**CSDA 6010 Analytics Practicum**

**Used Smartphone Price Analytics**

Reishekesh Reddy Inavola

Data Analytics, Webster University

Prof. Dr. Willie Rivers

17th September 2025

**Table of Contents:**

* **Introduction………………………………………………………………. 2**
* **Business Problem and Goals……………………………………………… 3**
* **Analytical goals and Approaches**: **……………………………………….. 4**
* **Data Exploration and Preprocessing……………………………………... 5**
* **Data Understanding………………………………………………. 5**
* **Attribute Definitions………………………………………………. 6**
* **Missing Data and Zero Value Handling…………………………. 7**
* **Outliers……………………………………………………………. 9**
* **Data Exploration…………………………………………………… 10**
* **Correlation of variables …………………………………………… 17**
* **Multi-Collinearity………………………………………………….. 18**
* **Predictor analysis and relevancy………………………………………….. 18**
* **Data Engineering and Transformation……………………………………. 19**
* **Data Partition …………………………………………………………….. 20**
* **Model Selection ………………………………………………………………. 20**
* **Model fitting and test accuracy……………………………………………… 22**
* **Models’ performance………………………………………………………… 25**
* **Model evaluation…………………………………………………………….. 28**
* **Observation and conclusion………………………………………………… 28**

**Used Smartphone Price Analytics**

**Executive Summary**

The quick growth of the used smartphone market evolves as a result of the rapid growth in technology this challenges to sellers and buyers. Price inconsistency and variability is confusing of phones that could be over-priced or undervalued and the sellers often lack control and miss out on money and buyers lost confidence in the transaction. The research study under analysis has discussed these issues by examining 3,454 used smartphones with 15 various features such as brand, operating system, memory, camera, battery, days used, etc. The missing data and outlier were eliminated, after which the exploratory data analysis (EDA) was conducted to discover patterns and correlation and multicollinearity. Then after regression models were used to predict resale value estimation and classification models were used to categorize smartphones into high price and low price groups.

The results show Regression model effectively captures complex relationships between features, new phone price, hardware specifications (Camera, screen size and storage) and brand reputation as major price drivers. This allows seller to benchmark resale price and optimize trade ins values, offering transparency to customers. Decision tree classification enhances more interpretability when it forms different if-then rules suggesting that heavier devices (meaning big screens and batteries) and more storage go to premium pricing with strong brand effects. These models collectively give businesses valuable insights to differentiate their premium offering, create consumer confidence and to set consistent and profitable resale prices. In this manner, these types of solutions provide businesses with a competitive advantage by applying analytical market solutions to the real-time demands in the used smartphone market.

**Introduction**

Mobile phones have been integrated in the modern life, as a result of the fast changing technology. Since people have been changing their technology and switching to new phones because of technological advancements, the second hand phones market is expanding. Even though a new group of customers has emerged in this industry and it is not easy to set fair price to the used smartphone, inconsistent pricing is always the issues to businesses lost profits, lack of trust in resale market, sellers have no idea on how to put the used prices, it has been volatile, this rise in customer dissatisfaction. The relation between features of the phone which determine the price and resale value is not that straightforward.

The aim of this study is to uncover possible problems of multi-collinearity, and to determine the dependence between the target variable (price) and the explanatory variables (phone characteristics) using the exploratory data analysis (EDA). In the research, we utilize classification models to categorize the smartphones into high and low-priced for premium sales segment and regression models is constructed to estimate the prices of used smartphones for selling to the customers and buying back.

The data employed in this analysis comprises 3,454 sales transactions of used smartphones containing 15 features (brand, operating system, number of days used and camera pixels strength). Data preprocessing is done to the dataset to accommodate outliers, missing values and then models are built.

The techniques involved are:

Missing value preprocessing.

Exploratory data analysis (EDA) to determine new trends and patterns.

Machine Learning (ML) Models price predicts and classifies by price as high and low.

Deciding the model based on the model evaluation metrics.

The project provides the used Smartphones sector new data insights and specific resale price predictions, sellers set competitive profitable resale price, customers avoid over price under value phones and improves trust and business transparency. These outcomes support better decision making for both buyers and sellers.

**Business Problem and Goals:**

Used smartphones pricing is not consistent and not clear which makes it hard for sellers to fix a fair price and attract customers. Many times, sellers miss the right price point and lose advantage in market. Buyers also get confused with phones that are undervalued or overpriced. There is a need for data driven way to estimate resale prices and give transparency for both buyers and sellers.

**Business Problem**: It is challenging to set price for used phones, without consistent pricing business face losses, difficult in attracting customers and reduce trust in resale market. Customers feel dissatisfied over price volatile and reduces purchases.

**Business Goal**: Enable seller to set consist and competitive price through data driven insights. This increases profits and attracts more customers. Improve trust in the resale market.

**Business Problem**: Struggling to determine the buying price of used phones, often results in over paying for undervalued phones loss in profit or under paying to customers creates dissatisfaction discourage future trade-ins. Negative to business profits and customers loyalty.

**Business Goal**: offer easy and straightforward price data to the sellers buying used phones, ensuring phones are bought at right price. This allows businesses to resell at higher margin, keeping customers satisfied with the trade-in values.

**Business Problem**: Struggling to classify used phone high end or low end model. Lack of clarity leads to missed opportunity for higher margin on premium phones. Customers may feel uncertain about the value of device they are purchasing.

**Business Goal**: Through data insights classify phone’s based of features, enable sellers to differentiate high end phones, sell at premium margin and at same time provide customer transparent understanding of phone value.

**Analytical goals and Approaches**:

To identify the characteristics of phone affecting the used price, building the models, preparing the data EDA (exploratory data analysis), cleaning and building Machine Learning models regression and classification.

**Analytics Goal**: Understanding how smartphone features screen size, rear and front camera, internal memory affecting used phone price. This help sellers highlight value adding features and set competitive prices.

**Analytics Approach:** Performing EDA (Exploratory data analysis) to identity trends, patterns, correlations between variables and distributions. EDA is essential in businesses as it reveals which feature impacts the resale price and allow business to focus on customer value.

**Analytics Goal**: Price models should be accurate and reliable so addressing issues of multi collinearity and outliers. Without handling issues businesses risk price decisions on misleading signals.

**Analytics Approach**: Statistical technique such as correlation matrices is used to detect multi collinearity, Preprocessing the dataset by adjusting the values to improve model. Ensures stable models are built and improving trust in price recommendation.

**Analytics Goal**: Providing sellers with reliable prediction of smart phone prices based on the phone features, allowing sellers to set consistent and profitable prices.

**Analytics Approach**: Developing regression model to predict the smart phone prices, Regression directly measures how each feature contributing to predict the price value. RMSE and MAE are used to ensure accuracy and business reliability.

**Analytics Goal**: Classifying smartphone into high or low priced category. Classification helps businesses to sell premium (high) value phones at higher margin and guiding customers on device position.

**Analytical Approach**: Implementing classifying model to categorize phone into price segments. Classification is justified as it translates pricing data into categories (high vs low), allowing sellers to identify premium opportunity and customers make informed purchase decisions. Model performance is evaluated on accuracy, F1 score and confusion matrix.

**Data Exploration and preprocessing**

To understand and prepare the data for model training we undergo the preprocessing steps.

* **Data Understanding**: The data set consist of 15 attributes describing the characteristics of smartphone device\_brand, OS, days\_used, normalized\_used\_price, battery. The dataset has 3455 rows. Target variable ‘**Normalized\_used\_price**’, The main business problem is to predict the used phone prices, the dataset contains all the variables to analyse phone features driving the price.
* **Attribute Definitions**:

**Device\_brand**: it is a categorical variable, it consists of brand name of the smart phone.

**OS**: It is the categorical variable, it tells what kind of operating system is running on the smart phone.

**Screen\_size**: It is a numerical variable. It tells the phone size of the display in centimetric.

**4G**: It is a categorical variable it tells does the mobile network supports, 4G or not.

**5G**: It is a categorical variable it tells does the mobile network supports, 5G or not.

**Rear\_camera\_mp**: It is a numerical variable it tells the rear camera mega pixel for photo capturing.

**Front\_camera\_mp**: It is a numerical variable it tells the front camera mega pixel for photo capturing.

**Internal\_memory**: It is a numerical variable that describes the phone storage.

**RAM**: it is a numerical variable, it describes the phone memory, it is used to quick access the applications.

**Battery**: It is a numerical variable that describes the battery capacity in mAh (milliamp-hours).

**Weight**: It is a numerical variable that describes the weight of the phone in grams.

**Release\_year**: It is a numerical variable, that indicates in which year the phone is launched.

**Days\_used**: It is a numerical variable, it describes how long the phone was used by someone.

**Normalized\_used\_price**: It is numerical variable, a price metric for the used phones (target variables).

**Normalized\_new\_price**: It is a numerical variable, a price metrics for the brand new phone.

**Price\_Class:** It is a categorical variable consist of HIGH or LOW based on normalized\_used\_price, top 30% as HIGH and bottom 70% as LOW.

* **Data Cleaning**

Categorical variables represent group not continuous values, if not taken care r considers as text so variable converted to factor ‘device\_brand’, ‘OS’, ‘4g’ and ‘5g’.

* **Missing Value**

Handling NA (Missing values) values in rear\_camera\_mp, I have grouped front camera specifications for all the rear\_camera\_mp value and imputated using median (robust to outliers) of that row for NA values of rear camera, for every front camera there is at most not more than 3 or 4 values of rear camera so by this I imputate the nearest value of actual rear\_camera\_mp. Handling NA values in front\_camera\_mp, there are two missing values in front camera where the rear\_camera\_mp is 12.2 so I filtered the rear camera with 12.2 and imputated NA with the median of front camera, Same group based approach was used in front camera likewise in rear camera.

Handling Na values in internal\_memory, imputated with median (robust to outliers), smart phone storage is fixed with few storage options (16gb,32gb…) the internal memory should match with ram (0.02 and 0.03) for missing values so 32gb is derived from median as best, mean gives 42gb which can influence from outliers. Handling NA values in ram, for the 4 missing values of ram the internal memory is (0.06,0.1) so we grouped based on internal memory and imputated with ram median of internal memory (0.06), both ram values of internal memory (0.06,0.1) are almost similar so imputated with ‘0.03’.

Missing battery values imputated using median, most of the devices vary from release year and screen size aspect to battery changes, mean influenced by outlier’s median is the best option. Handling NA values in weight, it has some outliers, mean is influenced by outliers and median is the best option as smart phones weights should be less and easy to carry, median satisfies this issue.

* **Checking Zero**

Based on the check of zero value, the majority of the attributes do not contain any zero entries. The front camera is the only column that has approximately 1.1 percent zero value and also no other features such as the screen size, RAM, battery, internal memory, prices and days used have any zero value. This demonstrates that the zero is not a frequent issue in the data set.

Most of the Zero values from front camera are from the year 2013 to 2020 that implies they are not too old phones with no front camera, typo error has been done or data wrongly recorded. Median (robust to outliers) value is imputated to the camera less than 1, so the data be realistic with device configuration. Rear camera has 70 fields of less than 1 value and they are from the same year as of front camera, same approach has done to imputate almost zero values so data be realistic with device configuration.

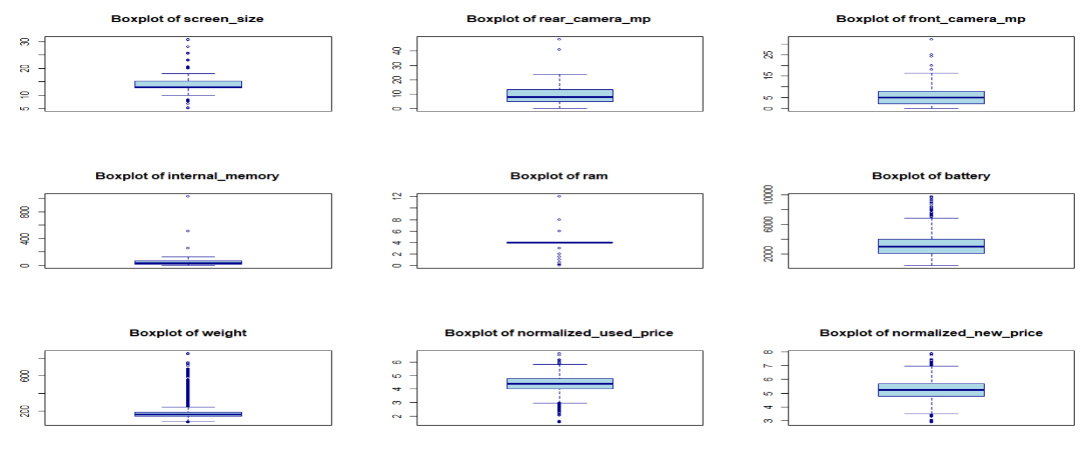
* **Outliers**

Outliers are evident in the boxplots in most of the attributes. Screen size, most of the values are around 6 inches but there are a few extreme values of more than 15 inches appears to be tablets rather than smartphone, including unrealistic values would mislead when smartphone price prediction and mislead businesses about customer demand in smartphone segment.

Megapixels of rear and front cameras also indicate the existence of devices with very high camera values if left unfiltered they would inflate the prediction price, causing sellers to overprice ordinary phones, lose customers to competitors over fare price. The extreme values of internal memory 500 GB and RAM of more than 12 GB that are not compatible with most smartphones, if not filtered business risk trade in values based on non-existence data, undermines price accuracy and customer trust.

Outliers present in battery of 8000 mAh and weight of 600 grams unrealistic values leads to push in predictive price, the buyback price increases, leads to loss in resale margin(compromised). Price outliers of normalized used price and normalized new price skew the model and businesses end up setting unprofitable high price (inventory remain unsold) and low prices (loss in revenue).

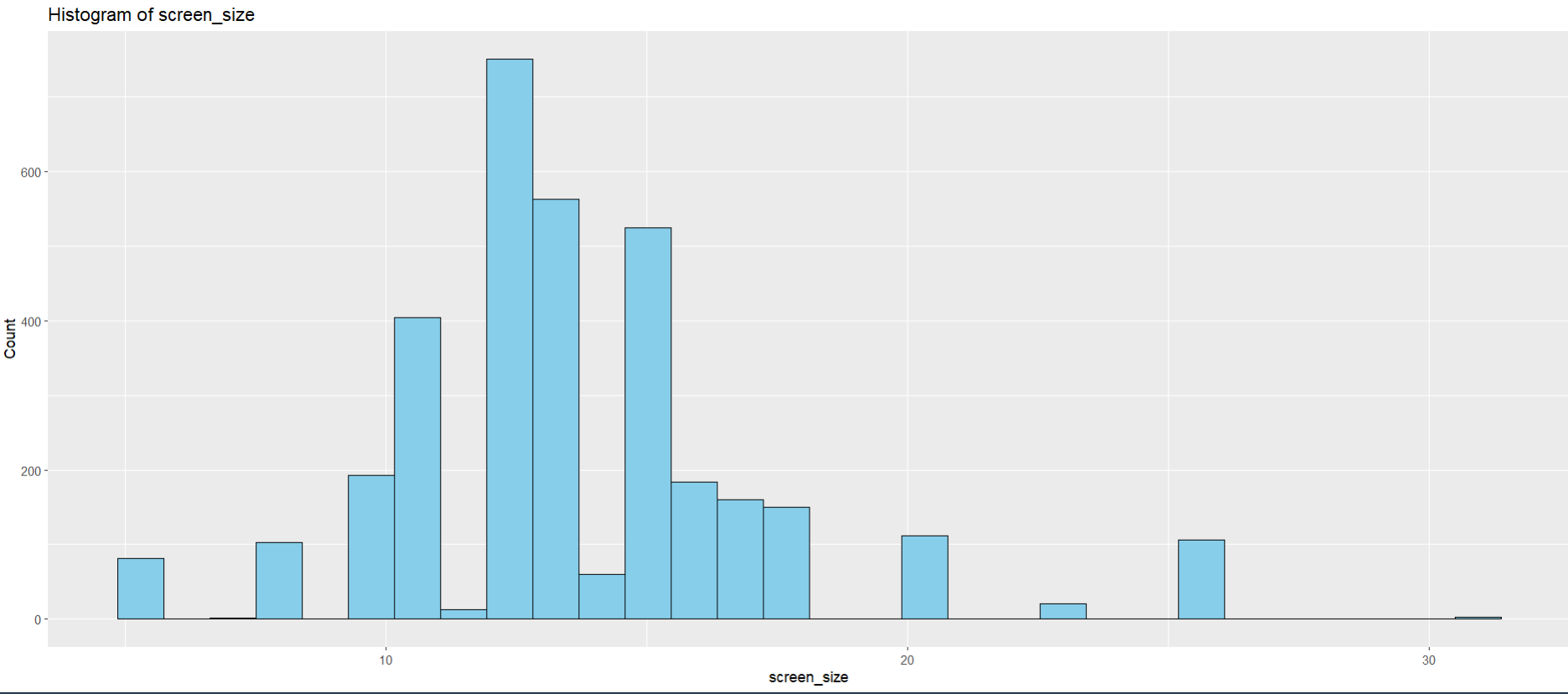
To protect pricing accuracy, simple methods will help us to address these outliers. Since extreme unrealistic values are probably erroneous entries, this can be filtered out of the data. The other strategy is to log transform skewed variables such as price and memory to diminish the effect of extreme values. In this manner the dataset will be balanced, clean and prepared to do correct modelling.



* **Data Exploration**:

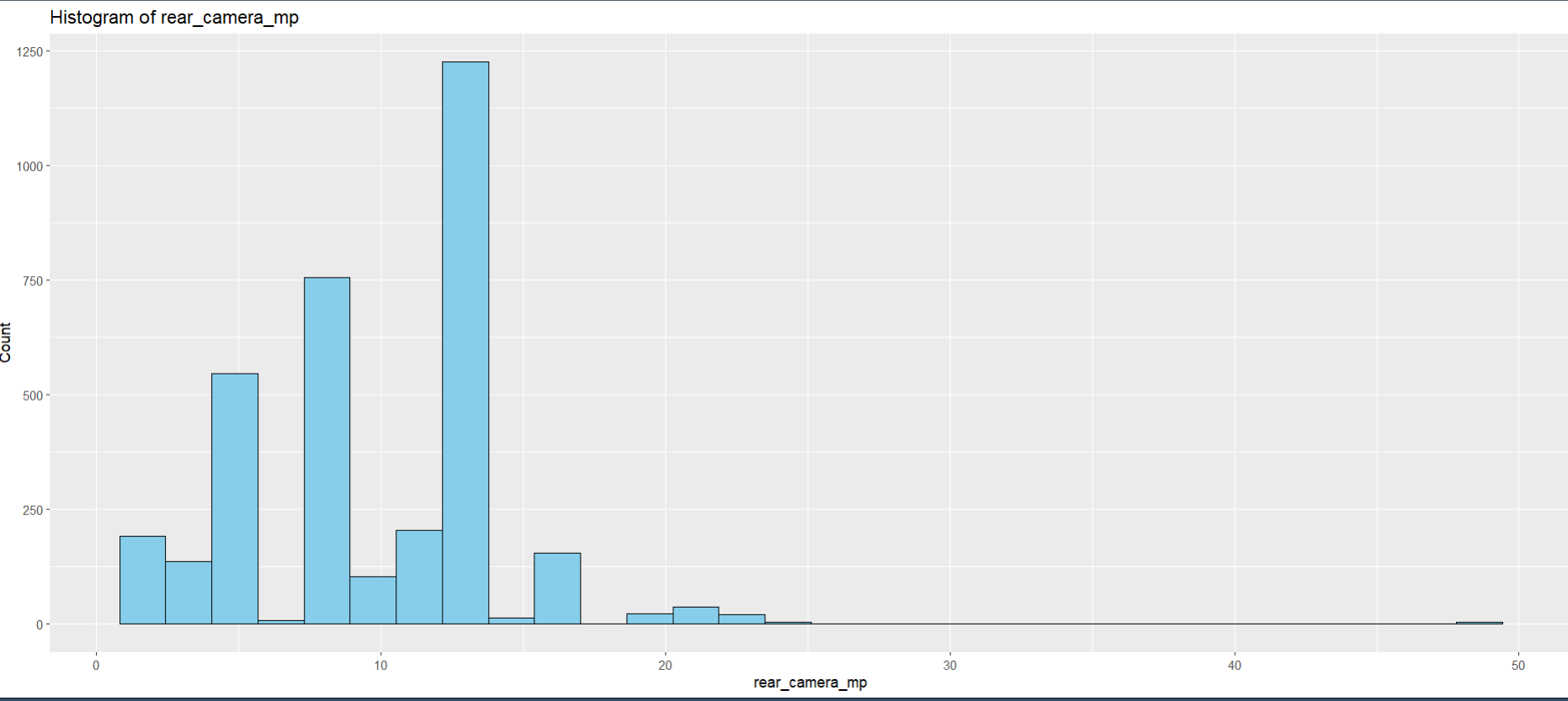
In order to understand the distribution of the data, it was calculated with the help of the summary statistics shows how key features such as the screen size, RAM, battery, weight, days used, and normalised used price. These metrics shows overall key features customers looking in the resale market. Almost all smartphones had valid values for days used. The dataset has a good amount of data, days used help in depreciate the buyback price and help in resale value.

Most screen sizes are between 12 to 15cm, RAM was mainly centred to 4, and battery capacity for most devices were between 2100 to 4000 mAh, this indicates most of the phones in the market are mid range and not too expensive or too low priced phones. Almost all devices had 4G support and only a smaller number had 5G, market is transitioning slowly from 4G to 5G network. For categorical features brand and OS Android is the majority, while Apple formed the second group, the market is driven by android operating systems. Data is consistent with real word smartphone specification, ensures insights applicable to real world market condition and business decision making.



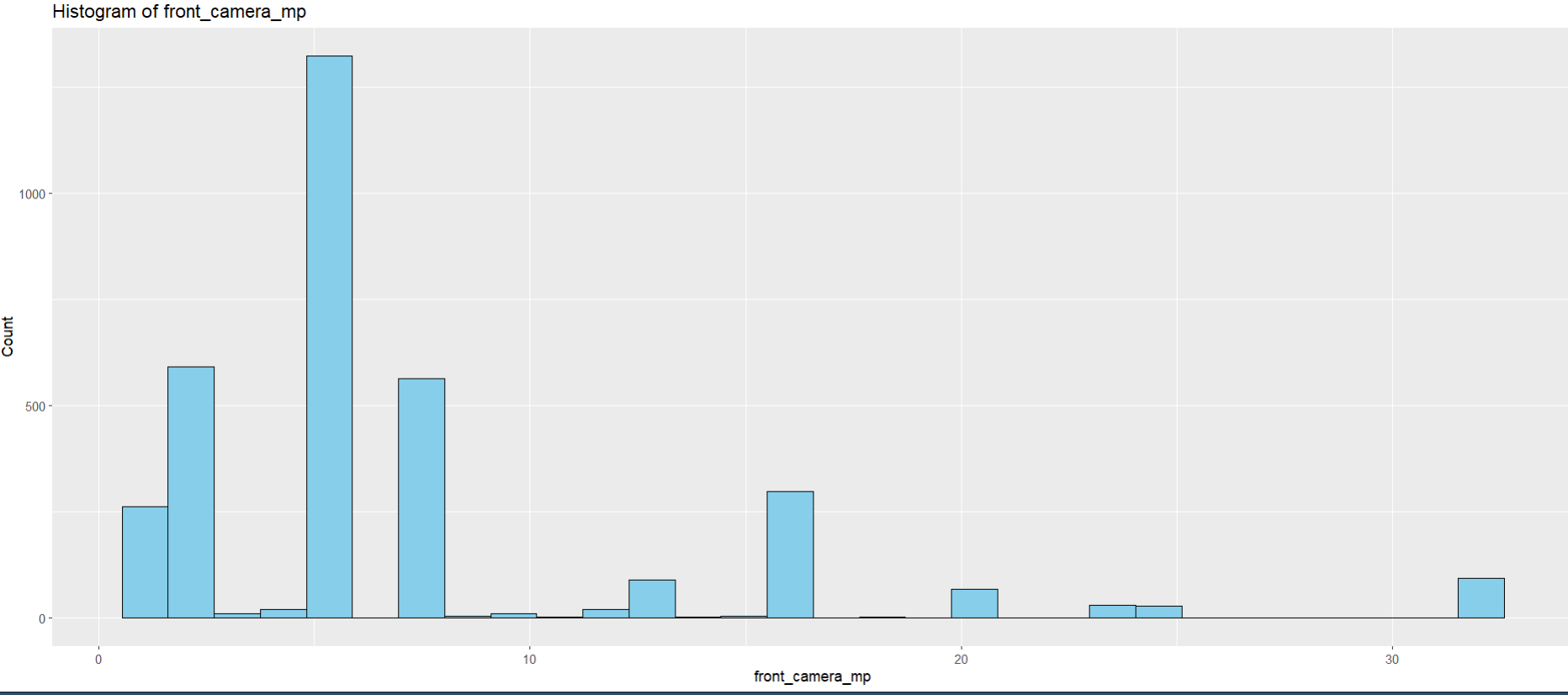
The histogram of smartphones shows most fall between 12 to 15cm, with a small number below 10 mostly old phones in 2013 and some phones with more than 20cm screen size can be newly launched fold phones in 2020, there is a peak at 13cm where most phones lie. The distribution is right skewed. There are some extreme outliers reaching 30 cm which are unrealistic most probably tablets.

Most resale smartphones fall between 12 to 15 cm so this is considered as consumer demand. Larger screen 15 to 17 cm considered as premium phones or high value.



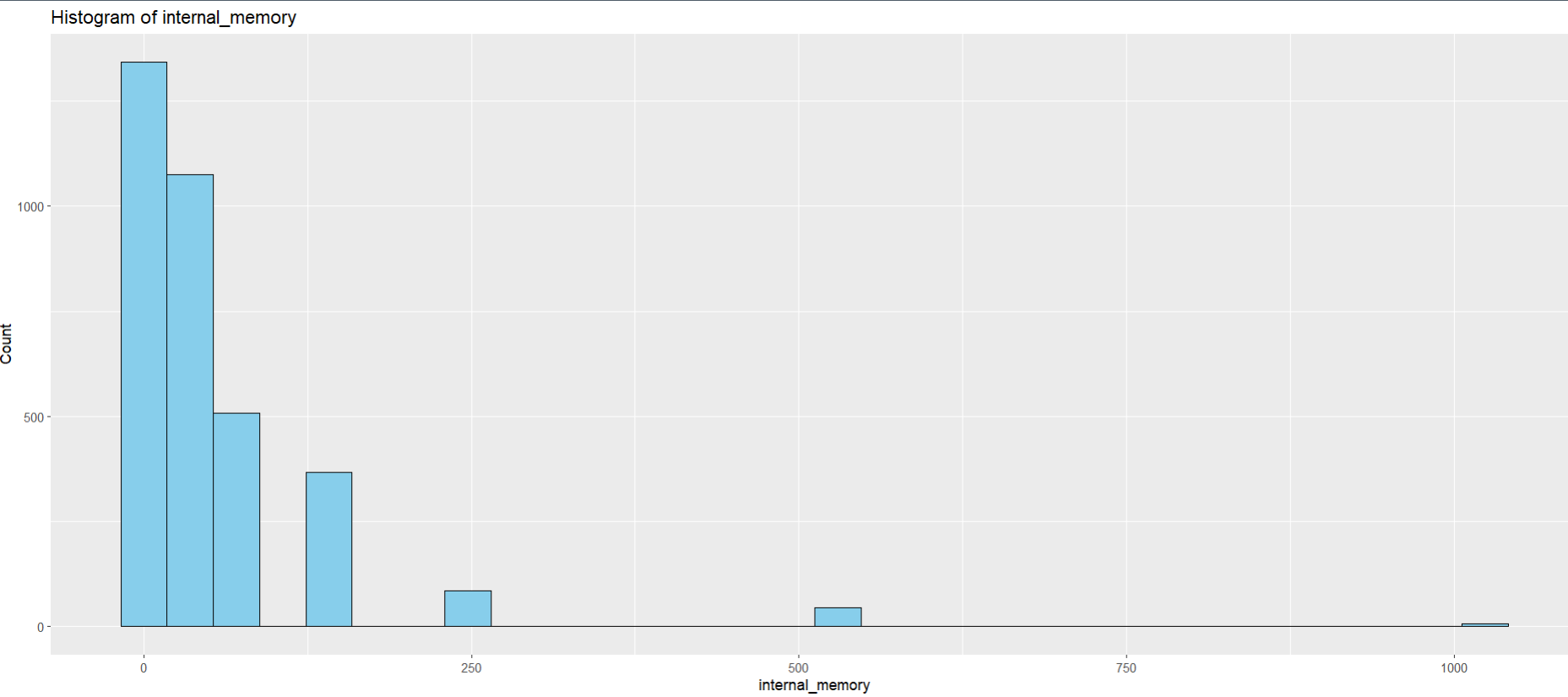
Most rear cameras fall between 5 to 13mp (over 1200 phones) and another frequency is between 6 to 8mp (over 500 phones). There is a sharp peak at 12mp where all the phones lie. There are few devices lower than 3mp probably old phones and some lie above 28 and almost 48 there are new generation premium phones. The distribution is moderately skewed.

Most used resale phones have 6 to 13 MP where market is standard, having 20 MP or higher is considered as premium or high vale phones. Phones under 5 MP appears less likely to be bought by customers comparing to market standards.



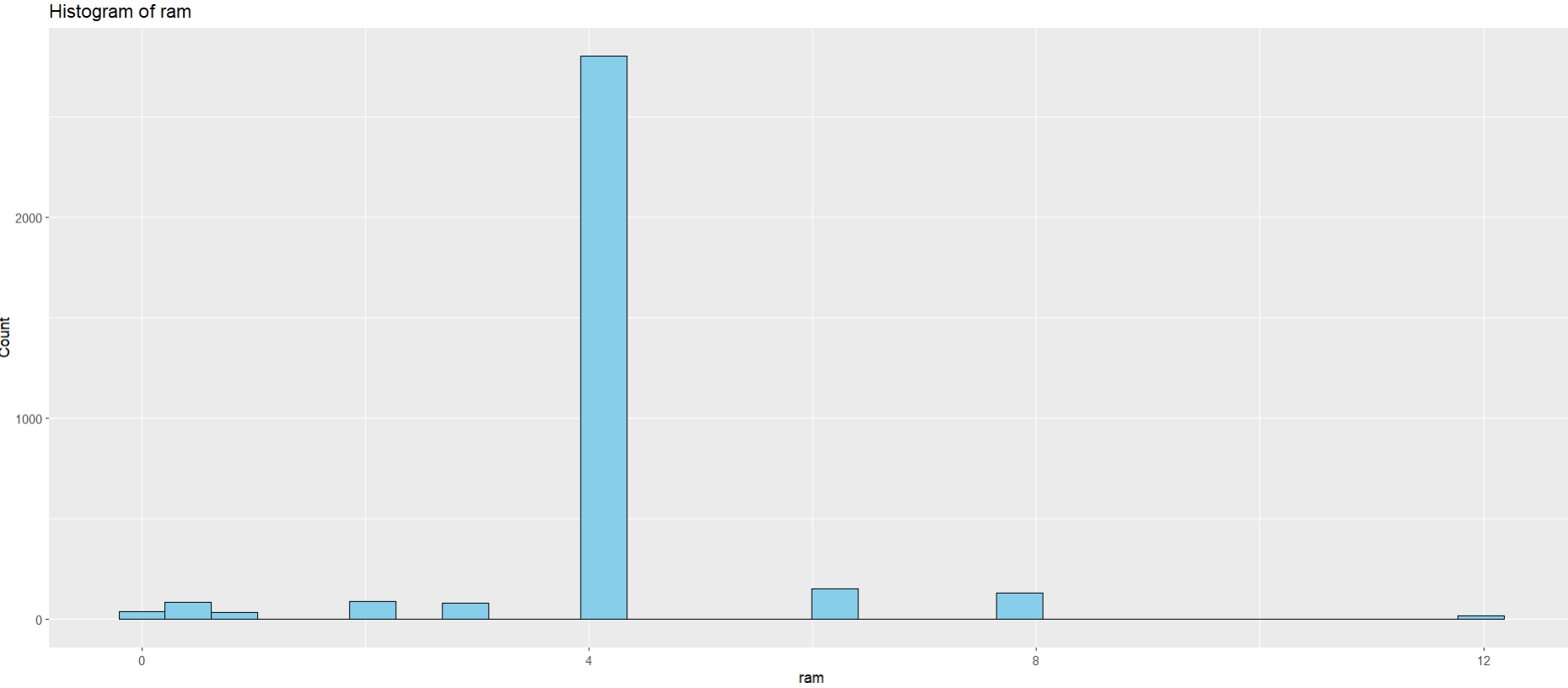
Most front cameras lie between 2 to 8mp at 4mp there is a sharp peak where most phones lie, 4mp makes most common front camera specification. Few devices lie between 20 to 32mp premium category new generation phones, few devices below 2mp that represents older phones (2013). The plot is highly skewed at lower end. Majority of the phones lie below 10 mega pixels.

Resale market is dominated by 5MP mid range, this indicates market baseline. Devices with more than 8 MP is considered as premium phones, can be positioned at high margin.



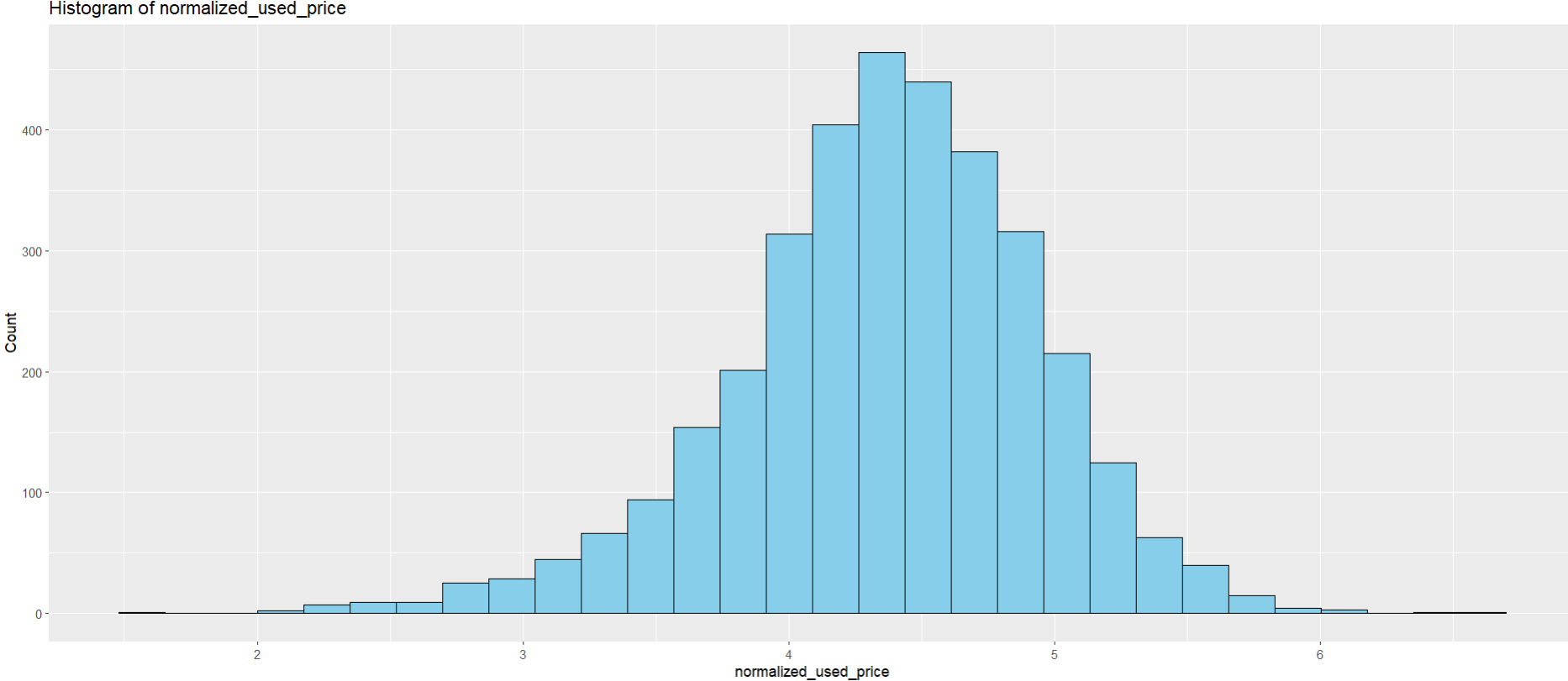
Majority of smartphones lie between 16 to 64gb, some have higher storage 256, 512, 1024gb which are premium segment phones, the data is highly right skewed. This suggest most of used phone market is mid range (16 to 64 GB). Few devices fall around 128 to 256 GB indicating newer to premium phones.

Resale market is dominated by mid range of 16 to 64 GB that indicates customers prioritize affordability over high storage. Extreme storage indicates flagship phones if genuine or some outliers.



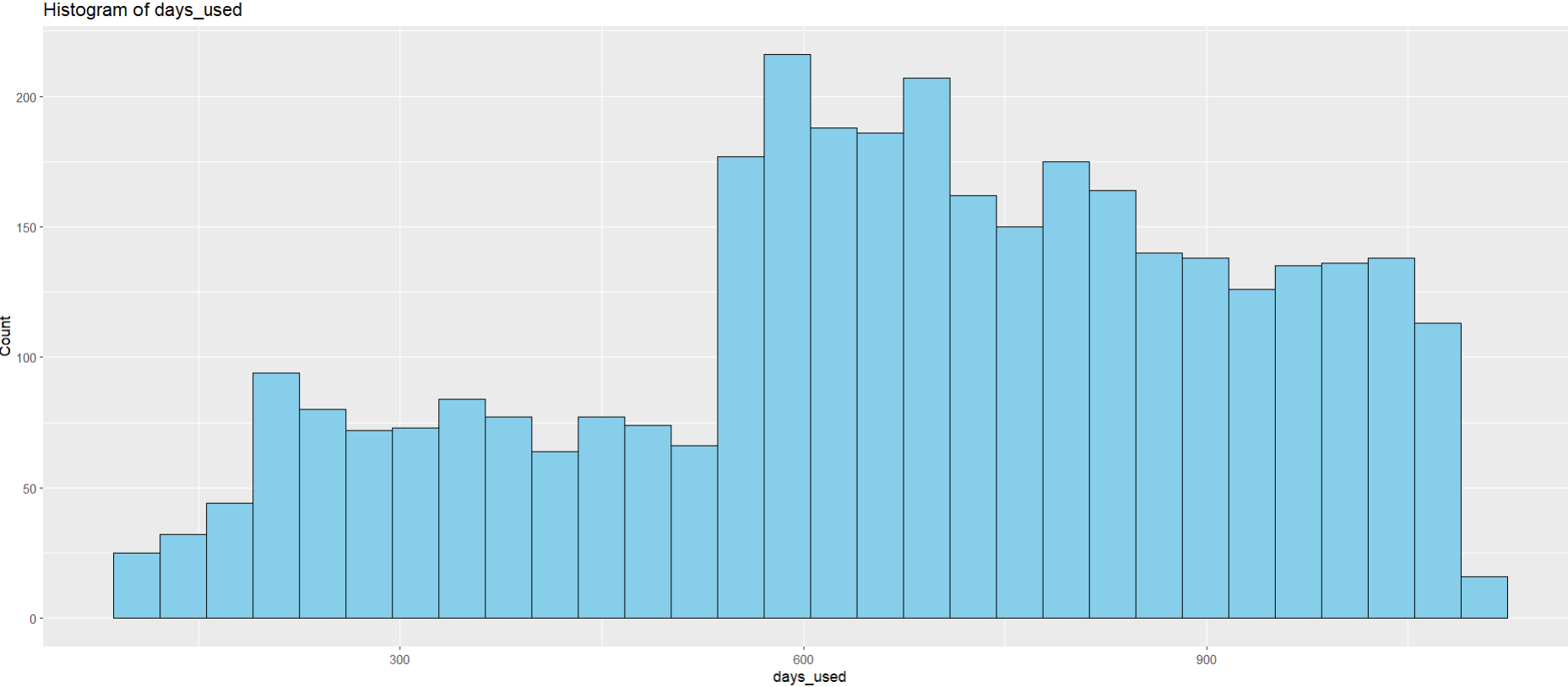
The distribution is strongly clustered at 4GB, where most phones lie, some phones has less than 2 GB pretty old phones, some lie in 6 GB, 8 GB are considered as high end model, 12 GB as flagship.

In the used phones resale market 4 GB is considered as baseline memory in RAM, 6 GB 8GB and 12 GB are considered as premium segment phones.



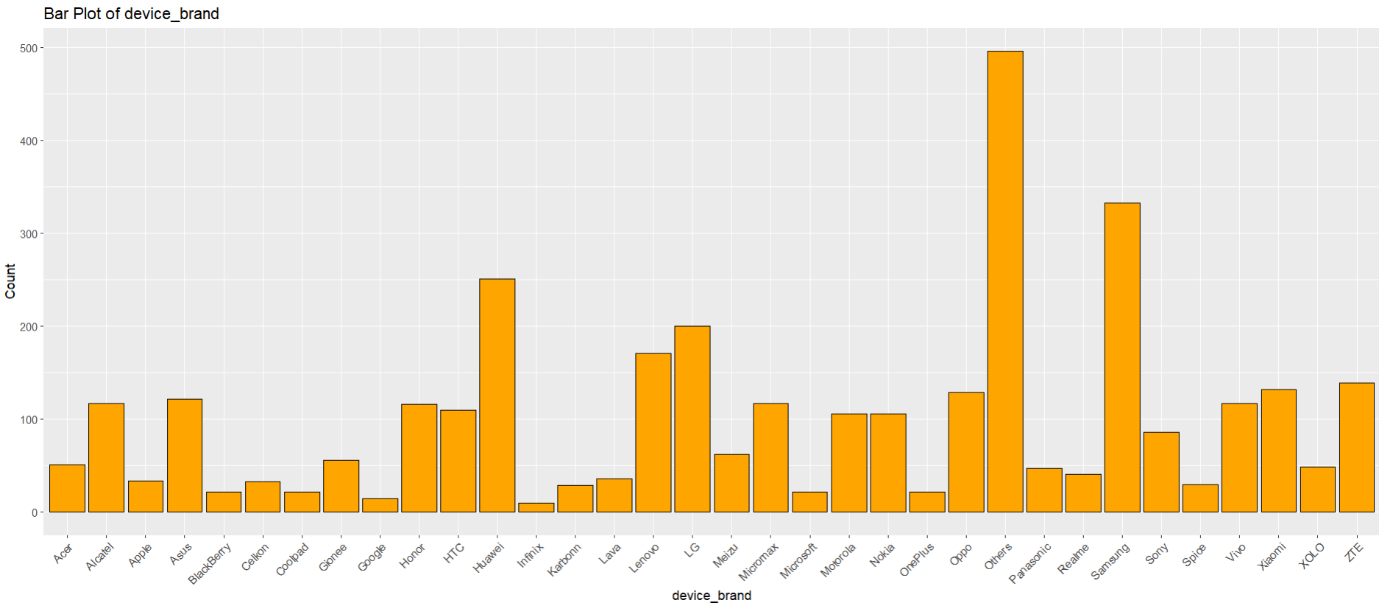
The plots show bell distribution, most phone fall in the price range of 4 and 5 and sharp peak at 4.5 common price point. Some fall below 2 indicates old phone, some are above 6.5 tell they are premium phones. Price range extends from 2 to 6 showing the range of used phone values.

Used phone price clusters around the middle range, buyers and sellers can use 4 to 5 as benchmark fair value. Sellers price below the range of 4.5 can be considered as undervalue devices, while price above may be considered as flagship or premium phones.



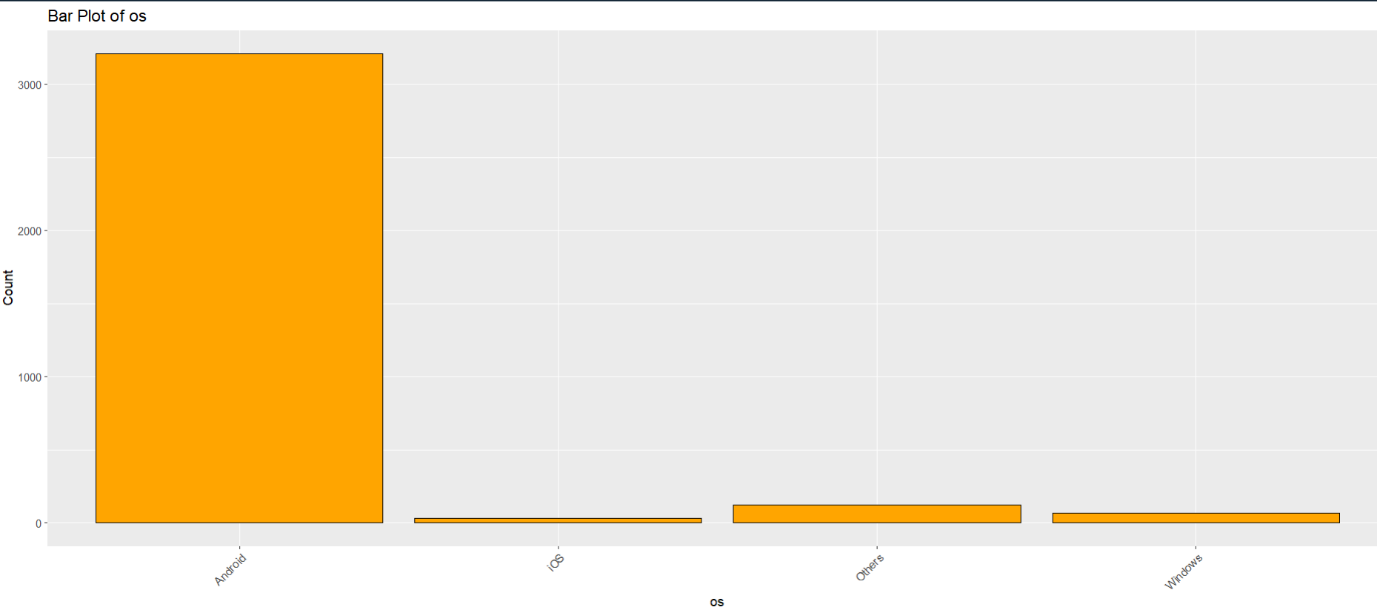
The days used fall between 500 to 900 days, noticeable peak is between 600 to 700 days. All phones are resold after 2 years of use, some sold under 300 days, few devices used for more than 1000 days (3 years and more).

Sellers can expect most resale happens after 1.5 years of use. Phones used less than 200 and sold can be flagged as defective or financial need. Phones sold after 900 days are heavily depreciated very few prefer old phones.

****

Thedistribution is diverse where Samsung, Huawei, Oppo, and Lenovo contribute a huge portion of phones, Apple, google and OnePlus appears much less (low selling phones)**,** othercategoryhas a huge contribution, consist of many small brands.

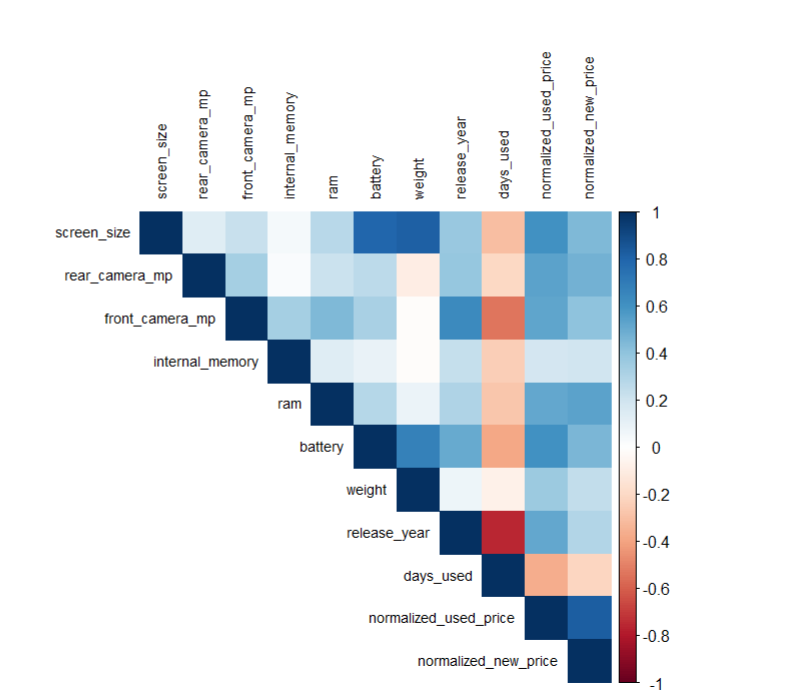
Oppo, Huawei and Samsung dominate the resale market so pricing strategy should prioritize these brands. Brands with lower presence may not be profit, no margin in sale.



The plot shows clear dominance of android, iOS and windows have a small share, the plot is heavily skewed towards android, reflects real world market trend of android.

The resale market is android driven primarily. However, iOS devices few in proportion often retain higher margin. Other OS and windows have minimum resale value and market demand.

* **Correlation of variables**



From the heat map we can clearly say the smartphones are strongly related, this allows us to feature grouping set consistent prices rather than independently treating features. Screen size, weight, and battery have high positive correlation, bigger screen size have bigger battery, this tends to gain in weight, this helps to classify premium phones vs low range phones and enables high margin placing on premium phones. Real and front camera are moderately correlated, higher real mega pixel has higher front camera. Days used is negatively correlated to release year, this helps us in deriving buying back price newly release models fetch good buyback price, new phones are used less. Normalized used price and normalized new price is strongly correlated, helps us in deriving resale price considering new price as anchor. Heatmap highlights clusters of related features and provides visual for collinearity, this phenomenon helps is price prediction in more transparent and easier to explain customers.

* **Multicollinearity**

The correlation matrix gives a clear sign of multicollinearity on some variables. Screen size, battery, and weight (0.79) strongly correlated, they provide over lapping information and this cause redundancy. From business perspective seller don’t have to emphasize separately on all the three features, considering one implies a phone with bigger screen should have bigger battery to support and tends to more weight can capture price driver without bothering customers with redundant information. Normalized used price and normalized new price have high correlation (0.82) market follows on new price, use this as benchmark for competitive pricing and maintaining margins. Release year and days used are strongly negatively correlated (-0.75) these results address the multi collinearity, this allows old released models which are used more days have less resale value and bare margin.

**Predictor analysis and relevancy**

The analysis shows normalized\_new\_price is the strong driver of used price (r≈0.83) this indicates the market follows original launched price, sellers can position resale price referring to original price, core specs battery (0.61), screen\_size (0.60), rear\_camera\_mp (0.54), front\_camera\_mp (0.53), RAM (0.52), and release\_year (0.51) add significant value, this tells us customers look for performance good mega pixel quality and launched year relative to purchase date by this sellers can justify higher reselling price, days\_used reduces the used price value (r≈−0.36) sellers should look after for less used phones to make high margin of profits, weight contributes minimum value and internal\_memory has negligible effect, sellers should not focus on this features for pricing.

Battery, screen\_size and weight are highly correlated, keeping battery as single representative, buyers often look for bigger battery for longer usage of devices can be a selling message, higher battery implies bigger screen\_size and weight. The categorical analysis shows all the variables have strong relevance 4g (p≈2.8e-272), device\_brand (p≈4.9e-208), os (p≈2.96e-102), and 5g (p≈3.85e-83) significantly explain the used price, brands like apple network type 5G bring higher resale value market thrives for better technology. The predictor analysis gives the important variables for building the models, aligning business insight features (normalized\_new\_price, screen\_size and rear\_camera…) ensure model statistically and strategically valid, these variables ensure alignment with the business goal by building the models with suitable variables selection.

Classification based on features like brand, OS, 5G ... helps businesses to segment their inventory into premium High value and budget value phones. Premium phones sold for higher margin and budget phones attract more customers by pricing strategy.

Selected features play a crucial role in model accuracy and business relevancy, normalized\_new\_price reflects how buyers benchmark against used price, RAM camera quality, screen size captures performance buyers willing to pay, Brand OS and 5G network classifies premium segment, aligning predictions with consumer demand. Model not only provides used phone price effectively but also provides actionable insights.

**Data engineering**

Outliers present in the dataset, so to reduce outliers, I filtered an ideal phone specification of screen size, battery and weight, above that all are considered as tablets or computers, records are removed 11 out of 3454. Ideal phone specification Screen size - 26 cm, battery-5500 mAh and weight 253 grams, in most cases phones do not cross this specification. Screen size is considered as primary factor and at most value is taken.

Any phone specification under screen size and battery (11 and 1000) are considered as smart watches, records are removed 14 out of 3443, a screen size below 11cm are too small to use regular functions of smart phone and the battery capacity below 1000 mAh won’t be sufficient for regular functioning of phone.

A new column Price\_Class is created and added to the dataset, based on normalized\_used\_price top 30% are classified as HIGH and bottom 70% are classified as LOW, From the data exploration most of the phone specifications are mid-range, considering most of the market is budget segment phones (70%) and the rest premium segment (30%).

**Data partitioning methods**

We split dataset into 80% training and 20% test sets using stratified sampling on the target (normalized\_used\_price) to preserve the price distribution across splits. The model learns patterns form training data only, test data is used to evaluate the model. stratified split preserves the outcome distribution in both splits, reduce variance in evaluation. A fixed random seed is set to reproduce the same split. This strategy is transparent, simple balancing enough data for training the model (80%).

For classification modelling the data is split into train\_cal and test\_cal data frame with the new column of Price\_Class. Data was split into 80% training(train\_cal) and 20% test(test\_cal).

**Model selection**

**Multiple Linear Regression** (MLR): is a simple and interpretable for predicting used smartphone prices. It quantifies the contribution of phone features such as RAM, battery, days used to the resale value. This makes transparency for explaining business owners in price logic.

Trade off: It assumes linear relationship between variables, which not holds true.

**Decision Tree Regression:** It captures nonlinear relationships and produces easy rules that understands. This provides business objective of providing interpretable pricing decisions.

Trade off: Trees are sensitive to small changes in data and overfits if not pruned properly.

**Random forest Regression**: builds multiple tress, aggregate their results. It improves generalization and avoids over fitting as compared to single tree. Performs well on mixed data and identify specifications which drive most.

Trade off: It is less interpretable compared to single tree.

**Naïve Bayes**: Naive bayes offers strong baseline model for classification because it handles categorical variables without much preprocessing. it works by learning simple probability of each feature and combines them to make a decision. It is easy to explain and transparent.

Trade off: Naive bayes assumes each feature as independently, in real world it is not true.

**Decision Tree Classification** (CART): It produces if then rules that can be easily explained to business owners, it makes transparency in explaining how phone is classified as high or low priced model.

Trade off: Highly sensitive to noise and variation.

**Random Forest Classification**: Improves upon single tress by ensembles (bootstrap), provides accuracy and robustness to data imbalance. It highlights important predictors which help in business decisions.

Trade off: It has low interpretability when compared to single tree, it is hard to explain decision path to the end user.

**Model fitting and test accuracy**

* **Multiple linear Regression**:

Considering the important predictors for building the model, normalized\_new\_price, battery, rear\_camera\_mp, front\_camera\_mp, ram, release\_year, days\_used, device\_brand, os, X4g, X5g. Screen\_size, battery and weight are Multi collinear that means they all carry same properties, bigger screen size leads to bigger battery and bigger battery leads to weight so to avoid redundancy we considered only battery, it has strongest correlation with normalized\_used\_price.

Multiple linear regression is a formula based model, it combines all factors to predict the model, factors (battery + normalized\_new\_price + ram...), each factor has weight that tells us how much it increases or decreases the results.

Comparing the predicted values with the actual values, "RMSE: 0.25" On average the predicted normalized\_used\_price deviates 0.25 units away from the actual resale price, model is making small error when predicting resale price.

"R-squared: 0.82" the model explains 0.82% variation, that means it captures 82% of key factors that drive resale value (batter, new price...).

**Decision Tree Regression:**

Decision tree works like flowchart for prediction, it asks simple yes or no question step by step, for example new phone in 2020, if yes it goes further, if no it goes further. At each path the model gives predicted value. We can see the exact rule the model uses, it explains how the product is valued at each level, make prediction, easy to understand.

The decision tree automatically selects the predictors it tests all the options and keep the best that explains the price difference.

Comparing the predicted values with the actual values. “RMSE: 0.29” It explains the model predicted price is off by 0.29 units from the actual price.

“R-squared: 0.76” decision tree explains 76% variation it captures key factors affecting the predicted price.

**Random forest Regression:**

Random forest regression doesn’t rely on one single decision tree it builds many single decision trees and averages the results. Each tree works like a set of if then rule, which can be interpreted easily. We can measure the importance of each factor contributing to the prediction.

Each decision path can be easily understood, making it more transparent. Multi collinearity do not affect the model.

Comparing the predicted values with the actual values. “RMSE: 0.22” On an average the model’s prediction is off by 0.22 units from the actual price.

‘’ R-squared: 0.86” 86% of the variation is captured by the Random Forest Regression, it means model captures most of key factors driving the normalized\_used\_price.

**Decision Tree Classification** (CART):

The model starts at the top with the important question that separates splits into high price vs low price, each split rule is clear that can be considered as business rule, we can trace the branches based on rules and how classification is done, variables affecting the model.

Comparing the predicted values with the actual values, the model tested on 686 phones, out of it 166 are correctly identified as high price, 442 as low price, there are some mistakes of 40 high price classified as low (wrongly), 38 low prices classified as high (wrongly).

The model prediction of high price is true by 81% (precision, how many phones we called high price were truly high), model successfully finds out 80% of actual high price phones in the data (recall). Model is reliable in identifying high price phones. F1 score 0.81 it means the model is reliable in identifying both high price and low price.

**Random Forest Classification:**

Random forest classification builds many trees and averages the results it combines many tress reduces the mistake compared to just one. It tells which features are important for the classification.

Comparing the predicted values with the actual values, out of 686 predictions the model correctly predicted 152 as high and 444 as low price, the accuracy of the model is 86% it predicted 86% correctly out of all the predictions. When the model classified high price and its truly high price 81% (Precision). The model finds out 74% high price only (recall), it misses 26% of actual high price. The model predicts correctly 89% of low price phones out all the predictions(precision). Out of all actual low price the model correctly identifies 92%(recall).

**Naïve Bayes**:

Naïve Bayes works on the conditional probability principle. It calculates how likely it belongs to a class based on the features. Each feature contributes its own probability so we can trace which feature is influencing the final prediction. The result is weighted probability, so we can say why the model picked certain class.

Comparing the predicted values with the actual values, the model gets 85% of prediction correct, identifying correct High price and low Price (Accuracy), the model is good at identifying low price 90% (Recall). The model predicts only 74% of high price correctly (sensitivity).

**Models’ performance**

**Multiple Linear Regression (MLR)**

RMSE: 0.252

R²: 0.824

Interpretation: The model explains 82% of variation (captures key features driving the price), prediction deviates 0.25 units from the normalized\_used\_price, showing strong accuracy. MLR is interpretable and highlights each feature contributing to the prediction. model is transparent and identifies key price drivers, but misses complex interactions between features.

**Decision Tree Regression**

RMSE: 0.291

R²: 0.765

Interpretation: The decision tree explains 72% variation (captures independent variables affecting the price), the predicated value deviates by 0.29 units from the actual normalized\_used\_price. Model provides less accurate but gives clear business rules, which is highly interpretable. Model is useful for generating transparent decision rule but accuracy is compressed as compared to others models.

**Random Forest Regression**

RMSE: 0.224

R²: 0.861

Interpretation: Random forest explains 86% variation (captures factors driving the price) and lowest RMSE 0.22 predicted price deviates very less as compared to rest of the model. It captures complex, non linear interaction between features by generating multiple decision tree and averaging it. Most accurate but less transparent than MLR and decision tree.

**Conclusion**: The Multiple linear regression and decision tree are preferred when transparency and interpretability are considered but models capture less variation explaining feature on price predication. Random forest is the best choice when predicted accuracy is the first priority.

**Decision Tree Classification**

Metrics:

Accuracy = 90.2%

Sensitivity = 80.6%

Specificity = 92.1%

Precision = 81.4%

Decision Tree provides clear business rule for classification it performs really good at identifying low price phones (Specificity) and captures most high price phones (sensitivity) and the overall accuracy is 90% which is solid as compared to other models.

**Random Forest Classification**

Metrics:

Accuracy: = 89.6%

Sensitivity = 73.8%

Specificity = 92.5%

Precision = 80.8%

Random Forest Classification increases slightly more on predicting low price phone correctly (Specificity) but compromises on catching high price (Sensitivity) and it not as transparent as decision tree.

**Naive Bayes Classification**

Metrics:

Accuracy: = 85.2%

Sensitivity = 74.3%

Specificity = 89.8%

Precision = 75.7%

Naive Bayes is the most transparent model, classification is based on probability, accuracy is lower as compared to decision tree and Random Forest classification, gives clear insights of how each feature contributes, misses out catching high price phones 74%(Sensitivity).

**Conclusion**: Decision tree provides clear rule of classification and gives best balance of accuracy and transparency, Random Forest misses out the high price phones (considering the high price phones as low price) generating losses, naive bayes is transparent but weak in performance and also misses out in catching high price phones.

**Model evaluation**

Random Forest Regression: It gives the strongest performance among the regression model, captures 86% most of the key factors driving the price. The average prediction error is relatively small on the normalized\_used\_price. The model addresses the challenges of setting accurate price to the used phones. The model combines many decision trees it handles complex patterns and avoid overfitting. the only issue with the model is not as transparent as MLR or decision tree, exact decision process is hidden. Still the model gives accuracy, reliability and strength.

Decision Tree Classification performs strongly, good balance between accuracy and interpretability, the model is good at classifying low price phones (92%) and capturing high value phones (81%) and the classification rule is easily interpreted, useful for business situation achieves accuracy 90% and provides clear rule for classification.

**Observation and Conclusion**

Random Forest Regression model captures complex relationships between features, IncMSE metric tells variable importance, when variable is shuffled how much the model error increases. Normalized\_new\_price (~45%) used market heavily tied to new price, sellers should benchmark new price when setting resale value for pricing strategy and buyers always compare to launched price.

Camera, screen size and storage these hardware specifications strongly influence the resale price, highlighting these features attract customers and justify the resale value of the product. Battery health and days used reduce the value, influence the resale price so sellers offering warranties can build trust, increase in sales. Device brand strongly influence the resale price, popular brands (Samsung and apple) can be sold at higher margin.

Random forest Regression model highlights important features that drive the used phone market, for trade -ins (buying back) looking at the high important features can predict value of used phones, sellers can purchase at low price valuable specifications phones and resale them at high margin, quickly specifications are in demand for used market. This makes model highly reliable in suggesting resale price and buying price.

Decision Tree classification was built to classify phones into high or low price segment. The models are easily interpretable and traceable if-then rule behind classification. The root node confirms majority of the phone classification is low priced segment. The decision tree split starts at the most important variable, normalized new price as the important driver, expensive phones most likely have high resale value, sellers should use as anchor point for classifying high value phones (greater than 5.5).

If internal memory is greater than 48 GB then it is a high chance of premium phone and lower memory phones depends weight and brand to classify in premium segment, heavy weights are classified as premium phones (due to bigger screen size and battery). Brands like Microsoft, Nokia, Oppo, Samsung, Sony and other classify as premium phones due to its reputation. Some brands Apple, Huawei, LG, Motorola, HTC, etc. are classified as low price due to their low depreciation effect. In some instance device with lesser price models (normalized new price) with high mega pixel front camera with reputed devices brand are classified as premium phones. The model provides rule-based insights, it helps business to avoid overpricing on low value phones and capture high margin on premium phones.

Classification is done on price range, higher priced model tends to be in premium segment because of high specifications, storage is considered as second important aspect that classifies into premium segment and weight is considered as it ties with battery and screen size. With low price range, good camera quality and reputed brand classifies into premium segment. Each factor contributes, with certain specifications of camera quality, brand reputation we classify into premium phones. Considering all aspects of specifications makes sense in classifying premium phones.

**Supporting Business Problem 1**

The random Forest Regression model provides data driven price recommendation, sellers can set competitive price and balance in profitability based on specifications driving the market.

**Supporting Business Problem 2**

Feature importance can give clear idea of what features drive purchase decisions, this ensure buying phone at right cost and in demand specifications and reselling at profitable price.

**Supporting Business Problem 3**

Decision tree classification offers transparent rule to phone segment, this enables sellers to classify high price phone or low price phone, captures high margin when selling high value phones.

The combined use of regression and classification model provides complete decision support for used phones, addressing business challenges and aligns with the goal.

**Suggestion**

In real world specification front camera mega pixel is always smaller than rear camera mega pixel, the dataset contains some entries where front camera is greater than rear camera. This indicates data entry error or exceptional outlier.