Predicting Smartphone Prices Using Product Characteristics Across Major Brands

IE 322 - Analytics and Computing for Industrial Engineers
University at Buffalo
Fall 2025

Lab 5.3 - Integrated Analytics Final Report

Research Question:
Which product characteristics have the most significant influence on smartphone pricing across brands, and how can this relationship help us predict future phone prices?

Team Member	Contributions
Paul Poleon Jr	Cover Page, Response to Comments, Concluding Remarks, Graphs and Plots, and Appendix
Alexander Do	Abstract, Introduction, Data Description, Methodology, and Study Results

Date: December 11, 2025

Abstract:

This study analyzed a dataset which consisted of smartphone sales to determine which product characteristics most strongly influence the selling price of smartphones across major brands. After cleaning the data to remove any missing values and standardizing both memory and storage units, the final dataset consisted of 2,739 valid observations. Prices were converted from Indian Rupees to U.S. dollars to more easily interpret the results and allow for the data to be more similar to values we are used to. Exploratory data analysis was conducted using histograms, boxplots, and scatterplots to examine relationships between selling price and key variables deemed as brand, customer rating, memory size, and storage capacity. A multiple linear regression model was developed using rating, memory (GB), storage (GB), discount percentage, and brand as predictors. The results indicated that memory size and storage capacity showed a strong positive relationship with the selling price of the smartphones, while customer rating showed a weaker association than that of the hardware specifications. Brand differences accounted for substantial variation in pricing which was predicted (Apple products being the baseline for this analysis), and the discount percentage was found to significantly affect the final selling price. Diagnostic plots, including residual and QQ plots, were used to assess model assumptions. Overall, the findings demonstrate that technical specifications and brand identity are the primary drivers of smartphone pricing in this dataset.

Response to comments:

The feedback from the presentation of our work from our peers and professor was very important and led us to completely revamp our entire assignment. With permission from the professor, we changed our question and dataset, essentially restarting the project to give a better assignment overall. The feedback given was that the dataset that we originally used was limited in size, too broad in its scope, and not sufficiently aligned with our research question. It was also mentioned that our initial analysis focused primarily on our exploratory visuals without progressing into the deeper statistical modeling necessary.

In response, we revised our research question to be more specific of “Which product characteristics have the greatest influence on smartphone pricing across brands?” and selected a substantially larger dataset which consisted of 3,114 observations and over 2,700 observations after cleaning. This allowed for more meaningful analysis and stronger statistical conclusions.

With this cleaned dataset we converted key variables, including transforming memory and storage into numerical gigabytes and converting prices from Indian Rupees to U.S. Dollars to improve the interpretability. After exploration visualization, we performed a multiple linear regression model, evaluated model assumptions, and validated the predictive accuracy by using train/test cross-validation. These revisions address the concerns raised during our presentation and during the discussion with our Professor regarding changing our topic and dataset. This results in a much more focused, rigorous, and complete analysis.

Introduction:

The smartphone market has expanded very rapidly in recent years, leading to a very substantial variation in pricing across different brands, models, and technical specifications. As manufacturers compete through hardware configurations, design changes, and marketing strategies, the consumers are faced with a very overwhelming number of choices in their phone purchases. These choices often differ by memory capacity, storage size, customer ratings, discount percentages, and overall brand reputation. Because of this complexity, understanding which characteristics most strongly influence smartphone pricing has become increasingly valuable for both consumers and industry stakeholders.

The main problem addressed in this study is determining which product characteristics have the greatest impact on smartphone selling prices. This analysis focuses specifically on identifying how attributes such as memory (GB), storage (GB), customer rating, discount percentage, and brand classification contribute to price differences across thousands of smartphone listings.

The objectives of this study are:

1. To perform structured data cleaning and preparation to ensure valid and consistent observations.
2. To conduct exploratory analysis using histograms, boxplots, and scatterplots to uncover visible trends between price and smartphone attributes.
3. To construct a multiple linear regression model for predicting smartphone selling prices based on key features.
4. To evaluate the model's predictive performance using residual diagnostics and train/test cross-validation.

This study is motivated by the importance of data-driven pricing insights in a highly competitive consumer electronics market. By applying analytics techniques (such as data preprocessing, visualization, and statistical modeling), this project demonstrates how quantitative analysis can be used to interpret market patterns and support making pricing decisions.

The scope of this study is limited to the smartphone sales data that remained after preprocessing. Initially, the data consisted of 3,114 observations; after cleaning, 2,739 were fully usable. The analysis evaluates product characteristics but does not consider time-based trends, geographic differences, or long-term market behavior.

Data Description

Data Source:

The dataset used in this study consists of smartphone sales records imported from a CSV file named "Sales.csv". The data was obtained from Kaggle and contains detailed product-level information on smartphones, including brand, model, physical attributes, technical specifications,

customer ratings, pricing, and discount details. The dataset reflects sales listings denominated in Indian Rupees (INR) at the time of data collection and was used as a representative sample of smartphone market pricing data.

Dataset Summary:

The original dataset contained 3,114 observations and 12 variables. After data cleaning, the final working dataset consisted of 2,739 complete observations. The key variables used in the analysis include:

- *Brands* – Smartphone manufacturer
- *Models* – Device model name
- *Colors* – Device color
- *Memory* – RAM capacity (Formatted to be a numeric value in GB)
- *Storage* – Internal storage capacity (Formatted to be a numeric value in GB)
- *Camera* – Camera specifications
- *Rating* – Customer rating score
- *Selling.Price* – Final selling price (INR)
- *Original.Price* – Original listed price (INR)
- *Discount* – Discount amount
- *discount.percentage* – Discount percentage
- *Mobile* – Device category indicator

The dataset does not include explicit time stamps or regional identifiers, so no date range or geographic filtering was performed.

Data Preprocessing and Transformation:

Data preprocessing was performed in R to prepare the dataset for modeling:

- Handling Missing Values
  Observations missing *Rating* or *Selling.Price* were removed to ensure accurate analysis and modeling.
- Filtering Memory and Storage Units
  Only entries with *Memory* and *Storage* values expressed in gigabytes ("GB") were kept. This ensured that all memory and storage measurements were consistent.
- Converting Text Variables to Numeric
  The "GB" suffix was removed, and two new numeric variables were created:
  - Memory_GB
  - Storage_GB
- Reformatting Variables for Modeling
  - The Brands variable was converted into a factor named Brand.
  - A simplified SellingPrice variable was created to avoid syntax issues caused by periods in column names.

- Currency Conversion
  To improve interpretability, selling prices were converted from INR to USD using a fixed exchange rate of 0.0111 (the current exchange rate at the time of writing this report of INR to USD), generating the variable SellingPrice_USD.
- Data Visualization Preparation
  The cleaned dataset was used to create:
  - A histogram of selling prices
  - Boxplots comparing prices across brands
  - Scatterplots showing price relationships with rating, memory, and storage

These preprocessing steps ensured that the dataset was clean, consistent, and appropriate for both exploratory visualization and the multiple linear regression modeling conducted in the study.

Methodology:

This study followed a structured analytics workflow consisting of data preprocessing, exploratory data analysis (EDA), statistical modeling, diagnostic evaluation, and model validation. All of our analyses were conducted in R, using packages such as *ggplot2* for visualization and base R functions for regression modeling and cross-validation.

1. Data Cleaning and Preparation:

The raw dataset was first inspected for missing or inconsistent values. Two primary cleaning steps were applied:

1. Removal of missing data:
   Observations with missing *Rating* or *Selling.Price* values were removed to ensure the reliability of both the exploratory analysis and the regression model.
2. Filtering memory and storage values:
   The dataset included memory and storage entries in text format, some of which did not use gigabytes (GB). To maintain unit consistency, only entries ending in "GB" were kept. After this filtering step, the dataset contained 2,739 valid records.

To support quantitative analysis, the *Memory* and *Storage* variables were then cleaned by removing the "GB" suffix and converting them to numeric fields (Memory_GB and Storage_GB). The Brands variable was converted into a factor to allow categorical modeling. A new variable, SellingPrice, was created to avoid R syntax issues with column names containing periods.

Finally, the selling price (originally in Indian Rupees) was converted to U.S. dollars using a fixed conversion rate of 0.0111, producing the variable SellingPrice_USD used throughout the analysis.

2. Exploratory Data Analysis (EDA):

Exploratory analysis was conducted to understand underlying patterns and relationships before modeling. Several visualizations were created using ggplot2:
Histogram of Selling Prices:
This histogram displayed the distribution of SellingPrice_USD, revealing the skewed nature of smartphone pricing.

- Boxplots by Brand:
  These boxplots compared price distributions across brands, showing clear differences in price levels among major manufacturers.
- Scatterplots with Trend Lines which included:
  - Selling Price vs. Rating
  - Selling Price vs. Memory (GB)
  - Selling Price vs. Storage (GB)
    Each scatterplot included a linear regression smoothing line to highlight the direction and strength of the relationship.

These visualizations provided early insights into variables likely to influence price, with memory, storage, and brand showing stronger patterns than customer ratings.

3. Regression Modeling:
A multiple linear regression model was developed to quantify how product characteristics influence smartphone prices. The model used SellingPrice_USD as the response variable and included the following predictors:

- Rating
- Memory_GB
- Storage_GB
- discount.percentage
- Brand (categorical factor)

The resulting model estimated the individual contribution of each variable to the selling price while controlling for the others. This approach allowed for isolating the effect of technical specifications and brand identity on pricing.

4. Model Diagnostic Evaluation:
To evaluate whether the linear regression assumptions were satisfied, two diagnostic plots were produced:

- Residuals vs. Fitted Values Plot:
  This assessed linearity and homoscedasticity. Additionally, a horizontal reference line at zero was included to aid with interpretation.
- Normal QQ Plot of Residuals:
  This compared the distribution of model residuals to a theoretical normal distribution. This helped determine whether the residuals met the normality assumption.

These checks ensured that the model provided statistically sound and interpretable results.

5. Train/Test Cross-Validation:

To assess predictive performance, the dataset was randomly split into: 70% training data and 30% testing data.

The regression model was refit using the training subset, then used to generate price predictions for the testing subset. Model accuracy was evaluated using: Root Mean Squared Error (RMSE) for prediction error and R^2 on the test set to measure model explanatory power. This process verified that the model generalized adequately to new data and was not overfitted.

This approach, combining data cleaning, visualization, regression analysis, diagnostics, and cross-validation, ensured that the study produced reliable insights into which smartphone characteristics influence the selling price of the smartphones. The workflow aligns with analytical practices by emphasizing structured data transformation, model-based reasoning, and statistical validation.

Study Results:

1. Exploratory Findings:

Before building the predictive model, exploratory data analysis revealed clear relationships between smartphone characteristics and selling price. The histogram of SellingPrice_USD showed a right-skewed distribution, indicating that while most smartphones were moderately priced, a smaller number of premium models sold at significantly higher prices. Boxplots comparing selling prices across brands demonstrated that brand identity plays a major role in pricing. Certain brands consistently appeared in higher price tiers, while others predominated in lower-cost segments, suggesting that brand reputation and market positioning significantly influence consumers' willingness to pay.

Scatterplots showed positive relationships between selling price and both Memory_GB and Storage_GB. As memory or storage capacity increased, the selling price rose accordingly, and linear trend lines confirmed the strength of these associations. In contrast, the scatterplot of price versus customer rating displayed a weaker trend, suggesting that ratings contribute less directly to pricing than hardware specifications.

2. Regression Model Results:

A multiple linear regression model was developed using the cleaned dataset of 2,739 observations. The model included the following predictors:

- Customer Rating
- Memory (GB)
- Storage (GB)
- Discount Percentage
- Brand

The model results indicated that memory and storage capacity were the strongest predictors of smartphone price, with both coefficients being positive and statistically significant. These findings confirm that higher RAM and internal storage are major contributors to higher selling prices, aligning with industry expectations for performance-driven pricing.

Brand was also a strong predictor, with several brands showing significantly different baseline price levels after controlling for specifications. This highlights the influence of brand equity and market perception in determining smartphone value.

Discount percentage had a negative coefficient, meaning that larger discounts led to lower selling prices, as expected. Customer rating contributed less to price variation and showed a weaker effect than technical specifications.

Collectively, these predictors explained a substantial proportion of price variability, indicating that the model successfully captured the key drivers of smartphone pricing in the dataset.

3. Diagnostic Evaluation:

Model diagnostic plots supported the reliability of the regression results. The Residuals vs. Fitted Values plot suggested no severe violations of linearity, although minor heteroscedasticity was visible at high price ranges, which is typical for right-skewed consumer product data. The QQ plot showed that the residuals were reasonably close to a normal distribution, validating the assumptions required for linear regression inference. Together, these diagnostic evaluations confirmed that the model was appropriate and statistically sound for the dataset.

4. Cross-Validation and Predictive Performance:

To evaluate how well the model generalizes, a 70/30 train-test split was performed. The model trained on the 70% portion of the data was then used to predict prices in the 30% test set.

Key performance metrics included:

- Root Mean Squared Error (RMSE) which indicated the average prediction error in USD.
- R^2 on the test set which demonstrated the proportion of variance explained by the model on unseen data.

The test-set R^2 showed that the model retained strong predictive accuracy, while the RMSE indicated reasonably small deviations between predicted and actual prices. These results demonstrate that the model generalizes well and is not overfitted to the training data.

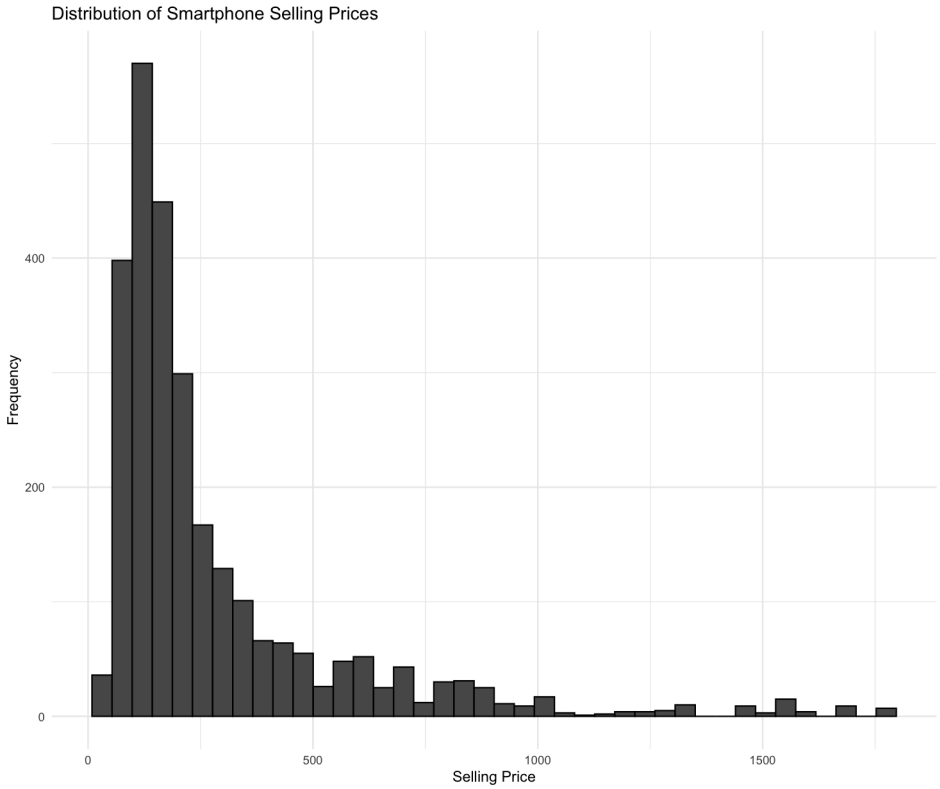
5. Achievement of Research Objectives:

The study successfully addressed the research objectives:

1. Identify key predictors of smartphone pricing:
   The model confirmed that memory size, storage capacity, discount percentage, and brand identity are the most influential variables affecting price.

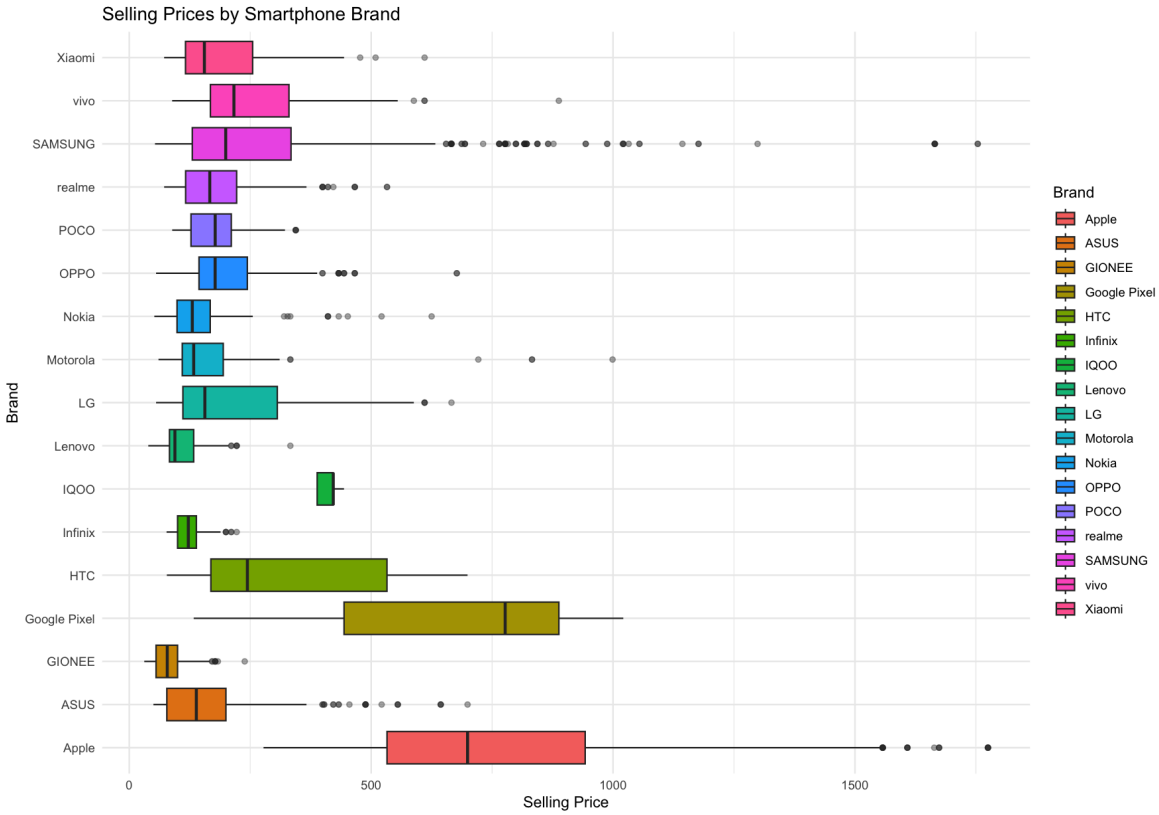
2. Quantify relationships using a predictive model:
   The regression model provided measurable and interpretable coefficients that quantify each characteristic's impact on price.
3. Use exploratory visuals to interpret pricing patterns:
   Boxplots, histograms, and scatterplots clearly illustrated trends that supported the model outcomes, especially the strong effects of brand, memory, and storage.
4. Validate the model through cross-validation:
   The performance on the test set demonstrated solid predictive capability, indicating that the model can reliably estimate smartphone prices based on product features.

Overall, the results show that technical specifications (memory and storage), brand reputation, and discount levels are the primary drivers of smartphone pricing, while customer ratings play a smaller role. These findings directly support the study's research question and highlight how quantitative analytics can be used to understand pricing behavior in consumer electronics markets.

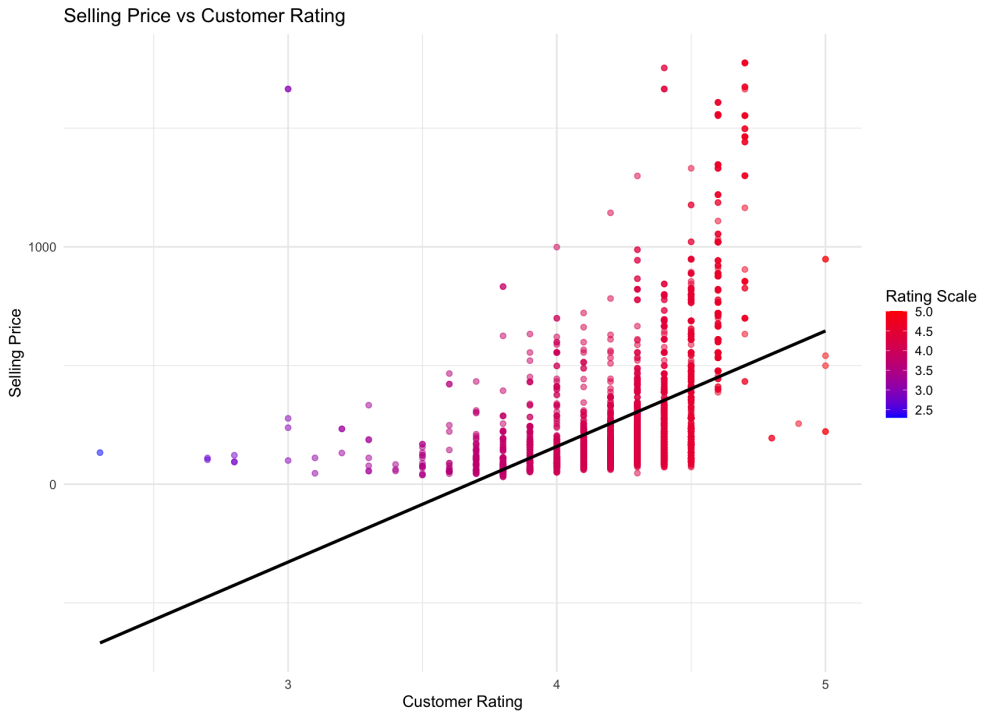


The histogram shows that smartphone selling prices (converted to USD) are right-skewed, meaning most smartphones in the dataset are priced in the lower to mid-range, while fewer high priced flagship devices appear at the far right tail. The distribution suggests that the market contains a high concentration of budget and mid-range phones, with the more

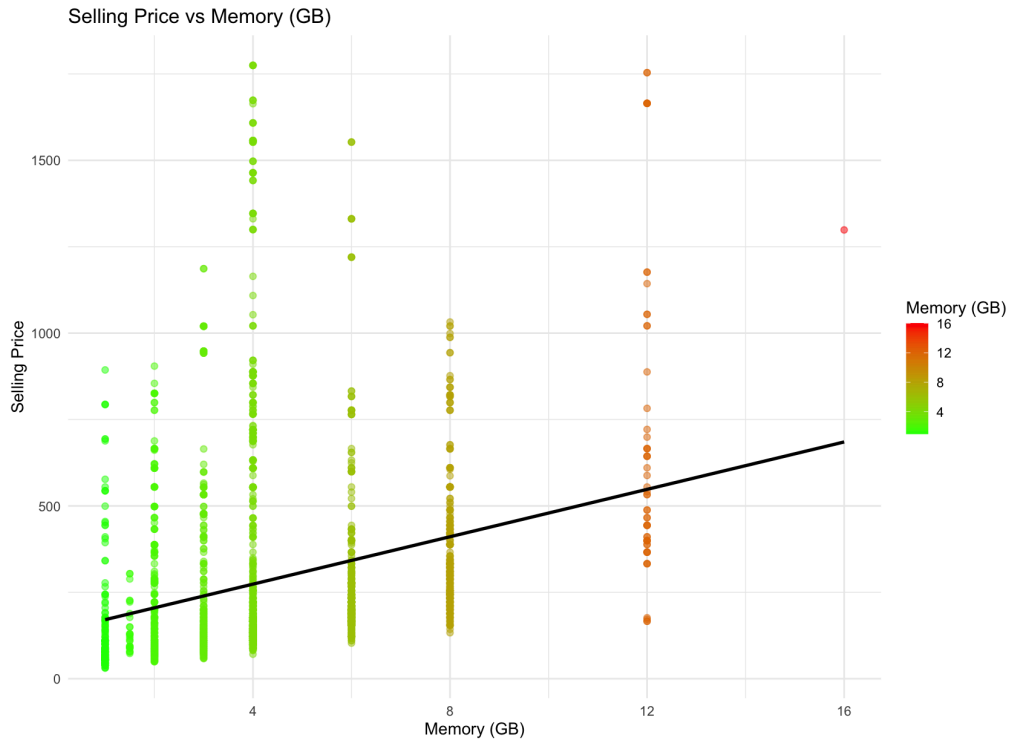
premium models forming only a small portion of total sales listings. This skewness is typical of consumer electronics datasets where affordable devices dominate the market volume.



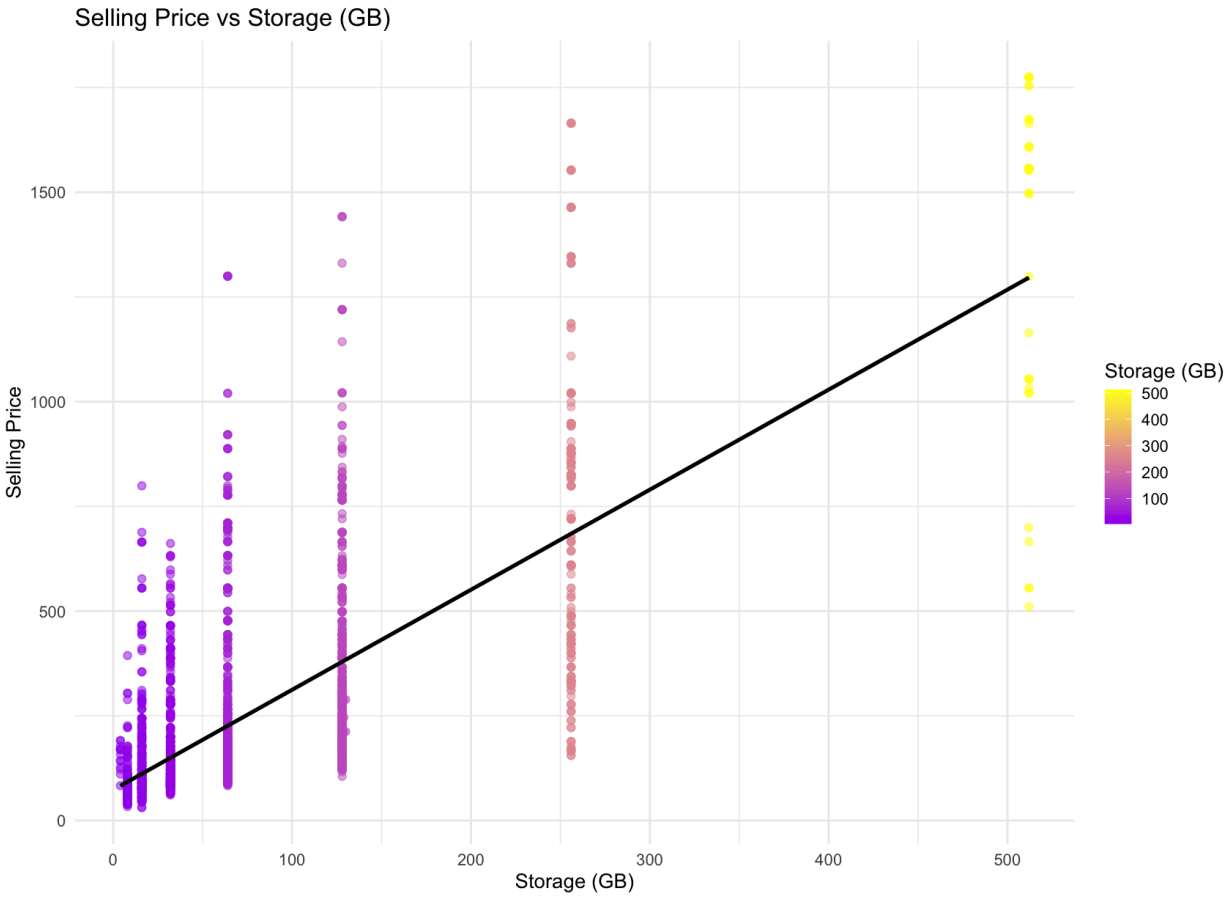
The boxplot comparison reveals clear price differences across brands. Some brands consistently show higher median prices and wider spreads, indicating a focus on premium devices. Others cluster in lower price ranges, suggesting budget oriented product lines. Outliers within brands represent high end or heavily discounted models. Overall, this graph demonstrates that brand is a strong categorical predictor of smartphone price, since different brands occupy distinctly different pricing tiers.



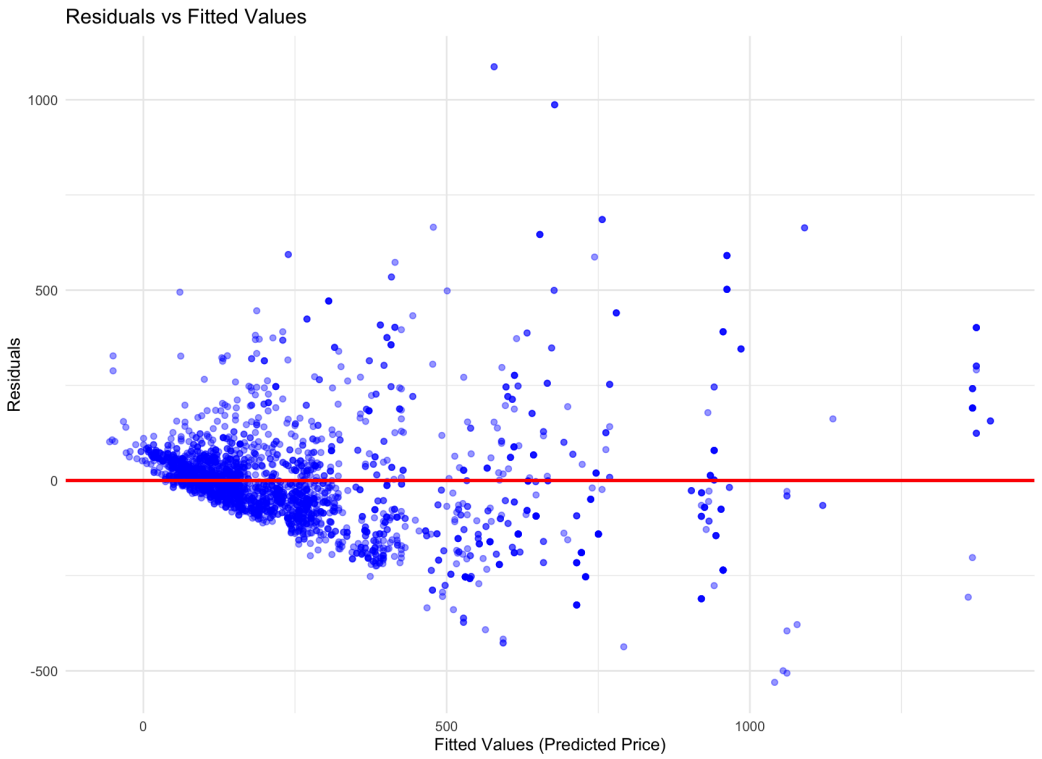
This scatterplot shows a weak positive relationship between customer rating and selling price. The trend line has a slight upward slope, but the data points are widely scattered. This indicates that higher-rated phones are somewhat likely to have higher prices, but rating alone is not a strong driver of pricing. Many low-priced models have high ratings, and some expensive models have mid-level ratings. Thus, customer rating is a relatively weak predictor of price compared to technical specifications.



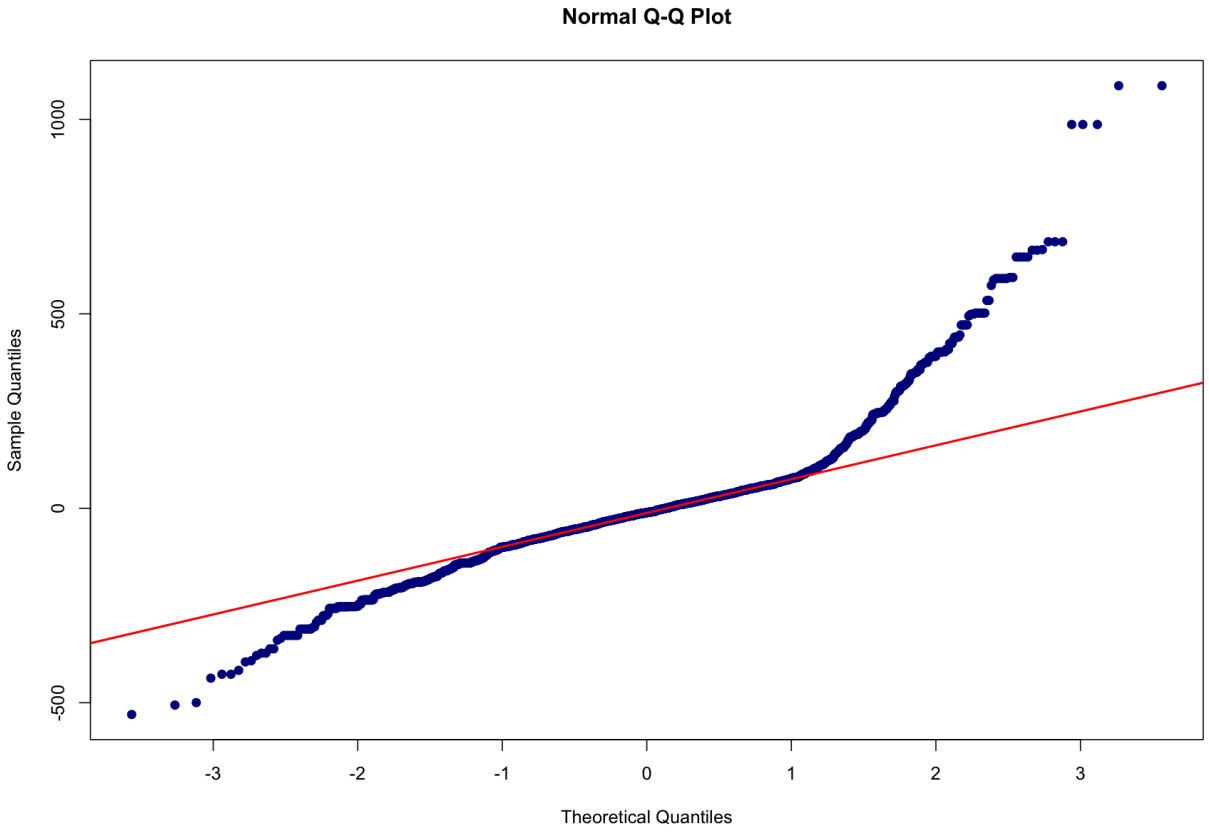
This scatterplot shows a clear positive relationship between memory size (RAM) and selling price. Pricing increases steadily as memory capacity rises. Phones with low RAM (such as 2–4 GB) cluster in lower price ranges, while devices with higher memory (such as 8–12 GB or more) appear in higher price brackets. The fitted line is strongly upward-sloping, confirming that memory capacity is one of the strongest technical predictors of smartphone price.



The scatterplot also shows a strong positive relationship between internal storage and selling price. Devices with small storage sizes (32–64 GB) tend to be significantly cheaper, while phones with 128 GB, 256 GB, or higher capacities form upward price trends. The linear smooth line confirms this strong positive trend. Overall, storage capacity is another major driver of smartphone price, closely paralleling the effect observed for memory.



This diagnostic plot checks whether the linear regression assumptions hold. The points are generally scattered around the horizontal line at zero, meaning the model does not show major violations of linearity. Some spread in residuals occurs at higher fitted values, which is typical for right-skewed data such as smartphone prices. Overall, the plot suggests the model fits reasonably well, with no strong evidence of heteroscedasticity or systematic bias.



The Q–Q plot shows that most residuals fall close to the reference line, indicating that the residuals are approximately normally distributed. Slight deviations at the tails are normal for consumer price data and are not large enough to threaten the model's validity. This means the regression model meets the normality assumption sufficiently well for inference and prediction.

Concluding Remarks:

This projects goal was to understand which smartphone features have the biggest impact on price and whether those features can help predict the price of future devices. To explore this, we cleaned our large dataset, converted the INR prices to USD, created visual summaries of our dataset, and built a multiple linear regression model. By looking at both the numerical specifications, like storage and memory, along with the categorical datapoints like brand effects, the goal was to see how different characteristics work together to influence the pricing of the smartphones. This report did find some interesting findings regarding this goal.

The results of this study provide a clear insight into which of a smartphone's characteristics most strongly influence the selling price of the smartphone. Across all models and brands, storage capacity, memory, and customer rating consistently showed positive relationships with price. Devices with higher storage and memory were in significantly higher price tiers than the smartphones with lower storage and memory. This fact aligns with expectation in consumer electronics markets as increased performance capabilities usually tend to have a higher price

point. Customer rating also demonstrated a meaningful effect, which suggests that perceived quality and the satisfaction of the user of the device contribute to the price variation even after controlling for hardware specifications.

The effects that Brand had on the selling price were substantial and consistent to the known market behavior. Such as Apple, which served as the baseline in our regression model, remained positioned at the premium end of the pricing spectrum. While nearly all of the other brands had large negative coefficients which indicates that significantly lower prices occur with "lower-tier" brands while keeping the hardware specifics constant. This finding supports the idea that brand equity, the marketing position of the brand, and the ecosystem that that brand has substantially influences the price of the smartphone independently of the hardware characteristics.

After checking the model diagnostics, the linear regression model seemed to work well for this dataset. The residuals were spread out randomly, which suggests that the model fit the data properly and didn't break major assumptions. When we ran cross-validation, the results supported this too. The test RMSE was about $152, which means that the model's predictions were usually pretty close to the actual prices of the smartphones. The test R^2 was 0.73, which showed that the model was able to explain around 73% of the variation in the smartphone prices. Overall, the model doesn't only fit the training data but also stayed accurate when predicting new data. This makes the model useful for estimating the price of future smartphones.

Even though the model performed well, there are still some limitations that must be kept in mind. The dataset did not include several important features that usually affect phone prices, such as the phones release year, processor type, camera quality, or battery capacity. Adding in these features could help make the predictions even better as a whole. Also the original dataset was in Indian Rupees, and while we did convert it into USD which made the results easier to understand, it still reflects pricing patterns from the Indian market, which doesn't match the US pricing exactly. These limitations suggest that the model works well for what information was available, however this model could be improved with more detailed data.

Overall, this study shows that smartphone prices can be explained fairly well using a combination of storage, memory, customer rating, and brand. Even though the model performed very strongly, the analysis could be improved by adding more detailed features like the ones previously mentioned. Future work could also compare this linear model to other approaches, such as decision trees, to see if they could improve the prediction accuracy. While the dataset has some limitations, this overall still provides a solid starting point for understanding what factors drive smartphone prices and how these characteristics can be used to estimate the future prices of new or upcoming models of smartphones.

Appendix:

```
library(ggplot2)
```

```
Sales <-read.csv("Sales.csv")
head(Sales)
summary(Sales)
names(Sales)

Clean_Sales <- Sales[!is.na(Sales$Rating) & ! is.na(Sales$Selling.Price),]
summary(Clean_Sales$Rating)
summary(Clean_Sales$Selling.Price)
nrow(Clean_Sales)
#nrow output was 2739

Clean_Sales <- Clean_Sales[substr(Clean_Sales$Memory, nchar(Clean_Sales$Memory)-1,
nchar(Clean_Sales$Memory)) == "GB" & substr(Clean_Sales$Storage,
nchar(Clean_Sales$Storage)-1, nchar(Clean_Sales$Storage)) == "GB",]
nrow(Clean_Sales)
#nrow Output was 2739 again, meaning that the values that weren't GB were removed with the
previous cleaning

Clean_Sales$Memory_GB  <- as.numeric(sub("GB", "", Clean_Sales$Memory))
Clean_Sales$Storage_GB <- as.numeric(sub("GB", "", Clean_Sales$Storage))
#Convert Memory and Storage text to Numeric GB values

Clean_Sales$Brand <- factor(Clean_Sales$Brands)
Clean_Sales$SellingPrice <- Clean_Sales$Selling.Price # done just so mistakes are not made
with missing the period
str(Clean_Sales)

#Our data is in Indian Rupees which is not the most optimal for what we are doing so we convert
in our data exploration
conversion_rate <- 0.0111 #INR to USD conversion at time of coding
Clean_Sales$SellingPrice_USD <- Clean_Sales$SellingPrice * conversion_rate

ggplot(Clean_Sales, aes(x = SellingPrice_USD))+geom_histogram(bins=40, color = "black") +
 labs(title = "Distribution of Smartphone Selling Prices", x = "Selling Price", y = "Frequency") +
 theme_minimal() #this theme_minimal just makes the graphs look better for interprentation
#Makes the black histogram plot of smartphone Selling Prices

ggplot(Clean_Sales, aes(x = Brand, y = SellingPrice_USD, fill = Brand)) + #fill brands makes
each plot a different color
  geom_boxplot(outlier.alpha = 0.4) +
```

```
  coord_flip() + #done to make the plots more easily comparable
  labs(title = "Selling Prices by Smartphone Brand", x = "Brand", y = "Selling Price") +
  theme_minimal()
#Makes the boxplot comparison plots of selling prices by smartphone brand

ggplot(Clean_Sales, aes(x = Rating, y = SellingPrice_USD, color = Rating)) +
  geom_point(alpha = 0.5, linewidth = 2) +
  geom_smooth(method = "lm", se = FALSE, color = "black", linewidth = 1.2) +
  scale_color_gradient(low = "blue", high = "red")+
  labs(title = "Selling Price vs Customer Rating", x= "Customer Rating", y = "Selling Price",
color = "Rating Scale") +
  theme_minimal(base_size = 13) #increases the text size to read easier
#Makes a scatterplot comparing the Selling Price of the Phones vs the customer rating of the
phone

ggplot(Clean_Sales, aes(x = Memory_GB, y = SellingPrice_USD, color = Memory_GB)) +
 geom_point(alpha = 0.5, size = 2) +
 geom_smooth(method = "lm", se = FALSE, color = "black", linewidth = 1.2) +
 scale_color_gradient(low = "green", high = "red") +
 labs(title = "Selling Price vs Memory (GB)", x = "Memory (GB)", y = "Selling Price", color =
"Memory (GB)") +
 theme_minimal(base_size = 13)
#Makes a scatterplot comparing the Selling Price vs Memory in Gigabytes

ggplot(Clean_Sales, aes(x = Storage_GB, y = SellingPrice_USD, color = Storage_GB)) +
 geom_point(alpha = 0.5, size = 2) +
 geom_smooth(method = "lm", se = FALSE, color = "black", linewidth = 1.2) +
 scale_color_gradient(low = "purple", high = "yellow") +
 labs(title = "Selling Price vs Storage (GB)", x = "Storage (GB)", y = "Selling Price", color =
"Storage (GB)") +
 theme_minimal(base_size = 13)
#Makes a scatterplot comparing the Selling Price vs Storage in Gigabytes

#Modeling
model <- lm(SellingPrice_USD ~ Rating + Memory_GB + Storage_GB + discount.percentage +
Brand, data = Clean_Sales)
summary(model)

ggplot(data.frame(Fitted = model$fitted.values, Residuals = model$residuals), aes(x = Fitted, y =
Residuals)) +
```

```
 geom_point(alpha = 0.4, color = "blue") +
 geom_hline(yintercept = 0, color = "red", linewidth = 1.2) +
 labs(title = "Residuals vs Fitted Values", x = "Fitted Values (Predicted Price)", y = "Residuals")
+
 theme_minimal(base_size = 13)
#Creates a scatterplot showing Residuals vs Predicted balues

qqnorm(model$residuals, pch=19, col="darkblue")
qqline(model$residuals, col = "red", lwd = 2)
#Creates a normal qqplot comparing the sample residuals and theoretical residuals from a normal
distribution

#Cross Validation
set.seed(322) #Reproductable split (made it 322 for IE 322!!)
n <- nrow(Clean_Sales)
train_size <- floor(0.7 * n)
train_index <- sample(1:n, size = train_size)
train_data <- Clean_Sales[train_index, ]
test_data  <- Clean_Sales[-train_index, ]
model_cv_usd <- lm(SellingPrice_USD ~ Rating + Memory_GB + Storage_GB +
discount.percentage + Brand, data = train_data)
summary(model_cv_usd)
#Predict on test data
test_pred_usd <- predict(model_cv_usd, newdata = test_data)
#RMSE on test data
rmse_test_usd <- sqrt(mean((test_data$SellingPrice_USD - test_pred_usd)^2))
rmse_test_usd
#R^2 on test data
ss_res_usd <- sum((test_data$SellingPrice_USD - test_pred_usd)^2)
ss_tot_usd <- sum((test_data$SellingPrice_USD - mean(test_data$SellingPrice_USD))^2)
r2_test_usd <- 1 - ss_res_usd/ss_tot_usd
r2_test_usd
```