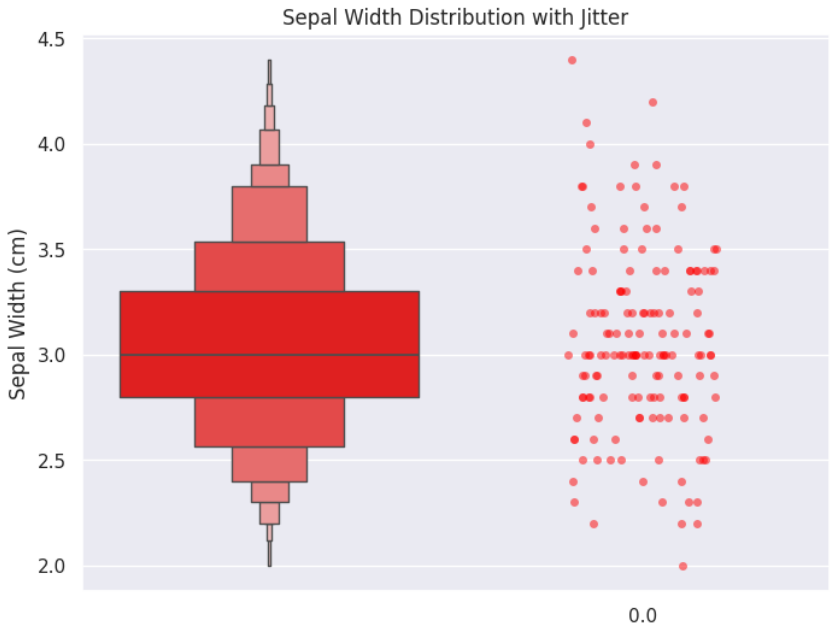
How to interpret boxen plot with jitters



A. Interpretation of the Boxen Plot Components:

- **Vertical Axis (Y-axis):** Represents the range of values for "Sepal Width (cm)," spanning from approximately 2.0 to 4.5.
- **Boxes:** Multiple boxes represent different letter-value intervals, showing the distribution's shape, particularly in the tails. The innermost (darkest red) represents the median and the middle 50% of the data. Subsequent, lighter boxes represent wider central proportions.
- **Width of the Boxes:** Indicates the density of data within each letter-value interval. Wider boxes signify more data points in that quantile range.
- **Jitters (Individual Data Points):** The scattered red points to the right of the Boxen plot represent each individual observation of "Sepal Width (cm)". The horizontal spread (jitter) is added to prevent overlap and show the density of data at different width values.

B. Interpreting the "Sepal Width (cm)" Distribution:

- **Central Tendency:** The median (innermost box) is around **3.0 cm**.

- **Spread:** The IQR (represented by the innermost box) spans roughly from **2.8 cm to 3.3 cm**. The increasing width of the outer boxes shows how the data spreads out in the more extreme quantiles. The overall range of the data (as seen from the jitters) is approximately from 2.0 cm to 4.4 cm.
- **Shape:** The Boxen plot appears somewhat symmetrical around the median, although the wider lower boxes might suggest a slightly heavier lower tail compared to the upper tail.
- **Tails:** The multiple outer boxes provide a detailed view of the distribution in the tails, indicating how the frequency decreases as we move away from the center.
- **Individual Data Points (Jitters):**
  - The jitters visually confirm the central tendency and spread shown by the Boxen plot. The highest density of points is around the median.
  - They reveal the presence of individual data points across the entire range, including the tails.
  - The jitters don't show any clear distinct clusters or multimodality in this distribution.
  - We can see data points extending to the extremes of the distribution, which correspond to the outer letter-value intervals in the Boxen plot.

Combining a Boxen plot with jitters is particularly useful in the following scenarios for univariate (and also for comparing across categories) analysis:

- **Large Datasets Where Showing Individual Points is Still Desirable:** When you have a large dataset, a simple scatter plot of all points can become overwhelming. Adding jitters alongside a Boxen plot allows you to see the overall distribution summarized by the Boxen plot while still getting a sense of the density and spread of individual data points.
- **Understanding Distributional Shape and Individual Variation:** The Boxen plot provides a detailed view of the distribution's shape, especially in the tails, and the jitters add the layer of showing where individual data points fall within that shape. This can help in understanding if the distribution is smooth or if there are gaps or clusters at a finer level.

- **Identifying Outliers in the Context of the Overall Distribution:** While the Boxen plot helps identify extreme values through its letter-value intervals, the jitters make these individual outliers visually explicit and show their distance from the main body of the data.
- **Comparing Distributions with More Granularity:** When comparing the distribution of a numerical variable across different categories, using Boxen plots with jitters side-by-side allows you to see both the summarized shape for each group and the underlying distribution of individual data points within each group. This can reveal subtle differences that might be missed by looking at either visualization alone.
- **Assessing the Robustness of Summary Statistics:** By seeing the spread of individual points around the median and quartiles (as represented by the Boxen plot), you can get a better sense of how representative these summary statistics are of the underlying data.

In essence, adding jitters to a Boxen plot enhances its interpretability by bridging the gap between a statistical summary of the distribution (provided by the Boxen plot) and the raw data points. It allows you to see both the forest (overall shape and quantiles) and the trees (individual observations). This combination is particularly powerful when you want to understand the nuances of a distribution in a moderately large to large dataset without losing sight of the individual data points.