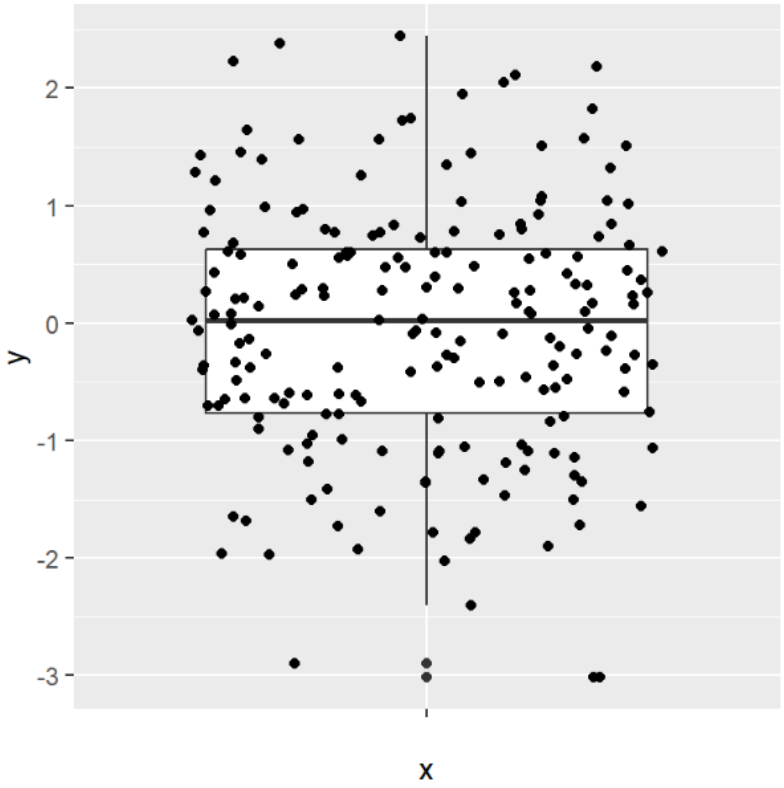
How to interpret Box plot with jitters – numerical variable



A. Interpretation of the Box Plot Components:

- **Box (Interquartile Range - IQR):**
    - The **bottom edge of the box** (Lower Quartile - Q1) is approximately at **-0.5**. 25% of the data points have a value less than or equal to -0.5.
    - The **top edge of the box** (Upper Quartile - Q3) is approximately at **0.6**. 75% of the data points have a value less than or equal to 0.6.
    - The **length of the box (IQR = Q3 - Q1 = 0.6 - (-0.5) = 1.1)** shows the spread of the middle 50% of the data.
- **Line Inside the Box (Median):**
    - The **line within the box** (Median - Q2) is approximately at **0**. This is the middle value of the dataset.

    - **Position relative to the box:** The median (0) is slightly above the center of the box (which spans from -0.5 to 0.6). This suggests a slight **positive skew** in the central 50% of the data.

- **Whiskers:**
    - The **upper whisker** extends to approximately **1.8**.
    - The **lower whisker** extends to approximately **-2.5**.
    - **Length of the whiskers:** The lower whisker is noticeably longer than the upper whisker, suggesting a **negative skew** in the overall distribution. The data extends further towards lower values.
- **Points Outside the Whiskers (Outliers):**
    - There are several **points plotted outside the whiskers**, both above the upper whisker (around 2.2) and below the lower whisker (around -2.8 and -3). These are potential **outliers**.

B. Interpretation of the Jitters (Individual Data Points):

The **jittered points** represent each individual observation of the numerical variable 'y' for the single category 'x'. The horizontal spread (jitter) is added purely for visualization purposes to prevent overlapping points and show the density of the data at different values.

- **Density:** Areas with a higher concentration of jittered points indicate where most of the data values lie. You can see a higher density of points within the box, as expected for the central 50% of the data.
- **Spread:** The jitters visually reinforce the overall spread of the data, from the lowest to the highest values.
- **Outliers:** The jitters clearly show the data points that fall outside the range defined by the whiskers, making the outliers more apparent and their values discernible.
- **Shape:** While the box plot gives a summary of the shape, the jitters provide a more granular view of the distribution. You can get a better sense of whether the distribution is unimodal, bimodal, or has other specific characteristics by looking at the spread of the individual points.

Box plots with jitters are particularly useful in the following scenarios for univariate (and also for comparing across categories) analysis:

- **Visualizing the Distribution While Showing Individual Data Points:** This combination allows you to see the summary statistics (median, quartiles, outliers) provided by the box plot *and* the actual distribution of the data points. This can reveal nuances in the distribution (e.g., clusters, gaps) that might be hidden by the summarized box plot alone.
- **Small to Moderate Datasets:** When you have a dataset that isn't so large that the jitters become an overwhelming mass, this visualization provides a good balance between summary and detail.
- **Identifying Outliers and Understanding Their Position:** The box plot clearly flags outliers, and the jitters show their exact values and how far they are from the main distribution.
- **Assessing Skewness and Spread with More Detail:** While the box plot indicates skewness and spread, the jitters can provide a more intuitive visual confirmation of these characteristics. For example, you can see if the points are more densely packed on one side of the median.
- **Comparing Distributions with More Granularity (Bivariate/Multivariate):** When comparing a numerical variable across different categories, using box plots with jitters side-by-side allows you to see both the summary statistics for each group and the underlying distribution of individual data points within each group. This can reveal differences in spread, skewness, and the presence of specific data clusters that might be missed by just comparing the boxes.
- **Highlighting Potential Multimodality:** If the jittered points show distinct clusters within a single box or whisker range, it might suggest that the underlying distribution is multimodal, a feature that might not be immediately obvious from the box plot alone.

In essence, adding jitters to a box plot enhances its utility by providing a richer view of the data distribution beyond the summary statistics. It's a powerful way to combine the benefits of a concise summary with the detailed information of individual observations. This makes it a preferred choice when you want to understand the overall distribution and identify potential patterns or issues at a more granular level than a standard box plot allows.