

***Project 2 - USED SMARTPHONE PRICE PREDICTION AND  
CLASSIFICATION***

## *Table of Contents*

1	Introduction .....	4
2	Business Opportunity .....	4
3	Analytical goals .....	4
3.1	Goal-1: Price Prediction.....	4
3.2	Goal-2: Price Classification .....	4
4	Data Preprocessing .....	4
4.1	Dataset description.....	4
4.2	Attributes Definition .....	5
5	Data Exploration.....	7
5.1	Check for Missing Values .....	7
5.2	Handling missing values .....	8
5.3	Check for Zeros.....	9
5.4	Summary Statistics for numerical attributes .....	10
5.5	Outliers.....	10
5.6	Categorical Attributes .....	12
5.7	Numerical Attributes Distributions .....	14
5.8	Target attributes distribution .....	17
6	Data Analysis .....	17
6.1	Relationship between Battery and Weight .....	17
6.2	Relationship between days_used and normalized_used_price .....	18
6.3	Unveiling Price Dynamics: A Tale of New and Used Mobile Devices Across Top Brands 20	
7	Data Transformation .....	21
7.1	Converting Categorical OS Information to Binary Representation .....	22
7.2	Creating categorical column from Days_used attribute.....	22
8	Feature Selection Methods .....	22
8.1	Correlation Analysis.....	22
8.2	Feature Selection Using Linear Regression with stepAIC .....	23
8.3	RFE: (Recursive feature selection) .....	23
8.4	Dimension Reduction : .....	24
9	Data Partitioning:.....	25
10	Model Selection: .....	26

10.1	For Goal 1: Price Prediction .....	26
10.1.1	Fitting Linear Regression model:.....	26
10.1.2	Model Evaluation:.....	26
10.1.3	Using step function to the linear model: .....	27
10.1.4	Comparison of performance evaluation:.....	27
10.1.5	Fitting Regression Tree model: .....	28
10.1.6	Model Evaluation:.....	29
10.1.7	Model Selection: Comparison of Linear Regression and Regression Tree Models: 29	
10.2	For Goal 2: Price Classification.....	30
10.2.1	Fitting Classification Tree Model: .....	30
10.2.2	Model Evaluation:.....	31
10.2.3	Fitting KNN (K-nearest neighbors) model: .....	32
10.2.4	Model Evaluation:.....	32
10.2.5	Models Comparison for Classification: .....	33
11	Model evaluation (of the selected models) on holdout dataset:.....	34
11.1	Goal-1 Prediction (Linear Regression): .....	34
11.2	Goal-2 Classification (K-Nearest Neighbors - All Predictors): .....	35
11.3	Results from Prediction and classification:.....	36
12	Features Impacting Mobile Device Prices based on Model Analysis are: .....	37
13	Conclusion: .....	37
14	Business Recommendations: .....	37
15	Executive Summary .....	38

## ***1 Introduction***

The project is about investigating the price of a used mobile phone while considering all the features impacting the consumer decision. The used mobile phone price is predicted using the machine learning algorithms and will classify the used mobile phones into two categories i.e., “Low” and “High”. This investigation helps the customers(stakeholders) to find which features are impacting the price of the mobile. By analyzing a bunch of data, we want to understand the relationships between these features and the prices of phones. It's like peeking into the secrets behind why phones cost what they do.

## ***2 Business Opportunity***

The business goal is to understand the price of used mobile phones and to figure out why people pay more for certain used phones. By using machine learning algorithms, we can predict and categorize the prices of these phones. This helps businesses know which features are popular, making it easier for them to sell more expensive phones and improve their overall sales. This can be good for the business because it helps them understand what features make people willing to pay more.

## ***3 Analytical goals***

### **3.1 Goal-1: Price Prediction**

The goal is to predict the price of used mobile phones using all the features or predictors.

### **3.2 Goal-2: Price Classification**

The goal is to classify the predicted price in goal-1 into two categories “Low” and “High”. If the predicted price is above average, then the mobile phone is considered as High price or else it is considered as Low price.

## ***4 Data Preprocessing***

### **4.1 Dataset description**

The dataset named "Used Mobiles" contains information about 3,454 used mobile phones. It includes both numerical and categorical attributes. The numerical attributes include details such as screen size, camera specifications (rear and front), internal memory, RAM, battery capacity, weight, release year, days used, and normalized prices. The categorical attributes cover information about the device brand, operating system (OS), and 4G/5G support. This dataset gives us a lot of information to explore and understand what makes a used mobile phone more or less expensive. It's like a big list where each row is a different phone, and each column tells us something specific about that phone.

## 4.2 Attributes Definition

**Device\_brand:** This attribute refers to the brand of each mobile phone, such as Honor, Samsung, or LG. This categorical attribute helps us identify which brand the mobile phones belong to, offering insights into brand preferences among users.

**OS:** The "os" attribute represents the operating system of each mobile phone, indicating whether it uses Android, iOS, or another operating system. This categorical attribute provides information about the software platform on which the mobile phone operates.

**Screen\_Size:** The "screen\_size" attribute is a numerical attribute, representing the size of the screen for each mobile phone. This value is measured in inches and provides insights into the physical dimensions of the phone's display. Larger screen sizes are often associated with devices optimized for media consumption and productivity.

**X4g and X5g:** These two variables indicate whether a mobile phone is compatible with 4G, 5G, or both. Both attributes are in binary format. ("yes" and "no").

Let's break down the meaning and implications of these variables:

### **X4g (4G Compatibility):**

If X4g is set to "yes," it means the mobile phone is compatible with 4G networks.

If X4g is set to "no," it indicates that the mobile phone does not support 4G connectivity.

### **X5g (5G Compatibility):**

If X5g is set to "yes," it means the mobile phone is compatible with 5G networks.

If X5g is set to "no," it indicates that the mobile phone does not support 5G connectivity.

### **Combinations:**

If both X4g and X5g are set to "yes," the phone is compatible with both 4G and 5G networks.

If X4g is "yes" and X5g is "no," the phone is only compatible with 4G.

If X4g is "no" and X5g is "yes," the phone is only compatible with 5G.

If both X4g and X5g are set to "no," the phone does not support either 4G or 5G connectivity.

**rear\_camera\_mp:** This attribute represents the resolution of the rear camera in megapixels, indicating the level of detail the camera can capture. Higher values suggest better image quality and clarity, contributing to improved photography and video recording capabilities. It is a numerical attribute.

**front\_camera\_mp:** The front\_camera\_mp attribute denotes the resolution of the front (selfie) camera in megapixels, influencing the quality of selfies and video calls. A larger value indicates a higher resolution front camera, potentially leading to better-quality self-portraits and video communication. It is a numerical attribute.

**internal\_memory:** Internal\_memory signifies the built-in storage capacity of the mobile device, measured in gigabytes (GB). More extensive internal memory allows users to store a greater number of apps, files, and media on their device. It is a numerical attribute.

**ram:** RAM (Random Access Memory) is represented by the ram attribute, indicating the device's short-term memory capacity in gigabytes. Higher RAM values contribute to smoother multitasking and improve overall performance of the mobile device. It is a numerical attribute.

**battery:** Battery represents the capacity of the device's battery in milliampere-hours (mAh), influencing the device's overall battery life. A larger battery capacity generally results in a longer duration between charges. It is a numerical attribute.

**weight:** Weight denotes the mass of the mobile device in grams, impacting its portability and user comfort. Lighter devices are generally more convenient for daily use and carrying. It is a numerical attribute.

**release\_year:** Release\_year indicates the year when the mobile device was launched, providing information on its age and technological features in relation to newer models. Newer release years may suggest more advanced technology and features.

**days\_used:** Days\_used represents the number of days the mobile device has been in use, offering insights into its usage history. This attribute can be relevant for assessing the device's condition and potential wear and tear.

**normalized\_used\_price:** Normalized\_used\_price signifies the relative price of the device after a certain period of usage. This numerical attribute provides insights into how the value of a phone changes over time, allowing us to understand the depreciation or appreciation in price based on its condition and usage.

**normalized\_new\_price:** Normalized\_new\_price represents the relative price of each mobile device when it is brand new. This numerical attribute helps us understand the initial market value of a phone, serving

as a reference point for comparison with the `normalized_used_price`. The `normalized_new_price` provides insights into how much value a phone loses or retains over time as it is being used.

### Sample Data: Consists of Numerical and Categorical data.

Figure 1 shows a dataset comprising both numerical and categorical attributes. The numerical attributes include `screen_size`, `rear_camera_mp`, `front_camera_mp`, `internal_memory`, `ram`, `battery`, `weight`, `release_year`, `days_used`, `normalized_used_price`, and `normalized_new_price`. On the other hand, the categorical attributes encompass `device_brand`, `os`, `X4g`, and `X5g`.

```
> head(numerical_attributes)
screen_size rear_camera_mp front_camera_mp internal_memory ram battery weight release_year days_used
1      14.50           13           5           64      3    3020    146         2020      127
2      17.30           13          16          128      8    4300    213         2020      325
3      16.69           13           8          128      8    4200    213         2020      162
4      25.50           13           8           64      6    7250    480         2020      345
5      15.32           13           8           64      3    5000    185         2020      293
6      16.23           13           8           64      4    4000    176         2020      223

normalized_used_price normalized_new_price
1          4.307572          4.715100
2          5.162097          5.519018
3          5.111084          5.884631
4          5.135387          5.630961
5          4.389995          4.947837
6          4.413889          5.060694

> head(categorical_attributes)
device_brand  os X4g X5g
1      Honor Android yes no
2      Honor Android yes yes
3      Honor Android yes yes
4      Honor Android yes yes
5      Honor Android yes no
6      Honor Android yes no
```

Figure 1. Sample Data

## 5 Data Exploration

### 5.1 Check for Missing Values

In this dataset, some information is missing for certain features like 'rear camera megapixels,' 'front camera megapixels,' 'internal memory,' 'RAM,' 'battery,' and 'weight' as shown in figure 2. This means that we don't have complete details for these aspects in some cases.

Dealing with missing values is important because it can affect our understanding of the data. One way to handle this is by filling in the missing values with reasonable estimates, like using the average value for that feature. However, we need to be careful about how we handle these gaps, as it can impact our analyses or predictions.

	Missing_Values
device_brand	0
os	0
screen_size	0
supports_4G	0
supports_5G	0
rear_camera_mp	179
front_camera_mp	2
internal_memory	4
ram	4
battery	6
weight	7
release_year	0
days_used	0
normalized_used_price	0
normalized_new_price	0

*Figure 2. Missing Values*

## 5.2 Handling missing values

To address missing values in the 'rear\_camera\_mp' variable of the 'Used\_Mobiles' dataset as shown in figure 2, employed the MICE (Multiple Imputation by Chained Equations) package in R. This approach considers relationships between variables and predicts missing values based on observed information. By fitting a linear model that considers various features such as 'screen\_size,' 'front\_camera\_mp,' 'internal\_memory,' 'ram,' 'battery,' 'weight,' 'release\_year,' and 'days\_used,' to impute missing 'rear\_camera\_mp' values more accurately.

The MICE package pooled results from multiple imputed datasets, providing a comprehensive summary of the imputation process. The final dataset was created, incorporating the imputed values for 'rear\_camera\_mp' and replacing the missing values as shown in figure 3. This meticulous approach ensures a more reliable and complete dataset for subsequent analyses, contributing to a more accurate understanding of mobile device characteristics.



	Missing_Values
device_brand	0
os	0
screen_size	0
supports_4G	0
supports_5G	0
rear_camera_mp	0
front_camera_mp	0
internal_memory	0
ram	0
battery	0
weight	0
release_year	0
days_used	0
normalized_used_price	0
normalized_new_price	0

*Figure 3. Handling Missing Values*

### 5.3 Check for Zeros

Figure 4 shows the occurrence of zero values in the 'front\_camera\_mp' column indicates that some mobile devices do not possess a front camera. While the numeric value is zero, it conveys valuable information about the absence of a front camera rather than being a data anomaly. Recognizing this, I handled these zero values by interpreting them as a meaningful representation of devices without front cameras.

	Zeros_Count
device_brand	0
os	0
screen_size	0
supports_4G	0
supports_5G	0
rear_camera_mp	0
front_camera_mp	39
internal_memory	0
ram	0
battery	0
weight	0
release_year	0
days_used	0
normalized_used_price	0
normalized_new_price	0

*Figure 4. Zeros Count*

## 5.4 Summary Statistics for numerical attributes

The summary statistics paint a vivid picture of the diverse mobile devices in our dataset as shown in figure 5. Screen sizes span from 5.08 to 30.71 inches, showcasing the array of choices consumers have. The camera capabilities vary widely, emphasizing the differences in megapixels for both rear and front cameras. Memory and processing power, indicated by internal memory and RAM, show a substantial range, catering to different user needs.

Battery capacities range from 500 mAh to 9720 mAh, reflecting varying power specifications. Device weight varies from 69.0 to 855.0 grams, highlighting the diversity in physical attributes. The dataset encompasses mobile phones released between 2013 and 2020, providing a temporal spread of devices. Usage durations ('days\_used').

```
> summary(numerical_attributes)
```

screen_size	rear_camera_mp	front_camera_mp	internal_memory	ram	battery	weight
Min. : 5.08	Min. : 0.08	Min. : 0.000	Min. : 0.01	Min. : 0.020	Min. : 500	Min. : 69.0
1st Qu.:12.70	1st Qu.: 5.00	1st Qu.: 2.000	1st Qu.: 16.00	1st Qu.: 4.000	1st Qu.:2100	1st Qu.:142.0
Median :12.83	Median : 8.00	Median : 5.000	Median : 32.00	Median : 4.000	Median :3000	Median :160.0
Mean :13.71	Mean : 9.46	Mean : 6.554	Mean : 54.57	Mean : 4.036	Mean :3133	Mean :182.8
3rd Qu.:15.34	3rd Qu.:13.00	3rd Qu.: 8.000	3rd Qu.: 64.00	3rd Qu.: 4.000	3rd Qu.:4000	3rd Qu.:185.0
Max. :30.71	Max. :48.00	Max. :32.000	Max. :1024.00	Max. :12.000	Max. :9720	Max. :855.0
	NA's :179	NA's :2	NA's :4	NA's :4	NA's :6	NA's :7

release_year	days_used	normalized_used_price	normalized_new_price
Min. :2013	Min. : 91.0	Min. :1.537	Min. :2.901
1st Qu.:2014	1st Qu.: 533.5	1st Qu.:4.034	1st Qu.:4.790
Median :2016	Median : 690.5	Median :4.405	Median :5.246
Mean :2016	Mean : 674.9	Mean :4.365	Mean :5.233
3rd Qu.:2018	3rd Qu.: 868.8	3rd Qu.:4.756	3rd Qu.:5.674
Max. :2020	Max. :1094.0	Max. :6.619	Max. :7.848

Figure 5. Summary Statistics for numerical data

## 5.5 Outliers

While examining the dataset, noticed some extreme values in attributes like battery capacity, internal memory, and device weight as shown in figure 6 and figure 7. However, these values might not necessarily be errors but rather reflect the diverse features of different mobile phones. For example, some phones may have larger batteries or more internal memory, and heavier weights could be due to specific designs or functionalities. Since these variations are common in the mobile phone market, I've chosen not to treat these values as outliers.

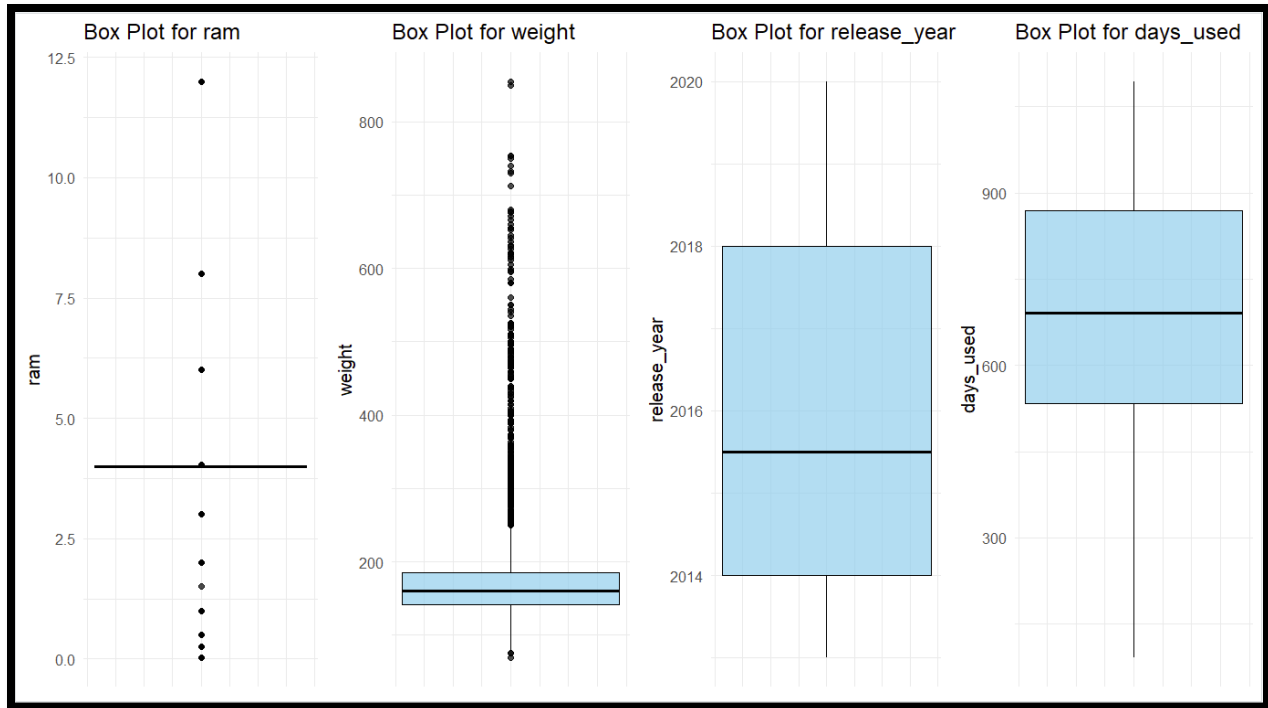


Figure 6. Boxplots

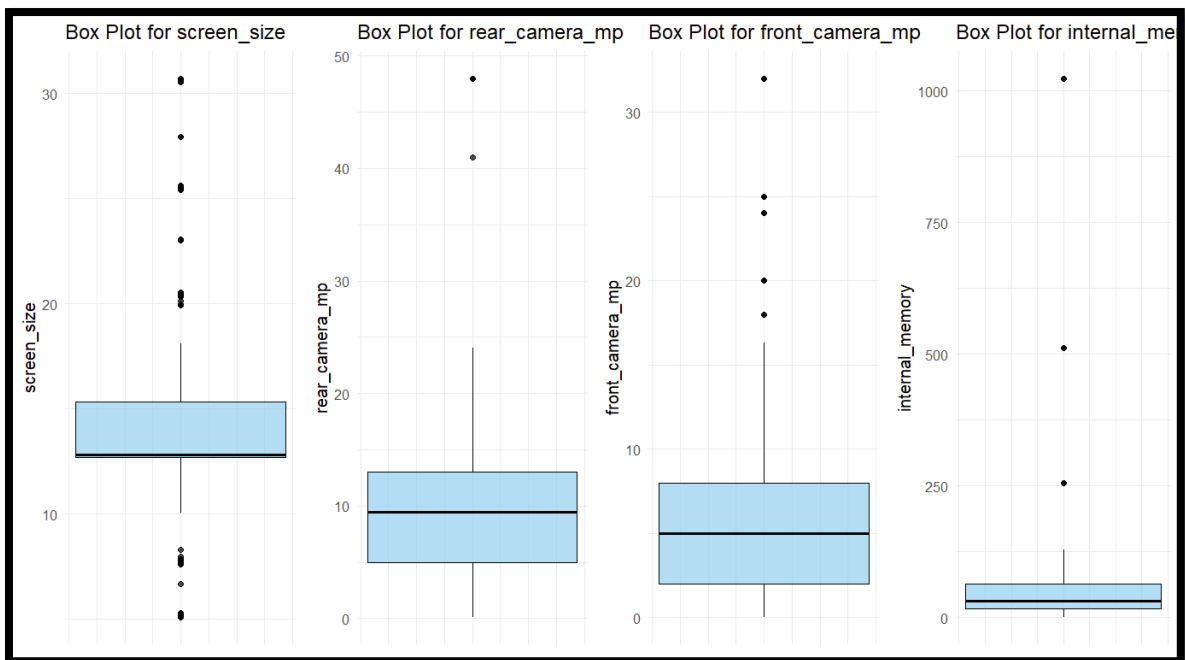


Figure 7. Boxplots

## 5.6 Categorical Attributes

### Device brand

In the bar plot, the term "Others" is used to group together all the different brands that are not specifically listed individually as shown in figure 8. The higher frequency in the "Others" category indicates that there are many different brands in the dataset, each with a lower occurrence compared to the more prominent brands like Samsung, Huawei, and LG. This simplification helps us see the general distribution of brands more easily.

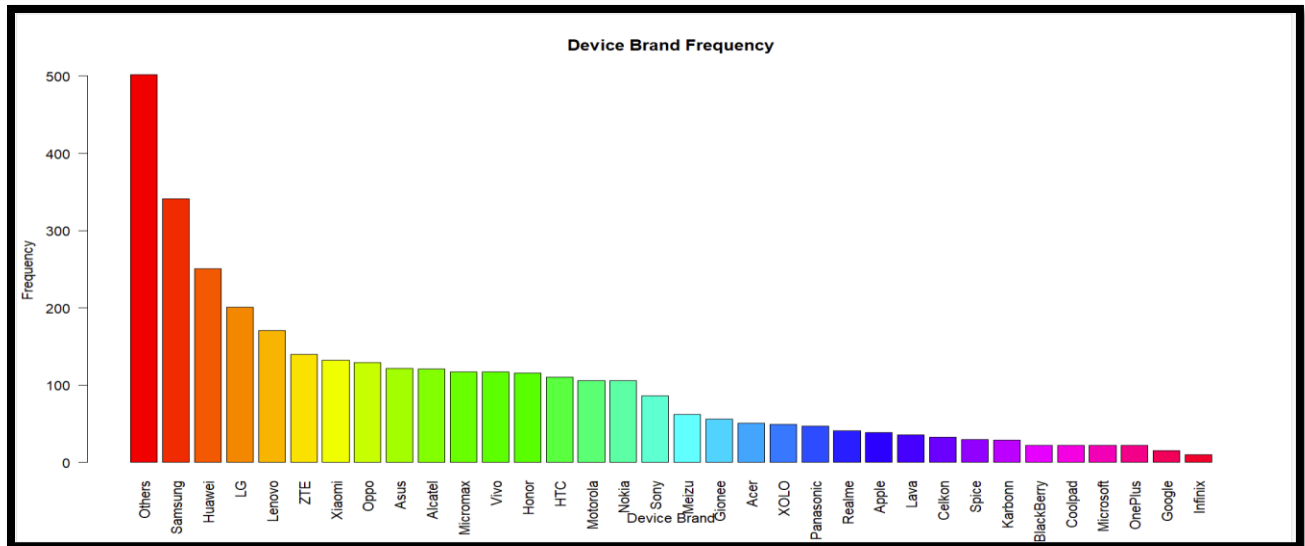
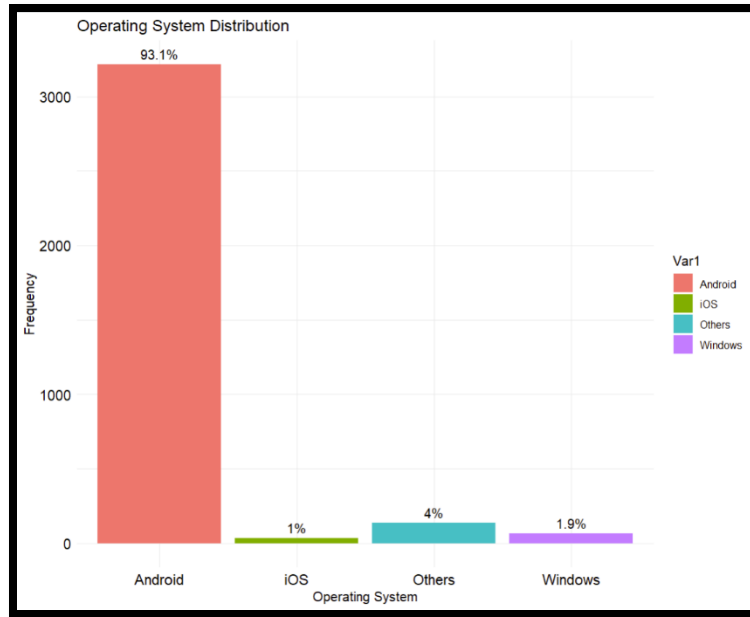


Figure 8. Device brand frequency

### OS (Operating System)

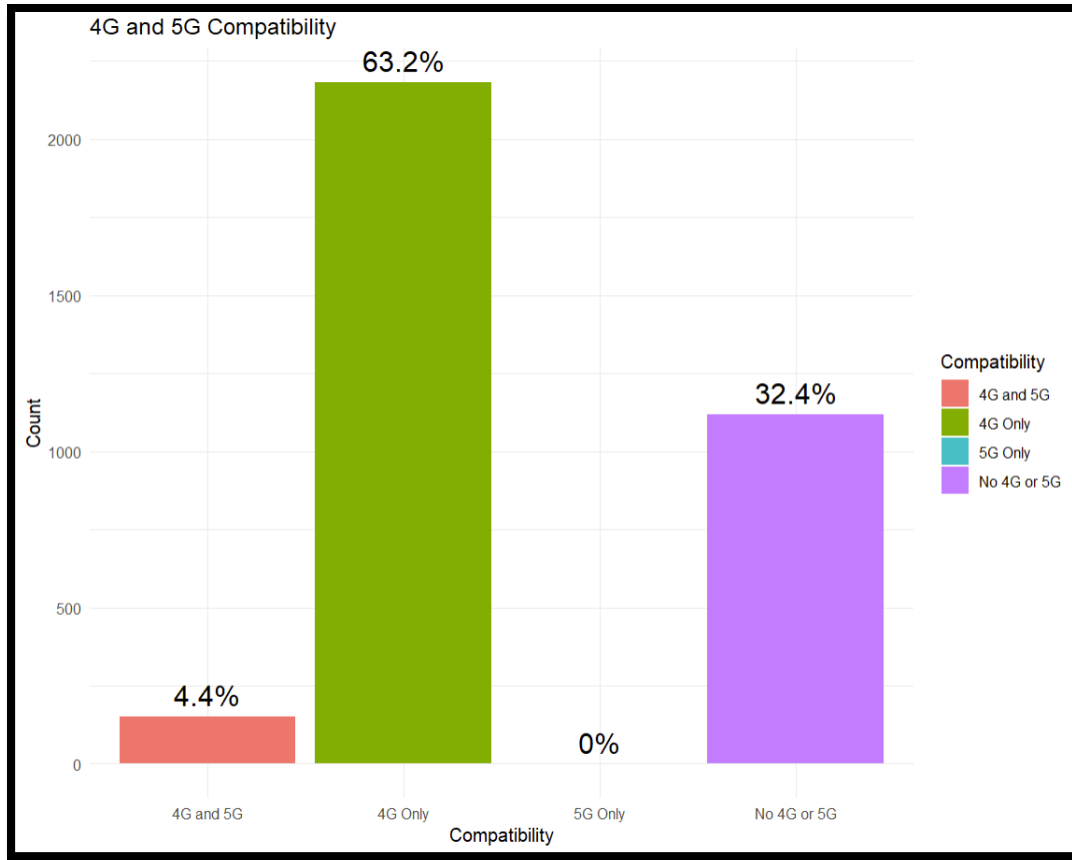
Looking at figure 9, it's evident that the majority of phones in our dataset use the Android operating system, making up a significant 93.1%. Apple's iOS, found in iPhones, is present in only 1% of the phones. A small 4% falls into the 'others' category, showcasing less common operating systems. Windows, a less popular choice, runs on a minimal 1.9% of the phones. Essentially, most people seem to prefer phones with Android, while other operating systems have a smaller share of our data.



*Figure 9. Operating System Distribution*

#### **4G and 5G compatibility**

The figure 10 shows different types of mobile phones based on how they connect to the internet. Most phones, about 63.2%, support only 4G. There are no phones that are made just for 5G. A small group, 4.4%, can connect to both 4G and 5G. Surprisingly, about one-third of the phones in the dataset (32.4%) don't support the faster 4G or 5G networks. This means these devices might rely on older network technologies or could be limited in terms of cellular data capabilities. This gives us a good picture of the different ways phones in the dataset connect to the internet.



*Figure 10. 4G and 5G Compatibility*

## 5.7 Numerical Attributes Distributions

The distribution of rear camera megapixels in the dataset reveals that the most common resolutions are 5 MP, 8 MP, or 12 MP. The distribution is skewed to the right, implying that while a majority of phones have lower megapixel counts, there exist a notable number of devices equipped with higher-resolution rear cameras. Similarly, the distribution of front camera megapixels indicates prevalent resolutions of 0 MP, 4 MP, 6 MP, 8 MP, or 15 MP on most phones. This distribution is right-skewed, signifying that the majority of phones tend to have lower megapixel counts for front cameras, but a significant subset features higher-resolution front cameras as shown in figure 11.

In terms of internal memory, the majority of phones exhibit smaller capacities. The distribution is right skewed with a tail extending towards higher values, illustrating that while most phones possess less internal memory, there are notable exceptions with substantial internal storage. Likewise, the distribution of RAM showcases a prevalence of phones with smaller RAM capacities. The distribution is right-skewed, indicating that while most phones lean towards lower RAM values, there are noteworthy instances of devices equipped with higher RAM capacities.

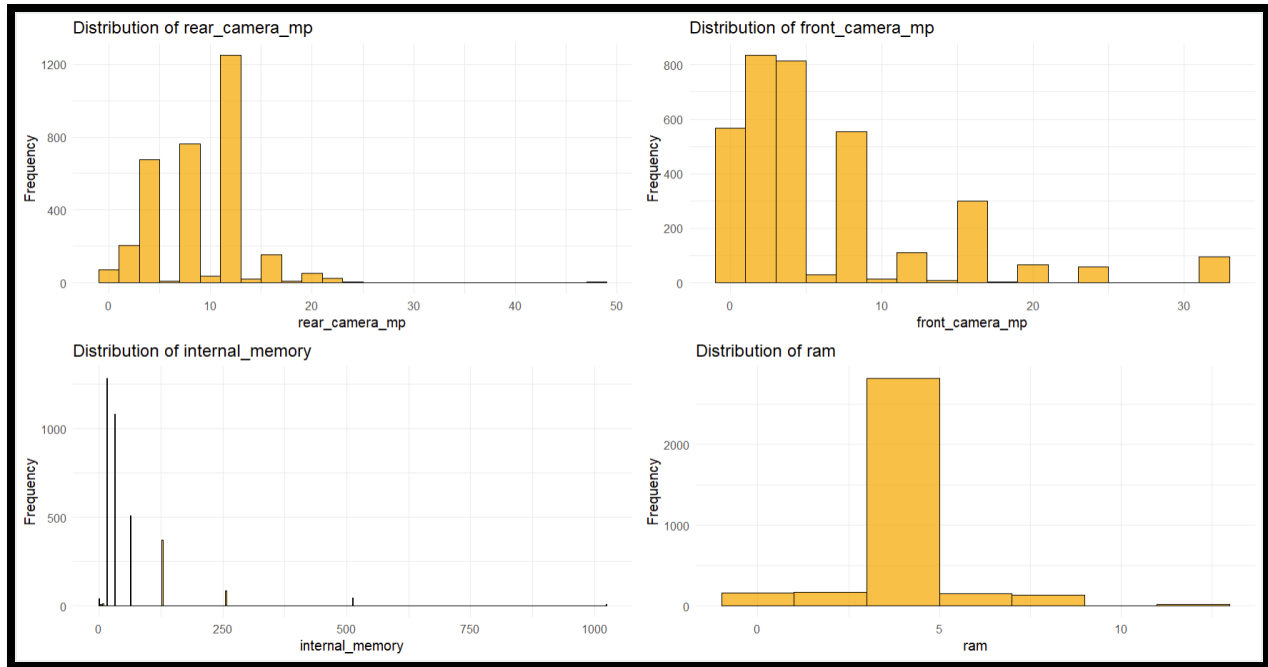


Figure 11. Histograms

### Distribution of battery

The histogram of battery capacities illustrates a right-skewed distribution, indicating that a majority of mobile devices have lower battery capacities as shown in figure 12. The peak of the distribution is observed in the range of 1500 to 3000, suggesting that mobiles with moderate battery capacities are more prevalent in the dataset. This information implies that while there are some devices with higher battery capacities, the majority fall within the lower to mid-range capacities.

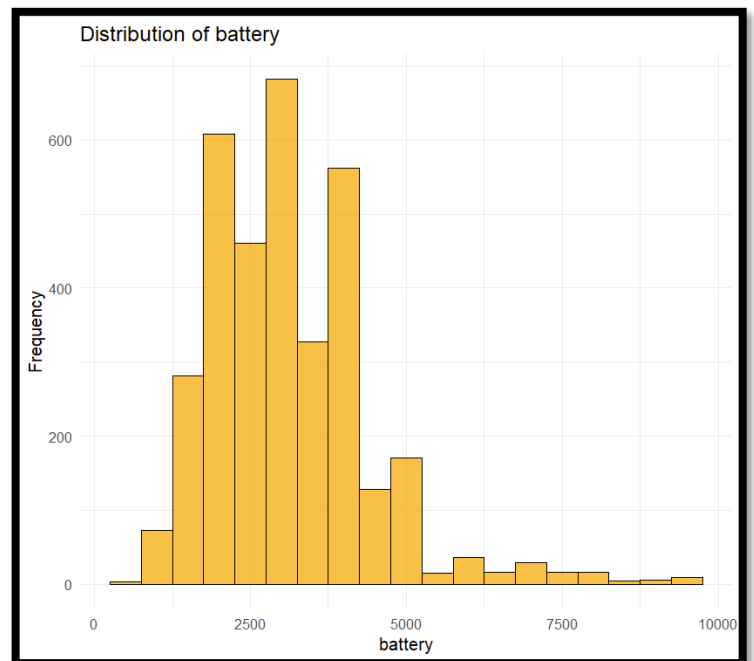


Figure 12. Battery Distribution

### Distribution of weight

The histogram depicting the distribution of weights showcases a right-skewed pattern in figure 13, signifying that a significant portion of mobile devices have lower weights. The peak of the distribution is concentrated in the range of 100 to 200, indicating that the majority of mobiles in the dataset are relatively lightweight.

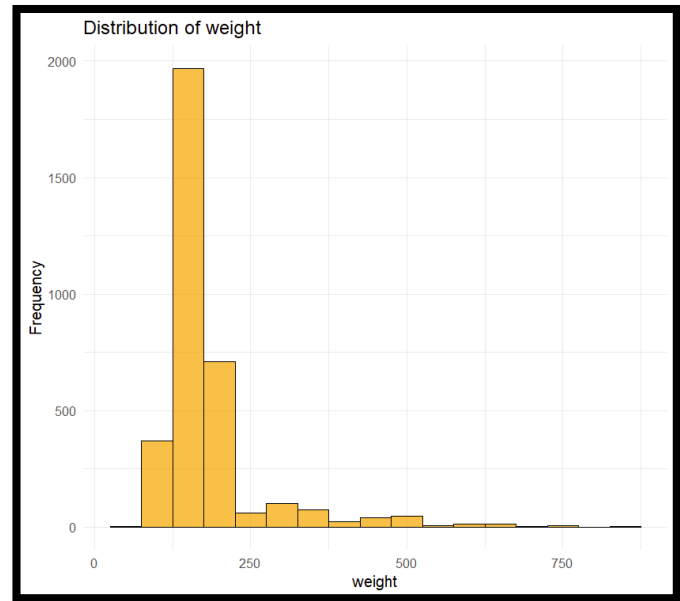


Figure 13. Weight Distribution

### Distribution of days\_used

The histogram for the "days\_used" attribute reveals that the distribution is skewed towards the left as shown in figure 14. This implies that the majority of mobile phones in the dataset have been used for a duration ranging from 600 to 1100 days. The left-skewed distribution indicates that while many phones fall within this usage duration, there are fewer instances of phones with extremely short periods of use. In summary, the plot highlights a common range of usage days among the sampled mobile devices.

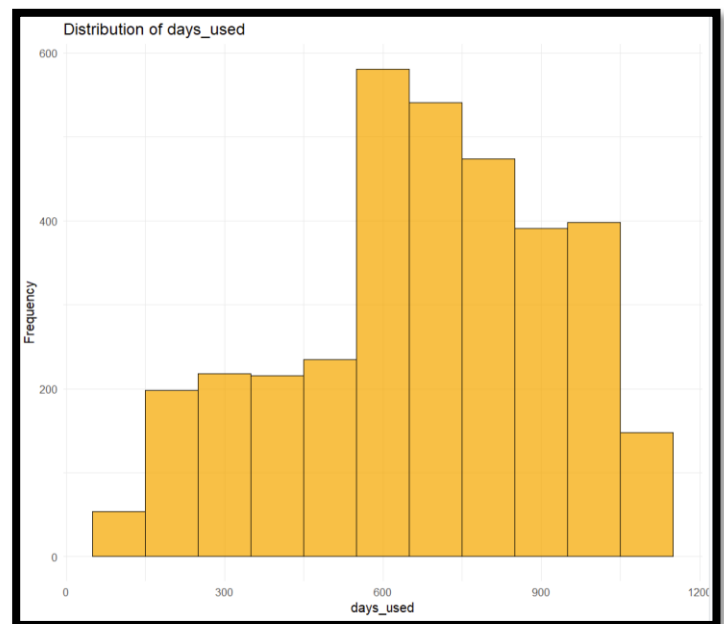


Figure 14. Distribution of days\_used



## 5.8 Target attributes distribution

The histograms for "normalized\_used\_price" and "normalized\_new\_price" indicate that the majority of used phones fall within the price range of 4 to 6. And new phones fall within the price range of 5 to 6. The skewness for used phones is slightly left shows that the used price for used phones is slightly lower than the new phones as shown in figures 15 and 16.

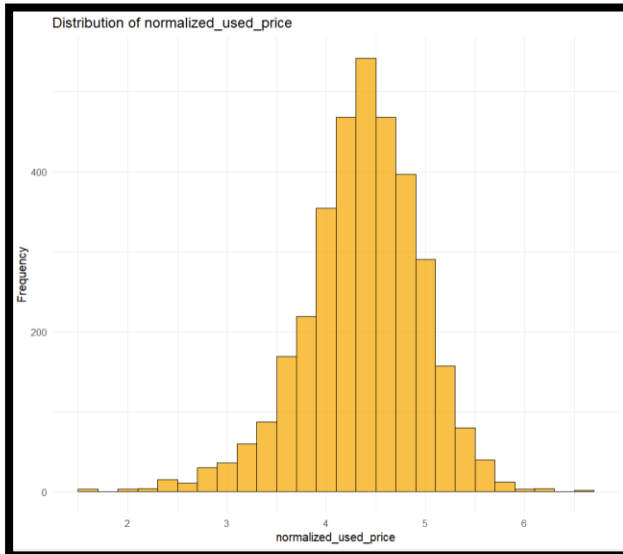


Figure 15. Distribution of normalized\_used\_price

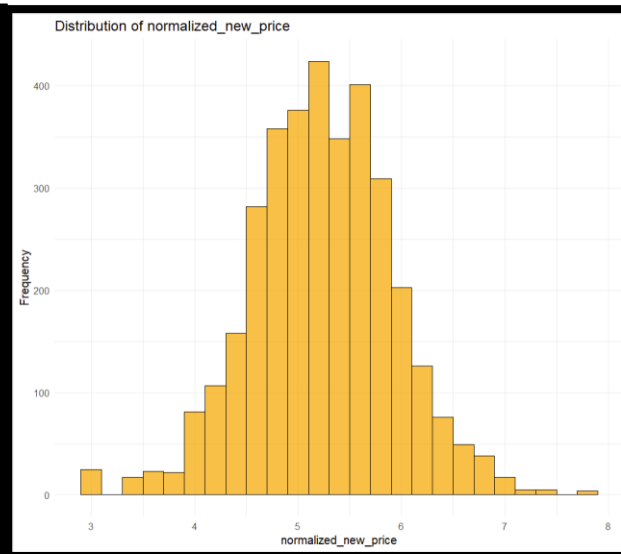
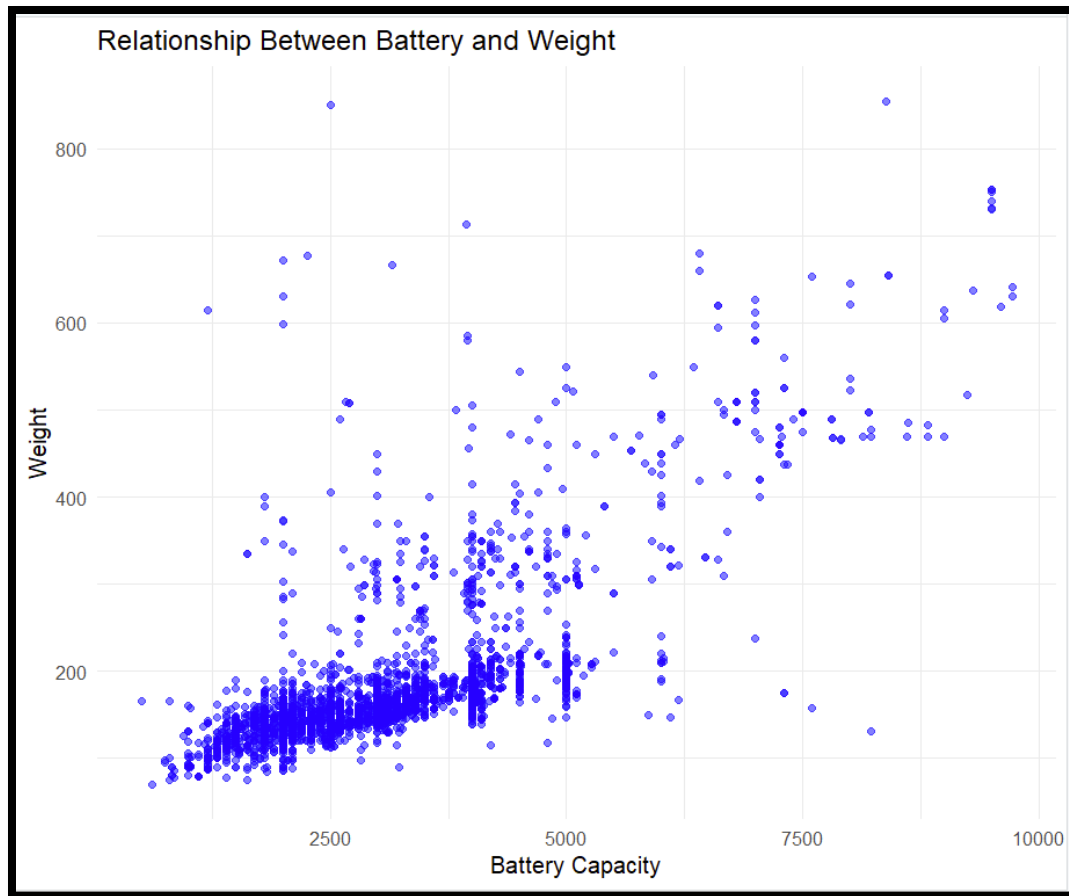


Figure 16. Distribution of normalized\_new\_price

## 6 Data Analysis

### 6.1 Relationship between Battery and Weight

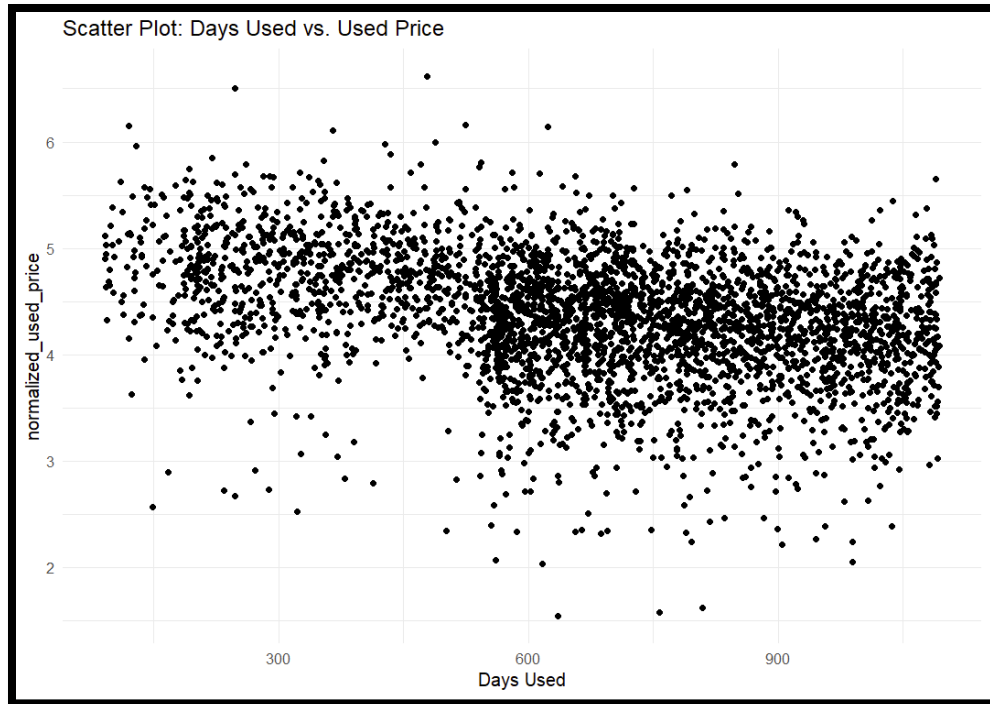
The scatter plot figure 17 clearly shows that when mobile phones have larger batteries, they also tend to be heavier. So, if a consumer wants a phone with a bigger battery that lasts longer, keep in mind that it might be a bit heavier. This information can be helpful for people who want to find the right balance between battery life and the overall weight of their mobile device.



*Figure 17. Scatterplot between Battery and Weight*

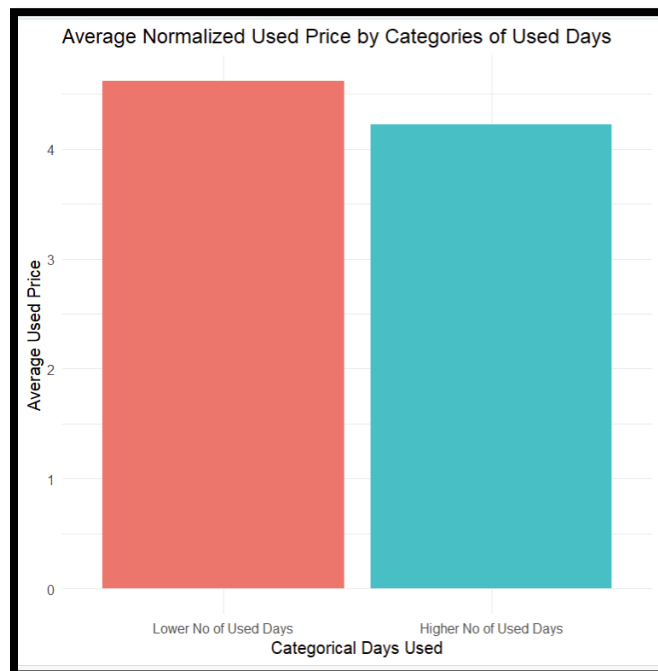
## 6.2 Relationship between days\_used and normalized\_used\_price

The scatter plot figure 18 illustrates a noticeable trend resembling a line, suggesting a potential relationship between the number of days a mobile device has been used (`days_used`) and its corresponding normalized used price (`normalized_used_price`). The clustering of data points in a pattern that resembles a line implies that as the usage duration increases, there may be a discernible impact on the normalized used price of mobile devices. This visual representation encourages further exploration into the quantitative relationship between these two attributes.



*Figure 18. Scatterplot between Days Used and Used Price*

We can visualize this more convenient way by Classifying Used Mobile Devices: Lower vs. Higher Number of Used Days as shown in figure 19.



*Figure 19. Bar plot for Average Normalized Used price by Categories of Used days*

As the number of days a mobile device is used goes up, we see a small drop in used device prices. This means that if a device has been used for a longer time, its price tends to decrease slightly. Though the effect is not significant, it indicates a minor connection between extended use and a slight reduction in device prices.

### 6.3 Unveiling Price Dynamics: A Tale of New and Used Mobile Devices Across Top Brands

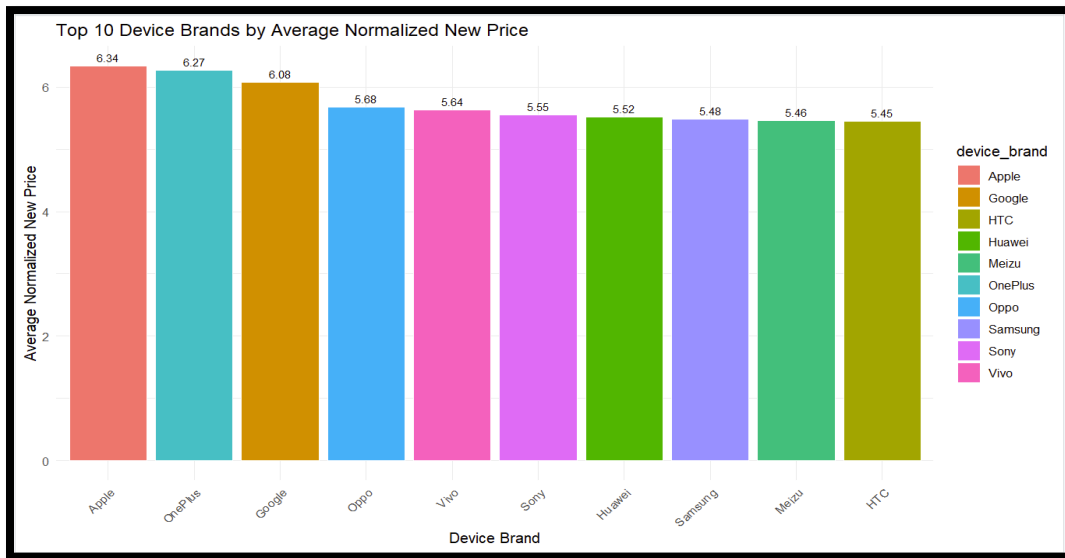


Figure 20. Top 10 Device Brands by Average Normalized Used Price

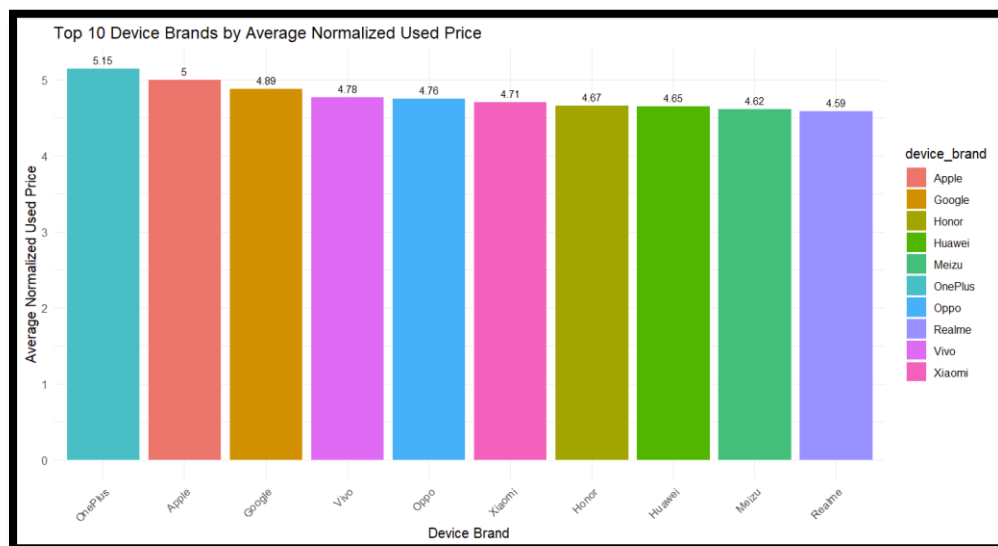


Figure 21. Top 10 Device Brands by Average Normalized New Price

After looking at figure 20 and figure 21, some interesting patterns emerged. For brands like OnePlus, the average price when they first release a new phone is around \$6.27. But, after people use the phone, the average price drops to about \$5.15 for a used device. Similarly, for Apple, the new phone starts at an average price of \$6.34, but the used one is around \$5.00. The used device price drop for used Apple phones seems to be bigger than for used OnePlus phones.

Surprisingly, Sony and Samsung, even though not in the top 10 for average used prices, tell a different story. It seems like when Sony and Samsung release a new phone, they start at a higher price compared to others. However, as people use these phones, their value drops more compared to brands in the top 10. This suggests that Sony and Samsung might see a bigger drop in value after people use their phones, even though they initially release them at a higher price.

## 7 Data Transformation

In the data transformation phase, focused on improving the clarity and readability of the dataset by renaming specific column names and changed the names of columns related to 4G and 5G connectivity from 'X4g' and 'X5g' to 'supports\_4G' and 'supports\_5G,' respectively. This adjustment was made to provide more meaningful and descriptive column names, contributing to a better understanding of the data. Importantly, it's essential to note that these changes in column names did not involve any alterations to the actual data values; they were solely aimed at enhancing the dataset's organization and interpretability.

**Before renaming:**

```
> colnames(Used_Mobiles)
[1] "device_brand"      "os"      "screen_size"
[4] "X4g"              "X5g"      "rear_camera_mp"
[7] "front_camera_mp"   "internal_memory" "ram"
[10] "battery"           "weight"    "release_year"
[13] "days_used"        "normalized_used_price" "normalized_new_price"
```

**After Renaming:**

```
> colnames(Used_Mobiles)
[1] "device_brand"      "os"      "screen_size"      "supports_4G"
[5] "supports_5G"       "rear_camera_mp" "front_camera_mp"  "internal_memory"
[9] "ram"              "battery"    "weight"           "release_year"
[13] "days_used"        "normalized_used_price" "normalized_new_price"
```

## 7.1 Converting Categorical OS Information to Binary Representation

In the data transformation process, utilizing the `fastDummies` library to create dummy variables for the "os" column in the dataset, specifically, for different operating system categories. This involves converting categorical OS information into binary form, where each OS category gets represented by a separate binary column. By doing this, the original categorical OS column is replaced with these dummy variables, allowing for a more effective representation of OS-related information in a format suitable for machine learning models.

## 7.2 Creating categorical column from Days\_used attribute

I introduced a new categorical column for 'days\_used,' distinguishing between 'Lower' and 'Higher' usage durations, to explore its impact on used mobile device prices. It's essential to note that this categorical column is specifically created for analysis purposes and has not been incorporated into any model for predicting prices. The original 'days\_used' column remains unaltered and continues to serve as a reference.

# 8 Feature Selection Methods

## 8.1 Correlation Analysis

The correlation matrix reveals the strength and direction of relationships between pairs of numerical attributes as shown in figure 22. There is a strong positive correlation (0.81) between screen size and battery capacity, suggesting that larger screen sizes tend to be associated with higher battery capacities. And, for battery and weight (0.71). Additionally, the correlation matrix shows a negative correlation (-0.75) between release year and days used, indicating that newer phones are likely to have been used for a shorter duration. These correlation coefficients provide valuable insights into the interdependence of various features in the dataset.

However, based on this correlation analysis alone, we cannot conclusively decide to reduce the number of dimensions in our dataset. The relationships between variables are indicative, but further analysis and techniques like dimensionality reduction

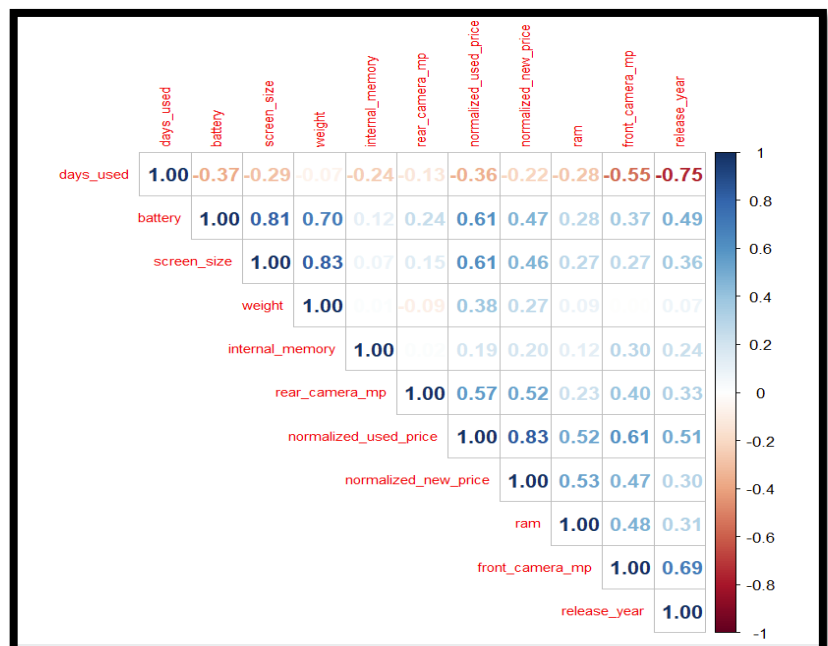
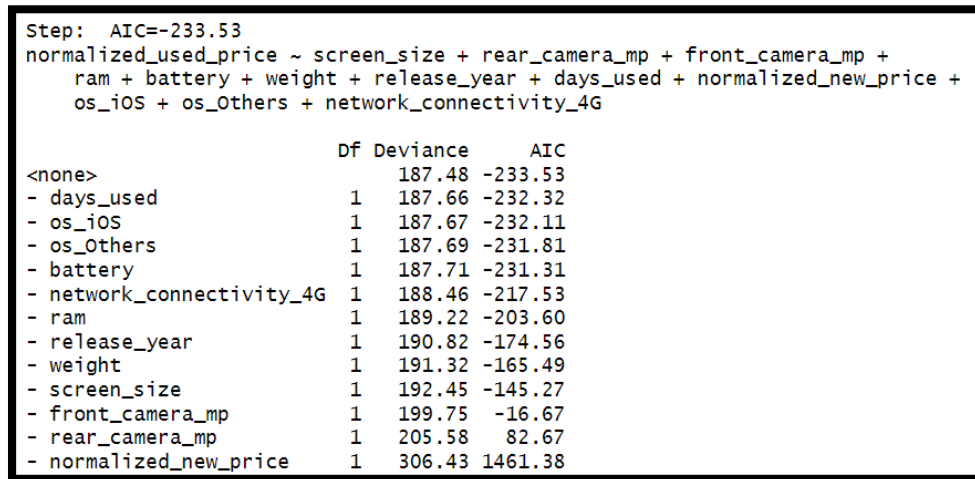


Figure 22. Correlation Analysis

methods would be needed to make informed decisions about feature selection.

## 8.2 Feature Selection Using Linear Regression with stepAIC

The stepwise backward selection process using AIC for building a predictive model of `normalized_used_price`. The final model includes variables like `screen_size`, `supports_4G`, `supports_5G`, `rear_camera_mp`, and others. The AIC-driven selection suggests that the inclusion of these variables optimally balances model complexity and goodness of fit. The process indicates that further removal of variables does not significantly enhance the model's performance.



The image shows a screenshot of an R console window. At the top, it says 'Step: AIC=-233.53' followed by the formula for the selected model: `normalized_used_price ~ screen_size + rear_camera_mp + front_camera_mp + ram + battery + weight + release_year + days_used + normalized_new_price + os_iOS + os_Others + network_connectivity_4G`. Below this is a table with three columns: 'Df', 'Deviance', and 'AIC'. The rows represent different models, starting with '<none>' and then listing variables to be removed one by one. The AIC values decrease as more variables are included, reaching a minimum of -233.53 for the full model.

	Df	Deviance	AIC
<none>		187.48	-233.53
- days_used	1	187.66	-232.32
- os_iOS	1	187.67	-232.11
- os_Others	1	187.69	-231.81
- battery	1	187.71	-231.31
- network_connectivity_4G	1	188.46	-217.53
- ram	1	189.22	-203.60
- release_year	1	190.82	-174.56
- weight	1	191.32	-165.49
- screen_size	1	192.45	-145.27
- front_camera_mp	1	199.75	-16.67
- rear_camera_mp	1	205.58	82.67
- normalized_new_price	1	306.43	1461.38

Figure 23. Linear Model Using StepAIC

## 8.3 RFE: (Recursive feature selection)

The recursive feature selection process, using a 10-fold cross-validated approach, assessed model performance for different numbers of variables. The results indicate that including 15 variables in the model yields the best performance as shown in figure 24, with the lowest Root Mean Squared Error (RMSE), highest R-squared value, and lowest Mean Absolute Error (MAE). The top 5 contributing variables for this optimal model are `normalized_new_price`, `weight`, `screen_size`, `rear_camera_mp`, and `internal_memory`. These variables are identified as the most influential in achieving accurate predictions for the given task.

Recursive feature selection						
Outer resampling method: Cross-Validated (10 fold)						
Resampling performance over subset size:						
Variables	RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD Selected
5	0.2290	0.8507	0.1784	0.01529	0.01766	0.010317
8	0.2254	0.8542	0.1769	0.01415	0.01494	0.009721
10	0.2229	0.8573	0.1756	0.01373	0.01401	0.009175
12	0.2225	0.8579	0.1754	0.01332	0.01301	0.008900
15	0.2218	0.8588	0.1751	0.01355	0.01340	0.009351 *
16	0.2220	0.8585	0.1749	0.01380	0.01375	0.009217
The top 5 variables (out of 15):						
normalized_new_price, weight, screen_size, rear_camera_mp, internal_memory						

Figure 24. Result of RFE

The Recursive Feature Elimination (RFE) plot indicates that as the number of variables increases, the Root Mean Squared Error (RMSE) decreases, showcasing a clear and easily interpretable trend of improved model performance with a larger set of variables.

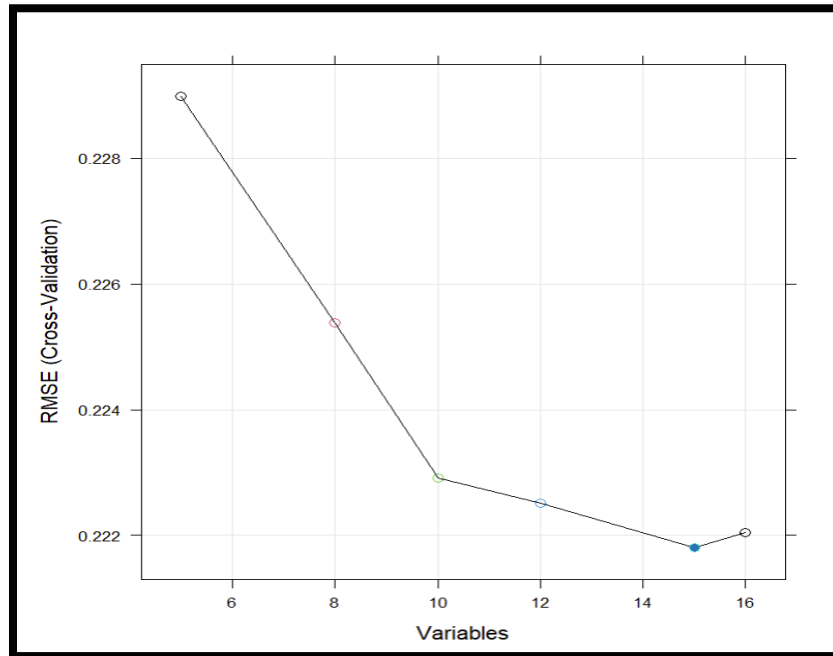


Figure 25. RMSE vs Variables

## 8.4 Dimension Reduction :

In the process of dimension reduction, the decision was made to exclude the 'device\_brand' attribute, which encompasses 35 brand names and is categorical. Including this attribute would require creating 35 dummy variables, adding complexity to the model, and potentially leading to misinterpretations of the data.



Upon reviewing the Recursive Feature Elimination (RFE) results for remaining attributes from figure 24, it is observed that the Root Mean Squared Error (RMSE) values for models with 15 and 16 variables doesn't make much difference in terms of prediction accuracy.

Additionally, the results from the Linear Model using Stepwise Akaike Information Criterion (AIC) revealed 13 predictors as shown in figure 23. Given that the dataset comprises all essential features of mobile devices crucial for prediction, the decision was reaffirmed to keep all predictors. This approach aims to maintain a comprehensive representation of features without further reduction, aligning with the goal of achieving accurate predictions in the model.

## 9 Data Partitioning:

In this dataset comprising 3454 observations, the data partitioning involves dividing it into three subsets: the training set, validation set, and holdout (test) set. The training set, comprising approximately 40% of the data (1381 observations), is utilized for training the machine learning model. The validation set, consisting of around 35% of the data (1207 observations), is employed for fine-tuning the model's hyperparameters. Lastly, the holdout set, comprising approximately 25% of the data (866 observations), remains untouched during the model-building process and serves as an independent evaluation set to assess the model's generalization performance on previously unseen data as shown in figure 25.

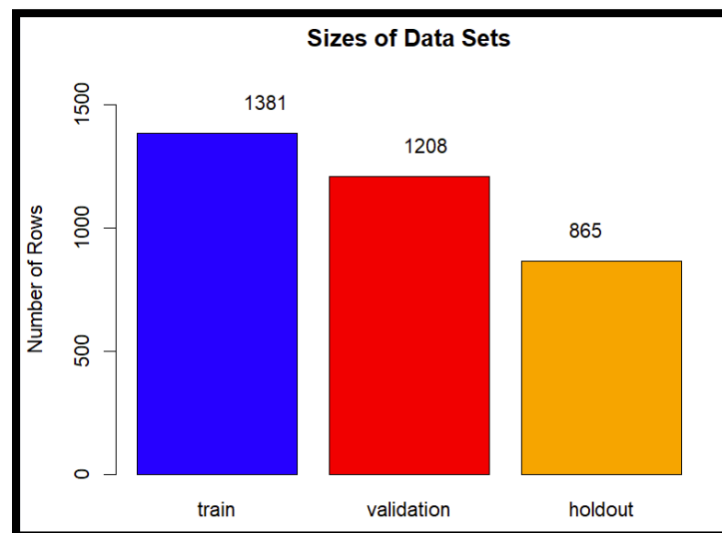


Figure 25. Data Partitions

## ***10 Model Selection:***

### **10.1 For Goal 1: Price Prediction**

For the primary goal of predicting used mobile device prices, two regression algorithms are planned for implementation:

#### **10.1.1 Fitting Linear Regression model:**

A Linear Regression model predicts numeric outcomes by establishing a linear relationship between independent and dependent variables, using coefficients to quantify the impact of each predictor. It leverages this linear equation to make predictions, providing a straightforward and interpretable approach for regression tasks.

The linear regression model was employed to predict normalized used prices for mobile devices. The model's residuals exhibit a distribution with a minimum value of -1.48 and a maximum of 1.21. The interquartile range (IQR) spans from -0.14 to 0.17, indicating a relatively balanced spread around the median. Notable coefficients with statistical significance include 'screen\_size,' 'rear\_camera\_mp,' 'front\_camera\_mp,' 'weight,' 'release\_year,' and 'normalized\_new\_price.' These factors are identified as impactful contributors to the predicted prices. Some coefficients, such as 'internal\_memory,' 'ram,' 'battery,' and certain OS and network connectivity categories, appear statistically insignificant as their p-values exceed the 0.05 significance threshold.

#### **10.1.2 Model Evaluation:**

The model achieved a high level of prediction accuracy, as evidenced by an Adjusted R-squared value of 0.8433. This metric indicates that approximately 84.33% of the variability in normalized used device prices is explained by the model. While generating predictions on the validation data, a warning message was observed, indicating a rank-deficient fit. This suggests potential multicollinearity issues or linear dependencies among predictor variables.

The model's performance was evaluated using the RMSE metric, resulting in a value of approximately 0.238. RMSE measures the average magnitude of errors between predicted and observed values. Additionally, the Mean Absolute Error (MAE) was calculated, yielding a value of around 0.185. MAE represents the average absolute differences between predicted and actual values. In conclusion, the linear regression model demonstrates strong predictive capabilities for used device prices, with a focus on key features like screen size, camera specifications, weight, release year, and normalized new price. Despite some insignificant predictors and a warning about potential rank deficiency, the model provides valuable insights and accurate predictions for normalized used prices.

### 10.1.3 Using step function to the linear model:

The stepwise regression technique was employed to enhance the linear regression model for predicting normalized used prices of mobile devices. The selected predictors, including screen size, camera specifications (rear and front megapixels), RAM, weight, release year, normalized new price, and network connectivity for 4G as shown in below table, were determined based on the Akaike Information Criterion (AIC). This process led to a more streamlined and effective model, with an adjusted R-squared value of 0.8438, signifying that approximately 84.38% of the variation in normalized used prices is accounted for by the chosen predictors.

	Df	Sum of Sq	RSS	AIC
none>			77.193	-3965.2
network_connectivity_4G	1	0.281	77.474	-3962.1
ram	1	0.494	77.686	-3958.4
weight	1	1.024	78.217	-3949.0
release_year	1	1.829	79.021	-3934.8
screen_size	1	2.660	79.853	-3920.4
front_camera_mp	1	4.608	81.801	-3887.1
rear_camera_mp	1	6.154	83.346	-3861.2
normalized_new_price	1	63.038	140.231	-3142.7

When applied to the validation dataset, the refined model demonstrated robust predictive performance, reflected in a Root Mean Squared Error (RMSE) of approximately 0.2379 and a Mean Absolute Error (MAE) of around 0.1854. These metrics underscore the model's accuracy in estimating normalized used prices, confirming the success of the stepwise regression in addressing multicollinearity and refining the predictive capabilities of the linear regression model.

### 10.1.4 Comparison of performance evaluation:

Model	Adjusted R-squared	RMSE	MAE
Linear Regression	0.8433	0.2379	0.1854
Stepwise Linear Model	0.8438	0.2379	0.1854

Table 1. Performance Evaluation

The comparison of performance evaluation between the original Linear Regression model and the Stepwise Linear Model indicates that both models have similar accuracy in predicting normalized used device prices as shown in table 1. While the Stepwise Linear Model shows a slightly higher adjusted R-squared value of 0.8438, the differences in RMSE (0.2379) and MAE (0.1854) are marginal compared to the original Linear Regression model, which has an RMSE of 0.2379 and MAE of 0.2379 on the validation dataset. Therefore, considering the minimal improvement and for simplicity, I have selected the original Linear Regression model for predicting normalized used prices.

#### 10.1.5 Fitting Regression Tree model:

A Regression Tree is a predictive modeling algorithm that partitions the dataset into hierarchical, tree-like structures based on feature values. It recursively splits the data into subsets, optimizing for reduced variance in the target variable within each partition. Predictions are made by traversing the tree, starting from the root, and navigating to the leaf node corresponding to the input feature values, where the average target variable value of the training samples in that leaf is used as the prediction.

The Regression Tree model was constructed using the rpart library as shown in figure 26, revealing significant predictors influencing normalized used mobile device prices. Notably, 'normalized\_new\_price,' 'screen\_size,' 'battery', 'rear\_camera\_mp' and 'front\_camera\_mp' emerged as pivotal variables within the model's decision-making process. Employing a maximum depth of 5 to prevent overfitting, the tree structure adeptly captured intricate, non-linear relationships between various features and normalized used prices.

Variable importance		
normalized_new_price	screen_size	battery
30	13	11
rear_camera_mp	front_camera_mp	internal_memory
10	10	6
network_connectivity_other	release_year	days_used
6	4	3
weight	ram	
3	2	

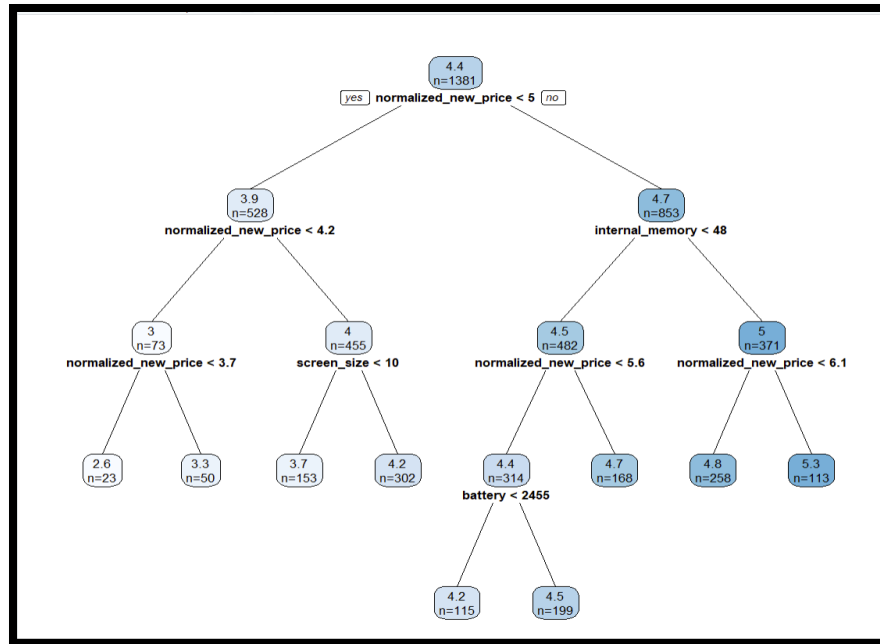


Figure 26. Regression Tree

#### 10.1.6 Model Evaluation:

After model fitting, apply the model on validation data which will predict the used price. Upon evaluation of the model's performance on the validation dataset, the Mean Absolute Error (MAE) was computed at 0.2323, denoting the average absolute difference between predicted and actual prices. Additionally, the Root Mean Squared Error (RMSE) measured at 0.2958 gauges the model's precision in predicting normalized used prices. However, despite these promising metrics, further attempts to enhance the model's efficiency through pruning, with a specified Complexity Parameter (cp) of 0.01, did not yield discernible improvements. The pruned tree structure retained the same performance metrics as the original, suggesting that the initial model already struck an optimal balance between complexity and predictive accuracy.

#### 10.1.7 Model Selection: Comparison of Linear Regression and Regression Tree Models:

Both the Regression Tree and Linear Regression models highlight similar key predictors for normalized used mobile device prices. Notably, 'normalized\_new\_price,' 'screen\_size,' and 'battery' are consistently identified as crucial variables in both models, emphasizing their significant impact on predicting device prices.

Model	RMSE	MAE
-------	------	-----

Model	RMSE	MAE
Linear Regression Model	0.2379	0.1854
Regression Tree Model	0.2958	0.2323

*Table 2. Model Comparison*

When comparing the performance of the Linear Regression and Regression Tree models, the Linear Regression model emerges as the preferred choice due to its superior predictive accuracy as shown in table 2. The Linear Regression model achieves a higher adjusted R-squared value of 0.8438 and demonstrates enhanced precision on the validation dataset with a lower RMSE of 0.2379 and MAE of 0.1854, outperforming the Regression Tree model with an RMSE of 0.2958 and MAE of 0.2323. Consequently, for the purpose of predicting normalized used prices, the Linear Regression model is selected for its improved performance and simplicity.

### **Selected Model for Goal-1: Linear Regression Model.**

## **10.2 For Goal 2: Price Classification**

The used prices of mobile devices have been transformed into a categorical variable, distinguishing between 'High' and 'Low' categories. This conversion enables the construction of Classification Tree models that leverage features to efficiently classify mobile devices based on their used price categories. The models aim to predict whether a used device falls into the 'High' or 'Low' price range, facilitating insightful analysis and decision-making in the context of used mobile device pricing.

For Goal 2, which involves classifying used mobile device prices, two classification algorithms are utilized:

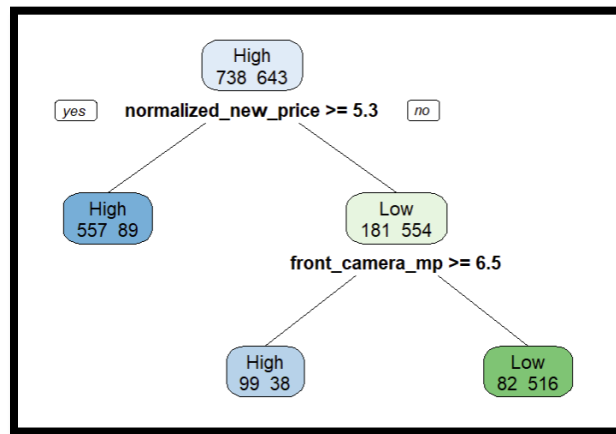
### **10.2.1 Fitting Classification Tree Model:**

A Classification Tree is a predictive model that facilitates decision-making by hierarchically partitioning data into subsets based on feature values, ultimately assigning each observation to a specific class. It helps understand the main factors influencing prices and simplifies the process of identifying why some devices are priced higher or lower.

A Classification Tree model was constructed to predict the categorical used price of mobile devices (High or Low) using features like 'normalized\_new\_price,' 'front\_camera\_mp,' 'battery' and others. The tree effectively splits the dataset based on feature values, creating decision nodes that classify

devices into High or Low-price categories. The key variables influencing the classification were identified, with 'normalized\_new\_price' having the highest importance.

The initial tree had a complexity parameter (cp) of 0.5801, and it was pruned to prevent overfitting, resulting in a pruned tree with  $cp = 0.01$ . The pruned tree maintained similar predictive performance but with increased interpretability.



*Figure 27. Pruned Classification Tree*

### 10.2.2 Model Evaluation:

When applied to the validation dataset, the Classification Tree achieved an accuracy of 83.03%, outperforming a naive classification based on the prevalence of High and Low prices. The confusion matrix indicated that the model had high sensitivity (88.08%) and specificity (77.68%), showcasing its ability to correctly identify both High and Low-priced devices from figure 28. The positive predictive value was 80.68%, highlighting the reliability of the model in predicting High-priced devices. Overall, the Classification Tree demonstrated effectiveness in categorizing mobile devices based on their features into High or Low-price segments.

Confusion Matrix and Statistics		
	Reference	
Prediction	1	0
1	547	131
0	74	456
Accuracy : 0.8303		
95% CI : (0.8079, 0.8511)		
No Information Rate : 0.5141		
P-Value [Acc > NIR] : < 0.00000000000000022		
Kappa : 0.6594		
McNemar's Test P-Value : 0.00009184		
Sensitivity : 0.8808		
Specificity : 0.7768		
Pos Pred Value : 0.8068		
Neg Pred Value : 0.8604		
Prevalence : 0.5141		
Detection Rate : 0.4528		
Detection Prevalence : 0.5613		
Balanced Accuracy : 0.8288		
'Positive' class : 1		

Figure 28. Confusion Matrix from classification tree

### 10.2.3 Fitting KNN (K-nearest neighbors) model:

The k-Nearest Neighbors (k-NN) models were trained and evaluated for classifying the price category of used mobile devices using the provided data. Two models were built: one using all predictor variables and another using a subset of predictors selected through stepwise feature selection (screen size, rear camera megapixels, front camera megapixels, RAM, weight, release year, normalized new price, and network connectivity).

### 10.2.4 Model Evaluation:

For the k-NN model trained with all predictors, the best performing model with 'k' set to 9 achieved an accuracy of 85.68% on the validation dataset. The confusion matrix revealed that out of 635 instances of 'High' price category, 542 were correctly classified, while out of 572 instances of 'Low' price category, 493 were correctly classified. The model demonstrated sensitivity of 87.28% and specificity of 83.99%, indicating its ability to effectively classify both price categories.



**KNN Model with all Predictors**

Confusion Matrix and Statistics		
Reference		
Prediction	1	0
1	542	94
0	79	493
Accuracy : 0.8568		
95% CI : (0.8358, 0.8761)		
No Information Rate : 0.5141		
P-Value [Acc > NIR] : <0.0000000000000002		
Kappa : 0.7131		
McNemar's Test P-Value : 0.2871		
Sensitivity : 0.8728		
Specificity : 0.8399		
Pos Pred Value : 0.8522		
Neg Pred Value : 0.8619		
Prevalence : 0.5141		
Detection Rate : 0.4487		
Detection Prevalence : 0.5265		
Balanced Accuracy : 0.8563		
'Positive' Class : 1		

**KNN Model with important predictors**

Confusion Matrix and Statistics		
Reference		
Prediction	1	0
1	554	96
0	67	491
Accuracy : 0.8651		
95% CI : (0.8445, 0.8838)		
No Information Rate : 0.5141		
P-Value [Acc > NIR] : <0.0000000000000002		
Kappa : 0.7296		
McNemar's Test P-Value : 0.0283		
Sensitivity : 0.8921		
Specificity : 0.8365		
Pos Pred Value : 0.8523		
Neg Pred Value : 0.8799		
Prevalence : 0.5141		
Detection Rate : 0.4586		
Detection Prevalence : 0.5381		
Balanced Accuracy : 0.8643		
'Positive' Class : 1		

Similarly, the k-NN model trained with selected predictors yielded an accuracy of 86.51% on the validation dataset with 'k' set to 9. In this model, out of 620 instances of 'High' price category, 554 were correctly classified, while out of 558 instances of 'Low' price category, 491 were correctly classified. This model exhibited sensitivity of 89.21% and specificity of 83.65%, indicating slightly improved performance compared to the model using all predictors.

Overall, both k-NN models demonstrated strong predictive performance in classifying the price category of used mobile devices, with the model trained on selected predictors showing slightly better accuracy and sensitivity. These models can be utilized to effectively categorize mobile devices based on their price range, facilitating decision-making processes for buyers and sellers in the mobile device market.

**10.2.5 Models Comparison for Classification:**

Models	Accuracy	Sensitivity (High)	Specificity (Low)
Classification Tree	83.03%	88.08%	77.68%
KNN All Predictors	85.68%	87.28%	83.99%
KNN with Selected Predictors	86.51%	89.21%	83.65%

*Table 3. Classification Model Comparison*

The comparison of the Classification Tree and K-Nearest Neighbors (KNN) models reveals that all models performed reasonably well in predicting the categorical used prices of mobile devices. Among

them, the KNN model utilizing selected predictors achieved the highest accuracy at 86.51%, surpassing both the Classification Tree (83.03%) and the KNN model with all predictors (85.68%). This model exhibited superior sensitivity (89.21%) and specificity (83.65%), indicating its effectiveness in correctly classifying both High and Low-price categories as shown in table 3. Notably, sensitivity is crucial in correctly identifying High-priced devices, while specificity ensures accurate classification of Low-priced devices.

Choosing the K-Nearest Neighbors (KNN) model with All Predictors as the best means that even though KNN model with only important predictors showed a slight improvement in accuracy, it's important to include all available information for a more comprehensive understanding. This ensures the model is well-equipped to make accurate predictions on new data, aligning with the goal of providing reliable insights for categorizing mobile devices based on their price range.

## ***11 Model evaluation (of the selected models) on holdout dataset:***

### **11.1 Goal-1 Prediction (Linear Regression):**

Model	RMSE	MAE
Linear Regression	0.2258	0.1778

The selected Linear Regression model, trained and evaluated on the holdout dataset, displayed excellent predictive accuracy. The Root Mean Squared Error (RMSE) was calculated at 0.2258, indicating the average magnitude of errors between the predicted and actual normalized used mobile device prices. Additionally, the Mean Absolute Error (MAE) measured at 0.1778, highlighting the model's ability to make predictions with a high degree of precision. These results affirm the reliability of the Linear Regression model in predicting the normalized used prices of mobile devices and its suitability for applications in the broader mobile device market.

## 11.2 Goal-2 Classification (K-Nearest Neighbors - All Predictors):

Metric	Value
Overall Accuracy	86.13%
Sensitivity (Recall)	87.50%
Specificity	84.54%
Kappa Value	0.7209

For classification purposes, the K-Nearest Neighbors (KNN) model, utilizing all available predictors, exhibited robust performance on the holdout dataset. The model achieved an overall accuracy of 86.13%, effectively classifying mobile devices into High and Low-price categories. The confusion matrix revealed a high sensitivity of 87.50%, indicating the model's proficiency in identifying High-priced devices. Specificity stood at 84.54%, demonstrating the ability to correctly classify Low-priced devices. With a Kappa value of 0.7209, the model showcases substantial agreement beyond what might occur by chance.

The KNN model's high AUC value of 0.855 means it's good at distinguishing between High and Low-priced mobile devices, confirming its reliability in categorizing them accurately. Overall, the KNN model, considering all predictors, proved effective in classifying used mobile devices based on their price categories.

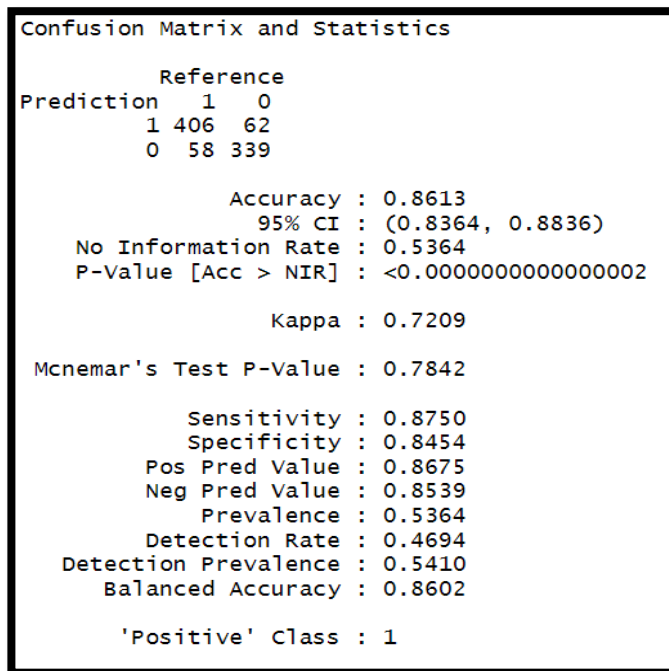


Figure 30. Confusion Matrix

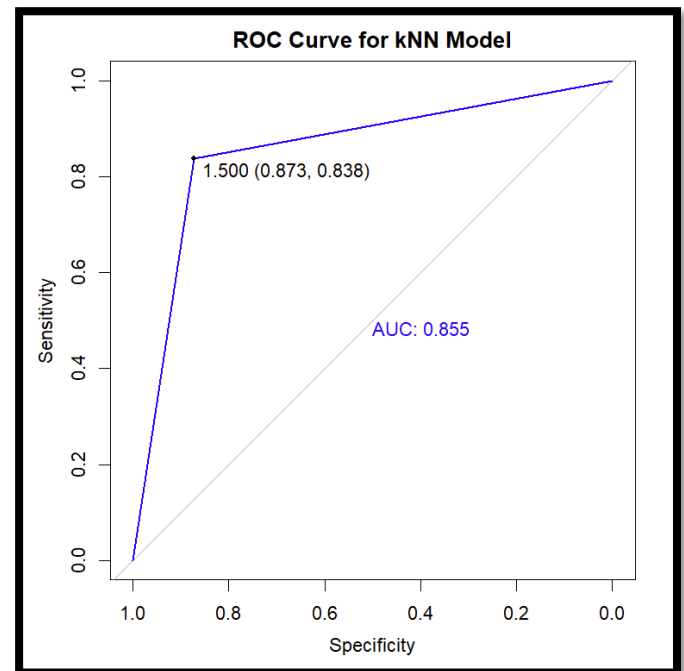


Figure 31. ROC Curve

### 11.3 Results from Prediction and classification:

The evaluation highlights the linear regression model's moderate accuracy, with an RMSE of 0.2257773 and MAE of 0.177837, while the kNN model shows stronger predictive performance, achieving an accuracy of 0.8566 alongside robust sensitivity and specificity metrics of 0.8728 and 0.8379, respectively.

Model	Metric	Value
Linear Regression	RMSE	0.2257773
	MAE	0.177837
k-Nearest Neighbors (kNN)	Accuracy	0.8566
	Sensitivity	0.8728
	Specificity	0.8379

## ***12 Features Impacting Mobile Device Prices based on Model Analysis are:***

- a) RAM (Random Access Memory)**
- b) Screen Size**
- c) Release Year**
- d) Network Connectivity**

The above features significantly impact mobile device prices. Devices with higher RAM, larger screen sizes, newer release years, and advanced network connectivity tend to command higher prices, reflecting consumer preferences for performance, display quality, technological advancements, and connectivity options. Businesses should prioritize these features in product development and marketing strategies to meet consumer demands effectively and optimize pricing strategies for profitability.

## ***13 Conclusion:***

In conclusion, this project dives into the world of used mobile phone prices, exploring various features that influence consumer decisions. Using machine learning algorithms, we aim to predict and categorize prices into "Low" and "High." This not only aids consumers in understanding what makes certain phones more expensive but also helps businesses tailor their strategies based on popular features.

The in-depth exploration and analysis of the mobile device dataset have culminated in the development of robust predictive models. The Linear Regression model, selected for its interpretability and accuracy, effectively predicts normalized used prices, achieving an impressive Adjusted R-squared value of 0.8438. It demonstrates precision, with a Root Mean Squared Error (RMSE) of 0.2258 and Mean Absolute Error (MAE) of 0.1778 on the holdout dataset.

Furthermore, the K-Nearest Neighbors (KNN) classification model excels in classifying used mobile devices into High and Low-price categories, exhibiting a remarkable accuracy of 86.13% on the holdout dataset, coupled with high sensitivity and specificity. These models collectively offer valuable insights for stakeholders, empowering informed decision-making in the dynamic realm of mobile device pricing.

## ***14 Business Recommendations:***

### **Feature Prioritization for Product Buying:**

To enhance feature prioritization for product buying, businesses should incorporate the processor as a key feature in the dataset. By leveraging insights from predictive models and including the processor's impact on mobile device prices, businesses can effectively prioritize features that drive higher prices. This addition enables more informed decision-making in product development, allowing

companies to align their resources strategically to meet consumer preferences and enhance competitiveness in the market.

#### **Dynamic Pricing Strategies:**

Implement dynamic pricing strategies based on predictive models to optimize revenue generation. By leveraging machine learning algorithms, businesses can adjust prices in real-time according to market demand and consumer behavior, maximizing profitability while remaining competitive.

#### **Inventory Management and Stock Optimization:**

Use predictive models to forecast demand for different mobile device models and variants. Optimize inventory levels and product assortment to meet customer demand while minimizing excess stock and associated costs.

#### **Competitive Benchmarking and Market Positioning:**

Continuously monitor and analyze competitor pricing strategies and market trends. Benchmark against industry peers and adjust pricing and marketing strategies accordingly to maintain a competitive edge and capture market share.

Implementing these strategic recommendations will empower your business to adapt to the dynamic landscape of the mobile device industry, enhance operational efficiency, and drive sustainable growth and profitability.

## ***15 Executive Summary***

### **Sri Vamshi Polela**

Date: March 20<sup>th</sup>, 2024

#### **Business Opportunities:**

This project focuses on predicting and categorizing the prices of used mobile phones using machine learning algorithms. The business goal is to understand the factors influencing mobile phone prices, helping businesses tailor their strategies, and assisting consumers in making informed decisions. The analysis includes data preprocessing, exploration, and transformation, revealing insights into the relationships between various features.

#### **Approach:**

Our investigation unveiled valuable market trends, consumer preferences, and pricing patterns. Android devices, notably from brands like Samsung and Huawei, dominate our dataset, while a noteworthy positive correlation between normalized new and used prices implies a parallel increase. To

ensure robust modeling, we executed predictor analysis, data transformation, and dimension reduction, with careful partitioning into training, validation, and test sets.

In preparation for model selection, options such as Linear regression and regression tree have been considered, aligning with our goals. As we move forward, our aim is to empower stakeholders in the mobile phone industry with actionable insights derived from these models, enabling informed decisions to optimize pricing strategies by understanding the correlation between new and used prices. Focusing efforts on the dominant Android market, acknowledging brand prominence like Samsung and Huawei, and considering the impact of screen size and battery capacity on prices are crucial for effective market positioning. Implementing differentiated marketing for high and low-priced categories, staying attuned to market dynamics, and leveraging predictive models for strategic planning can collectively empower stakeholders to enhance profitability and navigate the ever-evolving mobile phone market landscape.