Conclusions and Write Up for Air Quality Changes Associated with Coronavirus

Introduction and Conclusions

My project's purpose was to determine the way air quality behaved during the first part of coronavirus stay at home orders and shutdown and what effect coronavirus stay at home orders and shutdown had on air quality generally. I used the airnow daily peak data files from the airnow air quality archive for my resources,testing, and analyzing. I used Python 3 with Jupyter Notebook on top of Anaconda in order to extract, store and analyze the data, finally plotting my conclusions and the relevant data as well as writing txt files for other non visual analysis and testing. I came to the conclusion that in general the air quality has improved (lower particle count for each measure) due to the coronavirus shutdowns with the possible exception of the PM10-24hr measure. I also concluded that a linear model is not appropriate or meaningful when describing air quality behavior during the first part of coronavirus stay at home orders and shutdown and that it does not seem another type of model (exponential, quadratic, power law, etc) would be more appropriate. I wrote several programs that included the programming elements of iteration, ordered and unordered variables (lists, dictionaries, tuples, and sets), line splitting, functions and abstraction, statistics, and plotting to accomplish my project's purpose.

WorkFlow and Programs

Line Graphs and Scatter

The first program I wrote plotted line graphs for 2018, 2019 and 2020 for one city and each measure over a time period representative of coronavirus shutdown: March through April (the data available at the time). The city I chose to plot was Los Angeles. I was hoping to see a sudden downspike in the measure values but there did not seem any major effects. The line plots were not very conclusive but showed that generally it seemed there was better air quality in 2020 than in previous years over the time period. However with just one city I could not make any general or population conclusions.

Next I modified the above program to plot not only line graphs but scatterplots, least squares regression lines, residual plots, and histograms of the residuals. My idea was to model the behavior of the air quality after the shutdowns to see if the air quality got worse or better as the shutdowns progressed. I chose a different time period for these plots than the line graphs choosing March 19 through the end of April generalizing the time period for shutdown to the first state stay at home order in California. The scatter plots would show the emissions in regards to time and the least squares regression lines would model if there was a linear association after

the initial shutdown using the correlation coefficient ‘r’. The residuals and histogram of the residuals would help along with the correlation coefficient to determine if a linear model was an appropriate one. I finally added a feature that allowed the user to choose between line and scatter and then choose a location from a given list generated by the data and plot the respective graphs. Generally it seemed with the line graphs that the air quality in 2020 was better than in previous years. Also it seemed that there were generally very weak positive correlations in the scatterplots and less commonly weak negative correlations in the scatterplots. While the residuals showed no clear patterns their histograms were not often symmetric or approximately normal. Based on the scatterplots it also generally seemed like no model was appropriate for the data, not just a linear one. While it seemed I could gather some conclusions from this program there was not a lot of evidence and I could not generalize the conclusions for several reasons. Some of the data came from locations not in the United States so the month generalizations may not have been appropriate because of the difference in time between the United States and other countries' responses to the virus. Also because of the generalization in start date some scatter plots and fitted lines could be misrepresenting some United States locations who had shutdown orders later. Some areas may generally have better air quality due to having a smaller or less dense population or due to location.

To address these problems I modified the above program to output and save the figures and analysis for major cities inside the United States that I determined were included in the data files. Also because the first stay at home state order was from California I chose three locations from that state to plot and analyze so it had more weight than the other locations. I also chose Santa Fe as well as Albuquerque, even though it is smaller, because of the tourism and resulting traffic under normal conditions. The rest of the locations were large cities present and well represented in the data files. This program saved these figures for analysis. With all of these results, I was able to then conclude that generally for these larger cities the air quality was better in 2020 than in previous years based on the line graphs and that there wasn’t a linear trend or very appropriate model for the air quality after shutdown.

Difference in Means and Standard Deviations for 2019-2020

To further statistically address the general differences of levels of air quality between 2020 and 2019 I created a program that found the common cities between the years that recorded each measure, then iterated through those measures and years pulling the respective data into a dictionary titled with the measure, keys for the cities, then recorded a list of the means and standard deviations for each city. It then subtracted those two distributions from each other (2019-2020) recording those values and finally plotting the differences in means, the greatest differences in means, and the difference in standard deviation (as a measure of spread/variance) for each measure and for the cities as bar graphs. Positive values in the subtracted distributions would indicate lower means in 2020 and negative values subtracted distributions would indicate higher means in 2020. The larger standard deviations of the subtracted distributions have a higher variance and spread making it more difficult to make conclusions about unbiased/biased estimators. This program saved these figures for analysis. I was able to conclude from the figures that since most differences in means were positive and

had a greater magnitude of difference than the negatives generally the average air quality was better (the means for measure values were lower) in 2020 than 2019.

T-tests for a Mean of Differences

Finally, using similar methods to the program directly above I extracted data for the differences in distributions for each measure and cities and making sure the conditions for inference were met, I performed a t-test for a mean of differences (paired t-test between two populations/samples; in this case each city was there in 2019 and 2020) to determine if there was statistically convincing evidence that the population mean of each respective measure for cities similar to those in the sample for 2020 was lower (better air quality) than the population mean of each respective measure for cities similar to those in the sample for 2019 at the 0.05 significance level. I performed this t-test for 5 measures except PM10-24hr as it did not meet the conditions for inference (the population was too low, less than 300). The program writ these conclusions to a txt file. The t-tests for CO-8hr, Ozone-1hr, Ozone-8hr, and SO2-24hr all found convincing evidence that the population mean of each respective measure for cities similar to those in the sample for 2020 was lower(higher air quality) than the population mean of each respective measure for cities similar to those in the sample for 2019 at the 0.05 significance level. The t-test for PM2.5-24hr however did not find convincing evidence that the population mean of SO2-24hr for cities similar to those in the sample for 2020 was different than the population mean of SO2-24hr for cities similar to those in the sample for 2019 at the 0.05 significance level. So I can generally conclude that globally, locations with access to air quality measuring equipment had lower population means in 2020 for CO-8hr, Ozone-1hr, Ozone-8hr, and SO2-24hr and not for PM2.5-24hr. I therefore can conclude that air quality has improved due to coronavirus shutdown and stay at home orders for most measures.