Puja Roy 4/8/22

CET 4900 - OL60

Internship Journal Entry#11

Throughout this week of my internship, I continued analyzing the Astoria dataset collected from the New York Urban Hydro-meteorological Testbed (NY-uHMT) weather station. Since I worked on writing a Python script that removes the outliers in the data, the data became simplified. Likewise, the Python pandas profiling report assisted me in generating a report based on crucial statistics on the dataset. The Python pandas profiling report stated that there are some missing values in the Astoria dataset. For this reason, I conducted research on how to handle missing data by replacing mean and median values for certain columns in the dataset.

I imported the necessary python libraries in Jupyter notebook and loaded the dataset by assigning it to the most common variable for storing dataframes: df. Then, I wrote used the df.replace() method to replace NaN values with 0 instead. It is also necessary to indicate the values that you want to replace by adding them in brackets inside parentheses of the df dataframe. On the other hand, if you want to replace values of a specific column, you must add curly braces instead of the square brackets. From there on, you must specify what values you want to change. Most importantly, data cleaning is an important approach when it comes to analyzing and filtering datasets. In other terms, when a dataset is preprocessed, it is essential to drop duplicates and remove NaN values by either dropping them or replacing them.

Data science can be fun and creative to work around with, but it takes time to get used to what the dataset is primarily about. To add on, memorizing most of the Python methods can be difficult which is why research is helpful for allowing me to understand the syntax of the Python programming language efficiently.