Titanic Data Analysis and Visualization Project

Group Project

Due: 2024-01-10

Project Objectives

1. **Data Exploration:**
   - Load the Titanic dataset (`titanic.csv`) using a library like pandas.
   - Explore the dataset to understand its structure and contents.
2. **Data Cleaning:**
   - Handle missing values (e.g., imputation or removal).
   - Check for duplicates and handle them if necessary.
   - Convert categorical variables into numerical format or vice versa if needed.
3. **Descriptive Statistics:**
   - Calculate basic statistical measures (mean, median, mode, standard deviation, etc.) for relevant columns.
   - Generate summary statistics for key variables like age, fare, etc.
4. **Data Visualization:**
   - Utilize `pandas`, `numpy`, and `matplotlib` for data visualization.
   - Suggestions:
     - Create visualizations to showcase insights, such as:
       * Survival rates based on different factors (class, gender, age group).
       * Age distribution of passengers and passenger classes.
       * Age distribution of passengers according to passenger classes survived and genders.
       * Number and percentages of passengers in each class with appropriate figures.
       * Number and percentages of passengers in each class with appropriate figures according to passengers survived.
       * Number and percentages of passengers in each class with appropriate figures according to passengers genders.
       * Embarkation points.

   Visualization Types Examples:

   - Include a bar chart, pie chart, histogram, boxplot, and other advanced appropriate plots comparing the survival rates by class.

   Visualization Tips:

   - Use color effectively to highlight key information.
   - Provide clear labels and titles for each visualization.
   - Consider using subplots to compare multiple visualizations.

5. **Teamwork:**
    - Assign specific tasks to each team member (e.g., one or two for data cleaning and statistical analysis, one for visualization, one for writing, and good presentation).
    - Encourage collaboration and sharing of findings within the team.
6. **Report Writing:**
    - Each team member contributes to the final report.
    - Include sections on data cleaning steps, statistical analysis, and insights from visualizations.
    - Provide clear and concise interpretations of the visualizations.
7. **Presentation:**
    - Schedule a presentation session where each team presents their findings.
    - Emphasize the importance of clear communication and visualization during the presentation.
8. **Competition:**
    - Assess the reports and presentations based on criteria such as clarity, depth of analysis, and creativity in visualization.
    - Consider providing a small prize or recognition for the winning team.

Resources

- Titanic dataset (`titanic.csv`)
- Python libraries: pandas, matplotlib, numpy, etc.

Evaluation Criteria

- Depth and accuracy of statistical analysis and visualizations.
- Creativity and clarity of visualizations.
- Quality of the final report and presentation.
- Correct, concise (efficient) and clear coding

Timeline

Week 1: Data Exploration and Cleaning

- Each team member should contribute to the code here.

Week 2: Statistical Analysis and Initial Visualization

- Each team member should contribute to the code here.

Week 3: Finalize Visualizations, Report Writing, and Presentations

- Each team member should contribute to the code here.

Deadline: January 10, 2024

Additional Tips

- Encourage regular check-ins and progress updates to ensure each team is on track.
- Provide resources or short tutorials on data cleaning, statistical analysis, and Matplotlib if needed.
- Emphasize the importance of collaboration and communication within the teams.
- Allocate some time for a practice presentation session before the final presentation.