APPLE STORE REVIEWS

Statistical Analysis

Presented by Ethen D’Costa

AGENDA

- Introduction
- Problem Statement
- Questions to Solve
- Analysis of Ratings
- Range and Inter Quartile Range(IQR) of Purchase Amount
- Variance and Standard Deviation of Likes
- Correlation Between Likes & Ratings
- Distribution of App Ratings
- Hypothesis Test for Instagram vs WhatsApp Ratings
- Sampling Distribution and Central Limit Theorem (CLT)
- Conclusion

INTRODUCTION

I am a detail-oriented and analytical fresher with a strong foundation in data analysis, visualization, and problem-solving. I am proficient in tools like Python, SQL, Excel, Power BI, and Tableau, with hands-on experience in creating dashboards and performing exploratory data analysis (EDA). I excel at identifying trends, drawing insights, and presenting data-driven recommendations. I am eager to apply my technical and analytical skills to real-world business challenges and contribute to organizational success.

PROBLEM STATEMENT

As a data Analyst you will apply key statistical concepts to analyze a dataset of Apple Store product reviews. Your objective is to perform descriptive and inferential statistical analysis to draw meaningful insights from the data.

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

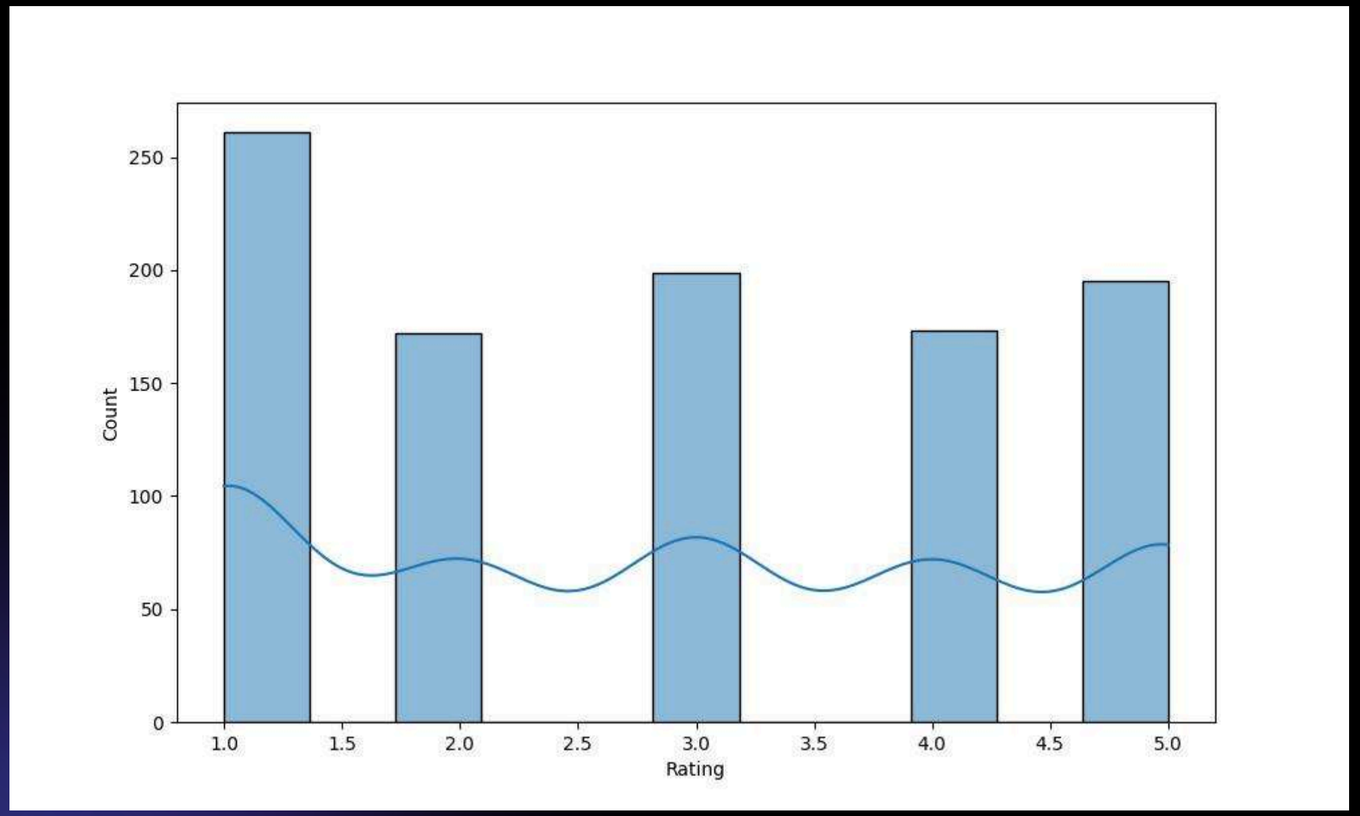
The mean rating is slightly below 3, indicating that, on average, the ratings tend to lean toward the lower end of the scale. This is evident in the graph, as the highest bar is for a rating of 1, which significantly pulls the average down.

✔ Median : 3.0

The median rating is 3.0, meaning that half of the ratings are below 3, and half are above. This aligns with the graph, where the counts of ratings appear to be more evenly distributed around 3, despite the peak at 1.

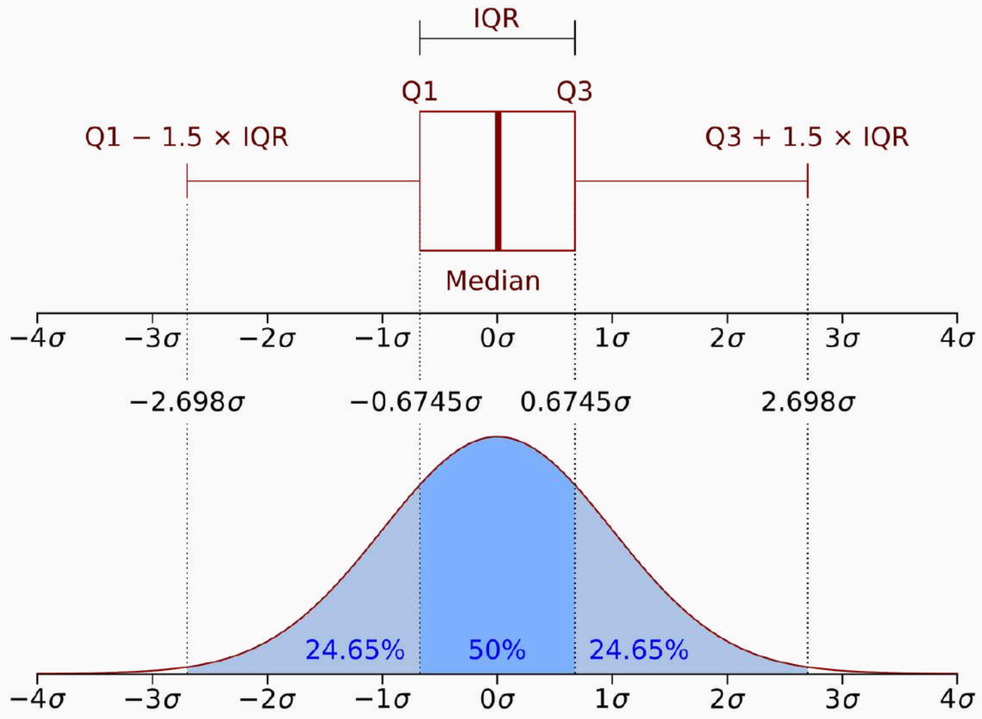
✔ Mode : 1

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.

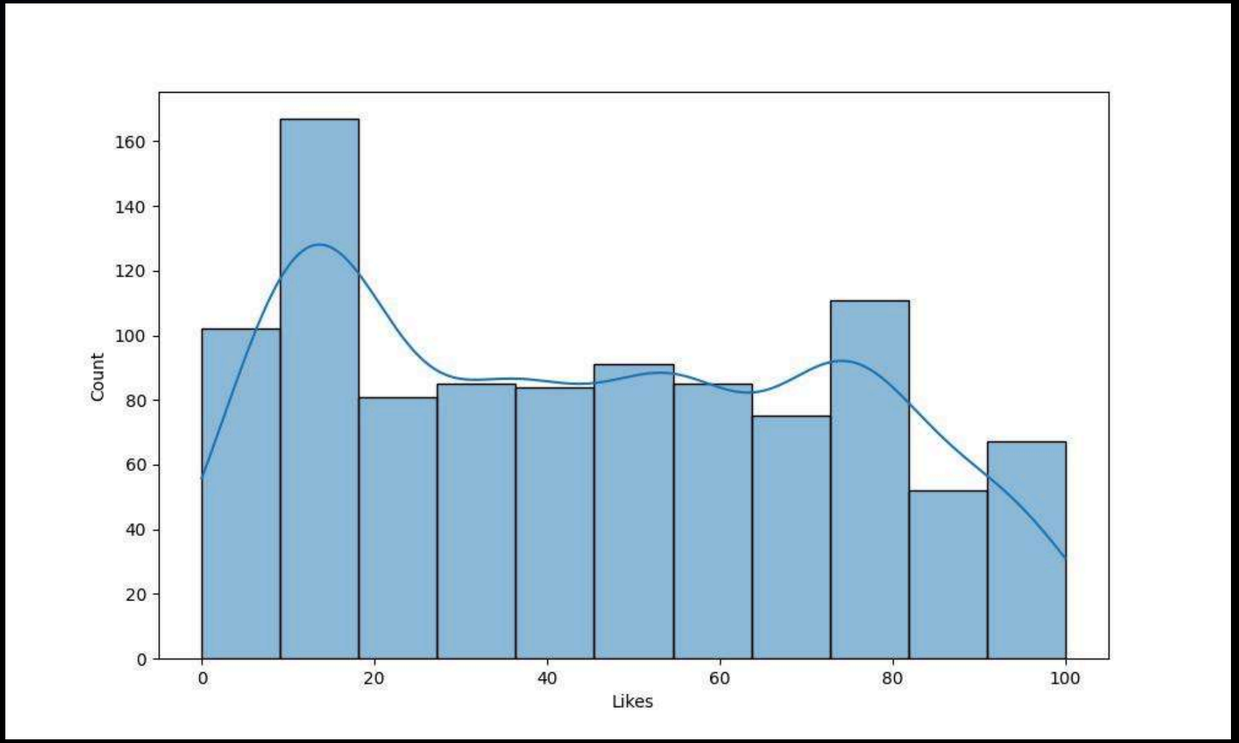
VARIANCE AND STANDARD DEVIATION ON LIKES

Variance : 822.85

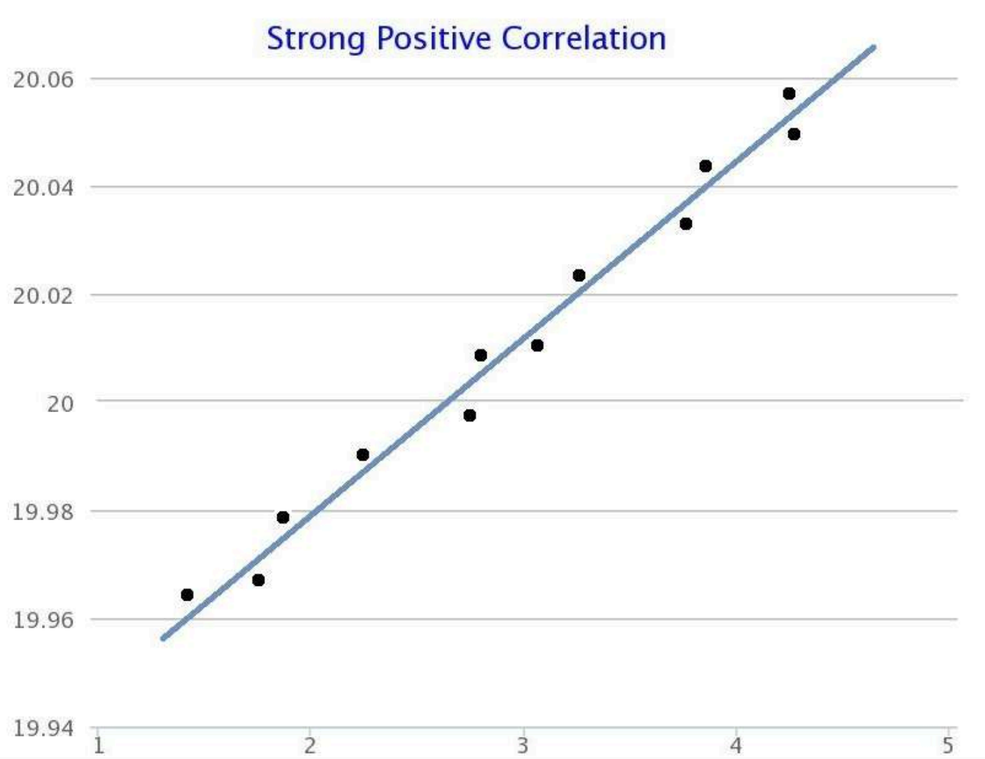
This high variance value indicates that the number of likes varies significantly from the mean. It suggests a wide spread in the data, with some observations having a high number of likes and others a low number, which is reflected in the uneven peaks across the graph.

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

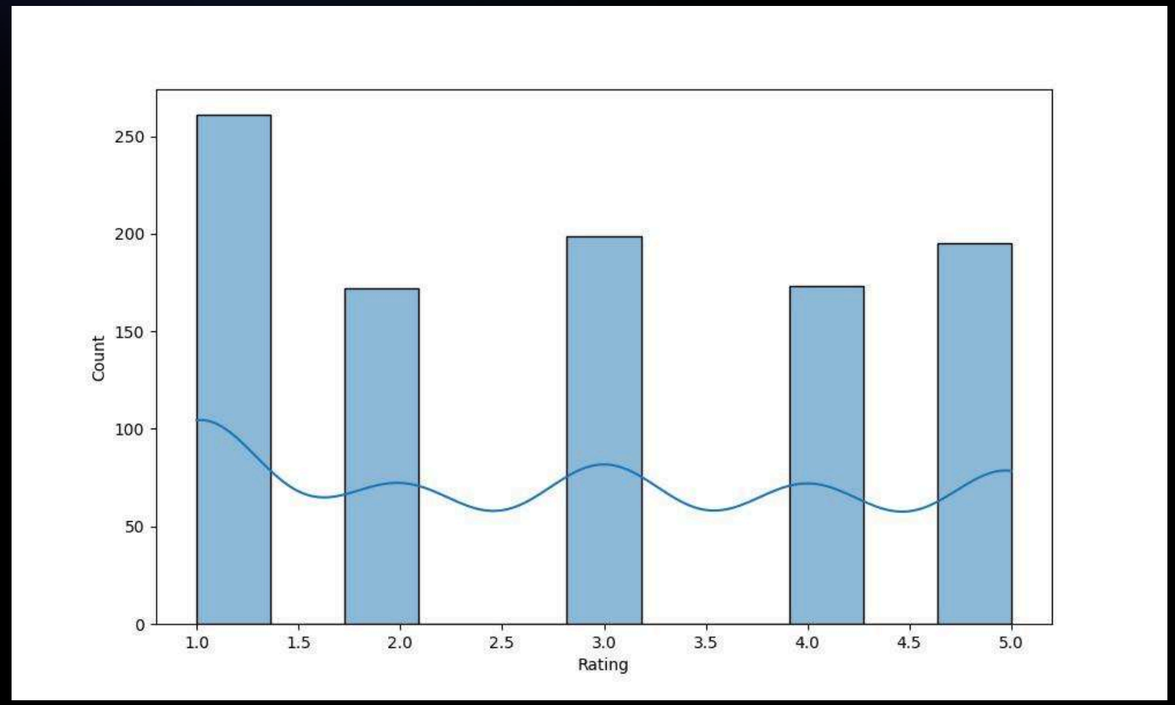
- A value above 0.8 generally represents a strong correlation, meaning there is a clear and predictable trend between the two variables. For instance, if the rating of a product goes up, the number of likes it receives also tends to rise, and vice versa.

- 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 average rating of Instagram is equal to or less than the average rating for WhatsApp.

Alternative Hypothesis (H_1):

The average rating of Instagram is greater than the average rating for WhatsApp.

- If the p-value is less than 0.05, you reject the null hypothesis, which would mean Instagram's average rating is significantly higher than WhatsApp's, according to the data.
- If the p-value is greater than or equal to 0.05, you fail to reject the null hypothesis, indicating that there is no significant difference in ratings between Instagram and WhatsApp.

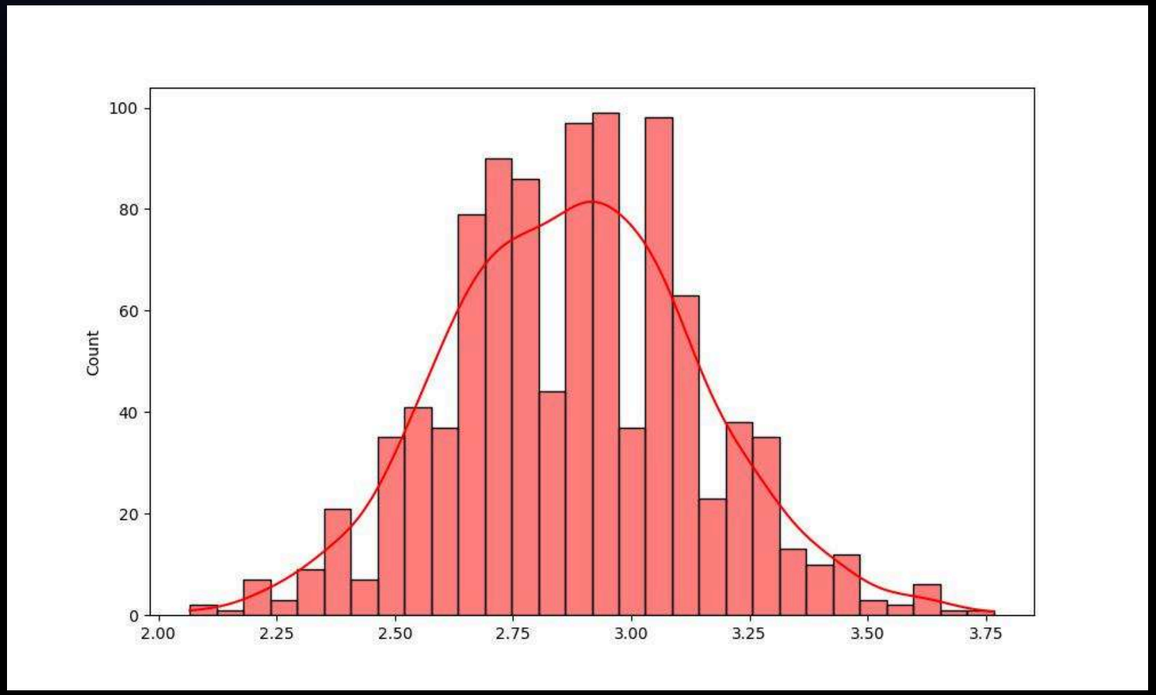
T-STATISTIC: -0.7967

P-VALUE: 0.7868

- Since the p-value (0.7868) is much larger than the significance level of 0.05, we fail to reject the null hypothesis.
- This means that there is no significant difference between the ratings for the two platforms based on the given data.

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 indicates a moderate level of user satisfaction, while the mode of 1 highlights a notable proportion of low ratings.
- Spending Behavior: The range and interquartile range (IQR) of purchase amounts reflect moderate variability in user spending habits.
- Diverse Engagement: A significant variance in the number of likes demonstrates varying levels of user engagement with app reviews.
- Correlation Insights: A strong positive correlation (r=0.8425r = 0.8425r=0.8425) between likes and ratings suggests that higher-rated reviews tend to attract more likes.
- Hypothesis Testing: Statistical analysis revealed no significant difference in ratings between Instagram and WhatsApp.
- Central Limit Theorem: The application of the Central Limit Theorem ensured reliable inferences, supporting data-driven decisions to enhance app features and improve user satisfaction.

THANK YOU :)

Connect With Me

 www.linkedin.com/in/ethendcosta

 ethendcosta5@gmail.com

 https://github.com/EthenDcosta5