Visualize Air Pollution Data of California

Haotian Zhang, Qianhui You, Jing Peng, and Fan Pan

University of Southern California, Los Angeles CA 90089, USA,
zhan559,qianhuiy,jingp,fanpan@usc.edu,
home page: http://scf.usc.edu/~fanpan/project2/index.html

Abstract. Air pollution has long been a concern in California. As the public become increasingly health conscious, people will have a desire to be more informed about the quality of the air we breathe everyday. While the hourly measurement of air quality across the country is fully available on Environmental Protection Agency (EPA)s website, information transparency is not achieved until its presented in a digestible form to a broader audience. The purpose of this project is to apply various data visualization forms and tactics to present recent air pollution status in California. We used visual elements including heat map, bar chart, line chart and sunburst diagram, to visualize data from various aspects.. . .

Keywords: Data Visualization, Air Pollution, Heat Map, Bar Chart, Line Chart, Sunburst Diagram

1 Introduction

Air pollution is one of the most serious environmental concern in California. The severe physical consequences of air pollution make it necessary to educate and inform people with air quality issue. Data visualization is an ideal way to present this issue since it is more interactive, memorable and timely. In our project, we combined a variety of visualization formats and tactics to introduce air quality in California from different perspectives. This project is made to be digestible to average people and at the same time provide additional deeper information to sophisticated users.

1.1 Background

The official documentation of air pollution in California can be dated back to early 1940s, when the first episodes of smog was recognized in Los Angels in 1943. People were recorded to suffer from different symptoms such as respiratory discomfort and vomiting.[1] Since then, Los Angeles, as well as many other industrial cities in California, has been notoriously known as one of the most polluted cities in the U.S. Fortunately, air pollution has been better controlled in recently years with more restricted regulations on both industrial and residential emissions. Yet, as California remains as the most polluted state in the U.S., its still necessary to help the public gain more knowledge on the issue through informative visualization.

1.2 Potential Audience

The primary audience of our visualization is California residents, especially those who live in the urban areas. We use different tactics and design to make the visualization more comprehensible and intuitive to even people with minimum quantitative background. While the visualization we chose aims to help average users, the information presented also carries enough depth to attract more sophisticated audience, such as environmentalists and public agents, who might use it as a supplemental tool to analyze the issue.

2 Functions and Visualization

We believe function should guide visualization. Therefore, when we designed the website based on our dataset, we started from defining the user journey and a hierarchy of information. The journey unrolls in four stages: an overview of geographical distribution through a heat map; quickly obtain ad-hoc information via a real-time dashboard; high-level insights from three groups of charts demonstrating a range of perspectives on the issue throughout a longer timeframe; additional health information for user who have specific needs. In addition, interactivity is enabled across all sections to engage users to explore and tell their own stories.

2.1 Dataset

The data we used to build the visualizations comes from EPAs database[2]. We processed the raw data to make it more meaningful and visually presentable.

The EPA Criteria Gases database provides hourly measurement of four types of air pollutants CO, NO2, SO2 and Ozone from stations across the country since 1980. We chose 2016 and California as our focus. From the 195 stations in 49 counties in California, we selected 14 counties that are located in the most populated areas and have records for all four air pollutants.

In the original data, sample measurements for different pollutants are presented in ppm (parts per million) or ppb (parts per billion). The absolute measurements make it difficult to compare between pollutants. For example, although one ppm of CO might not be a serious concern, one ppm of Ozone could be fatal. To standardize the data, we take the absolute measurement of different pollutants and divide them respectively with their pollution standards from National Ambient Air Quality Standards (NAAQS).[3] The end results are comparable pollution indexes, such that a value of 0.9 for any pollutant means its 90% close to being a pollution.

The two components that are most relevant to our visualization are time and location. In order to provide insights from various angles, we summarized data based on year, season, day of week, county and a combination of the above and used different summary data for different sections of the visualization.

2.2 Geographic Visualization

A **heat map** is the first visualization to greet the users in the journey. It intuitively shows the locations of the stations and the pollution level of the entire state. Heat map is an ideal way to provide a quick overview since it requires very little brain processing to understand the information. When moving the mouse over or clicking on each of the locations, we will direct users to the next level of information, which is the real-time dashboard.

2.3 Relevance

The **real-time dashboard** consists of three components: a line chart, a pie chart and a bar chart. The chart types we chose are very high in familiarity to serve the goal of providing quick understanding of current air quality in current location selected. This dashboard will answer some ad-hoc questions users might have to help plan their immediate activities. To further ensure relevance, the bar chart only shows four adjacent counties as a comparison.

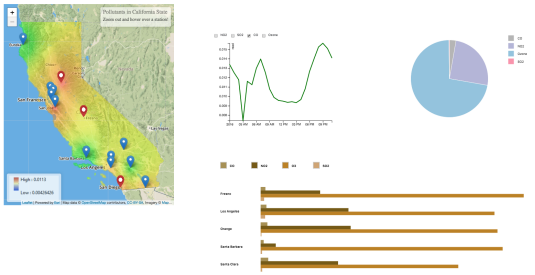

Fig. 1. Real-time Dashboard

2.4 Insight

The insight stage includes three chart groups, each provides summary information on a main dimension. The information presented in this section is aggregated to reveal consistent patterns throughout the year and help users make long-term, strategic planning.

Geographical Comparison. A **bar chart** is here to allow users to compare the air quality in all 14 counties. Users have the flexibility to add or remove one or more pollutants to amplify the similarity or difference.

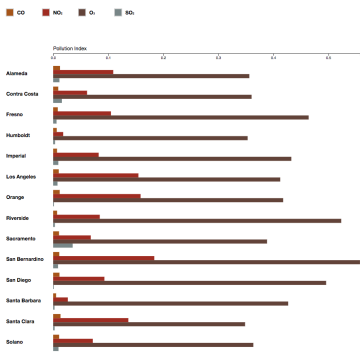

Fig. 2. Bar Chart

Pollutants Comparison. To compare how much does each of the pollutant contribute to the overall air quality, we used the **sunburst diagram** to show the respective proportions in different seasons. Besides the primary chart, small multiple also makes it easier to compare pollution in each day of week in different seasons.

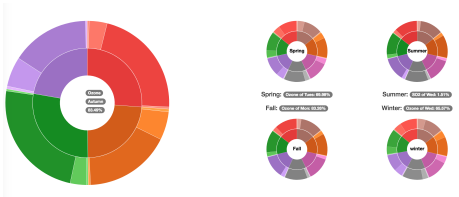

Fig. 3. Sun Burst

Trends Comparison. The **line chart** displays hourly trends at different aggregation levels. As the main graph shows the yearly/seasonal average of each hours pollution index, small multiple creates a quick way to spot differences between days of week.

2.5 Additional Knowledge

The last section of the visualization provides additional physical harm information to users who have specific interests or needs. Users can interact with the **infographic** to find out how can air pollution affect each organ. This supplemental knowledge adds flavor to our visualization and creates a sense of emergency and relevance to draw peoples attention to the issue.

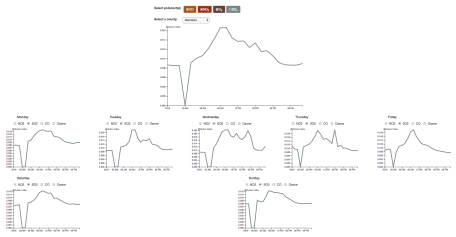

Fig. 4. Line Chart

3 Development Technology

For data wrangling part, we used R for data processing, which includes using lubridate to process timestamps, dplyr to filter and summarize data, and ggplot2 to do exploratory analysis to make further decision on data processing.

For web development part, we primarily used Bootstrap to build the frame of our website and CSS to style it.

For visualization part, we implemented graph drawing, transition and interaction using JavaScript with d3 and jQuery.

4 Discussion and Conclusion

Visualization is a functional art. Therefore, its crucial to start with function. Starting with a well planned user journey gave us a clear vision on what we want to achieve and made deployment and design guided. We also benefited a lot from peoples feedbacks, which helped us identify where peoples confusions were and how we could make this journey smoother. Due to technical limitations, there are still a few functions that we thought would add value to the website but couldnt be fulfilled with our current knowledge. For example, although we accomplished the visualization part of the real-time dashboard, we are still not able to feed real-time data. This will be something to continuously work on when we have more advanced knowledge.

References

1. California Air Resources Board (2017) Key Events in the History of Air Quality in California.
   https://www.arb.ca.gov/html/brochure/history.htm Accessed 20 Nov 2017
2. Environmental Protection Agency (2017) Pre-Generated Data Files.
   https://aqs.epa.gov/aqsweb/airdata/download_files.html Accessed 11 Nov 2017
3. Environmental Protection Agency (2017) NAAQS Table
   https://www.epa.gov/criteria-air-pollutants/naaqs-table Accessed 11 Nov 2017