Introduction to Data Science

Midterm Project - Summer 22-23

Project Description

Apply data preparation steps (which can be applied) and do the univariate data exploration for the given dataset. In this project, we are going to use a modified version of Titanic dataset ("Titanic – Modified.xlsx") which can be downloaded from the AIUB Portal. The original dataset can be found in the following link where the dataset description is available as well (you may need to log-in to download this dataset).

https://www.kaggle.com/datasets/ibrahimelsayed182/titanic-dataset?resource=download .

Project Deliverables

- Submit the implemented R program (R file or Text file) in the MS Teams. During the VIVA session, you will bring this implemented program and we may ask you to execute the program.
- Submit the report in the MS Teams. See the instruction section below for the report details. **Please bring the printed copy of the submitted report during the VIVA session.**

Instructions

- The submission deadline for all deliverables is **July 18, 2023** (you must submit the assignment before **10:00 AM**). There will be a penalty for the late submission.
- At the beginning of the report, write a short note about the dataset. You will get the dataset details from the above link provided for the dataset.
- For each implemented code segment in the R program, provide the code and its output along with their description in the report. In the description part, only write the content (do not write unnecessary content) that is sufficient to understand the code and its output.
- **Comments are not allowed in the R program.**
- The following topics can be focused to think about the project. **Note that the project is not limited to these topics where these are mentioned to get an idea about how to proceed with the project.**
    - In the data exploration step, we can check if there is any scope to apply the histogram and standard deviations to the dataset.
    - We can see if there is any scope to use the annotations to represent the data in a meaningful way.
    - If there exist any invalid data/outliers in the data set, then use appropriate approach to handle those values.
    - If there are any missing values in the dataset, then we will apply all applicable methods from the available options to handle the missing values.