Apple Store Reviews

Statistical Analysis

Presented By : Sangeeta Rajput

Problem Statement

Understanding user ratings, purchases, and review interactions is crucial for app performance evaluation. This study aims to analyze the central tendency (mean, median, and mode) of app ratings and determine the best representative measure. The spread of purchase amounts will be examined using range and interquartile range (IQR). Variance and standard deviation of likes on reviews will be computed to assess engagement dispersion. Correlation analysis will determine the relationship between likes and ratings. A distribution plot will reveal the skewness of app ratings and its implications for user satisfaction. A hypothesis test will compare the average ratings of Instagram and WhatsApp at a 95% confidence level. A sampling distribution of ratings will be created to demonstrate the Central Limit Theorem. The insights will guide app developers in improving user experience and business strategies.

Sangeeta Rajput

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?

The calculated values for app ratings are:

- Mean: 2.869
- Median: 3.0
- Mode: 1

Since the mode (1) is significantly lower than the mean and median, it suggests that a large number of ratings are clustered at the lower end. The median (3.0) is likely the best measure of central tendency because it is less affected by extreme values (outliers) compared to the mean.

Sangeeta Rajput

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?

The computed values for Purchase_Amount are:

- Range: 19.97
- Interquartile Range (IQR): 10.19

Interpretation:

- Range measures the total spread of the data by subtracting the minimum from the maximum value. A large range suggests high variability in purchase amounts.
- IQR represents the middle 50% of the data, reducing the influence of outliers. A higher IQR indicates that purchase amounts are more spread out, while a lower IQR suggests consistency among most users.

Sangeeta Rajput

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?

The computed values for the number of likes received on reviews are:

- Variance: 822.85
- Standard Deviation: 28.69

Interpretation:

- Variance measures how much the data points deviate from the mean in squared units.
- Standard Deviation (SD) is the square root of variance and represents the typical spread of likes from the average.
- A higher SD (28.69) suggests that the number of likes varies significantly across reviews, meaning some reviews get very few likes while others get a lot.

Sangeeta Rajput

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

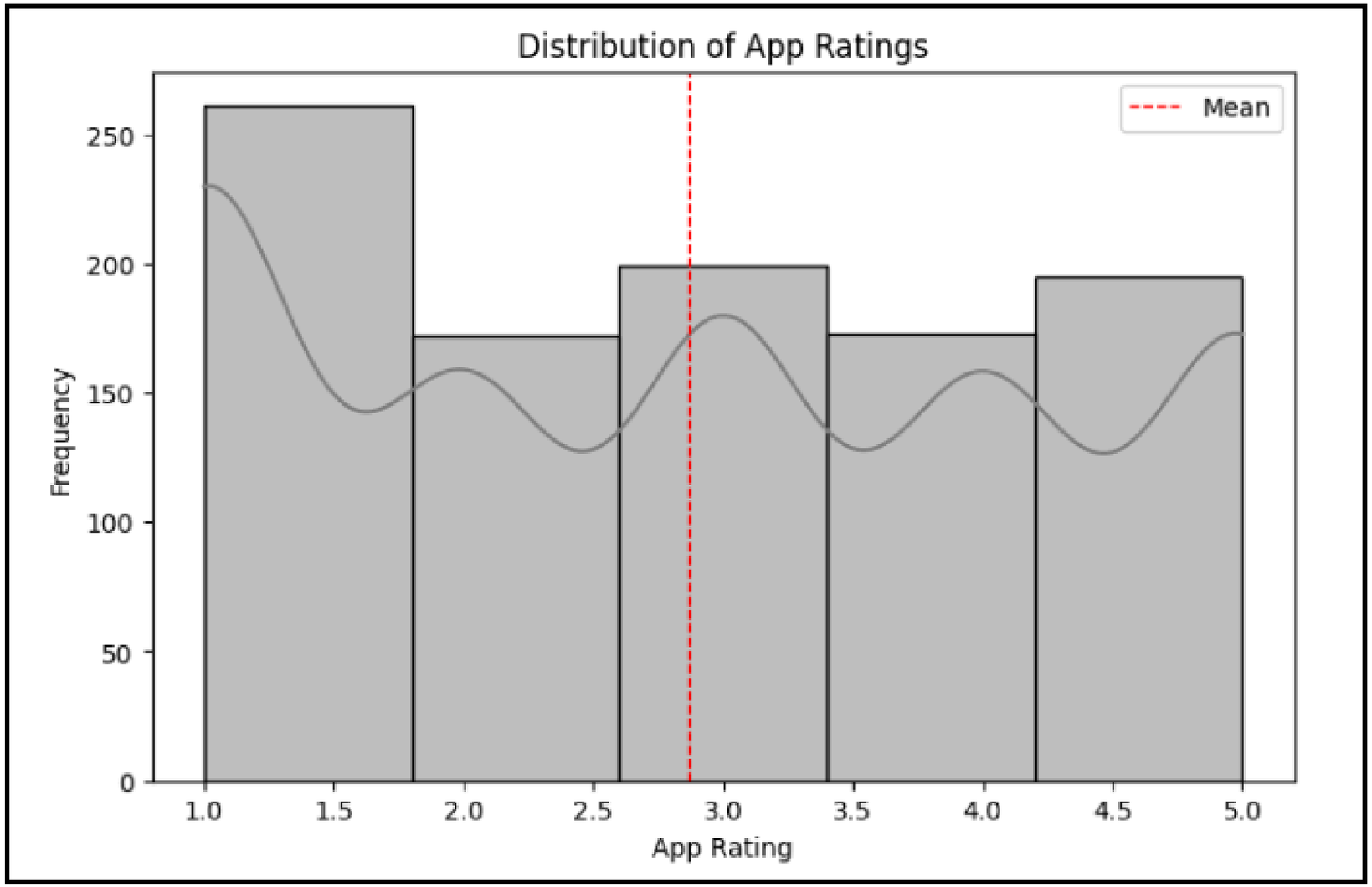
The computed correlation between Likes and Rating is 0.84, indicating a strong positive correlation.

Interpretation:

- Since the correlation is close to 1, it suggests that higher ratings tend to receive more likes.
- This positive relationship implies that users are more likely to engage (like a review) when it reflects a higher rating.

in **Sangeeta Rajput**

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



Sangeeta Rajput

The skewness of the app ratings is 0.10, which is close to 0, indicating that the distribution is approximately symmetrical with a slight positive skew.

Interpretation:

- A slight positive skew means that there are a few higher ratings (e.g., 4 or 5) that slightly pull the average upward.
- Since the distribution is nearly symmetrical, it suggests that most users rate the apps fairly evenly, meaning the user satisfaction is balanced, with no extreme dissatisfaction or overwhelming positivity.

Sangeeta Rajput

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.

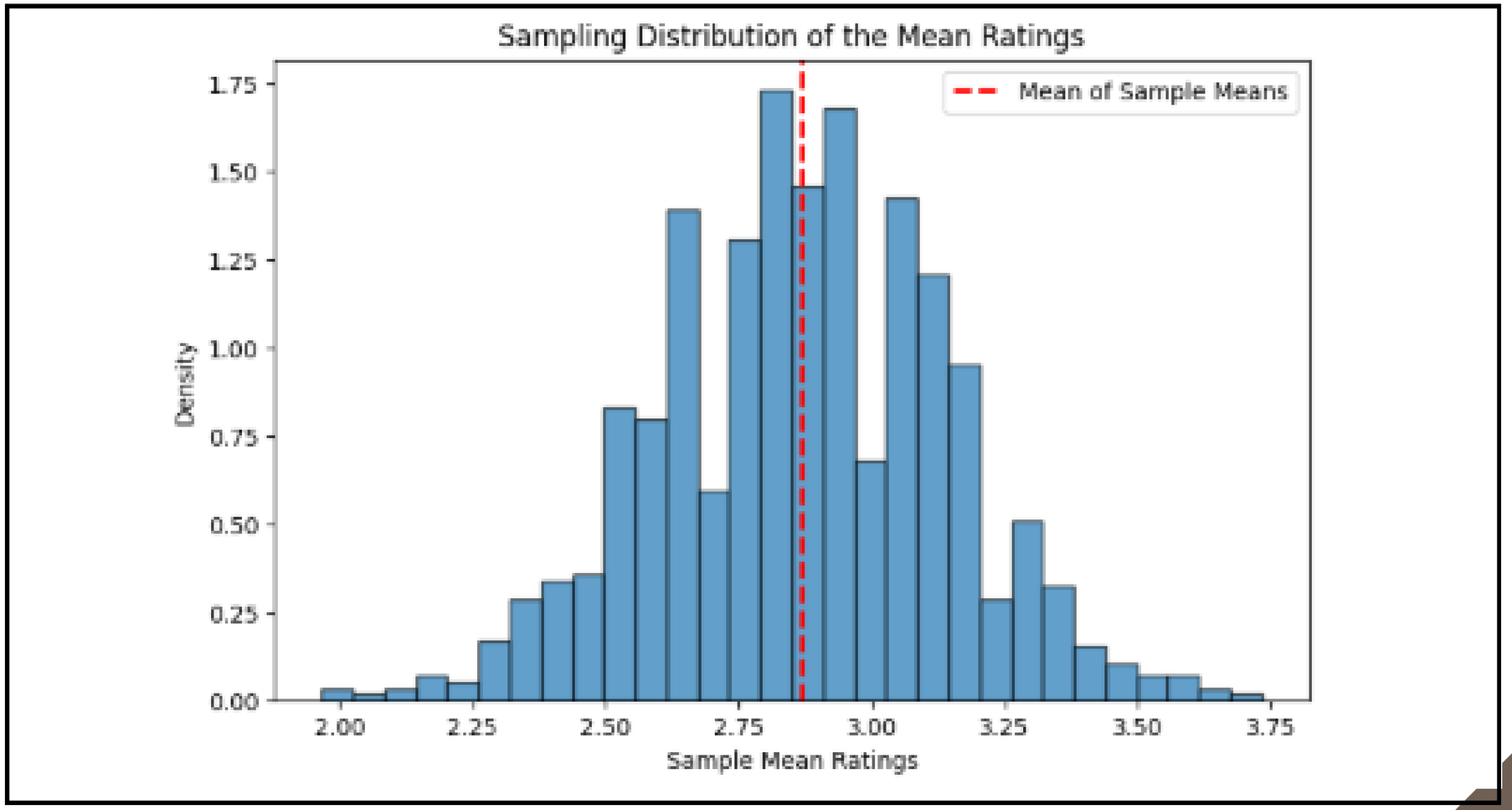
Hypothesis Test Results:

- Mean Instagram Rating: 2.77
- Mean WhatsApp Rating: 2.93
- t-statistic: -0.80
- p-value: 0.79

Since the p-value (0.79) is much greater than the significance level (0.05), we fail to reject the null hypothesis. This means there is no significant evidence to suggest that Instagram’s average rating is higher than WhatsApp’s. In fact, WhatsApp's mean rating is slightly higher.

Sangeeta Rajput

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.



Sangeeta Rajput

Explanation of the Central Limit Theorem (CLT):

1. **Sampling Distribution:** The histogram above represents the distribution of sample means from 1,000 randomly drawn samples (each of size 30).
2. **Key Observations:**
    - The shape of the distribution is approximately normal, even though individual ratings might not be normally distributed.
    - The mean of the sample means (≈ 2.88) is close to the overall population mean.
    - The spread (standard deviation) of the sample means is smaller than that of individual ratings.

in **Sangeeta Rajput**

Relation to CLT:

- The CLT states that when we take sufficiently large random samples from any population (regardless of the population's original distribution), the distribution of the sample means will approach a normal distribution.
- The standard deviation of the sample means (standard error) is lower than the standard deviation of the population, meaning sample means fluctuate less than individual data points.

Sangeeta Rajput

Thank You

Sangeeta Rajput