CSC-405/605 – Team 6
Fall 2020

Task 2

Note - All our COVID datasets were downloaded on September 15, 2020.

COVID-19 datasets and variables

1. Covid_confirmed_usafact.csv - The dataset gives us the total number of confirmed COVID-19 cases every day per county in the United States beginning January 22, 2020.

Variable Name	Data Type	Description
countyFIPS	Integer	A unique id associated with each county in the United States. It is a 5-digit code, first 2 codes are the state FIPS code and the remaining are county FIPS codes.
County Name	Text	County Name
State	Text	State Name
StateFIPS	Integer	A two digit unique Id for each state in United States
Dates	Integer	It contains the number of COVID-19 confirmed cases in specific county.

2. Covid_county_population_usafacts.csv - This dataset gives us the population of each county in United States.

Variable Name	Data Type	Description
countyFIPS	Integer	A unique id associated with each county in the United States. It is a 5-digit code, first 2 codes are the state FIPS code and the remaining are county FIPS codes.

County name	Text	County name
State	Text	State Name
Population	Integer	Population of the county.

3. Covid_deaths_usafacts.csv - The dataset gives us the total number of deaths counted every day from COVID-19 per county in the United States beginning January 22, 2020.

Variable Name	**Data Type**	**Description**
countyFIPS	Integer	A unique id associated with each county in the United States. It is a 5-digit code, first 2 codes are the state FIPS code and the remaining are county FIPS codes.
County Name	Text	County Name
State	Text	State Name
StateFIPS	Integer	A two digit unique Id for each state in United States
Dates	Integer	It contains the number of COVID-19 deaths in the specific county.

Preliminary intuitions from the data

1. Deaths in each county seem to be directly proportional to the confirmed cases in each county. Higher the number of confirmed cases, higher is the death rate.
2. Counties with higher populations can have a higher number of confirmed cases as well as deaths.
3. Confirmed COVID cases and deaths seem to have an increasing trend which is also evident from the trends analysis that we did for the last one week of confirmed cases and death.
4. There appears to be a general positive trend between the county population and number of COVID-19 cases. Counties with higher deaths may have a population with more elderly.

5. The data appears to be on a positive trend across states with higher populations meaning that the confirmed number of cases does not appear to be slowing down.

Enrichment Datasets

1. **Definitive_Healthcare_Hospital_Beds_0914.csv – Completed by Ritu Joshi**

- ***Section in the report describing the enrichment data and data type - variable dictionary.***
  This dataset is intended to be used as a baseline for understanding the typical bed capacity and average yearly bed utilization of hospitals reporting such information. The date of the last update received from each hospital may be varied. While the dataset is not updated in real-time, this information is critical for understanding the impact of a high utilization event, like COVID-19. The data is organized per county.

Variable Name	Data Type	Description
X	Decimal	Longitude
Y	Decimal	Latitude
Hospital Name	Text	Name of the hospital
Hospital Type	Text	Type of hospital e.g. VA hospitals, Children's hospital etc.
HQ_Address	Text	Street address of the hospital
HQ_City	Text	City of the hospital
HQ_State	Text	State of the hospital
HQ_ZIP_Code	Text	ZIP code of the hospital
County_Name	Text	County name
State_Name	Text	State Name
FIPS	Integer	Federal Information Processing Standards(FIPS) codes are numbers which uniquely identify geographic areas. The number of digits in FIPS codes varies depending on the level of geography. State-level FIPS codes have

		two digits, county-level FIPS codes have five digits of which the first two are the FIPS codes of the state to which the county belongs.
State_FIPS	Integer	State-level FIPS codes have two digits
Cnty_FIPS	Integer	County-level FIPS codes have five digits of which the first two are the FIPS code of the state to which the county belongs
Num_Licensed_Beds	Integer	It 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.
Num_Staffed_Beds	Integer	Staffed bed 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.

Num_ICU_Beds	Integer	These include ICU beds, psychiatric ICU beds and Detox ICU Beds
Adult_ICU_Beds	Integer	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.
Pedi_ICU_Beds	Integer	These are a combination of neonatal, pediatric and premature ICU beds
Bed_Utilization	Decimal	It 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_Capac	Integer	This metric is computed by subtracting the "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.
Avg_Ventilator_Usage	Integer	Average number of patients on ventilator per week based on an analysis of 2016-2019 Medicare & commercial claims volumes by Definitive Healthcare.

- ***How can you merge the data with the primary COVID-19 dataset? Identify the individual variable which map between the datasets.***

Hospital beds dataset has multiple hospital records per county information. The dataset contains “FIPS” code as the unique identifier for each county. We can merge the hospital beds information with COVID-19 dataset by applying inner join between two tables 1. “Covid-19 merged dataset” and 2. Hospital beds on columns “countyFIPS” from table 1 and column “FIPS” from table 2.

- ***Describe how your enrichment data can help in the analysis of COVID-19 spread. Pose initial hypothesis questions.***

  Hospital beds dataset along with COVID data can help us understand and answer alot of questions related to deaths and confirmed cases. My initial hypothesis questions are as follows:

  1. Are the deaths and confirmed cases in each county inversely proportional to the number of ICU beds in the county? In other words we can say that if a county has a low number of ICU beds then the chances of deaths in that county may be higher.
  2. Can we say from the merged dataset that the count of deaths is lower in counties where the bed utilization is higher?
  3. Can we hypothesize that “Potential_Increase_In_Bed_capacity” is directly proportional to the population of the county?
  4. Can we hypothesize that more the ventilator usage in a county, higher are the chances of critical patients leading to an increasing number of deaths.

2. **ACS Economic Dataset – Completed by Peter Yuan**

- ***Section in the report describing the enrichment data and datatype - variable dictionary.***

  This is an economic dataset group by each county in the United states. I have picked some features from the whole dataset which includes the health insurance coverage, employment status , the occupation distribution ,commuting status and the income distribution.

Variable Name	Data Type	Description
GEO_ID	String	This is ID concatenated by string “0500000US” and countyFIPS same as in core-dataset
NAME	String	Geographic Area Name
DP03_0001E	Integer	The Population 16 years and over
DP03_0024E	Integer	The population of Workers 16 years and over who worked at home
DP03_0026E	Integer	Civilian employed population 16 years and over
DP03_0027E	Integer	Civilian employed population 16 years and over in the field of Management, business, science, and arts occupations
DP03_0028E	Integer	Civilian employed population 16 years and over in the field of Service occupations
DP03_0029E	Integer	Civilian employed population 16 years and over in the field of Sales and office occupations
DP03_0030E	Integer	Civilian employed population 16 years and over in the field of Natural resources, construction, and maintenance occupations

DP03_0031E	Integer	Civilian employed population 16 years and over in the field of Production, transportation, and material moving occupations
DP03_0075E	Integer	The number of Total Families.
DP03_0076E	Integer	The number of Total Families whose INCOME AND BENEFITS (IN 2018 INFLATION-ADJUSTED DOLLARS) is Less than $10,000
DP03_0077E	Integer	The number of Total Families whose INCOME AND BENEFITS (IN 2018 INFLATION-ADJUSTED DOLLARS) is from $10,000 to $14,999
DP03_0078E	Integer	The number of Total Families whose INCOME AND BENEFITS (IN 2018 INFLATION-ADJUSTED DOLLARS) is from $15,000 to $24,999
DP03_0079E	Integer	The number of Total Families whose INCOME AND BENEFITS (IN 2018 INFLATION-ADJUSTED DOLLARS) is from $25,000 to $34,999
DP03_0080E	Integer	The number of Total Families whose INCOME AND BENEFITS (IN 2018 INFLATION-ADJUSTED DOLLARS) is from $35,000 to $49,999

DP03_0081E	Integer	The number of Total Families whose INCOME AND BENEFITS (IN 2018 INFLATION-ADJUSTED DOLLARS) is from $50,000 to $74,999
DP03_0082E	Integer	The number of Total Families whose INCOME AND BENEFITS (IN 2018 INFLATION-ADJUSTED DOLLARS) is from $75,000 to $99,999
DP03_0083E	Integer	The number of Total Families whose INCOME AND BENEFITS (IN 2018 INFLATION-ADJUSTED DOLLARS) is from $100,000 to $149,999
DP03_0084E	Integer	The number of Total Families whose INCOME AND BENEFITS (IN 2018 INFLATION-ADJUSTED DOLLARS) is from $150,000 to $199,999
DP03_0085E	Integer	The number of Total Families whose INCOME AND BENEFITS (IN 2018 INFLATION-ADJUSTED DOLLARS) is from $200,000 or more
DP03_0095E	Integer	Civilian noninstitutionalized population
DP03_0096E	Integer	Civilian noninstitutionalized population with health insurance coverage
DP03_0097E	Integer	Civilian noninstitutionalized population with private health insurance

DP03_0098E	Integer	Civilian noninstitutionalized population with public health insurance coverage
DP03_0099E	Integer	Civilian noninstitutionalized population without health insurance coverage

- ***How can you merge the data with the primary COVID-19 dataset? Identify the individual variable which map between the datasets.***

  The index variable GEO_ID of the enrichment dataset consists of the string "0500000US" and the index variable countyFIPS of the core dataset. Therefore, I could merge two datasets according to the GEO_ID and countyFIPS.

- ***Describe how your enrichment data can help in the analysis of COVID-19 spread. Pose initial hypothesis questions.***

  I choose the variables in the enrichment dataset which reflect 4 different features.

  1. The employment and unemployment percentages of the population 16 years and over.

     From this feature, I make a hypothesis that the higher employment percentage the easier to spread the COVID-19. In this experiment, I will deduct the population who work at home.
  2. The occupation distribution of the employed population.

     From this feature, I make an assumption that occupations with many contacts may be at higher risk of being infected and spreading COVID-19. For example, Service occupations.
  3. The family income distribution

     From this feature, I make an assumption that the high percentage of higher-income of the family may not affect the possibility of infection, but it may lead to a lower death possibility.
  4. The health insurance coverage distribution

From this feature, I make a hypothesis that the higher percentage of coverage of private health insurance may lead to a lower death possibility.

3. **ACS SEX AND AGE** - Quinn Tjin-A-Soe

Enrichment Datasets

ACSST1Y2018.S0101_data_with_overlays_2020-09-12T231309.csv

I will be looking at the enrichment dataset of the ACS SEX AND AGE dataset. This data set provides all counties in the US and their corresponding age groups amongst the total population, male population, and female population.

Variable Dictionary

GEO_ID	string	The geographic area name with "0500000US" appended to the beginning of the unique ID.
NAME	string	The county name
S0101_C02_002E	decimal	The percent of the total population of US citizens from ages under 5
S0101_C02_003E	decimal	The percent of the total population of US citizens from ages 5 to 9
S0101_C02_004E	decimal	The percent of the total population of US citizens from ages 10 to 14
S0101_C02_005E	decimal	The percent of the total population of US citizens from ages 15 to 19
S0101_C02_006E	decimal	The percent of the total population of US citizens from ages 20 to 24
S0101_C02_007E	decimal	The percent of the total population of US citizens from ages 25 to 29
S0101_C02_008E	decimal	The percent of the total population of US citizens from ages 30 to 34
S0101_C02_009E	decimal	The percent of the total population of US citizens from ages 35 to 39
S0101_C02_010E	decimal	The percent of the total population of US citizens from ages 40 to 44

S0101_C02_011E	decimal	The percent of the total population of US citizens from ages 45 to 49
S0101_C02_012E	decimal	The percent of the total population of US citizens from ages 50 to 54
S0101_C02_013E	decimal	The percent of the total population of US citizens from ages 55 to 59
S0101_C02_028E	decimal	The percent of the total population of US citizens of ages 60 and older

- ***How can you merge the data with the primary COVID-19 dataset? Identify the individual variable which map between the datasets.***

  Both the GEO_ID and the county FIPS correspond to zip codes in the US, and therefore, can be merged together.

- ***Describe how your enrichment data can help in the analysis of COVID-19 spread. Pose initial hypothesis questions.***

  By looking at the the age groups of people who have contracted COVID-19 and the cases of deaths, we can compare the two datasets and find a pattern between vulnerable age groups who are at higher risk of dying from COVID-19, as well as, age groups that trend towards higher cases.

4. **Employment Enrichment Dataset-** completed by Amantii Samson

This dataset displays the level of employment and earnings of each state and county within the United States for the first fiscal quarter of 2020. It also provides a breakdown of the earnings based on the industry within a specific state or county and also expresses all these values for the entire United States.

Variable Dictionary:

Variable Name	Data Type	Description
Year	String	The year for which the employment data was collected

Quarter(Qtr)	Integer	The fiscal quarter for which the employment data was collected.
Area Type	String	Describes the nation, state, or county for which the employment data was collected.
State Name	String	Name of the state for which the employment data was collected.
Area	String	Area within state(statewide or county) for which employment data was collected.
Ownership	String	Status of ownership of certain areas of employment within each state/county area.
Industry	Integer/String	Specific type of industry that each area of employment belongs to.
Status Code	String	Describes an instance when an industry has no available data for Q1 other than the establishment count.
Establishment Count	Integer	Determines amount of a specific industry within the area.
January Employment	Integer	Amount of people employed in a certain state/county during the month of January.
February Employment	Integer	Amount of people employed in a certain state/county during the month of February.
March Employment	Integer	Amount of people employed in a certain state/county during the month of March.
Total Quarterly Wages	Integer	An accumulation of the wages made during the first fiscal quarter (January, February, and March).
Average Weekly Wage	Integer	Average amount of money a person in each state/county makes per week.

Employment Location Quotient Relative to U.S.	Decimal	Determines the concentration of employment within a certain area relative to the employment concentration of the entire U.S.
Total Wage Location Quotient Relative to U.S.	Decimal	Determines the total wages in a specific area relative to the total wages of the entire U.S.

- ***How can you merge the data with the primary COVID-19 dataset? Identify the individual variable which map between the datasets.***

  The employment enrichment dataset can be merged with the COVID-19 dataset using the geo locations for counties. A merge between the two datasets may also be possible using the January, February, and March employment counts in the employment dataset and the data for the first three months of the year from the confirmed cases and deaths datasets.

- ***Describe how your enrichment data can help in the analysis of COVID-19 spread. Pose initial hypothesis questions.***

  Using the employment enrichment dataset can help in analyzing the COVID-19 spread because a county or state with a higher concentration of employment means that more people are coming in contact with one another, thus resulting in a quicker and larger spread of COVID-19.

- ***Hypothesis* <u>questions</u>**:

  - Since the employment dataset applies only to the first fiscal quarter (first three months) of 2020 does that mean that there will be little to no connections between the spread of COVID-19 and employment?
  - Do states or counties with higher concentrations of employment during the first fiscal quarter of 2020 result in earlier confirmed cases of COVID-19?
  - Do states or counties with higher concentrations of employment during the first fiscal quarter of 2020 result in a quicker and larger spread of COVID-19?

5. **Educational Attainment Enrichment Dataset -** Completed by ***Alejandro Penaloza***

- This Dataset represents the population of people in specific age ranges and their level of education attained. It goes from less than high school degree all the way to graduate or professional degree. It's data from the 2019 year and is broken down into all counties in the United States

Variable Dictionary

Geo ID	Text	Geographic Area ID
Name	Text	County and State Name
S1501_C01_001E	Integer	Population 18 to 24 years
S1501_C01_002E	Integer	Population 18 to 24 years, Less than high school graduate
S1501_C01_003E	Integer	Population 18 to 24 years, High school graduate (includes equivalency)
S1501_C01_004E	Integer	Population 18 to 24 years, Some college or associate's degree
S1501_C01_005E	Integer	Population 18 to 24 years, Bachelor's degree or higher
S1501_C01_006E	Integer	Population 25 years and over
S1501_C01_007E	Integer	Population 25 years and over, Less than 9th grade
S1501_C01_008E	Integer	Population 25 years and over, 9th to 12th grade, no diploma
S1501_C01_009E	Integer	Population 25 years and over, High school graduate (includes equivalency)
S1501_C01_010E	Integer	Population 25 years and over, Some college, no degree
S1501_C01_011E	Integer	Population 25 years and over, Associate's degree

S1501_C01_012E	Integer	Population 25 years and over, Bachelor's degree
S1501_C01_013E	Integer	Population 25 years and over, Graduate or professional degree

- ***How can you merge the data with the primary COVID-19 dataset? Identify the individual variable which map between the datasets.***

  In the first column of the enrichment dataset, is a GEO_ID variable that can be mapped and merged alongside the corresponding countyFIPS of the main dataset.

- ***Describe how your enrichment data can help in the analysis of COVID-19 spread. Pose initial hypothesis questions.***

  Based on the variables given and analysed in the enrichment dataset, I could say that this could help in the analysis of COVID-19 spread through reviewing specific age groups and what their highest level of educational attainment is.

- ***Hypothesis Questions:***
  - Are counties with a high population of people with a college degree educational attainment seeing more or less confirmed cases of COVID-19?
  - Are counties with a high population of people with less than a high school degree educational attainment seeing more or less confirmed cases of COVID-19?
  - What can this data say about how information about the spread of COVID-19 is being shared and distributed? Is there a connection/correlation between which areas are being distributed information alongside confirmed COVID-19 cases?