Exploring the COVID-19 Dataset and The American Community Survey Dataset

Kimmi Sin & Brinnah Welmaker

Datasets

JHU Daily Reports COVID-19 Dataset:

- COVID-19 fatality ratios, incident rates, active, recovered, confirmed, and death cases
- Location information admin2, fips, province_state, country_region, latitude, longitude, combined_key

US Census Bureau's American Community Survey (ACS) Dataset:

- US population estimates and percentages of the total population for social, economic, housing, and demographic data from 2015-2019 (error margins included)

Areas of Interest

A closer look into the US experience of the COVID-19 pandemic.

- As income was getting harder to maintain, how were Americans affected as inferred by the population distribution across different housing situations?
- What were the COVID numbers for different ages, education levels, and ancestry?

Staging Tables

COVID-19:

- 6 tables, separated by season and year
- Columns with non-conforming datatypes were kept as STRING types
    - I.e. covid_spring2020.last_update

ACS:

- 4 tables by the type of data they contain
    - Social, demographics, housing, economics
- Data was not queryable due to the 2 axis data structure

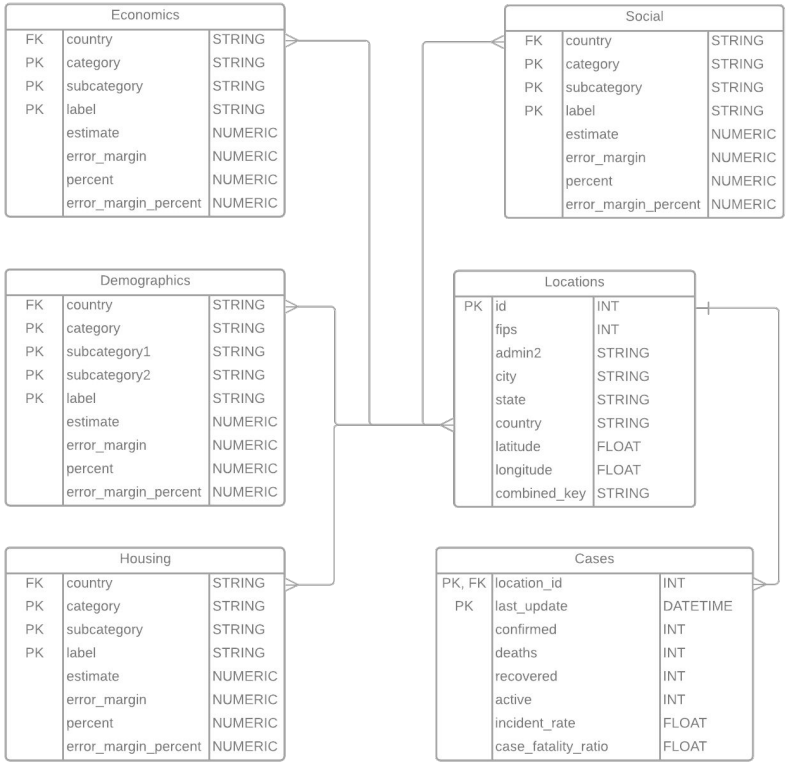
Model Tables

COVID-19:

- Combined all COVID staging tables into one table, then split the columns into either the Locations or Cases table

ACS:

- No staging tables were merged
- The 2 axis data structure was translated into a category and subcategory system



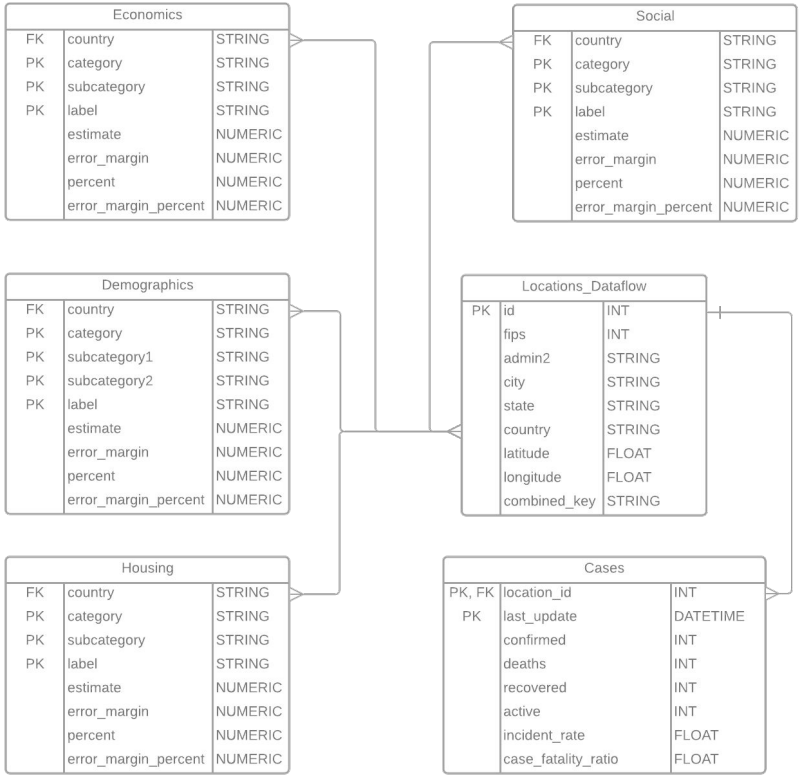
Beam Pipelines

COVID-19:

- Removed duplicate Location records based on longitude and latitude
- Standardized US state names

ACS:

- No pipeline written

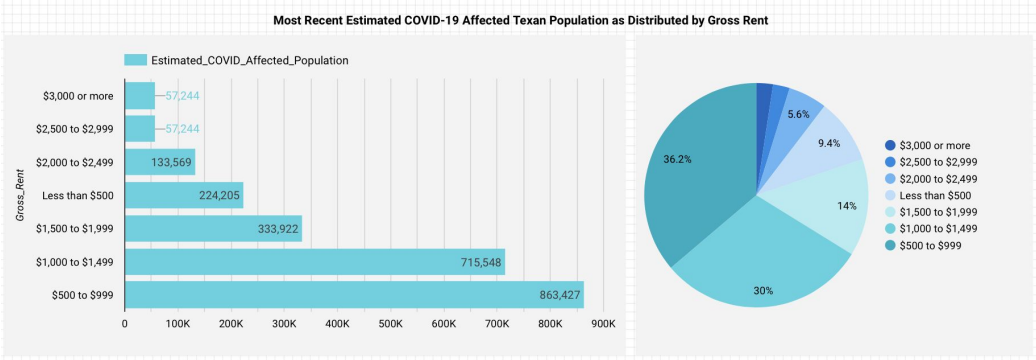


SQL Queries & Data Visualizations

Display the population of Texans affected by COVID-19 based on the average incident rate of the most recent updated Texan cases, and distribute the population by the gross rent they pay. (Assuming that all Texans occupy housing units that require rent, that the percentages in the housing table is applicable to Texans as is, and that the total Texas population is 8.74% of the US's total population)

```
%%bigquery
SELECT label AS Gross_Rent, (h.percent/100)*
                            (((SELECT AVG(incident_rate)
                                FROM datamart.cases
                                WHERE last_update=(SELECT MAX(last_update)
                                                   FROM datamart.cases AS c
                                                   JOIN datamart.locations_Dataflow AS l ON c.location_id=l.id
                                                   WHERE l.state='Texas' AND l.country='US'))/100000)
                            *
                            ((SELECT estimate FROM datamart.demographics WHERE label='Total population')*8.74/100)) AS Estimated_COVID_Affected_Population
FROM datamart.housing AS h
WHERE subcategory='gross rent' AND percent IS NOT null
ORDER BY Gross_Rent
```

	Gross_Rent	Estimated_COVID_Affected_Population
0	1,000to1,499	715547.538817
1	1,500to1,999	333922.184781
2	2,000to2,499	133568.873912
3	2,500to2,999	57243.803105
4	$3,000 or more	57243.803105
5	500to999	863427.363506
6	Less than $500	224204.895496

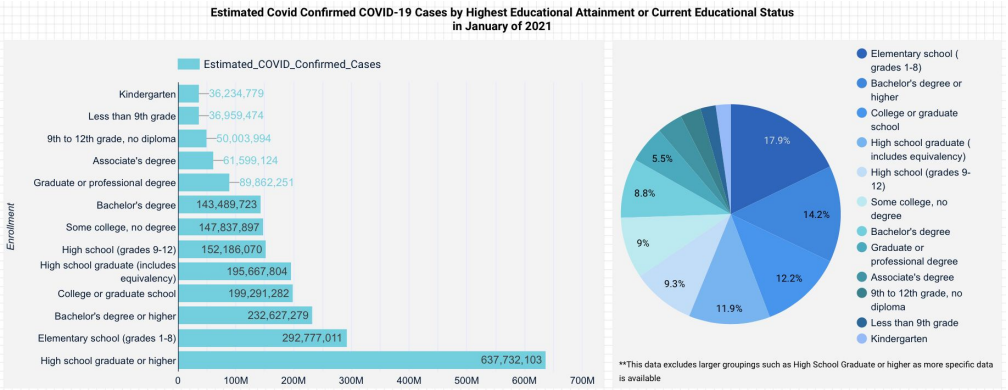


SQL Queries & Data Visualizations

Display the estimated number of confirmed US COVID-19 cases during January of 2021 per current educational status (those in school) as well as highest educational attainment (those out of school), ordered by ascending number of confirmed cases

```
%%bigquery
SELECT s.label AS Enrollment, (s.percent/100)*
                              (SELECT SUM(c.Confirmed)
                                  FROM datamart.cases AS c
                                  JOIN datamart.locations_Dataflow AS l ON c.location_id=l.id
                                  WHERE l.country='US' AND c.last_update between '2021-01-01' AND '2021-01-31 23:59:59') AS Estimated_COVID_Confirmed_Cases
FROM datamart.social AS s
WHERE (s.subcategory='School Enrollment' OR s.subcategory = 'Educational Attainment') AND s.percent IS NOT null
ORDER BY Estimated_COVID_Confirmed_Cases
```

	Enrollment	Estimated_COVID_Confirmed_Cases
0	Kindergarten	36234778.600000000
1	Less than 9th grade	36959474.172000000
2	9th to 12th grade, no diploma	50003994.468000000
3	Associate's degree	61599123.620000000
4	Graduate or professional degree	89862250.928000000
5	Bachelor's degree	143489723.256000000
6	Some college, no degree	147837896.688000000
7	High school (grades 9-12)	152186070.120000000
8	High school graduate (includes equivalency)	195667804.440000000
9	College or graduate school	199291282.300000000
10	Bachelor's degree or higher	232627278.612000000
11	Elementary school (grades 1-8)	292777011.088000000
12	High school graduate or higher	637732103.360000000

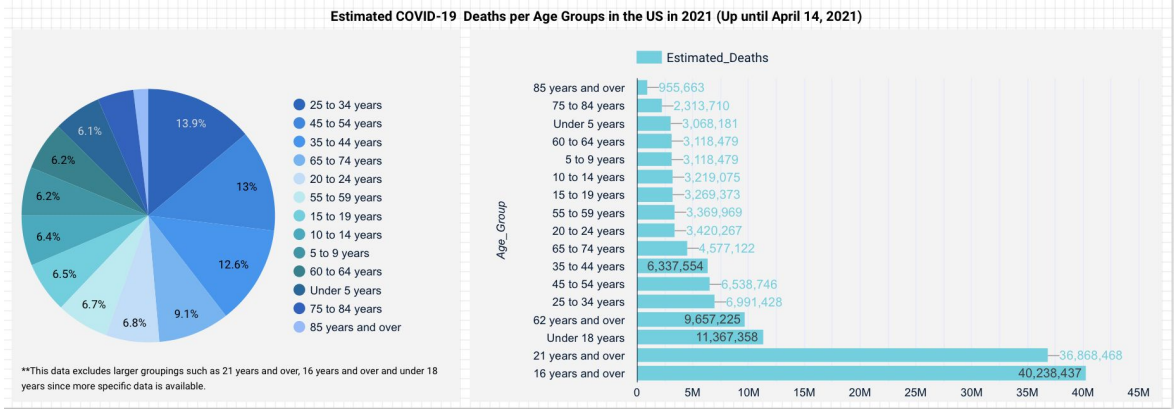


SQL Queries & Data Visualizations

Display the estimated number of deaths per age group in the US during 2021 (up until April 14, 2021), ordered by estimated deaths in descending order

```
%%bigquery
SELECT d.label AS Age_Group, (d.percent/100)*
                                    (SELECT SUM(deaths)
                                    FROM datamart.cases AS c
                                    JOIN datamart.locations_Dataflow AS l ON c.location_id=l.id
                                    WHERE l.country='US' AND last_update between '2021-01-01' and '2021-04-14 23:59:59') AS Estimated_Deaths

FROM datamart.demographics as d
WHERE (subcategory1 = 'age' AND subcategory2 is null AND d.percent is not null)
ORDER BY Estimated_Deaths DESC
```



	Age_Group	Estimated_Deaths
0	16 years and over	40238436.800000000
1	21 years and over	36868467.718000000
2	Under 18 years	11367358.396000000
3	62 years and over	9657224.832000000
4	25 to 34 years	6991428.394000000
5	45 to 54 years	6538745.980000000
6	35 to 44 years	6337553.796000000
7	65 to 74 years	4577122.186000000
8	20 to 24 years	3420267.128000000
9	55 to 59 years	3369969.082000000
10	15 to 19 years	3269372.990000000
11	10 to 14 years	3219074.944000000
12	5 to 9 years	3118478.852000000
13	60 to 64 years	3118478.852000000
14	Under 5 years	3068180.806000000
15	75 to 84 years	2313710.116000000
16	85 years and over	955662.874000000

Challenges & Future Improvements

- ACS data was nested
- COVID-19 Tables had inconsistencies
- Time accurate ACS Data, as well as more extensive
- ACS dataset could have been replaced with another dataset with information by state
    - Alt: another dataset for non-US countries to make comparison studies
- More specific case data