Assignment 02

Data Wrangling and EDA

- **Total Marks: 4.0 %**
- This assignment must be completed in **teams of 2 or otherwise, individually**.
- Only one member should submit.
- The file names should follow the format as: **Name1(ERP1)_Name2(ERP2)**
- **Deadline:** Sunday, 6th April, 2024 @11.55pm

Objective

The primary objective of this assignment is to assess your proficiency in data wrangling, cleaning, and exploratory data analysis (EDA). This involves demonstrating your ability to effectively clean data, conduct thorough statistical analyses, and present your findings in a logical, professional manner. Additionally, a bonus goal is to develop a reusable API for data cleaning and EDA, which can serve as a valuable tool for future projects.

Plagiarism Note

In case of potential plagiarism in notebooks or analysis, both teams involved will receive a zero.

Grading Rubric

Strategy for catering inconsistencies/data entry errors and its execution	0.5 %
Strategy for catering missing values and its execution	1.5 %
Univariate Analysis (including Outlier Analysis)	1.0 %
Bivariate/Multi-variate Analysis and Statistical Testing	1.0 %
Total	**4.0 %**

Submission Requirements

- Only one member should submit.
- The file names should follow the format: **Name1(ERP1)_Name2(ERP2)**
- **Required Files:**
  a. Python Notebook (with all analysis– make use of *markdown* and *comments*).
  b. Submit the original dirty data file.
  c. Submit the clean data file.

If due to large file size, the LMS restricts the upload - then make sure to upload the python notebook on LMS and attach the data files (dirty and clean) as an online drive link.

Task

1. **Select a dataset** of your choice from here https://tinyurl.com/BI-Assignment02-Datasets
2. **Acquire some basic background knowledge** about dataset to understand its context and relevance (if needed).
3. **Data Inconsistencies**
   a. Develop a strategy to identify and correct data inconsistencies and data entry errors.
   b. Write your strategy in the notebook as markdowns.
   c. Provide brief interpretations of the corrections made.
   d. Majority marks are for this interpretation and for the strategy (not for Python code)
4. **Missing Values**
   a. Develop a strategy to handle missing values for each column separately.
   b. Write down your strategy in the notebook before you execute it.
   c. Ensure to make use of missingno library for missing value type identification and strategy building.
   d. Write down the interpretation of your results in your notebook as much as is possible. This should be brief, e.g., one sentence to describe a visual output, or 2-3 sentences summarizing a sequence of results etc.
   e. Majority marks are for this interpretation and for the strategy (not for Python code)
5. **Univariate Analysis**
   a. Histograms, boxplots, density plots of important numerical columns
   b. Frequency histograms of important categorical data
   c. Focus on outlier analysis, anomaly detection (if applicable)
   d. Provide brief interpretations of the findings, focusing on insights rather than code.

6. **Bivariate/Multi-variate Analysis**
   a. Perform analysis using various statistical tests as studied including ANOVA, T-test, Tukey, Chi-squared, and correlation heatmaps.
   b. Optionally, explore additional techniques like clustering or regression.
   c. Provide brief interpretations of the findings.

7. **Bonus Task [0.5%]: API Development (Optional) -** Develop a reusable API consisting of personalized functions for data cleaning and exploratory data analysis (EDA). This API should be capable of handling common data wrangling tasks and performing statistical analyses. If your API is deemed robust, useful, and well-structured, you may earn 0.5% bonus marks. Include a brief report explaining your API's structure, usability and capability to work with other datasets.

8. **Disclosure of AI Usage –** At the end of your notebook, using a markdown disclose where and how you made use of any AI tool. Your interpretations in the notebook should be in your own words. Make sure to be ready for a viva after the submission of the assignment.

At the end of the notebook, the checker should clearly understand the data through your statistical analysis and the cleaning process through your wrangling activities. Demonstrating a complete understanding of the data will increase your chances of receiving better marks.
Good luck!