COVID DATA ANALYSIS
Stage-1

-Gowtham Ravella

Dataset: Demographic and Housing Estimates

Section in the report describing the enrichment data and datatype - variable dictionary

The ACS Demographic and Housing Estimates enrichment data contains variables that describe many aspects of housing and population of various age groups. The various data that are in the dataset are Population based on different ages, Total housing units, People eligible for voting who are segregated into male and female.

Variable Dictionary

Name	Definition	Data Type
Geography	Data related to FIPS code for countys	Int64
Geographic Area Name	Consists of County and state name	Object
SEX AND AGE!!Total population	Total male, female population	Object
Estimate!!SEX AND AGE!!Total population!!Male	Total male population	Object
Estimate!!SEX AND AGE!!Total population!!Female	Total female population	Object
Estimate!!SEX AND AGE!!Total population!!Under 5 years	Total population under the age of 5	Object
Estimate!!SEX AND AGE!!Total population!!5 to 9 years	Total population between the age of 5 and 9	Object
Estimate!!SEX AND AGE!!Total population!!10 to 14 years	Total population between the age of 10 and 14	Object
Estimate!!SEX AND AGE!!Total population!!15 to 19 years	Total population between the age of 15 and 19	Object
Estimate!!Total housing units	Total housing units	Object
Estimate!!CITIZEN, VOTING AGE POPULATION!!Citizen, 18 and over population	Population who can vote	Object
Estimate!!CITIZEN, VOTING AGE POPULATION!!Citizen, 18 and over population!!Male	Male population who can vote	Object
Estimate!!CITIZEN, VOTING AGE POPULATION!!Citizen, 18	Female population who can	Object

and over population!!Female	vote	

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

1. Initially, the data was very large with many Nan values and dummy data. So, dropped the margin, annotate, percent of error columns. Changed the column name from default values to values that were in 1st row.
2. Renamed the 1st two columns (i.e. Geography to countyFIPS and Geographic area name to County Name) so that they can be merged with super covid data.
3. Modified the values in Geography so that they can be merged with countyFIPS values in the covid data.
4. Also, split the Geographic area name into county name and stripped the state name, to use it to map as well.
5. Finally, merging the demographic data with the covid data using countyFIPS and county name.

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

The enrichment data can be useful to gain insights into how these characteristics are related to the spread of covid-19, which impacted the housing and which age group/race of people was mostly impacted.

Hypothesis questions.

1. In areas where the housing is dense, will that affect the spread of Covid-19?
2. Does the number of Covid-19 cases depend on the race of people?
3. The people who spread Covid-19 more are Male or Female?