**Report of the analysis**

**Step 1: Initial Data Ingestion**

* Constructed a Initialconfig.json file with data links from1947 to current date in 2017
* Used those data links to download complete data in .csv format.
* Uploaded the .csv file to amazon S3 bucket
* If you will rerun the code, instead of downloading the initial data again, the process will check whether initial data is already present on amazon S3 or not. If initial data is present, the process will move on to the next step of daily data ingestion

**Step 2: Daily Data Ingestion:**

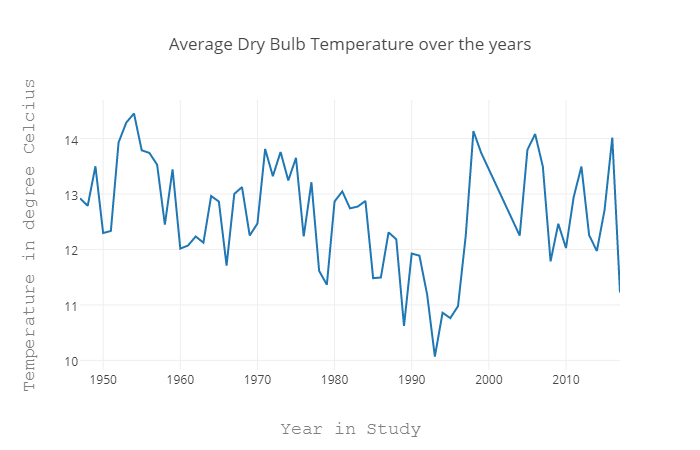
* This step will run after the initial data ingestion step.
* The new data from the received response will be appended to the initial data
* This step will create a new file with the max date from the complete appended data
* The new file will be uploaded to amazon S3 bucket

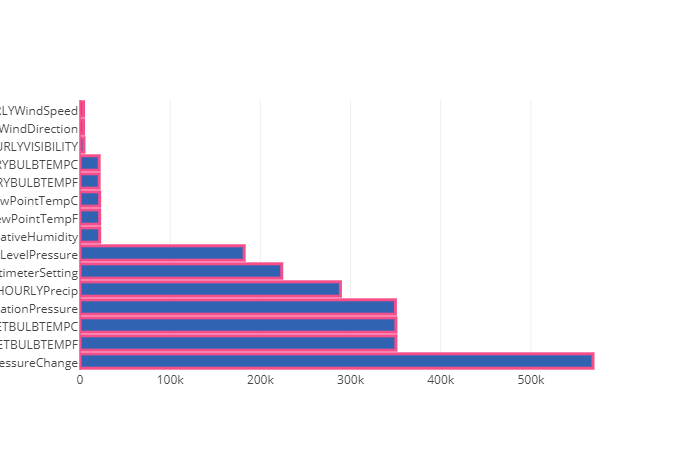
**Step 3: EDA on RAW data & Clean data:**

**1. Dry Bulb Temp Analysis**

* The dry-bulb temperature (DBT) is the temperature of air measured by a thermometer freely exposed to the air but shielded from radiation and moisture.
* DBT is the temperature that is usually thought of as air temperature, and it is the true thermodynamic temperature.

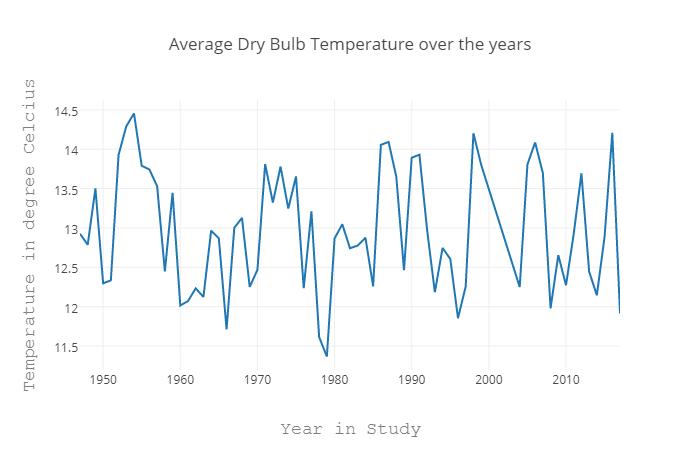
**Raw Data graph:**

****

* As we explored through the data, we observed that there are lot of blank and null values in the column
* Here is the graph for null values in each column. ****

**Clean Data graph:**

* + - For cleaning the dry bulb temp column, we followed the process of Interpolation.
    - As the values in temperature column was consistent and the temp in the column was gradually increasing over the years. So, for replacing blank we need to use **interpolation** where those values will be replaced by nearest suitable value.
    - Removed the tailing characters like 's' and 't' to the numeric values and created a separate column with the meaning of those characters
    - Here is the graph for dry bulb temp after cleaning.

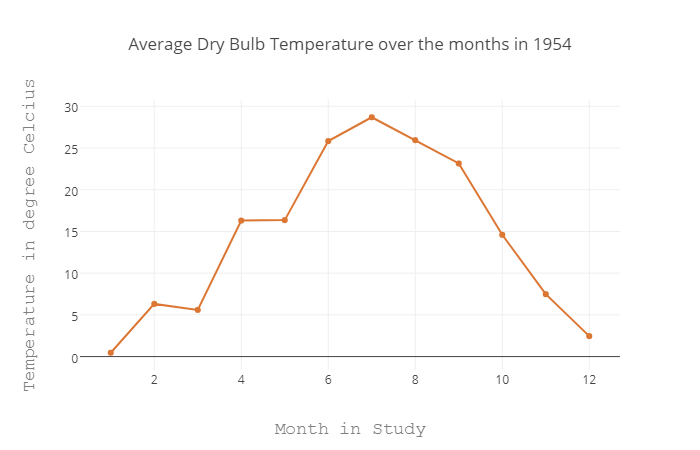


**Conclusion:**

* We can see two important results from above plotted graph:
  + 1954 is the hottest year as on average the temperature is highest as compared to other years
  + 1979 is the coldest year as on average the temperature is lowest as compared to other years
* In the below two graphs we dig deeper into month level to see the temperature variation.
* As you can see the coldest year has changed and is correct one as the history suggest years between 1978 to 1984 were the coldest in Illinois
  1. **Monthly analysis of the hottest and coldest year**

### Raw Data graph:

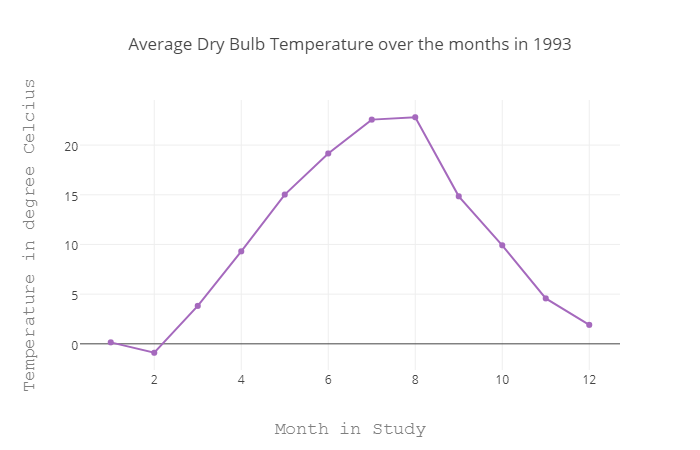
**Clean Data graph:**



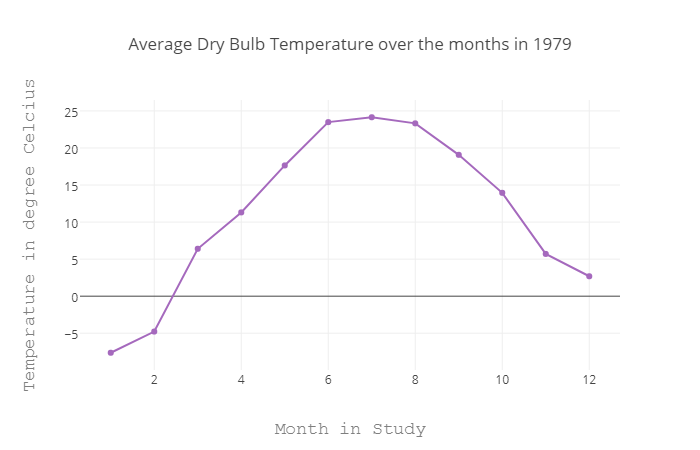
* Most important observation is that there are very less missing values for year 1954 for dry bulb as the graph remains the same for monthly plotting.
* From above graph we can see that:
* The temperature variation is not smooth or we can say that slope is not gradual for most of the months.
* Due to these sudden rise of temperature between months, the temperature rise could have felt more to the people.
* The below graph shows that during 1954 the average wind speed was 9.4 mph, which was 5th highest recorded average wind speed till date.
* Combine this with the wind speed during those months, the real feel temperature would have felt more.

### For coldest year, we got 2 different years, before cleaning its 1993 and after cleaning its 1979:

**Raw Data graph:**

****

**Clean Data graph:**

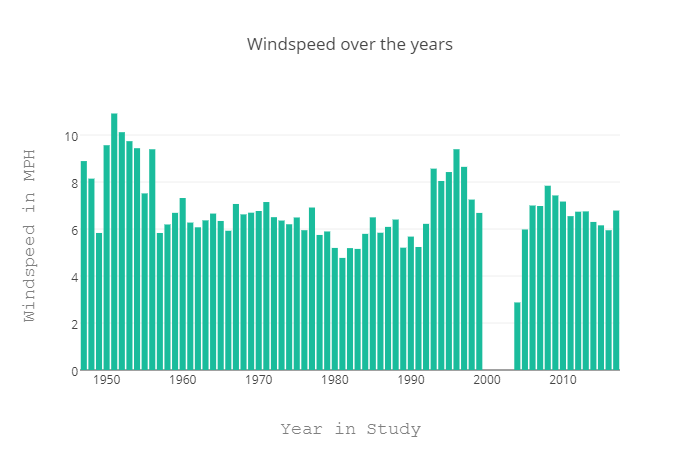
****

**Conclusion:**

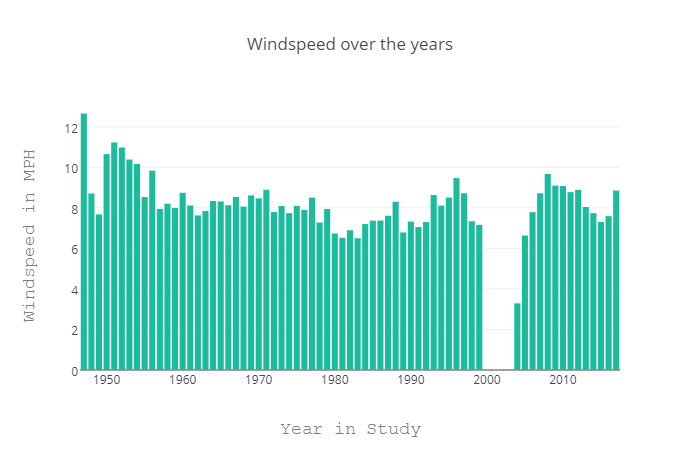
* The temperature variation is smooth or we can say that slope is gradual for most of the months.
* Due to these gradual rise of temperature between months, the temperature decrease could have felt more to the people.
* As from the wind speed graph, we can see that during 1979 the average wind speed was 8.6 mph, which was way above the average wind speed as compared with average wind speed in adjacent years.
* Combine this gradual decrease in temperature with the wind speed during those months, the real feel temperature would have dropped more.

### 3. Wind Speed Analysis

**Raw Data graph:**

****

**Clean Data graph:**

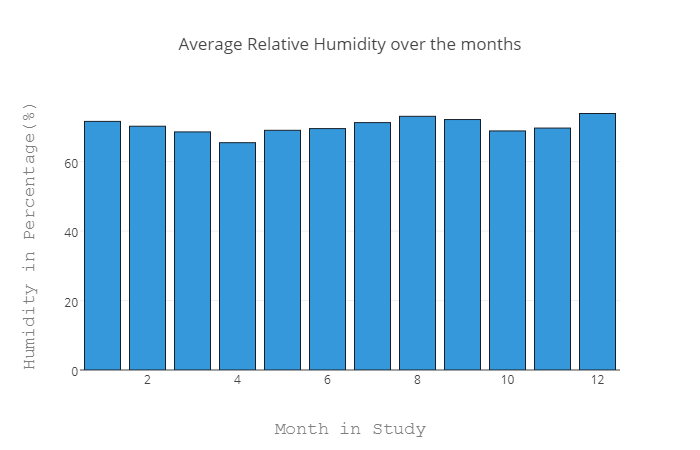
****

**Conclusion:**

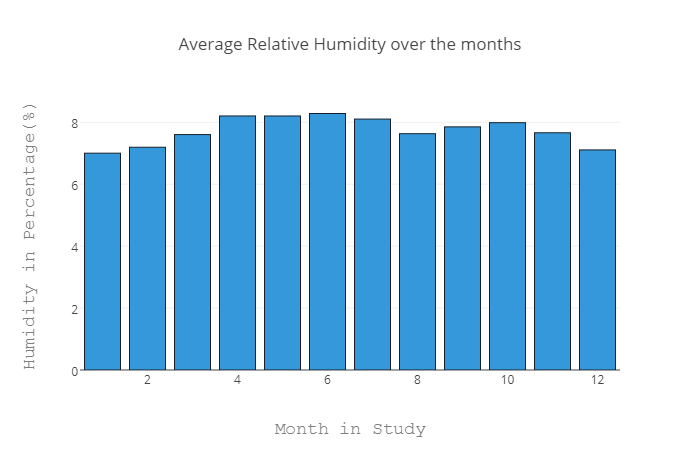
* As you can see Chicago is known for its windy nature, so, the speed is in high range
* There are no significant changes in data for this column after cleaning.
* The gap in graph represents the data for those years is not present in dataset.

### 4. Average Dew Point Temperature and Humidity over the months

**Raw Data graph:**

****

**Clean Data graph:**

****

**Conclusion:**

* As we saw in raw data, it is showing January and December are the most humid months, which is wrong as the temperature in those months in Illinois is very cold. Clean data is giving the correct results.
* The average humidity remains the same throughout the year.
* The top three most humid months each year in ascending order are as follows:
  + May
  + June
  + July
* After wrangling, the data remains pretty much the same so, the graphs are almost similar before and after the wrangling process

s

### 5. Dew point temperature

**Raw Data graph:**

### 

**Clean Data graph:**

### 

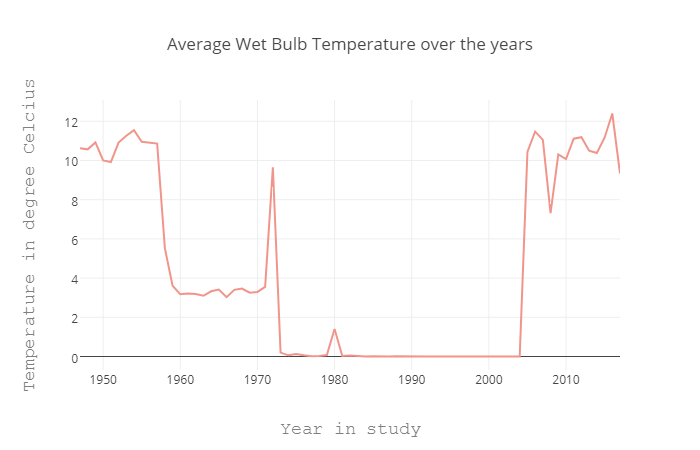
**Conclusion:**

* The data remains almost similar before and after the wrangling process
* Relative humidity is highest in the months of January, August and December.
* The average dew point temperature is lowest in January, quite low in December and pretty high in August.
* We can assume below two things from above two observations that:
  + It snows heavily in the months of January and December
  + It rains heavily in the month of August

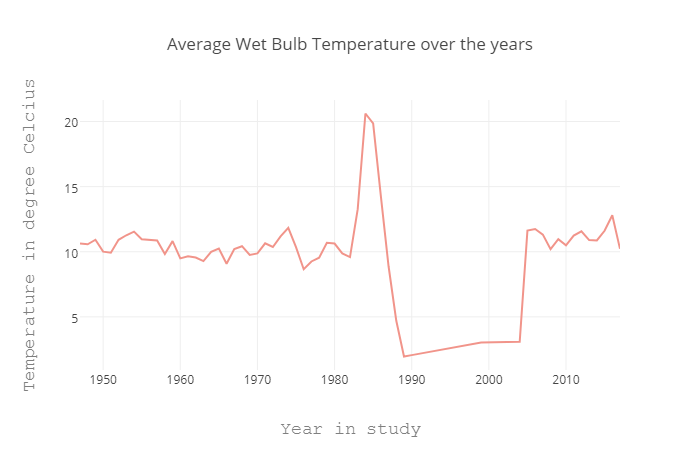
### 6. Wet bulb temperature

* The wet-bulb temperature (WBT) is the lowest temperature that can be reached by evaporating water into the air.
* The difference between the dry bulb temperature and wet bulb temperature determines how much dry the air is.
* If DBT-WBT is large, then the air has lower relative humidity

**Raw Data graph:**

****

**Clean Data graph:**

****

**Conclusion:**

* There is sudden rise and fall of wet bulb temperature.
* Ideally a wet bulb temperature increases with the relative humidity and we have seen that relative humidity remained almost the same throughout the year. So, there is something wrong with the data provided.
* As you can see there is a significant change here in analysis as we filled the missing values between 1997 to 2003. We are getting a straight line because of the average values.

**Step 4: Dockerized Before and After Wrangling Process:**

* Created base Python Docker image to upload and download data from amazon S3
* Image names are as follows:
  + beforewrangling
  + afterwrangling
* Created log files inside Docker image for each run of Docker image
* Published the Docker image to Docker hub
* To pull the Docker images from Docker hub use:
  + docker pull prashantvksingh/assignment1

**Step 5: Automated the complete process using Luigi:**

* Created 6 tasks:

DataIngestion\_Initial

DataIngestion\_Normal

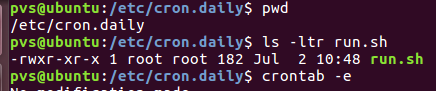
DataWrangling

UploadtoS3\_Initial

UploadtoS3\_Raw

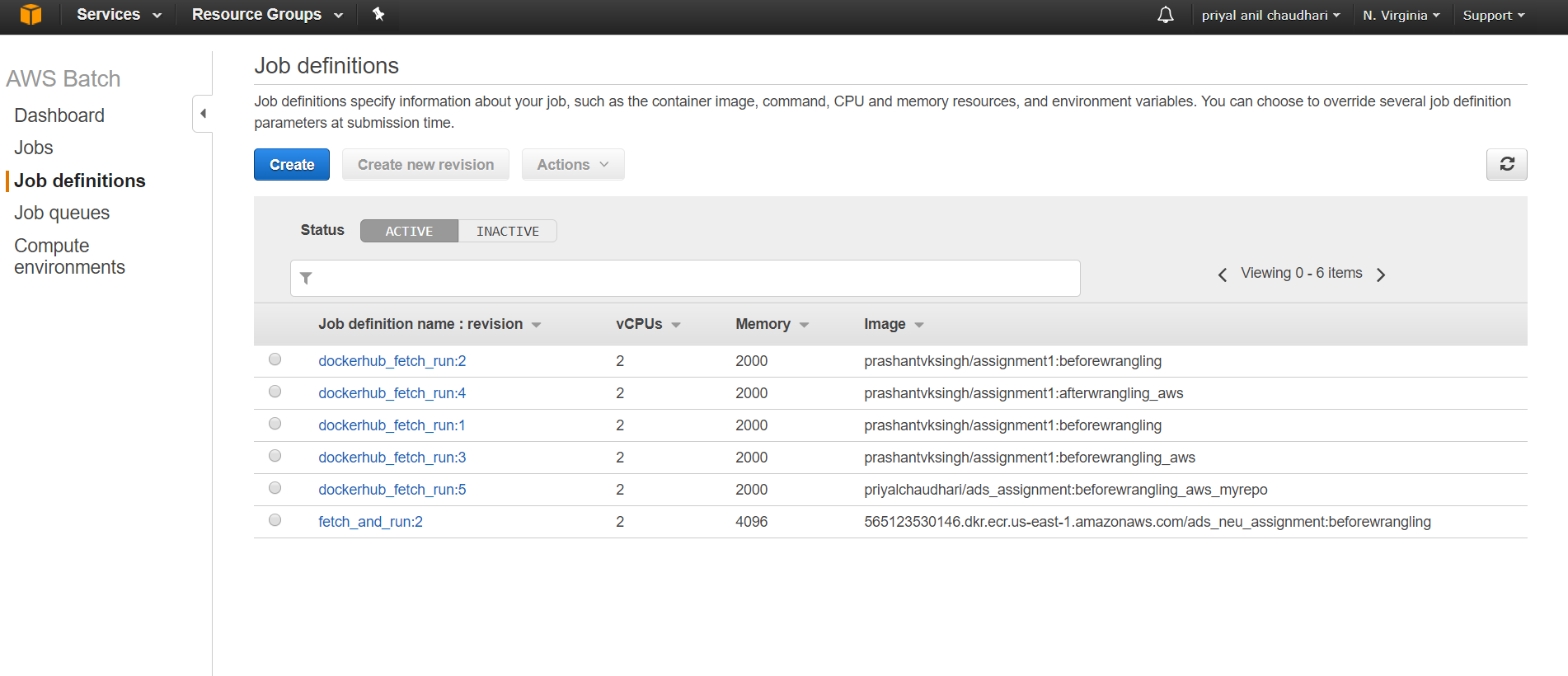
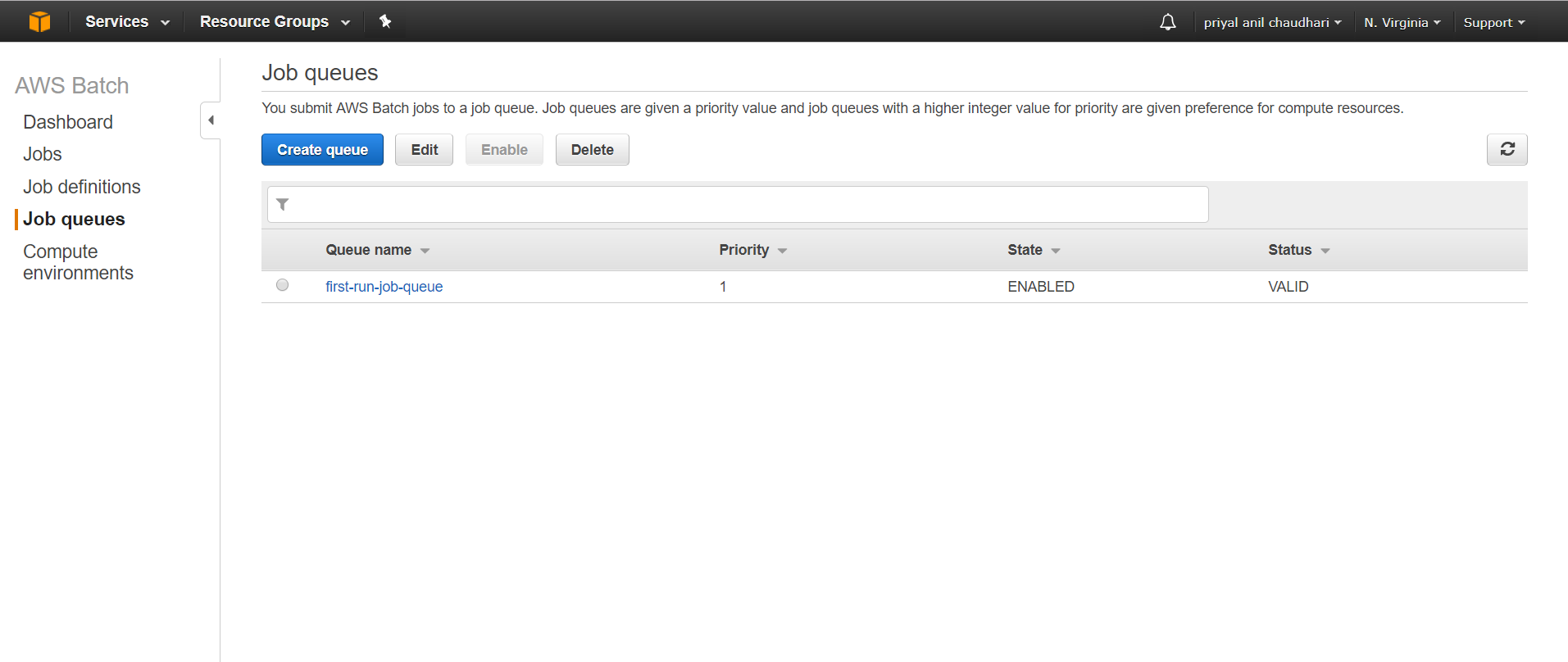
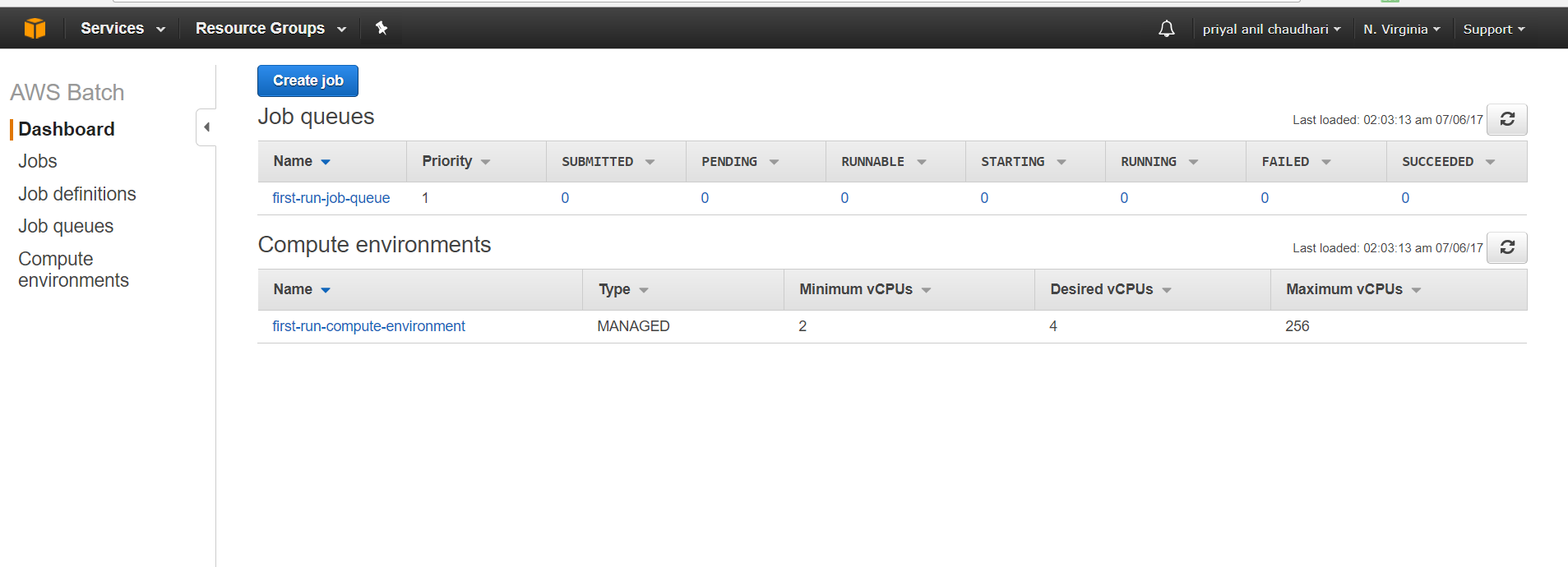
UploadtoS3\_Clean

* Opened the crontab using *crontab -e*
* Created daily cron job using crontab, to trigger last task which in turn will trigger the other depending tasks
* Placed the *run.sh* file in */etc/cron.daily* directory
* Cron job gets triggered daily at 11:00 AM EST
* Screenshot is as follows:





**Amazon EC2 job screenshots :**

****