Analyzing Relationship Between Air Quality Index and Average Temperature in the United States from 2000-2019

Daniel Kim, Maximillian Jacob, Zach Fenton

Github repository: UC-Berkeley-I-School/Project2_Kim_Fenton_Jacob (github.com)
Primary Datasets:
AQI data - Download Files | AirData | US EPA
Temp data - washingtonpost/data-2C-beyond-the-limit-usa: The Washington Post's analysis of NOAA climate change data for the contiguous United States (github.com)

Introduction

Climate change, a significant global issue, is amplifying environmental health risk factors such as extreme temperatures, air pollution, and allergens, causing a rise in cardiovascular and respiratory diseases[1]. Since the late 1800s, the Earth's average surface temperature has risen approximately 2 degrees Fahrenheit, primarily due to the accumulation of greenhouse gasses in the atmosphere, a consequence of increased carbon dioxide emissions and other human activities. This warming has been most dramatic over the past four decades, with the seven most recent years experiencing record high temperatures. Particularly, 2016 and 2020 stand as the warmest years on record[2].

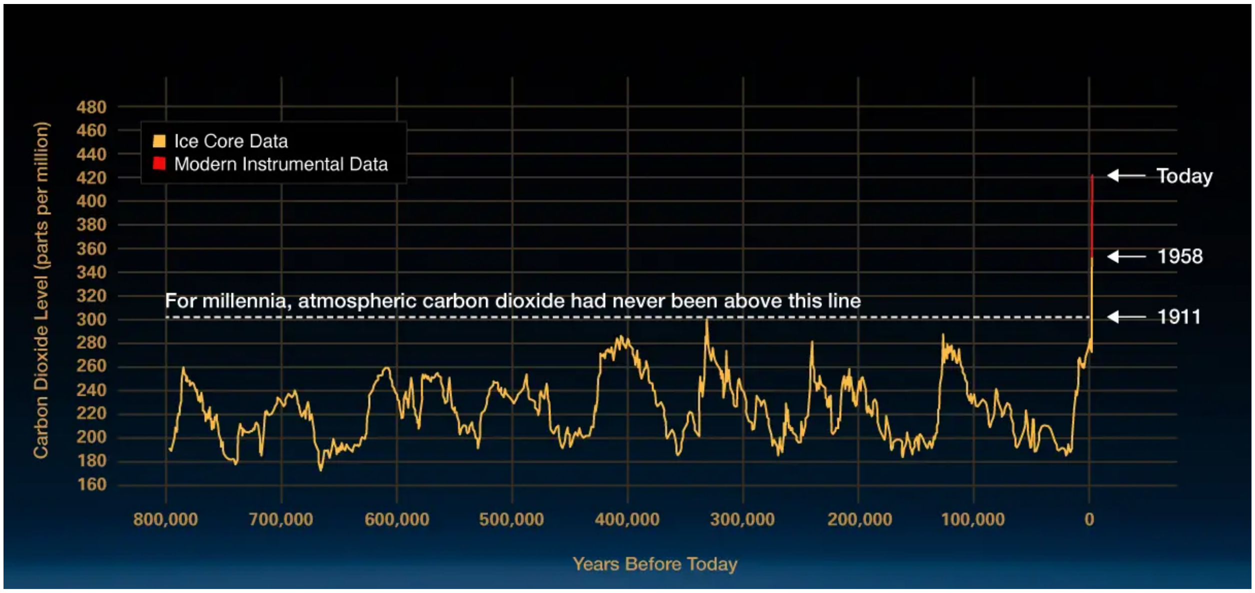


Figure 1: This graph shows the remarkable increase in CO_2 levels over the past century.[9]

As global temperatures continue their upward trajectory, the detrimental impacts on human health and the environment become increasingly apparent. This growing concern highlights the urgent need to understand and counteract the impacts of climate change on a global scale.

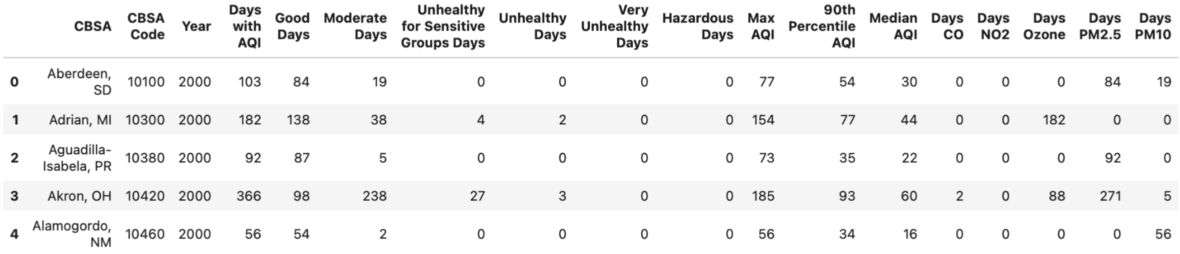
In the face of these challenges, our study seeks to uncover the intricate interplay between air quality, temperature, and geographical location. Our objective is to decipher any potential underlying patterns and potentially pinpoint areas most vulnerable to air quality deterioration. Understanding how temperature and location contribute to pollution levels may offer a unique insight to address climate change. While not providing direct solutions, doubtless, our findings will contribute to the development of effective air quality management plans, facilitate the identification of susceptible regions, and encourage the adoption of sustainable practices.

Data

In order to execute this analysis, we utilized two datasets: AQI data from the EPA (Environmental Protection agency) and annual temperature data from the Washington post.

The first data source was collected from the US Environmental Protection Agency ambient air quality monitoring program[4]. We aggregated the data spanning from the year 2000 up to and including 2019. This data provided us with important information for our analysis (see Table 1), specifically the following columns:

- CBSA Code (Core-Based Statistical Areas)
  - Defined by the U.S. Government Office of Management and Budget as geographical locations of at least 10,000 people.
  - Used for alignment
- Year
- Median AQI
  - This is the primary value we will be using to determine correlation.
- Various particulate data (i.e. CO, NO_2)

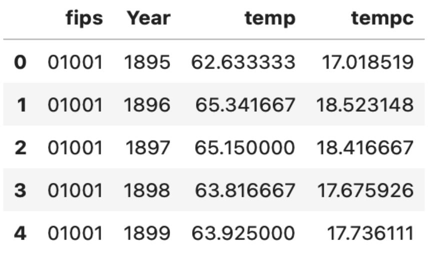


	CBSA	CBSA Code	Year	Days with AQI	Good Days	Moderate Days	Unhealthy for Sensitive Groups Days	Unhealthy Days	Very Unhealthy Days	Hazardous Days	Max AQI	90th Percentile AQI	Median AQI	Days CO	Days NO2	Days Ozone	Days PM2.5	Days PM10
0	Aberdeen, SD	10100	2000	103	84	19	0	0	0	0	77	54	30	0	0	0	84	19
1	Adrian, MI	10300	2000	182	138	38	4	2	0	0	154	77	44	0	0	182	0	0
2	Aguadilla-Isabela, PR	10380	2000	92	87	5	0	0	0	0	73	35	22	0	0	0	92	0
3	Akron, OH	10420	2000	366	98	238	27	3	0	0	185	93	60	2	0	88	271	5
4	Alamogordo, NM	10460	2000	56	54	2	0	0	0	0	56	34	16	0	0	0	0	56

Table 1. Dataframe head of EPA AQI data

The second data source was collected from the Washington Post[5]. This data shows average temperature by year starting from 1895 to the present. (see Table 2).

- Fips (Federal Information Processing Standards)
  - Similar, but not equal to CBSA, so further alignment was needed.
- Year
- temp
  - Average temperature for the year (in Fahrenheit).
- tempc
  - Average temperature for the year (in Celsius).



	fips	Year	temp	tempc
0	01001	1895	62.633333	17.018519
1	01001	1896	65.341667	18.523148
2	01001	1897	65.150000	18.416667
3	01001	1898	63.816667	17.675926
4	01001	1899	63.925000	17.736111

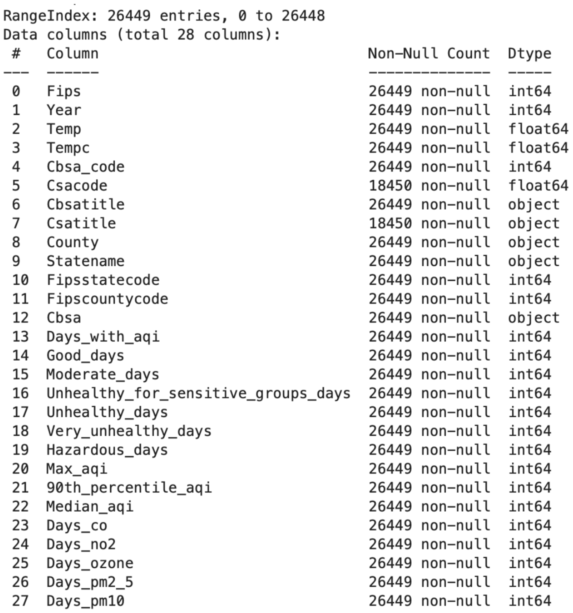
Table 2. Dataframe head of washington post Temperature data

Data Alignment, Cleaning, and Sanity Check

Our initial goal in cleaning the data was to ensure alignment and merging of the two sets was possible. To do this, we used a data set from the National Bureau of Economic Research[6] to directly convert from CBSA to fips and performed the following:

1. Load Air Quality Index (AQI) data set for each year from 2000 to 2019.
2. Concatenate individual years to obtain a single dataframe
3. Load CBSA to fips conversion table
4. Load temperature data
5. Merge temperature data and CBSA to fips data on fips columns.
6. Merge this dataframe with the AQI dataset on CBSA codes.

We then had a dataframe of 32 columns, and 26449 entries in each column. We made the following assumptions based on this final dataframe:



```
RangeIndex: 26449 entries, 0 to 26448
Data columns (total 28 columns):
 #   Column                               Non-Null Count  Dtype
---  ------                               --------------  -----
 0   Fips                                 26449 non-null  int64
 1   Year                                 26449 non-null  int64
 2   Temp                                 26449 non-null  float64
 3   Tempc                                26449 non-null  float64
 4   Cbsa_code                            26449 non-null  int64
 5   Csacode                              18450 non-null  float64
 6   Cbsatitle                            26449 non-null  object
 7   Csatitle                             18450 non-null  object
 8   County                               26449 non-null  object
 9   Statename                            26449 non-null  object
 10  Fipsstatecode                        26449 non-null  int64
 11  Fipscountycode                       26449 non-null  int64
 12  Cbsa                                 26449 non-null  object
 13  Days_with_aqi                        26449 non-null  int64
 14  Good_days                            26449 non-null  int64
 15  Moderate_days                        26449 non-null  int64
 16  Unhealthy_for_sensitive_groups_days  26449 non-null  int64
 17  Unhealthy_days                       26449 non-null  int64
 18  Very_unhealthy_days                  26449 non-null  int64
 19  Hazardous_days                       26449 non-null  int64
 20  Max_aqi                              26449 non-null  int64
 21  90th_percentile_aqi                  26449 non-null  int64
 22  Median_aqi                           26449 non-null  int64
 23  Days_co                              26449 non-null  int64
 24  Days_no2                             26449 non-null  int64
 25  Days_ozone                           26449 non-null  int64
 26  Days_pm2_5                           26449 non-null  int64
 27  Days_pm10                            26449 non-null  int64
```

1. '**metropolitandivisioncode'** and **'metropolitandivisiontitle'** have 90% 'NaNs'. For our analysis, this is not relevant. Both columns can be dropped.
2.'metropolitanmicropolitanstasis' column is not relevant for current analysis. This column can be dropped.
3. 'centraloutlyingcounty' is not relevant for current analysis. This column can be dropped.
4. 'Countycountyequivalent' can berenamed to 'county'.
5. Columns that were used to merge data (CBSA code, fips, cbsacode) were kept to maintain integrity of the dataframe and for any future troubleshooting processes.

Table 3. Table info of combined data set for sanity check

This data set consists of only the lower 48 states with the addition of the District of Columbia, limiting our analysis to this geographical boundary

It was originally assumed that the 'county' column represented a specific county in that state; however, when reviewing the EPA documentation, it was determined that this represents the county in which the EPA reporting site is located. We did notice that the count for each state was not equal, this would cause some states to carry a heavier weight when averaging. This will be accounted for by averaging each state first, and then across all states.

Initial Data Exploration

Now that we have a single dataframe and have checked its validity, we can begin the exploration. We began by looking at the average yearly temperature and the yearly median AQI by state (Figure 2 and Figure 3).

As can be seen in the 'Average Yearly Temperature by State' graph, it is difficult to interpret a trend. There is one notable event that occurs in 2012 where the temperature spikes by 2° Fahrenheit (on average). After exploring the causes, we determined this to be a heat wave in North America[7].

The 'Yearly AQI by State' graph does show a decreasing trend in Median AQI by state. It also reveals that Colorado started as the lowest out of all the states, but rapidly increased to be within family with the other states.

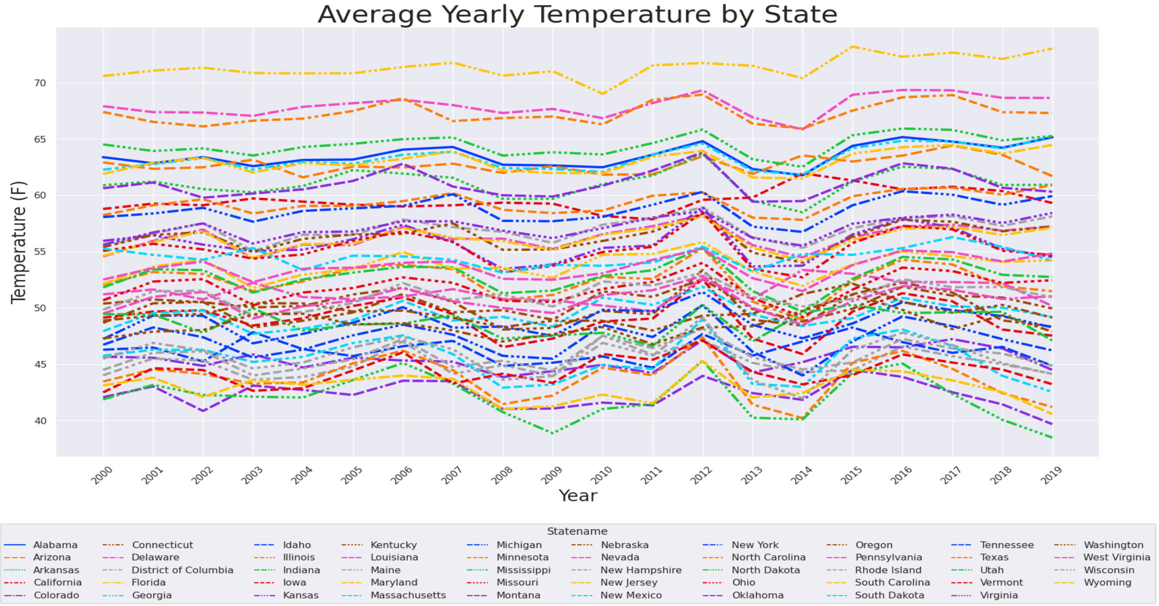


Figure 2. Average temp vs Year by state

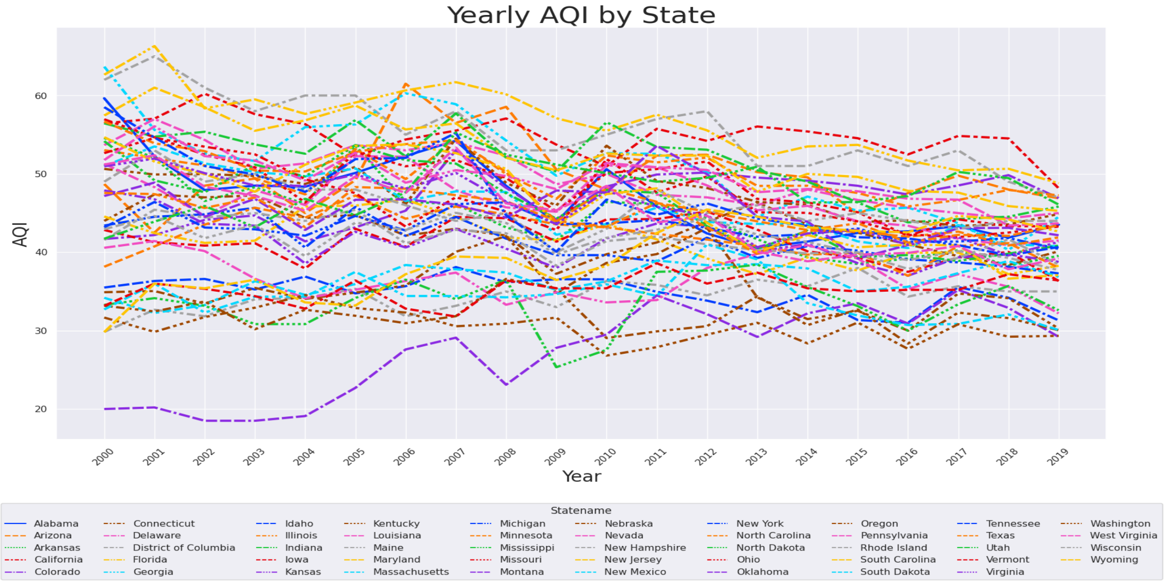


Figure 3. Average AQI vs Year by State

From Figure 4, we can see that the Median AQI shows a decreasing trend while the Average temperature across the US shows an increasing trend, albeit not as big. This is opposite from what we had expected.

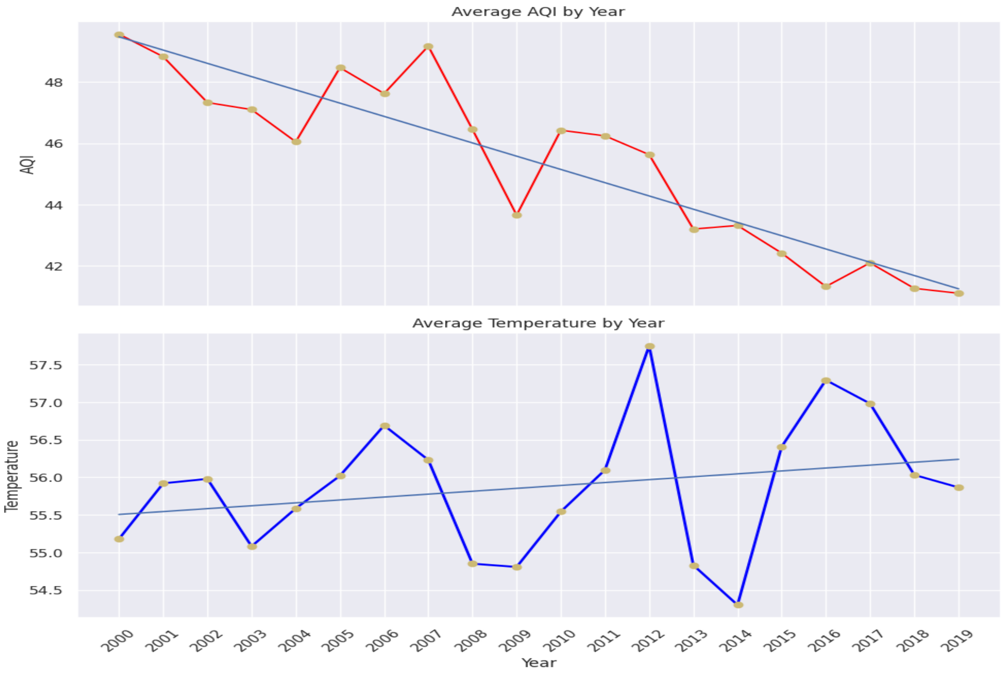


Figure 4. Average AQI and Temperature vs Year

Analysis

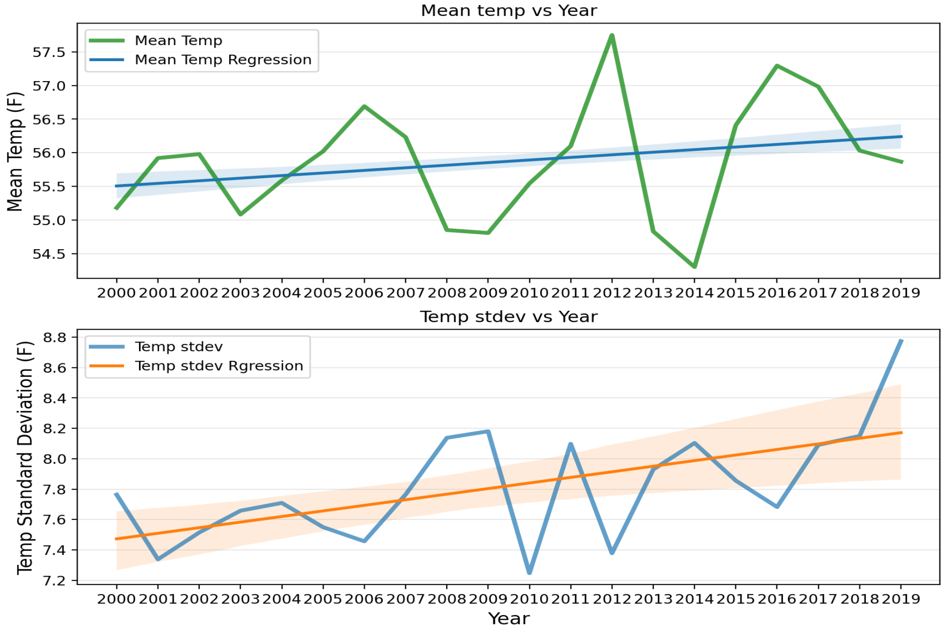


Figure 5. Mean Temperature and Standard deviation of temperature by year.

Time Dependence of Temperature

An exploration on the average temperature data across the US reveals two interesting finding as can be observed in figure 5:

1. There is an increase in mean temperature of about 0.5F across the US in the 19 year period analyzed.

2. The temperature swings experienced in the US within a year shifts more significantly as time passes

The temperature increase confirms the current scientific consensus that temperatures are increasing across the US. The periodicity may be attributed to some time-localized events such as La Nina in 2012 which could be one of the contributors to the temperature increase. One potential next step is to explore what causes these swings in temperature and see if there's any intervention action that could be pulled from these events (ie. man-made ocean currents, cloud formations, etc.)

Although this analysis will not explore the direct relationship between man-made climate change and temperature due to the limitations of the scope and data set, it can be seen from figure 5 that there is not only an increase in temperature over the years, but an increase in the variation per year. These swings in temperature can cause more extreme climate events that lead to unstable conditions around the US. with implications for agriculture (growing temps), tourism, and real estate.

One potential area of exploration is to explore the peak and min years in terms of mean temp in order to gain a better understanding on what causes extreme temperature years and how to control them.

General Relationship between AQI and Temperature

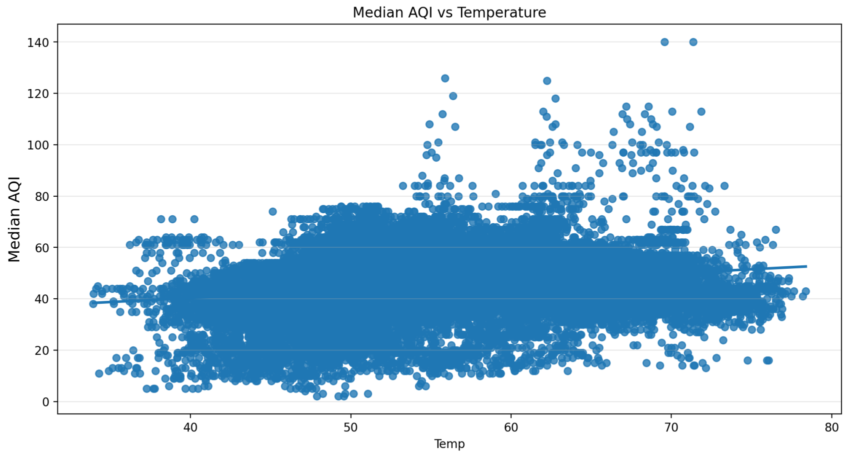


Figure 6. Median AQI vs Temperature - no clear correlation is observed.

Although there is no direct correlation that can be seen between temperature when considering the bulk data-set as partially observed in figure 6, a geographic analysis reveals a trend. We begin our analysis by categorizing the states into their respective geographical regions in the US: South, West, Northeast, and Midwest. To achieve this, we utilize a scatterplot to present a graphical representation of the data. In this plot, each point symbolizes the mean value of the median Air Quality Index (AQI) of all the counties in each respective state (Figure 7).

Examining the graph from figure 7 , we notice potential patterns associated with different regions:

Southern States: Southern states typically display higher AQI and temperature values compared to the median. Given their geographical location closer to the equator, these states naturally have warmer climates, which not only contribute directly to higher temperatures, but may also indirectly influence AQI. The warm temperatures facilitate the formation of certain pollutants like ground-level ozone[12]. Moreover, many of these states contain significant urban and industrial areas, which are primary sources of pollutants[12].

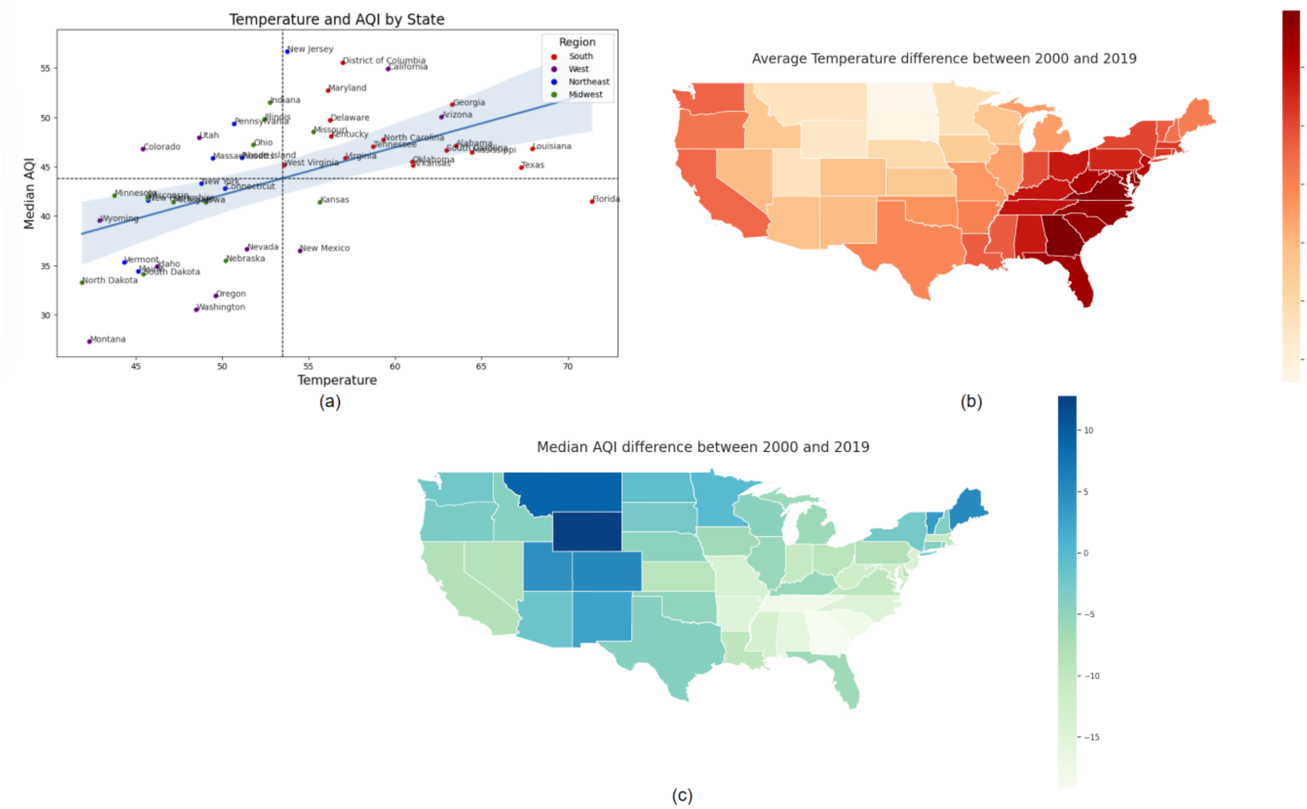


Figure 7. (a) This graph shows the correlation between AQI, Temperature, and states, in quadrant format. It can be seen that there are tendencies of states from the regions to occupy a specific quadrant. (b) This graph shows the average temperature changes of the US states between 2000 and 2019. (c) This graph shows the Median AQI difference of the US states between 2000 and 2019.

Western and Midwestern States: On the contrary, the Western and Midwestern states, often found in higher latitudes or altitudes, display lower AQI and temperature values. Factors such as lower population density, lesser pollution from human activities, and large expanses of natural land, which help absorb or reduce pollutants, may contribute to their lower AQI levels[13]. However, specific areas in the West, like

California, may experience poor air quality due to wildfires, which significantly increase particulate pollution.

Northeastern States: Northeast states usually have moderate climates and hover around the median values for both temperature and AQI. They host large urban areas producing significant pollution, but also have robust environmental regulations in place to manage and reduce pollution levels. Furthermore, geographical features like coastal winds can help disperse pollutants and improve air quality[14] .

These trends highlight the complex interplay of environmental, geographic, demographic, and many other unknown factors in influencing AQI and temperature relationships across regions.

Visualization of 19 change in Average temperature and Median AQI in US

US state-by-state visual indication of changes in average temperature and median AQI can be seen below. Similar to what is discussed above, there are regions of the US that have shown larger changes when compared. Seen in Figure 7(b), the Southeastern US shows a large increase in average temperature between 2000 and 2019. This increase in temperature in this specific region can be attributed to increased El Nino Southern Oscillation, tropical weather systems, and differences in atmospheric regions of the Earth[15].

As for median AQI changes, Figure 7(c) shows a trend of increased median AQI in the western US, specifically the rocky mountain region. As discussed above, this region of the US is surrounded by various mountain ranges. Due to this, coastal winds which have the effect of reducing pollutants, are stopped, which can cause pollutants to accumulate.

The two figures below further support our recommendation for allocation of resources.

Conclusion

Our extensive analysis has revealed a potentially distinctive patterns between temperature and Air Quality Index (AQI) across the U.S. Notably:

- **Geographical and Environmental Interplay**: Southern states have higher AQI values due to warmth and urbanization, while cooler Western and Midwestern states generally show lower AQI with the exception of California and Arizona. Northeastern states maintain median AQI levels potentially influenced by their temperate climates.
- **Potential Temperature Shifts**: Average temperatures in the U.S. have noticeably changed, highlighting the widespread effects of global climate change. Over the 19-year period, there appears to be a consistent warming trend. Additionally, a global event such as 2012's La Nina may have an important contribution to the variation in this observation.
- **Topographical Potential Influence on AQI**: The Western U.S., particularly the Rocky Mountain region, is witnessing a surge in AQI levels. The unique topographical features of mountain ranges may act as barriers, hindering the natural dispersal of pollutants by coastal winds.

Overall, our results highlight how geography, climate, and human actions affect U.S. temperature and air quality. Future research should focus on understanding these patterns at a deeper level and exploring ways to counteract negative temperature changes.

Final Recommendation

Based on our findings, a recommendation for potential resource allocation:

1. **Studies on periodicity in temperature data** - if extreme points can be understood, then additional mitigation measures may be discovered
2. **Prioritize focus on Southeast US for temperature studies**, and **Midwest for AQI studies** due to the relative higher increases in both metrics across these regions
3. **Exploration into adaptive infrastructures -** for instance, the integration of heat-resistant materials for roads and enhanced cooling systems might counteract the repercussions of extreme temperatures.
4. **Examination of temperature's influence on species' behaviors** - understanding how fluctuations affect breeding and migration could prevent imbalances in local ecosystems.

References

1. https://ehjournal.biomedcentral.com/articles/10.1186/s12940-020-00681-z
2. https://www.ncei.noaa.gov/monitoring
3. https://climate.nasa.gov/evidence/
4. Download Files | AirData | US EPA
5. washingtonpost/data-2C-beyond-the-limit-usa: The Washington Post's analysis of NOAA climate change data for the contiguous United States (github.com)
6. Public Use Data Archive | NBER
7. 2012 North American heat wave - Wikipedia
8. https://ehjournal.biomedcentral.com/articles/10.1186/s12940-020-00681-z (Source A)
9. https://www.ncei.noaa.gov/monitoring (Source B)
10. https://climate.nasa.gov/evidence/ (Source C)
11. https://www.epa.gov/air-trends/trends-ozone-adjusted-weather-conditions#:~:text=Variations%20in%20weather%20conditions%20play,cool%2C%20rainy%2C%20or%20windy. (Source Ozone)
12. https://www.jstor.org/stable/26225507 (Source Southern)
13. https://worldpopulationreview.com/state-rankings/state-densities (Source population density)
14. https://www.clarity.io/blog/air-quality-measurements-series-wind-speed-and-direction#:~:text=Coastal%20areas%20or%20regions%20with,have%20originated%20in%20the%20area. (Source wind pollute disperse)
15. Dec 22, 2016, Climate Impacts in the Southeast, EPA.gov, https://19january2017snapshot.epa.gov/climate-impacts/climate-impacts-southeast_.html#:~:text=Climate%20change%20is%20causing%20increases%20in%20temperature%20across,to%208%C2%B0F%20by%20the%20end%20of%20the%20century.

Appendix A: More Supporting Graphs

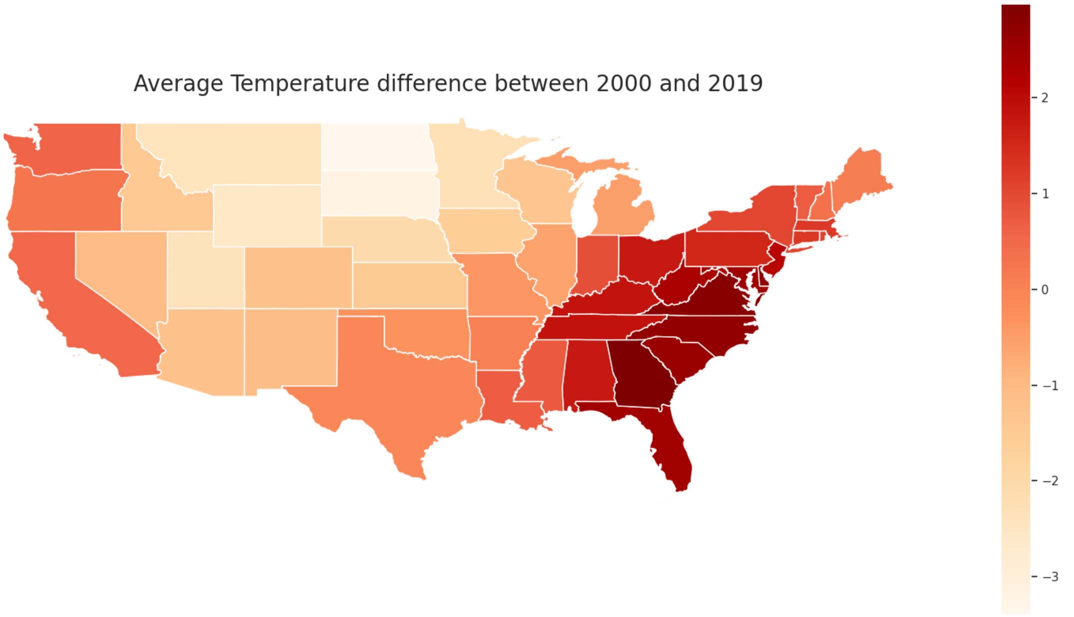


Figure A.1

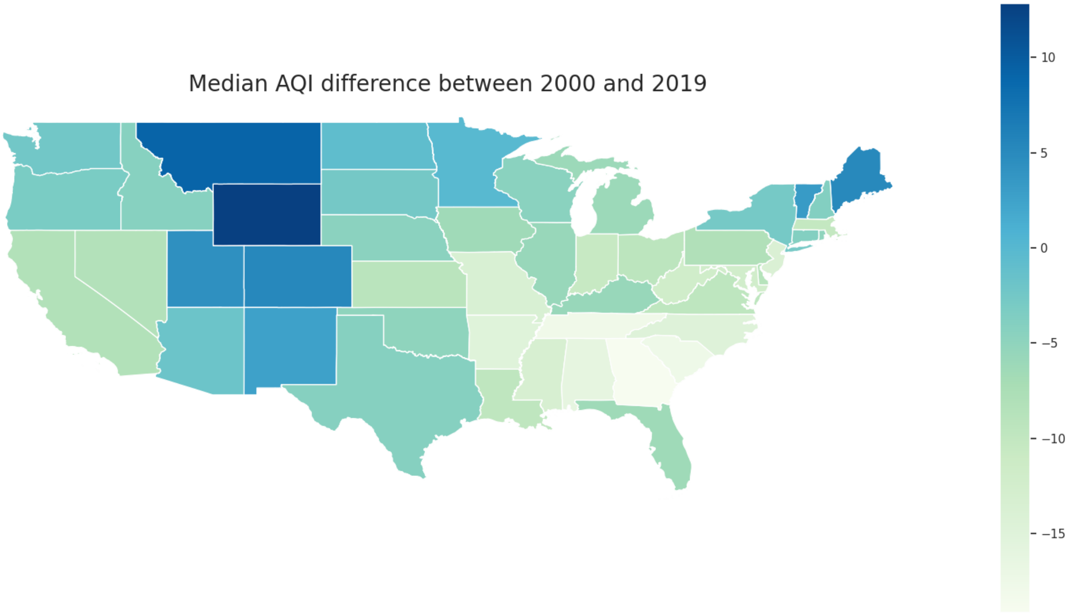


Figure A.2

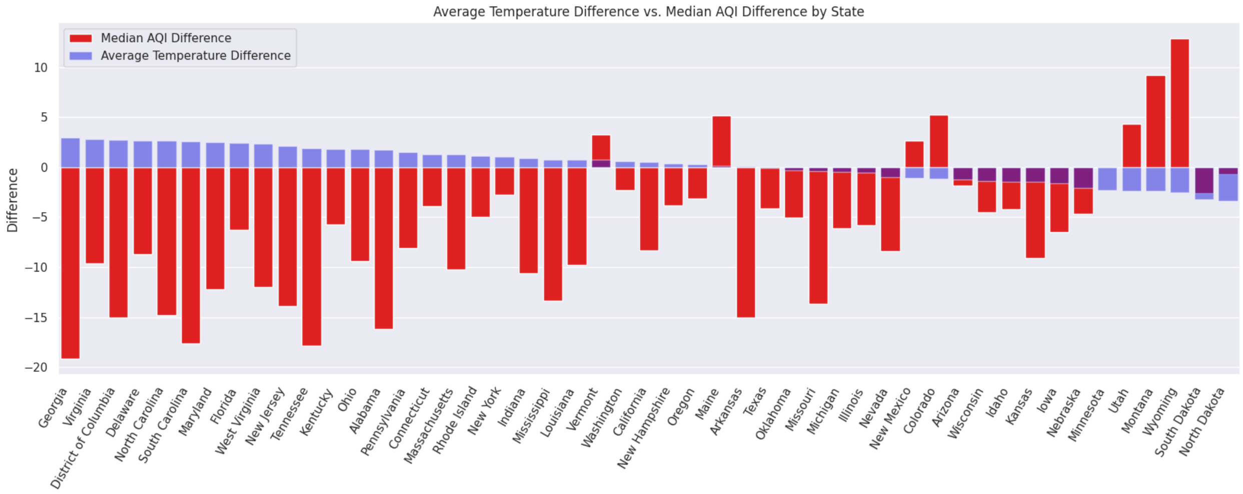


Figure A.3

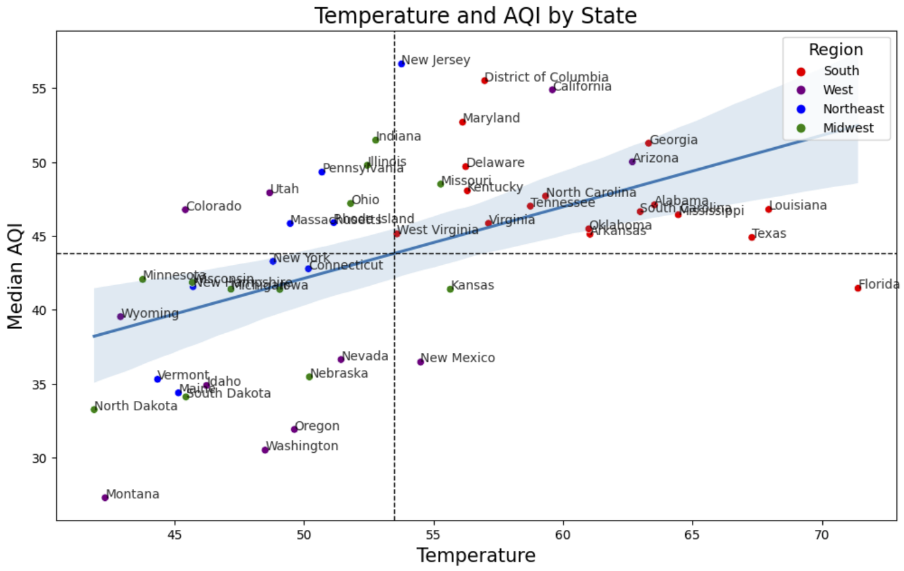


Figure A.4 This graph shows the relative location of states in Median AQI vs Temperature chart

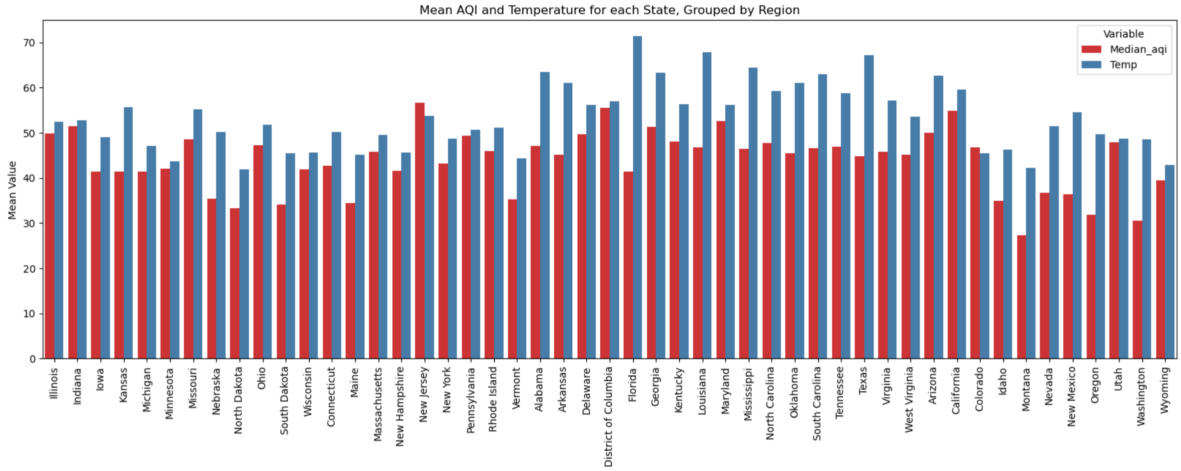


Figure A.5 This graph shows the aggregated mean value of the Median AQI and Temperature bar chart with respect to each state.

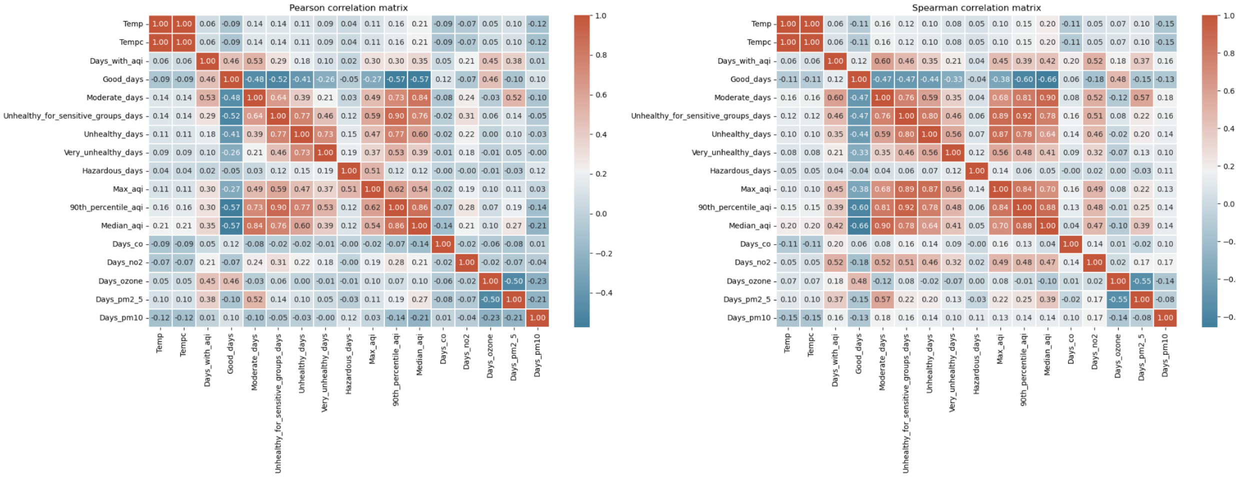


Figure A.6. The charts above show the Pearson correlation chart on the left side, and the Spearman correlation matrix on the right.

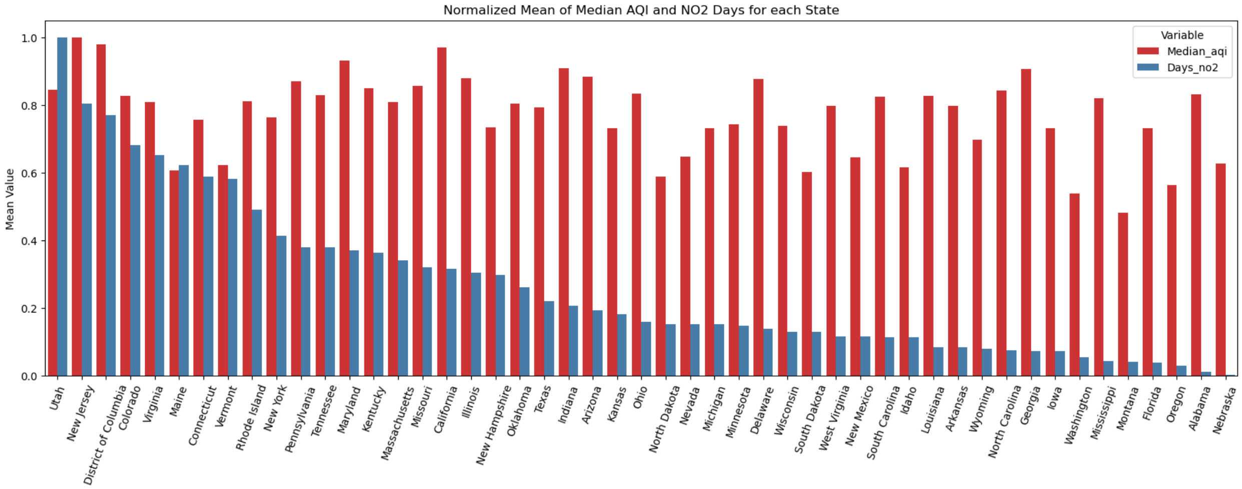


Figure A.7 This bar plot shows the bar plot of mean value of the Median AQI and number of days where NO_2 is the primary pollutant for each state.