Stage 1 Report

Covid-19 Dataset and Datatype:

Variable Name:	Date Type:	Description:
countyFIPS	Integer	Unique index that identifies County Name.
County Name	String	Name of Counties in the United States.
State	String	Name of States in the United
stateFIPS	Integer	Unique identifier that identifies States
Population	Integer	Number of People living in named County.
Date(1/22/20-9/14/20)	Integer	Number of confirmed Cases/Deaths of selected date.

The three COVID-19 datasets of cases, deaths, and population are the three main datasets of analyzing the spread of COVID-19. The “County by Population” dataset includes important variables such as “County Name”, “States”, “Population”, and “countyFIPS”. This dataset shows the population of each county and will give us insight on how COVID-19 spreads in different populated counties when merged with the other datasets. The “Cases” dataset shows individual county cases per day since January 22, 2020 to present day. The “Deaths” dataset is very similar to the cases dataset as it shows individual county deaths that are caused by COVID-19 per day since January 22,2020. Each column of both “cases” and “deaths” dataset represent the days of this year and gives us a better way to analyze the trend of COVID-19. We can use the dates to calculate the trend of cases if they are increasing or

decreasing. All three datasets have a “countyFIPS” variable which matches a specific county. Since all datasets have the variable “countyFIPS”, “County Name”, “State”, and “StateFIPS”. We can merge the three datasets into a super COVID-19 dataframe by linking these variables together in a python notebook.

Enrichment Data Analysis:

Demographic(Sanam Khalili):

Variable Name	Data type	Description
countyFIPS	Integer	Unique index that identifies County Name.
Estimate!!SEX AND AGE!!Total population!!Under 10 years	Integer	Total population of under 10 yrs
Estimate!!SEX AND AGE!!Total population!!10 to 19 years	Integer	Total population of 10-19 yrs
Estimate!!SEX AND AGE!!Total population!!20 to 34 years	Integer	Total population of 20-34 yrs
Estimate!!SEX AND AGE!!Total population!!35 to 44 years	Integer	Total population of 35-44 yrs
Estimate!!SEX AND AGE!!Total population!!45 to 54 years	Integer	Total population of 45-54 yrs
Estimate!!SEX AND AGE!!Total population!!55 to 64 years	Integer	Total population of 55-64 yrs
Estimate!!SEX AND AGE!!Total population!!65 to 74 years	Integer	Total population of 65-74 yrs
Estimate!!SEX AND	Integer	Total population of 75-84

AGE!!Total population!!75 to 84 years		yrs
Estimate!!SEX AND AGE!!Total population!!85 years and over	Integer	Total population of 85 yrs and over
Estimate!!SEX AND AGE!!Total population!!65 years and over!!Female	Integer	Total population of female over 65 yrs
Estimate!!SEX AND AGE!!Total population!!65 years and over!!Male	Integer	Total population of male over 65 yrs
Estimate!!SEX AND AGE!!Total population!!Female	Integer	Total population of female
Estimate!!SEX AND AGE!!Total population!!Male	Integer	Total population of male
Estimate!!RACE!!Total population!!One race!!White	Integer	Total population of whites
Estimate!!RACE!!Total population!!One race!!Black or African American	Integer	Total population of African Americans
Estimate!!HISPANIC OR LATINO AND RACE!!Total population	Integer	Total population of Hispanics
Estimate!!RACE!!Total population!!One race!!Asian	Integer	Total population of Asians

This dataset includes the following information for each county of states : total population of sex, total population of each age group and total population of different reces. For choosing important information from the demographic dataset, I calculated the correlation between each column of the demographic dataset with total number of cases and total number of deaths. Then, I selected several of them. This table shows the columns that I selected from

the available information. The main reason and criteria to select these columns is to relize which age group, sex and race are more prone to COVID19. Moreover, I was interested to find any relationship between gender and death rate with respect to COVID19 fatality. This could also help us find out if the counties with higher percentage of young people compared to elderly, have lower confirmed cases and deaths due to COVID19. The same inquiry applies to the influence of race on confirmed cases and deaths caused by COVID19. In order to merge the selected dataset with the main covid dataset, county FIPS is used because it is unique for each county.

Social (Isaac Taylor):

Variable Dictionary

Variable Name:	Date Type:	Description:
COUNTY_FIPS	Integer	An identification code for a specific county
GEO_ID	String	A geographic identifier that contains a county's FIPS code and additional data.
Name	String	The geographic name and state of a location.
DPO2_0001E	Integer	Estimate of the total households in a county
DP02_0002E	Integer	Estimate of total family households in a county
DP02_0005PE	Integer	Estimate of the total Married-couple family households with children under 18 years
DP02_0001E	Integer	Estimate of nonfamily households where householder lives alone
DP02_0015E	Float	Average household size
DP02_0037E	Integer	Number of women 15-50

		years old who had birth in the past 12 months
DP02_0057E	Integer	Estimate of people enrolled in college or graduate school

DP02_0123E	Integer	Ancestry Total population American
DP02_0123E	Integer	Ancestry Total population Arab
DP02_0125E	Integer	Ancestry Total population Czech
DP02_0126E	Integer	Ancestry Total population Danish
DP02_0127E	Integer	Ancestry Total population Dutch
DP02_0128E	Integer	Ancestry Total population English
DP02_0129E	Integer	Ancestry Total population French (except Basque)
DP02_0130E	Integer	Ancestry Total population French Canadian
DP02_0131E	Integer	Ancestry Total population German
DP02_0132E	Integer	Ancestry Total population Greek

DP02_0133E	Integer	Ancestry Total population Hungarian
DP02_0134E	Integer	Ancestry Total population Irish
DP02_0135E	Integer	Ancestry Total population Italian

DP02_0136E	Integer	Ancestry Total population Lithuanian
DP02_0137E	Integer	Ancestry Total population Norwegian
DP02_0138E	Integer	Ancestry Total population Polish
DP02_0139E	Integer	Ancestry Total population Portuguese
DP02_0140E	Integer	Ancestry Total population Russian
DP02_0141E	Integer	Ancestry Total population Scotch-Irish
DP02_0142E	Integer	Ancestry Total population Scottish

DP02_0143E	Integer	Ancestry Total population Slovak
DP02_0144E	Integer	Ancestry Total population Subsaharan African
DP02_0145E	Integer	Ancestry Total population Swedish
DP02_0146E	Integer	Ancestry Total population Swiss
DP02_0147E	Integer	Ancestry Total population Ukranian
DP02_0148E	Integer	Ancestry Total population Welsh
DP02_0149E	Integer	Ancestry Total population West Indian

Merging the data with the primary COVID-19 dataset:

The Social and COVID-19 datasets both have the county fips variable. This means that the two datasets can be merged by essentially performing a left join where the COVID-19 dataset is on the left the Social dataset is on the right. By doing this, all the values from the Social dataset will be joined together with their respective match in the COVID-19 dataset.

Enrichment Data Description:

The subset of data that has been selected from the social dataset contains information about some of the different types of households, a fertility estimate, a higher education enrollment estimate, and a variety of different ancestral population estimates for specific geographical areas . This information is likely to be meaningful in regards to the analysis of COVID-19 in different counties and states because it describes some of the social and cultural aspects that show how and where groups of people live. Furthermore, some hypotheses that can be formulated based on this dataset include the following:

1. Greater numbers of traditional family households in an area will indicate less cases than areas with a larger number of single householder households.
2. Areas with larger average household sizes will positively correlate to more COVID-19 cases.
3. The higher the fertility rate of an area, the more COVID-19 cases there will be.
4. The higher the number of college students in an area, the more COVID-19 cases there will be.
5. People of different Ancestral backgrounds will show varying degrees of COVID-19 deaths when compared to each other.

Since this data is helpful in showing more or less how groups live across specific U.S states and counties, this information should be helpful to determine which population(s) of people may be more at risk.

Economic(Ali Altamimi):

Variable Name:	Date Type:	Description:
DP03_0001E	Integer	Estimate!!EMPLOYMENT STATUS!!Population 16 years and over
DP03_0008E	Integer	Estimate!!EMPLOYMENT STATUS!!Civilian labor force
DP03_0018E	Integer	Estimate!!COMMUTING TO WORK!!Workers 16 years and over
DP03_0026E	Integer	Estimate!!OCCUPATION!!Civilian employed population 16 years and over
DP03_0032E	Integer	Estimate!!INDUSTRY!!Civilian employed population 16 years and over
DP03_0046E	Integer	Estimate!!CLASS OF WORKER!!Civilian employed population 16 years and over
DP03_0047E	Integer	Estimate!!CLASS OF WORKER!!Civilian employed population 16 years and over!!Private wage and salary workers
DP03_0048E	Integer	Estimate!!CLASS OF WORKER!!Civilian employed population 16 years and over!!Government workers
DP03_0027E	Integer	Estimate!!OCCUPATION!!Civilian employed population 16 years and over!!Management, business, science, and arts occupations
DP03_0028E	Integer	Estimate!!OCCUPATION!!Civilian employed population 16 years and over!!Service occupations
DP03_0029E	Integer	Estimate!!OCCUPATION!!Civilian employed population 16 years and over!!Sales and office occupations
DP03_0030E	Integer	Estimate!!OCCUPATION!!Civilian employed population 16 years and over!!Natural resources,

		construction, and maintenance occupations
DP03_0031E	Integer	Estimate!!OCCUPATION!!Civilian employed population 16 years and over!!Production, transportation, and material moving occupations
DP03_0032E	Integer	Estimate!!INDUSTRY!!Civilian employed population 16 years and over
DP03_0033E	Integer	Estimate!!INDUSTRY!!Civilian employed population 16 years and over!!Agriculture, forestry, fishing and hunting, and mining
DP03_0034E	Integer	Estimate!!INDUSTRY!!Civilian employed population 16 years and over!!Construction
DP03_0035E	Integer	Estimate!!INDUSTRY!!Civilian employed population 16 years and over!!Manufacturing
DP03_0036E	Integer	Estimate!!INDUSTRY!!Civilian employed population 16 years and over!!Wholesale trade
DP03_0037E	Integer	Estimate!!INDUSTRY!!Civilian employed population 16 years and over!!Retail trade
DP03_0038E	Integer	Estimate!!INDUSTRY!!Civilian employed population 16 years and over!!Transportation and warehousing, and utilities
DP03_0039E	Integer	Estimate!!INDUSTRY!!Civilian employed population 16 years and over!!Information
DP03_0040E	Integer	Estimate!!INDUSTRY!!Civilian employed population 16 years and over!!Finance and insurance, and real estate and rental and leasing
DP03_0041E	Integer	Estimate!!INDUSTRY!!Civilian employed population 16 years and over!!Professional, scientific, and management, and administrative and waste management services
DP03_0042E	Integer	Estimate!!INDUSTRY!!Civilian employed population 16 years and over!!Educational services, and health care and social assistance
DP03_0043E	Integer	Estimate!!INDUSTRY!!Civilian employed population 16 years and over!!Arts, entertainment, and recreation, and accommodation and food services
DP03_0044E	Integer	Estimate!!INDUSTRY!!Civilian employed

		population 16 years and over!!Other services, except public administration
DP03_0045E	Integer	Estimate!!INDUSTRY!!Civilian employed population 16 years and over!!Public administration

The economic dataset has a lot of different data that can help us monitor the covid-19 cases and death. Depending on how many people are working in a specific industry that can give us a clue on how the covid-19 transfer. The covid19 dataset and economic enrichment data have the FIPS in common where we can make the connection between the two datasets.

Describing the variables:

We can count the number of people who go to work and depending on the industry we can connect the number of cases and then compare it with the number of employment and see if there is a carolations. The variables are the number of employment, employment based on industry and class.

Housing(Ali Altamimi):

Variable Name:	Date Type:	Description:
GEO_ID	String	It contain the fips number
DP04_0001E	Integer	Total housing units
DP04_0002E	Integer	Occupied housing units
DP04_0003E	Integer	Vacant housing units
DP04_0004E	Integer	Homeowner vacancy rate
MORTGAGE STATUS!!		
DP04_0090E	Integer	Owner-occupied units
DP04_0091E	Integer	Housing units with a mortgage

DP04_0092E	Integer	Housing units without a mortgage
GROSS RENT		
DP04_0126E	Integer	Occupied units paying rent
DP04_0127E	Integer	Less than $500
DP04_0128E	Integer	$500 to $999
DP04_0129E	Integer	$1,000 to $1,499
DP04_0130E	Integer	$1,500 to $1,999
DP04_0131E	Integer	$2,000 to $2,499
DP04_0132E	Integer	$2,500 to $2,999
DP04_0133E	Integer	$3,000 or more
DP04_0135E	Integer	No rent paid
GROSS RENT AS A PERCENTAGE OF HOUSEHOLD INCOME (GRAPI)		
DP04_0136E	Integer	Occupied units paying rent
DP04_0137E	Integer	Less than 15.0 percent
DP04_0138E	Integer	15.0 to 19.9 percent
DP04_0139E	Integer	20.0 to 24.9 percent
DP04_0140E	Integer	25.0 to 29.9 percent
DP04_0141E	Integer	30.0 to 34.9 percent
DP04_0142E	Integer	35.0 percent or more

The housing dataset has a lot of different data that can help us monitor the covid-19 cases and death. Depending on how many people are working in a specific industry that can give us a clue on how the covid-19 transfer. The covid19 dataset and housing enrichment data have the FIPS in common where we can make the connection between the two datasets.

Describing the variables:

We can count the number of people who live in houses and see how many of them are paying rent or not. From that we can check if the pandemic affected their rent payment.

Employment(Nadia):

Data	**DataType**	**Data description**
countyFIPS	**Integer**	**Counties FIPS code**
Establishment Count	**Integer**	**Quarterly establishment counts for a given quarter**
January Employment	**Integer**	**Employment level for the first month of the quarter**
February Employment	**Integer**	**Employment level for the second month of the quarter**
March Employment	**Integer**	**Employment level for the first month of the quarter**

The employment data set gives basic information about the employment level of a county based on the three months in a given quarter. For the analysis, four variables are relevant as they give useful information that can be related to the COVID dataset. From these variables, we can see how COVID19 affects the level of employment since it started (e.g: how many employees lost their job).

Hospital Beds(Harinder Badesha):

Variable Name:	Data Type:	Description:
FIPS	**Integer**	Unique identifier which identifies counties
STATE_NAME	**String**	Name of State in the United States.

COUNTY_NAME	**String**	Name of county.
NUM_LICENSED_BEDS	**Integer**	maximum number of beds for which hospitals hold a license to operate.
NUM_STAFFED_BEDS	**Integer**	Number of beds that are currently maintained in the hospital.
NUM_ICU_BEDS	**Integer**	Number of Intensive Care Unit Beds.
ADULT_ICU_BEDS	**Integer**	additional intensive care beds to supplement an influx of patients, used in emergency situations.
PEDI_ICU_BEDS	**Integer**	combination of neonatal, pediatric and premature ICU beds
BED_UTILIZATION	**Integer**	Total Patient Days/ Bed Days available.
Potential_Increase_In_Bed_Capac:	**Integer**	Provides insight to a scenario if bed capacity needs to be increased and computes the potential number of increases.
AVG_VENTILATOR_USAGE:	**Integer**	Average number of patients on a ventilator per week.

The Hospital Beds enrichment dataset has a lot of information on the number of beds and bed usage per hospital in the United States. With this dataset, we can analyze how severe covid cases are in each county. We can analyze and see if there is any correlation between the number of hospital beds licensed in a hospital and covid deaths. If a county has a lower amount of hospital beds, are the covid deaths higher than usual? What about the opposite? Does a higher number of staffed beds equal to lower covid deaths in that county? Does higher ventilar usage equal to lower amount of covid deaths? These are the questions we can ask

when we look at the hospital beds dataset alongside the covid-19 dataset. When merging the two datasets together, we can look at the FIPS in the hospital beds dataset which equates to countyFIPS for the covid-19 dataset. In the hospital bed dataset, We would have to change the variable name “FIPS” to “countyFIPS” so it can merge with the COVID-19 dataset with no issues. Since the hospital beds dataset has numerous hospitals in one county, we have duplicate county names with several data within. We would have to add all the integer variables above to only one county. What we can do is organize the hospital beds dataset by hierarchically indexing the county name, FIPS, and the state. We would have all the counties' names and unique FIPS under one state and then group every other integer variable with pandas by adding them up within one county name and FIPS and then we can successfully merge the two datasets together.