Stacey Wheeler
DATA512_A5
Due: November 10, 2021

A5: Extension Plan

1. **Motivation/problem statement:** *Why are you planning to do this analysis? Why is it potentially interesting and useful, from a scientific or practical perspective? What do you hope to learn? Your motivation statement should explain how/why the problem is human centered. That most of the COVID-19 data reflect people as data points does not make a problem immediately human centered.*

   Seasonal influenza continues to infect thousands of people every year. For many, this preventable illness is deadly. In order to control the spread of COVID-19, most state and local governments in the US put in place restrictions on gatherings and public spaces, enforced masking, and eventually encouraged residents to get vaccinated against the virus. One of the positive effects of these restrictions was that in trying to control the spread of COVID-19, they also minimized the spread of the seasonal flu.

   Before the pandemic, the idea of restricting people's access to public spaces and engagement with the economy would have been considered a ridiculous proposition. But COVID-19 forced us to go to extreme measures to keep ourselves and others safe. Now we have the chance to see how these restrictions changed the landscape for another common and preventable disease by comparing the rates of flu transmission during times with high restrictions to times where no such restrictions were in place.

2. **Research questions and/or hypotheses:** *These describe what you hope to discover or determine. Keep in mind this should be related to the Common Analysis that you conducted in A4. There are lots of possible questions related to the pandemic, and the A4 data set has day-by-day data for your county. For example you might ask:*
   - *Example research question: How was the local economy influenced by the pandemic?*
   - *Example hypothesis: Every 10% change in confirmed COVID cases resulted in a 5% decrease in employment for service sector workers in <CITY>.*

   I am interested in learning how the COVID-19 pandemic impacted transmission of influenza. It makes sense that the measures put in place to control the spread of COVID-19 were also effective at mitigating transmission of influenza. My hypothesis is that when looking at the weekly incidence of influenza during the '19-'20 and '20-'21 flu seasons, we will see at least a 90% decrease from the two seasons prior.

3. **Data used**: *What additional data do you plan to use, and why? Summarize what is represented in the dataset; Link to the dataset, and specify license/terms of use; Highlight any possible ethical considerations to using this dataset (and say why or why not). Briefly explain how this data will allow you to answer your question and how it expands on the Common Analysis.*

The CDC publishes data about the flu every season, which I can download from the National, Regional, and State Level Outpatient Illness and Viral Surveillance dashboard. I am able to download a csv of weekly data for the state of Michigan containing number of specimens tested, number of specimens positive for influenza, and percent positive by influenza type. The Weekly US Influenza Surveillance Report contains summary level information about influenza around the US. I will also use the datasets from A4 about confirmed COVID cases and mask requirements. I also found a report that summarizes COVID related restrictions in Michigan over the duration of the pandemic. The report is not in tabular form, but I can use it to code the time frames I am interested in comparing to understand how COVID precautions prevented the spread of the flu.

The flu data sourced from the CDC is free and available for use as a public service.

4. **Unknowns and dependencies**: *Are there any factors outside of your control that might impact your ability to address the supplementary questions? The purpose of this section is to get you thinking, in a practical sense, about your ability to answer the supplementary research question in the time allotted.*

   The main challenge with addressing this question is regarding granularity of data. The flu data is available only on a state level, but the COVID case data and mask data is broken down by county. For COVID cases, I can aggregate by state instead of county, but the mask mandates differ by county. The report about state-level lockdown restrictions will be helpful to codify requirements that were implemented for the entire state, and hopefully Michigan used a tier structure similar to what we experienced in Washington which will make the codification much simpler.

5. **Methodology**: *Describe how you plan to investigate this phenomenon. Don't just describe what your analytical methods are (e.g. "ordinary least squares", "student's t-test", "heatmap visualization", or "recurrent neural network"), it's critical to justify why these are appropriate methods for gathering and analyzing your data, or presenting your findings. You are expected to be thorough here: please describe to the best of your ability the entire series of gathering, analysis, and presentation methods you plan to use.*

   I plan to plot the rate of flu transmission on a time series plot from 2017 through 2021. Hopefully this graph will reveal some obvious changes in flu transmission once COVID restrictions went into place. I want to see how similar transmission rates were during the ‘17-’18 and ‘18-’19 seasons, and average them together to predict how ‘19-’20 and ‘20-’21 would have gone had it not been for COVID restrictions. I can then use a t-test to see if the actual transmission rate during COVID seasons was within a 95% confidence interval of the expected value. If it outside this interval, I can be confident that COVID restrictions were effective at controlling transmission of the flu.

6. **Timeline to completion**: *You should list each of the milestones and significant tasks you will need to complete the extended analysis. You might need to: collect data, build a model, test or analyze the model, visualize the results, and document your process. Think of the tasks that you need to complete. You should then provide a timeline that would allow you to complete the analysis on deadline. Keep in mind that you have a presentation (Assignment A6) and final report (Assignment A7) to complete as well. You'll need some time for those activities.*

   I will complete the bulleted tasks ON/BEFORE the date given (i.e. the first 3 bullets will be completed by 11/19)

   Friday November 19:

   - Ingest flu data, COVID data, mask data
   - Filter COVID/mask data to MI only and aggregate to get weekly values instead of daily
   - Join COVID and mask data to flu data on week#

   Friday November 26:

   - Codify the weekly data with COVID restrictions that were in place
   - Create time series for COVID and FLU, color coded by COVID restriction and/or mask rules in place

   Tuesday November 30:

   - Perform t-test on flu rate during COVID years compared to non-COVID years

   Tuesday December 7: A6 Presentation DUE

   - Create presentation

   Tuesday December 14: A7 Report DUE

   - Write report