COVID-19 Data Analysis Report

Jason Manning
Francis Perez
Jamison Valentine
Raiana Zaman
Xinrui (Sam) Zhang

CSC 405-01: Data Science

University of North Carolina Greensboro

Fall: 2020

Summary:

The goal of this project is to develop an analytical framework that can be used to study patterns of COVID-19's effect and spread throughout the United States.

Task 2 Section 1: Dataset description and Data type variable dictionary

Covid_county_population_usafacts.csv		
This file contains a list of counties in the U.S.A with their related populations.		
Variable	**Data Type**	**Description**
county FIPS	int64	The FIPS id of the county. A Unique code for each county of the U.S.A
County Name	object/string	The name of the county.
State	object/string	The name of the state.
population	int64	The population of the county.

covid_confirmed_usafacts (known cases).csv		
This file contains a list of counties in the U.S.A with their related confirmed cases of Covid-19. The file contains ever growing columns of dates with a running total of confirmed cases.		
Variable	**Data Type**	**Description**
county FIPS	int64	The FIPS id of the county. A Unique code for each county in the U.S.A
County Name	object/string	The name of the county.
State	object/string	The name of the state.
state FIPS	int64	The FIPS id of the state. A Unique code for each state in the U.S.A
1/22/2020 1/23/2020 **…** 9/7/2020	int64	The running total of confirmed cases starting on 1/22/2020.

covid_deaths_usafacts.csv		
This file contains a list of counties in the U.S.A with their related confirmed cases of Covid-19. The file contains ever growing columns of dates with a running total of deaths.		
Variable	**Data Type**	**Description**
county FIPS	int64	The FIPS id of the county. A Unique code for each county in the U.S.A
County Name	object/string	The name of the county.
State	object/string	The name of the state.
state FIPS	int64	The FIPS id of the state. A Unique code for each state in the U.S.A
1/22/2020 1/23/2020 **…** 9/7/2020	int64	The running total of deaths starting on 1/22/2020.

Task 2 Section 2: Preliminary Intuitions:

Upon giving the data a precursory look, potential patterns appear. Are densely populated areas affected more as there tends to be more known cases. A logical deduction would be that the more people directly interacting in a more confined space could easily lead to an increase in infections. Also, how people travel seems to have an effect. Popular vacation spots, high traffic airports, and any other areas of tourism could increase the number of infections as, again, large groups of people are clustered in a specific area. The more crowded an area, the greater chance that anyone infected with COVID-19 can spread it to many other people. Another potential hazard of crowded areas is if there are several people infected, it can increase the viral load of those infected and worsen symptoms, possibly increasing the number of deaths. Finally, the question arises of whether the sex of a person affects the spread of the coronavirus, particularly in older populations. More men seemed to die from the coronavirus than women. Could this be an example of a prevalence based on the chromosomes we carry, or is it just an example of women being smarter than men? A deeper analysis of the data will demonstrate if there is a correlation for any of our preliminary intuitions.

Enrichment Sets

Jamison

Dataset: Selected Social Characteristics (DP02)

Variable	Type	Description
GEO_ID / id	Text/String	Unique identifier for specific counties
Name / Geographic Area Name	Text/String	County Name, State Name
Households by type	Numerical (int)	Population estimates broken down into the following categories: Family households (families), Married-couple family, Male householder - no wife, Female householder no husband, Nonfamily
Relationship	Numerical (int)	Population estimates of different roles in relationships: householder, spouse, child, other relatives, nonrelatives
Marital Status	Numerical (int)	Never married, “Now married, except separated”, separated, widowed, divorced
Fertility	Numerical (int)	Number of women 15 to 50 years old who hard a birth in the past 12 months
Grandparents	Numerical (int)	Estimates for groups involving grandparents and their roles or level of responsibility
School Enrollment	Numerical (int)	Estimates based on current level of enrollment
Education Attainment	Numerical (int)	Estimates for highest level of education for people age 25 and up
Veteran Status	Numerical (int)	Number of civilian veterans age 19 and over

Disability Status of the Civilian Noninstitutionalized Population	Numerical (int)	Breakdown of disability status for 3 age groups: under 18, 18-64, 65 and older
Residence 1 Year Ago	Numerical (int)	Estimates of population based on past year's residency
Place of birth	Numerical (int)	Native vs. foreign born
US citizenship status	Numerical (int)	citizenship status
Year of entry	Numerical (int)	Year of entry: before 2010 vs after 2010
World region of birth of foreign born	Numerical (int)	Distribution of foreign born population based on regions: Europe, Asia, Africa, Oceania, Latin America, Northern America
Language spoken at home	Numerical (int)	Primary language for those age 5 and over
Ancestry	Numerical (int)	Ancestry divided into 20+ various bins
Computers and internet use	Numerical (int)	Estimates of people with computer in home plus estimates for those with internet

This enrichment data can be used to better understand which demographics are most affected by the COVID-19 spread and/or may contribute most to the spread. This can be achieved by analyzing the groups independently then by looking for combined effects. A few of the questions to ask include:

- Is there a larger percentage of people infected amongst those with a lower level of education attainment?
- Can the lack of computer or internet access be a factor in the spread?
- Are single people more likely to contract COVID-19 than those who are married?

Dataset: Selected Economic Characteristics (DP03)

Variable	Type	Description
id	Text/String	Unique identifier for specific counties
Geographic Area Name	Text/String	County Name, State Name
Employment status	Numerical (int)	Estimates for associated categories: 16 years and over, civilian labor force, unemployment rate, females 16 years and over, children under 6, children 6-17
Commuting to work	Numerical (int)	Various methods of transportation
Occupation	Numerical (int)	Types of occupation
Industry	Numerical (int)	Various industries
Class of worker	Numerical (int)	private/salary, government, self-employed, unpaid family workers
Income and benefits (2018 inflation-adjusted dollars)	Numerical (int)	Various bins for income amounts including various forms of government assistance
Health insurance coverage	Numerical (int)	Categories including combinations of age and employment status
Percentage of families below poverty last past 12 months	Numerical (int)	Percentages of various groups considered to have been below poverty level past year

This enrichment set provides critical information such as employment, commuting methods, health insurance coverage and more. It will be helping in answering the following questions:

- Do areas that reflect a higher percentage of families that have recently had incomes below the poverty line exhibit a higher number of cases or deaths by the spread?
- Does the lack of health insurance tend to be accompanied with a higher degree of vulnerability?
- Is there any correlation between average income and percentage of population affected within a county?
- Do industries and work classes that characterize certain counties have a significant impact on the overall spread within that county?

Merging with COVID-19 Dataset

Merging can be done by extracting the countyFIPS out of the "id" column. First, the "0500000US" prefix must be truncated from each entry in the "id" column. Next, the column data type must be changed to "int64" in order to match the datatype of the "countyFIPS" column of the main COVID-19 data set.

Xinrui Zhang
Dataset: Census Demographic ACS

Variable	Data Type	Description
id	Numerical and letters(String)	County FIPS id
Geographic Area Name	String	County name
Total population	Numerical(int)	Total population of County in 2018
Total population Male	Numerical(int)	Total population of male of County in 2018
Total population Male(percent)	Numerical(float)	Total population of male in percent of County in 2018
Total population Female	Numerical(int)	Total population of female of County in 2018
Total population female(percent)	Numerical(float)	Total population of female in percent of County in 2018
Sex Ratio(males per 100 females)	Numerical(float)	Sex ratio of males per 100 females of County in 2018
Total population under certain ages	Numerical(int)	Total population under certain ages of County in 2018
Total population under certain ages(percent)	Numerical(float)	Total population under certain ages in percent of County in 2018

Total population median age	Numerical(float)	Median of ages of County in 2018

Merging based on the last five digits of column “id”(County FIPS).
Since the demographic data column “id” has a data type of string and the super data column “countyFIPS” has a data type of integer, I have to first convert the “countyFIPS” column to data type string and then add leading zeros to make every data five digits. Then I can create a new column in the demographic data and name it “countyFIPS” in order to perform merge.

The enrichment data will be helpful for analyzing how confirmed cases and deaths related to people’s age and sex. Are older people or younger people more likely to be infected by COVID-19? Or are females or males more likely to be infected?

Francis Perez: Employment Dataset

Variable Dictionary

This file contains the quarterly census of employment and wages for each county in the USA.		
Variable	**Data Type**	**Description**
Area Code	object/string	County FIPS code
St	object/string	State FIPS code
Cnty	float64	3-character County FIPS code
Own	int64	1-character Ownership code
NAICS	int64	4-character Industry code (SuperSector)
Year	int64	4-digit year
Qtr	int64	quarter (always A for annual)
Area Type	object/string	Category of the given area. (State, County,Nation, etc…)
St Name	object/string	State name
Area	object/string	Area title associated with the area's FIPS code
Ownership	object/string	Ownership title associated with the ownership code
Industry	object/string	Industry title associated with the industry code
Status Code	object/string	Status code, or disclosure code ('N' for not disclosed)
Establishment Count	object/int64	Quarterly establishment counts for a given quarter
January Employment	object/int64	Employment for month
February Employment	object/int64	Employment for month
March Employment	object/int64	Employment for month
Total Quarterly Wages	object/int64	Total Wages
Average Weekly Wage	object/int64	Average Weekly Wage
Employment Location Quotient Relative to U.S.	float64	Employment Location Quotient Relative to U.S.
Total Wage Location	float64	Total Wage Location Quotient Relative to U.S.

Quotient Relative to U.S.		

How To Merge With Team Super Covid File:

The data variables that I intend to use for merging the two dataframes is the "Area Code" from the employment data file and the countyFIPS from the team super file (superCOVID-19datafame.csv). The employment data file does have rows that I will not be using, the only rows that will be using for the merge are the rows labeled with the "Area Type" column of value "County." I will break out the employment numbers by Industries of: federal government, state government, local government, private sector good-producing, and private sector services. Due to the employment data file containing the information in a row format for each county, it will not be possible to use the "Area Code" without pivoting the rows to columns to produce rows with only one "Area Code" or "countyFIPS" to merge with the main team file. The columns I will use are, Area Code, Area Type, Ownership, Industry, January Employment, February Employment. and March Employment. At this point I do not believe the wages data offers significant value on determining the spread of Covid-19.

Describe the Employment Data:

The Quarterly Census of Employment and Wages (QCEW) data file provides information on the employment and wage statistics for the United States and its counties. The file contains a list of counties and data for January, February, and March months.

Initial hypothesis questions:

I believe that the employment data can play a significant role in determining the spread Covid-19, due in part to the number of jobs that may require an employee to work in close proximity to one another. Such is the case for jobs that are in manufacturing. Could the type of employment, such as manufacturing or service based job help increase the spread of Covid-19? Could the counties or states with different percentages of each have lower or higher cases of confirmed and deaths due to covid?

Raiana Zaman:
Enrichment Data: Hospital Beds Dataset (Definitive_Health_USA_Hospital_Beds.csv)

1.Enrichment data and datatype - variable dictionary.

Variable	Data Type	Description
FLIP	Numerical(float64)	County FIPS
COUNTY_NAME	String	County name
HQ_STATE	String	State
HQ_CITY	String	Name of the city
STATE_FIPS	Numerical(float64)	State FIPS
NUM_LICENSED_BEDS	Numerical(float64)	the maximum number of beds for which a hospital holds a license to operate
NUM_STAFFED_BEDS	Numerical(float64)	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
NUM_ICU_BEDS	Numerical(int)	are qualified ICU based on definitions by CMS, Section 2202.7, 22-8.2. These beds include ICU bed
ADULT_ICU_BEDS	Numerical(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
PEDI_ICU_BEDS	Numerical(float64)	are a combination of neonatal, pediatric and premature ICU beds.
Bed Utilization Rate	Numerical(float64)	is calculated based on metrics from the Medicare Cost Report: Bed Utilization Rate = Total Patient Days (excluding nursery days)/Bed Days Available

2. I will use “countryFIPS” from Covid data set(superCOVID-19datafame.csv)and “FIPS” from the Hospital data set (Definitive_Health_USA_Hospital_Beds.csv) to merge the datasets .

3.Initial hypothesis:

If the number of confirmed case higher than NUM_LICENSED_BEDS is a Country, death number would be higher

Jason Manning

Enrichment Dataset: Selected Housing Characteristics (DP04)

Variable	Type	Description
GEO_ID/ id	Text/String	CountyFIPS, StateFIPS
Name/ Geographic Area Name	Text/String	County Name, State Name
Total Housing Units	Numerical (int)	Total homes, apartments, etc.
Occupied Housing Units	Numerical (int)	Total of homes with occupants
Vacant Housing Units	Numerical (int)	Total of homes without occupants
Homeowner Vacancy Rate	Numerical (float)	Proportion of the homeowner inventory that is vacant for sale.
Rental Vacancy Rate	Numerical (float)	Proportion of the rental inventory that is vacant for rent.
Total 1-unit detached	Numerical (int)	Total standard homes
Total 1-unit attached	Numerical (int)	Total townhouses, duplexes, etc. (shared wall must go from ground to roof)
Total 5 to 9 units	Numerical (int)	Total apartments, and similar dwellings
Total 20 or more units	Numerical (int)	Total apartments, and partments
Total Mobile Homes	Numerical (int)	Total mobile homes
Total Rooms: 1	Numerical (int)	Total of homes with only one room

Total Rooms: 4	Numerical (int)	Total of homes with four rooms
Total Rooms: 9 or more	Numerical (int)	Total of homes with nine or more rooms
Average Household size -- Owner occupied	Numerical (float)	Average number of occupants per owner occupied home
Average Household size -- Renter occupied	Numerical (float)	Average number of occupants per renter occupied home
Owner-occupied unit Value: Median (dollars)	Numerical (int)	Median value in dollars for homes occupied by owners
Housing unit with a mortgage: Median (dollars)	Numerical (int)	Median value in dollars of homes with mortgages
Housing unit paying rent: Median (dollars)	Numerical (int)	Median value in dollars of homes that pay rent.

-After extracting the countyFIPS and separating state and county, I was able to merge the datasets on those three variables.

-The housing data consists of variables such as size of a home, value of a home, rent or own, household size, and vacancy. This data can be used to deduce the type of neighborhood someone lives in or the income level of a family. Hypothetically, could the living conditions contribute to the possibility of being infected? Do lower income families living in apartments have a higher chance of being infected than a higher income family living in a more expensive single family home?

Task 3: Calculate COVID-19 data trends for last week of the data. Are the cases increasing, decreasing, or stable. Each student chooses a State (such as North Carolina, Virginia, New York, etc.) to analyze.

** Please refer to individual notebooks

Jamison: Georgia
Xinrui Zhang: Indiana
Francis Perez: North Carolina
Raiana Zaman: New York
Jason Manning: Washington