Exploratory Data Analysis and Visualization Final Assessment

# Task Description

In this assessment, you will perform exploratory data analysis (EDA) on a real-world dataset and create visualizations to gain insights and answer specific questions about the data. You are required to complete the tasks outlined below using R programming language (Python acceptable) and submit your code along with the generated visualizations and a brief summary of your findings.

# Dataset

For this assessment, you will use the "Titanic" datasets, which contain information about passengers aboard the Titanic, including their demographic information, ticket details, cabin information, and survival status. There are two datasets for this analysis, one called “Titanic Survival.csv”, and one called “titanic\_info.csv” located in the data folder.

# Tasks

* Load the necessary libraries.
* Load two datasets into R/ Python (of your choice, but be consistent).
* Join two datasets on “name”, “sex”, and “age” columns.
* Perform initial data exploration by examining the dataset's structure, summary statistics, and missing values.
  + Provide one summary table of all variable
  + Provide one pivot table to show some valuable insights (e.g. sex ~ survived)
  + Report how many missing values are there in the data
* Clean the data by handling missing values and transforming variables if necessary.
  + Report how you cleaned and prepared the data set and the reason of your procedure.
  + Produce a cleaned dataset in csv format.
* Conduct exploratory data analysis to answer the following questions:
  + What was the survival rate of passengers on the Titanic?
  + How does the survival rate vary based on gender?
  + Are there any significant differences in survival rates across different passenger classes?
  + Did passengers' age have an impact on their chances of survival?
  + Are there any notable correlations between variables in the dataset?
    - Provide a correlation heatmap
* Create at least one appropriate visualization for each of the following:
  + bar plots, histograms, scatter plots

to support your findings and provide insights into the data.

* Perform one regression analysis with selected variables (linear or logistic regression), explain variable selection, interpret the model, and assess the model.
* Perform one clustering analysis, explain variable selection, interpret the model, and assess the model
* Write a brief summary of your findings and observations based on the EDA and visualizations.

# Submission Guidelines

* Write a R/Python script that includes your code for each task.
* Write a report with all summary tables, questions answered, visualizations and interpretations, modeling and model summaries, conclusion and insights. Either in slides, word doc or PDF is ok, but Apple Pages is not acceptable.
* Include comments in your code to explain your approach and any assumptions made.
* Save your code file, visualizations, and summary of findings in a single folder.
* Submit your folder as a compressed file (e.g., ZIP or RAR) to your instructor or evaluator.

# Grading Criteria

* Proper loading of the dataset and required libraries: 1 point
* Load data and perform data joining correctly: 4 point
* Proper data exploration, summary, and summary of missing values: 3 points
* Cleaned data set: 7 points
* Accurate answers to the questions and the correlation heatmap: 25 points (5 points for each question)
* Appropriate visualizations and interpretations: 15 points
* Regression analysis: 15 points
  + Variable selection (3 points)
  + Model summary (3 points)
  + Model interpretation (5 points)
  + Model assessment - is it a good model? And why. (4 points)
* Cluster analysis: 15 points
  + Variable selection (3 points)
  + Model summary (3 points)
  + Model interpretation (5 points)
  + Model assessment - is it a good model? And why. (4 points)
* Clear and concise summary of findings: 10 points
* Code readability and structure: 5 points
* Bonus (for any additional insights or creative visualizations such as the association rule that we covered): 10 points

Good luck with your assessment!