Project Stage 1

This document was written and edited by:
Nikitha Narsing
Amulya Yadagani
Rashmi Hassan Udaya Kumar
Serena Wisnewski
Eric Cortes Aguilera

University of North Carolina at Greensboro

Covid-19 Datasets

Number of Cases

Description:

This dataset has the count of cases of each day in every county in all the states from 22nd January 2020-14th September 2020.

Datatype-variable dictionary: Covid_confirmed_usafacts dataset

Name	Datatype	Description
countyFIPS	Series(unique id) int	Each county has a unique Id.
County Name	object	County name in each state
State	object	State names in the country
stateFIPS	Int	Each state has a unique id
Different dates	Int	All the other columns are dates ranging from 22Jan to 14September 2020(they have the count of cases on each day)

Number of Deaths

Description:

This dataset has the count of deaths of each day in every county in all the states from 22nd January 2020-14th September 2020.

Datatype-variable dictionary: Covid_death_usafacts dataset

Name	Datatype	Description
countyFIPS	Series(unique id) int	Each county has a unique Id.
County Name	object	County name in each state
State	object	State names in the country
stateFIPS	Int	Each state has a unique id
Different dates	Int	All the other columns are dates ranging from 22Jan to 14September 2020(they have the count of deaths on each day)

County Population

Description:

This dataset contains the population in each county.

Datatype-variable dictionary: Covid_county_population_usafacts dataset:

Name	Datatype	Description
countyFIPS	Series(unique id) int	Each county has a unique Id.
County Name	object	County name in each state
State	object	State names in the country
Population	int	Number of people in each county

Preliminary Intuitions

- Inspecting the data from Jan 22 2020 to Sep 14 2020, we observe that the first covid-19 cases were detected in the month of March 2020 in most of the counties.
- We can observe that the states with large population have more number of covid-19 cases and deaths
- Through a brief overview of the data it appears as if the spread of covid-19 grew slow to begin with then exponentially increased soon after
- The analysis of the last month of the data shows that the covid-19 spread, both deaths and cases has started to decrease
- Los Angeles county in California has the highest number of cases and deaths till date.

Enhancement Datasets

Demographics Dataset - Rashmi Hassan Udaya Kumar

Description

The Demographic ACS Enrichment dataset is produced by the American Community Survey (ACS) for the year 2018 and contains the population and demographic information estimates of the counties present in the United States of America.

Demographic information estimates are based on the following categories:

- **Total Population** : Total population estimate for each county is provided
- **Sex** : For each county there is estimate of total population of male and total population female.
- **Age** : For each county, estimate population count is provided for age groups ranging from 5 years to 85 years
- **Race** : For each county there is population count estimate for different races including White, Black or African American , Asian, American indian and Alaskan Native and Native hawaiian and Pacific Islander
- **Citizen Voting age population** : County wise population estimate of people above 18 years of age

Data Type-variable dictionary:

SL.No.	Variable Name	Data Type	Variable Description
1	GEO_ID	Object	This ID is 14 characters in length and the last 5 characters represent the county FIPS
2	NAME	Object	Geographic Area Name which includes the name of county and state
3	DP05_0001E	int64	Estimate of Total population
4	DP05_0002E	int64	Estimate of Male Total population
5	DP05_0003E	int64	Estimate of Female Total population
6	DP05_0004E	int64	Estimate of sex ratio for total population (males per 100 females)
7	DP05_0005E	float64	Estimate of Total population with age under 5 years
8	DP05_0006E	int64	Estimate of Total population with age from 5 to 9 years

9	DP05_0007E	int64	Estimate of Total population with age from 10 to 14 years
10	DP05_0008E	int64	Estimate of Total population with age from 15 to 19 years
11	DP05_0009E	int64	Estimate of Total population with age from 20 to 24 years
12	DP05_0010E	int64	Estimate of Total population with age from 25 to 34 years
14	DP05_0011E	int64	Estimate of Total population with age from 35 to 44 years
15	DP05_0012E	int64	Estimate of Total population with age from 45 to 54 years
16	DP05_0013E	int64	Estimate of Total population with age from 55 to 59 years
17	DP05_0014E	int64	Estimate of Total population with age from 60 to 64 years
18	DP05_0015E	int64	Estimate of Total population with age from 65 to 74 years
19	DP05_0016E	int64	Estimate of Total population with age from 75 to 84 years
20	DP05_0017E	int64	Estimate of Total population with age from 85 years and over
21	DP05_0018E	float64	Estimate of Median age in years of Total population
22	DP05_0037E	int64	Estimate of White race in Total population
23	DP05_0038E	int64	Estimate of Black or African American race in Total population
24	DP05_0039E	int64	Estimate of American Indian and Alaska Native race in Total population
25	DP05_0044E	int64	Estimate of Asian race in Total population
26	DP05_0052E	int64	Estimate of Native Hawaiian and Other Pacific Islander race in Total population

How to Merge:

We can merge the enrichment dataset with the primary covid-19 dataset since both the datasets have data that is distributed across counties present in the USA.In the primary covid-19 dataset, each county is identified by a unique "**countyFIPS**".In the enrichment dataset each county is identified by a unique "**GEO_ID**" which contains the "countyFIPS" in the last 5 characters. These last 5 characters exactly match the "countyFIPS" in the primary covid-19 dataset. So we can merge the two datasets based on the "countyFIPS".We can use inner join to merge the two data sets as there is a majority of matching information in both the datasets.

Hypothesis:

Enrichment data has demographic information for the counties in the United States. The demographic information has the total population distributed across male vs female, different age groups, different races. Below are some of the questions we can pose based on the information in the enrichment dataset.

1. Can we correlate say male to female ratio to the number of covid-19 cases to find whether the infection rate is more in either males or females?
2. Can we find correlation of the covid-19 infection rate with different age groups and different races?
3. Did the summer break for school children have an influence on the increase in the number of covid-19 cases across different counties?
4. Did the stay at home order reduce the number of covid-19 cases among the working class of the population?
5. Do the states with busy international airports have more covid-19 cases?
6. Does high population rate increase the spread of covid-19?
7. How is the rate of spread in the counties with a high percentage of working professionals?
8. Is the infection more prevalent in the elderly population?
9. What is the distribution of mortality rates across the different races of population?
10. Is the mortality rate in elderly population higher when compared to the younger population?

Social and Economic Datasets - Serena Wisnewski

Description:

This dataset collected by the American Community Survey in 2019 has population education demographics information by county in the United States. This information can be found here https://data.census.gov/cedsci/table?q=dp&tid=ACSDP1Y2018.DP02

Datatype-variable dictionary: EducationEnrichment_data_with_overlays

Name	Datatype	Description
FIPS	Series(unique id) int	Last 5 digits of county GEO_ID
GEO_ID	Series(unique id) int	Each county has a unique Id.
Name	object	Geographical Area name "Name County, State"
DP02_0053E	Int64	SCHOOL ENROLLMENT: Population 3 years and over enrolled in school Population
DP02_0054E	Int64	SCHOOL ENROLLMENT: Population 3 years and over enrolled in school Nursery School/Preschool
DP02_0055E	Int64	SCHOOL ENROLLMENT: Population 3 years and over enrolled in school Kindergarten
DP02_0056E	Int64	SCHOOL ENROLLMENT: Population 3 years and over enrolled in school Elementary School (1-8)
DP02_0057E	Int64	SCHOOL ENROLLMENT: Population 3 years and over enrolled in school High School (9-12)
DP02_0058E	Int64	SCHOOL ENROLLMENT: Population 3 years and over enrolled in school College or Graduate School
DP02_0059E	Int64	EDUCATIONAL ATTAINMENT: Population 25 years and over Population
DP02_0060E	Int64	EDUCATIONAL ATTAINMENT: Population 25 years and over Less than 9th grade
DP02_0061E	Int64	EDUCATIONAL ATTAINMENT: Population 25 years and over 9th-12th grade, no diploma
DP02_0062E	Int64	EDUCATIONAL ATTAINMENT: Population 25 years and over High school graduate/GED
DP02_0063E	Int64	EDUCATIONAL ATTAINMENT: Population 25 years and over Some college, no degree
DP02_0064E	Int64	EDUCATIONAL ATTAINMENT: Population 25 years and over Associate's degree
DP02_0065E	Int64	EDUCATIONAL ATTAINMENT: Population 25 years and over Bachelor's degree
DP02_0066E	Int64	EDUCATIONAL ATTAINMENT: Population 25 years and over Graduate or professional degree
DP02_0067E	Int64	EDUCATIONAL ATTAINMENT: Population 25 years and over High school graduate or higher
DP02_0068E	Int64	EDUCATIONAL ATTAINMENT: Population 25 years and over Bachelor's degree or higher
DP02_0151E	Int64	COMPUTERS AND INTERNET USE: Total households

DP02_0152E	Int64	COMPUTERS AND INTERNET USE: Total households with a computer
DP02_0153E	Int64	COMPUTERS AND INTERNET USE: Total households with broadband internet subscription

How to Merge:

This data can be merged with the project's primary COVID-19 data on EducationEnrichment_data_with_overlays FIPS = Covid_county_population_usafacts dataset countyFIPS

Hypothesis:

Counties with higher education attainment saw a decrease in new cases earlier than counties with lower education attainment

Counties with higher percentage of the population with a broadband internet subscription saw initial cases earlier than counties with lower percentage of the population with broadband internet subscriptions

Housing Dataset - Eric Cortes Aguilera

Description:

This dataset contains the housing information the American Community Survey obtained for each county for the year 2019. Information like listed below corresponding to the number of house owned units, rented units, and average household size per each unit for each county in the United States.

Data Type-Variable dictionary:

ACSDP1Y2019.DP04_data_with_overlays_2020-09-17T112419

Name:	Datatype:	Description:
GEO_ID	Series(unique id) int	Unique ID corresponding to geographical location of each county
Name	object	The county name with state name following
DP04_0046E	float	Estimate!!HOUSING TENURE!!Occupied housing units!!Owner-occupied
DP04_0047E	float	Estimate!!HOUSING TENURE!!Occupied housing units!!Renter-occupied

DP04_0048E	float	Estimate!!HOUSING TENURE!!Occupied housing units!!Average household size of owner-occupied unit
DP04_0049E	float	Estimate!!HOUSING TENURE!!Occupied housing units!!Average household size of renter-occupied unit

How to Merge:

To properly merge this dataset with the Covid-19 dataset a couple steps need to be taken. First, the column names need to be replaced with more comprehensible names making sure the county name column matches the county name column of the Covid-19 dataset. Second, the row corresponding to the position 0 needs to be deleted since the column names were replaced with the data from this row. Third, the county name column items need to be changed to lowercase and the state name needs to be removed. Finally, a merge can be made using the county name column.

Hypothesis:

Counties with a higher number of owner-occupied housing units then renter-occupied housing units will have a lower number of covid-19 cases as well as deaths.

Counties with a higher average of household sizes will have a higher number of covid cases as well as deaths due to the number of people in close proximity on a daily basis.

Employment Dataset - Nikitha Narsing

Description:

Employment data is collected by ‘The Quarterly Census of Employment and Wages (QCEW)’ program every year and this dataset contains the employment data of all the states(as well as individual counties) from Jan-March for the year 2020.

Datatype-variable dictionary:employment

Name	Datatype	Description
Area	Object	Contains the county name and state name
Area\nCode	Object	It is the distinct id of each county.
St Name	Object	State name
Ownership	object	Different ownerships to see the employment details of Federal, State, Local government or Private sectors.
Own	Int64	Unique code for each ownership
Industry	Object	To see the employment details of each private industry individually(Manufacturing, Goods producing etc) and Govt as a whole.
Establishment count	Int64	Number of people employed in each establishment.
January, February and March Employment	Int64	Number of people who were employed in the months of 2020(Jan,Feb and March) in each industry in a county.
Total Quarterly Wages	Int64	Total quarterly wages in a county based on 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 Relative to U.S.	Float64	Wage location quotient compared to the total country.

How to Merge:

We can first preprocess the data of the employment dataset; we need to fix the inconsistencies in the Area Code column(there are leading 0's for 4 digit area codes, removing them). And then we can choose only required variables which would be used for analysis. I would remove some rows which are not up to date or which I feel I may not use for the Covid analysis or the rows with inconsistent or missing data.

Then I would merge the employment dataset using the 'Area Code' variable with the 'County Name'(these both are the unique county id columns) variable in the Covid super dataset to get an emp_covid dataset.

Hypothesis:

I will take the population of each county with the employee dataset and calculate the employment rate in each county. This would help me in knowing the employment rate and Covid cases/deaths connection. Then with the deaths/cases, it would help me to know which counties, industries or ownerships were highly affected and which were not.

1. With high employment rate counties there must have been an increase in Covid cases in the initial months as those counties would be crowded and people would be going to their workplaces daily.

2. High employment rate may indicate less death rate as there will be more Medical facilities in the counties and people would be recovered.
3. Highly paid counties would also have less deaths as they would have knowledge on how to deal with the disease like quarantine, take medicines.
4. More employment rate counties would have low cases in later months as they start using masks, use sanitizers, any precautions needed.
5. We can take a few industries into consideration, and see which industry is more affected by the pandemic(number of cases/deaths). Like the manufacturing industry should be running so there may be cases which would not decrease in that industry.
6. Business management related industries might have reduced cases as they can work from home.

Hospital Beds Dataset - Amulya Yadagani

Description:

The Hospital Beds dataset is provided by Definitive Healthcare and made available by Esri's Geospatial Cloud which contains 6621 rows and 23 columns as of 09/14/2020. This dataset is primarily used as a baseline to know the available facilities (Number of beds) and their average usage (Bed Utilization rate, Avg Ventilator usage) in the hospitals at all counties across the USA.

Data Type-Variable dictionary: Hospital_Beds

Name	Datatype	Description
OBJECTID	int64	Unique identifier for this dataset
HOSPITAL_NAME	object	Name of the hospital
HOSPITAL_TYPE	object	Type of facility if it is a Term Acute Care Hospital, Critical Access Hospital, Psychiatric Hospital, Long Term Acute Care Hospital, Rehabilitation Hospital, VA Hospital, Children's Hospital, Department of Defense Hospital, Religious Nonmedical Health Care Hospital
HQ_ADDRESS	object	Address of the hospital

HQ_CITY	object	Name of the City where the hospital is located
HQ_STATE	object	2 letter State code of the hospital
HQ_ZIP_CODE	int64	Zip Code of the hospital
COUNTY_NAME	object	Name of the County where the hospital is located
STATE_NAME	object	Name of the State where the hospital is located
STATE_FIPS	int64	2 digits unique FIPS code representing the State in the USA
CNTY_FIPS	int64	3 digits FIPS code representing the County in a particular state in USA
FIPS	int64	The combined 5 digits unique FIPS code of State and County
NUM_LICENSED_BEDS	int64	Number of beds for which the appropriate state agency licenses a facility
NUM_STAFFED_BEDS	int64	Number of beds that are available for which the staff is on hand to attend to the patient who occupies the bed
NUM_ICU_BEDS	int64	Number of ICU Beds
ADULT_ICU_BEDS	int64	Number of Adult ICU Beds
PEDI_ICU_BEDS	int64	Number of Pediatric ICU Beds
BED_UTILIZATION	float64	Average yearly bed utilization computed as Total Patient Days (excluding nursery days)/Bed Days Available
Potential_Increase_In_Bed_Capacity	int64	Estimation of an increase in bed capacity by subtracting Number of Staffed Beds

		from Number of Licensed beds
AVG_VENTILATOR_USAGE	int64	Average number of patients on a ventilator per week

How to Merge:

The Hospital_Beds Enrichment dataset has a "FIPS" column which is a 5 digit number representing unique code for each county and the COVID dataset has a unique 5-digit number for each county in "countyFIPS" column. Preprocessed the Hospital_Beds Enrichment dataset by removing rows having "FIPS" value as NAN and dropped a few columns which might be repetitive after merge. Using columns "countyFIPS" and "FIPS" from both datasets we can merge the COVID dataset with the Hospital_Beds dataset using an inner join.

Hypothesis:

The increase in "bed utilization rate" can mean fewer hospital resources are available for people who are COVID positive resulting in an increase in COVID spread and deaths.

The number of COVID confirmed cases & deaths might see a downward trend in counties that have more "number of Hospitals" compared to counties that have lesses hospitals.

Counties with fewer populations have fewer hospital resources available because of which the "bed utilization rate" is equal to or almost 100% that could lead to an increase in the number of deaths.