Sarvesh Sridher, Rongxin Liu, Will Hutcheon, Sergio Guerra

Stage I Task 2 Report

The COVID-19 data comes in three datasets which focus on confirmed cases, deaths, and total county populations. The three datasets have several variables in common which include county FIPS code, county name, state FIPS code, and state. The COVID cases dataset and the COVID deaths dataset contain hundreds of columns that specify different dates from 1/22/2020 to 7/23/2023. In these columns, it specifies the number of cases and deaths in the county on that day respectively. Below are the data dictionaries for the following datasets:

Confirmed Cases Data Dictionary:

Name	Definition	Data Type	Possible Values
countyFIPS	Unique code that identifies counties	Integer	01008, 01002
County Name	Name of the county	String	Autauga County, Coosa County
State	State that the county resides in	String	North Carolina, Arizona
StateFIPS	Unique code that identifies states and serves as prefix to countyFIPS	Integer	01, 02, 03
Dates	There are columns that each have a date from 1/22/2020 to 7/23/23 that specify the number of cases on that day	Integer	100, 12309, 410

COVID Deaths Data Dictionary:

Name	Definition	Data Type	Possible Values
countyFIPS	Unique code that identifies counties	Integer	01008, 01002
County Name	Name of the county	String	Autauga County, Coosa County
State	State that the county resides in	String	North Carolina, Arizona
StateFIPS	Unique code that identifies states and serves as prefix to countyFIPS	Integer	01, 02, 03
Dates	There are columns that each have a date	Integer	100, 12309, 410

	from 1/22/2020 to 7/23/23 that specify the number of deaths on that day		

County Populations:

Name	Definition	Data Type	Possible Values
countyFIPS	Unique code that identifies counties	Integer	01008, 01002
County Name	Name of the county	String	Autauga County, Coosa County
State	State that the county resides in	String	North Carolina, Arizona
Population	Number of people in county	Integer	100, 12309, 410

There were also several enrichment datasets used to provide further context for the COVID datasets. One of the datasets included the social characteristics of the people living in the US counties. Some of the variables included in the data include Geographic code, Geographic Area, and other social characteristics of the people living in a certain county. Some of the social characteristics include household size, citizenship status, language spoken, and ancestry. This enrichment data provides details about how a population lives and under which conditions. It discusses their communities and can help draw conclusions between those characteristics and the rate at which they faced COVID.

This data is also able to be combined with the COVID-19 dataset because of the geographic code, which contains the county FIPS code. Since this variable is present in both datasets, they can be easily merged based on this code. Merging this data will help the analysis because it allows us to connect faster rates of spread and death with specific social attributes. Some initial hypothesis questions could include:

- Does household size affect the rate of COVID spread?
- Are people of certain ancestries more susceptible to the COVID-19 infection?
- Are people with computer use more knowledgeable about the disease and less likely to be infected?

Will’s Enrichment Dictionary

Name	Definition	Data Type	Possible Values

Area Code	5-character FIPS code	object	int64, US000
St	2-character State FIPS code	object	(Blanks), int64, US
Cnty	3-character County FIPS code	float64	(Blanks), int64
Own	1-character Ownership code	int64	0-5
NAICS	4-character Industry code (SuperSector)	int64	10, 101, 1011, 1012, 1013, 102, 1021, 1022, 1023, 1024, 1025, 1026, 1027, 1029
Year	4-digit year	int64	2023
Qtr	1-character quarter (always A for annual)	int64	1
Area Type	Category of the given area. (State, County, Nation, etc.…)	object	Nation, State, County, MSA
St Name	Multi-character State name	object	Alabama, Alaska, Arizona, Arkansas, California, Colorado, Connecticut, Delaware, Florida, Georgia, Hawaii, Idaho, Illinois, Indiana, Iowa, Kansas, Kentucky, Louisiana, Maine, Maryland, Massachusetts, Michigan, Minnesota, Mississippi, Missouri, Montana, Nebraska, Nevada, New Hampshire, New Jersey, New Mexico, New York, North Carolina, North Dakota, Ohio, Oklahoma, Oregon, Pennsylvania, Rhode Island, South Carolina, South Dakota, Tennessee, Texas, Utah, Vermont, Virginia, Washington, West

			Virginia, Wisconsin, Wyoming
Area	Area title associated with the area's FIPS code	object	object in the form of a String
Ownership	Ownership title associated with the ownership code	object	Federal Government, Local Government, Private, State Government, Total Covered
Industry	Industry title associated with the industry code	object	(10 Total, all industries), 101 Goods-producing, 1011 Natural resources and mining, 1012 Construction, 1013 Manufacturing, 102 Service-providing, (1021 Trade, transportation, and utilities), 1022 Information, 1023 Financial activities, 1024 Professional and business services, 1025 Education and health servics, 1026 Leisure and hospitality, 1027 Other services, 1029 Unclassified
Status Code	Status code, or disclosure code ('N' for not disclosed)	object	(Blanks), N
Establishment Count	Annual establishment count for a given year	int64	int64
January Employment	Employment level for the first month of the quarter. The name of the month is displayed. (EX: January)	int64	int64
February Employment	Employment level for the second month of the quarter. The name of the month is displayed. (EX: February)	int64	int64

March Employment	Employment level for the third month of the quarter. The name of the month is displayed. (EX: March)	int64	int64
Total Quarterly Wage	Total quarterly wage level for the given quarter	int64	int64
Average Weekly Wage	Average weekly wage based on the 12-monthly employment levels and total annual wage levels	int64	int64
Employment Location Quotient Relative to U.S.	Employment Location Quotient Relative to U.S.	float64	float64
Total Wage Location Quotient Relative to U.S.	Total Wage Location Quotient Relative to U.S.	float64	float64

Wills Task 2 Question Answers:

The individual variable I will use to map between my enrichment data set and the shared data set will be the countyFIPS.

Are employees working in certain industries more susceptible to covid infection and death? The employment enrichment data helps analyze covid spread by allowing us to correlate rates of covid infections and deaths to an area's employment distribution. This insight can reveal trends among covid infections relative to certain jobs and industries. I hypothesize that we will observe industries exhibiting higher rates of covid infections and deaths than others, giving rise to a question of why workers in that industry appear more susceptible to the effects of covid than workers in other industries.

RongXin

	Name	Definition	Data Type	Possible Values
	County	5-character FIPS code	Object	FIPS code
	Sex	Male/Female	String	Male/Female
	Ages	Ages of the population	Int64	1-99
	Race	Different type of Race	String	White,Black or African American,Asian
	Citizen, voting Age	Citizen that is over 18 years	Int64	U.S Citizen
	Population	Population of the State	Int64	Total of the number in the state

To merge the enrichment data with the COVID-19 dataset, we will use geographic identifiers. In this specific case, the “CountyFIPS” code serves as an ideal common variable for merging. This code is a unique identifier for each county in the United States, ensuring accurate matching between the datasets. And we are using pandas to combine and merge using the “CountyFIPS”

Is there a positive correlation between population density and the rate of COVID-19 transmission? The census enrichment data enables an analysis of the population in each county, allowing us to examine whether counties with larger populations experience higher rates of COVID-19 transmission compared to those with smaller populations. I hypothesize that countries with larger populations will exhibit higher rates of COVID-19 infections and deaths, prompting an investigation into how and why higher population density might influence the increased transmission rates of COVID-19.

Sarvesh Sridher

Names	Definition	Data Type	Possible Values
State	Shows the states	String	Delaware, Florida, Alaska
Country	Shows the country in the state	String	Kent County, Charlotte County
Candidate	Holds the candidates names	String	Joe Biden, Donald Trump, Write-ins
Party	Shows what party the candidate is applying for	String	DEM, REP, LIB, GRN, WRI
Total_votes	Contains the amount of votes casted for each candidate	Integers	405, 100, 005, 293
Won	Shows which candidate has won	Boolean	True, False

We merged our primary Covid-19 data by using the variable 'County Name' in Covid-19 dataset and 'county' in the president candidate dataset to merge the datasets. This will specify to use the left dataframe of 'County Name' and the right dataframe 'county' to use for merging.

Question: What is the likely hood that a person voting for a certain political candidate has covid?

Hypothesis: I hypothesize that we will observe a group of democrats, republicans, green party, etc. and check if one of the candidates have a higher likelihood of getting covid. If one of the candidates in the democrat group, or example, has covid, we can conclude that the people that voted for this candidate have a higher change of getting covid. The president candidate enrichment data helps analyze the covid spread by allowing us to see the rates of the covid infections and deaths in the counties.