Exploratory Data Analysis (STAT-S - 670)

Final Project

April / 08 / 2022

Team Members:

Rajesh Sharma - rajeshar

Ashish Patidar - apatida

INTRODUCTION

A mobile phone consists of various components and the addition of each component adds to the overall cost of a mobile phone. Since there are various qualities of a component, a $500 smartphone can have a 16MP primary camera and a $1000 smartphone can have a 16MP camera, but, intuitively, that the camera quality of a $1000 phone will be much better than the $500 one.

This point motivated us to analyze the components that impact the price of a mobile phone and create a logistic model using the components analyzed to categorize a mobile phone into a price category of the low, medium, high, and very high.

We will try to answer three questions after analyzing the data:

1. Which is the primary component in the mobile phone affecting its price?
2. Which kind of model can be used for this classification?
3. Which combination of components gives us the best accuracy for price prediction?

DATASET DESCRIPTION AND PRE-PROCESSING

Our dataset is available on Kaggle in the form of a CSV file. This dataset contains 2000 rows and 14 columns. These 14 columns are as follows:

- Battery_power: the measure of battery in mAh
- Clock_speed: the measure of clock speed in hertz
- pc(primary camera): megapixel of the camera
- Ram: the measure of ram in MegaBytes
- Int_memory: the measure of internal memory in GigaBytes
- Px_height: pixel resolution height
- Px_length: pixel resolution length
- Sc_height: screen height in cms
- Sc_length: screen length in cms
- N_cores: number of processor cores
- Four_g: binary variable: 0 for four_g not present, 1 for four_g present
- Dual_sim: binary variable: 0 for dual_sim not present, 1 for dual_sim present
- Touch_screen: binary variable: 0 for touch_screen, not present, 1 for touch_screen present
- Price: Categorical variable which classifies mobile phones into four categories: low, medium, high, and very high

Our dataset does not have any null values for any feature. Therefore we did not remove any row from the dataset.

Since there are 13 features to build a model, it is too much to visualize and analyze. Therefore, we used our intuition and research to combine the relative variables.

- Sc_length and sc_width are combined using the Pythagoras theorem to calculate the diagonal length of the screen.
- Px_width and px_length are multiplied
- by N_cores and clock_speed are multiplied and termed as performance as they both impact the performance of mobile phones.

ANALYSIS AND VISUALIZATION

---

7 continuous and 3 binary predictor variables(which one)?

We did the univariate and bivariate analysis and both were categorized on the price column.

Univariate Analysis:

For univariate analysis, We did count plots (Figure 1-3) for binary variables and categorized them based on the price of univariate analysis. After seeing the plots, we can say that each price category has almost the same count for each binary variable.

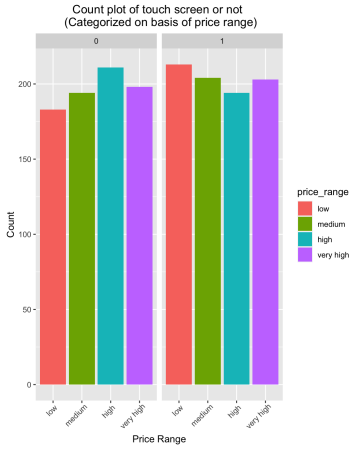

Figure 1

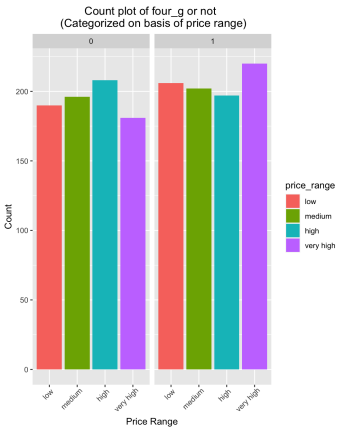

Figure 2

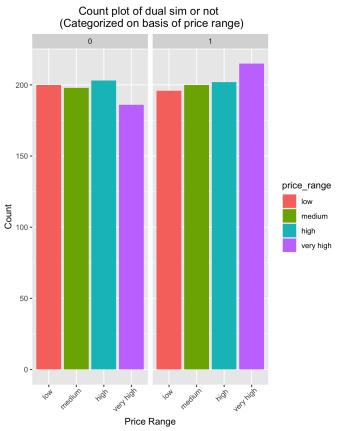

Figure 3

We plotted density plots for continuous variables(Figure 4-10) categorized on the price variable. Except for the RAM, the density distribution of the other continuous variable is the same for all price ranges. But for RAM, the distribution of different price categories can be distinguished from each other. As can be seen from the Density plot (Figure 4) of the RAM variable, the low price range is distributed from 0 to 2000 and the very high range is distributed from 2000 to 4200.

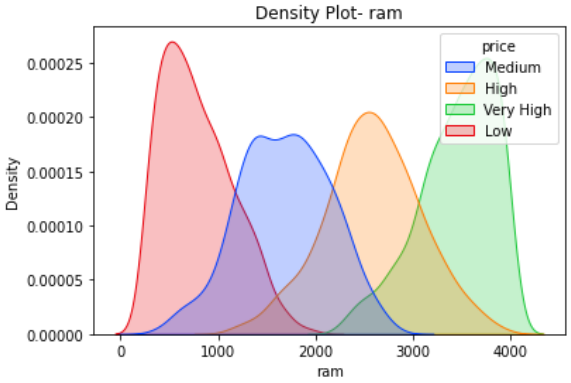

Figure 4

Figure 5

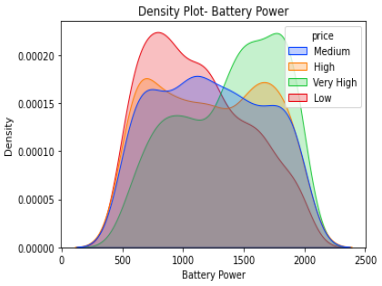

Figure 6

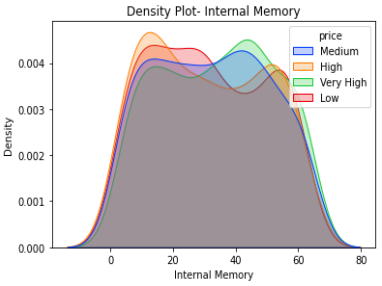

Figure 7

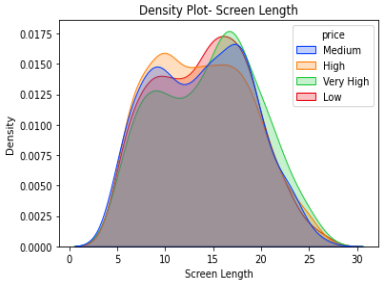

Figure 8

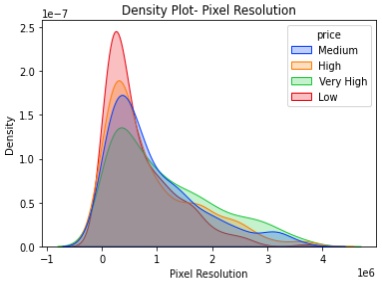

Figure 9

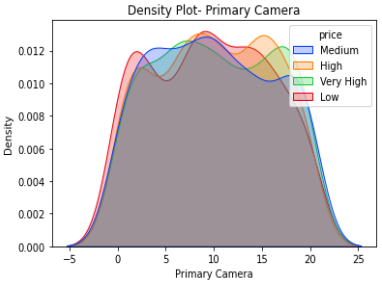

Figure 10

Bivariate Analysis:

- ## Correlation Plot & Pair Plots

  After univariate analysis, we plotted correlation plots(Figure 11-12) of all variables and from that, we can conclude that RAM is the only predictor variable that is highly correlated with our response variable price. The other variables have very little or almost zero correlation values. But if we consider correlation values with only the response variable, after RAM, battery power and pixel resolution have the highest values; although their values are very low compared to other variables, it is still noticeable. So we can say that both these variables, battery power, and pixel resolution, can be interesting to combine with RAM for price prediction.

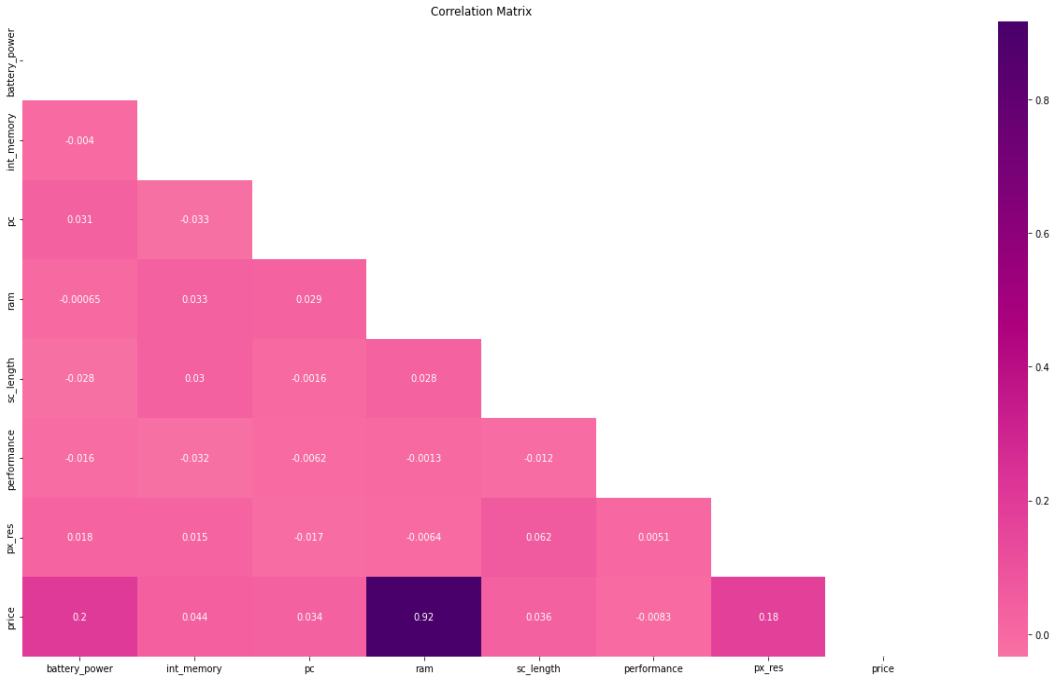

Figure 11

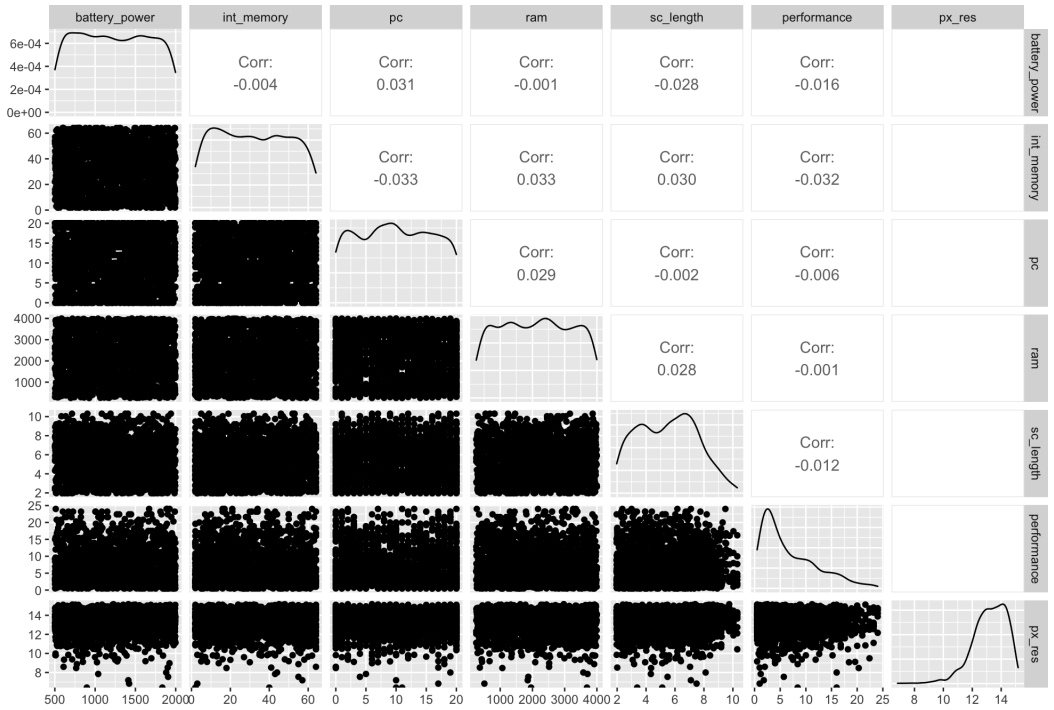

Figure 12

From the above pair plots(Figure 12) we can infer that there is no correlation between the variables.

- Mean vs price for each variable

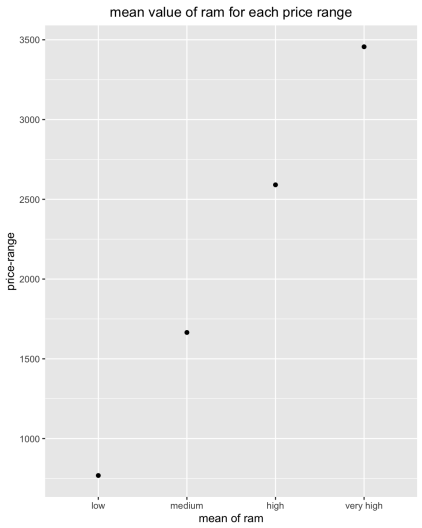

Figure 13

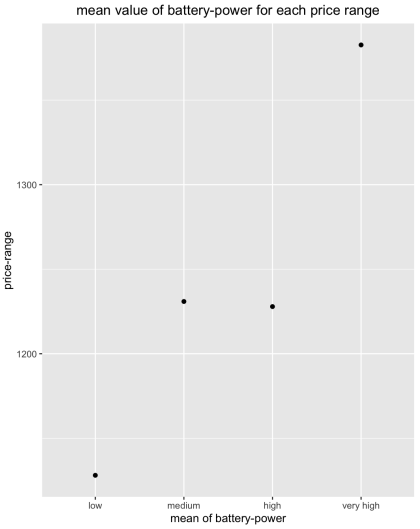

Figure 14

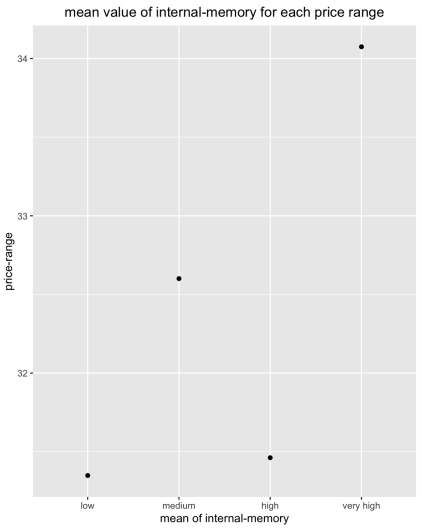

Figure 15

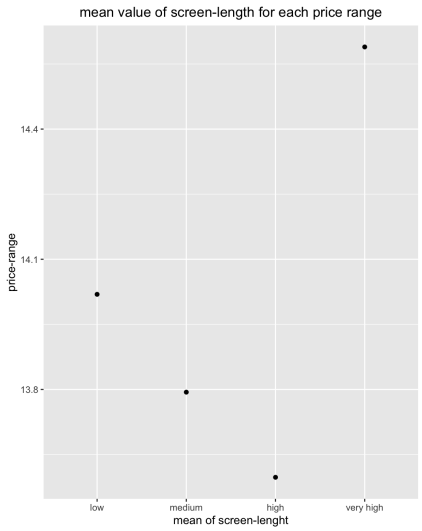

Figure 16

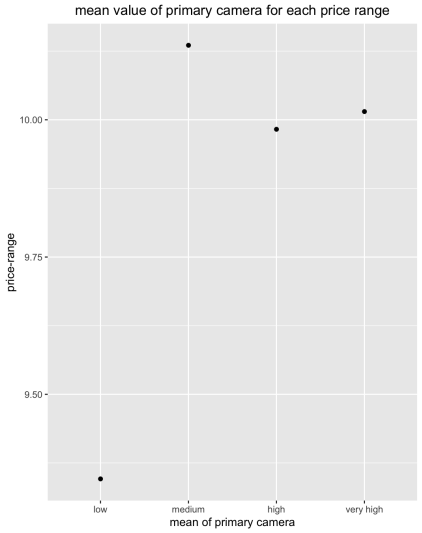

Figure 17

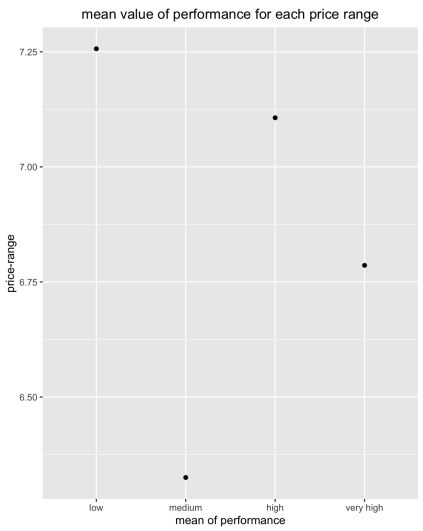

Figure 18

From the graphs(Figure 13-28) we can see the trend of mean values for each continuous variable with the price range. For RAM, the mean trend(Figure 13) is clear and is in increasing order. Price and mean of RAM are in a linear relationship. Similarly the mean of battery power(Figure 14) and mean of pixel resolution(Figure 15) are also almost in a linear relationship with price. And for the rest of the variables, it is hard to define any particular trend.

- Bivariate interaction of each continuous variable with another
  - ☐ **Relation between RAM vs All**

    Plotting the scatter plot of RAM with all other variables gives us some interesting insights. After visualizing plots(Figure 19-22) of RAM with performance, screen length, internal memory, and primary camera, we can conclude that if one keeps RAM constant, he can get a mobile phone with a better(or higher) specification

of these variables within the same price range. In the plots, the smoother line is almost parallel to the x-axis, which suggests the same trend. For example, to get a mobile phone in the lower price range, we have to choose RAM between 200 to 1000 MB, and for a primary camera, we also have a choice from 0 to 20 MegaPixel.

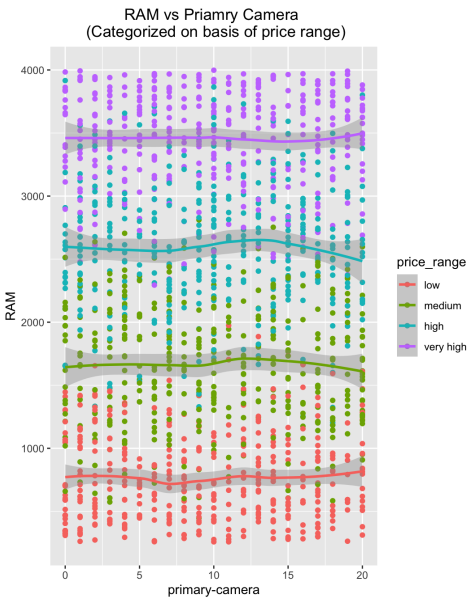

Figure 19

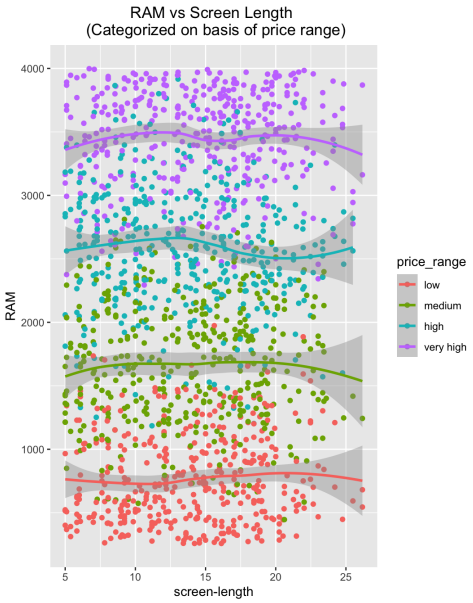

Figure 20

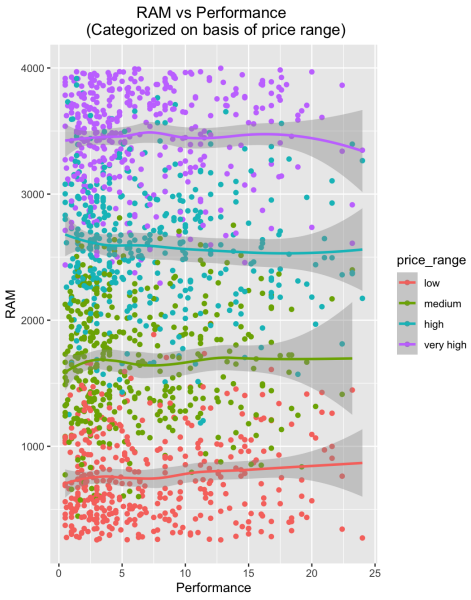

Figure 21

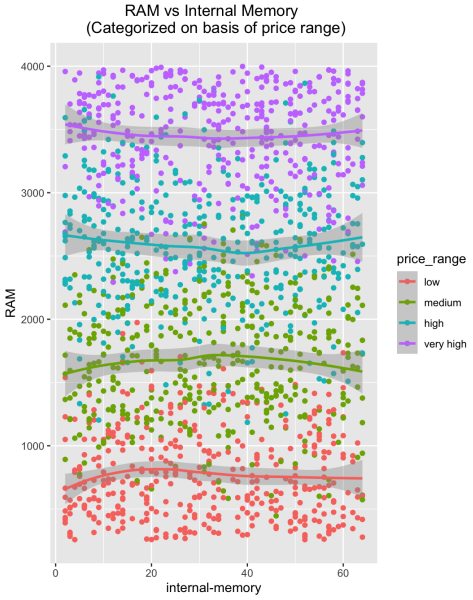

Figure 22

But that's not the case with battery power and pixel resolution variables. The plot(Figure 23-24) of RAM with Battery Power and Pixel Resolution suggests that if we want to buy a mobile phone with higher battery power and pixel resolution but within the same price range, we must compromise on the RAM requirements. As in the plots, the smoother line is inclined downwards, which suggests the same trend; for example, to get a mobile phone in the low price range, we can get 1000 MB of ram and 500 MAH battery power, but to get a mobile phone of 2000 MAH battery power we have to pick mobile of 700 MB RAM to remain in the same price range.

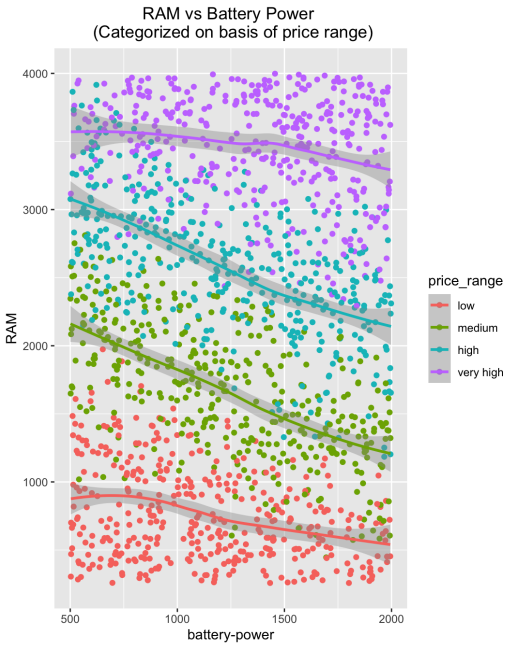

Figure 23

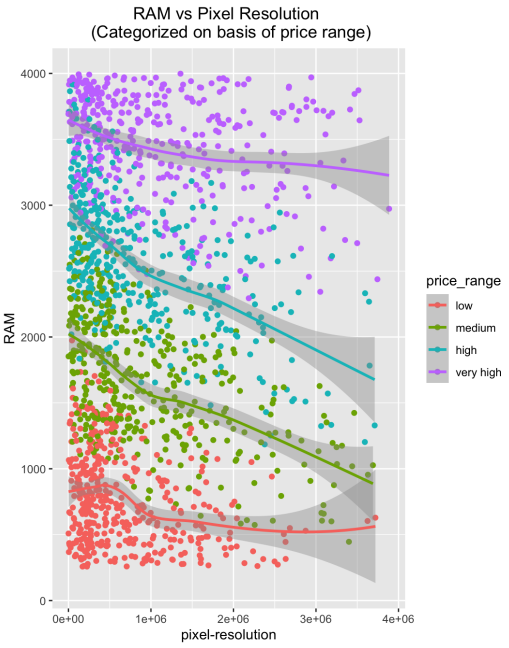

Figure 24

But looking at the plots(Figure 23-24) carefully, we can infer that one can get both high RAM and high battery power or pixel resolution within the same price range for the very high price range.

☐ Relation between rest of the variables

The Scatter plot of all other variables(Figure 25-42) without RAM did not give much inference for all models. Few of the plots(Figure 25-30) are below and the rest of them(Figure 31-42) are in the appendix section. We can see that there is no particular trend for any price range, all category points are scattered all around.

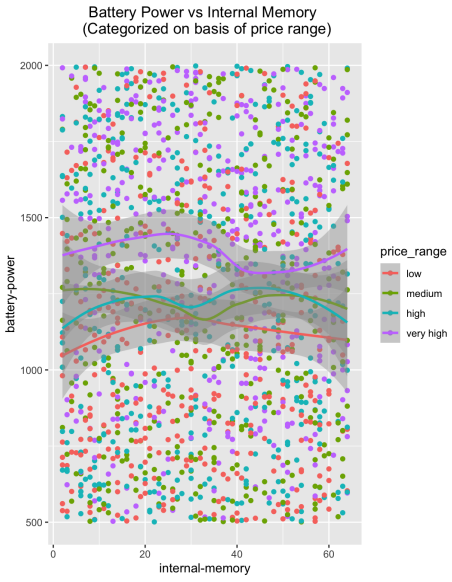

Figure 25

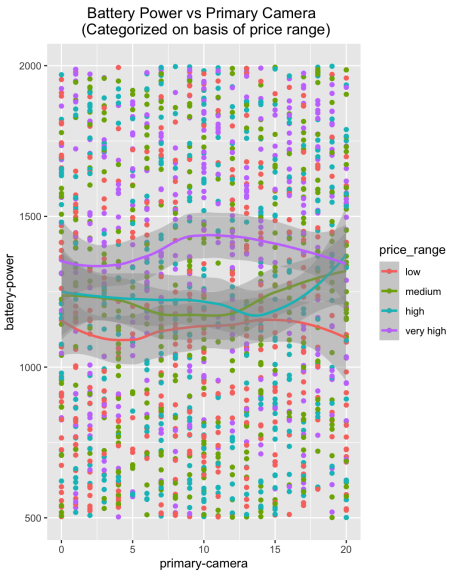

Figure 26

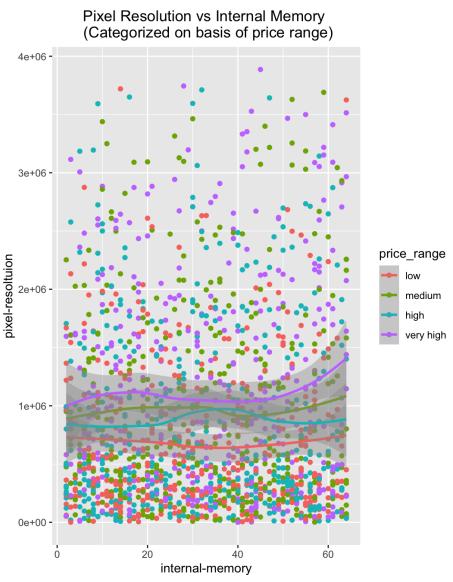

Figure 27

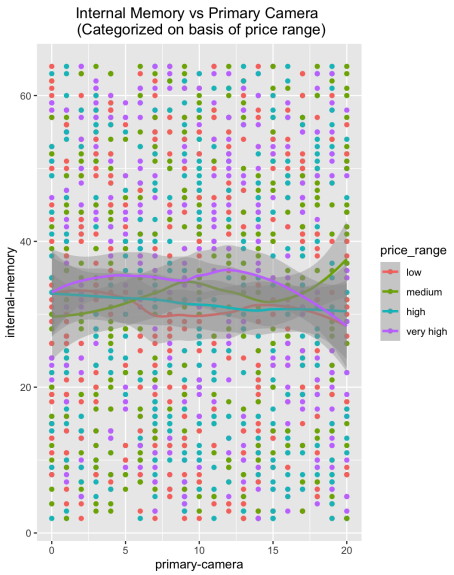

Figure 28

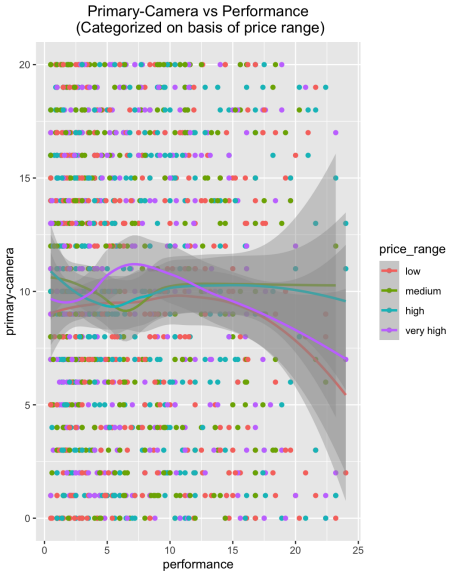

Figure 29

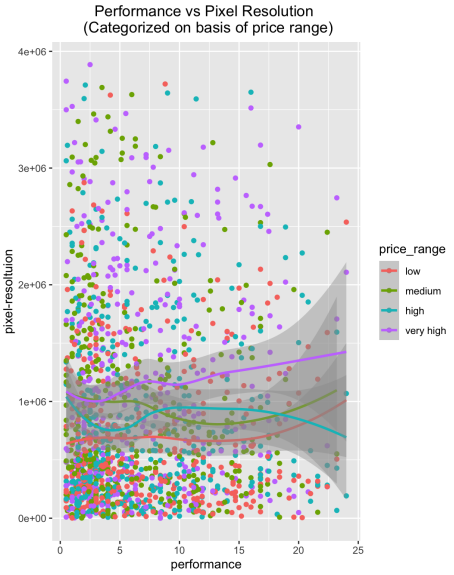

Figure 30

MODELS AND ACCURACY

Our response variable price range is a categorical variable, categorized into low, medium, high, and very high. So we have a few choices of models available for classification, like the proportional odds logistic model and the multinomial logistic model.

Since our data is artificially generated and has an equal number of mobile phones for all the price categories, we cannot implement a proportional odds logistic model as the coefficient values are either 0 or very close to zero. The other choice we have is the multinomial logistic model for which we used the **nnet** library of R.

From the above analysis, it is pretty clear that RAM is the most important component for the classification of mobile phones on the basis of the price range. Apart from RAM components like battery power and pixel resolution can also act useful.

On the basis of our analysis, we have created several multinomial logistic models and predicted the price range of both test and train data, and checked the accuracy of the model on both datasets.

The below table shows the predictors, response, training data accuracy, and test data accuracy for important predictors we have analyzed:

Predictor(s)	Response	Training data accuracy	Test data accuracy
RAM	Price Range	76.06%	74.25%
Battery Power	Price Range	31.37%	32.75%
Pixel Resolution	Price Range	29.81%	29.75%
Ram + Battery Power	Price Range	82..25%	82.25%
Ram + Pixel Resolution	Price Range	80.88%	80.00%
Ram + Battery Power + Pixel Resolution	Price Range	93.44%	91.75%
All variables(continuous)	Price Range	93.88%	91.25%

From the table, we can see that training data accuracy and test data accuracy are almost the same so our models are not overfitted or under-fitted.

We can see from the table that RAM is the primary predictor, and Battery power and Pixel resolution are individually not of much use but when combined with RAM they increase the accuracy of the model by a good margin.

For our test data, the model with RAM, battery power, and pixel resolution is giving better or approximately the same accuracy compared to the model with all the variables. This shows that the combination of RAM, battery power, and pixel resolution is the best combination for price classification based on our data. Adding more components to this combination is just increasing unwanted noise in our model.

INTERPRETATION OF ANALYSIS AND RESULT

Correlation between all the predictors we have is very low or almost close to 0, i.e. our predictors do not have any kind of linear relationship with each other. There is a more complex kind of relationship between them.

Coming to the questions we considered, RAM is the most important or primary component for mobile price classification. The multinomial logistic model performs well for the components we selected and the combination of RAM, battery power, and pixel resolution is the best combination to classify price range using multinomial logistic models.

LIMITATIONS

Since, our dataset is artificial we cannot be sure that our model will predict the correct price range for real-world mobile phones and as the data is artificial we cannot fit the proportional odds logistic model which may have given us a better prediction result. We cannot be sure about the prediction from the proportional odds logistic model but we would have better interpretability of the model.

FUTURE SCOPE

Other machine learning classifiers like K Nearest Neighbor, Random Forest, or a more complex classifier can be used to create the model but this will reduce the interpretability of the model.
We can try to gather real-world data for mobile phones and try to analyze that and create a model based on that concerning the real world.
We can even use techniques like LDA for a better selection of features to improve the efficiency of the model.

REFERENCES

- https://jfukuyama.github.io/teaching
- https://www.kaggle.com/datasets/iabhishekofficial/mobile-price-classification?select=train.csv
- https://stats.oarc.ucla.edu/r/dae/multinomial-logistic-regression/#:~:text=Multinomial%20logistic%20regression%20is%20used,combination%20of%20the%20predictor%20variables.
- https://rpubs.com/jpmurillo/153750

APPENDIX

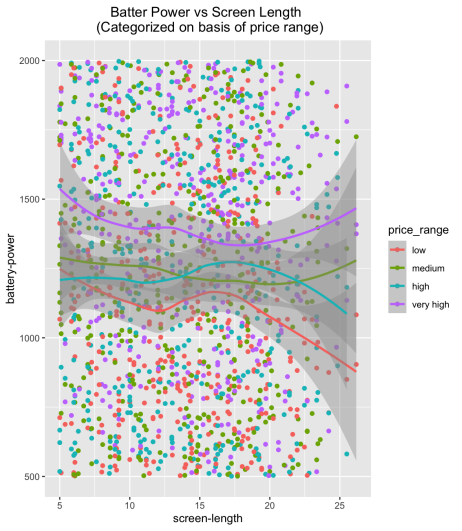

Figure 31

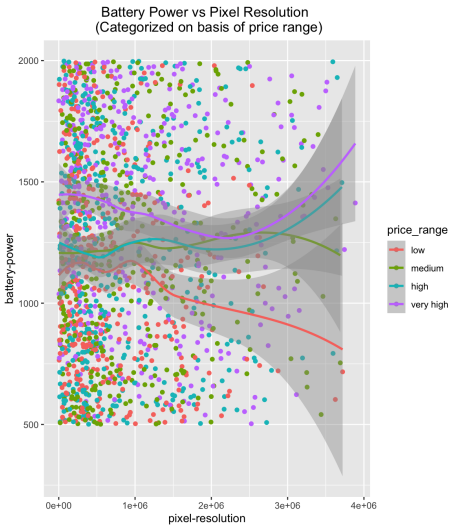

Figure 32

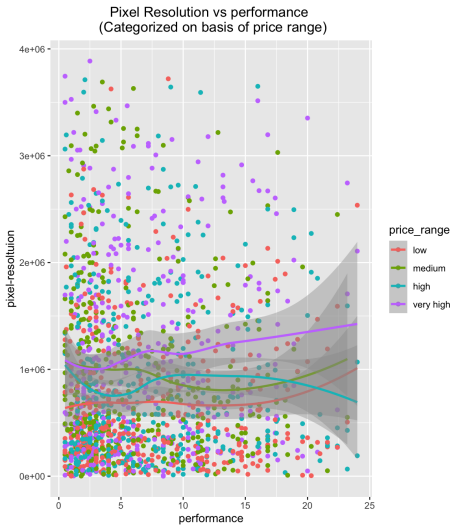

Figure 33

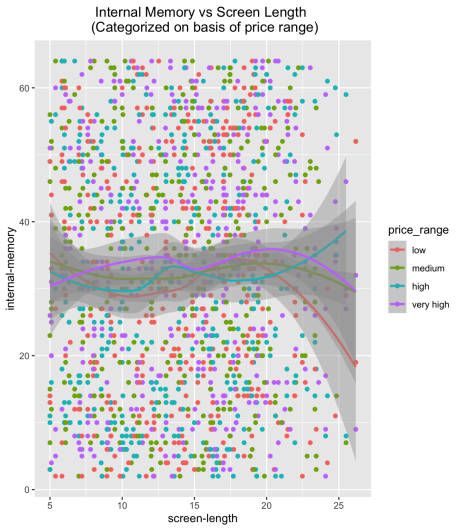

Figure 34

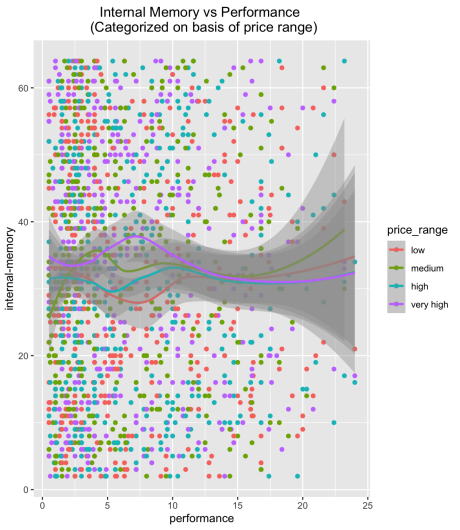

Figure 35

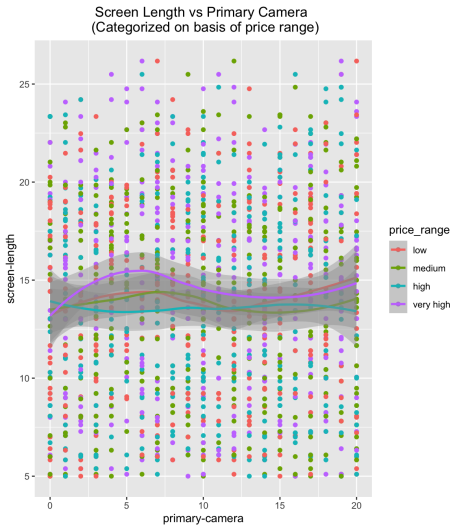

Figure 36

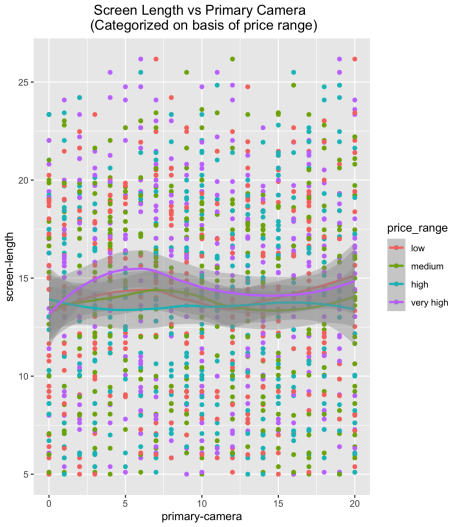

Figure 37

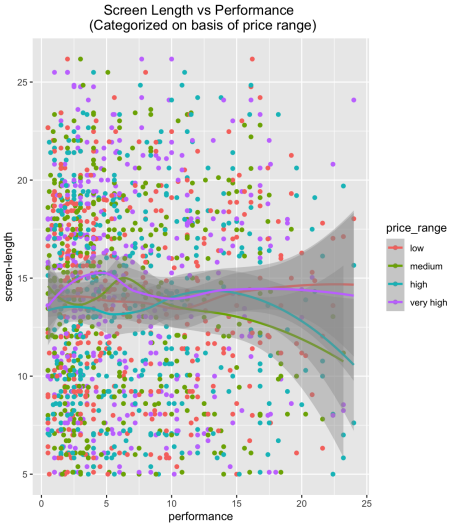

Figure 38

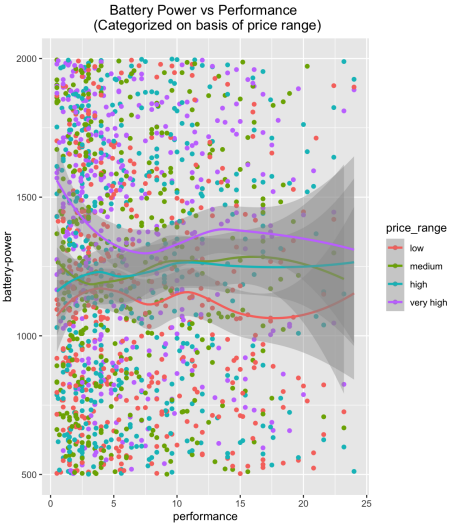

Figure 39

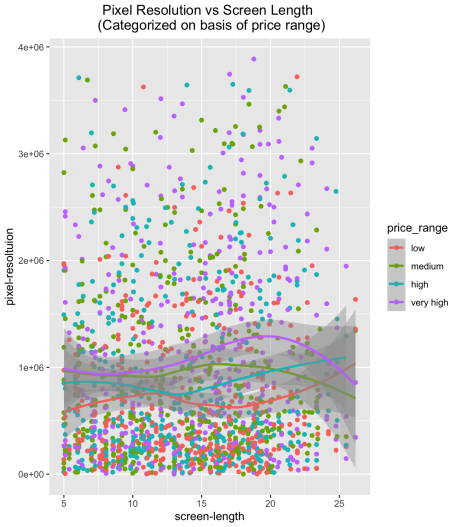

Figure 40

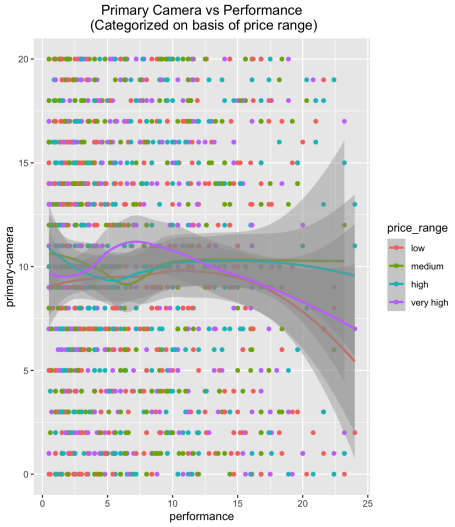

Figure 41

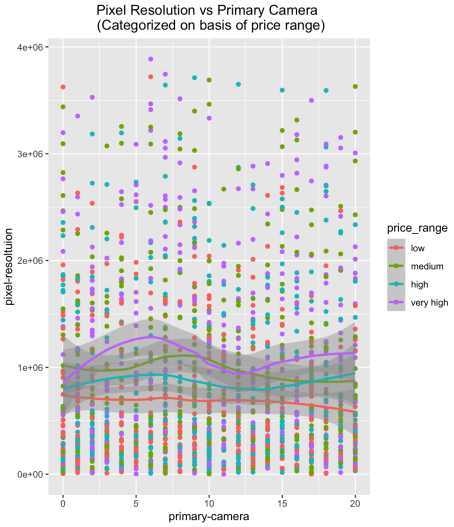

Figure 42