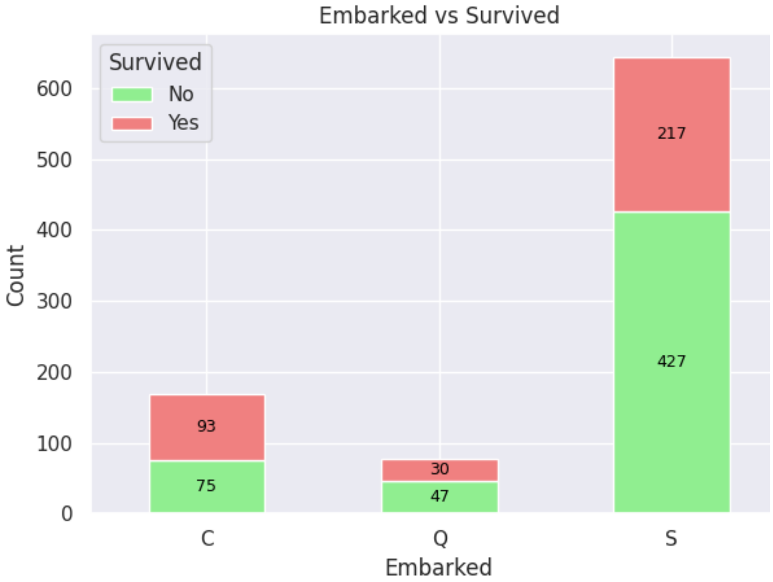
How to interpret stacked bar chart for bivariate analysis



A. Understanding the Components of a Stacked Bar Chart:

- **Horizontal Axis (X-axis):** Represents one of the categorical variables, in this case, "Embarked," with three categories: C, Q, and S.
- **Vertical Axis (Y-axis):** Represents the "Count" or frequency of passengers.
- **Stacked Bars:** For each category on the x-axis ("Embarked"), there is a single bar that is divided into segments. Each segment represents the count of the second categorical variable ("Survived") within that "Embarked" category.
  - **Light Green Segment:** Represents the count of passengers who did **not survive** (Survived = No) from that port of embarkation.
  - **Light Red Segment:** Represents the count of passengers who **survived** (Survived = Yes) from that port of embarkation.
- **Labels:** The numbers within each segment indicate the exact count for that combination of categories.

- **Legend:** The legend in the top-left corner clarifies which color corresponds to "No" (did not survive) and "Yes" (survived).

B. Interpreting the Relationship Between Embarked and Survived:

By examining the stacked bars for each port of embarkation, we can understand how survival rates varied across these ports:

- **Port C (Cherbourg):**
    - The light green segment (Did Not Survive) has a count of 75.
    - The light red segment (Survived) has a count of 93.
    - For passengers who embarked at Cherbourg, the number of survivors was higher than the number of those who did not survive.
- **Port Q (Queenstown):**
    - The light green segment (Did Not Survive) has a count of 47.
    - The light red segment (Survived) has a count of 30.
    - For passengers who embarked at Queenstown, the number of those who did not survive was higher than the number of survivors.
- **Port S (Southampton):**
    - The light green segment (Did Not Survive) has a count of 427.
    - The light red segment (Survived) has a count of 217.
    - For passengers who embarked at Southampton, the number of those who did not survive was significantly higher than the number of survivors.

C. Overall Interpretation:

The stacked bar chart clearly shows that the port of embarkation had an influence on the survival outcome. Passengers who embarked at Cherbourg (C) had the highest proportion of survivors relative to non-survivors among the three ports. Passengers who embarked at Queenstown (Q) had a lower proportion of survivors. Passengers who embarked at Southampton (S), which had the largest number of passengers overall, had the lowest proportion of survivors.

Stacked bar charts are the best choice for visualizing the relationship between two categorical variables when you want to:

- **Show the composition of each category of one variable based on the counts of the categories of the other variable.** In this case, for each port of embarkation, we see the breakdown of survivors and non-survivors.
- **Compare the absolute counts within each sub-category across the main categories.** We can directly compare the number of survivors (red segments) across the three ports.
- **Highlight the total count for each main category.** The total height of each bar represents the total number of passengers who embarked at each port.
- **Make it relatively easy to see which sub-categories are dominant within each main category.** For example, for Port S, the "Did Not Survive" segment is much larger than the "Survived" segment.
- **Present a clear visual representation of a contingency table.** The stacked bars effectively display the joint frequencies of the two categorical variables.

However, be mindful of a limitation: While stacked bar charts show the absolute counts well, it can be harder to directly compare the *proportions* of the sub-categories across different main categories, especially if the total counts for the main categories are very different. For comparing proportions, a **normalized stacked bar chart** (where each bar has the same total height representing 100%) might be more suitable.

In summary, the stacked bar chart is effective for showing the counts and composition of one categorical variable conditioned on another, providing insights into the relationship between them.