Puja Roy 3/25/22

CET 4900 - OL60

Internship Journal Entry #9

Throughout this week of my internship, I worked on detecting and removing outliers in the Astoria dataset that was collected from the New York Urban Hydro-Meteorological Testbed (NY-uHMT) weather station. I downloaded the data as a CSV file in Excel and saved it as a comma delimited csv file for preparation of data analysis. Then, I imported and processed the data in Jupyter Notebook to analyze and remove the outliers in the dataset. While analyzing the data, I analyzed that the AST_AirTF column which is the column storing the information for Astoria's air temperature consisted of outliers. I observed the outliers by plotting the data using matplotlib and using boxplots to verify where the outliers were located. I conducted research on which techniques to leverage to remove the outliers immediately. However, this became a difficult and lengthy process since there are 8,246 rows and the data was not clean enough to spot the exact values of the outliers. I wrote a for loop using Interquartile range to replace outliers with null values. I also wrote a Python script to replace the outliers with median values. However, the code did not completely remove the outliers. I continued research on the best ways to remove the outliers.

```
26-Jan-2018,07:30:00,28.98356,30.86388,0,0,0,0,0
26-Jan-2018,07:45:00,29.2418,32.79153,0,0,0,0,0
26-Jan-2018,08:00:00,-4.730743,27.01772,0,0,0,0,0
26-Jan-2018,08:15:00,26.51103,30.21064,0,0,0,0,0
26-Jan-2018,08:30:00,27.26928,30.46705,0,0,0,0,0
26-Jan-2018,08:45:00,27.43411,30.96156,0,0,0,0,0
26-Jan-2018,09:00:00,28.07147,30.59526,0,0,0,0,0
26-Jan-2018,09:15:00,28.34345,32.70607,0,0,0,0,0
```

Figure 1 – Astoria Dataset that displays some outliers

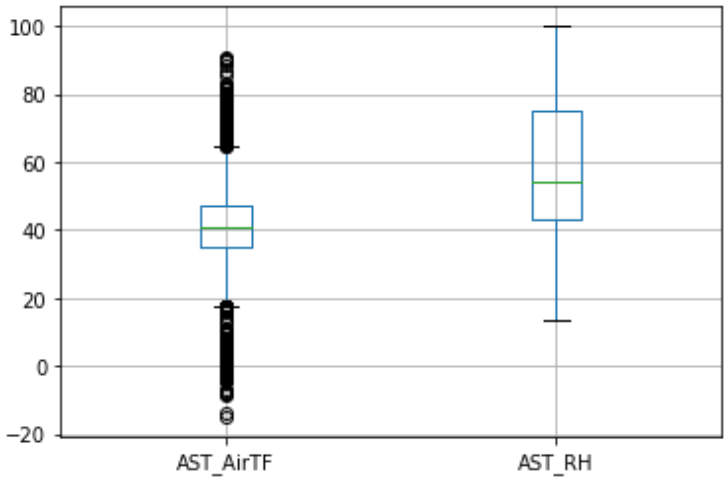

Figure 2 – Boxplot displaying outliers present in the AST_AirTF Column

```
In [120]: # For loop using IQR to replace outliers with a NULL value

          # Calculate the first and third quartile (Q1 and Q3).
          # Further, evaluate the interquartile range, IQR = Q3-Q1.
          # Estimate the lower bound, the lower bound = Q1*1.5
          # Estimate the upper bound, upper bound = Q3*1.5
          # Replace the data points that lie outside of the lower and the upper bound with a NULL value.

          for x in ['AST_AirTF']:
              q75,q25 = np.percentile(df.loc[:,x],[75,25])
              intr_qr = q75-q25

              max = q75+(1.5*intr_qr)
              min = q25-(1.5*intr_qr)

              df.loc[df[x] < min,x] = np.nan
              df.loc[df[x] > max,x] = np.nan

          # Used numpy.percentile() method to calculate the values of Q1 and Q3
          # Replaced the outliers with numpy.nan as the NULL values
```

Figure 1 – For loop code using IQR to replace outliers with a NULL value

```
In [121]: # Check the sum of null values or missing values using the below code:
          df.isnull().sum() # Sum of count of NULL values/outliers in each column of the dataset:

Out[121]: AST_date_time_1        0
          AST_date_time_2        0
          AST_AirTF              2
          AST_RH                 2
          AST_Rainfall_Tot       2
          AST_VWC1             398
          AST_VWC2             398
          AST_VWC3            1392
          AST_VWC4             399
          dtype: int64

In [122]: # Drop the null values (if the proportion is comparatively less)
          # drop the null values using pandas.dataframe.dropna() function
          df = df.dropna(axis = 0)

In [123]: median = df.loc[df['AST_AirTF']<75, 'AST_AirTF'].median()
          df.loc[df.AST_AirTF > 75, 'AST_AirTF'] = np.nan
          df.fillna(median, inplace = True)
```

Figure 4 – Python code for replacing outliers with median values

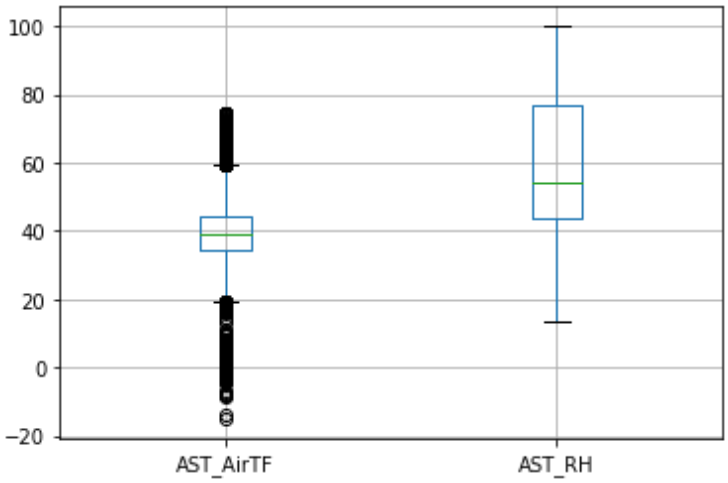

Figure 5 – Boxplot that displays outliers that are still present

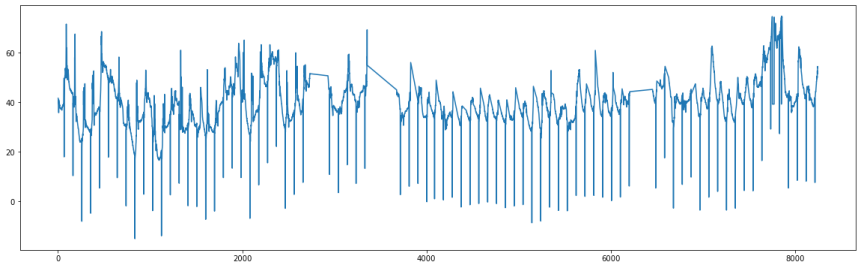

Figure 6 – Graph using matplotlib that displays outliers present

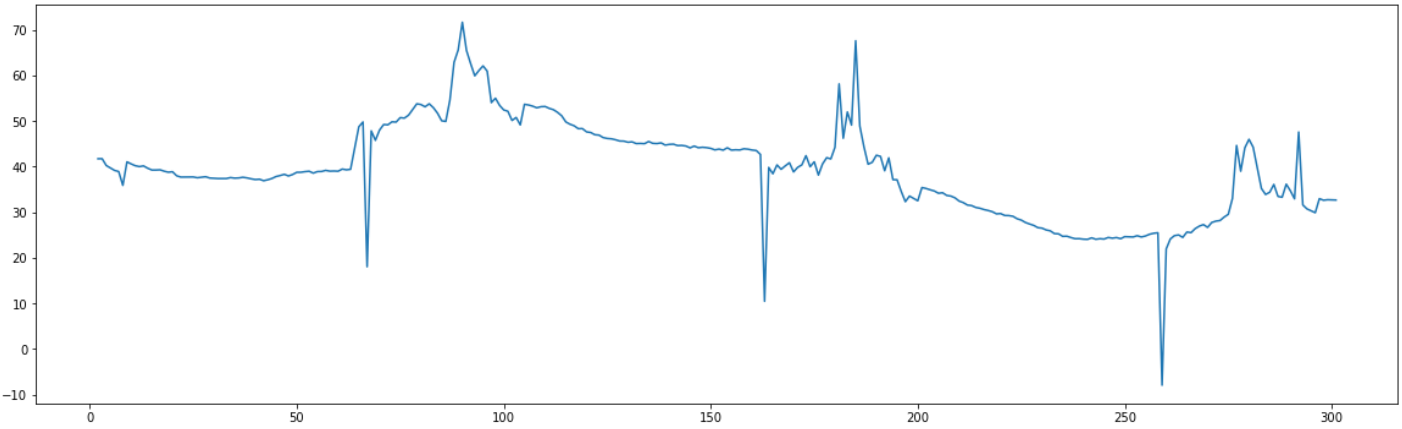

Figure 7 – Graph displaying spikes of temperature values along with outliers