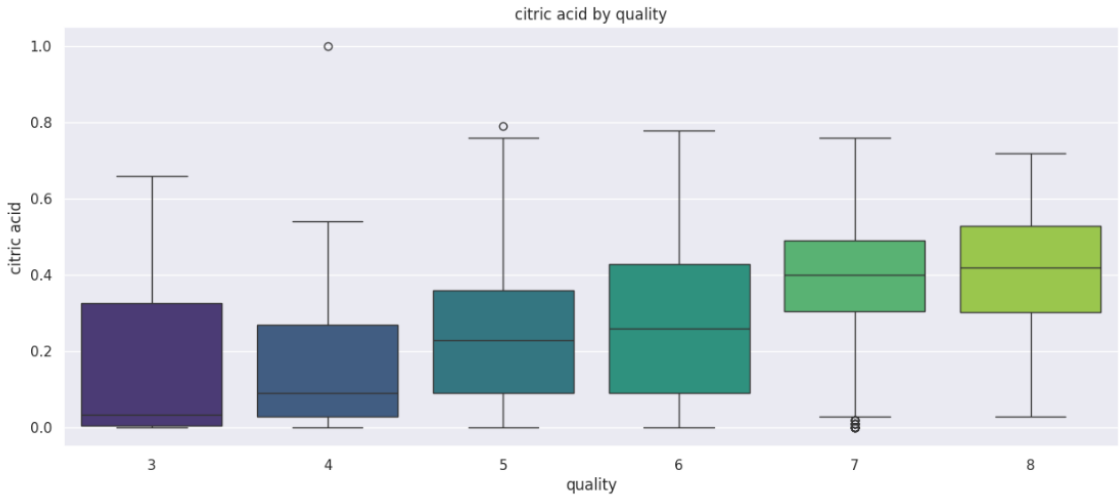
How to interpret Box plot for Bivariate analysis – numerical vs categorical variable



A. Understanding the Components of a Box Plot:

- **Categorical Axis (X-axis):** Represents the categories of the categorical variable, "quality," with levels ranging from 3 to 8.
- **Numerical Axis (Y-axis):** Represents the range of the numerical variable, "citric acid," from 0 to 1.
- **Boxes:** For each category of "quality," there is a box that summarizes the central tendency and spread of the "citric acid" values:
    - The **bottom edge** of the box represents the 25th percentile (Q1). 25% of the data points in that quality category have a "citric acid" value below this line.
    - The **top edge** of the box represents the 75th percentile (Q3). 75% of the data points in that quality category have a "citric acid" value below this line.
    - The **line inside the box** represents the median (Q2), which is the middle value of the data. 50% of the data points are below this line, and 50% are above.
    - The **height of the box** (from Q1 to Q3) represents the interquartile range (IQR), which contains the middle 50% of the data.
- **Whiskers:** Lines extending from the top and bottom of the box. They typically extend to 1.5 times the IQR from the edges of the box. Data points outside the whiskers are considered potential outliers.

- **Outliers:** Individual points plotted outside the whiskers. These are values that are unusually high or low compared to the rest of the data within that quality category. In this plot, we see a few outliers for qualities 3, 4, and 5.

B. Interpreting the Relationship Between Quality and Citric Acid:

By examining the box plots for each quality level, we can infer how the distribution of "citric acid" varies with "quality":

- **Quality 3:** Shows a lower median "citric acid" level compared to higher quality wines. The IQR is relatively small, suggesting less variability in "citric acid" for this quality. There's one notable high outlier.
- **Quality 4:** Has a slightly higher median "citric acid" than quality 3, but still lower than higher qualities. The IQR is also relatively small, and there's one high outlier.
- **Quality 5:** Shows a wider spread of "citric acid" values (larger IQR) and a slightly higher median than qualities 3 and 4. There's one high outlier.
- **Quality 6:** Has a median "citric acid" level similar to quality 5, but with a slightly smaller IQR.
- **Quality 7:** Shows a noticeable increase in the median "citric acid" level compared to lower qualities. The IQR is also relatively large, indicating greater variability. There are a few low outliers.
- **Quality 8:** Has the highest median "citric acid" level among all quality categories. The IQR is moderate.

C. Overall Interpretation:

There appears to be a general trend of increasing "citric acid" content with increasing wine "quality," as indicated by the upward shift in the median values. However, there is also considerable overlap in the distributions, and the variability within each quality level differs. Higher quality wines (7 and 8) tend to have a higher central tendency of "citric acid" compared to lower quality wines (3, 4, and 5).

Box plots are the best choice for visualizing the relationship between a numerical and a categorical variable when you want to:

- **Compare the distribution of a numerical variable across different categories.** They provide a concise summary of the central tendency (median), spread (IQR), and range of the data for each category.
- **Identify differences in the median and spread of the numerical variable across the categories.** You can easily see if the central values differ and if the variability is different between groups.
- **Detect outliers in each category.** Box plots clearly highlight data points that are unusually high or low within each group.

- **Get a quick visual summary of the key statistical properties of the numerical variable for each category.** This is more informative than just looking at means or standard deviations.
- **Compare multiple groups side-by-side in a compact format.** This makes it easy to see patterns and differences across all categories of the categorical variable.
- **Assess the skewness of the distribution within each category.** The position of the median within the box and the length of the whiskers can provide clues about the symmetry or skewness of the data.

In summary, box plots are excellent for comparing the distributions of a numerical variable across different groups defined by a categorical variable, highlighting central tendency, spread, and outliers in a visually efficient manner.