Group 4

Name: Sakshi Patel

Selected Dataset Name: Census Demographic ACS

ACS (American Community Survey) Demographic and Housing Estimates

About Dataset:

This is American Community Survey (ACS) produces population, demographic and housing unit estimates for 2020. The 2020 Census provides the official counts of the population and housing units for the counties

Source:

U.S. Census Bureau, 2016-2020 American Community Survey 5-Year Estimates

Executive Summary:

This report documents the process of cleaning, merging, and analyzing COVID-19 and demographic data from U.S. counties. The main focus is on the COVID-19 pandemic trends for 2020, particularly in Arizona, as well as enriching the COVID-19 data with demographic information from the American Community Survey (ACS) to better understand the spread of the virus. The steps outlined include data preparation, cleaning, and merging to create a comprehensive dataset. The enriched dataset helps explore various social, economic, and demographic factors that may have influenced the virus's transmission. This analysis allows us to pose key hypotheses regarding the relationship between demographic variables and the spread of COVID-19.

Introduction:

The COVID-19 pandemic has had profound effects on public health, the economy, and society worldwide. To understand its spread and impact, it is essential to analyze not only the reported cases and deaths but also the demographic context of the affected areas. This project aims to clean and merge COVID-19 case and death data with demographic data from the ACS. By analyzing trends in different regions, particularly in Arizona, and enriching the data with demographic information, this report seeks to gain deeper insights into how factors such as population density, age, and socioeconomic status might have influenced the spread and severity of the virus.

Visualization of dataset for one county:

Label (Grouping)	Estimate	Margin of Error	Percent	Percent Margin of Error
SEX AND AGE				
Total population	65432			
Male	32379			
Female	33053			
Sex ratio (males per 100 females)	98.0			
Under 5 years	3776			
5 to 9 years	4581			
10 to 14 years	5598			
15 to 19 years	4966			
20 to 24 years	4632			
25 to 34 years	8411			
35 to 44 years	7934			
45 to 54 years	6257			
55 to 59 years	3501			
60 to 64 years	4754			
65 to 74 years	6913			
75 to 84 years	3166			
85 years and over	943			
Median age (years)	35.8			
Under 18 years	16961			
16 years and over	50686			
18 years and over	48471			
21 years and over	45281			
62 years and over	13911			
65 years and over	11022			
18 years and over	48471			
Male	23902			
Female	24569			
Sex ratio (males per 100 females)	97.3			
65 years and over	11022			
Male	4945			
Female	6077			
Sex ratio (males per 100 females)	81.4			
RACE				
Total population	65432			
One race	61959			
Two or more races	3473			
One race	61959			
White	12705			
Black or African American	320			
American Indian and Alaska Native	48041			
Cherokee tribal grouping	N			
Chippewa tribal grouping	N			
Navajo tribal grouping	N			
Sioux tribal grouping	N			
Asian	368			
Asian Indian	N			
Chinese	N			
Filipino	N			
Japanese	N			
Korean	N			
Vietnamese	N			
Other Asian	N			

Native Hawaiian and Other Pacific Islander	13			
Native Hawaiian	N			
Guamanian or Chamorro	N			
Samoan	N			
Other Pacific Islander	N			
Some other race	512			
Two or more races	3473			
White and Black or African American	161			
White and American Indian and Alaska Native	295			
White and Asian	60			
Black or African American and American Indian and Alaska Native	85			
Black or African American and Some Other Race	0			
Race alone or in combination with one or more other races				
Total population	65432			
White	15587			
Black or African American	566			
American Indian and Alaska Native	49031			
Asian	445			
Native Hawaiian and Other Pacific Islander	N			
Some other race	3367			
HISPANIC OR LATINO AND RACE				
Total population	65432			
Hispanic or Latino (of any race)	4728			
Mexican	3288			
Puerto Rican	144			
Cuban	0			
Other Hispanic or Latino	1296			
Not Hispanic or Latino	60704			
White alone	12328			
Black or African American alone	318			
American Indian and Alaska Native alone	47177			
Asian alone	368			
Native Hawaiian and Other Pacific Islander alone	13			
Some other race alone	0			
Two or more races	500			
Two races including Some other race	43			
Two races excluding Some other race, and Three or more races	457			
Total housing units	29042			
CITIZEN, VOTING AGE POPULATION				
Citizen, 18 and over population	23738			
Male	23738			
Female	24358			

In Dataset column name format example:
Percent Margin of Error!!RACE!!Total population!!One race!!Asian!!Other Asian

Enrichment data and datatype - variable dictionary:

You can find the variable dictionary as file name **ACS Demographic and Housing Estimates_Variable dictionary.csv** where we can see the each column name with it's datatype value.

Work:

1. Preparing the Dataset for Cleaning

Before initiating the data cleaning process, three essential datasets were uploaded:

1. covid_deaths_usafacts.csv – Data on COVID-19 deaths by county.
2. covid_confirmed_usafacts.csv – Confirmed COVID-19 cases by county.
3. covid_county_population_usafacts.csv – Population data for U.S. counties.

2. Data Cleaning

For the COVID-19 deaths, confirmed cases, and population datasets, the following data cleaning steps were carried out:

- **Removal of Invalid Rows**: Rows where countyFIPS had a value of 0 were removed, as these are not valid county codes and represent areas with a population of 0. This step was crucial for ensuring data consistency and relevance.
    - *COVID Deaths*: 3142 rows, 1269 columns.
    - *COVID Cases*: 3142 rows, 1269 columns.
    - *Population*: 3144 rows, 4 columns.

3. Merging the COVID-19 Data

Next, the datasets were merged based on the countyFIPS column using an inner join, ensuring that only rows with matching county codes across all datasets were retained. This resulted in the creation of the super_covid19_dataframe.csv, which combined COVID-19 cases, deaths, and population data:

- **Final Dataset**: 3142 rows and 2535 columns.

4. Analyzing COVID-19 Data for 2020

The merged dataset was then filtered to focus solely on COVID-19 data from the year 2020. This reduced the dataset to 3142 rows and 695 columns. The filtered data was essential for further trend analysis.

5. COVID-19 Trends for the Last Week of 2020 (Arizona)

Arizona's data was analyzed to observe COVID-19 trends for the last week of 2020, broken down by county:

- **Counties with Increasing Cases**: Maricopa, Pima, Pinal, Yuma.
- **Counties with Stable Cases**: Apache, Cochise, Coconino, etc.

The data was visualized using line plots for each county to show the trends in COVID-19 cases.

6. Enrichment Data: ACS Demographic and Housing Estimates:
To enhance the analysis, demographic data from the ACS Demographic and Housing Estimates dataset was included. This data provided information such as population estimates, sex ratios, and more.

7. Cleaning ACS Data

The ACS dataset was cleaned as follows:

- **Removal of Puerto Rico Data**: Rows where Geography column values began with 0500000US7 were excluded.
- **Prefix Removal**: The Geography column's prefix (0500000US) was removed to match the countyFIPS codes from the COVID-19 dataset.
- **Column Renaming**: The Geography column was renamed to countyFIPS.
- **Data Type Correction**: The countyFIPS column's data type was converted to an integer to ensure compatibility for merging with other datasets.
- **Dropping Unnecessary Columns**: An unnamed column containing NaN values was removed.

After cleaning, the ACS dataset had 3143 rows and 358 columns.

8. Merging the Enriched Dataset

Finally, the cleaned ACS demographic dataset was merged with the COVID-19 dataset using an outer join on column **countyFIPS** to ensure that all relevant records from both datasets were included:

- **Final Merged Dataset**: 3144 rows and 1052 columns.

This final dataset, merge_Enrichment_data.csv, combined COVID-19 data with demographic information, providing a comprehensive view of the pandemic's impact across U.S. counties.

9. Enrichment Data's Role in COVID-19 Spread Analysis

Demographic factors like population density, age distribution, and socioeconomic conditions can affect the transmission rate and mortality of COVID-19 in a region. For instance:

- **Population Density and Housing**: The number of housing units, particularly in relation to the population size, can provide insights into population density and crowding, both of which are factors that increase the likelihood of COVID-19 spread.
- **Age Distribution**: Areas with a higher elderly population might experience higher mortality rates since COVID-19 poses a greater risk to older adults.
- **Sex Ratios and COVID-19**: The dataset provides information about the sex ratio, which can be used to analyze if certain trends in the virus's transmission. For example, men were initially found to have a higher risk of severe outcomes from COVID-19.

10. Initial Hypothesis Questions:

The enriched dataset allows us to pose several hypothesis questions for future analysis:

1. **Does higher population density correlate with a higher rate of COVID-19 cases?**
2. **Are counties with a larger elderly population experiencing higher COVID-19 death rates?**
3. **Does sex ratio influence the COVID-19 death rate?**

Conclusion

This project successfully cleaned, merged, and analyzed COVID-19 data for 2020, focusing on Arizona while also enriching it with demographic data from the ACS. The merged dataset provides a comprehensive view of both COVID-19 trends and the demographic context, offering deeper insights into how factors such as population density and age may influence the spread of the virus. By posing relevant hypotheses, this analysis paves the way for more detailed studies on the role of demographics in the COVID-19 pandemic.