Final Project Part 2

Myranda Swartzwelter

2/20/2022

Introduction

This final project is on the use of the Census Bureau's Household Pulse Survey and Center for Disease Controll's COVID-19 data could be used to understand COVID-19 case rates (data from the New York Times).

The Household Pulse Survey dataset is a collection of survey responses to a survey from the U.S. Census Bureau designed to understand how people's daily lives were impacted by the COVID-19 Pandemic.

How to Import and Clean Data

There are many nuances to the Household Pulse Survey datasets. To start, each file available in the datasets is organized by survey week and contains upwards of 75,000 individual responses. I don't want to aggregate on an individual survey level, so I need the data aggregated into counts of responses by U.S. state and metropolitan area (the results are from the largest 15 metro areas in the U.S.). Luckily the U.S. Census Bureau already does this, and the data is available to download. However, for reach response there are 182 variables across several different categories. Instead of trying to do an analysis over every variable, I'm going to focus on a few key categories: education, employment, and vaccine hesitancy.

Week over week, there are also different numbers of respondants, so rather than aggregating the data by raw counts, I'll aggregate to percentage of respondants for each category.

Another nuance of the Household Pulse Survey is the timing at which the survey took place and how it changed. The survey implementation was split into 3 phases. When started in April 2020, during the first phase, the survey was collected and disseminated on a weekly basis. There is then a gap in data from July 21, 2020 through August 19, 2020 between when phase 1 ended and phase 2 started. For phase 2, the data was collected and disseminated on a two-week basis. There is another gap in data from March 30, 2021 - April 13, 2021 and again from July 6, 2021 - July 21, 2021. The implentation again changed for phase 3.3 and the Census Bureau took a two week on two week off approach starting in December 2021. There is a gap in the data between phases 3.2 and 3.3 between October 12, 2021 and December 1, 2021.

Both the NYT Covid case number dataset and the CDC Vaccine dataset are aggregated on a daily, by U.S. state. Due to the difference in aggregations between the three data sets and the inconsistent timing of the survey, I'm going to do aggregates of the responses, vaccine, and case data by month and U.S. State.

Additionally, the NYT Covid file has full state names whereas the Vaccine dataset and survey dataset have state abbreviations, so the states will have to be mapped.

What does the final data set look like?

Data Dictionary

Below is an example of the final data set for the state of Alabama. The variables are as follows:

- Date: 1st date of month of data
- State: 2 letter abbreviation of state

- Covid Cases: New Covid Cases this week
- Vaccines doses: Vaccine doses administered
- vac2: Percent that self-reported getting the vaccine
- pct_dnt_ned_vac: Percent of respondants who have not recieved the vaccine that believe they don't need the vaccine
- pct_concrnd_se: Percent of respondants who have not recieved the vaccine that are concerned about side effects
- pct_wait_and_see: Percent of respondants who have not recieved the vaccine that are planning on 'Waiting and Seeing'
- pct_wrk_pst_7_days: Percent of respondants who have worked in the past 7 days
- pct_concerned_covid: Percent of respondants who are concerned about COVID-19
- pct_laid_off: Percent of respondants who have not worked in the past 7 days that were laid off due to the Coronavirus pandemic
- pct_wrk_vlntr_outside_home: Percent of respondants who have worked or volunteered outside of their home in the last 7 days
- pct_no_childcare: Percent of respondants who need childcare who did not have it in the past 7 days.
- pct_spn_7_days: Percent of respondants who reported having at least some difficulty with household spending in the past 7 days.
- pct_ed_change: Percent of respondants who had plans for secondary education and changed them

Table 1: Sample of Final Data Set

Date	State	covid_cases	vaccine_doses	vac2	pct_dnt_ned_vac	pct_concrnd_se	pct_wait_and_see	pct_wrk_pst_7_days	pct_concerned_covid	pct_laid_off	pct_wrk_vlntr_outside_home	pct_no_childcare	pct_spn_7_days	pct_ed_change
4/1/20	IL	143	0	0	0	0	0	0.763	0.689	0.021	0.325	0.445	0.540	0.230
5/1/20	IL	219	0	0	0	0	0	0.523	0.721	0.045	0.467	0.473	0.565	0.368
6/1/20	IL	114	0	0	0	0	0	0.586	0.632	0.078	0.540	0.426	0.582	0.353
7/1/20	IL	698	0	0	0	0	0	0.524	0.540	0.091	0.565	0.247	0.457	0.396
8/1/20	IL	2651	0	0	0	0	0	0.593	0.565	0.247	0.582	0.327	0.498	0.632
9/1/20	IL	2284	0	0	0	0	0	0.543	0.582	0.327	0.457	0.385	0.437	0.697

Questions for Future Steps

Now that we have the data in a final format, there are several questions to inform more about the relationship. * How do survey question answers relate to COVID-19 case counts? + Are there any questions that help predict case counts? * How do survey question answers relate to COVID-19 vaccination rates? + Are there any questions that help predict COVID-19 vaccination rates? * How do COVID-19 case counts and vaccination rates relate to each other? * How do these relationships vary by state? * How do these relationships vary by region?

What information is not self-evident?

It is not immediately self evident if any of the answers are related to case counts or vacination rates. It is also no self evident if the case counts and vaccination rates are related to each other.

What are different ways you could look at this data? How do you plan to slice and dice the data?

We can look at this data as an aggregate of the U.S, or split apart by state or region. The Household Pulse survey contains demographic information, and so does the vaccination data, so if we were to re-design our final data set, we could also break apart these analyses by age, race/ethnicity, etc. However, for the first analysis, I plan to look at the data by question or question category, and by U.S. state. An interesting data set that you could join to the final data set is general party affiliation - if the state tends to go "Red" or "Blue".

How could you summarize your data to answer key questions?

I can use regression techniques to understand what the relationships between the variables are. Then once I've done the analysis for the questions around survey and vaccination or COVID-19 case counts by state, I'll have to run another analysis to understand how these trends vary by state.

What types of plots and tables will help you to illustrate the findings to your questions?

- A summary of the final dataset and variables included in analysis
- A correlation table between variables
- A plot of vaccine data by state
- A plot of Coronavirus case count data by state
- Table Summaries of the regression models
- Plots of the residuals
- Plots of predicted value vs true values to show quality of models

Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.

Yes, I plan on using regression models to understand the relationship between the variables and then use those models to try to predict future case counts and vaccine rates.

Questions for future steps.

After these questions, I think it's be interesting to investigate the demographic information with the survey and see how demographics can influence case counts, vaccine rates, or survey answers. Additionally I think understanding how political affiliation impacts survey answers, case counts or vaccine rates would be an interesting question for future steps.