CDS 1020-1: Introduction to Computational Data Science
Homework 1

To understand a dataset, it is necessary to perform basic operations of data characterization and manipulation. With Python we can call numerous functions to understand our data prior to analysis. Here you are expected to employ exploratory analysis techniques in Python using libraries such as Pandas and Matplotlib.

Assigned: Sunday, January 29, 2023, 11:59pm
Deadline: Sunday, February 5th, 2023, 11:59pm

20 points

Part I: Prepare Your Python Environment

1. Install a Python IDE (e.g., JupyterLab, Spyder, PyCharm) on your personal machine. This can be done by downloading the distribution called Anaconda on your machine. Anaconda is an open-source Python and R distribution software, offering the aforementioned Python libraries or IDEs. Once installed, you can access the individual IDEs for computing.

Part II: Python Review

2. What is the difference between a list, tuple and a dictionary?
3. Demonstrate the immutability of a tuple (i.e., generate an error statement and take a screenshot of your entire interface, including the name of your Python file).
4. Define a function with a relevant name that performs two mathematical operations using two parameters.

Python III: Conduct Exploratory Analysis on the Pima Dataset

The Pima Indians Diabetes dataset originates from the National Institute of Diabetes and Digestive and Kidney Diseases. This dataset consists of several features including number of pregnancies and BMI. Please follow the steps below to explore the dataset.

5. Load the dataset into your Python environment as a Pandas data frame.
6. Retrieve information about the dataset (dimensions like number of rows and columns, row names and column names). Describe the features of the dataset: are they categorical or numerical?
7. Compute summary statistics for a feature of your choice (e.g., mean, median, mode, standard deviation, minimum, maximum, range). Are there any outliers?
8. Calculate the median BMI for the age range 20-50 years old (closed interval) by generating a box plot.

9. Generate a histogram to visualize the distribution of pregnancies.
10. Generate a scatter plot of blood pressure against BMI. Is there a relationship based on observation? Compute the Pearson's correlation coefficient. Is there a relationship based on this metric?

Submission:
Submit a standard Python file (.py) with your code using the name "[last name]_HW1" to Canvas. For visualizations and text answers, please use a document (.pdf) with appropriate numbering. The submission to Canvas will consist of these two files.