**Performance Analysis of Code**

**Sravan Kumar Borra**

**Capstone**

**Lines of code: 105**

**Time take for each Operation:**

* Data loading into Zeppelin – 10 Sec.
* Time taken for correlation and time series – 16 Sec.

**Bottlenecks:**

While cleaning the data I had some problem with missing values in the data set. So I replaced those missing values by taking the average of all the values. This gave me appropriate values for correlation calculation.

While performing correlation process, loading all the data files consumed more time. So I loaded the CSV files separately for each dataset.

**Optimization:**

All the data sets provided are in JSON format. So I cleaned the data and saved to CSV file. There are **seven** different data sets in weather data. So cleaning the data for each data set will consume time and makes the process slow. So write a loop for loading the data set where the code is optimized and took less than half of the time for loading the data separately. This makes my code optimized and efficient.

While plotting time series, the values presented in the data are not sorted. So the results are not accurate and it will not give sorted values. So I separately sorted the data set in CSV file separately and converted the strings into Date Time format. All this data manipulation will be useful for time series plotting and performing regression, SVM in the future.