Brantley, Kawabata, Urgitus
CSC405 FA20
Project Stage I Report

Introduction

In this report we will be looking at the COVID-19 data for the year 2020 in the USA. We have three primary datasets which include the number of cases, deaths, and the total population for each county. Aside from population by county, each primary data set contains information for each day, starting from 01/22/2020, and ending on the most recent data upload. Using this data, we will be able to track COVID-19 on a granular level.

Preliminary Intuitions about Covid Data

We speculate that an increase of covid cases will lead to an increase in covid deaths. This increase in deaths will occur several weeks after the increase in cases since that's how long the virus takes to act. These spikes in death may become less pronounced as time goes on, due to advances in our understanding of the virus as well as our ability to treat patients. It's likely that there will be consistent rising and falling of new cases as the US is implementing a sort of 'touch and go' policy when it comes to reopening. That is to say we have been reacting to case-spikes by tightening restrictions and loosening them during case-lulls, which leads to another spike.

In the tables below we are displaying the variable dictionary for all the datasets involved in our project.

Primary COVID-19 Data

COVID Deaths

Variable Name	Data Type	Description
County FIPS	int	Federal Information Processing code standard to identify a US county
County Name	String	*
State	String	*
State FIPS	int	Federal Information Processing code standard to identify US state or territory
Total Deaths by Date	int	number of deaths caused by Covid on that day

COVID Cases

Variable Name	Data Type	Description
County FIPS	int	Federal Information Processing code standard to identify a US county
County Name	String	*
State	String	*
State FIPS	int	Federal Information Processing code standard to identify US state or territory
Total Cases by Date	int	number of total COVID cases on that day

Population By County

Variable Name	Data Type	Description
County FIPS	int	Federal Information Processing code standard to identify a US county
County Name	String	*
State	String	*
Population	int	Total population for the given county

Enrichment Data

Hospital beds

Variable Name	Data Type	Description
X & Y Coordinates	double	*

OBJECTID	int	*
Hospital Name	String	*
HQ Address	String	*
HQ Address 1	String	*
HQ City	String	*
HQ State	String	*
HQ Zip Code	int	*
County Name	String	*
State Name	String	*
State FIPS	int	Federal Information Processing code standard to identify a US state
County FIPS	int	Federal Information Processing code standard to identify a US county
FIPS	int	Federal Information Processing code standard
Number of Licensed Beds	int	is the maximum number of beds for which a hospital holds a license to operate; however, many hospitals do not operate all the beds for which they are licensed. This number is obtained through DHC Primary Research. Licensed beds for Health Systems are equal to the total number of licensed beds of individual Hospitals within a given Health System
Number of Staffed Beds	int	is defined as an "adult bed, pediatric bed, birthing room, or newborn ICU bed (excluding newborn bassinets) maintained in a patient care area for lodging patients in acute, long term, or domiciliary areas of the hospital." Beds in labor room, birthing room, post-anesthesia, postoperative recovery rooms, outpatient areas, emergency rooms, ancillary departments, nurses and other staff residences, and other such areas which are regularly maintained and utilized for only a portion of the stay of patients (primarily for special procedures or not for inpatient lodging) are **not** termed a bed for these purposes. Definitive Healthcare sources Staffed Bed data from the Medicare Cost Report or Proprietary Research as needed. As with all Medicare Cost Report metrics, this number is self-reported by providers. Staffed beds for Health Systems are equal to the total number of staffed beds

		of individual Hospitals within a given Health System. Total number of staffed beds in the US should exclude Hospital Systems to avoid double counting. ICU beds are likely to follow the same logic as a subset of Staffed beds
Number of ICU Beds - ICU (Intensive Care Unit) Beds	int	are qualified ICU based on definitions by CMS, Section 2202.7, 22-8.2. These beds include ICU beds, psychiatric ICU beds, and Detox ICU beds
Adult ICU Beds (Legacy)	int	In an emergency situation, hospitals may use additional intensive care beds to supplement an influx of patients. This number consists of all ICU beds, burn ICU beds, surgical ICU beds, or trauma ICU beds minus any pediatric, premature or neonatal ICU beds
Pediatric ICU Beds	int	are a combination of neonatal, pediatric and premature ICU beds
Average Ventilator Usage	int	is the average number of patients on a ventilator per week based on an analysis of 2016-2019 Medicare & commercial claims volumes by Definitive Healthcare
Bed Utilization Rate	double	is calculated based on metrics from the Medicare Cost Report: Bed Utilization Rate = Total Patient Days (excluding nursery days)/Bed Days Available
Potential Increase in Bed Capacity	int	This metric is computed by subtracting *"Number of Staffed Beds from Number of Licensed beds" (Licensed Beds – Staffed Beds)*. This would provide insights into scenario planning for when staff can be shifted around to increase available bed capacity as needed
Hospital Types	String	**Short Term Acute Care Hospital (STAC)** o Provides inpatient care and other services for surgery, acute medical conditions, or injuries o Patients care can be provided overnight, and average length of stay is less than 25 days **Critical Access Hospital (CAH)** o 25 or fewer acute care inpatient beds o Located more than 35 miles from another hospital o Annual average length of stay is 96 hours or less for acute care patients o Must provide 24/7 emergency care services o Designation by CMS to reduce financial vulnerability of rural hospitals and improve access to healthcare **Religious Non-Medical Health Care Institutions** o Provide nonmedical health care items and services to people who need hospital or skilled nursing facility care, but for whom that care would be inconsistent with their

		religious beliefs **Long Term Acute Care Hospitals** o Average length of stay is more than 25 days o Patients are receiving acute care - services often include respiratory therapy, head trauma treatment, and pain management **Rehabilitation Hospitals** o Specializes in improving or restoring patients' functional abilities through therapies **Children's Hospitals** o Majority of inpatients under 18 years old **Psychiatric Hospitals** o Provides inpatient services for diagnosis and treatment of mental illness 24/7 o Under the supervision of a physician **Veteran's Affairs (VA) Hospital** o Responsible for the care of war veterans and other retired military personnel o Administered by the U.S. VA, and funded by the federal government **Department of Defense (DoD) Hospital** o Provides care for military service people (Army, Navy, Air Force, Marines, and Coast Guard), their dependents, and retirees (not all military service retirees are eligible for VA services)

We will merge the COVID-19 data with the hospital data by matching the countyFIPS for each data set. This will allow us to see the hospital data side by side with the primary data. Using this combined data, we can see how hospitals have and will be affected by COVID-19. Looking at the allowed hospital beds and average ventilator usage for each county, we will be able to predict how well these hospitals are equipped to handle this situation. Is there a potential that some hospitals can't handle this? Will there be overflow to nearby hospitals?

Housing and Demographics

Variable Name	Data Type	Description
GEO_ID	string	A fourteen character string that summarizes a geographic location.
countyFIPS	int	A value made up of up to five digits that is used to identify a state and some

		county.
total_housing	int	The total number of housing units within a county.
male_population	int	Total number of males within the population.
female_population	int	Total number of females within the population.
sex_ratio	float	Proportion of male population when compared to female.
under_18	int	Total number of the population under 18.
over_18	int	Total number of the population over 18
62_and_over	int	Total number of the population that is either 62 or older.
median_age	float	The median age of a population.

We will be able to merge the housing data with the primary Covid-19 dataset through feature engineering. We will take the last five digits of each of the GEO_ID's and place this data in a new column which will represent the countyFIPS. The individual variable that maps between the two datasets is countyFIPS. Now that both datasets have countyFIPS as a column we will use this attribute to merge the two datasets.

The housing enrichment dataset will help in our analysis of Covid-19 spread by allowing us to focus on three main attributes. Those attributes being total housing, gender, and age. Total housing will allow us to estimate how many people live within each household, on average, for a given county. This gives insight into the spread of the virus because people who live together are sharing many of the same areas. This reduces to extended periods of time in close proximity which allows easier spread of viruses. The gender attribute will help in determining the density of either gender for a population. Gender differences can give insight into the spread of Covid-19 due to differences in

biology, behaviour, and attitudes (Bwire). Finally, the age attribute will allow us to determine the proportion of a population that is one of the three: under 18, young/middle aged, or elderly. This can help in our analysis of the spread of Covid-19 in two ways. First, age can bring about a sense of caution, we seek to protect the elderly or the young. Closing of schools, for example, may cause low spread between children. Our second example of how age influences spread of Covid-19 involves how the eldery have a higher likelihood of having pre-existing conditions that leads to hospitalizations. This places them in a position to spread the disease widely within hospitals (Older Adults and COVID-19). To conclude, we have seen that the three attributes, total housing, age, and gender have large impacts on the spread of Covid-19 and we will see their effects more clearly when we analyze these relationships.

Some questions we would like to answer are as follows. First, how does the spread of COVID-19 change through housing density? Are counties that have lower housing densities experiencing higher or lower number of cases, and vice versa, are higher housing densities correlated to higher or lower number of cases. Next, does the spread of Covid-19 change with various ratios of gender density? Are males more likely to transmit Covid-19, or are females? And finally, how does age affect the spread of Covid-19. Are children the super spreaders, or is it the young/middle aged adults, or perhaps it is the eldery? I hypothesize that those counties which have higher proportions of people living together, those with populations with a higher ratio of males, or those whose populations contain the highest percentage of elderly will have a higher number of Covid-19 cases.

Social

Variable Name	Data Type	Description
GEO_ID	string	A string that uniquely identifies each county. The last 5 digits are the countyFIPS
% of total households with a computer	float	*
% of total household with a broadband subscription	float	*

% popluation age 3+ enrolled in highschool	float	*
% popluation age 3+ enrolled in college	float	*
% population age 1+ living in the same house as last year	float	*
% population age 25+ graduated highschool	float	*
% population age 25+ graduated college - bachelors or higher	float	*
% of foreign-born population not a U.S. citizen	float	*
% households with male householder, no wife present, w/ family	float	*
% households with female householder, no husband present, w/ family	float	*
% household with one or more people of 65	float	*
Average household size	int	*
Average family size	int	*
% Males over 15, now married, except separated		Out of all males 15 and over, the % that are currently married, excluding those who are married but separated
% Females over 15, now married, except separated		Out of all females 15 and over, the % that are currently married,

		excluding those who are married but separated

To merge this data with the primary covid data, you must locate a common variable. Though the two data sets don't directly have any variables in common, the last five digits of the GEO_ID on the enrichment data can be extracted to match the countyFIPS field on the covid data. So if you convert the GEO_ID column to countyFIPS, you can merge the two tables through the countyFIPS variable.

The variables I choose cover several different types of metrics. One thing that I thought could be useful was measuring the education level of a county. It's possible that more highly educated places are more likely to trust science and practice social distancing. The *amount of people in college or highschool* could also lead to more cases since there will be more people in proximity to each other. I also included the amount of *households with computers and internet subscriptions.* It seems reasonable that counties with more internet access could stay better informed and up to date on their states policies and current recommendations. Since the virus affects the elderly most frequently, the *number of households with at least one person over 65* seemed like a natural thing to measure. Also *average household size* and *average family size* could definitely have an impact on covid cases.

Some other metrics that I chose may be less obviously tied to Covid but I would still be interested to see if there was any correlation. For example one category I chose was *percent of family households with a single mother/father* which i can speculate may affect covid cases because being a single parent requires going to work under any circumstances. Even though this seems like a stretch, it will still be interesting to see if there is a connection.

* Self explanatory variable

<u>Citations</u>

Bwire, George M. “Coronavirus: Why Men Are More Vulnerable to Covid-19 Than Women?” SN comprehensive clinical medicine. Springer International Publishing, June 4, 2020. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7271824/.

“Definitive Healthcare: USA Hospital Beds.” *COVID-19 Resources*, 2020, coronavirus-resources.esri.com/datasets/1044bb19da8d4dbfb6a96eb1b4ebf629_0/data?geometry=21.093%2C-16.820%2C-46.055%2C72.123.

US Coronavirus Cases and Deaths. 14 Sept. 2020, usafacts.org/visualizations/coronavirus-covid-19-spread-map/.

“Older Adults and COVID-19.” Centers for Disease Control and Prevention. Centers for Disease Control and Prevention, September 11, 2020. https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/older-adults.html.