COVID-19 and the Weather: A data visualization

Nick P Palacio

Department of Computer Science, University of Nebraska at Omaha

Capstone Project Proposal

Dr. Brian Dorn, Dr. Rex Cammack, Dr. Brian Ricks

COVID-19 and the Weather: A data visualization

On January 9th, 2020 the World Health Organization (WHO) announced that a coronavirus related pneumonia had been spreading in Wuhan, China. The US confirmed its first coronavirus case on January 21st, 2020 (American Journal of Managed Care, 2021). By March 11th, 2020 the WHO had declared the COVID-19 pandemic. Since then the US has experienced several waves of increased infection rates that have varied in severity across the country.

There is precedent to think that the COVID-19 virus spreads more easily in certain weather conditions. According to the CDC, COVID-19 can spread from human to human via respiratory droplets in the air, and the virus is known to spread more easily indoors where there is less air ventilation (CDC, 2020). Dr. Fauci, who serves as the director of the US National Institute of Allergy and Infectious Diseases, spoke about the potential connection between COVID-19 and the weather in April 2020 on ABC's Good Morning America saying:

> There is precedent with other infections like influenza and some of the common more benign coronaviruses that when the weather gets warmer that the virus goes down, that its ability to replicate, to spread, it doesn't like warm, moist weather as much as it likes cold, dry weather. But having said that, one should not assume that we are going to be rescued by a change in the weather. (AP, 2020)

Influenza is another respiratory illness that is spread via respiratory droplets in the air. It is well established that influenza spread is influenced by the weather, which Dr. Fauci alludes to above (Huang et al., 2017; Roussel et al., 2016). Roussel et al. (2016) studied the role of weather on seasonal influenza spread in France. Their study found 2 groups of 3 climatic variables that had a significant impact on seasonal influenza spread at the intra-annual scale. The first group of variables was average temperature, absolute humidity, and daily variation of absolute humidity. The second group of variables was sunshine duration, relative humidity, and daily variation of relative humidity. The impact of these groups of variables on seasonal influenza spread was found to be relatively low, between 3% – 6%. While the coronavirus is certainly not the same thing as the flu it does spread in a very similar manner. This makes the relationship between COVID-19 transmission and the weather worth exploring.

There has been some research published already exploring the relationship between weather and COVID-19. However, results from these studies have been mixed. One literature review published in the *International Journal of Environmental Research and Public Health* analyzed the current available literature on the association between weather and COVID-19 incidence (McClymont & Hu, 2021). This literature review looked for relevant studies on COVID-19 and weather by searching PUBMED, Web of Science and Scopus databases. The 23 articles selected for this review were epidemiological studies that evaluated the relationship between weather variables and COVID-19 transmission up to October 1st, 2020. All 23 articles included temperature in their study. 18 of the 23 studies reported a significant correlation between temperature and COVID-19 incidence. However, of these 18 studies 11 reported a negative correlation while the remaining 7 reported a positive correlation. 16 of the 23 articles included humidity in their assessment. Of these 16, 12 reported significant associations between humidity and COVID-19 incidence. However, of these 12, 4 reported a positive correlation, 6 reported a negative correlation and 2 reported an optimal range of humidity for new cases.

Another study published in the same journal highlighted an issue with the existing research on COVID-19 and weather. Jamshidi et al. (2020) said that existing research on this association only considers weather variables during analysis. In this study instead of just looking at weather variables and their impact on COVID-19 transmission they looked at other important factors such as mobility, homestay, population, and urban density. For their weather variable they used equivalent temperature which is a combination of temperature and humidity. The study evaluated the impact of equivalent temperature on COVID-19 transmission using different scales such as global, regional, US state and US county. At the global scale this study found contradictory patterns between equivalent temperature and COVID-19 infection rates. From January to July 2020 the USA, Italy and India showed a positive correlation between the two while China, Brazil and Australia had a negative correlation. At the US county scale equivalent temperature was found to have a contributing factor of <3%. This study recommended using finer scale weather data when incorporating it into a study given how much weather can vary across a country or region. They concluded that weather on its own was a non-influential factor in COVID-19 transmission. Instead, it said that other factors such as urban density and mobility of the population influenced COVID-19 transmission much more than weather.

One limitation of both studies is the data that they had to work with. The first research article discussed was received for peer review in November 2020. The second article was received in September 2020. This means that both articles were working with limited COVID-19 data, specifically missing out on spikes that were seen in the United States during the November 2020 – January 2021 time frame. That notwithstanding, these articles highlight the fact that there is an ongoing debate right now in the scientific community around weather’s role in the COVID-19 pandemic.

In this project I am to build a visualization tool that allows explanation of this relationship. The intended user for my project would be a middle school scientist because this open debate in the scientific community presents a unique opportunity to engage students. According to the Nebraska Department of Education (2017), by the 7th grade students should be able to understand evidence for how different factors contribute to the weather and climate. Students should also understand the scientific process for asking questions and carrying out investigations by gathering evidence. Given the right tools, teachers could leverage this debate to engage students in the scientific process by tasking them to perform their own investigation into the same question of weather's role in the COVID-19 pandemic. My visualization would equip a teacher with a tool that students could use to explore this relationship. An activity like this would make the students think critically and ask questions about the data and what conclusions can, or cannot, be drawn.

In 2018, Lee and Wilkerson studied data use by middle and secondary school students. One of the things they looked at was how teachers can best support students working with data. One of their recommendations for teachers' use of data in the classroom was that data should be leveraged in the context of meaningful scientific pursuits. My project falls in line with this guidance because students would be asked to participate in an open debate in the scientific community and draw their own conclusions using evidence they gather using the tool.

A study by Linn et al. (2006) found evidence that visualization technologies can improve student learning outcomes while they learn scientific concepts. From a high level, this study compared assessment results for two groups of students who received different curriculum. One group received a normal curriculum while the other group received curriculum that included visualizations of scientific phenomena in order to help illustrate it. They found that both groups of students performed equally well on multiple choice assessment questions. However, the group that received the curriculum that included the visualizations performed significantly better on assessment questions that required the student to provide their own explanations. Questions that require the student to provide their own explanations are better able to discriminate varying levels of knowledge integration, making these findings significant. While my visualization does not try to explain any particular scientific phenomena like heat transfer or a chemical reaction it does provide students a visual representation of a couple scientific phenomena, disease spread and weather.

Existing research has been aimed at proving or disproving weather's effect on the pandemic. My project aims to allow a user to explore this relationship on their own as opposed to establishing whether one exists or not. My proposed project is a web application that would allow a user to explore the relationship between weather and COVID-19 in different parts of the United States by interacting with a map and several charting widgets that would plot weather and COVID-19 infection data side by side.

Related Work

In this section I will compare some existing COVID-19 data visualizations to highlight work that is currently out there as well as some gaps in that work. I will also highlight some existing literature on how students interpret graphs as well as best practices for presenting graphs to students. The visualizations I selected for evaluation were found by doing my own research on the Internet. I wanted to find visualizations that came from a trustworthy organization and provided views into similar data points that I wanted to use, specifically confirmed cases by location. I looked at COVID-19 data visualizations from John Hopkins University of Medicine (John Hopkins, 2021), the COVID Tracking Project at the Atlantic (The COVID Tracking Project, 2021) and the Institute for Health Metrics and Evaluation (IHME) at the University of Washington (Institute for Health Metrics and Evaluation, 2021). I compared the visualizations that these organizations offered along several dimensions. Specifically, I looked at the following: How granular is the COVID-19 data? Which COVID-19 data points are visualized? Does it offer a spatial view? How configurable are the visualizations?

The results of this comparison can be seen in Table 1.

Table 1

Comparison of Existing COVID-19 Data Visualizations

Organization	**Granularity of COVID-19 Data**	**COVID-19 Data Points**	**Any Spatial View**	**Configurability of the Visualizations**
John Hopkins	County, State and Country	Confirmed Cases, Deaths, Tests, Hospital Use	Yes, map of US with counties	Minimal, can toggle the data point plotted
COVID Tracking Project	State and Country	Confirmed Cases, Deaths, Tests, Hospital Use	Yes, map with hospital use data, a few cartograms	Moderate, can set date range and if data is normalized
IHME	State and Country	Confirmed Cases, Deaths, Tests, Hospital Use	Yes, most data points can be viewed on a map	Moderate – High, can set date range, if data is normalized and if data should be 7-day rolling averages

Now I will summarize my findings and discuss how this relates to my visualization. Only one of the organizations, John Hopkins, offered COVID-19 data at the county level in the US. Given that my visualization will show weather and COVID-19 data together, the location granularity of this data becomes more important. Weather in any state can vary greatly across different locations in that state (Jamshidi et al., 2020). Therefore, my visualization will use county level COVID-19 data. All the organizations offered the same COVID-19 data points in their visualizations (cases, deaths, etc.). For my purposes of allowing a user to compare COVID-19 infection rates to weather patterns I will only be using confirmed COVID-19 case counts.

Since weather and COVID-19 infection rates both have a spatial dimension, a spatial view for my visualization is warranted. This is consistent with the existing visualizations I have looked at, all 3 provided some sort of spatial view for the COVID-19 data. This is why I will be displaying a map to the user that they can interact with in order to view data at their location of interest. These organizations offered a variety of levels of configurability in their visualizations. Given that the purpose of my visualization is to allow a user to explore the data on their own I will offer a high level of configurability in my visualization in order to allow a user to visualize the data in a few different ways.

These existing COVID-19 data visualizations are limited in a couple ways. They are not built with any specific group or learnability purpose other than presenting information, presumably to the general public. The purpose of my visualization will be to engage students to think critically about weather's role in the COVID-19 pandemic. As I will discuss later, visualization for learning requires special design considerations. Another limitation of existing visualizations is they are only concerned with displaying COVID-19 data. My visualization will allow a user to explore COVID-19 data alongside several weather data points a user can choose from.

Given my goal of allowing a user to investigate the relationship between two variables, COVID-19 infection and the weather, a scatterplot graph might be appropriate. There has been research done investigating how students interpret graphs as well as best practices for providing graphs to students for their interpretation. A literature review by Hoeffner and Shah (2002) looked at the cognitive literature on how people understand graphs. This paper looked at 3 factors that influence a viewer's understanding of a graph: the visual characteristics of the graph, a viewer's knowledge about graphs, as well as a viewer's knowledge about the data in the graph. The paper synthesizes these findings into recommendations for how to best present graphs to students. One of their recommendations was to represent the same data in multiple formats. This helps students' understanding when there are multiple quantitative facts to communicate about the data. I have 3 quantitative facts about the data I wish to communicate for a given US county and date range: the trend of COVID-19 infections, the trend of several weather data points, and the covariance of COVID-19 infections with each weather data point. Given this, I will provide a scatterplot graph that communicates the covariance of COVID-19 infections and a weather data point that a user could select from a predefined list. I will also provide individual line graphs of COVID-19 infections and each weather data point that will communicate the trend of each variable on its own. Another recommendation from this paper was to be careful about the density of the data points, specifically for scatterplots, because users often mentally exaggerate how correlated 2 variables are in a scatterplot that is very dense with data points. A graph can become denser by either adding data points or shrinking its size. This means that I will need to be careful to not try to plot too many data points on the scatterplot I provide depending on its size.

Methods

Data Sources

For my COVID-19 data source I will be using one of the datasets generated and maintained by the New York Times hosted on GitHub (The New York Times, 2021). This data source provides several datasets that can be downloaded via GitHub. There is also documentation about the datasets that can be viewed on GitHub to understand how they are structured. I will be using the us-counties.csv dataset. This dataset contains a full history of cumulative COVID-19 cases and deaths by county by day in the US going all the way back to January 1st, 2020. I evaluated two other sources for COVID-19 data before selecting the New York Times dataset. One of them came from the COVID Tracking Project published by *The Atlantic*. This data source provided an API as well as files you can download. However, it only had COVID-19 data at the state level. Given the location sensitive nature of both weather and COVID-19 data, state level data will not suffice. Weather in any state can vary greatly depending on location so I wanted county level data. The other data source I evaluated came from the Center for Systems Science and Engineering at Johns Hopkins University. This data source was also hosted on GitHub where the dataset files can be downloaded. This data source is very similar to the New York Times data source in that it provides case counts by county in the US. It also provides good documentation. This data source would work for my project as well. In the end I had to pick one so I went with the New York Times.

For my weather data I will be using an API from Weather Source. Weather Source is a technology company that provides a suite of products that help businesses leverage weather and climate data. On March 16th, 2020 Weather Source opened their API for free to any researchers exploring the relationship between weather and the COVID-19 pandemic. Their Weather History API exposes many different weather data points that can be queried with a date range along with latitude and longitude, or zip code. Data can be returned in an hourly or daily format. For my purposes I will be retrieving average temperature, average relative humidity and average absolute humidity in a daily format.

User Stories

Table 2 below contains the list of user stories for my project. User stories are descriptions of a feature of a piece of software told from the perspective of the user who desires the feature. Their typical format is 'As a *<type of user>*, I want *<some feature>* so that *<some reason>*'.

Table 2

User Stories

	User Stories
US.1	As a student, I can search for a county of interest on a map in order to begin to investigate weather's role in the pandemic for this county.
US.2	As a student, I can search for a county of interest by name and view a list of matches in order to begin to investigate weather's role in the pandemic for this county.
US.3	As a student, I can select my county of interest on a map and view COVID and weather data for that county.
US.4	As a student, I can view the number of daily COVID cases in my county of interest over a date range in order to understand the trend of COVID-19 cases for this date range for this county.
US.5	As a student, I can view daily average temperature in my county of interest over a date range in order to understand the trend of average temperature for this date range for this county.
US.6	As a student, I can view daily absolute humidity in my county of interest over a date range in order to understand the trend of absolute humidity for this date range for this county.

US.7	As a student, I can view daily relative humidity in my county of interest over a date range in order to understand the trend of relative humidity for this date range for this county.
US.8	As a student, I can view a scatter plot of the number of daily COVID-19 cases in my county alongside any of my weather data points over the date range in order to visualize the relationship between the 2 data points for this date range.
US.9	As a student, I can view the correlation coefficient with the scatterplot in order to understand the strength of the relationship between the 2 variables in the scatterplot.
US.10	As a student, I can adjust the date range for all data points.
US.11	As a student, I can view all the same daily data points rolled up to weekly averages in order to smooth out the data.
US.12	As a student, I can view all the same daily data points converted to 7 day rolling averages in order to smooth out the data.
US.13	As a student, I can share/save my current views of the data with another person by sharing my current URL in order to share evidence of my conclusions.

A couple user stories to point out are US.11 and US.12. Enabling the user to aggregate the data to weekly averages and 7 day rolling averages help to smooth out the data from large variations that can occur day to day in the COVID-19 dataset. This is important because there can be large spikes that occur in the COVID-19 datasets due to fluctuations in tests performed for a particular county on any given day. Aggregating helps to smooth out these spikes.

Another user story to point out is US.13. Enabling the user to share a link that contains their current application configuration will allow them to gather evidence they can use to support their conclusion. The URL would contain the current selected county, date range, and all the chart settings they have selected.

UI Design

Figures 1 – 6 contain mockups for what the UI design of my project will look like.

Figure 1

Home page

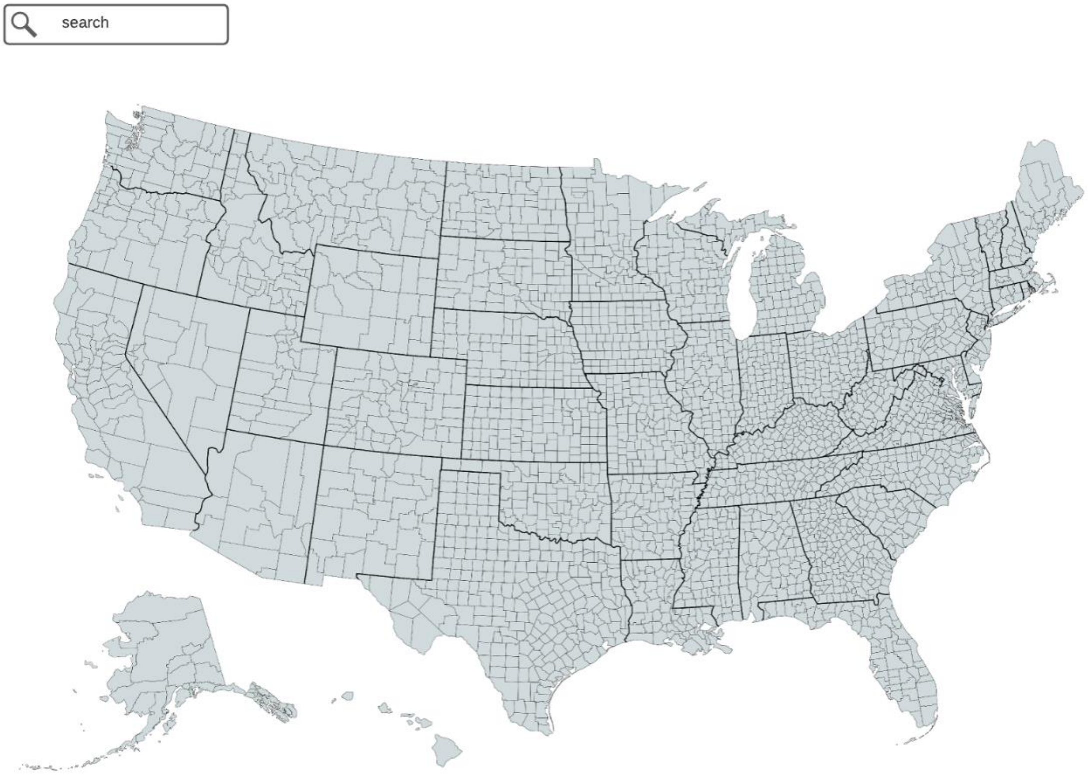
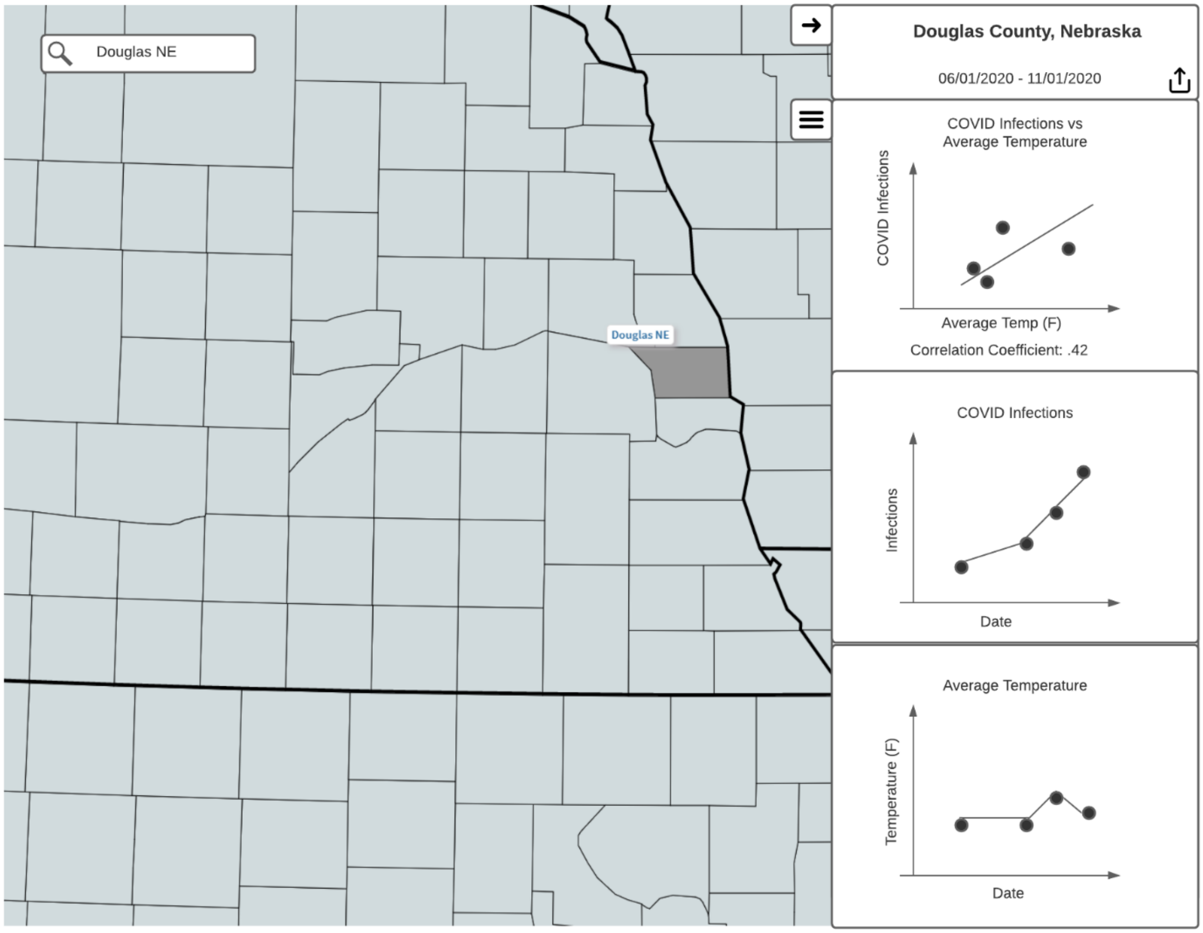


Figure 1 shows what the home page of the application could look like. This is a map that the user can interact with in order to zoom to or search for their county of interest.

Figure 2

County Selected



Douglas NE

Douglas County, Nebraska

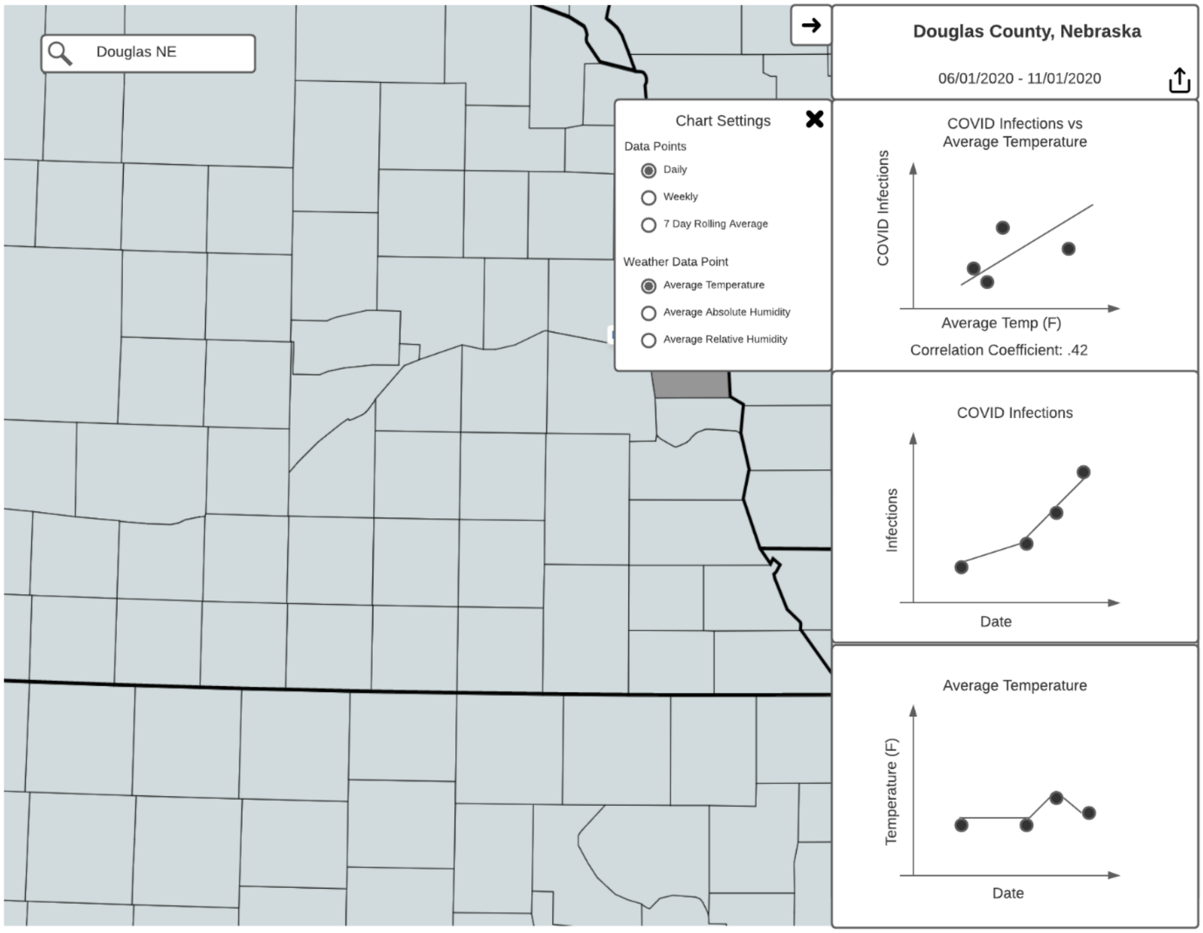
06/01/2020 - 11/01/2020

Correlation Coefficient: .42

Figure 2 shows the panel on the right that would show up once a county is selected, either through the search bar or by clicking on the map. This figure is related to US.3, US.4, US.8, and US.9.

Figure 3

Chart Settings Open



Douglas NE

Douglas County, Nebraska

06/01/2020 - 11/01/2020

Chart Settings

Data Points

● Daily

○ Weekly

○ 7 Day Rolling Average

Weather Data Point

● Average Temperature

○ Average Absolute Humidity

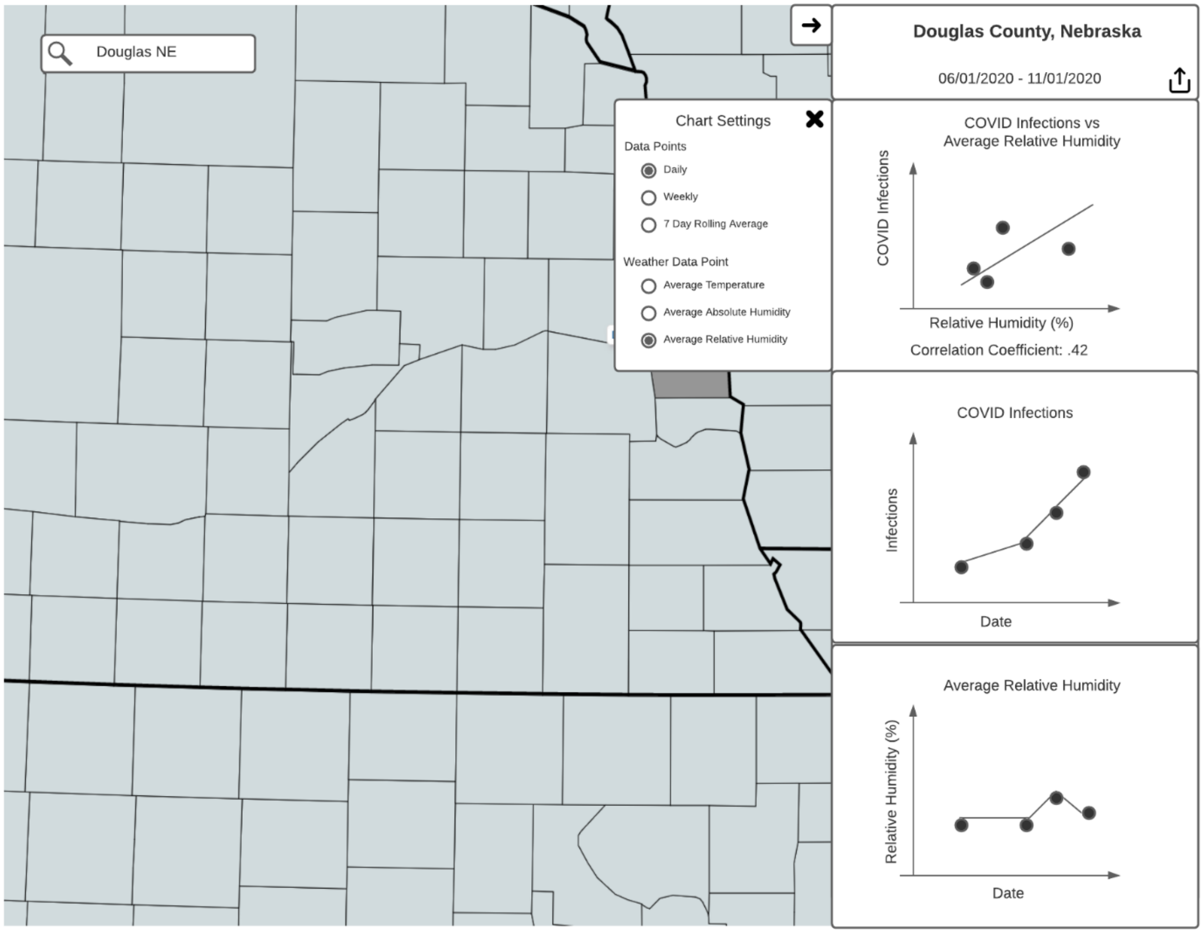
○ Average Relative Humidity

Correlation Coefficient: .42

Figure 3 shows what the chart setting menu could look like. There are options there to change the data points from daily to weekly averages or 7-day rolling averages. This is related to US.5, US.6, and US.7.

Figure 4

Weather Data Point Changed



Douglas NE

Douglas County, Nebraska

06/01/2020 - 11/01/2020

Chart Settings

Data Points

● Daily

○ Weekly

○ 7 Day Rolling Average

Weather Data Point

○ Average Temperature

○ Average Absolute Humidity

● Average Relative Humidity

Figure 4 shows what happens when a user changes the ‘Weather Data Point’ chart setting from ‘Average Temperature’ to ‘Average Relative Humidity’. Changing this setting updates the bottom chart as well as the scatterplot on top.

Figure 5

Share Button Clicked

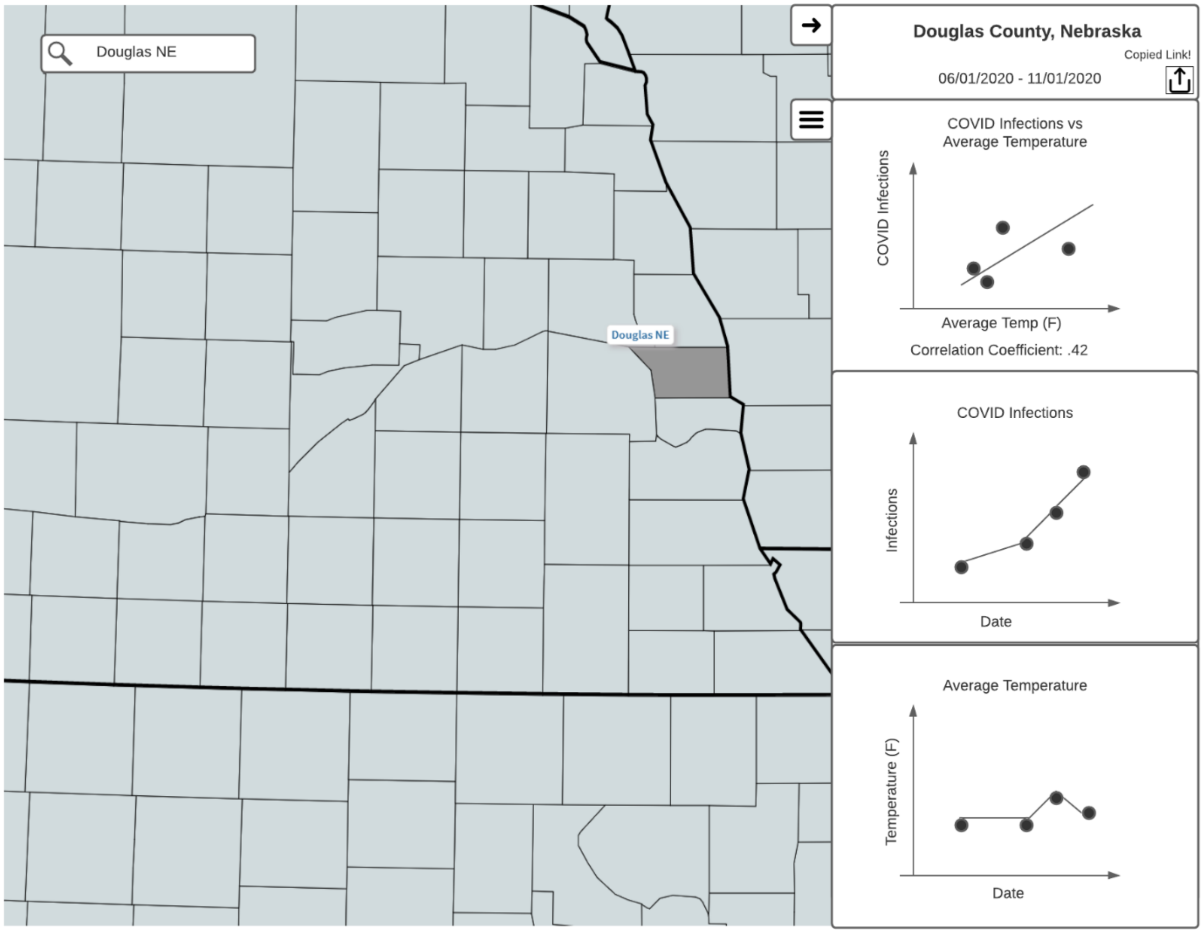


Figure 5 shows the sharing feature. Students would be able to share URLs of the application in different configurations in order to show someone else what they are seeing.

Figure 6

Chart Panel Closed

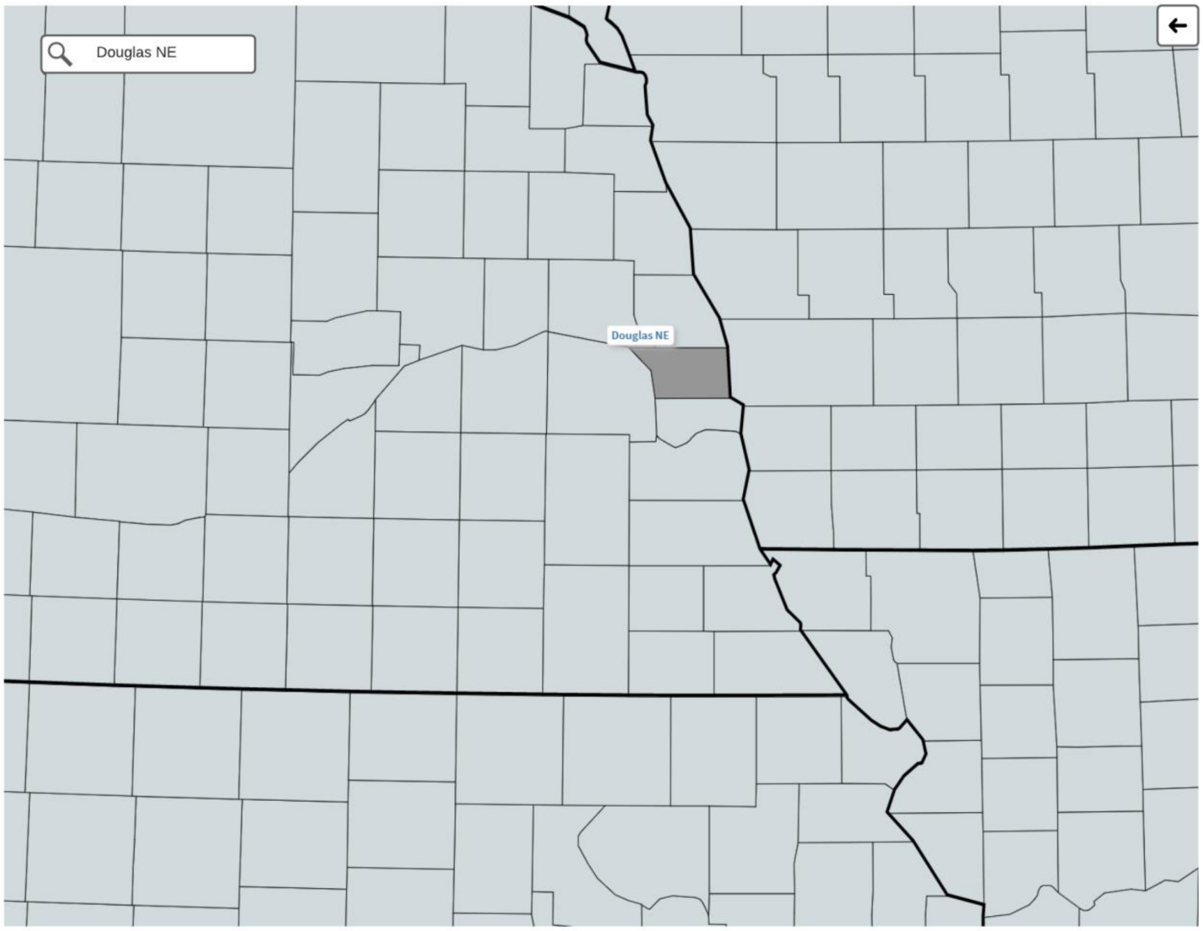


Figure 6 shows what the screen looks like after a user closes the panel on the right.

Data Preparation

As previously mentioned, my data source for COVID-19 infections provides cumulative cases by county by day. In order to implement US.11 and US.12 from Table 2 I will need to transform this data. For US.11 I will need to convert the daily case count for a county over a date range into a weekly average case count for that county over that same date range. This can be done by averaging the case count every seven days in order to get the average case count for the week that the 7 days represents.

For US.12 I need to convert the daily case count for a county over a date range into a 7-day rolling average. This can be defined as:

$$7DayAverage[x] = \frac{\sum_{n=x-7}^{x-1} dataByDay[n]}{7} \; (1)$$

Where *x* is the day we are calculating for and $dataByDay$ is an array of the confirmed cases in a county by day. The same transformations detailed above will be used for the weather dataset where I start with daily weather data points that can be rolled up to weekly and 7-day rolling averages.

Resources

For my project I will create an ASP.NET Core Web Application that will serve up an Angular frontend application. The ASP.NET Core application will also provide the API endpoints for the Angular frontend. The weather endpoints in my application will use the WeatherSource API behind the scenes. The COVID-19 endpoints in my application will use an Azure SQL Database that will be loaded nightly with the latest CSV dataset from the New York Times. This ASP.NET Core Web Application will be hosted in an Azure App Service via a free student account that I am able to register for. This nightly load of the CSV data into an Azure SQL database can be done with a WebJob inside of my Azure App Service. This nightly load could also be done with an Azure Function setup with a scheduled trigger.

Timeline

Figure 7 below shows a Gantt chart representing the timeline of project deliverables as user stories completed. I would aim to be presenting my project the week of July 19th – 22nd, 2021.

Figure 7

Gantt Chart of Project

	5/10	5/17	5/24	5/31	6/7	6/14	6/21	6/28	7/5	7/12	7/19	7/22
Setup												
Create code repository	🟧	🟧										
Setup CI/CD Pipelines	🟧	🟧										
Mapping Functionality												
US.1		🟧	🟧									
US.2			🟧	🟧								
US.3				🟧	🟧							
Chart Functionality												
US.4					🟧	🟧						
US.5						🟧	🟧					
US.6						🟧	🟧					
US.7						🟧	🟧					
US.8							🟧	🟧				
US.9								🟧	🟧			
US.10								🟧	🟧			
US.11									🟧	🟧		
US.12									🟧	🟧		
US.13									🟧	🟧		
Project Write-up										🟧	🟧	🟧
Project Presentation										🟧	🟧	🟧

References

America Journal of Managed Care. (2021, January 1). A Timeline of COVID-19 Developments in 2020. https://www.ajmc.com/view/a-timeline-of-covid19-developments-in-2020

AP. (2020, April 9). Dr. Fauci: Don't assume coronavirus fades in warm weather. ABC7 New York. https://abc7ny.com/6089537/

CDC. (2020, October 28). COVID-19 and Your Health. Centers for Disease Control and Prevention. https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/how-covid-spreads.html

Huang, X., Mengersen, K., Milinovich, G., & Hu, W. (2017). Effect of Weather Variability on Seasonal Influenza Among Different Age Groups in Queensland, Australia: A Bayesian Spatiotemporal Analysis. The Journal of Infectious Diseases, 215(11), 1695–1701. https://doi.org/10.1093/infdis/jix181

Institute for Health Metrics and Evaluation. (2021, March 20). IHME | COVID-19 Projections. Institute for Health Metrics and Evaluation. https://covid19.healthdata.org/

Jamshidi, S., Baniasad, M., & Niyogi, D. (2020). Global to USA County Scale Analysis of Weather, Urban Density, Mobility, Homestay, and Mask Use on COVID-19. International Journal of Environmental Research and Public Health, 17(21), 7847. https://doi.org/10.3390/ijerph17217847

John Hopkins. (2021, March 20). Coronavirus Resource Center. Johns Hopkins Coronavirus Resource Center. https://coronavirus.jhu.edu/

Lee, Victor R, and Michelle H Wilkerson. "Data Use by Middle and Secondary Students in the Digital Age: A Status Report and Future Prospects," n.d., 43.

Linn, M., Lee, H.-S., Tinker, R., Husic, F., & Chiu, J. (2006). Teaching and Assessing Knowledge Integration in Science. Science (New York, N.Y.), 313, 1049–1050. https://doi.org/10.1126/science.1131408

McClymont, H., & Hu, W. (2021). Weather Variability and COVID-19 Transmission: A Review of Recent Research. International Journal of Environmental Research and Public Health, 18(2), 396. https://doi.org/10.3390/ijerph18020396

Nebraska Department of Education. (2017). Nebraska's College and Career Ready Standards for Science. https://cdn.education.ne.gov/wp-content/uploads/2017/10/Nebraska_Science_Standards_Final_10_23.pdf

Roussel, M., Pontier, D., Cohen, J.-M., Lina, B., & Fouchet, D. (2016). Quantifying the role of weather on seasonal influenza. BMC Public Health, 16, 441. https://doi.org/10.1186/s12889-016-3114-x

Shah, P., & Hoeffner, J. (2002). Review of Graph Comprehension Research: Implications for Instruction. Educational Psychology Review, 14(1), 47–69. https://doi.org/10.1023/A:1013180410169

The COVID Tracking Project. (2021, March 20). Charts. The COVID Tracking Project. https://covidtracking.com/data/charts

The New York Times. (2021). Coronavirus (Covid-19) Data in the United States. Retrieved March 7, 2021, from https://github.com/nytimes/covid-19-data.