Castle Leonard
November 10th, 2021
Analysis Extension
DATA 512: Human-centered Data Science

Motivation

From the A4 analysis, I came to a hypothesis that the strictness of adherence and regularity of compliance with COVID-19 precautions has more to do with social forces and human tendencies than state-wide orders. And that if anything, prevailing public opinion may induce masking mandates more so than the reverse. In particular, I wondered how much of the spike in new cases during the winter of 2020 into 2021 was the result of the weather changing, as is often blamed, versus the social dynamics among close friends and families that may encourage lax attitudes around masking and other precautions. So for my extended analysis, I'd like to identify to what extent the weather explains the increase in infection rate of COVID-19 in Worcester, Massachusetts over time.

Understanding the circumstances under which people increasingly spread infectious disease is as human-centered a problem as there is as it attempts to understand the behaviors and choices of people and can inform new guidance and perhaps enable people to make different choices for the betterment of public health if they're more aware of the pitfalls. In this way the analysis can attempt to work with people for the benefit.

Research questions & hypotheses

I've always heard that diseases spread more rapidly from person-to-person in winter because the poor weather keeps people indoors and in closer proximity to others where air circulation is reduced. It's worth noting that this pandemic is a bit different than a standard flu season in terms of awareness and extent of precautions. However, I'm still curious to learn to what extent the weather contributes to increases in infection rate or if there are any clear patterns between aspects of the weather and new confirmed cases after some interval of delay.

My hypothesis is that increasingly uncomfortable weather (temperatures beyond some threshold of discomfort or other impactful weather events, such as forms of precipitation) will correlate with a bell-shaped curve of infection rate. I am curious to learn the thresholds of weather that lead to increased or decreased infection rate.

Data

I intend to join the daily confirmed cases data from the Kaggle repository of John Hopkins University COVID-19 data (RAW_us_confirmed_cases.csv) that we used in A4 with a Local Climatological Data (LCD) from the Climate Data online data tool by the National Oceanic

and Atmospheric Administration (NOAA) (https://www.ncdc.noaa.gov/cdo-web/datatools). The data order I submitted was from January 15, 2020 to November 8th, 2021 for each of the two weather stations in Worcester county, the first: WORCESTER, MA US, and the second: FITCHBURG MUNICIPAL AIRPORT, MA US. The two stations report hourly readings on pressure, humidity, temperature, and precipitation. The data is not listed under a license as you'd see by companies or academic institutions, but rather states that "Climate Data Online (CDO) provides free access to NCDC's archive" (https://www.ncdc.noaa.gov/cdo-web/) and may be governed for commercial use under Resolution 40 (https://community.wmo.int/resolution-40). Once aggregated to daily metrics this data will easily join with daily confirmed cases data and potentially shed light on the interactions between the two.

Unknowns and dependencies

It is not clear if this level of granularity of weather will be sufficiently nuanced to bear out the interaction at a whole county level or if the weather across the two stations will vary much.

Additionally the relationship between weather and human behavior, let alone virus behavior, is likely complex and will require building out a variety of features to determine which are effective predictors.

Methodology

I will engineer many features surrounding min and max temperature with rolling averages, difference of temperature in the day, change from across periods, amount of precipitation, etc. and use that to train a multivariate regression model for infection rate at different intervals of delay. Given the complexity of the problem, I do not anticipate an extremely good predictor, but more so an idea of 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.

Timeline

Stage	**Start by**	**Complete by**
Prep Data	Nov 11	Nov 16
Feature Engineering	Nov 14	Nov 18
Model Interactions	Nov 19	Nov 24
Analyze, Visualize Results	Nov 25	Nov 27
Revise Analysis, if needed	Nov 28	Nov 30

Document Process	Dec 1	Dec 3
PetchaKutcha Presentation	Dec 4	Dec 7
Final Report	Dec 8	Dec 14