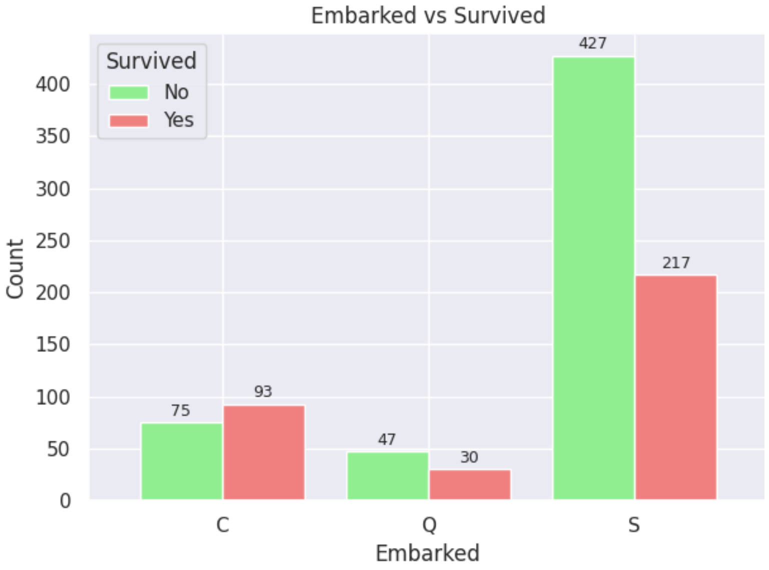
How to interpret clustered bar chart for bivariate analysis



A. Understanding the Components of a Clustered 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.
- **Clusters of Bars:** For each category on the x-axis ("Embarked"), there are multiple bars, each representing a category of the second categorical variable ("Survived"). [1]
    - **Light Green Bars:** Represent the count of passengers who did **not survive** (Survived = No) from that port of embarkation.
    - **Light Red Bars:** Represent the count of passengers who **survived** (Survived = Yes) from that port of embarkation.
- **Labels:** The numbers above each bar indicate the exact count for that specific 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 clusters of bars for each port of embarkation, we can understand how the number of survivors and non-survivors compares across these ports:

- **Port C (Cherbourg):**
    - The light green bar (Did Not Survive) has a count of 75.
    - The light red bar (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 bar (Did Not Survive) has a count of 47.
    - The light red bar (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 bar (Did Not Survive) has a count of 427.
    - The light red bar (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 clustered bar chart clearly illustrates the distribution of survival outcomes for passengers from each port of embarkation. It allows for a direct comparison of the number of survivors and non-survivors within each port. We can easily see that Cherbourg had the highest number of survivors relative to non-survivors, while Southampton had the lowest. Queenstown had a small number of passengers overall, with more not surviving than surviving.

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

- **Compare the counts (or proportions, if normalized) of the sub-categories of one variable directly across the different categories of the other variable.** In this case, we can easily compare the number of survivors (red bars) across the ports and the number of non-survivors (green bars) across the ports.
- **Show the absolute counts (or proportions) of each combination of the two categorical variables.** Each pair of bars represents a specific combination (e.g., Survived=Yes and Embarked=C).
- **Make it easy to see the differences in counts (or proportions) for each level of the second categorical variable within each level of the first categorical variable.** For each port, we can readily compare the height of the "Survived" bar to the height of the "Not Survived" bar.
- **Provide a clear visual representation of a contingency table, especially when the number of categories for each variable is not too large.**

In contrast to stacked bar charts:

- Clustered bar charts make it easier to compare the sizes of the sub-categories (e.g., the "Yes" bars across different "Embarked" categories) because they are placed side-by-side with a common baseline.
- Stacked bar charts are better at showing the total count for each main category.

In summary, the clustered bar chart is effective for comparing the counts of different categories of one variable across the categories of another variable, providing a clear view of the joint distribution of the two categorical variables.