Final Project Proposal (Part 1)

Myranda Swartzwelter

2/13/2022

Introduction

The Coronavirus pandemic has impacted everyone around the world in some way or another. For most of us, there are obvious changes to daily life and routine – wearing a mask, social distancing, consistently getting tested, etc. but there have also been large changes to the way we live our lives. For example, many employers went remote, or students turned to online school. In order to understand how much life changed for Americans during the pandemic, the Census Bureau launched the Household Pulse survey. I'd like to combine the data from this survey, vaccination rates, and Coronavirus case, death and recovery counts to understand how surges of cases or vaccination rates in certain states changed the way Americans lived.

Research questions

1. For Colorado, when Coronavirus cases increased, did more Coloradans report a change in post-Secondary education plans?
2. In the United States, did the number of households that reported doing online schooling increase with case counts and decrease with vaccinations?
3. Is there a relationship between case counts and American's reporting being more social than previous weeks?
4. What is the relationship between workers reporting working onsite, the report of increased case counts and vaccination rates? Do they vary by state?
5. Can we predict the number of cases in future weeks in a metro area based on the reported answers to education and work related questions on the survey in that metro area?
    - Is there another variable in the survey that works better for prediction?

Approach

In order to answer the above questions, I plan on segmenting the data out by date and time period for each data set, specifically selecting the large metro areas well represented in the surveys. Then I'll join the data sets based on the location where the vaccines or cases were reported, and the areas from where the survey respondents resided. Understanding that there is likely a delay in the count of cases and survey answers, I'll have to determine what the best time period is to account for that delay. Then I'll look for a correlation between the various survey answers and vaccination rates and the case counts for a given area for the time period. Once I determine what may have a strong correlation and determine whether that is consistent across the time periods, I'll do a regression analysis to determine whether we can predict future case counts based on the answers to the questions on the survey or the vaccination rates or predict answers to the survey based on the number of case counts in an area.

How your approach addresses (fully or partially) the problem.

Coronavirus cases and restrictions are a bit like the chicken and the egg scenario - we may see less restrictions in the survey answers which leads to an increase in cases later. An increase in cases may then cause more restrictions in future survey answers. I believe my approach will allow me to measure the various 'levers' a metro area or state can 'pull' to decrease their case counts, and understand what's better – to encourage employees to work from home, or to encourage vaccinations? Understanding these relationships has been a question from public health officials and the public for the past 2 years, so while I don't believe that I'll fully be able to answer all of the questions, I'm hoping to get some insights into the relationships in question. I think using the correlation and regression analyses will provide these insights.

Data (Minimum of 3 Datasets - but no requirement on number of fields or rows)

1. Public access files for the Census Bureau's Household Pulse survey
2. NYT US State Covid data
3. Vaccination data through the CDC's Vaccination Data Tracker

Required Packages

Possible required packages:

- Dplyr for manipulating data
- GGM for correlation analysis
- Ggplot2 for creating visualizations
- Purrr for functional programing
- Stringr for data cleaning and manipulation of strings

Plots and Table Needs

- A summary of each dataset and variables included in analysis
- A correlation table
- A plot of vaccine data
- A plot of Coronavirus case count data
- Table Summaries of the regression models
- A plot of the residuals
- A plot of predicted value vs true value

Questions for future steps

Once this work has been completed, it will be interesting to see how states and regions compare to each other. I think there's a possibility that the relationship between survey answers, case counts, and vaccine rates will vary because of the variability of the cultural makeup of the regions and or states. I've heard it said colloquially that the biggest determination of vaccine rates is political makeup of an area, and I'd be curious to see whether that is true or not. I'd also be interested to know if there are trends or predictors other than region or politics – potentially urban vs. rural is a better predictor. The final question I would propose is whether this is also true when compared across countries – how similar is the US to other countries in this regard, and what does that say about other cultural differences?