**The COVID Conundrum, An Attempt to Understanding COVID-19 Data Sources**

Group 11: Pankaj Tahiliani, Patricia Mobolade, Jennifer Darby, Tobias Judd

The COVID19 coronavirus outbreak was reported in December 2019 with it originating in Wuhan, China. This pandemic has brought together the world’s scientists, global health professional, business and political leaders, and data scientists to accelerate understanding and to help contain the pandemic. The outbreak has impacted the lives of billions of people and as of today the US has had 6.79M cases and 199K deaths. This has resulted in both the US Federal and State governments in implementing health policies and travel restrictions to contain the spread of the coronavirus, however, these actions have led to a disruption of everyday life while causing frustration and anxiety among citizens.

Group 11 collaboratively approached this topic and wanted to improve our own understanding of the subject and investigate the following issues:

1. What geographies are most affected by COVID Cases?
2. Does wearing a mask impact COVID numbers?
3. How does mobility correlate with the COVID rate over time?
4. Does median income impact the rate?
5. Integrate and leverage several data sources including New York Times and Dynata, Weather API, Apple Mobility, Google Maps API, and Census Data.

The project proposed and its outcome was based upon several different sources of information and represents a comprehensive attempt to understand the limitation of each dataset and merge these sources together to provide a better insight into the daily numbers. Given the unique challenges and time / resource limitations it was impossible to do an exhaustive search since there were data limitations, not all or some targeted data required a paid subscriptions, and or new data and formats are being added given the intense global focus on the subject matter. Our overall project approach included:

* Leverage the US Census Data that offered a wide range of public information. This information included population, demographics, income, and insight into work / commute habits pre-pandemic. Unfortunately, this information was last revised for 2018 and an update for 2019 is in process of being released. For either year the data is organized at a Zip Code level.
* COVID Cases are reported daily at a state and county level which can be linked to a Federal Information Processing Standard Codes (FIPS Code). A FIP Code consist of multiple Zip Codes and allowed us to link Census data to COVID Cases.
* A survey done by the New York Times in junction with Dynata asked people on their mask wearing habits. The firm asked a question about mask use to obtain 250,000 survey responses between July 2 and July 14 and was organized on a FIP Code level with daily results.
* The Weather API was used to used to collect the longitude and latitude coordinates for each primary key within the dataset. We additionally wanted to look at daily weather patterns per location of interest, however, the historical weather feature required a subscription service which prevented this analysis from being investigated further.
* An Apple Mobility API was used to gain insight in daily mobility behavior of mobile device owners using Bluetooth Technology.

Despite this data coming from different public domain sources, it was possible to combine the data in an efficient and logical manner. However, the “freshness” and up to date of each data set was something of concern given when a change occurred on a source versus when that change appeared in the dataset (i.e. Census data lag, none unique individual information). Likewise, the “granularity” of information was a challenge from the start since we are dealing with a pandemic that impacts individuals, however, reported data is captured at a much higher level (i.e. FIPS Level information for Covid Cases or Mask Wearing Information). The use of our project data resulted in the following analysis to be performed:

* Data analysis portion to make sure that we were capturing actual new cases versus a start of reporting which resulted in a different time basis. This resulted in us narrowing the investigation to focus on the time between May – September versus including when a global pandemic was declared in March.
* Heatmap generation of Top 100 Counties within the US and the marking of the Top 5 based on case rate of change.
* Statistical analysis and comparison within the dataset of behavioral and mobility factors between May 31st and September 8th.

The ability to understand Covid depends on the ability to understand a complex mechanism of factors, however, no dataset known to exist has addressed and captured all of the components related to the exposure, transmission, and infection process. Moreover, the variability in exposure and transmission within the US population has illustrated the susceptibility and multifactorial nature of the disease and its subsequent impact on the local population. General surrounding data limitation were:

1. Reporting over Time - The virus began to spread through the United States over time. The world was not ready for an event such as this and with it being a novel virus, tests did not exist. Once it was declared a pandemic, the reporting of the data and testing began to accelerate. Therefore, appropriate times were selected to eliminate sparse data density due to limited reporting versus rates.
2. Differing Geographies - This analysis was limited by the level of accounting for the variables. For instance, COVID data is reported for the United States only down to the county level, whereas the Census data for median income is reported to zip code level. FIPS and zip code tables, and sample datasets were used to align datasets.
3. Different Time Elements - The COVID data was reported daily for all counties in the United States and the Census Data is reported annually. The census data used was legacy data for 2018. The mobility data was a precursory data point for COVID rates, so adjusting the time periods for comparative analysis was needed. It was moved two weeks forward to align with rates reporting due to COVID incubation period.
4. Causal Elements - The granularity of the data prevented an in-depth investigation of real cause. Exposure would be a large contributing factor, but it is difficult to quantify. Understanding correlation is not necessarily causation, other impacts could have critical impacts to the data. For example, the analysis presented questions regarding the data distribution, exposure, population density and other causal elements.

In conclusion area for improvement include:

* The project study was impacted and influenced in the granularity of data, how the data is reported over time for each data source and differing of geographies and boundary areas
* There were multiple assumptions made in calculating the analysis, including the data was normally distributed, the data was independent, the data was homogenous, and the standard deviations were assumed equal.
* There was not enough evidence to prove that mask wearing patterns have impacted the transmission relationship between the month of March and September. (*See Appendix A for more details*)
* Unfortunately, mobility data did not correlate with COVID case rates given the granularity of information and the incubation period for the virus could be up to two weeks. (*See Appendix B for more details*)
* It was not possible to do a direct comparison of COVID Cases versus Median Income, since COVID data is reported for the United States down to the county level and Census data for median income is reported to zip code level.
* Finally, a sample dataset of cases in Florida was identified that reported cases down to the zip code. Florida was chosen as it had a large data density of 680,000 cases reported. The initial analysis was conducted using symmetrical bins based upon categorical distributions of median income in $10,000 increments. The analysis demonstrated a direct correlation between median income and rate visually on the box plots. As income increased in each boxplot, rate increased. (*See Appendix C for more details*)

In closing, this project has been very educational and beneficial for the team and has illustrated the challenges associated within the data analytic space. More importantly, if a thorough understanding is desired for qualifying Covid transmission a thorough study needs to be framed in which a complete population is monitored with data being collected daily at the individual level for an extended duration.

**Appendix A**

Based on the results obtained from the survey we were interested in finding if there were any consistencies in the mask wearing habits amongst the people in the top 100 counties (case-wise) in March and September 2020. An independent t-test was conducted with a random sample of 25 counties from these two populations.

H0: There is no change in mask wearing patterns in the two populations

HA : There is a change in the  mask wearing patterns in the two populations

The objective of this test was to check whether the mask wearing patterns influence the spread of covid-19. The statistical analysis on these two samples showed that 73% people in March and 70% people in September on average, answered as “always wearing a mask when expected to be within 6 feet of distance in public”. Similarly, 2% and 3% on average answered as “never wearing a mask when expected to be within 6 feet of distance in public” for the month of March and September, respectively.

In both cases the p-value was found to be greater than .05 which means we fail to reject the null hypothesis. In conclusion, there was not enough evidence to prove that the mask wearing patterns have changed in these two populations. As a result, it can be said that mask wearing patterns do play a role in the spread of covid based on this analysis, provided there are other potential factors that were not explored in this project.

**Appendix B**

**Appendix C**

Ho Income has no impact on COVID rates and cases

Ha The p-value of 0.13 demonstrates we cannot reject the null hypothesis

Understanding if income levels of the population impact the COVID rates were limited by accounting levels of the data categories down to differing geographic increments.

COVID data is reported for the United States down to the county level.

Census data for median income is reported to zip code level.

A sample dataset of cases in Florida was identified that reported cases down to the zip code. Florida was chosen as it had a large data density of 680,000 cases reported. The initial analysis was conducted using symmetrical bins based upon categorical distributions of median income in $10,000 increments. This distorted the distribution of data. A calculation was used to determine data quartiles. Then the the dataset was divided into appropriate bin increments.

The analysis demonstrated a direct correlation between median income and rate visually on the box plots. As income increased in each boxplot, rate increased.

More statistical tests were conducted to quantify correlation. The higher the income, the higher the rate represented by y=0.09x + 18,544. The Pearson correlation coefficient of 0.05 indicated a wide variation around the line of the best fit. With a large distribution of COVID rates between $20,000 and $40,000 median income range.