Castle Leonard
December 14th, 2021
A7 Project Report
DATA 512: Human-centered Data Science

Table of Contents

Introduction

This course project consisted of four assignments (A4 through A7), where:

- Assignment A4: Common Analysis set the stage for the subsequent assignments. In A4, I attempted to address the question “How did masking policies change the progression of confirmed COVID-19 cases from February 1, 2020 through October 15, 2021?” for Worcester COunty, Massachusetts.
- Assignment A5: Extension Plan required that I pose my own human centered data science question that extends the work in A4: Common Analysis. I wanted to answer the question “To what extent does weather explain the increase in infection rate of COVID-19 in Worcester, Massachusetts over time?” by discovering “which pieces of information are most useful and what percentage of the variability in the covid case fluctuates can be accounted for by weather metrics”.

- Assignment A6: was to present a PetchaKucha style presentation of the completed project. Both the .pdf and .pptx versions are available as part of the complete documentation in a public github repo (https://github.com/CastleA/data-512-a7) .
- Assignment A7: Final Report requires a written submission of a project report.

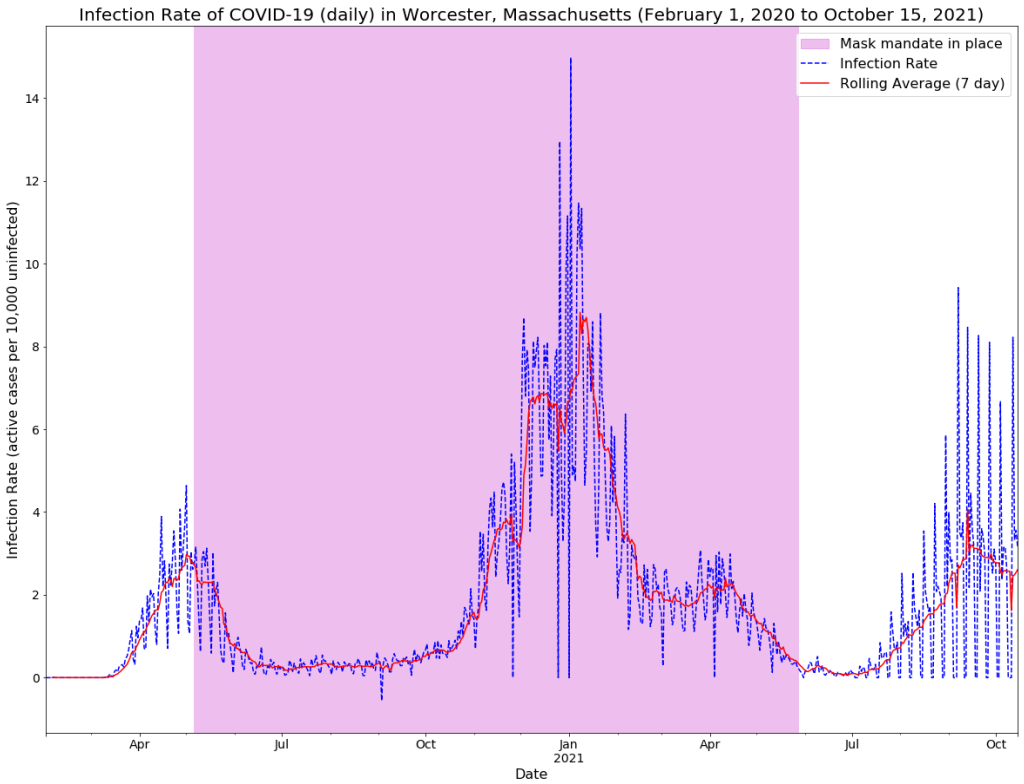
I will begin by introducing the analysis done for A4: Common Analysis as it is the basis of the following A5 analysis.

A4: Common Analysis

The research question to answer for the given county of Worcester, Massachusetts was:

> How did masking policies change the progression of confirmed COVID-19 cases from February 1, 2020 through October 15, 2021?

Using data from a Kaggle repository of John Hopkins University COVID-19 data (RAW_us_confirmed_cases.csv) containing daily reports of confirmed COVID-19 cases and a CDC dataset of mask mandates (masking mandates by county), I constructed a plot of time series of COVID spread in the context of mask mandates, shown below:



The figure itself tracks the daily infection rate (y-axis) over time, from February 1st, 2020 to October 15th, 2021 (x-axis). Infection rate is defined as the number of new cases per 10,000 uninfected population. The uninfected population uses the fixed population from the 2020 census and subtracts the number of active cases on that day. Cases are

considered active starting two days before the test is confirmed and for the following 8 days, under the assumption that people with positive tests have been infected for some time prior and the test results also contain some delay. The 10 day active infection assumption is the rough midpoint between the shortest and longest 7 and 14 days estimates of infectiousness. The pink shaded portion of the graph shows the time period in which masking mandates were in place for all indoor, public spaces and outdoor spacing when distancing could not be maintained. Lastly, to better see the slope of the infection rate a smoothed, rolling average is plotted in red over top of the raw data. The 7 day day rolling average was smoothest because it aligned with the weekly reporting patterns that had zero reported cases on some days.

The most interesting pattern is that the initiation of the masking mandate is a peak of infection, which could imply that the mandate stemmed the tide of infections, and that concludes at a trough before the infection rate again grows from July into October. It seems likely that the confounding factor would be public fear of COVID-19 which would move politicians to put measures in place and may coincide with greater caution and adherence with advice to isolate as much as possible. By the end of May, the lowered infection rate, an impatience with COVID safety measures, and a desensitization of the public to the dangers could explain why the mandate was lifted. Second most notable is that there was a dramatic peak of infection within the mask mandate at the beginning of January. This may be explained by the winter and the holiday season in the US. More time indoors and perhaps more exposure through social gatherings and lenience with masking around friends and family may explain the phenomenon.

There are too many potential confounders to say for certain, but It appears that masking policies may be effective at reducing the rate of infection when aligned with public beliefs, other associated precautions, and incentives.

A5: Extension Plan

In my discussion of findings in A4, I noted that “there was a dramatic peak of infection within the mask mandate at the beginning of January” and theorized that it could be explained by “the winter and the holiday season in the US” due to “more time indoors and perhaps more exposure through social gatherings and lenience with masking around friends and family”. For A5, I decided to go beyond this idle mulling and attempt to confirm for myself if the prevailing wisdom that illness, especially colds and flu, are driven by winter weather or the avoidance of it (i.e. staying indoors during uncomfortable weather) is supported by the data. I also considered the impact of human psychology around fall and winter holidays as well as diminishing patience with and capacity for adhering to strict safety precautions, however sourcing data to support such a claim proved quite challenging. I decided to attempt to capture the influence of weather on people as well as viruses with the idea that the amount of variation that

remained unaccounted for would stand in for the many other complex and difficult to quantify factors, including human behavior and psychology. The implication would be that we've been letting ourselves off the hook for not following health precautions more thoughtfully by casting the cold and flu season as a natural and unavoidable phenomenon, not unlike the common attitude towards automobile deaths. With this in mind, I set out to address the research question "To what extent does weather explain the increase in infection rate of COVID-19 in Worcester, Massachusetts over time?" by discovering "which pieces of information are most useful and what percentage of the variability in the covid case fluctuates can be accounted for by weather metrics".

Background/Related Work

The background for my A5 research question came from years of absorbing the commonly repeated explanation that cold and flu season is due to winter weather and the physical and behavioral changes that it encourages. I had begun to take this knowledge for granted, however due to my increased education in bias and data I found myself beginning my new practice of trying to come up with confounding variables rather than taking a correlation at face value (with implied causation, no less). I also recalled the reporting done around Thanksgiving as an anticipated "super-spreader event" and wondered if we as a society have been blaming our winter illness woes wholly on the uncontrollable outside force of weather as a way to deflect responsibility for our own role. From this hunch, I hypothesized that weather data would not be able to account for more than 33% of the variability in new COVID cases. This was not a formal or testable hypothesis, but rather a way to document my expectations and avoid HARKing.

I was also reminded of the plot point of Pride & Prejudice where Jane catches a cold from riding in a downpour to dine at Netherfield. I don't have the highest opinion of medical science from the 1800s, but this too is not an uncommon belief. I performed some light research on the effect of weather on the body to confirm and in "What's the link between cold weather and the common cold?" came across the latest understanding that cold, dry air both benefits virus spread and weakens humans. This became the basis of my expectation that humidity and temperature would be the most impactful weather metrics on the spread of COVID-19.

Methodology

I began by cleaning, aggregating, and joining the weather data from the National Oceanic and Atmospheric Administration (NOAA) (https://www.ncdc.noaa.gov/cdo-web/datatools). I then engineered many features surrounding temperature, precipitation, snowfall, humidity, and wind speeds, including rolling averages, differences in the day or week to week. I also derived season feature

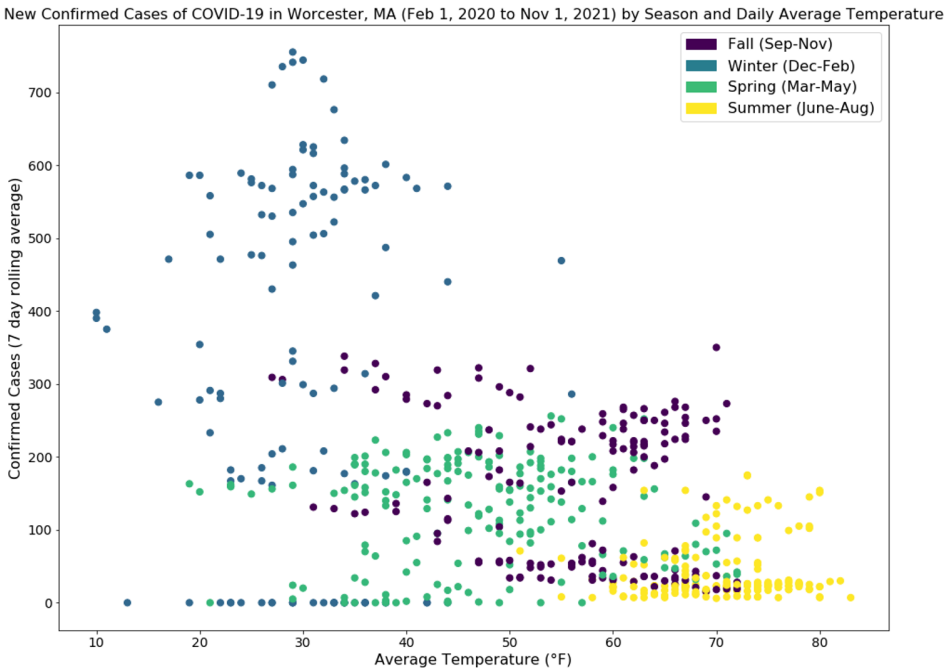
based on the mapping of Winter, December to February, Spring, March to May, Summer, June to August, and Fall, September to November, to try to capture the overall effect of seasonality more than just changes in weather. This metric would account for any calendar patterns such as breaks from school or increased rates of vacation or visitation. I then trained several multivariate regression models to predict the number of new cases using subsets of the raw and derived weather and season data. I took care to remove features that were overly correlated with each other, for example minimum, maximum, and average temperature all had correlations above 0.9.

I opted for this modeling technique because I was interested in the influence of each predictor variable in the form of the coefficient, even as a descriptive metric as opposed to predictive. I also began with the belief that weather would only be part of the story with the spread of infectious disease and the R^2 metric lent itself conveniently since it is inherently a measure of proportion.

With respect to ethics, I am not aware of any risk of harm due to the analysis of COVID case data and the weather. Whereas analyses involving vaccination effectiveness, unemployment rate, or crime rates could misrepresent the interaction in a way that influences policy or paints an unfair or misleading picture of real problems, the weather is not within our power to affect or tied closely with political bodies where discrimination can become a concern.

Findings, Discussion, & Implications

To begin I simply wanted to observe the interaction between average daily temperature, number of new cases, and control for season.



I was interested in separating out winter weather from the calendar season of winter, since they are often conflated when being named as the cause of cold and flu season. The scatter plot shows a clear separation of high cases numbers in winter even for similar temperatures in Fall and Spring. Meanwhile, summer is bunched rather tightly in the opposite corner as expected while spring and fall commingle in between. This is when I began to suspect that weather, especially temperature, would be too correlated with season such that they could serve as confounders to each other. I shaped my analysis to account for this by testing those sets of variables separately and then together.

Analysis 1

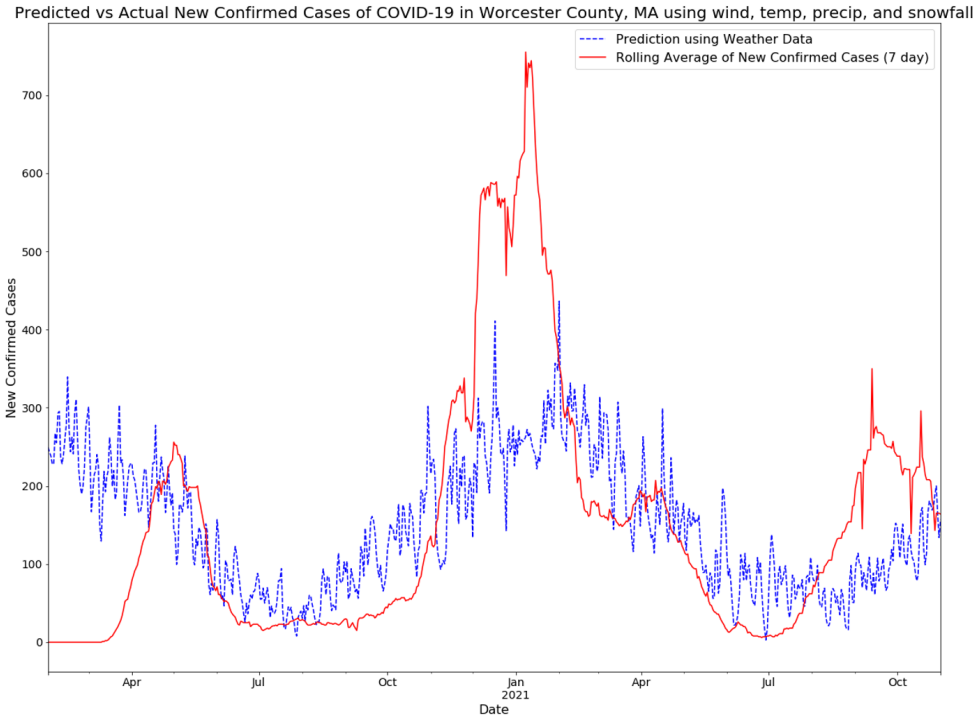
As a baseline, I first trained a multivariate linear model using the cleaned, but unprocessed weather data. The results showed that only about 28% of the variability of the response variable, the seven day rolling average of new confirmed cases, was explained. Additionally, only the average daily temperature was significant, but given the previously noted concerns of correlation with season this warranted scrutiny.

OLS Regression Results

Dep. Variable:	RollingAverage	R-squared:	0.277
Model:	OLS	Adj. R-squared:	0.273
Method:	Least Squares	F-statistic:	60.28
Date:	Mon, 13 Dec 2021	Prob (F-statistic):	4.22e-43
Time:	20:50:17	Log-Likelihood:	-4020.9
No. Observations:	634	AIC:	8052.
Df Residuals:	629	BIC:	8074.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	414.2841	27.662	14.977	0.000	359.963	468.605
Average daily wind speed (mph)	-0.3086	1.601	-0.193	0.847	-3.453	2.836
Precipitation (in)	0.2677	14.057	0.019	0.985	-27.336	27.871
Snowfall (in)	8.9337	5.921	1.509	0.132	-2.693	20.561
Average temperature (°F)	-4.9517	0.346	-14.316	0.000	-5.631	-4.272

To understand what the model had learned, I then visualized the predictions overlayed on the actual response variable. The approximation was a somewhat rough sinusoidal curve that could not account for the exponential tendency of infection curves.



Analysis 2

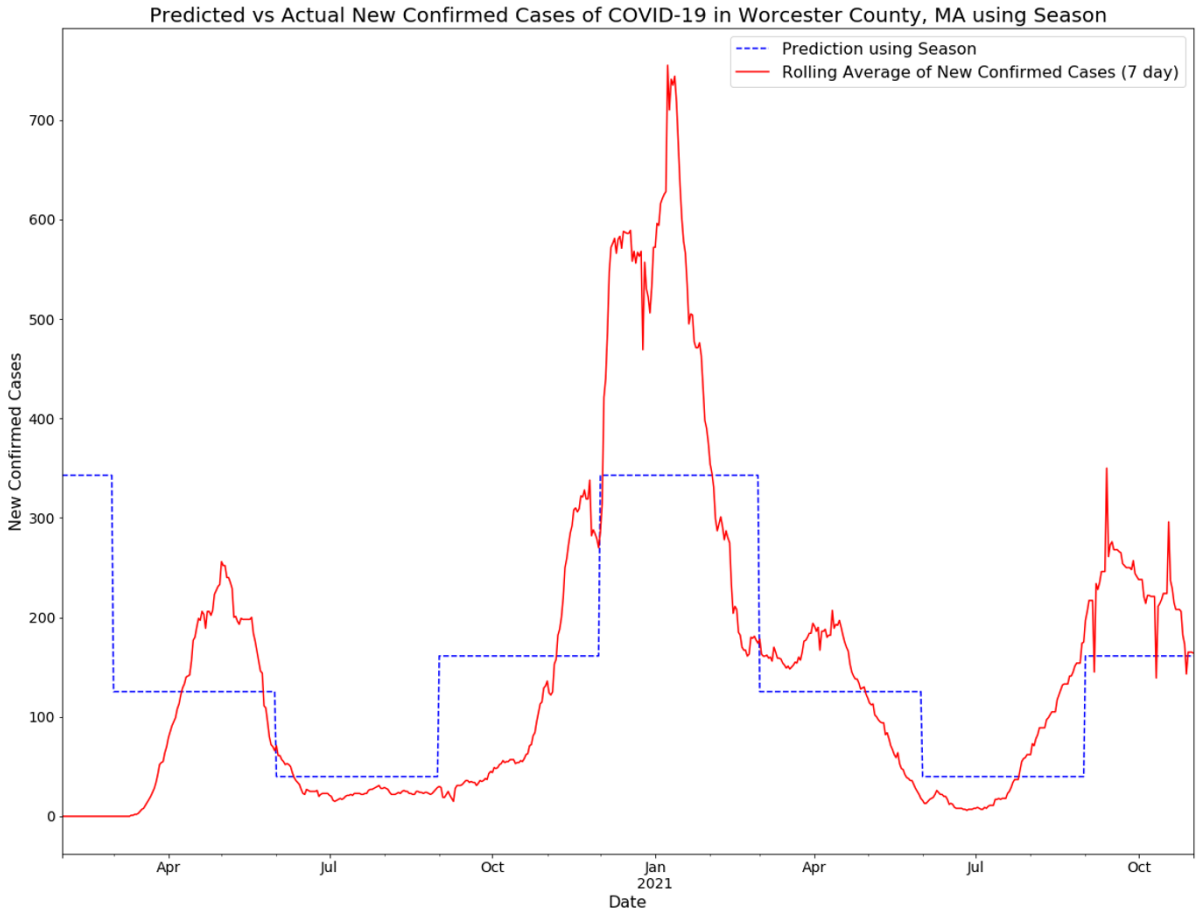
Similarly to Analysis 1, I trained another baseline multivariate linear model using only the one hot encoded seasons and leaving summer to correspond to the baseline. The results showed improvement with about 43% of the variability of the response variable, explained. This would be more compelling with more years of data, but they were all significant nonetheless and aligned intuitively with the chart from A4.

OLS Regression Results

Dep. Variable:	RollingAverage	R-squared:	0.433
Model:	OLS	Adj. R-squared:	0.431
Method:	Least Squares	F-statistic:	160.5
Date:	Mon, 13 Dec 2021	Prob (F-statistic):	2.72e-77
Time:	20:50:19	Log-Likelihood:	-3943.8
No. Observations:	634	AIC:	7896.
Df Residuals:	630	BIC:	7913.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	39.9565	9.000	4.440	0.000	22.283	57.630
isFall	121.1958	13.405	9.041	0.000	94.872	147.519
isWinter	314.7565	14.512	21.690	0.000	286.260	343.253
isSpring	85.3859	12.727	6.709	0.000	60.392	110.379

The visualization of the predictions overlayed on the actual response variable were unvarying and steeper than in Analysis 1, but still reminiscent of a sinusoidal curve.



Analysis 3

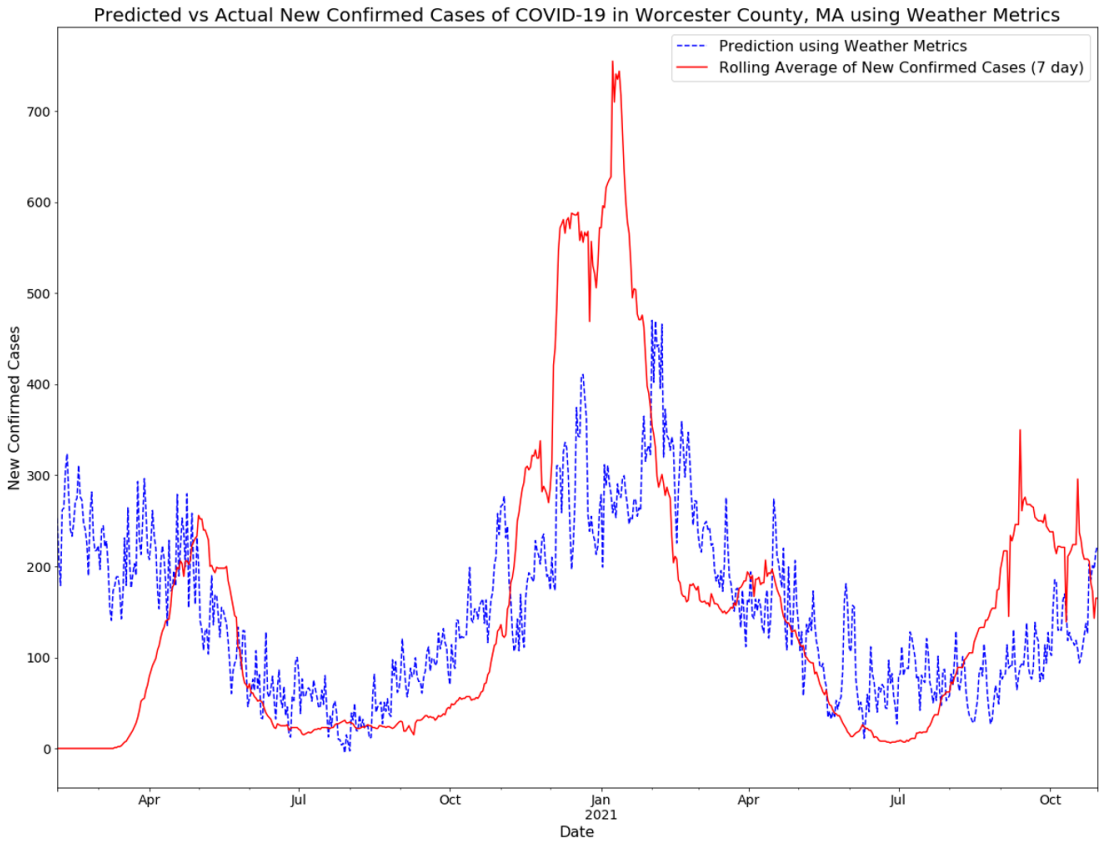
In Analysis 3, I trained another multivariate linear model using derived weather metrics that I suspected may influence human and virus behavior. The results showed improvement with about 33% of the variability of the response variable, explained. This demonstrated that weather data, when processed thoughtfully does have some predictive power towards infection spread.

OLS Regression Results

Dep. Variable:	RollingAverage	R-squared:	0.452
Model:	OLS	Adj. R-squared:	0.442
Method:	Least Squares	F-statistic:	46.58
Date:	Mon, 13 Dec 2021	Prob (F-statistic):	6.48e-74
Time:	20:50:21	Log-Likelihood:	-3933.3
No. Observations:	634	AIC:	7891.
Df Residuals:	622	BIC:	7944.
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	10.1564	59.765	0.170	0.865	-107.209	127.522
Glaze or rime	-104.9523	38.533	-2.724	0.007	-180.622	-29.282
Average relative humidity	0.2834	0.378	0.750	0.453	-0.458	1.025
AverageTempWeek	0.4477	0.754	0.594	0.553	-1.032	1.928
ChangeInAverageTempWeek	-1.2834	0.860	-1.493	0.136	-2.972	0.405
AveragePrecipWeek	23.8079	32.822	0.725	0.468	-40.647	88.262
AverageSnowWeek	28.6680	14.136	2.028	0.043	0.907	56.429
ChangeInTempDay	-1.4369	0.991	-1.450	0.147	-3.383	0.509
TotalPrecipWeek	-2962.6181	3759.242	-0.788	0.431	-1.03e+04	4419.725
isFall	118.8680	16.192	7.341	0.000	87.070	150.666
isWinter	318.0849	32.591	9.760	0.000	254.083	382.087
isSpring	103.3613	22.112	4.675	0.000	59.939	146.784
TotalSnowfallWeek	0	0	nan	nan	0	0

The visualization of the predictions overlayed on the actual response variable align much more closely this time.



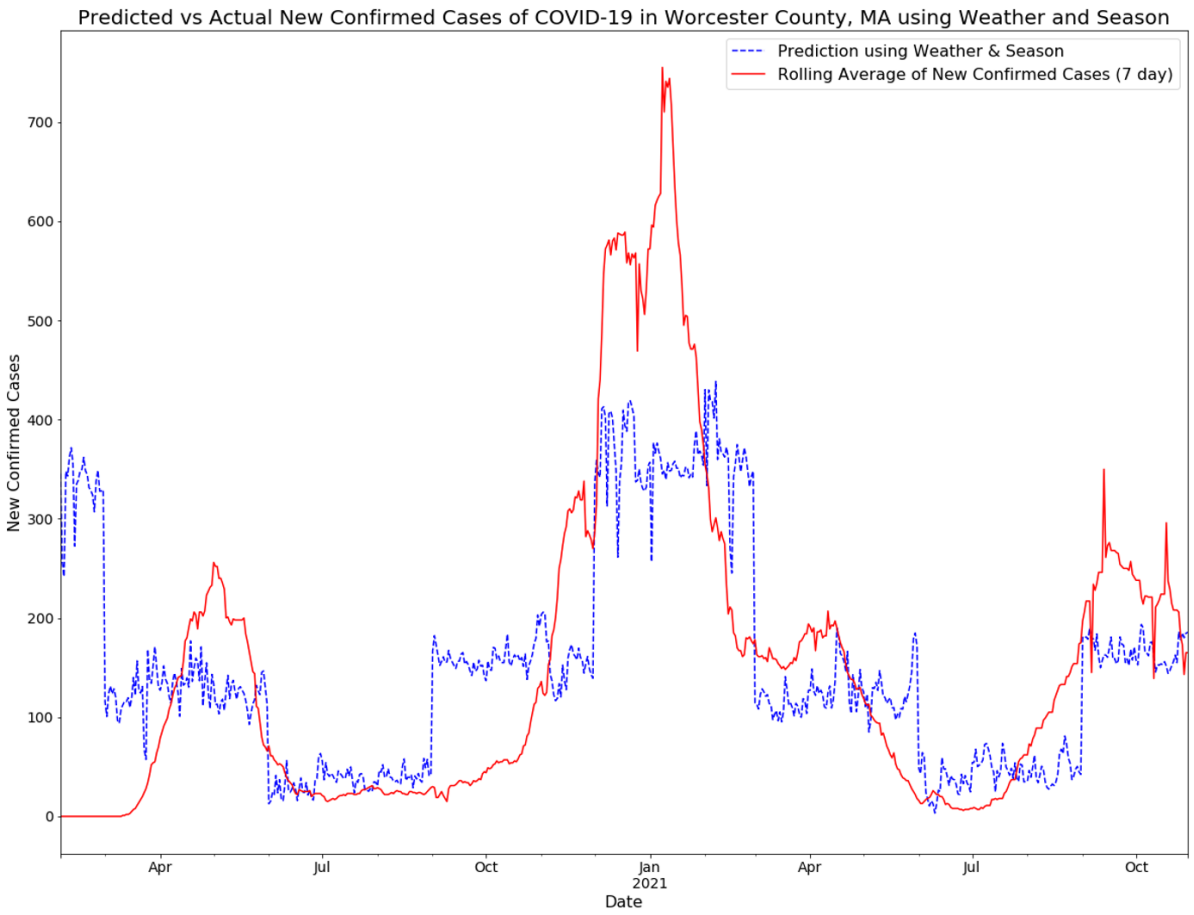
Analysis 4

Lastly, I trained another multivariate linear model using derived weather metrics that I suspected may influence human and virus behavior in addition to one hot encoded seasons. The results showed the best result yet with about 44% of the variability of the response variable, explained. The importance of weather data did drop though, indicating the seasonality is still a strong influencer.

OLS Regression Results

Dep. Variable:	RollingAverage	R-squared:	0.435
Model:	OLS	Adj. R-squared:	0.429
Method:	Least Squares	F-statistic:	68.96
Date:	Mon, 13 Dec 2021	Prob (F-statistic):	1.34e-73
Time:	20:50:20	Log-Likelihood:	-3942.6
No. Observations:	634	AIC:	7901.
Df Residuals:	626	BIC:	7937.
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	76.6685	41.025	1.869	0.062	-3.894	157.231
Average daily wind speed (mph)	0.1449	1.464	0.099	0.921	-2.730	3.019
Precipitation (in)	1.2623	12.485	0.101	0.919	-23.256	25.780
Snowfall (in)	5.0639	5.281	0.959	0.338	-5.308	15.435
Average temperature (°F)	-0.5415	0.537	-1.009	0.313	-1.596	0.513
isFall	113.0197	15.523	7.281	0.000	82.536	143.504
isWinter	289.5032	26.293	11.011	0.000	237.871	341.136
isSpring	71.9198	18.275	3.935	0.000	36.033	107.807

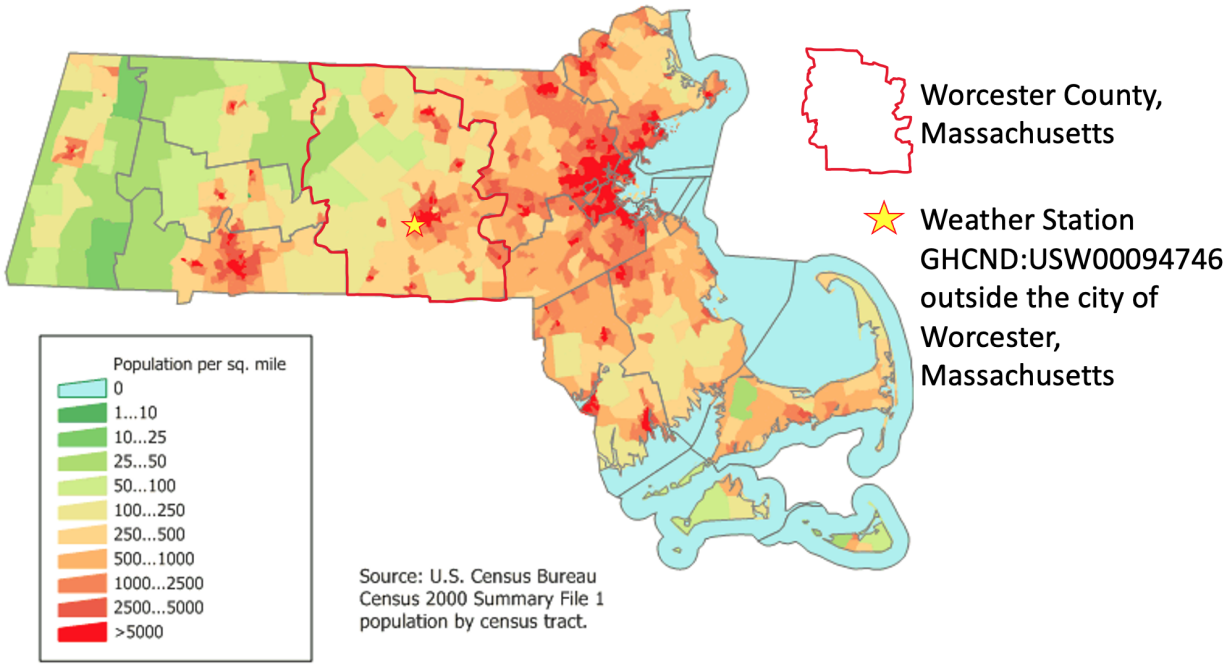


Limitations

Limitations of multivariate linear regression are that it assumes a linear interaction between the predictor and response variables when that is not a given. I debated whether to introduce variations of my features to match the fact that infection curves are noticeably exponential, however I decided not to as I understood that the exponential curve is based on infection models (ratios of infected to uninfected, etc.) and I did not have a compelling theory as to why weather would be contributing to exponential change as well. Additionally, this model technique is based on mean values, which as an aggregation can be misleading. For this reason it can also be sensitive to outliers. Beyond that, the model has an assumption of independence, which between days and across metrics is not the case. Lastly, I considered introducing interaction factors between temperature and humidity, but without proper standardization that tactic can be misleading and I had wanted to understand coefficients in relation to familiar units.

With respect to data, I opted to use the weather station with the most complete data and supportable location in the context of my analysis based on population density.

However supportable the location may be, by a rough weighted average of population density, it is still localized and may not be representative across the county.



Likewise, the COVID data set showed signs of retroactive modifications to the cumulative case numbers, which would throw off accurate day to day counts. It also did not include tracking on the weekends, which introduced weekly seasonality into the counts and needed to be mitigated via a seven day rolling average, which could affect the soundness of the model.

Finally, the formulation of my analysis was to quantify the proportion of spread in illnesses that can be explained by weather data, however this technique could only supply a lower bound based on the quality of feature engineering. This makes my already informal hypothesis that it would not account for more than a third of the variation even less credible. Additionally, in retrospect, I should have cropped the data so as not to include February and the parts of March when COVID testing was not yet reporting, which likely skewed the models.

Conclusion

In this A5 analysis, I set out to address the research question “To what extent does weather explain the increase in infection rate of COVID-19 in Worcester, Massachusetts over time?” by discovering “which pieces of information are most useful and what percentage of the variability in the covid case fluctuates can be accounted for by weather metrics”.

The findings were more descriptive than conclusive and seemed to point at calendar seasons as well as seasonal weather having an influence on spread of infection. The takeaway being to know the risks associated with each season statistically and to adjust your behavior accordingly. In a COVID world, we can no longer continue to oversimplify our beliefs about seasonal peaks in illness as expected and unavoidable.

References

What's the link between cold weather and the common cold?
https://www.medicalnewstoday.com/articles/323431#prevention

Data Sources & Descriptions

'COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University' sourced through Kaggle dataset 'COVID-19 data from John Hopkins University'

Link: https://www.kaggle.com/antgoldbloom/covid19-data-from-john-hopkins-university/activity

Documentation: Original repo - https://github.com/CSSEGISandData/COVID-19

Licensed under: Attribution 4.0 International (CC BY 4.0)

Contains: Confirmed cases and deaths by US county daily, both as cumulative and new cases per day.

Quirks: Negative new cases (or decreasing cumulative values) are a known issue documented in the discussions on kaggle. Within my county only one day contained a negative value of new cases (September 3rd, 2020). However if we are right to interpret the discrepancy as a retroactive correction being applied then on a day when there were sufficient cases to exceed the correction it could still be positive.

U.S. State and Territorial Public Mask Mandates From April 10, 2020 through August 15, 2021 by County by Day

State and territorial executive orders, administrative orders, resolutions, and proclamations are collected from government websites and cataloged and coded using Microsoft Excel by one coder with one or more additional coders conducting quality assurance. These data are derived from publicly available state and territorial executive orders, administrative orders, resolutions, and proclamations ("orders") for COVID-19 that expressly require individuals to wear masks in public found by the CDC, et al. [Excerpt from - https://data.cdc.gov/Policy-Surveillance/U-S-State-and-Territorial-Public-Mask-Mandates-Fro/62d6-pm5i]

Link:
https://data.cdc.gov/Policy-Surveillance/U-S-State-and-Territorial-Public-Mask-Mandates-Fro/62d6-pm5i

Documentation:
https://data.cdc.gov/Policy-Surveillance/U-S-State-and-Territorial-Public-Mask-Mandates-Fro/62d6-pm5i

Licensed under: Public domain

Contains: State and county identifiers, date, binary field for masks required, source, url, and citation for the order.

Quirks: Binary field Face_Masks_Required_in_Public equal to 'Yes' is defined by potentially multiple policies. Defined as 'a requirement for individuals operating in a personal capacity to wear masks 1) anywhere outside their homes or 2) both in retail businesses and in restaurants/food establishments.' More explicit details are available via url to the corresponding order.

Masking survey by "The New York Times and Dynata"

This data comes from a large number of interviews conducted online by the global data and survey firm Dynata at the request of The New York Times. The firm asked a question about mask use to obtain 250,000 survey responses between July 2 and July 14, enough data to provide estimates more detailed than the state level. (Several states have imposed new mask requirements since the completion of these interviews.)

Specifically, each participant was asked: *How often do you wear a mask in public when you expect to be within six feet of another person?*

This survey was conducted a single time, and at this point we have no plans to update the data or conduct the survey again. [Excerpt from - https://github.com/nytimes/covid-19-data/tree/master/mask-use]

Link: https://github.com/nytimes/covid-19-data/tree/master/mask-use

Documentation: https://github.com/nytimes/covid-19-data/tree/master/mask-use

Licensed under: https://github.com/nytimes/covid-19-data/blob/master/LICENSE

Contains: Proportion of each county that is estimated to wear their mask "in public when [they] expect to be within six feet of another person" either never, rarely, sometimes, frequently, or always.

Quirks: Snapshot data, uses weighted proportions of 200 nearest respondents to aggregate to census tracts and then weights by population to create county level data.

Global Historical Climatology Network - WBAN:94746 Station

The Global Historical Climatology Network daily (GHCNd) is an integrated database of daily climate summaries from land surface stations across the globe. GHCNd is made up of daily climate records from numerous sources that have been integrated and subjected to a common suite of quality assurance reviews.

GHCNd contains records from more than 100,000 stations in 180 countries and territories. NCEI provides numerous daily variables, including maximum and minimum temperature, total daily precipitation, snowfall, and snow depth. About half the stations only report precipitation. Both record length and period of record vary by station and cover intervals ranging from less than a year to more than 175 years. [Excerpt from - https://www.ncei.noaa.gov/products/land-based-station/global-historical-climatology-network-daily]

Link: Climate Data Online: Dataset Discovery - https://www.ncdc.noaa.gov/cdo-web/datasets

Climate Data Search Online - https://www.ncdc.noaa.gov/cdo-web/search

(Listed under 'Daily Summaries' in both sources)

Documentation:
https://www1.ncdc.noaa.gov/pub/data/cdo/documentation/GHCND_documentation.pdf

Licensed under: Public domain, NOAA's data is "available to the public without restriction on use"

Contains: See description above

Quirks: Data availability day-to-day and station-to-station is unpredictable and must be carefully assessed when selecting which source(s) and fields to use.

Local Climatological Data - WBAN:94746 Station

Local Climatological Data (LCD) consist of hourly, daily, and monthly summaries for approximately 950 U.S. Automated Surface Observing System (ASOS) stations, as well as observations collected every 20 minutes from around 1,400 U.S. Automated Weather Observing System (AWOS) stations. Data is also available for U.S. Climate Reference Network (USCRN) stations and a small number of stations at international U.S. facilities. [Excerpt from - https://www.ncei.noaa.gov/products/land-based-station/local-climatological-data]

Link: https://www.ncdc.noaa.gov/cdo-web/datatools/lcd

Documentation:
https://www1.ncdc.noaa.gov/pub/data/cdo/documentation/LCD_documentation.pdf

Licensed under: Public domain, NOAA's data is "available to the public without restriction on use"

Contains: Measurements on precipitation, wind, pressure, humidity, snowfall, visibility, etc. aggregated to the hour, day, and month.

Quirks: Alphanumeric flags may be included with or instead of measurement values designating potential issues or special cases. Therefore particular cleaning is required for these fields to overcome data type issues in pandas. Data availability day-to-day and station-to-station is unpredictable and must be carefully assessed when selecting which source(s) and fields to use.