Analysis of US Daily Air Quality Index 1980-2021

Department of Applied Data Science
San Jose State University
DATA 230: Data Visualization
Professor: Andrew H. Bond
May 17, 2022

By
Shrey Bishnoi
015945501

Table of Content

1.Introduction

1.1 Purpose of this Document

The purpose of this document is to provide a detailed project description of the analysis of daily US air Quality Index since 1980 which is designed to understand air quality patterns over the time and future air quality prediction by creating visualization on Tableau to get the brief statistics and analysis of AQI. This document consists of background information of the project, organization plan, dataflow architecture, development process, deliverables, charts and visualizations.

1.2 Intended Audience

This paper will be used as a guideline during the project. All project stakeholders are the intended audiences for this project:

- Professor Andrew Bond
- TA Ritanjali Jena
- Shrey Bishnoi

1.3 Scope

The goal of this project is to do a quick analysis on AQI data by using Tableau desktop to create some visualizations. This paper details the project plan and methods for utilizing Tableau to analyze daily US AQI data. The aim of the project, the project's organization plan, data gathering, and programs utilized are all included in the paper. The amount of data gathered for analysis is enormous, and it provides a concise snapshot of the air quality situation in the United States, including how clean or filthy the air is, and what health impacts may be a concern, particularly for ground-level ozone and particle pollution.

1.4 About The dataset

1.4.1 Definitions

This Dataset contains the following attributes:

- **State Name:** The name of the state where the monitoring site is located.
- **County Name:** The name of the county where the monitoring site is located.
- **Defining Parameters:** The name or description assigned in AQS to the parameter measured by the monitor. Parameters may be pollutants or non-pollutants.
- **Date**: The calendar date with which the AQI values are associated.
- **AQI:** The Air Quality Index for the day for the pollutant, if applicable.
- **Category:** AQI is divided into 6 category levels(Good, Moderate, Unhealthy for Sensitive Groups, Unhealthy, Very Unhealthy, Hazardous) describing the air quality which corresponds to various health issues.
- **Longitude:** The monitoring site's angular distance east of the prime meridian measured in decimal degrees.
- **Latitude:** The monitoring site's angular distance north of the equator measured in decimal degrees.

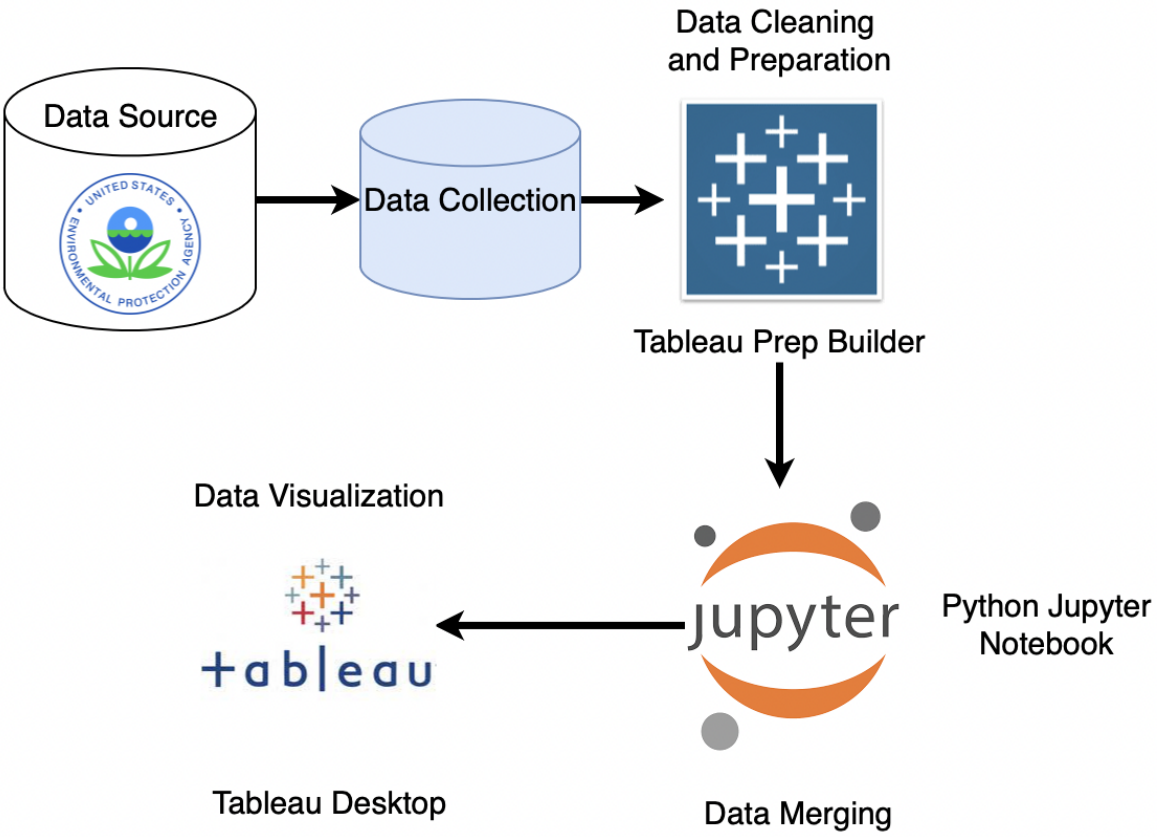
2. Background and Objectives

Air quality concentrations can change from year to year, impacted not just by pollution emissions but also by natural phenomena such as dust storms and

wildfires, as well as seasonal weather changes. Some pollutants are directly discharged into the atmosphere, while others are generated in the atmosphere because of chemical interactions. When NOx and VOC emissions react in the presence of sunshine, ground-level ozone is produced. Air pollution has several negative effects on human health and the environment. The information I'll be utilizing comes from the Environmental Protection Agency, and it contains yearly reports on air quality indexes from 1980 to 2021 in several US metro regions, as well as geographic data for the collection locations. This dataset has over 10 millions items with 8 properties, including geographical data like longitude and latitude. On Tableau, this dataset will be used to generate insights and geographic visualizations.

Tableau will aid in historical evaluation and visualization, as well as analysis of air quality patterns over time and future air quality predictions based on the visualization. Analyzing the factors which have improved the Air Quality since past years. As we progress through this course, I'll aim to learn and use more different sorts of visualization themes, such as heatmaps, to provide useful data. Following the completion of this project and the creation of various visualizations, it will be easier to convey the story about the air quality index patterns and future inferences.

3. Architecture and High level design



- The data for this project comes from a daily recorded AQI in the United States, which is managed by the US Environmental Protection Agency (EPA) to provide a simple and uniform manner to report daily air quality conditions. It is an open source government website where you can get data about environmental factors.
- The data collected from EPA is in csv format. There was a lot of historical data but we selected to use the data for our project from 1980 to 2021.
- Once the initial data is collected, all of the data is imported into tableau prep builder, where the data preparation has been done, cleaning the data, removing the outliers, handling inconsistent data and creating bins for AQI levels.

- After the data preparation the data is then exported to Python as a data frame so as to merge files and alter some columns using Python pandas library and converted into csv format.
- Finally the processed and formatted data file was imported into Tableau Desktop to perform analysis and visualization to get the brief insights about the data.

4.Organization

4.1 Project group

I, Shrey Bishnoi (015945501) studying at San Jose State University have done this project Analysis of Daily US AQI 1980-2021.

5. Development Process

The project's procedure begins with data collection. We wanted historical data on the US AQI for analysis. The information was gathered from the US Environmental Protection Agency (EPA), which had all of the data from the official database on the Air Quality Index that had been collected throughout the years.

Individual csv files were entered into tableau prep builder after the data had been obtained. Tableau prep builder is a Tableau tool that uses a graphical user interface to clean and process data. Data types for the columns in each file were changed according to the data it contained, such as converting the data type of the state and county columns from string to geographical location, and so on.

Multiple rows have been removed because they were having many missing values. So in order to do proper analysis we have to get rid of null values to get the

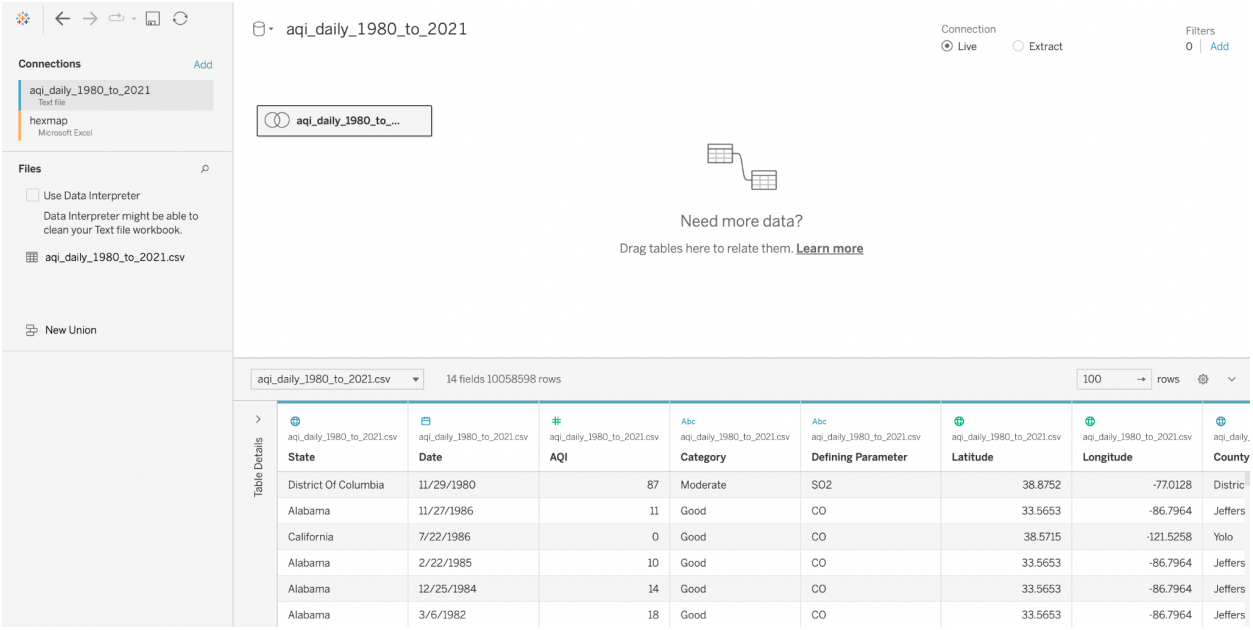
accurate results. Moreover the data has millions of rows so there were some outliers as well, to make the data in structure format we removed outliers.

All of the data files were imported into a pandas dataframe using Python once the data had been cleansed and was ready to be combined. The rationale for choosing Python was that the tableau prep builder was using a lot of resources and time, whereas python was much quicker as the data had 10 million rows which was difficult to load into prep builder.

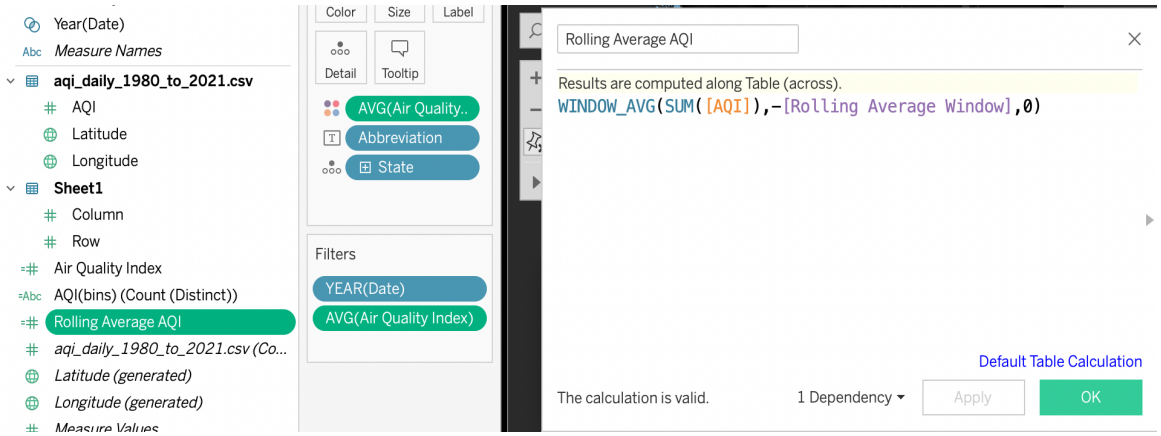
The csv files are imported into a python dataframe as shown in the diagram below. Once the data frames were created, data processing was done such as cleaning the data, removing duplicate rows and outliers and renaming some of the columns, handled inconsistent data by changing the field format to match across all data frames, which is required for combining the data.

```
In [1]: import pandas as pd

In [6]: df1 = pd.read_csv('aqi_daily_1980_to_2021.csv')
        df2 = pd.read_csv('aqi_yearly_1980_to_2021.csv')

In [8]: df1.columns

Out[8]: Index(['State Name', 'Date', 'AQI', 'Category', 'Defining Parameter',
               'Latitude', 'Longitude', 'County Name'],
              dtype='object')

In [9]: df2.columns

Out[9]: Index(['State', 'County', 'Year', 'Days with AQI', 'Good Days',
               'Moderate Days', 'Unhealthy for Sensitive Groups Days',
               'Unhealthy Days', 'Very Unhealthy Days', 'Hazardous Days', 'Max AQI',
               '90th Percentile AQI', 'Median AQI', 'Days CO', 'Days NO2',
               'Days Ozone', 'Days SO2', 'Days PM2.5', 'Days PM10', 'Latitude',
               'Longitude'],
              dtype='object')
```

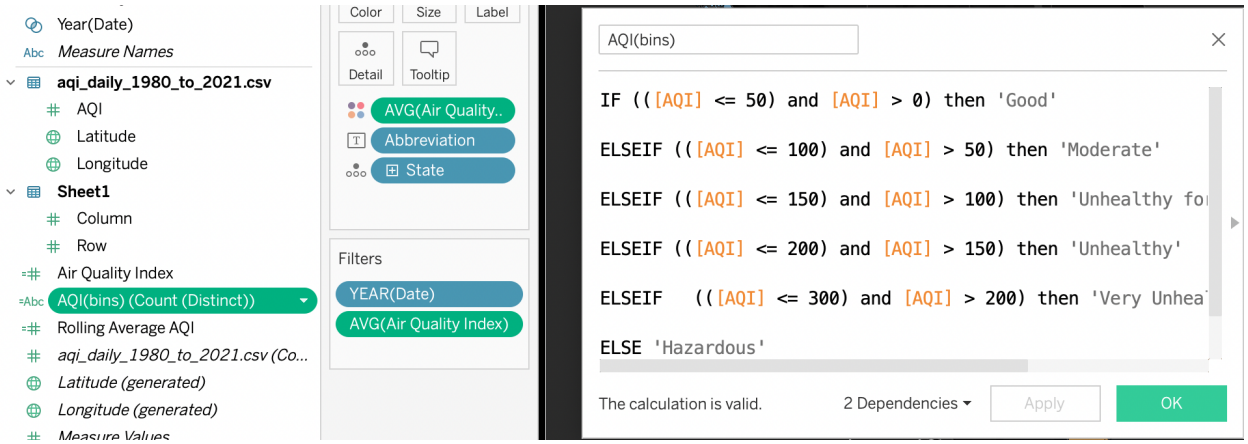
After merging the data files, Tableau was used to do analysis and produce visuals. The image below displays a sample of data that has been loaded into Tableau and is ready to be visualized.



Created some calculated fields on the data to get more insights. We calculated rolling average AQI to generate the average AQI values. There is a column called rolling average window which contains average values.



As I mentioned earlier, there are 6 categories in AQI(Good, Moderate, Unhealthy for Sensitive Groups, Unhealthy, Very Unhealthy, Hazardous) so we wanted to label all recorded AQI values into their respective categories. For this, we created 6 bins with different ranges as shown in the table below so that whenever you see any AQI value it will show which category it belongs to.



AQI Quality Values	AQI Category	Actions to Protect your Health
(<50)	Good	AQI value in this category is between 0-50 and air quality is considered satisfactory and poses little or no health risk.
(50-100)	Moderate	Air quality is acceptable ;however individuals who are very sensitive to air pollution may experience adverse health effects. People who are unusually sensitive to ozone or particle pollution may experience respiratory symptoms
(100-150)	Unhealthy for Sensitive Groups	People with lung or heart disease, older adults, children, and people participating in activities that require heavy or extended exertion may experience adverse health effects.
(150-200)	Unhealthy	Everyone may begin to experience adverse health effects and members of sensitive groups may experience more serious health effects.
(200-300)	Very Unhealthy	This would trigger a health alert signifying that everyone may experience more serious health effects.
(>300)	Hazardous	This would trigger health warnings of emergency conditions. The entire population is more likely to be affected.

6. Deliverables

The deliverables of this project are:

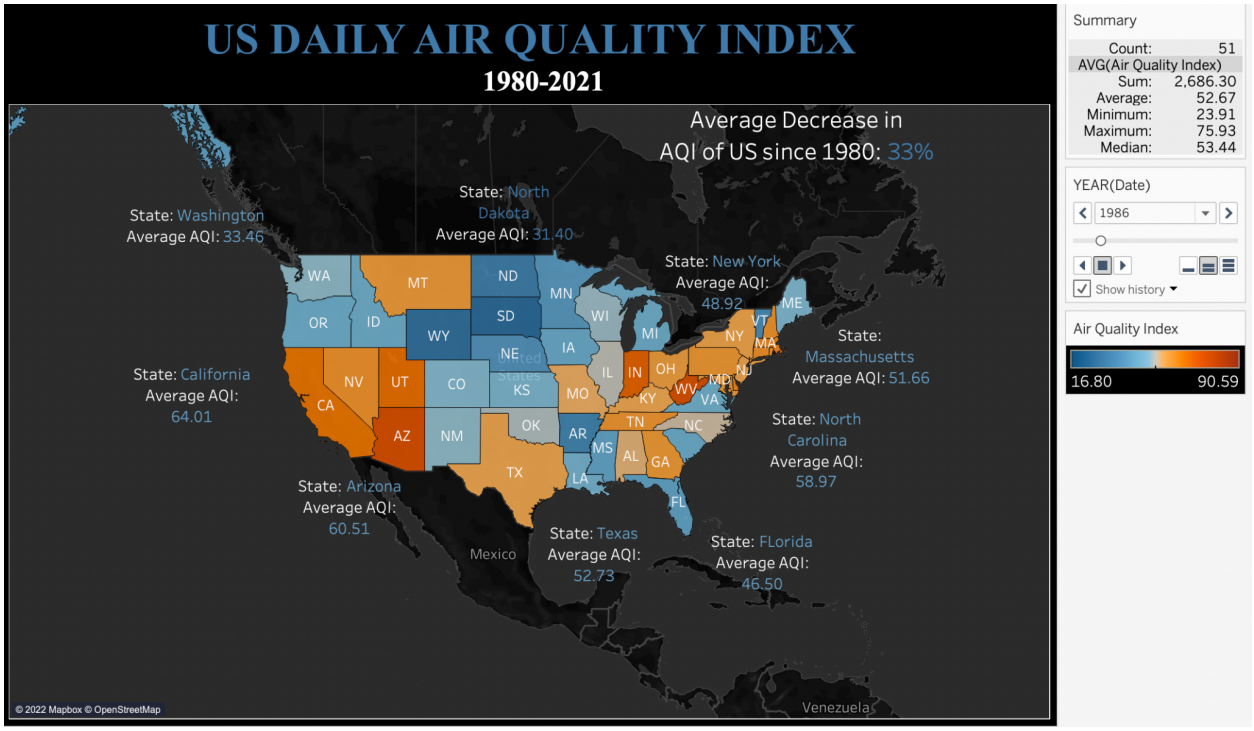
1. Tableau Dashboard: It gives brief analysis and visualization showing on multiple dashboards
2. Project Presentation: Detailed overview of the project which has been given in the class.
3. Final Project report: will be submitted on Tuesday May 17th which includes everything about the project like its background, architecture, project organization plan, development process and description of all the visualizations.
4. Code: Will be submitted by Tuesday May 17th, 2022.

All these deliverables will be uploaded on Github.

7. Github

Github Link: https://github.com/ShreyBishnoi96

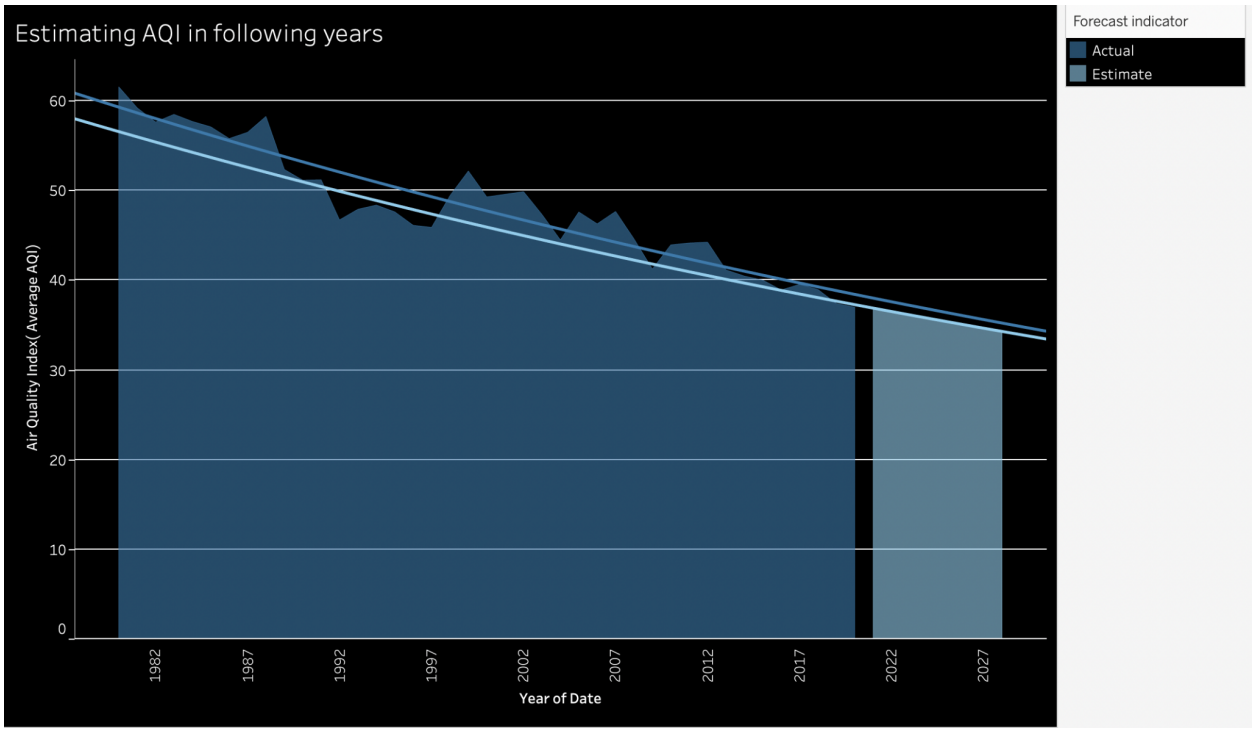
8. Charts



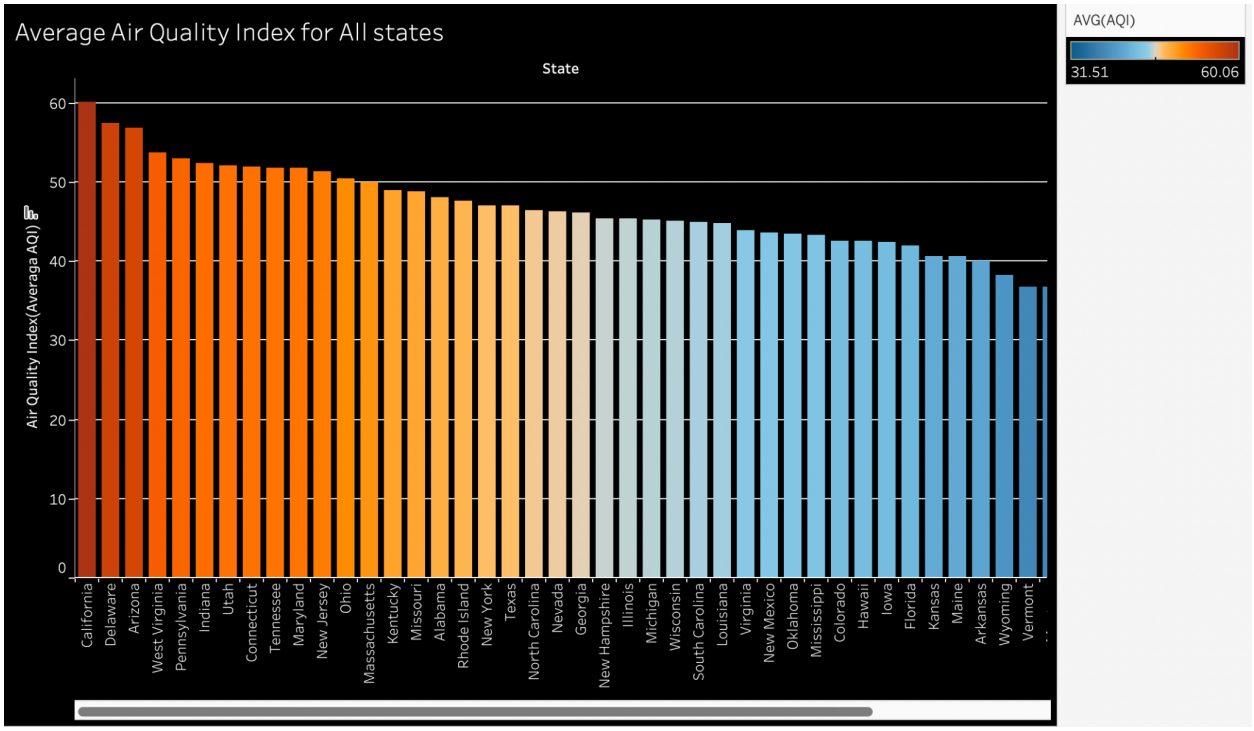
The overall emissions of the six main pollutants (PM2.5 and PM10, SO2, NOx, VOCs, CO, and Pb) decreased by 78 percent between 1980 and 2021. This breakthrough was made although economic indices in the United States remained robust. The quality of the air we breathe has improved dramatically as a result of the emission reductions. This is due to the Clear Air Act, which was enacted in 1970 as a regulation to enhance air quality.

The average decrease in AQI of the US since 1980 is 33% because of the Clear Air act. It is shown in the above visualization that the average AQI of all the states in 2020. Like New York had an average AQI 48.92 in 2020 which is considered as good and not at all harmful for people and the average AQI in the entire US was 52.67 in 2020.

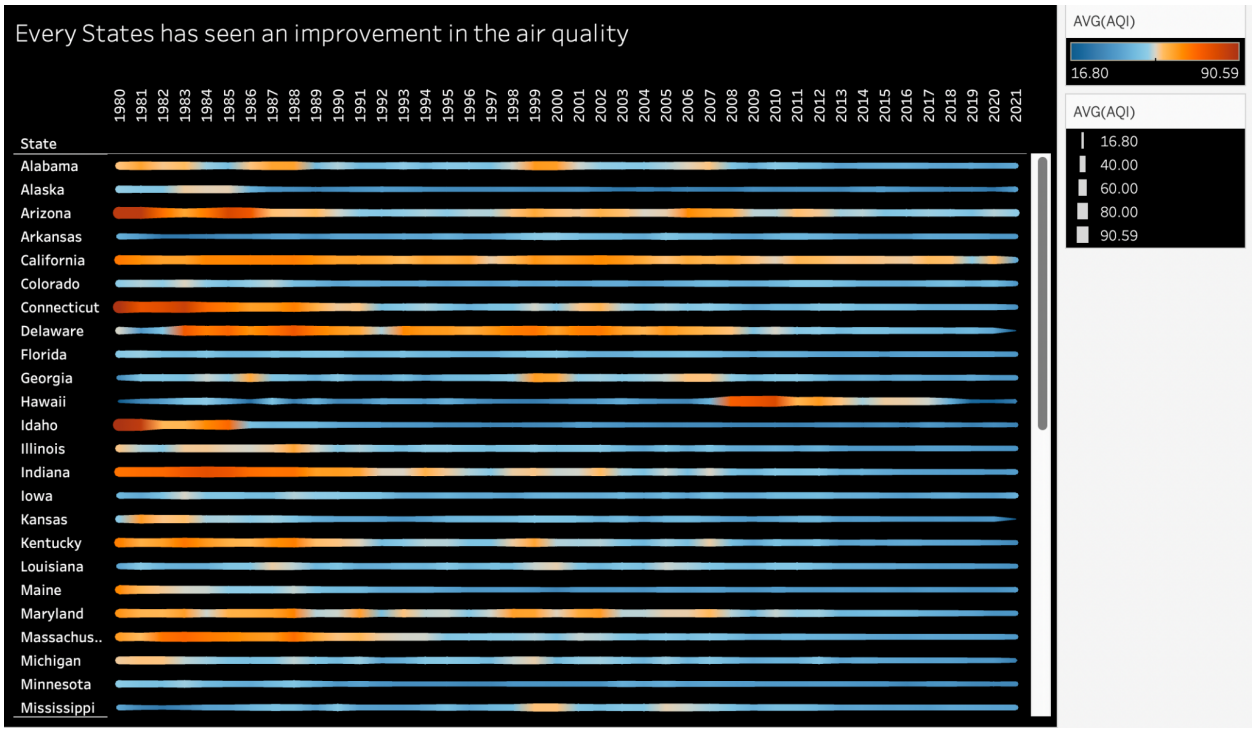
The above Visualization is a geospatial sheet which shows the average air quality index of every state in the United States. If you hover over any of the states it will show you the average AQI of that state.



The above figure is an area plot of average air quality index recorded since 1980. It is seen in the graph that the average AQI level has decreased since 1980 maybe because of the Clear Air Act implemented in the United States. Predicting from the previous historical data of AQI we can estimate that the AQI level might improve in the following years. The reason might be because everyone is moving towards electric cars and automation.



It can be seen from the above graph that overall California has the worst average air quality index whereas North Dakota has the best air quality. First and foremost reason is because California is a dry and hot state whereas North Dakota is a cold state. The above plot shows all the states in the United States in the order of worst air quality to the best air quality.



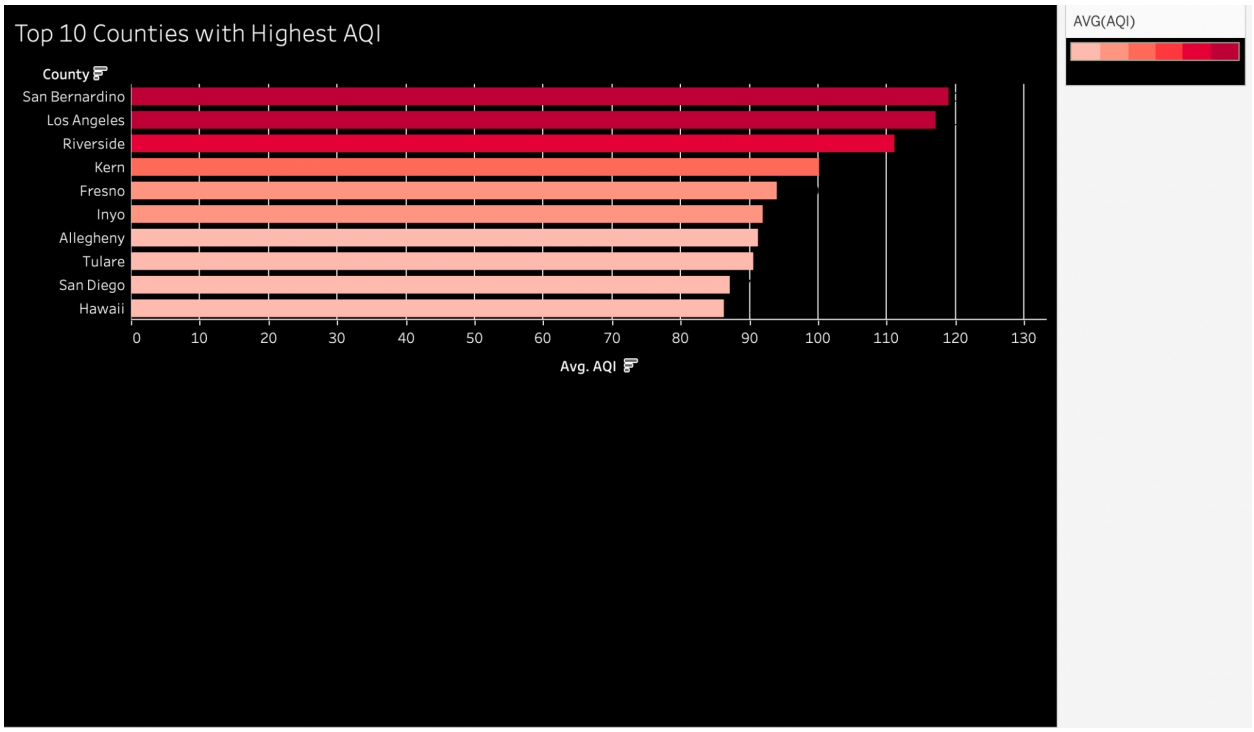
The above graph shows the average AQI in all the states from 1980 to 2021. Below are the few reading from the above plot.

• Almost all states had significantly worse air quality in the 1980s.

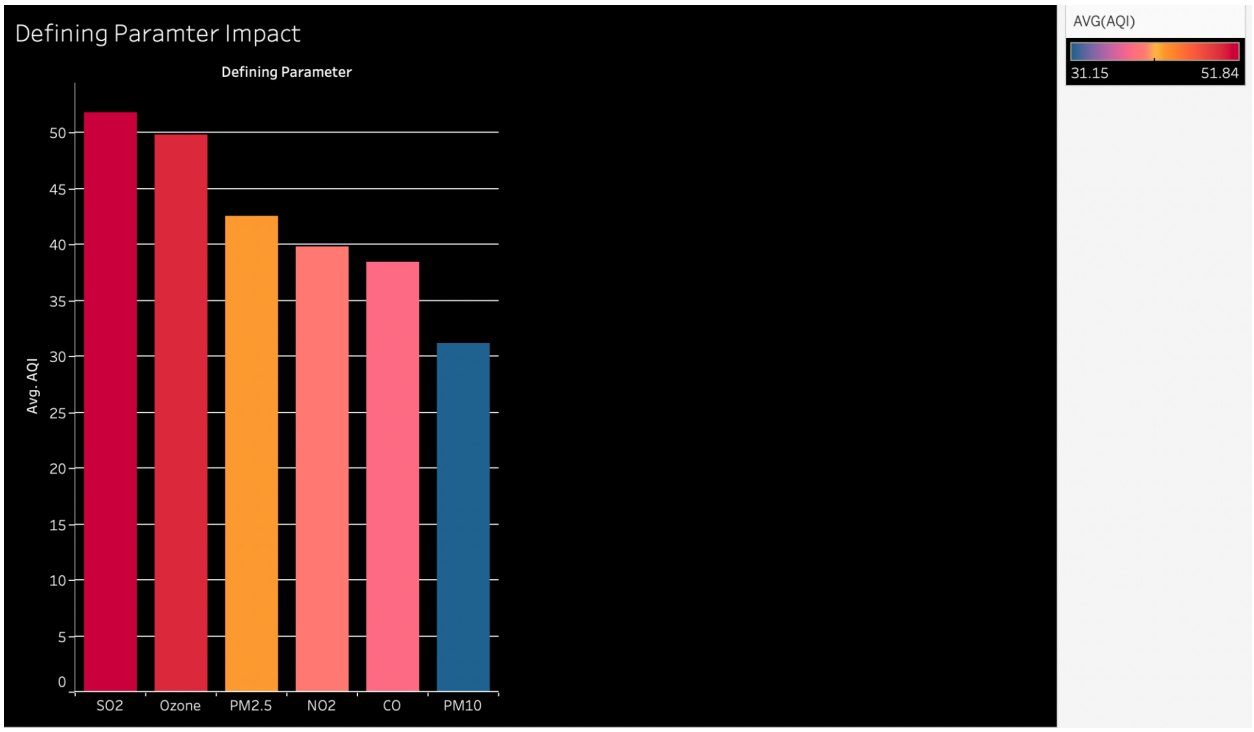
• California has consistently had poor air quality compared to other states.

• Wyoming, Vermont, South Dakota,Minnesota and Virginia have consistently maintained relatively good air quality compared to other states.

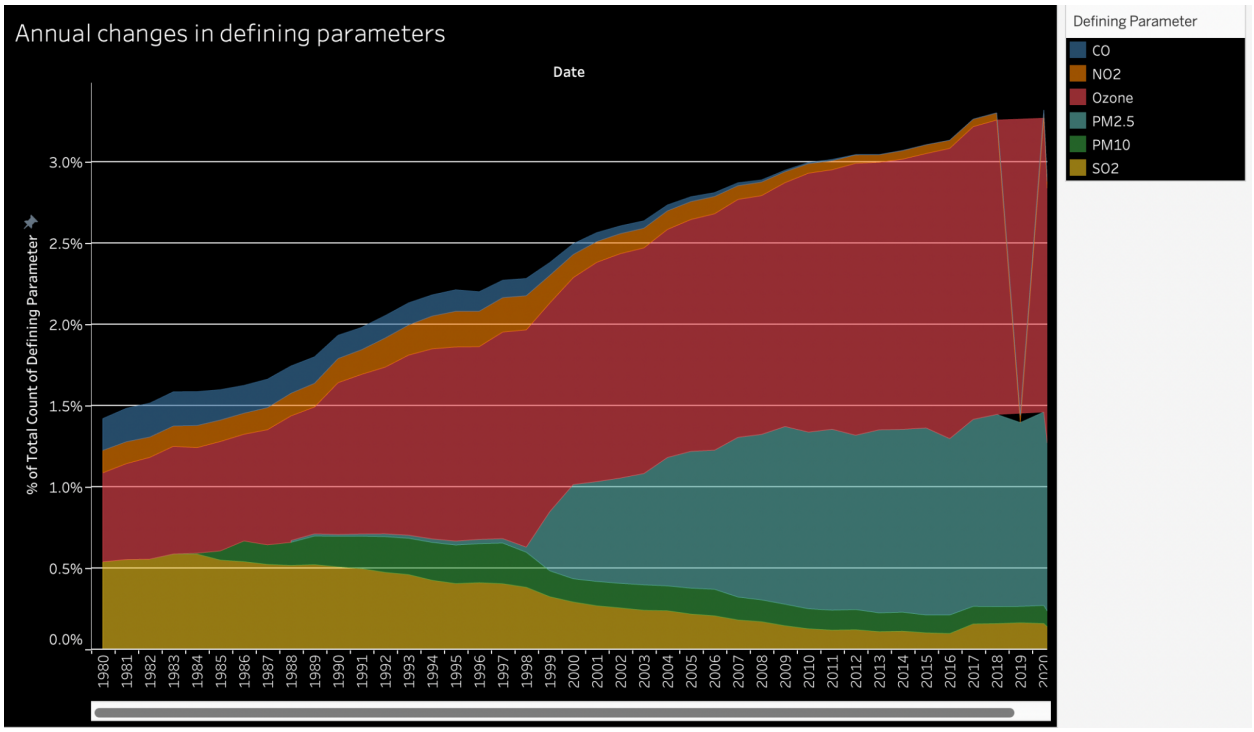
• Hawaii suffered significantly reduced air quality starting in 2008 and continuing through the next decade. Research indicates this is largely due to increased volcanic activity following the Kilauea Overlook vent eruption. This produced large amounts of volcanic smog in the form of sulfur dioxide (SO2).



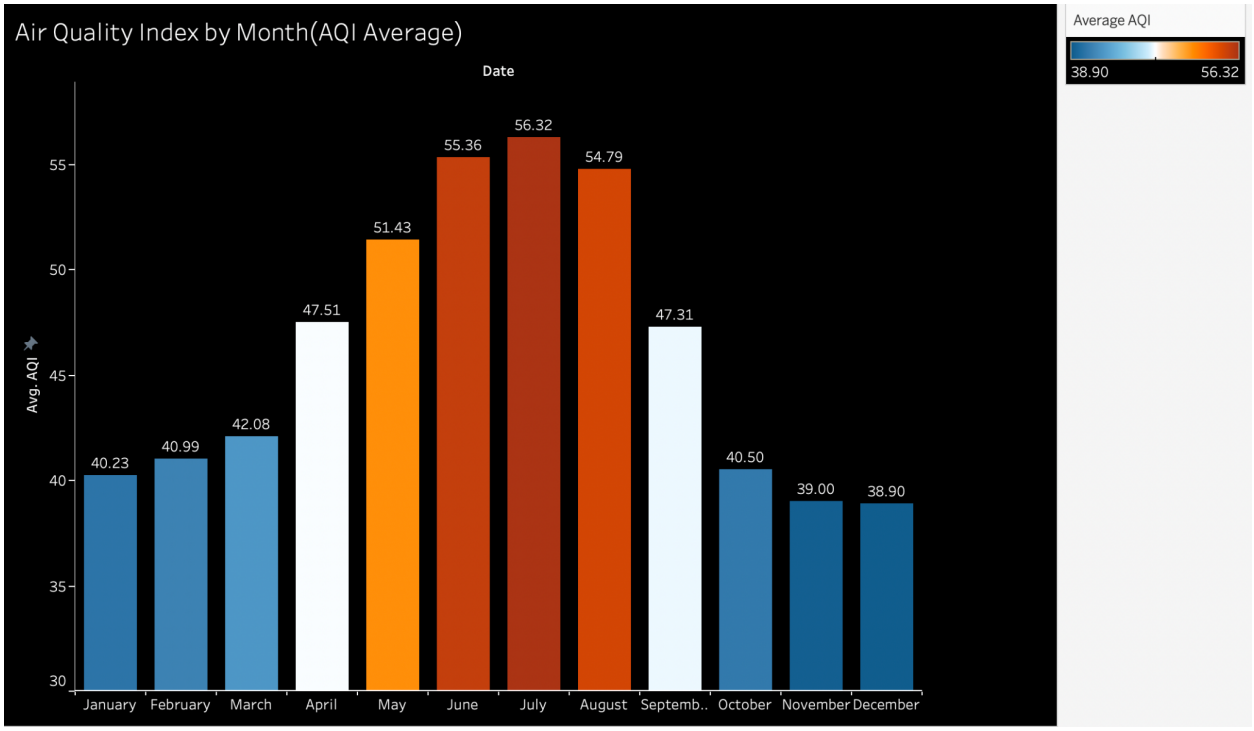
The above bar plot shows the top 10 counties in the United States with the Highest AQI level. It is seen in the above graph that San Bernardino, Los Angeles, Riverside, Kern and Fresno counties have the highest AQI. San Bernardino has the highest AQI of 119.52 and the second highest is Los Angeles with the AQI of 117.07. Most of these top counties are in California. California has the worst air quality as compared to all other states.



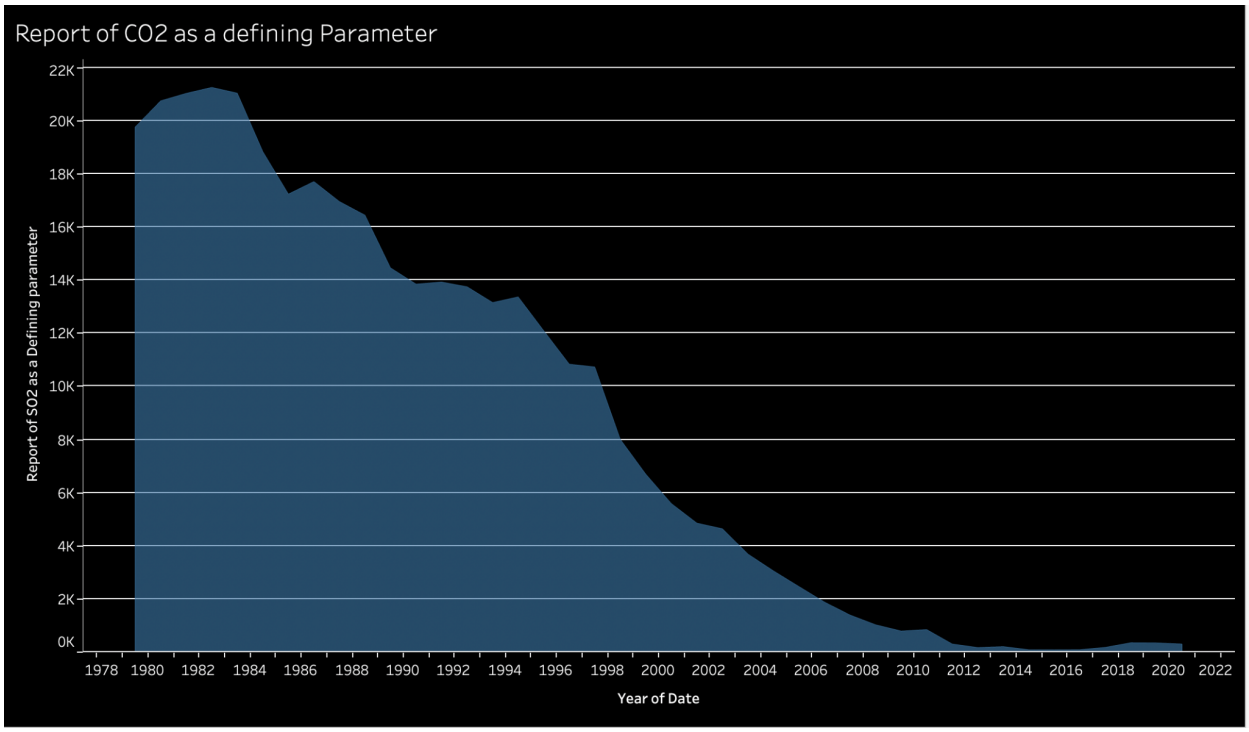
In the above bar plot it is shown that there are 6 defining parameters for air quality that are SO2, Ozone, PM2.5, NO2, CO and PM10. It is noticed that days in which Sulfur Dioxide was the defining parameter had the worst air quality of average AQI of 51.62 whereas days in which PM10 was the defining parameter had the best air quality of an average AQI of 31.15. This might be the case because for SO2 the largest source is the burning of fossil fuels which worsen the air quality.



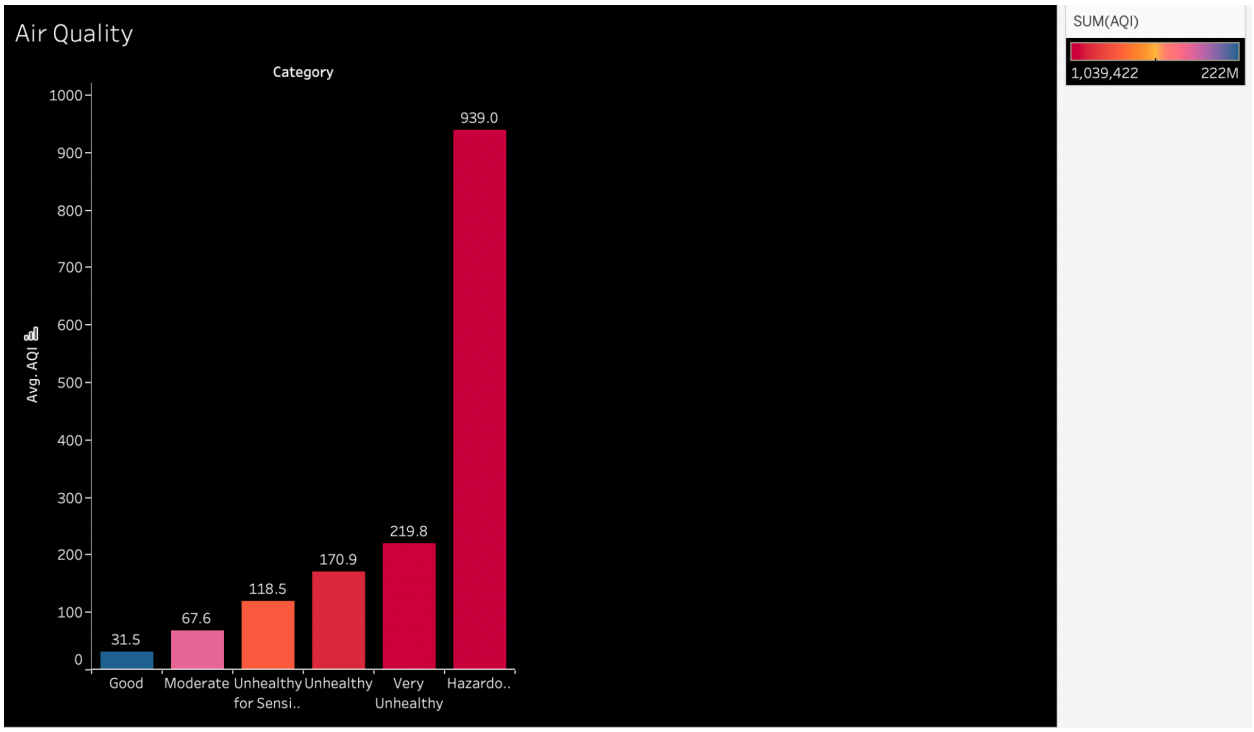
The above area graph shows the annual changes in the defining parameters of the air quality and it is seen that the red area which refers to the ground ozone level shows an increment since 1980 which might lead to worse air quality and can affect people as well.



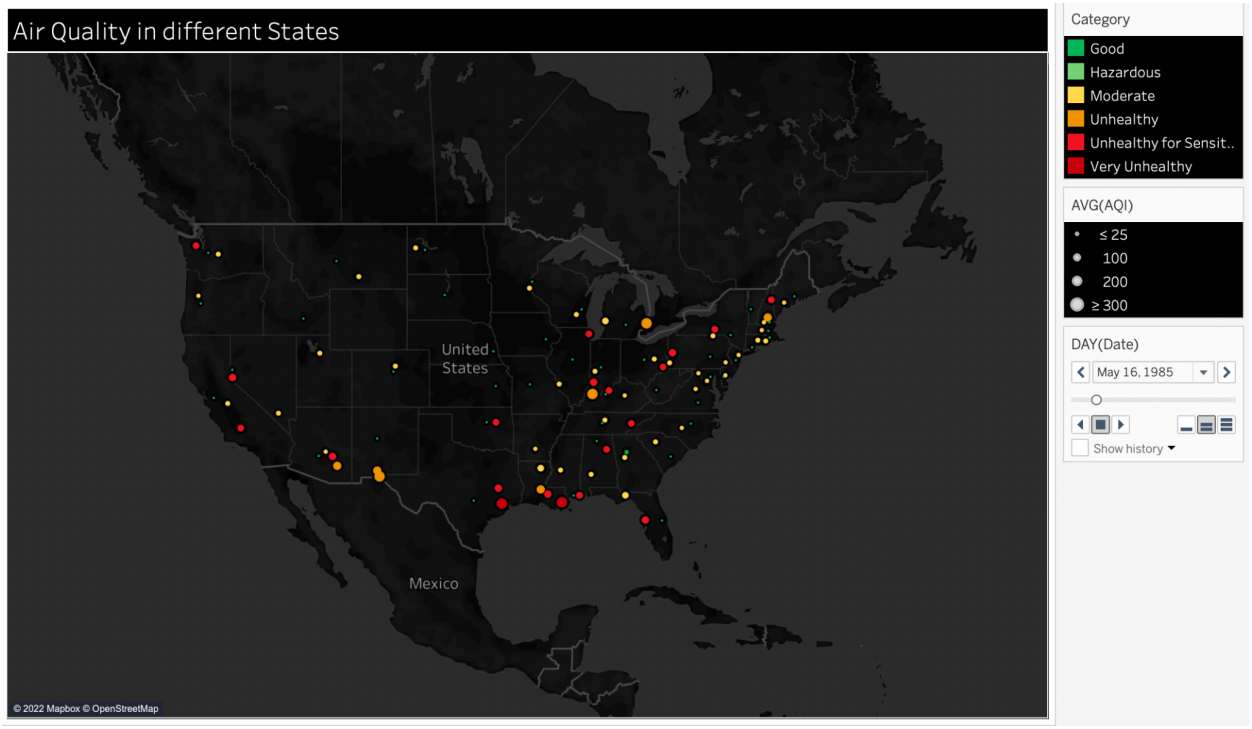
The above graph shows an average AQI level in each month. As we know that the air quality goes poor in summer, it is seen in the above plot that in the month of May, June, July and August has the highest average AQI level in the entire year. Few reasons might be the temperature and its a summer break so people usually travel a lot in this period by their cars etc which in turn deteriorates the air quality in these months.



The above area graph shows a drastic change in the sulphur dioxide as a defining parameter. It is seen that SO2 have been reduced since 1980.

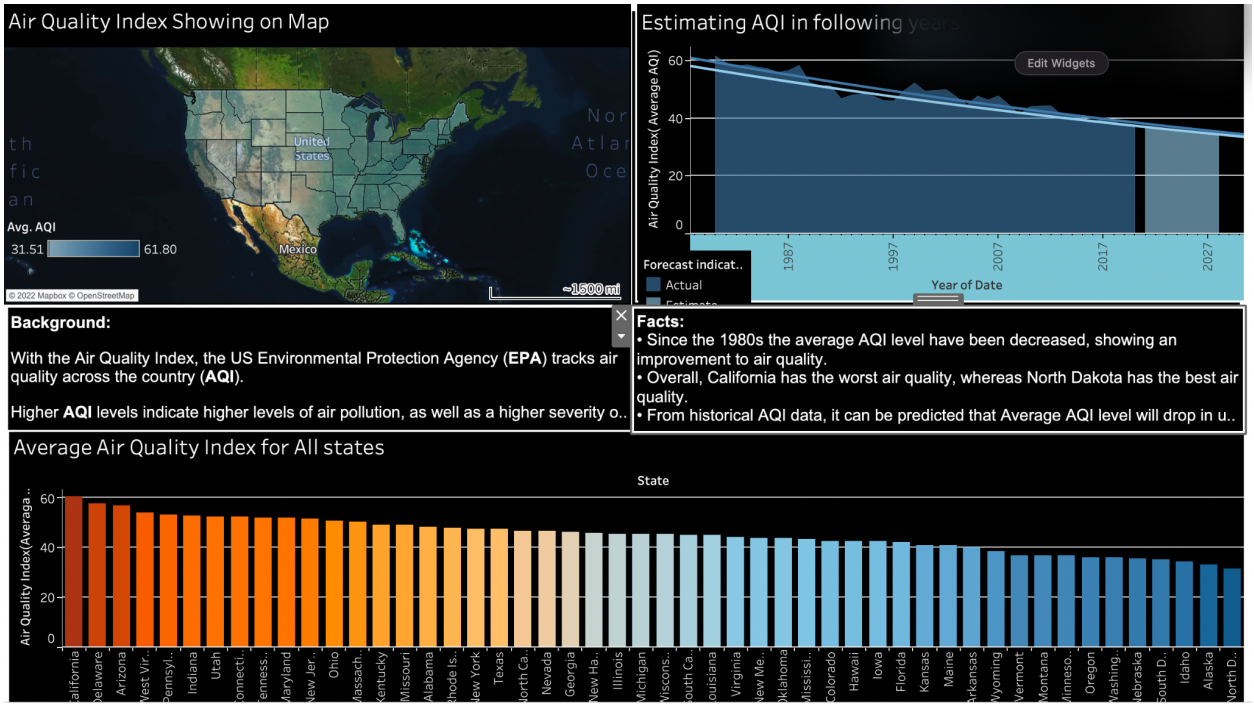


This figure is a histogram showing the categories of an AQI. I have created 6 bins which refer to these categories which are good ranging from 0-50, moderate AQI ranging from 50-100, unhealthy for sensitive people ranging from 100-150, unhealthy having a range of 150-200, very unhealthy have a range of 200-300 and hazardous refers to an AQI which is above 300. So all the AQI values belong to either of these 6 categories.

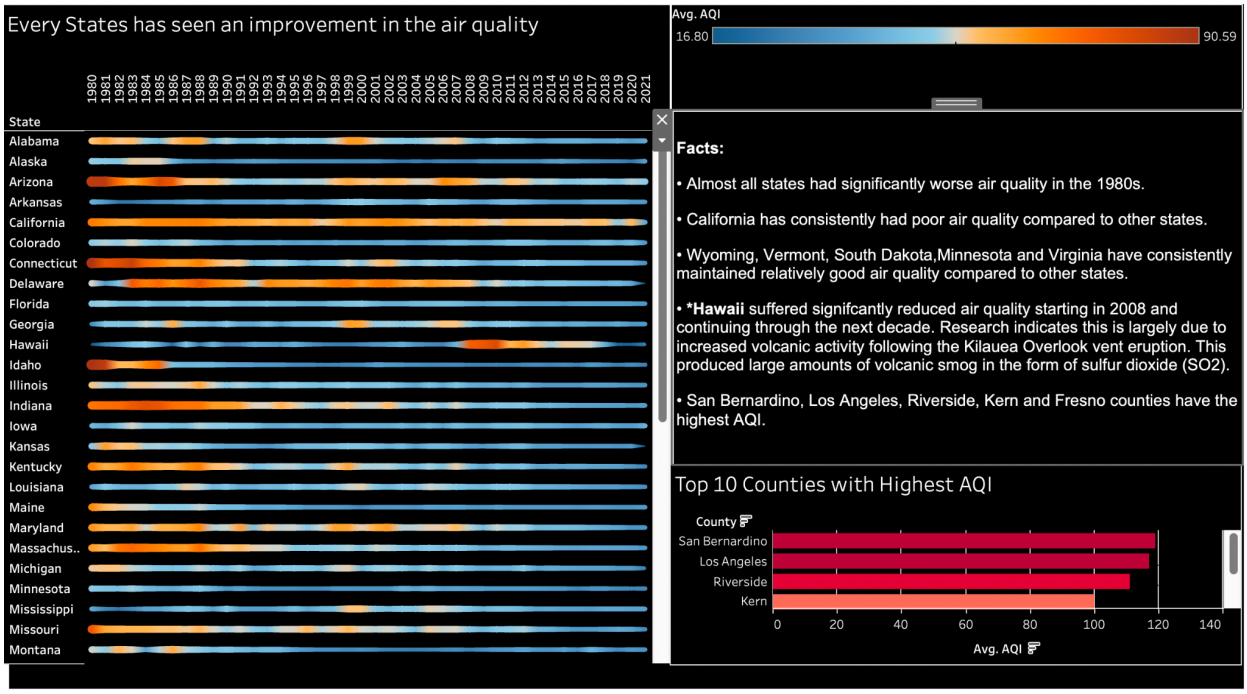


The above geographical map shows air quality in different states. The AQI categories have color coding to distinguish between them and size of the bubble tells you what category of the air quality. This map can give you brief details of the AQI on every date since 1980. If you hover over any bubble and select the date from the right hand side you will get the recorded AQI reading on that specific date in the specific state and county.

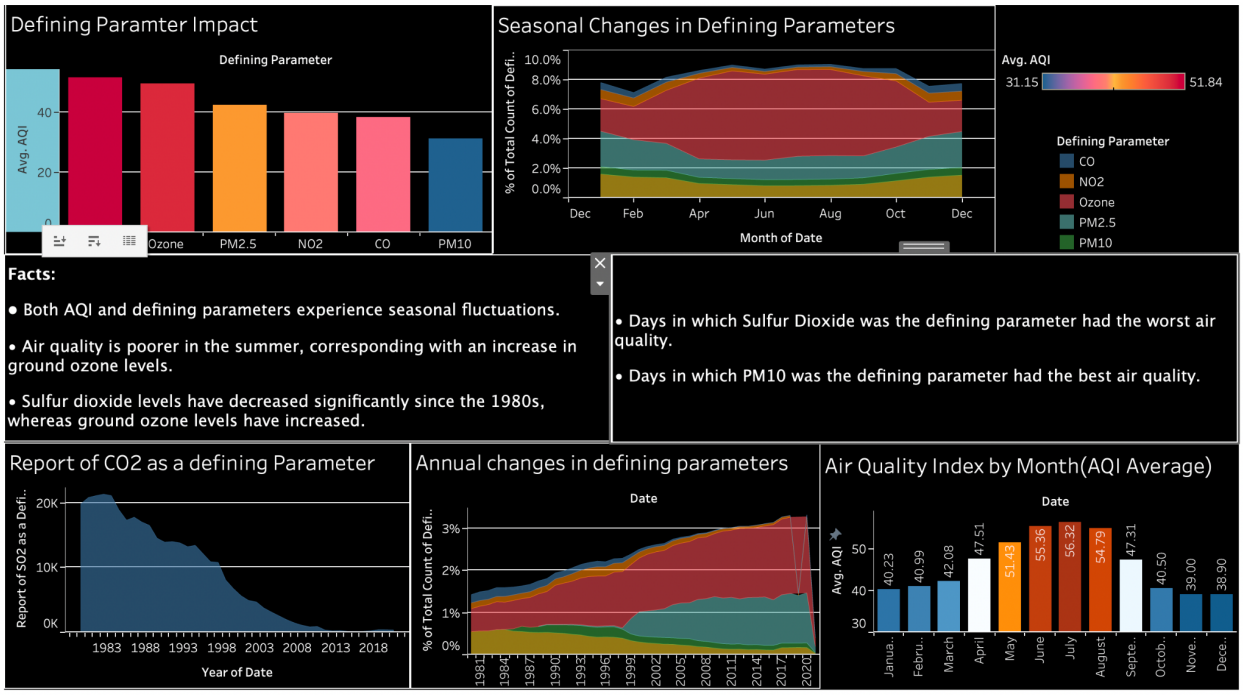
9. Dashboards



This dashboard was created using multiple charts such as geospatial, histogram and area chart representing AQI levels in different states. This dashboard shows you the changes in air quality since 1980 and what can be predicted through this historical data for AQI. It also shows the average air quality index in the specific state if you hover over on any state on the map. It also gives you a brief overview of AQI in every state and ordering them ranging from worst air quality to the best. As we move from left to right from dark orange to blue it shows states with the air quality and we can easily identify which state has the best air quality. It shows overall air quality patterns over the years using area plots.



This dashboard shows the overall information about the AQI changes over the past years in every state and shows the trend as well. It tells us which states had the worst air quality in the 1980s and how the trend changed in these years. Comparing all the states based on the average AQI level over these years we can conclude that only few states have maintained the consistency in the air quality. The above dashboard also gives a brief overview of the top counties in the US which has the worst average air quality.



This dashboard includes various charts and figures which tells about parameters impacting the AQI level and what are the changes in them annually and which are the months having the worst air quality. There are 6 defining parameters in AQI such as CO, NO2, Ozone, PM2.5, PM10, SO2 and shows which parameter has days with worst and best air quality. It also shows the changes in parameters over the past year and describes the trend as well. It brief about the air quality index in each month and shows that average AQI level goes high in summers.

10. References

https://www.epa.gov/outdoor-air-quality-data/air-data-basic-information

https://aqs.epa.gov/aqsweb/airdata/FileFormats.html

https://www.epa.gov/outdoor-air-quality-data

https://www.epa.gov/outdoor-air-quality-data/air-quality-index-daily-values-report

https://www.epa.gov/air-research/air-quality-and-climate-change-research