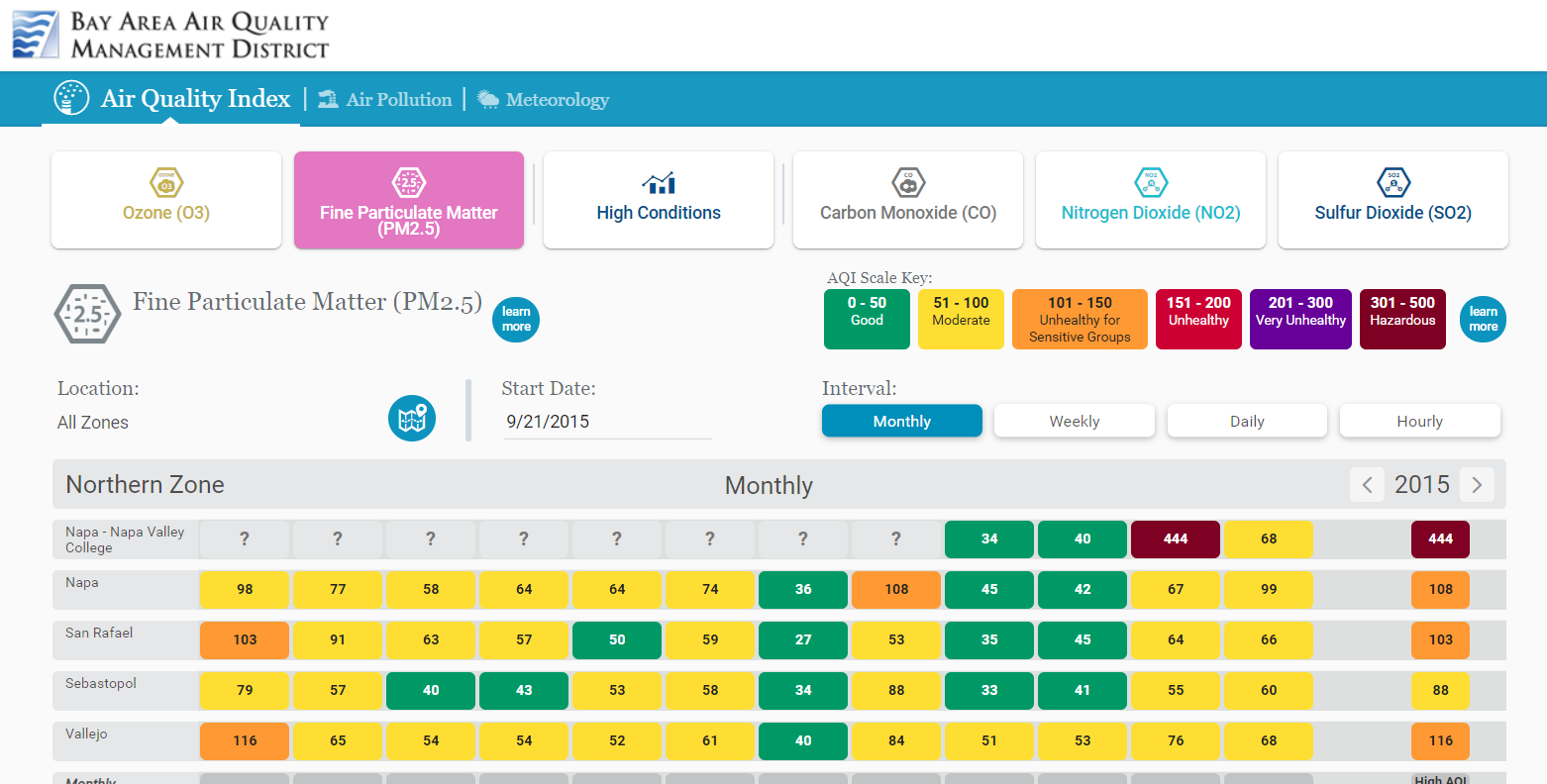
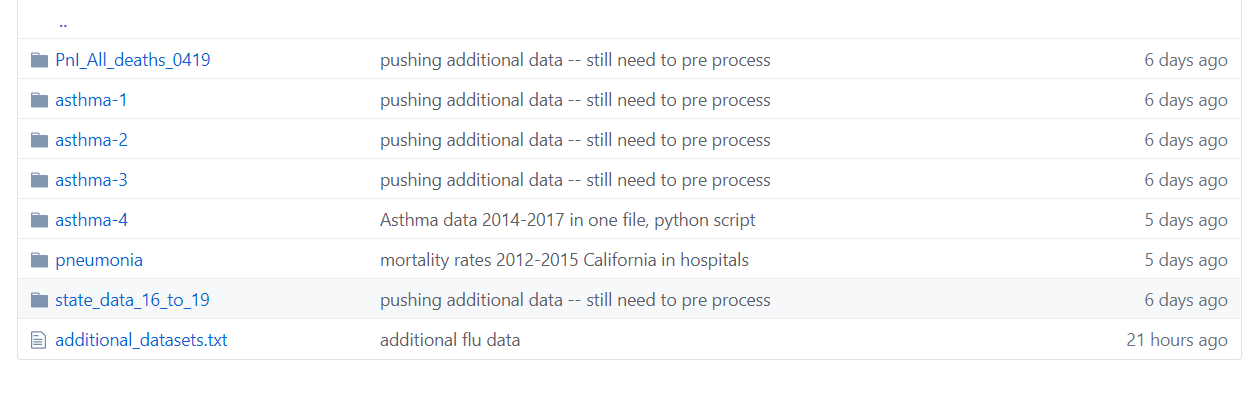
Accomplished So far:

* Data collection

1. Air Quality Data: We have collected Fine Particulate Matter (PM2.5) data from Bay Area Air Quality Management site. We have written web scraping python script to extract data from this site. Our script also does the initial cleansing of data and build it required structured csv format.



1. Purple Air Data
2. Air borne diseases data:



1. All death
2. Asthma1
3. Asthma2
4. Asthma3
5. Asthma4: We have collected this data set from tracking california site writing python script

* Data preprocessing and understanding the data: For data preprocessing and understanding, we are trying to apply the methods that we are learning in our data mining course. At first, each type of data set is being analyzed individually to understand the relation among variables, to identify important attribute and to remove unnecessary features. Then we are trying to find correlation between air quality data and air borne diseases data.
* Initial data visualization: Pair plot, Correlation plot, swarm plot
* KDD process: Data integrated from various data source and team involved in data selection.
* Data visualization: Data pre-processing to clean up the data relevance to the problem
* Literature study for prior model on similar kind of problem.

Findings and Results so far:

* The PM2.5 – fine particle index is one of the key parameters contribute for the air quality index.
* From the plot, few counties have poor air quality compared to other county

Difficulties:

* Time constraint is a major difficulty. All 4 of us are full time employees.
* Data collection part was very challenging. We had to communicate with multiple area experts for latest data and to understand the data
* Finding up to date data for all of the different datasets as we are collecting data from different sources.
* Correlation with two different domain data – air quality and disease might be challenging

Remaining:

* Model design:
* Model tweaking: