Puja Roy 3/11/22

CET 4900 - OL60

Internship Journal Entry #7

Throughout this week of my internship, I continued working on analyzing and visualizing the Bay View site dataset of real-time data collected from the New York Urban Hydro-Meteorological Testbed (NY-uHMT) weather station. The dataset is very large and messy with various elements of data scattered. While I was analyzing the Bay View site data, I noticed that there were many outliers. Outliers are data points that range from values further from normal observations of the data. In simple terms, it is when a value is plotted outside and is smaller or larger than most values in the dataset. Usually, outliers occur because of:

I. Experimental measurement errors
II. Sampling problems
III. Natural Variations

I noticed outliers in the Bay View dataset while I was using Python pandas to visualize the data. I created time series graphs of Bay View's Air temperature which is the BAY_AirTF column. As shown below in Fig 1-2, I used Python Matplotlib to create the time series graphs.

```
In [19]: df.BAY_AirTF.plot(figsize=(20,5), title= "BAY_AirTF")
         plt.xlabel("TIMESTAMP")
         plt.ylabel("Air Temperature Fahrenheit")
         plt.show()
```

Figure 1 – Python Code for time series analysis of Bay View's air temperature

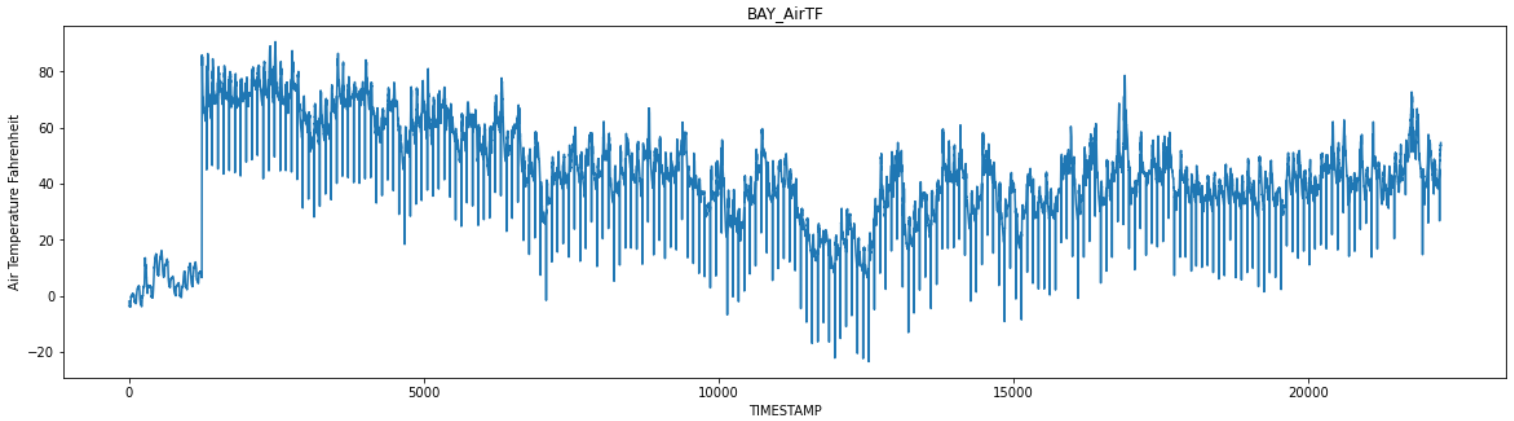

Figure 2 – Time series graph of Bay View's air temperature displaying outliers (light blue lines)

The time series graphs display extra blue light lines that range further from the plotted values shaded in deep blue. I analyzed that the outliers in this dataset were caused primarily by experimentation errors that was collected by the New York Urban Hydrometeorological Testbed Array weather station. Shown below in Fig 3-4, I also created a time series graph of Bay View's relative humidity, air temperature and total rainfall to identify any significant trends.

```
In [26]: df.BAY_RH.plot(figsize=(20,5), title= "BAY_RH") #(Site name_Relative Humidity)
         df.BAY_AirTF.plot(figsize=(20,5), title= "BAY_AirTF") #Air Temperature Fahrenheit
         df.BAY_Rainfall_Tot.plot(figsize=(20,5), title= "BAY_Rainfall_Tot") # Site view Rainfall Total
         plt.title("uHMT Weather Data of BAY RH, AirTF & Rainfall_Tot")
         plt.xlabel("TIMESTAMP")
         plt.ylabel("Relative Humidity, Temperature & Rainfall Total")
         plt.show()
```

Figure 3 – Python code for time series analysis of Bay View's air temperature, relative humidity, and total rainfall

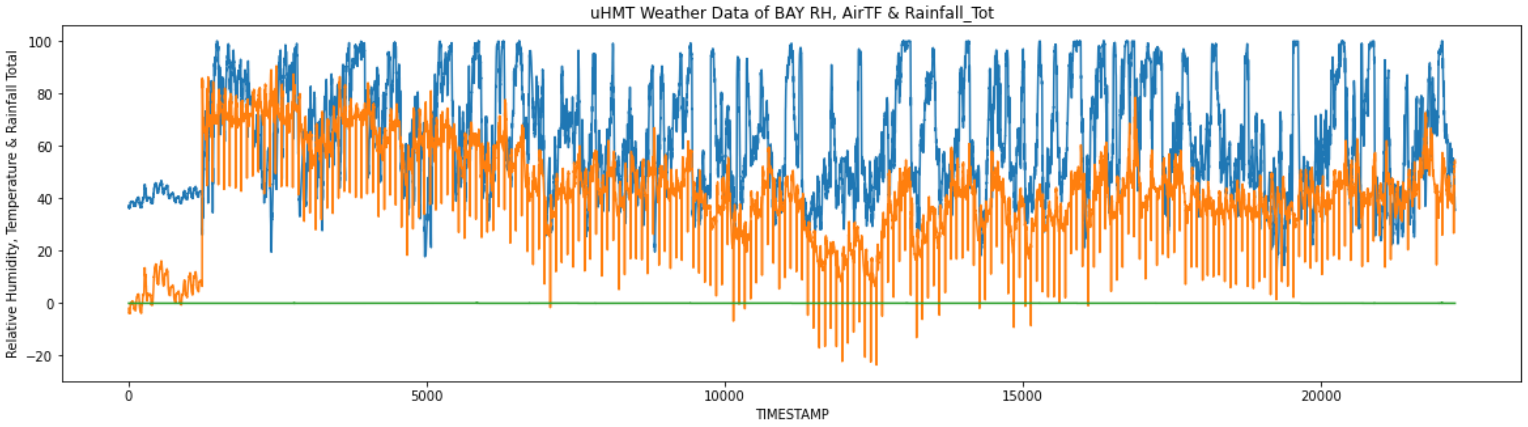

Figure 4 – Time series graph of Bay View's air temperature, relative humidity, and total rainfall displaying outliers

Bay view's relative humidity did not display any signs of outliers and the total rainfall is constant. To handle outliers in a dataset, it is necessary to remove them. I conducted research on the best techniques to detect and handle outliers. In addition, I analyzed the Astoria site dataset and noticed additional outliers in the dataset. My research proved that there were mostly outliers in the Air temperature columns than the rest of the columns. I was curious to observe the values of the outliers, so I imported the Astoria dataset in Excel to examine which parts of the data the outliers were located in. Fig 5 below shows the data in which outliers are located and replacing the values.

AST_date_time_1	AST_date_time_2	AST_AirTF	AST_RH	AST_Rainfall_Tot	AST_VWC1	AST_VWC2	AST_VWC3	AST_VWC4	AVERAGE OF AIR TEMP		
NaT	NaT	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!			
NaT	NaT	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!			
22-Jan-2018	15:45:00	41.69786	81.72929	0	0	0	0	0			
22-Jan-2018	16:00:00	41.70884	81.87734	0	0	0	0	0			
22-Jan-2018	16:15:00	40.21433	82.8053	0	0	0	0	0			
22-Jan-2018	**16:30:00**	39.67313	83.0495	0	0	0	0	0		39.67313	83.0495
22-Jan-2018	**16:45:00**	39.14016	85.54492	0	0	0	0	0		39.14016	85.54492
22-Jan-2018	**17:00:00**	38.89566	87.0834	0	0	0	0	0	**38.91763**	38.89566	87.0834
22-Jan-2018	**17:15:00**	35.86268	88.22808	0	0	0	0	0		35.86268	88.22808
22-Jan-2018	**17:30:00**	41.01653	91.6942	0	0	0	0	0		41.01653	91.6942
22-Jan-2018	17:45:00	40.56599	91.99944	0	0	0	0	0		**38.91763**	**87.12002**
22-Jan-2018	18:00:00	40.17038	92.40849	0	0	0	0	0			
22-Jan-2018	18:15:00	39.97533	92.25586	0	0	0	0	0			
22-Jan-2018	**18:30:00**	40.12643	91.61483	0	0	0	0	0		40.12643	91.61483
22-Jan-2018	**18:45:00**	39.62093	90.45641	0	0	0	0	0		39.62093	90.45641
22-Jan-2018	**19:00:00**	39.22532	89.34377	0	0	0	0	0	**39.49346**	39.22532	89.34377
22-Jan-2018	**19:15:00**	39.22532	89.33156	0	0	0	0	0		39.22532	89.33156
22-Jan-2018	**19:30:00**	39.26928	89.0141	0	0	0	0	0		39.26928	89.0141
22-Jan-2018	19:45:00	38.96159	89.13009	0	0	0	0	0		**39.49346**	**89.95213**
22-Jan-2018	20:00:00	38.74455	89.18504	0	0	0	0	0			
22-Jan-2018	20:15:00	38.84895	91.12643	0	0	0	0	0			
22-Jan-2018	**20:30:00**	37.96159	93.06935	0	0	0	0	0		37.96159	93.06935
22-Jan-2018	**20:45:00**	37.67587	95.62429	0	0	0	0	0		37.67587	95.62429
22-Jan-2018	**21:00:00**	37.68687	96.51105	0	0	0	0	0	**37.74565**	37.68687	96.51105
22-Jan-2018	**21:15:00**	37.70609	96.59042	0	0	0	0	0		37.70609	96.59042
22-Jan-2018	**21:30:00**	37.69785	97.10934	0	0	0	0	0		37.69785	97.10934
22-Jan-2018	21:45:00	37.53301	98.02967	0	0	0	0	0		**37.74565**	**95.78089**

Astoria | Bay_View | Sheet1

Figure 1 – Astoria Dataset in Excel being calculated to replace outlier values with the average air temp values