Statistical Distribution and Boxplot Explanation

Heider Jeffer

Boxplot Explanation

The **boxplot** used in the code visualizes the statistical distribution of engagement levels for each activity of a given stakeholder. In this case, the boxplot will help to understand the spread, central tendency, and potential outliers of the engagement levels.

A **boxplot** (also known as a **box-and-whisker plot**) is used to display the distribution of a dataset. It shows the following:

- **Median (Q2)**: The middle value of the dataset (50th percentile).
- **First Quartile (Q1)**: The 25th percentile, or the median of the lower half of the data.
- **Third Quartile (Q3)**: The 75th percentile, or the median of the upper half of the data.
- **Interquartile Range (IQR)**: The range between the first quartile (Q1) and the third quartile (Q3). It represents the middle 50% of the data.
- **Whiskers**: These lines extend from the first quartile (Q1) and third quartile (Q3) to show the range of data. The whiskers typically extend to 1.5 $\times$ IQR from Q1 and Q3, beyond which points are considered as outliers.
- **Outliers**: Data points that fall outside the whiskers are considered outliers and are typically marked as dots.

Formulae for Boxplot Construction

Given a dataset of engagement levels for a particular activity, here are the key steps/formulae involved in calculating the statistics used in a boxplot:

- **Median (Q2)**:

  $$\text{Median} = \text{middle value of sorted data}$$

  If the number of data points is odd, it's the middle number. If even, it's the average of the two middle values.

- **First Quartile (Q1)**:

$$Q1 = \text{median of the lower half of the dataset}$$

- **Third Quartile (Q3)**:

$$Q3 = \text{median of the upper half of the dataset}$$

- **Interquartile Range (IQR)**:

$$\text{IQR} = Q3 - Q1$$

- **Whiskers**: The whiskers extend to:

$$\text{Lower whisker} = \max(\text{Minimum value}, Q1 - 1.5 \times \text{IQR})$$

$$\text{Upper whisker} = \min(\text{Maximum value}, Q3 + 1.5 \times \text{IQR})$$

  Any data points outside these whiskers are considered outliers.

Numerical Example

Let's walk through a simple numerical example based on the `df_activity_engagement` data, using one activity for a single stakeholder (say **Patients** and their activity **Participation in Care**).

Assume we have the following engagement levels for **Patients** in the activity **Participation in Care** over 12 months:

$$[0.6, 0.65, 0.7, 0.55, 0.6, 0.7, 0.75, 0.6, 0.8, 0.65, 0.7, 0.55]$$

Step 1: Sort the Data

Sort the data in ascending order:

$$[0.55, 0.55, 0.6, 0.6, 0.6, 0.65, 0.65, 0.7, 0.7, 0.7, 0.75, 0.8]$$

Step 2: Calculate the Median (Q2)

There are 12 values, so the median is the average of the 6th and 7th values:

$$\text{Median} = \frac{0.65 + 0.65}{2} = 0.65$$

Step 3: Calculate the First Quartile (Q1)

The first quartile is the median of the lower half of the data (first 6 values):

$$\text{Lower half: } [0.55, 0.55, 0.6, 0.6, 0.6, 0.65]$$

The median of this half is the average of the 3rd and 4th values:

$$Q1 = \frac{0.6 + 0.6}{2} = 0.6$$

Step 4: Calculate the Third Quartile (Q3)

The third quartile is the median of the upper half of the data (last 6 values):

$$\text{Upper half: } [0.65, 0.7, 0.7, 0.7, 0.75, 0.8]$$

The median of this half is the average of the 3rd and 4th values:

$$Q3 = \frac{0.7 + 0.7}{2} = 0.7$$

Step 5: Calculate the Interquartile Range (IQR)

$$\text{IQR} = Q3 - Q1 = 0.7 - 0.6 = 0.1$$

Step 6: Calculate the Whiskers

The lower whisker extends to:

$$\text{Lower whisker} = \max(\text{Minimum value}, Q1 - 1.5 \times \text{IQR}) = \max(0.55, 0.6 - 1.5 \times 0.1) = \max(0.55, 0.45) = 0.55$$

The upper whisker extends to:

$$\text{Upper whisker} = \min(\text{Maximum value}, Q3 + 1.5 \times \text{IQR}) = \min(0.8, 0.7 + 1.5 \times 0.1) = \min(0.8, 0.85) = 0.8$$

Step 7: Identify Outliers

The outliers are any data points that fall outside the whiskers (below 0.55 or above 0.8). In this case, there are **no outliers**, as all data points fall within the whiskers.

Final Distribution

- **Median (Q2)**: 0.65
- **First Quartile (Q1)**: 0.6
- **Third Quartile (Q3)**: 0.7
- **Interquartile Range (IQR)**: 0.1
- **Lower Whisker**: 0.55
- **Upper Whisker**: 0.8
- **Outliers**: None

Boxplot Interpretation

For the activity **Participation in Care**:

- The **box** will span from **0.6** (Q1) to **0.7** (Q3), with the **median line** at **0.65**.
- The **whiskers** will extend from **0.55** (lower) to **0.8** (upper).
- All the data points will fall within the whiskers, so there will be **no outliers**.

Visualization in the Code

The **boxplot** will show a box from **0.6** to **0.7**, a line at **0.65** (median), and whiskers extending from **0.55** to **0.8**. If we had more data with more variation, the whiskers and box might expand or outliers could be detected, showing how the engagement levels fluctuate for the activity.

Conclusion

This is how the boxplot visualizes the distribution of engagement levels for an activity. By repeating this process for all activities and stakeholders, we can get a comprehensive view of how engagement varies across the stakeholders and their activities, identifying trends, stability, and outliers.