perplexity

Step-by-Step Explanation: Titanic Data Cleaning & Analysis

Below is an easy-to-understand, detailed breakdown of the process we followed to clean and analyze the Titanic dataset, with clear explanations of what's happening at each stage.

1. Importing Libraries & Loading Data

- **Imports:** We loaded necessary Python libraries:
    - `pandas` and `numpy` for data handling and calculations.
    - `seaborn` and `matplotlib.pyplot` for making plots and graphs.
    - `os` for creating folders to save files.
- **Loading Data:** We loaded the Titanic data (`train.csv`) into a DataFrame (a table-like structure) using pandas.

2. Checking & Understanding the Data

- **First Look:** We printed the first few rows to see what the data looks like.
- **Info:** We checked data types (like numbers or text) and saw where values might be missing.
- **Summary Statistics:** We printed counts, averages, and ranges for each column.
- **Missing Data:** We counted how many values are missing in each column.

3. Data Cleaning

- **Filling Missing Values:**
    - *Age*: Some age values were missing, so we filled them with the median (the middle value).
    - *Embarked*: Sometimes passengers' port of embarkation was missing, so we filled it with the most common port.
    - *Cabin*: Too many 'Cabin' values were missing, so we removed this column altogether.
- **Changing Data Types:** We told Python that some columns are categories (like 'Sex' and 'Embarked'), which helps with analysis.
- **Feature Engineering:** We created a new column called `FamilySize` by adding up the number of siblings/spouses, the number of parents/children, plus 1 (the passenger themself). This helps see if traveling with family affected survival.

4. Exploratory Data Analysis (EDA)

4.1 Univariate Analysis (One Column at a Time)

- **Survival Count:** We counted how many survived and how many didn’t, then made a bar chart.
- **Age Distribution:** We showed how passenger ages are spread out using a histogram.

4.2 Bivariate Analysis (Comparing Two Columns)

- **Survival by Sex:** We compared survival rates for men and women using a bar chart.
- **Survival by Passenger Class:** We compared survival rates in 1st, 2nd, and 3rd class.
- **Age vs Survival:** We checked how age affected survival by making a violin plot (shows age distributions for survivors and non-survivors).

4.3 Correlation Heatmap

- We looked at how different numeric features (like age, fare, family size) relate to each other and to survival. This is shown in a heatmap, with brighter colors for stronger relationships.

5. Saving the Plots

- Instead of just showing the plots on the screen, we saved each graph as a PNG image in a folder called `png_outputs`. This is useful for sharing results or including them in reports.

6. Insights Discovered

- **Women and children** survived at higher rates than men.
- **First class passengers** survived more often than those in lower classes.
- **Younger passengers**, especially infants, tended to survive more.
- **Larger families** had different survival chances compared to solo travelers.
- **Paying higher fares** was linked to higher survival chances.

7. What You Can Do Next

- Try more detailed visualizations or calculations.
- Use this cleaned data to train machine learning models to predict survival.
- Share your findings with others by including the saved graphs in your presentations or reports.

In summary:
You loaded the data, fixed missing or messy values, explored the data using basic statistics and colorful graphs, and saved your visualizations for future use. This process is a foundation of all data science work and helps you get familiar with what's in your data before making any predictions or conclusions.

⁂