Section in the report describing the COVID-19 dataset and datatype - variable dictionary.

Description of Data:

In the dataset “covid_confirmed_usafacts”, each row represents a single county in the USA. The number of daily confirmed COVID cases are recorded in the column under the date they were recorded. “covid_deaths_usafacts” uses the same format, but records COVID deaths. “covid_county_population_usafacts” records the population of each county in the USA.

Data Type - Variable Dictionary

covid_confirmed_usafacts			
Name	Definition	Data Type	Required?
countyFIPS	5 digit code that uniquely identifies all counties	Text	Yes
County Name	The name of the county	Text	Yes
State	The name of the state	Text	Yes
stateFIPS	2 digit code that uniquely identifies all states	Text	Yes
Date (YYYY-MM-DD)	All other columns enumerate the number of COVID cases confirmed on a single date. The date is listed with the YYYY-MM-DD format.	Number	Yes

covid_county_population_usafacts			
Name	Definition	Data Type	Required?
countyFIPS	5 digit code that uniquely identifies all counties	Text	Yes
County Name	The name of the county	Text	Yes
State	The name of the state	Text	Yes
Population	Number that describes the population within a county	Number	Yes

covid_deaths_usafacts			
Name	Definition	Data Type	Required?
countyFIPS	5 digit code that uniquely identifies all counties	Text	Yes
County Name	The name of the county	Text	Yes
State	The name of the state	Text	Yes
stateFIPS	2 digit code that uniquely identifies all states	Text	Yes
Date (YYYY-MM-DD)	All other columns enumerate the number of COVID cases confirmed on a single date in the corresponding county.. The date is listed with the YYYY-MM-DD format.	Number	Yes

Individual work:

William Harper

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

This dataset describes the quarterly census of employment and wages. Each row represents an industry in a particular area.

US_St_Cn_MSA			
Name	Definition	Data Type	Required?
Area Code	5-character FIPS code	Text	Yes
St	2-character State FIPS code	Text	No
Cnty	3-character County FIPS code	Text	No
Own	1-character Ownership code	Text	Yes
NAICS	4-character Industry code (SuperSector)	Text	Yes
Year	4-digit year	Number	Yes
Quarter	1-character quarter	Text	Yes
Area Type	Category of the given area. (State, County, Nation, etc.…)	Text	Yes
St Name	Multi-character State name	Text	No
Area	Area title associated with the area's FIPS code	Text	Yes
Ownership	Ownership title associated with the ownership code. (ex: Privately owned, owned by local, state, or federal governments)	Text	Yes
Industry	Industry title associated with the industry code	Text	Yes
Status Code	Status code, or disclosure code ('N' for not disclosed)	Text	No
Establishment Count	Quarterly establishment counts for a given quarter	Text	Yes

January Employment	Employment level for the first month of the quarter. The name of the month is displayed.	Text	Yes
February Employment	Employment level for the first month of the quarter. The name of the month is displayed.	Text	Yes
March Employment	Employment level for the first month of the quarter. The name of the month is displayed.	Text	Yes
Total Quarterly Wages	Total quarterly wage level for the given quarter	Text	Yes
Average Weekly Wage	Average weekly wage based on the 12-monthly employment levels and total annual wage levels	Text	Yes
Employment Location Quotient Relative to U.S.	Location quotients compare the concentration of an industry within a specific area to the concentration of that industry nationwide. Location quotients are ratios that allow an area's distribution of employment by industry, ownership, and size class to be compared to a reference area's distribution. The U.S. is used as the reference area for all LQs within the files. The reference industry is always the all-industry, all-ownerships total for the local area, and for the nation. If an LQ is equal to 1, then the industry has the same share of its area employment as it does in the nation.	Text	Yes
Total Wage Location Quotient Relative to U.S.	Similar to Employment Location Quotient, but comparing wages instead of raw employment numbers.	Text	Yes

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

Although these columns are hidden by default, every county is listed with its county and state name. Although the data that represents every county is broken into different industries, there is a row that describes all industries that can be added to the end of the merged dataset.

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

Places with high employment in industries that require lots of human contact like hospitality are more likely to spread disease. I hypothesize that, in places with high employment, COVID will spread from person to person more easily, and the data may reflect that.

Mawuli Agboklu: <u>Presidential Election Results Enrichment Dataset</u>

- **<u>Section in the report describing the enrichment data and datatype - variable dictionary.</u>**

This dataset describes how the different states and counties in the United States voted in the 2020 election. It has 6 rows depicting state, county, candidate, party, total_votes and won.

Name	Definition	Data Type	Required
State	Name of state	Text	Yes
County	Name of county within the states	Text	Yes
Candidate	Names of all presidential candidates	Text	Yes
Party	Three letter word abbreviation of parties which contested the elections	Text	Yes
Total_Votes	Total votes won by each candidates in each county	Integer	Yes
Won	Whether the candidate won the county or not	Boolean	Yes

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

State, county, candidate and won can be merged with the primary covid-19 dataset at any point of the primary covid-19 dataset except the beginning.

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

States which are more consecutive in their political ideology are likely to see a spike in the spread of the covid-19 virus than those who are liberal in their political ideology. My hypothesis is that, given the difficulty of the republican party to appreciate the safety protocols issued by the CDC and the WHO in managing the covid-19 virus, states which are pro-republican are likely to ignore the safety instructions which will lead to a spike in spread of the virus in those states.

Ege Keser - Presidential Election Results (Political Leanings)

This dataset represents the political leanings of the counties all around North Carolina based on the results of the 2020 Presidential Election.

Name	Definition	Data Type	Required
state	Name of the state	Text	Yes
county	Name of the county	Text	Yes
candidate	Name of the people ran for presidency	Text	Yes
party	Abbreviation of political parties that candidates represented	Text	Yes
total_votes	Number of votes the specific candidate got	Integer	Yes
won	Whether the candidate won the presidency race or not	Boolean	Yes

Based on the variables we have in our political leaning dataset, we can use county variable for merging operation with the main COVID-19 dataset. County will help us to match the geographical locations. We will also be using the won and candidate variables. By looking at both at the same time, we can understand which candidate received the most votes in that county.

Using the political leaning dataset, we can analyze the correlation between COVID-19 statistics and political ideas like conservatism and liberalism. My hypothesis is that the counties in North Carolina where Joe Biden had more votes than any other candidates will have less number of cases since the voters of Joe Biden follow the COVID-19 safety regulations more closely compared to the voters of Donald Trump and possibly other candidates.

Leena Godbole - Consumer Price Index of Gasoline

Consumer Price Indexes - Gasoline			
Name	Definition	Data Type	Required
SeriesID	This queries the gasoline data from the larger dataset	Text	No
Year	Year	Text	No
Period	Internal key used in the BLS to refer to specific months	Text	No
Label	Date of data observation in year / month format (eg. 2023 Aug)	Text	Yes, will use to combine tables then drop
Value	The price of Gasoline for that period	Integer	Yes

Andrew Van Es - Census Demographic ACS Information

This dataset describes the age ranges and demographics of the population of the United States.

Name	Definition	Data Type	Required
County	Name of County	Text	Yes
State	Name of State	Text	Yes
Sex and Age (Total Population)	Total Population of the county of Male populations, Female populations, and their age ranges	Int	Yes
Sex and Age (Male)	Total Male population of the county	Int	Yes
Sex and Age (Female)	Total Female population of the county	Int	Yes
Sex and Age (Age Ranges)	Population within the age range, 5 to 9 through 85 years and over	Int	Yes
Race (Total population)	Total population of the United States	Int	Yes
Race (Demographics)	This encompasses multiple columns of information of the estimated demographics on the population of the United States	Int	Yes
Hispanic or Latino and Race (Total population)	Total population of estimated Hispanic and Latino populations within the United States	Int	Yes
Hispanic or Latino and Race (Total population)	This encompasses multiple rows of Hispanic or Latino populations based on demographics	Int	Yes
Citizen, Voting Age population (Citizen, 18 and over population)	This shows the total population of the United States who are of voting age estimates.	Int	Yes

Citizen, Voting Age population (Male)	This column shows the estimated Male population that is of voting age.	Int	Yes
Citizen, Voting Age population (Female)	This column shows the estimated Female population that is of voting age.	Int	Yes

We can merge this Data set with the main covid data set by the County and State variables.

With this information we can learn about covid 19 infection levels between the different age ranges of people within a population. It can also be used to find different infection levels between different demographics of people. I hypothesize that the younger age ranges would have higher levels of infection since children may lack basic sanitary knowledge. I think that higher age ranges could also suffer from more cases and deaths due to having weaker immune systems.