Karina Sirabian 10/20/20

**Activity 7: Leading you own data analysis**

1. During this project, I would like to improve my skills in interpreting data and being able to generate insightful connections/correlations between variables. I would also like to improve how I communicate my results to an audience both in an written description and producing graphs/charts that are easy for people to understand.
2. For this project, I would like to explore if COVID-19 effected air pollution. I will specifically explore the New York City region and how quarantine/the city shutdown impacted the air pollution because there were less people working, less public transportation, etc. I will also be looking at fine particle matter air pollutant (PM2.5), but I may also look at other air pollutants such as sulfur dioxide.
3. Questions: Has the current pandemic impacted the air pollution in the NYC area due to stay-at-home orders? Are the air pollution levels in NYC in 2020 (post NYC shut down) statistically different than in previous years? Could a predictive model from air pollution levels before 2020 possibly predicted air pollution levels in 2020? Hypothesis: Air pollution during the stay-at-home period in NYC is less than air pollution normally.
4. Using the air pollution link on the project sheet, I found air pollution data for different sites in the greater NYC Area. The air pollution data has daily measurements of the mean PM2.5 concentration (and also a corresponding daily aqi value) for different sites near NYC. Each csv file corresponds to a different year, so I will be comparing the data from 2020 (after quarantine rules in NYC) to data from previous years.
5. In April 2019, the mean daily amount PM2.5 in Morrisania was 4.65 μg/m^3 with a standard deviation of 2.15 μg/m^3, while in April 2020 the mean daily amount of PM2.5 in Morrisania was 4.06 μg/m^3 with a standard deviation of 1.95 μg/m^3.



1. I am still not positive on what I want to specifically look at; however, I have many ideas. I could compare just the air pollution levels in 2019 to 2020 and run a two sample t-test, or I can look at multiple years and perform an ANOVA test to see if 2020 is different than the other years. However, I may have to break the analysis up by month or even day and year instead of just looking at years because in January, February, and most of March people were not quarantining yet, so this wouldn’t be our experimental data. However, we could treat these first three months as a control group for 2020 because there is a possibility that 2020 could be different than 2019 but not because of people staying inside (additional buildings/factories, population increase, etc.) I will have to look up the exact data of when people began to quarantine and when schools/facilities were shut down in order to not mix up the control group of 2020 (when people were out/working) and the experimental group. Once I get the data of these two groups from 2020, I can compare it to previous years using either a two sample t-test or an ANOVA test. I can do this for multiple sites to look for trends of air pollution during the pandemic across the greater NYC area. Another idea I had was to look at air pollution data from a long time ago (maybe the past 30 years) up to 2019 and try to create a predictive model for the year 2020. I would do this by creating a regression model of data from before 2019 and then using this model to try and fit 2020. Then I would see if the 2020 data for each site falls within the uncertainty of the model (for each site) and judge how drastic of a change covid made on the air pollution of the greater NYC area.
2. I think it will be difficult to compare the results from different sites. For each site, I am planning on finding out if there is a statistical difference in air pollution during normal times to during the quarantine period, but I am not sure how to quantify this in order to compare it to different sites and which graph to come up with to display the information from multiple sites into one.