

Data Visualization and EDA of Air Quality Index in the United States

INFO-5709

Manasa Cherukupally

(11604712)

Introduction

Air pollution is one of the major concerns in the present world. The health of human beings is highly dependent on the air they breathe. If the air quality is not up to the mark, it may lead to major diseases like respiratory diseases, cardiovascular problems, and even cancer. It also has detrimental effects on ecosystems, crops, and buildings. Especially in children, breathing pure air is very important for the lungs and brain to function correctly. This issue is severe in some of the major areas in the United States which when not addressed immediately can become worse.

The causes of air pollution can be mainly. Particularly in densely populated areas and industrial regions, where emissions from transportation, factories, power plants, and other sources can lead to a deterioration of air quality.

Background

The **Air Quality Index (AQI)** is a measure used to know how polluted the air is in a particular location at a given time. It is typically calculated based on the concentrations of five major pollutants: ozone (O₃), particulate matter (PM_{2.5} and PM₁₀), nitrogen dioxide (NO₂), carbon monoxide (CO), and sulfur dioxide (SO₂). Air quality monitoring stations will monitor the concentration of these gases in the air.

The Air Quality Index (AQI) in the United States is measured every day throughout the year by the Environmental Protection Agency (EPA). The EPA operates several air quality monitoring stations across the country for continuously measuring the levels of air pollutants such as particulate matter (PM_{2.5} and PM₁₀), ozone (O₃), nitrogen dioxide (NO₂), carbon monoxide (CO), sulfur dioxide (SO₂).

These monitoring stations report the measurements to EPA for calculating the AQI for each region. EPA further conducts the annual review of air quality data collected at different locations and publishes it in an annual report called the "Air Quality Trends Report." This data is used for analysis and trends in the air quality at different locations.

Related work:

Air pollution is one of the global issues related to the environment that has gained attention in recent times. Measuring Air Quality Index (AQI) is one of the efficient ways of measuring air quality and its components. Air Quality can change over time. It is affected by several external factors and mostly by mistakes made by human beings. Several research activities are carried out on this topic to analyze the AQI. Many researchers have used AQI as a factor to analyze weather conditions all over the world. It is usually done through the metrics collected from the EPA stations.

The paper "Impact of Air Pollution on Running Performance" by Marika. et.al. discusses how air pollution can affect the running performance of individuals. The author discussed how air pollution affects the health and exercise performance of a person. The author also mentioned the different effects of air pollution. Various other factors like duration, pollution levels, and exercise intensity can influence the magnitude of impact according to the study. As part of the research and analysis carried out by the author, some of the important aspects of air quality are uncovered. Apart from that, the visualizations used by the author are effective in understanding the trends in air quality.

“Exploring actionable visualizations for environmental data: Air quality assessment of two Belgian locations” is an article written by Gustavo Carro, Olivier Schalm, Werner Jacobs, and Serge Demeyer. In this author has mentioned the difficulty in understanding the environmental data by non-experts. To overcome such difficulties, the author has proposed a solution of superposing the health risk-related information from the Air Quality Index data into different graphs. This is done by collecting data from monitoring stations in different regions in Belgium and pollution data obtained from the Sentinel-5p satellite which is spatially distributed. In this paper, the author has succeeded in visualizing the AQI data efficiently to draw a pattern from the data.

“A Spatio-Temporal Visualization Approach of PM10 Concentration Data in Metropolitan Lima“ is an article written by Alexandra. et.al discusses the visualization methods for spatio-temporal data related to the PM10 concentration in metropolitan Lima. In this article, the author mentioned a method that combines spatial and temporal visualization to analyze the data. This approach proposed by the author is useful for analyzing the patterns and trends of air quality. The author’s work in this research has also contributed to the environmental science field and provided an effective way for analyzing spatiotemporal data.

In the article "Trend Analysis of Air Quality Index in Catania from 2010 to 2014" by Lanzafame. et. al. discusses the trend of the air quality index in Catania, Italy from 2010-2014. In this paper, the author has taken different air pollutants like PM10, nitrogen dioxide, and ozone collected from the monitoring stations in that location. The author has used statistical methods to analyze the data and to find the trends in air quality over the period. This contribution by the author has helped the stakeholders to find the patterns and to implement certain strategies for air quality management in the Catania region.

The research carried out by several individuals in analyzing the air quality index in different locations over different periods has helped in analyzing some interesting trends in air quality. Several methods and approaches have come into existence to perform the analysis, but there is no ideal method to analyze the data. The approach chosen and the analysis techniques required depend on the type and period of the selected dataset.

Tools: Analytical tools like PowerBI and Tableau are used for Data Visualization.

Programming: For data visualization and some part of analysis python is used.

Dataset:

The dataset used for this research problem is collected from the official EPA (https://aqs.epa.gov/aqsweb/airdata/download_files.html#AnnualLinks to an external site.) website. The dataset consists of the annual AQI data from the past 5 years from different cities in the United States. There are 5013 rows and columns of the dataset in 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 PM2.5, Days PM10.

Sample dataset:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	State	County	Year	Days with AQI	Good Days	Moderate	Unhealthy for S	Unhealthy	Very Unhe	Hazardous	Max AQI	90th Perce	Median AC	Days CO	Days NO2	Days Ozon	Days PM2.5	Days PM10
2	Alabama	Baldwin	2022	141	119	22	0	0	0	0	96	52	40	0	0	114	27	0
3	Alabama	Clay	2022	58	50	8	0	0	0	0	64	52	27	0	0	0	58	0
4	Alabama	DeKalb	2022	242	225	17	0	0	0	0	97	48	37	0	0	224	18	0
5	Alabama	Elmore	2022	117	110	7	0	0	0	0	67	47	37	0	0	117	0	0
6	Alabama	Etowah	2022	179	140	39	0	0	0	0	93	58	42	0	0	76	103	0
7	Alabama	Jefferson	2022	182	89	91	2	0	0	0	105	74	51	1	0	67	114	0
8	Alabama	Madison	2022	181	134	47	0	0	0	0	97	60	43	0	0	71	107	3
9	Alabama	Mobile	2022	180	143	37	0	0	0	0	93	58	42	0	0	73	107	0
10	Alabama	Montgomery	2022	181	148	33	0	0	0	0	95	58	41	0	0	103	78	0
11	Alabama	Morgan	2022	181	136	45	0	0	0	0	97	60	43	0	0	83	98	0
12	Alabama	Russell	2022	180	130	50	0	0	0	0	96	63	44	0	0	70	110	0
13	Alabama	Shelby	2022	122	109	12	1	0	0	0	105	51	41	0	0	122	0	0
14	Alabama	Sumter	2022	180	161	19	0	0	0	0	79	51	34	0	0	74	106	0
15	Alabama	Tuscaloosa	2022	139	121	18	0	0	0	0	87	53	37	0	0	107	32	0
16	Alaska	Anchorage	2022	244	214	29	1	0	0	0	125	54	21	5	0	0	104	135
17	Alaska	Denali	2022	211	207	4	0	0	0	0	61	44	38	0	0	211	0	0
18	Alaska	Fairbanks	2022	201	140	45	8	7	1	0	280	85	39	23	0	89	86	3
19	Alaska	Juneau	2022	188	171	17	0	0	0	0	89	46	14	0	0	0	187	1
20	Alaska	Matanuska	2022	238	221	17	0	0	0	0	83	35	10	0	0	0	127	111
21	Arizona	Apache	2022	273	271	2	0	0	0	0	57	29	11	0	0	0	21	252
22	Arizona	Cochise	2022	273	203	69	1	0	0	0	101	64	45	0	0	195	0	78
23	Arizona	Coconino	2022	273	196	76	1	0	0	0	101	67	47	0	0	273	0	0
24	Arizona	Gila	2022	273	158	97	16	2	0	0	177	90	48	0	0	257	0	16
25	Arizona	La Paz	2022	267	187	79	1	0	0	0	130	71	45	0	0	248	5	14
26	Arizona	Maricopa	2022	203	160	140	55	13	6	0	264	122	74	0	0	236	17	54

There are 20 variables, and the dataset consists of both numerical and categorical variables. These variables help us to analyze the data and provide conclusions.

Table with variable description:

Variable	Description	Datatype
Days with AQI	Number of days in the year that is having the Air Quality Index value	Numerical
Days Good	Number of days in the year AQI value 0 through 50.	Numerical
Days Moderate	Number of days in the year AQI value 51 through 100.	Numerical
Days Unhealthy for Sensitive Groups	Number of days in the year AQI value 101 through 150.	Numerical
Days Unhealthy	Number of days in the year AQI value 151 through 200.	Numerical
Days Very Unhealthy	Number of days in the year AQI value 201 through 300.	Numerical
Days Hazardous	Number of days in the year having an AQI value 301 through 500.	Numerical
AQI Max	The daily AQI highest value in the year.	Numerical
AQI 90th %ile	AQI values during the year were less than or equal to the 90% value.	Numerical
AQI Median	Half of daily AQI values during the year that is <= median value	Numerical
Days Days Days Days PM2.5	CO NO2 O3 These columns give the number of days each pollutant measured which is the main pollutant.	Numerical

Days PM10		
year	Year when metric taken	Numerical
State	State of United States	Categorical
County	County for each state in the United States	Categorical

Data Preprocessing:

As part of data preprocessing, there are multiple tasks that need to be done to prepare data for analysis. That is cleaning, transforming, and preparing. Preprocessing data is very important. Failing to take this step may cause a wrong analysis.

For preprocessing data, the **PowerBI** tool is used. As the first step, data is loaded to transform it.

Data Cleaning:

The datatypes of some of the variables are changed to numerical as it is difficult to analyze the categorical variable with numeric data. Duplicate rows present in the data are removed. Since the dataset does not consist of any missing values, this step is skipped.

The sample data cleaning using PowerBI

The screenshot displays a PowerBI Desktop interface. The main area shows a data table with the following columns: State, County, Year, Days with AQI, Good Days, Moderate Days, and Unhealthy for Sensitive Groups. The table contains 28 rows of data, starting with Alabama (Baldwin, 2022) and ending with Arizona (Navajo, 2022). The right sidebar is open to the 'Query Settings' pane, which shows the 'APPLIED STEPS' section with 'Removed Duplicates' as the first step.

State	County	Year	Days with AQI	Good Days	Moderate Days	Unhealthy for Sensitive Groups
Alabama	Baldwin	2022	141	119	22	
Alabama	Clay	2022	58	50	8	
Alabama	DeKalb	2022	242	225	17	
Alabama	Elmore	2022	117	110	7	
Alabama	Etowah	2022	179	140	39	
Alabama	Jefferson	2022	182	89	91	
Alabama	Madison	2022	181	134	47	
Alabama	Mobile	2022	180	143	37	
Alabama	Montgomery	2022	181	148	33	
Alabama	Morgan	2022	181	136	45	
Alabama	Russell	2022	180	130	50	
Alabama	Shelby	2022	122	109	12	
Alabama	Sumter	2022	180	161	19	
Alabama	Tuscaloosa	2022	139	121	18	
Alaska	Anchorage	2022	244	214	29	
Alaska	Denali	2022	211	207	4	
Alaska	Fairbanks North Star	2022	201	140	45	
Alaska	Juneau	2022	188	171	17	
Alaska	Matanuska-Susitna	2022	238	221	17	
Arizona	Apache	2022	273	271	2	
Arizona	Cochise	2022	273	203	69	
Arizona	Coconino	2022	273	196	76	
Arizona	Gila	2022	273	158	97	
Arizona	La Paz	2022	267	187	79	
Arizona	Maricopa	2022	303	60	168	
Arizona	Mohave	2022	273	257	15	
Arizona	Navajo	2022	272	221	51	

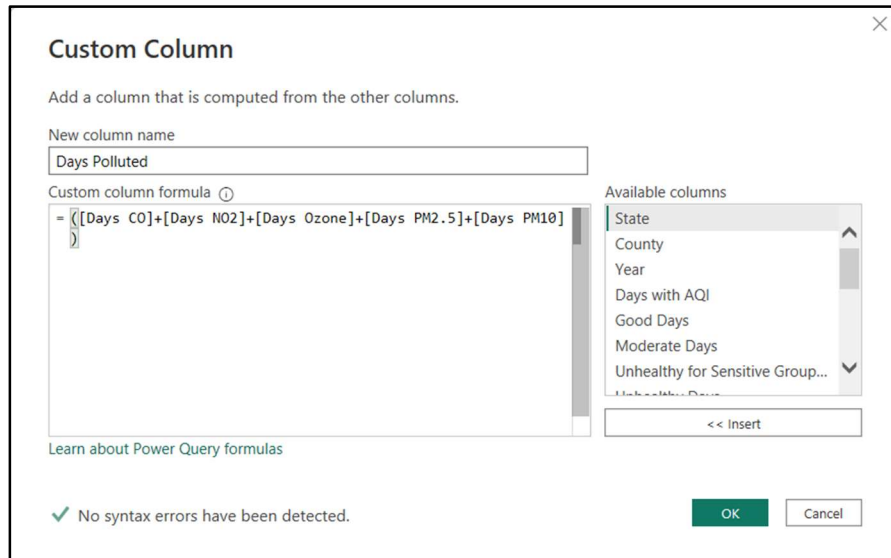
Feature Engineering:

Feature engineering is a technique of selecting or transforming data into features. This step helps in improving the efficiency of the analysis. The goal of this step is to extract meaningful information from the dataset.

Two features were introduced in the dataset.

1. One is **“Days Polluted”**. This feature is the sum of days that have been polluted with any one of the poisonous gases. This feature is extracted using **PowerBI**.

The **“Days Polluted”** feature was created in PowerBI with the sum of other features.



Custom Column

Add a column that is computed from the other columns.

New column name
Days Polluted

Custom column formula ⓘ
=([Days CO]+[Days NO2]+[Days Ozone]+[Days PM2.5]+[Days PM10])

Available columns
State
County
Year
Days with AQI
Good Days
Moderate Days
Unhealthy for Sensitive Group...

<< Insert

Learn about Power Query formulas

✓ No syntax errors have been detected.

OK Cancel

2. **Regions**. For some of the visualizations, It is difficult to work with states and counties. Hence a variable that categorizes the states based on their geographical location is introduced. This variable is implemented using the **Python** code.

```
# Define a dictionary that maps states and territories to their respective regions
state_regions = {
    'Alabama': 'South',
    'Alaska': 'West',
    'Arizona': 'West',
    'Arkansas': 'South',
    'California': 'West',
    'Colorado': 'West',
    'Connecticut': 'Northeast',
    'Delaware': 'South',
    'Florida': 'South',
    'Georgia': 'South',
    'Hawaii': 'West',
    'Idaho': 'West',
    'Illinois': 'Midwest',
    'Indiana': 'Midwest',
    'Iowa': 'Midwest',
    'Kansas': 'Midwest',
    'Kentucky': 'South',
    'Louisiana': 'South',
    'Maine': 'Northeast',
    'Maryland': 'South',
    'Massachusetts': 'Northeast',
    'Michigan': 'Midwest',
    'Minnesota': 'Midwest',
    'Mississippi': 'South',
    'Missouri': 'Midwest',
    'Montana': 'West',
    'Nebraska': 'Midwest',
    'Nevada': 'West',
    'New Hampshire': 'Northeast',
    'New Jersey': 'Northeast',
    'New Mexico': 'West',
    'New York': 'Northeast',
    'North Carolina': 'South',
    'North Dakota': 'Midwest',
    'Ohio': 'Midwest',
    'Oklahoma': 'South',
}
```

```

    'South Dakota': 'Midwest',
    'Tennessee': 'South',
    'Texas': 'South',
    'Utah': 'West',
    'Vermont': 'Northeast',
    'Virginia': 'South',
    'Washington': 'West',
    'West Virginia': 'South',
    'Wisconsin': 'Midwest',
    'Wyoming': 'West',
    'District Of Columbia': 'South',
    'Puerto Rico': 'South',
    'Virgin Islands': 'South',
    'Country Of Mexico': 'South'
}

# Read the CSV file into a pandas DataFrame
df = pd.read_csv('transformed_data.csv')

# Define a function that takes a state or territory name as input and returns its region
def get_region(state):
    return state_regions.get(state)

# Add a new column 'Region' to the DataFrame
df['Region'] = df['State'].apply(get_region)

# Save the updated DataFrame as a new CSV file
df.to_csv('updated_file.csv', index=False)

df.head(5)

```

Sample output:

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 PM2.5	Days PM10	Days Polluted	Region
Alapai	2022	273	249	24	0	0	0	0	87	49	42	0	0	273	0	0	273	West
Summa	2022	273	218	55	0	0	0	0	87	61	45	0	0	273	0	0	273	West
Lake	2022	181	169	12	0	0	0	0	87	50	40	0	0	181	0	0	181	South
Lasco	2022	181	155	26	0	0	0	0	87	51	42	0	0	181	0	0	181	South
Floyd	2022	201	185	16	0	0	0	0	87	48	36	0	0	201	0	0	201	Midwest
Nicks	2022	232	218	14	0	0	0	0	87	48	37	0	0	232	0	0	232	Midwest
Knox	2022	226	204	22	0	0	0	0	87	50	39	0	0	226	0	0	226	Midwest
mine	2022	121	108	13	0	0	0	0	87	51	41	0	0	121	0	0	121	South
arles	2022	122	105	17	0	0	0	0	87	54	42	0	0	122	0	0	122	South
ington	2022	104	92	12	0	0	0	0	87	51	40	0	0	104	0	0	104	Midwest

Data Exploration:

Exploring the mean, standard deviation, minimum and maximum values in dataset using python

df.describe()

	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
count	5012.000000	5012.000000	5012.000000	5012.000000	5012.000000	5012.000000	5012.000000	5012.000000	5012.000000	5012.000000	5012.000000
mean	2019.974461	300.260974	247.458300	49.035116	2.788907	0.786911	0.126097	0.065642	119.282123	55.847765	35.699721
std	1.410309	87.298245	78.816691	42.201505	7.871481	3.614987	1.624394	0.654323	147.815906	16.377853	10.125801
min	2018.000000	2.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	8.000000	5.000000	0.000000
25%	2019.000000	243.000000	199.000000	17.000000	0.000000	0.000000	0.000000	0.000000	83.000000	48.000000	33.000000
50%	2020.000000	357.000000	263.000000	38.000000	0.000000	0.000000	0.000000	0.000000	100.000000	54.000000	37.000000
75%	2021.000000	365.000000	313.000000	70.000000	2.000000	0.000000	0.000000	0.000000	129.000000	61.000000	41.000000
max	2022.000000	366.000000	365.000000	261.000000	103.000000	69.000000	74.000000	26.000000	7577.000000	215.000000	122.000000

Days CO	Days NO2	Days Ozone	Days PM2.5	Days PM10	Days Polluted
5012.000000	5012.000000	5012.000000	5012.000000	5012.000000	5012.000000
0.659816	4.717279	172.010774	112.160215	10.712889	300.260974
9.463571	22.840822	116.193409	107.358222	43.816761	87.298245
0.000000	0.000000	0.000000	0.000000	0.000000	2.000000
0.000000	0.000000	85.000000	2.000000	0.000000	243.000000
0.000000	0.000000	192.000000	97.000000	0.000000	357.000000
0.000000	0.000000	245.000000	174.000000	0.000000	365.000000
277.000000	365.000000	366.000000	366.000000	366.000000	366.000000

The dataset is huge with 5013 rows and 20 columns explaining the metrics of air quality annually across different states and counties in the unite states. This dataset is well organized and is collected from the government website (Environmental Protection Agency) and is accurate. Hence it is sufficient to answer the hypothetical questions related to the dataset.

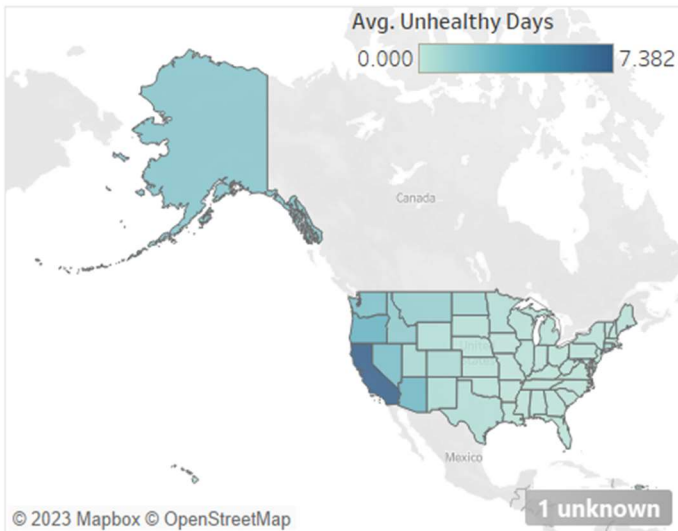
The Questions that we chose to answer from the given dataset are.

1. Which states in the United States have the most average unhealthy days from the year (2018-2022)?
2. Which states in the United States have good air days from the year (2018-2022)?
3. How does the trend in Air Quality Index(AQI) change over 5 years (2018-2022)?
4. Which gases are responsible for most the of pollution in different states and the days of pollution over the period(2018-2022)?

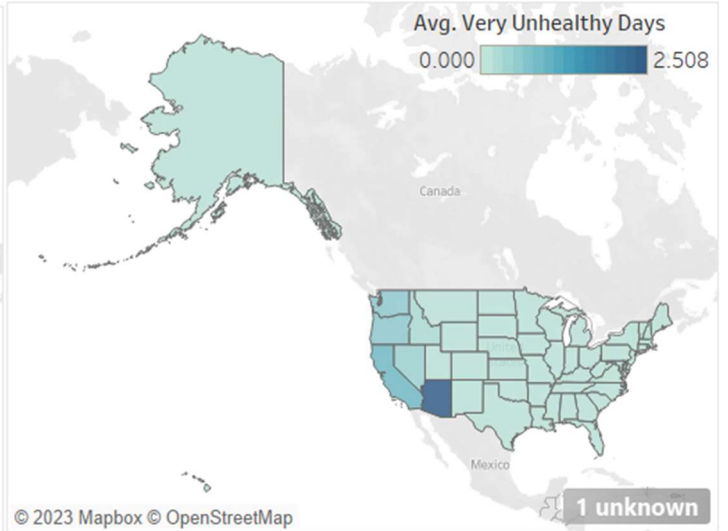
Data Visualization

Question 1: Which states in the United States have the most average unhealthy days from the year (2018-2022)?

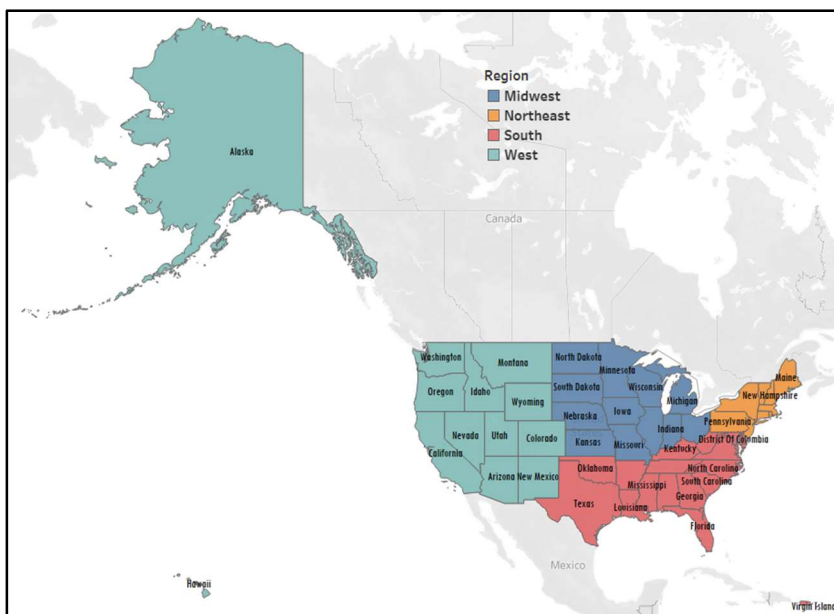
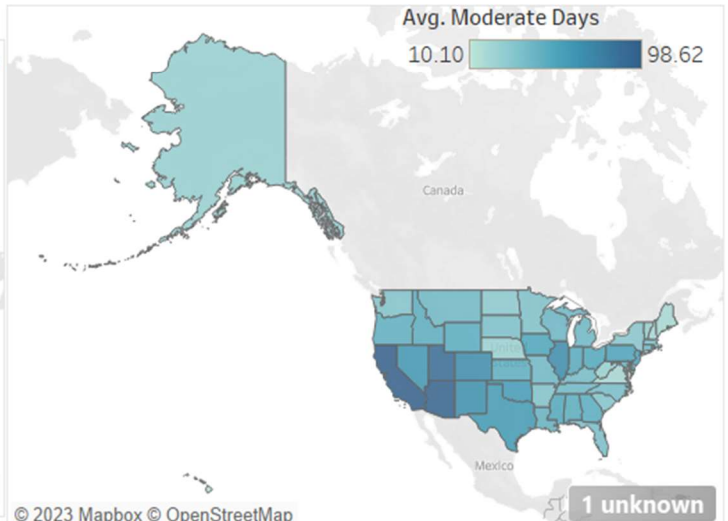
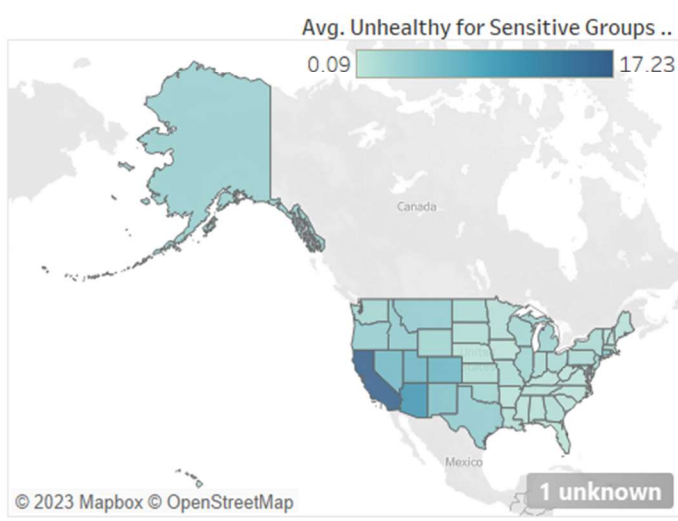
Avg Unhealthy days for each state(2018-2022)



Avg Very Unhealthy days for each state(2018-2022)



Avg Unhealthy for Sensitive days for each state(2018-2022) Avg Moderate days for each state(2018-2022)



Result:

The geographical maps above help us in analyzing the question. It can be inferred from that visualization that, California state has the average unhealthy days from the year(2018-2019) with 7.3 average days. State Arizona(2.5 days) has the highest number of very unhealthy days followed by California. California (7.23 days) again stands in the first position for unhealthy days for the sensitive people category and Average moderate days are high in the states of California(98 days), Utah(90 days), and Arizona(97 days).

It can be inferred from the maps that, overall average unhealthy days are more in California state followed by Arizona State. It can also be inferred from the graph that, states that belong to the **western region** of the United States have more unhealthy days when compared with other regions. It is followed by the **southern region**. This may be because these regions are highly populated and thus have higher pollution.

Design Choice:

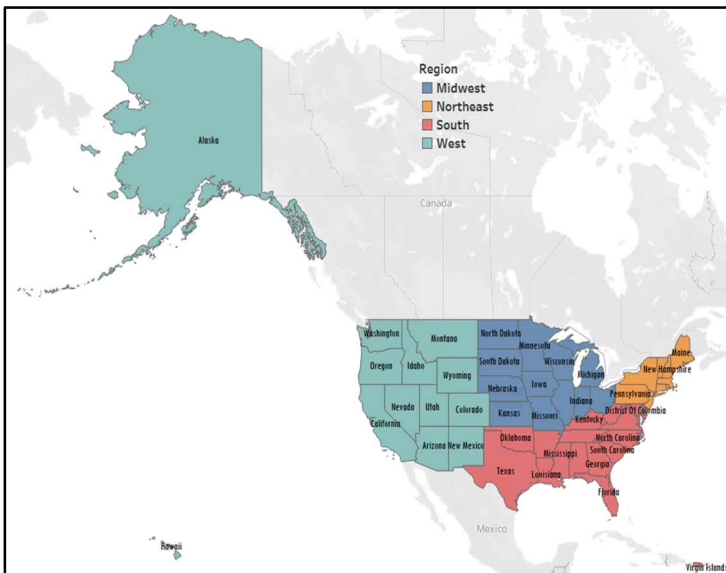
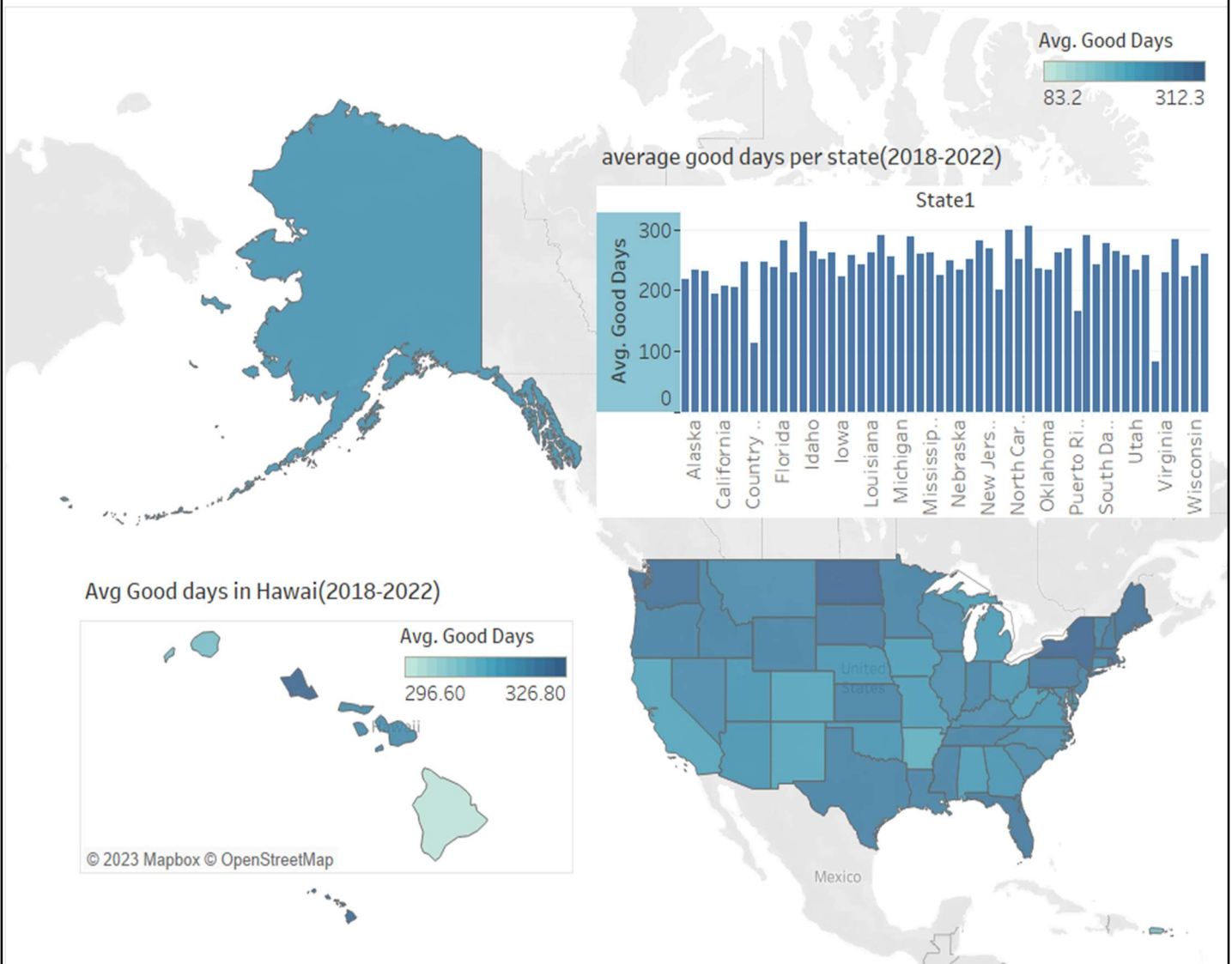
To answer the above hypothesis, **Tableau** is used. **Geographical maps** are used as the design element. **Dark blue** indicates higher average unhealthy days and **lighter blue** is used to indicate the lower average unhealthy days. Orange, red, blue, and teal colors are used to distinguish regions **Northeast, South, Midwest**, and **West** respectively. The reason for choosing the Geographical maps is because they give precise information regarding the geographical locations, and it will be easy to find the differences and compare the states on average unhealthy days.

Graphic Excellence Principle:

The above graphs follow the graphical excellence principles. As the geographical maps are well organized and states can be easily distinguishable with appropriate legends showing the metric scale.

Question2: Which states in the United States have good air days from the year (2018-2022)?

Avg Good days for each state(2018-2022)



Results:

To answer the above hypothesis, Geographical maps, and bar plot is used. From the above data visualization, it can be inferred that the Hawaiian Islands have the highest average number of good days with 312 days when compared with other states. Especially Honolulu city in the Hawaiian Islands has shown the highest average number of good days from the year (2018-2022). The least number of average good days is for the state Virgin Islands with 83.2 days. It is followed by the Puerto Rico islands. There is not much difference in the states when it comes to good days. But the western region has a lesser average number of good days when compared with other regions. The midwestern region shows more number average of good days.

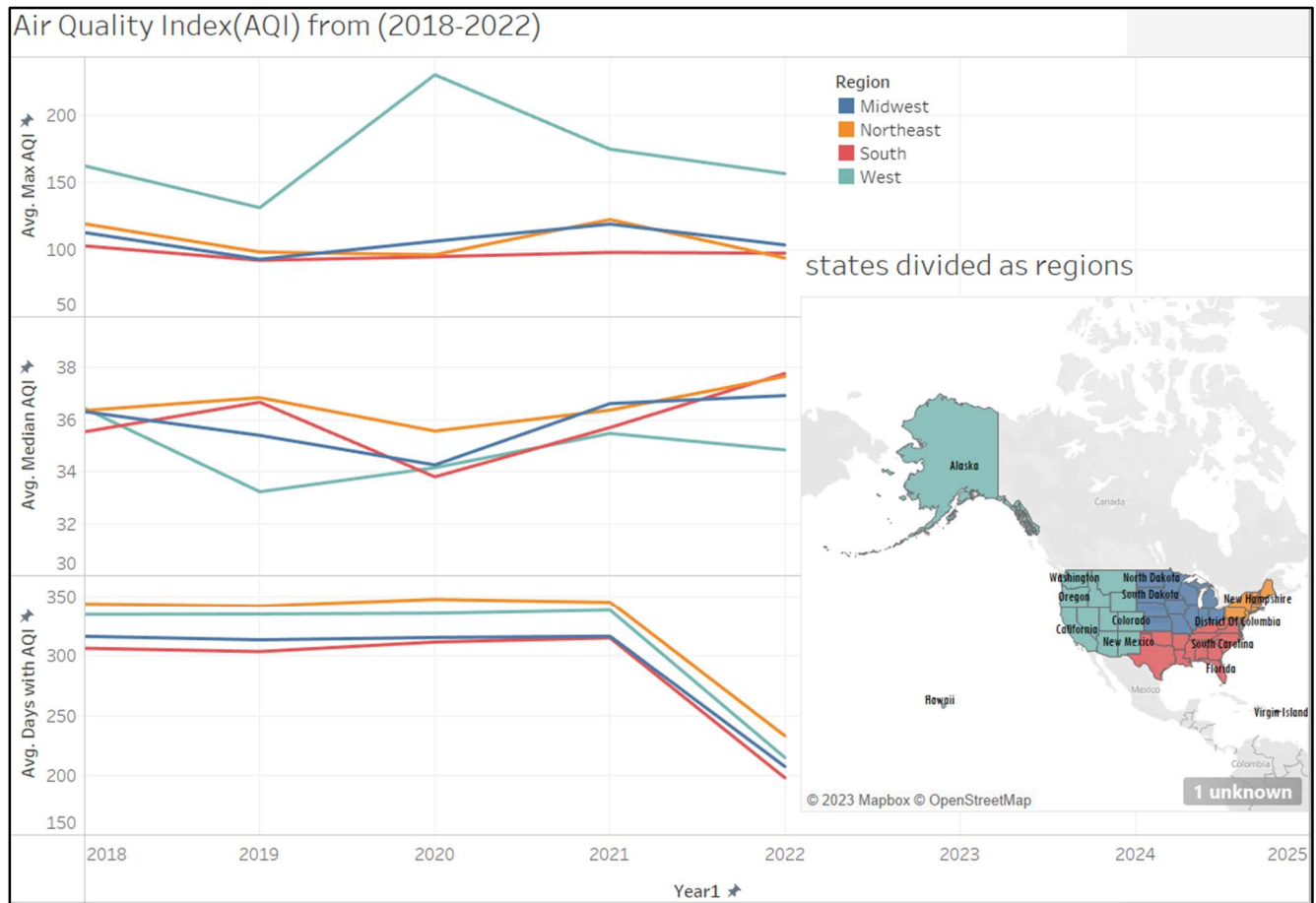
Design Choice:

To answer the above hypothesis, **Tableau** is used. **Geographical maps and Bar plots** are used as the design element. **Dark blue** indicates higher average good days and **lighter blue** is used to indicate the lower average good days. Orange, red, blue, and teal colors are used to distinguish regions **Northeast**, **South**, **Midwest**, and **West** respectively. The reason for choosing the Geographical maps is because they give precise information regarding the geographical locations, and it will be easy to find the differences and compare the states on average unhealthy days. Bar plot is used to compare the average good days of each state with metrics being the average days.

Graphic Excellence Principle:

It can be said that the above graphs follow the graphical excellence principles. As the geographical maps are well organized and states can be easily distinguishable with appropriate legends showing the metric scale. The bar plot is designed to analyze the accurate differences between the states as there are no drastic changes in average good days among states.

Question 3: How does the trend in Air Quality Index(AQI) change over 5 years (2018-2022)?



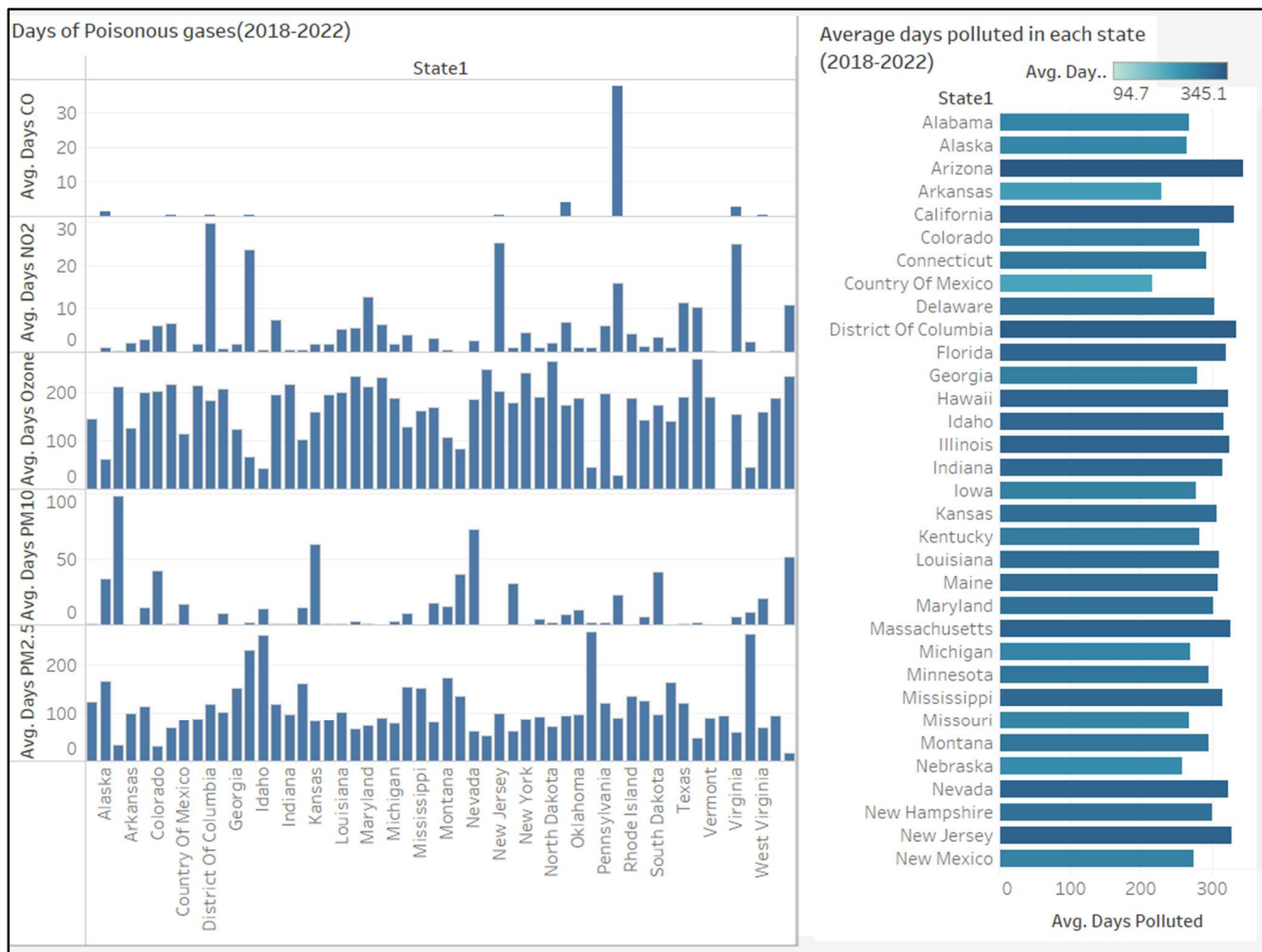
Results: To answer the above question, Line plots, and geographical maps are used. It can be inferred from the above visualization that; the western region has the highest average maximum air quality index(AQI) and the least for the southern region. The average median air quality index is higher in the Northeastern region and the low average median AQI is fluctuating between the western and southern regions. It can be observed from the graph that there is a slight increase in the average median AQI days in the year 2022 and the average days with AQI has been decreased from the year 2021 to 2022.

Design Choice: To answer the above hypothesis, **Tableau** is used. **Line plots and geographical graphs** are used as the design element. Orange, red, blue, and teal colors are used to distinguish regions **Northeast**, **South**, **Midwest**, and **West** respectively. The reason for choosing the line plot is to show the trend of average AQI days from the year 2018 to 2022. Line plot works best for trend analysis. The reason for choosing the Geographical maps is because they give precise information regarding the geographical locations, and it will be easy to find the differences and compare the regions.

Graphic Excellence Principle:

It can be said that the above line plot and maps follow the graphical excellence principles. As the geographical maps are well organized and states can be easily distinguishable. The line plots depict the trend of the AQI from the year 2018-2022 serving the purpose of the hypothesis.

Question 4: Which gases are responsible for most of the pollution in different states and the states that have the highest average days of pollution over the period(2018-2022)?



Results:

To answer the above question, a **Bar plot** is used. From the above plot, it can be inferred that average CO days were found in Puerto Rico(37). The average NO2 days(29 days) is found in the District of Columbia. Average Ozone days were found in Utah(270 days). Average PM10 days were found in Arizona and average PM2.5 days were found in Washington followed by Oregon(269 days). The average number of days polluted was found in Arizona(345 days), the District of Columbia(334 days) followed by California (331 days).

Design Choice: To answer the above hypothesis, **Tableau** is used. **Bar plots** are used as the design element. The reason for choosing the Bar plot is to show the states and the number of days each state is affected by poisonous gases from the year 2018 to 2022. It will be easy to find the differences and compare the states using a Bar plot.

Graphic Excellence Principle:

The above Bar plot follows the graphical excellence principles. As the Bar plot is well organized, comparisons can be easily understandable. A bar plot can be used to find the value of each individual with respect to another individual.

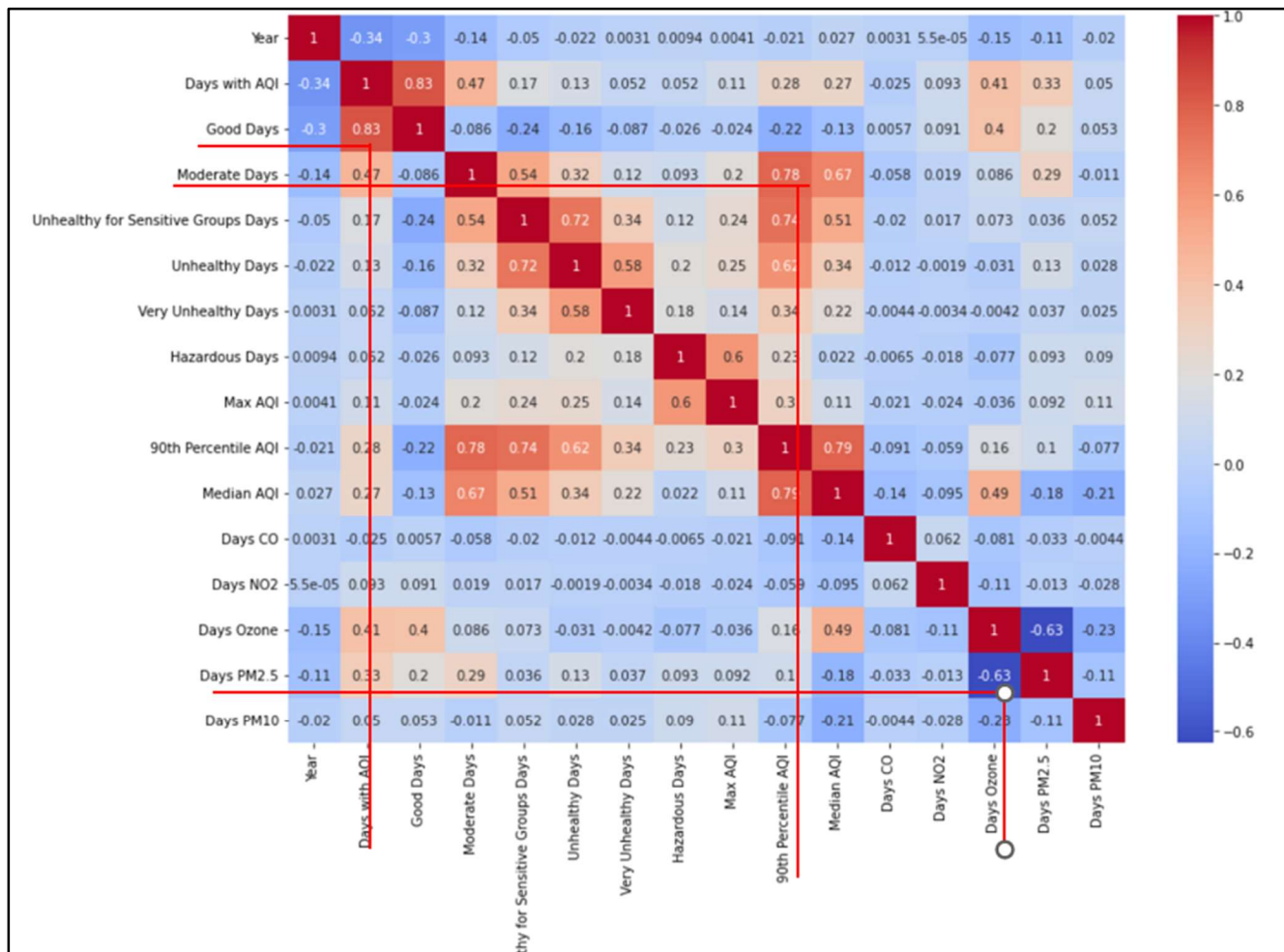
Python code with seaborn is used to find the correlation between the variables using a **heatmap**. This correlation is useful to analyze how each variable changes with respect to another variable. From the below correlation matrix, it can be inferred that good days are highly correlated with Day with Air Quality Index, Moderate days are highly correlated with 90th percentile AQI and days with ozone is negatively correlated with days of PM2.5

Python code:

```
import seaborn as sns
import matplotlib.pyplot as plt
# Create a correlation matrix

corr_matrix = df.iloc[:, :-1].corr()
fig, ax = plt.subplots(figsize=(14, 10))
# Create the correlation plot
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm')
```

Correlation matrix:



Prediction of a number of good days with Linear regression:

Python code with Scikit-learn module is used to predict the number of good days with respect to some of the variables like 'Max AQI', '90th Percentile AQI', 'Days CO', 'Days NO2', 'Days Ozone', 'Days PM2.5', 'Days PM10', 'Median AQI'. The data is split into 80% as training data and 20% as testing data.

The linear regression model is fitted to the training data. With the model built testing data is predicted. The R-squared score that is used to evaluate the model is **0.93**.

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
y=df['Good Days']
selected_cols = ['Max AQI', '90th Percentile AQI','Days CO', 'Days NO2', 'Days Ozone', 'Days PM2.5', 'Days PM10', 'Median AQI']
X = df[selected_cols]

# Create a train-test split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

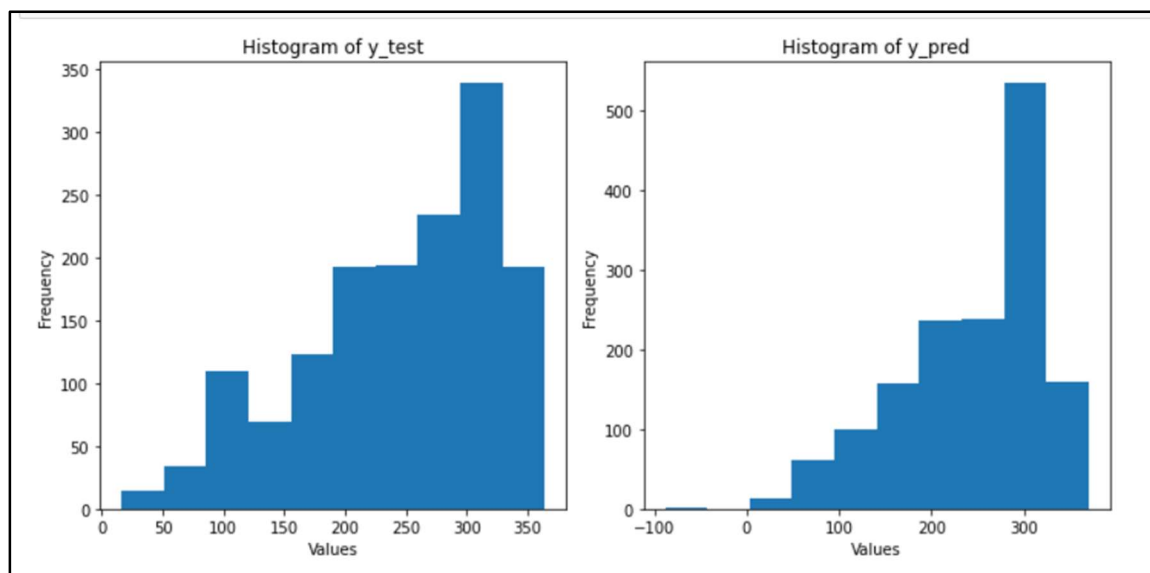
# Fit a linear regression model
lr = LinearRegression()
lr.fit(X_train, y_train)

# Make predictions on the test data
y_pred = lr.predict(X_test)

from sklearn import metrics
import numpy as np
# Compute the mean squared error (MSE)
mse = metrics.mean_squared_error(y_test, y_pred)
print("mean squared error {}".format(mse))
# Compute the root mean squared error (RMSE)
rmse = np.sqrt(mse)
print("root mean squared error {}".format(rmse))
# Compute the R-squared score
r2 = metrics.r2_score(y_test, y_pred)
print("R-Squared score{}".format(r2))

mean squared error 420.2024598801947
root mean squared error 20.49884045208886
R-Squared score0.93152013167175
```

Histogram showing the distribution of the testing data vs predicted data



Discussion:

The Air Quality Index (AQI) is a metric that can be used to check air quality. Especially since it is a standard way of measuring air pollution. The Environmental Protection Agency(EPA) of the United States is the website that measures the air quality Index. In this analysis, the EPA annual Air quality metrics from different counties in different states of the United States are taken for the period of 2018-2022. These metrics are analyzed with different visualizations like geographical maps for comparing the values with geographical locations, bar charts to compare the values, and line plots to analyze the trend in the values over the period of 2018-2019.

Since the data collected is annual data, this can be a limitation of the analysis. The minor transitions that occurred in the middle days of the year are not present in the day which can limit the in-depth exploration of the trends in the Air Quality Index. One more drawback is that the Air Quality Index is not calculated daily to get accurate annual metrics. It is calculated only for a few days in a year.

Future work:

As part of the future scope, more enhancements can be made to the Analysis of Air Quality Index. More data visualization techniques can be incorporated to understand the trend. Data at present is only for 5 years period (2018-2022), which is less. More data needs to be collected at least for ten years for better understanding. Two new features were added to the dataset for better analysis. There can be added more meaningful features that help in efficient analysis.

Conclusion:

To conclude, analysis of air quality is very important in the recent world where pollution is increasing along with the population. From the above visualizations, it can be analyzed that the states in the western region, especially states like California where the population is high are having unhealthy days. The Hawaiian state which is an island where the population is less has less air pollution and has seen more good air days in the past five years. Overall, This analysis has given us a broader picture of the air quality all over the united states and thus helps us to take necessary measures to reduce pollution and improve the air quality.

References

- Cusick, M., Rowland, S. T., & DeFelice, N. (2023). Impact of air pollution on running performance. *Scientific Reports*, 13(1), 1832.
- Carro, G., Schalm, O., Jacobs, W., & Demeyer, S. (2022). Exploring actionable visualizations for environmental data: Air quality assessment of two Belgian locations. *Environmental Modelling & Software*, 147, 105230.
- Lanzafame, R., Monforte, P., Patanè, G., & Strano, S. (2015). Trend analysis of Air Quality Index in Catania from 2010 to 2014. *Energy Procedia*, 82, 708-715.
- Encalada-Malca, A. A., Cochachi-Bustamante, J. D., Rodrigues, P. C., Salas, R., & López-Gonzales, J. L. (2021). A spatio-temporal visualization approach of pm10 concentration data in metropolitan lima. *Atmosphere*, 12(5), 609.
- Li, H., Fan, H., & Mao, F. (2016). A visualization approach to air pollution data exploration—a case study of air quality index (PM2. 5) in Beijing, China. *Atmosphere*, 7(3), 35.
- United States Environmental Protection Agency. <https://www.epa.gov/outdoor-air-quality-data/about-air-data-reports>