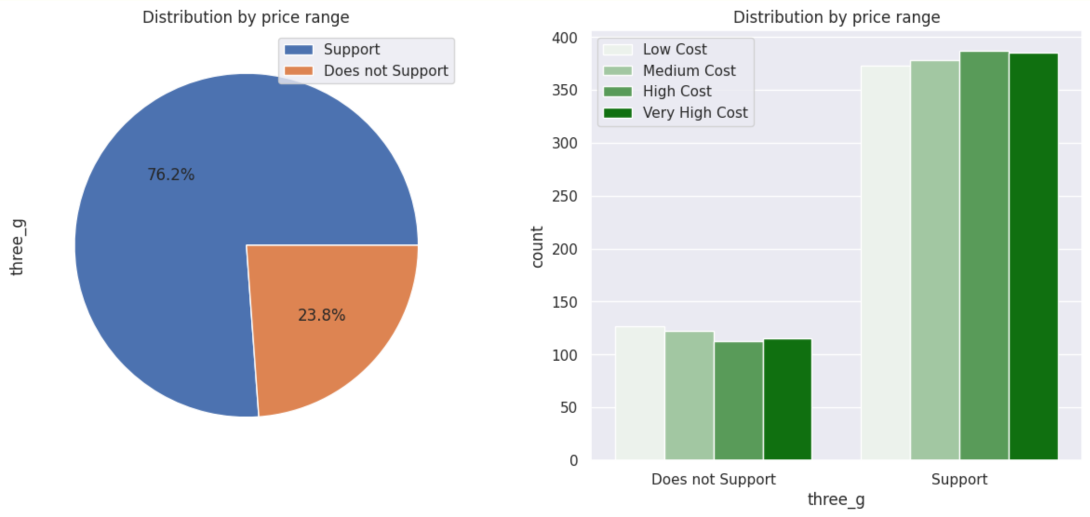
Screen Size



Screen Size shows little variation along the target variables. This can be helpful in predicting the target categories.

3G & 4G





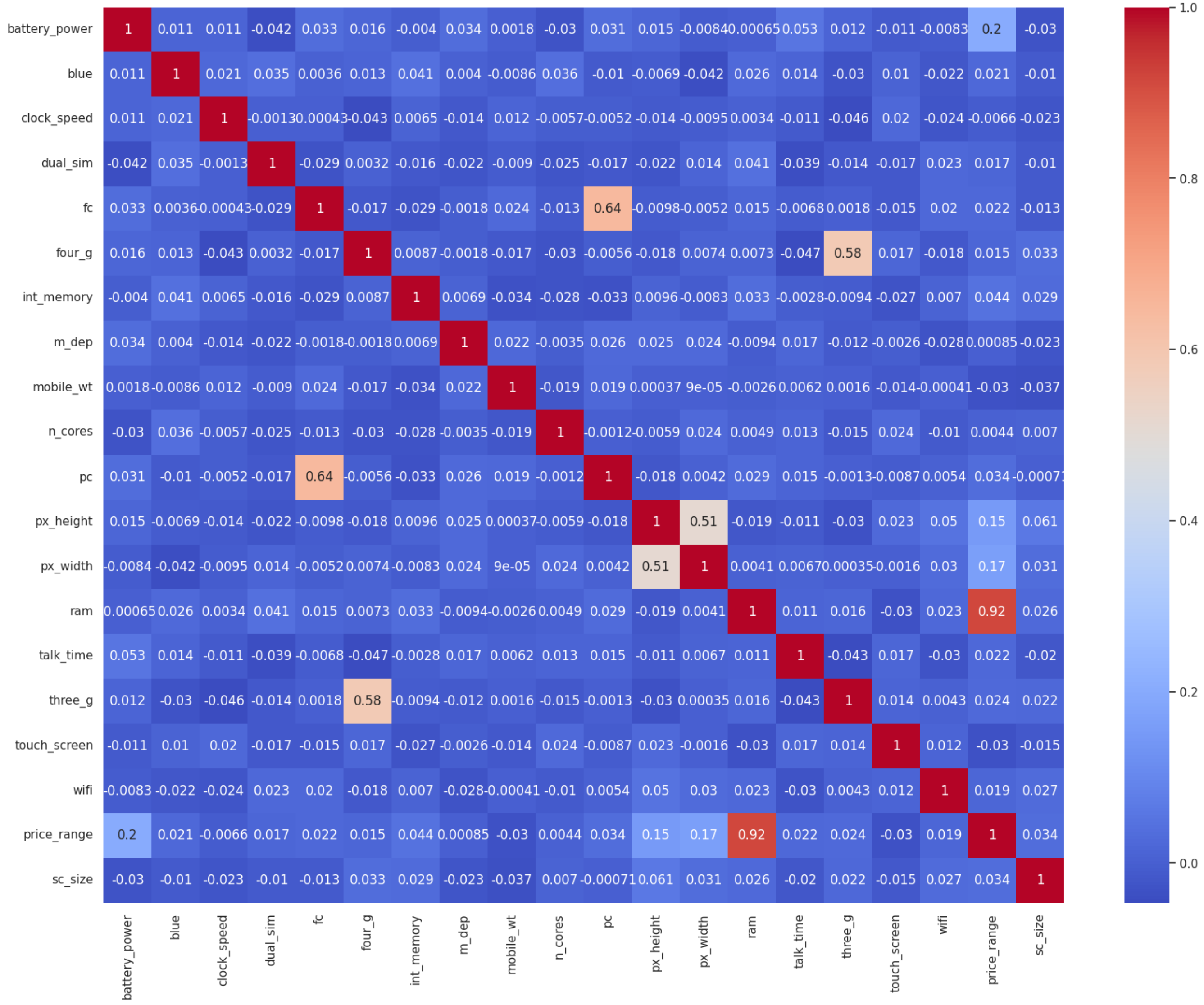
Feature 'three_g' play an important feature in prediction

Correlation 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.

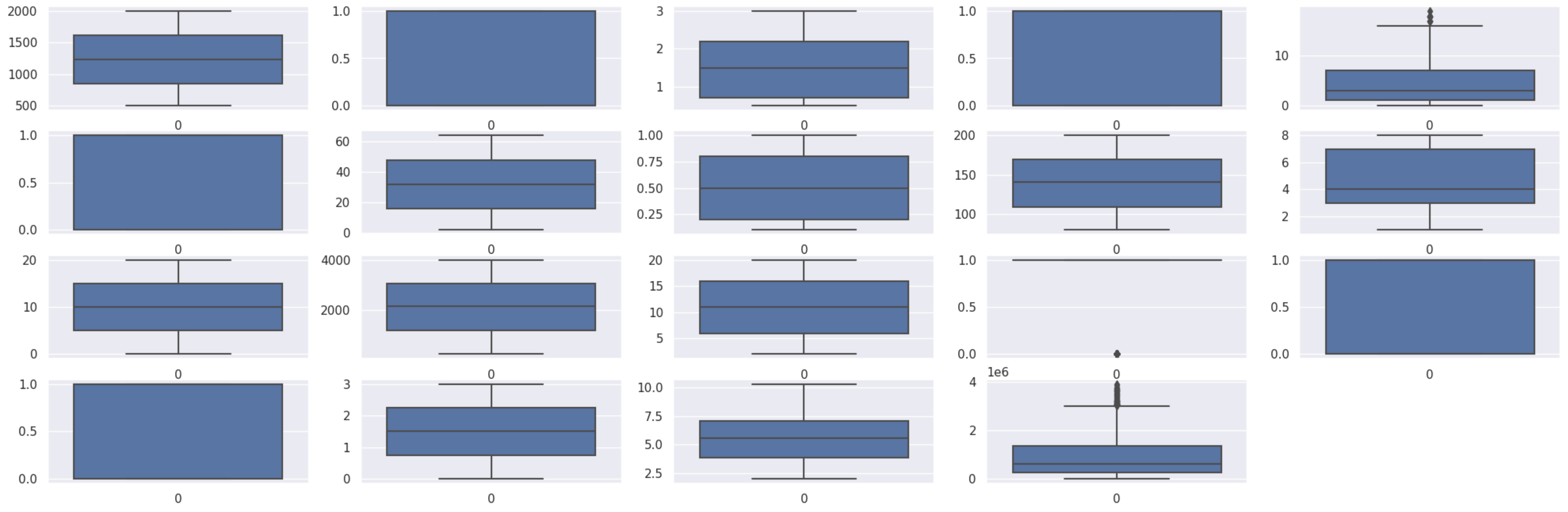
Also, if px_height increases, pixel width also increases, that means the overall pixels in the screen. We can replace these two features with one feature. Front Camera megapixels and Primary camera megapixels are different entities despite of showing colinearity. So we'll be keeping them as they are.



	battery_power	blue	clock_speed	dual_sim	fc	four_g	int_memory	m_dep	mobile_wt	n_cores	pc	px_height	px_width	ram	talk_time	three_g	touch_screen	wifi	price_range	sc_size
battery_power	1	0.011	0.011	-0.042	0.033	0.016	-0.004	0.034	0.0018	-0.03	0.031	0.015	-0.0084	0.00065	0.053	0.012	-0.011	-0.0083	0.2	-0.03
blue	0.011	1	0.021	0.035	0.0036	0.013	0.041	0.004	-0.0086	0.036	-0.01	-0.0069	-0.042	0.026	0.014	-0.03	0.01	-0.022	0.021	-0.01
clock_speed	0.011	0.021	1	-0.0013	0.00043	-0.043	0.0065	-0.014	0.012	-0.0057	-0.0052	-0.014	-0.0095	0.0034	-0.011	-0.046	0.02	-0.024	-0.0066	-0.023
dual_sim	-0.042	0.035	-0.0013	1	-0.029	0.0032	-0.016	-0.022	-0.009	-0.025	-0.017	-0.022	0.014	0.041	-0.039	-0.014	-0.017	0.023	0.017	-0.01
fc	0.033	0.0036	-0.00043	-0.029	1	-0.017	-0.029	-0.0018	0.024	-0.013	0.64	-0.0098	-0.0052	0.015	-0.0068	0.0018	-0.015	0.02	0.022	-0.013
four_g	0.016	0.013	-0.043	0.0032	-0.017	1	0.0087	-0.0018	-0.017	-0.03	-0.0056	-0.018	0.0074	0.0073	-0.047	0.58	0.017	-0.018	0.015	0.033
int_memory	-0.004	0.041	0.0065	-0.016	-0.029	0.0087	1	0.0069	-0.034	-0.028	-0.033	0.0096	-0.0083	0.033	-0.0028	-0.0094	-0.027	0.007	0.044	0.029
m_dep	0.034	0.004	-0.014	-0.022	-0.0018	-0.0018	0.0069	1	0.022	-0.0035	0.026	0.025	0.024	-0.0094	0.017	-0.012	-0.0026	-0.028	0.00085	-0.023
mobile_wt	0.0018	-0.0086	0.012	-0.009	0.024	-0.017	-0.034	0.022	1	-0.019	0.019	0.00037	9e-05	-0.0026	0.0062	0.0016	-0.014	-0.00041	-0.03	-0.037
n_cores	-0.03	0.036	-0.0057	-0.025	-0.013	-0.03	-0.028	-0.0035	-0.019	1	-0.0012	-0.0059	0.024	0.0049	0.013	-0.015	0.024	-0.01	0.0044	0.007
pc	0.031	-0.01	-0.0052	-0.017	0.64	-0.0056	-0.033	0.026	0.019	-0.0012	1	-0.018	0.0042	0.029	0.015	-0.0013	-0.0087	0.0054	0.034	-0.00071
px_height	0.015	-0.0069	-0.014	-0.022	-0.0098	-0.018	0.0096	0.025	0.00037	-0.0059	-0.018	1	0.51	-0.019	-0.011	-0.03	0.023	0.05	0.15	0.061
px_width	-0.0084	-0.042	-0.0095	0.014	-0.0052	0.0074	-0.0083	0.024	9e-05	0.024	0.0042	0.51	1	0.0041	0.0067	0.00035	-0.0016	0.03	0.17	0.031
ram	0.00065	0.026	0.0034	0.041	0.015	0.0073	0.033	-0.0094	-0.0026	0.0049	0.029	-0.019	0.0041	1	0.011	0.016	-0.03	0.023	0.92	0.026
talk_time	0.053	0.014	-0.011	-0.039	-0.0068	-0.047	-0.0028	0.017	0.0062	0.013	0.015	-0.011	0.0067	0.011	1	-0.043	0.017	-0.03	0.022	-0.02
three_g	0.012	-0.03	-0.046	-0.014	0.0018	0.58	-0.0094	-0.012	0.0016	-0.015	-0.0013	-0.03	0.00035	0.016	-0.043	1	0.014	0.0043	0.024	0.022
touch_screen	-0.011	0.01	0.02	-0.017	-0.015	0.017	-0.027	-0.0026	-0.014	0.024	-0.0087	0.023	-0.0016	-0.03	0.017	0.014	1	0.012	-0.03	-0.015
wifi	-0.0083	-0.022	-0.024	0.023	0.02	-0.018	0.007	-0.028	-0.00041	-0.01	0.0054	0.05	0.03	0.023	-0.03	0.0043	0.012	1	0.019	0.027
price_range	0.2	0.021	-0.0066	0.017	0.022	0.015	0.044	0.00085	-0.03	0.0044	0.034	0.15	0.17	0.92	0.022	0.024	-0.03	0.019	1	0.034
sc_size	-0.03	-0.01	-0.023	-0.01	-0.013	0.033	0.029	-0.023	-0.037	0.007	-0.00071	0.061	0.031	0.026	-0.02	0.022	-0.015	0.027	0.034	1

Outliers Removal

There are almost no outliers in the data



Feature Encoding

1. Creating copy of Data Frame for Modelling.
2. Creating list of final features which will be used in modelling.
3. Creating Sales as dependent variables and features as independent variable.
4. Train-Test Split

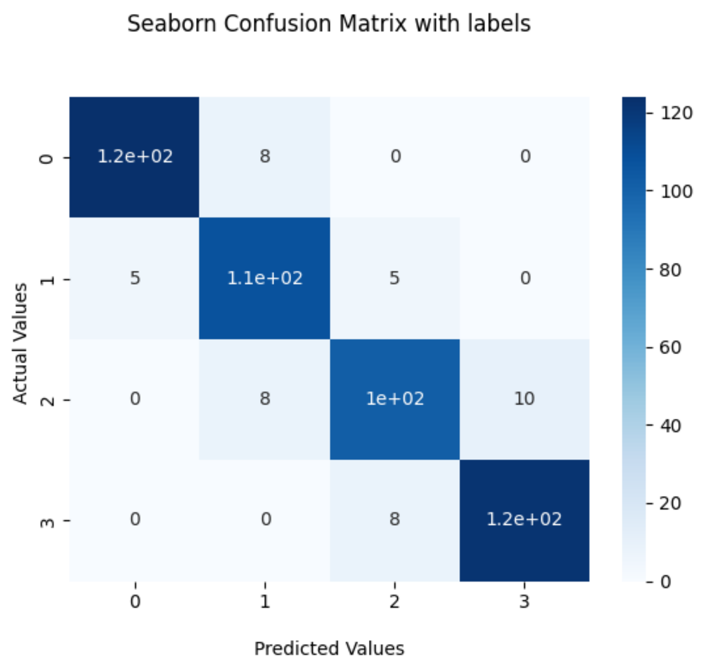
Models Implemented

1. Logistic Regression
2. Decision Tree
3. Random Forest Regression with Hyperparameter tuning
4. xgBoost with Hyperparameter Tuning
5. KNN classifier
6. Naïve Bayes
7. Support Vector Machine

Logistic Regression

Classification report for Logistic Regression (Test set)

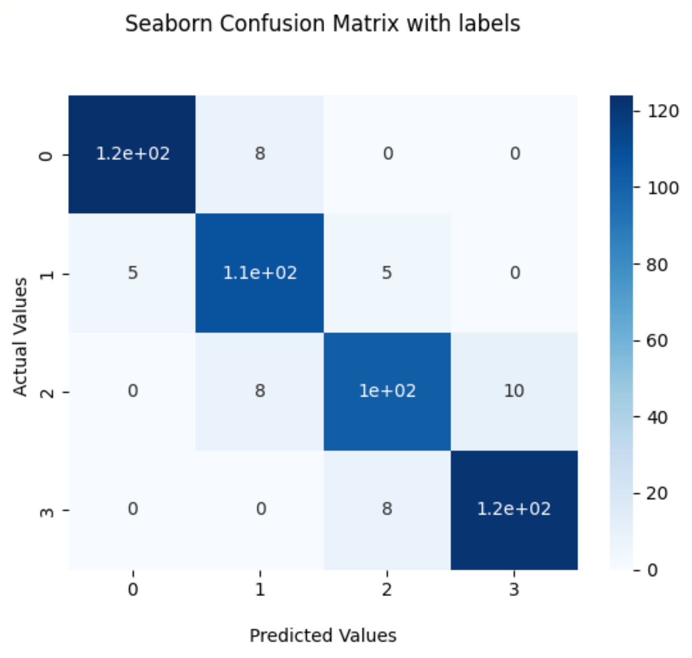
	Precision	Recall	F1-Score	Support
0	0.94	0.96	0.95	129
1	0.92	0.87	0.89	124
2	0.85	0.89	0.87	115
3	0.94	0.92	0.93	132
Accuracy			0.91	500
Macro Avg	0.91	0.91	0.91	500
Weighted Avg	0.91	0.91	0.91	500



Decision Tree(Hyperparameter Tuning)

Classification report for Decision Tree (Test set)

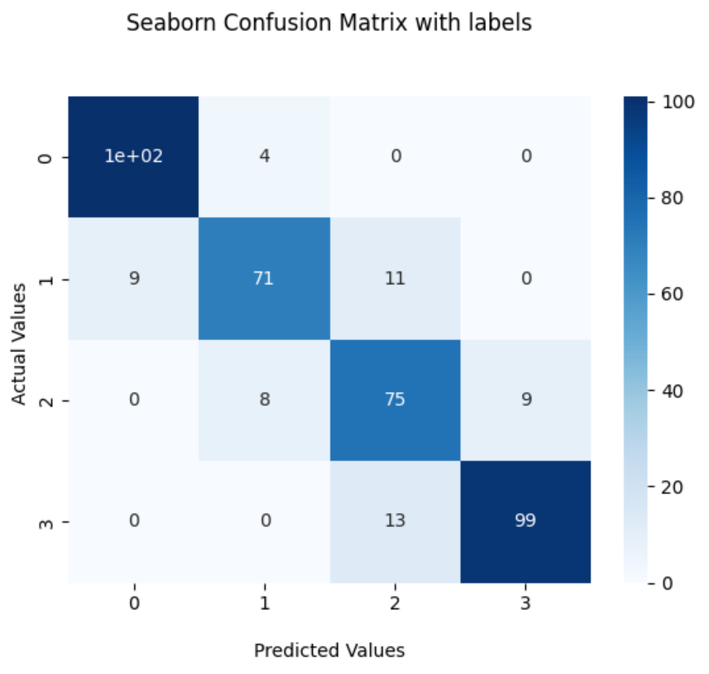
	Precision	Recall	F1-Score	Support
0	0.95	0.87	0.91	132
1	0.75	0.86	0.80	118
2	0.78	0.72	0.75	120
3	0.88	0.89	0.89	130
Accuracy			0.84	500
Macro Avg	0.84	0.84	0.84	500
Weighted Avg	0.84	0.84	0.84	500



Random Forest(Hyperparameter Tuning)

Classification report for Random Forest (Test set)

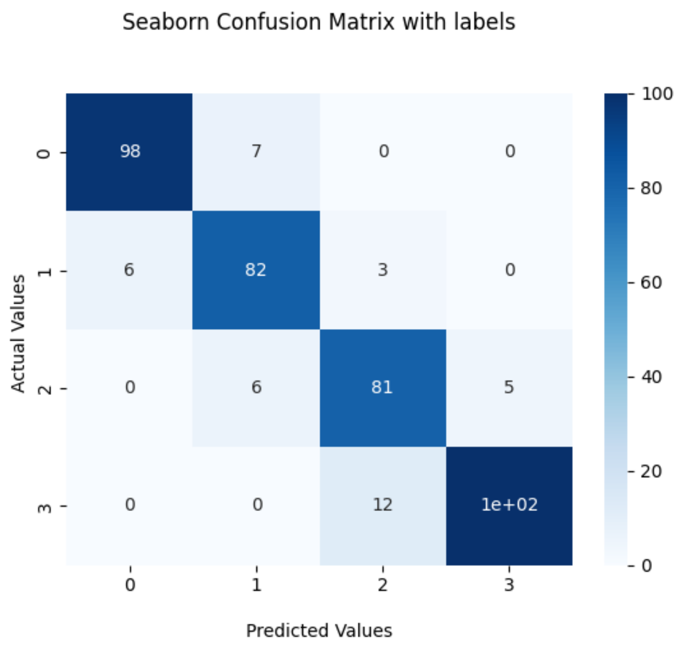
	Precision	Recall	F1-Score	Support
0	0.92	0.96	0.94	105
1	0.86	0.78	0.82	91
2	0.76	0.82	0.79	92
3	0.92	0.88	0.90	112
Accuracy			0.86	400
Macro Avg	0.86	0.86	0.86	400
Weighted Avg	0.87	0.86	0.86	400



xgBoost(Hyperparameter Tuning)

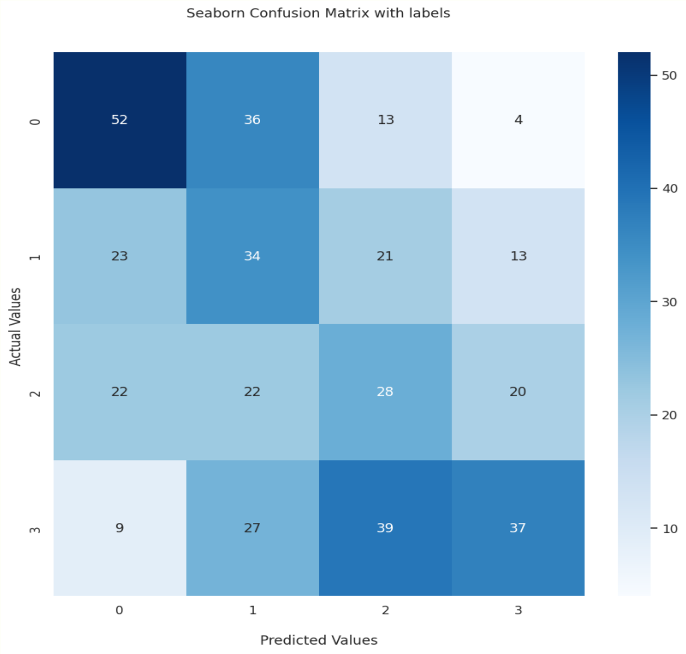
Classification report for xgBoost (Test set)

	Precision	Recall	F1-Score	Support
0	0.94	0.93	0.94	105
1	0.86	0.90	0.88	91
2	0.84	0.88	0.86	92
3	0.95	0.89	0.92	112
Accuracy			0.90	400
Macro Avg	0.90	0.90	0.90	400
Weighted Avg	0.90	0.90	0.90	400



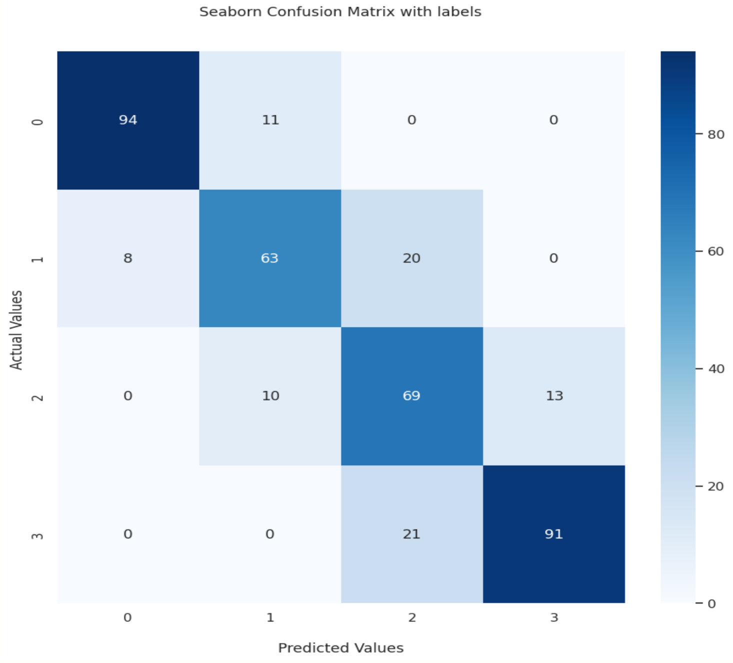
KNN classifier

	Precision	Recall	F1-Score	Support
0	0.92	0.90	0.91	105
1	0.75	0.69	0.72	91
2	0.63	0.75	0.68	92
3	0.88	0.81	0.84	112
Accuracy			0.79	400
Macro Avg	0.79	0.79	0.79	400
Weighted Avg	0.80	0.79	0.80	400



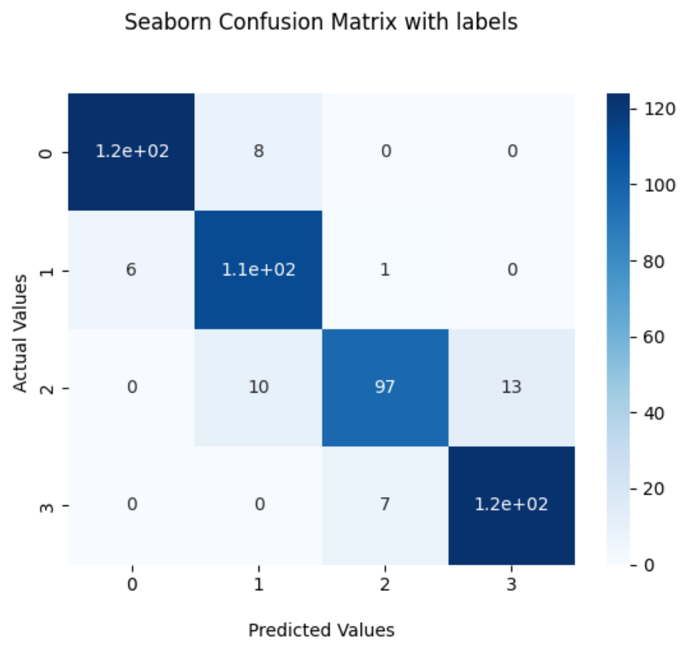
Naïve Bayes

	Precision	Recall	F1-Score	Support
0	0.92	0.90	0.91	105
1	0.75	0.69	0.72	91
2	0.63	0.75	0.68	92
3	0.88	0.81	0.84	112
Accuracy			0.79	400
Macro Avg	0.79	0.79	0.79	400
Weighted Avg	0.80	0.79	0.80	400

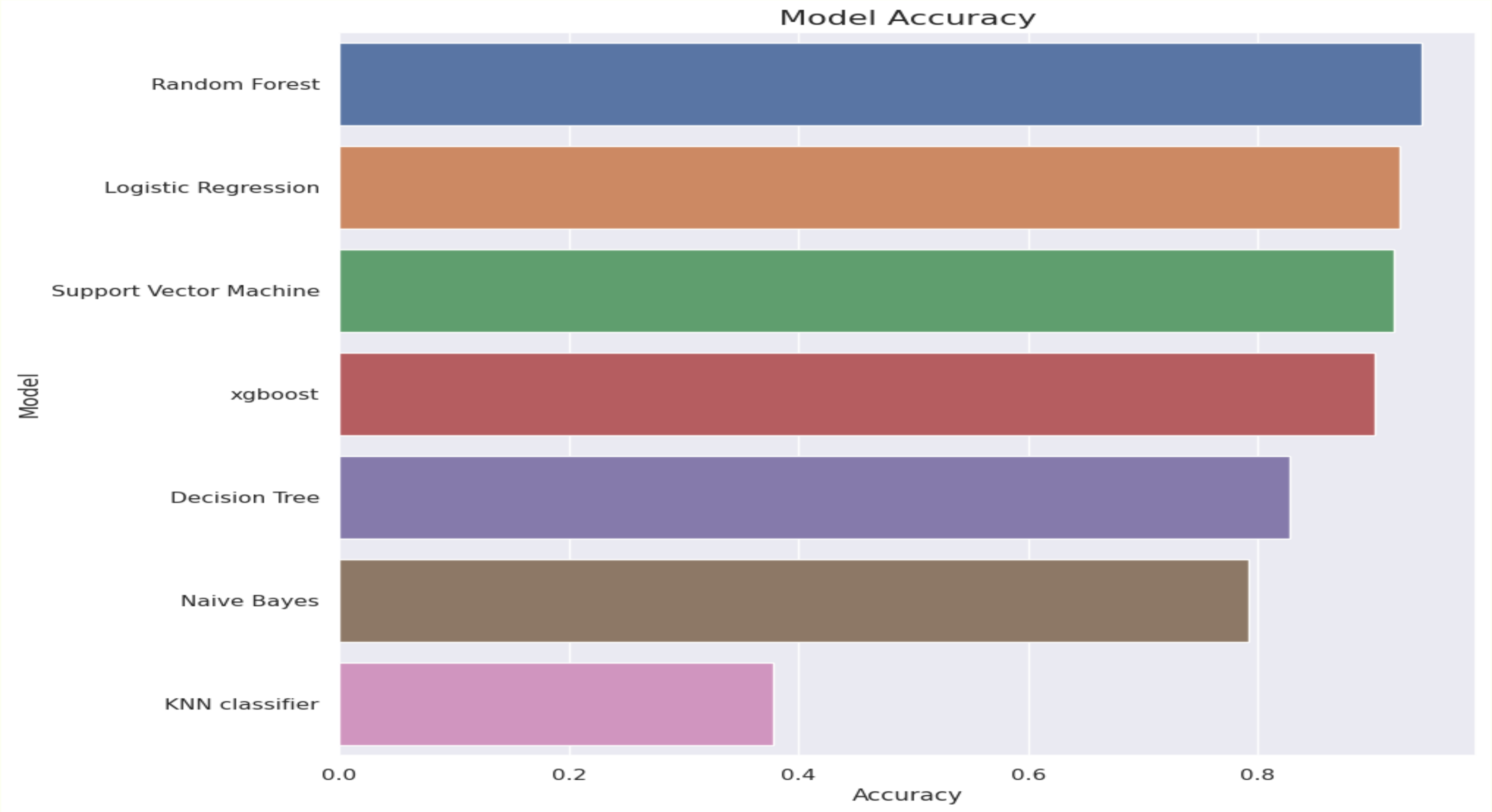


Support Vector Machine

	Precision	Recall	F1-Score	Support
0	0.95	0.94	0.95	132
1	0.86	0.94	0.90	118
2	0.92	0.81	0.86	120
3	0.90	0.95	0.92	130
Accuracy			0.91	500
Macro Avg	0.91	0.91	0.91	500
Weighted Avg	0.91	0.91	0.91	500



Model Performance



Conclusions of Modelling

1. The linear regression model is least accurate as it has very high coefficient of Assortment categories and Store type categories and it neglected features like customers , promotions which has positive correlation with sales , so we will use hyperparameter tuning to impose penalties on coefficients.
2. Decision Tree Model density distribution plot of sales varies highly with real data of sales.
3. Random Forest Regression has 99% accuracy for train data but 96% for test data, so this type of model cant be trusted , as the difference between train -test is very high
4. The most accurate models are Ridge , Lasso and Elastic-Net Regression , there train-test performances are almost similar and coefficient's are also similar.
5. The week of year line plot shows that Predicted Sales follows Actual Sales , with variation of mostly 700 dollars , except for last 2 week of the year

Thank You