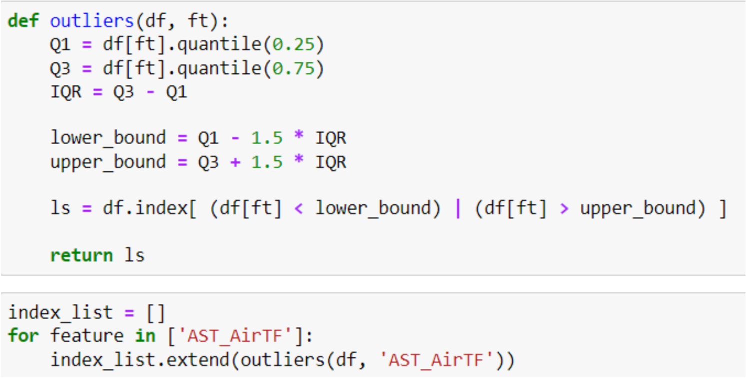
NOAA-CESSRST CENTER FOR EARTH SYSTEM SCIENCES AND REMOTE SENSING TECHNOLOGIES

Puja Roy 4/1/22

CET 4900 - OL60

Internship Journal Entry#10

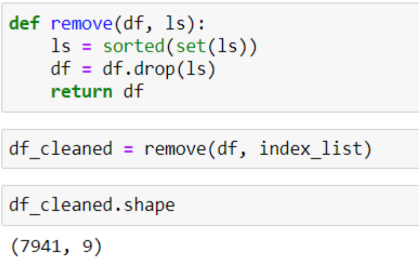
Throughout this week of my internship, I worked on detecting and removing outliers in the Astoria dataset that was collected from the New York Urban Hydro-Meteorological Testbed (NY-uHMT) weather station. I downloaded the data as a CSV file in Excel and saved it as a comma delimited csv file for preparation of data analysis. Then, I imported and processed the data in Jupyter Notebook to analyze and remove the outliers in the dataset by writing another Python code that helps to remove the outliers immediately. I had to clean the data by dropping the first 2 columns of the dataset since they were NaN values. I converted the AST_AirTF column values into float values. Visualizing the data assisted me in verifying whether the AST_AirTF column had outliers. Other ways to check for outliers is utilizing the boxplot method by installing the Python Seaborn packages/libraries. Then, I wrote a Python Script by defining a function that searches for and removes outliers using a for loop. Once the program ran, the outliers were removed by checking the shape of the overall cleaned dataset. Originally, there were 8,246 rows and after the program executed, the dataset showed 7,941 rows. This proofs that the outliers were removed from the dataset. I also wrote a short Python script to generate a profile report of the overall Astoria site dataset the dataframes.



```
def outliers(df, ft):
    Q1 = df[ft].quantile(0.25)
    Q3 = df[ft].quantile(0.75)
    IQR = Q3 - Q1

    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR

    ls = df.index[ (df[ft] < lower_bound) | (df[ft] > upper_bound) ]

    return ls
```

```
index_list = []
for feature in ['AST_AirTF']:
    index_list.extend(outliers(df, 'AST_AirTF'))
```

Figure 1 – Python Script for removing outliers



```
def remove(df, ls):
    ls = sorted(set(ls))
    df = df.drop(ls)
    return df
```

```
df_cleaned = remove(df, index_list)
```

```
df_cleaned.shape
```

```
(7941, 9)
```

Figure 2 – Python Script and function removing outliers

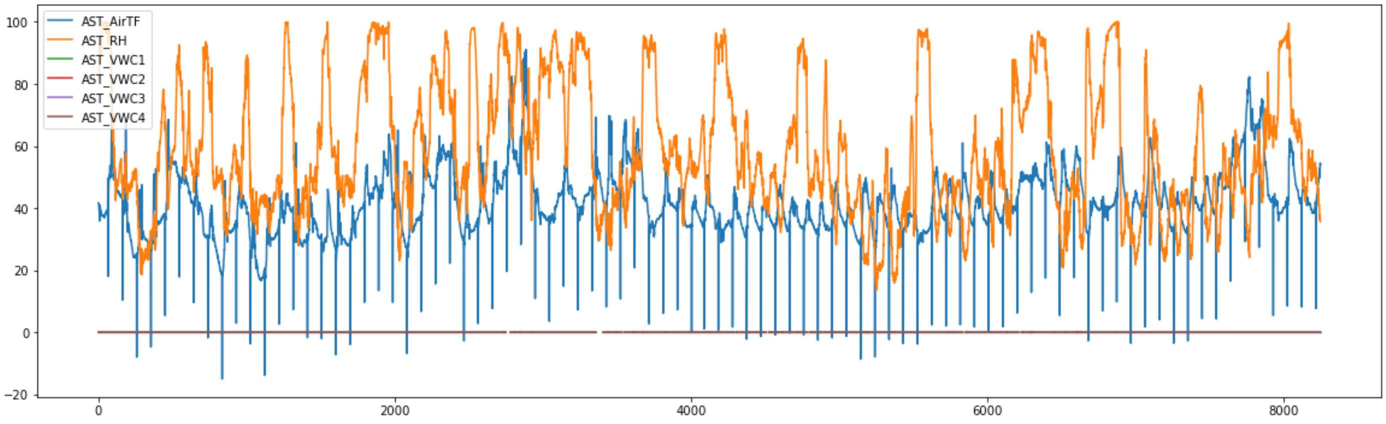


Figure 3 – Original uncleaned dataset displaying

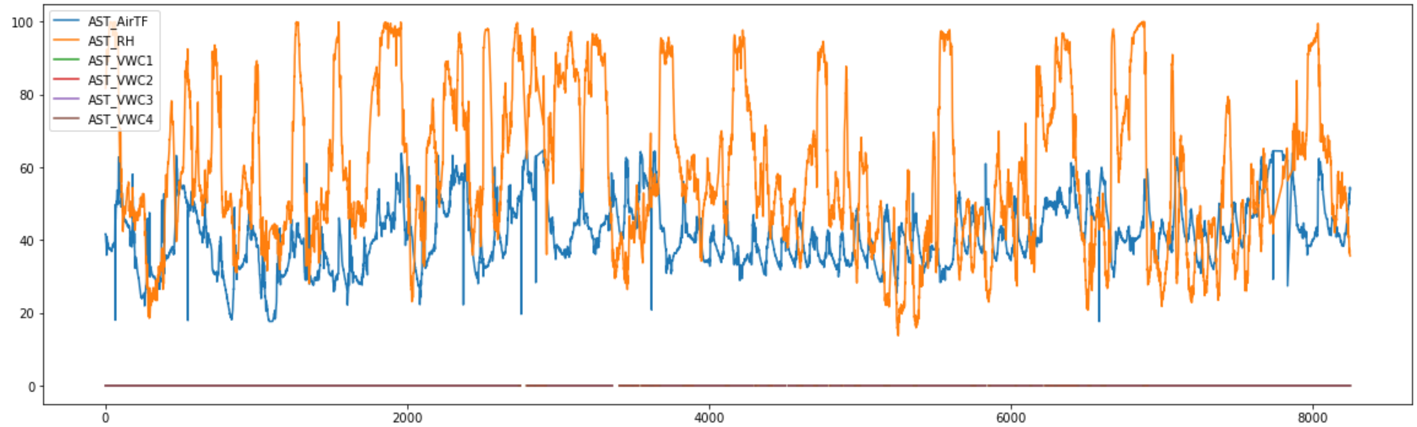
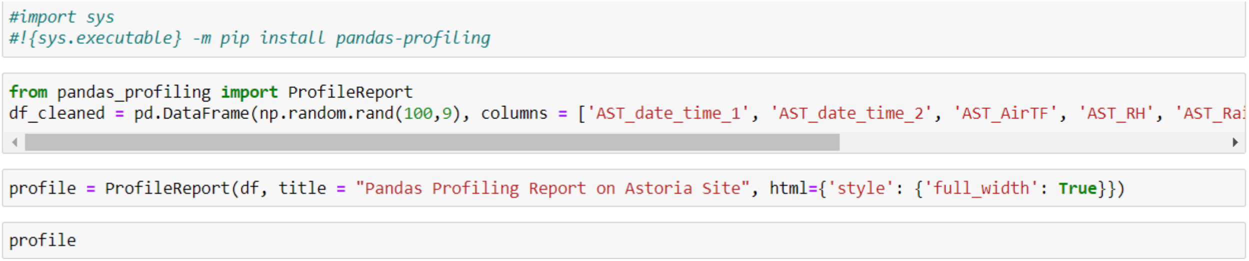


Figure 4 – Cleaned dataset of the Astoria site



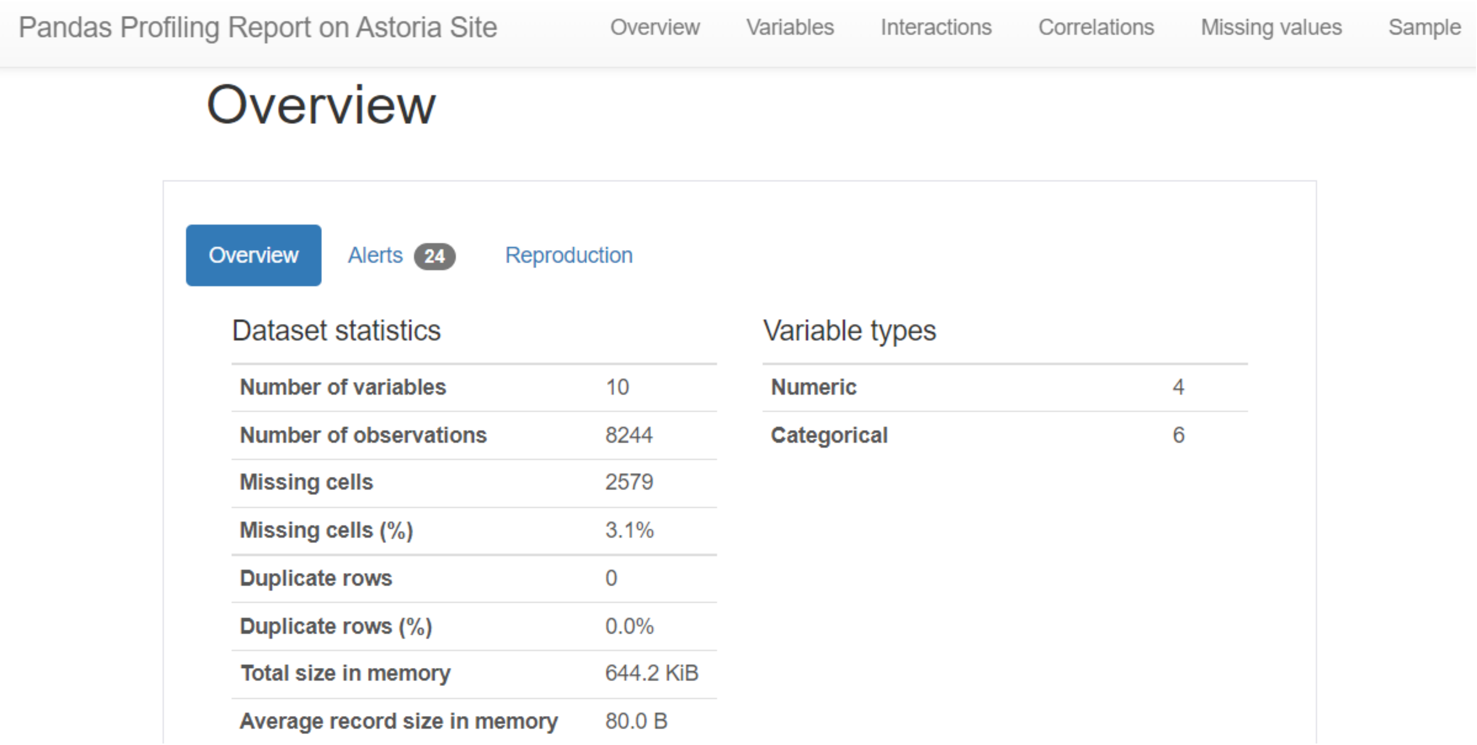
```
#import sys
#!{sys.executable} -m pip install pandas-profiling
```

```
from pandas_profiling import ProfileReport
df_cleaned = pd.DataFrame(np.random.rand(100,9), columns = ['AST_date_time_1', 'AST_date_time_2', 'AST_AirTF', 'AST_RH', 'AST_Rai
```

```
profile = ProfileReport(df, title = "Pandas Profiling Report on Astoria Site", html={'style': {'full_width': True}})
```

```
profile
```

Figure 5 – Python Pandas script to generate Pandas Profiling Report



Overview

Dataset statistics

Number of variables	10
Number of observations	8244
Missing cells	2579
Missing cells (%)	3.1%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	644.2 KiB
Average record size in memory	80.0 B

Variable types

Numeric	4
Categorical	6

Figure 6 – Pandas Profiling Report on Astoria Site displaying an overview of the dataset’s statistics