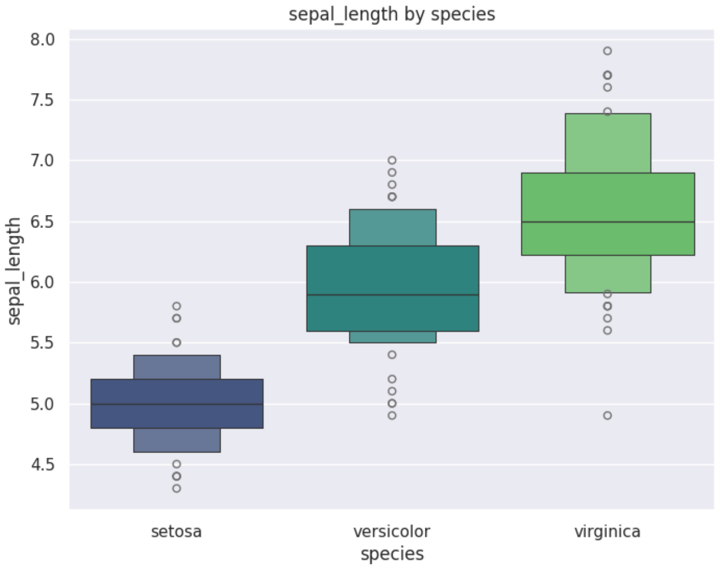
How to interpret Boxen plot for bivariate analysis



Let's interpret the boxen plot (also known as a letter-value plot) in the attached image, which visualizes the distribution of the numerical variable "sepal_length" across the different categories of the categorical variable "species" from the Iris dataset.

A. Understanding the Components of a Boxen Plot:

- **Horizontal Axis (X-axis):** Represents the categorical variable "species," with three categories: "setosa," "versicolor," and "virginica."
- **Vertical Axis (Y-axis):** Represents the numerical variable "sepal_length."
- **Boxes:** Instead of a single box representing the IQR, a boxen plot displays multiple nested boxes. Each box represents a different quantile range of the data:
    - The **innermost box** typically represents the median (50% of the data). [1]
    - The subsequent boxes represent progressively larger quantiles (e.g., 25-75%, 12.5-87.5%, 6.25-93.75%, and so on). These are often referred to as "letter values."

    - The number of boxes depends on the size and distribution of the data. More boxes are shown when the tails of the distribution are heavier.

- **Whiskers:** The lines extending from the outermost boxes (if present) indicate the spread of the remaining data, similar to the whiskers in a box plot, but their definition might vary slightly depending on the implementation (often related to a multiple of the IQR or other quantile-based measures).
- **Outliers:** Points plotted outside the whiskers represent values that are far from the central distribution.

B. Interpreting the Sepal Length Distribution by Species:

By examining the boxen plots, we can compare the sepal length distributions across the three species:

- **Setosa (Leftmost Boxen Plot):**
    - The innermost box (darkest blue) shows the median sepal length around 5.0 cm.
    - The nested boxes are relatively compact, indicating less spread in the central portions of the data.
    - The whiskers are short, and there are a few low outliers.
    - The overall shape suggests a relatively concentrated distribution.
- **Versicolor (Middle Boxen Plot):**
    - The innermost box (darkest teal) shows the median sepal length around 5.9 cm, which is higher than setosa.
    - The nested boxes are wider than those of setosa, indicating more spread in the central data.
    - There are some lower and upper outliers.
    - The shape suggests a more dispersed distribution compared to setosa.

- **Virginica (Rightmost Boxen Plot):**
    - The innermost box (darkest green) shows the median sepal length around 6.5-6.6 cm, the highest among the three species.
    - The nested boxes are the widest, indicating the most spread in the central data.
    - There are several lower and upper outliers.
    - The shape suggests the most dispersed distribution with potential skewness or heavier tails.

C. Overall Interpretation:

The boxen plot provides a more detailed view of the distribution's shape, particularly in the tails, compared to a standard box plot. It confirms the differences in median sepal length across the species, with setosa having the shortest, followed by versicolor, and then virginica. The increasing width of the nested boxes from setosa to virginica indicates increasing spread and potentially heavier tails in the sepal length distribution for these species. The presence and number of outliers are also clearly visualized.

Boxen plots are particularly useful when you want to:

- **Visualize the shape of the distribution, especially the tails and the presence of outliers, in more detail than a standard box plot.** The multiple letter-value boxes provide a better sense of the distribution's quantiles.
- **Compare distributions with potentially heavy tails or outliers across different categories.** The boxen plot gives a clearer picture of the extent and nature of these tails.
- **Work with large datasets where the detailed distribution, beyond just the IQR, is important.** The letter values provide more granular information about the spread of the data.
- **Identify potential differences in the shape of the distribution (e.g., kurtosis, multimodality - although multimodality might be better visualized with violin plots or histograms).** The pattern of the nested boxes can hint at these characteristics.

- **Provide a robust visualization that is less sensitive to outliers than simply plotting all data points.** The letter values focus on the central distribution.

In summary, boxen plots are a valuable alternative to standard box plots when you need a more nuanced understanding of the distribution's shape, especially the tails and the spread beyond the central quartiles. They are particularly helpful for comparing distributions with potential outliers or heavy tails across different categories.