Final Project Part 3

Myranda Swartzwelter

3/1/2022

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. Additionally, the Center for Disease Control has been tracking vaccine rates across America, and the New York Times has been tracking COVID-19 case rates. For my final project in Statistics for Data Science, I would like to investigate these datasets to see if there are trends in the data between vaccine rates, Coronavirus case rates, vaccine hesitancy among Americans, and employment, spending and education concerns.

Problem Statement

The Census Bureau's household data set was very large, so I picked specific variables around vaccine hesitancy, employement, education and spending in order to see how case rates and vaccine rates were impacted by state, and which, if any, of the variables were strongly correlated. By understanding this, we could recommend a path forward to decrease COVID-19 cases and increase vaccine rates.

How I addressed the Problem Statement

In order to address the problem at hand, I wanted to do a multivariate regression analysis. To begin to address the problem, I looked at the Household dataset and read through the data dictionary to determine which variables I thought were interesting and would fit my analysis well. (This will come up again in the limitations portion.)

After deciding which variables to analyze, I had to clean and assemble the data into a useful dataset. The Household Pulse survey was aggregated in a weekly or biweekly basis depending on when the survey was taken. It is also very large, and contains a variety of datatypes. I therefore had to re-aggregate the data into a useable dataset that had the timeframe and variable types I was looking for. Both the NYT Covid case number dataset and the CDC Vaccine dataset are aggregated on a daily basis, by U.S. state. Due to the difference in aggregations between the three data sets and the inconsistent timing of the survey, I decided to do aggregate the responses, vaccine, and case data by month and U.S. State. Finally, the NYT Covid file had full state names whereas the Vaccine dataset and survey dataset had state abbreviations, so the states had to be mapped correctly.

After cleaning and aggregating the data, I would do an analysis which I'll discuss in the next section.

Analysis

For my analysis, I wanted to answer several questions surrounding COVID-19 case and vaccination rates. To begin with, I would start by trying to answer the question: How do survey question answers relate to COVID-19 case counts in America? To answer this, I would begin by looking at the covariances and correlations between my variables like demographic information, in-person employment rates, number of

children enrolled in school, and rate of spending from the survey, and COVID-19 cases. I would perform a multivariate regression using the highest correlated variables. Using this knowledge, I would apply it to later surveys from the Household Pulse survey to see if the regression analysis was a good predictor for COVID-19 case counts.

After this analysis, I would preform a similar analysis for Vaccine rates to answer the question: How do survey question answers relate to COVID-19 vaccine rates in America? I would also preform a single variable regression to see what the relationship between COVID-19 case rates and COVID-19 vaccination rates to see if this was a better predictor than the survey answers. Finally, I would bring COVID-19 case rates into the initial 2 analyses to see if combining the data gave a better predictor than survey data alone.

Within the survey, there are reasons for vaccine hesitancy. Because these are categorical variables, I think an interesting analysis would be to use the demographic variables included within the survey to do a clustering analysis to see if there are trends between vaccine hesitancy reasons and demograpgics.

Implications

The U.S. and countries around the world have been trying to end the Coronavirus Pandemic since 2020. In order to do this, they need to decrease COVID-19 case counts, and the strategy to do this for many countries has been to increase vaccine rates. By understanding which factors may contribute to low case rates, we may be able to employ several strategies. An example would be if this analysis found that states that had lower in-person employment rates also had lower COVID-19 case rates, the government could recommend that companies allow all employers that are able to allow their employees to work from home to do so, and have it be statistically backed. Similarly, we would likely find that higher vaccine hesitancy is correlated with a lower vaccine rate within a state. If this is the case, state officials with high vaccine hesitancy could investigate those reasons and create a solution to reduce vaccine hesitancy. The final cluster analysis I proposed could be useful to identify why a perosn has vaccine hesitancy, and a strategy could be created around those reasonings. For example, if we found that low income populations had a high rate of vaccine hesitancy due to inability to locate a vaccine (one of the survey answers), a better strategy of vaccine deployment could be created. If instead, there was a high rate of vaccine hesitancy in the caucasian population due to a lack of education, a state could introduce strategies to provide that education for these communities.

Limitations

One large limitation of the Household Pulse Survey is that it is not a longitudinal study of the same respondants over time. This means that although we can create a sample and see if it is statistically significant, we can't necessarily due a survival analysis. I think an interesting analysis would be to do a survival analysis on factors from variables in the survey and when/if the respondant got COVID-19 (another survey answer). If we had the same respondants over time, we could track their answers to the question "Did you receive a positive COVID-19 test result in the last 2 weeks?" as the event (answering "Yes" to this question) and then cohort the respondants by different variables such as demographic information, employment information, student status etc. to understand which factors are higher risk for contracting COVID-19.

The other limitation of this dataset is the timing of the survey. Because it changes over the course of the survey, I had to aggregate it by month, but I think aggregating on a weekly scale would lead to better results.

There are also limitations in the data itself, especially in the NYT COVID-19 cases. Because of the nature of the availbility of testing at the beginning of the pandemic, likely this dataset undercounts case rates. Additionally, as the pandemic has gone on, and home tests are available, people have been relying less on the tests in labs that are reported to the NYT and more on home tests which are not reported to the NYT. As policy around testing has changed and test availability has changed, it's hard to udnerstand and predict just how much the dataset is undercounting COVID-19 cases by. Because of this, hospitalization data may be a better metric to use than COVID-19 case rates.

Finally, COVID-19 case and vaccine rates are a complicated subject, and it is difficult to predict a virus' behavior from survey results along. Even human behavior on whether someone will get the vaccine or not is not necessarily straightforward just by knowing their answers to a survey. Although this analysis may give us

a good idea of indicators of COVID-19 case rates and factors that correlate with vaccine rates, I don't believe it would ever be able to perfectly predict case or vaccine rates.

Final Remarks

The COVID-19 pandemic is an interesting subject not just in terms of the data available, but how much of an impact it had on the daily lives of Americans and people around the world. These datasets offer an insight into that impact and could provide leaders with strategies for ending the COVID-19 pandemic by increasing vaccination rates and decreasing case counts. Data Science is a powerful tool, and I am looking forward to the future to see all of the ways it will be used to inform world leaders on how to act. However, as with this study, it will be important to remember that while data can tell a really good story and act as a great informant to make decisions, leaders will have to balance that with the diversity of the world we live in and make ethical decisions keeping their people's best interest in mind.