**Used Device resale Category Classification, Normalized Used Price Prediction and Recommendations**

Rajesh Dasari Timmisetty

**Executive Summary**

BIG\_C, an e-commerce company specializing in selling used devices with high resale value, faces bankruptcy due to three key reasons. First, it's hard to predict whether a device will have a high or low resale value, as many factors affect it. Second, BIG\_C struggles to determine the current market value of the devices, leading to overspending when buying stock. Thirdly, the limited availability of used devices frequently results in customers being unable to find their desired items, leading to loss of customers for the business. To improve profits, the company decided to use predictive and clustering analyses on historical data. The backward selection logistic regression model was selected as the best model to classify whether a used device has high resale value or not, allowing BIG\_C to purchase only devices classified as having high resale value instead of buying all used devices. The regression tree model was chosen as the best model to predict the market value of used devices, preventing the company from overinvesting in devices that would lead to decreased profits. Cluster analysis is used to group used devices with high resale value and recommend similar devices to customers if their preferred item is unavailable. Knn model is considered as the best model to assign a used device to a cluster.

**Table of Contents**

1. Introduction …………………………………………………………………………….5
   1. Business Background …………………………………………………………………..5
   2. Business Goal…………………………………………………………………………...5
   3. Analytical Approach ……………………………………………………………………5
   4. About Dataset…………………………………………………………………………...6
2. Exploring and preprocessing the data…………………………………………………..7
3. Dimension Reduction …………………………………………………………………..37
4. Partition the Data……………………………………………………………………….38
5. Modelling ………………………………………………………………………………39
   1. Decision Tree.…….……………………………………………………………………39
      1. Measuring Performance.……………………………………………………………41
      2. Improving Model Performance……………………………………………………..42
   2. Logistic Regression ……….……………………………………………………………43
      1. Measuring Performance…………………………………………………………….45
      2. Improving Model Performance …………………………………………………….45
6. Classification Model Selection………………………………………………………….47
7. Modelling For Regression Task ………………………………………………………...48
   1. Linear Regression……………….……………………………………………………...48
      1. Measuring Performance……………………………………………………………..49
      2. Improving Model Performance……………………………………………………..50
   2. Decision Tree for Regression…………………………………………………………...51
      1. Measuring Performance……………………………………………………………..52
      2. Improving Model performance……………………………………………………..52
8. Regression Model Selection…………………………………………………………….52
9. Modelling for Clustering Analysis……………………………………………………...54
   1. Evaluating Model Performance…………………………………………………………55
   2. Improving k-means Model……………………………………………………………...56
      1. Retraining k-Means with Best cluster number……………………………………...57
      2. Evaluating Model Performance……………………………………………………..58
10. Classification Modelling to classify cluster…………………………………………….59
    1. KNN Model……………………………………………………………………..59
    2. Measuring the Performance……………………………………………………..60
11. Predicting New Data……………………………………………………………………61
12. Conclusion………………………………………………………………………………62
13. **Introduction**
    1. **Business Background**

BIG\_C, is an e-commerce company, sells used devices with high resale value. It buys them from people and sells them online. However, BIG\_C is facing losses for three reasons. First, it's hard to predict whether a device will have a high or low resale value, as many factors affect it. Second, BIG\_C struggles to determine the current market value of the devices, leading to overspending when buying stock. Thirdly, the limited availability of used devices frequently results in customers being unable to find their desired items, leading to loss of customers for the business.

* 1. **Business Goal**

The main business goal of BIG\_C is to overcome its challenges by utilizing predictive analysis and cluster analysis on historical data. Predictive analysis will help classify used devices with high resale value and estimate their market value. This allows the company to focus on buying only those devices with high resale value from original customers, investing less compared to market prices, and selling them for profits. Cluster analysis will group used devices with similar features into different clusters, enabling the recommendation of alternative devices to customers when their preferred device is out of stock. This enhances customer experience and mitigates the impact of limited stock on potential customer loss.

* 1. **Analytical Approach**

The main objective is to utilize predictive analysis and cluster analysis for classifying whether a used device has a high or low resale value and grouping the similar devices, estimating the market price of devices, and recommending similar used devices to customers based on the device features. To achieve these goals, first, a classification model is built to classify if a used device has high resale value or not. Then a regression model is used to estimate the market value of the used devices that have high resale value. Finally, clustering analysis is used on the high-resale-value devices to recommend similar used devices to customers if their preferred devices are out of stock or unavailable.

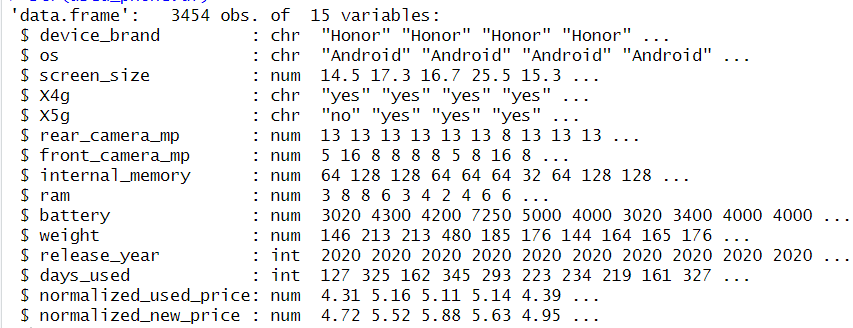
* 1. **About Data**

The data is historical data with features that may impact the resale value of a mobile. The dataset consists of 3454 rows and 15 columns. Now let’s try to understand the definition of each column.

* device\_brand : The brand of the used devices which is a categorical column.
* OS: The operating systems used by the mobile device which is a categorical column.
* screen\_size : The screen size of the device in centimeters which is a numerical column.
* 4g : A categorical column which represents weather the mobile device supports 4g or not.
* 5g: A categorical column which represents weather the mobile device supports 5g or not.
* rear\_camera\_mp : Which represents the back camera megapixel count for device. providing information about the camera's capability to capture detailed and high-resolution images.
* front\_camera\_mp : Which represents the front camera megapixel count for device. providing information about the camera's capability to capture detailed and high-resolution images.
* Internal\_memory : It represents the amount of digital storage capacity the device has for storing data, files, and applications. It is given in gigabytes.
* Ram: RAM, or Random Access Memory, is a type of volatile memory used in mobile phones and tablets to store data that is actively being used or processed by the device. This Column is a numeric column which represents the amount of the ram in the device which is measured in gigabytes.
* Battery : Which represents the battery capacity of the device in mah.
* Weight : Which represents the weight of the device in grams.
* release\_year : Year the device released.
* days\_used : Number of days the original user used the device before selling it to BIG\_C company.
* normalized\_used\_price : The resale price of the device which is in normalized format to bring all different used prices under one scale.
* normalized\_new\_price : The price of the new device which is normalized to bring all different prices under one scale.

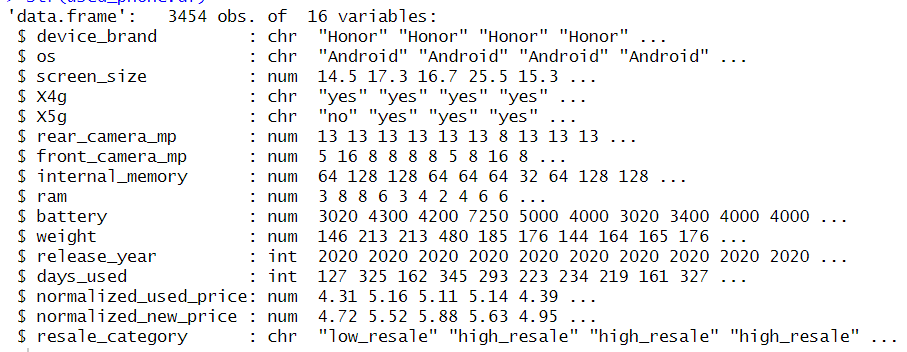
1. **Exploring and Pre-Processing Data.**

In this section the data is explored and preprocessed to make the data ready for building the models. First will check the structure of the data.



**Figure 2.1 structure of the data**

From figure 2.1, it becomes evident that there is no column that indicate the device is having high resale value or low resale value. So, a new column, resale\_category, will be created, containing two distinct values high\_resale and low\_resale. Assuming the available data represents the population data, it is considered that a normalized\_used\_price above the mean indicates a high\_resale value, while a value below the mean is categorized as a low\_resale value.



**Figure 2.2 Structure of the data after adding new column**

**Handling Categorical Variables**

Based on the structure analysis of the data shown in figure 2.2. Converting all the categorical columns into factors is appropriate to handle them correctly by the models. While using the logistic regression creating dummy variables can be handled automatically but while using knn, neural network dummy variables need to be created manually.

**Missing Values**

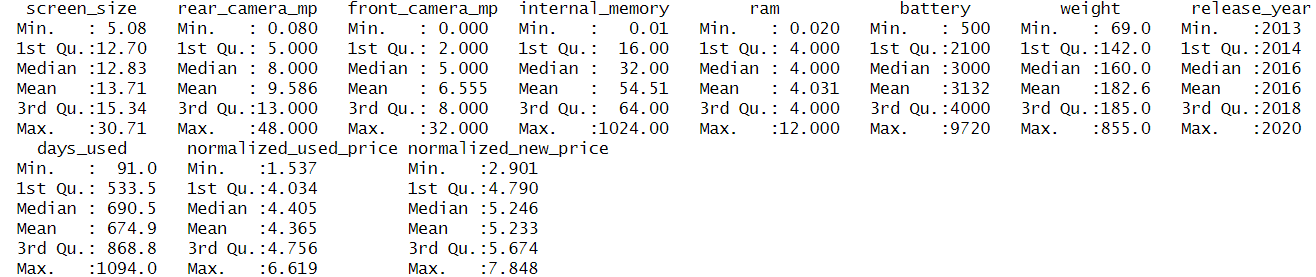
Using colsums(is.na()) in R helps identify null values in each column. There are 179 null values in rear\_camera\_mp, 2 in front\_camera\_mp,and 4, 4, 6, 7 in internal\_memory,ram,battery, and weight columns, respectively. There are various methods for replacing null values, the most common ways to replace them with the mean, median, or to simply delete them. However, these traditional approaches are generalized, and complex patterns in the data are not considered while replacing the values, which will potentially impact model performance. To address this issue, the k-nearest neighbor imputation method is used as it is known fact that all mobiles specification will be same for similar mobiles and k-nearest neighbor imputation method tries to replace with the values that are closest to the null values.

**Zero Values**

ColSums(data==0,na.rm = TRUE) in R can be used to get how many zero values in each column. There are 39 zero values in the column front\_camera\_mp. Based on the domain knowledge there is chance that devices with no front camera exists so these zero values can be used as it is.

Now let’s analyze the data of each feature and how it is impacting the target column resale\_category, normalized\_used\_price

**Summary Statistics**

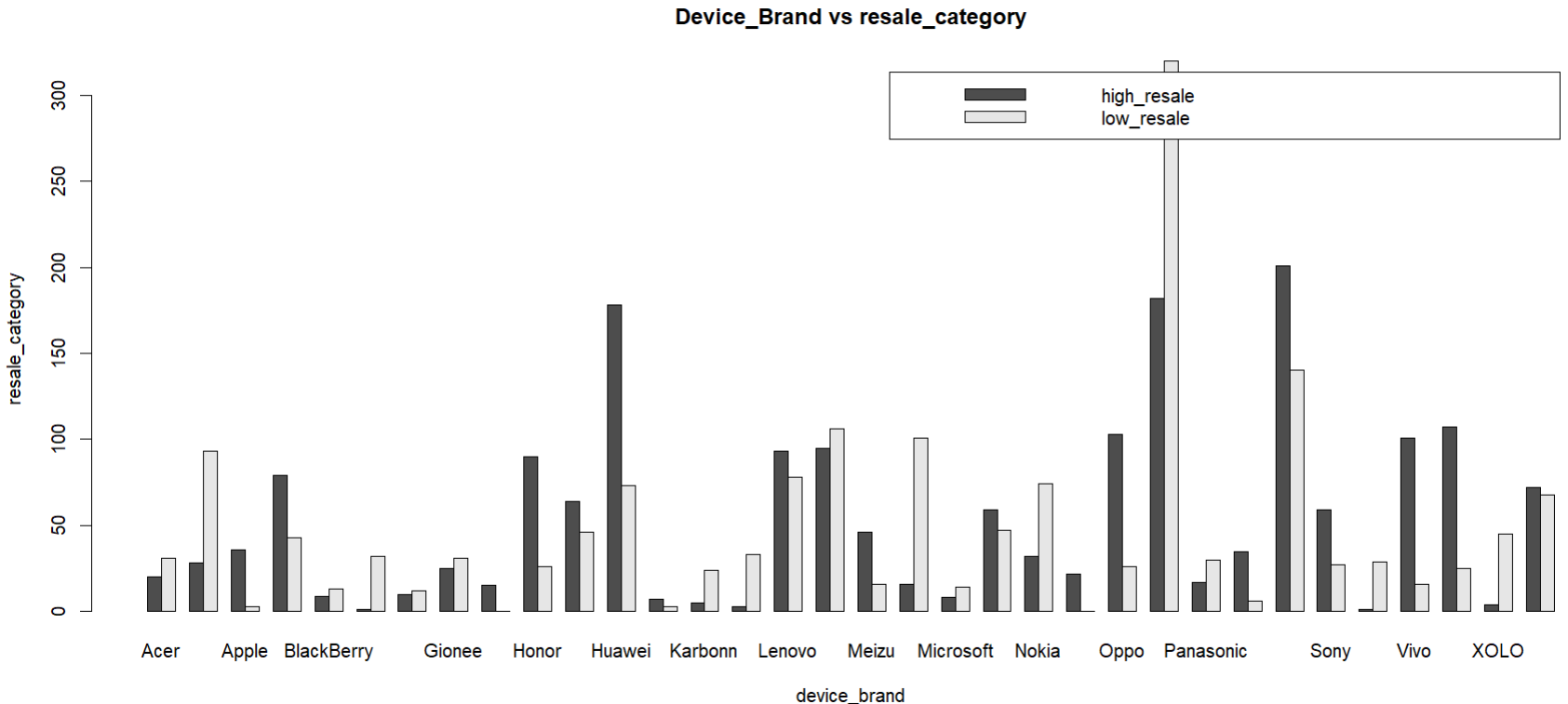
****

**Figure 2.3 Summary statistics for numerical columns**

From figure 2.3 it is clear that the different variables have very different range of values. All the null values are replaced and there are no extreme values in the data.

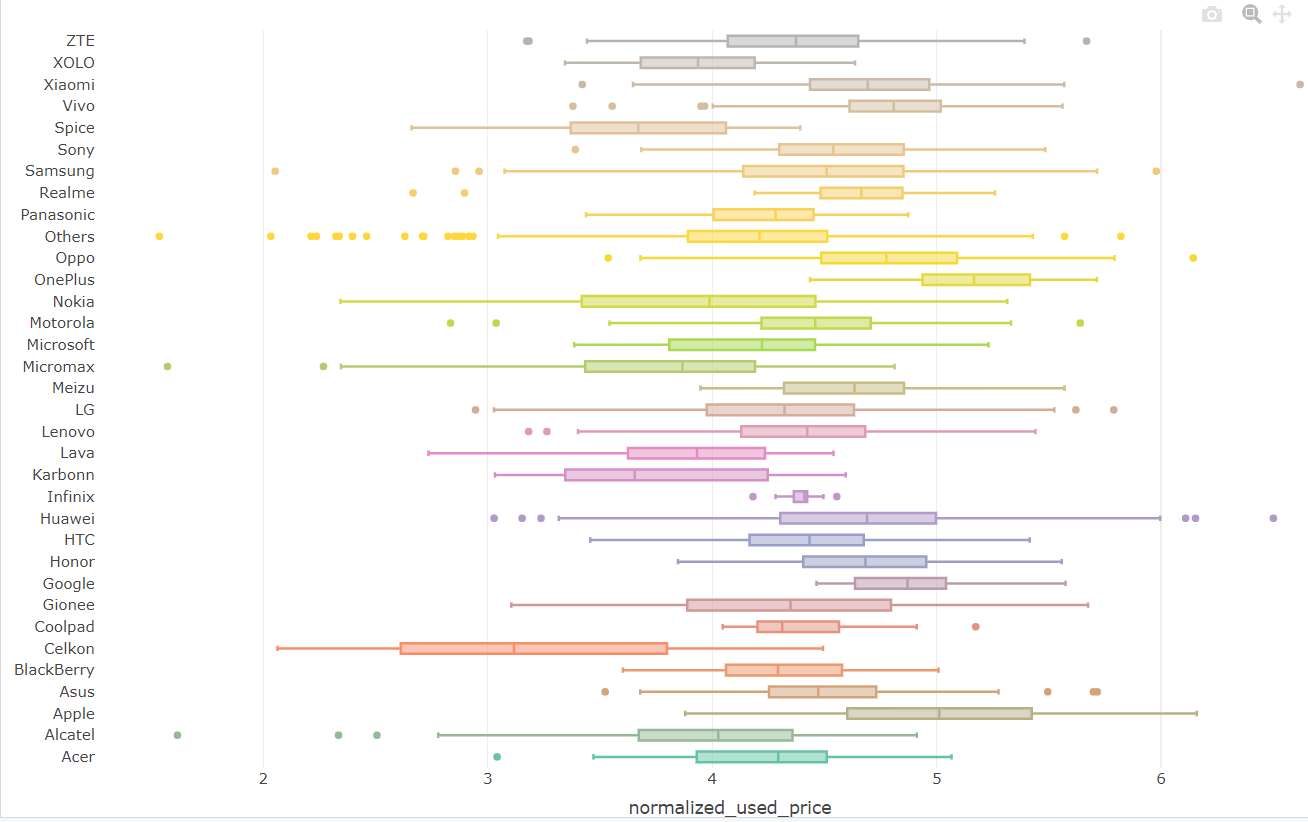
Now let’s analyze each column how it is impacting the target columns.

**Device\_Brand Column**

****

**Figure 2.4 device\_brand vs resale\_category**

Most of the Panasonic mobiles have low resale value and most of the hauwei branded devices and apple branded device are having high\_resale value. So, to gain more profits BIG\_C company needs to focus on selling hauwei and apple brand devices.

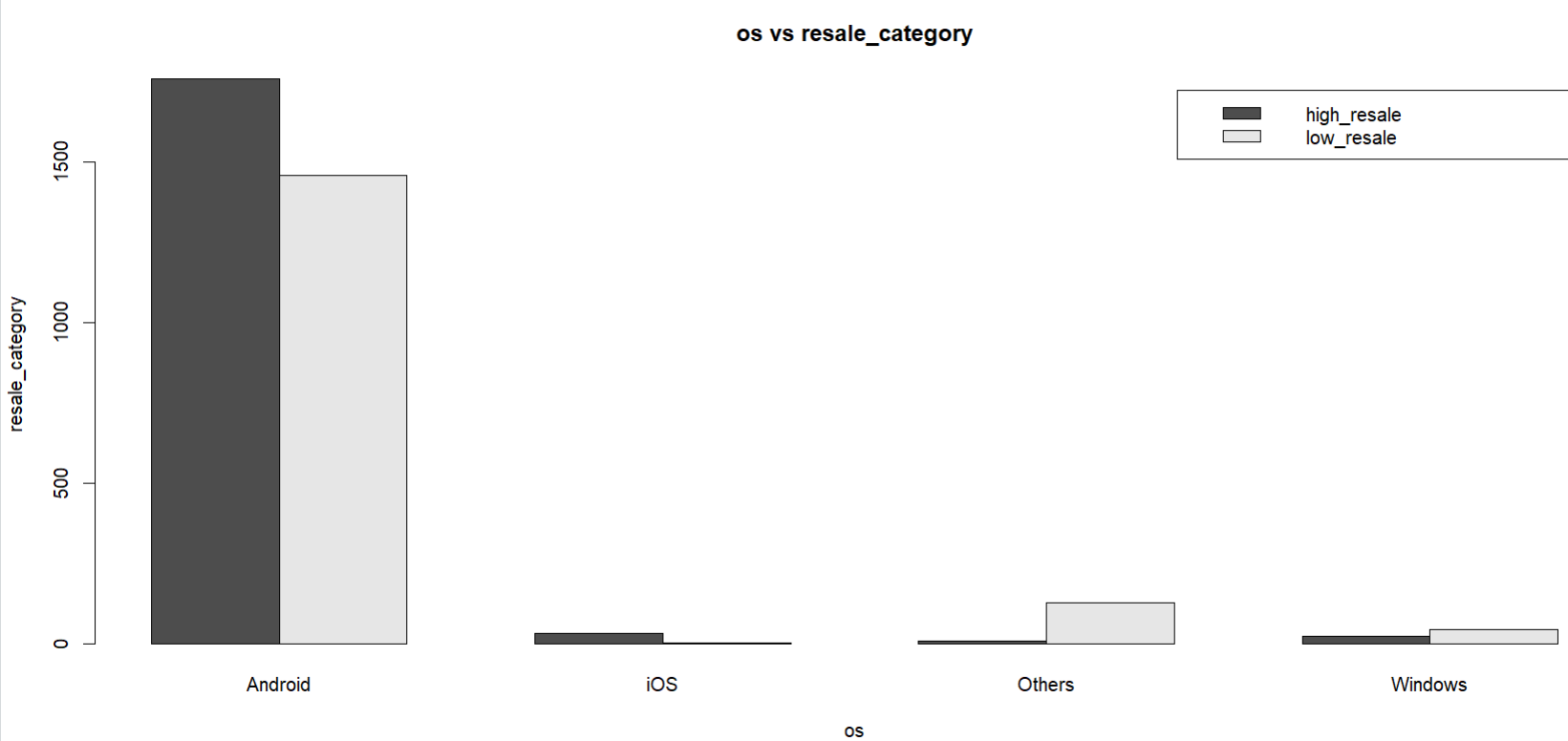


**Figure 2.5 device\_brand vs normalized\_used\_price**

From the figure 2.5, it is evident that nearly all device brands, except Celkon, Alcatel, and Micromax, have a normalized used price greater than 3. Therefore, BIG\_C company can anticipate obtaining a normalized resale price above 3 for all mobile brands, excluding Celkon, Alcatel, and Micromax.

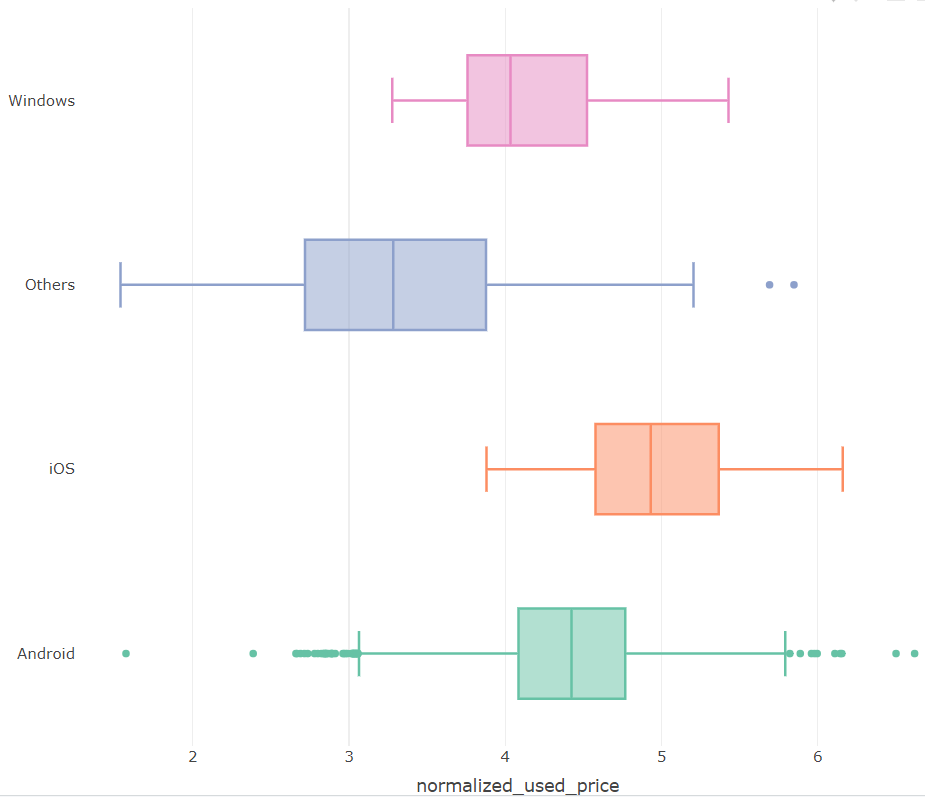
**Os column**

OS is categorical column which has mainly 4 distinct values android, iOS, others and windows. Now let’s try to understand how each os devices impacting the resale\_category and normalized\_used\_price

****

**Figure 2.6 OS vs resale\_category**

Based on Figure 2.6, it is evident that there are more Android type OS devices in the dataset compared to others. The majority of windows and others OS type devices have a low resale value, while most iOS, Android os type devices have a high resale value. So, the BIG\_C company need to focus on reselling IOS and Android os devices to gain more profit.

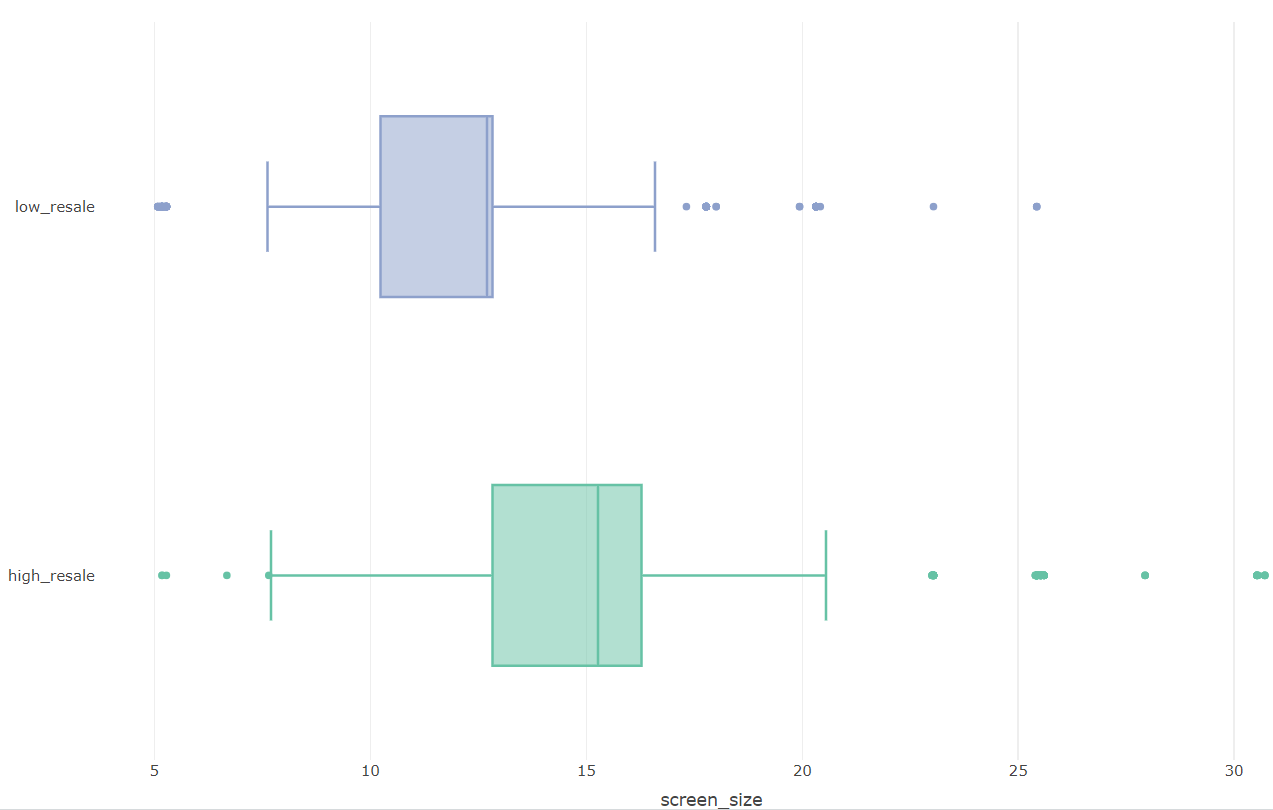


**Figure 2.7 OS vs normalized\_used\_price**

From figure 2.7 all of the IOS devices are having normalized resale price between 3.8 to 6.15. For windows it is between 3.2 to 5.4. For other os type it is in between 1.5 to 5.8 and finally android os type devices is in between 1.5 to 6.6. Most of the IOS devices have the resale price between 4.5 to 5.3 which is higher when compared to all other os type devices. So, the conclusion would be IOS type devices will be more profitable to the BIC\_C company.

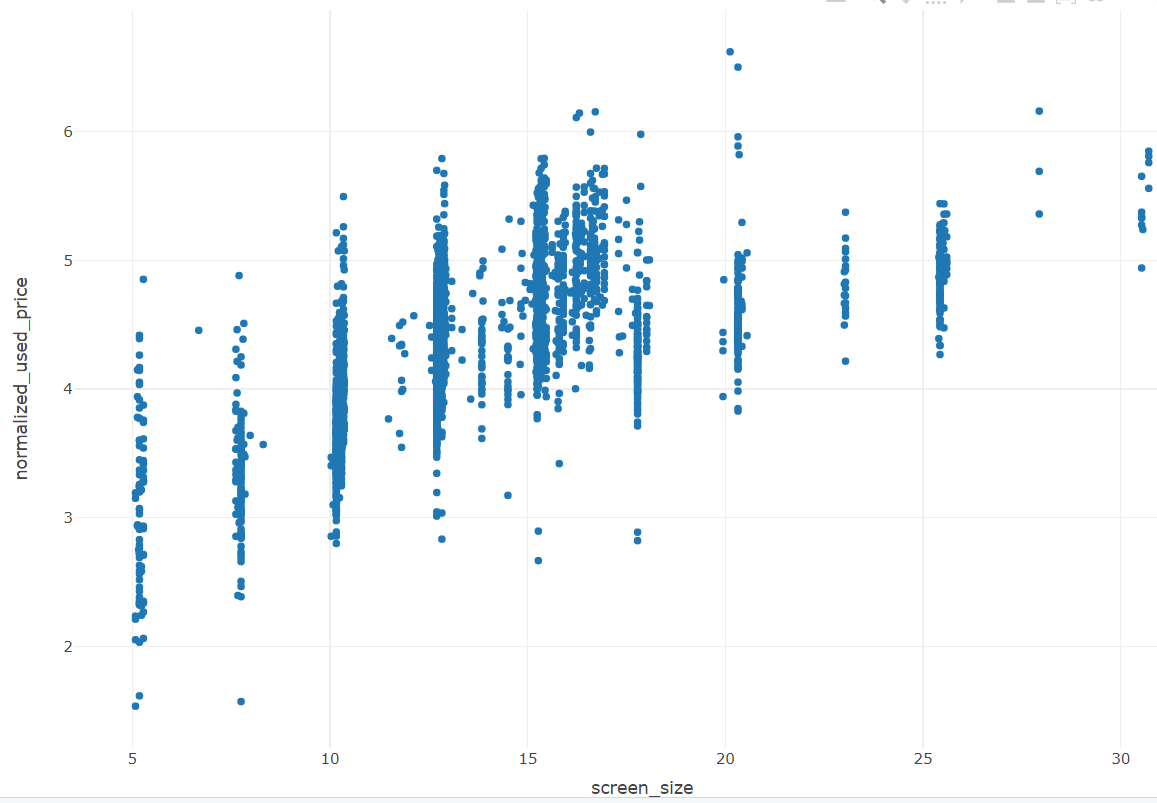
**Screen\_size Column**

Screen\_size is a continuous column which represents the screen size of the device and now lets try to understand how screen size is impacting the resale\_category and normalized\_used\_price



**Figure 2.8 screen\_size vs resale\_category**

From figure 2.8 it is clear that device with a screen size between 12.83 to 16.28 generally have a high resale value, while those with a screen size between 7.6 to 16.59 tend to have a low resale value. Additionally, all devices with a screen size greater than 27 exhibit a high resale value. In conclusion, devices with larger screen sizes have higher resale value. Therefore, BIC\_C company may consider prioritizing the sale of devices with higher screen sizes over those with smaller screen sizes to gain more profits.

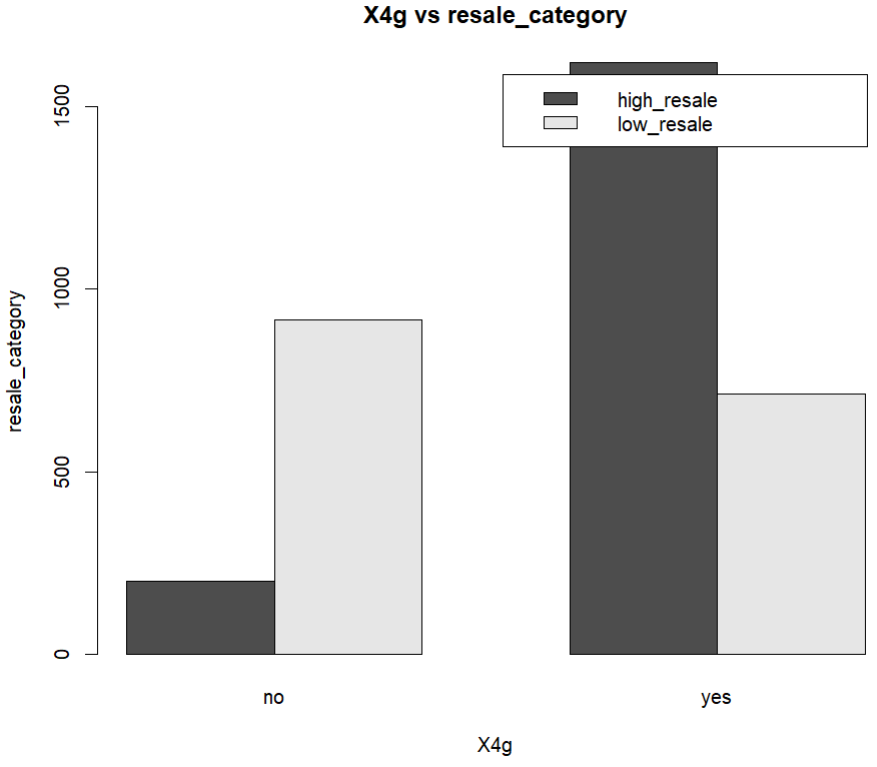


**Figure 2.9 screen\_size vs normalized\_used\_price**

From the figure 2.9, it is evident that there is a positive correlation between screen size and normalized used price. Generally, as the screen size of devices increases, the normalized used price also tends to increase. However, there are instances where the relationship is reversed. This suggests that there may be other factors influencing the normalized used price of the device.

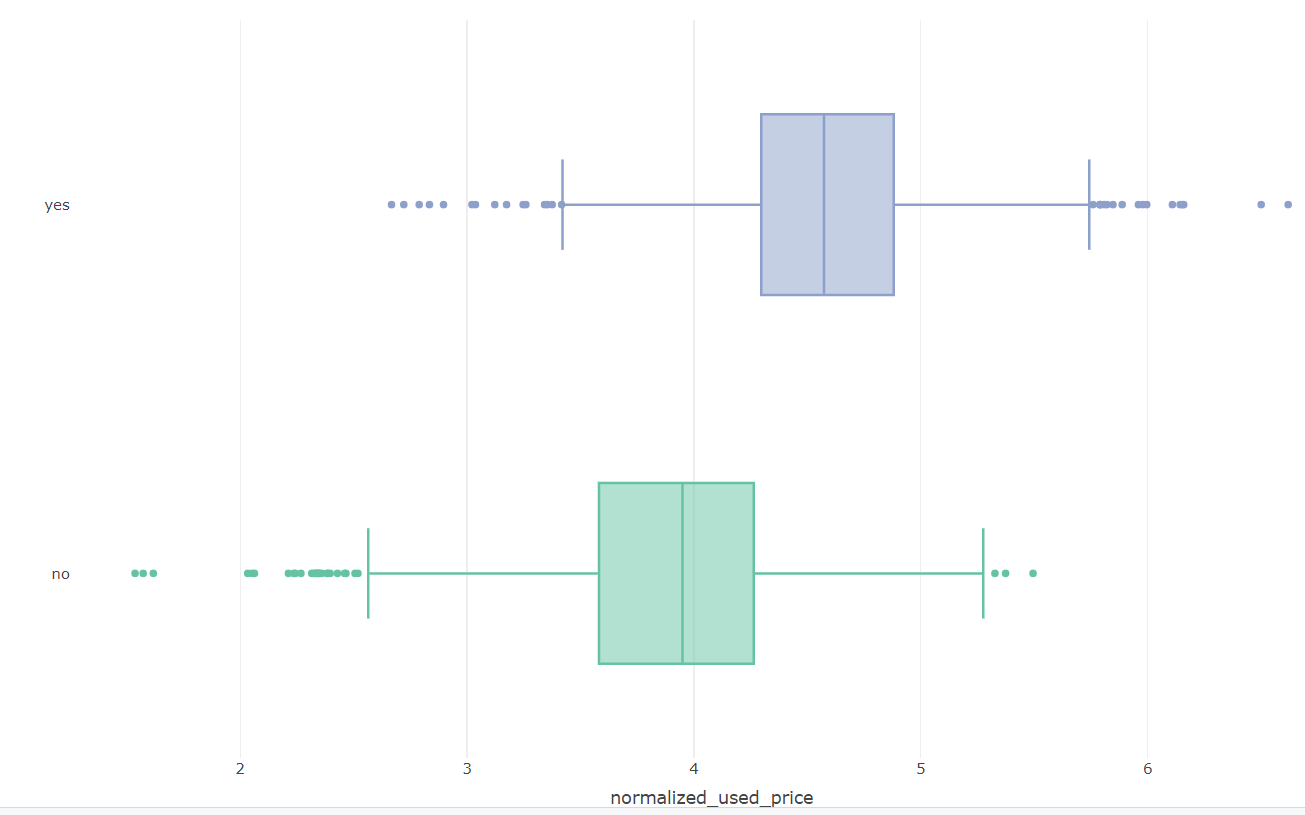
**X4g Column**

X4g column is categorical which tells whether the device supports 4g network or not. Let’s understand how X4g column impacting resale\_category.

****

**Figure 2.10 X4g vs resale\_category**

From figure 2.10 it is clear that devices which support 4g network has high\_resale value when compared to devices that does not support 4g network. So, the BIC\_C company need to focus on devices that support 4g.

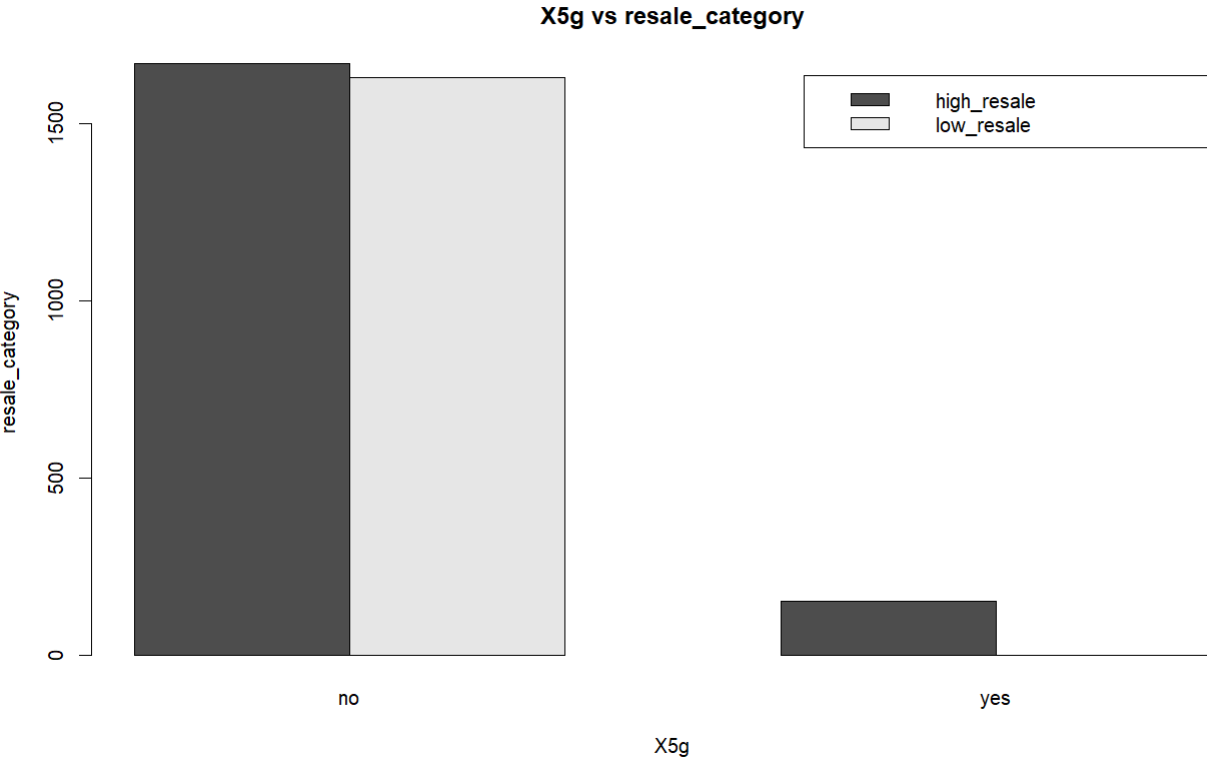


**Figure 2.11 X4g vs normalized\_used\_price**

From the figure 2.11 it is clear that the minimum and maximum normalized used price for devices with no 4g network support is in between 1.5 to 5.4 and for the devices with 4g network support is in between 2.6 to 6.6. most of the devices which are supporting the 4g network has higher normalized used price. So if BIG\_C company sells the devices which has 4g network support will have more profits.

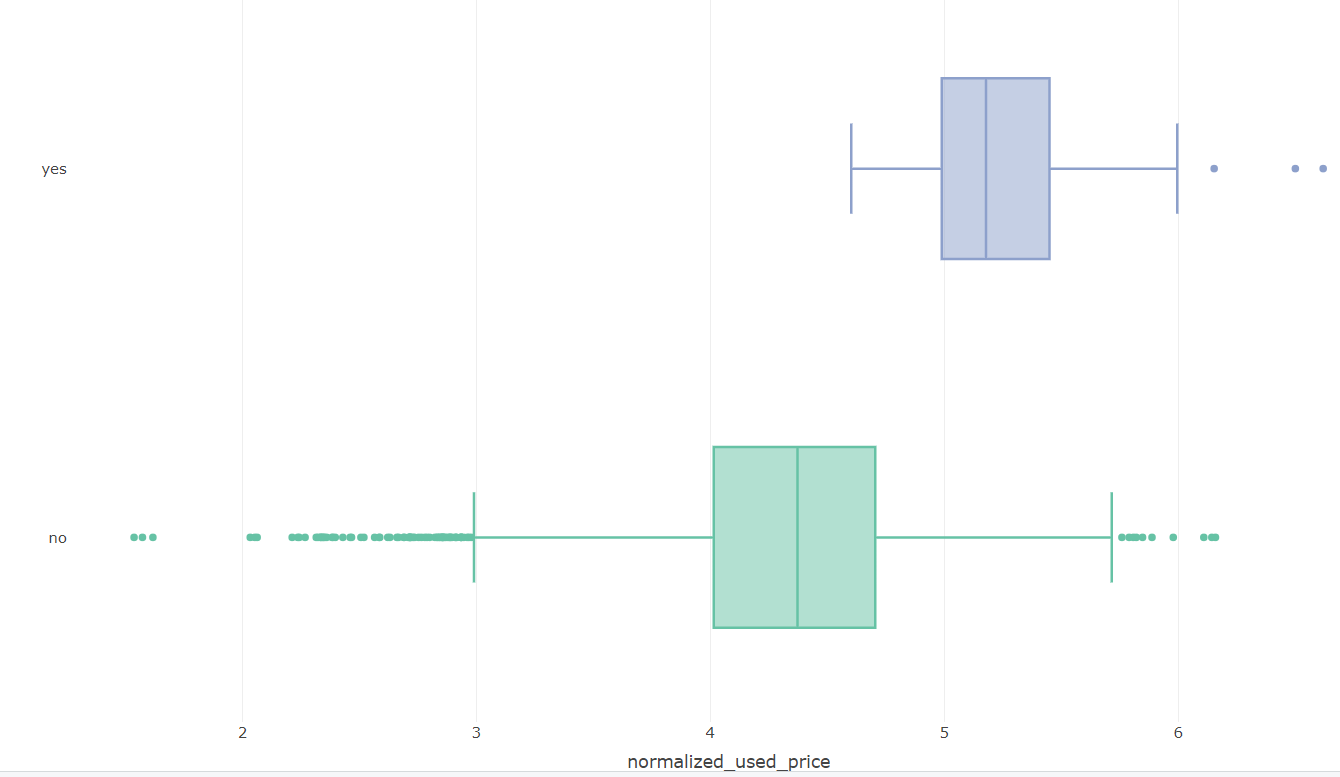
**X5g column**

X5g column is a categorical column which tells whether the device will support the 5g network or not.

****

**Figure 2.12 X5g vs resale\_category**

From figure 2.12, it is evident that nearly all devices supporting 5G networks exhibit a high resale value. Conversely, devices lacking 5G support show a relatively equal distribution between high and low resale value. So BIG\_C company can get more profits by selling the device that supports 5g network support.

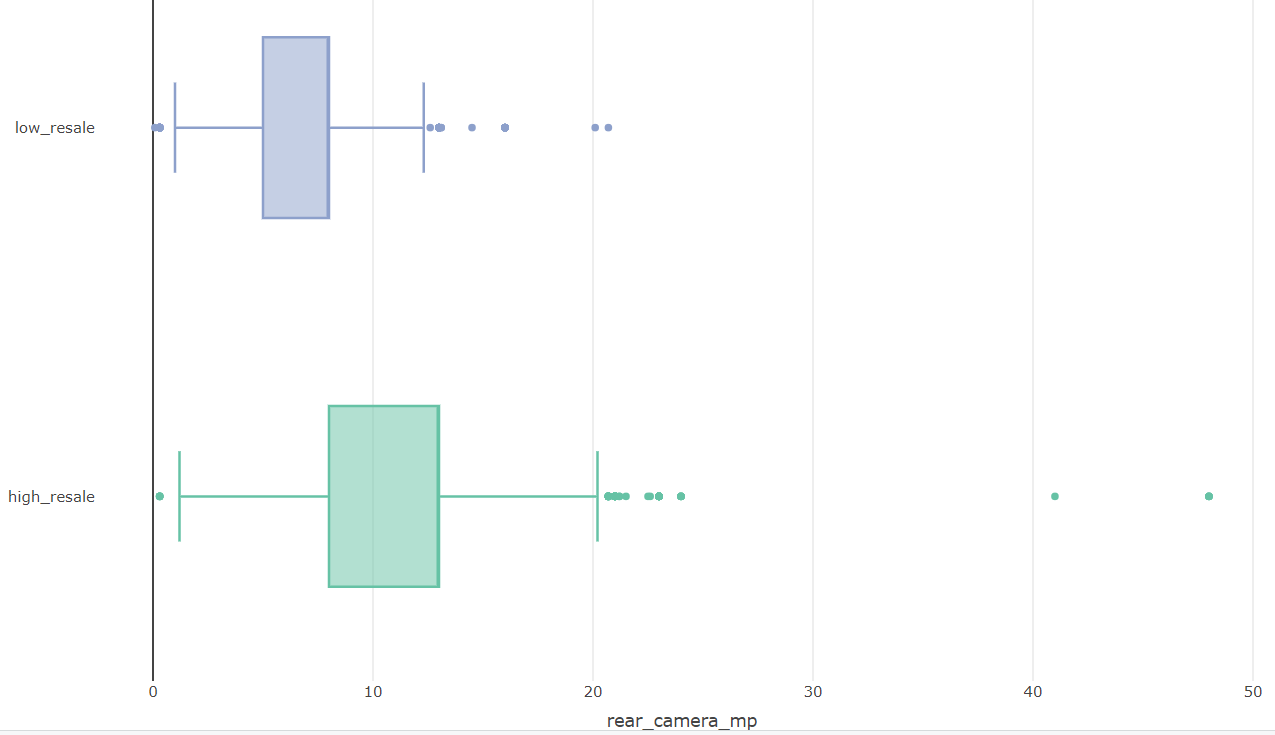


**Figure 2.13 X5g vs normalized\_used\_price**

From the figure 2.13 it is evident the minimum and maximum normalized used price for the devices that support 5g network is in between 4.6 to 6.6 and devices that does not support the 5g network is in between 1.5 to 6.1. So, all devices which support 5g will have higher normalized used price. So BIG\_C company will get more profits for reselling the 5g network supported devices.

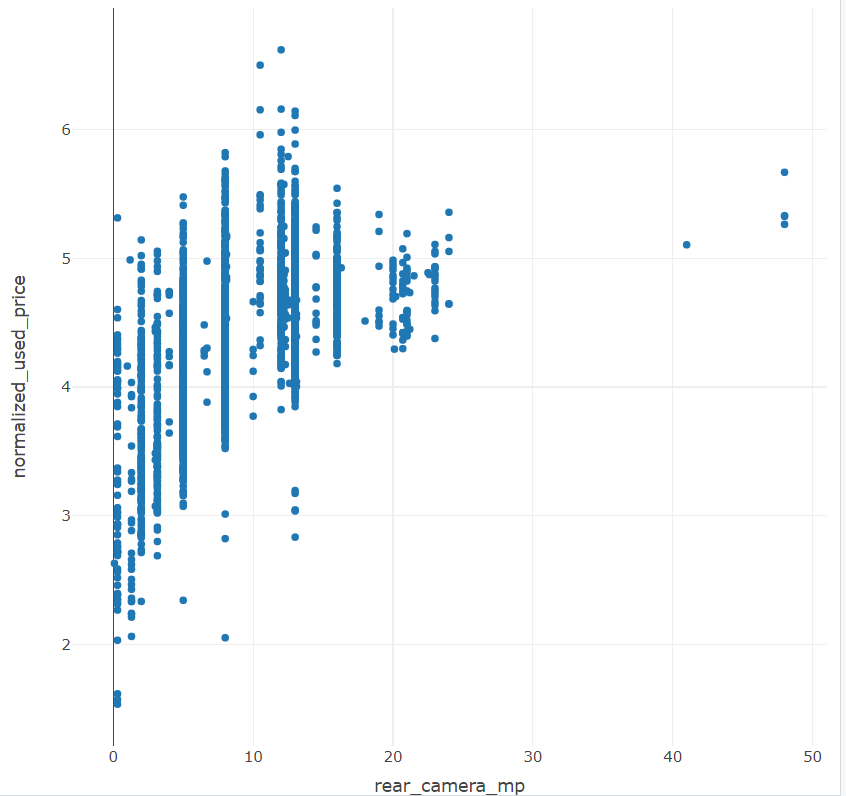
**rear\_camera\_mp Column**

rear\_camera\_mp column is numerical column which gives the back camera pixel information in a device. Let’s explore how the rear camera pixel impacts the resale\_category and normalized\_used\_price

****

**Figure 2.14 rear\_cemera\_mp vs resale\_category**

From figure 2.14 most of the devices that have rear camera pixels between 8 to 13 have high resale value and all devices that has rear pixel greater than 20.7 has high resale value. Most of the devices that have the pixel value between 5 to 8 has low resale value. BIG\_C company needs to prefer reselling the devices that higher pixel rear cameras to gain profits.

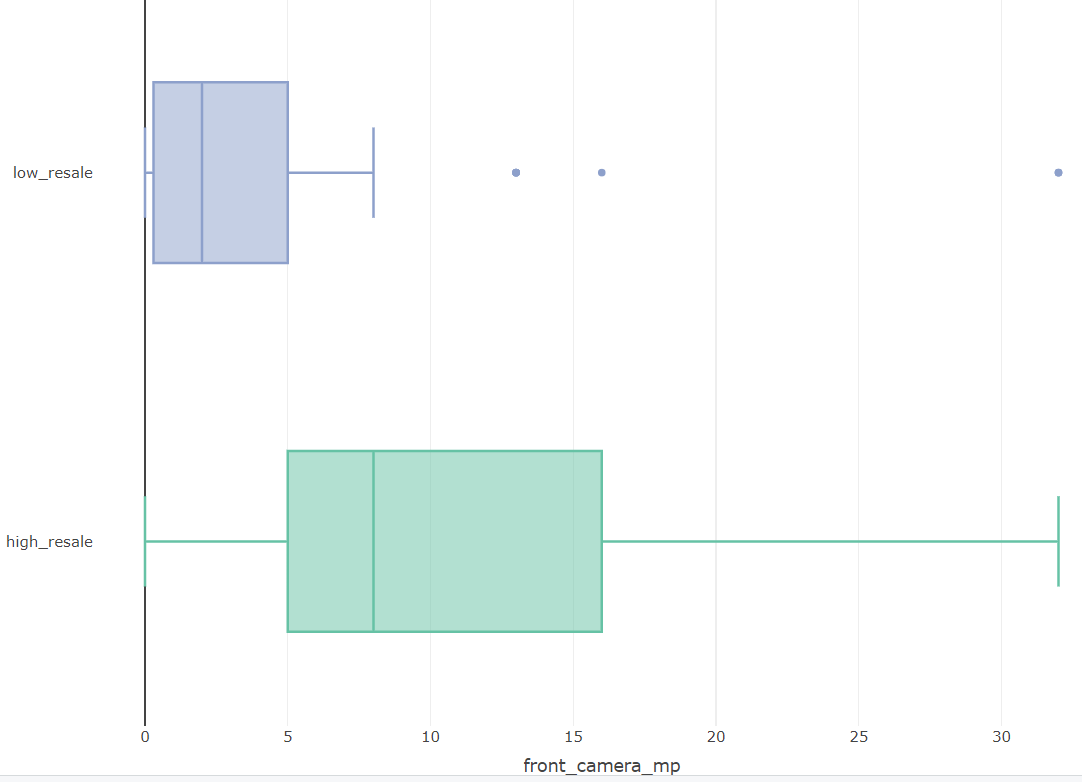


**Figure 2.15 rear\_cemera\_mp vs normalized\_used\_price**

From the figure 2.15 it is clear that rear\_cemera\_mp and normalized\_used\_price has positive correlation so if the rear camera pixels increase and the normalized used price also increases. If the BIG\_C company resale the devices with rear camera pixels greater than 16 the normalized used price will be greater than 4.18.

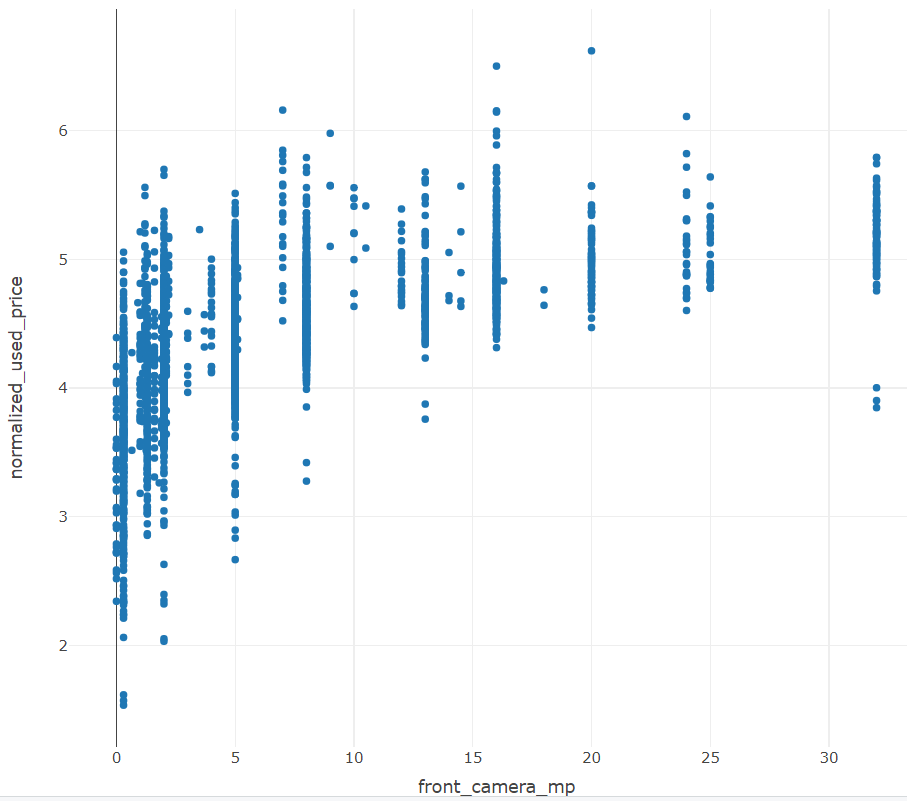
**front\_camera\_mp Column**

front\_camera\_mp column is numerical column which gives the front camera pixel information in a device. Let’s explore how the front camera pixel impacts the resale\_category and normalized\_used\_price

****

**Figure 2.16 front\_camera\_mp vs resale\_category**

From the figure 2.16 most of the devices that have the front camera pixels above 5 have high resale value. So BIG\_C company needs to prefer reselling the devices that has front camera more than 5 pixels to gain more profits.

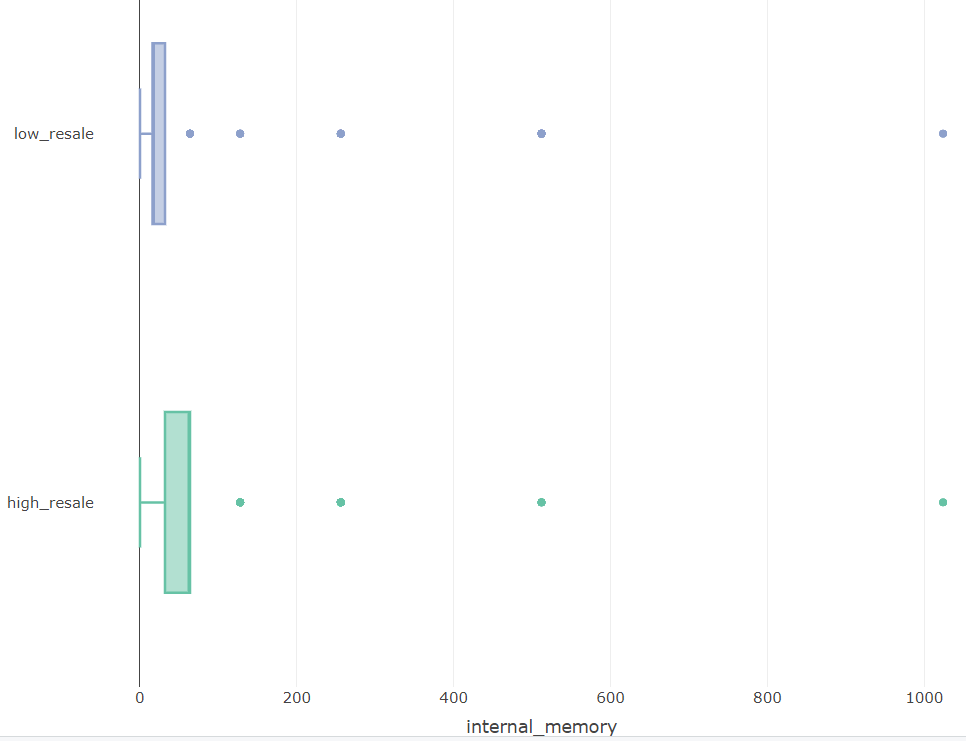


**Figure 2.17 front\_camera\_mp vs normalized\_used\_price**

From the figure 2.17, a clear positive correlation is observed between front camera megapixels and normalized used price. This suggests that as the front camera pixel count increases, the normalized used price also tends to rise. Notably, for most devices with front cameras featuring more than 8 megapixels, Company BIG\_C can expect to achieve a resale normalized price above 4.

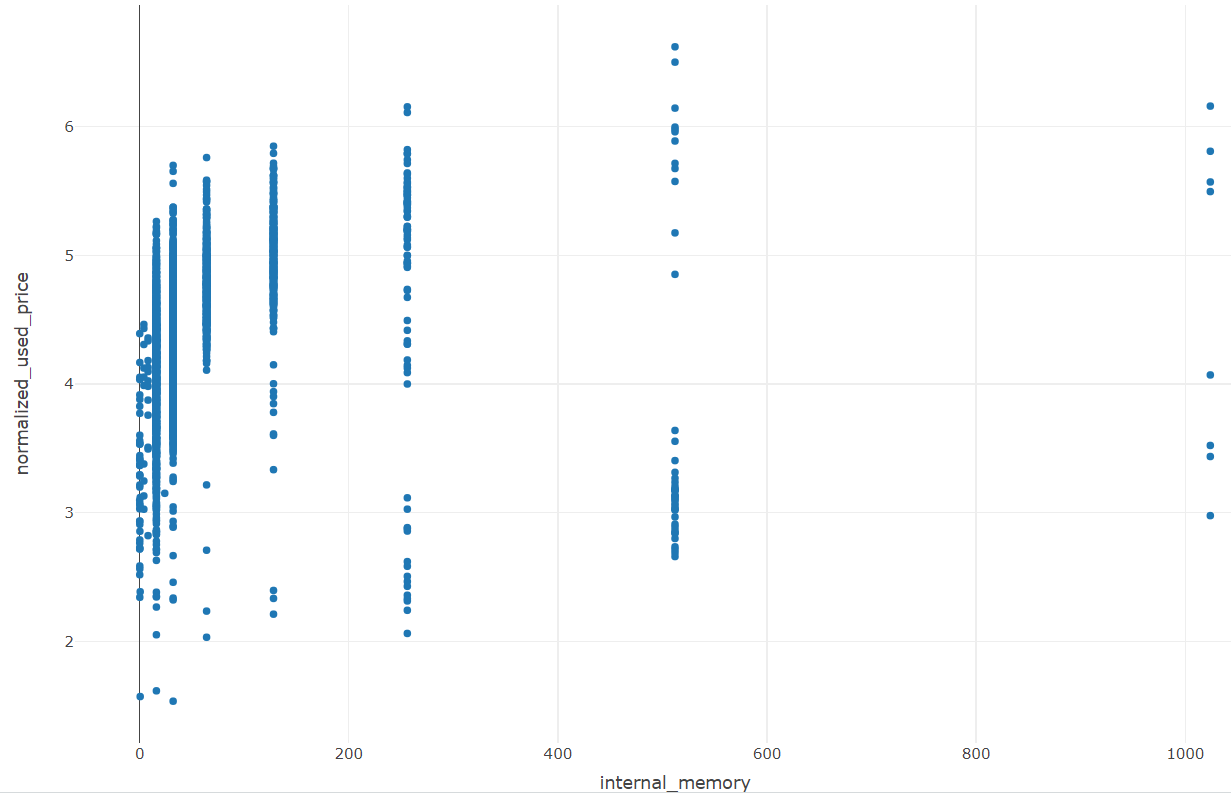
**internal\_memory**

internal\_memory is a numerical column which indicates the internal memory of the devices. Now let’s try to explore how the internal\_memory of a device is impacting its resale\_category and normalized\_used\_price.

****

**Figure 2.18 internal\_memory vs resale\_category**

From figure 2.18 it is clear that most of the devices that has the internal memory which is greater than 32 gigabytes have high resale value. So BIG\_C company can gain more profits by reselling the devices with the internal memory greater than 32 compared to those with less internal memory (below 32).

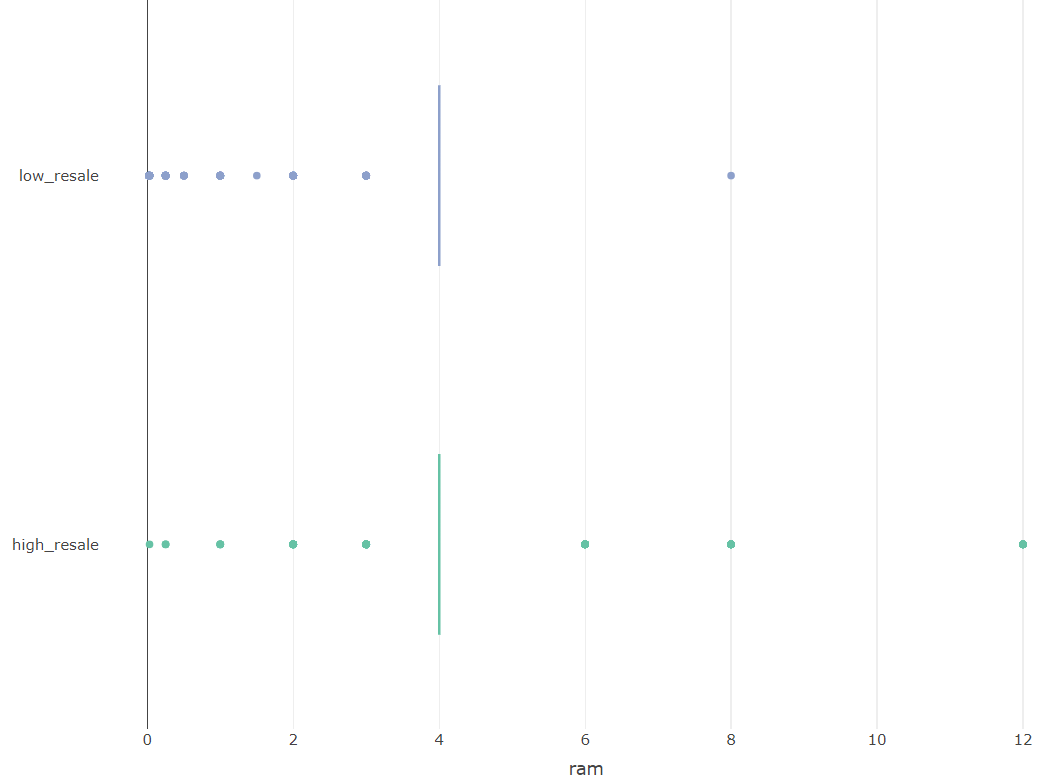


**Figure 2.19 internal\_memory vs resale\_category**

From the figure 2.19, it is clear that internal\_memory and normalized\_used\_price has no correlation and by reselling the devices with internal memory 0 gigabytes BIG\_C company can get normalized\_used\_price below 4.6.

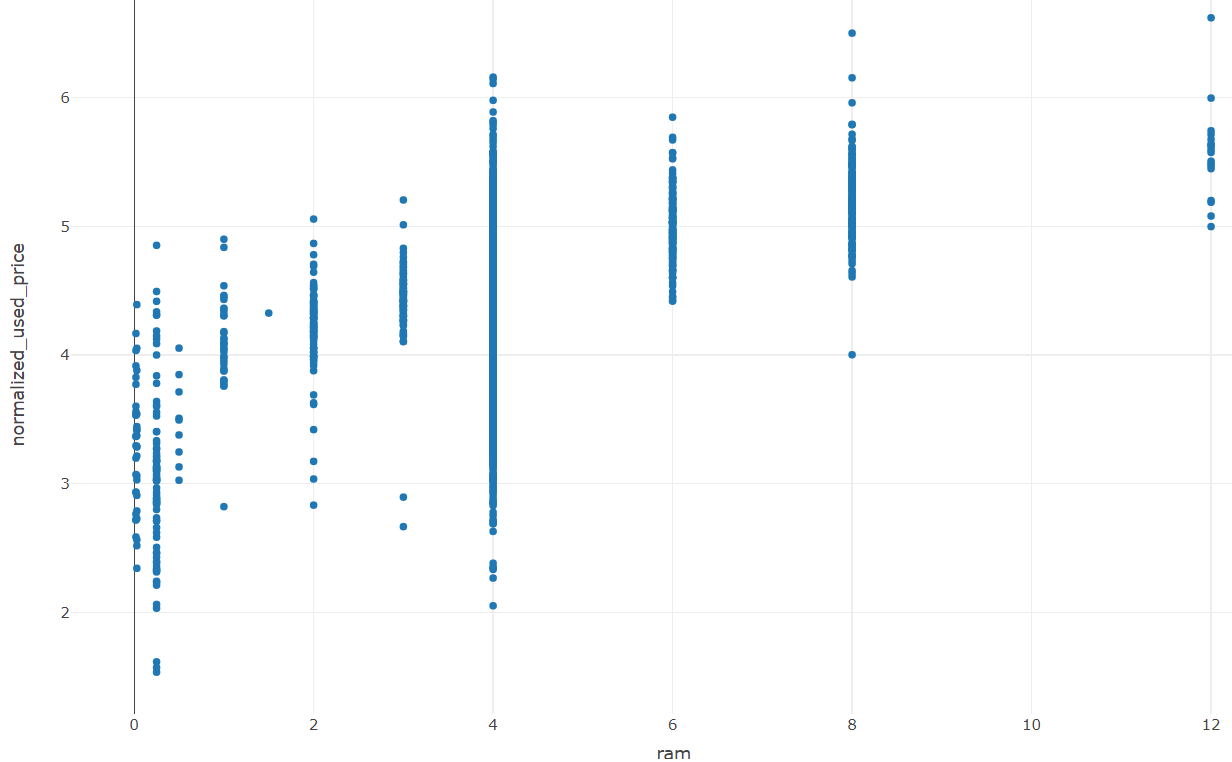
**Ram Column**

Ram column indicates about the ram of the devices. Let’s explore how ram of a device is impacting the resale\_category and normalized\_used\_price.

****

**Figure 2.20 ram vs resale\_category**

From figure 2.20 it is clear that all mobile devices with ram greater than 10 has high resale value category. So BIG\_C company will get more profits if it resells the devices with ram greater than 10.

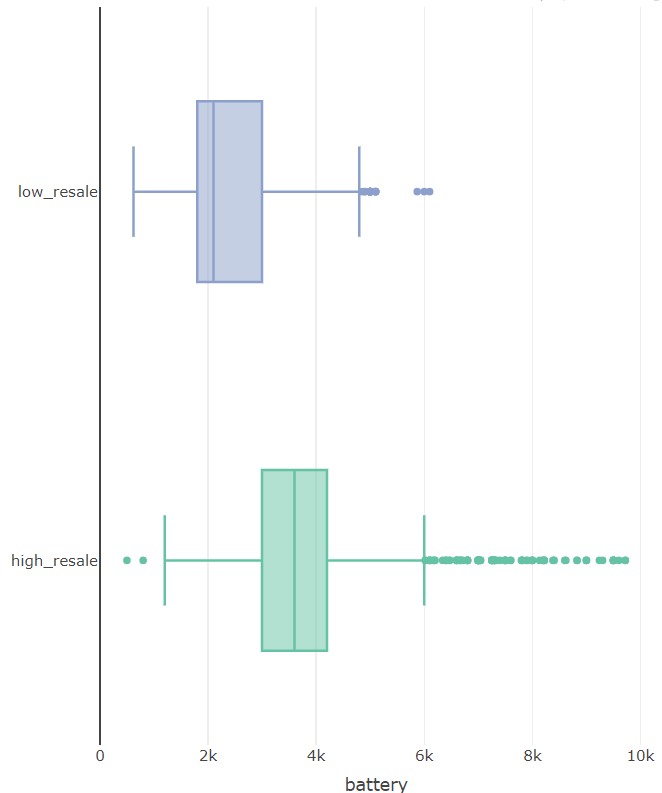


**Figure 2.21 ram vs resale\_category**

From the figure 2.21 it is clear that there is positive correlation between ram and normalized\_used\_price this indicates that if ram increases for a device than normalized\_used\_price will also increases.Almost all the devices with more than ram 6 gigabytes has the normalized\_used\_price greater than 4. Focusing on reselling devices with RAM greater than 6 could enable BIG\_C company to achieve a normalized price exceeding 4.

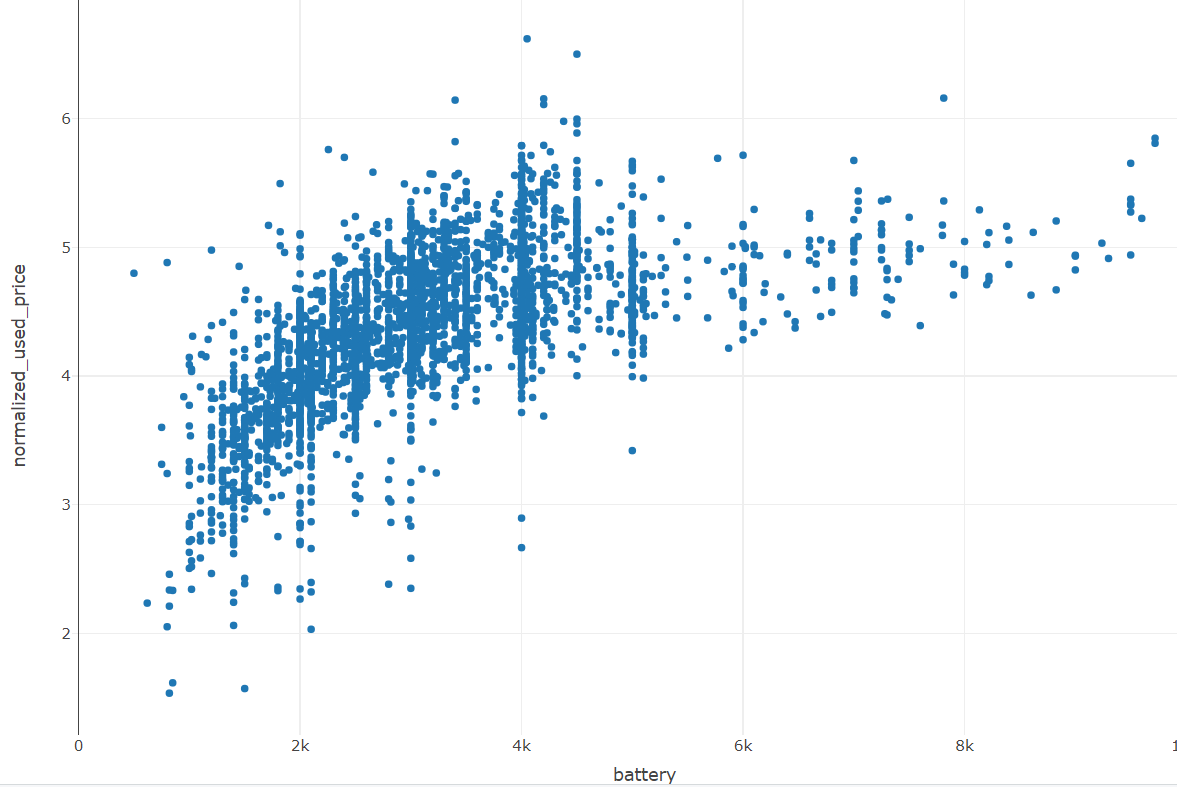
**Battery column**

Battery column which indicates the battery capacity of the devices. Now lets explore how battery impacting the resale\_category and normalized\_used\_price

****

**Figure 2.22 battery vs resale\_category**

From figure 2.22 most of the devices with battery capacity more than 3000mph will have high resale value. So, to gain profits BIG\_C company needs to focus on reselling the mobiles with battery capacity more than 3000mph.

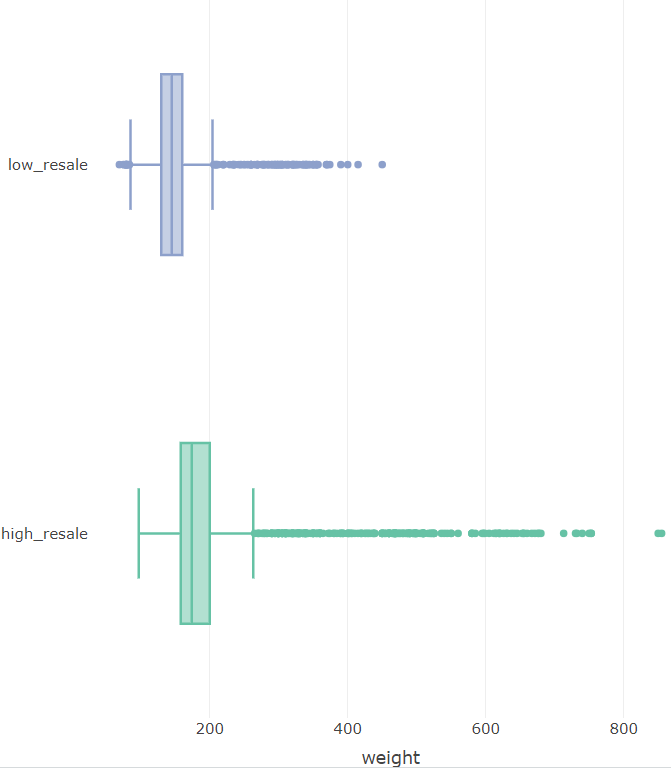


**Figure 2.23 battery vs resale\_category**

From the figure 2.23, it is clear that battery and normalized\_used\_price have positive correlation. Most of the mobiles with battery capacity more than 4000mph have the normalized\_used\_price greater than 4. So, if the BIG\_C company focus on reselling the devices with battery capacity more than 4000mph can get normalized resale price more 4.

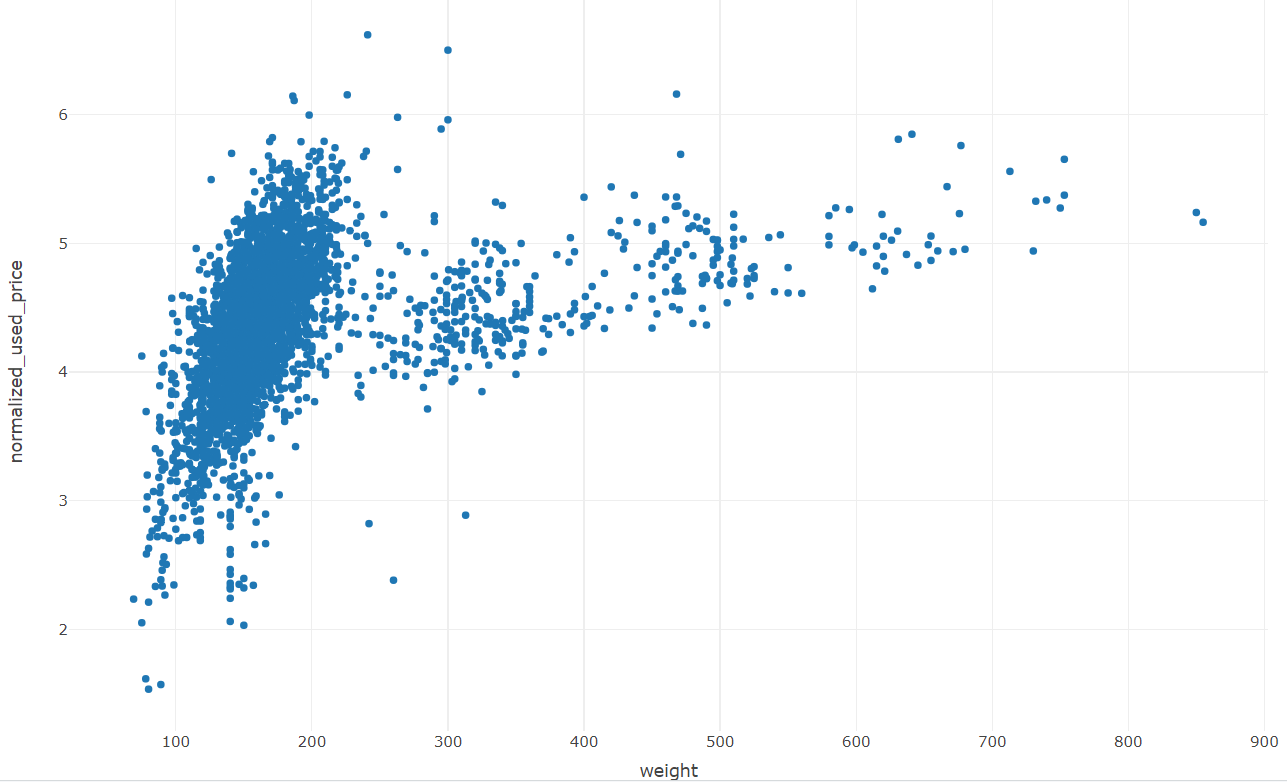
**Weight Column**

Wieght column represents the weight of the device in grams. Let’s explore how weight column impact the resale\_category.

****

**Figure 2.23 weight vs resale\_category**

From figure 2.23 it is clear that most of the devices with weight more than 180 grams has high resale value. So, the BIG\_C company can gain more profit if they focus on selling the devices with more weight.

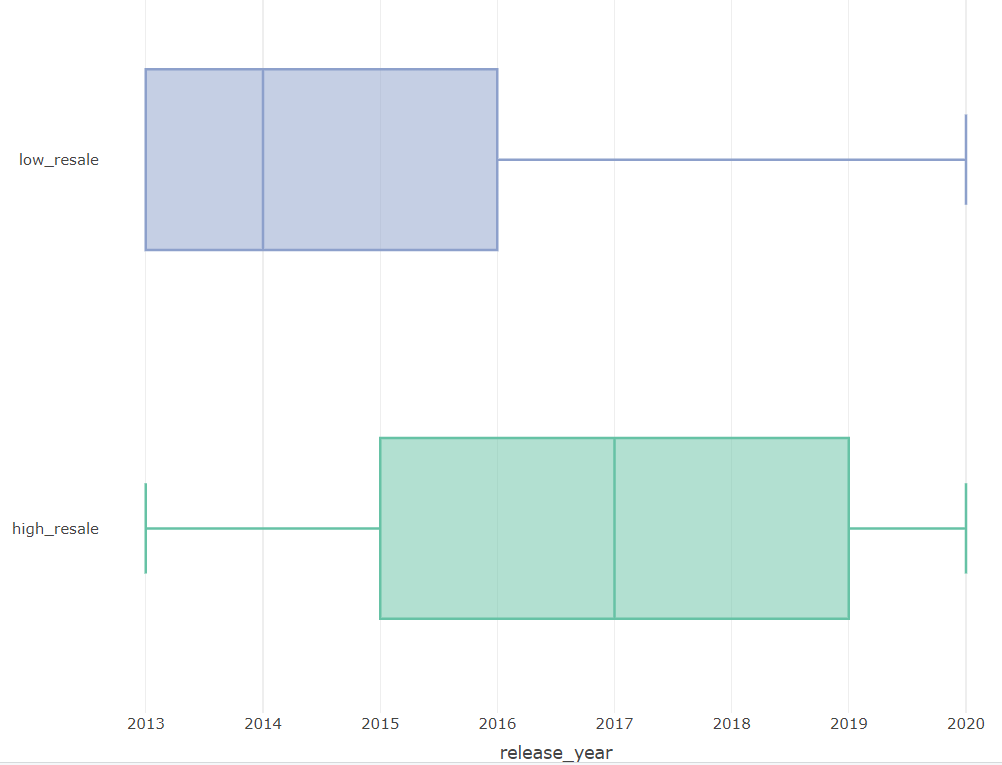


**Figure 2.24 weight vs resale\_category**

From the figure 2.24, it is apparent that there is a positive correlation between weight and normalized used price columns. Devices with a weight exceeding 200 grams tend to have a normalized used price above 4. Therefore, if BIG\_C focuses on selling devices with a weight above 200 grams, they can achieve a normalized resale price exceeding 4.

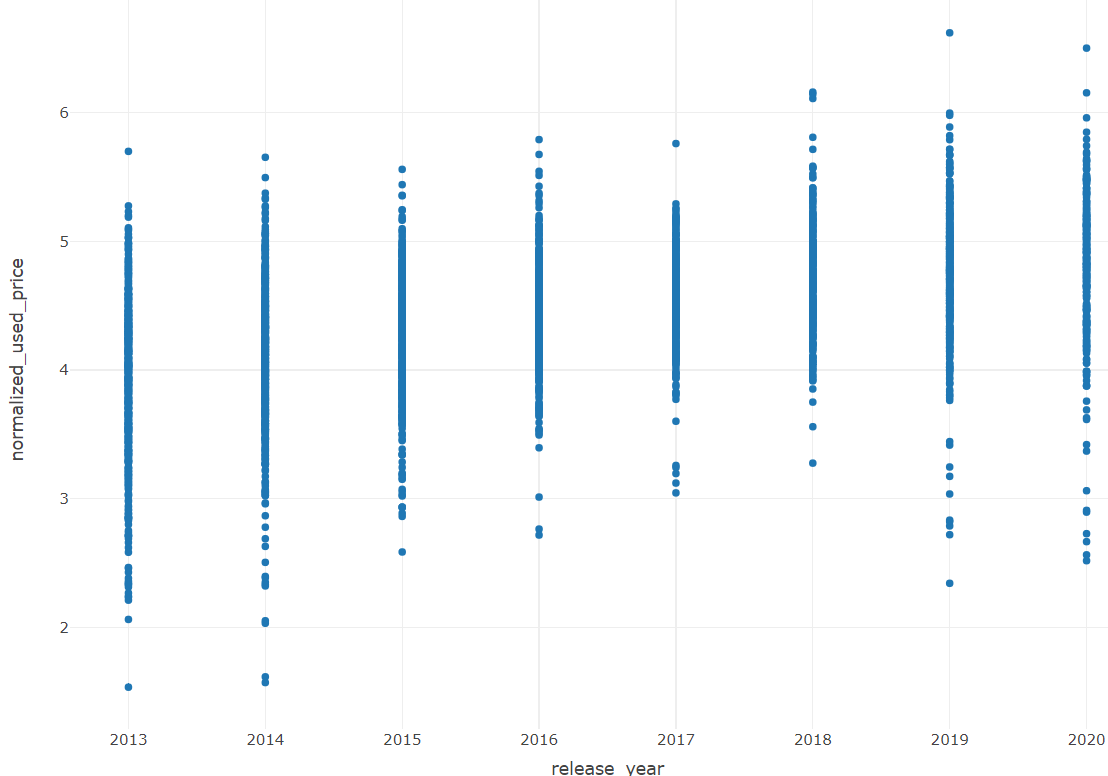
**release\_year Column**

Release year column indicates that year the device first released. It is a numerical column. Let’s try to find how release\_year is impacting resale\_category.



**Figure 2.25 release\_year vs resale\_category**

From figure 2.25 it is clear that most of the devices released after the 2014 have high resale value. So BIG\_C company can gain more profits by reselling the devices released after 2014.

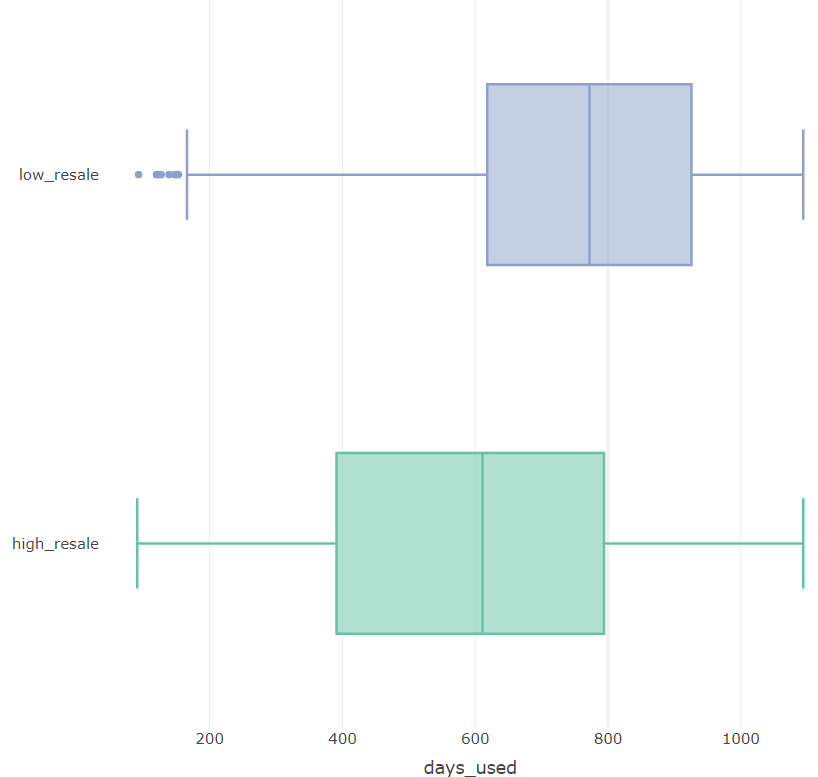


**Figure 2.26 release\_year vs normalized\_used\_price**

From the figure 2.26, it is apparent that there is a positive correlation between weight and release\_year columns. Most of the devices that released after year 2015 have a normalized used price above 3. Therefore, if BIG\_C focuses on selling devices that released after year 2015, can achieve a normalized resale price exceeding 3.

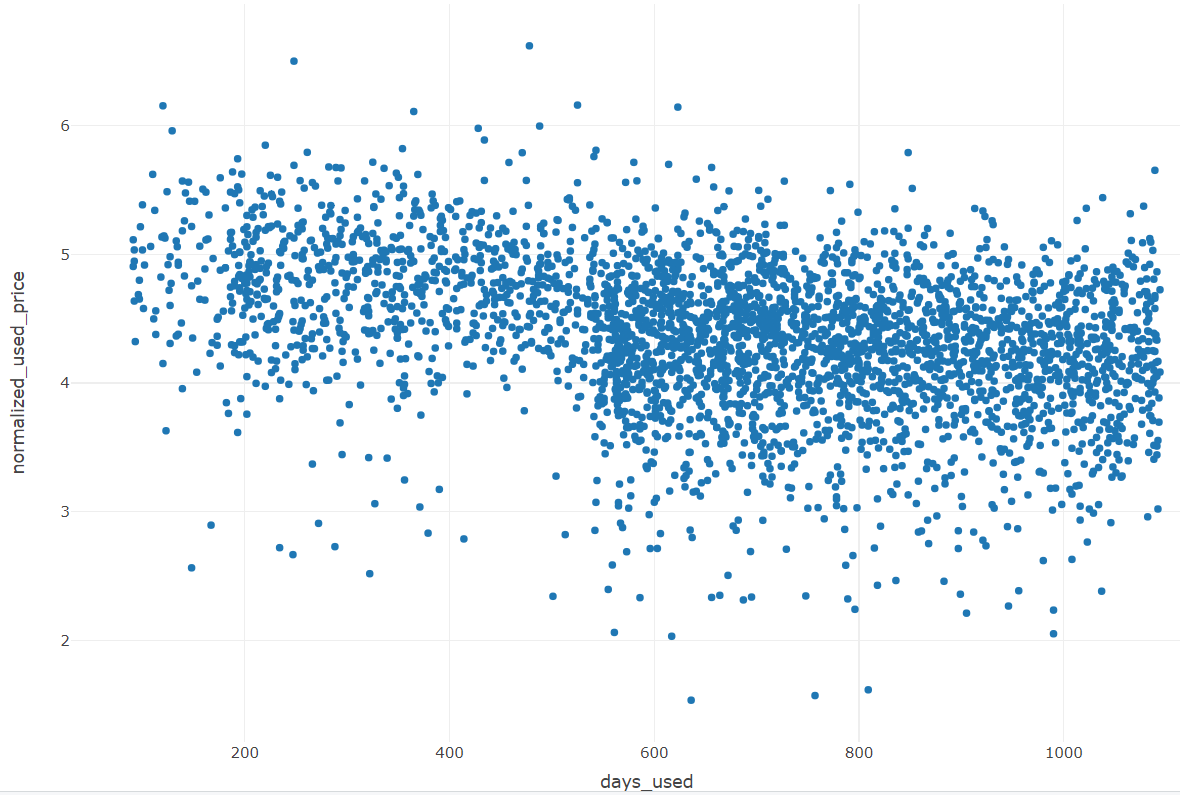
**days\_used Column**

days\_used column represents the number of the days the original customer used the device. Let’s explore how it is impacting the resale\_category.

****

**Figure 2.27 days\_used vs resale\_category**

From figure 2.16 it is clear that most of the devices that used less number of days have high resale value. So, to gain more profits BIG\_C company needs to focus on selling the devices that used less number of days by the original customer.

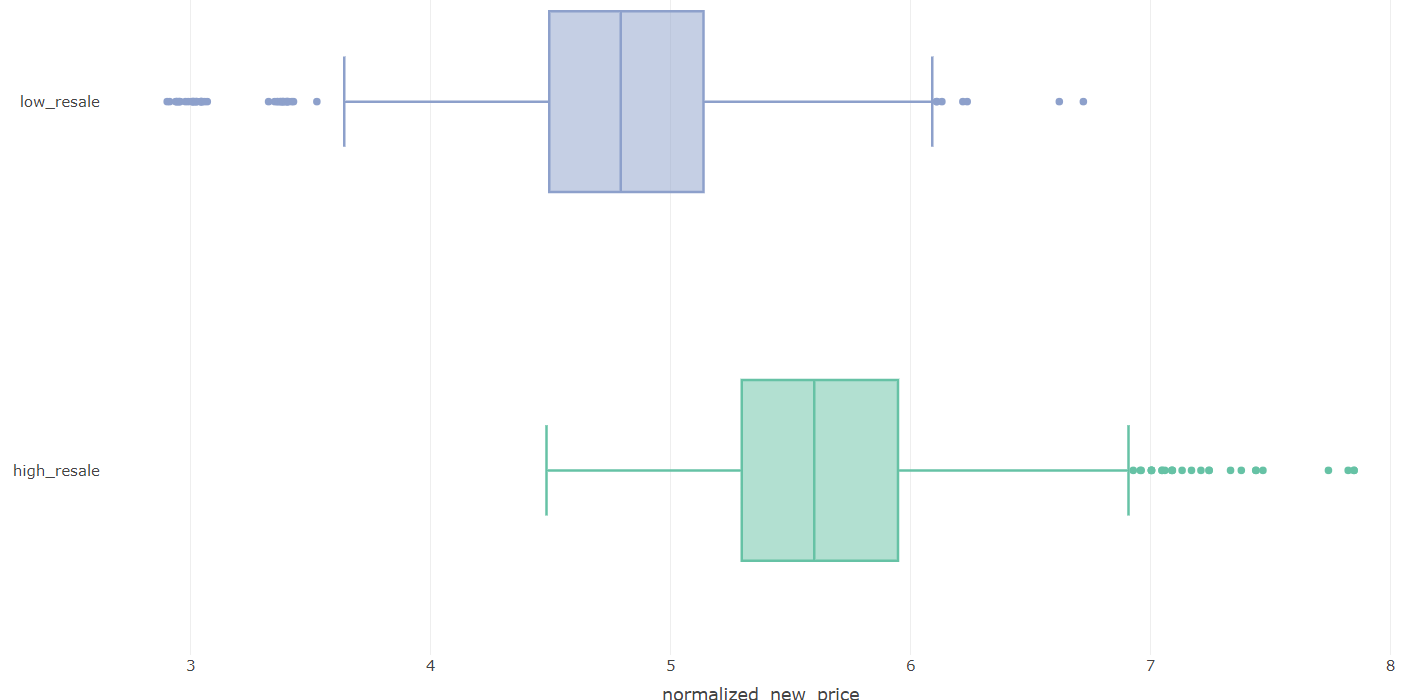


**Figure 2.28 days\_used vs normalized\_used\_price**

From the figure 2.28, it is evident that there is a negative correlation between days\_used and normalized used price. As the number of days the device is used by the original customer increases, the normalized used price tends to decrease. Notably, if the number of days used by the original customer is less than 600, the normalized used price is greater than 4. Therefore, if the company sells devices that are used by the original customer for not more than 600 days, they can achieve a normalized resale price exceeding 4.

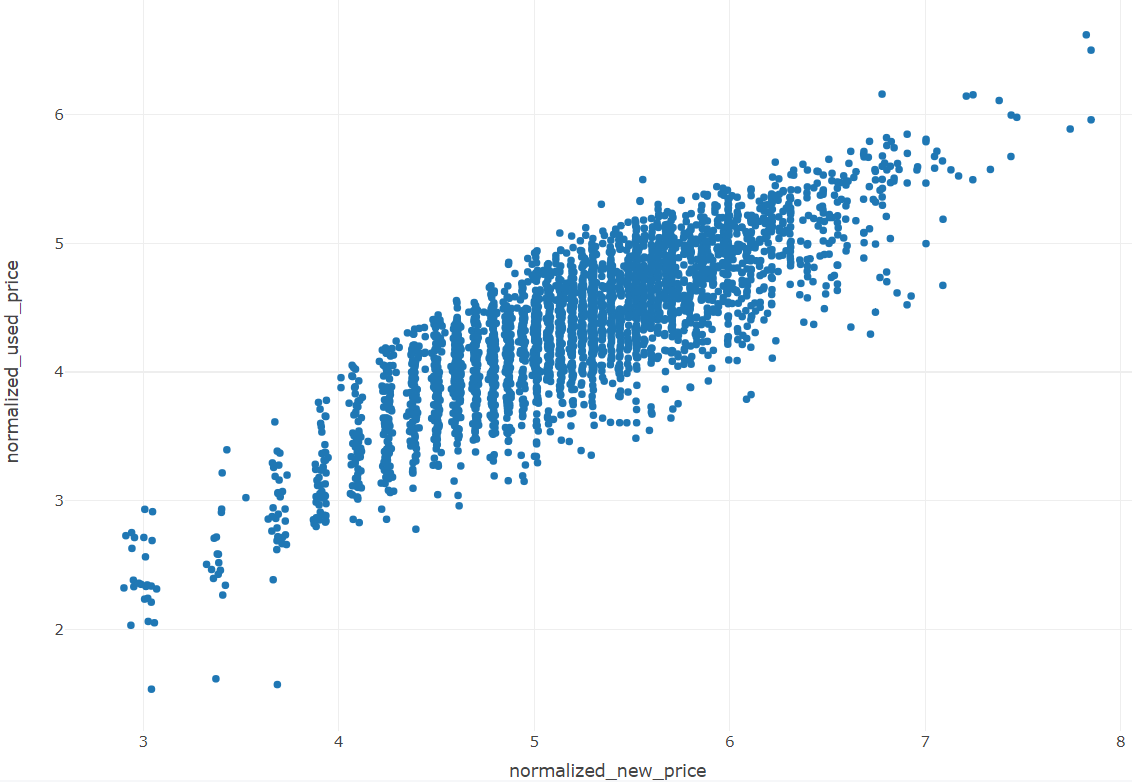
**Normalized\_new\_price**

This column represents the price of the new device.in normalized fromat. Let’s try to explore how the normalized price of the new device will impact the resale\_category.

****

**Figure 2.29 normalized\_new\_price vs resale\_castegory**

From figure 2.29 it is clear that if a device normalized new price is more than 4.5 then that device is having high resale. So BIG\_C company can get more profits if the company focuses on reselling the device that has new normalized price more than 4.5.

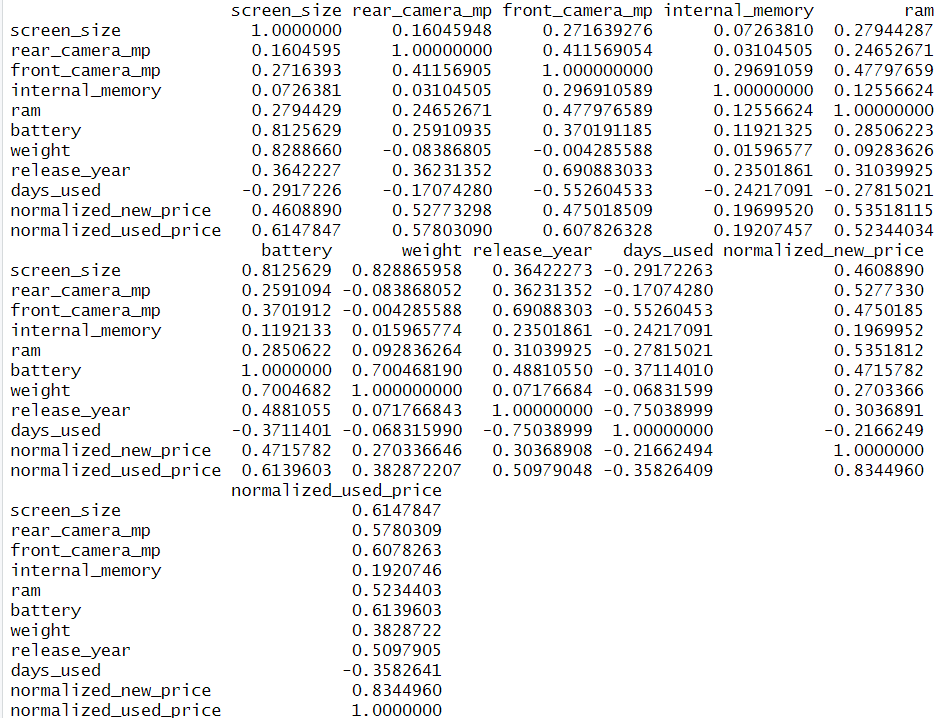


**Figure 2.30 normalized\_new\_price vs normalized\_new\_price**

From the figure 2.30 it is clear that normalized\_used\_price and normalized\_new\_price is positively correlated. if a device have normalized new price greater than 6 then normalized used price is 4. So If company focus on selling the devices that has normalized new price greater than 6 then company will normalized resale price greater than 4.

**Correlation Analysis**

Now let’s try to understand the correlation between all the numerical predictors.

****

**Figure 2.31 correlation matrix**

Findings of the correlations from figure 2.18 are

* Screen\_size has high positive correlation with battery and weight.
* Rear\_camera\_mp has high positive correlation with normalized\_new\_price.
* Front\_camera\_mp has high positive correlation with release\_year and high negative correlation with days\_used.
* Ram has high positive correlation with normalized\_used\_price.
* Battery has high positive correlation with screen\_size and weight.
* Weight has high positive correlation with screen\_size and battery
* Release\_year has high positive correlation with front\_camera\_mp and high negative correlation with days\_used.
* Days\_used has high negative correlation with front\_camera\_mp and release\_year.
* Normalized\_new\_price has high positive correlation with rear\_camera\_mp and ram

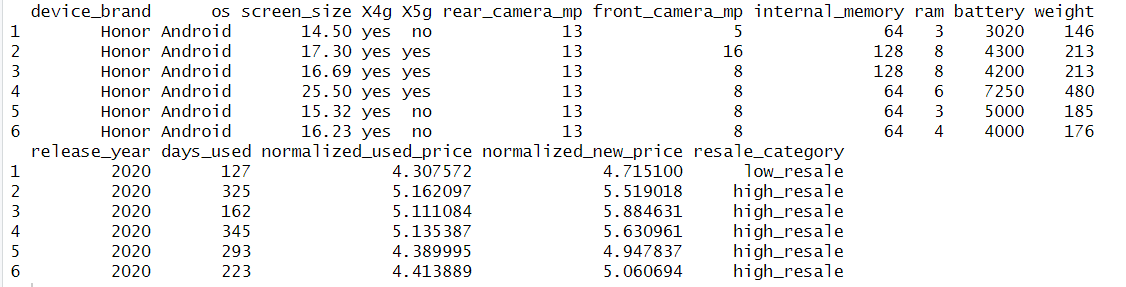
Columns with high positive or negative correlations indicate duplicative information between them.

All the columns except internal\_memory and days\_used have high correlation with target column normalized\_used\_price so all these columns are useful for predicting normalized\_used\_price.

**Outliers**

From the all the EDA analysis it is found that there are no outliers in the data.

**Normalizing and Rescaling Data**

****

**Figure 2.32 First 6 rows of the data**

Figure 2.32 clearly shows that the columns have values on different scales. When utilizing KNN models or neural networks, standardization or normalization of the data is necessary.

1. **Dimension Reduction**

The initial number of columns are very less and if logistic regression or knn and neural networks used dummy variables will be created which will expand the number of columns so it will be difficult for the BIG\_C company to collect the error free data so reducing the dimension is important.

**Practical considerations**

Based on domain knowledge and insights from EDA analysis for the classification task, the column normalized\_used\_price is not useful, as resale\_category is derived from it. For both classification and regression tasks, the normalized\_new\_price column is unnecessary because it only shows the original new price. In clustering, the focus is solely on device features, so the columns brand, original new price, and used price are not used..

**Dimension Reduction using correlation analysis**

All the columns that are high positive or negative correlations have the duplicate information. So, one of the columns can be removed to reduce the multicollinearity.

**Dimension Reduction using regression Models.**

Forward selection and backward selection algorithms utilize predictors that are statistically significant for predicting the outcome columns. Predictors that are not included in the forward or backward selection model can be removed.

**Dimension reduction using Decision Tree**

The resulting tree from the decision tree is used to identify important features. Predictors that are not included in the decision tree can be removed.

1. **Partition the Data**

As it is a supervised task, utilizing the complete data to build and test the model's performance may introduce optimism bias. This bias arises because the model may perform well with the current dataset, but it might not generalize effectively to real-world scenarios due to factors such as slight variations in the data which will be huge loss for the BIC\_C company. So the data needs to be partitioned into train, validation, and test sets.

The train partition is the largest partition containing the data to train various models. The same training data is used across different models. The validation partition is utilized to assess the predictive performance of each model and select the best model from the options available. The test partition is used to evaluate the performance of the chosen model with new data.

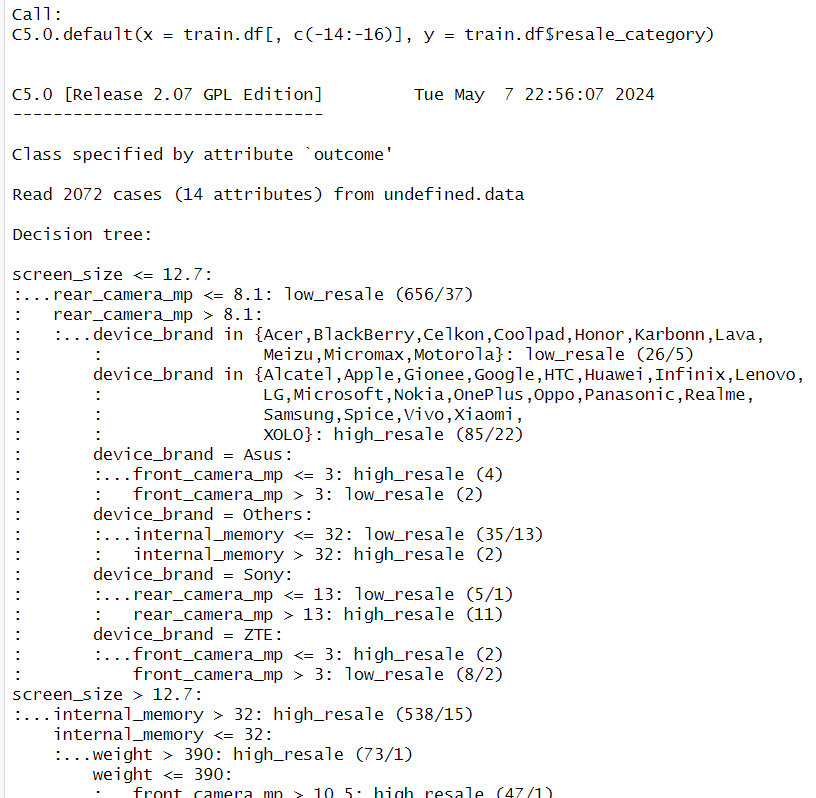
Partitioning the data into train, validation, and test sets is done randomly according to the proportions of 60%, 20%, and 20%, respectively. The records in the train, validation, and test partitions should be distinct. After partitioning the dataset, the train partition contains 2072 records, the validation partition contains 690 records, and the test partition contains 692 records.

1. **Modelling For Classification Task**

Now will try to build the classification model to classify if a used device has high\_resale value or low\_resale. Based on the classification the BIG\_C company can decide whether to buy the used device from the original customer or not. For modelling the target column will be the resale\_category and predictors will be all columns except normalized\_new\_price and normalized\_used\_price. The performance of the model can be evaluated by using the confusion matrix and here high\_resale category is considered as the important class because it is important to classify high\_resale category because classifying a low\_resale category mobile as high\_resale mobile will be a huge loss for the BIG\_C company.

* 1. **Decision Tree**

Decision tree is based on separating records into subgroups by creating split on predictors. These splits create logical if then prediction rules which are simple and easy to interpret. Decision trees can be applied for both classification and regression. There are multiple ways to fit the classification decision tree in one of the ways is using C5.0 library.

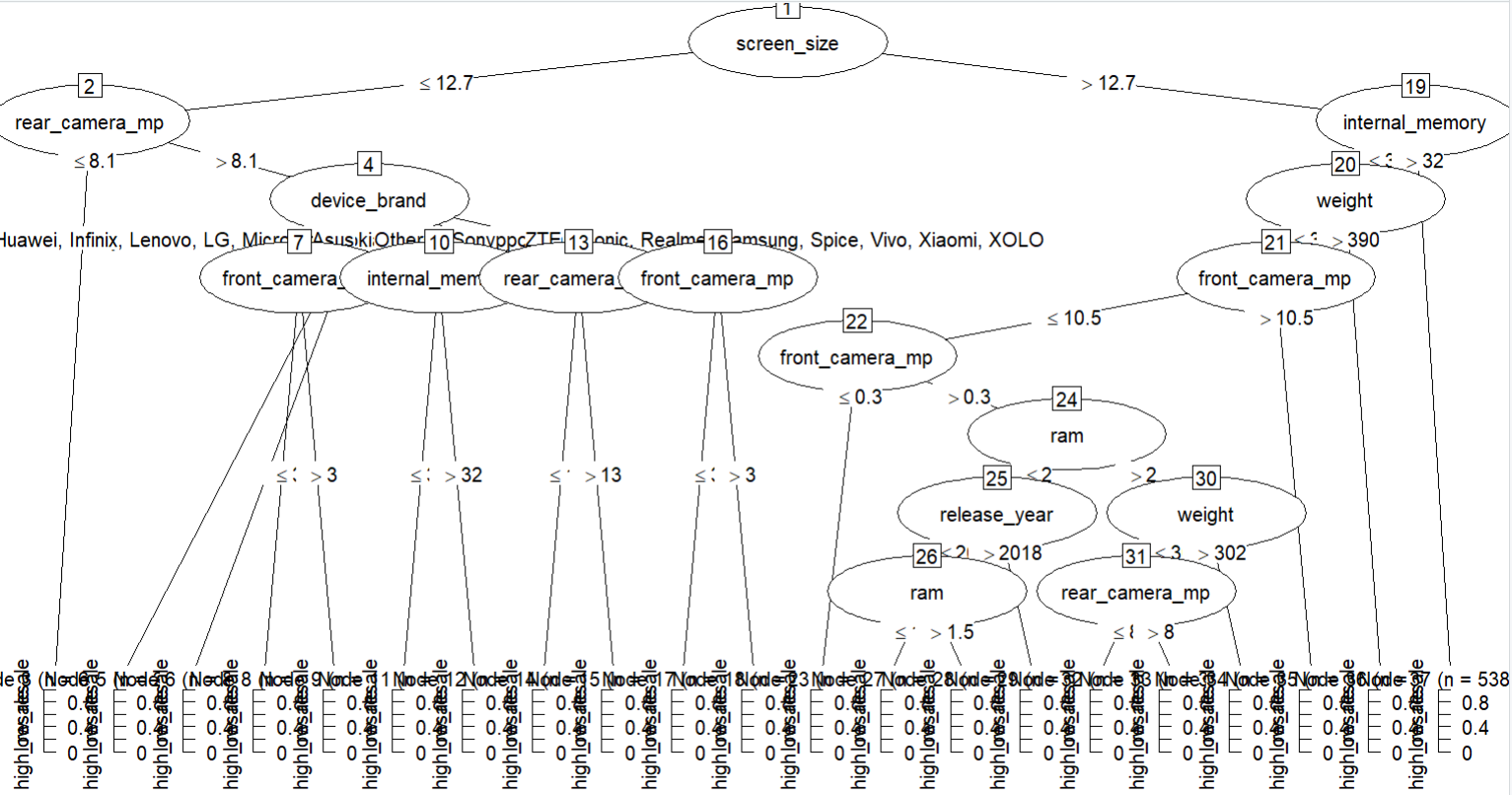
****

**Figure 5.1 Decision tree rules**

**Model Interpretation**

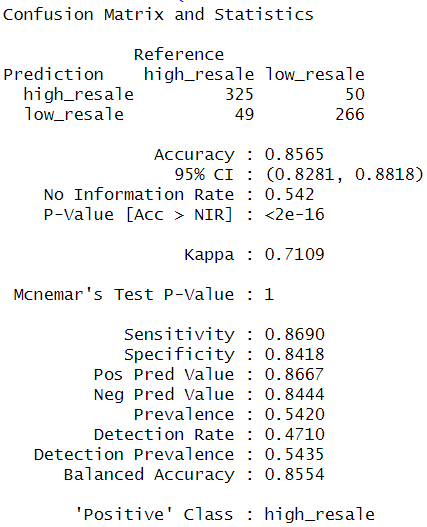
When classifying new records, the record will dropdown to the terminal node and based on the rules created by the decision tree which is shown in figure 5.1 the new record will be classified.

Using the plot function in R, one can visualize the tree structure.

****

**Figure 5.2 Tree structure**

* + 1. **Measuring Performance**

****

**Figure 5.3 confusion matrix on valid data**

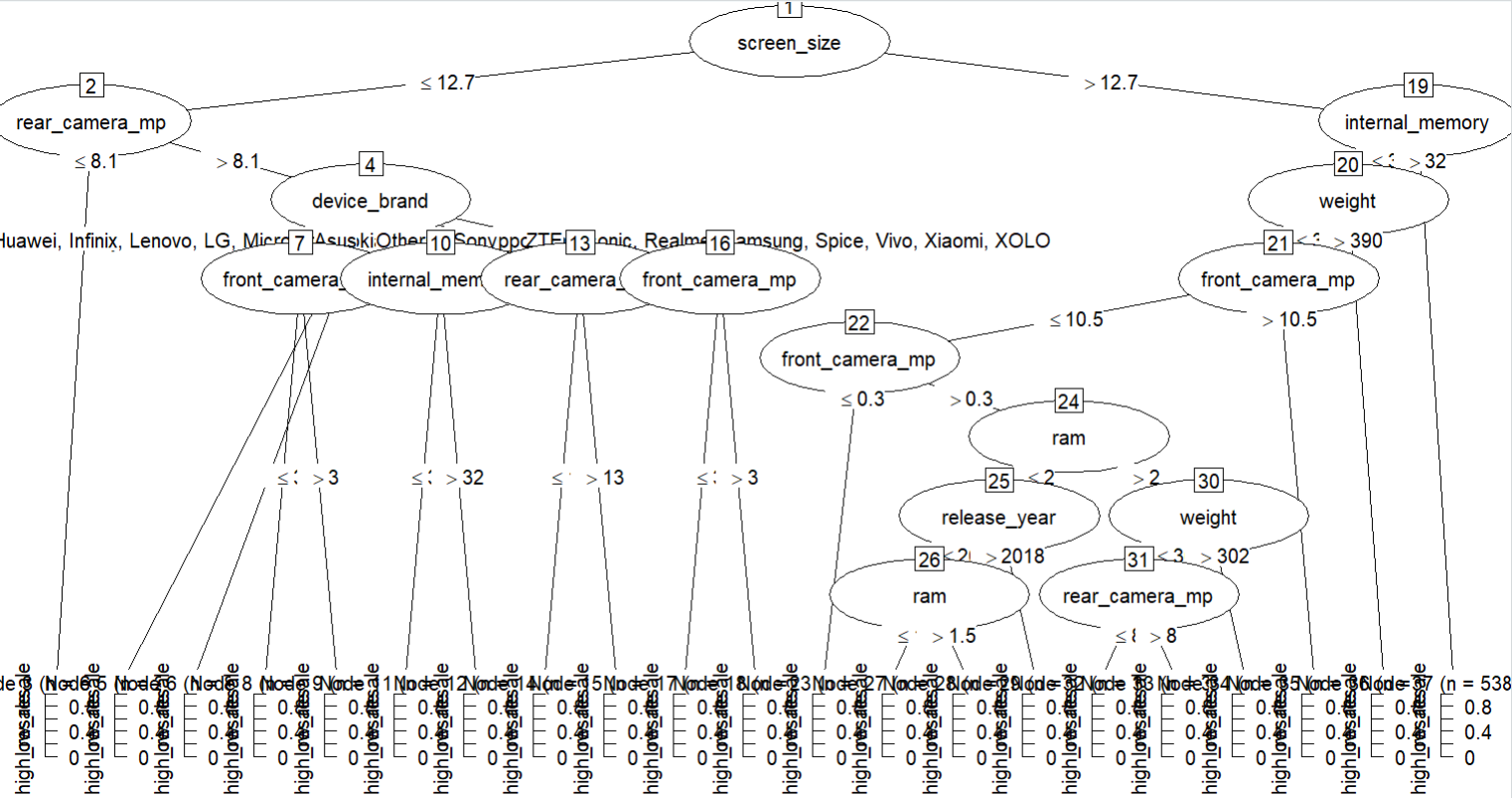
From the figure 5.3 the sensitivity rate is 86.90%, indicates that the model correctly classifies 86% of high-resale used devices in the validation set.

* + 1. **Improving Model Performance**

The performance of the decision tree model can be improved by using an extra parameter trail while building the model.

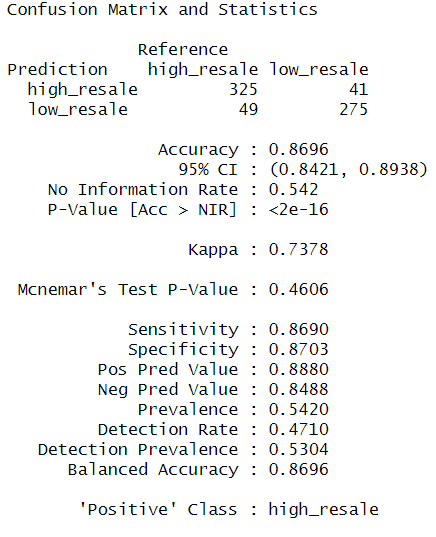
**Trails option 15**

After trying various trail options such as 6, 10, 15, and 20, it was found that trail option 15 has the highest sensitivity rate. Therefore, trail option 15 is deemed the best model among all other trail models.

****

**Figure 5.4 Tree Structure for trail 15**

**Measuring Performance**

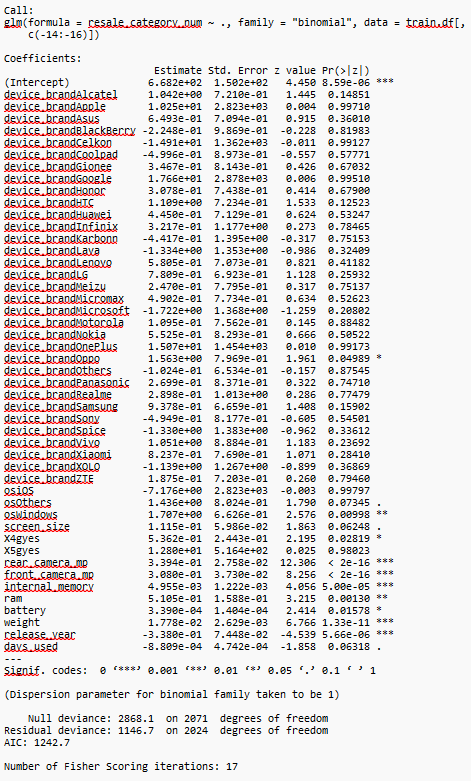
****

**Figure 5.5 confusion matrix on validation set for trail option 15**

From the figure 5.5 the sensitivity rate is 86.9 which indicates that 86.9% of the high\_resale used devices in the validation set is correctly classified.

* 1. **Logistic Regression**

Logistic regression provides estimates of propensities indicating the likelihood of each record belonging to each class. Subsequently, Use the threshold values to classify each case into one of the classes. In R GLM is used to fit the logistic regression. The GLM model works well if the target column has 0s and 1s, so before applying logistic regression, the target column resale\_category is converted to 1s and 0s by assigning high\_resale category to 1 and low\_resale to 0 in the train data.

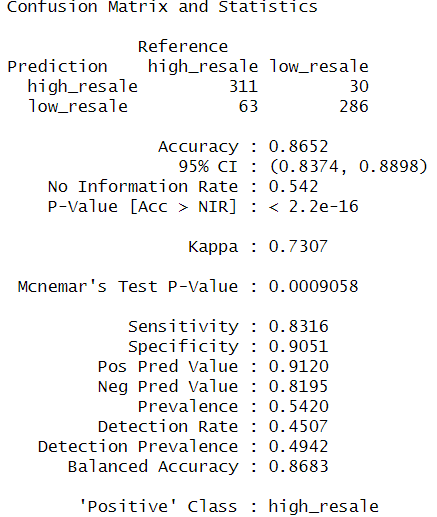
****

**Figure 5.6 Summary of the logistic regression**

**Model Interpretation**

Based on the model estimated equation the new records are classified as class 1(high\_resale) or 0 (low\_resale).Columns device\_brand, os, rear\_camera\_mp, front\_camera\_mp, screen\_size, x4g, x5g, internal\_memory, ram, battery, weight, release\_year, days\_used are most significant columns for classifying if a used device is belongs to class 1(high\_resale) or 0 (low\_resale).

**5.2.1. Measuring Performance**



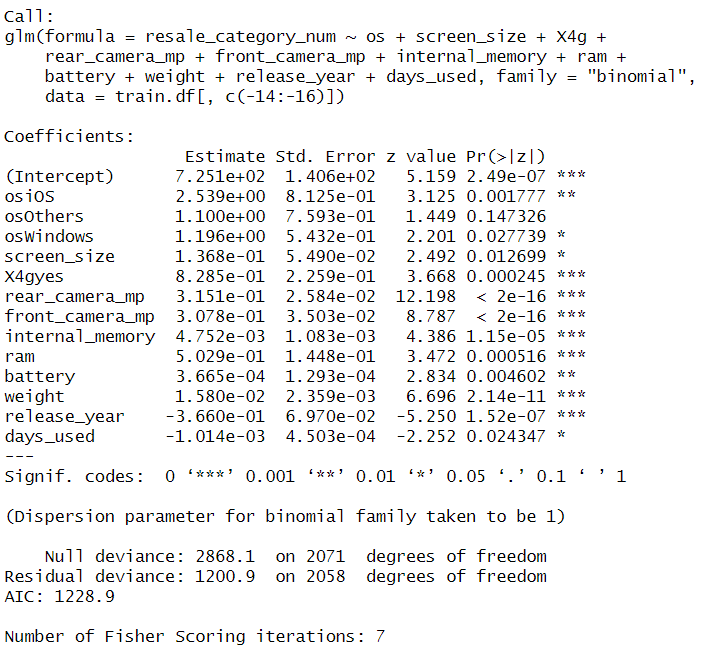
**Figure 5.7 confusion matrix of logistic regression on validation set**

From the figure 5.7 the sensitivity rate is 83.16% which indicates that 83.16% of the high\_resale used devices in validation set are classified correctly.

**5.2.2. Improving Model Performance**

In this section, the aim is to construct Backward selection models to assess whether there is an improvement in predictive performance compared to logistic regression.

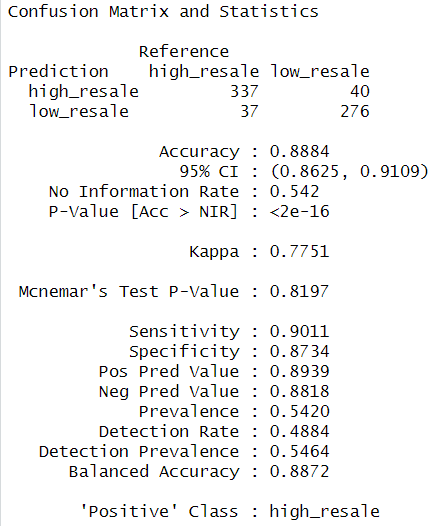
.

****

**Figure 5.8 summary of backward selection model**

Columns os, screen\_size, rear\_camera\_mp, front\_camera\_mp, x4g, internal\_memory, ram, battery, weight, release\_year, days\_used are most significant columns for classifying if a used device belongs to class 1(high\_resale) or 0(low\_resale).

**Measuring Performance**

****

**Figure 5.9 confusion matrix confusion matrix of backward selection logistic regression on validation set**

From the 5.9 the sensitivity rate is 90.11%, indicating that 90.11% of the high-resale used devices in the validation set are classified correctly.

1. **Classification Model Selection**

**Table 6.1 performance Comparison**

|  |  |  |
| --- | --- | --- |
| Model | Accuracy | Sensitivity |
| Decision Tree | 85.65% | 86.9% |
| Decision Tree trail 15 | 86.96% | 86.9% |
| Logistic Regression | 86.52% | 83.16% |
| Backward selection logistic regression | 88.84% | 90.11% |

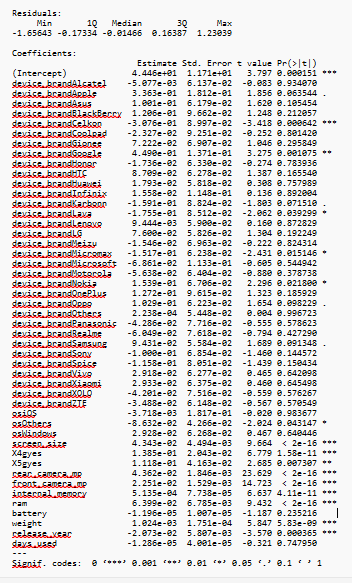
From table 6.1 the sensitivity rate of backward selection logistic regression model is higher than all other models. So, the backward selection logistic regression model is considered as the best model to classify the used device as high\_resale or low\_resale. So BIG\_C company can use the best classifier model to classify if the used device will have high\_resale or low\_resale. Based on the model output BIG\_C company selectively purchase only devices classified as having high resale value instead of buying all used devices.

1. **Modelling for Regression Task**

From now on, the prediction will focus on determining the market price of a used device. The normalized\_used\_price column is considered as the target column, and all remaining columns except resale\_category and normalized\_new\_price are considered as the predictors.

* 1. **Linear Regression**

The lm function is used to fit a multiple linear regression model. The multiple linear regression tries to form linear relationship between predictors and outcome column.

****

**Figure 7.1. Summary of the linear regression model**

**Interpretation of the Model**

The regression coefficients generated by the model are used to estimate the normalized\_used\_price of a used device. From the summary of the linear regression model shown in figure 7.1 it is clear that out of all predictors device\_brand, os, screen\_size, X4gyes, rear\_camera\_mp, front\_camera\_mp, internal\_memory, ram, weight, release\_year are most significant predictors to estimate the normalized\_used\_price of a used device.

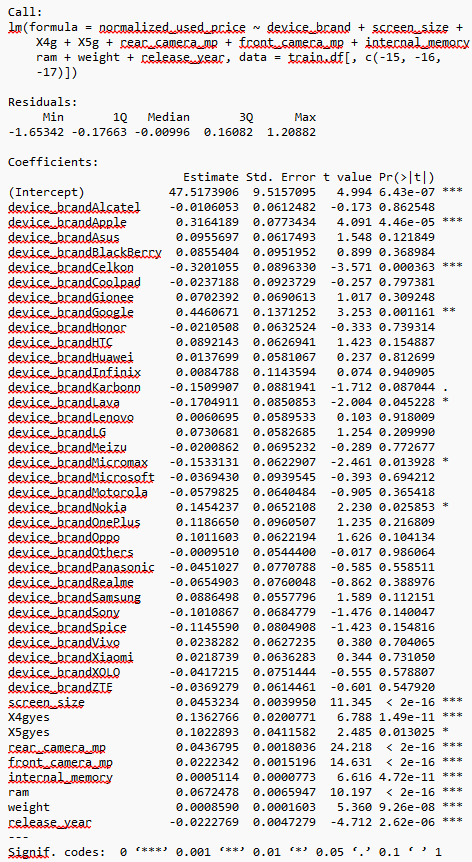
* + 1. **Measuring Performance**

To measure performance, RMSE is used. The model with lowest value of RMSE value is the best model.

The RMSE value for the validation data is 0.2930238.

* + 1. **Improving Model Performance**

In this section, the aim is to construct Backward selection models to assess whether there is an improvement in predictive performance compared to multiple linear regression model. Backward selection algorithm will start with all predictors and in each step will remove predictors one by one that are not important.



**Figure 7.2. Summary of the backward selection model**

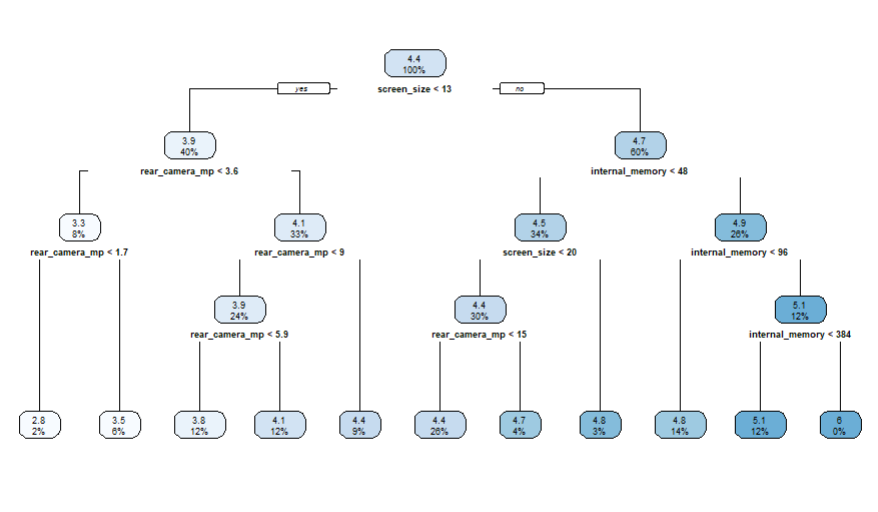
Out of all predictors only device\_brand , screen\_size , X4g , X5g , rear\_camera\_mp , front\_camera\_mp , internal\_memory , ram , weight , release\_year are most significant and used predictors to estimate the normalized\_used\_price of a used\_device.

**Performance measure**

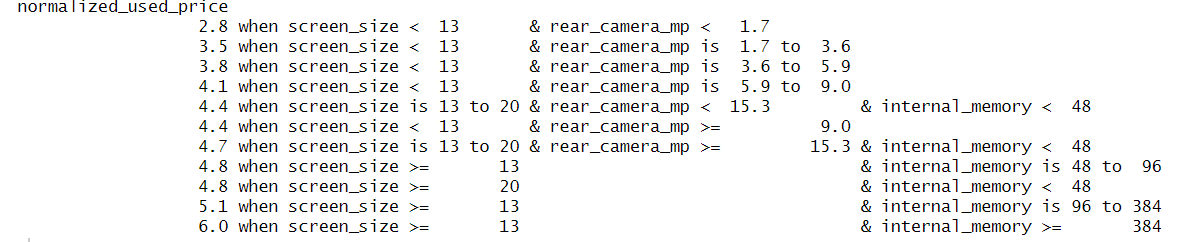
The RMSE of the backward selection model is 0.2940584.

* 1. **Decision Tree for Regression**

Regression tree is non-parametric so there is no assumption for the model. So, train data can be directly use to train the model. Regression tree can be applied in R by using RPART.



**Figure 7.3 Decision tree**

****

**Figure 7.4 Rules Generated by the Decision Tree**

**Interpretation**

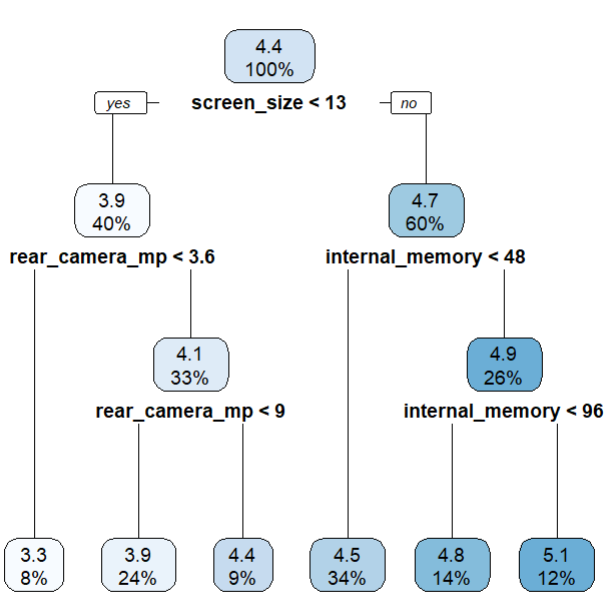
When predicting a used device normalized\_used\_price, the record will drop down to the terminal node based on the predictors' information, and the prediction will be the average normalized\_used\_price in the terminal node.

* + 1. **Measuring Performance**

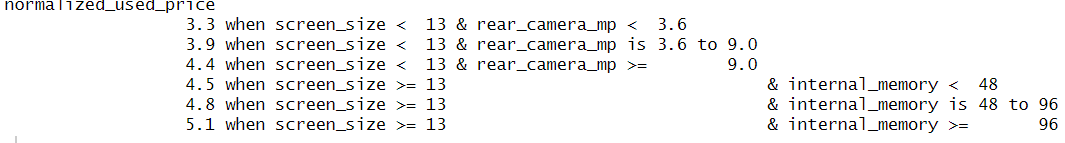
The RMSE value of the validation data is 0.283172.

* + 1. **Improving Model performance**

The regression tree model can be improved by adjusting the CP values and selecting the one that got less RMSE value. After trying all different CPs like 0.02, 0.03, 0.04, 0.05, 0.06, 0.07 ,0.08, 0.09, 0.10. The Best Cp value is 0.02.



**Figure 7.5 Decision tree for CP value 0.01**

****

**Figure 7.6 Rules for cp value 0.01**

**Performance measure**

The RMSE value for the validation data is 0.317542.

1. **Regression Model Selection**

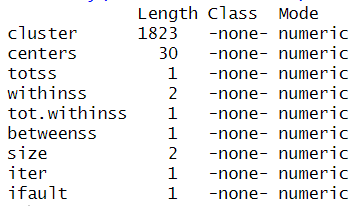
**Table 8.1 Performance Comparison**

|  |  |
| --- | --- |
| Model | RMSE on Validation Data |
| Multiple Linear Regression | 0.2930238 |
| Backward Linear Regression | 0.2940584 |
| Decision Tree (Regression Tree) | 0.283172 |
| Decision Tree (Regression Tree cp 0.02) | 0.317542 |

From table 8.1, it is evident that Decision Tree (Regression Tree) have RMSE values less than all other models, Therefore, Decision Tree (Regression Tree) can be considered as the best models. So by using the regression tree BIG\_C company can estimate the market price for a used devices which helps to preventing the company from overinvesting in devices.

1. **Modelling for Clustering Analysis**

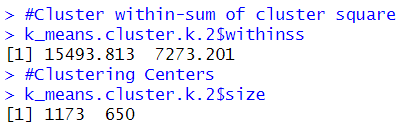
Going forward, the primary focus will be on modeling to recommend similar used devices to customers when their preferred used devices are out of stock or not available. The K Means model is suitable for clustering such devices. Since clustering is an unsupervised learning approach, the complete dataset is utilized to train the K-Means model. However, given that the BIG\_C company exclusively sells used devices with high resale value, only used devices with high resale value are filtered from the entire dataset to train the clustering models. The similarity of used devices is assessed using distance measures. As it is impossible to calculate distances for categorical columns, all categorical columns need to be converted into dummy variables. Before creating dummies x4g and x5g columns are converted into 1’s (yes) and 0’s (No) as these columns are already in dummies. Considering that distance depends on the scale of the data, the data needs to be standardized. In modeling, all features are used except device\_brand, normalized\_used\_price, normalized\_new\_price, and resale\_category. First, the k-means model is trained with k=2.

****

**Figure 9.1 summary of model**

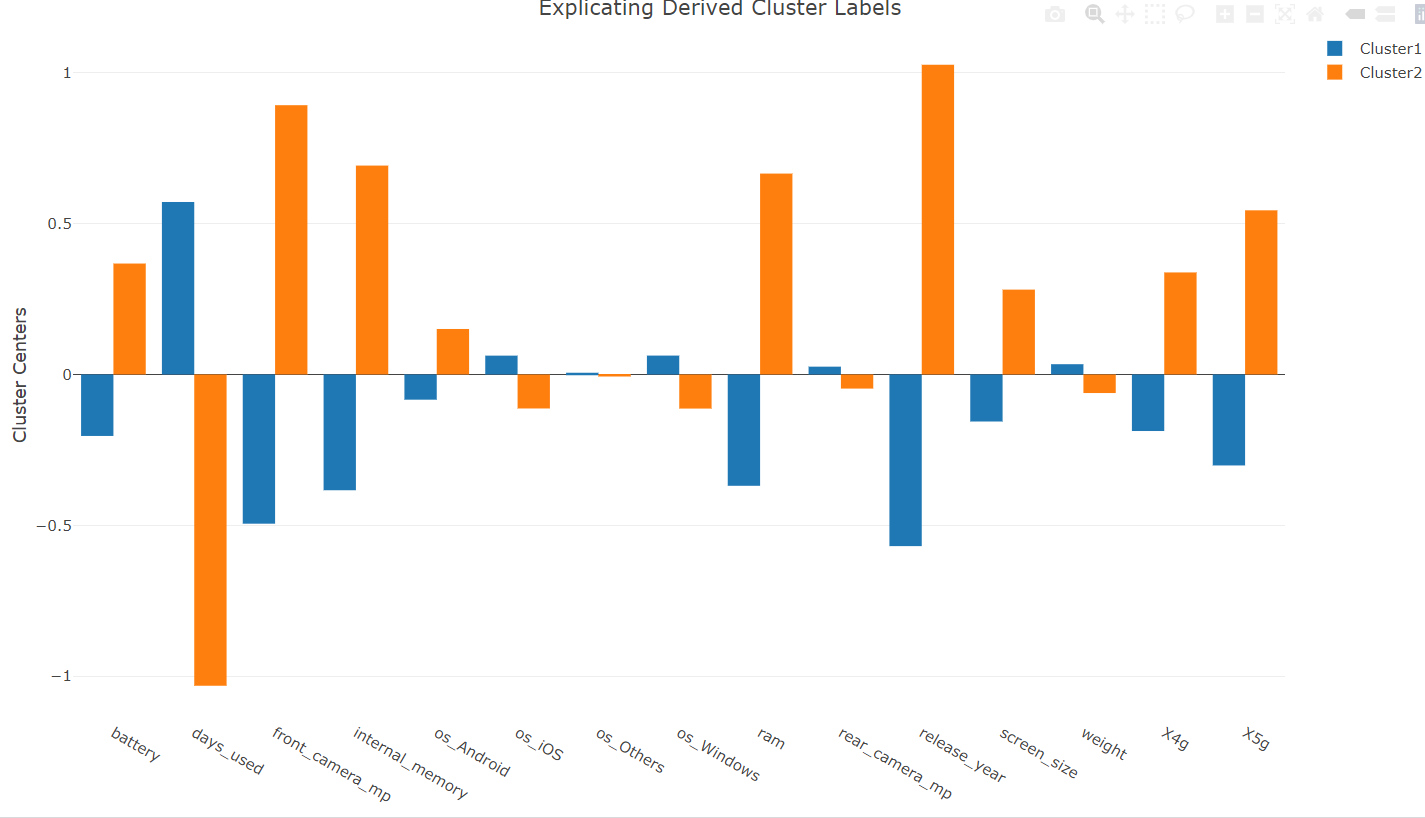
* 1. **Evaluating Model Performance**

There are many ways to evaluate the cluster performance. First let’s try to understand the size of each cluster



**Figure 9.2 cluster size and within sum of square**

From Figure 9.2, it is clear that Cluster 1, with 1173 records, has the largest within-sum squared deviation, while Cluster 2, with only 650 records, has a smaller deviation. This means that Cluster 2 is more homogeneous than Cluster 1. 1173 records in cluster 1 not clustered properly so, 2 Clusters are not appropriate

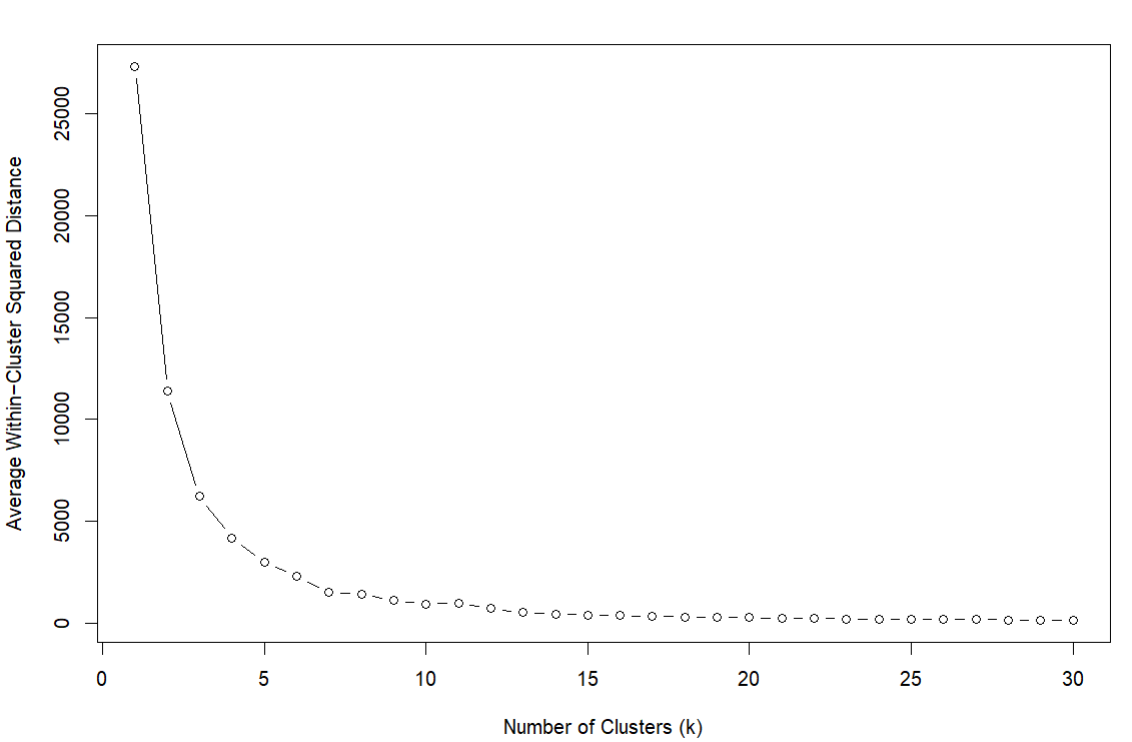


**Figure 9.3 bar plot of the clusters**

From Figure 9.3, it is clear that used devices that are in Cluster 2 has higher mega pixels of front camera, has released recently and used a smaller number of days.

* 1. **Improving k-means Model**

The k-means model can be improved by selecting different numbers of clusters, ranging from 2 to 30. The number of clusters should neither be too large nor too small. If the number of clusters is too large, the records become overly specific, losing their meaningfulness. Conversely, if the cluster number is too small, the records become excessively general.

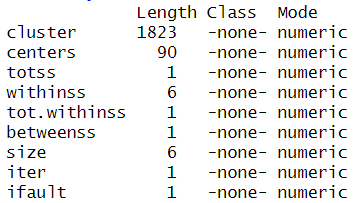
****

**Figure 9.4 Elbow curve for different clusters sizes**

From Figure 9.4, it is evident that among all the different numbers of clusters, Cluster size 6 is considered the best cluster size because after Cluster 6, the rate of decrease in the average within-cluster squared deviation starts to slow down.

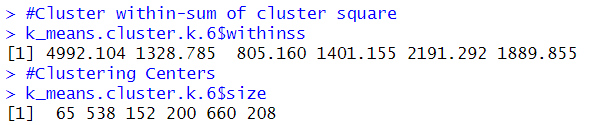
* + 1. **Retraining k-Means with Best cluster number**

Now the k-means model is trained with number of clusters 6

****

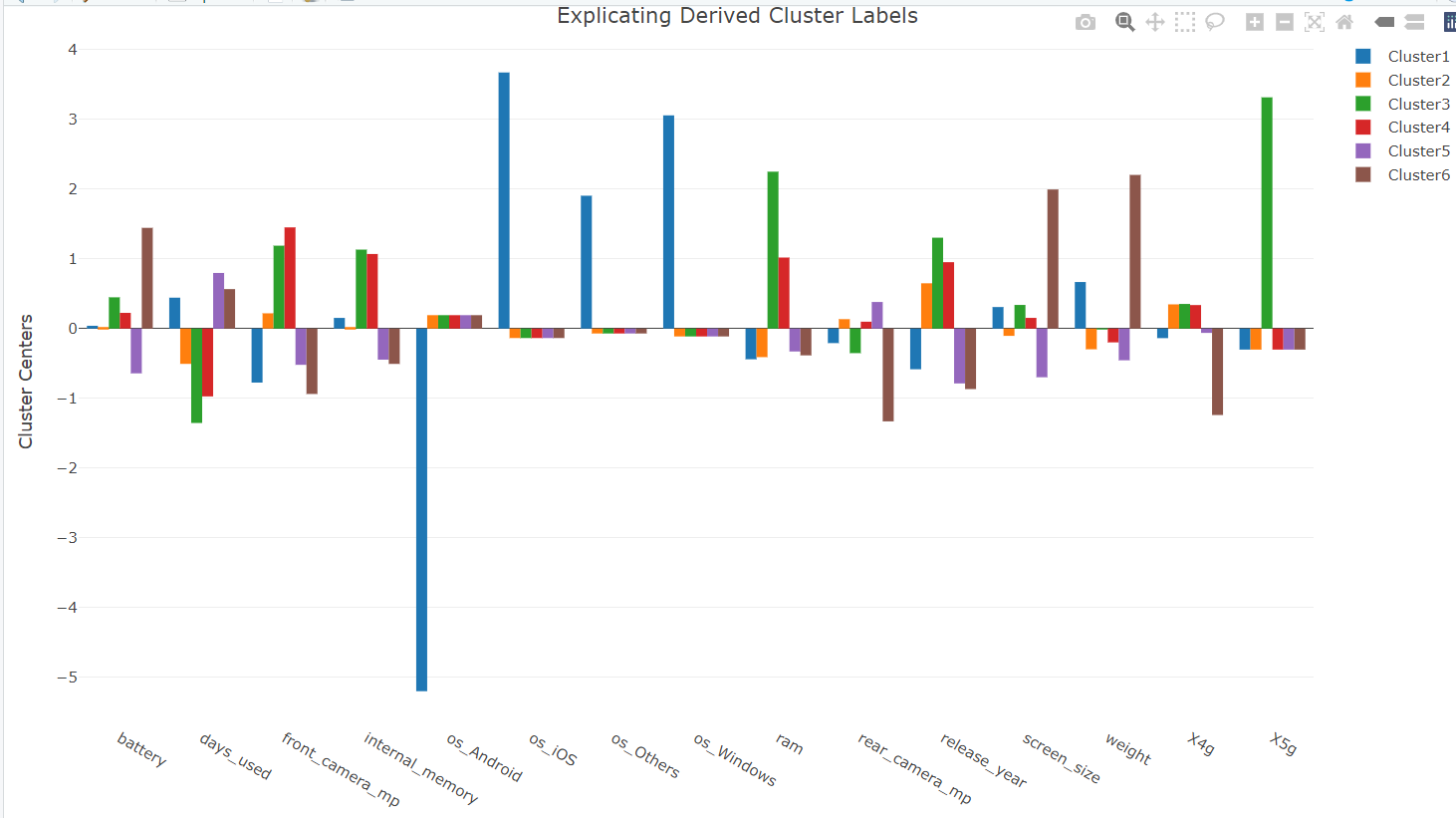
**Figure 9.5 summary of the model**

* + 1. **Evaluating Model Performance**



**Figure 9.6 cluster size and within sum of square**

From Figure 9.6, it is evident that among all clusters, Cluster 1 has the highest within-sum squared deviation compared to the other clusters. Cluster 3 exhibits the lowest within-sum squared deviation. Additionally, Cluster 5 has the highest number of records, whereas Cluster 1 has the lowest number of records

****

**Figure 9.7 plot of the clusters**

From Figure 9.7, it is clear that Cluster 1 comprises devices that supports ios, windows and other operating systems.

Cluster 2 has devices that has released recently and used a smaller number of days.

Cluster 3 has devices that supports mostly 5g network, has larger ram and internal memory and released more recently than cluster 3.

Cluster 4 has devices that has higher rear camera mega pixels.

Cluster 5 has the devices that used a greater number of days.

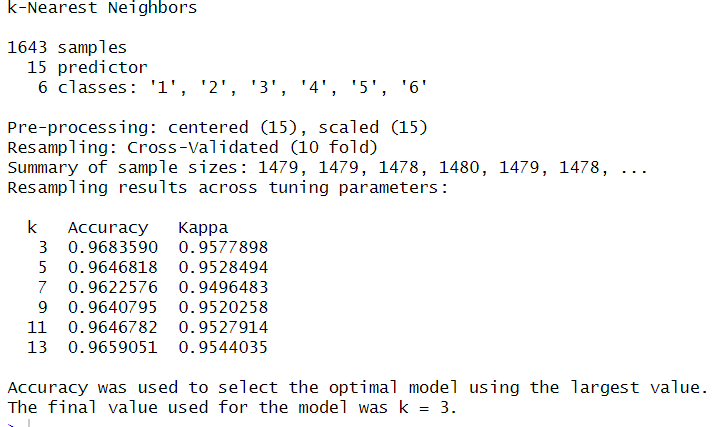
Cluster 6 has the devices that has higher battery capacity, higher screen size, low mega pixels of rear camera and high weight.

1. **Classification Modelling to classify cluster**

Now, a classification model is built to classify devices into their appropriate clusters. Since BIG\_C company only sells devices with high resale value, the classification model is trained using data that exclusively contains high resale value used devices. Since classification is a supervised task, the dataset needs to be partitioned into train and test sets. However, due to the limited number of records for devices with high resale value, only two partitions are created: train and test sets. To enhance model stability, cross-validation is employed during model training on the train data. This involves dividing the train data into multiple folds, with each fold alternately serving as validation data while the rest act as training data. Predictive performance is then assessed using the test partition.

* 1. **KNN Model**

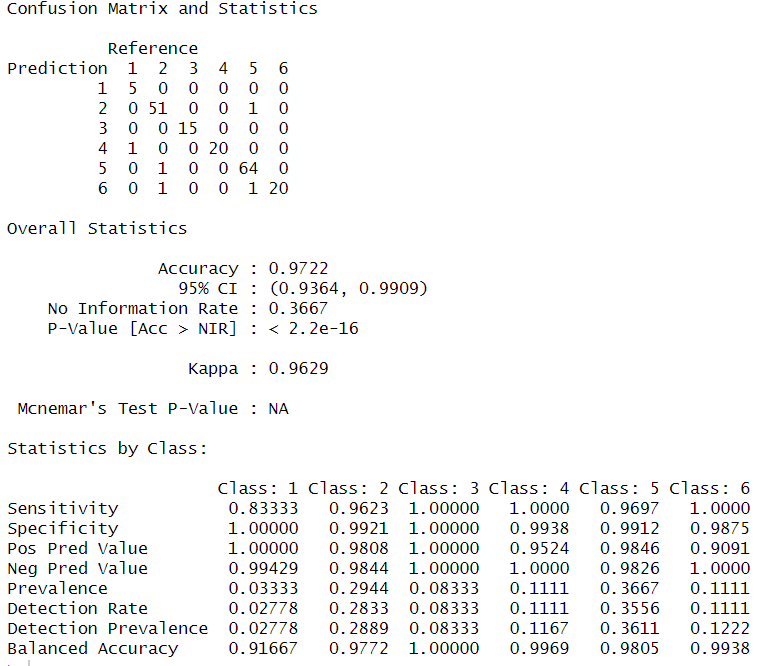
For classifying records into different clusters, the KNN model is considered the best option because it classifies records into clusters based on distance measures. When classifying new records, the model first identifies the n nearest neighbors in the training data that are similar to the new device. The new record is then classified based on the predominant class among these neighbors. All predictors are used to train the model, except for device\_brand, resale\_category, normalized\_used\_price, and normalized\_new\_price. The choice of the number of nearest neighbors significantly impacts the performance of the model. As the knn model works on distance measures all the categorical columns need to be converted into dummies. Before creating dummies x4g and x5g columns are converted into 1’s (yes) and 0’s (No) as these columns are already in dummies. Considering that distance depends on the scale of the data, the data needs to be standardized



**Figure 10.1 KNN model with best K**

* 1. **Measuring the Performance**

The predictive performance of the KNN model can be analyzed using test data. Here, accuracy is considered the most important metric because correctly classifying used devices into their respective clusters is crucial, rather than solely focusing on individual cluster classification.

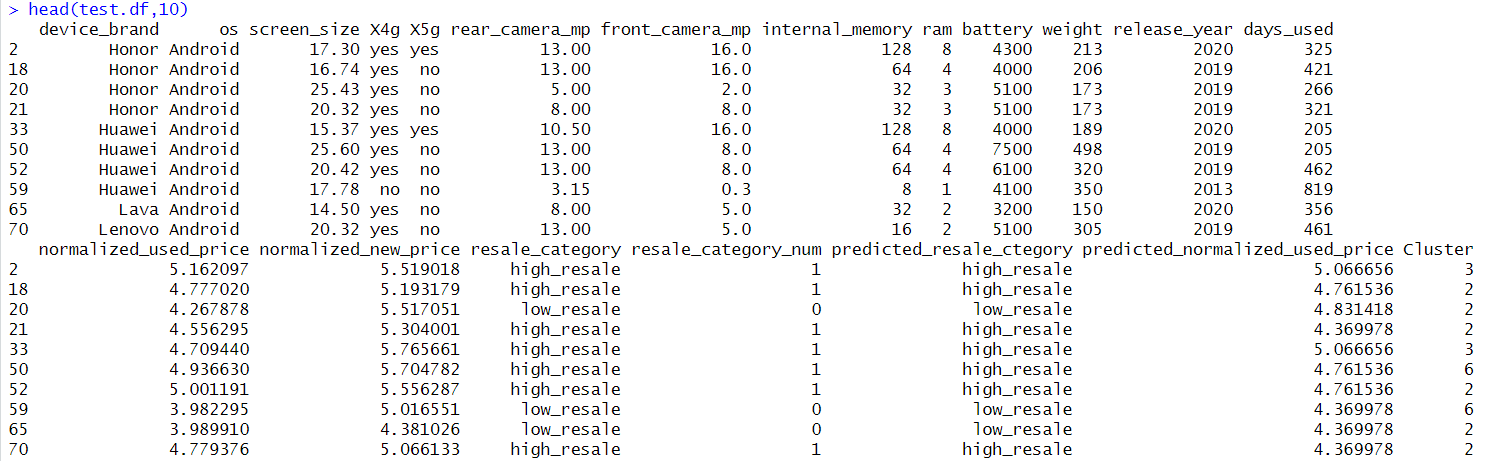
****

**Figure 10.2 confusion matrix of KNN on test data**

From the figure 10.2 it is clear that accuracy of the model on test data is 0.9722 which indicates that 97% of devices are correctly classified to its appropriate cluster.

1. **Predicting New Data**

In this section, all selected models are applied to the test data. Consider the test data as used devices that the BIG\_C company is willing to buy from the original customer. First, the best classification model backward selection logistic, is used to classify which used devices in the test set belong to the high resale category. Afterward, the best regression model regression tree, can be used to predict the market price of the devices. Consequently, the company can buy only the used devices classified as high resale, and the BIG\_C company can purchase used devices classified as high resale category for a price lower than the market price, then sell them to customers through its online platform for a higher price. Suppose customers are searching for a used device, the classification model KNN model, can be used to assign the used device that the customer wants, to a cluster. If the customer's preferred device is not available, then all the used devices in that cluster are recommended to the customer.



**Figure 11.1 Sample records in test data with all predicted values**

1. **Conclusion**

BIG\_C is an e-commerce company focusing primarily on selling used devices. The company acquires these devices from original users and resells them through its online platform. Currently, BIG\_C is facing loss for three reasons. To overcome these challenges, the company has decided to use predictive analysis on historical data. Backward selection classifier and regression tree models can be used to buy used devices from original users selectively for a price lower than the market price, which will increase profits rather than purchasing all used devices and overinvesting in them from the original owner. BIG\_C can use knn model to recommend similar used devices to a customer when the customer's preferred device is not available or out of stock, which will reduce customer loss to the company due to limited stock.