Alexander Adams

PPOL683 GIS in R

Fall 2021

Final Project Summary Document

Introduction

One of the most significant challenges of the COVID-19 pandemic has been how authority figures communicate information about the virus, its spread, and treatment and mitigation measures to the general public. While this issue is present in every public health crisis, in the context of the COVID-19 pandemic, the contours of this communication are shaped by certain novel factors. First, Americans, particularly younger Americans, are increasingly receiving significant amounts of their news information from social media sources. Second, national public health figures have become polarized and are distrusted by large sections of the population. Third, in an era where anyone can become a social media influencer, guidance from individual messengers is less effective than guidance from organizations.

Local public health department social media accounts present a possible solution to this issue. Social media channels may be more effective than traditional media and news sources at reaching underserved demographics. The fact that these departments operate at the local, community level, rather than at the state or federal levels, means that they are more immediately accessible and relevant to the average American. As organized government departments, they potentially have more credibility than singular health officers.

Thus the central question of my project is as follows: is there an identifiable difference with respect to COVID-19 outcomes between counties whose public health departments have social media accounts on major platforms (specifically Facebook and Twitter) and counties whose health departments do not have social media accounts?

Methods and Data

The data I used in this project came from multiple sources. The New York Times hosts a public GitHub repository with historical data on COVID-19 cases and deaths at the state and county levels. The Centers for Disease Control and Prevention publish a data set of county-level vaccination metrics, including data on completed vaccination sequences (two doses of an mRNA vaccine, or one dose of the Johnson and Johnson vaccine) and first doses administered. I used county-level race and population data from the 2020 U.S. census, available through the U.S. Census Bureau. For the main focus of this project, county-level health department social media accounts, I gathered the data manually through searches on Facebook and Twitter. I was then able to use the `twint` module for Python to scrape tweets from many (though not all) of the accounts. I was able to find Twitter accounts for approximately 875 counties, and Facebook accounts for approximately 2100.

The shapefile I used in this project was a U.S. counties shapefile, accessible through the `tigris` package. I modified the file using the `st\_union` function from the `sf` package to generate polygons of 87 multi-county health departments across different states, made up synthetic FIPS codes for those polygons, and then bound the rows to the existing shapefile. I then used the `ms\_simplify` function from the `rmapshaper` package to reduce the size of the shapefile.

Additional Considerations

At least 475 counties and county-equivalent jurisdictions across 16 states are aggregated into 87 multi-county health districts. Many of these districts encompass counties which are highly rural; for example, the Panhandle Public Health District in Nebraska contains ten counties with a combined population of less than 50,000 people in the northwest part of the state. The individual counties within these larger districts might not have social media accounts, but many of the overall districts have them. For the purposes of this project, those accounts were treated as belonging to both the larger health district and each constituent county within it. Additionally, I calculated aggregated statistics for each district (the total number of COVID-19 cases, deaths, and vaccinations across the member counties, as well as the total population of each district).

Findings

I initially calculated the Pearson’s *r* correlation coefficients for many of the variables in my data. While none of the correlations were particularly strong, there were negative correlations between the presence of a county health department Twitter account and the cumulative rate of COVID-19 cases and deaths in that district. There was also a minor positive correlation between the existence of a county health department Twitter account and the rate of completed vaccination sequences. After running a standard ordinary least squares multivariate regression, I found that this effect, while very small (less than 1%, closer to 0.5%), was statistically significant. I did not find a similar effect for the presence of a Facebook account; the correlation coefficients and model coefficients were indistinguishable from zero.

This effect becomes clearer when visualized. I included a line plot in my R Shiny application for this project which aggregates county COVID-19 case, death, and vaccination rates over time by the social media accounts of that county’s health department (e.g. Facebook only, Twitter only, both, or neither). Regardless of metric, the four lines follow the same general track. However, the average rates of cases and deaths for counties with Twitter accounts are visibly lower than the average rate of those with Facebook accounts or no social media. Counties which had both forms of social media had similar but slightly lower average COVID-19 case and death rates to those with only Twitter. Relative to counties with no health department social media, those with a Facebook page also had lower average case and death rates, though the difference was extremely small. The inverse is true for vaccination rates; counties with both Twitter and Facebook had higher rates of completed vaccinations and first doses administered than those with just Twitter or Facebook, and having one form of social media resulted in higher vaccination rates than having neither.

However, when I incorporated data on racial demographics into the multivariate regression models, the effects of having a Twitter account lost their statistical significance and became attenuated toward zero. This held regardless of the COVID-19 metric used as the dependent variable. The R2 values returned by each regression model were all smaller than 0.2, indicating that racial demographics and health department social media accounts collectively explain less than 20% of the variation in COVID-19 case, death, vaccination, and first dose rates. The R2 values when racial data was excluded were less than 0.05, meaning that the racial makeup of a county is a much more effective predictor of COVID-19 outcomes than whether or not that county’s health department has a social media account.

Conclusions

Based on the regression models and visualizations I produced during this project, I do not believe that county health department social media accounts have significant impacts on notable metrics of COVID-19. While the existence of a social media account may be associated with a positive effect on case rates, death rates, and measures of vaccination status, the fact that this effect disappears when additional controls are added (even if those controls are imperfect) suggests that health department social media presence is serving as a proxy for a different, unmeasured variable.

I believe that variable (or at least one potential variable) is political partisanship. A growing body of research in public health has found that jurisdictions with heavily Republican voting patterns have had higher rates of COVID-19 cases and deaths and lower rates of vaccination than Democratic areas. Republican elected officials have also refused to implement pandemic mitigation measures, and in many cases have actively blocked such measures from being implemented. I deliberately excluded partisanship from my analysis after reviewing the literature because I expected that it would overshadow any potential effects from health department social media, and I expect that controlling for partisanship would both zero out the effects of health department social media and increase the R2 value of the regression model significantly.

Ultimately, at this stage of investigation, I reject the current iteration of my hypothesis. Counties whose health departments have social media accounts do not experience statistically significant differences in COVID-19 outcomes from counties whose health departments do not have social media accounts, after controlling for the racial makeup of the county.

Limitations and Potential Future Steps

Due to time constraints, I was unable to incorporate data on specific social media metrics into my models or Shiny app. While I wrote an .R script to find various summary statistics for twitter engagement (mean number of likes, retweets, or replies, number of tweets since COVID-19 was officially declared a pandemic, etc.), those statistics were not used in this stage of this project, and as such present an opportunity for future research.

Another potential avenue of exploration is to collect and incorporate data on partisanship into my analysis. I deliberately excluded it from consideration here, but future projects could attempt to assess the nature of the relationship between past voting outcomes or voter registration counts and various dimensions of the COVID-19 pandemic.

A third limitation I encountered was more concrete: the app I wrote and submitted for this project is too large to be hosted on shinyapps.io with a free account. It runs locally without any issues, but with my current shinyapps.io account, I cannot deploy my work as an in-browser web app as it currently stands. I tried to reduce the app size, but was unable to shrink the app to be smaller than 1Gb. I believe this is due to the size of the data I used in the app, as well as the number and type of visualizations I included.