Data science project - Team 4

Stage 1

Analysis of Covid 19 cases and deaths by state in US

Covid – 19 Dataset

The COVID-19 dataset consists of three data sets - Number of cases, Number of deaths and Total population. All this data is based on county level. The number of cases dataset gives detailed information about confirmed cases per day at county level from the time covid pandemic just originated till date. The number of deaths dataset gives information on the number of people dying per day due to pandemic in each county. Total population dataset per se gives the total population living in each county in the United States.

Confirmed cases dataset (covid_confirmed_facts): This dataset has day to day information regarding all the confirmed covid cases that day for every county in the United States. The date starts from Jan 22nd when covid just emerged till August 16th 2021. It also contains corresponding state, stateFIPS and countyFIPS information for each county.

Variables and Data Type table:

Column (Attribute)	**Datatype**
County FIPS (unique code given to each county)	Int64
County Name (names of counties present in each state)	object
State (corresponds to state name where a county is present)	object
State FIPS (nothing but state code)	Int64
All the dates from Jan 22nd 2020 to August 16th, 2021 (It contains number of confirmed cases per day)	Int64

Preliminary Intuitions from Dataset:-

The evident intuition is that we can observe the total number of cases per day for a county. But also when we analyze deeper we can see the trends of increase or decrease patterns throughout the country. We can also limit the analysis to specific counties or specific states we are interested in. We analyzed some cumulative patterns and details are explained below.

- The graph (Figure -1) below shows the rate of increase in cases for all counties per month in an increasing order. After going through the data set, we observed that, at the initial stage i.e starting from 1/22/2020 to 3/18/2020, the cases were 0, and then there was a rise in the number of cases with increasing days.

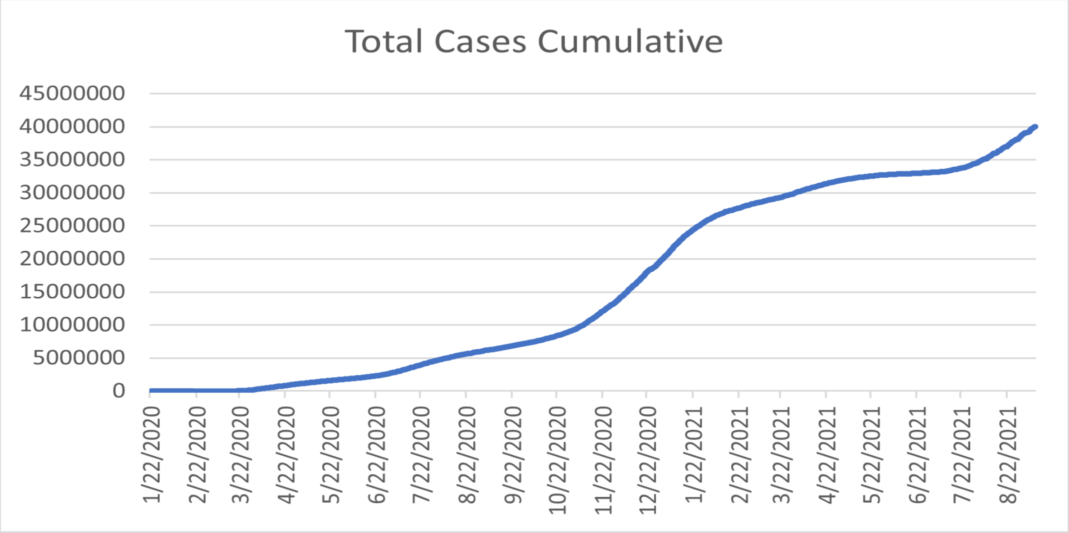


Figure - 1

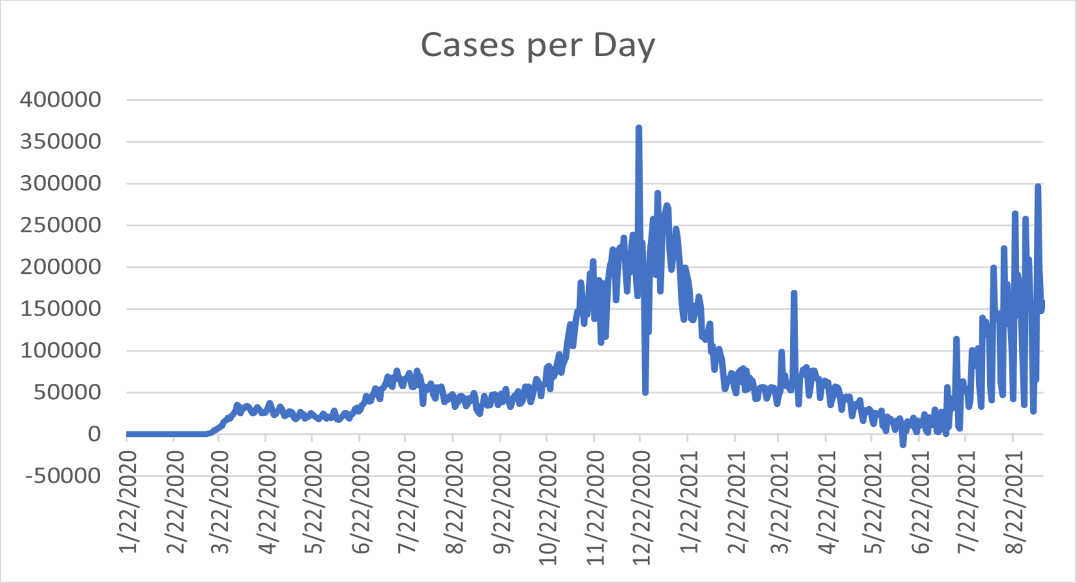


Figure - 2

The graph above shows the number of cases per day.

1. On observation, the graph (Figure -2) shows no cases till mid of March and slowly cases started rising from April 2020 steadily till June and slight increase between June 2020 and July 2020.
2. There was a sudden surge between October 2020 and reached its peak in December 2020 and again it got reduced in a similar way of increasing pattern till beginning of

July 2021 and at some point, it hit 0 cases per day as well. This is the phase where people got vaccinated and maintained social distancing.

3. Even if most of the people got vaccinated, from July 2021 to till date the trend of cases is increasing exponentially due to the delta variant.
4. There are few anomalies in this dataset. It has negative values, which are not supposed to show up as the data is showing only infected cases but not cured cases.

Death's dataset (covid_deaths_usafacts):

This dataset shows the number of deaths occurring across the US at county level. It also contains state, stateFIPS and countyFIPS information for corresponding county.

Variables and Data Type table:

Column	Datatype
County FIPS (unique code given to each county)	Int64
County Name (names of counties present in each state)	object
State (corresponds to state name where a county is present)	object
State FIPS (nothing but state code)	Int64
All the dates from Jan 22nd, 2020, to sep 10th 2021 (It contains number of death cases per day)	Int64

Preliminary Intuitions from Dataset:-

The below graph shows the number of deaths all over the country per month. The death cases follow the pattern of the increase of confirmed cases in Figure - 1. As the cases increased, death of people per month also increased proportionately. The death pattern is increasing because it was a new virus that attacked human kind and Doctors were just getting acquainted with virus behaviour and it's infection level. There was no proper medicine or cure for the people available.

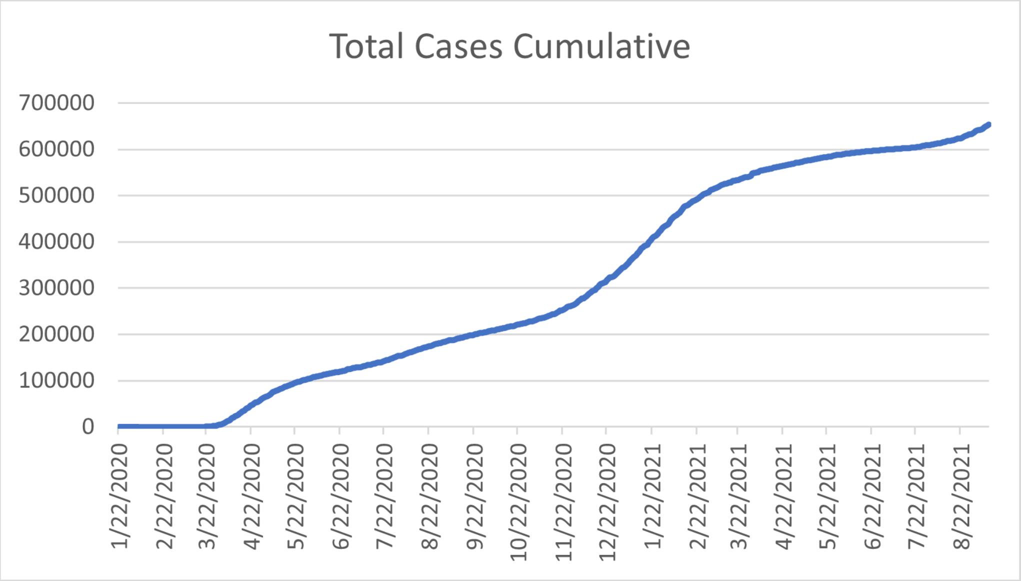


Figure – 3

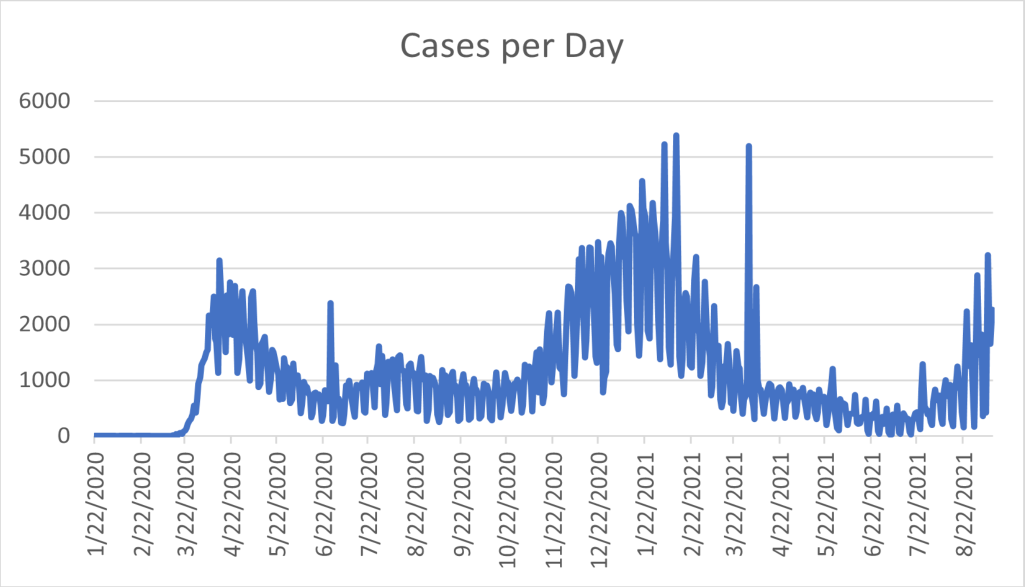


Figure – 4

The above cumulative graph shows there is a sudden increase of deaths from march 2020 and on average there is a rise since then to now. Even though there are some constant curves, overall, there is a high death rate.

1. No of deaths per day shows that there are zero deaths until march 2020 and from then there is an exponential raise in April and slowly no of deaths started decreasing with slight fluctuations till October 2020

2. In October 2020 there is an exponential raise of deaths again till the end of Jan 2021 and started decreasing from Jan through April this is when people started taking vaccinations
3. From April to July it was pretty much flat.
4. From the end of July when the delta variant kicked in, no. of deaths are increasing at a high rate.

Population dataset (covid_county_population_usafacts):

This dataset shows the total population of each county.

Variables and Data Type table:

Column	Datatype
County FIPS (nothing but county code)	Int64
County Name	object
State	object
Population	Int64

Preliminary Intuitions from Dataset:- This is a very critical dataset because it contains the total population of each county. It is very important in analysis of what percentage of people are affected by covid-19 in which county.

Heng Data Set (Hospital Beds Dataset)

Column	**Datatype**	**Description**
State	String	Name of the state
Inpatient beds	Int	Number of beds currently
Inpatient beds used	Int	Number of beds used currently
Inpatient beds used covid	Int	Number of beds used for Covid
53 more variables	All ints	Other data such as shortages, availability,and number of beds by the week

1. The data types for the entire file are all ints except for the State. The state is a String.
2. The variable that could merge these 2 data sets is using the State.
3. This dataset can help with Covid-19 spread because we can get a gauge of how many hospital beds are in use and see how many are not in use. You could possibly move

patients to different hospitals if they have room. My initial hypothesis is that the empty beds should be filled in every state for every case possible in the state. The more beds that are not empty the more patients could be holding on to the bed.

ACS Social, Economic, and Housing

Shipra Shanu

Task 2:

In the ACS Social, Economic and Housing Datasets, there are 3 separate datasets for each one of them for the year 2019.

Data columns and its attributes:

Housing Dataset

Variable name	Variable type	Description
GEOID/ID	INT	Geographical ID of given county name
NAME/Geographical Area name	STRING	County name and the state the county is in.
DP04_0001E to DP04_0143PM	INT	Unique ID referring to combination of multiple observations corresponding to below columns.
Estimate	INT	Estimated Housing Occupancy of each county
Margin of error	FLOAT	Margin of error that could have occurred during counting of various categories of total Housing units.
Percent Estimate	FLOAT	Percentage estimate for various Categories of Housing Occupancy
Percent Margin of error	FLOAT	Percentage margin of error relative to percent estimate for the various categories in Housing Occupancy.

Dataset description:

The data set provides details regarding estimated Total housing occupancy of a county, that could have occurred during counting of various categories of Housing units. i.e. values that could have been miscalculated or over

calculated, also determine what percent of a particular category is present given the housing occupancy, and percentage margin of error.
The initial intuition of the data set is explained as below.

1. Total housing occupancy is divided into various categories like Total housing units, occupied housing units, rental housing units, vacant housing units and so on. For each category, the dataset consists of the number of units of houses distributed.
2. The categories in the dataset are further divided based on “Year the housing was constructed”, “type of rooms in the house”, “number of rooms in the house” etc.
3. All categories are divided into further sub categories to give an inference of the distribution of housing units. One can get an idea of various types of housing units be it be rental or vacant or basis the type of rooms and even on the basis of the year it was built.

Social Dataset

Variable name	Variable type	Description
GEOID/ID	INT	Geographical ID of given county name
NAME/Geographical Area name	STRING	County name and the state the county is in.
DP02_0001E to DP02_0153PM	INT	Unique IDs referring to combination of multiple observations corresponding to below columns.
Estimate	INT	Estimated Households type of each county
Margin of error	FLOAT	Margin of error that could have occurred during counting of various categories of Total Households.
Percent Estimate	FLOAT	Percentage estimate for various Categories of Households.
Percent Margin of error	FLOAT	Percentage margin of error relative to percent estimate for the various categories in Households.

Dataset description:

The data set provides details regarding estimated Households type of each county, margin of error that could have occurred during counting of various categories of Total Households, Percentage estimate for various Categories of Households and the percentage margin of error relative to percent estimate for the various categories in Households.
The initial intuition of the data set is explained as below.

1. Total household type is divided into various categories like Married Couple family, family with children, cohabiting families and so on. For each category, the dataset consists of the number of household types.
2. The categories in the dataset are further divided based on the basis of their sex, their marital statuses, Fertility, Grandparents, School Enrollment etc.

3. All categories are divided into further sub categories to give an inference of the distribution of household type. One can get an idea of various types of household type be it be getting to know about the cohabiting families, estimation of population who have their kids enrolled in kindergarten etc.

Economic Dataset

Variable name	Variable type	Description
GEOID/ID	INT	Geographical ID of given county name
NAME/Geographical Area name	STRING	County name and the state the county is in.
DP05_0001E to DP05_0089PM	INT	These are Unique ID referring to combination of multiple observations corresponding to below columns
Estimate	INT	Estimated employment status for Population 16 years and over for various categories in that of of each county
Margin of error	FLOAT	Margin of error that could have occurred during employment status for Population 16 years and over for various categories of each county
Percent Estimate	FLOAT	Percentage estimate of Employment status for Population 16 years and over for various categories of each county
Percent Margin of error	FLOAT	Percentage margin of error calculated w.r.t margin of error of Employment status for Population 16 years and over for various categories of each county

Dataset description:

The data set provides details regarding estimated Employment status or Population 16 years and over for various categories in each county, margin of error that could have occurred during counting of various categories of Employment status. i.e. values that could have been miscalculated or over calculated, also determine what percent of a particular category is present given the employment status, and percentage margin of error.
The initial intuition of the data set is explained as below.

1. Employment Status is divided into first various categories of people like Females 16 years and over, Civilian labor force, people having ‘all parents in family in labor force’. For each category, the dataset consists of numbers corresponding to the employment statuses.
2. The categories in the dataset are further divided based on being “Employed”, “Unemployed”, “Commutation means”, “Class of workers” etc.
3. All categories are divided into further sub categories to give an inference of the distribution of employment status . One can get an idea of the variety of population employment related categories such

as if they are employed or which class of worker they belong to, also how they commute and many other factors.

Merging the data with the primary COVID-19 dataset

COVID-19 DATA SET	SOCIAL, HOUSING and ECONOMIC DATASET	How to Merge
covid_confirmed_usafacts [countyFIPS countyName State StateFIPS Number of confirmed cases per day]	**Merger of Social, Economic and Housing Dataset basis GEO_ID, County Name and State** [GEO ID, NAME- County Name + State]	We can merge confirmed_cases dataset with Social, Housing and Economic dataset using the GEO_ID which can be renamed as countyFIPS before merging as these attributes are common to both these tables. We can rename GEO_ID to countyFIPS since the data present in these columns are the same to both datasets.
covid_county_population_usafacts [countyFIPS countyName State StateFIPS Population]		We can merge confirmed_cases dataset with Social, Housing and Economic dataset using the GEO_ID which can be renamed as countyFIPS before merging as these attributes are common to both these tables. We can rename GEO_ID to countyFIPS since the data present in these columns are the same to both datasets.
covid_deaths_usafacts [countyFIPS countyName State StateFIPS Number of Deaths per day]		We can merge confirmed_cases dataset with Social, Housing and Economic dataset using the GEO_ID which can be renamed as countyFIPS before merging as these attributes are common to both these tables. We can rename GEO_ID to countyFIPS since the data present in these columns are the same to both datasets.

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

With the help of the Social, Housing and Economic dataset we get insights about the kind of population that lives in a county, their employment status and also their livelihood . We get an idea about different types of housing units, the type of housing units constructed their year and so on. All the information pertaining to Housing units can be acquired for each county. Similarly information can be fetched for each county in regards to economic status I.e. insights about the employment status, class of workers, how they commute and so on. With the help of Social dataset, One can get an idea of various types of household types, be it getting to know about the cohabiting families, estimation of population who have their kids enrolled in kindergarten etc.

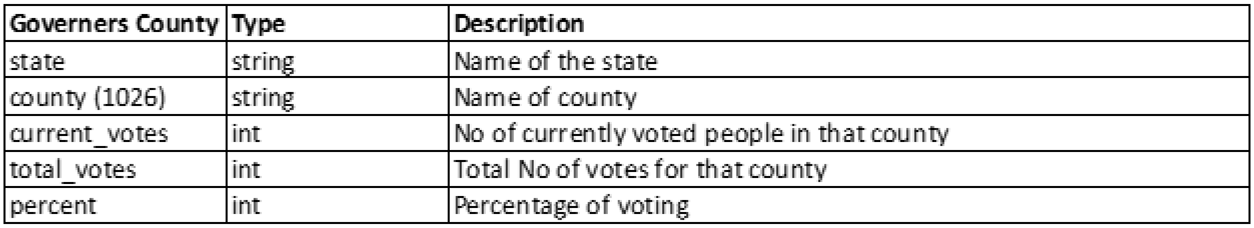
Presidential Election Results Enrichment Dataset

Poojitha Kalidindi

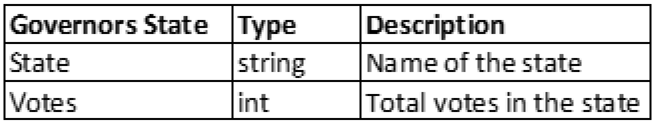
This dataset has 12 files based on Governors, Senate, President, House representatives.

- Governor’s data has 3 files and has state and county in common.
- Senate data has 3 files and have state, county, and total votes fields in common.
- President data has 3 files and have state, county, and total votes fields in common.
- House representative data has 3 files and have district and total votes fields in common.

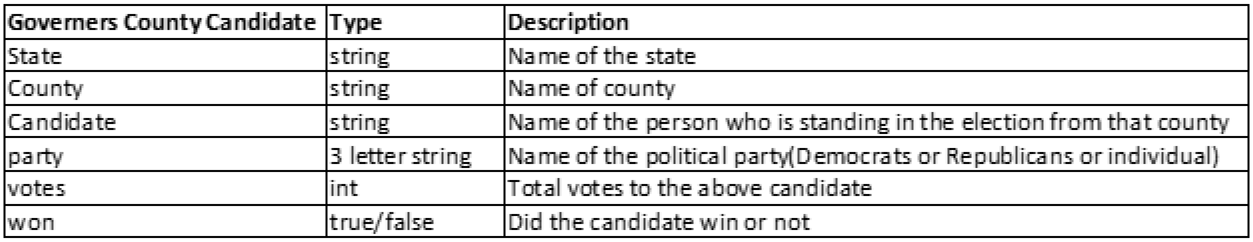
1. **Data types of Enrichment Dataset:**



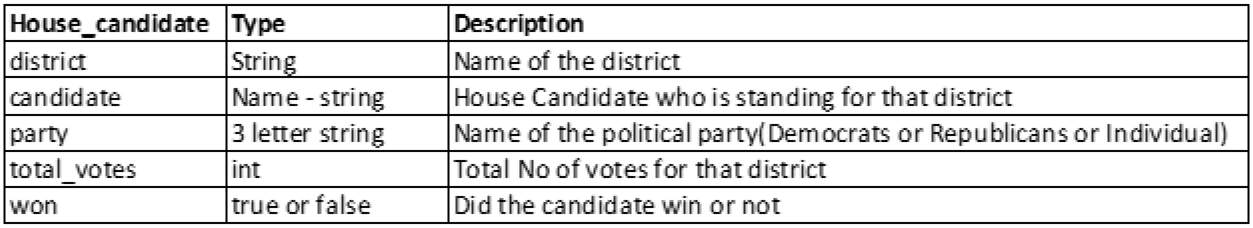
Governers County	Type	Description
state	string	Name of the state
county (1026)	string	Name of county
current_votes	int	No of currently voted people in that county
total_votes	int	Total No of votes for that county
percent	int	Percentage of voting



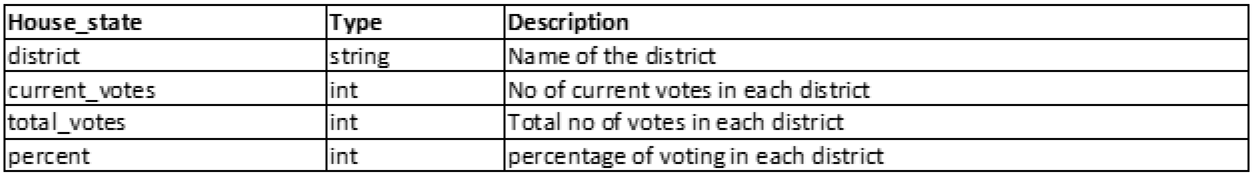
Governors State	Type	Description
State	string	Name of the state
Votes	int	Total votes in the state



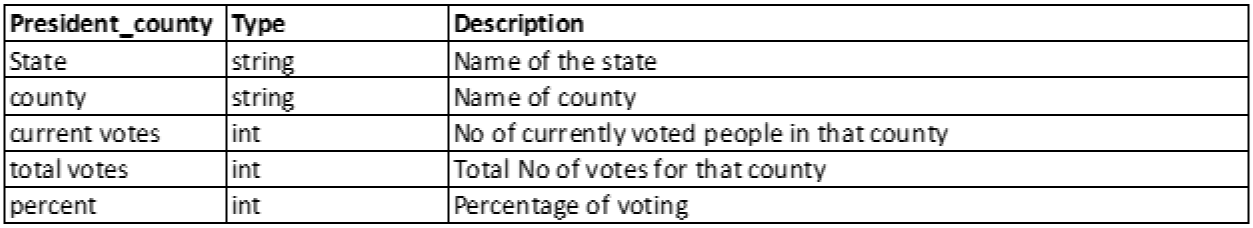
Governers County Candidate	Type	Description
State	string	Name of the state
County	string	Name of county
Candidate	string	Name of the person who is standing in the election from that county
party	3 letter string	Name of the political party(Democrats or Republicans or individual)
votes	int	Total votes to the above candidate
won	true/false	Did the candidate win or not



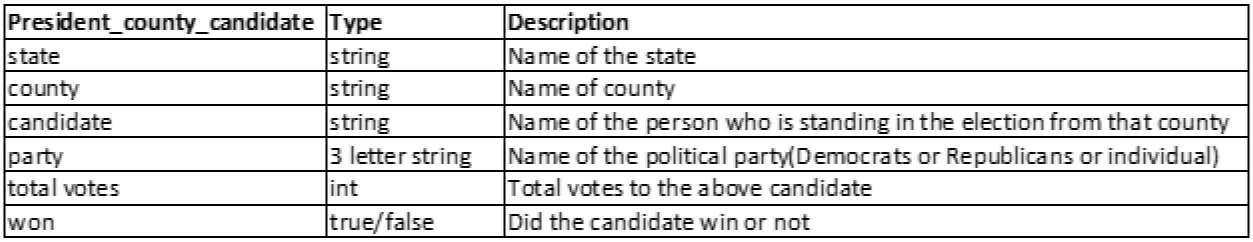
House_candidate	Type	Description
district	String	Name of the district
candidate	Name - string	House Candidate who is standing for that district
party	3 letter string	Name of the political party(Democrats or Republicans or Individual)
total_votes	int	Total No of votes for that district
won	true or false	Did the candidate win or not



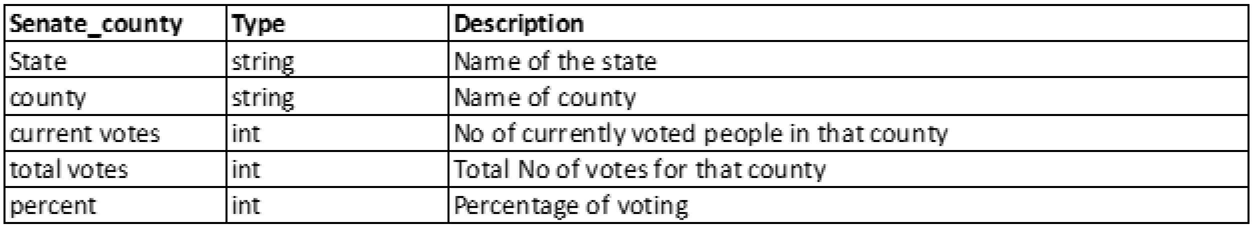
House_state	Type	Description
district	string	Name of the district
current_votes	int	No of current votes in each district
total_votes	int	Total no of votes in each district
percent	int	percentage of voting in each district



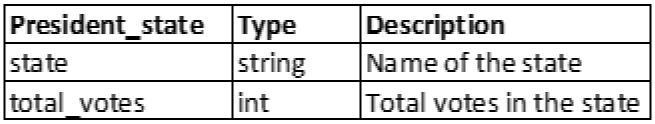
President_county	Type	Description
State	string	Name of the state
county	string	Name of county
current votes	int	No of currently voted people in that county
total votes	int	Total No of votes for that county
percent	int	Percentage of voting



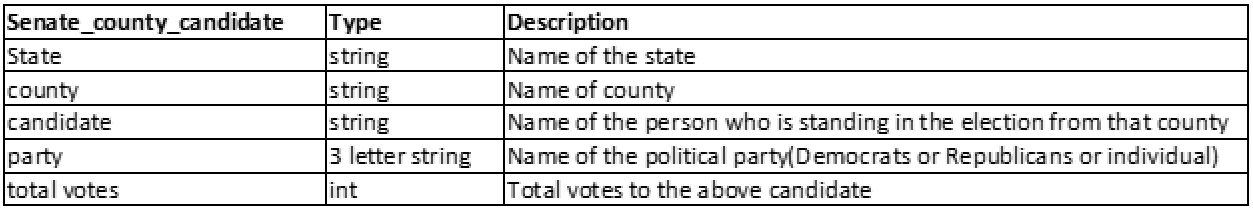
President_county_candidate	Type	Description
state	string	Name of the state
county	string	Name of county
candidate	string	Name of the person who is standing in the election from that county
party	3 letter string	Name of the political party(Democrats or Republicans or individual)
total votes	int	Total votes to the above candidate
won	true/false	Did the candidate win or not



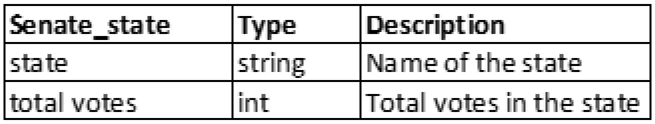
Senate_county	Type	Description
State	string	Name of the state
county	string	Name of county
current votes	int	No of currently voted people in that county
total votes	int	Total No of votes for that county
percent	int	Percentage of voting



President_state	Type	Description
state	string	Name of the state
total_votes	int	Total votes in the state



Senate_county_candidate	Type	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	3 letter string	Name of the political party(Democrats or Republicans or individual)
total votes	int	Total votes to the above candidate



Senate_state	Type	Description
state	string	Name of the state
total votes	int	Total votes in the state

2. The enrichment data of election results and Covid-19 dataset have the common variables of state and county names. 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 perform the merge.

3. Out of all the enrichment datasets for governor, senate, house and presidential elections that were provided, only the candidate level vote dataset has unique values as the county and state datasets are just a subset or summation of the data that we already have on the candidate level dataset. There are some unique values in the county dataset like current and total votes that provide us the completed voting percentage of the county but in this case almost all counties are at 100% as this data was after the completion of elections. Hence, that information cannot be utilized for identifying correlations.

Using a merged dataset of the Covid data and the election results data, we can look for a few correlations between them. We can hypothesize that the spread of covid-19 cases per capita is greater in counties where the percentage of a particular party's voters are greater. This could be chosen based on the policies and propaganda of the party. We could also Hypothesize that the covid cases are greater in counties with low voting percentages. This can be based on a general thought that people who are responsible to vote would also be responsible to practice social distancing. Alternatively, we could also Hypothesize that the covid cases are greater in counties with greater voting percentage as there are greater numbers of people gathering at the polling booths.

ACS Demographic and Housing Estimates

Deepa Jayanna

Task 2:

The *ACS demographic* dataset has two types of data set - one with 5-years of data and another with one year of data which gives county wise Demographics of the United States. I am using a 1-year dataset for the analysis.

Data columns and its attributes:

Variable name	Variable type	Description
GEOID/ID	object	Geographical ID of given county name
NAME/Geographical Area name	object	County name and the state the county is in.
DP05_0001E to DP05_0089PM	object	These are Unique ID referring to combination of multiple observations corresponding to below columns
Estimate	object	Estimated population of each county
Margin of error	object	Margin of error that could have occurred during census counting

Percent Estimate	object	Percentage estimate of a particular group relative to Total population
Percent Margin of error	object	Percentage margin of error calculated w.r.t margin of error

<u>Data description:</u>

The data set gives information about estimated Total population of a county, margin of error that might have happened during census counting i.e. values that could have been missed or over calculated, what percent of a particular category is present given the population count, and percentage margin of error.

The initial intuition of the data set is explained as below.

1. Total population is divided into various categories like sex and age, Race, Race alone or in combination with one or more race, HISPANIC or LATINO AND RACE and citizen, voting age population. Every category has a common term *Total population,* and that population is divided amongst the different subcategories within. At the end each category has the total housing unit variable which gives an idea of the number of houses the total population is living in category-wise.
2. All categories are again sub-divided into multiple categories. Example: sex and age main category is distributed into - Male, Female, again under 5 years, 5 to 9, 10 to 14 up to 75-84 years of age group people. Then, there is completely different categorization like 18 years or over - male, female, sex ratio and 65 years and over - male, female and sex-ratio of male to female per 100 females.
3. Similarly, all other main categories are divided into subcategories to give a clear picture of distribution of total population over these categories.

    - Based on Race categorization: The total population is divided into one race and two or more races. One race is again divided into multiple sections and so is two or more races.

    - Race alone category is divided among white, black, American Indian, Asian, Native Hawaiin and some other races.

· Hispanic or Latino race is distributed among Hispanic or latino, nonhispanic or latino and two or more races. There are many more divisions under these sub-categories.

· Citizen voting age population is divided into male and females who are citizens and 18 years over age.

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

COVID-19 DATA SET	**CENSUS DEMOGRAPHIC ENRICHMENT DATA SET**
covid_confirmed_usafacts Columns countyFIPS countyName State StateFIPS number of confirmed cases per day	**GEO_ID → countyFIPS** **NAME → countyName + State** We can merge confirmed_cases dataset with Demographic ACS using **GEO_ID → countyFIPS** because both attributes are common to both these tables. We can rename GEO_ID to countyFIPS since the data present in these columns are the same to both datasets.
covid_county_population_usafacts Columns: countyFIPS countyName State Population	**GEO_ID → countyFIPS** **NAME → countyName + State** We can merge covid_county_population with enrichment dataset using a common variable **GEO_ID → countyFIPS** (rename GEO_ID to countyFIPS). The data present in these columns are the same to both datasets.

covid_deaths_usafacts Columns countyFIPS countyName State StateFIPS number of deaths per day	**GEO_ID → countyFIPS** **NAME → countyName + State** We can merge covid_deaths_usafacts data set with Demographic ACS using **GEO_ID --> countyFIPS** because both attributes are common to both these tables. We should rename GEO_ID to countyFIPS since the data present in these columns is common to both datasets.

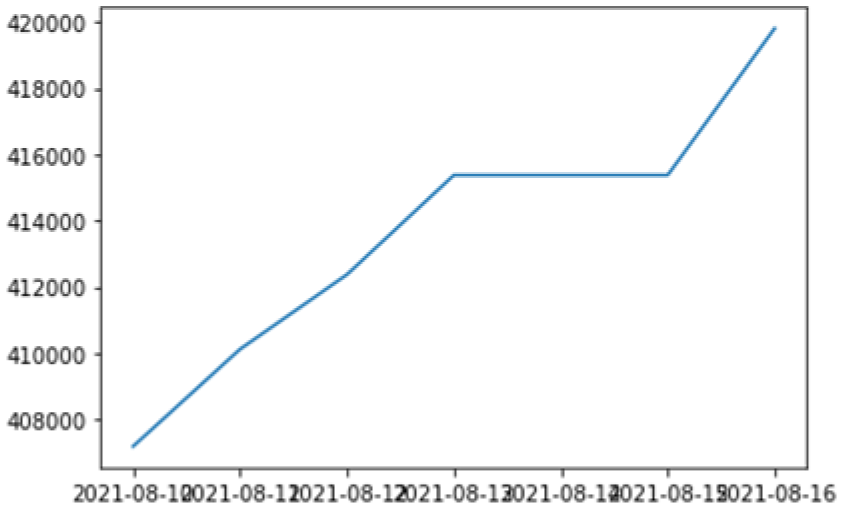
Describe how your enrichment data can help in the analysis of COVID-19 spread. Pose initial hypothesis questions.

From the above description, we know that the ACS Demographic dataset divides the total population into multiple categories. From sex and age category, we can find how the infection has impacted male, female, and what age group of people. Race categories can help to show what race of people are impacted or not. Whether the infection is impacting only one race or combined race people, Hispanic and Latino people, and People below and over 18 years. From the enrichment dataset, we know what category of people are concentrated in which counties and from COVID-19 data set we have infection count and death rate of each county, we could connect these datasets using countyFIPS as common variable and analyze what category of people are most or least infected.

Task 3:

Calculate COVID-19 data trends for last week of the data. Are the cases increasing, decreasing, or stable? Each student chooses a state to analyze.

I have chosen Arizona state to analyze the covid trend. I created a notebook and generated the matplotlib plot by taking Arizona state’s records of the last 7 days present in the data set. From the graph below, it is noticeable that at the beginning of the week for 4 days the cases were increasing day by day and from day 5-6 the cases remained constant, day 7 there was a surge in cases. In general, there is an increasing pattern.


420000
418000
416000
414000
412000
410000
408000
2021-08-10
2021-08-11
2021-08-12
2021-08-13
2021-08-14
2021-08-15
2021-08-16