Assignment 02

Data Wrangling and EDA

- **Total Marks: 4%**
- 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:** Monday, 18th March, 2024 @11.55pm

Objective

The main goal of this assignment is to gauge your data wrangling/cleaning and statistical analysis (EDA) skills, i.e., how much you can clean the data, how professionally you do it, and how much you understand the data through statistical analyses. So, do as much as you can but professionally and in a logical, controlled manner (not in a guns blazing manner). A by-product (bonus) goal is to allow you to develop your own API for data cleaning and EDA (which can be used later if needed).

Plagiarism Note

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

Grading Rubric

Strategy for catering missing values and its execution	1.5 %
Strategy for catering inconsistencies and data entry errors and its execution	0.5 %
Descriptive Statistics (Univariate)	1.0 %
Descriptive Statistics (Bivariate)	1.0 %
Total	**4%**

Task

1. **Select a dataset** of your choice from the uploaded list. (see end of document)
2. **Acquire some basic background knowledge** about the data (just to get an idea about the data, if needed)
3. **Develop a strategy for catering for missing values for each column separately.**
    a. Write down your strategy in the notebook for us to understand it before you execute it.
    b. Pen down the interpretation of your results as much as is possible. This should be minimal, e.g., one sentence to describe a visual output, or 2-3 sentences summarizing a sequence of results etc. Don’t forget the use of missingno module
    c. Majority marks are for this interpretation and for the strategy (not for Python code)

4. **Develop a strategy for catering inconsistencies and data entry errors and its execution.**
    a. Write down your strategy in the notebook
    b. Pen down the interpretation as much as is possible. Should be minimal (as above).
    c. Majority marks are for this interpretation and for the strategy (not for Python code)

5. **Execute univariate descriptive statistics**
    a. Histograms, boxplots, density plots of important numerical columns
    b. Frequency histograms of important categorical data
    c. Primary marks are for the interpretation (keep it brief)
    d. Focus on outlier analysis, anomaly detection (if applicable)

6. **Execute bivariate statistics**
    a. ANOVA and Chi-squared (as applicable)
    b. Correlation heatmaps
    c. You can also try out extra items like cluster analysis / regression etc. for your own learning.

7. **Bonus marks [0.5%]:** Development of an API. If we feel that you have developed a list of personalized data cleaning and statistical analysis functions i.e. an API, you can gain a bonus of 0.5% marks (depending on how robust or useful your API seems).

8. Note that, at the end of the notebook, the checker should have a clear idea about the data (through your statistical analysis) and how you cleaned it up (through wrangling activities). If the checker feels you have completely understood the data, then there are higher chances of getting better marks. Good Luck!

Submission Requirements

- Only one member should submit.
- The file names should follow the format: **Name1(ERP1)_Name2(ERP2)**

Required Files:

1. Submit your Notebook (all analysis within as specified above – make use of *markdown* and *comments*).
2. Submit the original dirty data file.
3. Submit the clean data file.

List of Datasets

Link to the datasets:

http://tinyurl.com/BI-EDA-Datasets