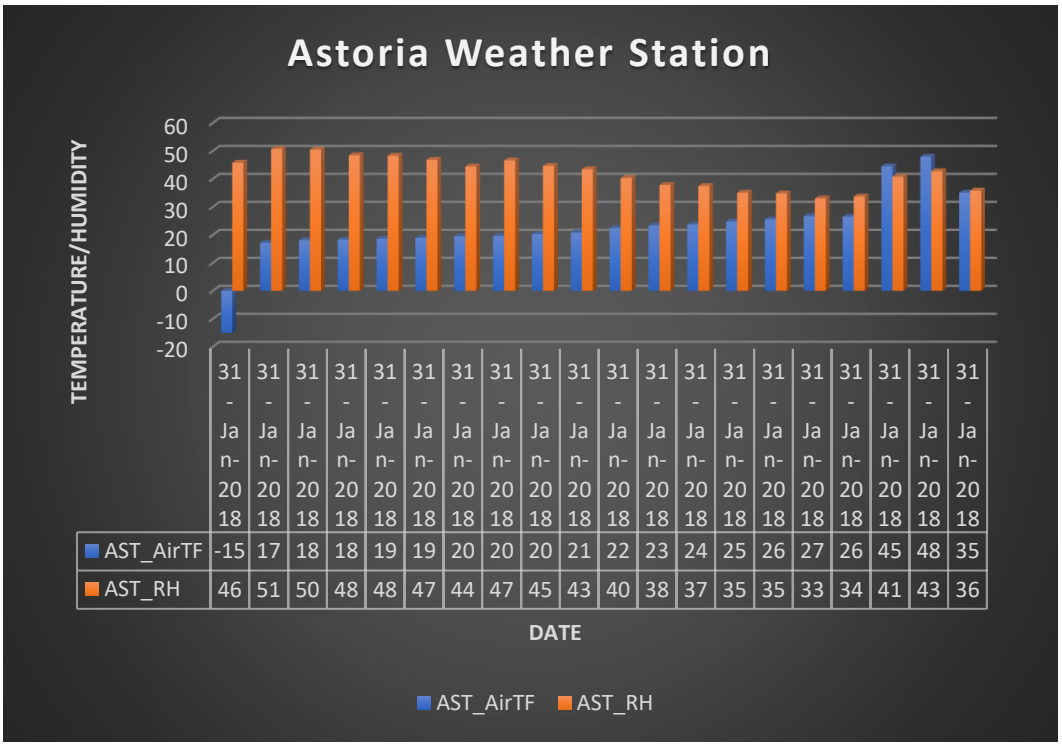
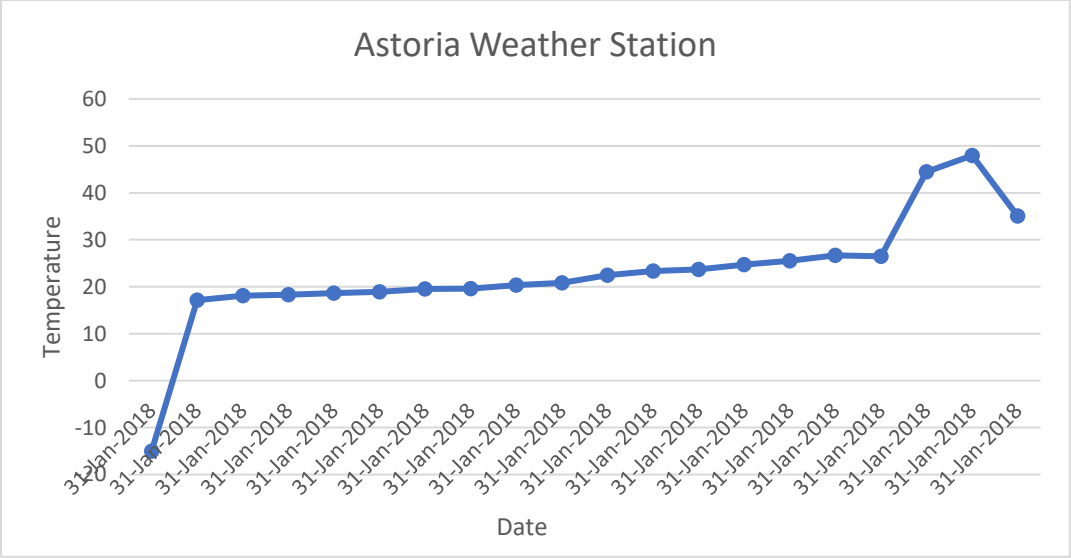
Puja Roy 4/15/22

CET 4900 - OL60

Internship Journal Entry#12

Throughout this week of my internship, I worked on simplifying the Astoria site data collected from the NY-uHMT weather station by taking the first 20 rows of data and loading it into a new dataset. This procedure was necessary since the dataset is extremely messy and I had to observe whether the rows or values in the AST_AirTF column was removed from the dataset when I removed the outliers using Python pandas. I also worked on creating visualizations of the new simplified dataset which is going to be used for testing purposes. As you can see below, there are still some outliers since I didn't remove the outliers for the simplified dataset yet. I created graphs and charts in Excel to display the number of temperature and humidity which occurred during January 2018. These graphs and charts will assist me in creating time series graphs of Astoria site's temperature and humidity using Python.





After I completed creating the graphs in Excel, I worked on removing the outliers for the sample dataset in Jupyter Notebook using Python pandas and NumPy. Since the sample dataset was small and simple to filter through, it was easy to detect and remove the outliers. I had to use the quantile method and define min and max thresholds which helped to detect and remove rows that consisted of outliers.

Detect outliers using percentile

In [131]:
```
max_thresold = df['AST_AirTF'].quantile(0.95)  #Anything above the output value will be considered an outlier
max_thresold
```

Out[131]: 44.6725775

In [132]:
```
df[df['AST_AirTF']>max_thresold]
```

Out[132]:

	AST_date_time_1	AST_date_time_2	AST_AirTF	AST_RH	AST_Rainfall_Tot	AST_VWC1	AST_VWC2	AST_VWC3	AST_VWC4
18	31-Jan-2018	12:30:00	47.9506	42.75033	0	0	0	0	0

In [133]:
```
min_thresold = df['AST_AirTF'].quantile(0.05)
min_thresold
```

Out[133]: 15.548662000000002

In [134]:
```
df[df['AST_AirTF']<min_thresold]
```

Out[134]:

	AST_date_time_1	AST_date_time_2	AST_AirTF	AST_RH	AST_Rainfall_Tot	AST_VWC1	AST_VWC2	AST_VWC3	AST_VWC4
0	31-Jan-2018	08:00:00	-15.00273	45.77231	0	0	0	0	0

Remove outliers

In [135]:
```
newdataset = df[(df['AST_AirTF']<max_thresold) & (df['AST_AirTF']>min_thresold)]
newdataset

# This line of code means that if the air temperature column is less than the max threshold and the air temperature is greater
# than the min threshold of the dataframe, then keep the values and remove the outliers that exceed the limit

# The outliers (row 0 & 18) were removed
```