Data Science Project Stage 1

Ashim Chalise, Erika Sudderth, Ilyass Sfar, Pratik Devkota, Sogol Ghotbi Taheri

CSC 405/605

Team Tasks Section

Covid 19 Dataset

COVID Cases - Data Dictionary

Description

This dataset contains a total number of confirmed COVID-19 cases per day per county across the United States. Each State in the US has one or more counties. All states and counties are assigned a unique FIPS code (Federal Information Processing Standard code) as an identifier. Each row represents data for one specific county where the data includes county FIPS, county name, state, state FIPS and confirmed number of cases from 01/22/2020 and onwards.

However, the dataset also contains one row per state identified with 0 as countyFIPS code. These rows (or data) represent statewide unallocated COVID-19 confirmed cases by states for each day since 01/22/2020, supposedly meaning that x number of cases were recorded but it was not sure which county each of those cases belong to. Value in any column with a particular date seems to represent the total number of confirmed COVID-19 cases till that date. However, rows with countyFIPS value of 0 do not seem to follow this pattern.

Data also shows that the first case of COVID-19 in the US was recorded on 22nd of January 2020 in King County of Washington (WA).

Variable Name	Datatype	Description
CountyFIPS	Integers	Unique County Identifier
County Name	String	Name of the counties in U.S.
State	String	Name of the states in U.S.
StateFIPS	Integers	Unique State Identifier
Cases	Integers	Number of cases in a day/county.

Population - Data Dictionary

Description

This dataset contains the total population per county across all states of the United States. Like covid_confirmed_usafacts dataset, this dataset also includes one row per state identified with 0(State unallocated) as countyFIPS code. However, it also includes two unique entries ‘Wade Hampton Census Area’ and ‘Grand Princess Cruise Ship’. They are unique in the sense that both of them are assigned with countyFIPS but the population is 0. Where every state has one entry with ‘Statewide Unallocated’ as County Name, New York includes two entries, an extra entry with ‘New York City Unallocated’ and 0 population. Data shows that Kalawao County of Hawaii (HI) has the lowest population of 86 and Los Angeles County of California (CA) has the highest population of 10039107.

Variable Name	Datatype	Description
countyFIPS	Integers	Unique County Identifier
County Name	String	Name of the counties in U.S.
State	String	Name of the states in U.S.
Population	Integers	Number of people in a county.

COVID Deaths - Data Dictionary

Description

This dataset contains a total number of confirmed COVID-19 deaths per day per county across the United States. The first death was recorded on the 6th of February 2020 in Santa Clara county, California as per our data. The maximum number of deaths is 18500 in Los Angeles county, California as per our data.

Variable Name	Datatype	Description
CountyFIPS	Integers	Unique County Identifier
County Name	String	Name of the counties in U.S.
State	String	Name of the states in U.S.
StateFIPS	Integers	Unique State Identifier
Deaths	Integers	Number of Deaths in a day/county.

Preliminary Intuitions of the COVID-19 Datasets

The first site (usafacts.org) shows processed data with 7-day averages of cases and deaths with the actual number by state. But to dig deeper, we need county level data which is provided by usafactsstatic.blob.core.windows.net/.../csv site. Our preliminary intuition is that this data is presented in a dataset that is not easily comprehended visually. We feel the data needs to be processed and presented in a more meaningful and organized way to allow any viewer to easily understand.

In total, new cases and deaths have rapidly increased since March 2020 in the United States. Based on the usafacts.org diagrams, we see that the number of deaths and known cases were at its highest in January 2021. The highest known average weekly number of cases was 246,164 on January 11th, 2021. The highest known average weekly number of deaths was 3307 on January 15th, 2021. In light of this, a possible hypothesis could be that more populous counties have seen a higher number of cases for long periods of time compared to less populous counties which have seen peaks of cases within shorter periods of time.

Individual Tasks Section

Hospital Bed Dataset - Ilyass Sfar

Description

This hospital bed dataset contains data about 6,634 medical facilities throughout the United States. It gives insight to how many hospital beds are available at a given time showing how full hospitals are. Other information is also given to give insight to how bad patients conditions are, such as ventilator usage.

I have removed some of the variables that are not useful, i am only listing the variables that remain and did not remove

Variable Name	Datatype	Description
OBJECTID	int	Index of data
HOSPITAL_NAME	String	Name of Hospital
HOSPITAL_TYPE	String	Type of hospital (ex. VA, Childrens, Psychiatric, Rehabilitation, short term acute, long term acute)
HQ_ADDRESS	String	Address of hospital
HQ_STATE	String	State abbreviation of hospital
HQ_CITY	String	City where hospital is located
HQ_ZIP_CODE	String	Zip code of hospital location
FIPS	int	Federal Information Processing Standards, this is used as a unique country identifier
NUM_LICESNSED_BEDS	int	Maximum number of beds a hospital is licensed to operate
NUM_STAFFED_BEDS	int	From the source of the data this is defined as,”adult bed, pediatric bed, birthing room, or newborn ICU bed (excluding newborn bassinets) maintained in a patient care area for lodging patients in acute, long term, or domiciliary areas of the hospital."
NUM_ICU_BEDS	int	Maximum number of intensive care unit beds, from the source of the data, is defined as CMS, Section 2202.7, 22-8.2.

ADULT_ICU_BEDS (legacy variable)	int	This is a legacy variable, this has the same value as NUM_ICU_BEDS, it holds the number of ICU beds
PEDI_ICU_BEDS	int	Combination of neonatal, premature and pediatric ICU beds
BED_UTILIZATION	int	From the source of the data this is calculated based on metrics from the Medicare Cost Report: Bed Utilization Rate = Total Patient Days (excluding nursery days)/Bed Days Available
AVG_VENTILATOR_USAGE	int	The average number of patients on a ventilator per week
Potenital_inscrese_in_bed_capacity	int	Predictive variable on potential increase in hospital patients, this is calculated as, *"Number of Staffed Beds from Number of Licensed beds" (Licensed Beds – Staffed Beds)"*

Preliminary Intuitions

The hospital bed data is very insightful to the severity of COVID in certain counties. The data includes 6,634 medical institutions, and allows insight into how bady a community is being affected by COVID, using how strained community medical resources are and how many patients require ventilators.

How to Merge

The hospital bed data can be merged with the COVID superset using the FIPS, which has a different name in this data set but means the same thing, also in the hospital bed data the FIPS is a float, so to merge it it would be turned into a int and merged on FIPS with the superset.

Importance of this Enrichment Data/Hypothesis

The hospital bed dataset and variables such as bed utilization and average ventilator can be used to give insight into how bady COVID is effecting a community. Given this my hypothesis with the hospital bed dataset is counties with less hospital resources specifically beds in proportion to the county population were hit harder by COVID, since there are less resources to go around and help people, this can also be done with ventilator usage to find the severity of cases.

ACS Social, Economic, and Housing Dataset. - Ashim Chalise

Description

The dataset has all the variables related to housing. It has variables like Occupied housing units, Vacant housing units, year built, total rooms, rented or owned, vehicle per house, fuel used, value, mortgage status etc. The dataset is very rich in the housing market of the United states and the dataset I have is at county level. It has a history of 5 years. The data also contains, Estimate, Margin of Error, Percent, Percent of Error for each variable. This dataset has a lot of categories and subcategories with county level information.

These data variables belong to the original data file.

Variable Name	Datatype	Description
GEO_ID	String	countyFIPS with additional variables.
Name	String	Name of the counties.
HOUSING OCCUPANCY	MULTIPLE***	Total number of units occupied.
UNITS IN STRUCTURE	MULTIPLE***	Total housing units.
YEAR STRUCTURE BUILT	MULTIPLE***	The year that the unit was built.
ROOMS	MULTIPLE***	Rooms per unit.
BEDROOMS	MULTIPLE***	Bedrooms per unit.
HOUSING TENURE	MULTIPLE***	Length of the lease.
YEAR HOUSEHOLDER MOVED	MULTIPLE***	The year that the current resident moved.
VEHICLES AVAILABLE	MULTIPLE***	The number of vehicles per unit.
HOUSE HEATING FUEL	MULTIPLE***	The fuel used to heat the house i.e. gas, electric.
SELECTED CHARACTERISTICS	MULTIPLE***	Lacking complete plumbing/kitchen/telephone facilities.
OCCUPANTS PER ROOM	MULTIPLE***	Number of occupants per room.

VALUE	MULTIPLE***	Value of the units.
MORTGAGE STATUS	MULTIPLE***	Units with and without a mortgage.
SELECTED MONTHLY OWNER COSTS (SMOC)	MULTIPLE***	Monthly costs for owners with a mortgage.
SELECTED MONTHLY OWNER COSTS AS A PERCENTAGE OF HOUSEHOLD INCOME(SMOCAPI)	MULTIPLE***	Monthly costs as a percentage of the income of the resident/owner.
GROSS RENT	MULTIPLE***	Grossing rent of the unit.
GROSS RENT AS A PERCENTAGE OF HOUSEHOLD INCOME (GRAPI)	MULTIPLE***	Grossing rent of the unit as a percentage of the income of the resident.

MULTIPLE*** - All of these data variables have multiple categories and subcategories with double, int and string as their datatype.

Preliminary Intuitions

This data set makes much more sense compared to the cases, deaths and population dataset. But, as this dataset is not mapped or plotted, it is difficult to understand what the trend of a particular category or subcategory is. With Margin of Error and Percentage Margin of Error, the statistical accuracy of this dataset is impeccable**.**

How to Merge

The dataset I chose had a lot of variables and I had the option to filter/choose a lot of it. I have downloaded county level data which is why I am going to use the countyFIPS to merge to the main dataset as this unique number has been very helpful in merging and understanding the dataset as a whole.

Importance of this Enrichment Data/Hypothesis

I feel this dataset is very important to get an understanding of the overall economy before and after Covid-19. It'll help us show the impact in the economy, and to recover and rebuild our economy, I believe this dataset can be really useful. This dataset can help the government analyze Covid and plan the specifics of the stimulus. Counties with houses that have more rooms per unit and less occupants per room should have less

Covid-19 cases compared to counties with houses that have less rooms per unit and more occupants per room. I'm going to try to gather facts to prove this hypothesis by comparing the dataset of confirmed cases and my enrichment dataset.

2020 Presidential Election Results - Erika Sudderth

This dataset contains the results of the 2020 United States presidential election

Variable Name	Datatype	Description
state	String	The state of the election results.
county	String	The county of the election results.
candidate	String	The presidential candidates that received votes in each county.
party	String	The political party of the candidate.
total_votes	Integer	The total votes each candidate received per county
won	Boolean	The results of the presidential election per county (true or false value of it a candidate won.

Preliminary Intuitions

Republican leaning counties might have more cases of and deaths from COVID- 19 per capita than any other party leaning counties. On the other hand, if larger cities have a higher Democratic leaning population, Democratic leaning counties might have more cases of and deaths from COVID-19 per capita due to the larger population.

How to Merge

In order to merge this data with the overall data involving the cases of and deaths due to COVID-19 in the United States and the population information, the variables 'county' and 'state' must be used. These are common variables between these four datasets, and will allow for linking the data together.

Importance of this Enrichment Data/Hypothesis

Different political parties are associated with different views. This includes the debate over public safety versus personal freedoms. Analysis of this dataset could relate the

majority political leaning of a county to the number of cases and deaths in order to see if they are correlated. Knowing this could lead to the ability to better predict the future outcome of the virus in certain areas and where to focus preventative measures, like the vaccine.

ACS Census Demographics - Pratik

Description

The ACS Census Demographics enrichment dataset is produced by the American Community Survey and is different from the official estimates produced by the Census Bureau's Population Estimates Program. These data are only estimates, based on a sample and are subject to some degree of uncertainty.
The dataset contains data that are divided into 6 major categories namely:

1. SEX AND AGE,
2. RACE,
3. Race alone or in combination with one or more other races,
4. HISPANIC OR LATINO AND RACE,
5. Total housing units and CITIZEN, and
6. VOTING AGE POPULATION.

Each category is further divided into subcategories. Each row in this dataset (except the first row) represents statistics for individual counties where the data is not only divided into above mentioned categories but also includes subcategories within the categories. Since this data is based on a sample, ACS has also provided a margin of error for each category. There is 90 percent probability that the true value of any category or subcategory lies between estimated value - margin of error and estimated values + margin of error.

There are a total of 3222 rows (including header) and 359 columns in this dataset. The first row is the header, the second row is the description of the header itself while the remaining rows are unique entries for each county or county equivalent. Each row includes unique GEO_ID, NAME and 357 columns of estimated data. That includes an estimate, margin of error, percent and percent margin for every subcategory.

Variable Name	Datatype	Description
GEO_ID	String	Unique id for county or county equivalent Eg. 0500000US01001. Here, '0500000US' is static in all rows with the following 5 numbers representing the unique county FIPS code.

NAME	String	County name (or equivalent) followed by ‘,’ followed by state name (or equivalent) Eg. Autauga County, Alabama
DP05_XXXXE	Integer	First row of the dataset includes the actual meaning of each value of XXXX. E.g. First row under column DP05_00001E says ‘Estimate!!SEX AND AGE!!Total Population’. This represents estimate of total population of category SEX AND AGE. ‘!!’ can be used to represent the hierarchy of categories and subcategories.
DP05_XXXXM	Integer	
DP05_XXXXPE	Float	
DP05_XXXXPM	Float	

How to Merge

All GEO_IDs have unique strings - ‘0500000US’ followed by a 5 digit number representing unique county FIPS code. So, by removing substring ‘0500000US’ from GEO_ID, converting the remaining characters to integer, we will have unique FIPS code for each row. We can achieve this by dropping the first row, applying lambda on the ‘GEO_ID’ column of our new dataframe where lambda will work for each x such that we take the last 5 characters of x and typecast it to integer. This will return a series which we can append to our new dataframe as new column data (under column name countyFIPS). This new dataframe can be merged with our super COVID-19 dataframe (dataframe achieved from merging cases, deaths and population dataframe) with inner join where the common column will be countyFIPS.

Importance of this Enrichment Data/Hypothesis

The ACS Census Demographics enrichment dataset can be used to understand the correlation between COVID-19 cases and different age groups, races and sexes. Based on the distribution of population across multiple age, race groups, we can visualize how the higher population distribution of certain age groups can contribute to the increase in the number of confirmed cases and/or deaths.

- Hypothesis 1:
  Counties or states with higher populations of millennial (age 24 to 35) are more likely to see a sudden rise or higher number of COVID-19 cases. Millennials are active groups and are more likely to go out and travel. This increases the risk of exposure to COVID-19 virus, ultimately leading to increased number of cases.

- Hypothesis 2:
  Counties or states with higher populations of old aged people (55 and above) are more likely to see a higher number of deaths from COVID-19. Old aged people have weaker immunity as compared to young aged people and thus are more likely to suffer from serious health consequences.

If these hypotheses are supported by the facts, these findings can help the counties better prepare for the consequences. For instance, if the above hypotheses are proved with facts and figures, counties with higher numbers of millennials can impose stronger restrictions on cross county and cross state travels in order to control the possible spread of COVID-19 virus. Similarly, counties with higher numbers of old aged people can improve their medical emergency facilities, give higher priority in providing medical assistance to old aged people.

2020 Employment dataset - Sogol

Description

This dataset includes the employment data for year 2020 and during the pandemic

Variable Name	Datatype	Description
Wage change	double	The changes in wage for employees
Contributions change	integer	Changes in contributions
Area FIPS	integer	Unique area identifier
quarter	integer	Quarter in a year
Area title	string	Name of the area
Month 1 employment level	integer	Employment level of first month
Month 2 employment level	integer	Employment level of second month
Month 3 employment level	integer	Employment level of third month

Preliminary Intuitions

Based on the data we have, the number of employees has decreased since coronavirus outbreak and also there is a huge contributions change between the first and second quarter of the year 2020, as an example the contribution change in first quarter is -6898192 and in the second quarter this number has been decreased to -9002009. There is also change in wage between quarter one and quarter two, the amount of wage has been increased and that's due to the less employees that have been at work or had a job. Last thing we want to consider is the employment level among the first 3 months of the year (January, February, and March) the employment level in February is the highest among these three months in quarter one but this number has decreased in quarter two and in quarter two the employment level in March is the highest.

How to Merge

To merge this data we should be concerned about the quarters, so there are a lot of differences between the first quarter and second quarter when we compare wages, employment level, contribution changes, etc.

Importance of this Enrichment Data/Hypothesis

It's important because employment has a close relationship with economy and this virus has caused to death and unemployment of many people who had a job and provided service in the United States, less contributions, and more wages (because there was less employee than usual therefore they should have been kept at work and even work more than usual) as the hypothesis question we want to consider a rural county in the US vs. a populous county as an example let's compare catron county(a rural) vs. maricopa county(a populous county) based on the data if we only look at the month 1 employment level of maricopa county the number is 2099851 and this number in catron county is 594 and only these two numbers can prove that the populous counties saw covid cases for a longer time than rural counties and it's because of the population and the need of job in populous counties.