Data Science Team 4

Project Report - 1

Team Members:

Aditi Darandale
Kyle Killworth
Manish Shah
Priyanka Budavi
Reetika Sarkar

Covid19 Dataset:

1. Confirmed covid cases:

- This dataset contains information about the total number of confirmed covid cases each day in each county of every state in the U.S.A.
- The dataset contains information from 22nd January 2020 to 7th February 2022.

Variable Name	Datatype	Definition
countyFIPS	int64	5-digit code used to uniquely identify the county
County Name	String	Contains name of the county
State	String	Contains name of the state
StateFIPS	int64	2-digit code used to uniquely identify the state
date	Int64	Dates showing no. of confirmed covid cases reported on a particular date

2. Covid Deaths cases:

- This dataset contains information about the total number of covid deaths recorded each day in each county of every state in the U.S.A.
- The dataset contains information from 22nd January 2020 to 7th February 2022.

Variable Name	Datatype	Definition
countyFIPS	int64	5-digit code use to uniquely identify the county
County Name	String	Contains name of the county
State	String	Contains name of the state
StateFIPS	int64	2-digit code used to uniquely identify the state
date	Int64	Dates showing no. of covid death cases reported on a particular date

3. County Population:

- This dataset contains information about population of each county in each state in the U.S.A.

Variable Name	Datatype	Definition
countyFIPS	int64	5-digit code use to uniquely identify the county
County Name	String	Contains name of the county
State	String	Contains name of the state
Population	int64	Contains total number of populations in each county

Preliminary Intuitions:

1. Confirmed covid cases:

When analyzing the confirmed cases of COVID-19 by county, we see that the confirmed cases start off slow in January and February 2020. By mid-March 2020, we can see that cases begin to exponentially increase in more populous counties in states on the east and west coasts, such as Nassau County in New York and Orange County in California. By early-to-mid April 2020, this trend has spread throughout the country. The number of new cases daily does not appear to start subsiding until early 2021, which coincides with the public administration of COVID-19 vaccines in the United States. This trend generally continues throughout Spring 2021, though we start to upticks in July and August 2021, which coincides with the arrival of the Delta variant of COVID-19 in the United States.

2. Covid Deaths:

For COVID-19 deaths by county, the trends we saw above can be found here as well. However, the trends occur a few weeks later than with COVID-19 cases. This seems logical, as people who end up dying from COVID do so a few weeks after contracting it. As such, we would expect daily COVID cases deaths to increase or decrease a few weeks after increases and decreases in daily COVID-19 cases.

3. County Population:

The need for the county population becomes apparent as we analyze confirmed cases and deaths. While the trends above are generally true for all counties, the raw numbers themselves vary greatly. Counties like Orange County in California have hundreds of thousands of confirmed cases, while others like Graham County in Arizona have less than 6,000. At first glance, this could lead someone to believe that certain counties may be under-reporting COVID-19 numbers. However, by bringing in the county population, we can see that Graham County only has a total population of less than 40,000 people. With that additional context in mind, 6,000 confirmed cases seem much more plausible. In order to account for these population discrepancies, we will be standardizing the COVID-19 confirmed cases and deaths. This will minimize the impact of population size and will put all counties on the same scale for analysis purposes.

Enrichment Datasets:

1. Census ACS Demographic Dataset: (MANISH SHAH)

This dataset provides information about various parameters of demographics at county level. The dataset contains many parameters that are not much useful and related in the analysis part, so only few selected columns have been used.

The information about the dataset is as follows:

It contains information about countyFIPS ID, county name, information about total population, information on male/female population, columns which contains information about the age ranges starting from under 5 years old to over 85 years old.

Below is the variable dictionary.

Name	Datatype	Definition
GEO - ID	string	Contains CountyFIPS at the end of the string
NAME	string	Contains county information
DP05_0001E	int	Estimate of total population
DP05_0002E	int	Estimate of total male population
DP05_0003E	int	Estimate of total female population
DP05_0005E	int	Estimate of total population under 5 years of age
DP05_0006E	int	Estimate of total population between 5 to 9 years of age
DP05_0007E	int	Estimate of total population between 10 to 14 years of age
DP05_0008E	int	Estimate of total population between 15 to 19 years of age
DP05_0009E	int	Estimate of total population between 20 to 24 years of age
DP05_0010E	int	Estimate of total population between 25 to 34 years of age
DP05_0011E	int	Estimate of total population between 35 to 44 years of age
DP05_0012E	int	Estimate of total population between 45 to 54 years of age
DP05_0013E	int	Estimate of total population between 55 to 59 years of age
DP05_0014E	int	Estimate of total population between 60 to 64 years of age
DP05_0015E	int	Estimate of total population between 65 to 74 years of age
DP05_0016E	int	Estimate of total population between 75 to 84 years of age
DP05_0017E	int	Estimate of total population over 85 years of age
DP05_0034E	int	Estimate of total population by race

Merging: The demographic data can be merged with the covid dataset using the county name present in the NAME variable, and also GEO-ID which contains countyFIPS in the last 5 digit.

Initial hypothesis: It is observed that children and elderly are more exposed to covid rather than young & middle-aged people. So, the counties with higher population of children or elderly will have higher covid cases and death counts. Using the age and sex category, we can know which category are more infected by covid19. Using race information, we can know if covid19 affects some race more, what race are most affected by covid.

2. Enrichment dataset: ACS Social, Economic, and Housing (KYLE KILLWORTH)

Introduction:

The dataset I was tasked with was the ACS Social, Economic and Housing data from the US Census, by county. ACS Social contains information such as the number of households by various sub-categories, the number of individuals for a large variety of demographics (age, gender, race, etc.), and the number of individuals at certain levels of education. ACS Economic follows a similar structure, but it focuses on categories such as occupation, income level, and poverty level. Lastly, ACS Housing contains information specifically about the contents of each household. This includes the number of vehicles, the number of rooms, and the age of each home. As each file contains a large amount of categories to choose from, I have opted to select a few from the ACS Social and Economic datasets.

Variable Dictionary:

ACS Economic:

Variable Name	**Datatype**	**Definition**
countyFIPS	Object	5-digit code use to uniquely identify the county
County Name	String	Contains name of the county and state
Less than $10,000	int64	No. of individuals with income below $10,000
$10,000 to $14,999	int64	No. of individuals with income from $10,000-$14,999
$15,000 to $24,999	int64	No. of individuals with income from $15,000-$24,999
$25,000 to $34,999	int64	No. of individuals with income from $25,000-$34,999
$35,000 to $49,999	int64	No. of individuals with income from $35,000-$49,999
$50,000 to $74,999	int64	No. of individuals with income from $50,000-$74,999
$75,000 to $99,999	int64	No. of individuals with income from $75,000-$99,999

$100,000 to $149,999	int64	No. of individuals with income from $100,000-$149,999
$150,000 to $199,999	int64	No. of individuals with income from $150,000-$199,999
$200,000 or more	int64	No. of individuals with income $200,000
Below Poverty Level	nt64	No. of individuals below the poverty line the last 12 months

ACS Social:

Variable Name	**Datatype**	**Definition**
countyFIPS	Object	5-digit code use to uniquely identify the county
County Name	String	Contains name of the county and state
No Diploma	int64	No. of individuals with no diploma
High School	int64	No. of individuals with a high school diploma
Associate’s	int64	No. of individuals with an Associate’s degree
Bachelor’s	int64	No. of individuals with a Bachelor’s degree
Graduate	int64	No. of individuals with an advanced degree
Foreign Born	int64	No. of individuals born in another country

Merge: I will be using the countyFIPS variable to merge the two datasets onto the COVID-19 dataset. Some cleaning of the variable will be required, as the format isn’t quite the same as the format found in the COVID-19 dataset.

Initial Hypothesis:

As can be seen from the variable dictionary above, the ACS Economic variables I will be merging into the COVID-19 dataset are primarily focused on household income. As such, my first hypothesis question is: do the various levels of family income show a significant difference in covid cases and deaths for the county? Additionally, I will be bringing in a variable concerning the number of individuals below the poverty line. This leads to my second hypothesis: do counties with a higher number of people below the poverty line have a significant difference in COVID-19 cases and/or deaths?

As for the ACS Social variables, the ones to be merged in mostly contain information about how many people in the county are at various levels of education. So my third hypothesis is: do counties with greater numbers of higher educated individuals have significantly different COVID-19 cases and/or deaths? Lastly, I will be including a variable indicating how many people in the county were born internationally. Thus, my final hypothesis is: is there a significant difference in COVID-19 cases and/or deaths amongst counties with higher numbers of foreign-born individuals?

3. Enrichment dataset: Employment dataset (ADITI DARANDALE)

Introduction:

The dataset I am working with is the employment dataset. This dataset gives us an overview of the employments quarterly. This employment dataset finds the number of people working or employed in the area and the annual wages in that county.

Variable Dictionary:

Variable Name	Type of variable
year	Int
Quarter	Int
County_FIPS	Object
County_Name	Object
Quaterly Employment	Int
Month 3 Employment	Int
Month 3 Employment rank	Object
Average weekly wages	Int
Average weekly wages percent change quantity	Object
Average weekly wages rank	Object

Merge:

I am merging my enrichment dataset with the super dataset by using the ‘County_Name’, ‘County_Name’ is used in both dataset and we are using as a catalyst to merge the data.

Initial Hypothesis:

I am analyzing my enrichment employment dataset by relating with which county and state are working and how we can correlate with COVID-19 trends, My hypothesis is that If there are higher number of people working in county it will lead to higher number of covid confirmed cases. This can give us the spread of covid in that particular county in that state.

4. Political Leanings Dataset:
(REETIKA SARKAR)
Task 2

Deliverables:

- Describe the enrichment data and datatype - variable dictionary.

This dataset contains information about the results of the presidential elections conducted in the year 2020 in the USA. It shows the names of the presidential candidates contesting along with their political party and number of votes cast in their favor. From the original data files, I chose the total number of votes, name of the winning party and votes in favor of them as variables of interest.
There were separate csv files denoting county-wise total votes, and county-wise data on the total votes won by each political party (Republican, Democratic, Libertarian and Write-ins) and the party that won. These two datasets were filtered and merged to record the following variables:

Column/Variable Name	**Data type**	**Definition**
state	String	Name of state
county	String	Name of county
countyfips	Integer	County FIPS
total_votes	Float	Total votes won by each party in each county
party_won	String	The winning political party in a particular county
num_votes	Float	Votes won by the winning party per county

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

The given dataset has county names, county FIPS and state names that are in common with the merged COVID-19 dataset. Data on county-wise political affiliation can be merged with COVID-19 data based on these identifiers.

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

It is a common perception that political affiliation can influence an individual’s choice of following COVID related safety protocols and immunization against it, and thus could affect the COVID-19 outcomes significantly. The county-wise presidential election results (political affiliation) can be used to analyze association between COVID-19 cases (and deaths) and political leaning. My initial hypothesis is to assess whether preference for a particular political party is associated with an increase or a decrease in the number of COVID-19 cases (and deaths).

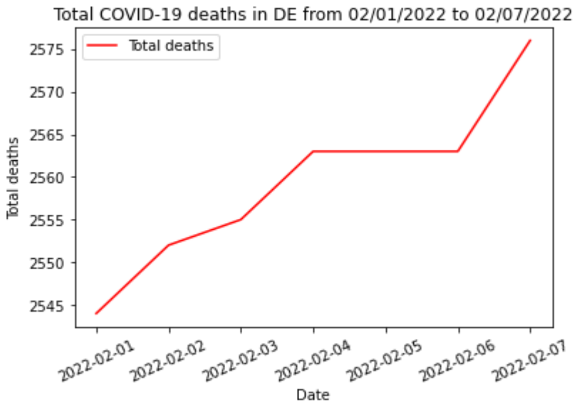
Task 3

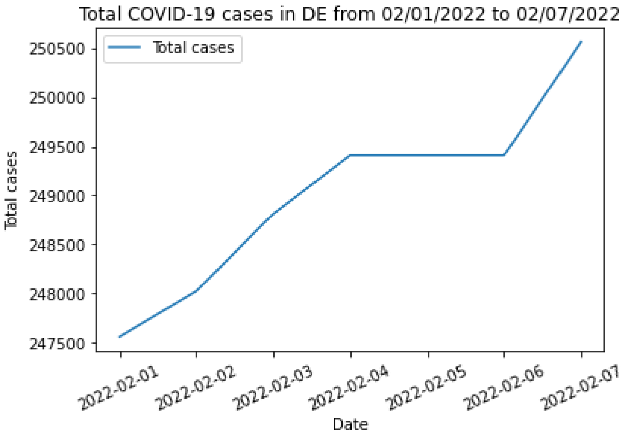
Deliverables:

- 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.

In general, the state of Delaware showed an increase in the number of COVID-19 cases and deaths for all its counties over the week starting February 1, 2022.

Plots depicting the pattern in the total number of COVID-19 cases and deaths in Delaware are shown below.





The total number of COVID-19 cases and deaths in Delaware show an increasing trend over the week and have a similar overall pattern of rise.

- Create a notebook to read the Enrichment data and display them on a notebook.
- Perform initial merges with the COVID-19 data using the variables in the Enrichment data.

I merged my enrichment dataset with COVID-19 merged superset using County FIPS and State using an inner merge. The merged dataset has additional columns showing the total number of votes cast in each county, the name of the winning party and the number of votes in favor of the winner. The merging has been shown in the Python notebook Political_linkage.ipynb uploaded to the group repository.

5. Enrichment dataset: Hospital Bed dataset (PRIYANKA BUDAVI)

Description:

The hospital bed dataset contains information about the hospital resources and the bed utilization values of both confirmed and suspected cases of COVID-19. It is a high dimensional dataset of 54 observations and 117 columns. Out of 117 columns I chose the below variables to define a new data frame which I find relevant to the covid dataset.

Variables that I can relate to the covid dataset:

Column Name	Datatype	Description
State	String	Name of the state
Inpatient beds	Int	Number of beds
Inpatient beds used	Int	The number of beds used
Inpatient beds used covid	Int	Beds used for covid
critical_staffing_shortage_today_yes	Int	Staff Information
critical_staffing_shortage_today_no	Int	Staff Information
staffed_icu_adult_patients_confirmed_and_suspected_covid	Int	Suspected covid and confirmed cases
staffed_icu_adult_patients_confirmed_covid	Int	confirmed covid cases
total_adult_patients_hospitalized_confirmed_and_suspected_covid	Int	Total confirmed and suspected cases
total_adult_patients_hospitalized_confirmed_covid	Int	Confirmed cases
total_staffed_adult_icu_beds	Int	Number of ICU beds
inpatient_beds_utilization	Float	Beds used

adult_icu_bed_covid_utilization	Float	Beds used for Covid
inpatient_bed_covid_utilization	Float	Beds used for Covid
percent_of_inpatients_with_covid	Float	% Of covid patients
adult_icu_bed_utilization	Float	Beds utilized by adults
reporting_cutoff_start	String	Day wise information
deaths_covid	Float	Number of deaths

1. The data types for the entire file are integers, float, or strings.
2. The variable that could merge the two data sets is using the column ‘State’.

Hypothesis:

The hospital bed dataset will help us understand how the pandemic has challenged the critical care capacity and put a strain on the healthcare system. If we compare the number of confirmed COVID-19 cases with the available hospital beds in the same area, we can conclude that patients who need extreme supportive treatment and could not be admitted to the hospital are left with no choice but to self-isolate and treat the virus at home. This undoubtedly would increase the spread of the virus as members of the family could be infected. Such infected persons may also spread it further to members of the community with whom they have contact. Thus, we see that there is an increase in the number of covid cases.

Task 3:

In this task the three datasets i.e., confirmed, deaths, population were merged to form a super dataset. Prior to that I performed analysis by selecting ALABAMA as my state of interest to check the trends in COVID cases and deaths. After plotting the graph with the dataset, I was able to understand that initial days of the week the cases were consistent but in the mid-week that is on 3rd Feb 2022 to 4th Feb 2022 there was a sudden spike in the number of covid cases and the deaths. I also noticed that there were no cases reported on the weekends and again there was a rise in covid cases. Thus, I conclude that there is a weekly increase in the covid cases and deaths.