Yuchuan (Helen) Ma, Guhui Zhang

COSC89.21

March 11, 2021

Relationship between temperature and COVID spread in the U.S.

1. Research Question

The relation between weather and the spread of COVID-19 has been suspected and corroborated by researchers since the pandemic's initial outbreak. Through observational studies focusing on 9 major cities around the world, Rouen et al. (2020) detected "a negative correlation between temperature and new daily cases growth rate with an average lag of 9.7 days." The study of Xie and Zhu (2020), with a concentration on Chinese cities, discovered that temperature rise in the last one or two weeks is positively associated with newly diagnosed cases of the day. In this project, we aim to examine the existence of similar relations in the United States context, with a primary focus on a selection of six states, including California, Texas, Florida, North Dakota, South Dakota, and Rhode Island. Furthermore, this study inspects the strength of the correlation of the temperate-spreading rate among counties inside these states. We chose them because as of March 10, 2021, the former 3 states have the most significant number of confirmed cases, and the latter 3 have the most significant number of cases per million of population.

2． Methodology and Data

2.1 Data Source and Collection

We use two major sets of data: 1) the COVID-19 infected cases data and 2) the weather/temperature data.

For the former, we rely on the Coronavirus Data in the United States dataset curated by The New York Times, available in both webpage form and on Github. The data tracks historical data of COVID infection within each county since January 21, 2020. We also refer to the API of The COVID Tracking Project of The Guardian, which provides historical data on the population of infected cases by state and of the entire U.S.

Furthermore, we scrape the daily COVID-19 updates provided by Worldometer to obtain the latest statistics (of March 10, 2021) by state.

We utilize an API from the National Oceanic and Atmospheric Administration and access its Global Historical Climatology Network Daily (GHCND) dataset for the latter weather data. The dataset offers historical weather data on a weather station level that can be queried with FIPS code, a federal geographic information code for counties, and county-equivalent, if available. We collect the average temperature of each day (TAVG) in Fahrenheit. Since two weather stations observe two TAVG values for a given day and location, their mean is taken to represent the overall mean temperature. Despite our effort, we are not able to obtain dew point data to calculate humidity because they are no longer offered in GHCND since 2010.

2.2 Analysis Method

We generate two columns of calculated data on the case number dataset per county for preprocessing: 1) number of daily increased cases and 2) COVID spreading/case increase rates in percentages. We choose to use the normalized spreading rates for correlation calculations (instead of raw numbers) to account for differences in geographic sizes and populations across counties. For the weather data, we take means of different day ranges to generate averaged 3-day, 5-day and 10-day temperatures.

We explore the correlations between the spreading rates and the three ranges of averaged temperatures for each county corresponding to a FIPS code within selected states. The correlation results are then collected into a data frame (one for each state). While Pearson's coefficient measures linear correlation, the other two compare the data's rank and thus capture the existence of a monotonic association between the spreading rates and average temperatures. To facilitate calculations, we add 0.0000001 to deal with zero-values in the spreading rates, i.e., when there are no increased cases during the given day, and further take the natural log to avoid floating-point errors.

We adopt K-Means clustering to cluster counties by their correlations between the spreading rates and averaged temperatures. Linear correlation is the focus of our study, and both Spearman's rho and Kendall's tau coefficients fail to demonstrate strong correlation, so we choose three Pearson's coefficients, each for a temporal range, as the model features. The Elbow/Knee method is used to determine the number of clusters for the model with the lowest Sum of Squared Error (SSE). Three-dimensional graphs with four color-labeled clusters (each data point representing a county) are generated at last.

3. U.S. Covid spread overview

Figure 1 shows an overview of the spread of COVID-19 in the United States from January 13, 2020 to March 7, 2021.

As evident from Figure 1, which graphs the number of daily increases of COVID-19 confirmed cases, there have been 3 significant transmission waves that peaked during April 2020, July 2020, and January 2021, respectively, and resulted in subsequent rises in the number of total confirmed cases in the following weeks. Since the beginning of 2021, the spreading rates of COVID-19 started to decrease and have dropped significantly by March 2021, which could be due to public vaccination and the developing group immunization. On the other hand, the number of total deaths keeps rising. Fortunately, it demonstrates a linear increase over the past year instead of a proportional growth to the number of cases.

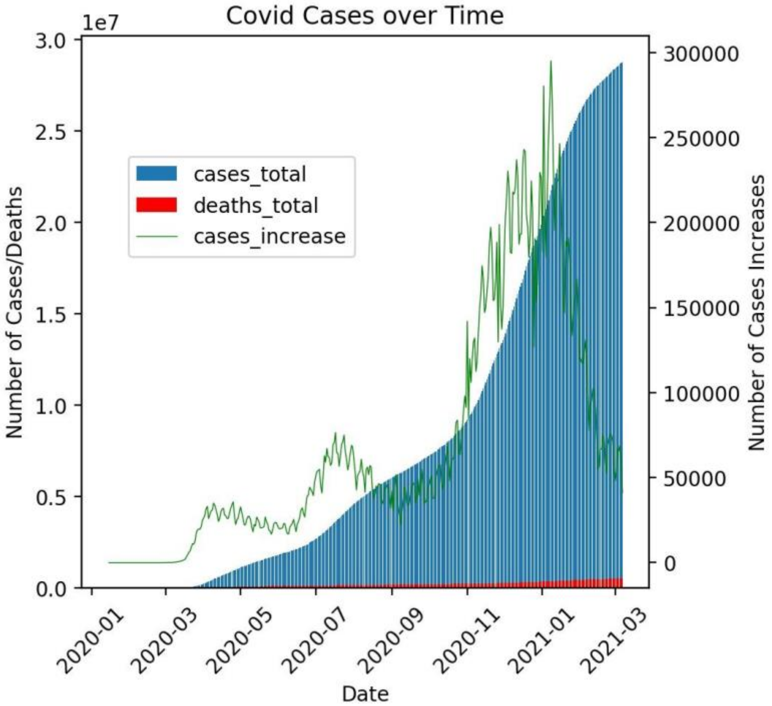


Figure 1: U.S. COVID-19 cases overtime

Given the striking increase of confirmed cases in the winter of 2020, there are reasons to believe that lower temperatures are associated with the intensification of transmission. After finding the states that COVID-19 most severely struck, we then dive into state-wise data for more microcosmic speculation. Table 1 and 2 show the top 5 states with the highest numbers of COVID-19 cases in total and cases per million population as of March 10, 2021. From the table results we generated, we choose to explore the top three of each: California, Texas, Florida, North Dakota, South Dakota, and Rhode Island.

Table 1: Top 5 states by number of total cases

#	USA State	Total Cases	Total Deaths	Total Recovered	Active Cases	TotCases Per1MPop	Deaths Per1Mpop
1	California	3,608,376	54,621	1,891,820	1,661,935	91,323	1,382
2	Texas	2,708,716	45,808	2,533,691	129,217	93,417	1,580
3	Florida	1,952,733	31,926	1,223,932	696,875	90,919	1,486
4	New York	1,745,965	48,726	839,031	858,208	89,750	2,505
5	Illinois	1,201,027	23,039	1,112,007	65,981	94,779	1,818

Table 2: Top 5 states by number of cases per population (million)

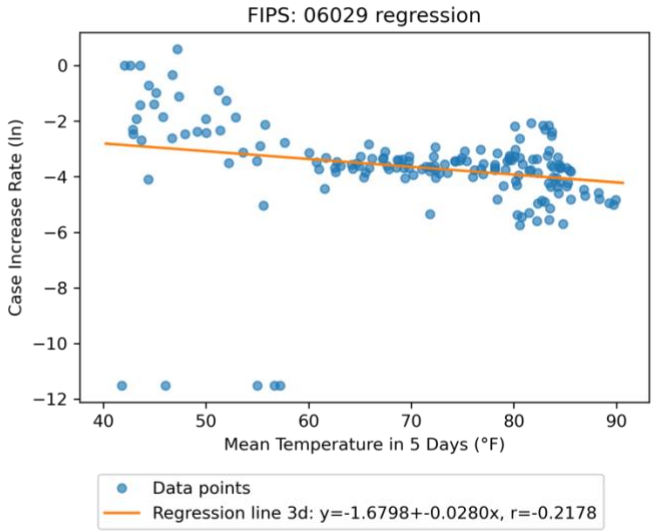
#	USA State	Total Cases	Total Deaths	Total Recovered	Active Cases	TotCases Per1MPop	Deaths Per1MPop
1	North Dakota	100,514	1,449	98,489	576	131,897	1,901
2	South Dakota	113,753	1,901	109,755	2,097	128,584	2,149
3	Rhode Island	129,277	2,556	7,936	118,785	122,033	2,413
4	Utah	375,669	1,990	359,997	13,682	117,178	621
5	Iowa	367,894	5,574	321,517	40,803	116,604	1,767

4. States

4.1 California

49 out of 58 counties in California have available data for both the weather and COVID-19 cases. The average Pearson correlations between the spreading rates and the average temperatures of preceding 3 days, 5 days and 10 days are -0.053295, -0.054861, -

0.044745, respectively, while the Kendal and Spearman correlations are all below -0.03. The Kern County (FIPS: 06029) shows the minimal value of and hence the strongest negative correlation between averaged temperatures of 5 days and the daily increase of cases with Pearson r=-0.42495. Its linear regression with R=-0.2178 is illustrated in Figure 2. On the other hand, Trinity County (FIPS: 06105) possesses the maximal correlation, indicating a potential positive relation. On closer examination of the data, we find that its low spreading rates, with zeros during most days, contributed to highly unrepresentative correlation coefficients and linear regression results. As shown in Figure 3, the regression line does not accurately reflect the relation of interest. Using K-means clustering on all 49 states, we choose 4 clusters by the Elbow Method. Figure 4 shows that although 4 clusters are generated and colored accordingly, most counties have either slightly positive or slightly negative correlations between averaged temperature with Pearson R values fluctuating around zero, which might be explained by two major waves of transmission in California, one in summer 2020 and the other in the winter (Figure 5).



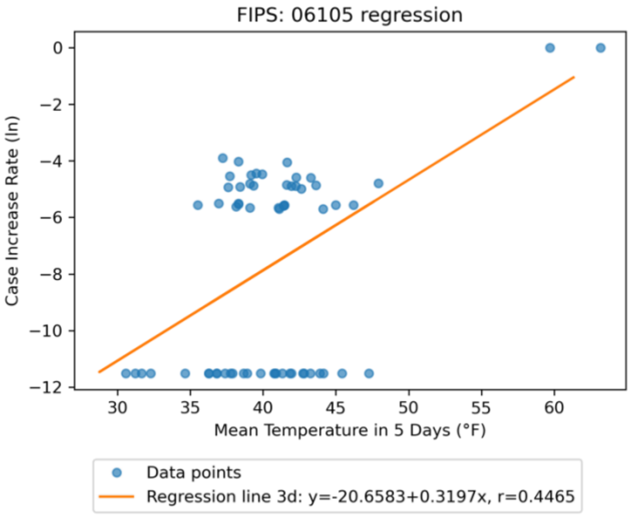
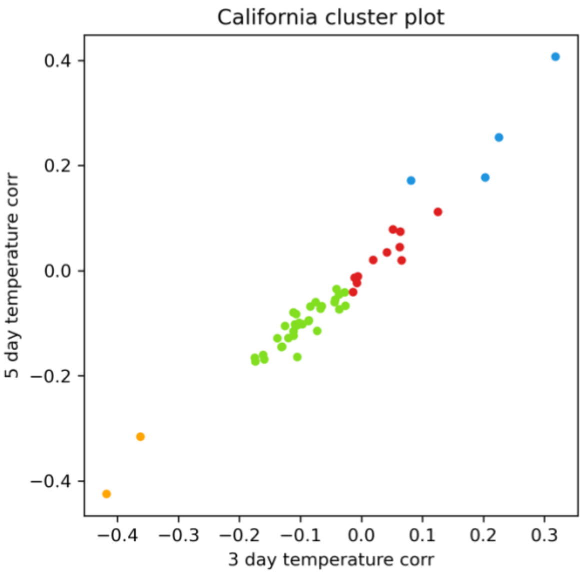


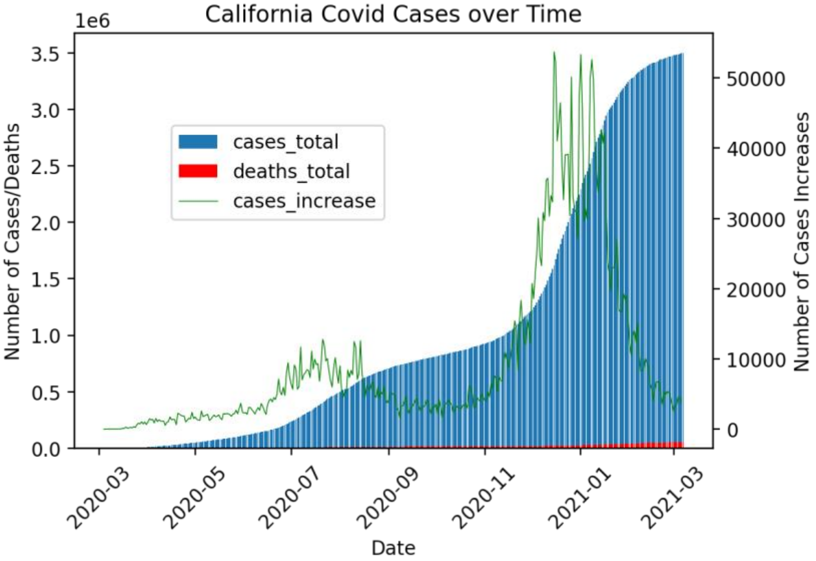
Figure 2 (left): linear regression of Kern, CA

Figure 3 (right): linear regression of Trinity, CA

Figure 4 (left): K-means clustering plot of California

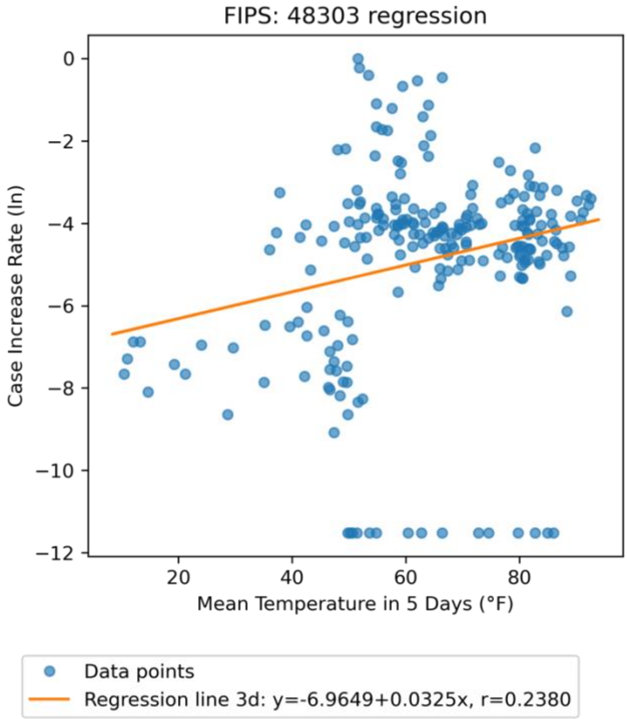
Figure 5 (right): California COVID-19 cases over time





4.2 Texas

We acquire weather and COVID-19 data for 60 out of 254 counties in Texas. The average Pearson correlations between the spreading rates and the average temperatures of preceding 3 days, 5 days and 10 days are 0.061755, 0.055751, 0.050872, respectively. Curiously, the average correlations are all positive, suggesting that higher temperatures facilitate the virus's spread. The Lubbock County (FIPS: 48303) has the smallest Pearson correlation of -0.04122 between averaged temperatures of 5 days and the daily increase of cases (Figure 6), while the Starr county (FIPS: 48427) is found with the maximal Pearson correlation value of 0.17629 (Figure 7). Neither of the two linear regressions lines fits the data well since the points are relatively spread out and might possess non-linear relationships. We choose to generate 4 clusters with K-means (Figure 8). Clusters are situated almost evenly apart along the linear line of 3- and 5-day temperature correlations. The clustering again corroborates that most counties show a positive correlation between the temperature and the COVID-19 spread, which might be related to the surge in COVID-19 confirmed cases in late October 2020, as demonstrated in Figure 9. Furthermore, while a relapse of transmission in the winter from 2020 to 2021 is observable, the spreading rate is lower than that in the summer due to the larger number of previously confirmed cases. Therefore, for Texas, higher temperatures become associated with the dissemination of COVID-19.



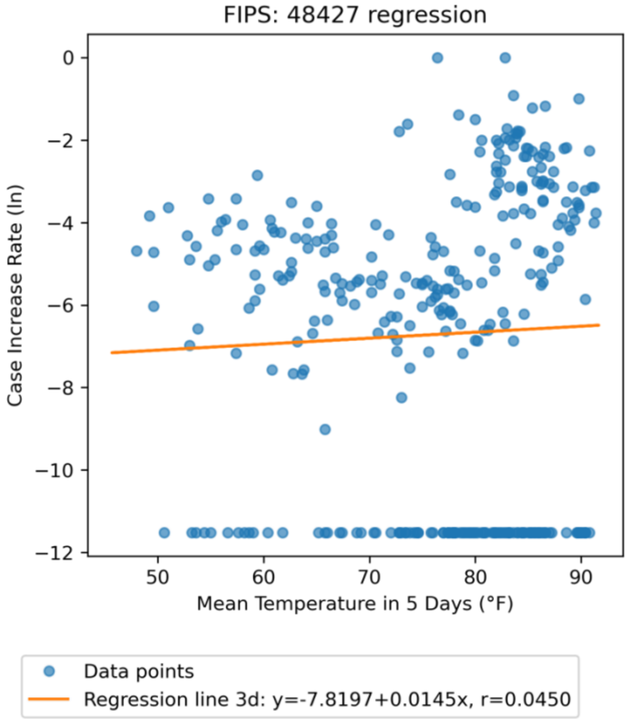
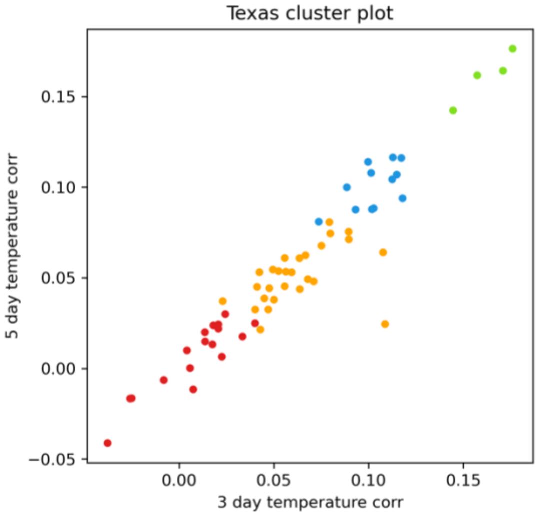


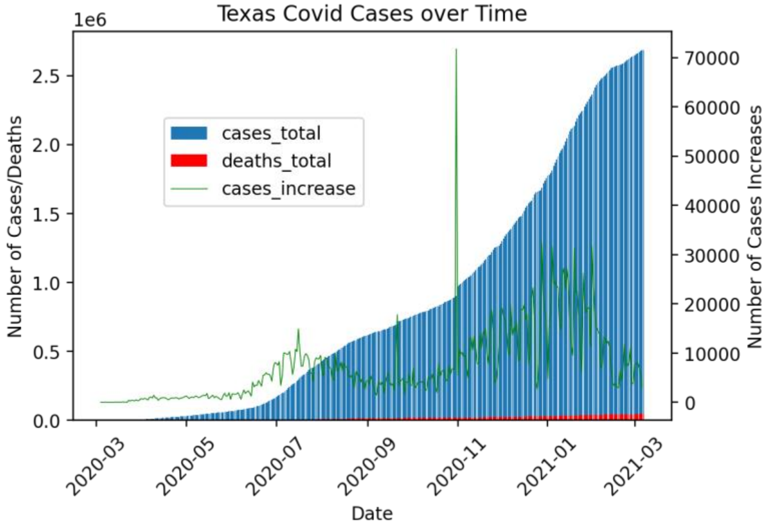
Figure 6 (upper left): linear regression of Lubbock, TX

Figure 7 (upper right): linear regression of Starr, TX

Figure 8 (lower left): cluster plot of Texas

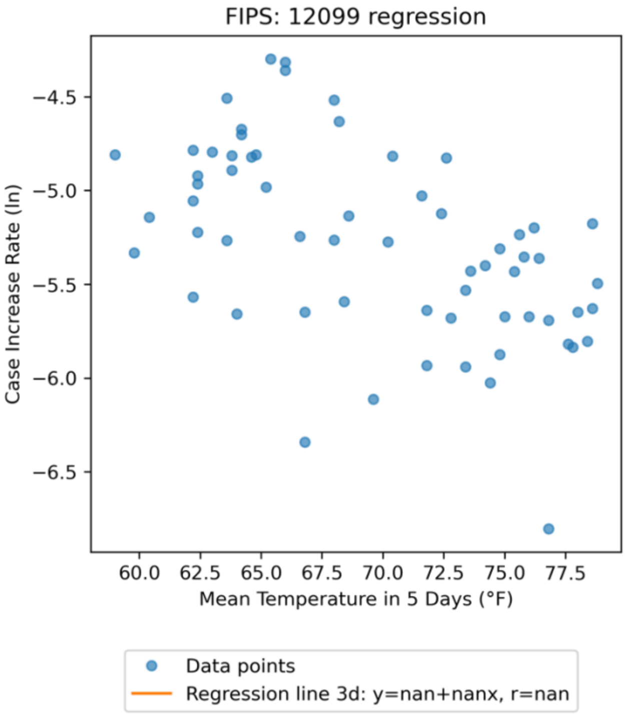
Figure 9 (lower right): Texas COVID-19 cases over time.





4.3 Florida

20 out of 67 counties in Florida provide data regarding both the weather and COVID-19 cases. The average Pearson correlations between the spreading rates and the average temperatures of preceding 3 days, 5 days and 10 days are -0.027277, -0.027873, -0.066876, respectively. The minimal and maximal correlations between averaged temperatures of 5 days and the daily increase of cases belong to the Palm Beach county (FIPS: 12099) with a value of -0.528329 (Figure 10) and the Baker county (FIPS: 12003) with a value of 0.170129 (Figure 11). Unfortunately, Palm Beach's data are so scattered that it is impossible to fit a linear regression line. Besides, the regression of Baker is ill-fitted as well. Figure 12 points out that most counties in Florida have very mildly positive correlations between temperatures and the COVID-19 spread, which might be explained by its similar trend of COVID-19 development to that of Texas (Figure 12). The two states are similar in weather as well.



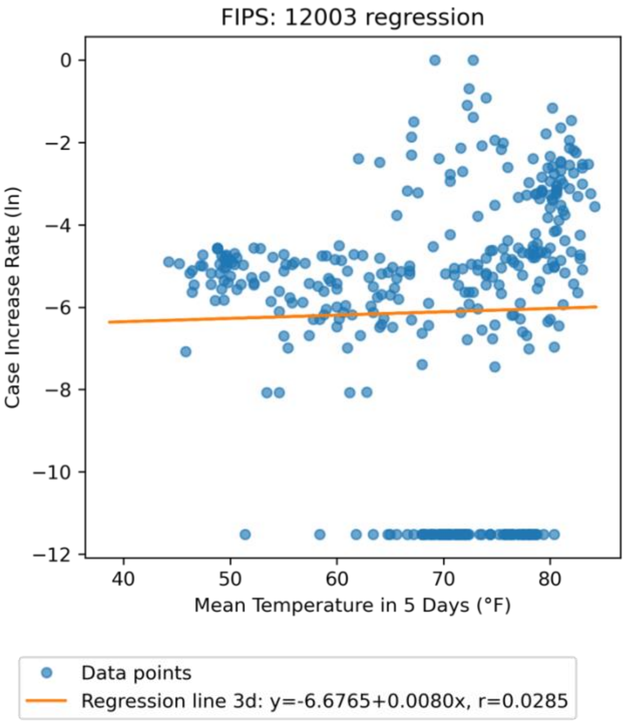
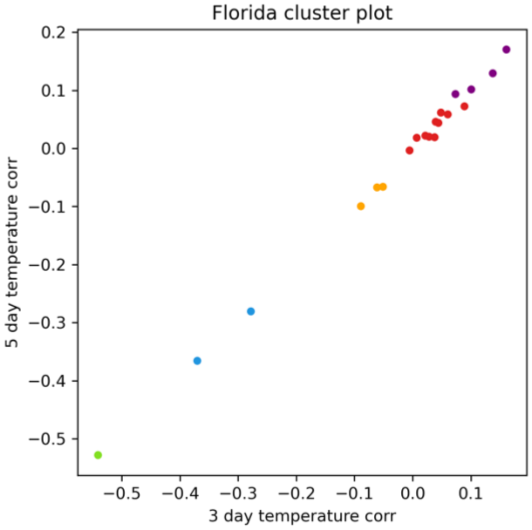


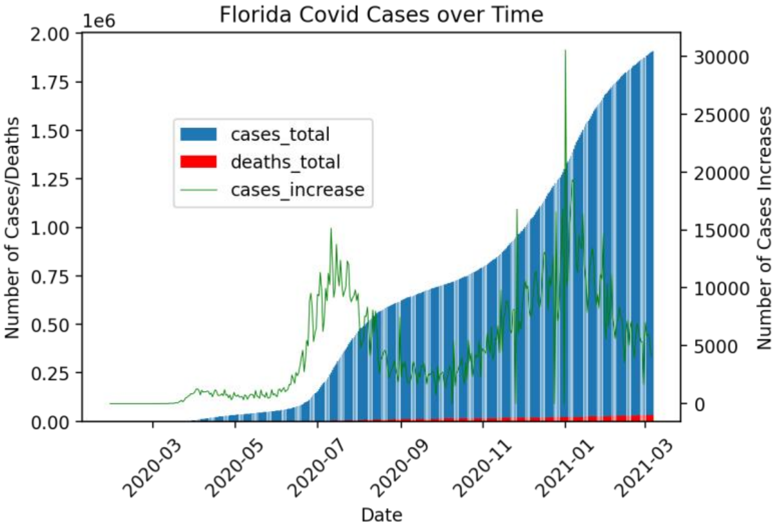
Figure 10 (upper left): linear regression of Palm Beach, FL

Figure 11 (upper right): linear regression of Baker, FL

Figure 12 (lower left): cluster plot of Texas

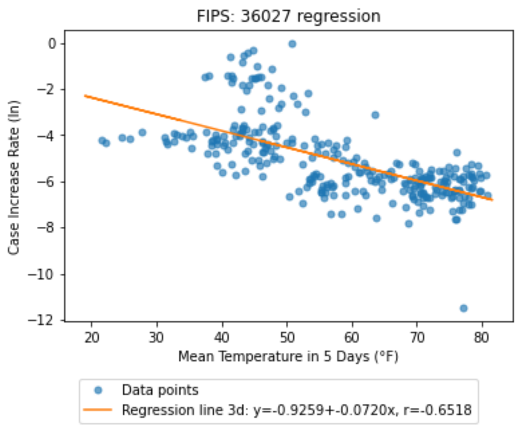
Figure 13 (lower right): Florida COVID-19 cases over time





4.4 New York

12 out of 62 counties in New York have accessible data for both weather and COVID-19 cases data. Averaged Pearson correlation coefficients between the spreading rates and the mean temperatures of the preceding 3 days, 5 days and 10 days are -0.107343, -0.109565, and -0.114671, respectively. Dutchess, New York (FIPS: 36027) shows the strongest correlation with a minimal value of Pearson coefficient R=-0.2150 between COVID-19 spreading rate and the averaged 5-day temperature (illustrated in Figure 14). For the same correlation calculation, Eerie county (FIPS: 36029) has the most significant correlation coefficient R=-0.0643, which only indicates a small negative correlation, as seen in Figure 15. With the K-means clustering and the elbow method, 4 clusters of counties are graphed in Figure 16. The majority of counties are clustered in the upper left corner, with mild negative correlations. Only a few counties demonstrate stronger negative correlations between the spreading rate and temperature. The larger cluster in red and other clusters with larger correlation coefficients could contribute to the spark of increased cases around January 2021 when the weather is generally colder in the north (in Figure 17). The rise around April 2020 might be more closely related to the disease's initial outbreak and many other complicated factors related (like people were unaware of it).



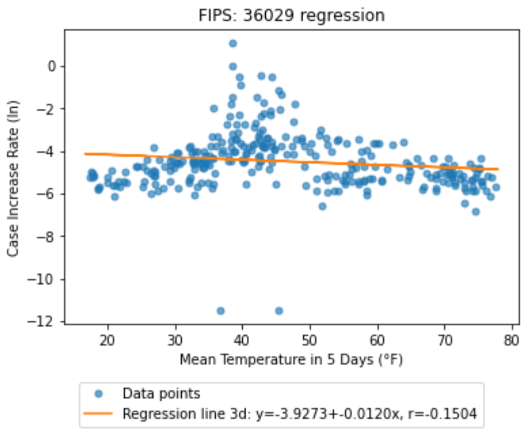
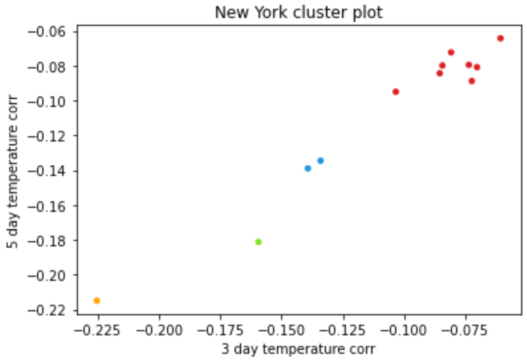


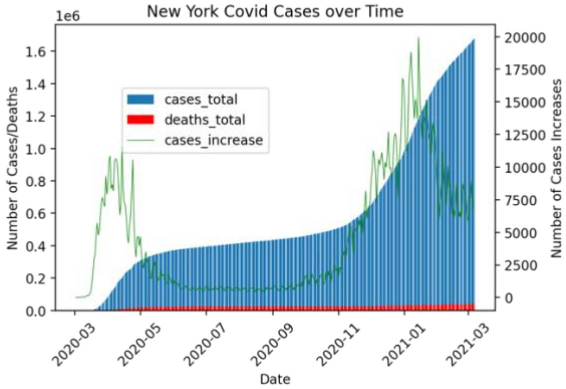
Figure 14 (upper left): linear regression of Dutchess, NY

Figure 15 (upper right): linear regression of Eerie, NY

Figure 16 (lower left): cluster plot of New York

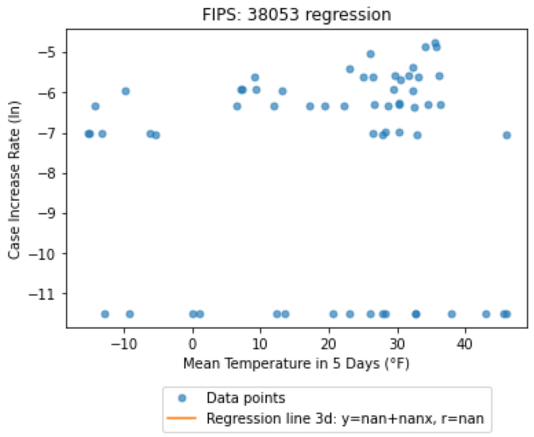
Figure 17 (lower right): New York COVID-19 cases over time





4.5 North Dakota

Temperature and COVID-19 cases data from 14 out of 53 counties in North Dakota are available in our datasets. The average Pearson correlation coefficients between the spreading rate and averaged temperature ranges of 3 days, 5 days, and 10 days are 0.077957, 0.073187, and 073418. North Dakota's data present slightly positive correlations like Texas's; however, the correlations' strengths are relatively weak. Interestingly, McKenzie county (FIPS: 38053), which has the largest correlation coefficients for the coefficient calculation between the spreading rate and 5-day-averaged-temperature R=0.2500, shows a positive correlation between the spreading rate and the averaged 5-day temperature as seen in Figure 18. However, no regression line fits the data, and this result could be ascribed to two reasons 1) the data is too scattered, and 2) zeros are present throughout the c ounty's COVID-19 spreading rate data. Cass (FIPS: 38017), the county with smallest correlation coefficient R=-0.0607, is also ill-fitted. The negative correlation does not match with a positive regression line in Figure 19. The scatter plot might present nonlinearity that our linear model fails to capture. We can tell that from the clustering result (K-means cluster n=4) in Figure 20, only 4 out of 14 counties present close-to-zero and slightly negative correlation results (clustered in blue); other counties still have a higher spreading rate when the temperature is higher. The clustering graph corresponds with the positive averaged correlation coefficients. The sharp rise in daily increased cases around November 2021 from Figure 21 might be able to account for this positivity in correlation, as the slope of change in increased cases in previous months before November is rather steep.



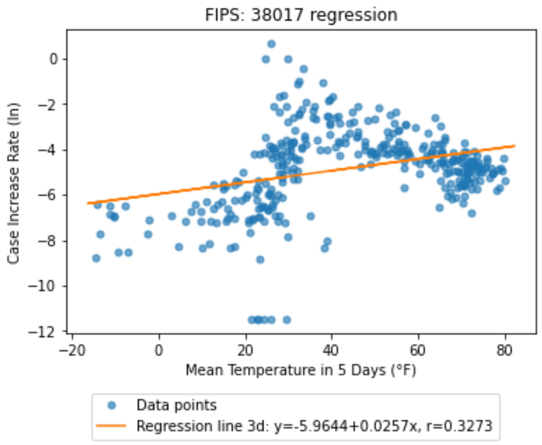
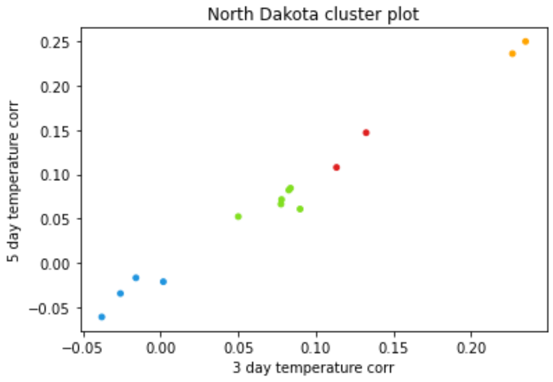


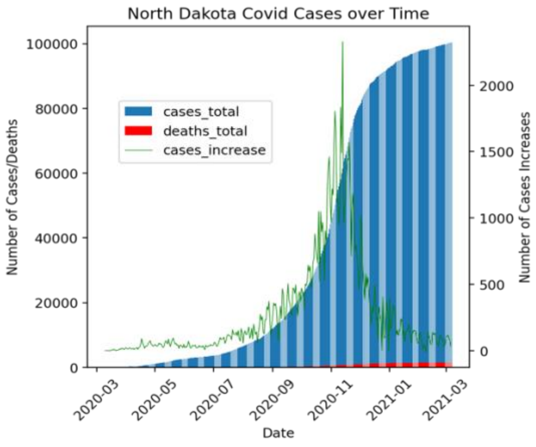
Figure 18 (upper left): linear regression of McKenzie, ND

Figure 19 (upper right): linear regression of Cass, ND

Figure 20 (lower left): cluster plot of North Dakota

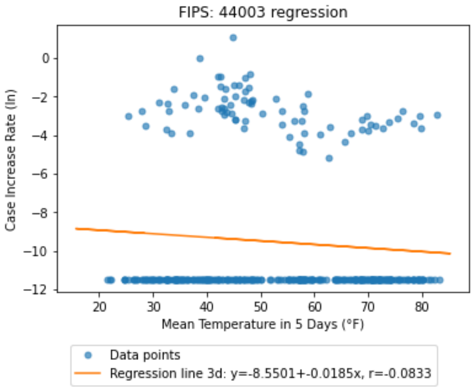
Figure 21 (lower right): North Dakota COVID-19 cases over time





4.6 Rhode Island

We obtain 1 out of 5 counties in Rhode Island that have both temperature and COVID-19 cases data. Since only one county's data is available, Kent, RI (FIPS: 44003), the state's average Pearson correlations, or the Kent county's correlation results between the COVID-19 spreading rate and the average temperatures of preceding 3 days, 5 days, and 10 days are -0.065737, -0.073165, and -0.078701 respectively. The correlations are slightly negative, and a corresponding negative regression line is plotted in Figure 22. However, we encounter many zero data points in the spreading rate dataset, which might have affected the regression line fitting. The K-means clustering is omitted here due to the limitation of our data. Although we only calculate correlations for one county, we could conjecture that if the other four counties have similar correlations between the spreading rate and the temperature, the rise of increased cases in Rhode Island around November 2020 to January 2021 in Figure 23 could be partially explained by this relationship.



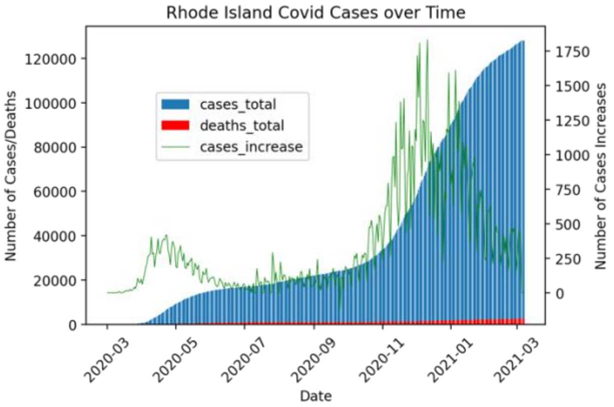


Figure 22 (left): linear regression of Kent, RI

Figure 23 (right): Rhode Island COVID-19 cases over time

5. Discussion and Conclusion

Variations are seen throughout the correlation calculation, including the strength of correlation and regressions. To explain this, we need to look at two relevant factors: specificity of the county and averaged time range. First, each county has its own demographic and geographic conditions. The spread of COVID-19 is also affected by the local government's prevention policy and residents' awareness on how to deal with the disease. Other factors not mentioned here can play a role as well. We can account for a comparable number of factors for the spreading rates, and the temperature is only one of them. However, we do see a general correlation between temperature and the COVID-19 spread on a state level, and the regressions can lend this further relationship credibility. Second, factors unknown to our particular study can also account for the difference in correlation strength regarding various periods. Such difference indicates that a temperature change needs time to influence the disease's transmission effectively. The period for this impact to take place fluctuates according to state and county. Such impact is mainly more evident after ten days than the first three and five days since most counties have more robust correlation coefficients for the ten-day temperature range and sometimes highly similar coefficients for the first three- and five-day correlation test.

Our modeling and analysis demonstrate the correlation between COVID-19 spreading rates and averaged temperature in the six states tested. The relationship is evident on a state level after averaging each county's correlation coefficient. The generalization is that such correlation could apply to all states in the United States. Lower temperature, to a certain degree, can help COVID-19 to spread faster. Higher temperature may halt the COVID-19's transmission to a certain degree. Along with this perspective, our general discovery is in line with the previous literature. Meanwhile, the spread of the pandemic disease relies on much more conditions than temperature, so the temperature is not the sole or primary factor determining the infection speed or trend.

References

Coronavirus (Covid-19) Data in the United States. The New York Times (2021). Retrieved 10 Mar 2021, from https://github.com/nytimes/covid-19-data."

"National Centers for Environmental Information." *National Climatic Data Center*, ncdc.noaa.gov/. Accessed 11 Mar 2021.

Rouen, A., et al. "COVID-19: relationship between atmospheric temperature and daily new cases growth rate." *Epidemiology & Infection* 148 (2020).

The COVID Tracking Project, covidtracking.com/. Accessed 10 Mar 2021.

" United States." *Worldometer Coronavirus Updates*, Worldometer, worldometers.info/coronavirus/country/us/. Accessed 10 Mar 2021.

Xie, Jingui, and Yongjian Zhu. "Association between ambient temperature and COVID-19 infection in 122 cities from China." *Science of the Total Environment* 724 (2020): 138201.