Report: Analysis of Dataset

Introduction

This report presents an analysis of a dataset, including an explanation of the dataset, an exploration of a chosen column, calculation of descriptive statistics, identification of the shape of the distribution, detection of outliers, and construction of confidence intervals for the mean and variance.

2. Chosen Column

The chosen column is named “User Rating” The selection of this column was based on its potential significance and relevance to the analysis objectives. I want to see how users rated books according to their Author and Price.

3. Mean of Column Data

The mean of the column data is calculated to understand the central tendency or average value. It provides insights into the typical value observed in the dataset.

4. Median of Column Data

The median of the column data is calculated to determine the middle value. Unlike the mean, the median is less sensitive to outliers and provides a robust measure of central tendency.

5. Variance, Standard Deviation, and Standard Error

The variance, standard deviation, and standard error provide measures of dispersion and variability within the dataset.

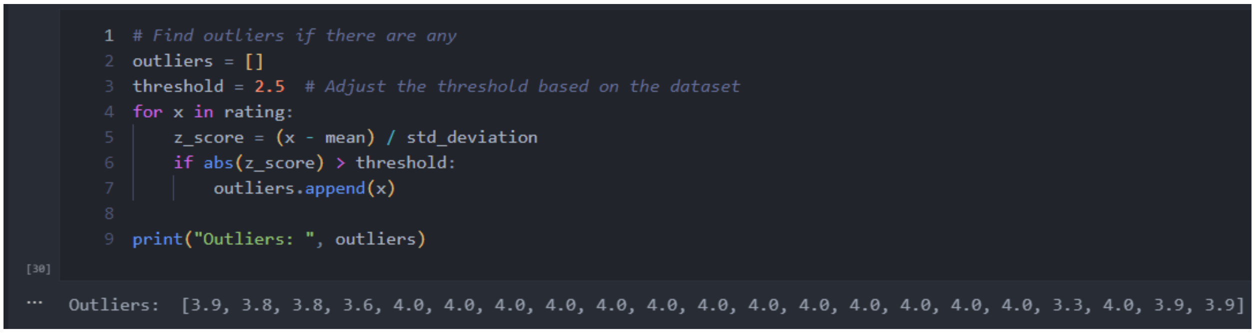
- Variance: It quantifies the spread of data points around the mean.
- Standard Deviation: It is the square root of the variance and provides a more interpretable measure of dispersion.
- Standard Error: It estimates the precision of the sample mean and reflects the variation between different samples

6. Shape of Distribution

The shape of the distribution provides insights into the underlying patterns and characteristics of the data.

7. Outliers

Outliers are data points that significantly deviate from the overall pattern observed in the dataset. Identifying outliers is important as they may indicate measurement errors, data anomalies, or unique observations.



```python
# Find outliers if there are any
outliers = []
threshold = 2.5  # Adjust the threshold based on the dataset
for x in rating:
    z_score = (x - mean) / std_deviation
    if abs(z_score) > threshold:
        outliers.append(x)

print("Outliers: ", outliers)
```

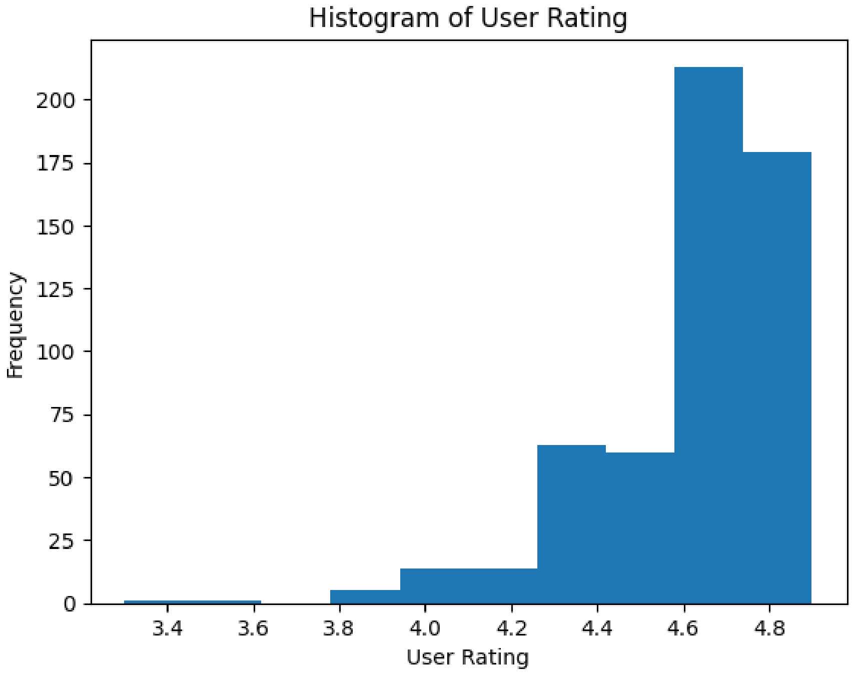
[30]

```
...  Outliers:  [3.9, 3.8, 3.8, 3.6, 4.0, 4.0, 4.0, 4.0, 4.0, 4.0, 4.0, 4.0, 4.0, 4.0, 4.0, 4.0, 4.0, 3.3, 4.0, 3.9, 3.9]
```

8. Histogram

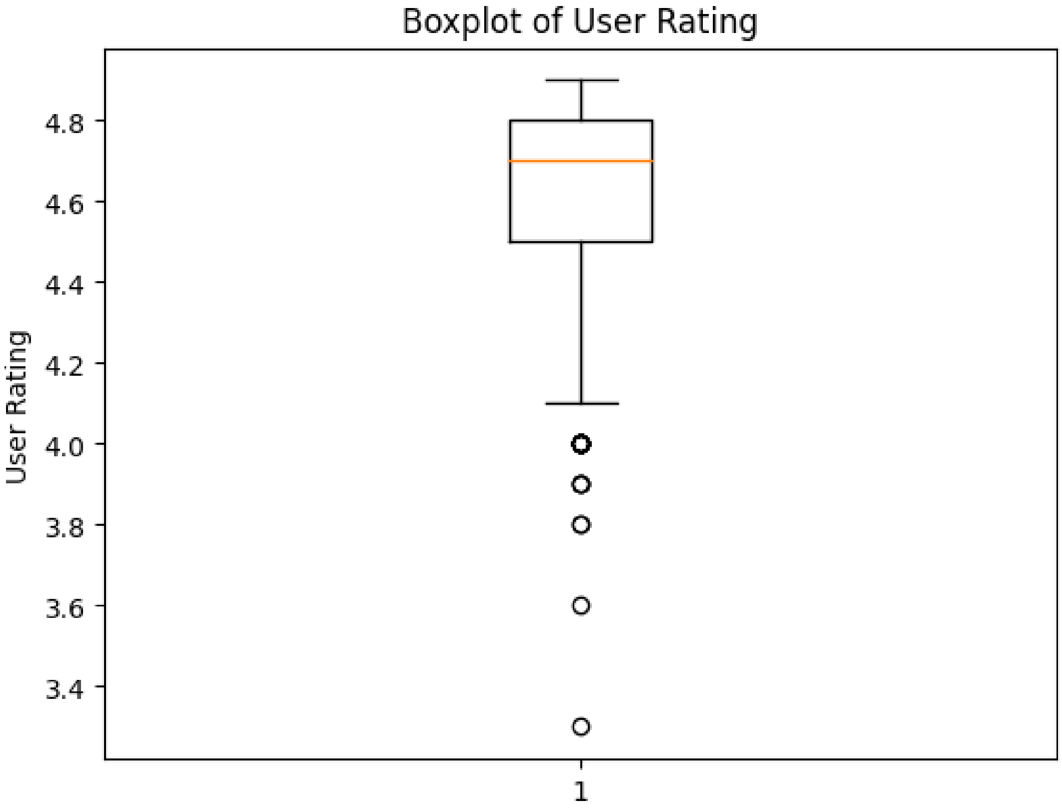
A histogram is a graphical representation that shows the distribution of data points in intervals, providing insights into the frequency and concentration of values within the chosen column.

This histogram shows that ratings generally like in 4.5 – 4.7 and there is few datas that its rating lower than 4.3.



9. Boxplot

A boxplot is a visual tool used to display the distribution of data through quartiles, median, and potential outliers. It provides a summary of the dataset's spread and identifies extreme values.



As we mentioned in the histogram, the boxplot shows that the median of the rating is 4.6... Range of ratings is 4.3 – 4.8. There are some outliers that its rating lower than 4.2.

10. Confidence Intervals for Mean and Variance

Confidence intervals provide an estimated range within which the population mean and variance are likely to fall.

11. Sample Size Determination

To estimate the population mean with a specific margin of error and confidence level, it is essential to determine an appropriate sample size. For this analysis, the desired margin of error is [specified margin of error] units, and the confidence level is [specified confidence level].

12. Conclusion

In conclusion, this report provided an analysis of a dataset, focusing on a chosen column. Descriptive statistics, including the mean, median, variance, standard deviation, and standard error, were calculated. The shape of the distribution was identified, and outliers were discussed. Histogram and boxplot visualizations were provided to gain insights into the data distribution and extreme values. Confidence intervals for the mean and variance were constructed to estimate the likely range for population parameters. Additionally, recommendations for determining an appropriate sample size for estimating the population mean with a specific margin of error and confidence level were given.

Mehmetcan Bozkuş – 2121221041