UNCG-CSC 405-605 (Data Science) Spring 2022

PROJECT - COVID DATA ANALYSIS

Team 3

INTRODUCTION:

In this project, we will learn how to preprocess and merge datasets to identify and calculate measures to draw meaningful and insightful data conclusions. To provide an accurate record of critical metrics related to the Covid19, our team will be working with a primary COVID-19 dataset developed by a team at U.S.A Facts in collaboration with state and local governments. The datasets we are using contain confirmed cases of COVID-19 per day over a specified period and confirmed deaths caused by COVID-19 over the same period, as reported by each county in every state in the United States of America. We will also use a third data set containing the population of each county to understand the rates of infections and deaths better when these are compared. Each member of our team will also be analysing separate enrichment datasets that they have chosen from various datasets. These datasets will aid in determining what factors may contribute to the spread of COVID-19 and the likelihood of an individual becoming infected with COVID-19 and dying because of infection. These datasets will then be combined with our primary dataset, as described above, to provide additional information for analysis.

Team Task

COVID-19 Datasets Variable Description

Covid-19 Confirmed Cases Dataset:

This dataset contains the count of confirmed Covid-19 cases in every county from all the states across the USA for every day starting from 01/22/2020 to 08/16/2021.

Datatype variable dictionary:

Name	Datatype	Description
countyFIPS	Integer	The unique ID for every county
CountyName	Object	Name of the county
State	Object	Name of the state
stateFIPS	Integer	The unique ID for every state.
Number of confirmed cases	Integer	Confirmed covid-19 cases from 01/22/2020 to 08/16/2021.

Number of Covid-19 deaths dataset:

This dataset contains the count of Covid-19 deaths in every county from all the states across the USA for every day starting from 01/22/2020 to 08/16/2021.

Datatype variable dictionary:

Name	Datatype	Description
countyFIPS	Integer	The unique ID for every county
CountyName	Object	Name of the county
State	Object	Name of the state
stateFIPS	Integer	The unique ID for every state.
Number of confirmed cases	Integer	Confirmed covid-19 cases from 01/22/2020 to 08/16/2021.

County Population Dataset:

This dataset contains the population of each county across the USA.

Datatype variable dictionary:

Name	Datatype	Description
countyFIPS	Integer	The unique ID for every county
CountyName	Object	Name of the county
State	Object	Name of the state
Population	Integer	Number of people in each county.

Preliminary Intuitions from the three Datasets:

The variables State, County name and County FIPS are common in all the three COVID-19 datasets. By merging these datasets, we can easily analyze the information and say which states and which counties in each state has more infections and fatalities per capita. We can get month-by-month statistics on infections and deaths, such as which months of the year (January 2020 to March 2021) have the most COVID-19 related infections and deaths.

Member Tasks

SAIPAVAN TADIKONDA (Team 3)

TASK-2

Census Demographic ACS:

The Census Demographic ACS dataset contains demographic information of all states of the United States by county level from the United States Census Bureau site for the year 2019. The demographic information includes population estimations based on sex, age (different age groups), races (different races), eligible voters based on sex.

Variable dictionary:

Column Variables	Data types	Variable description
countyFIPS	int64	Unique five-digit code for every county
County Name	object	Name of the county
State	object	Name of the State
Total population	object	Estimation of total population
Males	object	Estimation of total male population
Females	object	Estimation of total female population
Sex ratio (males per 100 females)	object	Estimated ratio of males per 100 females
Under 5years	object	Total population estimate of ages under 5years
5-9years	object	Total population estimate of ages from 5-9 years
10-14years	object	Total population estimate of ages from 10-14years
15-19years	object	Total population estimate of ages from 15-19years
20-24years	object	Total population estimate of ages from 20-24years
25-34years	object	Total population estimate of ages from 25-34years
35-44years	object	Total population estimate of ages from 35-44years
45-54years	object	Total population estimate of ages from 45-54years
55-59years	object	Total population estimate of ages from 55-59years
60-64years	object	Total population estimate of ages from 60-64years

65-74years	object	Total population estimate of ages from 65-74years
75-84years	object	Total population estimate of ages from 75-84years
85years and over	object	Total population estimate of ages 85years and over
Median age(years)	object	Median age (years)
Under 18years	object	Total population estimate of age under 18years
18years and over	object	Total population estimate of age 18years and over
18years and over Male	object	Total population estimate of age 18years and over (Male)
18years and over Female	object	Total population estimate of age 18years and over (Female)
Sex ratio (18years and over)	object	Estimated Sex ratio of 18years and over
White population	object	Total population estimate of White population
Black population	object	Total population estimate of Black population
Asian population	object	Total population estimate of Asian population
Voting population	object	Estimate of Voting population
Voting population (Male)	object	Estimate of Voting population (Male)
Voting population (Female)	object	Estimate of Voting population (Female)

Merge with Covid Dataset:

The census demographic dataset has around 358 columns. I considered the columns of population estimations based on sex, age groups, different races to merge with covid dataset which would be further helpful in performing analysis related to covid spread.

I observed GEO_ID and Name column variables in the enrichment dataset. The GEO_ID has the unique five-digit countyFIPS code and the Name column has both the County Name and the State name. These two columns in enrichment dataset are in common with the covid dataset. So, I extracted the countyFIPS code from GEO_ID using slice and I obtained the County Name and State by using split. Finally, using the countyFIPS, County Name, State I merged both enrichment dataset and covid dataset.

Importance of Census Demographic enrichment data in Covid analysis:

This enrichment data helps us to know which age groups are mostly affected across different counties. Also, which gender got mostly affected with the covid spread. It also helps us to know which races were affected with this covid spread. Some awareness can be shared among people through government, private organizations and NGOs with this analysis.

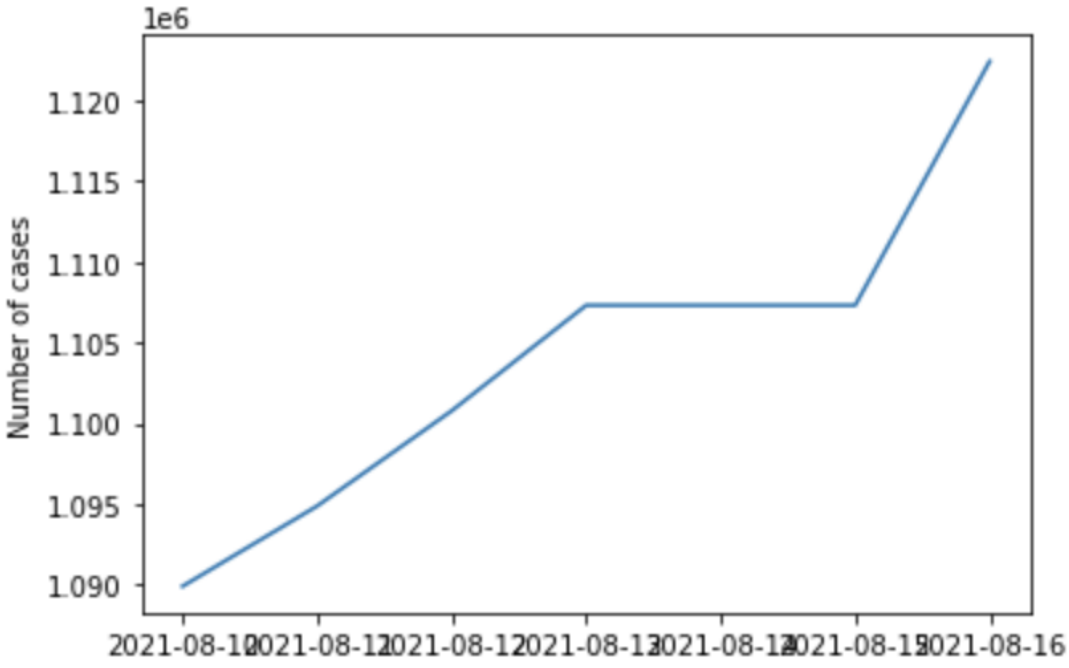
INITIAL HYPOTHESIS QUESTIONS:

1) Which age group was badly affected by this virus?
2) Which race was mostly affected with this virus?
3) Which gender was more prone to the virus?
4) Is there more death rate among the higher age groups?
5) Were there high cases among the children under 5 years?
6) Which county was mostly affected with this virus?
7) Which state has a smaller number of cases compared to other states?
8) Does the chance of getting infected increases with the increase in age?

TASK-3

NC state Last week covid cases trends:

I chose the North Carolina State for observing the last week trends of the covid cases i.e., from 10th August 2021 to 16th August 2021.



From the above plot we can say that the cases increased initially from 10th August 2021 to 12th August 2021 and cases remained stable for the next three days i.e., from 13th August 2021 to 15th August 2021. But on the last day i.e., on 16th August 2021 cases increased.

Aman Tej Vidapu, Graduate Student

ACS Economic Survey Data:

This enrichment dataset contains the economic information of every state by county level from the American Community Survey for the year 2019. The Information includes Modes of transportation used by people, Occupation details, Health insurance information, Households and their income and Per capita income.

Variable Dictionary:

~Mentioned datatypes are Initial datatypes.

Column Variable	**Datatype**	**Variable description**
GEO_ID/countyFIPS	Object/Int64	Geography id and the last five digits contains the fips code
County Name	Object	Name of county
State	Object	Name of State
Total people commuting to work	Object	Total Number of people commuting to work
Driving alone	Object	Total Number of people driving to work alone
Carpooled	Object	Total Number of people sharing rides while going to work
Only public transports	Object	Total Number of people commuting to work by public transport (other than taxi)
Walked	Object	Total Number of people commuting to work by walking
Other means	Object	Total Number of people commuting to work by other means
Work from home	Object	Total Number of people working from home

Mean travel time to work	Object	Mean time taken to travel
Total people has occupation	Object	Total Number of people having an occupation
Management&business&science&arts occupations	Object	Total Number of people having an occupation in Management, business, science and arts fields
Service occupations	Object	Total Number of people having an occupation in Service field
Sales and office occupations	Object	Total Number of people having an occupation in Sales and office fields
Natural resources, construction, and maintenance occupations	Object	Total Number of people having an occupation in Natural resources, construction, and maintenance fields
Production, transportation, and material moving occupations	Object	Total Number of people having an occupation in Production, transportation, and material moving fields
Private health insurance	Object	Number of people having private health insurance
Public health insurance	Object	Number of employed people having public health insurance
No health insurance object	Object	Number of employed people having no health insurance
Total households	Object	Total number of households in the county
Mean household income	Object	Mean household income of the county
Per capita income	Object	Per capita income for each county

How to Merge:

This particular enrichment data have more than hundreds of columns. So, we have to read and understand the data before merging with the super covid dataset. I considered estimates

of commuting to work, health insurance, occupations, household and per capita income-related columns to merge with the super covid dataset to perform further analysis. Also, thought about what hypothesis can be made by keeping these particular columns.

To merge, I used (County Name, State, countyFIPS) as common columns in both datasets. In enrichment data GEO_ID has the last five digits as county fips, has to extract there and NAME has both county name and state, has to split one column to two required columns. So that we can perform the merge.

Importance of Economic enrichment data in Covid analysis:

By using the household income, modes of transport, insurance status, occupation details combined with the super covid dataset, we can understand which economic state people are more prone to covid and who are the reason for the spread of the virus. So that, moderate rules can be implemented as per the observations made to stop the spread.

Hypothesis Questions:

1. Commuting time plays a role in the spread of covid and by which means of transport we are more prone to covid?
2. Which occupational fields are more prone to covid. Closed workspaces are more secured than onfield workspaces?
3. Counties with a fewer number of health insurances has more covid spread?
4. Does private/public insurance makes the difference in the spread and which insurance has more difference.
5. Counties with low mean household incomes are more prone to covid spread than high mean household incomes?
6. Does Per capita Income play a role in the spread of covid?
7. People who walk to work are more prone to covid spread?
8. Which occupation field has the potential to increase the covid spread.

Viveka Reddy Erram, Graduate Student

Employment Dataset-

Description on the Employment enrichment dataset:

This data set (Employment dataset) provides the information about the level of employment, ownership and establishment count by county level in every state in the United States for the month of January, February and March of the year 2021.By merging this dataset with COVID-19 dataset we can analyze and observe how the employment is affecting covid-19 cases based on the type of Industry, ownership, average weekly wages. Employment plays a key role while calculating the Covid-19 cases. Based on different categories of industrial employment also we see the increase and decrease in the covid-19 cases.

Variables and Data type table:

Name	Data type	Description
Area	object	Shows state name and area name

Area\nCode	int64	Unique ID for each county
Area Type	object	Has area in county
Year	int64	2021
Own	int64	Unique code for ownership
Ownership	object	This shows the employment details of Federal, state, local government or private sectors.
Industry	object	To see the employment details of each industry (Information, Manufacturing, trade)
State Name	object	Shows the name of the state
Establishment Count	int64	Number of people employed in each establishment.
January Employment	int64	Number of people who were employed in the month of January 2021.
February Employment	int64	Number of people who were employed in the month of February 2021.
March Employment	int64	Number of people who were employed in the month of March 2021.
Total Quarterly Wages	int64	Total quarterly wages in a county based on each industry.
Average Weekly Wage	int64	Average weekly wage of the county based on industry.
Employment Location Quotient Relative to U.S.	Float64	Employment Location quotient compared to the total country.
Total Wage Location Quotient to US	Float64	Wage location quotient compared to the total country.

Merging the employment data set with the COVID-19 dataset:

To merge the enrichment dataset with the COVID-19 data, first process the employment data set. Here we need to remove the inconsistencies in the Area Code column. Select the columns which are useful for the analysis of the data. Then I have merged the employment data set using 'Area Code' variable with the 'countyFIPS' variable in the large covid data set to get Employee covid dataset. I grabbed a sample random set of 500 rows from the data frame and converted it to a csv file.

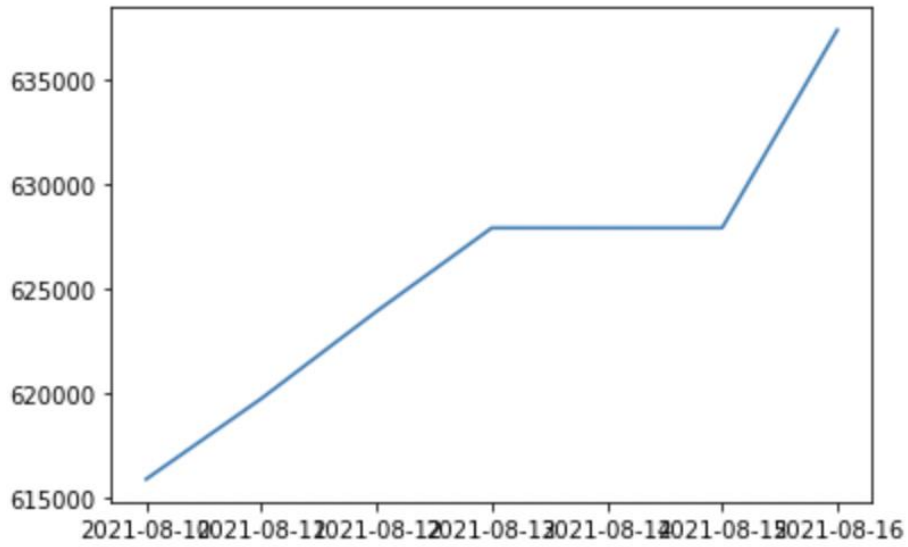
Hypothesis Questions:

I have used the employment dataset and the population of each county to compute the employment rate in each county. This would assist me in determining the employment rate and the link between Covid cases and mortality. Then, based on the deaths/cases, I'd be able to determine which counties, industries, or ownerships were severely affected and which were not.

- Unemployed people in various counties may have a higher number of COVID reports?
- Due to federal reporting requirements, counties with government control may have more reports of COVID -19 cases than counties with private ownership?
- There might be higher number of COVID -19 cases in the Counties which have trade and transportation as employment?
- If there are more medical centers in the areas with high employment rates, that may lead to fewer deaths due to COVID-19?
- Since they can work from home, business management-related industries may have fewer COVID-19 cases?

Analysis on COVID-19 Data set:

Firstly, Calculate the COVID-19 cases confirmed trend for the state "Alabama" for last week which is 2021-08-10 to 2021-08-16. The below graph shows the data visualization of COVID-19 cases trend for the state Alabama.



By analyzing the above graph, we can say that the deaths have constantly increased from the first day (2021-08-10) to 2021-08-13 later the number of deaths are stable from 2021-08-13 to 2021-08-15, from there we can see deaths has rocketed. So, we can say that the death cases have increased.

Varsha Veeramaneni, Graduate Student

Presidential Election Results (Political leanings) –

Description:

This enrichment data set provides information regarding the presidential elections 2020. The dataset has 11 files based on Governors, Senate, President, House representatives. The president county candidate data set provides information on the candidate who received the most votes in the county. This data set mentioned Only a few states. President county election candidate data set have the following variables, and their description is mentioned below.

Datatype Variable Dictionary:

Name	Datatype	Description
state	String	Name of the State
county	String	Name of county
candidate	String	Name of the Person who is standing in the election from that County
party	String	Name of the Political party which they are representing
total_votes	Integer	Number of the votes for the candidate
won	Boolean	Describes whether the candidate win or lost

How to Merge:

The state and county names are common variables in both the election results enrichment data and the Covid-19 dataset. Since the County names are not unique, as they are some repetitive names across different states, we would have to use a combination of state and county columns as the key to performing the merge.

Importance of Presidential Election Results dataset in the analysis of Covid19:

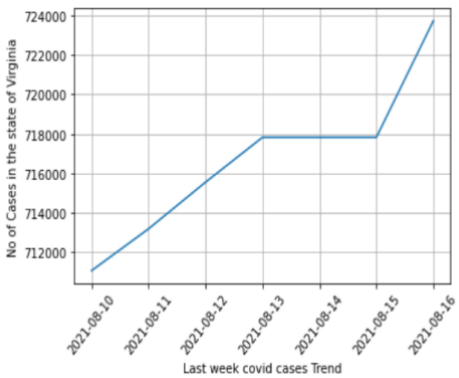
Presidential election results dataset can be useful to analyze how political equations resulted in COVID-19 spread. We can look for a few correlations between a merged dataset of the Covid data and the election results data. We can hypothesize that the spread of covid-19 cases per capita is more remarkable in counties where the percentage of a particular party's voters is more significant. May this could be chosen based on the party's policies and propaganda. We could also Hypothesize that the covid cases are greater in counties with more significant voting percentages as there are greater numbers of people gathering at the polling booths.

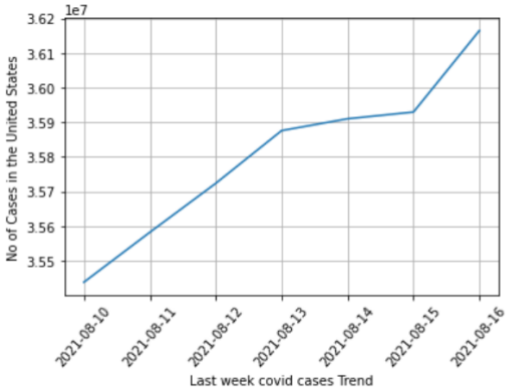
Hypothesis Questions:

- Whether Counties with high population has more covid spread?
- Is COVID-19 cases decreased electoral support for Donald Trump?
- Are the voting percentages decreased due to the rise in Covid cases?
- Is gatherings at the polling booths was the main reason for increasing Covid cases ?
- Covid cases are less in some counties with more significant voting percentages. Can this be based on a general thought that people responsible for voting would also be accountable for practicing social distancing?

Analysing COVID-19 data trends:

For calculating data trends of COVID-19, I observed the number of confirmed cases belonging to the state of Virginia during seven days between August 10th and August 16th. I chose to compare this with the last week's Covid cases data in the whole United States for a more robust analysis of the trends. By measuring these two against each other, I could determine the growth of Covid cases in Virginia state while comparing with all other states. Below are the line graphs depicting the trend for the number of cases over the last week throughout Virginia and the whole United States.





By Observing the above two graphs, I was able to determine that the number of cases saw an exponential increase from August 10th to August 13th. For the next two days, the growth was almost stable, and on the last day of the week, the number of covid19 cases increased drastically.

Lahari Chilakuri, Graduate Student

Hospital Beds Dataset –

Description:

The enrichment dataset I chose to analyze is Hospital Beds dataset. The size of the dataset as of today (i.e., 02/15/2022) is 399863 rows × 109 columns. This dataset provides the information about the information about the hospitals, total number of beds and the available ICU units in the hospital and their average usage in all the counties across USA.

Datatype Variable Dictionary:

Name	Datatype	Description
ccn	Integer	Certification number of hospitals.
hospital_name	Object	Name of the hospital
address	Object	Address of the hospital
city	Object	City in which the hospital is located.
Zip	Object	Zip code of the hospital's physical address.
Hospital_subtype	Object	Type (CA, Children's, long term etc..) of the hospital to which the observation belongs
fips_code	Integer	Unique ID used to identify the observations from county.
total_beds_7_day _average	Float	The average number of total beds in seven days
total_icu_beds_7_ Day_average	Float	Average Total number of ICU beds in seven days
total_adult_beds_7_ Day_average	Float	Number of adult hospital beds in seven days
Total_pediatric_beds_7 _day_average	Float	Number of pediatric hospital beds
Total_staffed_beds_7 _day_average	Float	Number of beds available with staff for the patient who occupies the bed
icu_beds_used_7_day	Float	Toatal number of ICU beds used

Initial intuitions of the dataset:

This dataset can be very useful in determining the reason why the covid cases are increasing/ decreasing. This can be achieved by analyzing the relationship between the total number of covid cases and the number of available ICU or hospital beds. For example, if there are more available hospital beds in a particular city/county/state, there is a smaller number of covid cases in that city/county/state. The other observation that can be made is, if there are not many hospitals in a particular city/county/state, then the number of covid cases can be higher than the average.

How to merge:

Here, the hospital beds dataset has “FIPS_code” column which is the unique five-digit code for each county. Similarly, the large covid dataset which was obtained by merging the three datasets in the team task also has “county_FIPS” column which is unique for each county. The hospital beds dataset has been pre-processed. The rows where the values of FIPS are NaN are dropped in the enrichment hospital beds dataset. By using FIPS_code and county_FIPS from the datasets, the hospital beds dataset can be merged with the large covid dataset using the inner join.

Hypothesis:

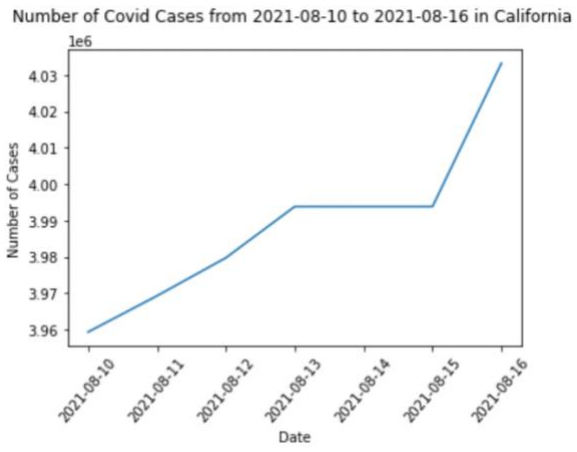
Below are some of the questions we can pose based on the information given in the dataset:

1. Counties with small population may have limited resources to the hospitals which leads to maximum utilization to the hospital beds, can this be the reason in raising the death rates?
2. Whether the counties which have the large number of hospitals eventually results the fewer confirmed covid cases?
3. Whether the number of confirmed covid cases are fewer in the areas which contains a greater number of hospitals when compared to the areas which contains a smaller number of hospitals?
4. As the hospitals are categorized as adult and pediatric, can we analyze the ratio of adults and the children who are confirmed to covid?

COVID-19 Data Trends:

To calculate the COVID-19 data trends, I chose California state to observe the total confirmed covid cases in that state in the last week i.e., from 10th Aug 2021 to 16th Aug 2021.

Below is the graph plot between the total number of covid cases in California state and the dates.



According to the graph, the number of covid cases are steadily increasing in the first four days i.e., from 10th Aug to 13th Aug. Then the number of covid cases got stabilized for the next two days. Later, on the last day, the graph sharply raised.

The graph was plotted by summing the number of covid cases on one day in the last week for every county in the California state. The total number of cases in California state in the last week is as follow:

Dates	No. of covid cases
08-10-2021	3959335
08-11-2021	3969262
08-12-2021	3979714
08-13-2021	3993812
08-14-2021	3993812
08-15-2021	3993812
08-16-2021	4033203