James Yang
DATA 512
11/8/22

Part 2 – Extension Plan

Problem Statement:

In America, there is a theory that the cold can make you sick. Whenever you get a runny nose, begin coughing, or feel week, the first instinct is to get warm. There are many studies that suggest that we don't get sick from being cold. However, these studies are mostly in regards to the cold or flu. The problem in which I am trying to expand upon is whether COVID-19 is impacted by the weather, and how the Bergen county in New Jersey population fairs with this theory that the cold can inherently make you sick. Does temperature impact human susceptibility to COVID-19? Or is the temperature a non-direct factor like it is for the cold or flu?

Research Questions

In this project, we hope to extend upon the existing research that weather has no impact on the cold/flu and whether that concept applies to COVID-19. Dr. Fecher says "You can't get sick from being cold in general, whether you are outside or inside. Can you get sick from being cold? Yes, but not in terms of a cold or the flu" [1]. This begs the question, does COVID-19 follow the same trend as the cold or flu? How are we going to determine whether it follows a trendline? What is the baseline?

Data Used

For this problem, I first pulled from the Rutgers climate database that contained the max, min, and average temperature for each county in New Jersey [2]. I cross checked this data with different weather websites for archived temperatures and determined that they were accurate through some averaging of sampled weeks. The site allows for open use of its data and requests for a reference as its constraint of usage. This data will allow for further analysis of weather impacts within the Bergen county, obtaining one of our most important features. This feature being the average temperature recorded in each month. This dataset expands on the common analysis by introducing a new factor of weather; this feature is independent of COVID-19, as the weather will not depend on COVID-19 cases.

Another dataset I used is the California Outpatient Influenza surveillance dataset that contains a set of flu confirmed cases in the past 20 years. Unfortunately, I was not able to find a New Jersey dataset on the flu, but this dataset should mimic the idea of flu cases found in New Jersey [5].

Unknowns and Dependencies:

When recording average temperature, there is a difference in the temperature displayed versus the temperature felt. There is a difference between air temperature and "feels-like" temperature; air temperature is the actual temperature outside and "feels-like" temperature is the wind and humidity combined with the air temperature [3]. This discrepancy can cause a little bit of unknown as the dataset doesn't explicitly say which type of temperature it is referencing.

Methodology:

We start by doing data cleaning with regards to 3 datasets; the first dataset is a Kaggle repository of John Hopkins University COVID-19 data, the second being a CDC dataset of masking mandates by county, and the third being a New York Times mask compliance dataset [4]. In order to standardize and clean the three datasets, we needed to link the raw us deaths with the state associated to the New Jersey state and the columns associated with the Bergen county. Another thing we needed to clean is the dates associated with each state and the FIPS value that links with the mask use by county. Lastly, the number of deaths and the date associated with each date needs to be standardized across each of the datasets so they can be interpretable. These are the inner workings of the figure 1.1 produced below.

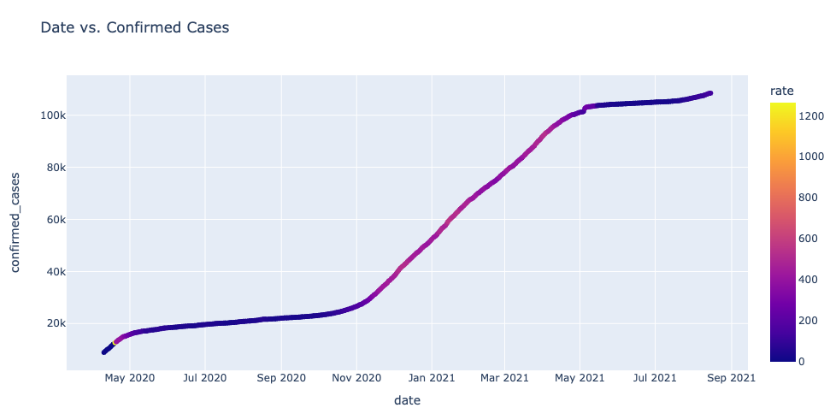

Figure 1.1: Depicts the visualization.

The visualization shows the dates vs the confirmed amount of cases with a corresponding rate on a color coded map of New Jersey in the county Bergen. The visualization itself contains all moments of masking policies in effect because New Jersey instantiated the mask policy from the beginning of COVID.

The next step is producing a rate progression of confirmed cases over a set amount of days. I have chosen 10 to average 3 rate changes per month. The rates increase a bit through mid November 2020 to May 2021 before flattening out again, creating a modal distribution in figure 1.2.

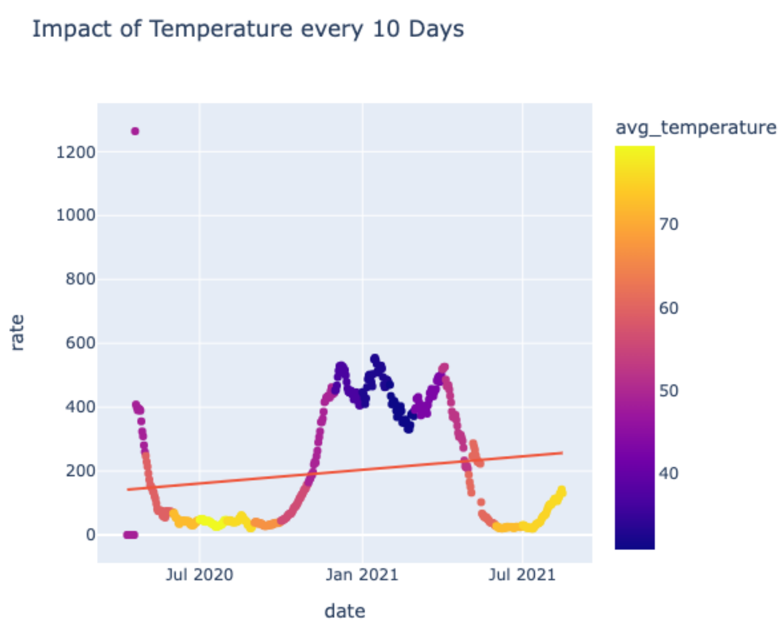

Figure 1.2: Depicts the rate progression of confirmed cases every 10 days in New Jersey, Bergen county.

When importing the temperature and transforming it to align with the dates associated with the cases, we notice an interesting trend with regards to the average temperature. In figure 1.3, we can see that the average temperature dramatically drops as the rate increases. We can see that the OLS trendline has the rate trending upwards with the scale of the same dates as the flu figure 1.5.

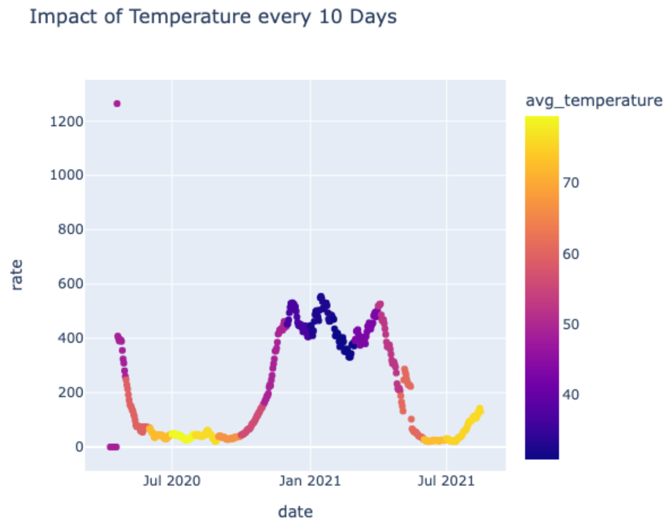

Figure 1.3: Depicts the impact of temperature every 10 days for New Jersey, Bergen.

When we compare with the confirmed cases of the flu in California, the t-test gives us a result that they are not similar. The p-value is less than 0.05 in the case of the null hypothesis being that they have the same trend. That means we reject the null hypothesis and believe that COVID-19 does not trend the same way that the flu does. This can also be reinforced through figure 1.2 vs. figure 1.5 and the difference in trend that they produce as well. However, this idea will need to be reinforced through more examples and datasets.

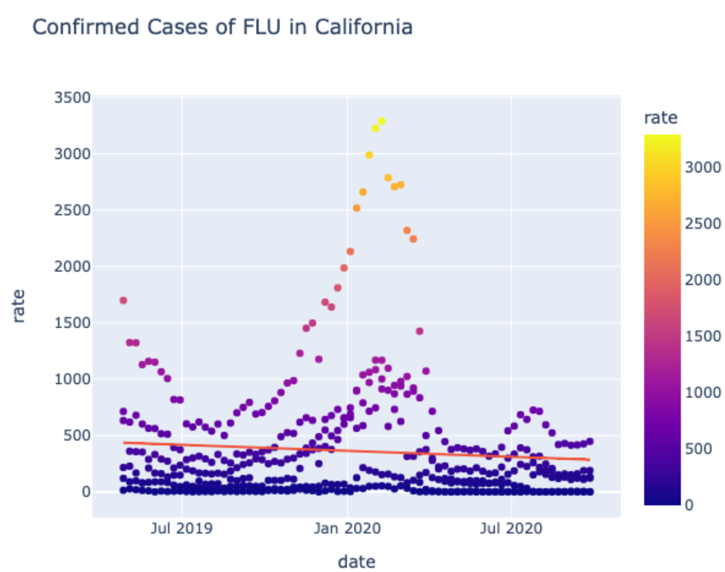

Figure 1.5: Confirmed flu cases in the state of California with 10 day rate vs. time.

Timeline to Completion:

Course Project 2	Setup an initial outcome (This paper here)	**11/10/22**
Data Clean for more sampled data	Get more sources of Flu and Common cold outbreaks to create an ANOVA rather than a t-test	**11/15/22**
Finish the integration of modeling	See the results of ANOVA and expand, otherwise produce results	**12/1/22**
Implement presentation	Create presentation for the project	**12/5/22**
Finish the project	Have everything completed in a repo	**12/12/22**

References:

[1]. "Do You Really Get Sick from Being Cold?" *UnityPoint Health*, https://www.unitypoint.org/livewell/article.aspx?id=9161c3a0-54dc-46d7-a633-e1d15a5227e2.

[2]. *ONJSC :: Historical Monthly Summary Tables*, https://climate.rutgers.edu/stateclim_v1/nclimdiv/.

[3]. ThermoPro. "Temperature vs. Feels like Temperature." *Thermopro*, 15 Dec. 2021, https://buythermopro.com/knowledge/temperature-vs-feels-like-temperature/#:~:text=The%20air%20temperature%20is%20the,and%20how%20we%20should%20dress.

[4]. https://www.kaggle.com/datasets/antgoldbloom/covid19-data-from-john-hopkins-university
https://data.cdc.gov/Policy-Surveillance/U-S-State-and-Territorial-Public-Mask-Mandates-Fro/62d6-pm5i
https://github.com/nytimes/covid-19-data/tree/master/mask-use

[5]. "Influenza Surveillance - Outpatient Influenza-like Illness Surveillance Data by Region and Influenza Season." *California Open Data*, https://data.ca.gov/dataset/influenza-surveillance/resource/78afd269-ea5b-41aa-9ca8-37af6bb35f5a.