FUNDAMENTALS OF DATA ANALYTICS

PROJECT REPORT: MOBILE PRICES DATA ANALYSIS

Prepared by Group - 4

**Salman Khan**

**Gowri Rohith Chirra**

**Ratnesh Chilamkurthi**

**Bhanuprasad Unnam**

**Manikanta Modugula**

**Rishi Vardhan Reddy Navaru**

**Professor: Ankur Arora**

**Table of Contents**

1). Introduction

2) . Dataset Description

3) . Data Cleaning and Preprocessing

4) . Data Summarization

5) . Hypothesis Testing

6). F- Test

7). Categorical data analysis

8). Linear Regression

9). Conclusion

10). Annexure (Graphs and histograms)

11). R-Script

**1. Introduction**

Mobile phones are an integral part of our daily lives, and understanding the factors that influence their prices can provide valuable insights for both consumers and manufacturers. In this project, we analyze a dataset containing various features of mobile phones, such as battery power, RAM, camera quality, and more. We aim to understand the relationship between these features and the price range of mobile phones.

The rapid evolution of mobile technology has led to a diverse range of mobile phones available in the market. These devices vary in terms of hardware specifications, features, and, consequently, their prices. For consumers, choosing the right mobile phone that meets their needs and budget can be a challenging task.

**2. Dataset Description**

According to the Data set given the data set consists of numerical data about mobile prices and there are some categorical data also inserted in the data set In the dataset we have 2000 observations of data and there are a total 21 variables including numerical 14 and 7 categorical data

•**Numerical data**: Numerical data, also known as quantitative data, is a type of data that is expressed in numerical form and is used for performing various mathematical calculations and statistical analyses. It represents measurable quantities and can take on a wide range of values.

First, we will investigate what dataset we have and what are the names, and these are the numerical data we have in the following data set.

|  |  |
| --- | --- |
| **COLUMN NAMES** | **DETAIL COLUMN NAMES** |
| ## 1 battery power | Battery Power in MAH |
| ## 2 clocks speed | Clock speed in GHz |
| ## 3 fc | Font Camera megapixel |
| ## 4 int\_memory | Internal Memory in GB |
| ## 5 m\_dep | Mobile Depth in cm |
| ## 6 mobile\_wt | Weight of Mobile phone in grams |
| ## 7 n\_cores | Number of cores |
| ## 8 pc | Primary Camera Mega Pixels |
| ## 9 px\_height | Pixel Height of the display |
| ## 10 px\_width | Pixel Width of the display |
| ## 11 ram | Random Access Memory in MB |
| ## 12 sc\_h | Screen Height in cm |
| ## 13 sc\_w | Screen Width in cm |
| ## 14 talk\_time | Talk Time in Hours |

**Categorical data**: Categorical data, also known as qualitative data, is a type of data that represents categories or labels rather than numerical values. It is used to group or classify items based on specific characteristics or attributes.

These are the categorical data that have been given in the data set.

|  |  |
| --- | --- |
| **COLUMN NAMES** | **DETAIL COLUMN NAMES** |
| ##1 blue | Bluetooth support (1 for yes, 0 for no). |
| ##2 dual\_sim | Dual SIM support (1 for yes, 0 for no). |
| ##3 four\_g | 4G support (1 for yes, 0 for no). |
| ##4 three\_g | 3Gsupport (1 for yes, 0 for no). |
| ##5 touch\_screen | Touchscreen support (1 for yes, 0 for no). |
| ##6 wifi | Wi-Fi support (1 for yes, 0 for no). |
| ##7 price\_range | Price range (0 - low cost, 1 - medium cost, 2 - high cost, 3 - very high cost). |

**3. Data Cleaning and Preprocessing**

In the data cleaning and preprocessing phase, we addressed the following:

Checked for missing values (none found in this dataset).

# Check for missing values

missing\_values <- sum(is.na(data))

print(paste("Missing values:", missing\_values))

By the analysis of mobile prices dataset, we got missing values “0”.

**Outlier Analysis:**

We performed an outlier analysis on the 'battery\_power' variable in the dataset and We started by creating a box plot to visualize the distribution of battery power. This helped me identify any potential outliers. Next, we calculated the quartiles (Q1 and Q3) and the interquartile range (IQR) to define the lower and upper bounds for outliers.

After applying the IQR method, we found that there were no outliers in the 'battery\_power' variable. All data points fell within the acceptable range, and no values were identified as outliers.

Therefore, we have done an outlier analysis for all the other variables and we found out there were no outliers in the entire data set as we determined by seeing the box plot diagram we can refer to figures 1 to 14 where all the data points fell within the range and no values were identified as outliers.

**4. Data Summarization**

We obtained summary statistics that include mean and median and Inter quartile ranges for numerical variables and created visualizations to better understand the data.

**Mean**: The mean is a known word for the average. It's a way to figure out what's typical or normal in a group of things.

* In everyday life, we use the means to understand what's typical. For example, if you're talking about the average battery capacity of a 10 mobile phones, you're using the mean. Or if you're figuring out the average mobile weight in the mobiles that is used by the people in 2023, that's also the mean.

**Median:** The median is like the middle point in a list of things, and it's a way to find what's right in the middle. If you take from the mobile prices data, the median of the battery power is 1226.0 Mah capacity.

**Quartile Range**: The "interquartile range" is a measure of how spread out the mobile data of a particular battery power capacity (the second and third groups). It tells you the difference between the earnings of the friend who earns more than 25% of your friends but less than 75% of your friends (the upper quartile of the middle group) and the friend who earns more than 25% but less than 75% of battery capacity.

We obtained summary statistics for numerical variables and Creating visualizations to better understand the data.

The summary statistics provided give a good overview of the numerical variables in our dataset. Some of the statistics and make observations of the dataset are included below:

1. Battery Power: The range of battery power is from 501 to 1998 with a mean value of approximately 1238.5. This suggests a wide variation in battery power across the mobile phones in the dataset.

2. RAM: The RAM values range from 256 to 3998 with a mean value of approximately 2124. RAM is an essential factor affecting a smartphone's performance and this wide range indicates significant variability.

3. Internal Memory: The internal memory (int\_memory) ranges from 2 to 64 with a mean of approximately 32.05. This indicates variation in internal memory capacity.

4. Front Camera (fc): The front camera ranges from 0 to 19 with a mean value of approximately 4.309. This variable represents the megapixels of the front camera.

5. Price Range: The target variable “price\_range” has values ranging from 0 to 3 indicating different price ranges or categories.

6. 4G and Wi-Fi Support (four\_g wifi): These binary variables have means around 0.52 indicating that about half of the phones support 4G and Wi-Fi.

These summary statistics provide a broad view of the dataset. To gain more insights you can explore relationships between variables through visualizations hypothesis testing and further analysis. If you have specific questions or hypotheses, you'd like to test please let me know and I can assist you further.

**Summary Statistics for Mobile Prices Data Set Variables**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Feature** | **Min.** | **1st Qu.** | **Median** | **Mean** | **3rd Qu.** | **Max.** |
| **battery power** | 501.0 | 851.8 | 1226.0 | 1238.5 | 1615.2 | 1998.0 |
| **clock\_speed** | 0.500 | 0.700 | 1.500 | 1.522 | 2.200 | 3.000 |
| **fc** | 0.000 | 1.000 | 3.000 | 4.309 | 7.000 | 19.000 |
| **int\_memory** | 2.00 | 16.00 | 32.00 | 32.05 | 48.00 | 64.00 |
| **m\_dep** | 0.100 | 0.200 | 0.500 | 0.5018 | 0.800 | 1.000 |
| **mobile\_wt** | 80.0 | 109.0 | 141.0 | 140.2 | 170.0 | 200.0 |
| **n\_cores** | 1.000 | 3.000 | 4.000 | 4.521 | 7.000 | 8.000 |
| **pc** | 0.000 | 5.000 | 10.000 | 9.916 | 15.000 | 20.000 |
| **px\_height** | 0.0 | 282.8 | 564.0 | 645.1 | 947.2 | 1960.0 |
| **px\_width** | 500.0 | 874.8 | 1247.0 | 1251.5 | 1633.0 | 1998.0 |
| **ram** | 256 | 1208 | 2146 | 2124 | 3064 | 3998 |
| **sc\_h** | 5.00 | 9.00 | 12.00 | 12.31 | 16.00 | 19.00 |
| **sc\_w** | 0.000 | 2.000 | 5.000 | 5.767 | 9.000 | 18.000 |
| **talk\_time** | 2.00 | 6.00 | 11.00 | 11.01 | 16.00 | 20.00 |

**Mode Data:**

Here we are going to know what the meaning of mode. The mode is the value that appears most frequently in a set of data.

For example, take real life sample and consider that you have a list of mobile phones, like various battery capacities which of the mobiles have common and same battery capacity will be shown. If you have taken categorical data, for example from the dataset if we take four g data of a mobile phone which is mentioned as 1 or 0 which gives mod data as 1 so we can say most of the phones or Four g phones. It’s the one the most on that list.

**Finding mode values for all numeric variables data**

|  |  |
| --- | --- |
| **Attributes** | **Values** |
| battery power | 1589.0 |
| clock speed | 0.5 |
| fc | 0.0 |
| Internal memory | 27.0 |
| m\_dep | 0.1 |
| mobile\_wt | 182.0 |
| n\_cores | 4.0 |
| pc | 10.0 |
| px\_height | 347.0 |
| px\_width | 874.0 |
| ram | 2227.0 |
| sc\_h | 17.0 |
| sc\_w | 1.0 |
| talk\_time | 7.0 |

**Finding Mode values of Categorical variables**

|  |  |
| --- | --- |
| **Attributes** | **Values** |
| blue | 0 |
| dual\_sim | 1 |
| four\_g | 1 |
| three\_g | 1 |
| touch\_screen | 1 |
| wifi | 1 |
| price\_range | 1 |

Above data represents the mode data of each categorical variable which by seeing the data we can conclude some of the specifications that are in the mobile phones installed most out of 2000 mobile phone samples

* We can see that most of the phones don’t have Bluetooth.
* But the remaining dual sim, four g, three g, touch screen, Wi-Fi has been installed in most of the phones.
* But if you observe the data of price range the price range is divided into three values (0,1,2,3) here ‘0’ represents very low price,’1’ = low price, ‘2’= medium price, ‘3’ =high price
* By the above mode values of the price range, the low-price mobile phones are most frequently consisting in the dataset.

The summary statistics provided give a good overview of the numerical variables in our dataset. Some of the statistics and make some observations of dataset are included below:

1. Battery Power: The range of battery power is from 501 to 1998, with a mean value of approximately 1238.5. This suggests a wide variation in battery power across the mobile phones in the dataset.

2. RAM: The RAM values range from 256 to 3998, with a mean value of approximately 2124. RAM is an essential factor affecting a smartphone's performance, and this wide range indicates significant variability.

3. Internal Memory: The internal memory (int\_memory) ranges from 2 to 64, with a mean of approximately 32.05. This indicates variation in internal memory capacity.

4. Front Camera (fc): The front camera ranges from 0 to 19, with a mean value of approximately 4.309. This variable represents the megapixels of the front camera.

5. Price Range: The target variable, "price\_range," has values ranging from 0 to 3, indicating different price ranges or categories.

6. 4G and Wi-Fi Support (four\_g, wifi): These binary variables have means around 0.52, indicating that about half of the phones support 4G and Wi-Fi.

These summary statistics provide a broad view of the dataset. To gain more insights, you can explore relationships between variables through visualizations, hypothesis testing, and further analysis. If you have specific questions or hypotheses you'd like to test, please let me know, and I can assist you further.

**5. Hypothesis Testing**

**Hypothesis Testing for Mean Battery Power**

Objective: To test whether the mean battery power of mobile phones in the dataset is equal to 1200 mAh.

1. Null Hypothesis (H0):

* H0: The mean battery power is equal to 1200 mAh.
* The null hypothesis represents the assumption that the mean battery power of mobile phones in the dataset is 1200 mAh.

2. Alternative Hypothesis (Ha):

* + Ha: The mean battery power is not equal to 1200 mAh.
  + The alternative hypothesis is a two-tailed statement indicating that the mean battery power differs from 1200 mAh in either direction.

3. Significance Level (α):

* + α = 0.05 (5%)
  + The significance level represents the probability of making a Type I error, which is the chance of incorrectly rejecting the null hypothesis when it is true. A common choice is 0.05, indicating a 5% risk of making a Type I error.

4. Data Analysis:

* + We calculated the mean battery power in the dataset, which was found to be 1238.5 mAh.
  + The sample data reflects the average battery power of the mobile phones under investigation.

5. Hypothesis Test:

* + We conducted Z-test to compare the mean battery power to the hypothesized value of 1200 mAh.

6. Z test

After performing the Z-test on the dataset, I found the following results:

* Null Hypothesis (H0): My initial assumption was that the mean battery power is equal to 1200.
* Alternative Hypothesis (Ha): However, based on the data and the Z-test, it appears that the mean battery power is not equal to 1200.

The p-value obtained from the test was approximately 8.848183e-05, which is significantly smaller than the chosen significance level (α = 0.05). This low p-value led me to reject the null hypothesis.

The test statistic (Z-score) was approximately 3.920183, indicating a substantial difference from the assumed mean value.

9. Conclusion

* + Based on the analysis, there is sufficient evidence to conclude that the mean battery power of mobile phones in the dataset is not equal to 1200 mAh.
  + The sample data suggests that the mean battery power is significantly different from the hypothesized value.
  + This result may have practical implications for mobile phone manufacturers and consumers, as it suggests that the actual mean battery power may deviate from the expected value of 1200 mAh.

**Hypothesis Testing Report for Mean Internal Memory.**

Is there enough evidence to conclude that the population's mean internal memory significantly differs from the following mean sample.

Objective: To test whether the mean internal memory of mobile phones in the dataset is equal to the sample mean of internal memory 35gb of mobile phones.

1. Null Hypothesis (H0):

* H0: The mean internal memory is equal to 35gb.
* The null hypothesis represents the assumption that the mean internal memory of mobile phones in the dataset is 35gb.

2. Alternative Hypothesis (Ha):

* + Ha: The mean Internal memory is not equal to 35gb
  + The alternative hypothesis is a two-tailed statement indicating that the mean internal memory differs from 35gb in either direction.

3. Significance Level (α):

* + α = 0.05 (5%)
  + The significance level represents the probability of making a Type I error, which is the chance of incorrectly rejecting the null hypothesis when it is true. A common choice is 0.05, indicating a 5% risk of making a Type I error.

4. Data Analysis:

* + We calculated the mean Internal memory in the dataset, which was found to be 32gb.
  + The sample data reflects the average Internal memory of the mobile phones under investigation.

5. Hypothesis Test:

* + We conducted a two-tailed t-test to compare the mean internal memory to the hypothesized value of 35gb.
  + The Z-test is an appropriate statistical test for comparing means when the population mean is different compared to sample mean.

6. Test Statistic (Z-statistic)

* + The calculated Z-statistic value is 1.6223.
  + The Z-statistic quantifies how many standard errors the sample mean is from the hypothesized population mean (35gb).

7. P-value

* + The p-value associated with the test statistic is 0.10524.
  + The p-value is the probability of obtaining a sample mean as extreme as, or more extreme than, the one observed in the sample if the null hypothesis is true.

8. Decision

* + Since the p-value 0.10524 is less than the significance level (α = 0.05), we reject the null hypothesis.
  + This decision indicates that there is strong evidence that the mean Internal memory is not equal to 35gb.

9. Conclusion

* + Based on the analysis, there is sufficient evidence to conclude that the Internal Memory p of mobile phones in the dataset is not equal to 35gb.

**Hypothesis Testing Report for RAM Comparison**

We performed a hypothesis test to investigate whether phones with 4G support have a significantly different average RAM compared to phones without 4G support.

Objective: To test whether there is a significant difference in the mean RAM between mobile phones with 4G support and those without 4G support.

1. Null Hypothesis (H0):

* H0: The mean RAM of mobile phones with 4G support is equal to the mean RAM of mobile phones without 4G support.
* The null hypothesis assumes no difference in the mean RAM between the two groups.

2. Alternative Hypothesis (Ha):

* Ha: The mean RAM of mobile phones with 4G support is not equal to the mean RAM of mobile phones without 4G support.
* The alternative hypothesis is two-tailed indicating that there is a difference in mean RAM between the two groups.

3. Significance Level (α):

* α = 0.05 (5%)
* The significance level represents the probability of making a Type I error and it is set at 0.05 indicating a 5% risk of making a Type I error.

4. Data Analysis

We separated the data into two groups: mobile phones with 4G support and mobile phones without 4G support.

5. Hypothesis Test

We conducted a two-sample t-test to compare the mean RAM between the two groups.

6. Test Statistic (t-statistic):

* + The calculated t-statistic is approximately 0.32713.
  + The t-statistic measures the difference in means relative to the variability within the groups.

7. P-value:

* + The p-value associated with the test statistic is approximately 0.7436.
  + The p-value is the probability of obtaining a sample result as extreme as or more extreme than the one observed if the null hypothesis is true.

8.. Decision:

* + Since the p-value (0.7436) is greater than the significance level (α = 0.05) we fail to reject the null hypothesis.
  + This decision suggests that there is no strong evidence to conclude that the mean RAM of mobile phones with 4G support is different from those without 4G support.\

11. Conclusion

* Based on the analysis we do not have sufficient evidence to claim a significant difference in the mean RAM between mobile phones with and without 4G support.
* This result may be of interest to mobile phone manufacturers and consumers as it indicates that the presence of 4G support does not significantly impact the mean RAM of mobile phones.

**6. F-Test**

**One-way anova**

One-way Analysis of Variance (ANOVA) is a statistical method used to assess whether there are any statistically significant differences between the means of three or more independent (unrelated) groups. In this analysis, the dependent variable is “ram,” and the independent variable is “price\_range,” which represents different price categories for mobile devices.

**ANOVA Results**

The ANOVA table summarizes the sources of variability in the data and helps determine whether the means of the groups are significantly different.

**1. ANOVA Table:**

Call:

aov(formula = ram ~ price\_range, data = data)

Terms:

price\_range Residuals

Sum of Squares 1978061205 374049364

Deg. of Freedom 1 1998

Residual standard error: 432.6799

Estimated effects may be unbalanced.

**2. Sum of Squares:**

The sum of squares represents the total variability in the dependent variable “ram” that is partitioned into two components: variability between the groups (price ranges) and variability within the groups (residuals).

Sum of Squares for “price\_range” is 1,978,061,205, and for “Residuals” is 374,049,364.

3. Degrees of Freedom

Degrees of Freedom (DF) represent the number of values in the final calculation of a statistic that are free to vary. For “price\_range,” DF is 1, and for “Residuals,” it is 1998.

4. Residual Standard Error:

The residual standard error is a measure of the variability of the observed response values around the mean. In this case, it is approximately 432.6799.

5. Interpretation:

According to the p-value that we obtained 2e-16 which is significantly lower than 0.05 significance level.

The small p-value associated with the “price\_range” term in the ANOVA table indicates that there is a statistically significant difference in the means of “ram” across different price ranges.

The residual standard error provides an estimate of the variability of individual data points around the fitted regression line.

The results of the one-way ANOVA suggest that there is a statistically significant difference in the mean “ram” values across different price ranges. This information is valuable for understanding how the RAM specifications of mobile devices may vary based on their price categories. Further post-hoc tests or pairwise comparisons could be conducted to identify which specific price ranges differ significantly from each other in terms of RAM.

**Two-way ANOVA**

The results of the two-way ANOVA are presented below:

Call: The two-way ANOVA was conducted using the formula `ram ~ dual\_sim \* four\_g` on the dataset.

Objective: To find in the analysis is there any significant impact of “dual\_sim” and “four\_g” on RAM specifications.

- Terms: The analysis considers the factors “dual\_sim,” “four\_g,” the interaction term “dual\_sim: four\_g,” and the residuals.

- Sum of Squares:

- For “dual\_sim” is 3,967,800.

- For “four\_g” is 121,346.

- For “dual\_sim:four\_g” interaction is 1,595,110.

- For residuals is 2,346,426,313.

- Degrees of Freedom (DF):

- For “dual\_sim” is 1.

- For “four\_g” is 1.

- For “dual\_sim:four\_g” interaction is 1.

- For residuals is 1996.

- Residual Standard Error: The standard deviation of the residuals is 1084.234.

- Estimated Effects May Be Unbalanced: This statement suggests that the design of the study may not have a balanced distribution across all levels of the factors.

Based on the P value that we got combined effect of dual sim and four got 0.2442 which higher compared to p-value both are significantly higher than 0.05.

Where if we take individually of dual sim and four g p values which both significantly higher

than the 0.05.

Interpretation:

1. Sum of Squares (SS): Represents the variability in the dependent variable (“ram”) attributed to the main effects of “dual\_sim” and “four\_g,” the interaction effect between them, and the residuals.

2. Degrees of Freedom (DF): Indicates the number of independent values or quantities that can be assigned without violating any constraints.

3. Residual Standard Error: Reflects the average amount that the observed values deviate from the fitted values.

4. Estimated Effects May Be Unbalanced: Implies that there might be an uneven distribution of observations among different levels of the factors.

The significant Sum of Squares for “dual\_sim,” “four\_g,” and their interaction, along with the large Residual Standard Error, indicates that there are no substantial effects of both “dual\_sim” and “four\_g” factors on the “ram” variable.

The P-value of combined is 0.2442 which is higher than significance value 0.05 based on this we can conclude that there is no significance impact of dual sim and four\_g on RAM specification. The interaction effect suggests that there is no combination of these factors may also influence the response variable. Overall, the two-way ANOVA results provide insights into the combined there is no significant impact of “dual\_sim” and “four\_g” on RAM specifications.

**7. Categorical data analysis**

In this categorical data analysis, we go into various features of a mobile phone dataset to gain insights into the distribution of key attributes. The dataset encompasses variables such as Bluetooth availability, dual SIM support, 4G capability, CPU core count, camera specifications, 3G support, touch screen presence, Wi-Fi availability, and price range categories. Through a careful examination of these categorical variables, we aim to unravel patterns, frequencies, and disparities in the dataset, shedding light on the prevalence of specific features among mobile phones.

**Analysis**

1. Bluetooth (blue): The analysis indicates that nearly half of the phones (approximately 990) possess Bluetooth functionality, while the rest do not.

2. Dual SIM (dual\_sim): Around 1019 phones are equipped with dual SIM support, revealing a balanced distribution, with 981 phones lacking this feature.

3. 4G Capability (four\_g): The dataset demonstrates a relatively even split regarding 4G support, with 954 phones without 4G and 1046 phones supporting this technology.

4. CPU Core Count (n\_cores): Further exploration of the distribution of CPU core counts will provide insights into the prevalent configurations in the dataset.

5. Primary Camera Megapixels (pc): Detailed examination of the distribution of primary camera megapixels will offer valuable information about the diversity of camera specifications.

6. 3G Support (three\_g): A substantial majority, comprising 1523 phones, lacks 3G support, while 477 phones include this feature.

7. Touch Screen (touch\_screen): Approximately 1100 phones feature a touch screen, while 900 do not, highlighting a significant prevalence of touch screen technology.

8. Wi-Fi Availability (wifi): The dataset portrays a noteworthy presence of Wi-Fi, with 1490 phones supporting this feature, while 510 phones lack Wi-Fi.

This comprehensive categorical data analysis lays the foundation for a careful understanding of the dataset, guiding subsequent exploratory analyses and data-driven decision-making processes.

**Chi-squared test**

We conducted a chi-squared test to examine the association between two categorical variables namely 'four\_g' and 'wifi' in the dataset. Initially, We created a contingency table to summarize the distribution of observations across the categories of these two variables.

After constructing the contingency table, we proceeded to perform the chi-squared test. The test yielded a test statistic (X-squared) of 0.55238 with 1 degree of freedom. Additionally, I obtained a p-value of 0.4573.

Upon interpretation the p-value of 0.4573 exceeds the commonly used significance level of 0.05. Consequently, we fail to reject the null hypothesis. This implies that based on the chi-squared test there is insufficient evidence to suggest a significant association between the 'four\_g' and 'wifi' variables. The findings indicate that any observed differences in the distribution of categories between these variables may be attributed to random chance rather than a meaningful relationship.

**8. Linear Regression**

Linear regression analysis is a statistical method used to model the relationship between a dependent variable and one or more independent variables. In this case, a simple linear regression model was applied to explore the connection between the “battery\_power” (dependent variable) and “talk\_time” (independent variable) in a dataset containing various mobile device features. The aim is to understand how changes in talk time might impact battery power.

**Analysis**

The linear regression model generated for the data is expressed as:

battery\_power = 1192.020 + 4.223\*talk\_time

1. Coefficients Interpretation:

* Intercept (1192.020): The intercept represents the estimated battery power when the talk time is zero. However, in the context of this model, this value may not have a practical interpretation as talk time is unlikely to be zero for mobile devices.
* talk\_time (4.223): The coefficient for talk time indicates that, on average, for each additional unit increase in talk time, the battery power is estimated to increase by 4.223 units.

2. Statistical Significance

The p-value associated with the coefficient for talk time is 0.0188, which is less than the conventional significance level of 0.05. This suggests that there is a statistically significant relationship between talk time and battery power.

3. Model Fit

The R square value, representing the proportion of the variance in battery power explained by talk time, is 0.002757. This is a very small value, indicating that the linear relationship between talk time and battery power does not explain much of the variability in battery power. Other factors not considered in this model might have a stronger influence.

While the statistical analysis indicates a significant relationship between talk time and battery power, the small R square value suggests that talk time alone may not be a strong predictor of battery power. To enhance the model's predictive power, it's recommended to explore additional relevant variables that could contribute to a more comprehensive understanding of the factors influencing battery power in mobile devices.

**9. Conclusion**

The analysis of the mobile phone data set provided a thorough exploration of various features and their relationships. The absence of outliers and missing values ensured the integrity of the dataset. Summary statistics highlighted the wide variability in key attributes like battery power, RAM, and internal memory.

Hypothesis testing revealed that the presence of 4G support did not significantly impact the mean RAM of mobile phones. Additionally, the battery power of mobile phones in the dataset was found to be significantly different from the assumed mean of 1200 mAh.

ANOVA and Two-Way ANOVA demonstrated significant differences in RAM across different price ranges, emphasizing the influence of pricing on device specifications. Categorical data analysis unveiled insights into the prevalence of features such as Bluetooth, dual SIM, and 4G.

Linear regression indicated a statistically significant relationship between talk time and battery power, though the small R square value suggested limited predictability and this analysis offers valuable insights for both manufacturers and consumers in understanding the dynamics of mobile phone features and pricing.

**8. Annexure**

**A blue rectangle with black lines

Description automatically generatedA graph of a power line

Description automatically generated 1)** **The graphical representation of the box plot and histogram for the following dataset**

**Figure -1**: ## battery power summary

## Min.: 501.0, ## 1st Qu.: 851.8, ## Median :1226.0 , ## Mean :1238.5, ## 3rd Qu.:1615.2 ## Max. :1998.0

**A graph of a speed

Description automatically generatedA chart with a green rectangle and black lines

Description automatically generated2) The graphical representation of median and mean of #Clock Speed**

**Figure -2**: ## Clock Speed summary

## Min.: 0.50000, ## 1st Qu.: 0.7000, ## Median :1.5000.0 , ## Mean :1.5225,

## 3rd Qu.:2.2000 ## Max. :3.000

1. A diagram of a box plot

   Description automatically generated**The graphical representation of median and mean of #Front Camera**

**Figure -3**: ## Front Camera summary A graph with a red line

Description automatically generated

## Min.: 0, ## 1st Qu.: 1.000, ## Median :3.000, ## Mean :4.3095, ## 3rd Qu.7.000 ## Max. :19.000

1. **A diagram of a memory summary

   Description automatically generatedA graph of a graph showing the internal memory

   Description automatically generatedThe** **graphical representation of median and mean of #internal Memory**.

**Figure -4**: ## Internal Memory summary

## Min.: 2.000, ## 1st Qu.: 16.000, ## Median :32.000 , ## Mean :32.0465,

## 3rd Qu.: 48.000 ## Max. :64.000

1. **A graph with a red line

   Description automatically generatedA blue box with black lines

   Description automatically generatedThe graphical representation of median and mean of #M\_depth.**

**Figure -5**: ## Main Camera Depth summary

## Min.: 0.1000, ## 1st Qu.: 0.2000, ## Median :0.5000, ## Mean :0.50175, ## 3rd Qu.:0.8000 ## Max. :1.0000

1. **A yellow box with black text

   Description automatically generatedA graph of a graph with a red line

   Description automatically generatedThe graphical representation of median and mean of #N\_Cores**

**Figure - 6**: ## N\_cores summary

## Min.: 1.000, ## 1st Qu.: 3.000, ## Median :4.000, ## Mean :4.5205, ## 3rd Qu.:7.000 ## Max. :8.000

1. **The graphical representation of median and mean of #Mobile\_Weight**

A graph of a graph with a red line

Description automatically generatedA blue and black line graph

Description automatically generated

**Figure -7**: ## Mobile \_Weight summary

## Min.: 80.000, ## 1st Qu.: 109.000, ## Median :141.000 , ## Mean :140.249, ## 3rd Qu.:170.000 ## Max. :200.000

1. **The graphical representation of median and mean of #Primary Camera**

A graph with a red line

Description automatically generatedA diagram of a box plot

Description automatically generated

**Figure -8**: ## Primary Camera summary

## Min.: 0.000, ## 1st Qu.: 5.000, ## Median :10.000, ## Mean :9.9165,

## 3rd Qu.:15.000 ## Max. :20.000

1. **The graphical representation of median and mean of #Px\_height**

A graph of a graph with a red line

Description automatically generatedA yellow and black box plot

Description automatically generated

**Figure -9**: ## Pixel Height summary

## Min.: 0.000, ## 1st Qu.: 282.750, ## Median :564.000 , ## Mean :645.108,

## 3rd Qu.:947.250 ## Max. :1960.00

1. **The graphical representation of median and mean of #Px\_Width**

A graph with a red line

Description automatically generatedA diagram of a box plot

Description automatically generated

**Figure -10**: ## Pixel Width summary

## Min.: 500.00, ## 1st Qu.: 874.750, ## Median :1247.000, ## Mean :1251.515,

## 3rd Qu.:1633.000 ## Max. :1998.000

1. **The graphical representation of median and mean of #Ram**

A graph with red line and numbers

Description automatically generatedA green and black box plot

Description automatically generated

**Figure -11**: ## RAM summary

## Min.: 256.000, ## 1st Qu.: 1207.500, ## Median :2146.500, ## Mean :2124.213, ## 3rd Qu.:3064.500 ## Max. :3998.00

1. **The graphical representation of median and mean of #Screen Height**

A graph with a red line

Description automatically generatedA blue and white box diagram

Description automatically generated

**Figure -12**: ## Screen Height summary

## Min.: 5.000, ## 1st Qu.: 9.000, ## Median :12.000, ## Mean :12.3605,

## 3rd Qu.:16.000 ## Max. :19.000

1. **The graphical representation of median and mean of #Screen Width**

A graph with a red line

Description automatically generatedA diagram of a box plot

Description automatically generated

**Figure -13**: ## Screen Width summary

## Min.: 0.00, ## 1st Qu.: 2.000, ## Median :5.000, ## Mean :5.767, ## 3rd Qu.:9.000 ## Max. :18.000

1. **The graphical representation of median and mean of #Talk Time**

A graph with a red line

Description automatically generatedA yellow and black box plot

Description automatically generated

**Figure -14**: ## Talk Time summary

## Min.: 2.000, ## 1st Qu.:6.000, ## Median :11.000 , ## Mean :11.011,

## 3rd Qu.:16.000 ## Max. :20.000

1. **Normal distribution Graph Visualization:**
2. **Normal Distribution of battery power, clock speed, Internal memory and mobile Weight**.

Median :1226.0 , ## Mean :1238.5A diagram of a normal distribution of clock speed

Description automatically generatedA diagram of a normal distribution of battery power

Description automatically generated , ##SD:439.418206

**Figure -15**: ## Normal Distribution Battery Power

## Median :1.5000.0 , ## Mean :1.5225, ## SD:0.8160042

**Figure -16**: ## Clock Speed summary

## A diagram of a normal distribution of internal memory Description automatically generated## Median :32.000, ## Mean :32.0465 ,##SD:18.1457

**Figure -17**: ## Internal Memory Normal Distribution

A diagram of a normal distribution of mobile weight

Description automatically generated

**Figure -18**: ## Mobile Weight Normal Distribution

**R-Script**

# Loading the required library

library(dplyr)

# Loading the dataset

data <- read.csv ("C:/USERS /USER-1/Mobile Price.csv")

head(data)

# Getting the structure of the dataset

str(data)

# Checking for missing values

missing\_values <- sum(is.na(data))

print(paste("Missing values:", missing\_values))

#OUTLIERS

# Loading the ggplot2 library

library(ggplot2)

# Creating a box plot for RAM

ggplot(data, aes(x = 1, y = ram)) +

geom\_boxplot(fill = "lightblue") +

labs(title = "Box Plot of RAM")

# Creating a box plot for Battery Power

ggplot(data, aes(x = 1, y = battery\_power)) +

geom\_boxplot(fill = "lightgreen") +

labs(title = "Box Plot of Battery Power")

# Creating a box plot for Internal Memory

ggplot(data, aes(x = 1, y = int\_memory)) +

geom\_boxplot(fill = "lightcoral") +

labs(title = "Box Plot of Internal Memory")

summary(data)

#Hypothesis Testing

# Calculate the correlation matrix

correlation\_matrix <- cor(data)

correlation\_matrix

# Performing a t-test

t\_test\_result <- t.test(with\_4g, without\_4g)

# Print the result

print(t\_test\_result)

# Null Hypothesis: The mean battery power is equal to 1200

# Alternative Hypothesis: The mean battery power is not equal to 1200

alpha <- 0.05

test\_result <- t.test(data$battery\_power, mu = 1200)

# P-value

p\_value <- test\_result$p.value

p\_value

# Test Statistic

test\_statistic <- test\_result$statistic

test\_statistic

# Checking if p-value is less than alpha

if (p\_value < alpha) {

cat("Reject Null Hypothesis: Mean battery power is not equal to 1200.\n")

} else {

cat("Fail to reject Null Hypothesis: Mean battery power is equal to 1200.\n")

}

# Separate data into two groups: with 4G and without 4G

with\_4g <- data[data$four\_g == 1, "ram"]

without\_4g <- data[data$four\_g == 0, "ram"]

# Performing F-test

f\_test\_result <- var.test(with\_4g, without\_4g)

# Printing the result

print(f\_test\_result)

# Null Hypothesis: The mean RAM of mobile phones with 4G support is equal to the mean RAM of mobile phones without 4G support.

# Alternative Hypothesis: The mean RAM of mobile phones with 4G support is not equal to the mean RAM of mobile phones without 4G support.

alpha <- 0.05 # Significance level

# Calculating the sample means

mean\_with\_4g <- mean(with\_4g)

mean\_without\_4g <- mean(without\_4g)

# Calculating the sample standard deviations

sd\_with\_4g <- sd(with\_4g)

sd\_without\_4g <- sd(without\_4g)

# Sample sizes

n\_with\_4g <- length(with\_4g)

n\_without\_4g <- length(without\_4g)

pooled\_sd <- sqrt(((n\_with\_4g - 1) \* sd\_with\_4g^2 + (n\_without\_4g - 1) \* sd\_without\_4g^2) / (n\_with\_4g + n\_without\_4g - 2))

# Calculating the standard error

standard\_error <- pooled\_sd \* sqrt(1/n\_with\_4g + 1/n\_without\_4g)

# Calculating the test statistic (z-score)

z\_score <- (mean\_with\_4g - mean\_without\_4g) / standard\_error

# Setting the significance level (alpha)

alpha <- 0.05

# Calculating the critical value for a two-tailed test

z\_critical <- qnorm(1 - alpha/2)

# Calculating the p-value for the two-tailed test

p\_value <- 2 \* (1 - pnorm(abs(z\_score)))

# Performing the hypothesis test

if (abs(z\_score) > z\_critical) {

cat("Reject Null Hypothesis: The means are not equal.\n")

} else {

cat("Fail to reject Null Hypothesis: The means are equal.\n")

}

# Print the test statistic and p-value

cat("Z-Score:", z\_score, "\n")

cat("P-Value:", p\_value, "\n")

# Defining the null hypothesis (H0) mean

null\_mean <- 1200 # Change this to the specific value you want to test against

# Significance level

alpha <- 0.05

# Sample data

sample\_data <- data$battery\_power

# Calculate the sample mean and standard deviation

sample\_mean <- mean(sample\_data)

sample\_sd <- sd(sample\_data)

# Sample size

n <- length(sample\_data)

# Calculate the test statistic (Z-score)

test\_statistic <- (sample\_mean - null\_mean) / (sample\_sd / sqrt(n))

# Calculate the p-value

p\_value <- 2 \* (1 - pnorm(abs(test\_statistic)))

# Check if p-value is less than alpha

if (p\_value < alpha) {

cat("Reject Null Hypothesis: Mean battery power is not equal to", null\_mean, "\n")

} else {

cat("Fail to reject Null Hypothesis: Mean battery power is equal to", null\_mean, "\n")

}

# Display the test results

cat("P-value:", p\_value, "\n")

cat("Test Statistic (Z-score):", test\_statistic, "\n")

#second part

# Performing one-way ANOVA

anova\_result <- aov(ram ~ price\_range, data = data)

# Printing the ANOVA table

print(anova\_result)

# Performing two-way ANOVA

twoway\_anova\_result <- aov(ram ~ dual\_sim \* four\_g, data = data)

print(twoway\_anova\_result)

# Creating a table for categorical data analysis

table\_result <- table(data$dual\_sim)

# Printing the table

print(table\_result)

# Categorical Data Analysis

# 1. 'blue' variable

table\_blue <- table(data$blue)

print("Categorical Data Analysis for 'blue' variable:")

print(table\_blue)

# 2. 'dual\_sim' variable

table\_dual\_sim <- table(data$dual\_sim)

print("Categorical Data Analysis for 'dual\_sim' variable:")

print(table\_dual\_sim)

# 3. 'four\_g' variable

table\_four\_g <- table(data$four\_g)

print("Categorical Data Analysis for 'four\_g' variable:")

print(table\_four\_g)

# 4. 'n\_cores' variable

table\_n\_cores <- table(data$n\_cores)

print("Categorical Data Analysis for 'n\_cores' variable:")

print(table\_n\_cores)

# 5. 'pc' variable

table\_pc <- table(data$pc)

print("Categorical Data Analysis for 'pc' variable:")

print(table\_pc)

# 6. 'three\_g' variable

table\_three\_g <- table(data$three\_g)

print("Categorical Data Analysis for 'three\_g' variable:")

print(table\_three\_g)

# 7. 'touch\_screen' variable

table\_touch\_screen <- table(data$touch\_screen)

print("Categorical Data Analysis for 'touch\_screen' variable:")

print(table\_touch\_screen)

# 8. 'wifi' variable

table\_wifi <- table(data$wifi)

print("Categorical Data Analysis for 'wifi' variable:")

print(table\_wifi)

# 9. 'price\_range' variable

table\_price\_range <- table(data$price\_range)

print("Categorical Data Analysis for 'price\_range' variable:")

print(table\_price\_range)

# Creating a scatter plot to visualize the relationship

ggplot(data, aes(x = talk\_time, y = battery\_power)) +

geom\_point() +

labs(title = "Scatter Plot of Battery Power vs Talk Time",

x = "Talk Time",

y = "Battery Power")

# Perform simple linear regression

linear\_model\_talk\_time <- lm(battery\_power ~ talk\_time, data = data)

# Print the summary of the regression model

summary(linear\_model\_talk\_time)