Analyzing COVID-19 Data to Inform Future Pandemic Responses in Texas

[Abdul Wasay, Om Sumedh Kharwadkar /Team 1]

Executive Summary: The ongoing analysis of the COVID-19 pandemic aims to explore the intricate relationships between demographic, socio-economic, and health factors and their influence on virus transmission and mortality rates. By examining data primarily from 2020 to 2022, this project addresses critical public health challenges, particularly in understanding how various variables such as population density, transit options, ethnicity, income, age, and vaccination rates correlate with COVID-19 outcomes. The significance of this project lies in its potential to inform future public health strategies and resource allocation, aiming to reduce the impact of the pandemic and enhance community resilience.

Key findings from the correlation analysis reveal that confirmed cases are positively associated with Hispanic populations and negatively correlated with income levels. Higher vaccination rates correlate with lower mortality rates and families with young children have minimal impact on transmission rates, suggesting targeted public health strategies could be developed without disproportionate focus on this demographic.

These insights are valuable for stakeholders, including public health officials, policymakers, and community organizations, as they prepare for future health crises. The project will facilitate evidence-based decisions on vaccine distribution, healthcare resource allocation, and targeted interventions contributing to more effective public health responses.

Table of Contents

1 Business Understanding

Business Understanding: Frame the Problem

The global COVID-19 pandemic highlighted the critical need for understanding population dynamics, healthcare resources, and socio-economic factors that influenced virus transmission and mortality rates. This project analyzes data across different time periods of the pandemic, primarily focusing on the first two years (2020–2022). By assessing key metrics such as infection rates, hospitalizations, and demographic trends, the goal is to develop insights that can inform public health strategies for both ongoing pandemic recovery and preparation for future pandemics.

Flattening the curve is important because it allows hospitals and healthcare facilities to manage the influx of patients more effectively, ensuring that there are sufficient resources—such as beds, ventilators, and medical staff—to treat those who are severely ill. By spreading out the number of infections over time, public health officials aim to minimize peak demand on healthcare services, thereby reducing the overall mortality rate.

It is crucial to look at data regarding virus spread, hospitalizations, and available resources to inform public health strategies. Analyzing key metrics such as infection rates and demographic trends enables stakeholders to develop targeted interventions, allocate resources efficiently, and implement preventive measures that can ultimately save lives.

Stakeholder Identification

In addition to public health officials and policymakers, secondary stakeholders such the Texas local government, non-governmental organizations (NGOs), and healthcare providers are crucial. These stakeholders need actionable insights to make community-level decisions, such as resource allocation (e.g., vaccines, medical supplies) and public awareness campaigns. By understanding local needs and vulnerabilities, they can deploy targeted interventions to reduce the impact of outbreaks on healthcare systems and economically vulnerable populations.

For example, policymakers could use the analysis to decide on:

- Vaccine distribution policies—Should vaccines be prioritized in urban or rural areas?
- Healthcare resource allocation—Which counties need more hospital beds, ventilators, or mobile testing units?
- Public health communication strategies—How should culturally specific outreach programs be designed to encourage preventive behaviors among diverse populations?

Key Questions to Address

1. **How do population density and urbanization correlate with COVID-19 deaths and spread?**
   Understanding the relationship between population density and virus transmission can help target urban areas for preventive measures. However, if the data indicates that urban centers already have significant restrictions in place, it may be prudent to focus more efforts on rural areas, where information may not disseminate as rapidly, and access to healthcare resources may be limited.
2. **What role do transit options play in the spread of COVID-19?**
   Identifying high-traffic transit areas can inform strategies to mitigate spread in these locations. For instance, in cities with high transit usage, such as New York or Chicago, did this contribute to a higher number of deaths per capita? Conversely, should attention also be directed toward cities like Dallas or Houston, which are predominantly car-based, to assess whether the spread was equally significant?
3. **How do ethnicity and cultural background affect COVID-19 outcomes?**
   Cultural factors may influence health behaviors and social responses to a pandemic. Is there any correlation between ethnicity and outcomes during COVID-19? Furthermore, can this correlation be linked to access to healthcare or socioeconomic status?
4. **What is the impact of income and wealth on COVID-19 deaths and spread?**
   Socioeconomic status can affect access to healthcare and living conditions, potentially leading to disparities in health outcomes. Does income level influence the spread of COVID-19, or is it primarily related to mortality rates due to better healthcare access?
5. **What is the effect of lockdown measures on the spread of COVID-19?**
   Evaluating the effectiveness of lockdowns can guide future public health responses. Over time, did health statistics level out among cities with and without lockdowns? Additionally, what impact did lockdown measures have on death rates? Understanding this can address the longstanding debate over whether shutting down the economy was justified.
6. **What are the age-related risks associated with COVID-19 spread and mortality?**
   Identifying vulnerable age groups can aid in prioritizing vaccination efforts and health resources. Should we have focused all our efforts on the aging population, or did the data indicate that spread and mortality rates were consistent across age groups?
7. **How do family behaviors involving children influence COVID-19 transmission?**
   Understanding how children, who were generally less affected by the virus, contribute to its spread can inform family-focused public health strategies. Did families with children have a greater impact on transmission rates, given that children could experience milder symptoms while spreading the virus to more vulnerable populations?

8. **What is the relationship between vaccination rates and COVID-19 spread and deaths?**
   Understanding the impacts of vaccination is crucial for developing effective strategies for future vaccine rollouts.
9. **How do health risk factors influence COVID-19 outcomes?**
   Identifying additional health risk factors, such as pre-existing conditions, can help target interventions to the most vulnerable populations.

Data Requirements

To address these questions, the following data will be needed:

- COVID-19 case and mortality data by demographic characteristics (age, ethnicity, income, education level)
- Population density metrics
- Public transit usage statistics
- Vaccination rates and their correlation with infection rates
- Socioeconomic data related to income and wealth
- Data on health behaviors, including smoking rates and the prevalence of other health conditions

2 Data Understanding

Data Source Descriptions

1. Global Mobility Report

- **Source**: Google
- **Link**: Global Mobility Report
- **Overview**:

  The Google Global Mobility Reports provide insights into how communities are moving in response to COVID-19. They aim to assist public health officials by offering aggregated, anonymized movement trends over time by geography and across various categories.
- **Key Data Points**:
  - **Transit Stations Visits**: Important for understanding changes in public transit usage.
  - **Workplace Visits**: Useful for assessing economic activity and potential transmission in work settings.
- **Expected Data Quality and Reliability**:

  High-quality data due to robust collection methods, although variations in user engagement with location services may affect representation. This dataset is going to be called the ***GMR dataset*** for the rest of the report.

2. USAFacts (COVID Infections/Deaths) + Census Data

- **Source**: USAFacts
- **Link**: USAFacts COVID Data
- **Overview**:

  This dataset provides comprehensive COVID-19 case and death counts by state and county, sourced from the CDC and state and local health agencies. It also includes Census data, offering essential demographic information useful for integrated analysis.
- **Key Data Points**:
  - **COVID-19 Case and Death Counts**: Detailed statistics on infections and fatalities by state and county.
  - **Population Demographics**: Essential for analyzing the impact of COVID-19 across different demographic groups.
- **Expected Data Quality and Reliability**:

  High-quality data from reputable sources like the CDC, but potential variations in reporting practices may exist. This dataset is going to be called the ***Census dataset*** for the rest of the report.

3. Texas State (COVID Infections/Deaths)

- **Source**: USAFacts
- **Link**: USAFacts COVID Data
- **Overview**:

  This dataset provides comprehensive COVID-19 case and death counts specifically by counties in Texas, sourced from the CDC and state and local health agencies.
- **Key Data Points**:
  - **County-Level COVID-19 Data**: Detailed statistics on infections and fatalities for Texas counties.
- **Expected Data Quality and Reliability**:

  High-quality data with possible variations due to reporting practices across localities. This dataset is going to be called the *TX dataset* for the rest of the report. However, this dataset was not used extensively because we used the USAFacts dataset that had more variables and filtered for Texas.

4. US Counties Health and Vaccination Data

- **Source**: US County Health Ranking Dataset
- **Link**: US SocioHealth Data
- **Overview**:

  This dataset integrates health, socio-economic, and weather conditions across the 3,142 counties in the U.S. to identify populations at greater risk for COVID-19.
- **Key Data Points**:
  - **County-Level Health Data**: Provides insight into the health of residents of that county.
  - **Vaccination Data**: Correlates vaccinations to the counties.
- **Expected Data Quality and Reliability**:

  High-quality data from reputable sources, but missing data for certain counties may occur. This dataset is going to be called the *Health dataset* for the rest of the report.

Data Selection, Description and Quality

We aim to divide the usage of data from each of these datasets into how they help answer our questions raised earlier. Each question is formulated in a way to relate to the number of cases and deaths related to the pandemic within that area. However for the sake of data selection and cleaning we will organize the following based on data source. Tables with variables organized by question are also available in the appendix. Given that this report only focus on Texas, the data relating to other states will be removed when analyzing for data quality.

Data Source	Variable Name	Description	Type	NAs	Outliers %	Negative Values	Duplicates	Total Values
Census Data	total_pop	Total population of the county	numeric	0	15.35433	0	0	254
	county_fips_code	Unique identifier for each county	integer	0	0	0	0	254
	commuters_by_public_transportation	Number of commuters using public transportation	numeric	0	15.35433	0	150	254
	commuters_by_car_truck_van	Number of commuters using cars, trucks, or vans	numeric	0	15.35433	0	1	254
	commuters_by_subway_or_elevated	Number of commuters using subways or elevated transit	numeric	0	11.02362	0	228	254
	white_pop	White population	numeric	0	12.59843	0	1	254
	black_pop	Black population	numeric	0	14.17323	0	34	254
	asian_pop	Asian population	numeric	0	18.11024	0	92	254
	hispanic_pop	Hispanic population	numeric	0	15.35433	0	2	254
	amerindian_pop	American Indian population	numeric	0	14.96063	0	104	254
	other_race_pop	Other racial groups	numeric	0	17.32283	0	165	254
	two_or_more_races_pop	Population identifying with two or more races	numeric	0	16.14173	0	50	254
	not_hispanic_pop	Non-Hispanic population	numeric	0	13.38583	0	0	254

	median_income	Median income of households	numeric	0	3.937008	0	3	254
	median_age	Median age of the population	numeric	0	1.181102	0	100	254
	families_with_young_children	Number of families with young children	numeric	0	15.35433	0	7	254
	high_school_diploma	Number of people with a high school diploma	numeric	0	11.02362	0	3	254
	bachelors_degree	Number of people with a bachelor's degree	numeric	0	15.35433	0	7	254
	graduate_professional_degree	Number of people with graduate or professional degrees	numeric	0	14.96063	0	17	254
	bachelors_degree_or_higher_25_64	Individuals aged 25-64 with a bachelor's degree or higher	numeric	0	15.35433	0	8	254
TX Data	county_fips_code	Unique identifier for each county	Duplicated from larger census file					
	county_name	Name of the county	Duplicated from larger census file					
GMR Data	sub_region_1	Name of the first sub-region (e.g., state or province)	character	0	NA	0	64161	64162
	sub_region_2	Name of the second sub-region (e.g., county or city)	character	0	NA	NA	63942	64162

	metro_area	Metropolitan area associated with the data	logical	64162	NA	0	64161	64162
	census_fips_code	FIPS code related to the census	integer	343	0	0	63942	64162
	transit_stations_percent_change_from_baseline	Percentage change in visits to transit stations compared to a baseline period	integer	34109	0.589134	19408	63972	64162
	date	Date at which the data was recorded	character	0	NA	0	63819	64162
	workplaces_percent_change_from_baseline	Percentage change in visits to workplaces compared to a baseline period	integer	4789	2.236526	55827	64044	64162
Health Data	total_population	Total population of the county	integer	0	14.06481	0	59891	60143
	area_sqmi	Area of the county in square miles	numeric	0	17.64461	0	59889	60143
	population_density_per_sqmi	Population density per square mile	numeric	0	14.50709	0	59889	60143
	fips	FIPS code for the county	character	0	NA	0	59889	60143
	percent_below_poverty	Percentage of the population below the poverty line	numeric	0	4.221605	0	59999	60143
	percent_unemployed_CDC	Percentage of unemployed individuals	numeric	0	5.297375	0	60044	60143
	eightieth_percentile_income	Income at the eightieth percentile	numeric	0	6.115425	0	59889	60143

	twentieth_percentile_income	Income at the twentieth percentile	numeric	0	3.2589	0	59890	60143
	income_ratio	Ratio of high-income to low-income households	numeric	0	2.761751	0	59889	60143
	stay_at_home_announced	Date when stay-at-home orders were announced	character	0	NA	0	60141	60143
	stay_at_home_effective	Date when stay-at-home orders took effect	character	0	NA	0	60141	60143
	date_stay_at_home_announced	Specific date for when stay-at-home was announced	character	0	NA	0	60142	60143
	percent_fair_or_poor_health	Percentage of the population reporting fair or poor health	numeric	0	7.279318	0	59889	60143
	percent_smokers	Percentage of smokers	numeric	0	0.927789	0	59889	60143
	percent_adults_with_obesity	Percentage of adults classified as obese	numeric	0	0.842991	0	59996	60143
	percent_physically_inactive	Percentage of physically inactive individuals	numeric	0	0	0	60012	60143
	percent_with_access_to_exercise_opportunities	Percentage of individuals with access to exercise opportunities	numeric	0	0	0	59895	60143
	percent_excessive_drinking	Percentage of individuals engaging in excessive drinking	numeric	0	0.766506	0	59889	60143

	percent_adults_with_diabetes	Percentage of adults diagnosed with diabetes	numeric	0	0.804749	0	60007	60143
	percent_vaccinated	Percentage of the population vaccinated	numeric	71	0.339192	0	60100	60143

Table 1: Variable description and quality statistics

There are no negative values in any of the numeric columns, and no missing values (NAs) are present in the listed census data variables. This is advantageous for analyses, as missing data can complicate interpretations and lead to biased estimates. Our primary concern regarding duplicates was with the FIPS code in the census file, but there are no duplicates found, indicating that each county has a unique record.

While **total_population**, **area_sqmi**, and **population_density_per_sqmi** also show high outlier percentages, these can be justified due to the significant variation among counties in Texas. Areas such as Dallas, Houston, and Austin have much higher populations compared to other regions.

Some variables exhibit significantly high percentages of outliers, which may indicate irregular values attributed to errors in the data. Variables such as **total_population**, **commuters_by_public_transportation**, **commuters_by_car_truck_van**, **commuters_by_subway_or_elevated**, **white_pop**, **black_pop**, **asian_pop**, **hispanic_pop**, **amerindian_pop**, **families_with_young_children**, **bachelors_degree**, **graduate_professional_degree**, and **bachelors_degree_or_higher_25_64** will be examined for their correlation with total population to determine if they fall outside expected ranges. Additionally, their distribution will be assessed using histograms.

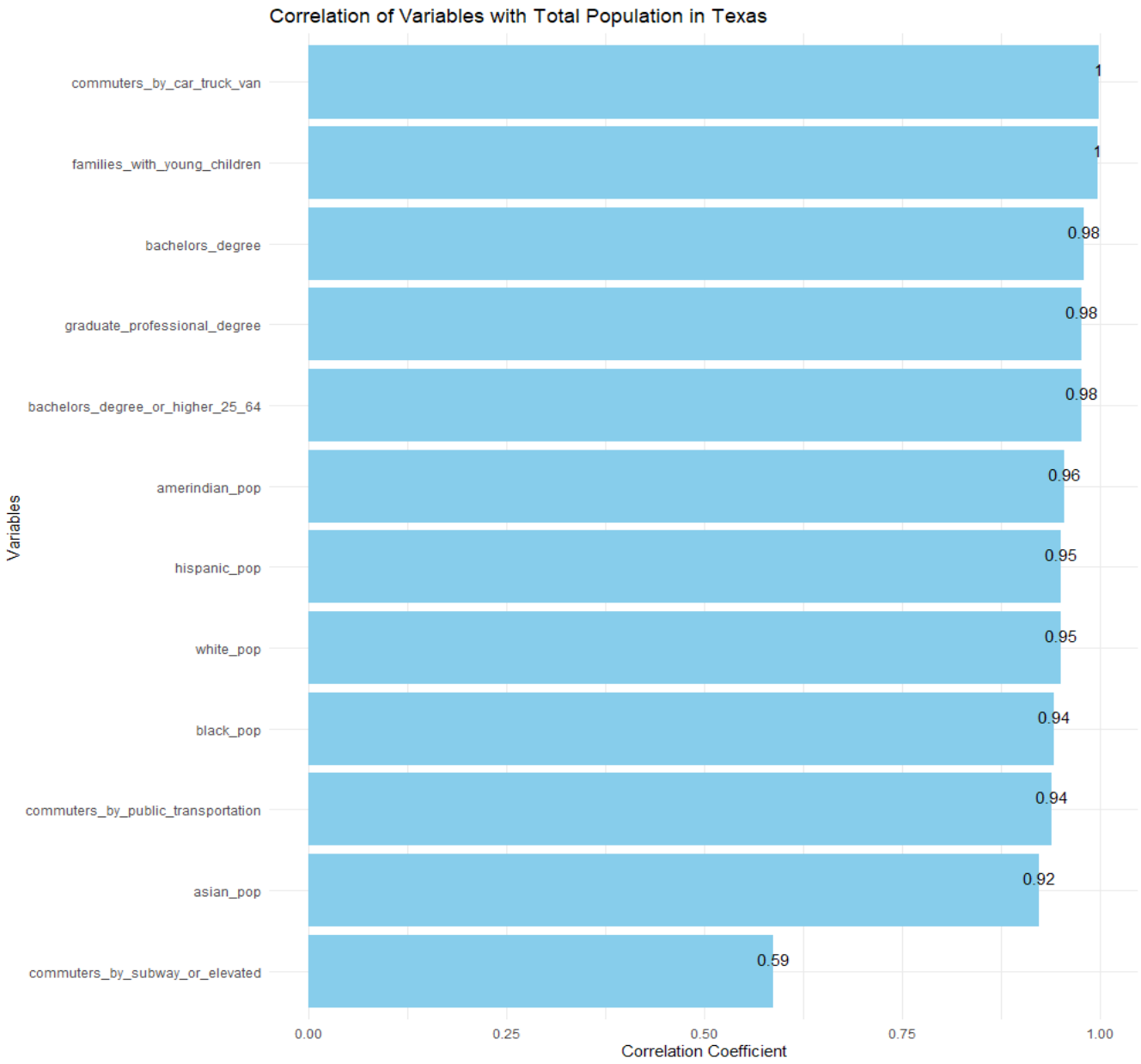


Illustration 1: Outlier variables correlation with county population

Given the correlation bar graph, we can see that all the variables tested exhibit over 90% correlation with the county's total population. This indicates that the outliers may be reflective of genuine population characteristics rather than data errors.

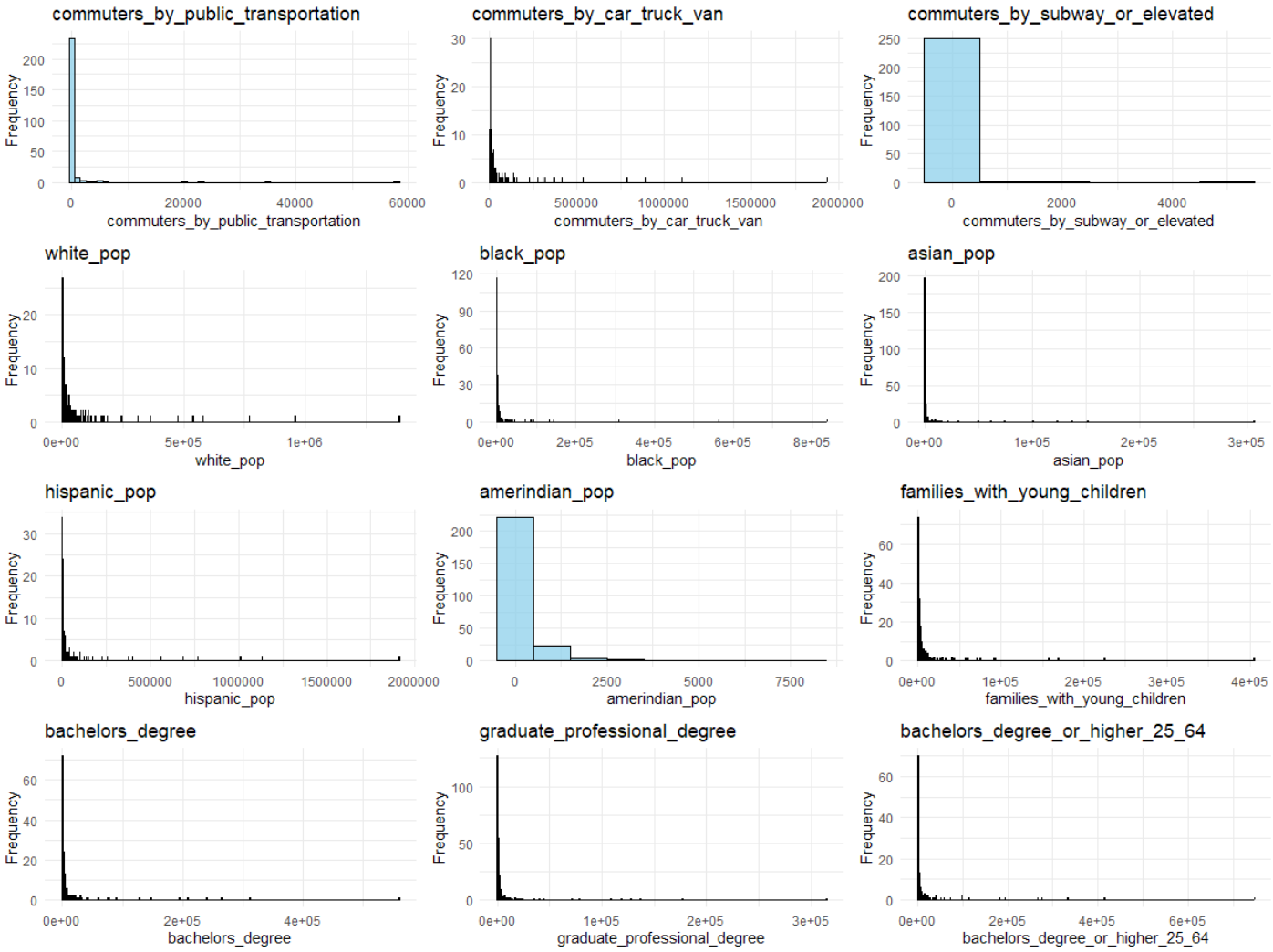


Illustration 2: Outlier variables histogram with county population bins of 1000

The histograms also show a significant left skew, with a large number of values concentrated in the initial bin of 1,000. This suggests that the majority of counties have populations clustered around lower values, highlighting the presence of a few counties with significantly higher populations that contribute to the outlier status.

Data Statistics

The interpretations for this section are provided as the insights in the right most column in the table below. They are based on the statistics derived from the selected variables.

Variable from Census data	Mean	Median	Variance	Range	Mode	Insights
total_pop	107951.2	18612.5	1.52E+11	74 - 4525519	74	Total population has a very large range, and a very large variance meaning there is a significant difference in county sizes from large to small.
commuters_by_public_transportation	737.4449	10	21857885	0 - 57933	0	Public transportation commuters indicate low usage, with mode of zero, which means Texas reliance on public transport is limited.
commuters_by_car_truck_van	44871.37	6537.5	2.79E+10	33 - 1930740	1796	
commuters_by_subway_or_elevated	36.93701	0	103196.4	0 - 4667	0	
white_pop	46281.47	9404	1.87E+10	55 - 1386576	9404	The white population is dominant in many Texas counties, but the distribution varies widely.
black_pop	12594.57	676	4.59E+09	0 - 838285	0	Counties with a significant Black population might require targeted outreach and resources, as minority communities have often experienced worse outcomes during the pandemic.
asian_pop	4814.862	73.5	6.71E+08	0 - 307109	0	The Asian population is relatively small but significant in a few counties.
hispanic_pop	42023.26	5068.5	2.97E+10	12 - 1910535	0	The Hispanic population is large in many Texas counties, with some counties having a majority Hispanic population. Public health strategies should consider language barriers and cultural differences to ensure effective communication and

						vaccination campaigns in these areas. Hispanic communities may also face challenges like limited healthcare access, which can affect outcomes.
amerindian_pop	259.3819	41	588213.2	0	365	Although smaller in number, the American Indian population is important to consider due to the historical healthcare disparities faced by Indigenous communities.
other_race_pop	154.1457	2.5	578293.2	0 - 9681	0	
two_or_more_races_pop	1744.122	201.5	42249928	0 - 62712	0	Public health campaigns should be inclusive and consider the complex identity and healthcare needs of individuals who identify with more than one race, especially as they may experience different socioeconomic or healthcare access issues.
not_hispanic_pop	65927.96	62	11419	62 - 2614984	62	
median_income	49894.34	48311	1.47E+08	24794 - 93645	42500	There is a moderate range in median income across counties.It can affect public health measures, such as social distancing or lockdown compliance, potentially influencing covid spread and outcomes.
median_age	39.01929	38.55	35.59263	25.8 - 57.5	39.01929	The median age can influence the severity of covid outcomes, as older populations are generally at greater risk.
families_with_young_children	9033.823	1244.5	1.17E+09	0 - 404641	63	Since children may act as vectors for spreading the virus in households and schools, counties with a higher number of families with young children could benefit from focused family-based health strategies and vaccination programs.
high_school_diploma	14130.69	3281	2.28E+09	3 - 559393	695	
bachelors_degree	12947.94	1289.5	2.59E+09	11	237	

graduate_professional_degree	6778.024	511.5	7.93E+08	0 - 314848	6778.024	
bachelors_degree_or_higher_25_64	16530.56	1289.5	4.58E+09	3 - 746724	138	
Variable	Mean	Median	Variance	Range	Mode	Insights
transit_stations_percent_change_from_baseline	-9.36333	-8	519.593	-221	0	The negative mean value indicates a significant drop in visits to transit stations compared to pre-pandemic levels. This suggests reduced public transportation use during the pandemic, likely due to lockdowns, remote work, or fear of virus transmission. The large range also shows that some counties experienced more drastic reductions than others.
workplaces_percent_change_from_baseline	-23.5368	-23	189.8938	-129	-18	The substantial decline in workplace visits indicates a widespread shift to remote work or reduced economic activity during the pandemic. This data highlights the significant impact of covid on businesses and employment, with many workplaces seeing reduced foot traffic.
Variable from health data	Mean	Median	Variance	Range	Mode	Insights
total_population	119469.8	20370	1.67E+11	76 - 4434257	10367	Texas counties vary widely in population size. This large variance highlights the urban-rural divide in population density, with a few counties having very large populations while the majority are smaller.
area_sqmi	1011.762	908.3914	396977.6	127.106 - 6183.759	1239.877	
population_density_per_sqmi	125.4584	24.65959	123354.9	0.113 - 2878.423	1499.099	There is a significant variation in population density, with some counties being highly urbanized while others are sparsely populated.
percent_below_poverty	16.86725	16.5	33.06407	1.8 - 37.9	17.2	The poverty rate is relatively high in many Texas counties, with some areas seeing almost 40% of their

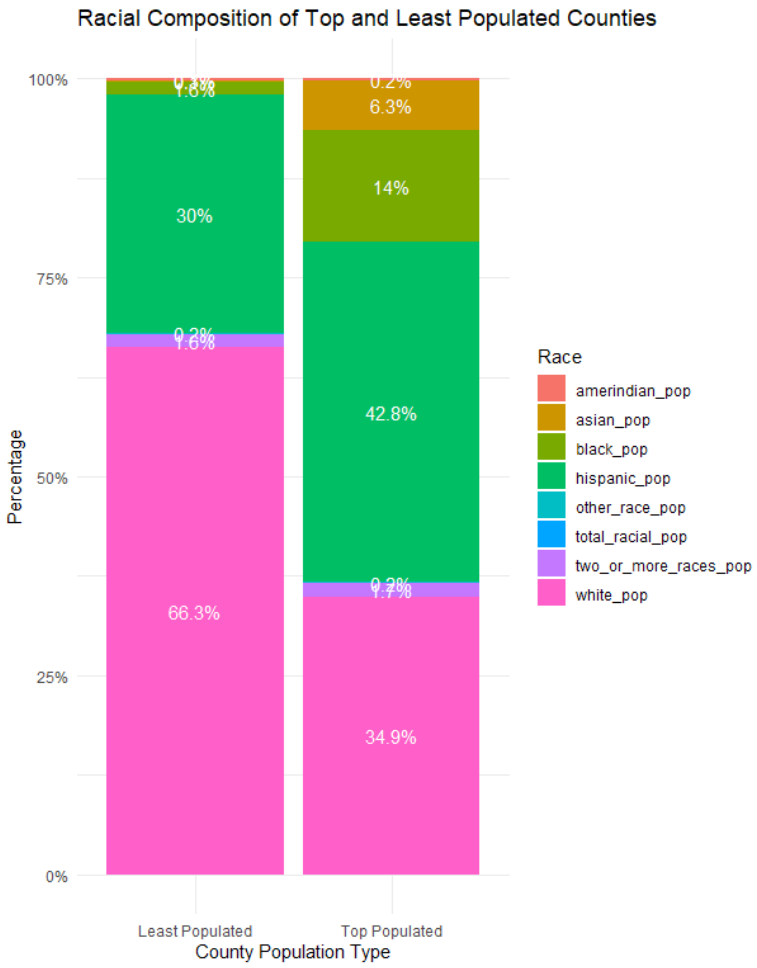
						population living below the poverty line.
percent_unemployed_CDC	6.429995	6	7.240308	0 - 18.4	6	Unemployment rates vary significantly, with some counties facing much higher unemployment during the pandemic.
eightyeth_percentile_income	101252.7	98523	4.15E+08	59219 - 178345	110196	
twentieth_percentile_income	22181.22	21232	34990610	22181.22 - 21232	22181.22	
income_ratio	4.681269	4.56706	0.482222	9750 - 43865	20375	
percent_fair_or_poor_health	20.59802	19.72772	23.84551	12.287 - 40.990	4.742264	
percent_smokers	14.98326	14.8993	2.337278	10.643 - 19.871	14.8993	Smoking, another risk factor which is severe, is prevalent in many counties. Understanding the correlation between smoking rates and covid outcomes could inform targeted health messaging and support for reducing risk in smokers.
percent_adults_with_obesity	31.4913	30.7	27.44873	21.6 - 47.3	28.6	The obesity rate is fairly high across the state, with more than 30% of the population in many counties classified as obese. This could help guide targeted health interventions in counties with higher obesity rates.
percent_physically_inactive	27.42912	27.2	21.92578	16.6 - 39.6	27.8	Poverty rates are critical in understanding the disparities in healthcare access and outcomes.
percent_with_access_to_exercise_opportunities	57.6936	60.17223	533.8134	0 - 97.566	0	
percent_excessive_drinking	17.98989	17.88092	3.23828	17.88092 - 17.98989	0	

percent_adults_with_diabetes	11.45414	10.7	21.3368	3.5 - 29.3	7.7	Diabetes rates are fairly high. Considering >10% of the population suffers from it.
percent_vaccinated	37.02822	39	90.27449	9-55	38	Fairly low range of vaccination rates across Texas counties.

Table 2: Variables with statistics

Visual Exploration

For the following section we looked at interesting variables to see how they vary across different Texas counties.



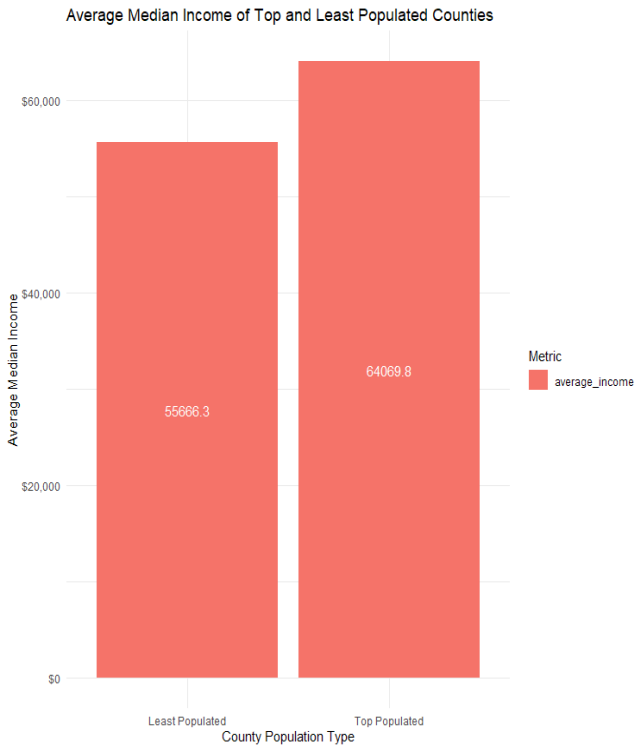


Illustration 3: Racial Composition of Top 10 Most Populated and Least Populated Counties

Illustration 4: Average Median Income of Top 10 Most Populated and Least Populated Counties

Illustration 3 was chosen to highlight the racial diversity in the most and least populated counties in Texas. Understanding the racial composition can provide insights into how demographic factors might correlate with health outcomes and social behaviors during the pandemic. We chose a stacked bar graph so that lower numbers from low population counties can still be interpreted and compared. The population of white people increased a lot more (from 34 to 66%) when moving to low population areas. Whereas the population of ethnically hispanic and ethnically black people followed the opposite trend (42% to 30%) and (1.6 to 14%) respectively.

For Illustration 4, we aimed to analyze how the population of a county may affect its median income. While we anticipated that lower-populated counties would exhibit lower median incomes, we were surprised by the relatively small difference of approximately $10,000 between the most populated and least populated counties. This suggests that despite the population size, income levels do not vary as dramatically as one might expect, indicating that perhaps income is not a major contributor to the accessibility of good healthcare, the location may be more important.

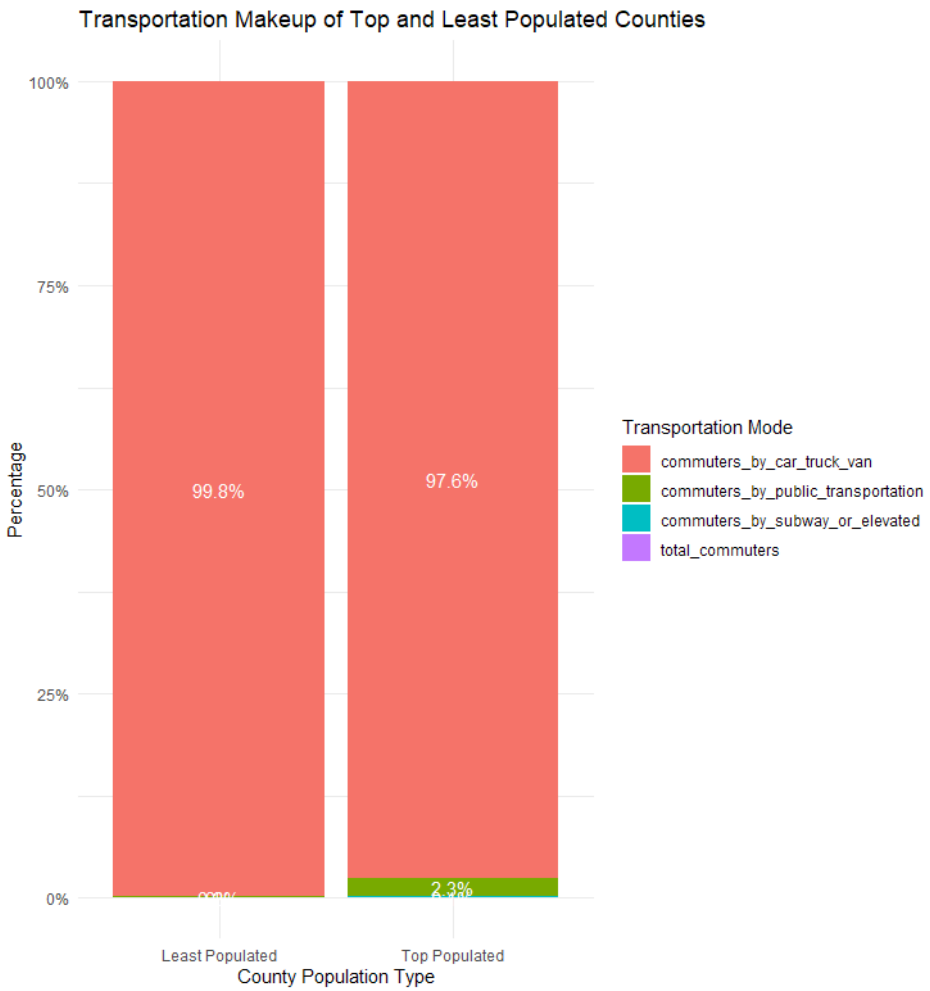
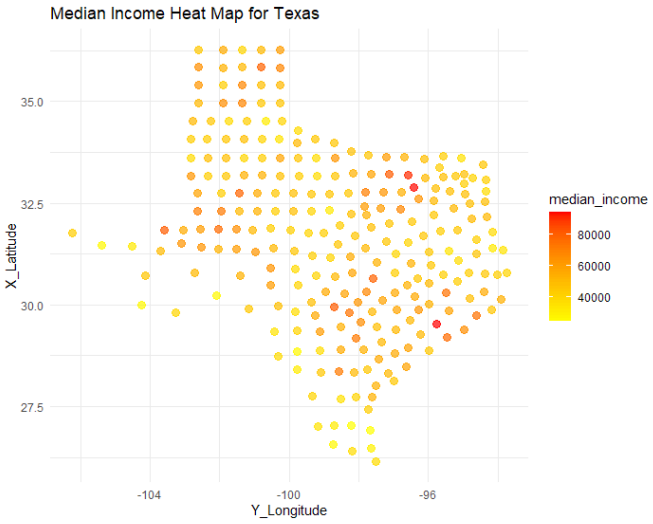


Illustration 5: Transportation of choice for of Top 10 Most Populated and Least Populated Counties

This stacked bar graph allows for a clear comparison of transportation choices between the most and least populated counties, which can impact COVID-19 transmission dynamics. Texas is not known for its public transportation, we expected public transport utilization to be fairly low for low populations, but important to note here is that it's almost just as low for high population counties. Meaning that this variable may not be important for our report.



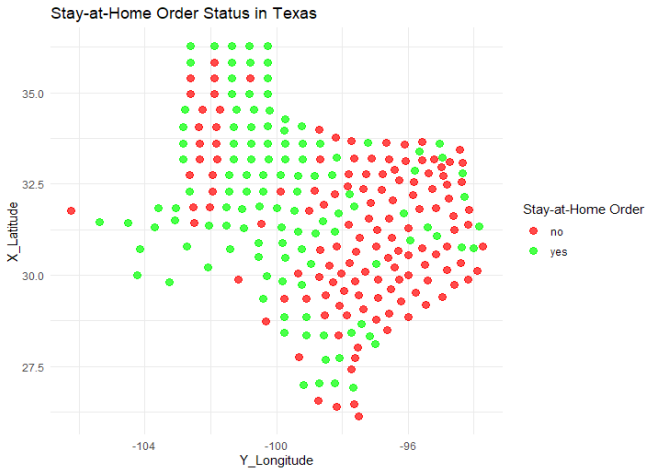
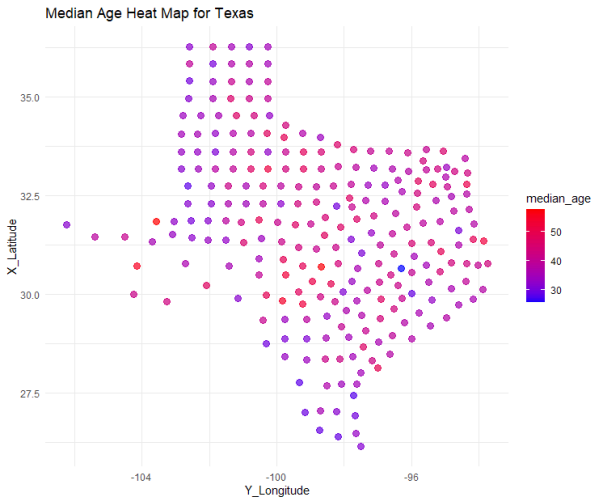


Illustration 6: Median Income Heatmap for Texas Counties

Illustration 7: Stay at Home Orders Heatmap for Texas Counties

The heatmap (illustration 6) was selected to visually represent the distribution of median income across Texas counties. We could assume that higher population centers like Dallas, Austin, San Antonio, Houston would have higher incomes, but it's interesting to see other areas that are not as population dense also having higher incomes. This will be worth exploring when creating clusters as higher incomes usually means access to better healthcare, but does it hold true when we move out of the main cities, where medical institutions are farther away.

This heatmap (illustration 7) above visualizes the implementation of stay-at-home orders across Texas, allowing for an assessment of how these measures varied geographically. It shows a clear indication that higher population areas like Dallas, Austin, San Antonio, Houston did not implement stay at home orders at the start of the pandemic. This means it's a worthy variable worth studying in our correlation matrix to see how it relates to the spread.



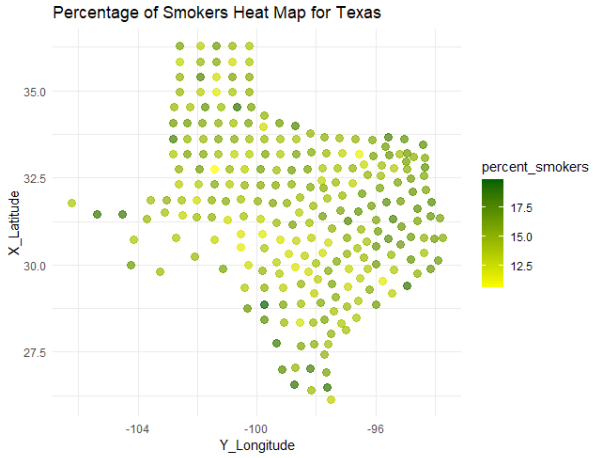


Illustration 8: Median Age Heatmap for Texas Counties

Illustration 9: Percentage of Smokers Heatmap for Texas Counties

The heat maps (illustration 8 and 9) above were chosen to see if there is any relationship with the high population density centers (Dallas, Austin, San Antonio, Houston) / low population density with the number of smokers and median age of the people living there. The map shows that high population density means that people smoke less. The same high density locations also have younger median populations. The median age heatmap helps us visualize the counties that may be more vulnerable to COVID-19. Identifying age demographics can help target vaccination and health resources effectively.

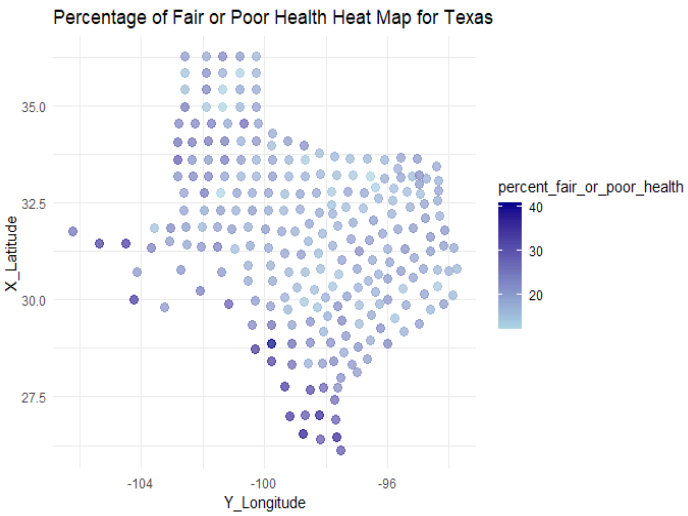


Illustration 10: Percentage of Fair to Poor Health Heatmap for Texas Counties

This heatmap (illustration 10) was chosen to visualize health status across Texas counties, helping to identify areas that may require additional health resources and interventions. The heatmap indicates that counties with a high percentage of residents reporting fair to poor health are primarily in the south and southwestern border, whereas healthier populations are concentrated in high population centers like Dallas, Houston, San Antonio and Austin. .

Relationship Exploration

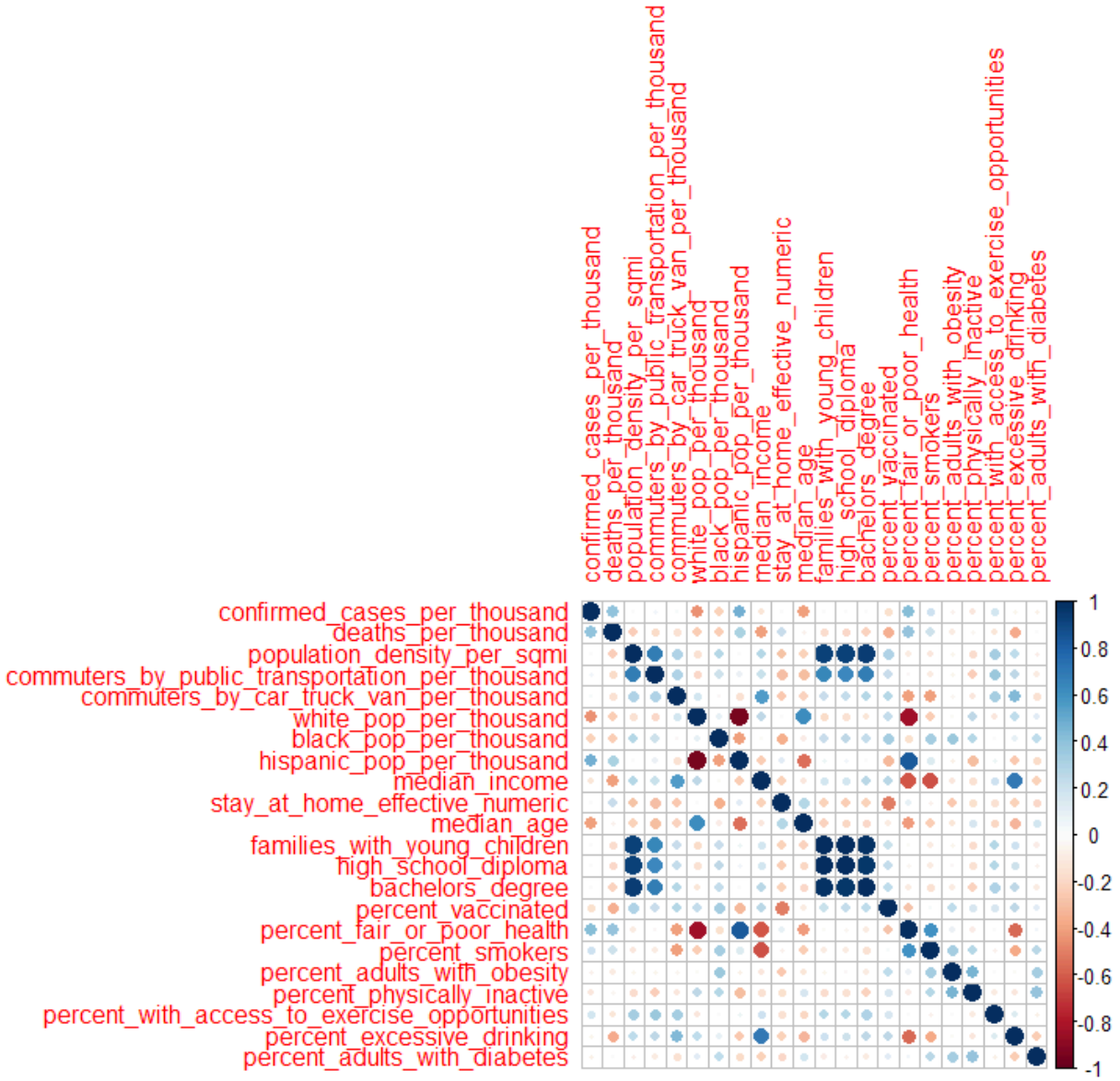


Illustration 11: Correlation Matrix of chosen Variables with Confirmed Covid Cases

A correlation matrix provides a comprehensive view of relationships between multiple variables and confirmed COVID cases. This visualization aids in identifying which factors are most strongly associated with case numbers. Interpretation of this illustration is below.

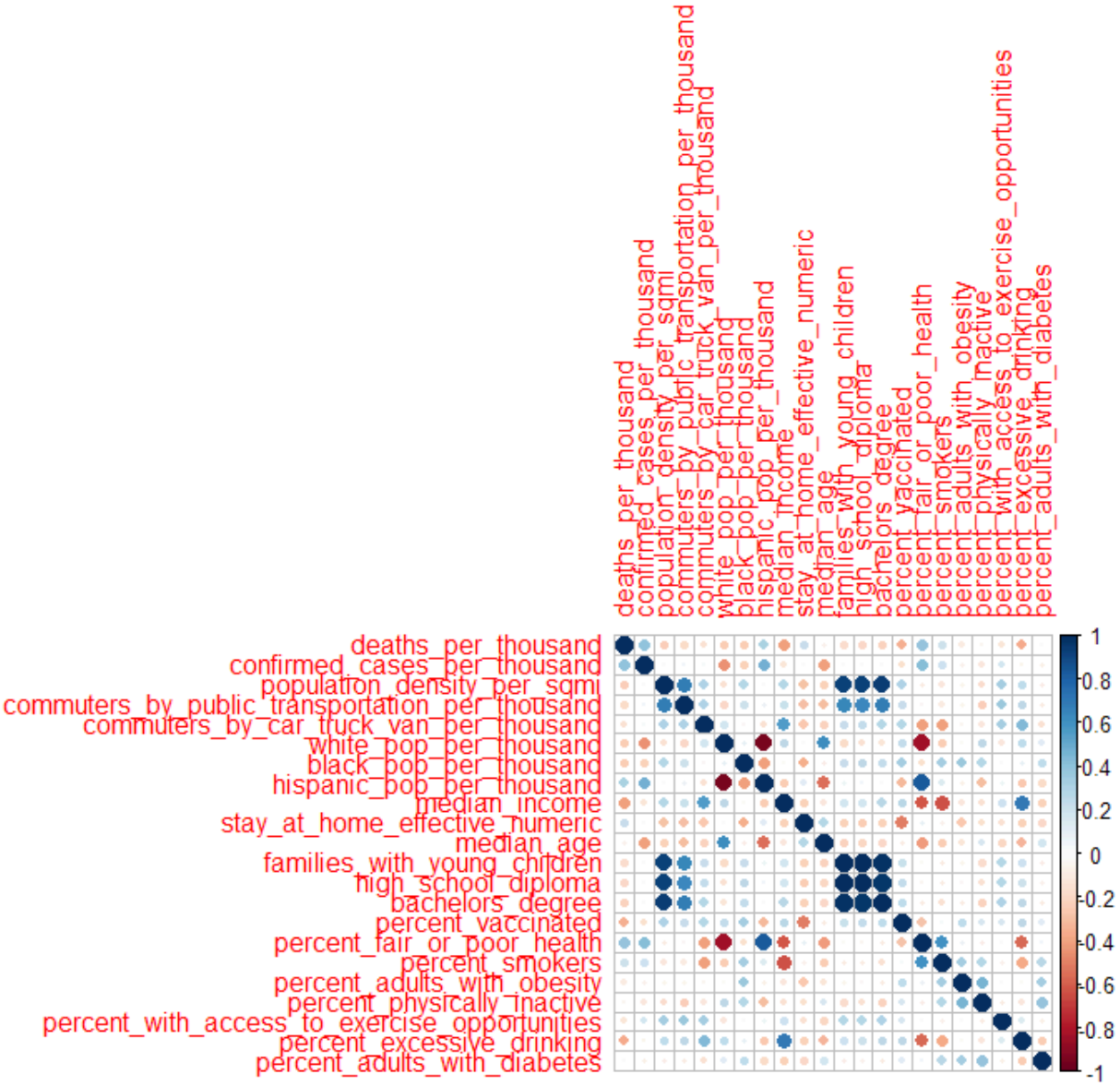


Illustration 12: Correlation Matrix of chosen Variables with Confirmed Covid Death

A correlation matrix provides a comprehensive view of relationships between multiple variables and confirmed COVID deaths. This visualization aids in identifying which factors are most strongly associated with deaths. Interpretation of this illustration is below.

Key Points from the Correlation Matrices:

1. Deaths and Confirmed Cases:
    - There is a moderate positive correlation between deaths per thousand and confirmed cases per thousand (0.39). This suggests that as the number of confirmed cases increases, the number of deaths also tends to increase, which is expected in a public health context.

2. Population Density:
    - Population density has a negative correlation with deaths (-0.24) and a very weak positive correlation with confirmed cases (0.02). This indicates that more densely populated areas may not necessarily have higher death rates, possibly due to better healthcare access or other factors.
3. Transit Options:
    - Commuters using public transportation show a weak negative correlation with deaths (-0.18) and a negligible positive correlation with confirmed cases (0.04). This might suggest that areas with more public transport options do not necessarily have higher death rates.
4. Ethnic Relations:
    - The correlation of hispanic population per thousand with deaths (0.31) and confirmed cases (0.47) indicates that higher proportions of the Hispanic population may be associated with more cases and deaths. The correlations for white (-0.25 for deaths and -0.45 for cases) and black populations (-0.25 for deaths and -0.23 for cases) suggest a negative relationship, indicating that higher proportions of these ethnicities might be linked with fewer cases and deaths.
5. Income:
    - There is a strong negative correlation between median income and deaths (-0.40) as well as confirmed cases (-0.13). This suggests that higher median incomes are associated with lower death rates, potentially reflecting better access to healthcare.
6. Lockdown:
    - The numeric representation of stay-at-home effective shows a weak positive correlation with deaths (0.22) but negligible correlation with confirmed cases (0.01). This might indicate that effective lockdown measures could be associated with lower death rates.
7. Age:
    - The median age has a weak negative correlation with deaths (-0.07) and a stronger negative correlation with confirmed cases (-0.41). This suggests that higher median ages may be linked to higher case rates, possibly due to increased vulnerability in older populations.
8. Children:
    - Families with young children show a negative correlation with deaths (-0.18) and a negligible positive correlation with confirmed cases (0.01). This might indicate that areas with more young families are not necessarily linked with higher death rates.
9. Education Levels:
    - There are negative correlations for both high school diploma (-0.19 for deaths) and bachelor's degree (-0.22 for deaths), suggesting that areas with higher education levels may have lower death rates.
10. Vaccination:

- The percent vaccinated shows a strong negative correlation with deaths (-0.35) and a weaker negative correlation with confirmed cases (-0.17). This indicates that higher vaccination rates are associated with lower death rates, consistent with vaccination's role in preventing severe outcomes.

11. Health Factors:
    - Percent fair or poor health has a strong positive correlation with deaths (0.39) and a positive correlation with confirmed cases (0.42), suggesting that areas with poorer health status experience more severe COVID-19 outcomes.
    - Percent smokers shows a positive correlation with both deaths (0.21) and confirmed cases (0.19), indicating a possible relationship between smoking prevalence and COVID-19 impacts.

The correlation matrix indicates several important relationships between COVID-19 deaths and confirmed cases with various demographic, socioeconomic, and health-related factors. These insights can help in further analysis and policy decisions aimed at mitigating the impact of COVID-19 in different communities. For example, improving healthcare access in lower-income or higher-density areas, increasing vaccination efforts, and supporting health education could be effective strategies.

3 Data Preparation

The dataset combined_data was created by merging multiple data sources relevant to Texas counties, including demographic, health, and socioeconomic information. The initial data was sourced from a census dataset that provided population and racial demographics, along with COVID-19 statistics such as confirmed cases and deaths. This was then combined with additional information regarding transportation, education, and health factors from various sources. Key variables were transformed to ensure comparability, such as converting counts into per-thousand measures to normalize data across counties of varying sizes. The final dataset has counties as individual rows and various attributes as columns, allowing for comprehensive analysis of factors influencing COVID-19 cases and outcomes. This dataset was used to answer questions in section 2 about correlations, that have told us which questions that we initially asked are relevant and worth creating models for and exploring further.

To enhance clarity and comprehension, the table displaying the top ten rows of the prepared dataset has been transposed. The csv file for this dataset is included in the submission zip.

total_pop	5532	1084	2932	3836	8145	3067	289	1591
county_fips_code	48195	48433	48079	48105	48307	48421	48269	48045
commuters_by_public_transportation	9	0	7	0	2	7	0	0
commuters_by_car_truck_van	2183	418	1130	1673	3281	1458	158	602

commuters_by_subway_or_elevated	0	0	0	0	0	0	0	0
white_pop	2950	787	1104	1264	5067	1712	212	964
black_pop	4	38	77	35	161	8	0	27
asian_pop	0	0	0	0	0	0	0	0
hispanic_pop	2569	246	1672	2537	2571	1327	77	513
amerindian_pop	6	0	6	0	1	7	0	0
other_race_pop	0	0	0	0	202	0	0	0
two_or_more_races_pop	3	13	73	0	143	13	0	15
not_hispanic_pop	2963	838	1260	1299	5574	1740	212	1008
median_income	40678	46786	37500	52310	42367	54961	56964	42500
median_age	34.9	51.9	33.5	35.5	43.5	39.8	32.7	46.4
families_with_young_children	451	75	272	406	589	230	23	84
confirmed_cases	673	134	216	472	467	118	11	114
deaths	17	4	12	13	12	11	0	7
total_population	5552	1233	2955	3836	8242	3069	274	1672
area_sqmi	919.8095	916.3122	775.0819	2807.323	1065.601	923.0352	910.8721	900.5631
population_density_per_sqmi	6.036033	1.345611	3.8125	1.366427	7.734603	3.3249	0.300811	1.73225
percent_below_poverty	20	18.9	19.6	22.7	18.4	13.6	3.3	25
percent_unemployed_CDC	2.5	6.9	7	2.8	4.2	1	4.1	5.9
eightieth_percentile_income	83438	101500	91417	108737	82214	105000	112900	64359
twentieth_percentile_income	21161	18250	14632	21040	18534	27500	35786	15222
income_ratio	3.943008	5.561644	6.247745	5.168108	4.435848	3.818182	3.154865	4.228025
stay_at_home_announced	yes	yes	yes	yes	yes	yes	yes	yes
stay_at_home_effective	yes	yes	yes	yes	yes	yes	yes	yes
date_stay_at_home_announced	3/31/2020	3/31/2020	3/31/2020	3/31/2020	3/31/2020	3/31/2020	3/31/2020	3/31/2020

percent_fair_or_poor_health	24.09282	17.358	29.51378	21.74686	20.1542	18.58456	17.41465	21.87081
percent_smokers	16.33528	13.46802	17.98534	12.93194	14.43336	13.42591	17.41465	15.7592
percent_adults_with_obesity	25	26.9	25.2	27.1	35	27.7	22.7	21.6
percent_physically_inactive	21.9	23	22.2	21.7	31.7	27.2	20.7	28.5
percent_with_access_to_exercise_opportunities	87.68929	53.08725	47.52159	18.31137	53.9901	37.60712	71.22786	84.18605
percent_excessive_drinking	17.24832	16.5449	15.39152	17.33939	17.00149	19.10179	16.68972	25.9
percent_adults_with_diabetes	6.4	7.5	16.7	5.1	13.1	4.2	11.2	9.1
percent_vaccinated	29	12	21	23	23	36	29	41
date	4/7/2020	7/18/2020	4/23/2020	6/5/2020	4/9/2020	4/16/2020	########	4/27/2020
lat	36.27744	33.17919	33.60418	30.72309	31.19888	36.27773	33.61655	34.53028
lon	-101.355	-100.253	-102.829	-101.412	-99.3475	-101.893	-100.256	-101.209

4 Recommendations

Further Exploration Needed:

1. **Hispanic Population**: The strong positive correlation (0.47 with cases and 0.31 with deaths) suggests that areas with higher Hispanic populations may experience more COVID-19 cases and deaths. Investigating the social determinants, healthcare access, and vaccination rates within these communities can yield important insights.
2. **Income Levels**: The negative correlation between median income and both confirmed cases (-0.13) and deaths (-0.40) indicates that lower-income areas may face greater challenges. Studying the impact of income on healthcare access, preventive measures, and overall community resilience against COVID-19 is essential.
3. **Vaccination Rates**: The negative correlation with deaths (-0.35) emphasizes the importance of vaccination efforts. Further analysis of vaccination accessibility, public health campaigns, and the impact of vaccination on various demographics could provide valuable information for future interventions.

4. **Health Factors**: The strong positive correlation of percent fair or poor health (0.39 with deaths, 0.42 with cases) and the positive correlation with smokers (0.19 with cases and 0.20 with deaths) warrant deeper investigation into health disparities and the role of pre-existing conditions in COVID-19 outcomes.
5. **Education Levels**: The correlation between educational attainment and death rates (negative correlations for high school and bachelor's degrees) suggests that educational interventions could be beneficial. Investigating how education impacts public health responses and individual health behaviors during pandemics could be valuable.

Questions to Drop or Deprioritize:

1. **Population Density**: The weak positive correlation with confirmed cases (0.02) and negative correlation with deaths (-0.24) suggests that population density alone may not be a significant factor in COVID-19 outcomes. Future analyses may focus on other variables instead.
2. **Transit Options**: The weak correlations with both confirmed cases (0.04) and deaths (-0.18) indicate that transit options (public transportation vs. cars) do not significantly impact COVID-19 spread or mortality. This area may be deprioritized in favor of more impactful factors.
3. **Children and Young Families**: The negligible correlations with confirmed cases (0.01) and negative correlation with deaths (-0.18) imply that the presence of families with young children does not play a significant role in COVID-19 outcomes. Thus, this line of inquiry can be dropped.

Next Steps

1. **Model Development**:
    - Create statistical models to test the relationships identified, particularly focusing on the variables with strong correlations. This may involve regression analysis to quantify the impact of each factor on COVID-19 cases and deaths.
2. **Cluster Analysis**:
    - Identify clusters of counties with similar characteristics (e.g., high Hispanic populations, low income) to target public health interventions effectively. This could involve using machine learning techniques such as k-means clustering or hierarchical clustering.
3. **Public Health Strategy**:
    - Develop actionable strategies based on the findings, particularly aimed at increasing vaccination rates, improving health education, and addressing healthcare access in lower-income communities.

5 Exceptional Work

For the exceptional work on this project, we are incorporating an additional data set beyond the three provided. We have analyzed this new data set using the same methodology applied to the original three, following the steps outlined in the Data Understanding phase. This involves a thorough examination and transformation of the new data, ensuring consistency in analysis across all sets. The new dataset called "health data" was chosen to include social, health, and vaccination data to the original 3 and it has been joined using the FIPS code into a new, consolidated data set. The variables we retain are those directly relevant to our research questions.

6 List of References

- Kaggle. (n.d.). *Risk factors of COVID-19 in the US*. Kaggle. Retrieved from https://www.kaggle.com/code/rayna3/risk-factors-of-covid-19-in-the-us
- USAFacts. (n.d.). *COVID-19 infections and deaths*. USAFacts. Retrieved from https://usafacts.org/
- Google. (n.d.). *COVID-19 community mobility reports*. Google. Retrieved from https://www.google.com/covid19/mobility/index.html
- Johns Hopkins University. (n.d.). *COVID-19 dashboard by the Center for Systems Science and Engineering (CSSE)*. Retrieved from https://coronavirus.jhu.edu/map.html
- Centers for Disease Control and Prevention (CDC). (n.d.). *COVID-19 data tracker*. CDC. Retrieved from https://covid.cdc.gov/covid-data-tracker

7 Appendix

Tables with variables used by data sources organized by the questions asked.

1. Population density and urbanization correlation with COVID-19

Data Source	Variable Name	Description
Census Data	total_pop	Total population of the county
	housing_units	Total number of housing units
	population_1_year_and_over	Population 1 year and over
	pop_5_years_over	Population 5 years and over
	county_fips_code	Unique identifier for each county

TX Data	county_fips_code	Unique identifier for each county
	county_name	Name of the county
	state	State in which the county is located
	state_fips_code	Unique identifier for the state
GMR Data	country_region_code	Code for the country or region
	country_region	Name of the country or region
	sub_region_1	Name of the first sub-region (e.g., state or province)
	sub_region_2	Name of the second sub-region (e.g., county or city)
	metro_area	Metropolitan area associated with the data
	census_fips_code	FIPS code related to the census
Health Data	total_population	Total population of the county
	area_sqmi	Area of the county in square miles
	population_density_per_sqmi	Population density per square mile
	fips	FIPS code for the county
Geometric Dataset	fips	FIPS code for the county
	shape_area	Area of the county

2. Transit options correlation spread of COVID-19

Data Source	Variable Name	Description
Census Data	commuters_by_public_transportation	Number of commuters using public transportation
	commuters_by_car_truck_van	Number of commuters using cars, trucks, or vans
	commuters_by_subway_or_elevated	Number of commuters using subways or elevated transit

GMR Data	transit_stations_percent_change_from_baseline	Percentage change in visits to transit stations compared to a baseline period
	workplaces_percent_change_from_baseline	Percentage change in visits to workplaces compared to a baseline period

3. Ethnicity and cultural background affect on COVID-19 outcome

Data Source	Variable Name	Description
Census Data	white_pop	White population
	black_pop	Black population
	asian_pop	Asian population
	hispanic_pop	Hispanic population
	amerindian_pop	American Indian population
	other_race_pop	Other racial groups
	two_or_more_races_pop	Population identifying with two or more races
	not_hispanic_pop	Non-Hispanic population

4. Income and wealth on COVID-19 deaths and spread

Data Source	Variable Name	Description
Census Data	median_income	Median income of households
	income_per_capita	Income per capita
	percent_income_spent_on_rent	Percentage of income spent on rent
	income_less_10000	Number of people earning less than $10,000
	income_10000_14999	Number of people earning between $10,000 and $14,999
	income_15000_19999	Number of people earning between $15,000 and $19,999
	...	... (continued for other income brackets)

Health Dataset	percent_below_poverty	Percentage of the population below the poverty line
	percent_unemployed_CDC	Percentage of unemployed individuals
	eightieth_percentile_income	Income at the eightieth percentile
	twentieth_percentile_income	Income at the twentieth percentile
	income_ratio	Ratio of high-income to low-income households

5. Lockdown measures on the spread of COVID-19

Data Source	Variable Name	Description
Health Dataset	stay_at_home_announced	Date when stay-at-home orders were announced
	stay_at_home_effective	Date when stay-at-home orders took effect
	date_stay_at_home_announced	Specific date for when stay-at-home was announced

6. Age-related risks associated with COVID-19 spread and mortality

Data Source	Variable Name	Description
Census Dataset	median_age	Median age of the population
	male_under_5	Population of males under 5 years old (for children)
	female_under_5	Population of females under 5 years old (for children)
	male_65_to_66	Population of males aged 65 to 66
	female_65_to_66	Population of females aged 65 to 66
	...	... (continued for other elderly population segments)

7. Family behaviors involving children influence COVID-19 transmission

Data Source	Variable Name	Description

Census Dataset	families_with_young_children	Number of families with young children
	children	Total number of children in the county
	two_parent_families_with_young_children	Two-parent families with young children

8. Education level correlate with health outcomes during the pandemic

Data Source	Variable Name	Description
Census Dataset	high_school_diploma	Number of people with a high school diploma
	bachelors_degree	Number of people with a bachelor's degree
	graduate_professional_degree	Number of people with graduate or professional degrees
	bachelors_degree_or_higher_25_64	Individuals aged 25-64 with a bachelor's degree or higher
Health Dataset	high_school_graduation_rate	High school graduation rate
	percent_no_highschool_diploma	Percentage of individuals without a high school diploma
	num_some_college	Number of individuals with some college education
	percent_some_college	Percentage of individuals with some college education

9. Vaccination rates and COVID-19 spread and deaths

Data Source	Variable Name	Description
Vaccination Data Set	total_vaccinations	Total number of vaccinations administered
	people_vaccinated	Total number of people who received at least one dose
	people_fully_vaccinated	Total number of people who are fully vaccinated
	total_vaccinations_per_hundred	Total vaccinations per hundred people
	people_vaccinated_per_hundred	Number of people vaccinated per hundred people

	people_fully_vaccinated_per_hundred	Fully vaccinated people per hundred people
	daily_vaccinations	Number of vaccinations administered daily
	daily_vaccinations_per_million	Daily vaccinations per million people
	share_doses_used	Percentage of distributed doses that have been administered
Health Dataset	percent_vaccinated	Percentage of the population vaccinated

10. Health risk factors influence COVID-19 outcomes

Data Source	Variable Name	Description
Health Dataset	years_of_potential_life_lost_rate	Rate of years of potential life lost
	percent_fair_or_poor_health	Percentage of the population reporting fair or poor health
	average_number_of_physically_unhealthy_days	Average number of physically unhealthy days reported
	average_number_of_mentally_unhealthy_days	Average number of mentally unhealthy days reported
	percent_low_birthweight	Percentage of low birthweight infants
	percent_smokers	Percentage of smokers
	percent_adults_with_obesity	Percentage of adults classified as obese
	percent_physically_inactive	Percentage of physically inactive individuals
	percent_with_access_to_exercise_opportunities	Percentage of individuals with access to exercise opportunities
	percent_excessive_drinking	Percentage of individuals engaging in excessive drinking
	percent_food_insecure	Percentage of food-insecure individuals
	percent_insufficient_sleep	Percentage of individuals reporting insufficient sleep
	percent_frequent_physical_distress	Percentage of individuals reporting frequent physical distress

	percent_frequent_mental_distress	Percentage of individuals reporting frequent mental distress
	percent_adults_with_diabetes	Percentage of adults diagnosed with diabetes
	percent_vaccinated	Percentage of the population that is vaccinated

8 Student Contributions

Both members worked on the R file using Google colab and wrote the report using Google Docs.