RSM 8101: Research Methods and Publications Test: Analyzing with R Software

Due date: Sunday, 28th March, 2024.
Method of handing in: R script uploaded to canvas

The test is out of 35, but there are 40 possible marks that you can get. The extra 5 marks are for installation and manipulation of R. Good luck!

Dataset: "Titanic: Machine Learning from Disaster"

Use R code and comments to answer the questions below
Dataset (only use train.csv): https://www.kaggle.com/c/titanic/data
Create a new R script file.

Answer each question in the R script, with your **explanation/answer** (as a comment), and show the **executable code** to accomplish the task your answer is based on.

1. Data Importation and Exploration. 3 marks

a. How can you import the Titanic dataset into R?
b. What are the key variables included in the Titanic dataset?
c. How would you display the structure and summary statistics of the Titanic dataset in R?

2. Data Cleaning and Preprocessing: 6 marks

a. Are there any missing values in the dataset? If so, how would you handle them using R?
b. How can you convert categorical variables, such as "Sex" and "Embarked", into factors in R?
c. Can you identify and remove any outliers in the dataset using R?
Hint: You may choose to analyze the "Fare" column, use a boxplot to show outliers

3. Exploratory Data Analysis (EDA): 9 marks

a. What is the distribution of passenger ages in the Titanic dataset, and how can you visualize it using R?
b. Is there a relationship between passenger class ("Pclass") and survival rate? How would you visualize this relationship in R?
c. Can you explore the survival rate based on gender ("Sex") and visualize it using R?

4. Is there a significant difference in survival rates between male and female passengers?
Hint: chi_square_test 7 marks

5. Visualization Techniques: 15 marks

a. How can you create a bar chart in R to compare the survival rates among different passenger classes?
b. What type of plot would you use in R to visualize the correlation between passenger age and fare paid?
c. Can you create a heatmap in R to visualize the correlation matrix of variables ("Pclass", "Age", "Fare") in the Titanic dataset? Hint: - You may need to handle missing, infinite or NaN values