Hema Lakshmi Prasanna Sanaka

Webster University

CSDA 6010 DATA ANALYTICS PRACTICUM

SPRING 2024

FINAL PROJECT (Case-2)

JP WANG

USED SMARTPHONE PRICE PREDICTION &

CLASSIFICATION

**TABLE OF CONTENTS**

1. INTRODUCTION……………………………………………….7
2. BUSINESS GOAL……………………………………………….7
   1. ANALYTICAL APPROACH……………………………….7
3. DATA PREPROCESSING……………………………………….8
   1. ATTRIBUTE DEFINITIONS…………………………………8
   2. DATA EXPLORATION……………………………………….9
   3. HANDLING CATEGORICAL VARIABLES……………….10
   4. CHECKING FOR MISSING VALUES………………………10

3.4.1. HANDLING MISSING VALUES……………………...11

3.5 CHECKING FOR ZERO’S……………………………………12

3.6 ADDING DEVIATION COLUMN……………………………12

3.7 DESCRIPTIVE STATISTICS…………………………………12

3.8 ADDING RESALE COLUMN……………………………….13

3.9 PREDICTOR ANALYSIS…………………………………….14

3.10 CORRELATION FOR ALL NUMERICAL COLUMNS……35

4.DIMENSIONAL REDUCTION………………………………….35

5.DATA TRANSFORMATION……………………………………36

6.DATA PARTITIONING METHOD……………………………………36

7. REGRESSION MODEL SELECTION…………………………………37

7.1 MULTIPLE LINEAR REGRESSION MODEL………………………37

7.2 BACKWARD STEPWISE LINEAR REGRESSION MODEL………41

7.3 DECISION TREE…………………………………………………….43

7.4 COMPARISION OF ALL REGRESSION MODELS……………….45

8.CLASSIFIER MODEL SELECTION………………………………….46

8.1 LOGISTIC REGRESSION………………………………………….46

8.1.1 PREPARING THE DATA FOR LOGISTIC REGRESSION……46

8.1.2 EVALUATING THE MODEL PERFORMANCE………………47

8.1.3 FORWARD STEPWISE LOGISTIC REGRESSION……………49

8.1.4 EVALUATING MODEL PERFORMANCE…………………….50

8.2 CLASSIFICATION TREE………………………………………….51

8.2.1 EVALUATING THE MODEL PERFORMANCE OF CLASSIFICATION TREE……………………………………………………55

8.2.2 CLASSSIFICATION TREE BY USING THE C5.0()……………56

8.2.3 EVALUATING THE MODEL PERFORMANCE FOR C5.0()…………………………………………………………………………..57

9. BEST CLASSIFIER MODEL SELECTION……………………………58

10.K-MEANS CLUSTERING……………………………………………….59

10.1 CLASSIFICATION MODELLING FOR CLUSTERING………….66

11. HOLDOUT MODEL PERFORMANCE……………………………….68

12. CONCLUSION………………………………………………………….70

**EXECUTIVE SUMMARY**

* Simba Mobiles, a newly established store specializing in used mobile devices sourced from wholesale markets, aims to optimize business operations through machine learning algorithms. To achieve this, the company embarked on building two predictive models and conducting clustering analysis.
* The first predictive model, a backward stepwise linear regression, was selected as the optimal approach for predicting the normalized used price of mobile devices. Leveraging this model's capability to identify important features, Simba Mobiles can make informed pricing decisions to maximize profitability.
* Second predictive model, a decision tree classifier was chosen to categorize mobile devices as having either high or low resale value in the market. By considering various device features, this model provides valuable insights for inventory management and sales strategies.
* Additionally, clustering analysis was performed to discern patterns among mobile devices and group similar ones together. A subsequent classification clustering model aids in identifying which cluster a mobile device belongs to base on its features, facilitating efficient restocking decisions.
* Key performance metrics, including predicted normalized used price, predicted resale value, and predicted cluster, serve as vital indicators for business decision-making. By leveraging these metrics, Simba Mobiles gains insights into market trends and optimizes investment strategies, ultimately driving business growth and profitability.

**INTRODUCTION:**

This project aims to understand how different feature affecting the price of used mobile phones and how it will make impact on the decision making of consumer. Before buying the used mobile we will look around the many features such as screen size, ram, battery, internal memory, OS, device brand etc…., about that phone and then will decide to purchase or not purchase. By analysing all these factors, can be able to provide the insights that can help to make decisions to providers and consumers while selling or purchasing the used mobile phones.

**2.BUSINESS GOAL:**

Simba Mobiles aims to maximize profitability and minimize financial losses by implementing data-driven strategies in its used mobile device business. Using machine learning techniques, Simba Mobiles intends to predict the normalized used prices of used mobile devices and categorize them into high or low resale segments. This categorization will consider various device features, such as brand, model, age, condition, and technical specifications. Clustering analysis will be used to group mobile devices with similar features, allowing Simba Mobiles to identify common characteristics among high and low resale segments. By leveraging these insights, Simba Mobiles can make more informed decisions about pricing, inventory management, and marketing strategies to drive business success.

**2.1 ANALYTICAL APPROACH:**

First, the data will be pre-processed and undergo predictor analysis to understand the relationships between variables and identify patterns. After this, data transformation will prepare the dataset for modelling. The data will be partitioned into three separate sets for building models.

Regression models will be constructed to predict the estimated normalized resale price of used mobile devices. Classification models will help determine whether a used mobile device has a high or low resale value in the market, aiding Simba Mobiles in identifying which used mobiles are in high demand for resale. Clustering techniques will group mobiles with similar characteristics, allowing for deeper exploration of patterns within the data. This approach can reveal specific combinations of features that contribute to high or low resale values. Focusing on both high and low resale value mobiles allows Simba Mobiles to make more informed inventory decisions. The results from these analyses enable business owners to identify the types of mobiles that are key to running a successful business, ultimately leading to increased profits.

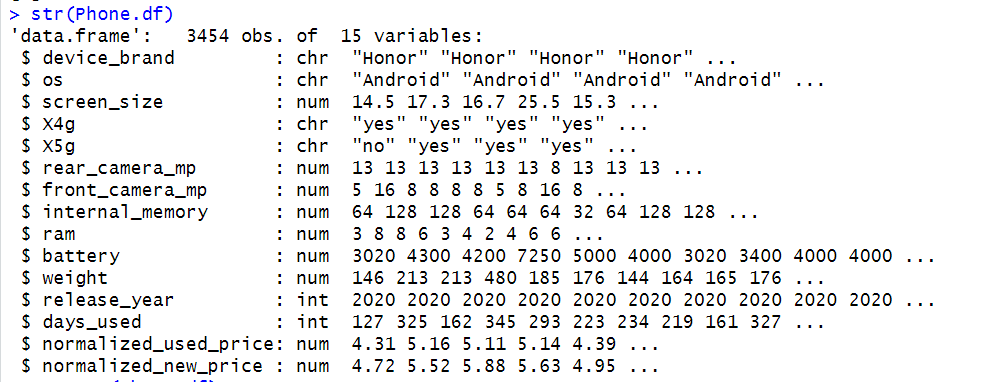
**3.DATA PREPROCESSING:**

**3.1 ATTRIBUTE DEFINITIONS:**

* **DEVICE BRAND:** This column represents the brand of the mobile**.**
* **OS:** This column represents the operating system running in the mobile which is android or iOS or others.
* **Screen size:** This column represents the length of the diagonal of the mobile screen which is measures in CM.
* **4G:** This column represents weather the mobile supports 4g service or not
* **5G:** This column represents weather the mobile supports 5g service or not.
* **Rear\_camera\_mp:** This column represents the resolution of the back camera which is measured in megapixels.
* **Front\_camera\_mp:** This column represents the resolution of the front camera which is measured in megapixels.
* **Internal-Memory:** This column represents the amount of built-in storage capacity in the mobile which is measured in gigabytes or terabytes.
* **RAM:** This column represents the random-access memory which is the volatile memory that the device uses for temporary storage while running the applications which is measured in gigabytes.
* **Battery:** This column represents the capacity of the mobile battery which is measured in milliampere per hour.
* **Weight:** This column represents the weight of the mobile based upon the brand which is measured in grams.
* **Released year:** This column represents the when the year mobile was released to the market.
* **Days\_Used:** This column represents the number of days mobile has been in use.
* **Normalized\_Used\_Price:** This column represents the adjusted or normalized prices of used phones, indicating a transformation from the original prices. This adjustment is likely intended to bring all prices within a specific range.
* **Normalized\_New\_Price:** This column represents the adjusted or normalized prices of new phones, indicating a transformation from the original prices. This adjustment is likely intended to bring all prices within a specific range, facilitating fair and meaningful comparisons.

**3.2 DATA EXPLORATION:**

In this dataset have total 3454 observations with 15 variables. All the data types in appropriate format no need to change any data types. 10 columns are in numeric and remaining 5 columns are categorical columns. Let’s have a look on the structure of the data then will get to know more insights about the data.



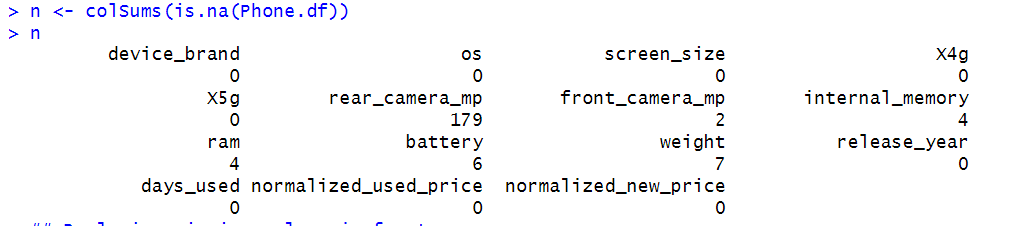
**Figure 3.2.1 Structure of the phones data**

**3.3 HANDLING CATEGORICAL VALUES:**

By examining the above screenshot, it is evident that the "char" datatype is present. If necessary, during modelling, it needs to be converted into factors or dummy variables based on the specific requirements of the models being used. Linear and logistic regression models automatically handle categorical columns during the modelling process. However, when employing K-nearest neighbours (KNN), it is necessary to convert categorical columns into dummy variables. On the other hand, decision trees handle categorical columns automatically without the need for additional conversions.

**3.4 CHECKING FOR MISSING VALUES:**

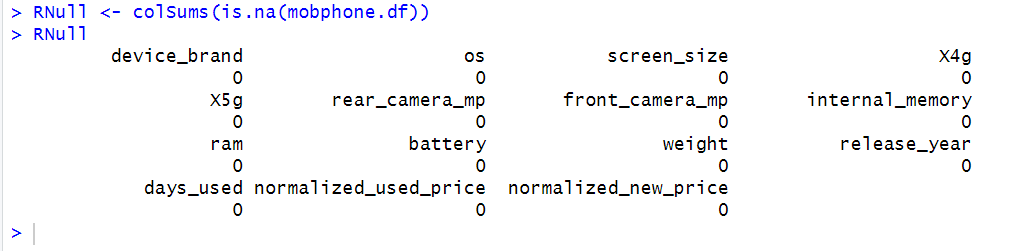
Checked for the null values in the data found total 202 missing values in different columns those are in rear\_camera\_mp we have 179 null values, in front\_camera\_mp we have 2 null values, in internal\_memory we have 4 null values, in ram we have 4 null values, in battery we have 6 null values, in weight we have 7 null values.



**Figure 3.2.2 Null values in column values**

**3.4.1 HANDLING MISSING VALUES:**

To address missing values in the data, traditional methods like replacing them with the mean or median are common, but these approaches can sometimes harm model performance. To avoid this, the k-nearest neighbours (kNN) imputation method was chosen, which relies on identifying the closest neighbours with similar features to fill in missing values. This required installing the VIM package. In this method, the imputed values are determined by examining mobile devices with similar characteristics, allowing for more accurate and contextually relevant replacements. This way, the missing values are replaced with values from similar mobiles, reducing the risk of introducing skewed or misleading data.

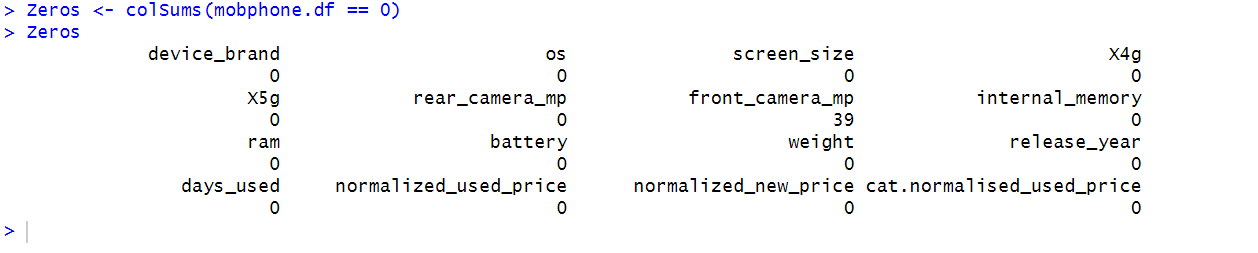


**Figure 3.2.3 After replacing the null values**

In the mobphone.df data frame, all missing values were replaced using the kNN imputation method. This approach is favoured because it ensures that the imputed values align with other mobile phones that share similar specifications, resulting in a more reliable and accurate dataset for analysis and modelling.

**3.5 CHECKING FOR ZERO’S:**

Checked for zeros in the data found total 39 zeros in front\_camera\_mp. Those are relevant zero’s because some mobiles do not have front camera in such condition consider as 0. So, there is no transformation required for those zeros in data.



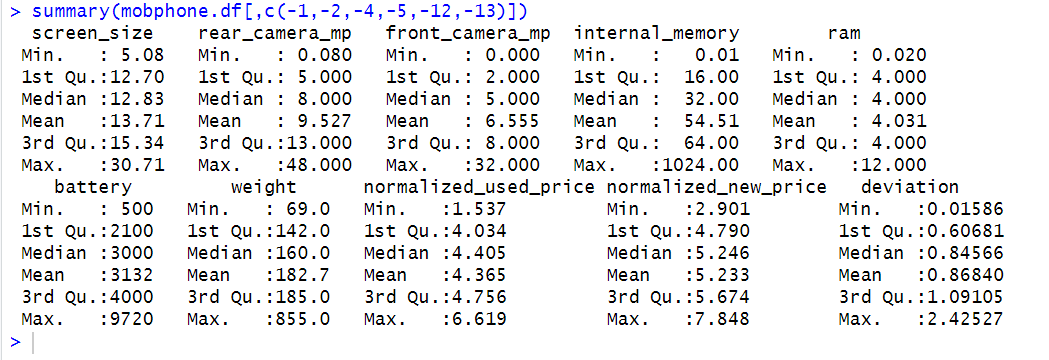
**Figure 3.2.4 Zero’s in column wise**

**3.6 Adding the Deviation column to the data:**

Create a Deviation column to capture the variance between the normalized\_new\_price and normalized\_used\_price. This addition will provide valuable insights, allowing for the observation of the extent of variation in mobile prices after being used for a certain period.

**3.7 DESCRIPTIVE STATISTICS:**

To get to know about the more insights about the numerical columns can have a look on the summary statistics of the data then will get to know about the distributions of the all the numerical variables and excludes all categorical column.

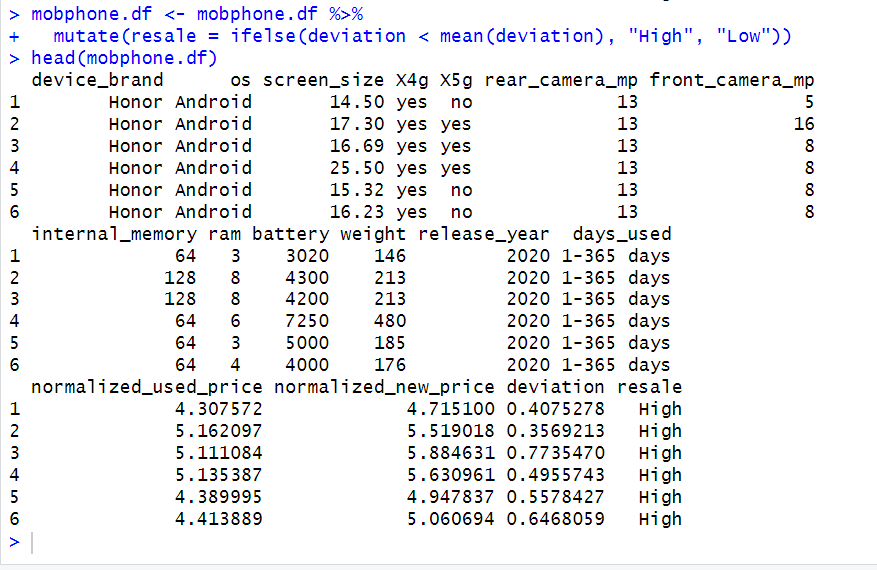


**Figure 3.2.5 Summary statistics of phone data**

By looking at figure 3.2.5 can be seen that the different variables have very different ranges of values. None of the variables have missing values. This summary provides an overview of the key statistics for each variable in the data offering insights into the range, central tendency, and distribution of the data. It can be useful for understanding the characteristics of the used mobile devices, facilitates in further analysis or modelling. Observing that mean of the internal\_memory is much larger than the median it is indicating right skew and mean of the weight is much larger than the median it is indicating that right skewed data.

**3.8 Adding Resale column:**

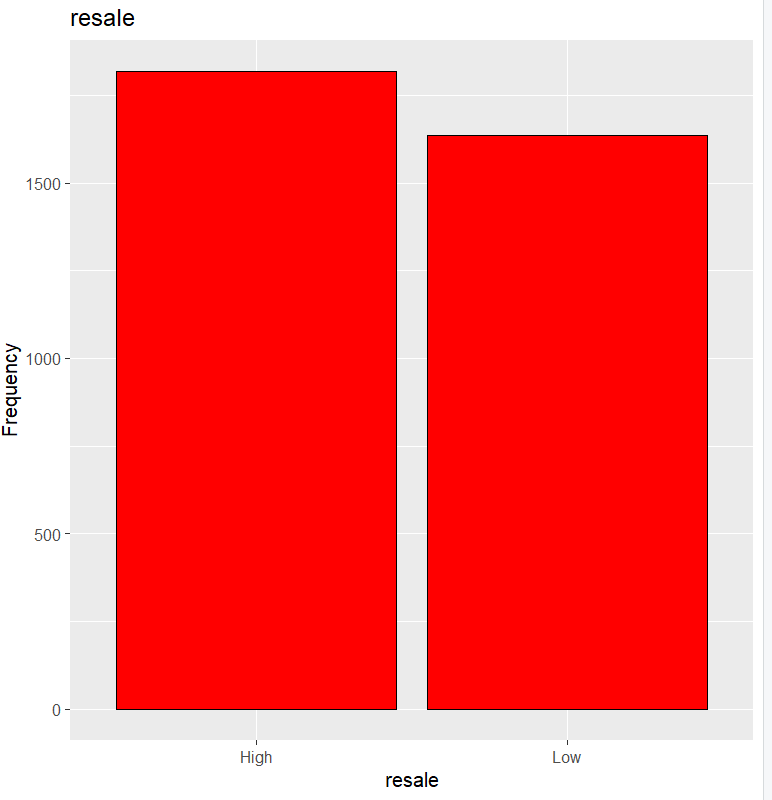
Creating a categorical column based on the mean of the deviation column is justified by the fact that it serves as a measure of central tendency and represents the average value across the entire dataset. This approach is generally less sensitive to outliers. Under this method, values above the mean in the deviation column are classified as Low Resale Value, indicating that these used mobiles have lower resale value due to their high price drop. Conversely, values below the mean are classified as High Resale Value, suggesting that these used mobiles experience a lower price drop, making them more appealing for resale in the market. This categorical column helps streamline the classification process for used mobile devices.



**Figure 3.2.6 Created a resale column**

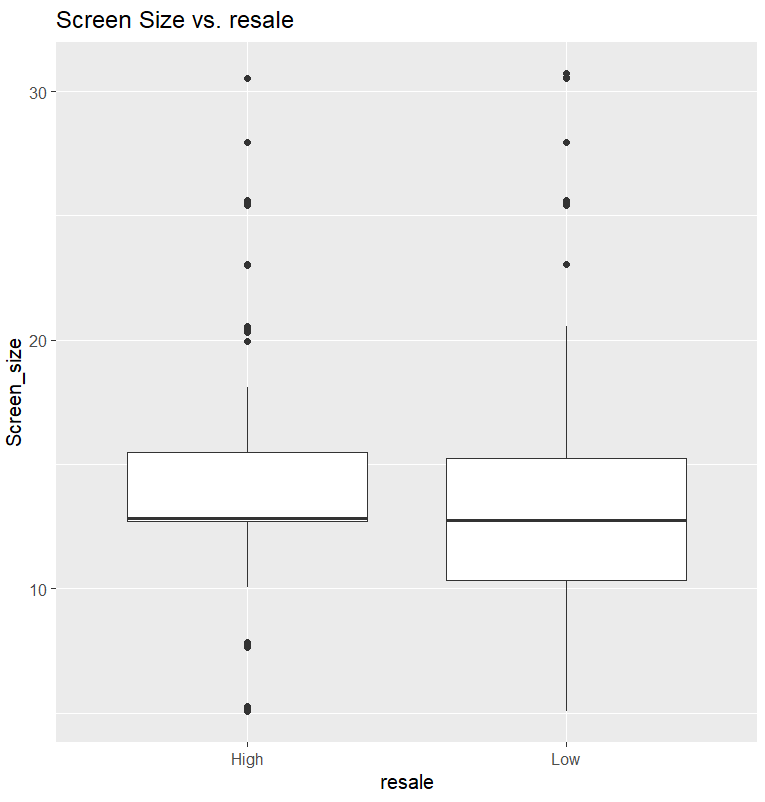
**3.9 PREDICTOR ANALYSIS:**

By doing the predictor analysis can get more details about the variables and relationship between the variables. First will have a look on the target variable how it is distributed.



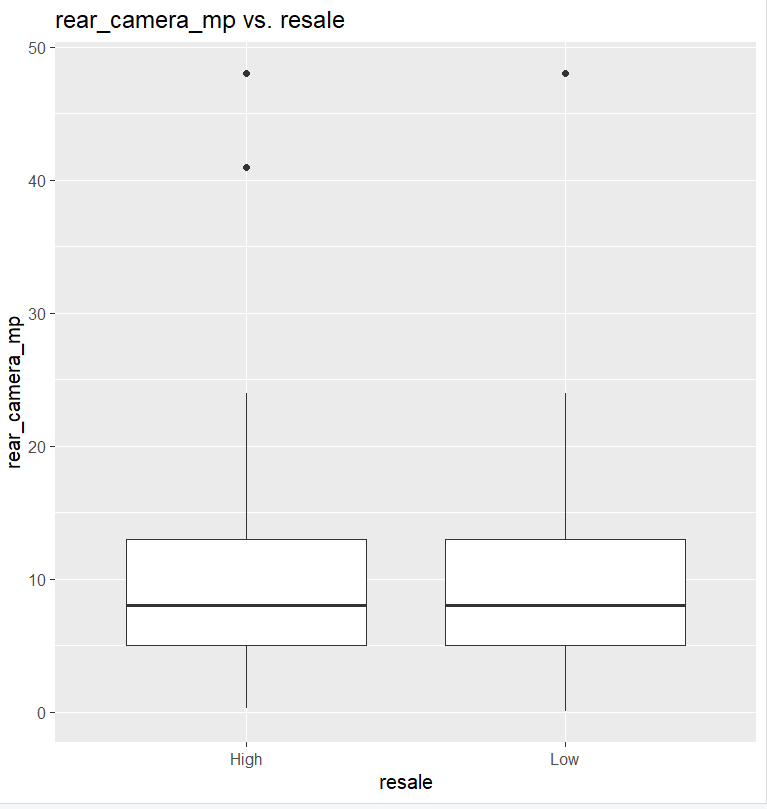
**Figure 3.9.1 Distribution of Resale**

As seen in Figure 3.9.1, used mobiles with low deviation are in greater demand in the resale market compared to those with high deviation. This suggests that most customers prefer purchasing used mobiles with lower deviation in the resale market. To increase profits, Simba Mobiles should focus on selling used mobiles with high deviation.



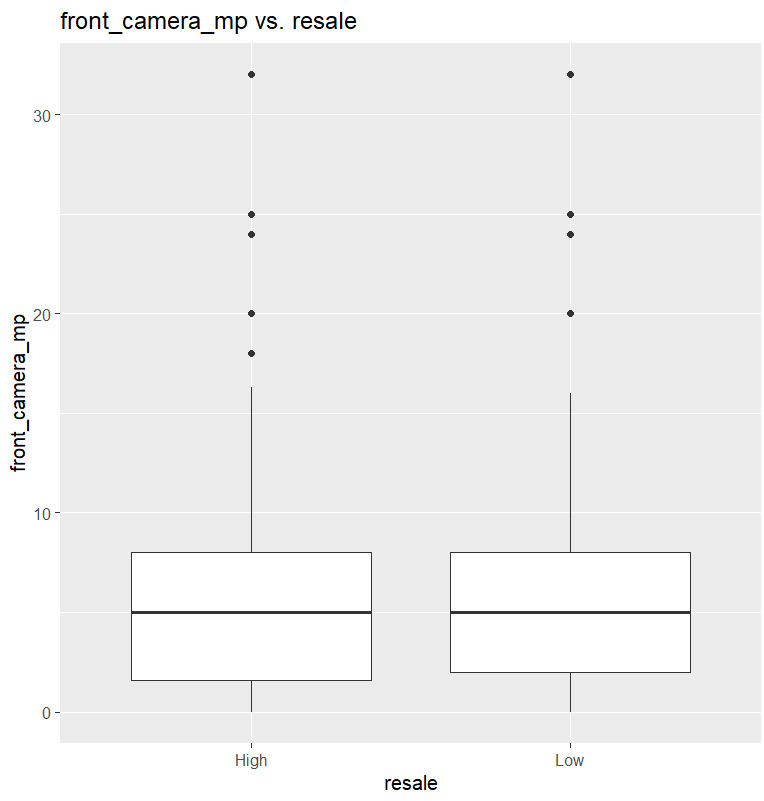
**Figure 3.9.2 Boxplot for screen size vs resale**

Figure 3.9.2 indicates that used mobiles with low deviation and screen sizes greater than 10 cm tend to have lower resale value in the market. In contrast, used mobiles with large screen sizes and high deviation are more likely to have high resale value. To maximize profits, Simba Mobiles should consider the impact of screen size and deviation when making decisions about which used mobiles to focus on in the resale market.



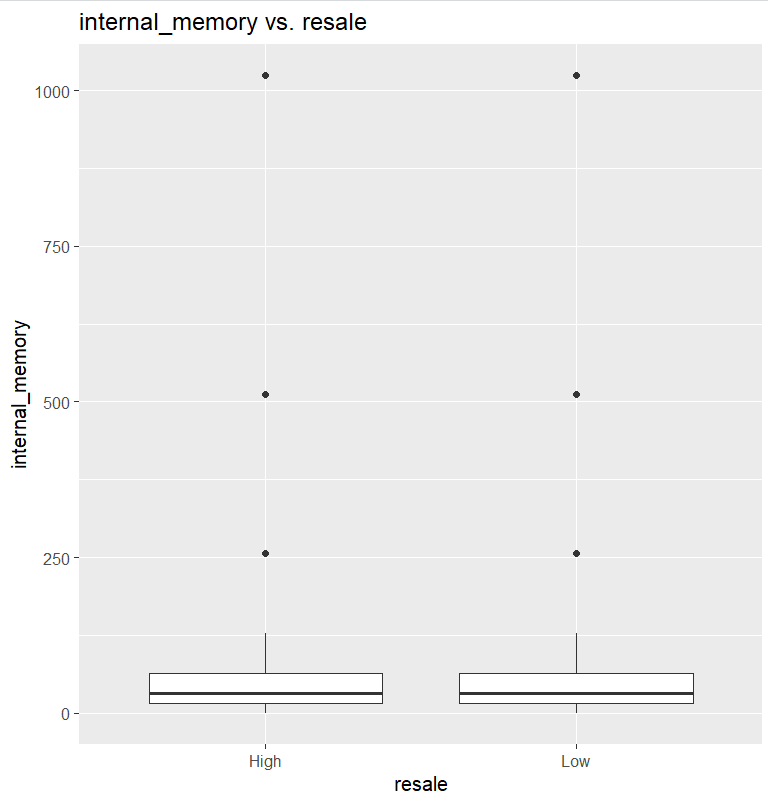
**Figure 3.9.3 Boxplot for rear\_camera\_mp vs Resale**

Figure 3.9.3 shows that mobile devices with higher-quality rear cameras tend to be more in demand in the resale market, depending on the overall condition of the device. Notably, used mobiles with both high and low deviation in rear\_camera\_mp, regardless of whether they have high or low pixel count, and those with low deviation in rear\_camera\_mp, whether high or low in pixels, tend to have similar resale values. This suggests that the resale market considers other specifications and the general condition of the mobile as key factors in resale value. To maximize profits, Simba Mobiles can benefit from purchasing used mobiles with varying rear camera qualities, whether high or low megapixels, when sourcing from the wholesale market. This approach allows the business to offer a wider range of options to customers, appealing to various preferences and potentially boosting sales.

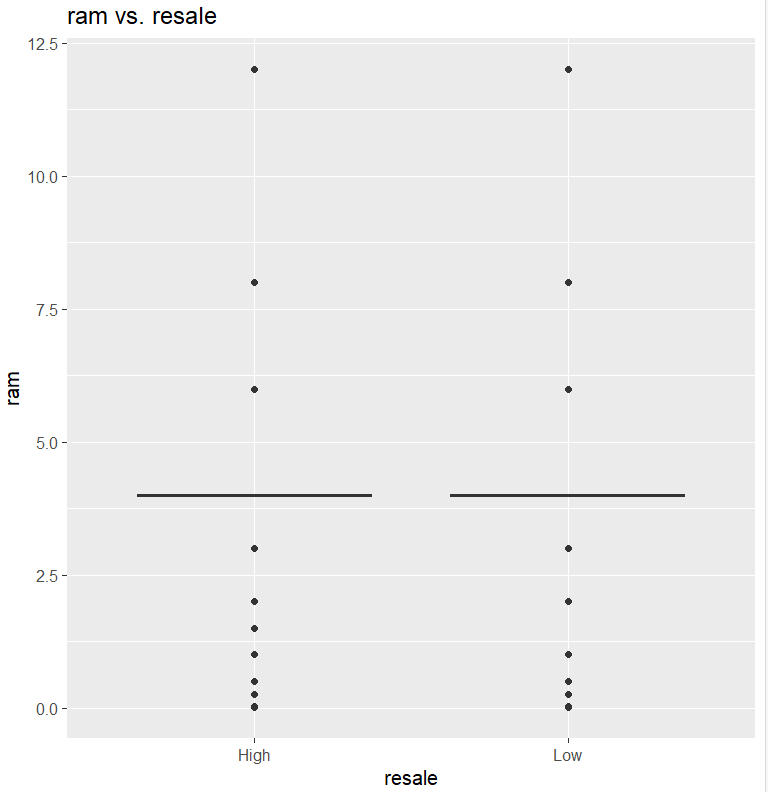


**Figure 3.9.4 Boxplot Front\_camers\_mp vs Resale**

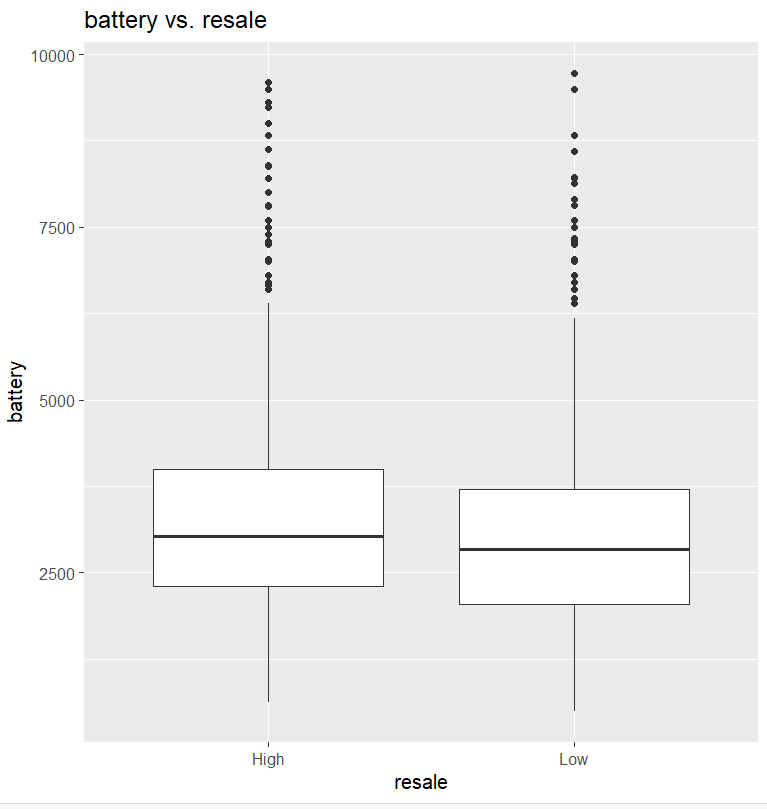
Figure 3.9.4 suggests that mobile devices with higher-quality front cameras tend to be in greater demand in the resale market, depending on the device's condition. Interestingly, used mobiles with high deviation in front\_camera\_mp, whether they have low or high pixel count, and those with low deviation in front\_camera\_mp, regardless of pixel count, seem to have similar resale values. This indicates that other factors, such as overall mobile specifications and condition, play a significant role in determining resale value. To maximize profits, Simba Mobiles should consider purchasing used mobiles with various front camera qualities, both high and low megapixels, when sourcing from the wholesale market. This strategy allows the business to cater to a broader customer base, with diverse needs and preferences, ultimately leading to higher sales and increased profitability.



**Figure 3.9.5 Boxplot for internal\_memory vs Resale**

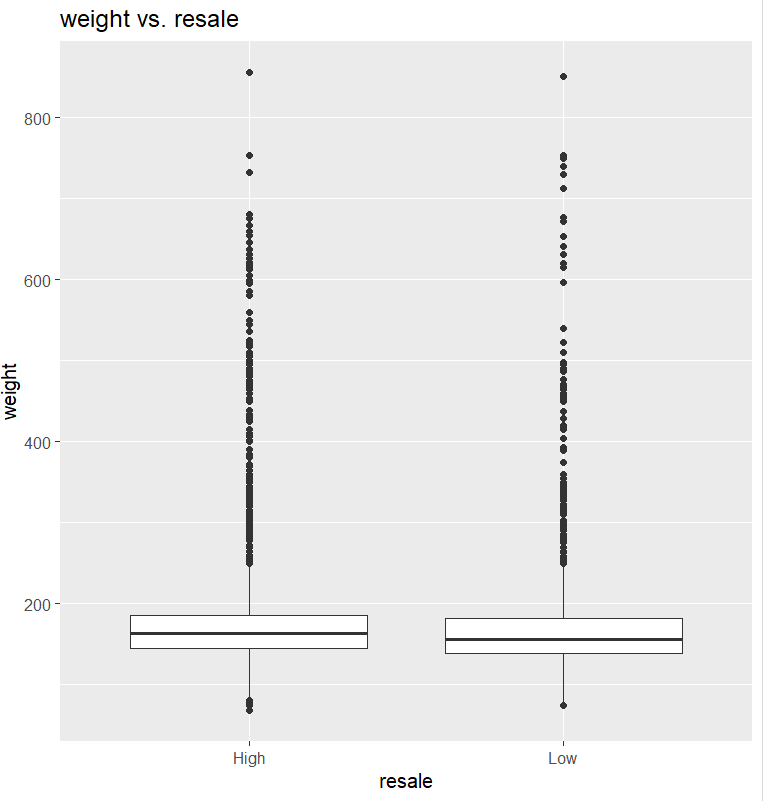
Figure 3.9.5 indicates that the resale value of used mobile devices is similar regardless of whether they have high or low internal memory deviation, or whether they have a large or small internal memory capacity. This suggests that in the resale market, the condition of the mobile device plays a more critical role in determining its resale value than the internal memory specifications.

**Figure 3.9.6 Boxplot for ram vs Resale**

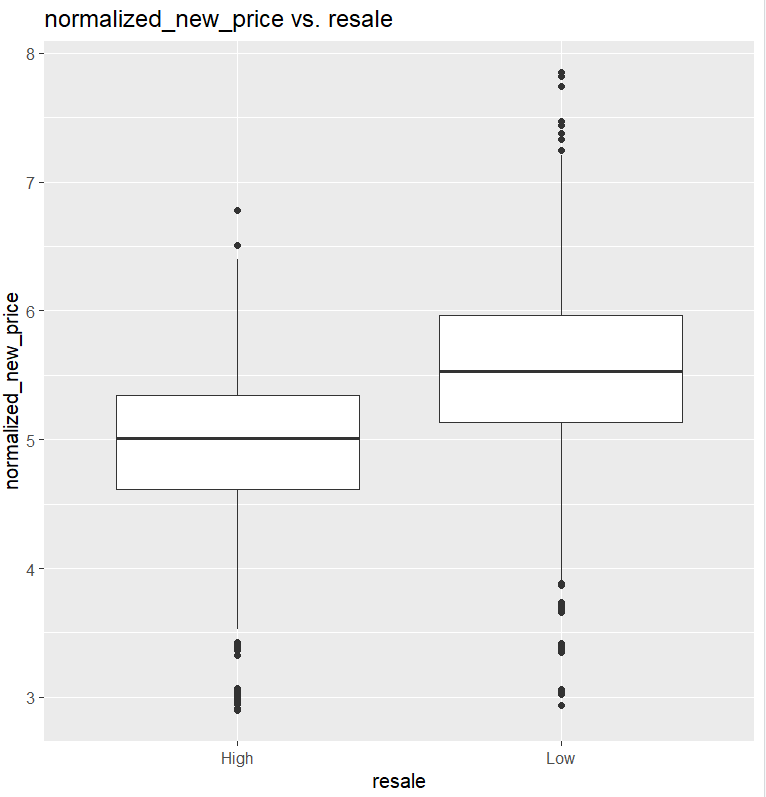
Figure 3.9.6 shows that the resale value of used mobile devices is similar regardless of whether the RAM has high or low deviation, and irrespective of the size of the RAM. This indicates that the condition of the mobile device is a more significant factor in determining its resale value than the variation or size of its RAM.

**Figure 3.9.7 Box plot for battery vs resale**

Figure 3.9.7 reveals that used mobiles with high deviation batteries of more than 5000 mAh have the highest resale value, compared to those with low deviation batteries of less than 5000 mAh. This suggests that used mobiles with high-capacity batteries and consistent performance are more valuable in the resale market. To maximize profits, Simba Mobiles should focus on acquiring used mobiles with high-capacity batteries that exhibit low deviation when sourcing from the wholesale market. This approach ensures they are selecting devices with the highest resale potential.

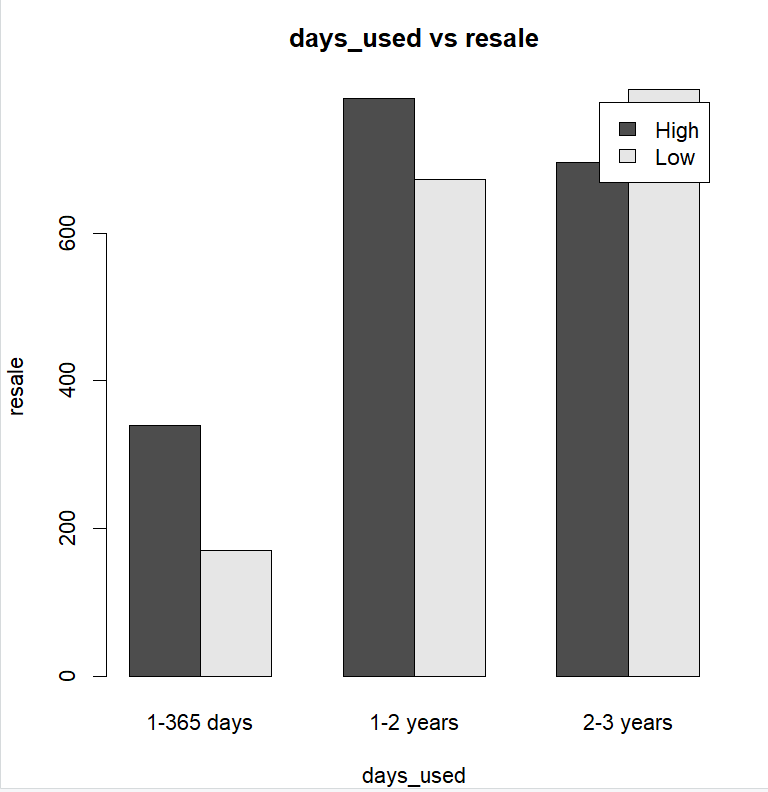


**Figure 3.9.8 Boxplot for weight Vs Resale**

Figure 3.9.8 indicates that the resale value of used mobiles is generally consistent regardless of whether they have low or high deviation of the used mobiles weight whether heavy or light. This suggests that the weight of the used mobile, and its variation, doesn't significantly impact its resale value in the market. 

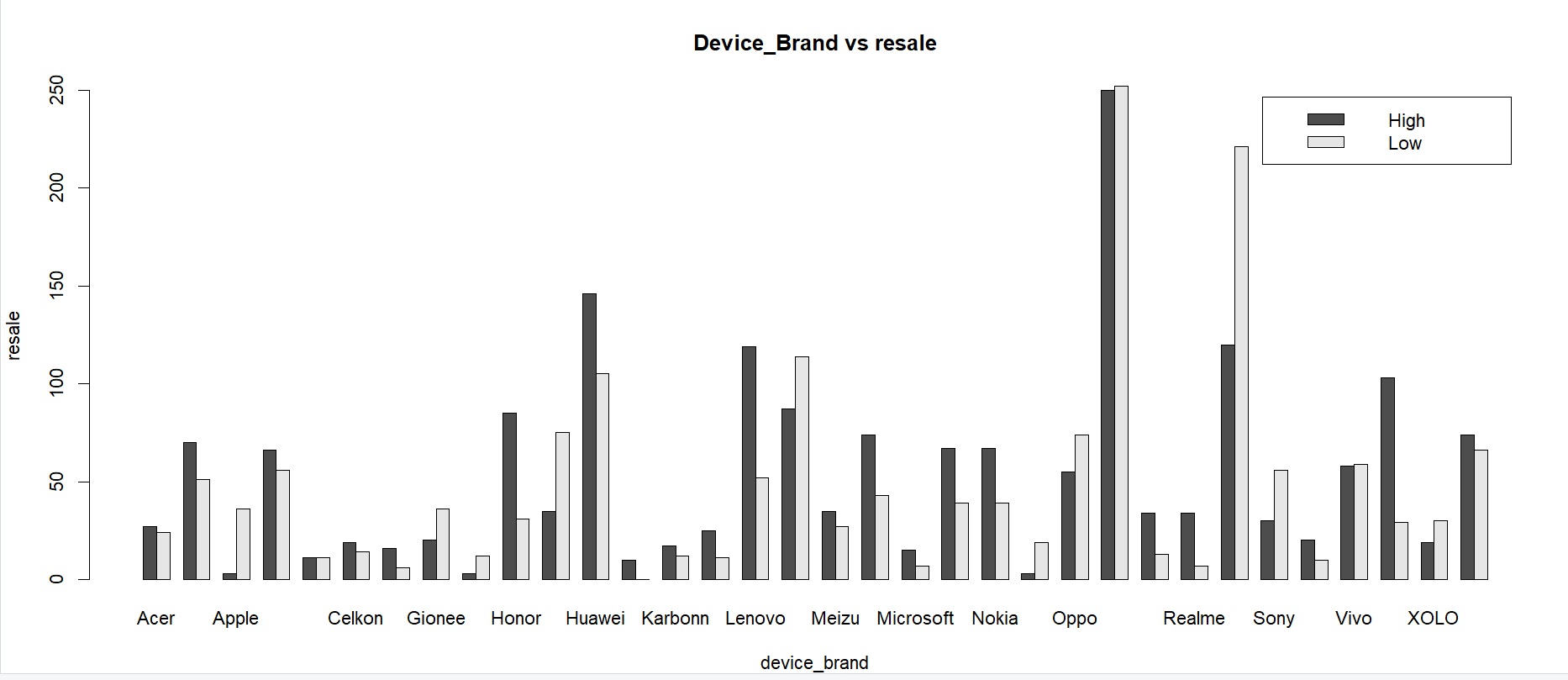
**Figure 3.9.9 Boxplot for normalized\_new\_price vs Resale**

Figure 3.9.9 shows that new mobiles with low deviation tend to have high resale value compared to those with high deviation in the resale market. To maximize profits, Simba Mobiles should focus on acquiring new mobiles with lower deviation when purchasing from the wholesale market, as these tend to yield greater resale value.



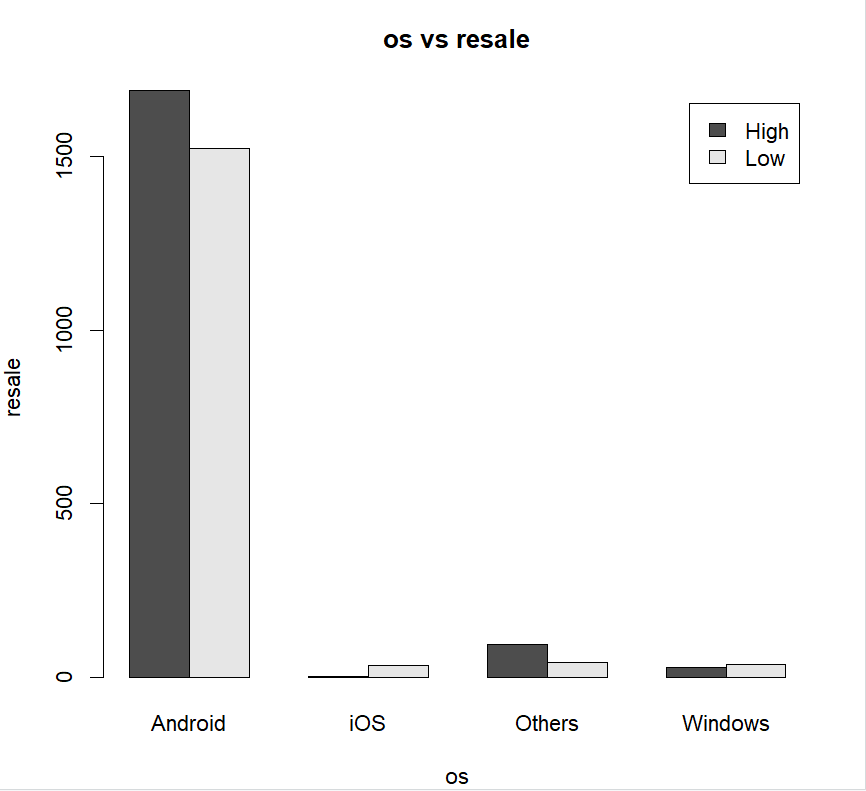
**Figure 3.9.10 Bar plot for Days\_used and Resale**

By looking at the above Figure 3.9.10, it can be seen that low deviation used mobiles have the highest resale value in the resale market within 1 year and 1-2 years used. It is also observed that high deviation used mobiles have the lowest resale value in the market within 2-3 years used mobiles. Therefore, to maximize profits, Simba mobiles need to focus on purchasing mobiles with fewer days used from the wholesale market.

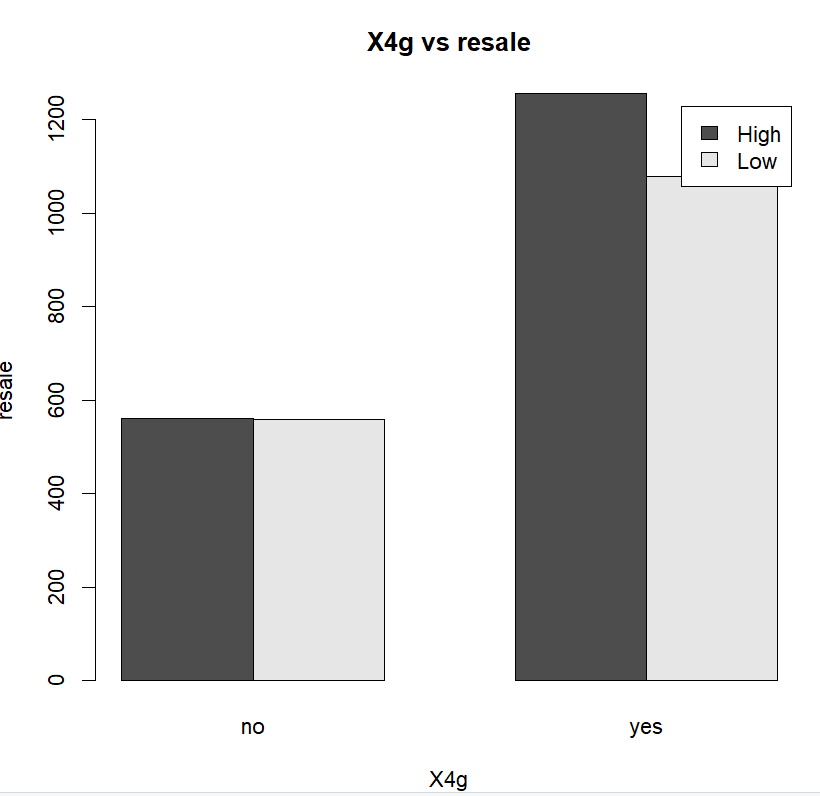


**Figure 3.9.11 Barplot of resale used price by device brand**

By looking at figure 3.8.11 can be seen that that the resale of apple and Huawei brands are the highest in the resale market. Remaining brands and Panasonic brands having low resale value is high when compared with the high resale value. So, to get more profits Simba mobiles need to focus on the high resale brands instead of focusing on the low resale brands.

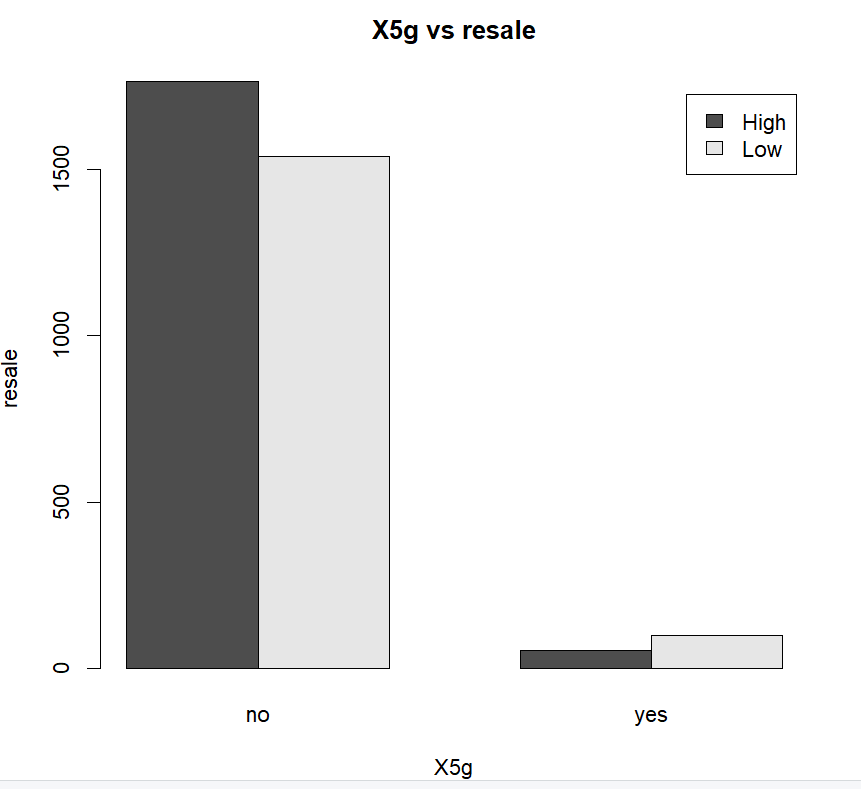


**Figure 3.9.12 Bar plot of resale used price by OS**

By looking at figure 3.9.12, can be seen that Android OS has the highest resale value compared to all other operating systems. Android exhibits both high and low resale values in the resale market. So, it can be clearly stated that Android has the highest resale value in the market when compared with all other OS in the resale market. To maximize profits, there is a need to focus on Android mobile devices in the resale market. However, iOS has high deviation with a low resale value in the resale market.

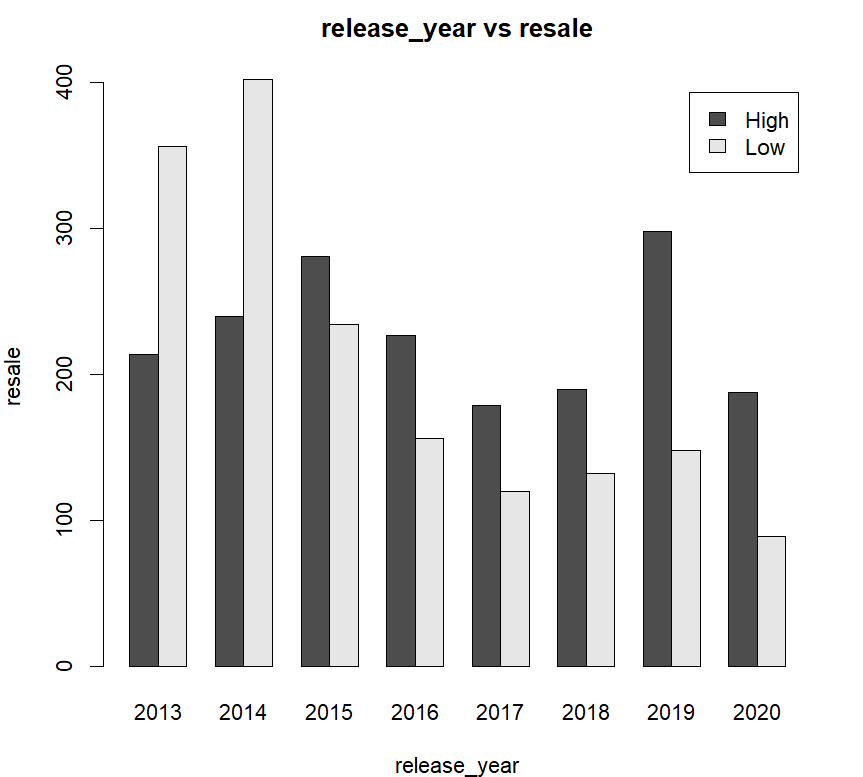
**Figure 3.9.13 Barplot of resale by X4g**

By looking at the above Figure 3.9.13, can be seen that devices which support 4G have the highest resale value among used mobile devices in the resale market when compared with others with low deviation. To maximize profits on mobile devices in resale, there is a need to focus on 4G-enabled mobile device.



**Figure 3.9.14 Barplot of resale vs X5g**

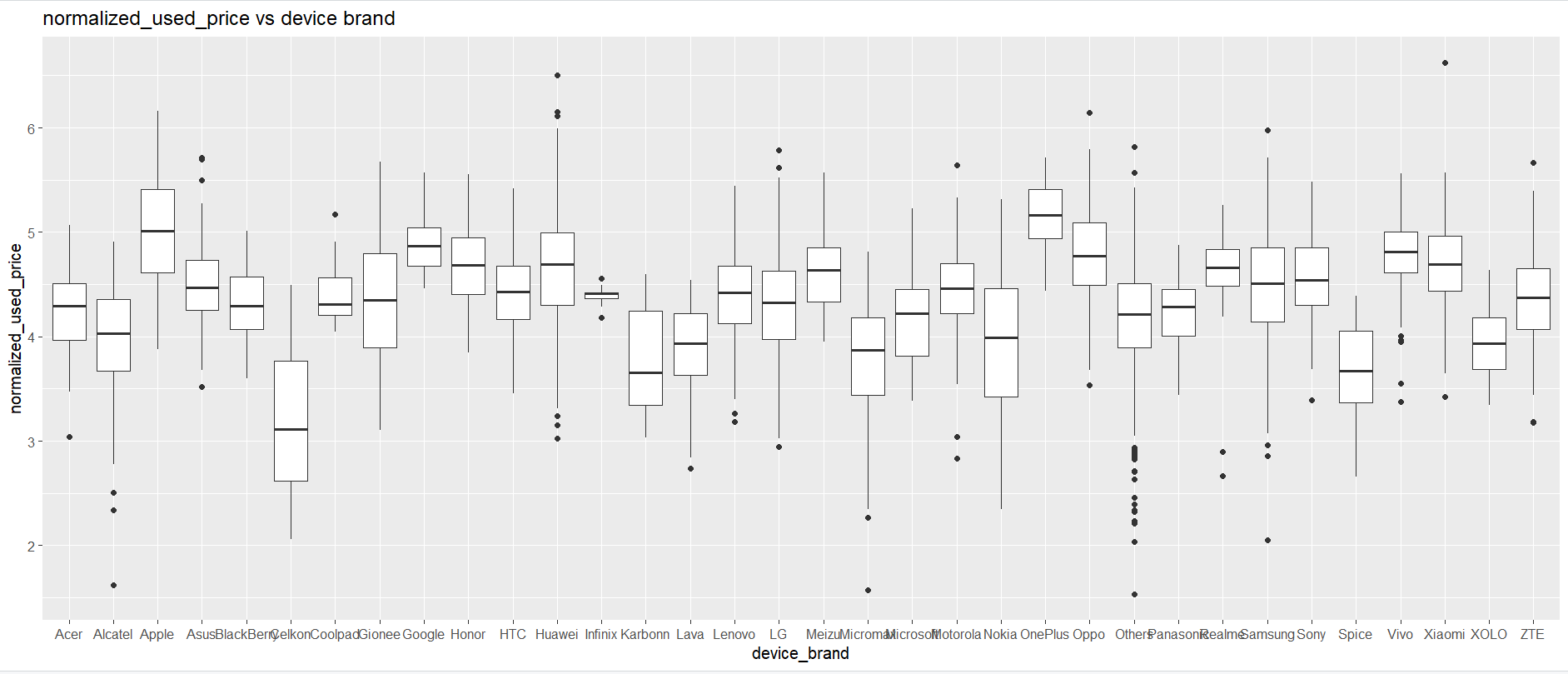
By looking at the above Figure 3.9.14, can be seen that mobile devices which support 5G service have the lowest resale value with high deviation compared to other mobiles that do not support 5G service in the resale market. So, to maximize profits, Simba mobiles need to focus on 5G service mobile devices with low deviation in the resale market.



**Figure 3.9.15 Barplot of Resale by release\_year**

By looking at the above Figure 3.9.15, it can be seen that mobiles used in recent years have a high resale value in the market compared to previous years. Observing that mobiles released in 2013 and 2014 have the lowest resale value when compared to all other years. It can be clearly stated that mobiles used in recent years have the highest demand in the resale market compared to previous years. So, to maximize profits, Simba mobiles need to focus on mobiles released in recent years in resale while purchasing from the wholesale market.

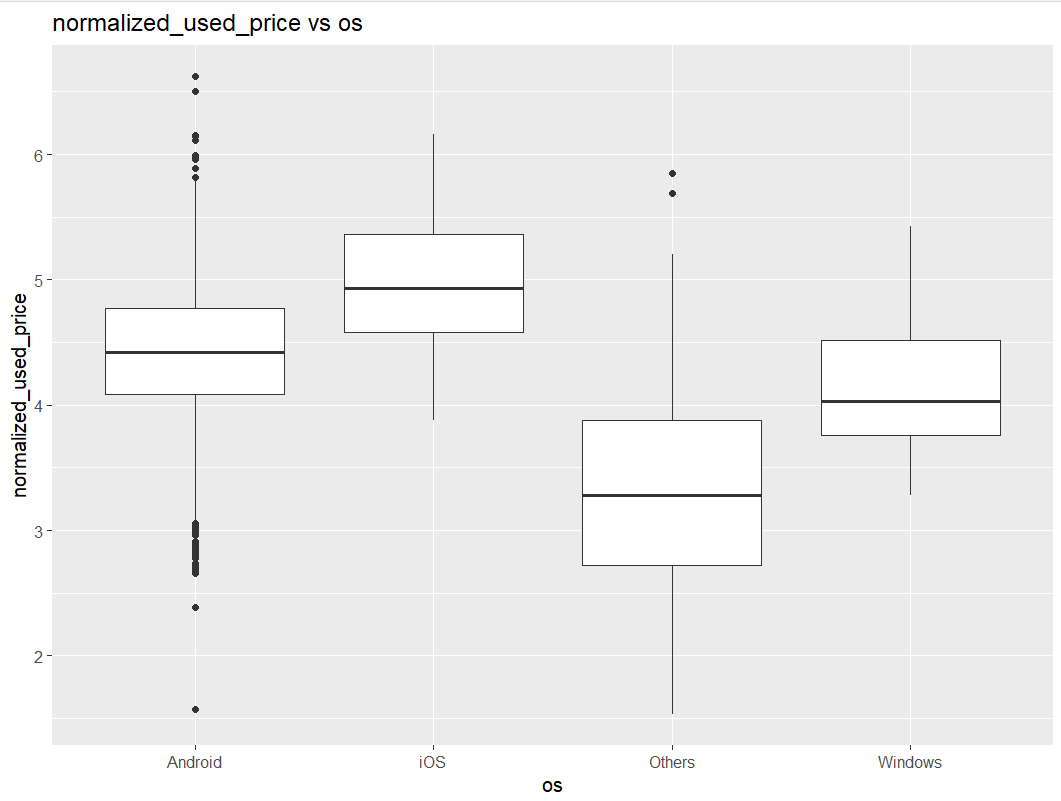
**Box plot of Device brand and Normalised used price:**

****

**Figure 3.9.16 Box plot of Device brand and Normalised used price**

By looking at the above figure 3.9.16 can be seen that apple brand has the highest normalised used price which is greater than 6, Hawaii brand has the second highest normalised used price which is equal to 6. Celkon has the lowest normalised used price among all brands.

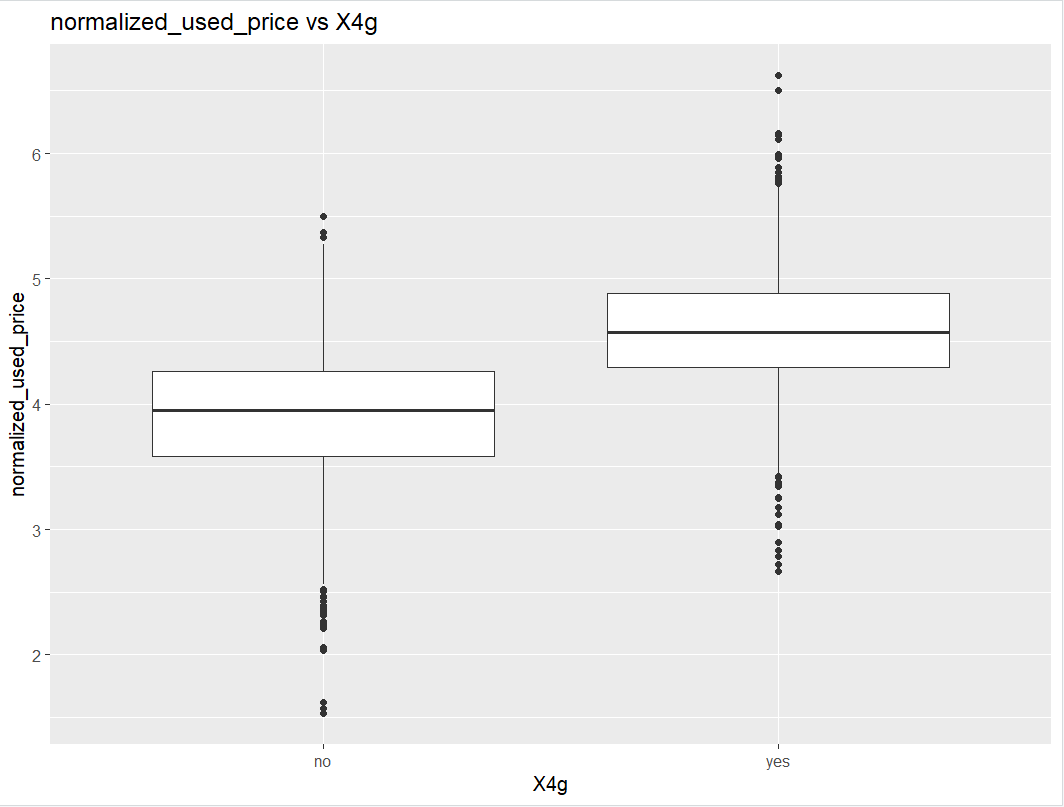
**Box plot of OS and Normalised used price:**

****

**Figure 3.9.17 Box plot of OS and Normalised used price**

By looking at above figure 3.9.17 can be seen that iOS and android has the highest normalised among all OS and others OS has the lowest normalised used price.

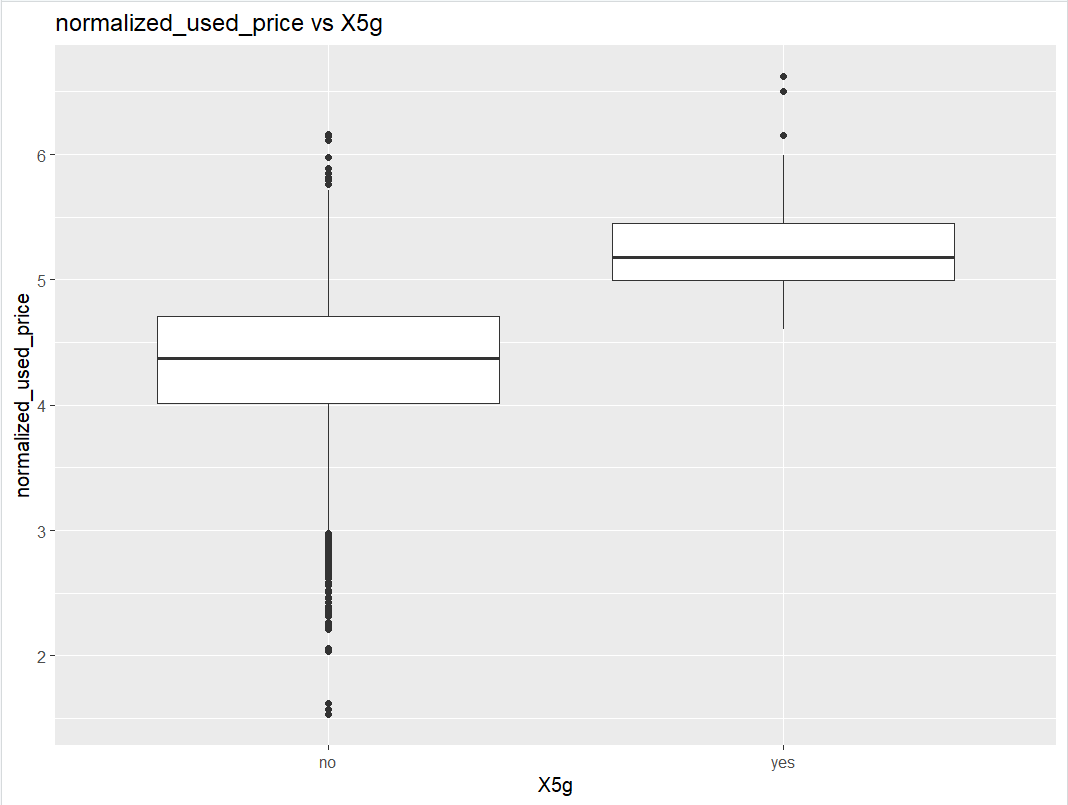
**Box plot of X4g and Normalised used price:**

****

**Figure 3.9.18 Box plot of X4g and Normalised used price**

By looking at the above figure 3.9.18 can be seen that most of the used mobile’s support 4g services to get more profits Simba mobiles need to focus on more 4g services used mobile devices while purchasing from the wholesale market.

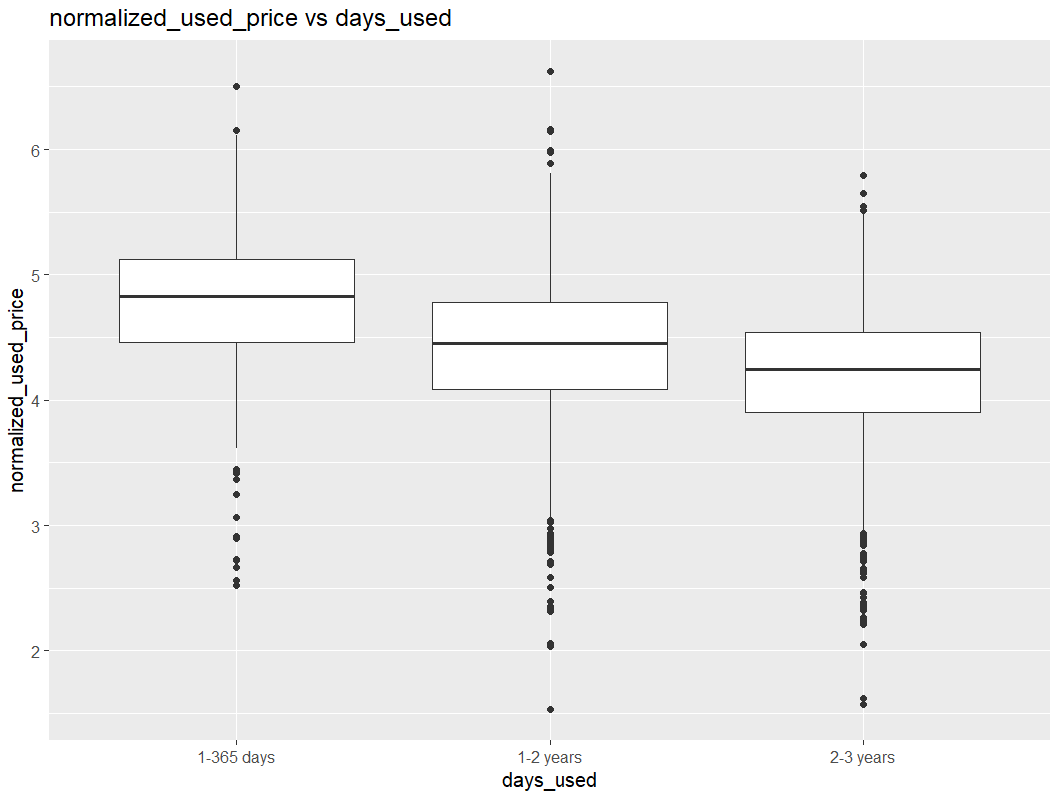
**Box plot of X5g and Normalised used price:**

****

**Figure 3.9.19 Box plot of X5g and Normalised used price**

By looking at the above figure 3.9.19 can be seen that used mobiles devices which supports 5g service has the highest used price when compared with which used mobiles not supporting the 5g service in the resale market. So, Simba mobiles need to focus on the 5g services used mobiles while purchasing from the wholesale market.

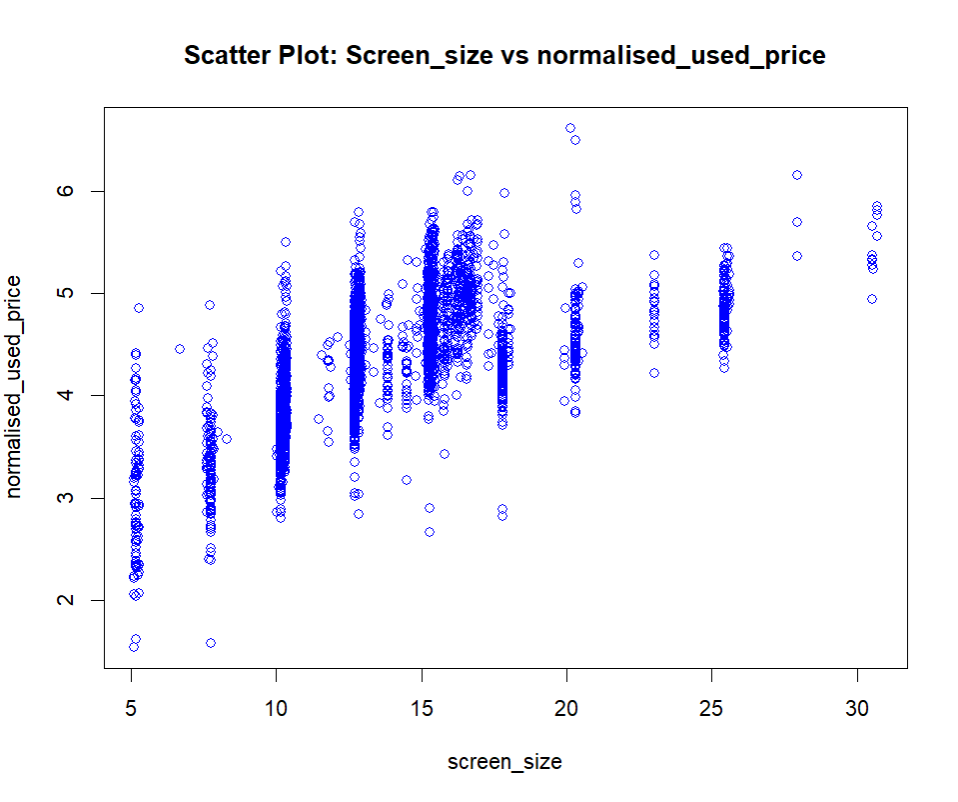
**Box plot of days used days and normalised used price:**



**Figure 3.9.20 Box plot of used days and normalised used price**

By looking at above figure 3.9.20 can be seen that less days used mobile devices has the highest normalised used price and more days used mobile devices has the lowest normalised price.

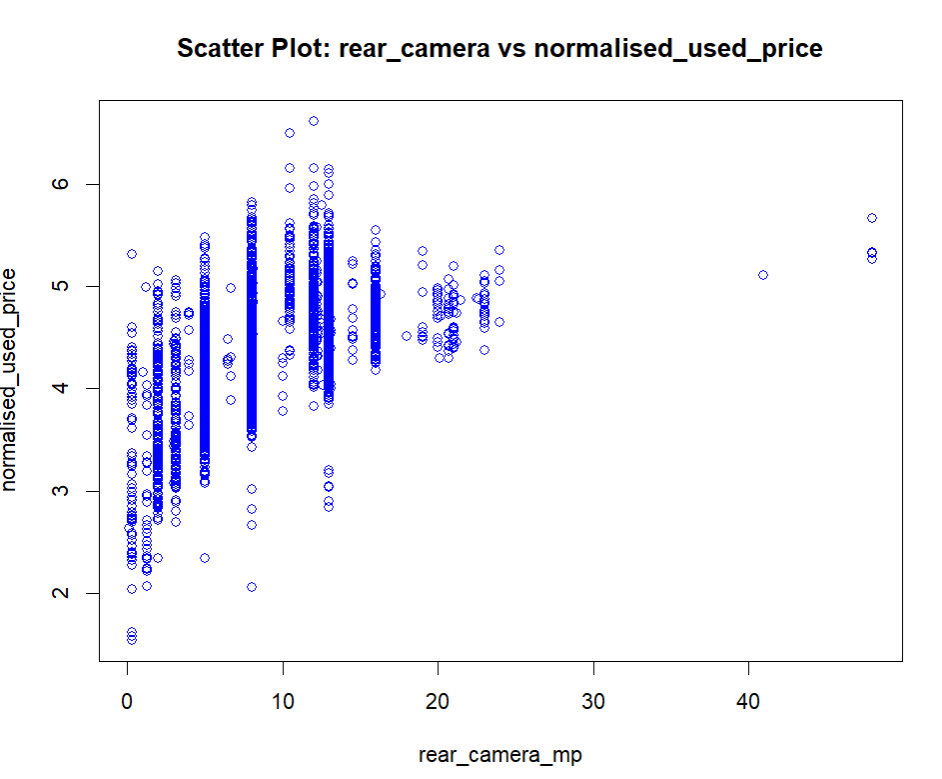
**Scatter plot of Screen size and normalised used price:**

****

**Figure 3.9.21 Scatter plot of Screen size and normalised used price**

By looking at above figure 3.9.21 can be seen that large screen used mobile devices has the high normalised used price and small size used mobile devices has the low normalised used price. So, to get more profits Simba mobiles need to focus on large screen size used mobiles while purchasing from the wholesale market.

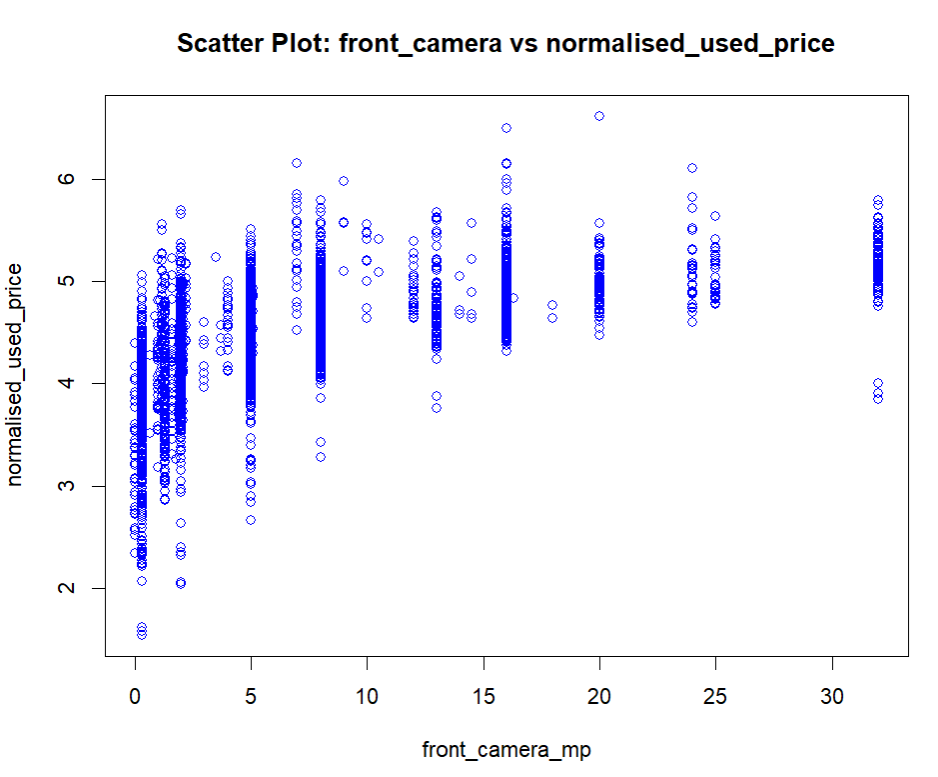
**Scatter plot of rear\_camera\_mp and normalised used price:**

****

**Figure 3.9.22 Scatter plot of rear\_camera\_mp and normalised used price**

By looking at the above figure 3.9.22 can be seen that high rear camera pixels used mobile devices has the high normalised used price and low rear camera pixels used mobile devices has the low normalised used price. Observed that some mobiles with high rear camera pixels have the low used price because of the condition of the camera. So, to get more simba mobiles need to focus on the good condition rear camera pixels used mobile devices while purchasing from the wholesale market.

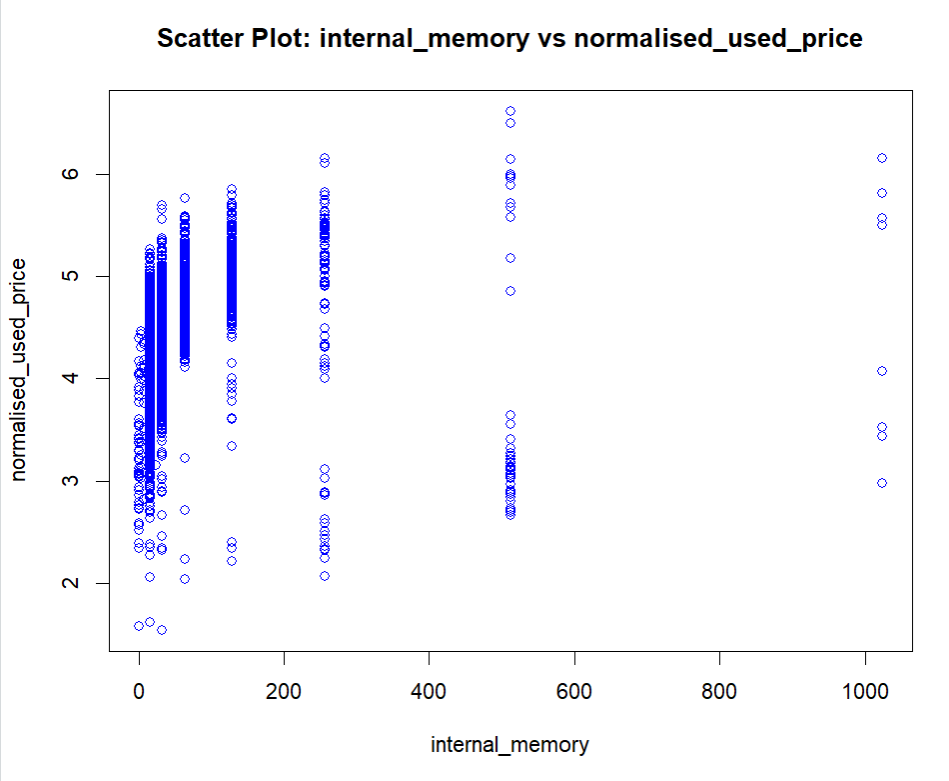
**Scatter plot of front\_camera\_mp and normalised\_used\_price:**

****

**Figure 3.9.23 Scatter plot of front\_camera\_mp and normalised\_used\_price**

By looking at above figure 3.9.23 can be seen that high front\_camera\_mp used mobile devices has the high normalised used price and low front\_camera\_mp used mobiles has the low normalised used price. So, to get more profits simba mobiles need to focus on high front camera used mobile devices.

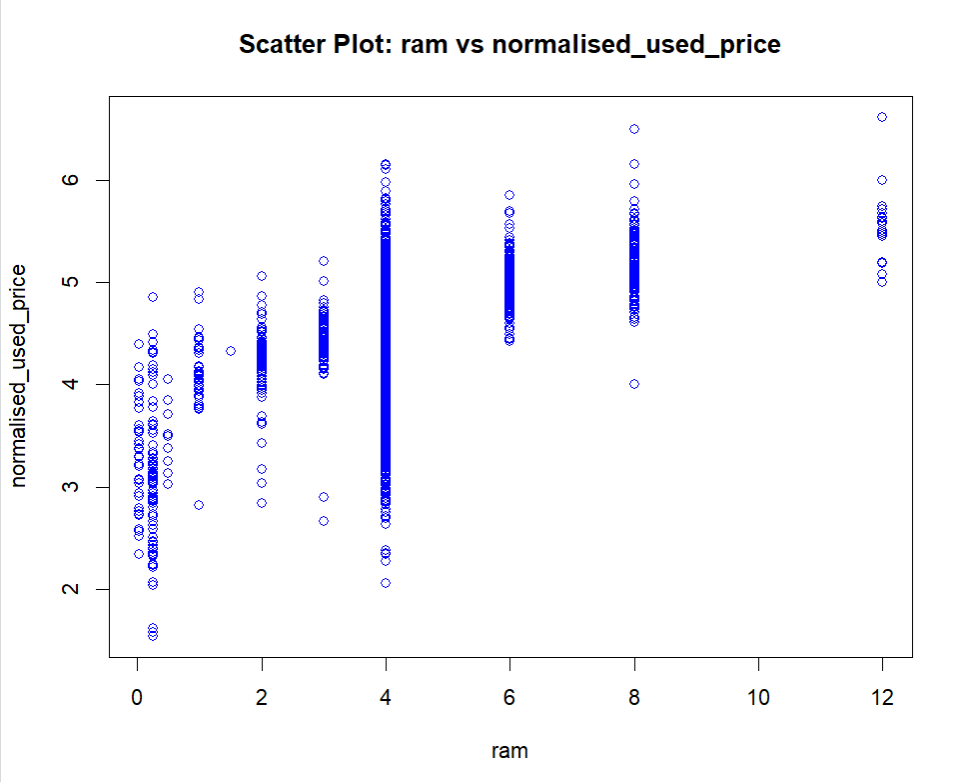
**Scatter plot of Interna\_memory and normalised\_used\_price:**

****

**Figure 3.9.24 Scatter plot of internal\_memory and normalised\_used\_price**

By looking at the above figure 3.9.24 can be seen that high internal used mobile devices have the high normalised used price and the used mobile devices with low internal memory have the lowest normalised used price. Observed that some mobiles with high internal memory used mobiles have the low used price may be the condition of the mobile. So, to get more profits Simba mobiles need to focus on the high internal memory devices while purchasing from the wholesale market.

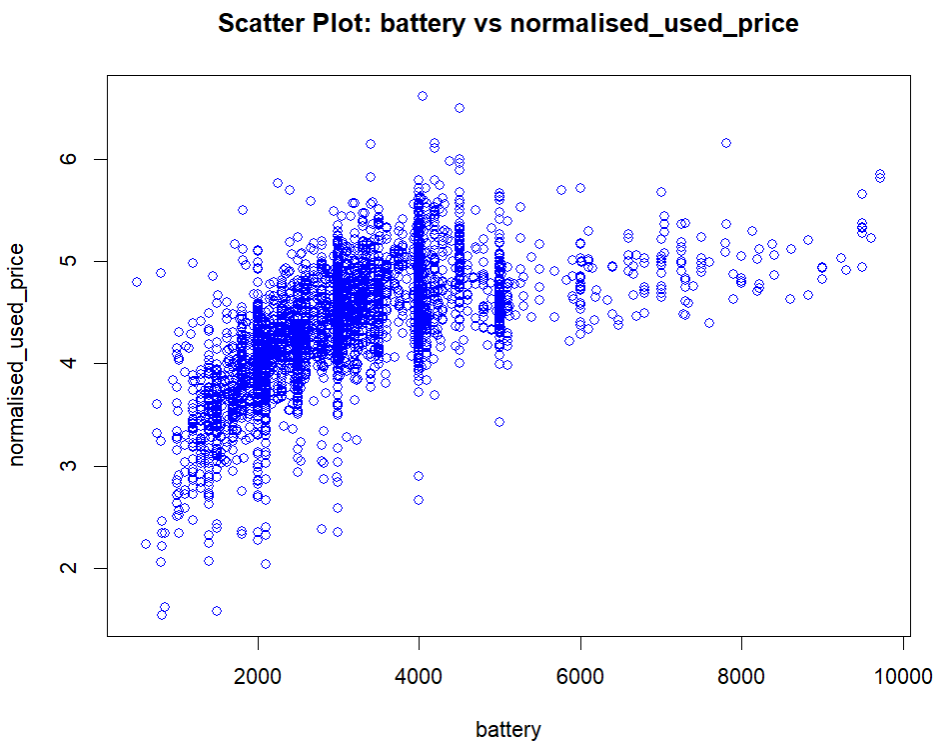
**Scatter Plot of ram and normalised\_used\_price:**



**Figure 3.9.25 Scatter plot of ram and normalised\_used\_price**

By looking at above figure 3.9.25 can be seen that high ram used mobile devices have the high normalised used price and low ram used mobile devices have the low normalised used price. To get more profits Simba mobiles need to focus on the high ram used mobile devices while purchasing from the wholesale market.

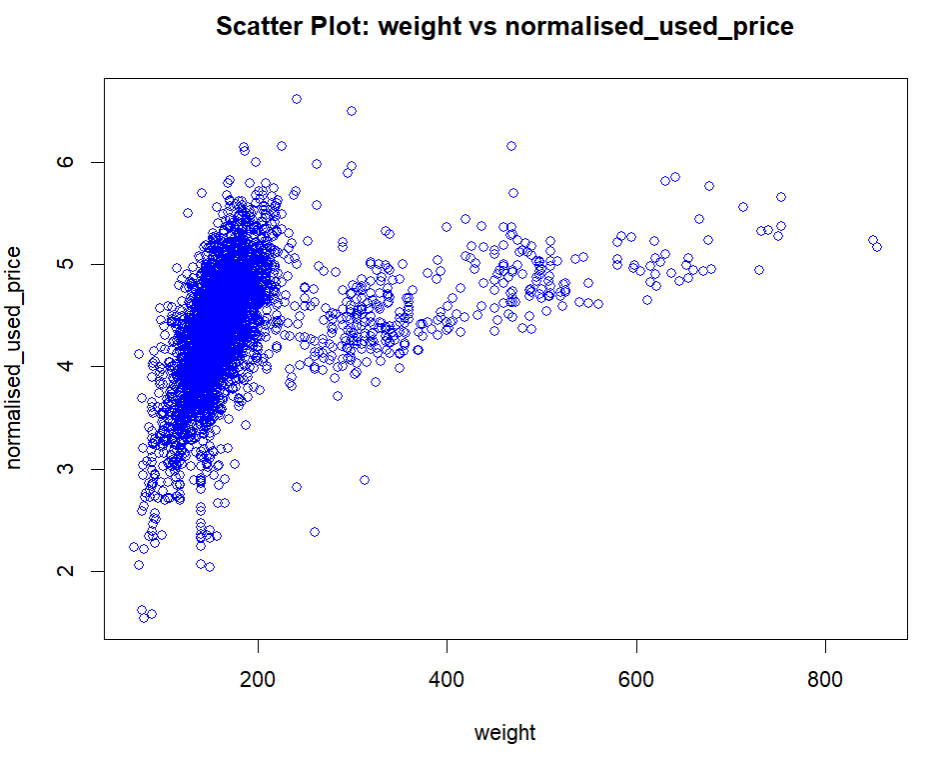
**Scatter plot of Battery and normalised\_used\_price:**

****

**Figure 3.9.26 Scatter Plot of Battery and Normalised\_used\_price**

By looking at the above figure 3.9.26 can be seen that high capacity of the battery more than the normalised used price is also more. Low-capacity battery used mobile devices has the low normalised used price. So, to get the more profits simba mobiles need to focus on the high-capacity battery used mobile devices while purchase from the wholesale market.

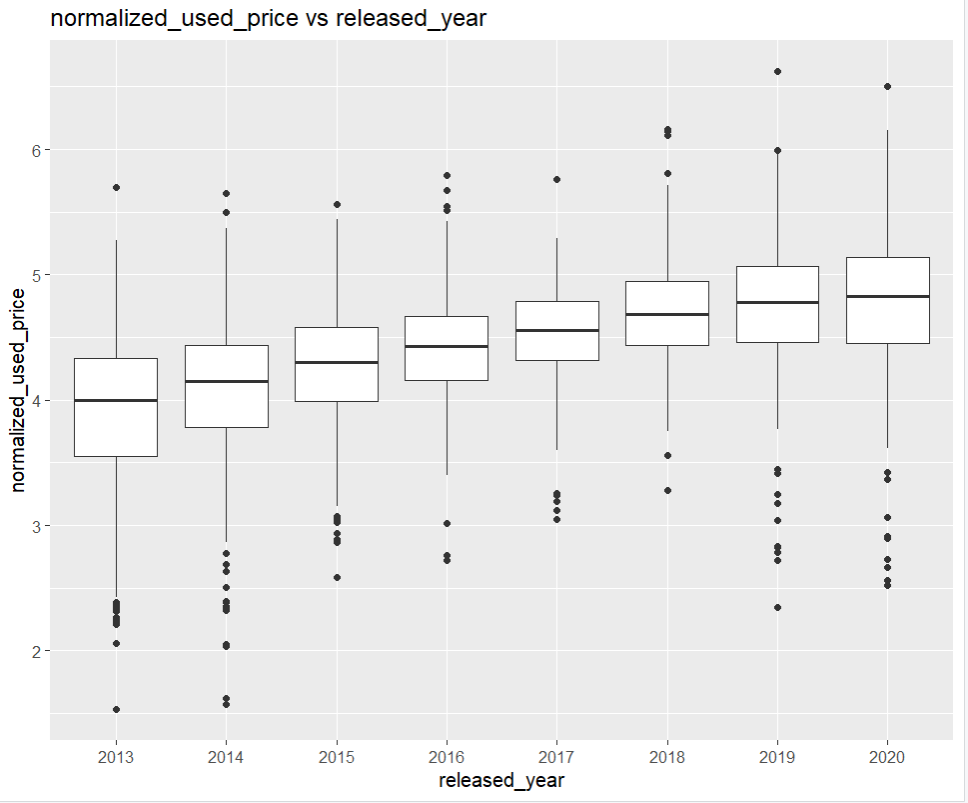
**Scatter plot of weight and normalised\_used\_price:**

****

**Figure 3.9.27 Scatter plot of weight and normalised used price**

By looking at the above figure 3.9.27 can be seen that less weight used mobile devices has the high normalised used price and high weight used mobile devices have the low normalised used devices. So, to get more profits simba mobiles need to focus on the lightweight mobiles while purchasing from the wholesale market.

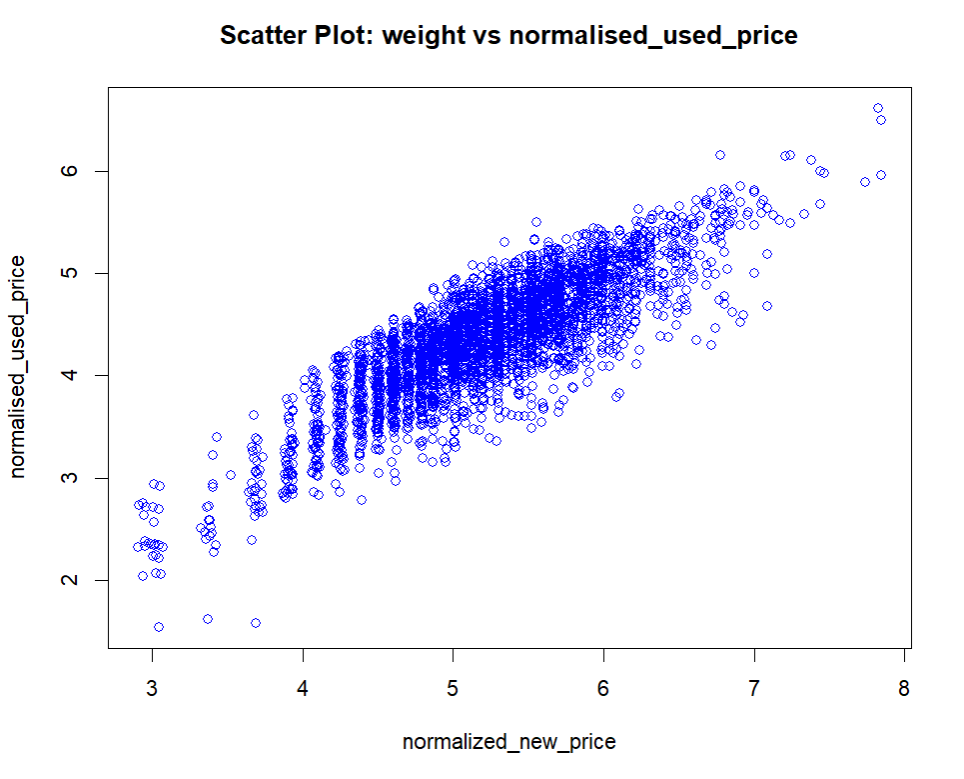
**Box plot of released year and normalised used price:**

****

**Figure 3.9.30 Box plot of released year and normalised used price**

By looking at the above figure 3.9.30 can be seen that recent released used mobiles devices have the highest normalised used price and at the time of 2013 released used mobile devices have the low normalised used price. So, to get more profits Simba mobiles need to focus on the recently released mobiles while purchasing from the wholesale market.

**Scatter plot of Normalised\_new\_price and normalised\_used\_price:**

****

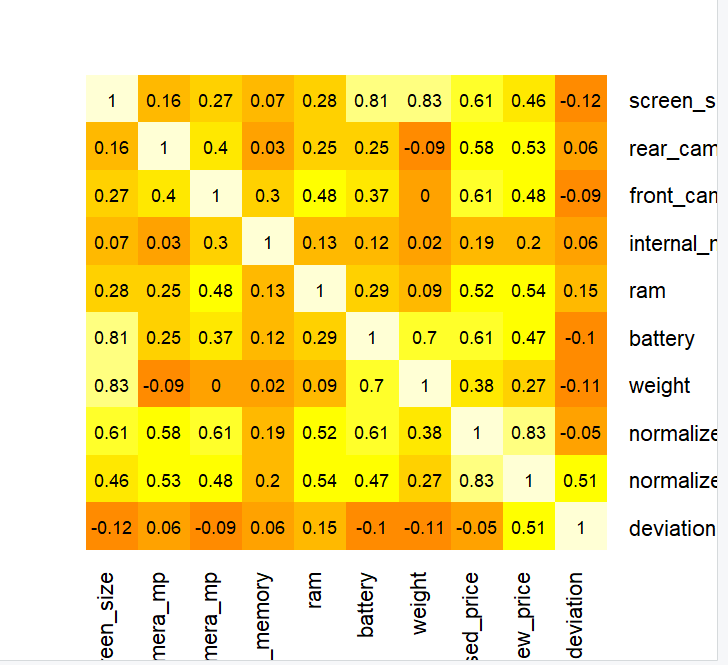
**Figure 3.9.31 Scatter plot of normalised new price and normalised used price**

Looking at the above figure 3.9.31, can be seen that high normalized new price mobile devices have high normalized used mobile devices, and low normalized new price mobile devices have low normalized used price. Both have a linear relationship between them.

After conducting predictor analysis, it was observed that several variables exhibit influences on the target column. The box plots provided insights into the degree of correlation between different columns, helping to understand the relationships among variables. Additionally, the box plots were effective in identifying the demand for different brands in the market. These visualizations collectively contribute to a comprehensive understanding of the data and its key patterns.

After the exploratory data analysis found that there are no outliers present in the data.

**3.10 Correlation for all numerical columns:**

****

**Figure 3.10.1 Correlation for all numerical columns**

By looking at the above figure 3.10.1 can be seen that battery and normalized\_used\_price is high positive correlation with each other, weight and normalized\_used\_price is also high positive correlation with each other. Screen\_size and battery are high positive correlation with each other. Screen\_size and weight are also highly correlated with each other. Days\_used and normalized\_used\_price are moderate negatively correlated with each other. Remaining all columns are also having some correlation between them.

**4.Dimensionality Reduction:**

In this business, the primary objective is to sell used mobile devices based on their characteristics and features. The dataset encompasses all relevant variables related to the target column, and each column holds its significance. Given the low-dimensional nature of the data, there is no need to employ dimensionality reduction techniques. The emphasis should be on harnessing the existing features, as they are all crucial to the goal of selling used mobile devices.

Variables exhibiting positive or negative correlations with others may indicate duplicate information. In such cases, one of the columns can be removed. Decision trees inherently consider the most important variables, and the remaining ones are automatically excluded during tree construction. After building the tree, one can observe the important features. This approach naturally leads to dimensionality reduction.

**5.DATA TRANSFORMATION:**

Data transformation is a crucial step in preparing data for machine learning models, and the nature of these transformations can vary based on the requirements of the selected model. For example, when applying a k-nearest neighbours (KNN) model, it may be necessary to scale features and create dummy variables. Conversely, when using logistic regression, converting categorical columns into factors might be essential before constructing the model. Therefore, the data preparation process is customized to meet the specific needs and assumptions of the chosen model. To simplify the analysis, the days\_used variable was categorized into three groups: within one year, 1-2 years, and 2-3 years.

**6. DATA PARTITIONING METHODS:**

To choose the best model that performs optimally in classifying the outcome variable of interest with the available data, it is crucial to avoid introducing optimism bias. This bias arises when the model is developed and assessed using the same dataset, potentially leading to issues in real-time scenarios. To mitigate overfitting problems, the data is divided into three partitions: train, validation, and holdout.

The training partition, being the largest, contains data used to build models for examination. This same training data is utilized to develop multiple models. The validation partition assesses the predictive performance of each model, allowing for model comparison and selection of the best one. In some algorithms, the validation partition may also be used to tune and improve the model. The holdout partition is then employed to assess the performance of the chosen model with new data.

The data is split randomly into three parts: train, validation, and holdout partitions, with the training partition comprising 50%, the validation partition 30%, and the holdout partition 20%. After this partitioning, the records in each set should be distinct. The training data consists of 1727 observations and 16 columns, the validation data contains 1036 observations and 16 columns, and the holdout data comprises 691 observations and 16 columns.

**7. REGRESSION MODEL SELECTION:**

Now, regression models will be built to predict the estimated normalized used price for the used mobile devices. Based on different features, regression models will be constructed to predict the normalized used price. By building predictive regression models, the best performing model can be selected based on metrics.

**PREPARING THE DATA:**

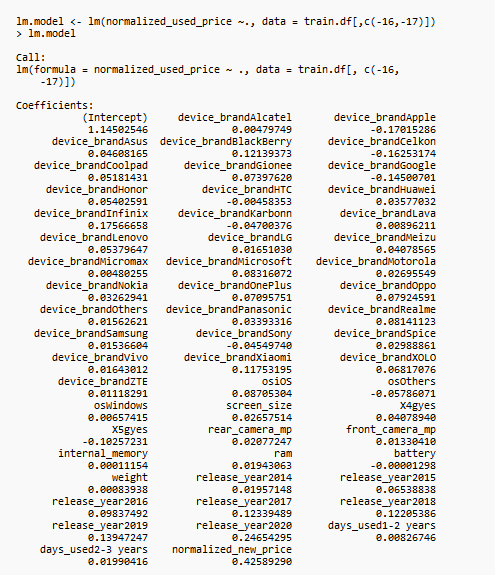
To build the predictive models, the data has already been divided into three partitions. Using these partitions, the model will be built. First, the data will be trained using the training data, and the model's performance will be tuned and evaluated using validation data to select the best regression models. In the final step, the model's performance will be evaluated with holdout data, which is new or unseen to the model.

**7.1 MULTIPLE LINEAR REGRESSION:**

Multiple linear regression model represents the relationship between the target and predictor variables, and it will assume the linear relationship between the target and predictor variables. By using all features as a predictors and normalised used price is a target variable. In linear regression no need to create the dummy variables in pre-processing while building the model it will automatically create. The estimated equation for the multiple linear regression model is

Normalised\_used\_price = b0+b1.x1+b2.x2+b3.x3+…………………bn.xn

Will use the lm() to build the linear model. First will train the data with the training model then will evaluate the model with the validation model. Let’s have a look on the lm model equation

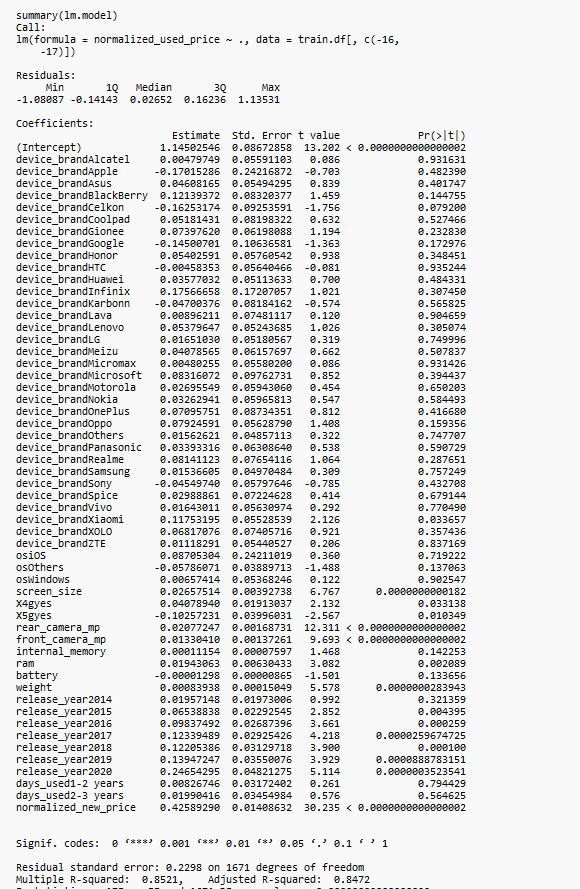


**Figure 7.1.1 Estimated linear model equation**

The estimated linear equation is

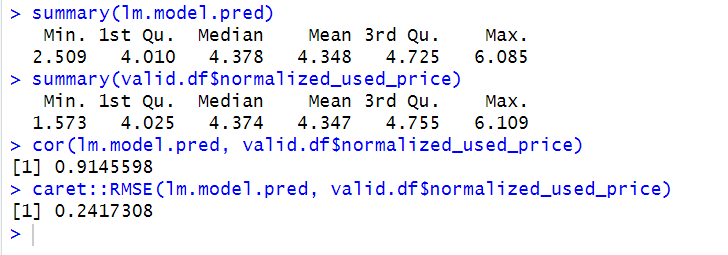
Normlaised\_used\_price = 1.14502546+ device\_brandAlcatel \* 0.00479749 + device\_brandApple \* ( -0.17015286) + device\_brandAsus \* 0.04608165 + device\_brandBlackBerry \* 0.12139373 + device\_brandCelkon \* -0.16253174 + device\_brandCoolpad \* 0.05181431 + device\_brandGionee \* 0.12139373 + device\_brandGoogle \* -0.14500701 + device\_brandHonor\* 0.05402591 + device\_brandHTC\* -0.00458353 +device\_brandHuawe \* 0.03577032 + device\_brandInfinix \* 0.17566658 + device\_brandKarbonn \* -0.04700376 + device\_brandLava \* 0.00896211 + device\_brandLenovo \* 0.05379647 + device\_brandLG \* 0.01651030 + device\_brandMeizu \* 0.04078565 + device\_brandMicromax \* 0.00480255 + device\_brandMicrosoft \* 0.08316072 + device\_brandMotorola \* 0.02695549 + device\_brandNokia \* 0.03262941 + device\_brandOnePlus \* 0.07095751 + device\_brandOppo \* 0.07924591 + device\_brandOthers \* 0.01562621 + device\_brandPanasonic \* 0.03393316 + device\_brandRealme \* 0.08141123 + device\_brandSamsung \* 0.01536604 + device\_brandSony \* -0.04549740 + device\_brandSpice \* 0.02988861 + device\_brandVivo \* 0.01643012 + device\_brandXiaomi \* 0.11753195 + device\_brandXOLO \* 0.06817076 + device\_brandZTE \* 0.01118291 + osiOS \* 0.08705304 + osOthers \* -0.05786071 + osWindows \* 0.00657415 + screen\_size \* 0.02657514 + X4gyes \* 0.04078940 + X5gyes \* -0.10257231 + rear\_camera\_mp \* 0.02077247 + front\_camera\_mp \* 0.01330410 + internal\_memory \* 0.00011154 + ram \* 0.01943063 + battery \* -0.00001298 + weight \* 0.00083938 + release\_year2014 \* 0.01957148 + release\_year2015 \* 0.06538838 + release\_year2016 \* 0.09837492 + release\_year2017 \* 0.12339489 + release\_year2018 \* 0.12205386 + release\_year2019 \* 0.13947247 + release\_year2020 + 0.24654295 + days\_used1-2 years \* 0.00826746 + days\_used2-3 years \* 0.01990416 + normalized\_new\_price \* 0.42589290 .

By utilizing this equation can be able to predict the normalised used price for each used mobile device. So, the Simba mobiles dealer will get a basic idea about the price of the used mobile device based on the features. Let’s have a look on the summary of the linear model

**** ****

**Figure 7.1.2 Summary of Linear Model**

By looking at above figure 7.1.2 can be seen that most significant features and R-squared is 0.8521 and adjusted R-squared is 0.8472.



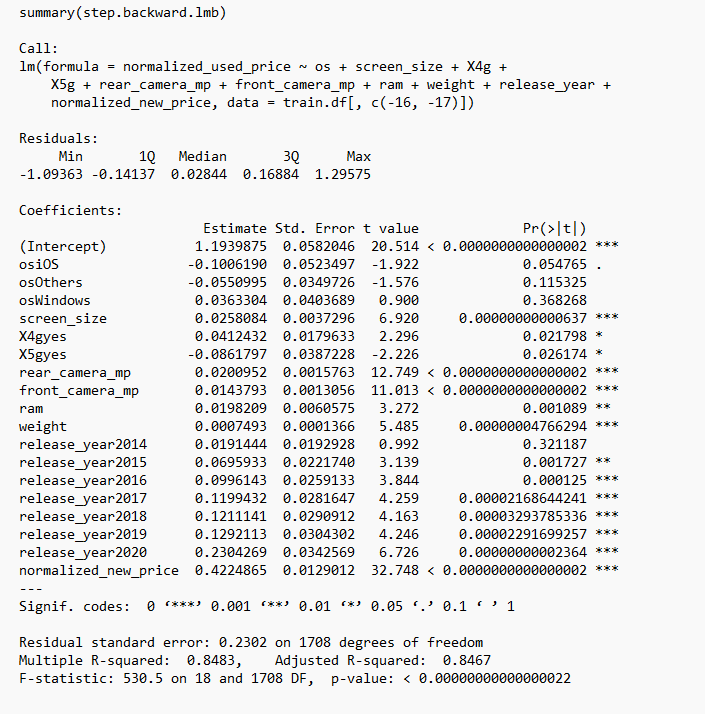
**Figure 7.1.3 Summary of Predicted and Actual values**

By looking at the above figure 7.1.3, can be seen that the summaries of actual and predicted values of the normalized used price do not have much difference between them. Therefore, the predictive power is relatively high, and the data is distributed uniformly. The correlation between the actual and predicted values is 0.9, and the RMSE value is 0.2417.

**7.2 BACKWARD STEP-WISE LINEAR REGRESSION:**

To improve the performance of the linear regression going to perform the backward stepwise regression in this it will starts with all predictors, and it will reduce the one-by-one predictors to get the accurate results.

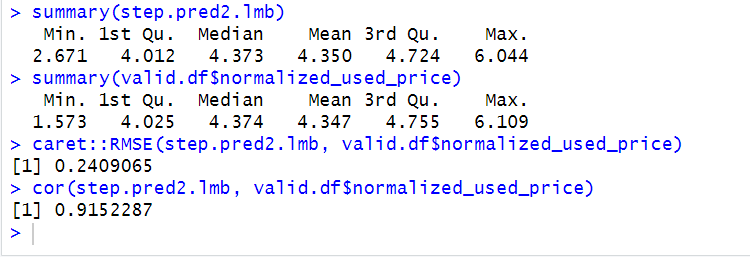
Let’s have a look on the summary of the backward stepwise linear regression model



**Figure 7.2.1 Summary of the Backward step wise linear regression**

By looking above figure 7.2.1 can be seen that most significant features, R-squared value is 0.8483 and adjusted R-squared value is 0.8467.

Let’s have a look on the summaries of the actual and observed values



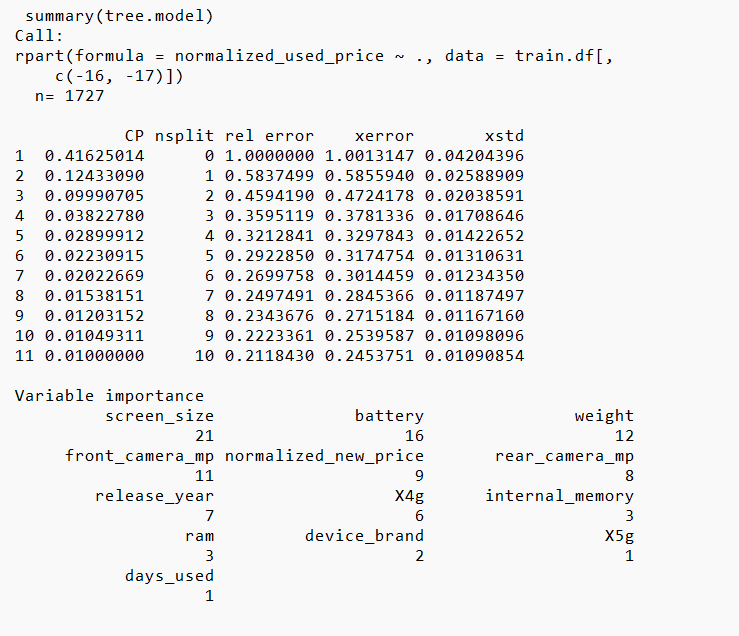
**Figure 7.2.2 Summaries of Actual and Predicted Values**

By looking at the above figure 7.2.2, can be seen that there is not much difference between the actual and predicted values. Therefore, it can be said that the predictive power is high, and the data is distributed uniformly. The RMSE value is 0.2409, and the correlation between the actual and predicted values is 0.91.

**7.3 DECISION TREE:**

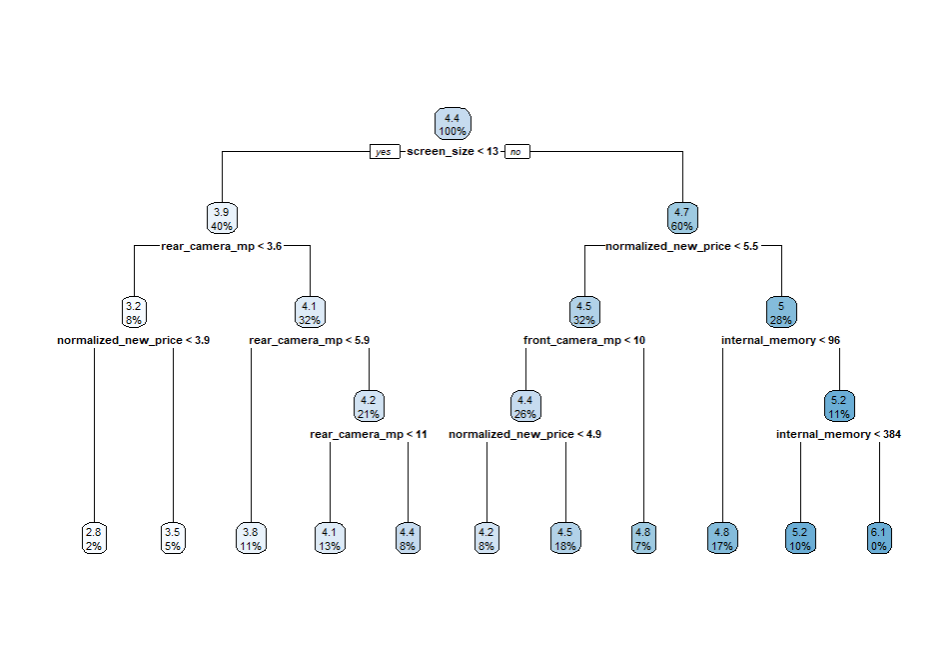
In regression tree prediction for new data are based on the rules represented on the tree in regression trees predictions are obtained by averaging the outcome values in the nodes. To access this regression tree we use r.part() function. It is based on the rules of the tree. In regression-tree prediction, node homogeneity is measured by various statistics such as variance, standard deviation, or absolute deviation from the mean.

Let’s have a look on the summary of the model



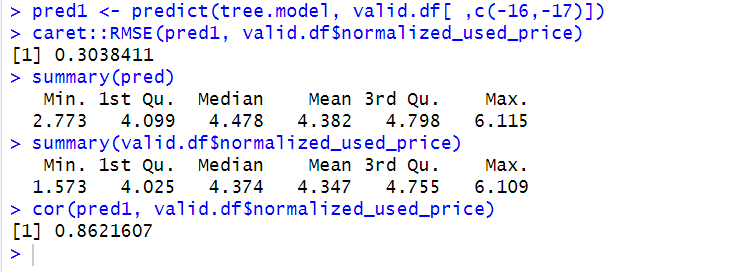
**Figure 7.3.1 Summary of the tree model**

By looking at the above figure 7.3.1 can be seen that importance of the variables and number of the splits. Let’s have a look on the plotted tree



**Figure 7.3.2 Plotted tree model tree**

By looking at above figure 7.3.2 can be seen that the nodes are splits based on the rules. Let’s have look on the summaries of actual and predicted values.



**Figure 7.3.3 Summaries of an Actual and predicted values**

By looking at figure 7.3.3, can be seen that there is not much difference between the actual and predicted values. This indicates more predictive power, and the data is distributed uniformly. The RMSE value is 0.3038, and the correlation between the actual and predicted values is 0.86, which indicates a strong correlation between them.

**7.4 Comparison of Regression Models:**

Let’s compare all RMSE values of the model and will select the best performance model based upon the lowest RMSE values.

|  |  |
| --- | --- |
| REGRESSION MODELS | RMSE |
| LINEAR REGRESSION MODEL | 0.2417 |
| BACKWARD STEPWISE LINEAR REGRESSION MODEL | 0.2409 |
| DECISION TREE | 0.3038 |

Based on the lowest RMSE value, the backward stepwise linear regression model emerges as the best regression model among all others. This model selectively chooses important variables to achieve accuracy while reducing unimportant ones. In the future, this approach will aid business owners in data gathering. Rather than collecting all variables, they can focus on the important ones, saving time and obtaining accurate results. By using this regression models, they can be able to predict the used price of the used devices so, instead of investing high amount money on used devices they can invest wisely.

**8. CLASSIFIER MODEL SELECTION:**

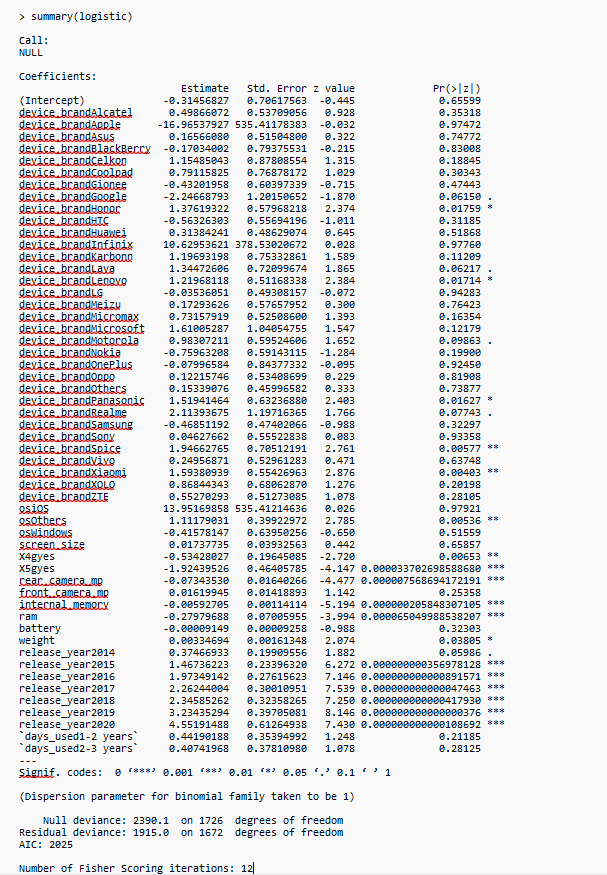
Now, we will build a classification predictive model to determine whether used mobile devices have high resale value in the market or not. Based on the model output, the company will make informed decisions about used mobile devices when purchasing from the original customer. The target column is resale, which indicates whether a mobile has high or low resale value in the market based on its features. We will use all predictors to build the model except normalized\_new\_price and normalized\_used\_price. The model's performance will be evaluated using the confusion matrix, with a focus on correctly identifying the class with high resale value in the market. To achieve this classification goal, we can use logistic regression, classification tree, KNN, and neural networks.

**8.1 LOGISTIC REGRESSION:**

Logistic regression is the most popular classification model. In this model, we use the glm() function. To access this function, we need to install and import the “Car” and “Caret” packages. Logistic regression extends the idea of linear regression to situations where the outcome variable is categorical. Logistic regression estimates probabilities, indicating the likelihood of each class. It uses a threshold value to classify the record into each class.

**8.1.1 PREPARING THE DATA FOR LOGISTIC REGRESSION**

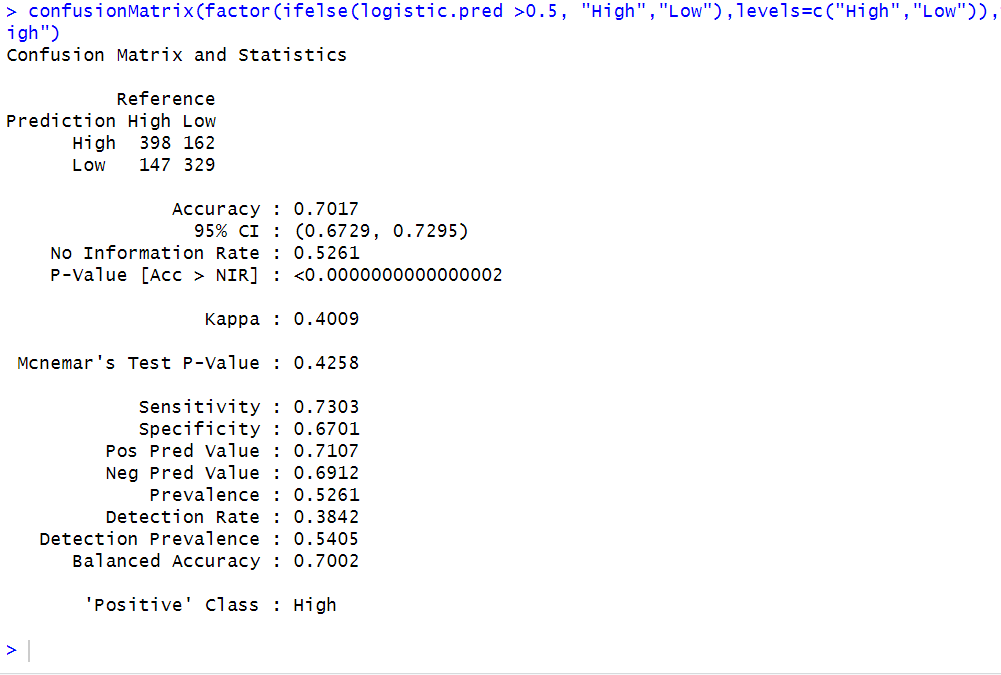
Before applying logistic regression, will encode the factors for categorical columns and then proceed to fit the model. First, we will fit the model with the training data and subsequently evaluate its performance with the validation data. Let's examine the summary of the logistic regression in Figure 8.1.1, which is presented in the form of linear output, to be later converted into probability. Based on the threshold value, it will assign the class.



**Figure 8.1.1 Summary of logistic regression model**

**8.1.2 EVALUATING THE PERFORMANCE:**

After fitting the model, will assess its performance with the validation data through a confusion matrix, which classifies records correctly and incorrectly based on the classifier data. From the summary can be able to see the most significance variables and the greater number of this symbol “\*” indicates the how much important these variables to predict the accurate results. Number of fishers scoring iterations is 12. The classifier uses a threshold value to categorize the class. In this case, the threshold value is set at 0.5, indicating that probabilities above the threshold represent high resale value for used mobiles in the market, while those below indicates low resale value. Using this classification, Simba Mobiles owners can make decisions about which used mobiles have high or low resale value in the market, allowing them to focus on specific used mobiles to maximize profits.

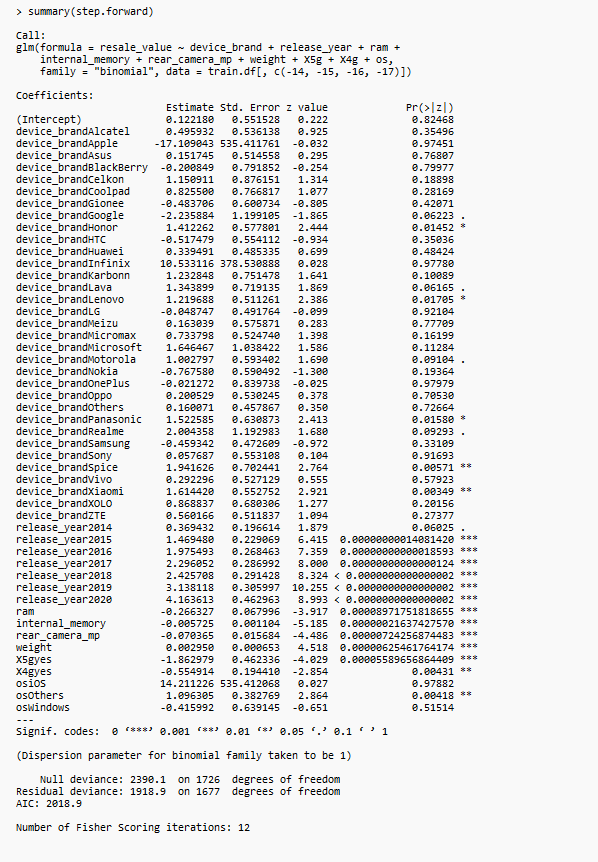


**Figure 8.1.2 Confusion Matrix of the Logistic Regression**

By examining the confusion matrix in Figure 8.1.2 for logistic regression, it shows an accuracy of 70.17%, a sensitivity of 73.03%, and a specificity of 67.01%. The sensitivity score is crucial because it reflects the model's ability to correctly predict the most important class, which in this case is "High Resale." A sensitivity of 73.03% indicates that the model has a 73.03% likelihood of accurately identifying mobile devices with high resale value in the market. This suggests a reasonable chance of predicting which used mobile devices will have high resale value. With this information, Simba Mobiles can make more informed decisions when purchasing used mobile devices from the wholesale market. The sensitivity score can guide the business in selecting mobiles with the best potential for resale, ultimately contributing to better profitability.

**8.1.3 FORWARD STEP WISE LOGISTIC REGRESSION:**

To improve the performance of logistic regression will perform the forward step wise logistic regression. In forward step wise logistic regression, it will start with 0 predictors and add one by one predictors to get the accurate results. Let’s have a look on the summary of the forward stepwise logistic regression.

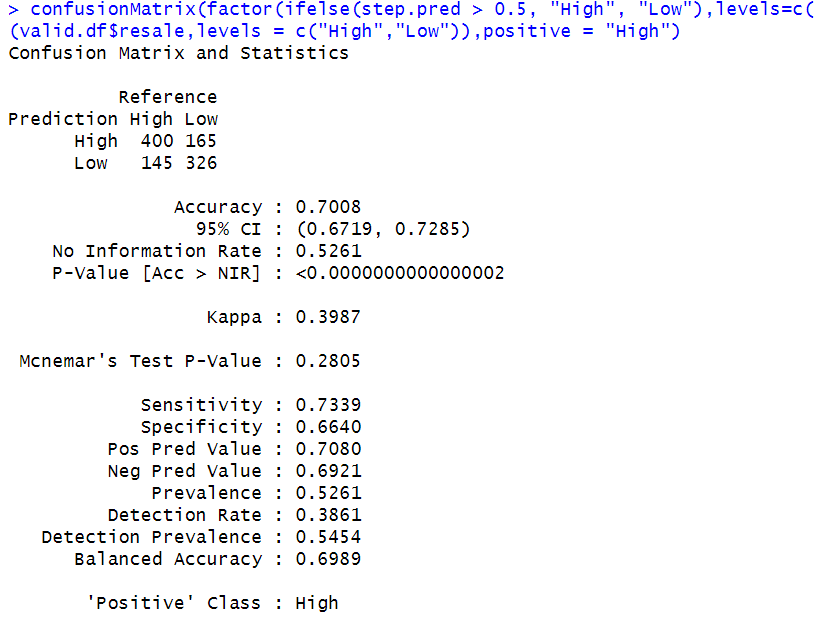


**Figure 8.1.3 Summary of the Forward Step Wise Logistic Regression**

By looking at the figure 8.1.3 can be able to see that summary of the forward stepwise logistic regression estimated equation and significance of the variables The greater number of this symbol “\*” indicates that how much important those variables to predict. Number of fishers scoring iterations are 12. By utilizing this estimated equation will calculate the probabilities for individual used mobile devices.

**8.1.4 EVALUATING MODEL PERFORMANCE:**

After fitting the model will assess the model performance with the validation data through a confusion matrix, which classifies records correctly and incorrectly based on the classifier data. Let’s have a look on the confusion matrix of the forward step wise logistic regression

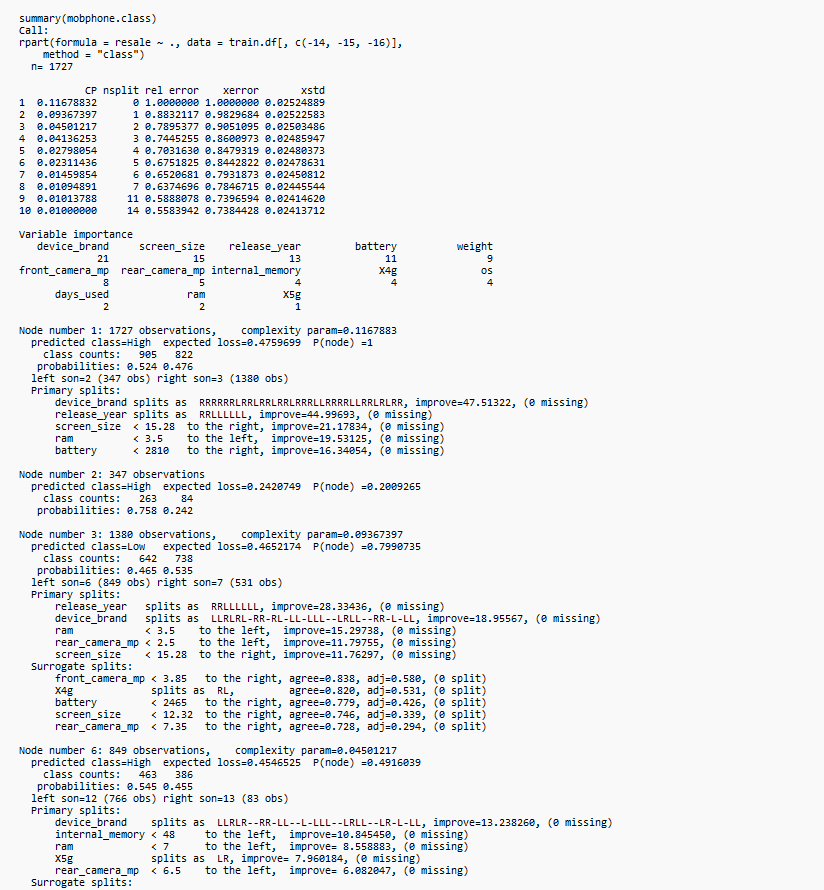


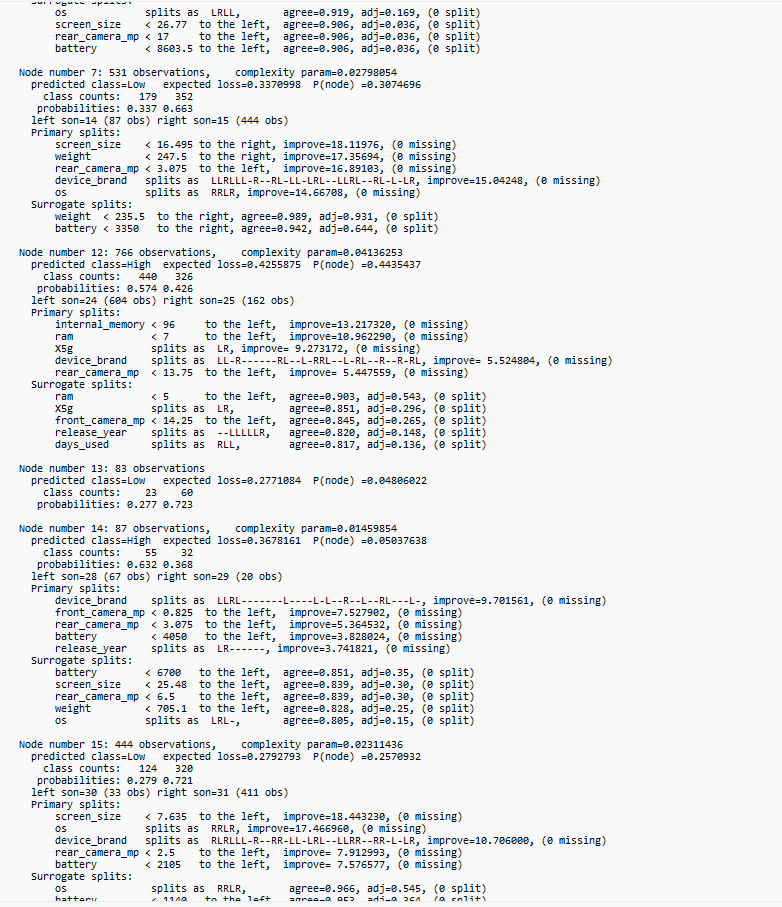
**Figure 8.1.4 Confusion Matrix of the Forward Stepwise Logistic Regression**

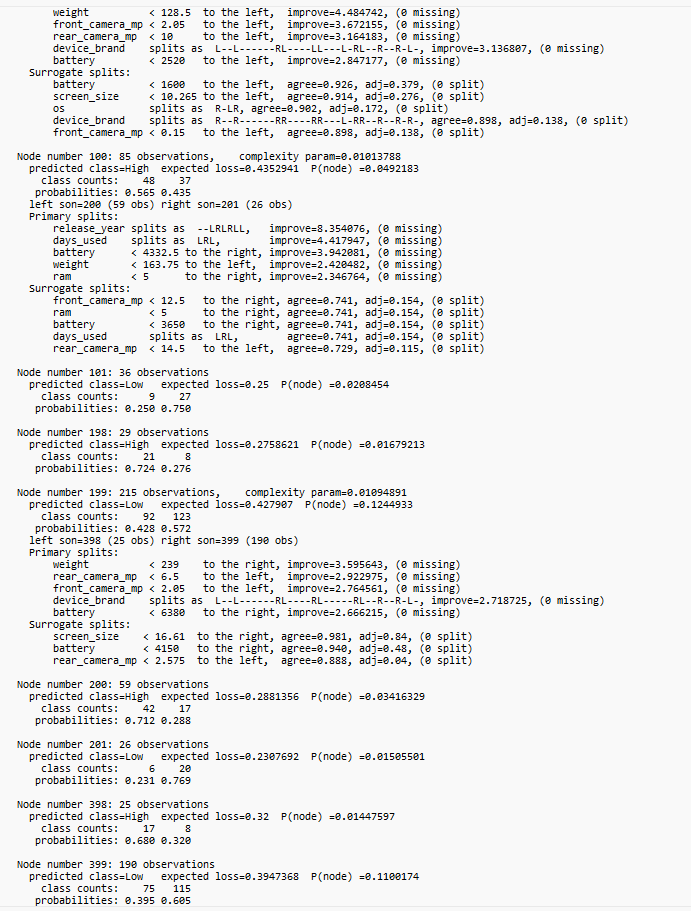
By examining the confusion matrix in Figure 8.1.4 for logistic regression, it shows an accuracy of 70.08%, a sensitivity of 73.39%, and a specificity of 66.40%. The sensitivity score is crucial because it reflects the model's ability to correctly predict the most important class, which in this case is "High Resale." A sensitivity of 73.39% indicates that the model has a 73.39% likelihood of accurately identifying mobile devices with high resale value in the market. This suggests a reasonable chance of predicting which used mobile devices will have high resale value. With this information, Simba Mobiles can make more informed decisions when purchasing used mobile devices from the wholesale market. The sensitivity score can guide the business in selecting mobiles with the best potential for resale, ultimately contributing to better profitability.

**8.2 Classification Tree:**

The classification tree is a widely used method for categorizing record outputs based on the rules represented on the tree. It considers most votes when classifying a new record. The process involves recursively splitting the dataset into rectangles, with each split determined by the most important variable at that point, aiming to make each rectangle as homogeneous or pure as possible.

To implement this, we utilize the 'rpart' and 'rpart.plot' packages in R to construct and visualize the classification tree. Initially, the model is built using the training data, and subsequently, the interpretation of the model output is examined.

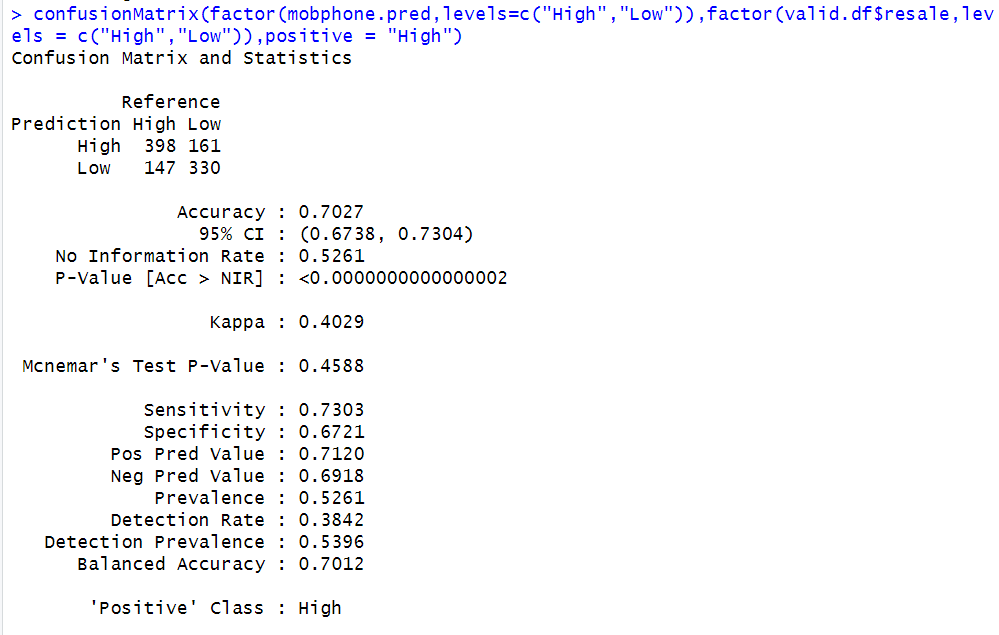




**Figure 8.2 Summary of the Classification Tree**

By looking at the above figure 8.2 can be able to observe the summary of the classification tree and variable importance.

**8.2.1 EVALUATING THE PERFORMANCE:**

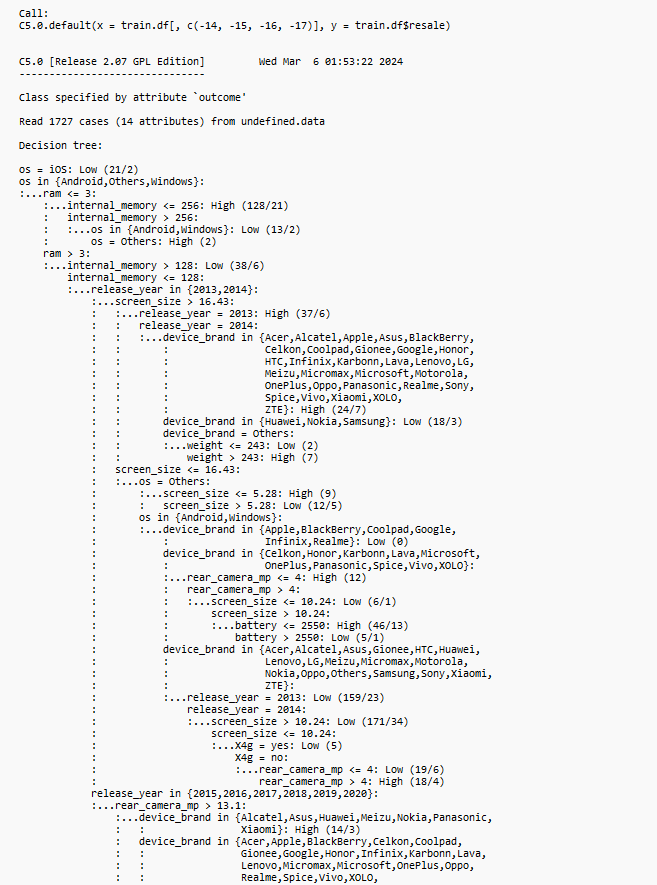
After constructing the model using the training data, the performance will be evaluated with the validation data using the confusion matrix. Assessing the model's performance is based on metrics from the confusion matrix, particularly focusing on sensitivity. Sensitivity measures the model's capability to correctly classify the positive class, signifying a higher ability to predict high resale used mobile devices in the resale market. By relying on this, the dealer can make decisions regarding which types of used mobiles possess the highest resale value, allowing them to concentrate more on those and maximize profits. Now, let's examine the confusion matrix.

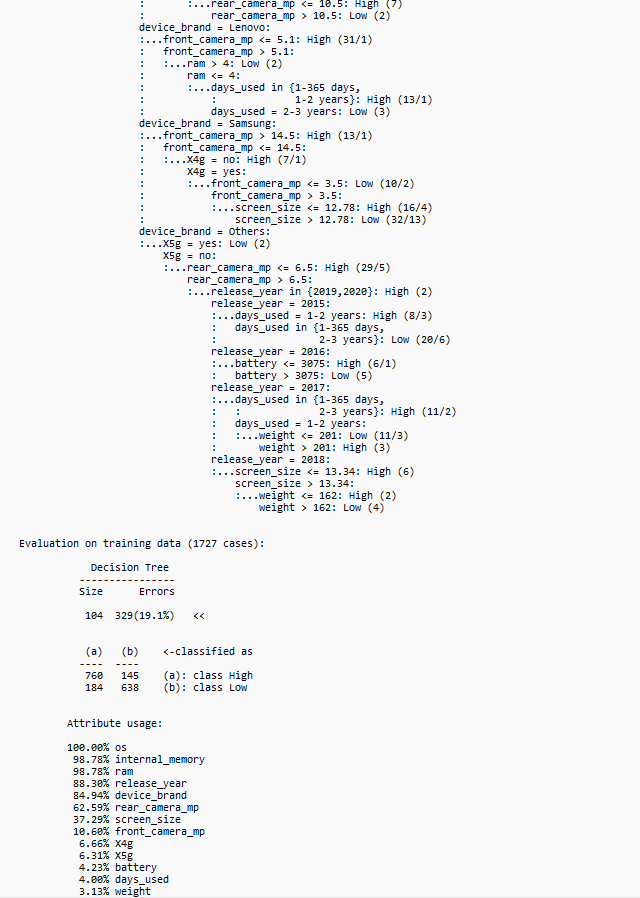
**Figure 8.2.1 Confusion matrix of the classification tree**

By examining the confusion matrix in Figure 8.2.1 for the classification tree, it can be observed that the accuracy is 70.27%, sensitivity is 73.03%, and specificity is 67.21%. Considering the sensitivity, it can be stated that there is a 73.03% chance of correctly classifying the positive class, indicating the likelihood of correctly classifying the high resale value in the market. Consequently, based on this result, the dealer can make informed decisions regarding used mobile devices to maximize profits.

**8.2.2 Classification Tree by using C5.0:**

This is another approach to in classification tree by using the C5.0. First will build the model with the training data then will interpret the results of the model

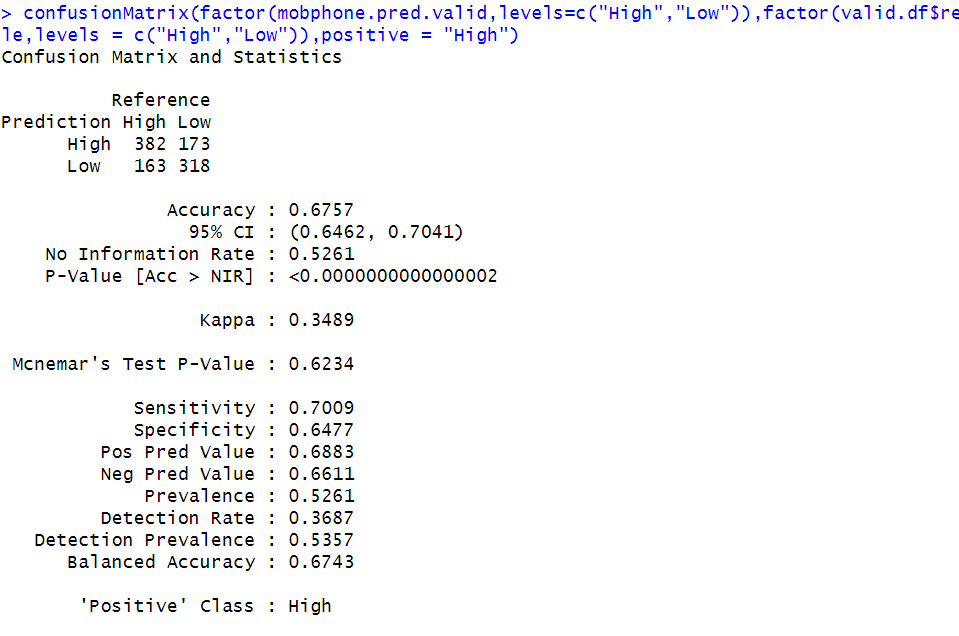




**Figure 8.2.2 Summary of the C5.0 model**

**8.2.3 EVALUATE THE PERFORMANCE USING C5.0:**

Now evaluate the performance with the validation data through the confusion matrix. Based on the confusion matrix metrics will evaluate the performance of the model. Let’s have a look on the confusion matrix.



**Figure 8.2.3 Confusion matrix of the C5.0 model**

By examining the confusion matrix in Figure 8.2.3 for the C5.0 model, it can be observed that the accuracy is 67.57%, sensitivity is 70.09%, and specificity is 64.77%. Based on the sensitivity results, it can be stated that there is a 70.09% chance of correctly classifying the high resale of used mobiles in the resale market.

**9. BEST CLASSIFIER MODEL SELECTION:**

|  |  |  |
| --- | --- | --- |
| **MODEL** | **ACCURACY** | **SENSITIVITY** |
| LOGISTIC | 70.17% | 73.03% |
| FORWARD STEP WISE LOGISTIC | 70.08% | 73.39% |
| DECISIONTREE USING rpart() | 70.27% | 73.03% |
| DECISIONTREE USING C5.0() | 67.57% | 70.09% |

**9. Comparison model selection table**

In selecting the best model, emphasis was placed on achieving high sensitivity. Upon reviewing the comparison table 9, it is evident that forward stepwise logistic regression exhibits the highest sensitivity at 73.39%. Consequently, the forward stepwise is chosen as the best classifier model among the available options. The rationale behind this selection lies in the forward stepwise logistic regression ability to consider the most crucial predictors in the data, as opposed to utilizing all predictors. This approach aligns with the future need for businesspeople to gather essential information rather than collecting extensive data. Using fewer data points, a high-performance model can be achieved.

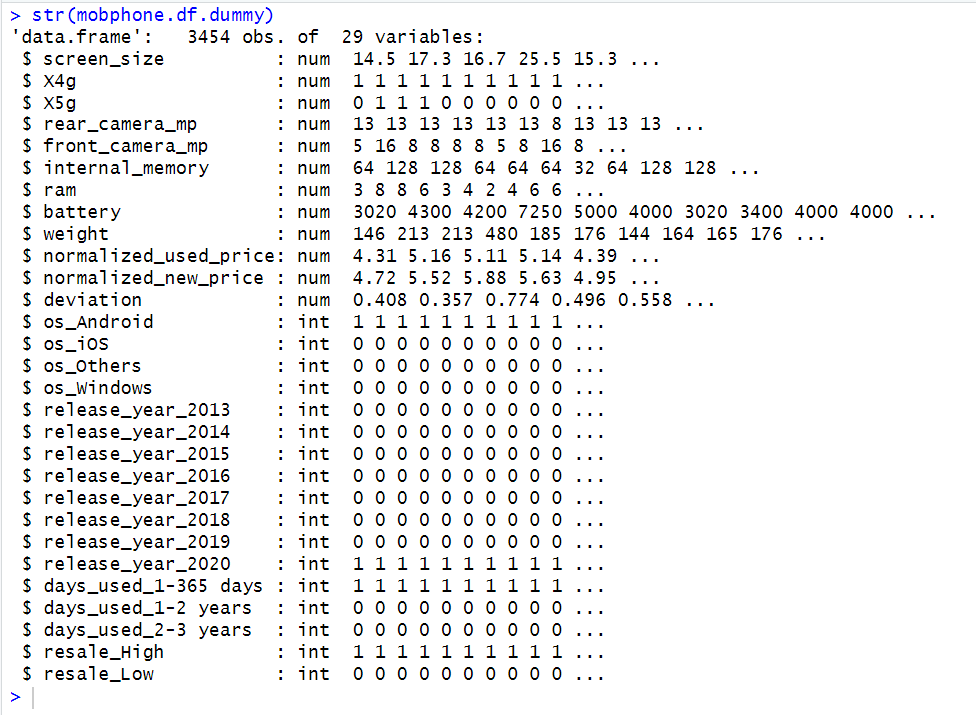
The forward stepwise logistic regression can be effectively employed to classify the probability of high resale values for used mobile devices in the market. This classification provides valuable insights for business dealer to make informed decisions regarding used mobiles, focusing on those with the potential for higher profitability. So, decision depends on the Simba mobiles owner which kind of used devices must be buy here assuming that they will buy both to run a successful business.

**10.k-Means Clustering:**

Cluster analysis is used to form groups or clusters of similar records based on several measurements made on these records. The key idea is to characterize the clusters in ways that would be useful for the aims of the analysis. In this context, clustering is used to form groups of similar mobiles based on the same specifications. This method helps understand the trends or patterns of the mobiles and is easy to interpret for the Simba Mobiles owner while restocking the used mobile devices from the wholesale market. By using this method, it is possible to understand the structure of the analysis based on the comprehensive measures of the used mobile devices. After grouping the used mobiles, it is possible to analyse which kind of mobiles has the high resale value in the resale market.

**Preparing the Data for Clustering:**

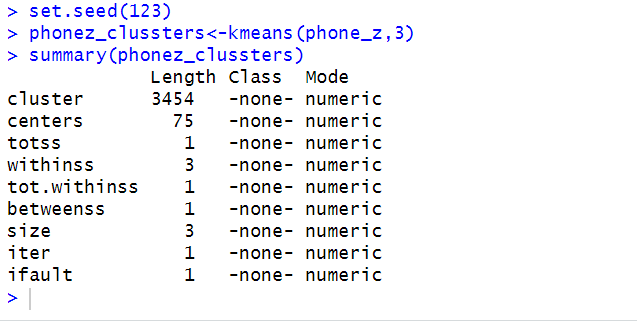
To perform cluster analysis on the used mobile devices, dummy variables will first be created for all the columns except for the device brand. Then, the entire data will be normalized before applying the clustering. This process converts all measurements to the same scale. Normalizing a measurement involves subtracting the average and dividing by the standard deviation. To avoid the duplicates converted 4g and 5g into 1 & 0 from yes and no.



**Figure 10.1 structure of the dummy variables in the data**

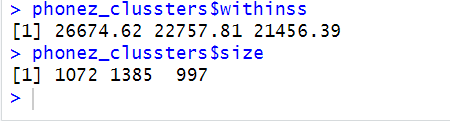
By looking at above figure 10.1 can be seen that dummy variables created for further analysis. Next, will perform the standardization so all features are unitless and follow the approximately standardized normal distribution.

Let’s have a look on the summary of the clusters



**Figure 10.2 Summary of the clusters**

Let’s have a look on the cluster withinss and size of the clusters

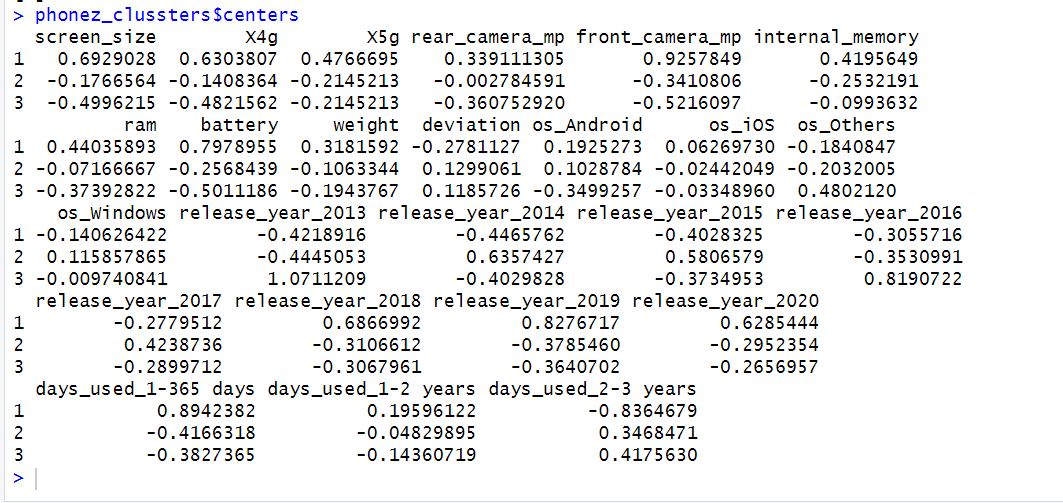


**Figure 10.3 Cluster withinss and Size**

The withinss of the cluster 1 is 26674.62, cluster 2 is 22757.81 and cluster 3 is 21456.39.

The size of the cluster 1 is 1072, cluster 2 is 1385 and cluster 3 is 997.

Let’s interpret the centres of each cluster



**Figure 10.4 Centres of clusters**

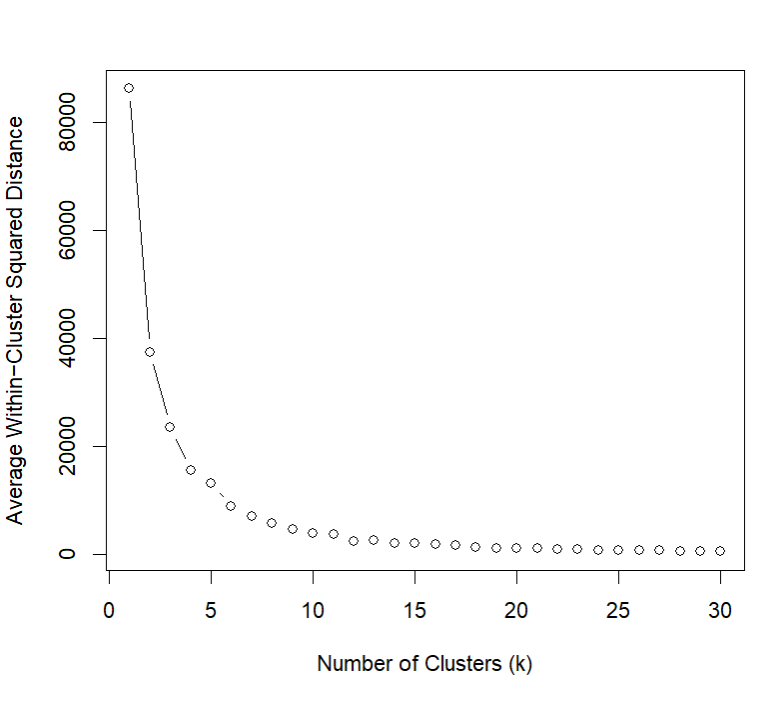
By looking at the above figure 10.4 can be seen that

This cluster 1 appears to represent newer, high-end used mobile devices with modern features and higher resale values. Given these characteristics, this cluster may attract customers seeking a like-new experience with advanced technology. Simba Mobiles should target this cluster for resale, as it presents a high demand used mobiles due to its appealing features and less used.

This cluster 2 represents older, less advanced used mobile devices with lower resale values. These devices may appeal to budget-conscious customers or those seeking basic functionality. Simba Mobiles could leverage this cluster to offer budget-friendly options, catering to a different market segment with more affordable prices.

This cluster3 seems to represent a middle-range segment with a balance of features and high resale values. These devices may appeal to customers seeking a balance between quality and affordability. Simba Mobiles could use this cluster to target a broader customer base, offering good-quality used mobiles at reasonable prices.

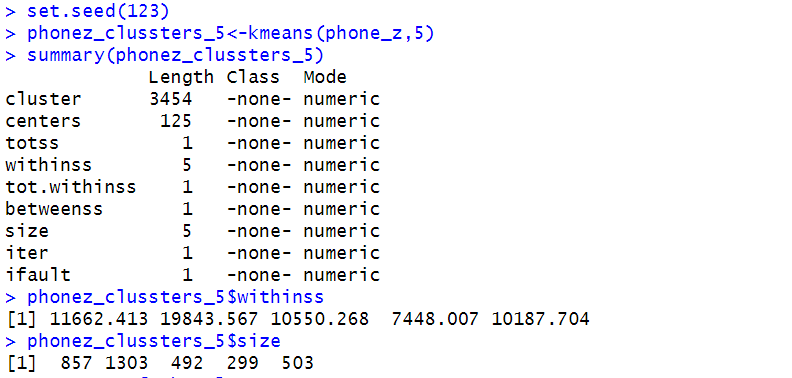
**Improving Cluster Performance:**



**Figure 10.5 Elbow Curve**

By looking at the above figure 10.5, can be seen that the best k=5 because after this point, the average within-cluster squared distances decrease gradually.

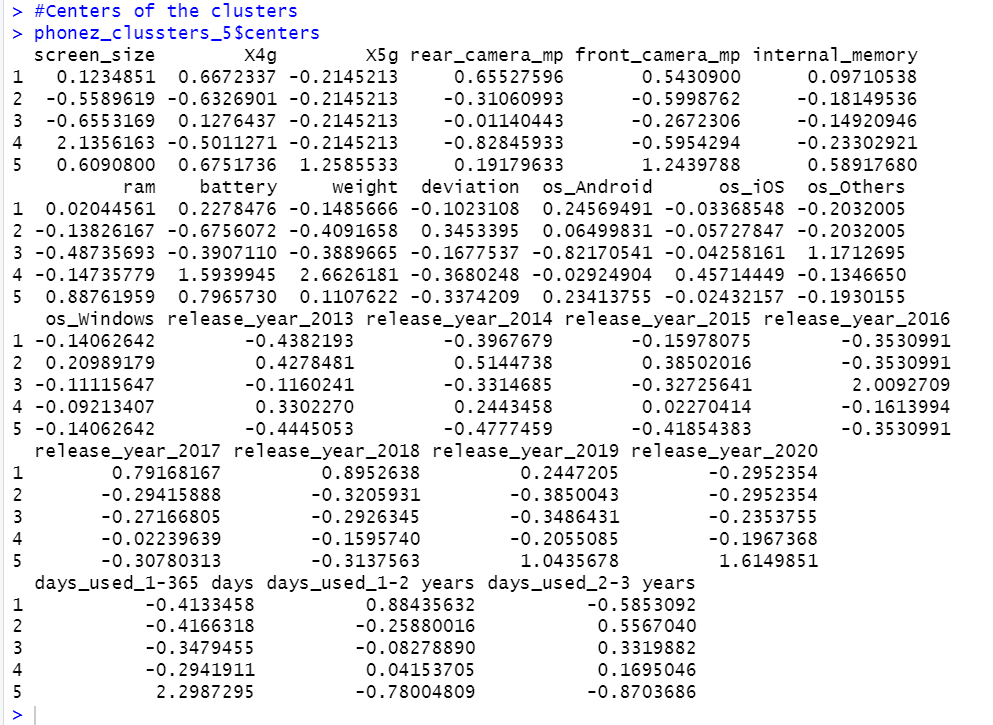
Let’s have a look on the summaries of the five clusters



**Figure 10.6 Summaries of 5 clusters**

By looking at the above figure 10.6 can be seen that withinss and size of each cluster. The withinss of cluster 1 is 11662.413, cluster 2 is 19843.567, cluster 3 is 10550.268, cluster 4 is 7448.007 and cluster 5 is 10187.704.

The size of the cluster 1 is 857 records, cluster 2 is 1303 records, cluster 3 is 492 records, cluster 4 is 299 and the cluster 5 is 503. Let’s have a look on the centres of each 5 clusters and their interpretation for each cluster.



**Figure 10.7 Centres of five clusters**

By looking at the above figure 10.7, This cluster is characterized by used mobile devices with 4G capability Android, and typically used for 1 to 2 years. These phones are relatively new and have medium specifications. Simba Mobiles can market these devices to customers seeking reliable phones with modern features at a reasonable price.

Cluster 2 has Android and Windows used mobile devices used within 2-3 years. These types of mobiles are basic devices. Simba Mobiles can focus on budget-conscious customers.

Cluster 3 has 4G service with other used mobiles used within 2-3 years. These are basic devices. Simba Mobiles can focus on customers who just need mobiles with basic features.

Cluster 4 has larger screen size with iOS-based used mobiles with good battery capacity, used within 1-2 years. So, Simba Mobiles can focus on customers who need big screen used mobiles with good battery capacity.

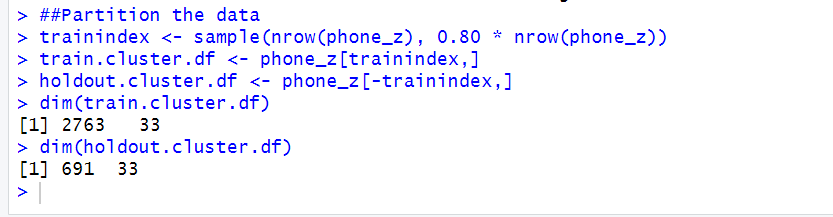
Cluster 5 has high specification used mobile devices used within 1-365 days, like new mobiles with more advanced features.

**10.1 CLASSIFICATION MODELLING FOR CLUSTERING:**

Now, classification will be performed for the clustering. Through clustering, patterns can be understood, and similar data points can be grouped in the data. Similar feature mobiles are grouped into clusters. Utilizing this cluster data, a classification model based on features can be built to classify used mobile devices into clusters. This analysis is more helpful to Simba Mobiles owners when purchasing from the wholesale market. In this, the KNN model will be used because it calculates the distance between points and provides accurate results.

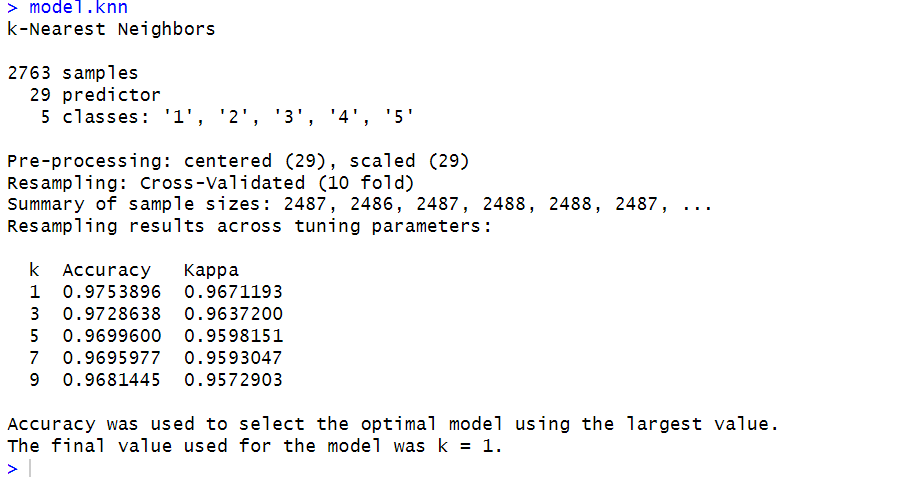
**Preparing the data for the classification model:**

Before building the classification model, dummy variables will be created, and the clustered column will be factored. After that, the data will be partitioned into two partitions, one for training data and the other for holdout data. Here, cross-validation method will be used to improve and stabilize the model performance. It will divide the data into sub-folds, and each time it will test with a new fold while the remaining folds are used for training.



**Figure 10.8 Data preparation for clustering classification model**

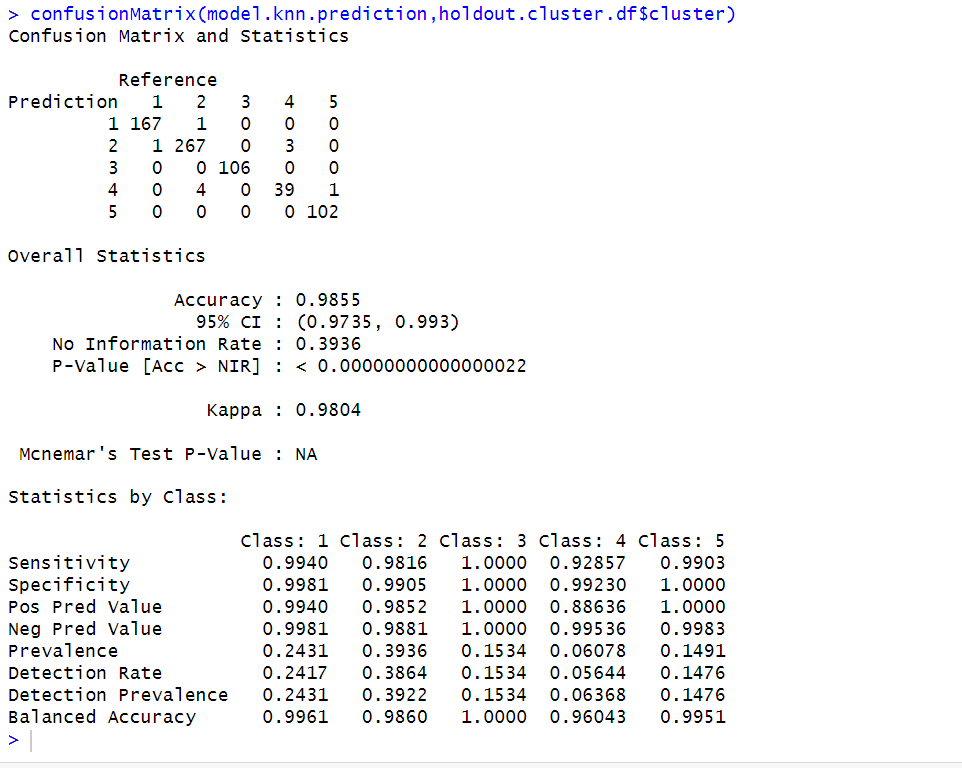
Let’s train the model with the training data and observe the results for that



**Figure 10.9 Clustering classification model**

By looking at the above figure 10.9 can be seen that have different accuracies for different k values. Initially took 9 k for observation. Here have 5 different classes.

Let’s evaluate the model performance with the holdout data and have a look on the accuracy of the model.

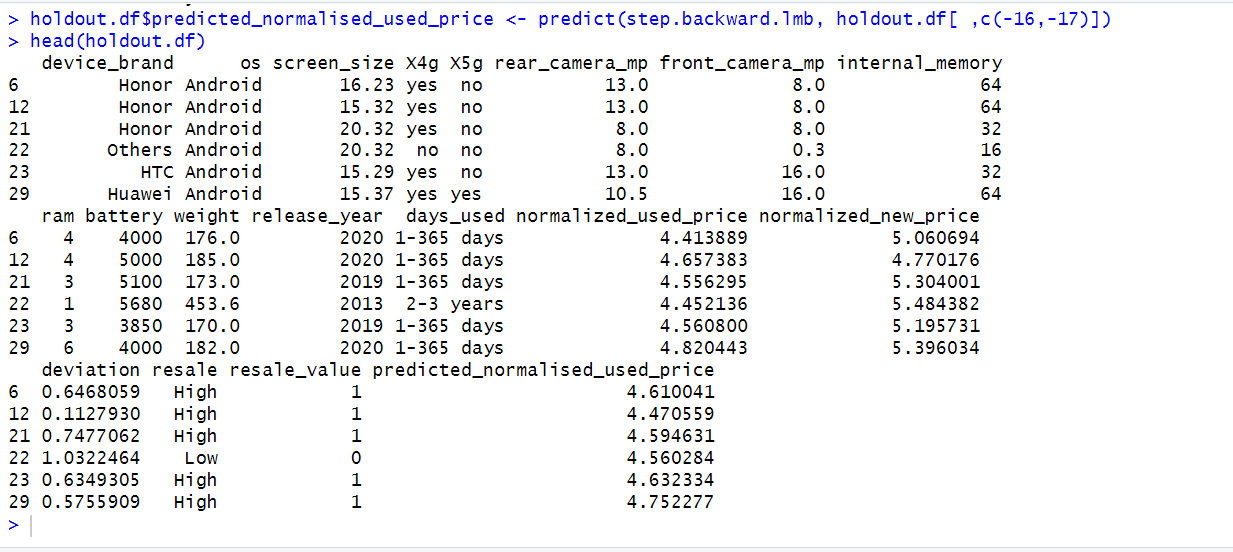


**Figure 10.10 Confusion matrix for the clustering classification model**

By looking at the above figure 10.10 can be seen that the accuracy of the model is 98.55 can say that best model performance and there is a high chance of predicting the accurate results. By using this analysis Simba mobiles owner will get more idea about the stocking while purchasing from the wholesale market and more chance to make a profit on the used mobile devices.

**11. HOLDOUT MODEL PERFORMANCE:**

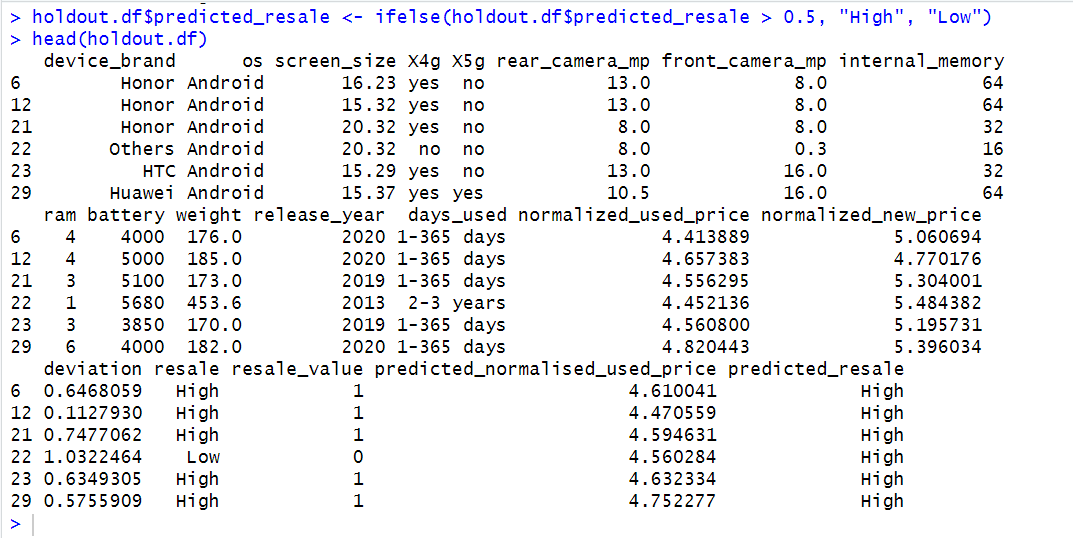
In this step, the model performance will be evaluated with the holdout data. This analysis will assist the Simba Mobiles owner when restocking new used mobile devices from the wholesale market. By providing inputs to this model, it will generate predicted price, predict resale value, and determine which cluster the mobile belongs to. This will give them an idea of which type of mobiles to buy from the wholesale market to improve their business and gain an overview of the patterns of used mobiles in the market. To assess the performance of the models, the holdout data will be applied by adding the predicted columns. Now, we will add the predicted normalized used price column and have a look on that



**Figure 11.1 Added Predicted normalised used price**

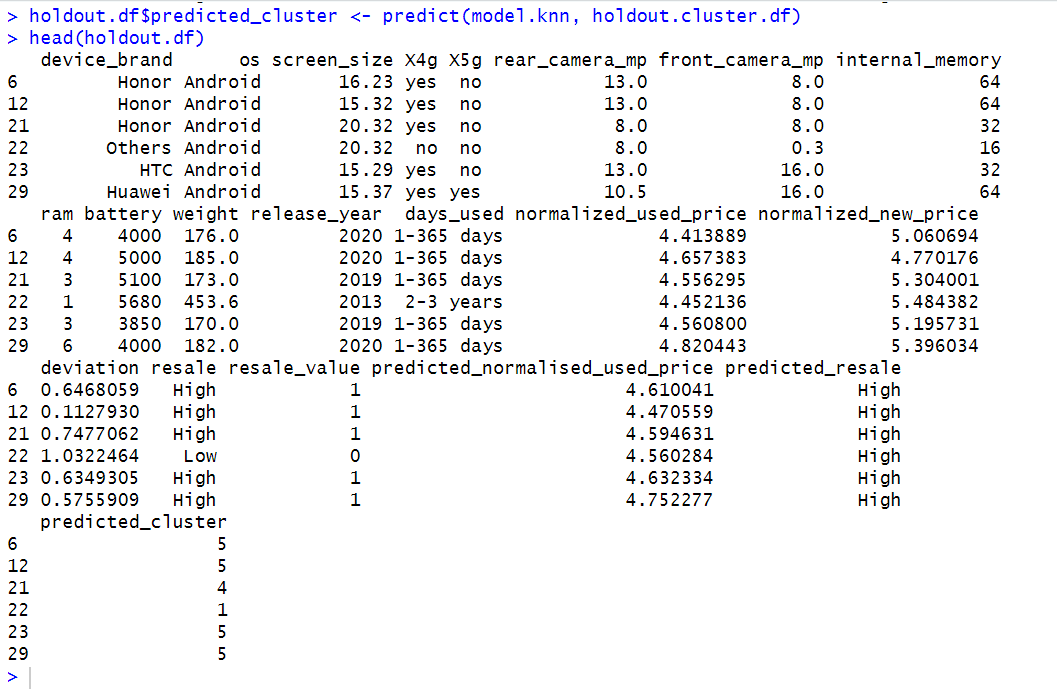
By looking at the above figure 11.1 can be seen that predicted normalised used price column added and can observe the predicted normalised used price for the new data which will provide the better understanding about the predictions about used mobile devices.

Now add another column that is predicted resale value for the used mobile devices to the holdout data and can be able to make the predictions on high and low resale of the used mobile devices.



**Figure 11.2 Added Predicted Resale Column to the Holdout Data**

By looking at the above figure 11.2 can be seen that added predicted resale column which will provide the predictions about the used mobile devices in the holdout data. So, based on the features can be able to predict the resale value of the used mobile devices.



**Figure 11.3 Adding Predicted Cluster to the Holdout Data**

By looking at the above figure 11.3 can be seen that predicted cluster column added to the hold out data. So, based on the features of the used mobile devices can be grouped into the cluster and can be able to predict the new used mobile devices belongs to which cluster.

**12. CONCLUSION:**

In conclusion, this analysis addresses how machine learning algorithms can predict the estimated normalized used price for used mobile devices in the market based on different features of the mobiles. It demonstrates how this prediction can be beneficial to Simba Mobiles owners while purchasing from wholesale market dealers. Various regression models were used to predict the estimated price, with the backward stepwise linear regression model identified as the most effective for making predictions on new data. This model aids Simba Mobiles in predicting the normalized used price for new used mobile devices. A classification model was built to classify whether a used mobile device has a high or low resale value in the market. The forward stepwise logistic regression model was selected as the best classifier due to its high sensitivity rate, enabling accurate predictions of high resale value used mobile devices. This classification analysis further supports Simba Mobiles in improving their business by making informed decisions on purchasing. Clustering analysis was conducted to understand patterns and relationships between features of used mobile devices. This helped identify which clustered mobiles have either high or low resale value in the market. The clustering analysis will help the Simba mobiles while restocking from the wholesale market with specific features mobiles as grouped. These insights provide valuable suggestions to Simba Mobiles owners regarding which features of used mobile devices tend to have higher resale value. By utilizing all these analyses, Simba Mobiles owners can make wise investments in used mobile devices when purchasing from the wholesale market, leading to a more profitable business.