Exploring the Relationship between Weather, COVID-19, and Tweets Sentiment

Yiyi Tong yt2239@nyu.edu
Jiraphon Yenphraphai jy3694@nyu.edu
Wenni Fan wt630@nyu.edu
Courant Institute of Mathematical Sciences
New York University

Abstract

This study examines the relationship between weather patterns and the spread of COVID-19 using MapReduce, Hive, and Tableau. Our results show that while a negative correlation between COVID-19 cases and factors such as temperature and air quality may exist, it is not significant due to the influence of other factors that affect social distancing behaviors. Furthermore, we found that the relationship between COVID-19 cases and sentiment expressed in tweets is not significant due to the presence of confounding variables. These findings highlight the need for further research to fully understand the complex relationship between weather and the spread of COVID-19.

1 Introduction

The effect of weather on the spread of the coronavirus is one of the most investigated research questions since the onset of the pandemic. Like other epidemic diseases, the trajectories in many countries show strong seasonal patterns with fewer cases during summer and more during winter(Ganslmeier et al. (2021)). The possible reason is that weather and infectious diseases are linked, with the potential for weather variability to favor the emergence of novel viruses and contribute to disease transmission, morbidity and mortality(McClymont & Hu (2021)). A global study of 166 countries (excluding China) reported a significant negative correlation between temperature and cases, where a 1 degree increase in temperature was associated with a 3.08% (95% CI: 1.53–4.63%) reduction in cases (Wu et al. (2020)). Other studies showed that the air quality is improved during the COVID-19 lockdown and high levels of outdoor and indoor air pollution can potentially aggravate the adverse health impacts that COVID-19 patients might experience and slow down their recovery(Adam et al. (2021)). Although lots of studies have provided empirical evidence for the relationship between COVID-19 contagion and weather, controversial conclusions can be drawn from various research. Since understanding global spatial and temporal patterns of COVID-19 transmission is vital in the control and prevention of future outbreaks, our study attempts to investigate deeper into the weather impact on COVID-19 incidence by quantitative analysis based on big data. It helps to provide possible suggestions for developing weather-based early warning system for COVID-19 transmission.

Sentiment analysis is a technique in natural language processing that is used to analyze text data and identify the emotional state of the writer. It is widely accepted that the sentiment expressed in tweets on social media platforms like Twitter is closely tied to the emotional state of the individual who wrote them. In the context of the COVID-19 pandemic, this means that people may be more likely to post critical or negative tweets when they are dissatisfied with the way the pandemic is being handled by their government or other authorities. With this in mind, our study aims to investigate the relationship between weather patterns, the spread of COVID-19, and sentiment expressed in tweets related to the pandemic. We use a combination of data analysis tools such as MapReduce, Hive, and Tableau to conduct our analysis and identify any potential correlations or trends.

2 Methodology

2.1 Data sources

2.1.1 Epidemiology

This dataset contains the information related to the COVID-19 infections for each location at each date and can be downloaded here. The table is in csv format and its schema is defined as follows:

Column	Description	Example
data	YYYY-MM-DD of each record	2020-11-24
key	location (country_state_county)	US_NY_36029
new_confirmed	new cases confirmed	1234
new_deceased	new death	12
new_recovered	new recoveries	23
new_tested	new COVID-19 tests	43
cumulative_confirmed	cumulative sum of confirmed cases	1234
cumulative_deceased	cumulative sum of death cases	1234
cumulative_recovered	cumulative sum of recovery cases	1234

Table 1: Table Schema

date	location_key	new_confirmed	new_deceased	new_recovered	new_tested	cumulative_confirmed	cumulative_deceased	cumulative_recovered	cumulative_tested
2020-01-01	AD	0	0			0	0		
2020-01-02	AD	0	0			0	0		
2020-01-03	AD	0	0			0	0		
2020-01-04	AD	0	0			0	0		
2020-01-05	AD	0	0			0	0		
2020-01-06	AD	0	0			0	0		
2020-01-07	AD	0	0			0	0		
2020-01-08	AD	0	0			0	0		
2020-01-09	AD	0	0			0	0		

Figure 1: Snippet of epidemiology dataset

This dataset is depicted in Fig. 1. This file is 520.9 MB in size and contains 12,525,826 records.

2.1.2 Google search trends

This dataset is available for download here. It's a compilation of Google searches for symptoms and health problems. The information includes hundreds of symptoms ranging from abdominal obesity to Hepatitis. The searches, like the epidemiology dataset, are mapped to each symptom and organized by date and region. The table schema (over 200 columns) is too large to fit in this report. Only the following columns are displayed:

Column	Description	Example
data	YYYY-MM-DD of each record	2020-11-24
key	location (country_state_county)	US_NY_36029
symptom_name	normalized search volumes of a symptoms	21.01

Table 2: Table Schema

This dataset is illustrated in Fig. 2. This file is 1.9 GB, and it contains 2,713,930 records. What does the score really means? It is the relative popularity of symptoms in searches within in that region for that day, compared to the maximum value.

2.1.3 Government response

This dataset is the summary of the government response to the COVID-19. It can be downloaded here. Only some part of table schema is shown as follows in Table 2.1.3:

date	location_key	search_trends_abdominal_obesity	search_trends_abdominal_pain	search_trends_acne	search_trends_actinic_keratosis	search_trends_acute_bronchitis
2020-01-01	AU	3.56	5.38	10.76	0.52	0.35
2020-01-02	AU	3.46	5.35	11.3	0.55	0.36
2020-01-03	AU	3.4	5.35	11.23	0.53	0.34
2020-01-04	AU	3.43	5.23	10.68	0.42	0.35
2020-01-05	AU	3.29	5.14	10.2	0.49	0.38
2020-01-06	AU	3.52	5.23	10.83	0.49	0.36
2020-01-07	AU	3.62	5.21	11.17	0.55	0.41
2020-01-08	AU	3.54	5.17	11.39	0.54	0.36
2020-01-09	AU	3.4	5.11	11.09	0.52	0.33

Figure 2: Snippet of google search trends dataset

Column	Description	Example
data	YYYY-MM-DD of each record	2020-11-24
key	location(country_state_county)	US_NY_36029
school_closing	Record closings of schools [0-3]	2
workplace_closing	Record closings of workplaces [0-3]	1
testing_policy	COVID testing policy [0-3]	2

Table 3: Table Schema

Let's take a look at the number that corresponds to each column. The closing of schools and universities is referred to as school closing. 0 denotes no measurements. 1 denotes suggested closings. 2 denotes the need for closure. 3 means the highest restriction. The rest of the numbers in the columns have similar meanings depending on the context.

date	location_key	school_closing	workplace_closing	cancel_public_events	restrictions_on_gatherings	public_transport_closing	stay_at_home_requirements
2020-01-01	AD	0	0	0	0	0	0
2020-01-02	AD	0	0	0	0	0	0
2020-01-03	AD	0	0	0	0	0	0
2020-01-04	AD	0	0	0	0	0	0
2020-01-05	AD	0	0	0	0	0	0
2020-01-06	AD	0	0	0	0	0	0
2020-01-07	AD	0	0	0	0	0	0
2020-01-08	AD	0	0	0	0	0	0
2020-01-09	AD	0	0	0	0	0	0

Figure 3: Snippet of government response dataset

This dataset is depicted in Fig. 3. This dataset has 303,970 records and is 17 MB in size.

2.1.4 Weather

This dataset is the daily weather data from 2019 to 2022 and contains the daily air quality index by county (AQI), daily temperature (TEMP), and daily barometric pressure (PRESS). It can be downloaded here. Some part of the schema is shown as follows in the table:

Column	Description	Example
Date	YYYY-MM-DD of each record	2020-11-24
StateName	location(full state name)	New York
AQI	How polluted the air is	40
Arithmetic Mean	Daily average value of temperature and barometric pressure	61.8125 and 757.9417

Table 4: Schema of the Weather Data

The AQI data ranges from 0 to 500 with the higher value indicating increasing air pollution including ozone, nitrogen dioxide, sulphur dioxide, and etc. The daily temperature and barometric pressure are averaged by the hourly data collected from the outdoor weather stations. The date and state name are

extracted from the three datasets and treated as the foreign keys. The total size of the data is 488 MB and a snippet of them is depicted in Fig. 4.

State Name	county Name	State Code	County Code	Date	AQI	Category	Defining Parameter	Defining Site	Number of Sites Reporting
New Mexico	Santa Fe	35	49	2022-01-01	0	Good	PM2.5	35-049-0021	1
Puerto Rico	Caguas	72	25	2022-01-01	0	Good	CO	72-025-0007	1
North Carolir	Northampton	37	131	2022-01-01	1	Good	NO2	37-131-0003	1
Puerto Rico	Bayamon	72	21	2022-01-01	2	Good	CO	72-021-0010	1
Virginia	Pittsylvania	51	143	2022-01-01	2	Good	NO2	51-143-0005	1
Illinois	Macoupin	17	117	2022-01-01	3	Good	NO2	17-117-0002	1
North Carolir	Wake	37	183	2022-01-01	3	Good	NO2	37-183-0021	2
Rhode Island	Kent	44	3	2022-01-01	3	Good	PM2.5	44-003-0002	1
Virginia	Norfolk City	51	710	2022-01-01	3	Good	NO2	51-710-0024	1
Texas	Karnes	48	255	2022-01-01	4	Good	NO2	48-255-1070	1
Texas	Wilson	48	493	2022-01-01	4	Good	NO2	48-493-1038	1
Arizona	Mohave	4	15	2022-01-01	5	Good	PM10	04-015-1003	1

State Code	County Code	Site Num	Parameter Code	POC	Latitude	Longitude	Datum	Parameter Name	Sample Duration	Pollutant Standard	Date Local	Units of Measure	Event Type	Observation Count	Observation Percent	Arithmetic Mean
1	53	1000	62101	1	31.0921	-87.5435	NAD 83.00	Outdoor Temperature	1 HOUR		2022-01-01	Degrees Fahrenheit	None	24	100.0	74.925
1	53	1000	62101	1	31.0921	-87.5435	NAD 83.00	Outdoor Temperature	1 HOUR		2022-01-02	Degrees Fahrenheit	None	24	100.0	61.8125
1	53	1000	62101	1	31.0921	-87.5435	NAD 83.00	Outdoor Temperature	1 HOUR		2022-01-03	Degrees Fahrenheit	None	24	100.0	34.425
1	53	1000	62101	1	31.0921	-87.5435	NAD 83.00	Outdoor Temperature	1 HOUR		2022-01-04	Degrees Fahrenheit	None	24	100.0	37.4125
1	53	1000	62101	1	31.0921	-87.5435	NAD 83.00	Outdoor Temperature	1 HOUR		2022-01-05	Degrees Fahrenheit	None	24	100.0	48.45
1	53	1000	62101	1	31.0921	-87.5435	NAD 83.00	Outdoor Temperature	1 HOUR		2022-01-06	Degrees Fahrenheit	None	24	100.0	56.220833
1	53	1000	62101	1	31.0921	-87.5435	NAD 83.00	Outdoor Temperature	1 HOUR		2022-01-07	Degrees Fahrenheit	None	24	100.0	42.058333
1	53	1000	62101	1	31.0921	-87.5435	NAD 83.00	Outdoor Temperature	1 HOUR		2022-01-08	Degrees Fahrenheit	None	24	100.0	50.1625
1	53	1000	62101	1	31.0921	-87.5435	NAD 83.00	Outdoor Temperature	1 HOUR		2022-01-09	Degrees Fahrenheit	None	24	100.0	64.966667
1	53	1000	62101	1	31.0921	-87.5435	NAD 83.00	Outdoor Temperature	1 HOUR		2022-01-10	Degrees Fahrenheit	None	24	100.0	47.8375
1	53	1000	62101	1	31.0921	-87.5435	NAD 83.00	Outdoor Temperature	1 HOUR		2022-01-11	Degrees Fahrenheit	None	24	100.0	41.854167
1	53	1000	62101	1	31.0921	-87.5435	NAD 83.00	Outdoor Temperature	1 HOUR		2022-01-12	Degrees Fahrenheit	None	24	100.0	43.966667
1	53	1000	62101	1	31.0921	-87.5435	NAD 83.00	Outdoor Temperature	1 HOUR		2022-01-13	Degrees Fahrenheit	None	24	100.0	47.275
1	53	1000	62101	1	31.0921	-87.5435	NAD 83.00	Outdoor Temperature	1 HOUR		2022-01-14	Degrees Fahrenheit	None	24	100.0	48.775
1	53	1000	62101	1	31.0921	-87.5435	NAD 83.00	Outdoor Temperature	1 HOUR		2022-01-15	Degrees Fahrenheit	None	24	100.0	52.079167
1	53	1000	62101	1	31.0921	-87.5435	NAD 83.00	Outdoor Temperature	1 HOUR		2022-01-16	Degrees Fahrenheit	None	24	100.0	37.7375
1	53	1000	62101	1	31.0921	-87.5435	NAD 83.00	Outdoor Temperature	1 HOUR		2022-01-17	Degrees Fahrenheit	None	24	100.0	38.866667

State Code	County Code	Site Num	Parameter Code	POC	Latitude	Longitude	Datum	Parameter Name	Sample Duration	Pollutant Standard	Date Local	Units of Measure	Event Type	Observation Count	Observation Percent	Arithmetic Mean
1	53	1000	64101	1	31.0921	-87.5435	NAD 83.00	Barometric pressure	1 HOUR		2022-01-01	Millibars	None	24	100.0	750.829167
1	53	1000	64101	1	31.0921	-87.5435	NAD 83.00	Barometric pressure	1 HOUR		2022-01-02	Millibars	None	24	100.0	750.991667
1	53	1000	64101	1	31.0921	-87.5435	NAD 83.00	Barometric pressure	1 HOUR		2022-01-03	Millibars	None	24	100.0	757.941667
1	53	1000	64101	1	31.0921	-87.5435	NAD 83.00	Barometric pressure	1 HOUR		2022-01-04	Millibars	None	24	100.0	756.995833
1	53	1000	64101	1	31.0921	-87.5435	NAD 83.00	Barometric pressure	1 HOUR		2022-01-05	Millibars	None	24	100.0	753.583333
1	53	1000	64101	1	31.0921	-87.5435	NAD 83.00	Barometric pressure	1 HOUR		2022-01-06	Millibars	None	24	100.0	752.179167
1	53	1000	64101	1	31.0921	-87.5435	NAD 83.00	Barometric pressure	1 HOUR		2022-01-07	Millibars	None	24	100.0	756.691667
1	53	1000	64101	1	31.0921	-87.5435	NAD 83.00	Barometric pressure	1 HOUR		2022-01-08	Millibars	None	24	100.0	758.008333
1	53	1000	64101	1	31.0921	-87.5435	NAD 83.00	Barometric pressure	1 HOUR		2022-01-09	Millibars	None	24	100.0	756.575
1	53	1000	64101	1	31.0921	-87.5435	NAD 83.00	Barometric pressure	1 HOUR		2022-01-10	Millibars	None	24	100.0	759.3
1	53	1000	64101	1	31.0921	-87.5435	NAD 83.00	Barometric pressure	1 HOUR		2022-01-11	Millibars	None	24	100.0	760.445833
1	53	1000	64101	1	31.0921	-87.5435	NAD 83.00	Barometric pressure	1 HOUR		2022-01-12	Millibars	None	24	100.0	757.504167
1	53	1000	64101	1	31.0921	-87.5435	NAD 83.00	Barometric pressure	1 HOUR		2022-01-13	Millibars	None	24	100.0	753.604167
1	53	1000	64101	1	31.0921	-87.5435	NAD 83.00	Barometric pressure	1 HOUR		2022-01-14	Millibars	None	24	100.0	752.783333
1	53	1000	64101	1	31.0921	-87.5435	NAD 83.00	Barometric pressure	1 HOUR		2022-01-15	Millibars	None	24	100.0	750.166667
1	53	1000	64101	1	31.0921	-87.5435	NAD 83.00	Barometric pressure	1 HOUR		2022-01-16	Millibars	None	24	100.0	749.15
1	53	1000	64101	1	31.0921	-87.5435	NAD 83.00	Barometric pressure	1 HOUR		2022-01-17	Millibars	None	24	100.0	753.695833

Figure 4: Snippet of weather dataset

2.1.5 Tweets

These datasets consist of the tweets IDs and sentiment scores of each tweet from October 01, 2019 to December 2022. We selected the time period from March 20, 2020 to October 02, 2022. Each csv file contains the tweets over a certain time period. The tweets were COVID-related and were selected by using over 90 different keywords and hashtags that are relevant to the pandemic. The datasets can be downloaded here. The schema of the datasets is defined as follows:

Column	Description	Example
first column	The id of the tweet	1240800606777990000
second column	The sentiment score of the tweet	0.2

Table 5: Schema of the Weather Data

The sentiment score is computed by TextBlob, which has a range of -1 to 1. We sampled about 50,000 tweets randomly from each day and put them into a new csv file for evaluation. The total data size is about 1.27 GB and a snippet of them is depicted in Fig. 5.

1240800606777990000	0
1240800606777990000	-0.4
1240800606777990000	0.3
1240800606777990000	-0.4
1240800606777990000	-0.1625
1240800606777990000	-0.1625
1240800606777990000	0.5
1240800606777990000	-0.005555556
1240800606777990000	0.366666667
1240800606777990000	0.5
1240800606777990000	-0.05
1240800606777990000	0
1240800606777990000	0.1
1240800606777990000	-0.352857143
1240800606777990000	-0.1625
1240800606777990000	0
1240800606777990000	0
1240800606777990000	-0.25

Figure 5: Snippet of tweet dataset

2.2 Design diagram

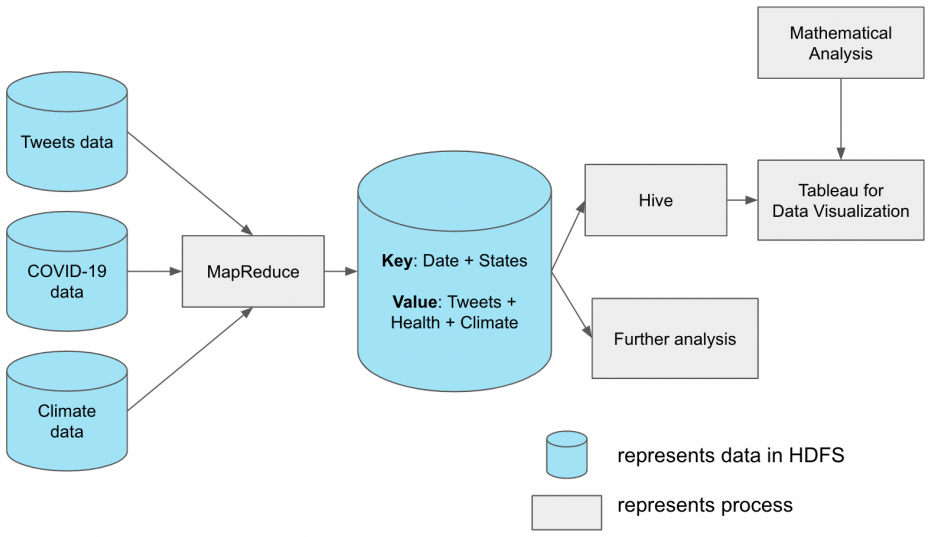

Figure 6: Pipeline

Fig. 6 demonstrates our general pipeline. The raw data is cleaned and profiled using MapReduce (Dean & Ghemawat (2008)). For the weather data, different mappers are implemented for AQI, temperature and pressure. And the join operations are performed at the reducer side with the time and location as foreign keys. The combiner is used to reduce the amount of data transferred through the network by averaging the data by states. And a custom partitioner is implemented to load balance the work so that each reducer is responsible for the data of each year.

For COVID-19 dataset, we only focus on cases in the United States, so our mapper filters out information from other countries. Our mapper reads the csv file line by line, just like a text file. Each column is separated by a comma. The mapper returns date and state as the foreign key and a tuple of all record attributes as the value. As reported in the data profiling, we use summarization patterns and top k pattern to obtain the data of interest.

For the tweets datasets, we focus on the date and the average sentiment score of tweets from that date. Therefore, for the mapper function, we extracted the date (month, day, year) information from the tweet id by doing some arithmetic computation and wrote the date as the key of the output. Then we wrote the sentiment score of each tweet as the value of the output. For the reducer function, we computed the average of the sentiment score for each day. We wrote the date as the output's key and the average sentiment score of tweets on that day as the output's value.

Date and states are the common output keys for our MapReduce jobs, and the values are tweets data, public health data, and weather data. They are then stored in the Hive table for queries and joins, with Tableau for interactive visualizations. Statistical analysis and mathematical models are applied to supplement our visualizations and insights.

2.3 Code challenge

We cannot split the string on each line of the COVID-19 data using the java split() method since it will discard the missing data and result in a dynamic array size, making it impossible to keep track of the data. We must create our own string splitting method. Additionally, we are using a k-sized minHeap to store the top-k Google search trends, where we discard the element at the top of the tree if it is less than the current value, and push it onto the tree if it is not. Both the mapper and the reducer use the same algorithm to process the data.

For the weather data, the main coding challenge lies in how to improve the efficiency of the MapReduce job. As the original data are collected by county while we need to average them by states, the count of values within each state is recorded to take advantage of the combiner. It helps to average the data firstly before transferring them to the reducer. To further load balance the work on each reducer, the outputs of the combiner are partitioned by the year as the data size is roughly the same each year except for the 2022.

For the tweets data, the main issue was to put a date column in the dataset. Since the dates of the tweets will be used as the keys for MapReduce and foreign keys for joining with other data sources, a date column was necessary. As the original datasets do not contain the dates of the tweets, an algorithm was used to extract the year, month, and day information from the tweets ID.

3 Results

3.1 Data visualization

3.1.1 COVID-19 data versus tweets data

We visualize our data using Tableau as stated in lecture 8. Fig. 7 depicts the emotive scores of tweets versus daily covid instances. As the points are distributed vertically, it is evident that there is no correlation between these two pieces of information. This lack of correlation is further supported by population correlation coefficient $\rho_{x,y}$ where x, y are random variables, ρ_x, ρ_y are their standard deviations. In this scenario, x represents the daily number of COVID-19 cases, while y represents the scores of tweets.

$$\rho_{x,y} = \frac{\mathbb{E}[(x - \mathbb{E}[x])(y - \mathbb{E}[y])]}{\sigma_x \sigma_y} \approx \frac{1}{N} \frac{\sum_{i=1}^{N} (x_i - \frac{1}{N} \sum_{j=1}^{N} x_j)(y_i - \frac{1}{N} \sum_{j=1}^{N} y_j)}{\sigma_x \sigma_y} \tag{1}$$

Following this formulation, $\rho_{x,y} \approx 0$. This adds to the evidence that there is no correlation at all.

We also analyze different government responses versus tweets' scores. This could reveal how people react to government answers. Fig. 8 is the visualization of government responses (facial mask covering policy, income support policy, testing policy, and work closing policy) and tweets' scores. The indicators of government responses are discrete between 0 and 3, with 0 representing no measure and 3 being the most stringent level of restriction. If there is a high connection, the graph should be seen as a ladder where people may be dissatisfied (negative scores) when there is no income support. The correlation score for mask policy, income support, testing policy, work closing versus sentimental scores are $0.1897, -0.0360, 0.2133, -0.0549$. Theses scores are quite low. It is unlikely to see the correlation.

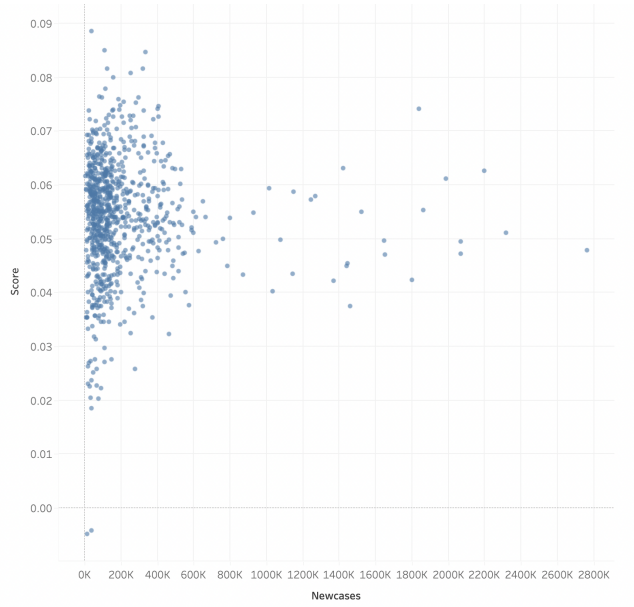

Figure 7: The visualization of daily covid cases vs sentimental scores

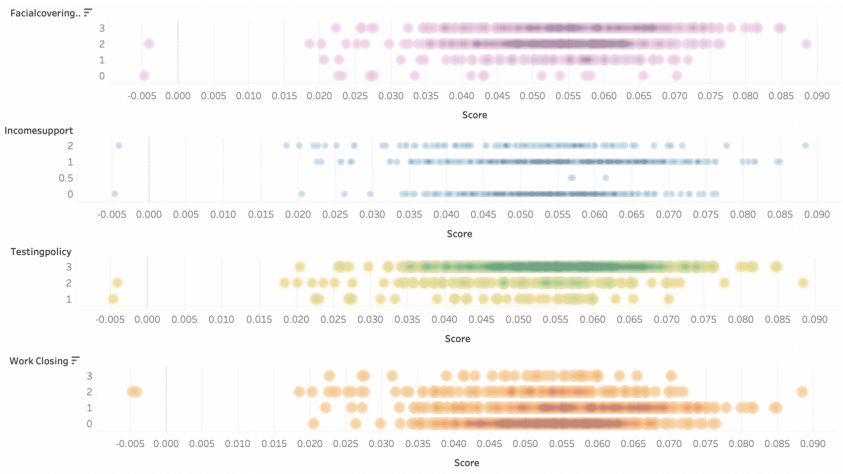

Figure 8: The visualization of government responses vs sentimental scores

We could also see how effective the government's pandemic response policy is. Figure 9 depicts the display of these data. Again, we don't see the correlation.

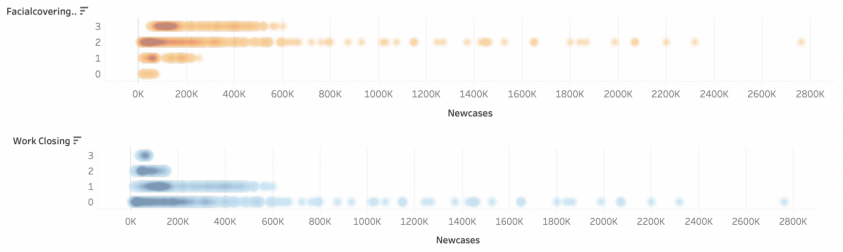

Figure 9: The visualization of government responses vs daily cases

3.1.2 COVID-19 data versus weather data

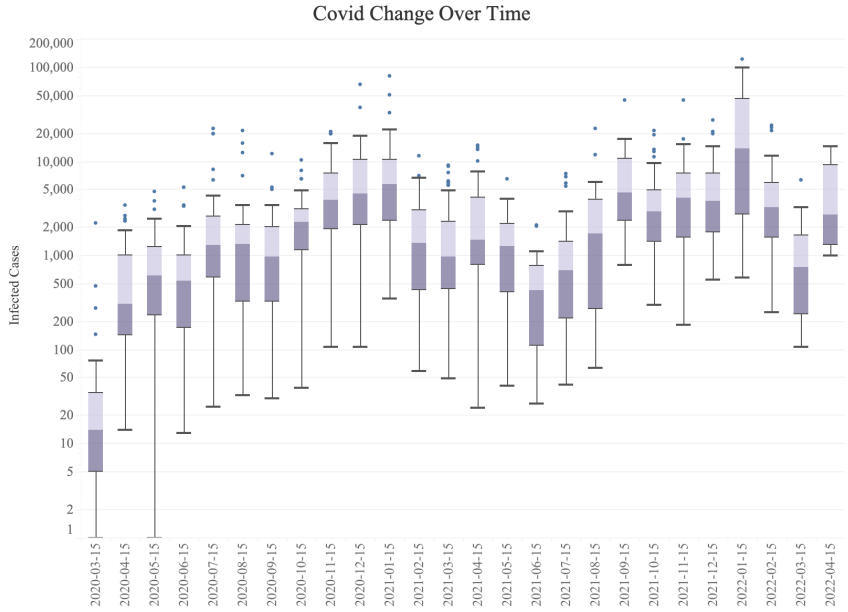

Figure 10: The COVID-19 new cases change over time

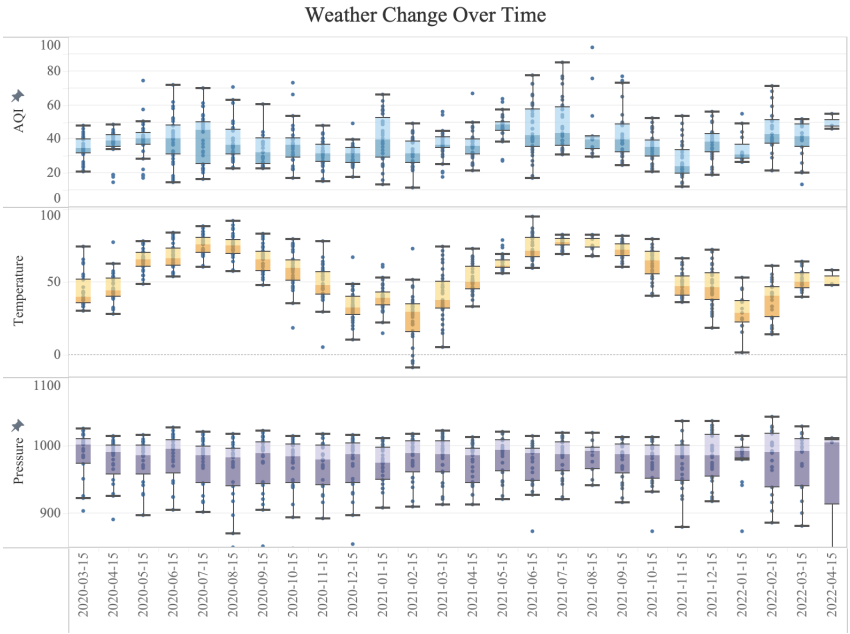

Figure 11: The weather change over time

As shown in Fig. 10, the boxes indicate the middle 50 percent of the data which is distributed over the states. Whiskers display all points within 1.5 times the interquartile range. Over the time, the new cases firstly increase steadily from 2020 January to 2021 January, and then decrease from 2021 February to 2021 June, and it fluctuates up again. As depicted in Fig. 11, the AQI and temperature show similar pattern with better air quality in winter and worse air quality in summer. As the new cases, temperature and AQI all demonstrate seasonal changes, the correlations between them are further explored. In contrast, as the average and spatial variation of air pressure remain similar

throughout the year, we assume that the barometric pressure has no significant impact on COVID-19 transmission. The possible reason is that though the pressure determines the formation of wind and the air circulation may have impact on the virus spread, the direct relationship between them cannot be reflected here.

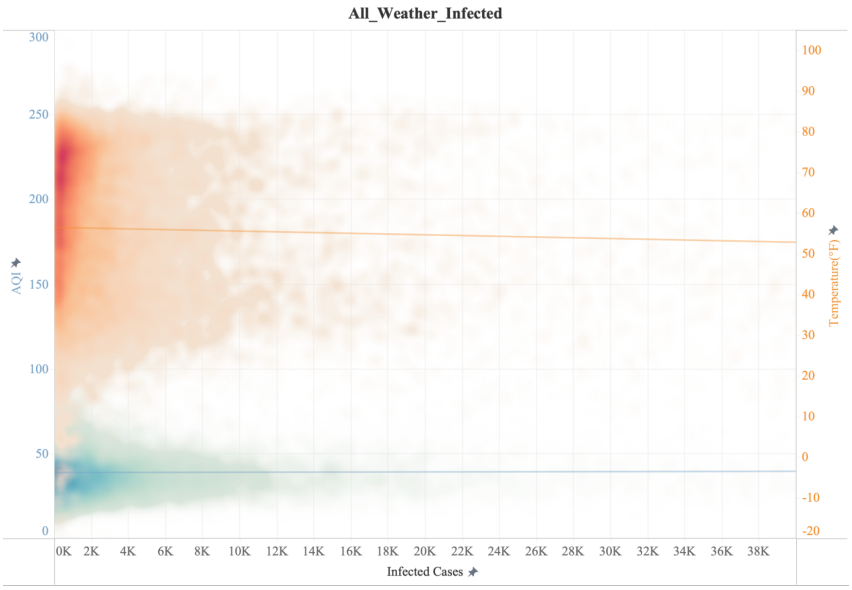

Figure 12: Infected cases versus weather data over all states

The correlations between infected cases and AQI, temperature are shown in 12. Although the regression lines indicate that the number of new cases decreases slowly with the increase of the temperature and AQI, the relationship is not significant due to a large number of points scattered away from the lines. To control for unobserved heterogeneity across states—such as cultural factors and regional-time-varying factors affecting the evolution of the pandemic such as the imposition of lockdown measures, mask requirements and other factors affecting social distancing, the relationship is further explored at regional granularity.

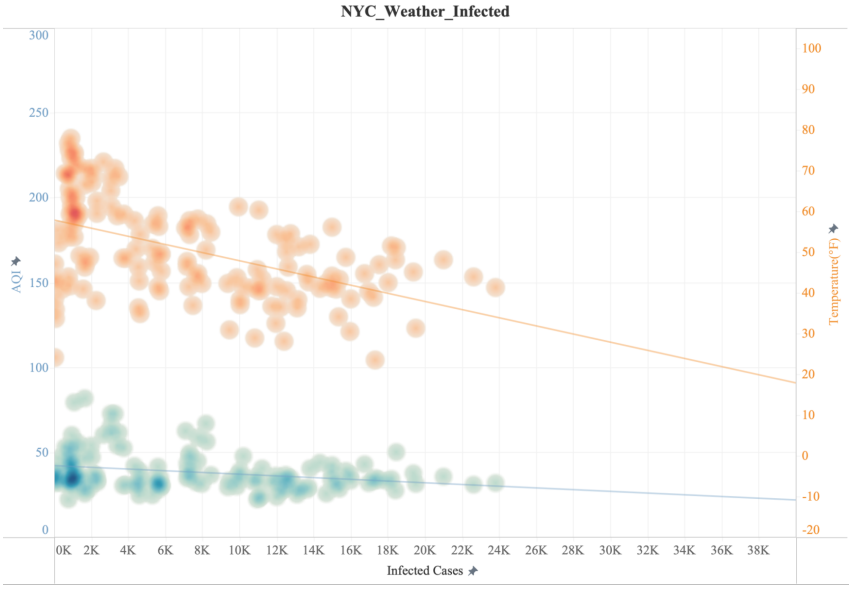

Figure 13: Infected cases versus weather data in New York City

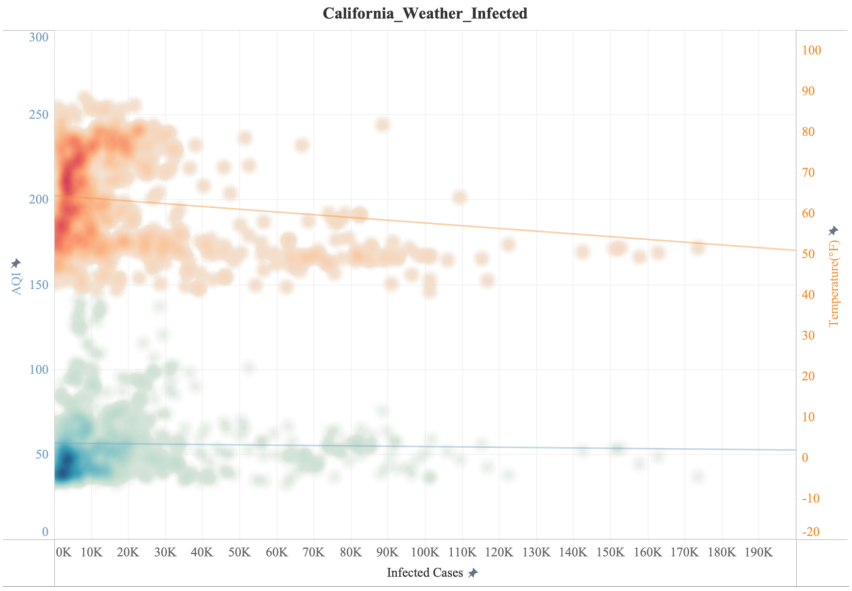

Figure 14: Infected cases versus weather data in California

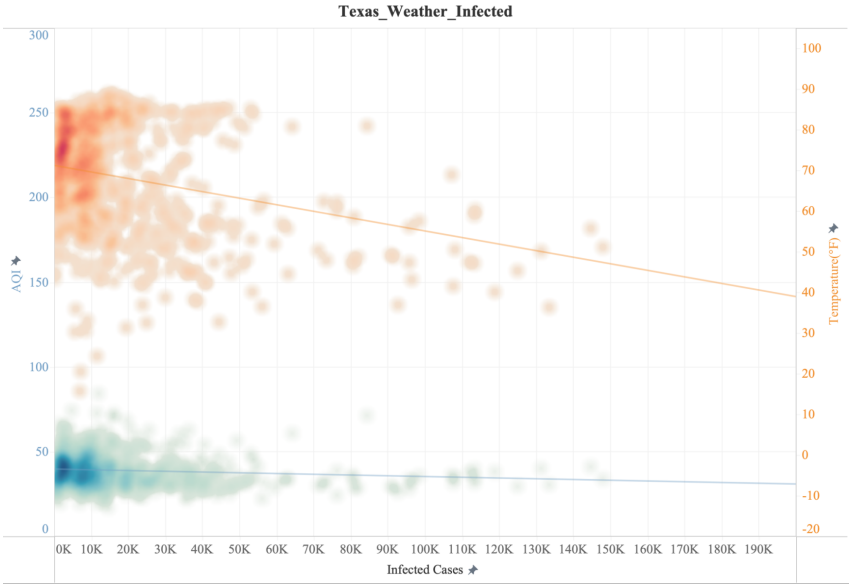

Figure 15: Infected cases versus weather data in Texas

The three states including New York City, California and Texas are selected to represent the weather variation in US. As shown from 13 to 15, the negative correlations between infected cases and temperature, AQI are more significant than those at country granularity especially in NYC and Texas. The Table 5 and Table 6 further quantify that the negative correlations are more significant within a state with higher R-square value and smaller slope value. And the p-values less than 0.05 prove the statistical significance.

The spatial distribution of infected cases and weather data is demonstrated in Figures 16 to 19. At different time, infected cases in the southern regions are greater than those in the northern regions despite of temperature and AQI. The possible reason is that the regional government response and local populations are varied and thus other factors such as social distancing and infection controls also have impact on virus spread. It again indicates that weather impact on COVID-19 should be

Measurements	All	NYC	California	Texas
Slope	-9.974e-05	-2.287e-04	-6.443e-05	-1.613e-04
R-square	0.0053	0.2602	0.0771	0.0997
P-value	< 0.0001	0.1395	< 0.0001	< 0.0001

Table 6: Relationship between Infected Cases and Temperature

Measurements	All	NYC	California	Texas
Slope	-2.583e-05	-3.343e-04	-1.586e-05	-4.405e-05
R-square	0.0003	0.0756	0.0019	0.0153
P-value	< 0.0001	0.0061	0.3392	0.0006

Table 7: Relationship between Infected Cases and AQI

analyzed at high spatial resolution. Comparing the results in winter and summer, the seasonal patterns are consistent with the above correlation analysis.

Covid and Temperature Map on 2021-01-25

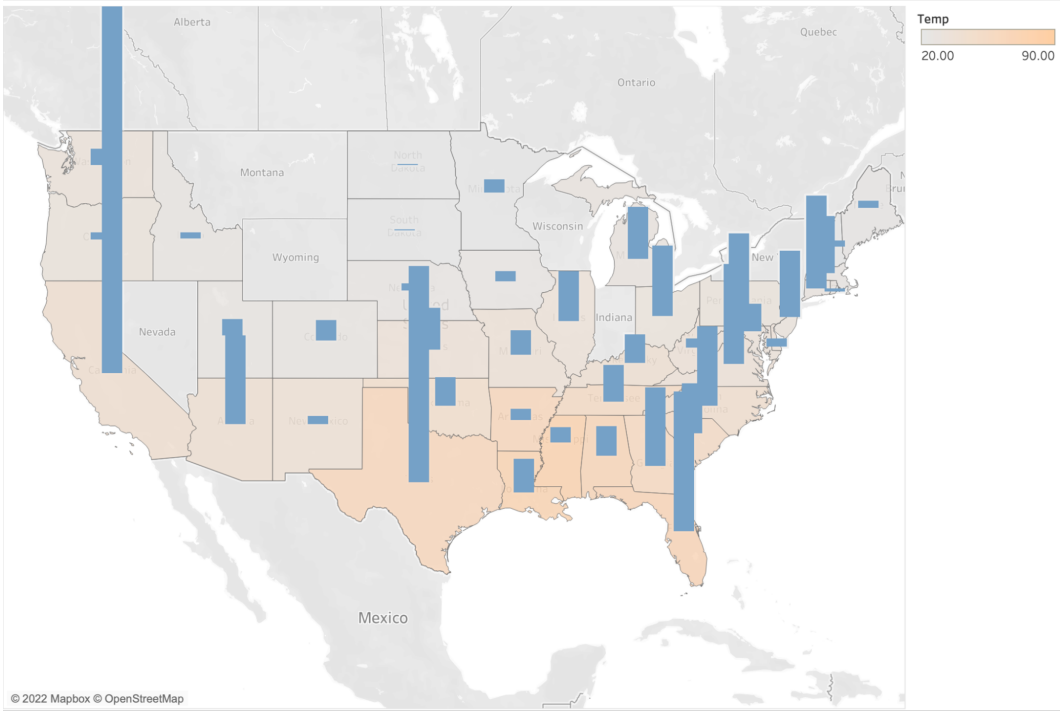

Figure 16: Spatial distribution of COVID and temperature on 2021-01-25

In general, the air pressure impact on COVID-19 transmission is not significant, while both the outdoor temperature and AQI have negative impact on virus spread. In our study, the number of infected cases is reduced with the increase of temperature and AQI. From an epidemiological standpoint, the survival and spread of a virus depends on the proper temperature of its environment. The cold dry air inhibits the innate immune response through damage to mucous membranes and slowing of mucociliary clearing (Lowen et al. (2007)). From a behavioral perspective, weather alters mobility levels, social distancing, and location of social gatherings, which in turn affects the spread of the virus across individuals (Kraemer et al. (2020)). In winter, people's outdoor activities are reduced and thus the probability of virus transmission is largely increased. In response to the reduced outdoor activities and lockdown, the emissions of primary air pollutants from major sources such as vehicular traffic and industries are at a lower level. At the same time, there are studies suggesting that poor outdoor air quality has the potential to increase the lethality of COVID-19 (Contini D. (2020)).

Covid and Temperature Map on 2021-07-26

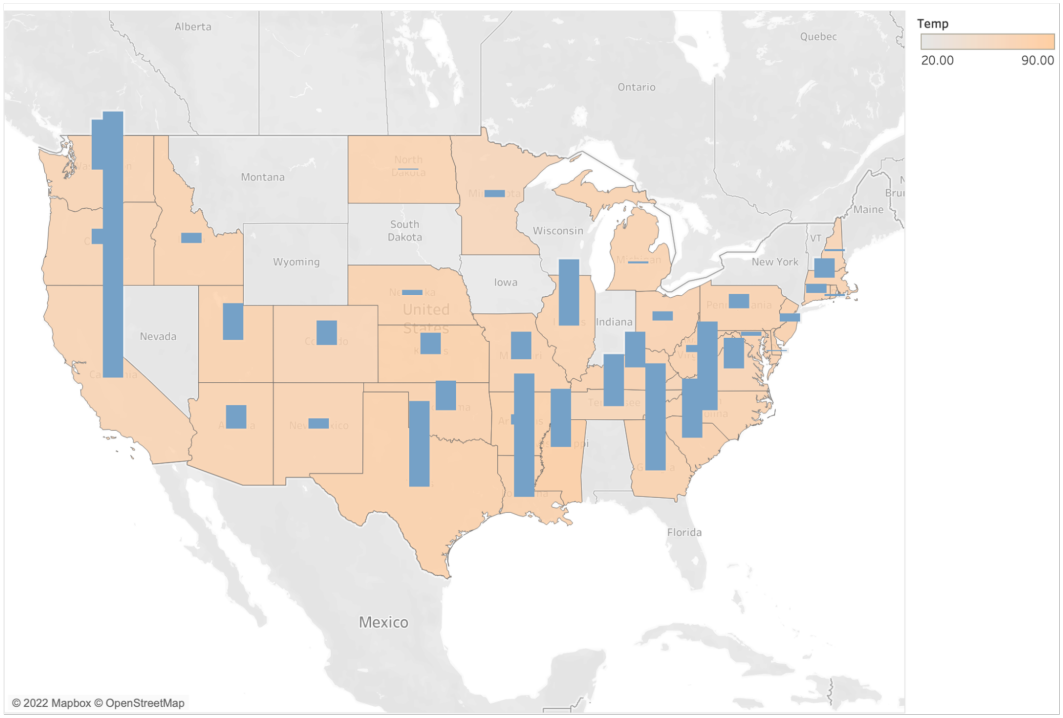

Figure 17: Spatial distribution of COVID and temperature on 2021-07-26

Covid and AQI Map on 2021-01-25

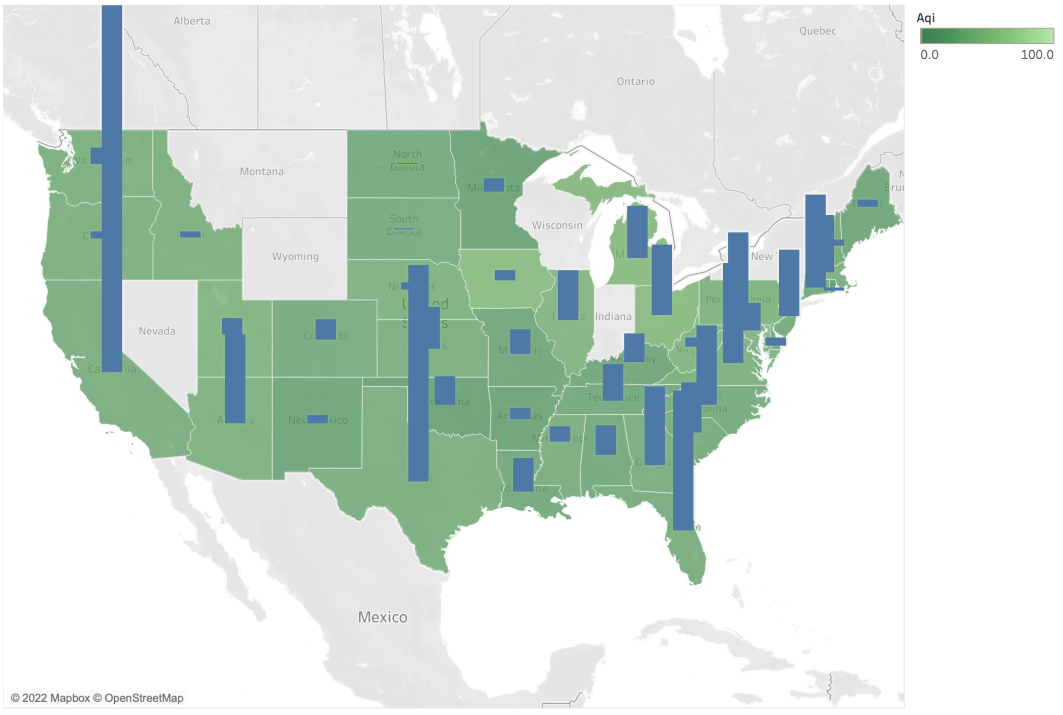

Figure 18: Spatial distribution of COVID and AQI on 2021-01-25

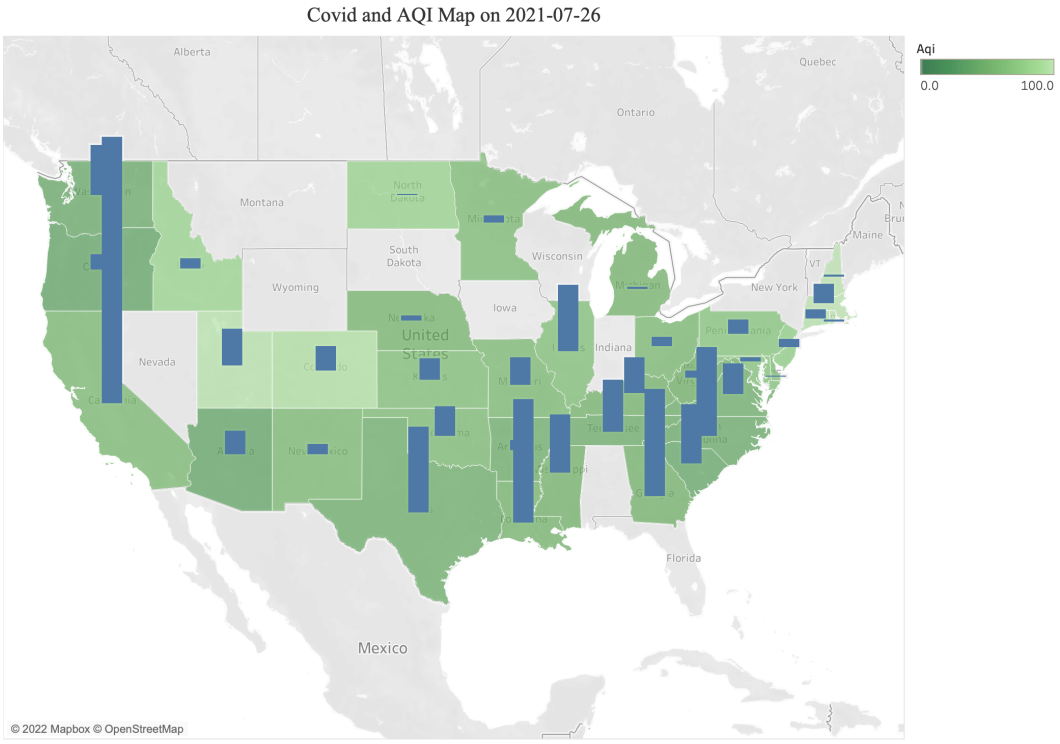

Figure 19: Spatial distribution of COVID and AQI on 2021-07-26

Therefore, the negative correlation between AQI and infected cases is found in our study due to the mutual interaction.

3.2 Machine learning

In addition to identifying potential correlations in our data, we can also derive insights from it by applying machine learning techniques such as regression and classification. However, because we did not find a strong association between the weather patterns, COVID-19 cases, and sentiment expressed in tweets, it does not make sense to use some features of our data to predict or classify the rest. Therefore, we will not be using regression and classification in our analysis.

Instead, we will be using a clustering approach to group our data based on the attributes derived from weather data and COVID-19 cases. This will allow us to see how the government might respond differently to different groups of data. To visualize our data in two or three dimensions, we will be using a technique called t-distributed stochastic neighbor embedding (tSNE) Van der Maaten & Hinton (2008), which is commonly used for mapping high-dimensional data to lower dimensions while preserving the similarity between points. This approach converts the similarity between points in high and low dimensions into a joint distribution, and then uses gradient descent to optimize the joint distribution in the low-dimensional space using KL divergence as the objective function.

In order to preserve the patterns in each data point, we have chosen to only use data points from each state that fall within a one-month range, so that the weather and COVID-19 cases do not change significantly. Figure 20 shows the tSNE plot of our data, with each color representing a different state. The plot shows that there are three distinct clusters, but the clustering does not keep all data points from the same state together in the same cluster. For example, the green points are dispersed among two clusters. This indicates that even though the data points are from the same state, they have different statistical properties. As a result, it may be difficult to draw any clear conclusions from this plot.

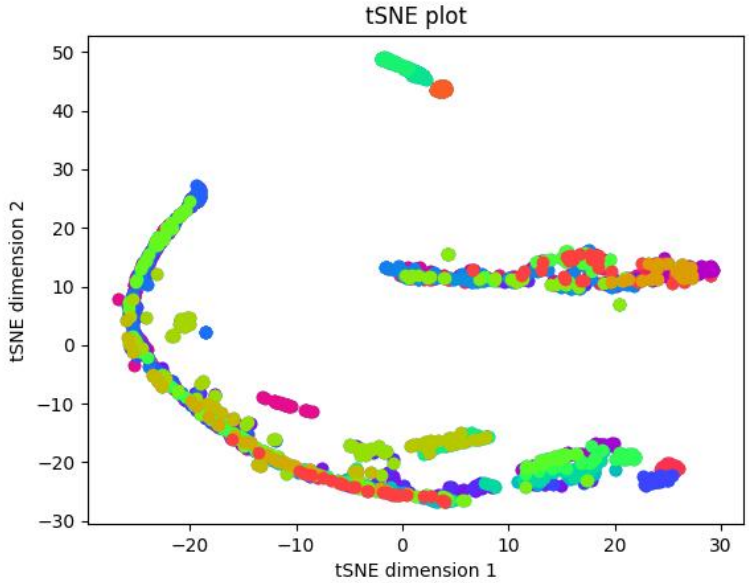

Figure 20: The tSNE visualization of our data

4 Discussions

4.1 Why is the correlation between COVID-19 cases and weather sometimes not significant?

Although the temperature was reported as significant in the greatest number of studies, with COVID-19 incidence increasing as temperature decreased, there are still controversies and the correlation sometimes is not significant in our study. We hypothesize that various factors may be overlooked. Firstly, epidemiological and social aspects are not factored out from the weather impact. For example, the COVID-19 vaccines, which were merely launched later in 2021 and reduce the number of infections, may disrupt the trend in 2020, even if the weather does not change significantly between 2020 and 2021. Secondly, time lags between weather and virus cases are not taken into account as there is delay between infection and reporting. Some research suggested a lag structure of 1 week on the weather variables while the results are similar(Michael Ganslmeier (2021)).

4.2 Why is the correlation between COVID-19 cases and tweets sentiment sometimes not significant?

There are several reasons why we may not have found a significant correlation between the sentiment expressed in tweets and the spread of COVID-19. One possible reason is that the location of tweet data cannot be determined, so the data we are analyzing includes tweets from all over the world. This means that signals from other countries may be mixed in with signals from the United States, resulting in a lack of correlation.

Another factor to consider is the sentiment score of the tweets. We observed that the majority of the tweets were positive, which is surprising given that we would expect people to be more negative during a pandemic. This was one of our doubts about the dataset of tweets, and it may be contributing to the lack of correlation we have found.

5 Summary

We explored the potential relationship between weather patterns, the spread of COVID-19, and the sentiment expressed in tweets related to the pandemic. Based on our analysis, we found that there is a negative correlation between COVID-19 cases and factors such as temperature and air quality,

although this relationship is not statistically significant due to the influence of other factors that affect social distancing behaviors. Additionally, we found that the correlation between COVID-19 cases and tweet sentiment is not statistically significant due to the presence of confounding variables. Overall, our findings suggest that further research is needed to fully understand the complex relationship between weather, COVID-19, and public sentiment.

References

Adam, M. G., Tran, P. T., and Balasubramanian, R. Air quality changes in cities during the covid-19 lockdown: A critical review. *Atmospheric Research*, 264:105823, 2021.

Contini D., C. F. Does air pollution influence covid-19 outbreaks? *Atmosphere*, 11:377, 2020.

Dean, J. and Ghemawat, S. Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.

Ganslmeier, M., Furceri, D., and Ostry, J. D. The impact of weather on covid-19 pandemic. *Scientific reports*, 11(1):1–7, 2021.

Kraemer, M. U., Yang, C.-H., Gutierrez, B., Wu, C.-H., Klein, B., Pigott, D. M., Group†, O. C.-. D. W., Du Plessis, L., Faria, N. R., Li, R., et al. The effect of human mobility and control measures on the covid-19 epidemic in china. *Science*, 368(6490):493–497, 2020.

Lowen, A. C., Mubareka, S., Steel, J., and Palese, P. Influenza virus transmission is dependent on relative humidity and temperature. *PLoS pathogens*, 3(10):e151, 2007.

McClymont, H. and Hu, W. Weather variability and covid-19 transmission: a review of recent research. *International journal of environmental research and public health*, 18(2):396, 2021.

Michael Ganslmeier, Davide Furceri, J. D. The impact of weather on covid-19 pandemic. *Scientific Reports*, 11(1):1–7, 2021.

Van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Wu, Y., Jing, W., Liu, J., Ma, Q., Yuan, J., Wang, Y., Du, M., and Liu, M. Effects of temperature and humidity on the daily new cases and new deaths of covid-19 in 166 countries. *Science of the Total Environment*, 729:139051, 2020.