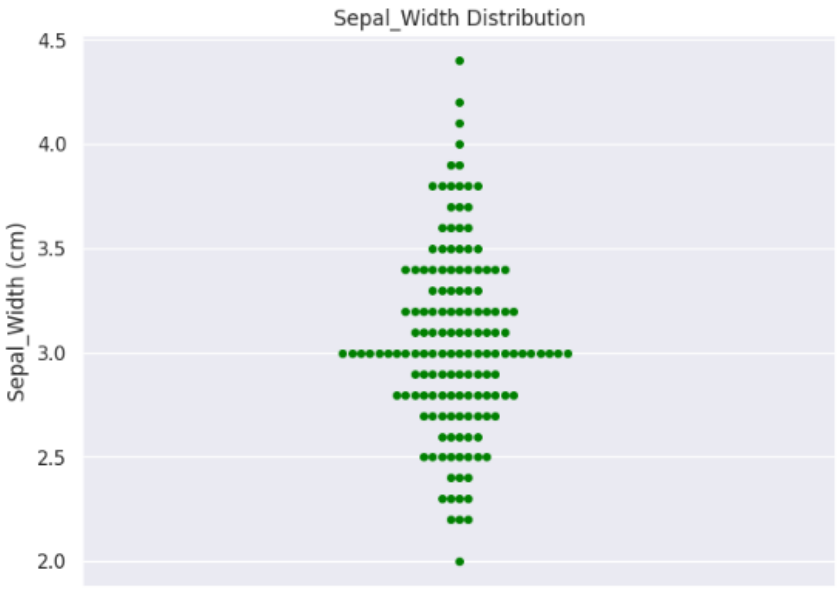
How to interpret Swarm plot?



A. Interpretation of the Swarm Plot Components:

- **Vertical Axis (Y-axis):** Represents the range of values for the numerical variable, "Sepal Width (cm)," spanning from approximately 2.0 to 4.5. Each dot's vertical position corresponds to its sepal width value.
- **Horizontal Axis (X-axis):** This axis represents the single category being analyzed (in this univariate case, there's no grouping). The dots are positioned horizontally in a way that avoids overlap, creating a shape that reflects the distribution's density at different sepal width values.
- **Dots (Individual Data Points):** Each dot represents a single observation in the dataset. Its vertical position indicates the value of "Sepal Width (cm)" for that observation, and its horizontal position is adjusted to prevent overlap with other points at similar values.

Interpreting the "Sepal Width (cm)" Distribution:

- **Central Tendency:** You can visually estimate the central tendency by observing where the widest part of the "swarm" of dots is located

along the y-axis. In this plot, the highest density of points appears to be around 3.0 cm.

- **Spread (Variability):** The vertical range covered by the dots indicates the overall spread or variability of the sepal width measurements. Here, the data ranges from approximately 2.0 cm to 4.4 cm.
- **Shape:** The shape formed by the arrangement of the dots provides a good visual representation of the distribution's density:
    - Wider "swarms" of dots at certain sepal width values indicate a higher frequency of those values. For example, there's a dense band of points around 3.0 cm.
    - Narrower "swarms" or fewer dots indicate lower frequencies. The density decreases as you move towards the lower and upper extremes.
    - The plot suggests a unimodal distribution with a peak around 3.0 cm. The distribution appears somewhat symmetrical around this peak.
- **Outliers:** Individual dots that are isolated and far from the main "swarm" can be identified as potential outliers. There are a few points at the lower end (around 2.0 - 2.2 cm) and a few at the higher end (around 4.0 - 4.4 cm) that are somewhat separated.
- **Density:** The horizontal spread of the dots at each y-value directly shows the density of data points at that specific sepal width.

Swarm plots are particularly useful for univariate analysis in the following scenarios:

- **Small to Moderately Sized Datasets Where Showing Individual Points and Density is Important:** Swarm plots excel at visualizing the distribution of individual data points while also providing a good sense of the data's density at different values, without the binning issues of histograms.
- **Comparing Distributions Across Categories (Bivariate/Multivariate):** Swarm plots are highly effective when used to compare the distribution of a numerical variable across different categories of a categorical variable. By plotting the swarms for each category side-

by-side (along the x-axis), you can easily see differences in central tendency, spread, and the detailed distribution of points. The non-overlapping nature makes it easier to compare the shape of the distributions.

- **Revealing Multiple Modes or Complex Distributions:** The way points arrange themselves in a swarm plot can sometimes reveal subtle patterns, multiple modes (peaks in density), or skewness that might be less apparent in a box plot or a simple strip plot with random jitter.
- **When Avoiding Overlap is Crucial:** Unlike a strip plot where many points at the same value can overlap, the point-placement algorithm in a swarm plot ensures that points are spread out to be visible, making it easier to see the number of observations at each value.
- **As a More Informative Alternative to Strip Plots:** Swarm plots address the main limitation of strip plots (overlapping points) while still showing individual data points. They provide a clearer picture of the distribution's shape and density.

In contrast to histograms, box plots, and violin plots:

- **Histograms:** Group data into bins, losing individual data point information and potentially obscuring the true shape depending on bin choice. Swarm plots show individual points and density directly.
- **Box Plots:** Provide a summary of the distribution but don't show the actual distribution of individual points or detailed shape. Swarm plots show the underlying data.
- **Violin Plots:** Show the estimated probability density, which can be excellent for shape, but might obscure individual points in denser areas. Swarm plots show individual points while still conveying density.

In summary, swarm plots are the best choice when you want to visualize the distribution of a numerical variable by showing individual data points without overlap, clearly displaying density, and effectively comparing distributions across categories, especially for small to moderately sized datasets where the detailed distribution and individual observations are important.