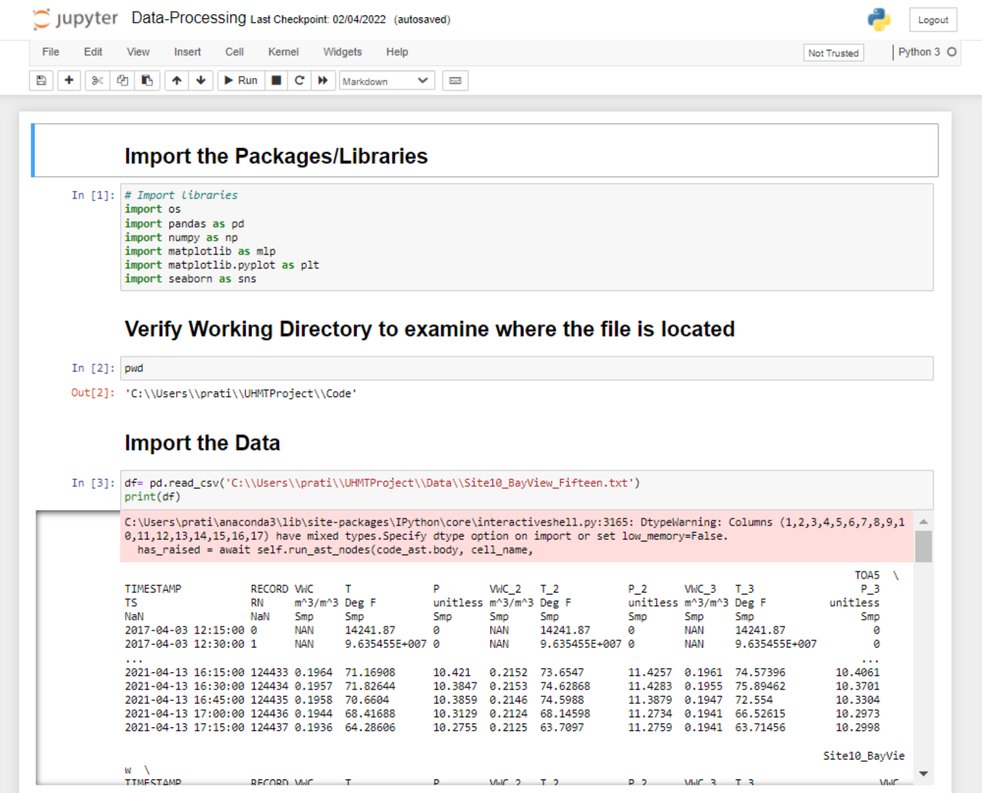
Puja Roy 2/18/22

CET 4900 - OL60

Internship Journal Entry #4

Throughout this week of my internship, I worked on analyzing large-scale datasets of real-time data collected from the New York Urban Hydro-Meteorological Testbed (NY-uHMT) weather station. Since there is 29+ large-scale datasets, it will take excessive time to analyze the data. The main objective of analyzing the large-scale datasets is to develop a Python script in Jupyter Notebook/Google Colaboratory to build a Weather App that automates the retrieval of weather stations and forecasts climate changes/natural hazards in NYC maps utilizing the New York Urban Hydro-Meteorological Testbed (NY-uHMT) data. Since data is not always organized and all over the place in Excel files/text files, it is necessary to clean, filter, examine and analyze the data to draw conclusions. My knowledge and skills of Python are being leveraged throughout my internship because I took EMT 1111 – Logic and Problem Solving during my freshman year of college. I gained exposure in learning how to program in Python and how it is mostly used for back-end programming. Most importantly, I am applying data science skills to my internship project because I am preparing the NY-uHMT data for analysis including cleansing, aggregating, and manipulating the data to visualize correlations and trends among various variables.

I downloaded all the data txt files that were collected by the NY-uHMT weather station. I used Python in Jupyter Notebook to import the data. For data analysis, I imported Python packages including pandas, numpy, matplotlib and seaborn. These packages are essential for reading and visualizing data files. Then, I printed the pwd command to verify the working directory in where the data files are located. Once the data was imported, I printed the dataframe which is declared df as a variable. The data for Bay View consists of 89,211 rows and 8 columns.



Import the Packages/Libraries

```
In [1]: # Import libraries
        import os
        import pandas as pd
        import numpy as np
        import matplotlib as mlp
        import matplotlib.pyplot as plt
        import seaborn as sns
```

Verify Working Directory to examine where the file is located

```
In [2]: pwd
Out[2]: 'C:\\Users\\prati\\UHMTProject\\Code'
```

Import the Data

```
In [3]: df= pd.read_csv('C:\\Users\\prati\\UHMTProject\\Data\\Site10_BayView_Fifteen.txt')
        print(df)
```

```
C:\Users\prati\anaconda3\lib\site-packages\IPython\core\interactiveshell.py:3165: DtypeWarning: Columns (1,2,3,4,5,6,7,8,9,1
0,11,12,13,14,15,16,17) have mixed types.Specify dtype option on import or set low_memory=False.
  has_raised = await self.run_ast_nodes(code_ast.body, cell_name,

                                                                                            TOA5  \
TIMESTAMP           RECORD VWC     T             P        VWC_2   T_2           P_2      VWC_3   T_3             P_3
TS                  RN     m^3/m^3 Deg F         unitless m^3/m^3 Deg F         unitless m^3/m^3 Deg F      unitless
NaN                 NaN    Smp     Smp           Smp      Smp     Smp           Smp      Smp     Smp             Smp
2017-04-03 12:15:00 0      NAN     14241.87      0        NAN     14241.87      0        NAN     14241.87          0
2017-04-03 12:30:00 1      NAN     9.635455E+007 0        NAN     9.635455E+007 0        NAN     9.635455E+007     0
...                                                                                                              ...
2021-04-13 16:15:00 124433 0.1964  71.16908      10.421   0.2152  73.6547       11.4257  0.1961  74.57396    10.4061
2021-04-13 16:30:00 124434 0.1957  71.82644      10.3847  0.2153  74.62868      11.4283  0.1955  75.89462    10.3701
2021-04-13 16:45:00 124435 0.1958  70.6604       10.3859  0.2146  74.5988       11.3879  0.1947  72.554      10.3304
2021-04-13 17:00:00 124436 0.1944  68.41688      10.3129  0.2124  68.14598      11.2734  0.1941  66.52615    10.2973
2021-04-13 17:15:00 124437 0.1936  64.28606      10.2755  0.2125  63.7097       11.2759  0.1941  63.71456    10.2998

                                                                                            Site10_BayVie
w  \
TIMESTAMP           RECORD VWC     T             P        VWC_2   T_2           P_2      VWC_3   T_3                VWC
```

Figure 1 – Python Script in Jupyter Notebook preparing for data analysis by importing Python packages and NY-uHMT data files