Hoang Nguyen

CmpSci 590V

Spire ID: 29071159

HW5 Report

1) Data choosing:

I chose the dataset about the pollution in US in 4 different substances: CO, SO2, NO2, O3. The dataset is a massive chunk of data that includes very detailed about the max value, min value, mean value, Air Quality Index (AQI) values for each substance by each hour of each day from 2000 to 2016, and also by each county of each state in US. These comprehensive data gave a wide range of values and variables to implement visualization, but also create many problems that I cannot anticipate, leading to failing to create most of desired visualization

2) Question need to answer:

I would like to compare the level of pollution between 2 chosen states by user, then using K-means algorithm to categorize these pollution levels to 3 different level (low, medium, and high polluted), using AQI values (the higher the value, the better the quality of air). I can only compare between 2 states, which is already a big problem for this data. There are 2 major problem to this data that become overwhelming obstacles.

_The data is huge and when take too long to load and too much space when saved in memory that sometimes crashed the program. My solution is to change database by getting only data in previous year 2016. The data was recorded early in 2016, so there only a few first month to compare.

_Even when many data entries, it still contains many holes (empty data), that create problems when getting data. For example, some states did not have report data for very long time. My solution is to split into 2 graphs to display 2 states' info. It is hard to compare and contrast when splitting 1 line graph to 2 separates, but because my implementation has problem when one state has data in a month and the other does not, I cannot fix it and resulting into separating them.

3) Work and Computation:

Without K-means, the computation is very simple, by taking the average AQI values to measure the overall AQI in each month of 2016 for each state. After viewing the trend in 2 graphs, I considered the average AQI of the whole year and decided where has better air quality to live.

4) The meant story: I used to have asthma in early childhood, which caused from bad environment rather than inheren. So I would like to have a website to compare environment (in this case is air quality) from one place to another, and decide whether to move and live there or not. I tried to implement the comparison and also the US map for visualizing the general level of Air Quality across US, and decided the best state to live (high AQI, but also consider the distance from current state user living in, the closer the chosen state, the easier to travel and moving). It is a very simple story, but I have not finished in time due to many problems occur when handling data too big and spare like this.

Link to my page: http://www-edlab.cs.umass.edu/~hsnguyen/