CMPS 460 Machine Learning – Spring 2024

Assignment #1 - Analysis of a dataset

Assigned: 08/02/2024 Due: **18/02/2024**

The purpose of this assignment is to practice using Python's packages Panadas, Matplotlib, and Seaborn to perform data analysis and data visualization.

Find an interesting dataset on https://www.data.gov.qa/ then analyze your selected dataset using **statistical** measures and relevant **visualization** charts.

Your analysis should include:

- Data cleaning: filtering out bad data, filling in missing values, reformatting the data, *etc*., as necessary.
- Compute measures of central tendency and measures of variability.
- Explore correlation between features.
- Exploratory uni/bi/multivariate data analysis.

Do not calculate a measure simply because you can. In your notebook, explain why you did the calculation and what insight you gained from the result. Your score will be based on a good choice of measures/graphs then deriving useful observations/insights.

Grading Rubric

Submit your work as **Jupyter Notebook** including a csv file of your dataset. Be sure to include useful markdowns and comments in your notebook to address the requirements listed in the table below.

Your submission will be scored according to these criteria:

Deliverable	Score
What is your dataset and why do you think it's interesting? Include a link to the dataset you have used.	5
List what data cleanup was required. Then perform the needed data cleaning.	5
Univariate analysis: measures of central tendency, measures of variability. Include interpretations and what insights did you discover from your analysis.	10
Univariate analysis using suitable charts (at least 3 charts). Include interpretations and what insights did you discover from your analysis.	40
Bi/multivariate analysis including computing correlations and suitable charts (at least 3 charts). Include interpretations and what insights did you discover from your analysis.	40