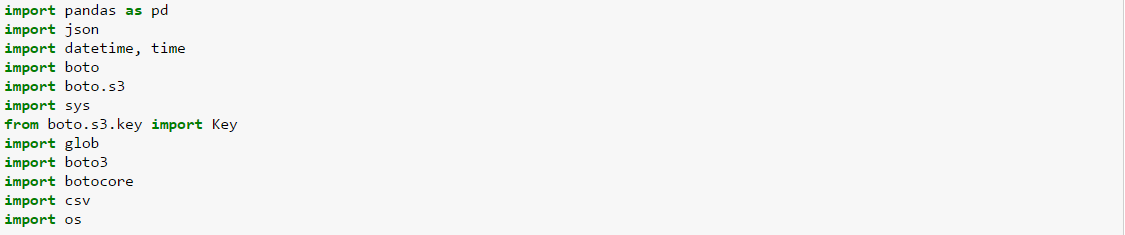
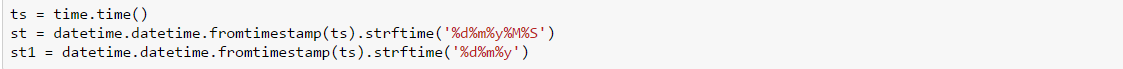
Data Ingestion

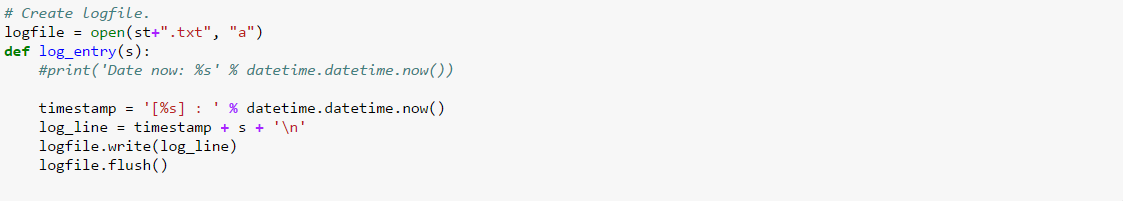
Imports for the dataingestion code:



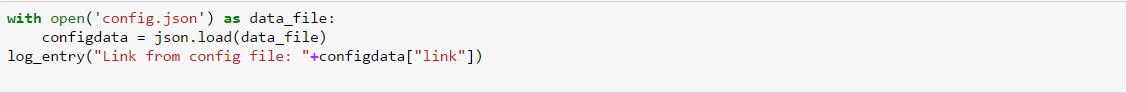
Code to fetch the timestamp



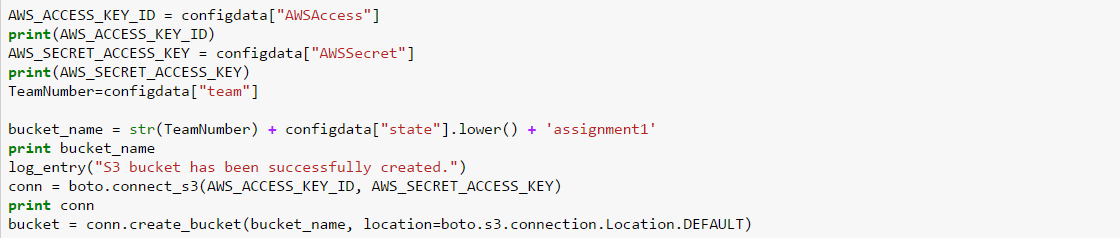
Code to create log files



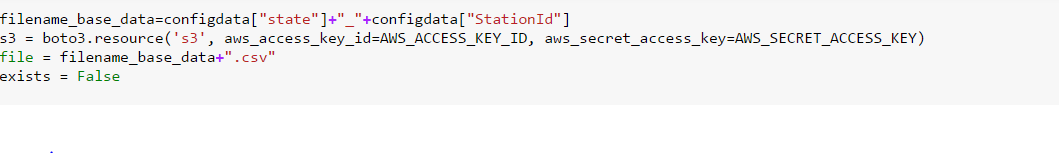
Code to open and load the config file provided

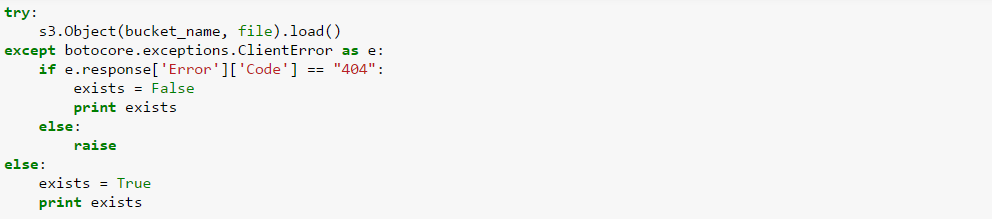


Code to create bucket on S3



Code to check for the base file that contains data for 10 years is available on S3 or Local. If the data exists on the local it will upload or else it will download the data from the links provided in the config, merge them and the upload it to s3

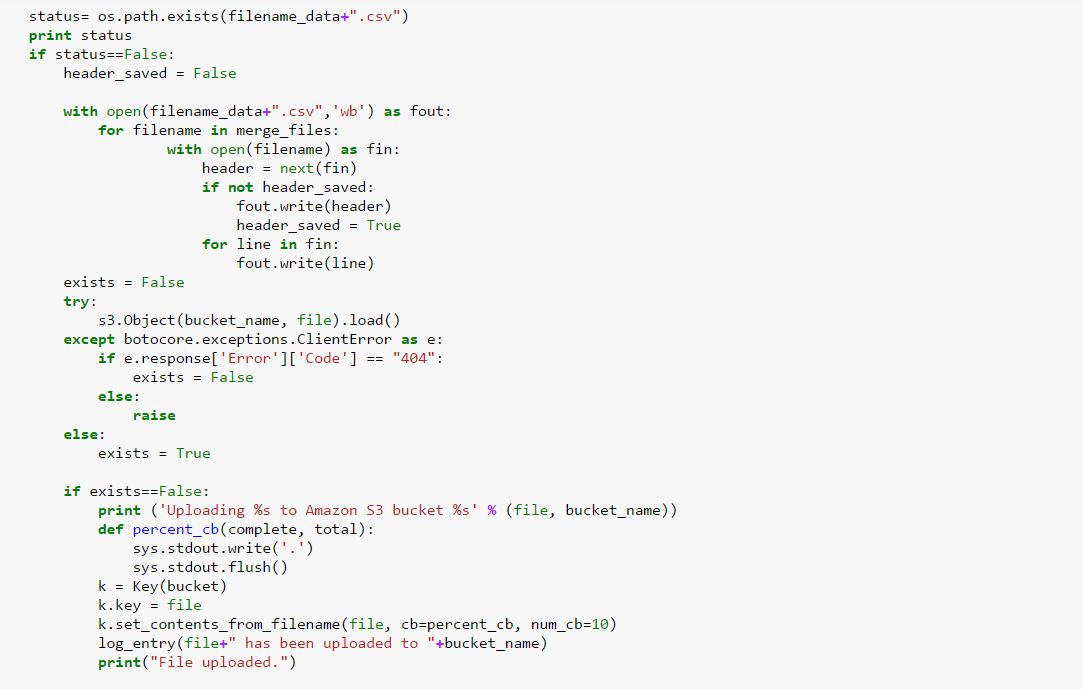




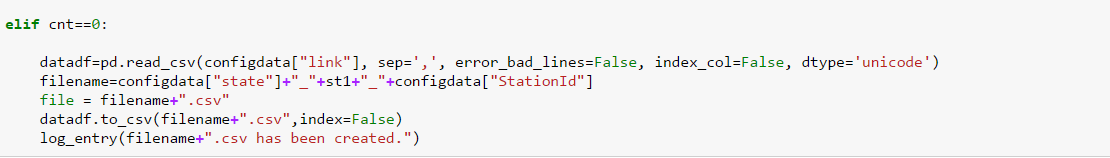


Code to upload data which is new the new link and adds only the remaining data which has not been added to the base file, creates a new merged file and is uploaded to the s3 bucket. If the file is on local but not on s3 bucket it will upload it / or if the data is not on the local it will download and create new file with new data and upload it again with the new time stamp









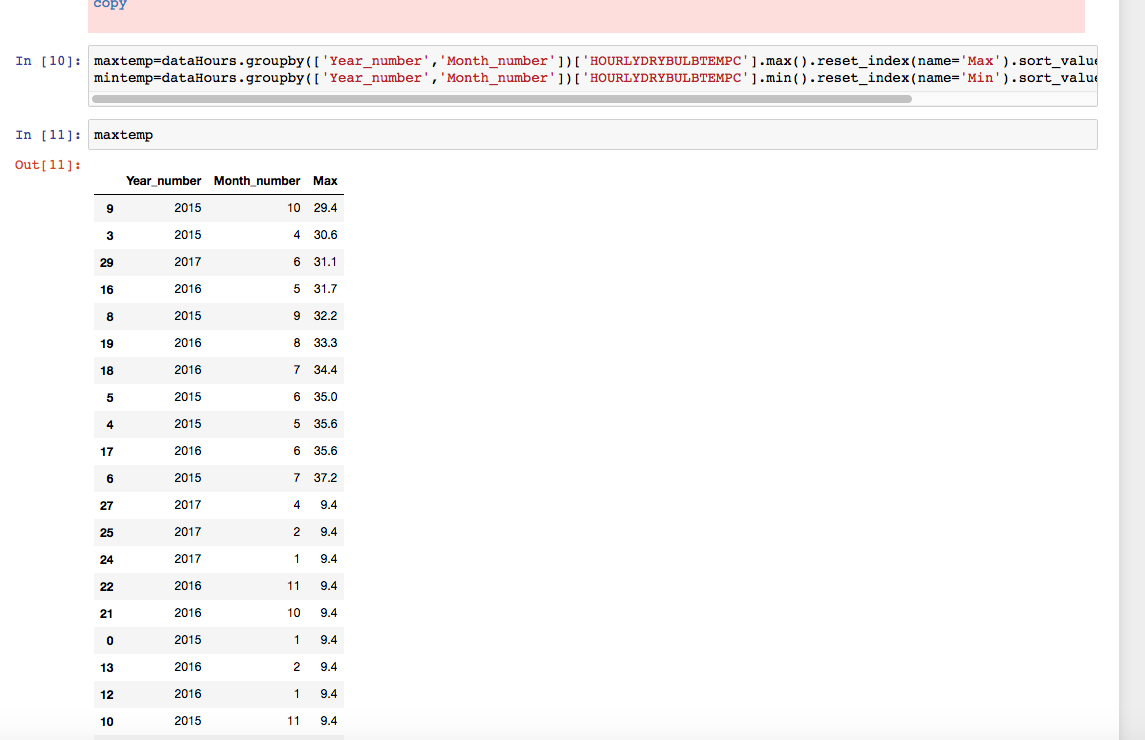
Raw Data EDA

I have tried to plot down few analytics :

Screen shots :

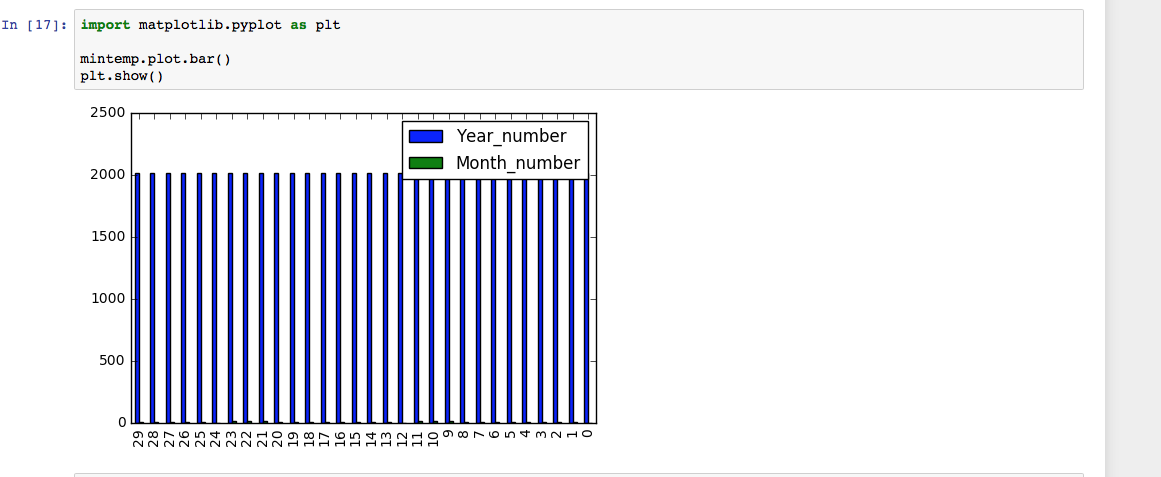
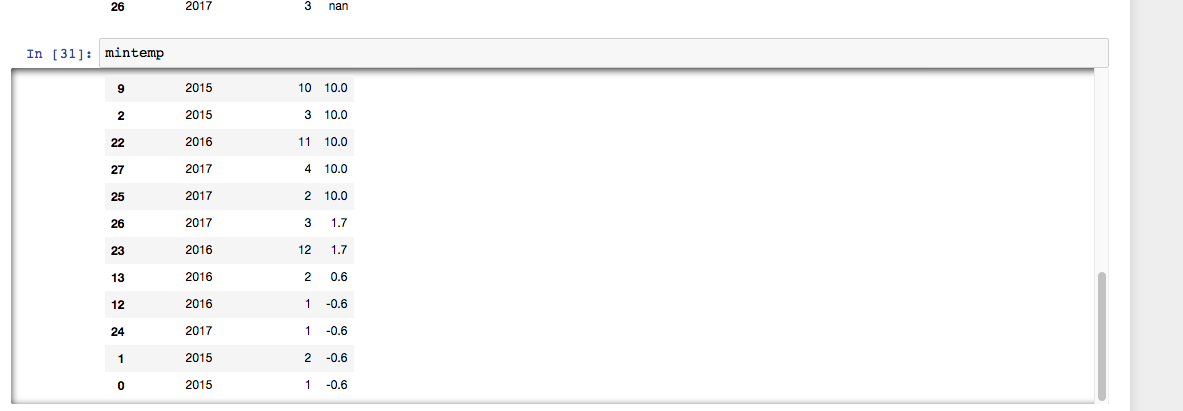
As in Raw EDA , we are importing the data form the Config Link as mentioned in the COnfig.json.

WE have found out the maximum temperature in the data sorted according to the Year and Month Number.   
We can clearly allocate and say about the highest peak of the temperature.



Similar analysis is done to find the minimum temperature from the data.

AS we can find from the plot and table that minimum temperature is occurring in the Year End Time.



Make File

It is DataPipeling tool. Which automate the task creating dependency

Run.sh

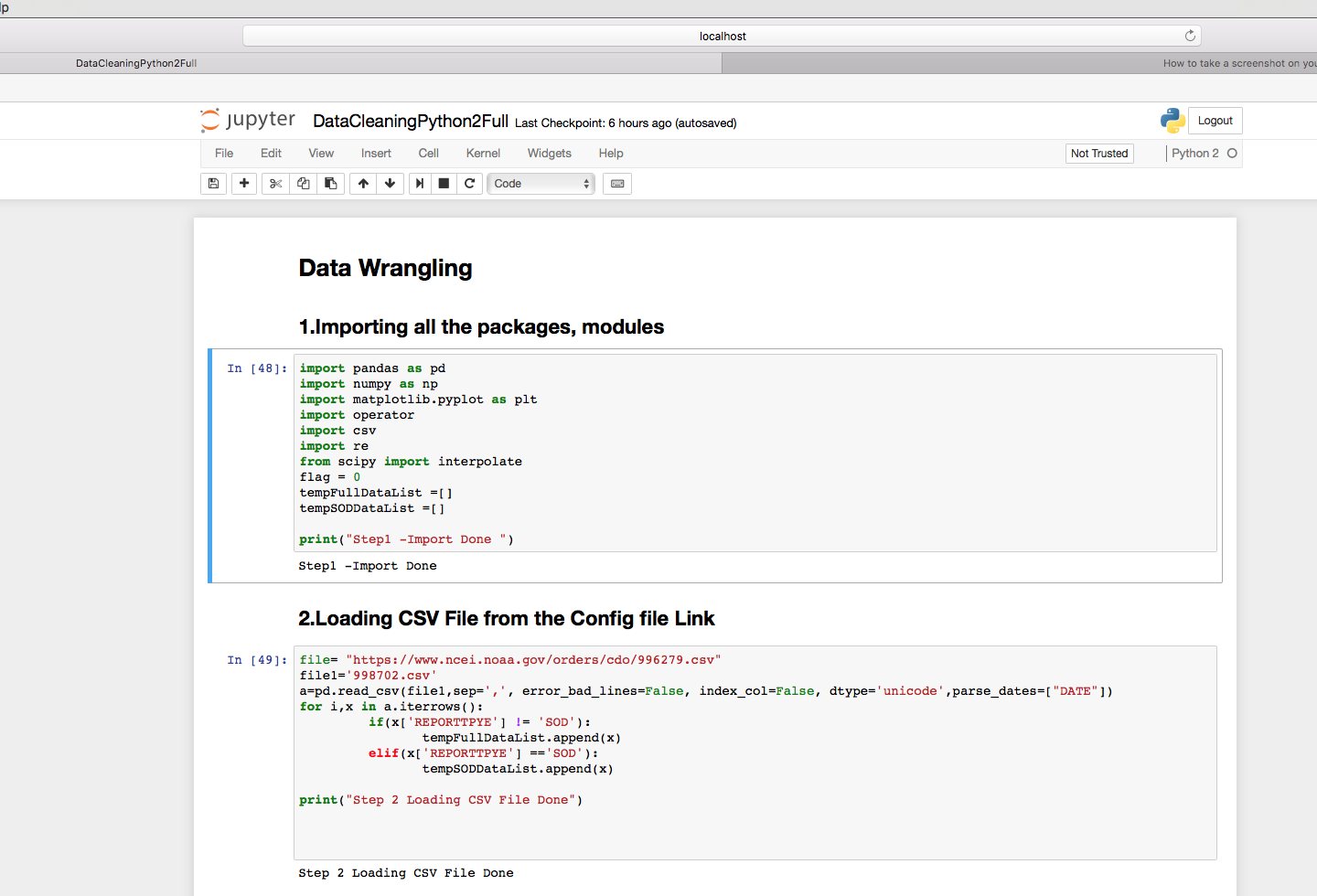
It will execute the make file when invoked by the docker container

Docker File

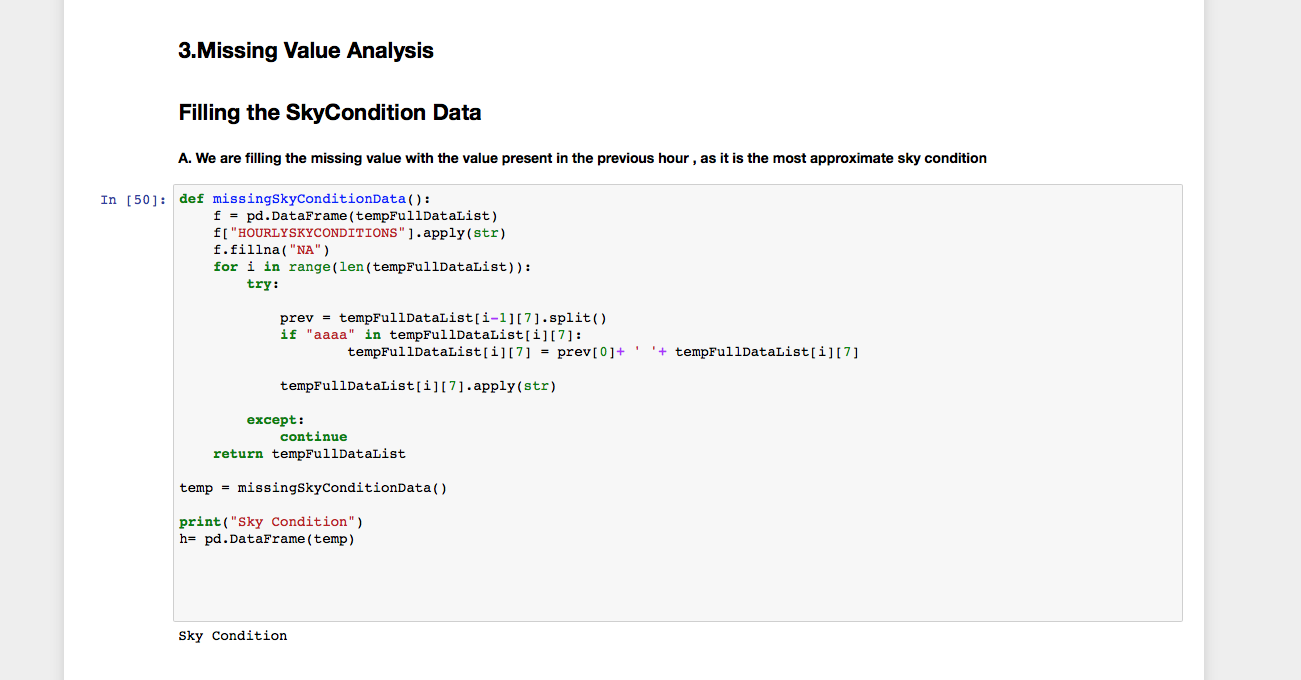
It invokes the Run.sh when the image is pulled from the Docker server Hub and executes on the Docker tools on any machine.

Clean Data EDA

1.Before Analysing we are Cleaning the data in the following step:

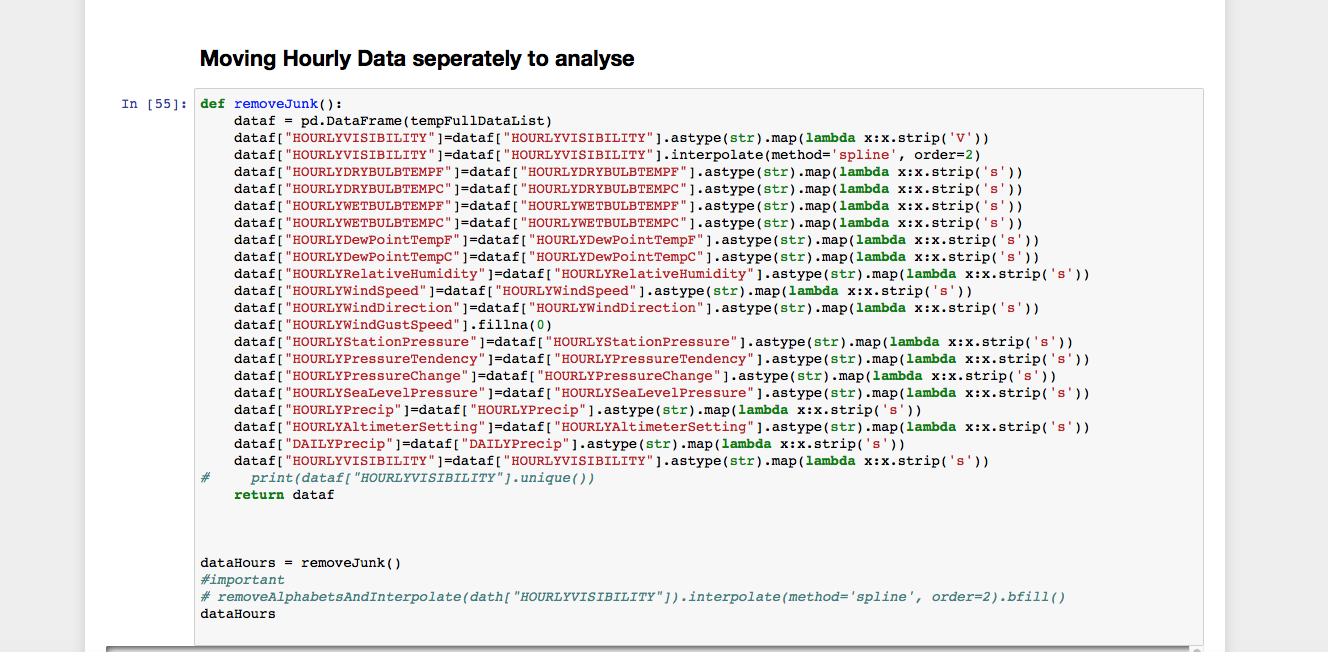


2.Populating the missing value in the Hourly Sky Condition Column.   
It is done on the basis of assumption that whenever the data is missing in the present column , we are taking the value present in the above column.  
It almost stated that the missing value of Sky condition would be almost similar to the previous hour.

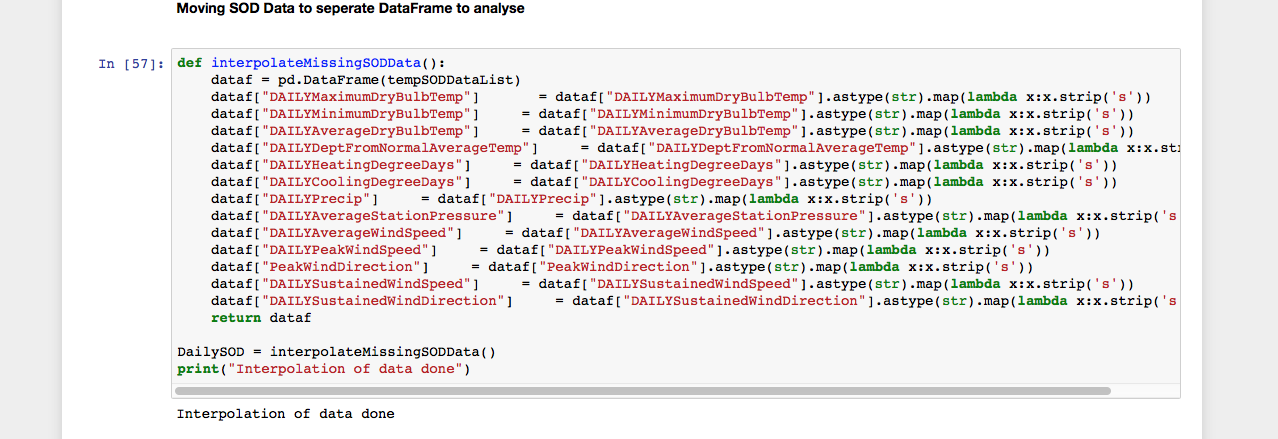


For Analysing our Data in more explorative way we are using the Bokeh Library to plot our graph.

Removing the junks from the file for hourly Data:

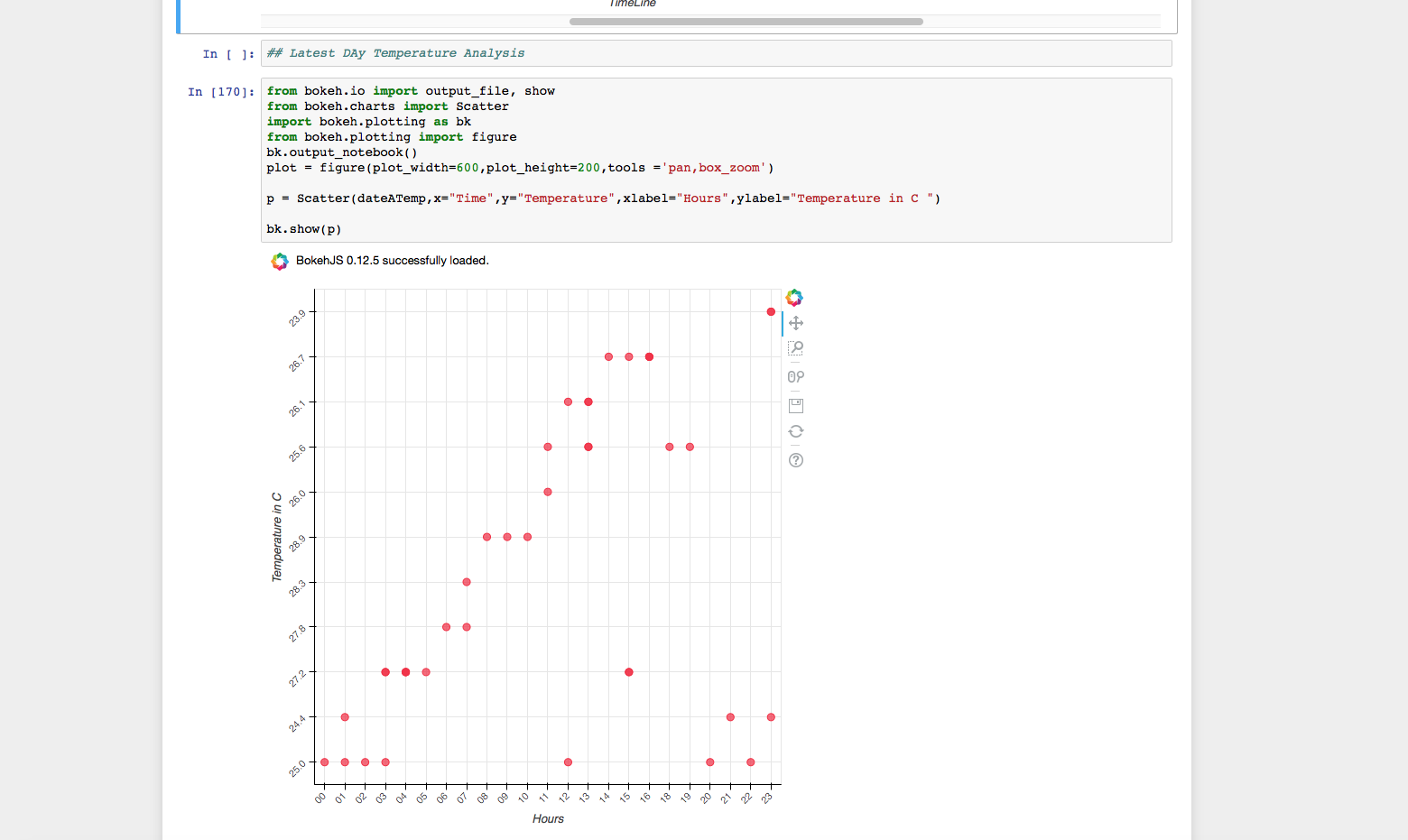


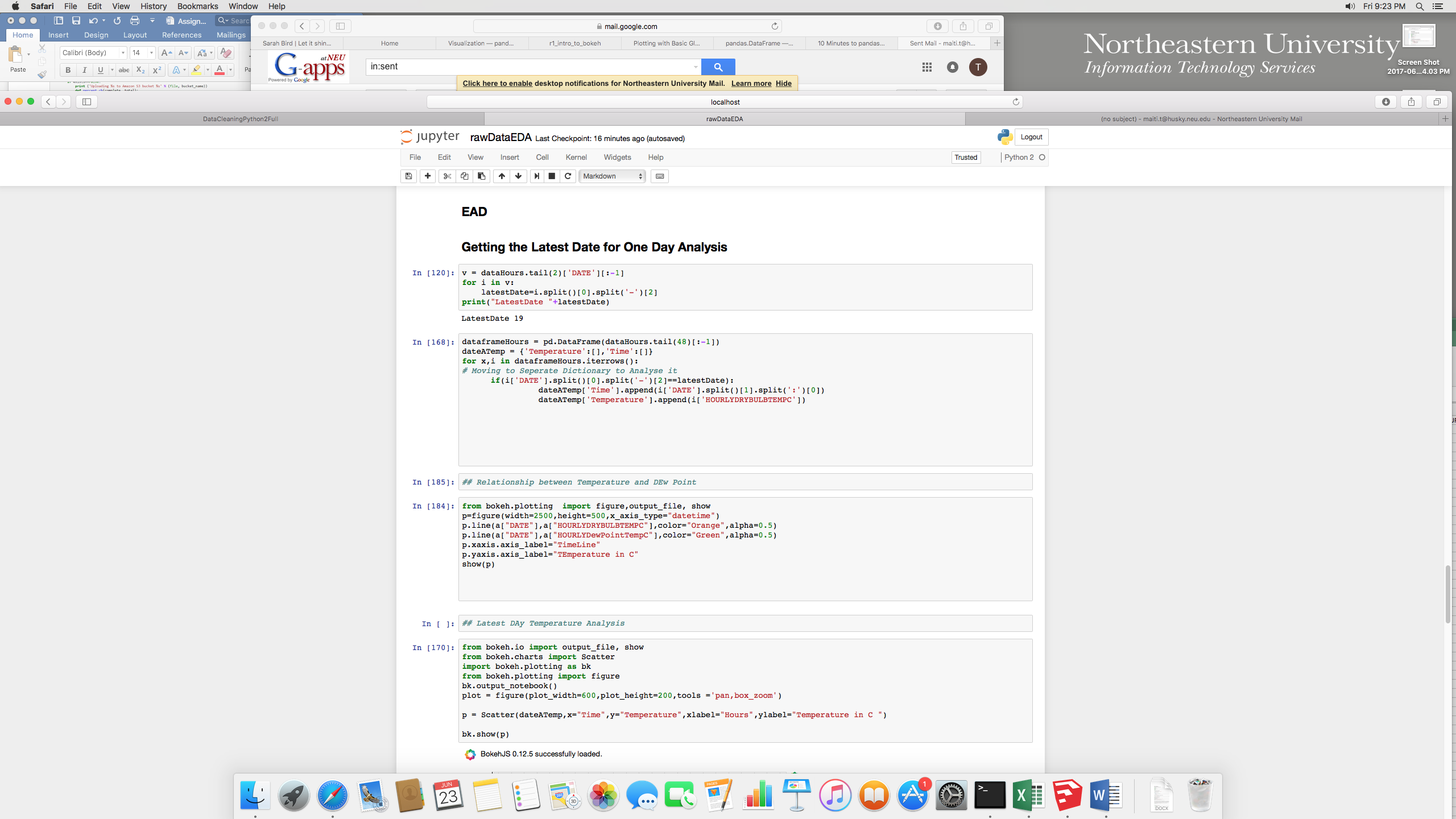
Removing the junks for Monthly Data



3.We are analysing the latest day analysis of last 24 hours on the temperature variance.  
We have written an function where it will catch all the required data of past 24 hours data.

Below are few other codes for the Bokeh Plot analysis.





4.We are tried to plot Relative Humidity as a Time Series analysis.

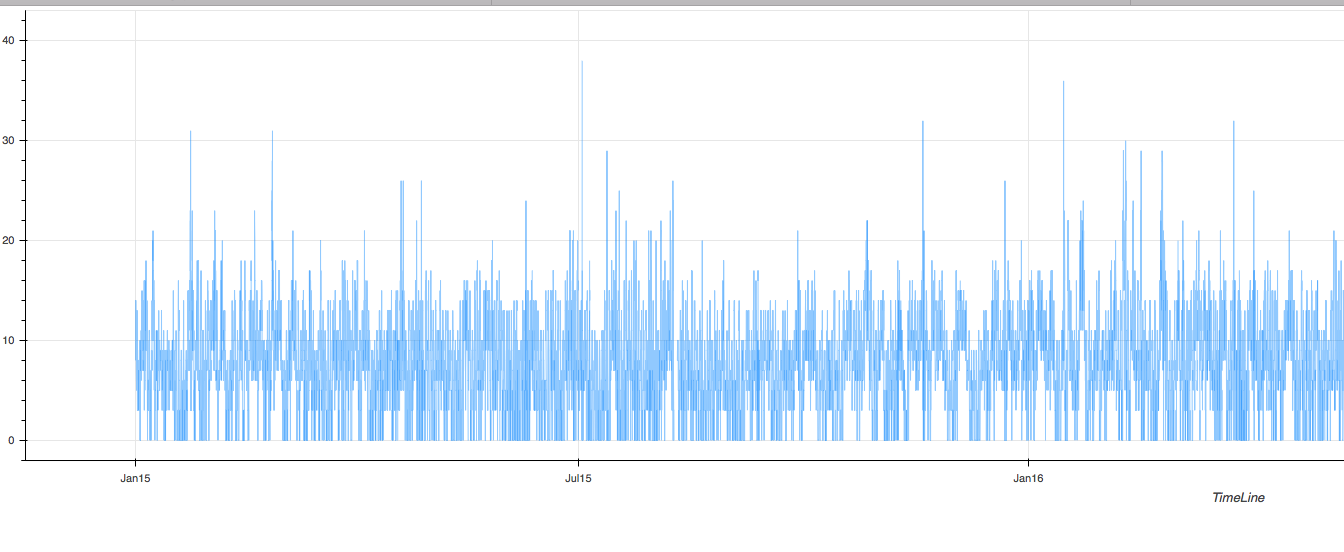
And as we can see from the graph Humidity have a lot of spikes , but still we can plot an average tendency to increase along the year from Jan to Jul Time period.



5. Analytics of DewPoint and Temperature :  
As we can infer from the Graph that Dew Point and the Temperature both mapped almost same.



6.Wind Speed Analytics L  
AS wecan infer form the graph that winf Speed follows similar patern of up and Down in a single day.



7.Analytics of Hourly SeaLevel Pressure , Altimeter Setting, Station Pressure

All the variables plot almost similar as we can infer all these follow similar fashion in Time series plot.



Read ME

Our Config File is Different from the Assignment . It is in the Following Format.

There should always be Link in the Base File.