Assignment 1 Mini Project

Understanding the data and Descriptive Analytics

Due date: 30.06.2023
[This assignment must be completed as a team of 4]
CORE ASSESSMENT
Worth: 8%

Summary

The main objective of this assignment is to understand the variables chosen from the datasets and their connections with each other. Your task in identify, collect and describe the **data of your choice** and perform various descriptive statistical analysis. The challenging part of this assessment is to finding an appropriate dataset to use. In addition, while you are working on the assignment, you are required to think carefully about the observational units (i.e. Rows) in the chosen datasets, they must be independent.

REPORT

The report should include:

1. Where the data is acquired from? (i.e. source of your data) and detail description of all the relevant variables. Identify and brief the observation unit (i.e. row) in the datasets.
2. Descriptive statistical analysis: Summarize the statistics with adequate explanations/ interpretation based on your learning in the lecture sessions. You can use Python or R programming language to perform statistics analysis. There is no requirement to provide definition of the statistics, but you must indicate what statistics tell about your chosen data. Examine the binary or other relationships as well as describing individual distributions. Prepare an appropriate table for summary. You are encouraged to use python packages which has functions to summarize data succinctly.
3. Data Visualization: Visualize the results with graphical display with brief explanation /interpretation. This task is to effectively summarize the data and can show the interesting features in the datasets. There is no requirement to visualize or picturize or tabulate for every variable. You are encouraged to report the checking data with the word like symmetric or bell shaped, but not “normal”.
4. Report any interesting observations or comments related to the chosen dataset for the project. Kindly report a) Is the data approximately what is been expected or did some of the results surprise your team? How did the sampling is conducted? If you think team got a representative sample of your population, brief it in the report.

Data Constraints

1. Your dataset must include at least 10 variables, with at least 4 independent quantitative and at least 4 independent categorical variables. (i.e. Label/ID/name does not count as a

categorical variable because can't summarize it or use it in any way.) Please make sure to use full variable names / descriptions in your sentences on the report and make your abbreviation clear to your evaluators.

2. You should have at least 100 independent cases / observations (ideal number of observations is 200-400). [Be wary of missing observations.

3. In your datasets, if it has time, be very careful with time (year) as a variable because it can be an indication that your observational units are not independent. The reason is we will be doing hypothesis testing as the next step, you need to indicate what population your data describes. If it is a census, then maybe it is representative of an even larger population? Discuss in detail the limitations of describing a larger population.

4. Idea of Data sources: Kaggle; or any open source publicly available datasets.

Report format

Report Structure

- Follow the paper template
- Cite and acknowledge sources correctly
- Make your Report to save in Word or PDF.

Presentation

- Prepare 5-7 slides which includes data set description, observation/ interpretations and visualization
- Presentation will be in final week of semester and marks will be evaluated based on the presentation by the team

Notes

- Do not print any warning or error messages. Only print code that is interesting and relevant
- to the reader
- Do not print lists of data.
- Be careful of overplotting. Use boxplots instead of scatterplots when appropriate. Use alpha=0.1 for transparent plotting symbols.
- No **linear models** for this assignment. No hypothesis tests or inference of any kind is expected. If you are curious about relationships in your data, it is possible you could run a t-test or a chi-squared test.
- Do not be tempted to turn in everything you do. Only turn in the interesting parts of the analysis. One of the hardest parts of being a consultant is figuring out what to tell the researcher

Marking Scheme & Submission

You must submit your Softcopy on the Google Class Room. Create a single zip file for your code and a working version of your program. Hard Copy of the report must be submitted at M249. You should name your source code and the zipped file as XXXXX_Ass1; where XXXXX is your team name.

The assignment is a one semester-long project, so your team would be the same for the entire semester. Team must have learning and communication benefits, The task for the assignment is to set up and use a GitHub repository. If you work in team, you must do so via a GitHub repository where the workload is shared and documented. In your assignment, please provide the name and location of the GitHub repository.

Marking Scheme

REQUIREMENTS	**MARKS**
Chosen data sets Descriptive statistical analysis Data visualization	20
Observations/Interpretations	40
Result evidences in programming	30
Paper and Presentation	20
INDIVIDUAL CONTRIBUTION	10
TOTAL	**30**