APPLE STORE REVIEWS DATA ANALYSIS

Presented by: Ajay Malik

Project Overview

This project uses statistical methods to examine product reviews from the Apple Store in order to learn more about app performance and user satisfaction. We will calculate measures of central tendency for ratings, evaluate the variability of purchase amounts, and use variance and standard deviation to examine review likes. A distribution plot will reveal skewness, and correlation analysis will look at the connection between likes and ratings. Instagram and WhatsApp ratings will be compared using a hypothesis test, and the Central Limit Theorem will be illustrated using a sampling distribution. Developers and stakeholders will be able to make wise decisions with the aid of these insights.

Problems

01 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?

02 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?

03 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?

04 Determine the correlation between the likes and the rating given. Is there a positive, negative, or no correlation between these variables?

05 Plot the distribution of the app ratings. Is the distribution positively or negatively skewed? What does this indicate about user satisfaction?

06 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.

RATING ANALYSIS

Mean

Mean (2.87): The mean is pulled down by the high frequency of low ratings (1s and 2s).

Median

Median (3.0): The median is the middle value and better represents the "typical" rating, as it is less skewed by outliers.

Mode

The mode (1) reflects the most common rating but does not represent the overall distribution well.

Conclusion: The median is the best measure of central tendency for this dataset due to the skewed distribution of ratings.

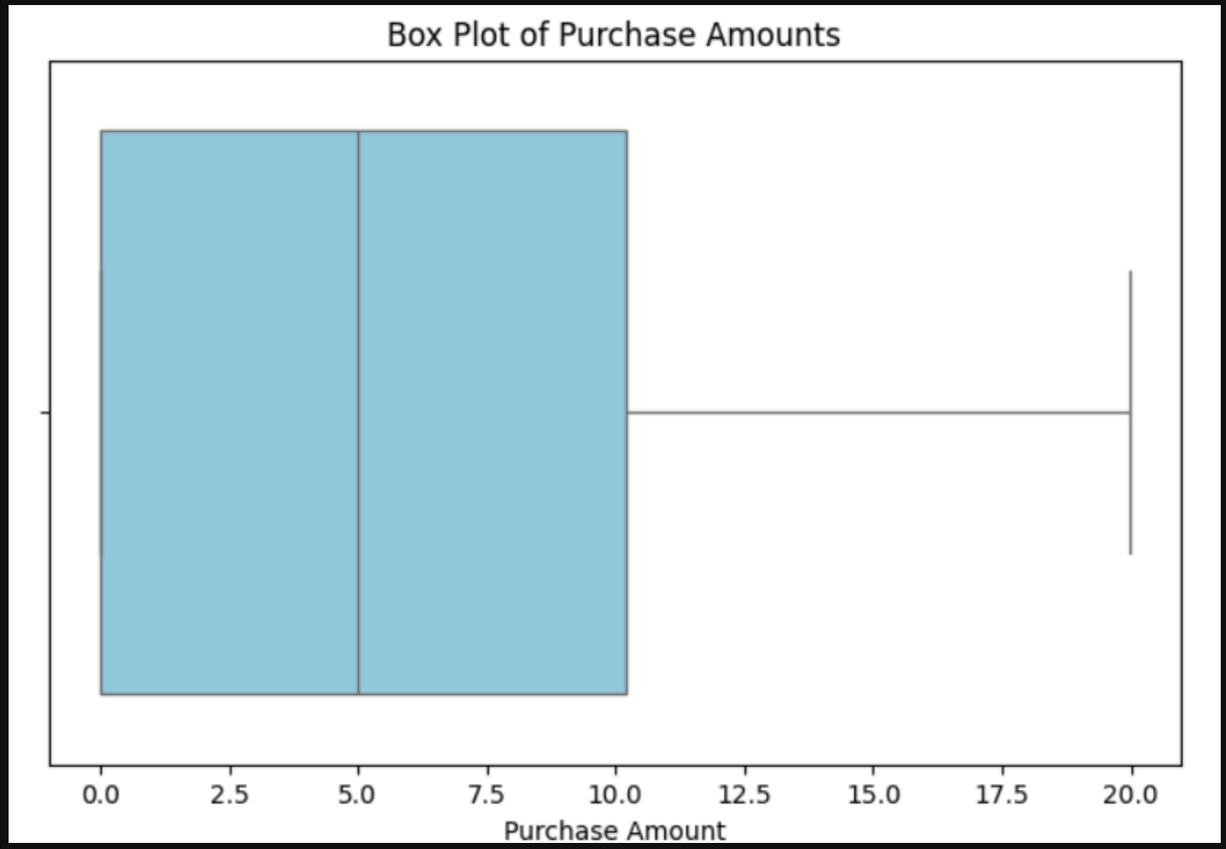
RANGE AND INQUARTILE RANGE(IQR)

Range: 19.97 (Difference between the highest and lowest purchase amounts)

Interquartile Range (IQR): 10.19 (Difference between the 75th percentile (Q3) and 25th percentile (Q1))

Interpretation:

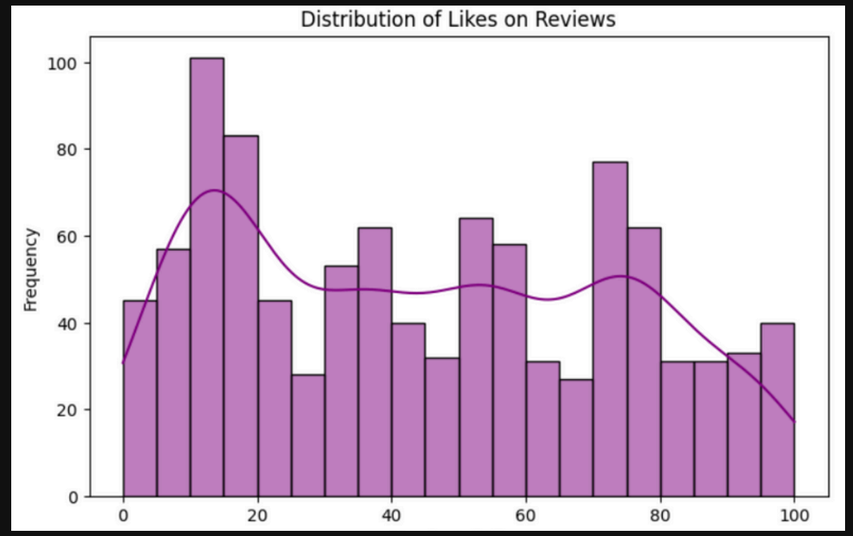
While the total spread of purchases is large (19.97), the middle 50% of transactions are relatively consistent (IQR = 10.19). The disparity between the two metrics highlights the influence of outliers or skewed tails on the dataset.



VARIANCE AND STANDARD DEVIATION OF LIKES

Variance (822.85): This indicates that the number of likes received on reviews has a large spread from the mean. Higher variance suggests more fluctuation in likes across different reviews.

The standard deviation (28.6854) represents the spread in the same units as the original data, highlighting significant variability in user engagement with reviews.

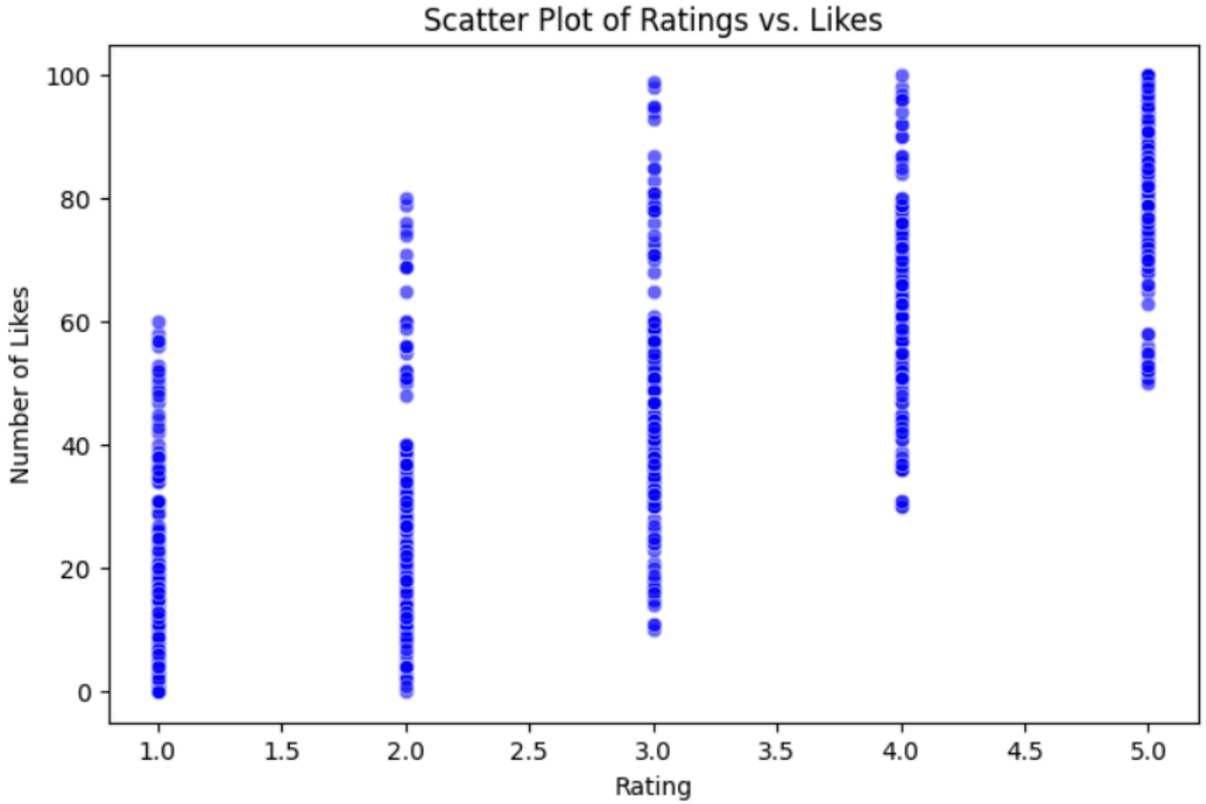


Conclusion: A high standard deviation of 28.69 suggests a wide variation in user engagement with reviews. This means that while some reviews received very few likes, others received significantly more, indicating diverse reactions among users

CORRELATION BETWEEN THE LIKES AND THE RATINGS

The correlation between likes and ratings is 0.84, which indicates a strong positive correlation.

Since the correlation is close to +1, it suggests that higher ratings tend to receive more likes on reviews.
This means that users are more likely to engage with and appreciate positive reviews.
The relationship is not perfect (not exactly 1), but it is quite strong.

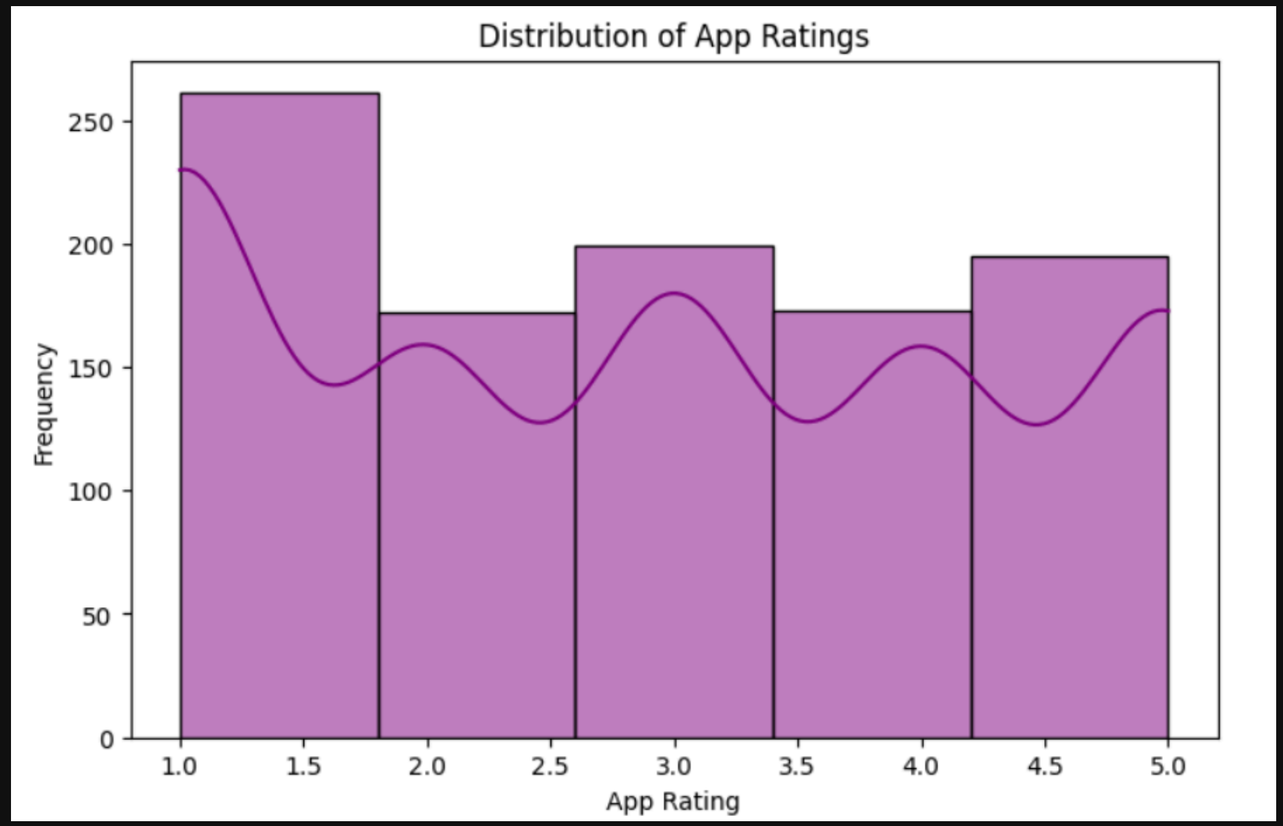


Here's the scatter plot showing the relationship between ratings and likes:
You can see a clear upward trend, confirming the strong positive correlation (0.84).
Higher ratings generally receive more likes on reviews, meaning users tend to engage more with positive feedback.

DISTRIBUTION OF THE APP RATINGS

Here's the histogram of app ratings:
Skewness Analysis:
The distribution appears negatively skewed (left-skewed) since there is a concentration of ratings at the lower end (1 and 2).
This means more users gave lower ratings, dragging the average down.

What Does This Indicate?
A negative skew suggests lower user satisfaction, with a significant number of users giving poor ratings.
There may be common issues (e.g., bugs, usability problems, or dissatisfaction) leading to these low ratings.



HYPOTHESIS TEST

FOR INSTAGRAM VS. WHATSAPP RATINGS

NULL HYPOTHESIS

The average rating for Instagram is less than or equal to the average rating for WhatsApp.
(μInstagram≤μWhatsApp)

ALTERNATIVE HYPOTHESIS

The average rating for Instagram is higher than the average rating for WhatsApp.
(μInstagram>μWhatsApp)

We'll use a 95% confidence level (α=0.05α=0.05). Let’s perform the test now. Hypothesis Test Results:
t-statistic: -0.80
p-value: 0.79

INTERPRETATION:

The average rating for Instagram is higher than the average ratingThe p-value (0.79) is much greater than 0.05, meaning we fail to reject the null hypothesis.

CONCLUSION:

This suggests that Instagram's average rating is not significantly higher than WhatsApp's rating at the 95% confidence level.
In fact, the negative t-statistic suggests that WhatsApp may even have a slightly higher average rating than Instagram, though not significantly.

SAMPLING DISTRIBUTION AND CENTRAL LIMIT THEOREM (CLT)

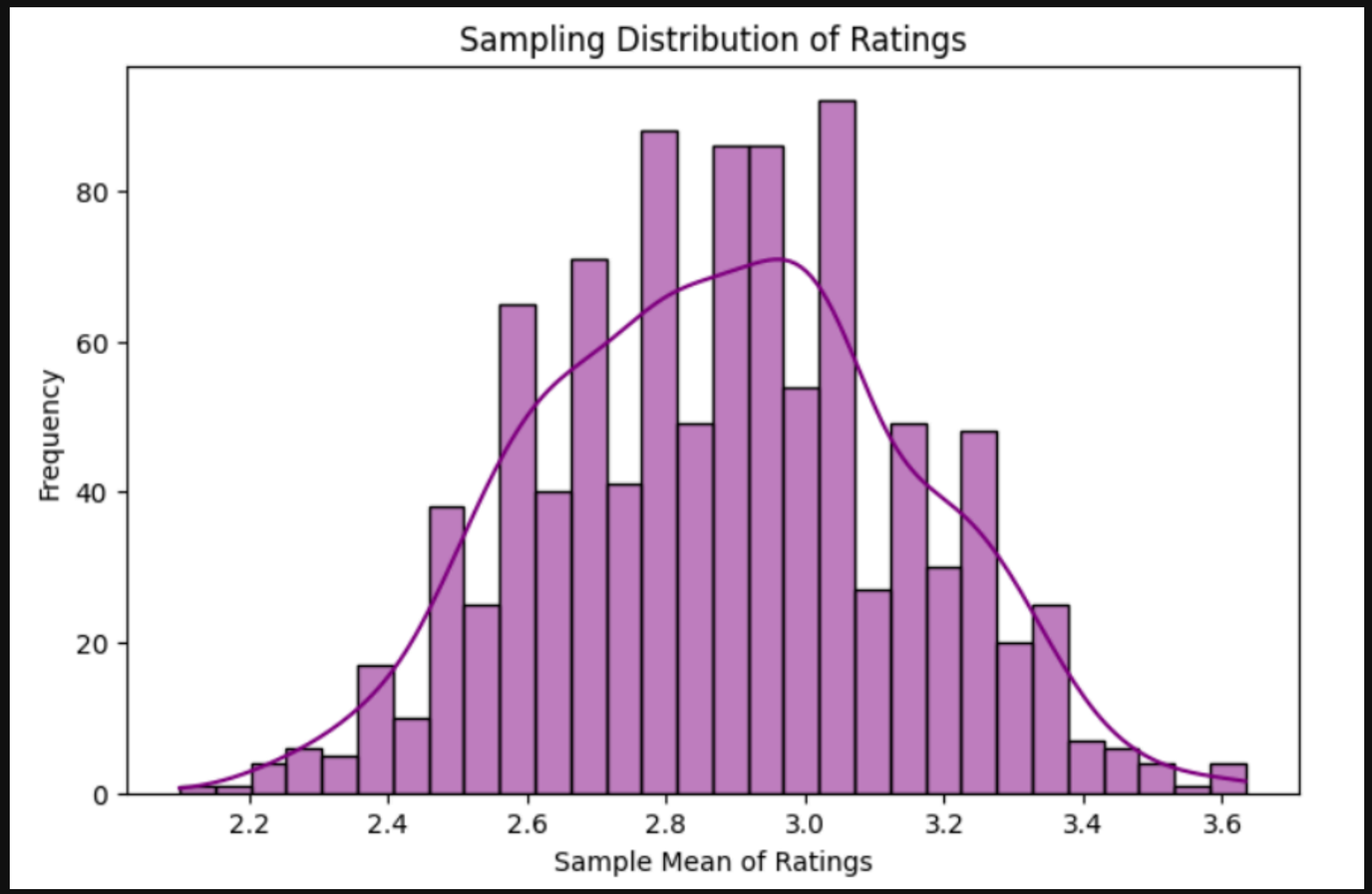
EXPLANATION & CENTRAL LIMIT THEOREM (CLT):

The sampling distribution of the mean follows a normal shape, even though individual ratings may not be normally distributed.
This occurs due to the Central Limit Theorem (CLT), which states that:

Regardless of the shape of the original population, the distribution of the sample means approaches a normal distribution as the sample size increases.

The mean of the sampling distribution is approximately equal to the population mean.

The standard deviation of the sampling distribution (standard error) is smaller than the population standard deviation.



KEY TAKEAWAYS:

The histogram confirms the normal shape predicted by CLT.
The more samples we take, the more precise our estimate of the population mean becomes.
This principle is the foundation for many statistical tests, confidence intervals, and hypothesis testing.

HYPOTHESIS TEST

KEY FINDINGS AND INSIGHTS

- User Satisfaction: The median rating of 3 reflects mixed satisfaction, with a significant number of low ratings (mode = 1).
- Spending Behavior: The range and IQR of purchase amounts indicate moderate variability in user spending habits.
- User Engagement: A high variance in likes suggests diverse engagement levels with app reviews.
- Correlation Analysis: A strong positive correlation (r = 0.8425) between likes and ratings shows that higher-rated reviews receive more likes.
- Hypothesis Testing: No significant difference was found between Instagram and WhatsApp ratings, implying similar user perceptions.
- Central Limit Theorem: The CLT validated reliable statistical inferences, aiding data-driven improvements in app features and user experience.

THANK YOU SO MUCH

Presented by: Ajay Malik