

Puja Roy

2/25/22

CET 4900 - OL60

Internship Journal Entry #5

Throughout this week of my internship, I continued working on analyzing the Bay View site dataset of real-time data collected from the New York Urban Hydro-Meteorological Testbed (NY-uHMT) weather station. Since the dataset contains rich content, I wrote Python pandas methods and functions in Jupyter Notebook to examine the data efficiently. I used the `DataFrame.head()` pandas method to return the top 5 default rows of the data frame. As shown below in Fig 2, I observed that there are columns for recording values of air temperature Fahrenheit, relative humidity, soil moisture and rainfall. Throughout my research, I had to examine the values, columns and other relevant information by utilizing other pandas functions and methods. The `df.info()` method prints information about the data frame. Similarly, the `df.describe()` method returns description of the data frame.

As you can see in Fig 1, there is a lot of information regarding climate change in Astoria, Queens. Analyzing the uHMT datasets will assist in predicting the weather and natural hazards such as droughts and storms which are becoming frequent and severe due to climate change. Since there are other variables and information to examine, it will take time to process the data using Python for data analysis and visualization. This data will eventually help me examine and draw conclusions based on the data.

```
In [4]: df.head()
```

```
Out[4]:
```

		TOA5		Site10_BayView		CR200X	
TIMESTAMP	RECORD	VWC	T	P	VWC_2	T_2	P_2
TS	RN	m^3/m^3	Deg F	unitless	m^3/m^3	Deg F	unitless
NaN	NaN	Smp	Smp	Smp	Smp	Smp	Smp
2017-04-03 12:15:00	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

```
In [5]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
MultiIndex: 89211 entries, ('TIMESTAMP', 'RECORD', 'VWC', 'T', 'P', 'VWC_2', 'T_2', 'P_2', 'VWC_3', 'T_3') to ('2021-04-13 17:15:00', 124437, 0.1936, 64.28606, 10.2755, 0.2125, 63.7097, 11.2759, 0.1941, 63.71456)
Data columns (total 8 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   TOA5                                  89211 non-null  object
1   Site10_BayView                        89211 non-null  object
2   CR200X                                89211 non-null  object
3   25223                                  89211 non-null  object
4   CR200X.Std.04                         89210 non-null  object
5   CR200X_12_25_rev2.CR2                89211 non-null  object
6   19761                                  89210 non-null  object
7   Fifteen                               89210 non-null  object
dtypes: object(8)
memory
```

Figure 1 – Examining the data frame using Python pandas functions and methods

In [5]: `df.info()`

```
<class 'pandas.core.frame.DataFrame'>
MultiIndex: 89211 entries, ('TIMESTAMP', 'RECORD', 'VWC', 'T', 'P', 'VWC_2', 'T_2', 'P_2', 'VWC_3', 'T_3') to ('2021-04-13 17:15:00', 124437, 0.1936, 64.28606, 10.2755, 0.2125, 63.7097, 11.2759, 0.1941, 63.71456)
Data columns (total 8 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   TOA5                                  89211 non-null  object
1   Site10_BayView                       89211 non-null  object
2   CR200X                               89211 non-null  object
3   25223                                89211 non-null  object
4   CR200X.Std.04                        89210 non-null  object
5   CR200X_12_25_rev2.CR2               89211 non-null  object
6   19761                                89210 non-null  object
7   Fifteen                              89210 non-null  object
dtypes: object(8)
memory usage: 20.6+ MB
```

In [6]: `df.columns`

Out[6]: Index(['TOA5', 'Site10_BayView', 'CR200X', '25223', 'CR200X.Std.04', 'CR200X_12_25_rev2.CR2', '19761', 'Fifteen'], dtype='object')

In [7]: `df.values`

Out[7]: array([['P_3', 'VWC_4', 'T_4', ..., 'RH', 'Rainfall_Tot', 'Snowfall_Tot'],
['unitless', 'm^3/m^3', 'Deg F', ..., '%', nan, nan],
['Smp', 'Smp', 'Smp', ..., 'Smp', 'Tot', 'Tot'],
...,
[10.3304, 0.2185, 66.81164, ..., 41.4637, 0.0, 0.0],
[10.2973, 0.2164, 63.62582, ..., 40.29307, 0.0, 0.0],
[10.2998, 0.2171, 62.0375, ..., 39.72377, 0.0, 0.0]], dtype=object)

In [8]: `df.describe()`

Out[8]:

	TOA5	Site10_BayView	CR200X	25223	CR200X.Std.04	CR200X_12_25_rev2.CR2	19761	Fifteen
count	89211	89211	89211	89211	89210.00000	89211	89210.0	89210.0
unique	61738	5119	20153	53794	33932.00000	51349	58.0	15.0
top	0	NAN	INF	0	43.64291	100	0.0	0.0
freq	1231	1231	1222	1231	19.00000	422	54316.0	56246.0

Figure 2 – Examining the information and values in the data frame