Process book
Air condition of major cities in California

Names: Spencer Fronberg, Qian Zuo, Mattia Grespan

December 1, 2018

1 Overview and Motivation:

Many environmental factors can affect our health, but the one that has the most impact to us is air pollution. We breathe air to live and what we breathe has a direct impact on our health. Environmental Temperature and humidity are also important factors that influence human's living standard.

In the past, research focused on only one issue of air pollution, temperature and humidity. However, these three issues influence human's living standards a lot. We want to create a visualization combining these issues together that informs the user about the relationship between them in terms of the air quality. Moreover, we want to provide recommendations in decision-making in terms of air condition quality.

In this project, we will collect and show data of these three aspects for major cities in California. We chose California because it is the most populous state in the United States and the third largest by area. Due to the fact that the state extends from the very south of the west coast up to the north region of the west coast, the climate ranges from polar to subtropical. These characteristics should guarantee considerable variations in the data for temperature, humidity and pollution across the state. Hence, we should get more noticeable results.

Additionally, California is the biggest hub of software engineering in the world. Therefore, we consider pertinent to give our colleagues some suggestions about such important aspects for human's health like fresh air, convenient temperature, and suitable humidity.

2 Related Work:

The following are links to some of the material that inspired us for this project:

1. Air Pollution: Invisible Killer. BBC World Services. The Real Story. Audrey de Nazelle - Imperial College London and Maria Neira - World Health Organization.

   This is a very interesting episode of a radio show from the BBC where they discuss in a clear way about the awareness of the risks posed by polluted air.

2. Research + Remote Pollution Sensor Telemetry.

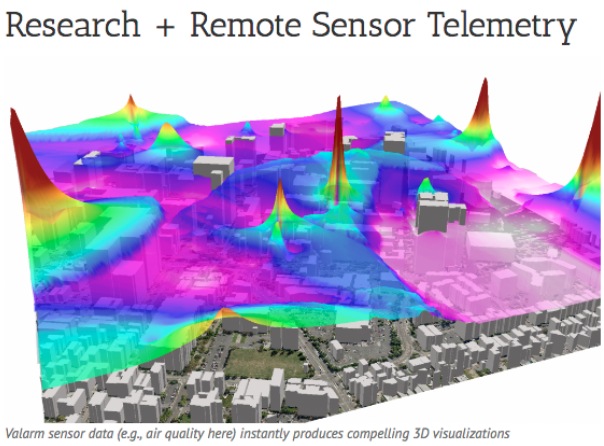

Figure 1: Valarm's company sensor data (Pollution)

This is a visualization of the air quality produced by sensor of the monitor industrial sensors Valarm. We found it interesting by the way it uses 3D encoding. This inspired us with a prototype version of our visualization.

https://www.valarm.net/research-and-academia/

3. Infoplease. Effects of Dry Air on the Body. https://www.infoplease.com/science-health/weather/effects-dr y-air-body

   We wanted to know more about the impact of lack of humidity on the body. We found this concise and well explained article.

4. We are considering different ways to encode friendly recommendations to the user about the levels of Pollution, Temperature, and Humidity. We found some good ideas like the following on Public Lab website: https://publiclab.org/notes/jiteovien/08-01-2018/air-quality-data-visualization-no-coding-necessary

Figure 2: Categorization of the Pollution levels

5. This example of a field shown during the Dataset Types lecture was also a good reference for us.

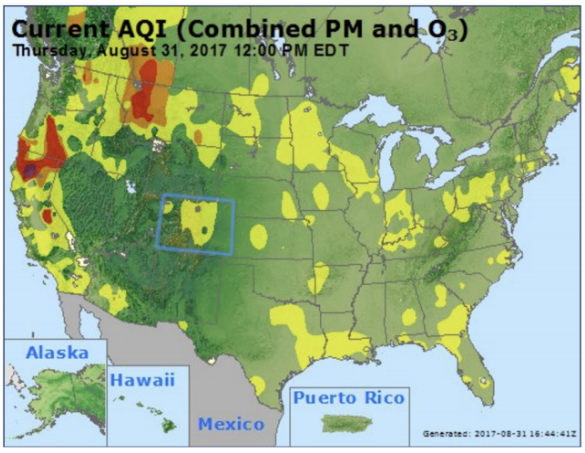

Figure 3: Air Quality map seen in class

6. One the most inspiring visualization we found online for our project was from the Air Pollution Quality Monitoring website *PurpleAir.com*. This website is key for our project because we also extracted the data from it. We would like to reproduce some of the visual encoding they use not only for pollution but also for temperature and humidity.

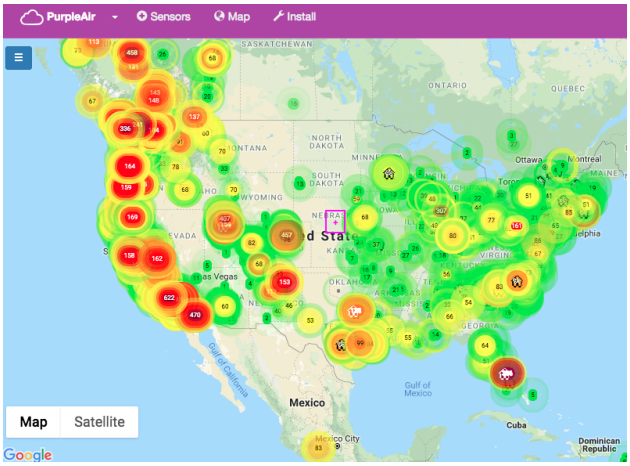

Figure 4: Purple Air visualization of North America Air Pollution Quality

3 Questions

- The first challenge/question we encounter at the beginning of this project was how to create a very friendly and informative visualization about the air conditions (temperature, humidity, and pollution). This visualization should help users to have a general idea about the air conditions in these cities in the simplest and interactive way possible.

- Were to find the right data about air quality condition?.
- Is there a noticeable relationship between pollution and temperature? What about pollution and humidity?.
- What would be the best time of the year to study the variation of these aspects of the air quality condition?.
- If we have to restrict our study to one state in the U.S. Which one would give us the best data to study these correlations?
- How to encode these tree aspects of air condition quality following the principles of design studied in class.

4 Data

We are downloading our data from the PurpleAir website (https://www.purpleair.com/sensorlist). Our data includes three different levels of pollution, but we are only focused on the PM2.5 particles level measured as CF ATM ug/m3. PM2.5 is a microscopic particle2.5 microns in width and almost 30 times smaller than the diameter of a human hair. When levels are high, PM2.5 particles form a haze in the sky, making their way into peoples respiratory tracts and reaching the lungs [3].

Our data also includes humidity(%) and temperature(F^o). Each entry in the data sets is recorded by sensors every one to two minutes. There are many different data sets that are available to us. We will focus on the vicinity of downtown (including it) areas for the major cities in California. There are about 20 to 30 sensors for each city. Our data is the format of csv files.

For the related statistics section, we obtained the population, number of registered vehicles, number of industries, and commute time for every city from the "DEPARTMENT OF MOTOR VEHICLES ESTIMATED VEHICLES REGISTERED BY CALIFORNIA COUNTIES" for the period of January 2017 through December 2017. Link, and the DATA USA website Link.

5 Data Processing

1. We had three factors to consider in order to find the best set of cities in California for this project.
   (a) The cities had to be well distributed all across the state. Ideally, from north to south and varying from big cities to rural cities in order to obtain a good variation in the values of the weather and pollution variables studied.

(b) Many cities in California do not have enough sensors to get an accurate average of the air condition in them. This is why we needed cities with at least 20 sensor working properly and constantly during the time frame our study is focused.

(c) Not every city in California has an available representation in geo.json maps for d3. This is why we had to choose cities that we could accurately show in our state map.

Under this constraints the optimal set of cities we found was: Sacramento, Eureka, San Francisco, San Jose, Bakersfield, Fresno, Los Angeles, and San Diego.

2. For each city, for each sensor data-set (between 5 and 10 csv files), we processed it to get the averaged pollution, humidity, and temperature in increments of one hour. Obtaining new one-hour averaged sensor data-sets (csv) corresponding to every sensor. We initially were going to have them in ten minute increments, but when we did that, our visualization took too long to display. So that is why we reduced it to one hour increments.

3. For every city, we created a new data-set with the averaged pollution, humidity, and temperature from all the corresponding one-hour averaged sensor data-sets obtained in the previous step. Obtaining eighth final data-sets containing the averaged value for each variable in increments of one hour for each city. These are the data-sets we are using for our visualization.

4. Given the big amount of data, the inconsistencies of availability of the data of the sensors, and the amount of time it takes to find a period of time where all the selected sensor have available data to analyze, we decided to use only the data for December 2017 which we checked and it was available for the set of sensor we are considering (see next section).

6 Exploratory Data Analysis

We used excel sheets and filtering, and averaging to analyze the data. Due to the difficulties in finding a set of sensors with recorded data on the same continuous period of time, we ended up having continuous data from 2-months, 1 and half and monthly from different years 2014-2018. After analyzing this data in excel, we found out that the period with better variations on the air aspects across all of the selected cities was November 28th 2017 to January 5 2010. Hence, we chose this period of time to be shown in our visualization and, in order to make the output more standard, we decided to use the period of time from December 1st to December 31st 2017.

7 Design Evolution

Given the references we found, we first discussed and sketched different possibilities for our visualization, then we created a first alternative digital prototype design.The following figures, Figure 5 and figure 6 show this work.

Some of the problems we found with these first approaches were:

- Occlusion in case of adding more cities to the visualization. Furthermore, the little bar charts and the name of the cities would not be readable.
- Different units. Even though the intention of the bar chart is to show simple proportions of each of the values of the air conditions (pollution in blue, humidity in green and temperature in purple), the units of each of them are different. This is not a very good way to encode this values at once. We tried to encode this three values in different ways but none of them convinced us (See sketches in appendix).
- In general one of the main problem we faced was to encode the three variables Temperature, Humidity and Pollution on the same map. We tried several options. Some of out different initial approaches to solve this can be traced in the sketches of the appendix.

7.1 Proposed visualization

We tried to fix some of this problem on the following visualization which was the one we considered as final on our project proposal. (Figure 7)

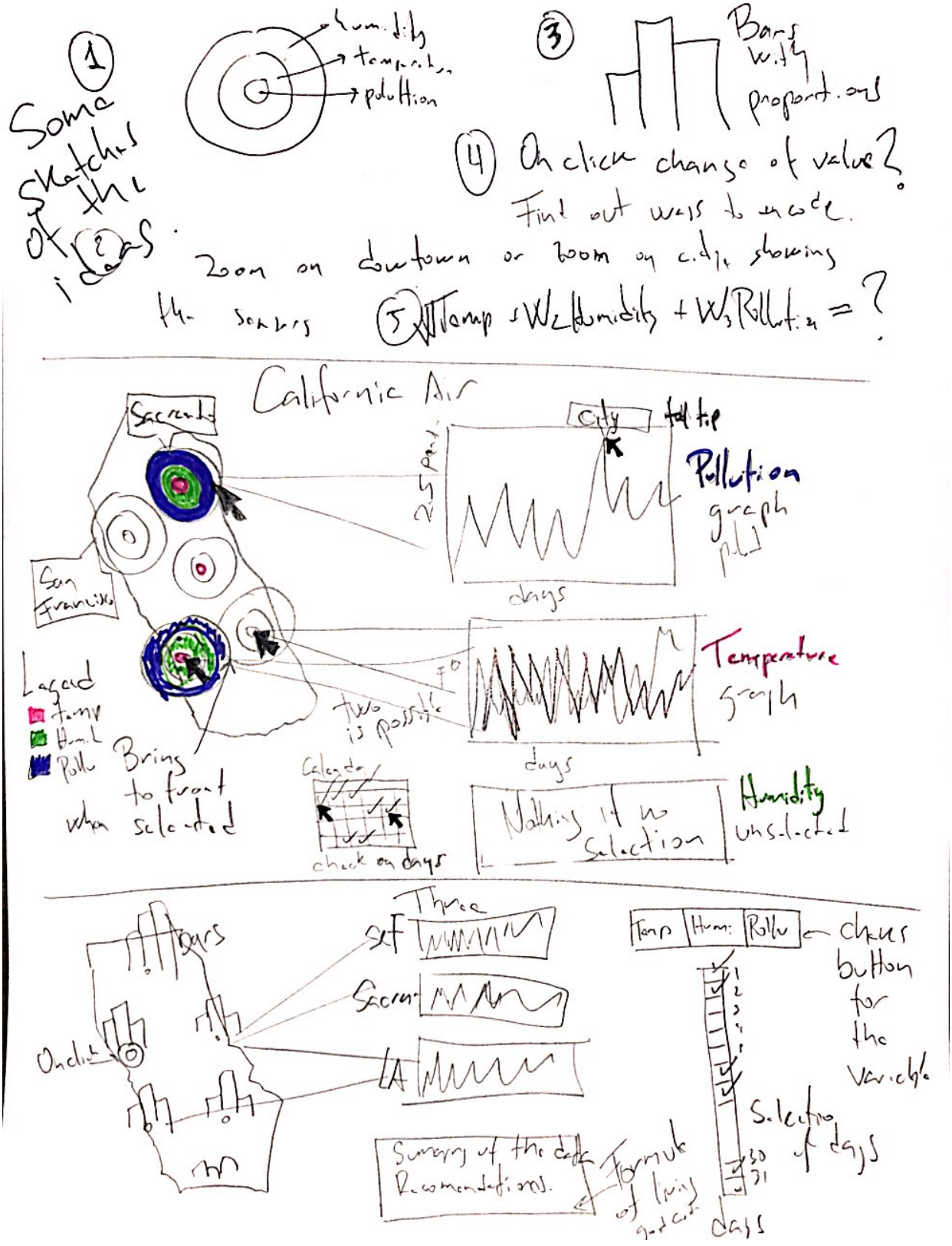

Figure 5: First prototype designs sketches

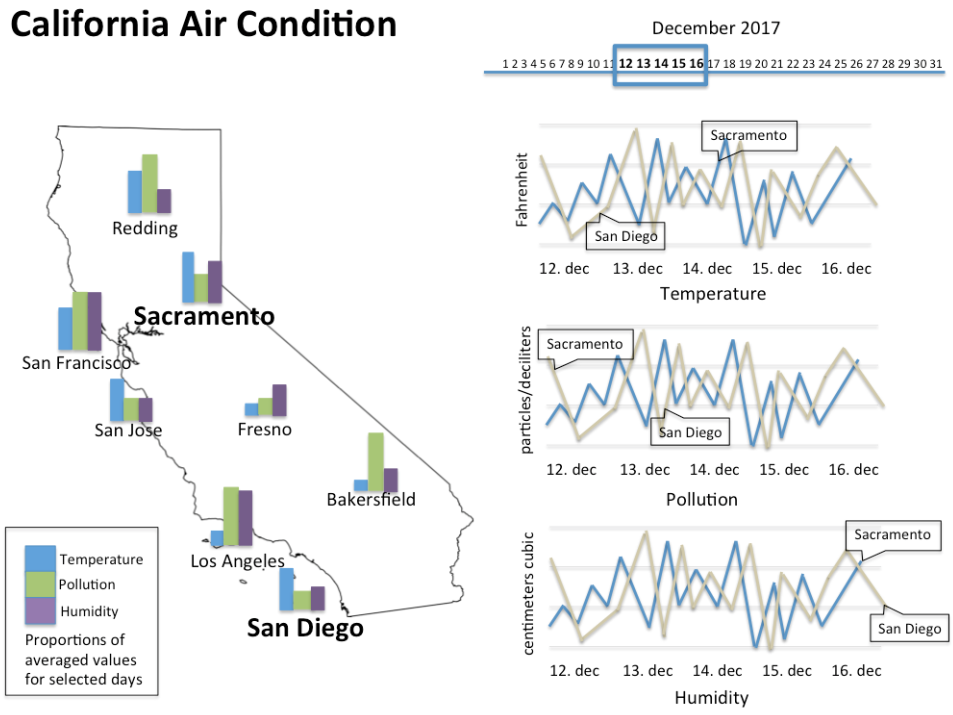

Figure 6: Prototype digital design

The decisions we proposed for our visualization the final design are the following:

1. Three buttons to select the condition of the air that the visualization will work with.

2. A brush bar to select any sequence of days. The visualization will use these information and generate the average of the values during that period of time.

3. Use the shape of the California state as an spatial region. This could be seen as a mark, but is also a channel to convey to give the user a better general sense of the location and distance between the cities the visualization is showing.

4. Circles to indicate each city (marks). This circles will be colored according to the value of the condition the visualization is currently showing. There is a color saturation bar with a specific hue gradient for each condition. Even though the saturation of color can be effective in showing the intensity of the selected air condition, we decided that the circles will also include the number of units that the color is encoding. Note that this last step is redundant but we think is important in order to the lecture of the data even easier to the user.

5. Once the user interactively clicks (or hover) on a city, a plot of the time in function of the current selected air condition will show up. This plot will use time by hours. In order to compare the selected air condition on different cities the user can click more than one city and the graph for both cities will show up on top of the other.

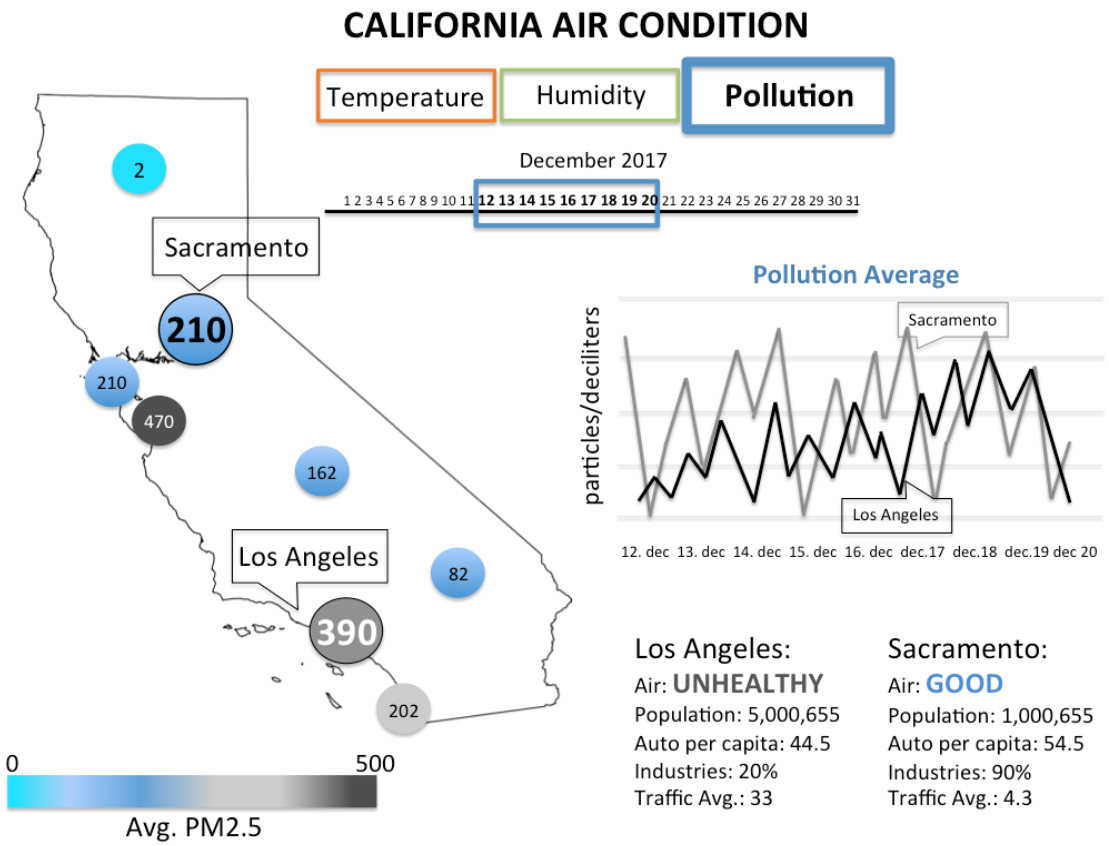

Figure 7: Final design proposed

6. We are displaying more information of the city: population, number of registered vehicles, traffic, etc., as well as the final description, number of industries, commute time. This is information summarizing the data for the current selected condition according to the data during the selected period of time.

7. We are using the averaged data from the chosen period of time, to calculate a summary of the condition of the Air based on the three factors we case about. This summary will be a categorical/rank channel with four possible values depending on the output number. We decided to include this following the Data and View Specification (Derivation) from the Taxonomy of interactive dynamics for visual analysis.

7.2 Peer review and suggested changes

We received suggestions from our peer students Maitrey Mehta, Sai Varun, and Shalin Parikh.

1. On the first proposal for our visualization, we were going to use the average data of the sensor on every city per day for every data variable: Pollution, Temperature, Humidity.

Our students peers found a problem on this approach and out intention of having a line chart plotting this data.

If the user select only one day, the plot would consist on only one point. In general, for a small number of days selected, the plot was not going to be very helpful for purpose of the visualization.

They suggested us to take the average of the sensor every hour or every 10 minutes. This would make our plot chart more meaningful and will show more precisely the behaviour of the variables.

2. Our line charts are going to show only the information of two selected cities, since we think that if the data from more than three cities is displayed we may deal with undesired occlusion. However, they suggested us to try to include at least one more city on the selection to encompass more data insight for the user.

3. Another suggestion was us to change the bar gradient color for the pollution visualization. Our choice was from light blue to black. They think this choice of colors was counter intuitive.

4. They liked our prototype visualization 3 so we also discussed ways to fixed the problem we encountered with it. They suggested us to use an area chart, similar to the "Bill Gastes" one used in class for design critique, to represent the proportion of each variable on top of every city.

7.3 Final visualization

Our final visualization includes all the proposed features with some extra/different approaches.

1. We followed and implemented all the suggestion from our peers review except adding the area chart from the prototype design.

2. Instead of comparing only two or three cities, we decided to allow the user to compare as many cities as needed. We did this by following the Data and View Specification from the Taxonomy of interactive dynamics for visual analysis.

3. We added a legend to properly label the data following the Tufte's integrity principles.

8 Implementation

We implemented a smart and well interactive website for California air conditions. We implemented the features of our visualization in the following chronological order:

1. Divided the website into three parts(title, map, wave chart) by using div tags.

Figure 8: website layout

2. Imported California State and cities map by using Geo.json data in order to avoid overlap problems.

3. Line charts interacting with the data of the city selected (Figure 9): We were having some problems with the loading time of these charts. That's the drawing chart function kept drawing functions. We used an array to keep all the statement of the city(selected: True, un-selected: False), then call the update function to refresh the charts.

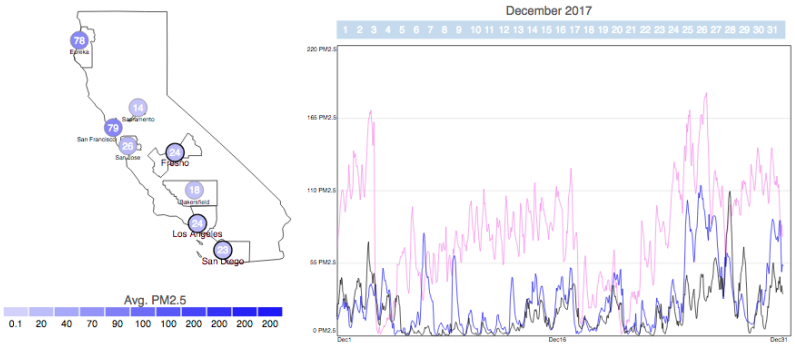

Figure 9: Line chart map interaction

4. We used D3 to help us implement the brush bar which is used to change the range of time to consider for the values of the visualization. Each time we change the brush bar which will trigger the update function to change the related data in map and chart. (Figure 10).

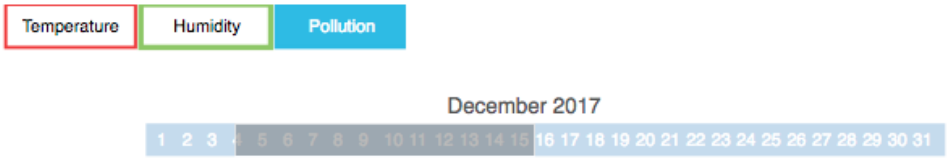

Figure 10: Brush bar and toggle buttons

5. We toggled the buttons to change the features of air condition which would trigger both the map and the wave charts updated. (Figure 10).

6. In order to make it more convenient to check the map, we created circles on the map.It has three functions: 1. the best way to figure out the select and unselected cities for the comparison processes which will cause the wave charts updated; 2. with the related color, we can compare its air condition according to its legend; 3. check the air condition value easily.(Figure 11).

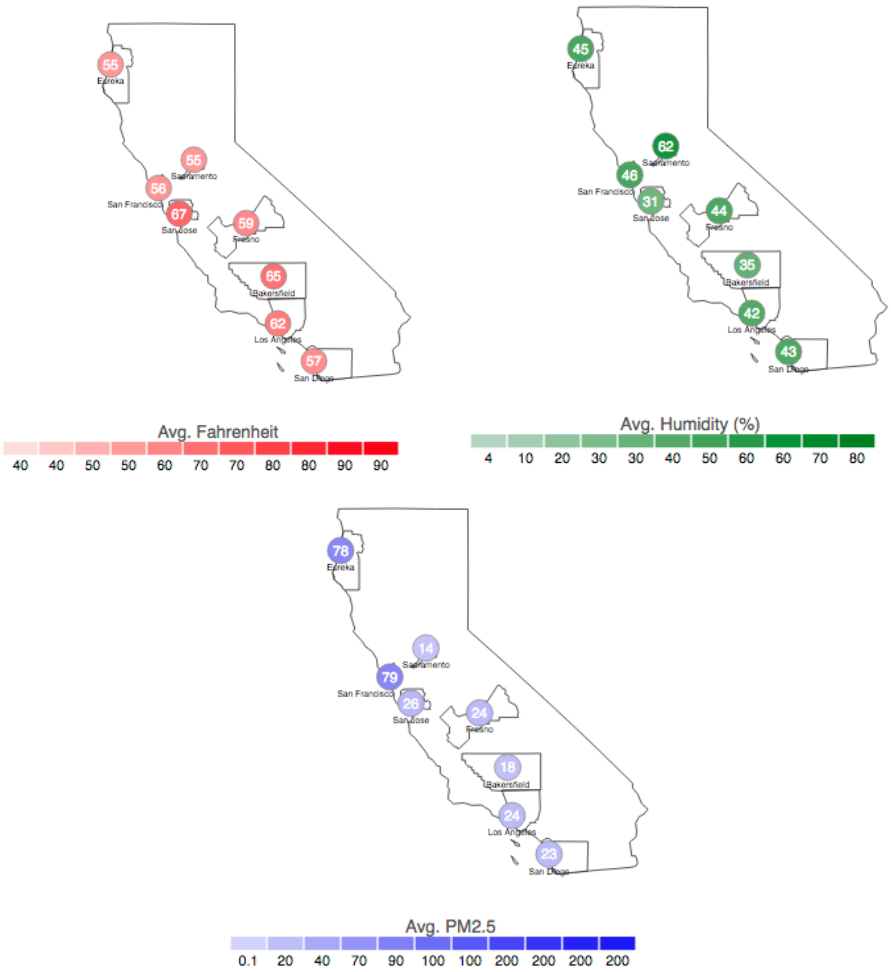

Figure 11: Map displays

7. Code for the city information area with suggestions for the user (Figure 12).

8. Gradient of the intensity of the aspect below each of the current map.

9. Line legend below the line charts for the comparison of cities (Figure 12).

Figure 12: City information and air condition

9 Evaluation

1. What did we learned from the data by using the visualization.

   - Even though San Diego is the second largest city in California in terms of population, its air quality is very good. The correlation between Pollution, Temperature and Humidity is almost optimal.
   - Cities with less population, less traffic and less industries like Bakersfield and Fresno do not have good air condition. This may have to do with geographical conditions.
   - In general, cities along the coast of California tend to have better air conditions. This may have to do with the fact that are the most populated.

2. Final visualization.

   The missing data from the sensors made our visualization very limited. But we think that achieved the goal of presenting the air condition data in a intuitively and friendly user oriented way by following the perceptual and design principles you learned in the course as much as it was possible.

   Some of the interaction between the map the info boxes below the line chart are not working properly.

3. Further Improvements.

   - Apply our visualization on data from several years, in order to find more significant trends and correlation between the Temperature, Humidity and Pollution.
   - This visualization could be expanded to other states.
   - Study the correlations between the aspects more in detail.
   - This is a visualization that summarizes the air condition in the past. We could try to add Machine Learning techniques in order to predict the air condition of the cities the user selects.

References

[1] Anton E. Kunst Casper W. N. Looman Johan P. Mackenbach. Outdoor Air Temperature and Mortality in the Netherlands: A Time-Series Analysis. American Journal of Epidemiology, Volume 137, Issue 3, 1 February 1993, Pages 331341, https://doi.org/10.1093/oxfordjournals.aje.a116680

[2] Infoplease. Effects of Dry Air on the Body. https://www.infoplease.com/science-health/weather/effects-dry-air-body

[3] The World Bank. IBRD.
Understanding Air Pollution and the Way It Is Measured. http://www.worldbank.org/en/news/feature/2015/07/14/understanding-air-pollution-and-the-way-it-is-measured