Apple Store Reviews

Statistical Analysis

Presented by
Thomas Sangala

INTRODUCTION

I am a highly analytical and detail-driven individual with a strong grounding in data analysis, visualization, and problem-solving. Proficient in tools such as Python, SQL, Excel, Power BI, and Tableau, I have practical experience in building dashboards and conducting exploratory data analysis (EDA). My expertise lies in uncovering trends, extracting meaningful insights, and delivering data-backed recommendations. I am passionate about applying my technical and analytical abilities to solve real-world business problems and support organizational growth.

QUESTIONS TO SOLVE

Perform the Following Statistical Analysis:

1. Calculate the mean, median, and mode of the app ratings in the dataset. Which measure (mean, median, or mode) best represents the central tendency of the ratings?
2. Find the range and interquartile range (IQR) of the Purchase_Amount in the dataset. How do these values help in understanding the spread of the data?
3. Calculate the variance and standard deviation for the number of likes received on reviews. What does the standard deviation indicate about the spread of the data?
4. Determine the correlation between the likes and the rating given. Is there a positive, negative, or no correlation between these variables?
5. Plot the distribution of the app ratings. Is the distribution positively or negatively skewed? What does this indicate about user satisfaction?
6. Perform a hypothesis test to determine if the average rating for Instagram is significantly higher than the average rating for WhatsApp. Use a 95% confidence level.
7. Take random samples of ratings from the dataset and calculate their means. Create a sampling distribution and explain how this relates to the Central Limit Theorem.

ANALYSIS OF RATINGS

- **Mean : 2.869**

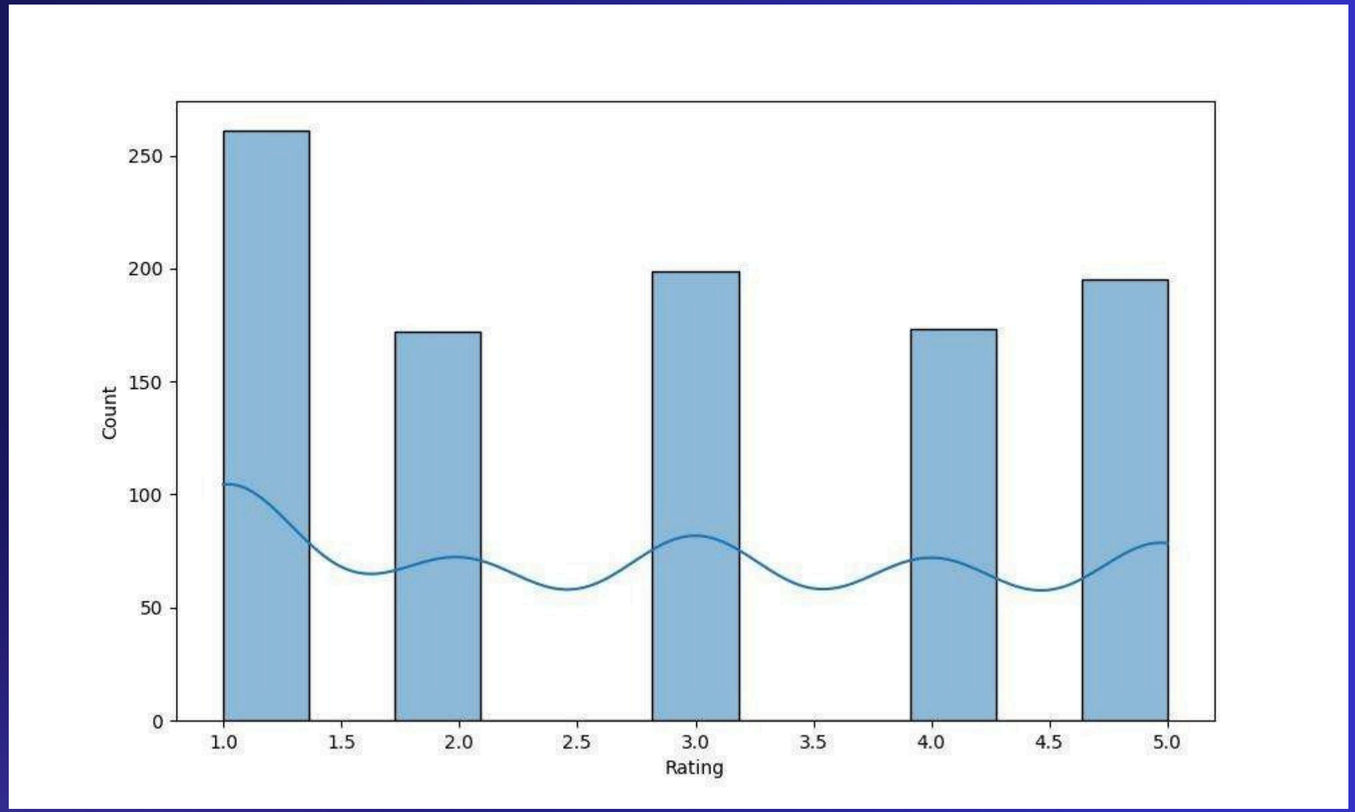
  The average rating is just under 3, suggesting that ratings generally skew towards the lower end of the scale. This is reflected in the graph, where the most prominent bar corresponds to a rating of 1, which heavily impacts and lowers the overall average.

- **Median : 3.0**

  The median rating is 3.0, indicating that 50% of the ratings fall below 3 and 50% are above. This matches the graph, where the ratings seem fairly balanced around 3, even though there is a noticeable peak at 1.

- **Mode : 1**

  The mode is 1, which is the most frequently occurring rating. This is reflected in the graph by the tallest bar at 1, showing that a significant number of users gave the lowest possible rating.



In the above distribution, Median is likely the best choice for central tendency, as it balances the skew and the multiple peaks.

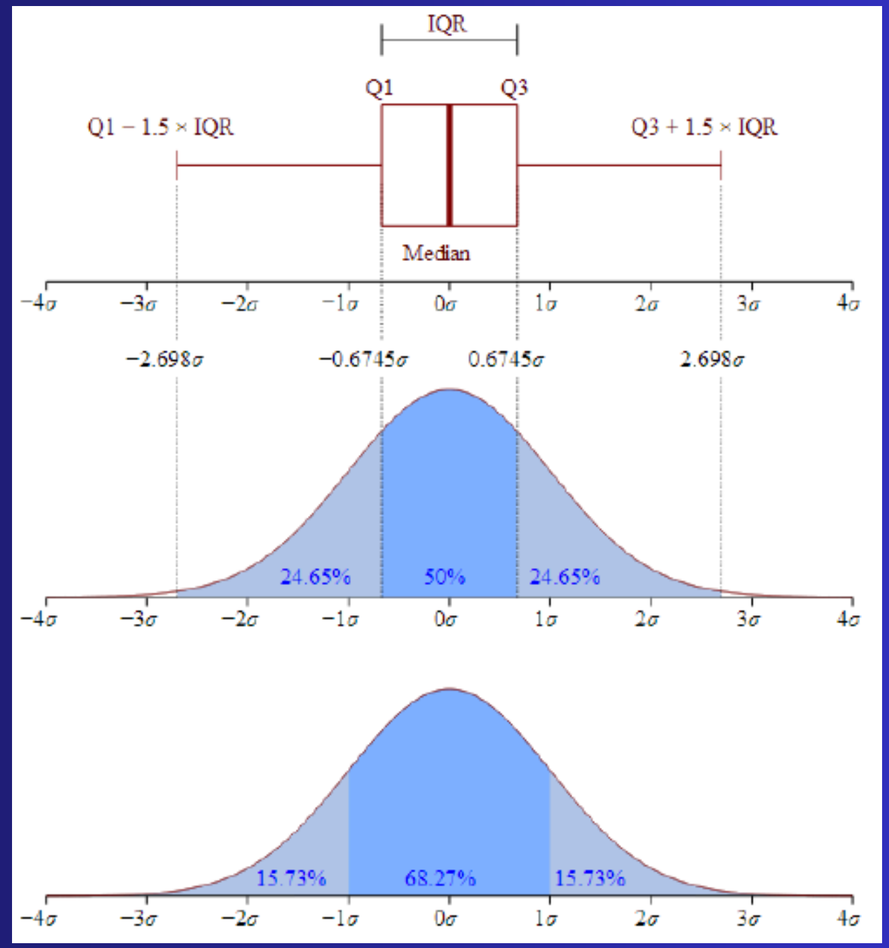
RANGE AND INTER QUARTILE RANGE (IQR)

Range :- 19.97

The range tells you the total span of values. A large range indicates a wide spread between the lowest and highest purchase amounts, while a small range indicates a more concentrated set of values.

Inter Quartile Range (IQR) : - 10.19

The IQR is useful because it focuses on the middle 50% of the data, ignoring outliers. It tells you how spread out the central portion of the data is. A small IQR indicates that the values are concentrated around the median, while a large IQR suggests a more dispersed dataset.



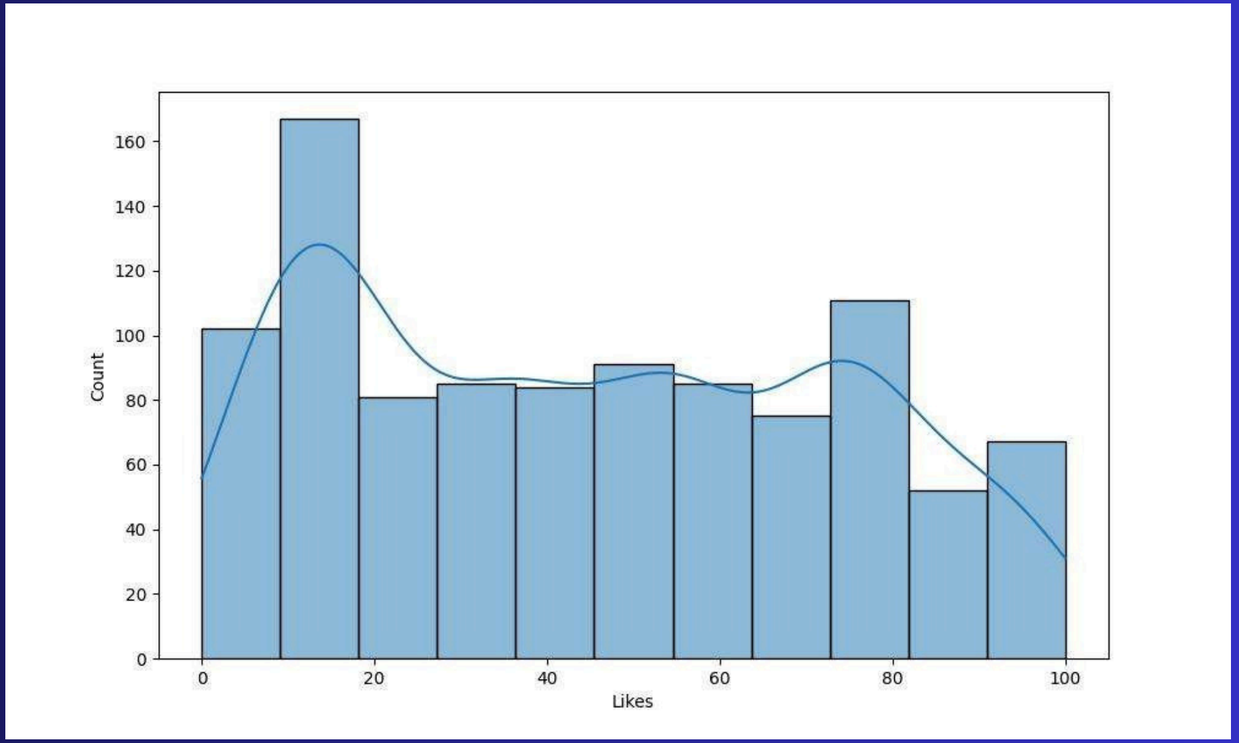
VARIANCE AND STANDARD DEVIATION ON LIKES

Variance : 822.85

This high variance value indicates that the number of likes varies significantly from the mean. It suggests a wide spread in the data, with some observations having a high number of likes and others a low number, which is reflected in the uneven peaks across the graph.

Standard Deviation : 28.69

The standard deviation of 28.69 means that, on average, the number of likes deviates from the mean by about 28.69. This large value supports the visual spread seen in the graph, where likes are distributed across a broad range, with some counts reaching higher values while others remain low.

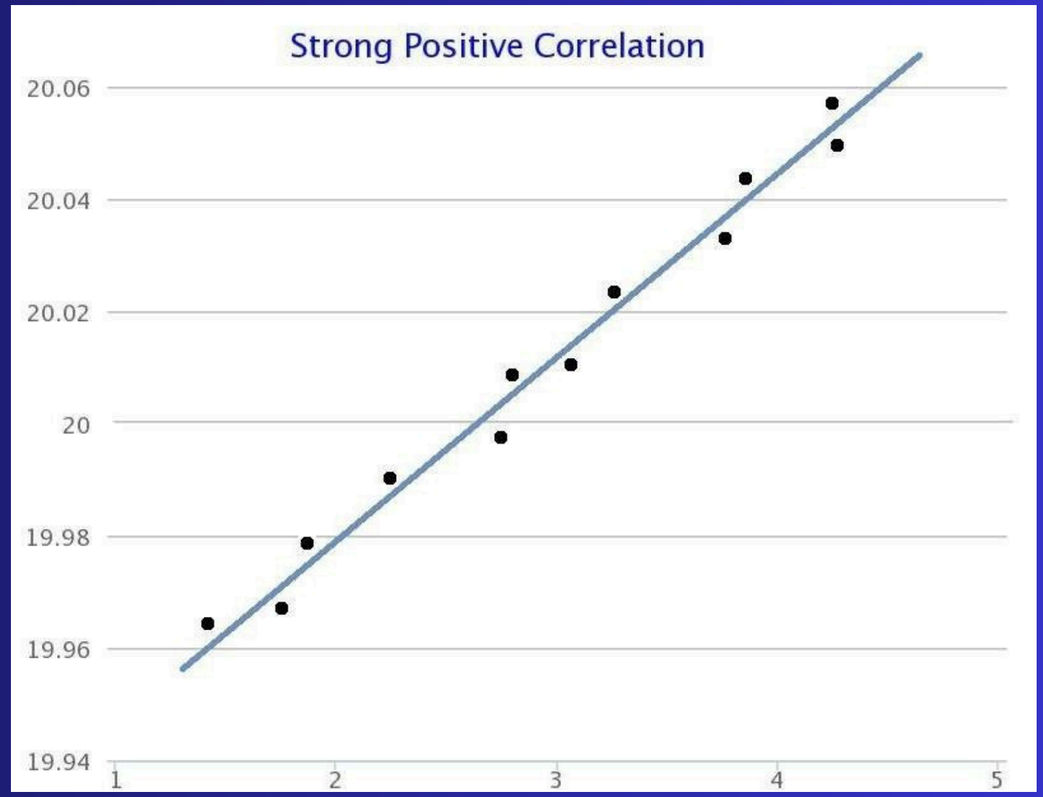


CORRELATION BETWEEN LIKES & RATINGS

Correlation Between Likes & Ratings : 0.8425

- A value above 0.8 generally represents a strong correlation, meaning there is a clear and predictable trend between the two variables. For instance, if the rating of a product goes up, the number of likes it receives also tends to rise, and vice versa.

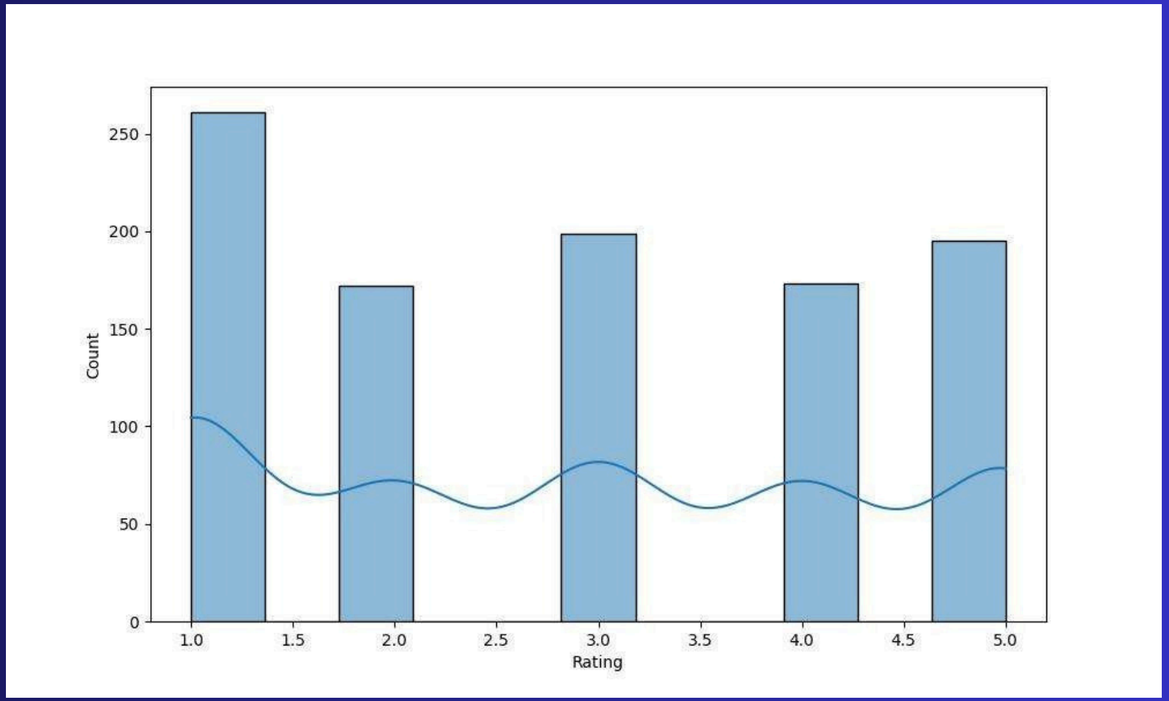
- This positive correlation could mean that products with higher ratings receive more likes, perhaps because users who are satisfied with the product are more likely to engage further by liking it. It also suggests that higher-rated products are likely perceived more favorably, which can lead to greater interaction, like and share behavior on the platform.



DISTRIBUTION OF APP RATINGS

The distribution is not clearly skewed in a traditional sense (positive or negative), but it suggests that there are two distinct groups of user ratings:

1. Positive feedback: Many users are highly satisfied (5.0 rating).
2. Negative feedback: A significant number of users are dissatisfied (1.0 rating).



User Satisfaction: The presence of peaks at both extremes of the scale suggests a polarized user base.

HYPOTHESIS TEST FOR INSTAGRAM VS WHATSAPP RATINGS

Hypothesis Testing Framework:

- Null Hypothesis (H_0): The mean rating of Instagram is less than or equal to the mean rating of WhatsApp.
- Alternative Hypothesis (H_1): The mean rating of Instagram is higher than the mean rating of WhatsApp.

- If the p-value is below 0.05, reject the null hypothesis, indicating Instagram's average rating is significantly higher than WhatsApp's. Otherwise, fail to reject it, suggesting no significant difference in their ratings.

T-STATISTIC: -0.7967
P-VALUE: 0.7868

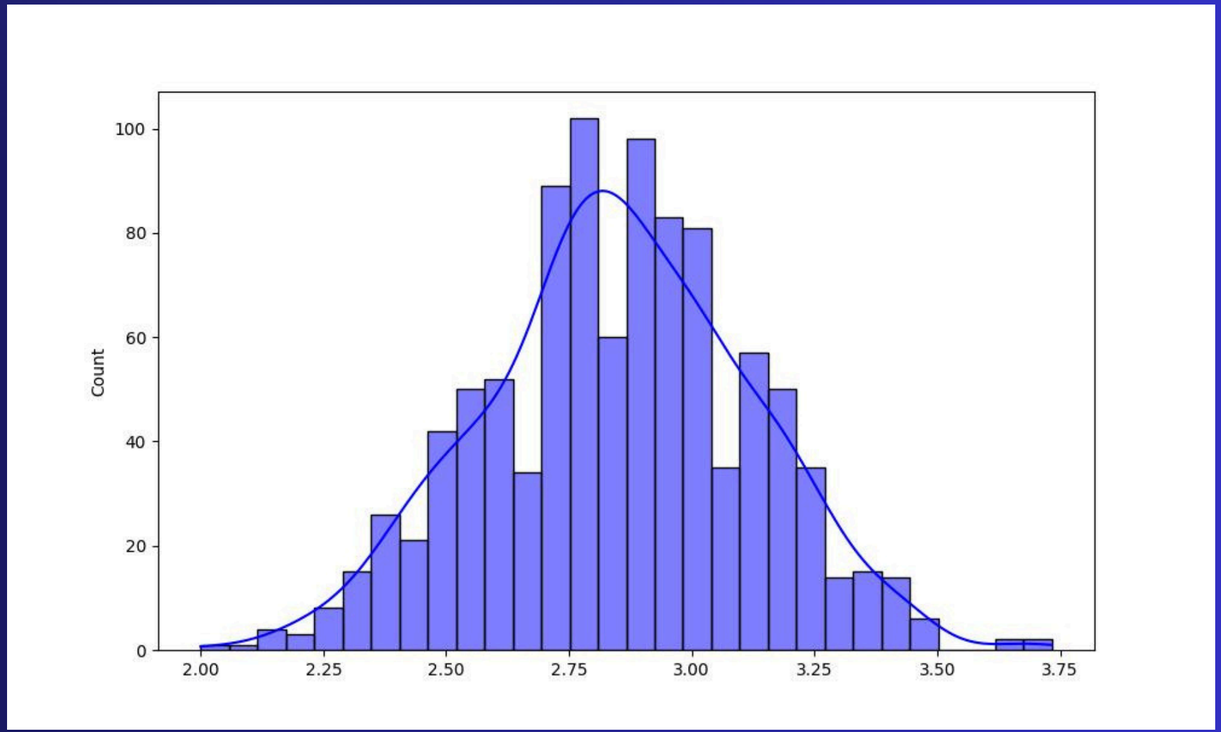
- If the p-value is below 0.05, reject the null hypothesis, indicating Instagram's average rating is significantly higher than WhatsApp's. Otherwise, fail to reject it, suggesting no significant difference in their ratings.

SAMPLING DISTRIBUTION AND CENTRAL LIMIT THEOREM (CLT)

- The histogram's shape suggests that the sample means are normally distributed.

  The distribution is centered around the population mean(2.869), and its spread is narrower than the original population due to the reduced variability of means.

  Regardless of the original distribution of the ratings, the sampling distribution of the mean will have a mean equal to the population mean.



CONCLUSION

- Mixed User Satisfaction: The median rating of 3 reflects an average level of user satisfaction, while the mode of 1 indicates a considerable number of users gave low ratings.
- Spending Patterns: The range and interquartile range (IQR) of purchase amounts show moderate variability in user spending behavior.
- Engagement Diversity: The high variance in the number of likes highlights differences in how users interact with app reviews.
- Correlation Findings: A strong positive correlation ($r = 0.8425$) suggests that reviews with higher ratings tend to receive more likes.
- Hypothesis Testing Outcome: Analysis showed no significant difference in the average ratings of Instagram and WhatsApp.
- Central Limit Theorem Application: The use of the Central Limit Theorem enabled accurate and reliable inferences, guiding data-driven strategies to improve app features and user experience.

Thank You