Final Project Step 3 - Eric Vance

Eric Vance

2/22/2022

Introduction

Given the state of the world today, I thought it would be important to analyze COVID-19 data, with the ultimate goal of determining what factors contributed to higher COVID cases and more importantly, the deaths associated with the virus. There has been a lot of talk about data ethics, methodologies, and results around the pandemic and it is important to cut through the misinformation and political discourse to find the true and unbiased results.

The data points that initially stood out to me were average age in the state, the health rankings of each state based on a variety of factors such as diet and exercise levels, and the vaccination rates of each state. These data points combined with a variety of others will help to identify what variables have the closest relationship with COVID deaths and hospitalizations.

Problem Statement

The problem that I chose to analyze boiled down to one concise thought would be: What are the largest factors contributing to COVID deaths, and what factors are largely inconsequential? Throughout the pandemic, there has been lots of talk about whether social distancing, mask-wearing, and other efforts has been an effective way of combating the virus. They are all important questions to ask – as these methods may be employed during future pandemics.

How I Addressed this Problem

I addressed this problem by first dropping any preconceived notion about how the virus spreads, who ultimately succumbs to it, and any other assumptions I had gained from two years of constantly hearing about it. I thought it was important to go into this experiment without any bias, so the bias would not leak into the experiment itself.

I then loaded the data into R-Studio and began some preliminary analysis. I looked at the heads of my datasets to see what the quality of the data looked like initially. I then checked for any null values and removed them – this decision was made by the fact that I still had plenty of the sample to work with after the null values were removed. I then decided whether or not I needed to join the datasets, do any mutations, or make add any custom variables. This was a long period of back and forth, because doing the actual analysis leads to realizations that certain things are needed or not needed, so this was a lengthy part of the process.

Analysis

I first took my data and joined everything together. This was to get the states, the date, which unfortunately was only as late as December 2020, and the count of hospitalizations, ICU stays, and deaths. All values were cumulative, so it was only important to have the final number. Any numbers from before would only have outdated data since everything was cumulative. The final COVID Death dataframe I ended up having looks like this:

```
##        date state positive hospitalizedCumulative inIcuCumulative
recovered
## 1  20201206    AK    35720                   799              NA
7165
## 2  20201206    AL   269877                 26331            2290
168387
## 3  20201206    AR   170924                  9401              NA
149490
## 4  20201206    AS        0                    NA              NA
NA
## 5  20201206    AZ   364276                 28248              NA
56382
## 6  20201206    CA  1341700                    NA              NA
NA
## 7  20201206    CO   260581                 14868              NA
13941
## 8  20201206    CT   127715                 12257              NA
9800
## 9  20201206    DC    23136                    NA              NA
16665
## 10 20201206    DE    39912                    NA              NA
18851
##    deathConfirmed
```

```
## 1              143
## 2             3462
## 3             2437
## 4               NA
## 5             6431
## 6               NA
## 7               NA
## 8             4143
## 9               NA
## 10             698
```

I then sorted each state by deaths to find that New York, California, Texas, and Florida had the highest number of deaths. My thought that population density contributes to higher numbers of deaths was put to the test here. I created a new variable based on the population of the state and the number of deaths that calculated the death rate. This was done simply by dividing the positive cases by the number of deaths. Florida and New York remained in the top, but California dropped out of the top in terms of death rate.

I then looked at the demographics dataset I downloaded from Kaggle for each state. This listed basic census information like age, race, occupation, and more. I plotted the ten states with the highest average age and noticed that both Florida and New York were in the top ten. This gave value to the idea that the higher the age, the more likely someone was to succumb to COVID-19.

To put this test to the theory, I ran a correlation table between the average of the state and the number of COVID deaths. It came back positively correlated between age and death, to the tune of a correlation coefficient of 0.62. While not perfect, it was still a positive correlation and a reasonably strong one at that. This confirms to me that age is a major factor in determining how someone will be affected by the virus.

So far, we have that age is a major factor, as well as the population density. Both of these were confirmed by the correlation tables that returned positive and strong correlations between COVID deaths and the aforementioned variables. Unfortunately, this is the latest data available and it preceded vaccines, so vaccination rate was not something that could be tested using the data.

Lastly, my final theory had to do with the average health of the people in the state vs the death rate in the state. I theorized that the healthier the state, the more they would be able to withstand the virus and thus have a lower death rate. I found a dataset that determined health by a variety of factors, the citizen's diets and exercise levels and how active they are in general. I then compared that to the death rate of each state. Sure enough, there were several states in the top 10 in the health rankings that were also in the bottom 10 of the

death rate. Hawaii, Vermont, and Puerto Rico were among them. However, correlation tests came back inconclusive, so it's tough to say for sure if this is statistically significant. This has concluded the analysis segment.

Implications

The implications of this test are important not only for now, but the future of pandemic research. If it can definitely be proven that population density, age, and health are three important factors when determining virus vulnerability, it may persuade people to change their lifestyles in case the next pandemic comes earlier than expected. On top of that, it may affect policy going forward in future pandemics. For example, if a pandemic began, it might be part of policy to separate older individuals from one another and from large population centers. This would prevent a repeat of what we saw at the beginning of COVID-19, where nursing homes and long term care facilities were devastated by the virus.

With the availability of data in today's modern age, it is crucial that lessons are learned from large-scale historical events such as the COVID-19 pandemic. While my analysis may just be a stepping stone to something larger, it is still important to isolate the factors that may help save people in the future from other viruses. Even if it is saving just 1% of people who have died, that's still in the tens of thousands of lives.

Limitations

Regrettably, there were several limitations when it came to this experiment and analysis. For one, the data I was able to get for free was relatively outdated. When it comes to real time pandemic statistics, obviously the more up to date the data is, the better. This data was from December 2020, which means it had about 9 months worth of initial data from the start of the pandemic in this country.

Because of the data being somewhat outdated, it meant that it was missing all sorts of vaccination information, since the vaccine came out after the final row of data. The vaccine signaled a massive shift in attitude towards the pandemic, as well as a large shift in who was more affected by the virus. For the sake of the experiment, it would have been nice to be able to filter populations based on vaccination status.

Lastly, my last limitation was myself, and my knowledge of the data science process. Having never run a larger-scale project like this one, I believe my inexperience and general lack of knowledge hindered the project a little bit. However, it was very important to my development that this project be done, as it helped to identify pain-points that I can work on mending in the future.

Concluding Remarks

In conclusion, this was a very fun, challenging, and interesting experiment. I think that I did a good job considering the limitations, and I would like to revisit this topic when more data comes out and I've honed my skills a little bit more. COVID-19 is a hot-button subject right now, as its dominated news cycles for the last two years. I think the analysis that I did could be useful for determining what to focus on to save as many lives as possible for when the next pandemic comes out. Though the scope of my analysis was limited, I think it would be a great stepping stone for a larger scale experiment. Thanks for reading, and for a great semester!