Tengli Fu (tf236)
Chenxi Su (cs2238)
John Hui (jmh547)
Project 3: INFO 5100

Chinese Air Quality by City: Breakdowns by Time and Geography

Note: A mobile version of our project is viewable here: http://chenxisu.github.io/DataVisualizationProject3/index.html

Overview

The premise of our visualization is to display Chinese Air Quality by city broken down by different time frames and geographies. The map visualization allows the user to see the scale of where the pollution is the worst (northeast), in between, and the least (south). The user can filter these to see highest cities. The user can also click on a city to see an individual breakdown of that city and segments the days into different pollution buckets and includes the highest and lowest AQI of each month. After analysis of the data, the three "Additional Findings" visualizations demonstrate some of our other key insights:

- Visualization 1: Peak effects of the January/December pollution effects are exacerbated the more polluted a city is.
- Visualization 2: Although one would potentially hypothesize there are differences in pollution for different days, there are virtually no differing effects between days of the week and pollution
- Visualization 3: Pollution is not necessarily consistent throughout the day. Beijing, Shenyang, and Chengdu demonstrate drops in the afternoon and rise in the night of PM2.5 data. Guangzhou and Shanghai do not.

Chinese AQI Pollution Data (Daily)

Our first dataset was scraped from the Chinese website, aqistudy.cn. We collected for the year 2015. This was definitely the most challenging part of the data collection, as there was no raw form of the data to download directly. We wrote a python script in order to scrape all 187 cities for every month and every day with over 70,000 rows of data. The script is attached in the appendix. This was especially challenging as the website is written in Chinese and thus there were issues with unicode conversion. The data then had to be processed and translated into English so that we could work with it for the visualization. The data includes dates, air quality index, range, descriptive quality, PM2.5, and other measures of quality. We ended up using the date, AQI, and descriptive quality for the cities.

Chinese AQI Pollution Data (Monthly)

This dataset was also taken from the Chinese website, aqistudy.cn. However, because the daily Chinese data scraper took a long time to build, we hadn't written this yet and thus scraped all of the monthly data manually. This is similar to the daily data, but instead includes summaries for each month of 2015. The data includes the month, the AQI, a range, descriptive quality, and other indicators.

U.S. Department of State Chinese PM2.5 Data (Hourly)

This data was taken from www.stateair.net and includes air quality monitoring from the U.S. State Department. This dataset was available in csv format. The data still had to be translated from Chinese to English. Additionally, the smallest time interval that the Chinese data was available was hourly, and thus we had to re-calculate all of the U.S. data to a daily basis in order to compare them. This data was also used to demonstrate

that the pollution levels of Chinese cities differ throughout the day rather than remaining constant.

Chinese City Latitude and Longitude Data
Additionally, in order to put our Chinese cities on the map, we manually collected the latitude and longitude for all of our Chinese cities, as there was no dataset available.

Map Visualization Mapping/Leaflet and Leaflet Sidebar
Leaflet.js was used to load and overlay the Chinese map, with the markers of the Chinese cities placed on top. These positions were based on the latitudes and longitudes of the Chinese cities. The Chinese map was from the Leaflet website, which was retrieved from OpenStreetMap's datasets. The map was originally created by MapBox. Leaflet was used to load the bottom layer and create floating windows on top of the map in addition to the leaflet-sidebar library that was used to create the sidebars to display the data. A filter with a range was designed if the user wanted to filter out cities that are less polluted to see those that are most polluted and observe that the less polluted ones are the ones that are filtered out first.

Map Information Visualization Mapping
There are two parts to the visualization included on the left sidebar. The first was a donut chart that used d3.layout.pie() to take the information from the daily csv file to breakdown the number of days and percentage of days that fall into each pollution category for each city. The color scale used was the same as before with AQI values before 50 as green, yellow below 100, orange below 150, light red below 200, dark red below 300, and purple below 400. Additionally, each segment of the donut chart can be clicked on, and the calculation for % of the total days is displayed.

The second visualization in the left sidebar is a bar chart that displays the lowest value and highest value of each month for each city. This uses the daily values to calculate the lowest and highest values for the month. This is also updated when a new city is clicked. The lowest values are green and the highest are red. This graph uses a linear scale for the AQI value and an ordinal scale for the month.

Graph Visualization Mapping
The graph visualization mappings all used data from the daily, monthly, and hourly pollution csv files. They were created as line graphs to demonstrate the trends between months, days, and hours. The three graphs all used linear scales.

In the first graph, three bands are used to demonstrate the top 5 highest yearly AQI for 2015, the median 5, and the bottom 5. These are segmented in order to show the distribution of the data so that one can get a summary of the overall trends of the data. Red represents the top 5, yellow represents the median 5, and green represents the bottom 5. Then the data is displayed across different months to show the December and January effects.

In the second graph, the same bands and thus same 15 cities are used. However, here the data displayed compares days of the week to one another. Even though one may expect there to be differences between pollution and the day of the week, this is done to demonstrate that there are minimal effects between different days of the week.

The third visualization uses the hourly US. PM2.5 data, PM2.5 is another indicator of pollution similar to AQI. This visualization represents the hourly differences for the day's average pollution for the five cities that were available. The user can select between different cities in order to see the graphs they are interested in learning more about.

The Story

The main point of our visualization is to visualize Chinese Air Quality by city segmented by different time frames and geographies. The map includes 187 Chinese cities color-coded based on their average annual Air Quality Index. A clear pattern is observed here as there seems to be less pollution in the south and increasing amounts approaching the northeast and northwest. As the user filters out cities with less pollution with the slider given, the most concentrated area is very clearly surrounding the northeast. The five cities with the highest average air quality index are Baoding, Dezhou, Hengshui, Xingtai, and Zhenzhou. The median 5 cities in terms of average air quality are Hangzhou, Dalian, Ma'anshan, Jiaozhou, and Changsha. The lowest 5 cities in terms of average air quality are Quangzhou, Xiamen, Yuxi, Haikou, and Sanya. Again, this pattern demonstrates that the most polluted cities are in the northeast, and they spread out from there with less pollution as you go farther. Additionally, when a user clicks on a city they can see what "rank" the city is so that they have an idea of where in falls in the 187 cities.

After a user clicks on a city, they can gain additional insights about the city that they clicked. The donut chart features the number of days that belong to a particular pollution severity as a proportion to days in a year. Again, one can see the trend as the farther you are from the northeast, the greater percentage you have of days that are not as polluted. Additionally, we provide the user with the lowest and highest AQI values for each month in a city. In high pollution cities, one can easily observe a more apparent parabolic-shape as December and January tend to be the most polluted months. In cities with less pollution, this is still apparent, however this effect isn't as obvious.

The third part of our visualization allows the user to see specific insights that we derived from the data. In visualization 1 one can see that the peak effects of the January/December pollution effects are exacerbated the greater the pollution is in a city. January and December tend to be the most polluted months, however for less polluted cities, the difference isn't as great and apparent as it is in cities with more pollution. The visualization 2 allows one to see that there is a lack of differentiation between the pollution levels between different days. Although one would potentially hypothesize there are differences in pollution for different days due to the number of cars that are used in a work week, there are virtually no differing effects between days of the week and pollution. In visualization 3 it's observed that pollution is not necessarily consistent throughout the day. In cities such as Beijing, Shenyang, and Chengdu, they demonstrate drops in the afternoon and rise in the night of PM2.5 data. Guangzhou and Shanghai do not demonstrate these differences.