Capstone Project - 3

Individual : Mobile Price Range Prediction

Member

Rajat Mishra

Let’s Catch The Defaulters

1. Defining problem statement
2. EDA and feature engineering
3. Feature Selection
4. Preparing dataset for modeling
5. Applying Model
6. Model Validation and Selection

The Dilemma

, I will use historical data to identify which phone features and specifications contribute to a higher or lower price range. By analyzing the data, I will create a predictive model that can identify which mobile phone models are relatively expensive or inexpensive.

To achieve this, I will break down the project into four parts. The first part is data collection, where I will gather relevant data on various mobile phone models and their corresponding prices. This data will include factors such as the phone's brand, processor, camera quality, storage capacity, and screen size.

The second part of the project is data cleaning, where I will remove any inconsistencies or errors in the data to ensure the model's accuracy. I will also transform the data into a format that is easy for the model to use.

The third part of the project is feature selection, where I will determine which factors have the most significant impact on a mobile phone's price range. By analyzing the data, I will identify which features are correlated with higher or lower prices and use them as inputs for the predictive model.

The fourth and final part of the project is model training, where I will use machine learning algorithms to develop a predictive model that can accurately predict mobile phone prices based on their features and specifications. This model will be trained on the historical data and tested on new data to ensure its accuracy and reliability.

- With this predictive model, potential buyers can use the phone's specifications to estimate the price range they should expect to pay. Additionally, mobile phone manufacturers and sellers can use this model to price their products competitively and increase their sales.

Data Pipeline

- Data processing-1: In this first part we’ve removed unnecessary features. Since there were nearly many columns with all null values.
- Data processing-2: In this part, we manually go through each features selected from part 1, And encoded the categorical features ,changed the columns containing date time values .
- EDA: In in this part, we do some exploratory data analysis (EDA) on the features selected in part-1 and 2 to see the trend.
- Create a model: Finally, In this last but not the last part, we create models. Creating a model is also not an easy task. It’s also an iterative process. we show how to start with a with a simple model, then slowly add complexity for better performance.

Data Summary

Data Summary

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.

Data Summary

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.

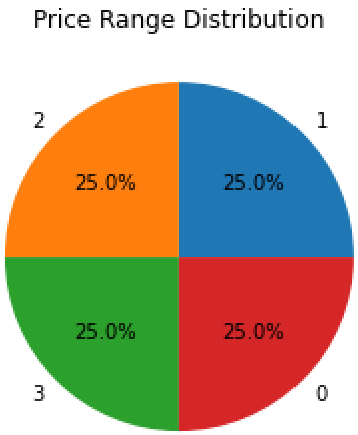
Sc_h - Screen Height of mobile in cm.

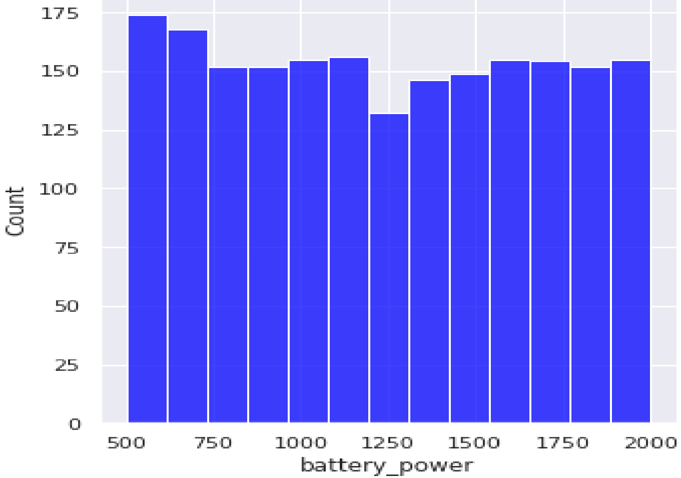
Sc_w - Screen Width of mobile in cm.

Talk_time - longest time that a single battery charge will last when you are.

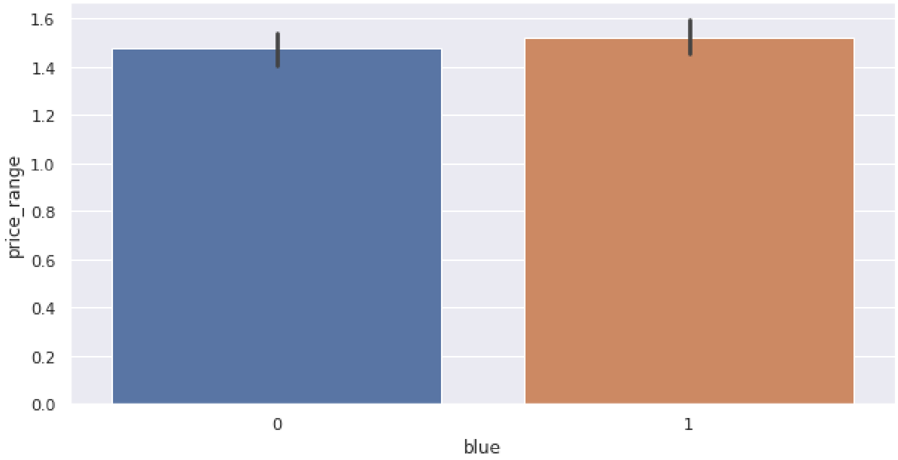
Three_g - Has 3G or not.
Wifi - Has wifi or not.
Price_range - This is the target variable with value of 0(low cost), 1(medium cost),2(High Cost),3(Very High cost).

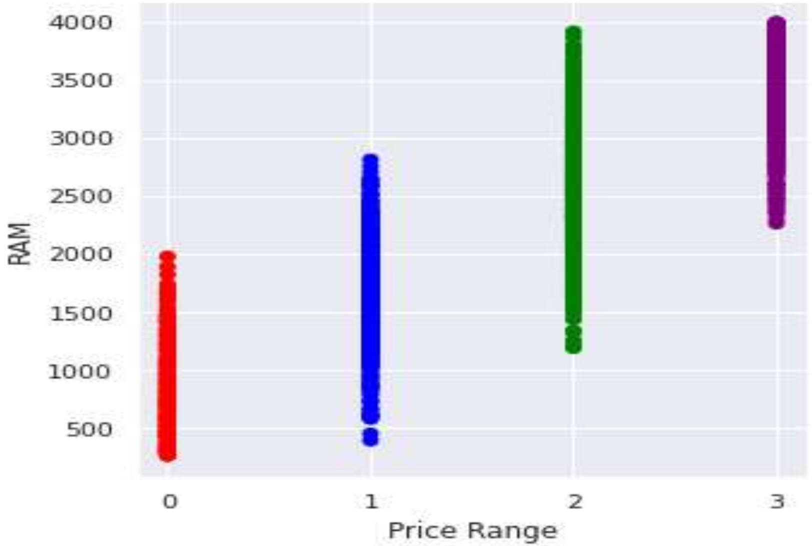
EDA



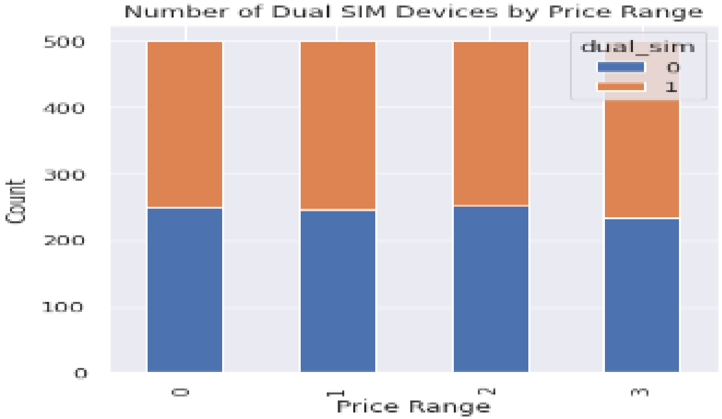


1.All category phones are distributed with equal price range
2.This plot visualizes how the battery capacity, measured in mAh, is distributed across the dataset. We can observe that the distribution of battery capacity is positively correlated with the price range of the mobile phones, as there is a gradual increase in the battery capacity as the price range increases. This suggests that there is a strong relationship between the battery capacity and the price of a mobile phone, and that consumers may be willing to pay more for a mobile phone with a higher battery capacity.

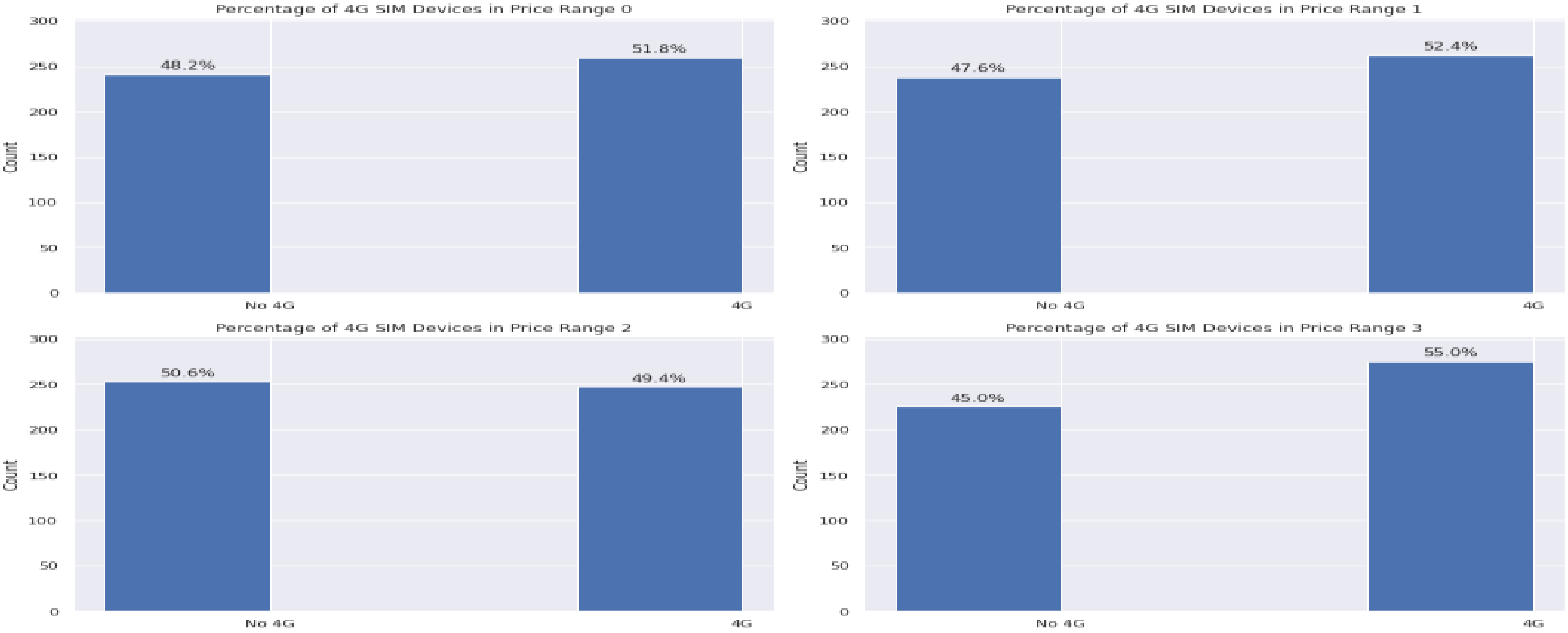




1.Almost half the devices have Bluetooth, and half don’t.
2.The scatter plot shows a clear positive correlation between RAM and price range, with the majority of the data points clustering towards the upper right corner. This suggests that as the price range increases, the amount of RAM in the device generally increases as well.



We can observe that upto low,medium,high almost it is same but for very high price range it is seen that it is found that the count is raised who using dual devices and count is increasing for dual devices.



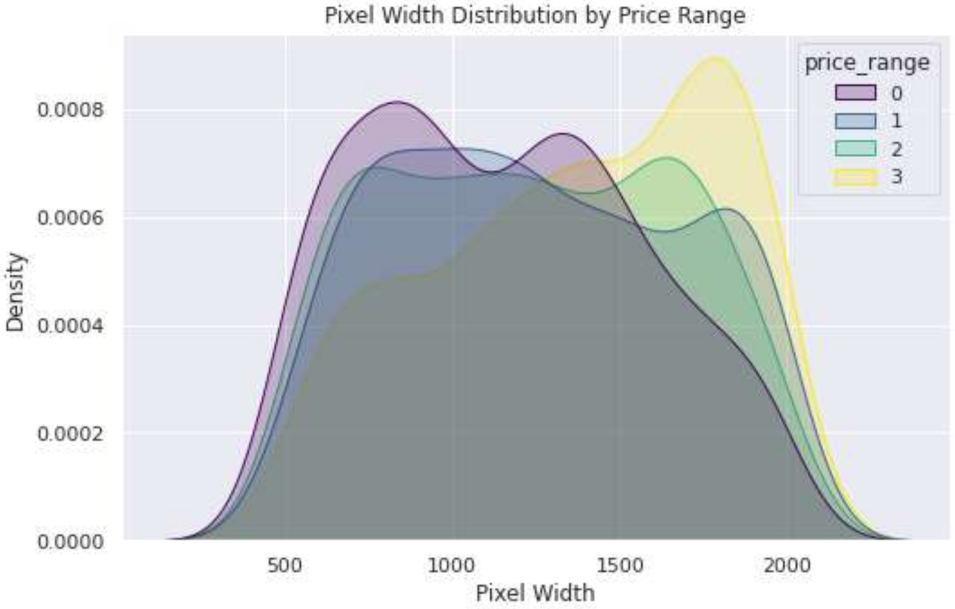
Percentage of 4G SIM Devices in Price Range 0

Percentage of 4G SIM Devices in Price Range 1

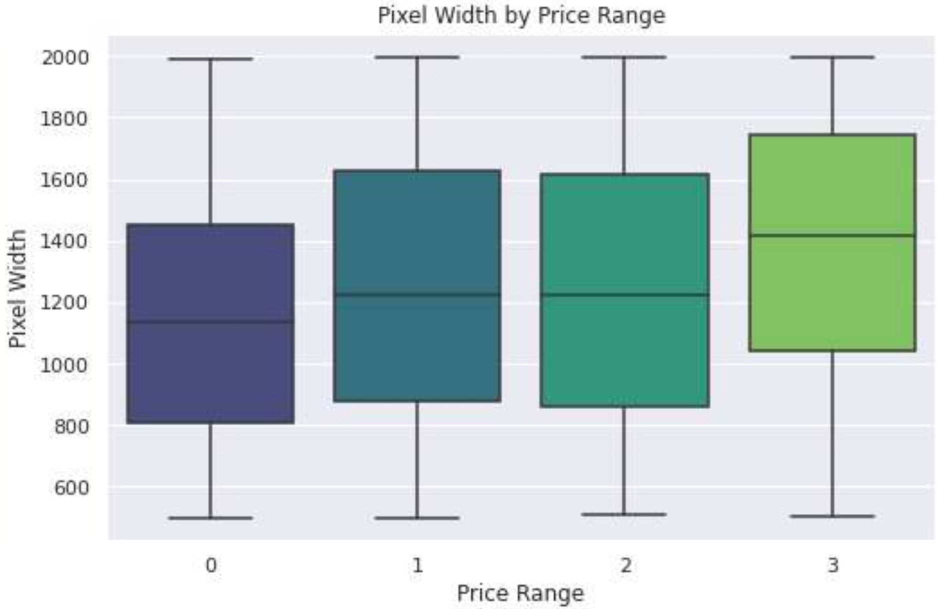
Percentage of 4G SIM Devices in Price Range 2

Percentage of 4G SIM Devices in Price Range 3

I have found that at low, medium,very high prices the mobile phones having sim in more numbers but at high prices it is showing slightly collapse.

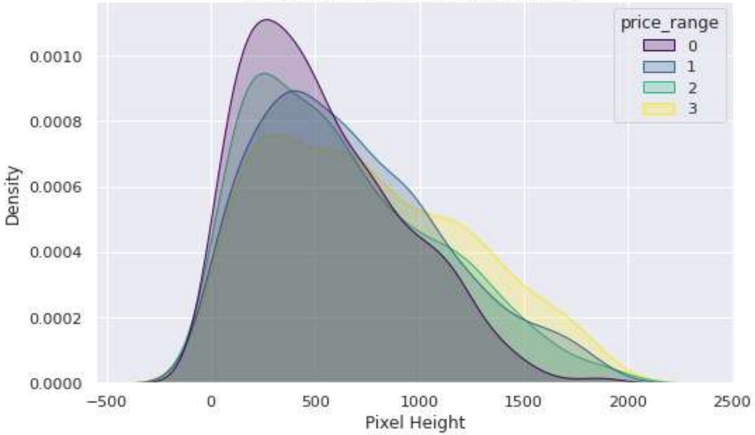


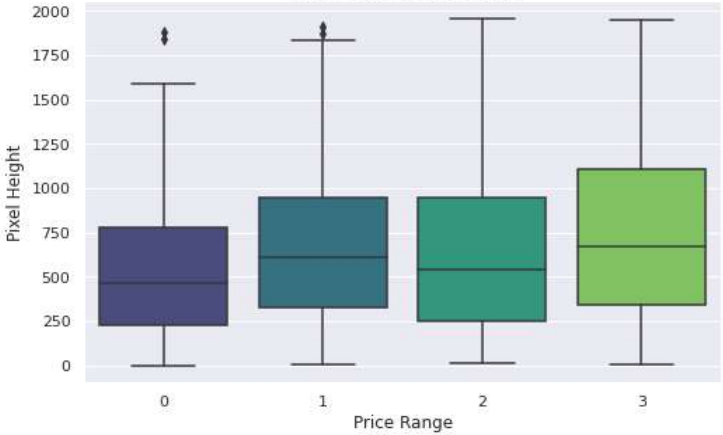
Pixel Width Distribution by Price Range



Pixel Width by Price Range

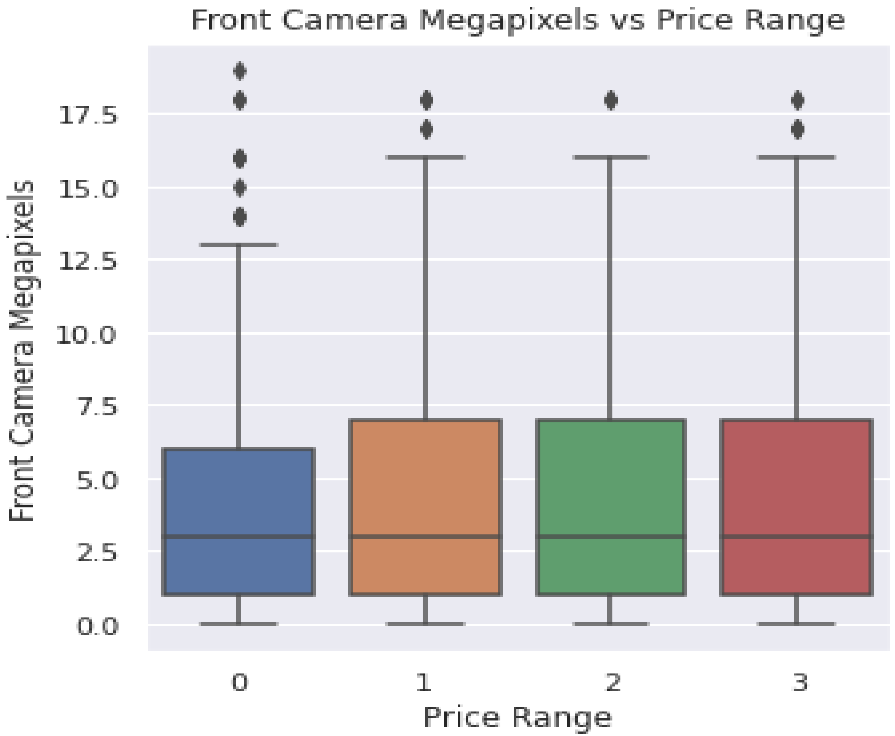
Futuristic Features


Pixel Height Distribution by Price Range


Pixel Height by Price Range

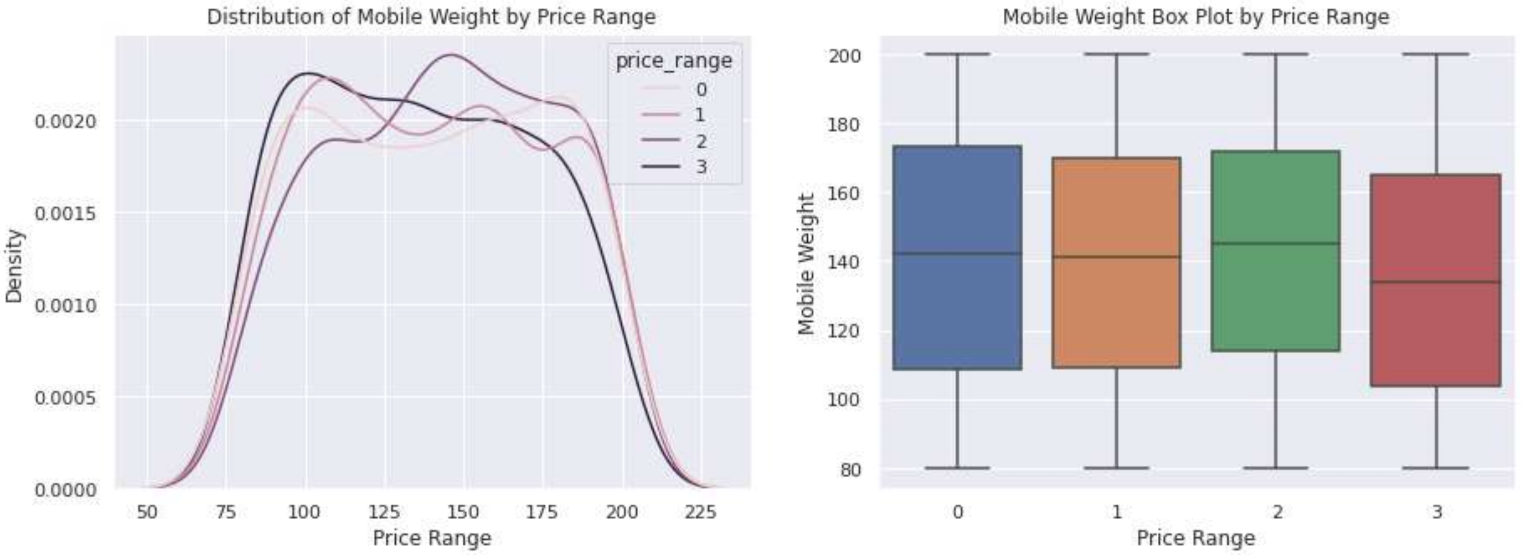
Based on the analysis of the pixel width distribution across different price ranges, it can be observed that there is not a continuous increase in pixel width as we move from low cost to very high cost mobile phones. In particular, mobile phones with medium cost and high cost have almost equal pixel width, indicating that this may not be the sole driving factor in deciding the price range of mobile phones. Other features such as processor, camera quality, storage capacity, and brand value may also play a significant role in determining the price range. Therefore, a holistic approach considering multiple factors is necessary for accurate pricing and positioning of mobile phones in the market.Pixel height is almost similar as we move from Low cost to Very high cost.little variation in pixel_height.

EDA



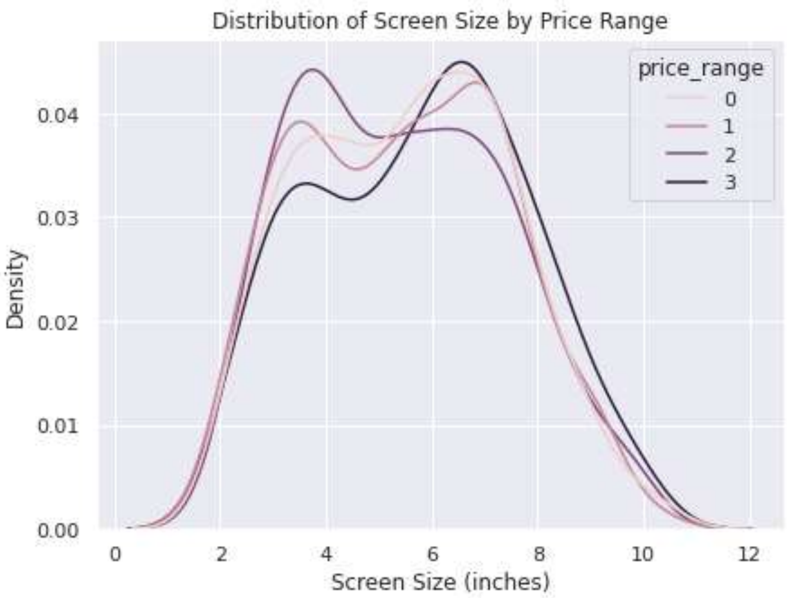
It is almost same impcact of price range in all categories.

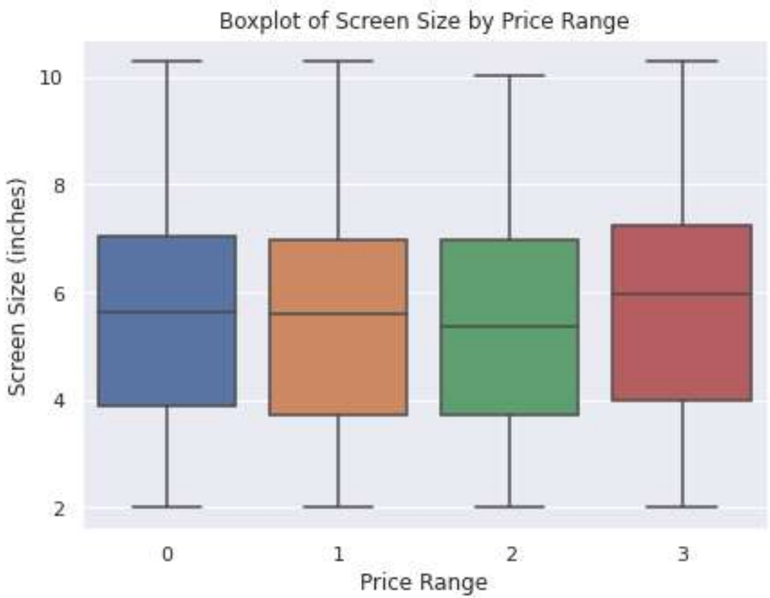
EDA (continued)



The distribution of primary camera megapixels across different target categories is relatively consistent, indicating that this feature may not significantly influence the price range of mobile phones. This consistency is a positive sign for prediction modeling, as it suggests that this feature may not be a major confounding factor in predicting the price range.

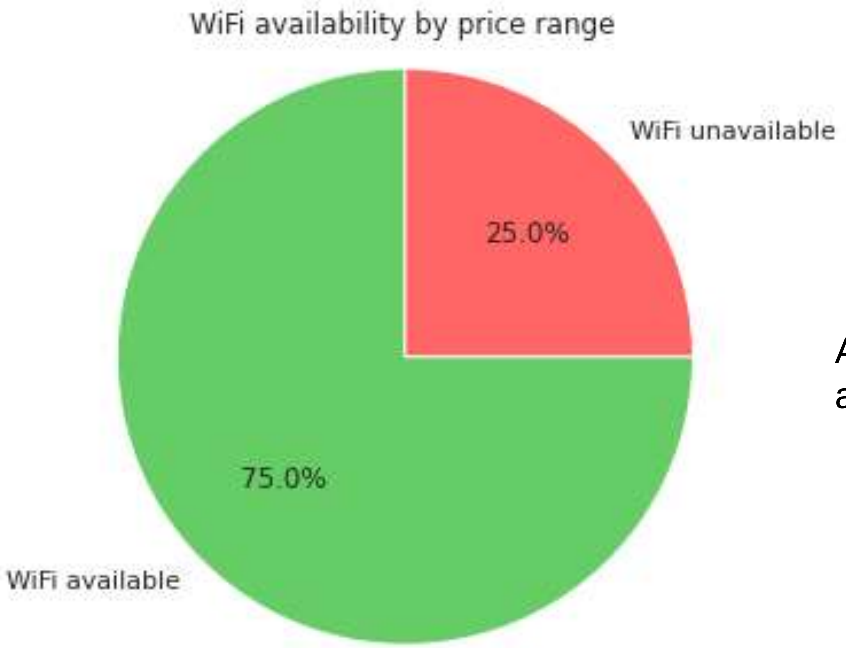
EDA **(continued)**


Distribution of Screen Size by Price Range


Boxplot of Screen Size by Price Range

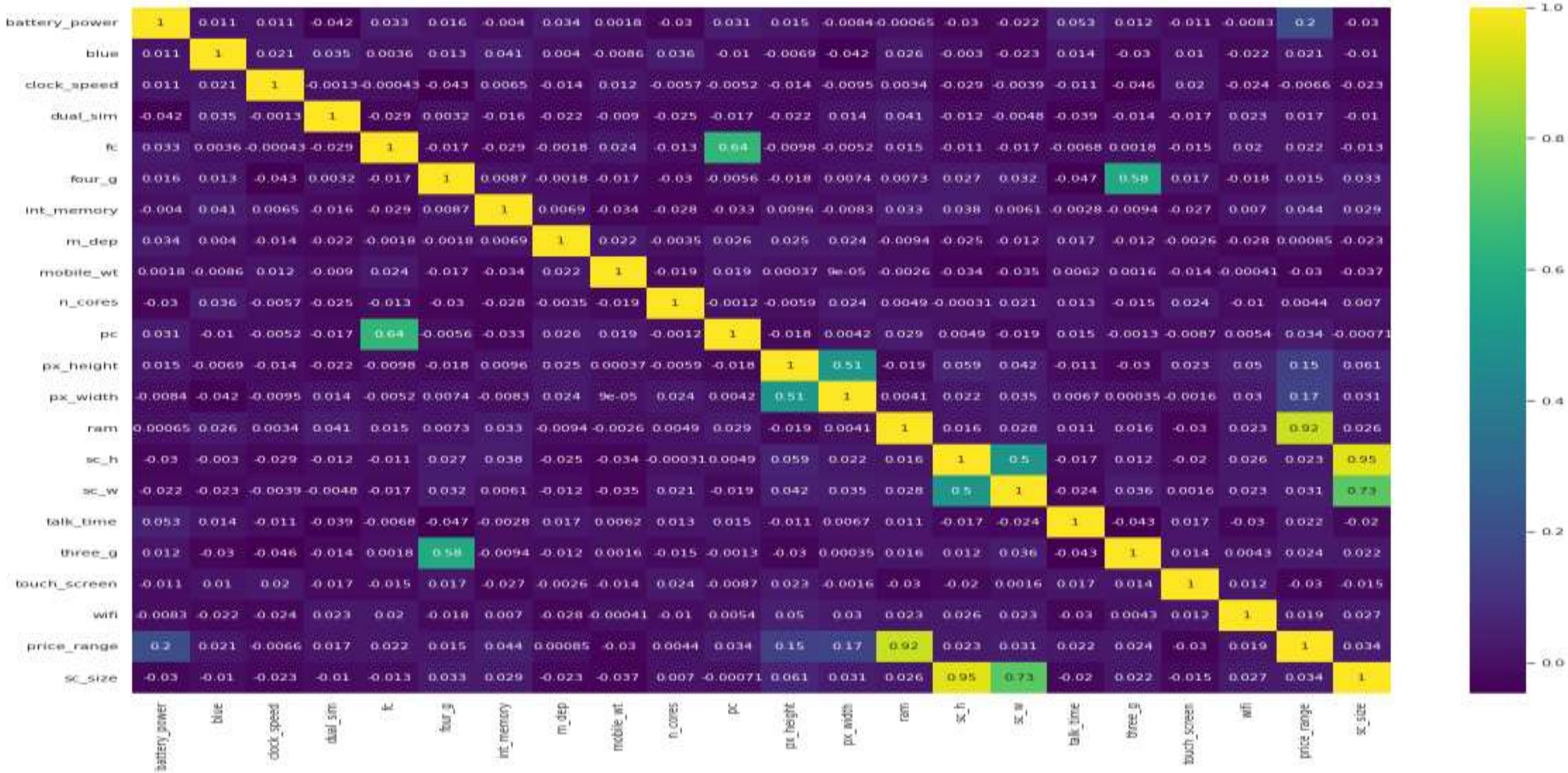
It can be observed that mobile phones with higher price ranges tend to be lighter in weight compared to lower price range phones.

EDA **(continued)**



Around in 25% the wifi is not available
and in 75% the wifi is availa

EDA (continued)



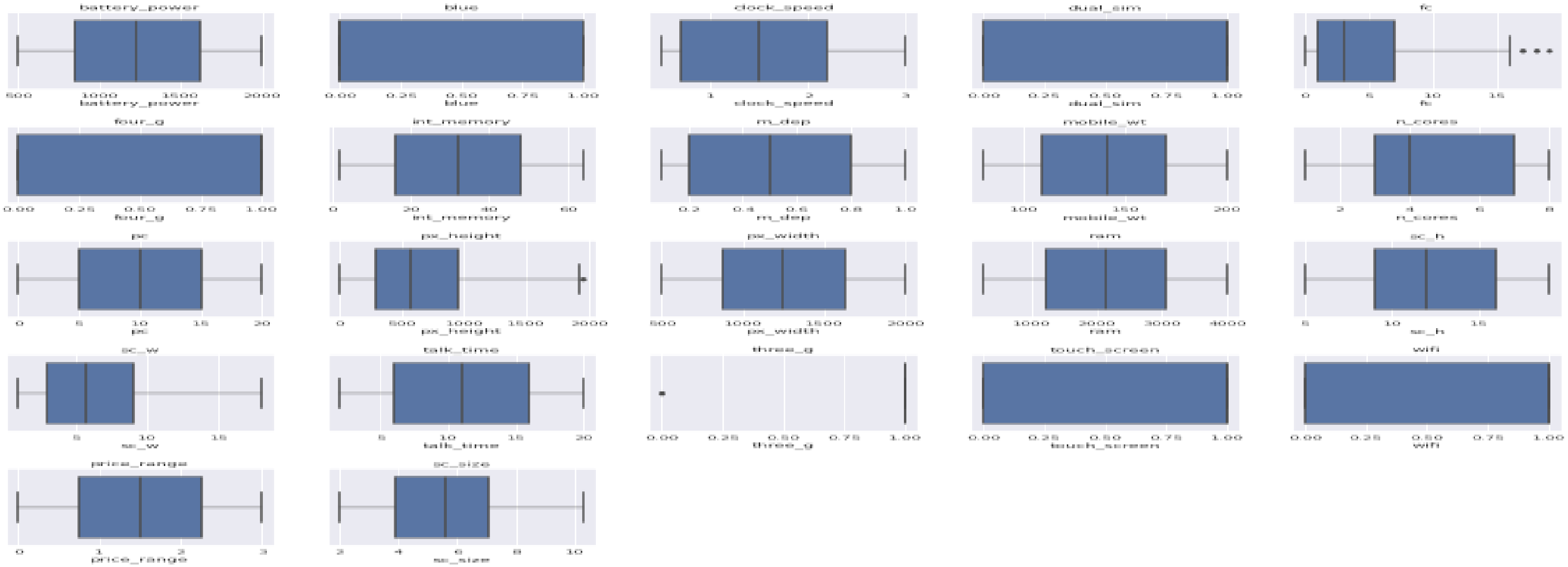
	battery_power	blue	clock_speed	dual_sim	fc	four_g	int_memory	m_dep	mobile_wt	n_cores	pc	px_height	px_width	ram	sc_h	sc_w	talk_time	three_g	touch_screen	wifi	price_range	sc_size
battery_power	1	0.011	0.011	-0.042	0.033	0.016	-0.004	0.034	0.0018	-0.03	0.031	0.015	-0.0084	-0.00065	-0.03	-0.022	0.053	0.012	-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	-0.003	-0.023	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	-0.029	-0.0039	-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	-0.012	-0.0048	-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	-0.011	-0.017	-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	0.027	0.032	-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	0.038	0.0061	-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	-0.025	-0.012	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	-0.034	-0.035	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	-0.00031	0.021	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	0.0049	-0.019	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	0.059	0.042	-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	0.022	0.035	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	0.016	0.028	0.011	0.016	-0.03	0.023	0.92	0.026
sc_h	-0.03	-0.003	-0.029	-0.012	-0.011	0.027	0.038	-0.025	-0.034	-0.00031	0.0049	0.059	0.022	0.016	1	0.5	-0.017	0.012	-0.02	0.026	0.023	0.95
sc_w	-0.022	-0.023	-0.0039	-0.0048	-0.017	0.032	0.0061	-0.012	-0.035	0.021	-0.019	0.042	0.035	0.028	0.5	1	-0.024	0.036	0.0016	0.023	0.031	0.73
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	-0.017	-0.024	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	0.012	0.036	-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	-0.02	0.0016	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	0.026	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	0.023	0.031	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	0.95	0.73	-0.02	0.022	-0.015	0.027	0.034	1

- **The high correlation between RAM and price_range is a positive sign for businesses as it indicates that RAM will be a major deciding factor in estimating the price range of a mobile phone.**

- **However, there are also some cases of collinearity in the data. Specifically, there is a correlation between the pairs of features ('pc', 'fc') and ('px_width', 'px_height'). These correlations make sense, as a phone with a good front camera is likely to have a good back camera, and an increase in pixel height typically corresponds with an increase in pixel width.**

- To address this collinearity, we could consider replacing the 'px_height' and 'px_width' features with a single feature representing the overall number of pixels in the screen. However, it is important to note that the 'fc' and 'pc' features should be kept separate, as they represent different aspects of the phone's camera capabilities (front camera megapixels vs. primary camera megapixels).

Outliers Handling



As we can see very less outliers are present so no need to remove

Model Implementation
Logistic Regression

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

           0       0.91      0.90      0.91       107
           1       0.69      0.76      0.72        83
           2       0.68      0.65      0.67        97
           3       0.85      0.84      0.84       113

    accuracy                           0.79       400
   macro avg       0.78      0.79      0.79       400
weighted avg       0.79      0.79      0.79       400
```

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

           0       0.93      0.88      0.90       421
           1       0.75      0.79      0.77       386
           2       0.73      0.79      0.76       379
           3       0.92      0.86      0.89       414

    accuracy                           0.83      1600
   macro avg       0.83      0.83      0.83      1600
weighted avg       0.84      0.83      0.83      1600

Cross-validation scores: [0.81   0.825  0.8375 0.81   0.8125]
Average cross-validation score: 0.8190000000000002
```

Xgboost

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

           0       0.91      0.91      0.91       105
           1       0.77      0.77      0.77        91
           2       0.66      0.76      0.71        92
           3       0.90      0.78      0.83       112

    accuracy                           0.81       400
   macro avg       0.81      0.81      0.80       400
weighted avg       0.82      0.81      0.81       400
```

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

           0       0.99      1.00      0.99       395
           1       0.99      0.98      0.99       409
           2       0.99      0.99      0.99       408
           3       1.00      1.00      1.00       388

    accuracy                           0.99      1600
   macro avg       0.99      0.99      0.99      1600
weighted avg       0.99      0.99      0.99      1600
```

```
Cross-validation score: 0.8150000000000001
Classification Report for XGBoost(Test set)=
              precision    recall  f1-score   support

           0       0.91      0.93      0.92       105
           1       0.76      0.76      0.76        91
           2       0.66      0.72      0.69        92
           3       0.89      0.80      0.85       112

    accuracy                           0.81       400
   macro avg       0.80      0.80      0.80       400
weighted avg       0.81      0.81      0.81       40
```

Random Forest Classifier

```
              precision    recall  f1-score   support

           0       0.92      0.94      0.93       105
           1       0.78      0.75      0.76        91
           2       0.63      0.72      0.67        92
           3       0.87      0.78      0.82       112

    accuracy                           0.80       400
   macro avg       0.80      0.80      0.80       400
weighted avg       0.81      0.80      0.80       400
```

```
              precision    recall  f1-score   support

           0       0.89      0.96      0.93       105
           1       0.80      0.73      0.76        91
           2       0.64      0.71      0.67        92
           3       0.86      0.79      0.83       112

    accuracy                           0.80       400
   macro avg       0.80      0.80      0.80       400
weighted avg       0.81      0.80      0.80       400
```

```
     precision    recall  f1-score   support

           0       0.91      0.97      0.94       395
           1       0.88      0.84      0.86       409
           2       0.88      0.84      0.86       408
           3       0.92      0.95      0.94       388

    accuracy                           0.90      1600
   macro avg       0.90      0.90      0.90      1600
weighted avg       0.90      0.90      0.90      1600
```

Conclusion

Based on the exploratory data analysis (EDA), we observed that the mobile phones in the dataset are divided into four different price ranges, each having a similar number of elements. Additionally, we found that approximately half of the devices have Bluetooth, while the other half do not. Furthermore, we noted that as the price range increases, there is a gradual increase in battery power, and RAM shows continuous growth from low-cost to very high-cost phones. Moreover, the costly phones tend to be lighter than the lower-priced ones.

Our analysis indicates that RAM, battery power, and pixel quality are the most significant factors affecting the price range of mobile phones. From our experiments, we concluded that logistic regression and XGBoost algorithms with hyperparameter tuning yielded the best results in predicting the price range of mobile phones.

In summary, the EDA revealed that the dataset consists of mobile phones grouped into four price ranges, with similar numbers of devices in each range, and a 50-50 distribution of Bluetooth. We also observed that RAM and battery power increase with the price range, and higher-priced phones tend to be lighter. Our experiments suggest that the most important factors affecting the price range of mobile phones are RAM, battery power, and pixel quality. Finally, we found that logistic regression and XGBoost algorithms, coupled with hyperparameter tuning, provide the best performance in predicting the price range of mobile phones.