Project 1 Write-up

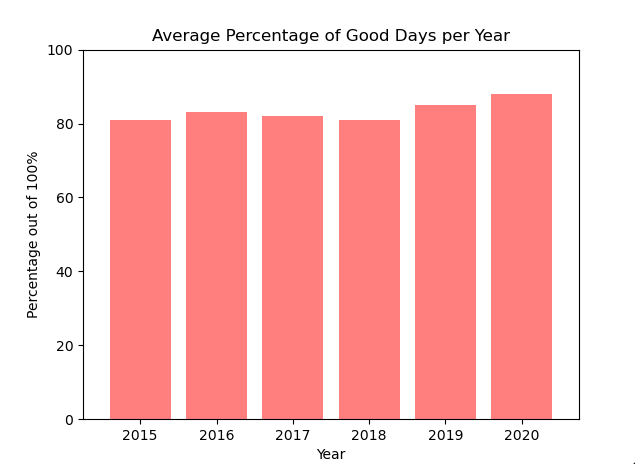
Group 2

UMN Data Analytics Bootcamp

**Covid-19’s Effect on Air Quality in the USA**

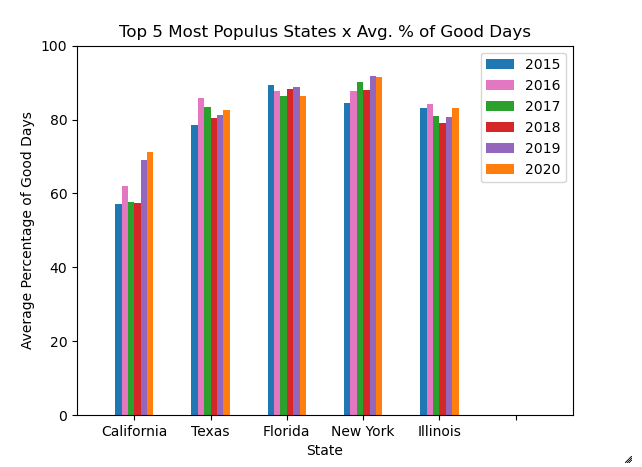
**I. Was there a significant increase in the number of “Good Days” in 2020 when compared to years prior?**

The main measurement system for AQI in the U.S. being a daily rating system between 50-500 and which outlines days in which were “Good”, “Moderate”, Unhealthy for Sensitive Groups”, “Unhealthy”, “Very Unhealthy”, and “Hazardous”. As it is the main measurement system and the basis of which the EPA evaluates Air Quality, it seemed like a safe choice to base our studies on. We didn’t realize at the time that the U.S. generally has good Air Quality, thus most days are considered “Good”. Additionally, the EPA doesn’t evaluate the AQI every single day, so some years have less days to go off and some have more. With those things considered, “Good Days” might not be the best way to evaluate the significance of 2020 to Air Quality scores, but nonetheless our study showed slightly more “Good Days” compared to the years 2015-2019.



Though the years were relatively close, 2020 saw the highest percentage of Good Days at 88%, indicating that quarantine might have had some significance in how many “Good Days” a given year had, but it seems too close to be conclusive. The range prior to 2020 goes from 80.75% to 84.85%, with no linear progression from years 2015-2019.

Taking this further, we conducted a similar test that analyzed the number of “Good Days” for the top five most populous states, which were California, Texas, Florida, New York, and Illinois according to the 2021 U.S. Census.



According to this data, California showed the lowest percentage of good days compared to other states analyzed. However, it also saw the most significant increase between years with 2019 having the highest increase but 2020 having the most days per year. Three out of the five states saw an increase in “Good Days” from 2019 to 2020 and the two that did not were Florida and New York City. All states did however see an increase between 2018 and 2019, possibly indicating that COVID-19 worries beginning in October may have affected the states with more people, though that is just a theory. 2018 - 2019 shows some significant jumps though, especially in California and New York. Additionally, New York shows the highest amount of “Good Days” being the only state to break 90% for one given year (2017).

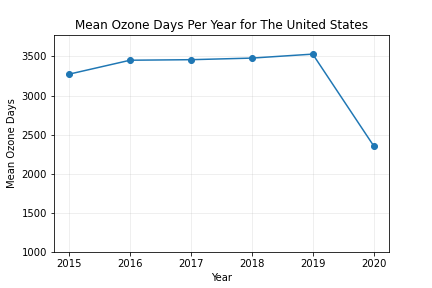
To answer the question, we did see an increase of “Good Days” between 2015-2019 and 2020, however the next question would be to determine whether or not it is “significant”.

**II. Was there a significant change in the number of “Bad Days” in 2020 compared to previous years?**

When we began analyzing the data, we noticed that there was not much data for “Bad Days” to analyze. That suggests that the air quality in the U.S. is generally good. So instead of analyzing “Bad Days” we decided to analyze “Moderate Days.” When looking at “Moderate Days” there does seem to be a downward trend in 2020. We thought it would be a good idea to take a closer look at a highly populated area as another way to compare. So, we isolated New York state and observed the same downward trend. Therefore, there does appear to be a change in “Moderate Days” from previous years. There does not appear to be much of a change when looking at “Days CO throughout the years. There are random spikes in “Days CO” throughout time.

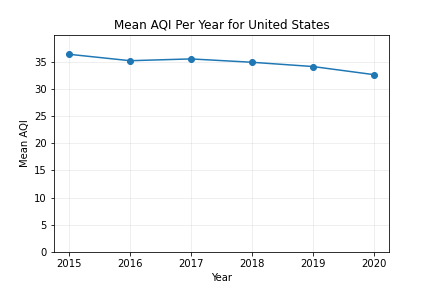
**III. Was there a change in Ozone emissions from 2015-2019 to 2020? Did we emit less Ozone pollution in 2020?**

When analyzing the change in Ozone emissions from 2015-2019 and comparing it to Ozone pollution in 2020 we found a significant decrease in 2020. Following trends, the days where Ozone was the main pollutant was growing from year to year until 2020, where it decreased by over 1,000 days alone across the United States. Ozone pollution that is caused by humans results mainly from different forms of transportation such as cars and planes, power plants, refineries, and chemical plants. When plotting the Ozone days year from year this is what we got.



**IV. Was there a significant change in the mean AQI year over year?**

What we were most interested in was if the mean AQI, or Air Quality Index, had changed significantly in 2020 compared to the previous years. Starting in 2017 the AQI was already starting to trend downwards, meaning that we were starting to see cleaner air year after year. In 2020 there was an accelerated drop in the AQI, although it was not as drastic a drop as we saw in Ozone pollution.

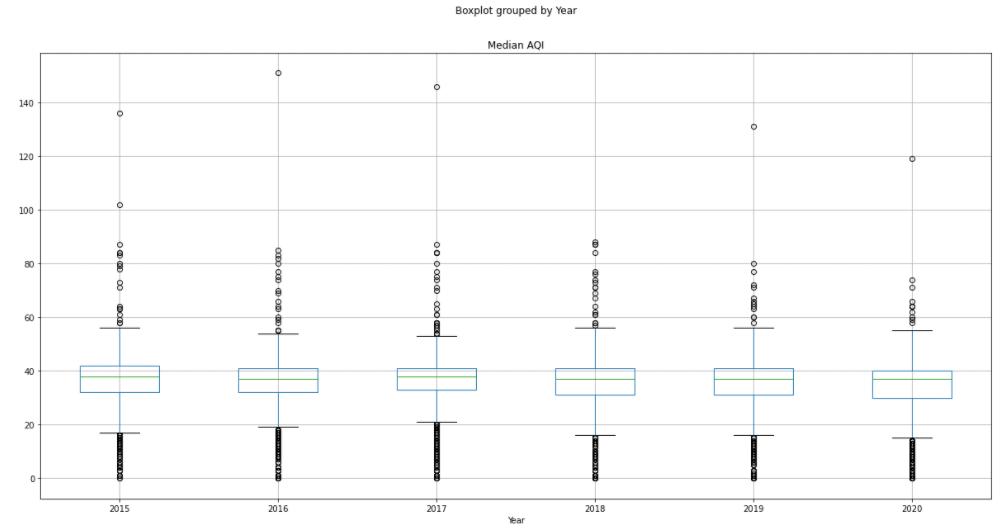


**V. Hypothesis Test:**

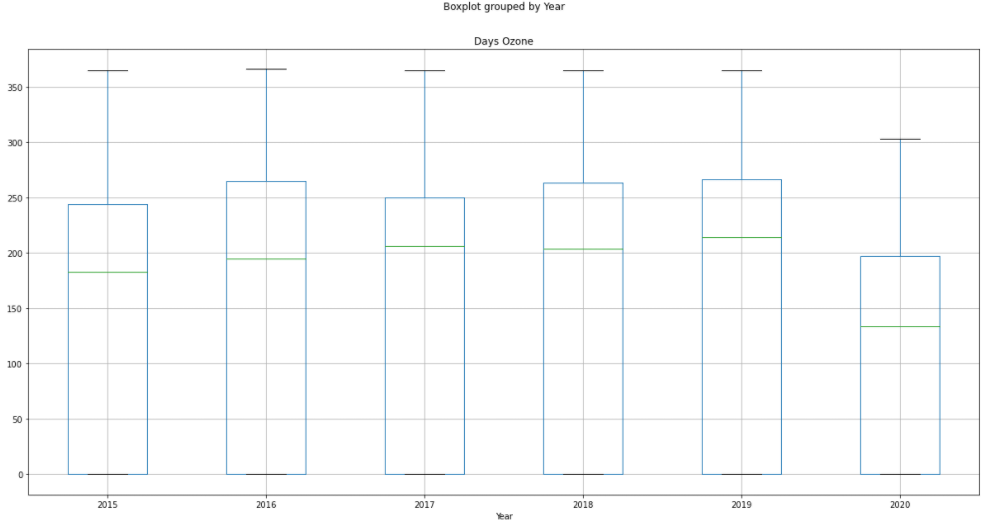
**Hypothesis:** In 2020 we expect to see cleaner air and less pollution because of Covid-19 and Quarantine.

**Null Hypothesis:** There was no change in Air Quality levels due to Covid-19 and Quarantine.

In running our main hypothesis test, we performed an Anova test which examined if there was any statistical significance that showed Median AQI becoming lower after 2020 had occurred. As a result, we received a pvalue of 0.10533. In removing 2020 data we got the pvalue to equal 0.56979 which not only removed the year we were focusing on, but also still did not allow us to reject a null hypothesis. We tested a series of different hypothesis tests including a secondary Median AQI x Year, and Days Ozone x Year. Nothing seemed to hold any significance, and if it did it was only after the data from 2020 was removed.



Median AQI x Year



Ozone Polluant Days x Year

**VI. Post Mortem**

If given more time and a different start to the project, we would have loved to look at multiple datasets. From the beginning we decided to go simple and then add to what we had if time would allow, but once it came time to add data, we had a hard time finding the data we were most interested in. When it came time to do Hypothesis Testing or Regression Lines, we tried performing both. Calculating lines of regression and correlation coefficients was difficult because our dataset didn’t give us many different facets to compare. It seemed we were just comparing each variable to time (year), in which a lot of the graphs were repetitive and didn’t seem to be effective. If we would have been able to find a dataset on transportation data, or data on the number of US residents who have switched to working from home, we could have tested data against each other to see if there were any correlations. We did not have the proper data to do so, unfortunately.