Data Mining Project 2

Diving Deeper into The Impact of COVID-19 on Texas' Counties

A Data Mining Analysis

By: Juan Carlos Dominguez, Salissa Hernandez, Leonardo Piedrahita

➢ Executive Summary:

This report addresses the ongoing challenge of understanding the spread of COVID-19 across Texas counties and how various factors—such as demographic characteristics, social distancing compliance, economic disparities, and access to remote work—contribute to infection rates and mortality. With the Texas Department of State Health Services (DSHS) as our primary stakeholder, this project aims to provide actionable insights to guide public health policy and resource allocation.

The importance of this work lies in the need for DSHS to identify high-risk areas and vulnerable populations, optimize interventions, and reduce COVID-19 transmission. The project highlights key trends, including how counties with lower income levels experienced higher case rates, the varying impact of social distancing across urban and rural areas, and the strong correlation between poverty, public assistance, and COVID-19 case and mortality rates.

Using advanced clustering techniques—including K-means, DBSCAN, Spectral Clustering, and Mean Shift—our report identified distinct county groupings that reveal patterns of vulnerability and resilience. For example, Cluster 3 captured urban counties with high public transit reliance and poverty rates, which saw elevated case and mortality levels, while Cluster 2 included wealthier counties with more remote work capacity and generally lower infection rates. These insights, reinforced by geospatial and demographic analysis, reveal regional disparities such as elevated death rates in eastern and southeastern Texas and the protective effects of rural isolation in some areas.

These findings equip DSHS with a data-driven foundation to prioritize resources and implement targeted, equitable strategies—helping reduce the public health burden, support vulnerable communities, and enhance resilience in the face of future outbreaks.

Table of Contents

1. Problem Description

1.1 Background

COVID-19 (coronavirus disease 2019), caused by the SARS-CoV-2 virus, is a respiratory illness that emerged in late 2019 and led to a global pandemic. It spreads primarily through respiratory droplets and airborne transmission, causing symptoms ranging from mild respiratory issues to severe complications such as pneumonia, organ failure, and death, particularly in high-risk populations [1]. In response, social distancing measures were enacted as a method to manage and control the spread of the disease. It refers to reducing close physical interactions between individuals to prevent viral spread. This includes staying at least six feet apart, avoiding large gatherings, and minimizing non-essential travel [2]. Measures such as social distancing, mask mandates, lockdowns, and remote work policies aimed to slow the virus's spread. The goal was to flatten the curve and reduce the peak number of cases, preventing healthcare systems from becoming overwhelmed and ensuring medical resources remained available for critical patients [3].

Analyzing COVID-19 data is essential for evidence-based decision-making in public health. Monitoring infections rates, hospital capacity, and medical resource availability helps in identifying outbreak hotspots. Tracking new cases allows health agencies to implement target restrictions before widespread community transmission occurs [4]. Assessing policy effectiveness can be utilized to evaluate trends before and after interventions to determine which strategies are most effective [5]. Data on hospitalizations can guide the appropriate resource allocations of ICU beds, ventilators, and medical personnel where they are needed most [6]. Lastly, disaggregated data by age, race, socioeconomic status, and pre-existing conditions can ensure at-risk groups receive priority care and vaccinations [7].

1.2 Stakeholder: Texas Department of State Health Services (DSHS)

The chosen stakeholder for this analysis is the Texas Department of State Health Services (DSHS). DSHS is responsible for monitoring and managing public health policies, coordinating COVID-19 responses, and allocating medical resources across all Texas counties. The agency collaborates with local public health departments, such as Dallas County Health and Human Services (DCHHS), to implement interventions at both the state and county levels [8]. DSHS is an ideal stakeholder for this analysis because it oversees statewide pandemic response efforts, making it responsible for policy decisions that impact all Texas residents. The agency allocates resources such as vaccines, ventilators, funding, across urban and rural counties with varying healthcare capacities. Given the variation in healthcare infrastructure and socioeconomic conditions across Texas, data-driven insights are necessary to optimize resource distribution and intervention strategies.

1.3 Research Questions & Clustering Approach

This study seeks to group Texas counties based on COVID-19 case trends, mobility patterns, and socioeconomic factors to identify similarities and differences in how regions experienced the pandemic.

The key research questions we need to answer include:

1. **Which Texas counties exhibit the highest COVID-19 infection and mortality rates per capita, and how do demographic factors (e.g., population size, income levels) correlate with these trends?**
    a. *Rationale*: This question aims to identify counties most severely affected by COVID-19, focusing on case and death rates normalized per capita. Understanding the relationship between these trends and demographic factors

like population size and income levels can help prioritize resources and tailor interventions for regions at the highest risk.

2. **How do commuting behaviors—such as public transportation usage and ability to work from home—differ between urban and rural counties, and how might these differences relate to COVID-19 case rates?**
    a. *Rationale*: Public transportation usage and the ability to work from home are key indicators of social distancing compliance and exposure risk. By analyzing these behaviors, this question explores how urban and rural counties differ in mobility patterns, helping assess the potential link between these behaviors and COVID-19 case rates.
3. **Which counties have the lowest per capita access to workers staying at home, and how does this relate to COVID-19 case rates?**
    a. *Rationale*: The ability to work from home is a crucial factor in minimizing exposure to COVID-19. This question identifies counties with lower rates of remote work, assessing the connection between this lack of workplace flexibility and higher COVID-19 case rates, ultimately highlighting areas in need of targeted interventions.
4. **How do poverty and public assistance rates correlate with COVID-19 case and mortality rates across Texas counties?**
    a. *Rationale*: This question explores the intersection of economic vulnerability and COVID-19 impact. By analyzing poverty and public assistance rates, it seeks to determine whether economically disadvantaged counties have experienced higher rates of infection and death, providing insight into areas that may require additional support and resources.
5. **How do COVID-19 case and death rates vary across counties with different socioeconomic profiles, and what disparities emerge across urban vs. rural regions?**
    a. *Rationale*: This question aims to uncover health disparities linked to socioeconomic factors such as income, poverty, and public assistance rates. It also explores differences between urban and rural counties, providing a comprehensive understanding of how underlying socioeconomic conditions contribute to varying COVID-19 outcomes.
6. **To what extent does income inequality (as measured by the Gini Index) correlate with disparities in COVID-19 outcomes across counties?**
    a. *Rationale*: Income inequality can have a significant impact on access to healthcare, living conditions, and overall health outcomes. By analyzing the relationship between the Gini Index and COVID-19 outcomes, this question seeks to quantify the role of structural inequality in shaping health disparities, supporting the design of more equitable health policies and interventions.

These questions are critical because they allow DSHS to identify and address COVID-19 hotspots to implement targeted interventions before widespread transmission occurs. It enables appropriate allocation of medical resources based on the current needs and capacity of local hospitals. DSHS can more effectively evaluate the effectiveness of public health measures to decide whether they should be intensified, relaxed, or restructured. It also allows DSHS to protect vulnerable populations by ensuring that healthcare services, such as vaccination and treatment, are prioritized based on demographic risk factors. By answering these questions, DSHS can make informed, data-driven decisions that directly improve public health outcomes across Texas counties.

DSHS plays a critical role in pandemic response by making data-driven decisions that directly impact COVID-19 transmissions, healthcare system strain, and mortality rates. DSHS can implement or adjust social distancing measures by analyzing infection trends and determining if it is appropriate to tighten or relax mask mandates, capacity limits, or stay-at-home orders. During the Delta variant surge of 2021, DSHS reinstated indoor mask mandates based on rising case counts, which led to a reduction in the growth rates of hospitalizations [9]. DSHS can also allocate medical resources efficiently by utilizing the hospitalization and ICU data to guide staffing and resource distribution across hospitals in Texas. In July 2020, DSHS used hospitalization data to redistribute ventilators to overwhelmed hospitals in southern regions of Texas [10]. DSHS can provide targeted vaccination and outreach programs by conducting demographic analysis to ensure that high-risk and underserved communities receive vaccines promptly. For example, DSHS launched a community-based vaccination clinic in Dallas to combat low vaccination rates and high case counts [11]. Lastly, DSHS can optimize their public health communication strategies to encourage compliance with safety measures. In December 2020, DSHS used Google Mobility Data to warn against holiday travel surges, which reduced post-holiday infection spikes [12]. Our stakeholder can conduct and refine each of these decisions. It directly affects COVID-19 outcomes by controlling transmission, reducing hospital burden, and saving lives.

1.4 Data & Methodology

The analysis will use county-level COVID-19 Census data to cluster counties based on socioeconomic factors and pandemic severity. This dataset includes key variables such as case and death counts, median income, poverty rates, public transit reliance, and remote work prevalence. These factors provide insight into how socioeconomic conditions influenced the spread and impact of COVID-19 across different regions. By analyzing these variables, the study aims to identify patterns in pandemic severity and assess how demographic and economic disparities contributed to variations in outcomes.

Clustering techniques such as **k-means and hierarchical clustering** will be applied to uncover groups of counties with similar characteristics. These clusters will help identify patterns in COVID-19 spread, evaluate the effectiveness of different public health measures, and highlight regions that may require additional support. Comparing cluster results across different feature subsets will provide a more comprehensive understanding of the factors driving COVID-19 outcomes in Texas.

The findings from this analysis will support DSHS in developing **targeted pandemic responses**, ensuring that interventions are tailored to the specific needs of different county groups. This data-driven approach will improve resource allocation, enhance public health strategies, and contribute to more effective management of future public health crises.

2. Data Collection & Data Quality

2.1 Data Source Description, Expected Data Quality & Reliability

For this analysis, the COVID-19 Census dataset was utilized. This dataset was cleaned and prepared in the previous project to ensure data quality and consistency before beginning the analysis, therefore we expect it to be reliable for our analysis. Given its size and complexity, it was essential to inspect the data for potential inconsistencies or gaps, which were addressed during the data cleaning process.

The COVID-19 Census dataset contains demographic and socioeconomic variables for Texas counties, including COVID-19 case and death counts, as well as information on population size, median income, food stamp usage, and public transportation reliance. This dataset consists of 3,142 rows and 259 columns and is designed to assess how various socioeconomic factors relate to COVID-19 outcomes. In the previous project, data cleaning steps included standardizing variable names, ensuring appropriate data types, and addressing missing values to optimize the dataset for analysis.

Rather than integrating multiple datasets, the COVID-19 Census dataset is analyzed independently to focus specifically on the relationship between socioeconomic and demographic factors and pandemic outcomes. This approach allows for a more detailed and precise analysis, ensuring that insights remain specific to socioeconomic influences on COVID-19 severity. By keeping this dataset separate, a clearer interpretation of its findings is possible before considering potential intersections with other external factors.

2.2 Clustering Objects

The objects for clustering in this study are **Texas counties**, as the goal is to identify regions with similar COVID-19 trends, mobility patterns, and socioeconomic factors. Clustering counties based on these factors will provide insights into variations in the pandemic's impact across different counties, helping the Texas Department of State Health Services (DSHS) allocate resources and refine intervention strategies.

2.3 Feature Selection

The selected variables for clustering shown in Table 1 were chosen based on their ability to capture key socioeconomic, demographic, and behavioral patterns that influence COVID-19 outcomes across Texas counties. As the goal of this analysis is to identify groups of counties that experienced similar pandemic trends and assess the role of socioeconomic factors in shaping these trends, each variable was carefully selected for its relevance in this context.

Table 1: Selected Features

Features	Type	Scale of Measurement	Relevance
county	factor	nominal	Identifies geographic unit for clustering
cases_per_100k	double	ratio	Confirmed COVID-19 cases per 100,000 residents (normalized for population)
deaths_per_100k	double	ratio	COVID-19 deaths per 100,000 residents
poverty	double	ratio	Percentage of people living below the poverty line
median_income	double	ratio	Median household income (economic health)
pct_on_food_stamps	double	ratio	Percentage of households receiving public assistance
commuters_by_public_transportation	double	ratio	Proxy for crowding & transmission risk
pct_work_from_home	double	ratio	Indicates ability to socially distance
income_per_capita	double	ratio	Reflects economic status on an individual level
gini_index	double	ratio	Income inequality measure
total_pop	double	ratio	Total population of the county

The inclusion of **COVID-19 cases and deaths per 100,000 residents** (`cases_per_100k`, `deaths_per_100k`) ensures that clustering reflects variations in infection severity and mortality, allowing for the identification of counties with similar pandemic trajectories. Since the spread and impact of COVID-19 were not uniform across all regions, incorporating these metrics helps differentiate counties based on their relative burden of disease. Economic indicators such as **poverty rate (`poverty`)**, **median income (`median_income`)**, and **income per capita (`income_per_capita`)** provide insight into financial stability, which has been linked to

healthcare access, susceptibility to severe illness, and overall resilience during the pandemic. **The percentage of households receiving food stamps (`pct_on_food_stamps`)** further strengthens this economic dimension by highlighting communities with higher dependence on public assistance, which may correlate with limited healthcare access and increased exposure risks. Mobility and behavioral variables were also integrated to assess differences in social distancing capacity and exposure risks. **The percentage of workers commuting via public transportation (`commuters_by_public_transportation`)** serves as a proxy for population density and potential transmission hotspots, while **the percentage of the workforce working from home (`pct_work_from_home`)** reflects an area's ability to reduce contact and minimize spread. The **Gini index (`gini_index`)**, which measures income inequality, was included to examine disparities that could influence pandemic outcomes, such as unequal access to healthcare, testing, and vaccinations. Finally, **total population (`total_pop`)** was retained to account for the scale of each county and its potential impact on healthcare demand and resource allocation.

By clustering counties based on these features, this analysis aims to uncover distinct regional patterns in COVID-19 outcomes, identify common socioeconomic risk factors among similarly affected counties, and provide insights that can inform targeted public health strategies. The selection of these variables ensures that the clustering results are meaningful, actionable, and aligned with the study's objective of understanding the socioeconomic dimensions of the pandemic's impact.

2.4 Statistical Summary of Selected Features

Table 2 presents a summary of the statistical properties for key numerical features within the dataset that will be utilized in the clustering process. The table highlights critical statistical measures such as the range, mode, mean, median, variance, and standard deviation, which reflect the distribution and variability of the data. These measures provide a foundational understanding of the original scale and spread of the features.

Table 2: Statistical Summary of Selected Features

Variable	Range	Mode	Mean	Median	Variance	SD
cases_per_100k	10,848.28	2,310.83	7,444.58	7,047.05	5,531,349.31	2,351.88
deaths_per_100k	409.28	354.61	186.68	176.69	7,212.95	84.93
poverty	16,420	4,344	3,514.25	2,211	12,258,036.85	3,501.15
median_income	49,568	24,800	48,382.29	47,264.50	77,028,900.31	8,776.61
pct_on_food_stamps	24.15	2.16	12.88	12.82	22.41	4.73
commuters_by_public_transportation	126	0	18.40	4	769.20	27.73
pct_work_from_home	7.32	0	3.16	3.12	2.22	1.49
income_per_capita	19,516	21,938	24,071.81	23,814	13,942,511.73	3,733.97
gini_index	0.14	0.44	0.45	0.45	0.00	0.03
total_pop	93,346	289	22,596.21	16,788.50	453,712,668.34	21,300.53

Key Observations

- **COVID-19 Impact:**
  The variable cases_per_100k exhibits a broad range, reaching up to approximately 10,800 cases per 100,000 residents. With a mean of ~7,445 and a median of ~7,047, the data suggest a right-skewed distribution, driven by counties with exceptionally high case counts. Similarly, deaths_per_100k shows considerable variability, with a maximum value of ~409 and a standard deviation of ~84.93, indicating a significant disparity in COVID-19 mortality rates across counties.

- **Socioeconomic Conditions:**
  The poverty variable demonstrates notable variation, with a mode of 4,344 individuals per county living below the poverty line and a mean of 3,514. The large variance (~12.2 million) suggests that while many counties share similar poverty counts, a few extreme outliers have disproportionately high numbers of people experiencing poverty. Median_income, with a range of approximately $49,568, also exhibits high variability, reinforcing the economic diversity across Texas counties. Meanwhile, pct_on_food_stamps shows a more consistent distribution, with a standard deviation of 4.73 percentage points, indicating that food assistance usage is relatively stable across counties.

- **Work and Transportation:**
  The dataset reveals that reliance on public transportation is minimal, as the variable commuters_by_public_transportation has a mean of approximately 18.4 commuters per

county, but a mode of 0, indicating that many counties report negligible or no public transit usage. The standard deviation of 27.73 suggests that while some counties have a small number of public transit users, a few counties have notably higher usage. Similarly, pct_work_from_home remains low across Texas counties, with a mean of 3.16%, reflecting the state's pre-pandemic occupational trends and predominantly rural landscape.

- **Income and Inequality:**
  Measures of income distribution and inequality highlight further disparities. income_per_capita displays a relatively tight interquartile range but exhibits variance due to a small subset of high-income counties. The gini_index, ranging from 0.44 to 0.45, indicates stable but persistent income inequality across the state.

- **Population Size:**
  County population sizes vary drastically, from fewer than 300 residents to over 93,000. This immense disparity underscores the necessity of population-adjusted metrics, such as per-100k rates, to ensure fair comparisons. Additionally, the variability in county populations explains potential data limitations, as smaller counties may lack sufficient data for certain calculations, such as death rates per 100,000 residents.

Interpretation
The statistical summary underscores the substantial heterogeneity among Texas counties in terms of health outcomes, economic conditions, employment patterns, and infrastructure. This variation justifies the application of clustering techniques to identify groups of counties with similar profiles, thereby facilitating more meaningful comparisons and targeted analyses. Moreover, understanding the distribution and skewness of variables like deaths_per_100k, poverty, and median_income will inform preprocessing choices, such as normalization or transformation, to optimize model performance.

Table 3 provides a summary of the same selected features, but this time the data has been **standardized** to prepare it for clustering. Standardization is a critical preprocessing step in clustering, as it ensures that all features contribute equally to the clustering process, regardless of their original scale or units. Many features, such as population size or number of cases, can have vastly different ranges and magnitudes compared to other features, such as median income or the percentage of people working from home. Without standardization, clustering algorithms might place undue importance on features with larger numerical ranges, potentially skewing the results.

By standardizing the data, each feature is transformed to have a mean of zero and a standard deviation of one. This process involves subtracting the mean from each value and dividing by the standard deviation. As a result, all features are now measured on the same scale, ensuring that no single feature dominates the analysis. The standardized values allow the clustering algorithm to better capture the relative relationships between the variables, ensuring that the clustering reflects true patterns and similarities within the data, rather than being influenced by the scale of individual features.

In Table 3, we can observe that each standardized feature now has a uniform standard deviation of 1, and their values have been adjusted accordingly to make them comparable. This standardization step is especially important when using distance-based clustering techniques like k-means, where the distance between data points (and therefore the clustering outcome) is

highly sensitive to the scale of the features. By standardizing the data, the clustering process becomes more robust and reliable, producing more meaningful groupings of the data.

Table 3: Normalized Statistical Summary of Selected Features

Variable	Min	1st Quantile	Median	Mean	3rd Quantile	Max
cases_per_100k	-2.183	-0.7423	-0.169	0	0.678	2.43
deaths_per_100k	-2.2	-0.7212	-0.118	0	0.663	2.621
poverty	-1.001	-0.712	-0.372	0	0.426	3.689
median_income	-2.687	-0.69	-0.127	0	0.504	2.961
pct_on_food_stamps	-2.2637	-0.695	-0.013	0	0.623	2.839
commuters_by_public_transportation	-0.663	-0.663	-0.519	0	0.301	3.88
pct_work_from_home	-2.125	-0.665	-0.028	0	0.659	2.794
income_per_capita	-2.866	-0.716	-0.069	0	0.650	2.360
gini_index	-2.439	-0.539	-0.1	0	0.603	2.436
total_pop	-1.047	-0.769	-0.273	0	0.46	3.335

2.5 Measures of Similarity & Distance

All selected features are continuous numerical variables measured on the ratio scale, which supports mathematical operations like standardization and distance calculation. To ensure equal contribution from each feature in clustering, we apply **Z-score standardization** (using `scale()`), which transforms each feature to have a mean of 0 and a standard deviation of 1. This standardization step ensures that features with different scales (e.g., population size vs. income) do not disproportionately affect the clustering outcome.

The appropriate distance measure for **K-means clustering** and **Principal Component Analysis (PCA)** is **Euclidean distance**, as both methods rely on minimizing squared distances. For **hierarchical clustering**, either **Euclidean distance** or **Manhattan distance** can be used. Euclidean distance is suitable for standardized data, ensuring that each feature contributes equally, while Manhattan distance is useful when dealing with more independent features or when the data contains outliers, as it is less sensitive to extreme values.

3. Data Exploration

After selecting the appropriate variables for our analysis, we proceeded with verifying the data quality. Data processing is crucial, as it ensures the removal of duplicated data, addresses outliers, and prevents the exclusion of important records, all of which can reduce the quality of data mining efforts. We approached the data quality inspection process in the same manner as variable selection, treating each dataset individually.

3.1 COVID-19 Cases Census Dataset Exploration

3.1.1. Verifying Data Quality

For the COVID-19 Cases Census dataset, we first assigned appropriate data types to each variable based on their nature and intended use. This step ensures more efficient analysis and guarantees that the data is correctly structured for modeling and computation.

Appropriate Data Types:

1. **Categorical Variables:**
   The following categorical variable was converted from string-based formats to categorical data types for computational efficiency, particularly for grouping, filtering, and statistical analysis:
   - `county` (factor) – Identifies Texas counties to facilitate county-level comparisons.

2. **Numerical Variables:**
   The following numerical variables were retained as float or integer types to ensure proper mathematical operations and statistical analysis:
   - `cases_per_100k` (double) – Represents the number of confirmed COVID-19 cases per 100,000 residents in each county, normalized for population size.
   - `deaths_per_100k` (double) – Represents the number of COVID-19-related deaths per 100,000 residents in each county.
   - `total_pop` (double) – Total population of each county, essential for per capita calculations.
   - `median_income` (double) – Median household income per county to evaluate the economic impact of COVID-19.
   - `pct_on_food_stamps` (double) – Percentage of households receiving public assistance or food stamps, indicating economic vulnerability.
   - `commuters_by_public_transportation` (double) – Measures the number of individuals commuting via public transportation, relevant for evaluating transmission risks.
   - `pct_work_from_home` (double) – Percentage of the labor force working from home, reflecting social distancing capacity.
   - `poverty` (double) – Percentage of the population living below the poverty line, highlighting socioeconomic disparities.
   - `income_per_capita` (double) – Average income per person, capturing wealth distribution.
   - `gini_index` (double) – Measures income inequality, which may influence health disparities.

3. **Justification for Keeping Data Types**

- Categorical variables were converted to factors to enhance computational efficiency, making filtering and grouping operations faster.
- Numerical variables were kept as doubles to ensure high precision for calculations and statistical analysis.

By assigning appropriate data types, the dataset is optimized for efficient analysis, particularly in examining the relationships between COVID-19 case trends and socioeconomic factors across Texas counties.

Handling Missing Values:

To ensure data completeness, we examined the dataset for missing values, as gaps in the data could lead to biased insights and misinterpretations of COVID-19 trends. Missing values can result from:

- Data collection inconsistencies
- Reporting limitations in certain counties
- Variations in data aggregation methodologies

After running a missing value check, **no missing data** was detected across any of the key variables. This indicates that each row contains a complete set of attributes, ensuring robust and reliable analysis.

Outcome:

- **No data was removed in this step.**
- **The dataset remains at 254 rows and 11 columns.**

This confirms that the dataset is fully complete, eliminating concerns about missing values affecting insights. With all variables intact, our analysis accurately represents COVID-19 case trends and socioeconomic factors across Texas counties.

Duplicate Check:

To ensure data quality, we examined the dataset for **duplicate rows**, as they could distort mobility trend analysis by **overrepresenting specific data points.** Duplicate rows can arise due to:

- **Data entry errors**
- **Redundant reporting**
- **Inconsistencies in data aggregation processes**

After running a duplicate check, **no duplicate rows were detected** in the dataset. This indicates that **each row represents a unique combination of date, location, and mobility metrics.**

Outcome:

- **No data was removed** in this step.
- The dataset remains at 254 rows.
- This confirms that the dataset is **free from redundancy**, ensuring that our analysis reflects **distinct and meaningful mobility trends** across Texas counties.

Outlier Detection and Removal:

To further refine the dataset and ensure that extreme values do not skew the analysis, we conducted an outlier detection process using the **Interquartile Range (IQR) method.** Outliers were identified for each key variable based in the values that fell:

- **Below Q1 - 1.5 x IQR**

- **Above Q3 + 1.5 x IQR**

where **Q1** and **Q3** represent the first and third quartiles, respectively.

Outliers Identified:
The analysis revealed the following number of outliers across key socioeconomic and COVID-19 related variables:

- **cases_per_100k**: 3
- **deaths_per_100k:** 5
- **poverty:** 31
- **median_income:** 10
- **pct_on_food_stamps:** 11
- **commuters_by_public_transportation:** 39
- **pct_work_from_home:** 15
- **income_per_capita:** 9
- **gini_index:** 9
- **total_pop:** 39

These extreme values likely stem from:

- Significant disparities in county population sizes
- Economic inequalities influencing public assistance and remote work trends
- Variability in COVID-19 case reporting between counties

Outlier Removal:
To maintain the integrity of the dataset, we removed these extreme values. As a result:

- **80 rows were removed due to outliers**
- **The dataset is now free from extreme anomalies**, ensuring a more reliable representation COVID-19 trends across Texas counties.

By removing these outliers, the dataset maintains **data consistency** and prevents extreme values from distorting insights in subsequent analyses.

Final Dataset Summary:

After processing the dataset by removing missing values and outliers, the final dataset was reduced from **254 rows to 174 rows**. This reduction resulted from filtering out:

- **Extreme anomalies** identified as outliers through the IQR method
- **Potential data inconsistencies** that could impact trend analysis

Why This Matters
By refining the dataset, we ensure that our analysis is based on cleaner and more reliable data, which:

- **Minimizes potential biases** caused by extreme values
- **Enhances accuracy** in detecting COVID-19 case trends and socioeconomic correlations across Texas counties
- **Improves overall data integrity,** leading to more meaningful and actionable insights

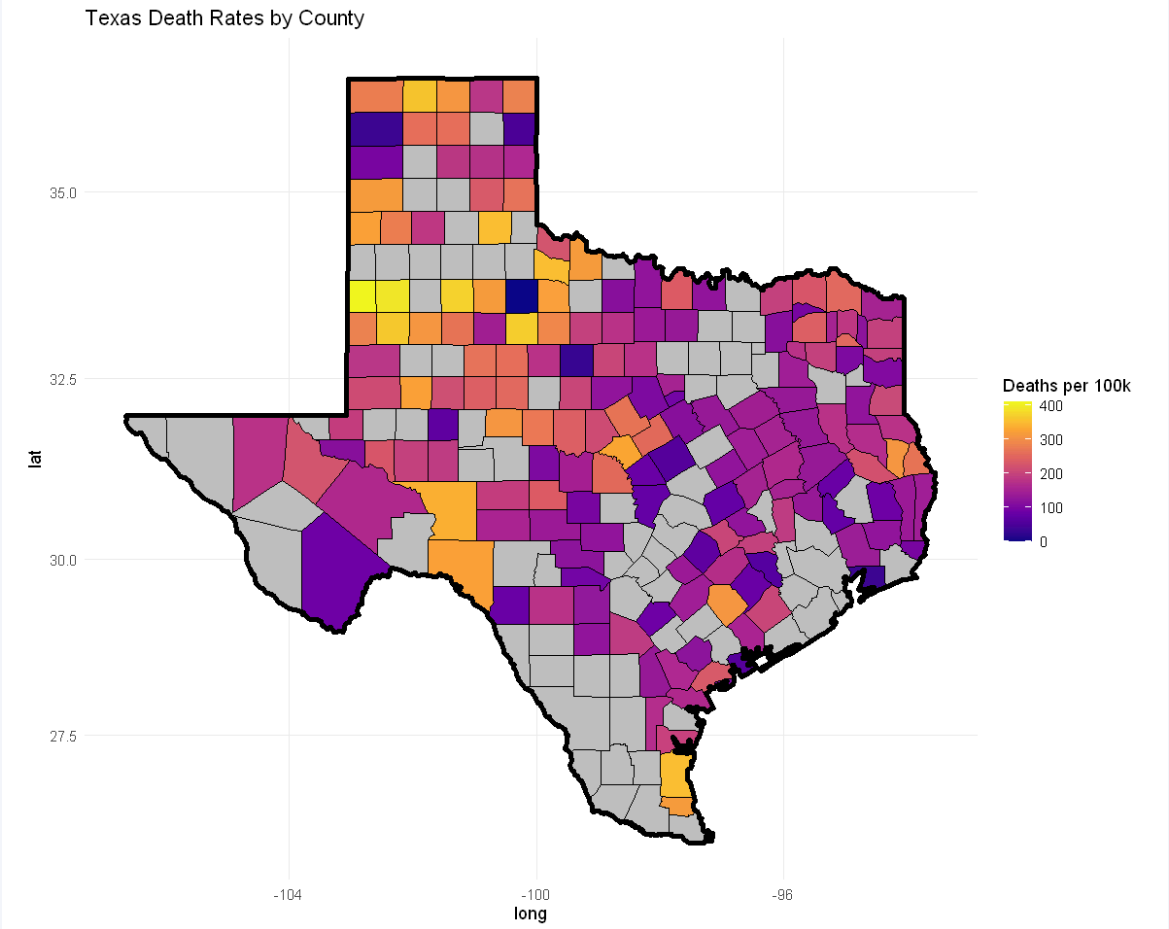
Retaining only **high-quality data points** allows us to **better understand** COVID-19 patterns **without distortions from errors or anomalies.**

3.1.2 Analysis & Visualization of Selected Features

Figure 1 presents a choropleth map illustrating the variation in COVID-19 death rates across Texas counties, measured as deaths per 100,000 residents. The map employs a plasma color scale to encode death rates, where darker shades (ranging from purple to yellow) signify higher death rates, and lighter shades represent lower rates. Counties shaded in gray indicate areas

where populations were too small to compute reliable per-capita death rates under the 100,000-resident threshold. This visualization offers a clear depiction of geographic disparities in pandemic impact across the state.

Figure 1:Texas Death Rates per 100k by County



Notable patterns emerge from the map, highlighting the uneven distribution of COVID-19 mortality. Higher death rates are concentrated in eastern and southeastern counties, including portions of the Houston metropolitan area and adjacent rural regions. This clustering may reflect multiple factors, such as higher population density, greater vulnerability among older or economically disadvantaged populations, and variations in public health interventions. In contrast, Central Texas, encompassing the Austin metro area and Hill Country, displays a mix of medium-to-high death rates. Meanwhile, many counties in western and southern Texas, particularly along the U.S.–Mexico border, are shaded gray due to insufficient population sizes for reliable rate calculations. However, some sparsely populated West Texas counties still exhibit elevated death rates, potentially indicating regional healthcare access challenges or localized outbreak patterns.

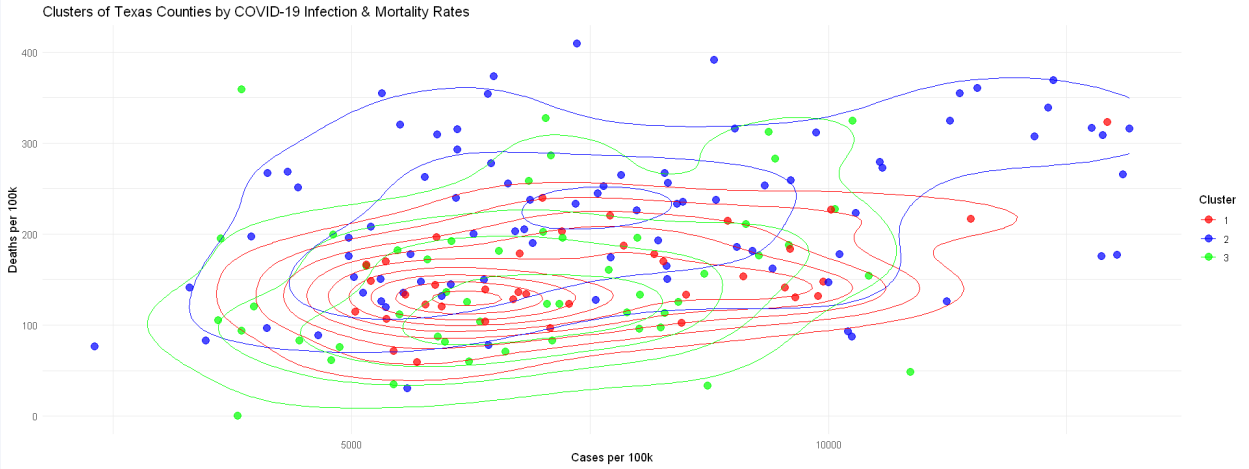
While this map effectively visualizes spatial disparities in COVID-19 mortality, it also underscores the limitations of per-capita rate calculations in low-population areas. The exclusion of gray-shaded counties from rate computations was done to ensure statistical validity rather than due to missing data. This omission should be carefully considered in the clustering process, as it may influence spatial balance and the overall model structure.

To address the broader research question regarding which Texas counties exhibit the highest COVID-19 infection and mortality rates per capita and how demographic factors contribute to these trends, subsequent analyses will integrate this spatial perspective with socioeconomic and demographic data. By clustering counties based on variables such as poverty rates, median income, health insurance coverage, racial and ethnic composition, and population density, we aim to uncover underlying patterns that drive disparities in COVID-19 outcomes. This approach will provide deeper insight into how different county profiles correspond to infection and mortality trends, informing potential public health interventions tailored to at-risk populations.

Identifying Counties with the Highest COVID-19 Infection and Mortality Rates

The scatter plot with density contours in Figure 2 provides key insights for the Texas Department of State Health Services (DSHS) in addressing the question of which Texas counties exhibit the highest COVID-19 infection and mortality rates per capita, and how demographic factors such as population size and income levels correlate with these trends. The plot categorizes counties into three distinct clusters—red, blue, and green—each reflecting different patterns in COVID-19 cases and deaths per 100,000 people as shown in Table 4.

Figure 2: Counties with Highest COVID-19 Infection & Mortality Rates



The red cluster represents counties with lower-to-moderate case rates and death rates. The density of points in this cluster suggests that these counties are more common in the dataset, indicating moderate exposure to COVID-19 but relatively balanced outcomes in terms of case and mortality rates. Demographic factors such as income levels, population density, and healthcare infrastructure may contribute to these counties experiencing less severe COVID-19 impacts. For DSHS, this cluster highlights regions that, while less affected, still warrant monitoring and preventive measures to avoid future surges. It may also reflect areas with better access to healthcare or younger populations that helped keep mortality rates low despite moderate infection levels.

The blue cluster, on the other hand, contains counties with relatively high death rates, even though their case rates can vary. Some of these counties exhibit extreme mortality values, pointing to higher death rates in the face of fluctuating infection levels. This suggests that these counties may have higher-risk populations, such as older individuals, or face healthcare system challenges that prevent them from managing severe cases effectively. The blue cluster may also correlate with areas experiencing higher poverty rates or systemic health disparities, which exacerbate the severity of COVID-19 outcomes. For DSHS, this cluster is crucial for identifying

counties with high mortality risks, enabling the prioritization of healthcare resources, testing, and vaccination efforts in regions with a greater need for support.

The green cluster displays a wider spread of data points, with some counties having high case rates but relatively lower death rates. This indicates that these counties may have managed to contain the mortality impact of COVID-19 despite experiencing a high number of infections. Factors such as better healthcare infrastructure, younger populations, or effective public health measures may have contributed to these outcomes. Additionally, these counties may have demographic characteristics like higher income levels or access to quality healthcare, which helped mitigate the severity of COVID-19's impact. For DSHS, identifying counties in the green cluster is important for understanding where continued public health efforts or healthcare system reinforcement may be most effective in reducing COVID-19-related deaths.

In summary, these three clusters—blue, red, and green—help DSHS identify regions in Texas with varying levels of vulnerability to COVID-19. The blue cluster highlights counties at high risk of mortality, possibly due to older populations, poverty, or strained healthcare systems, signaling a need for urgent intervention. The red cluster reflects regions with more moderate COVID-19 impacts, where continued monitoring and preventive measures remain necessary. Meanwhile, the green cluster identifies counties that, despite high case rates, have managed to keep death rates relatively low, possibly due to strong healthcare systems or younger populations. This clustering analysis allows DSHS to better understand how demographic factors—such as income levels, population size, and healthcare access—correlate with COVID-19's severity, ultimately helping to guide targeted resource allocation and interventions for the regions most at risk.

Commuting Behaviors & COVID-10 Case Rates in Urban vs. Rural Counties

When examining Figure 3 and Figure 4, respectively, for public transportation usage and remote work rates across clusters, we see clear differences in commuting behaviors among Texas counties. In **Cluster 1 (red),** counties had the highest percentage of residents using public transportation, while the percentage of people working from home remained at an average level. This suggests that counties in this cluster may have more urban characteristics, where public transit is more accessible and widely used, but remote work adoption was moderate. In **Cluster 2 (blue),** public transportation usage was lower compared to Cluster 1, and the percentage of people working from home was also slightly lower. This could indicate a mix of suburban or semi-rural counties where transit options are less common, and in-person work remained more prevalent. Lastly, **Cluster 3 (green)** had the highest percentage of residents working remotely but the lowest use of public transportation. This pattern suggests that counties in this cluster may be more rural or have higher-income populations with greater flexibility to work from home, reducing their reliance on public transit.

Figure 3: Public Transportation Usage Across Clusters

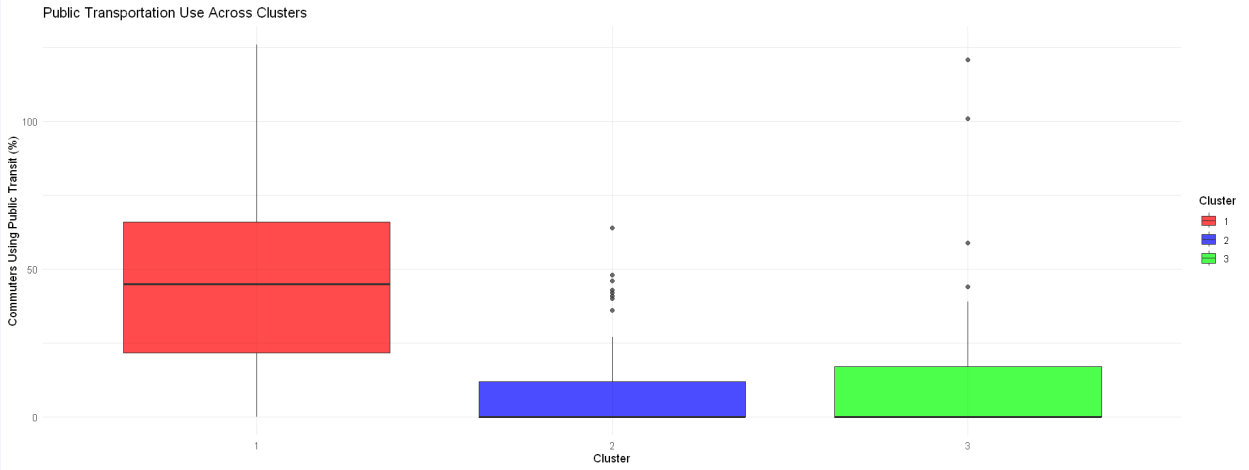
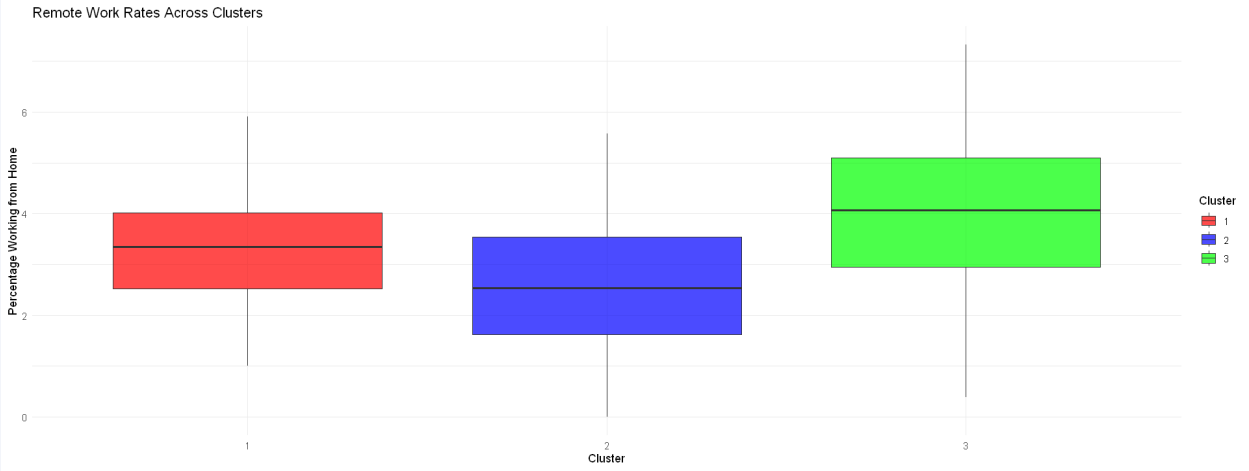


Figure 4: Remote Work Rates Across Clusters



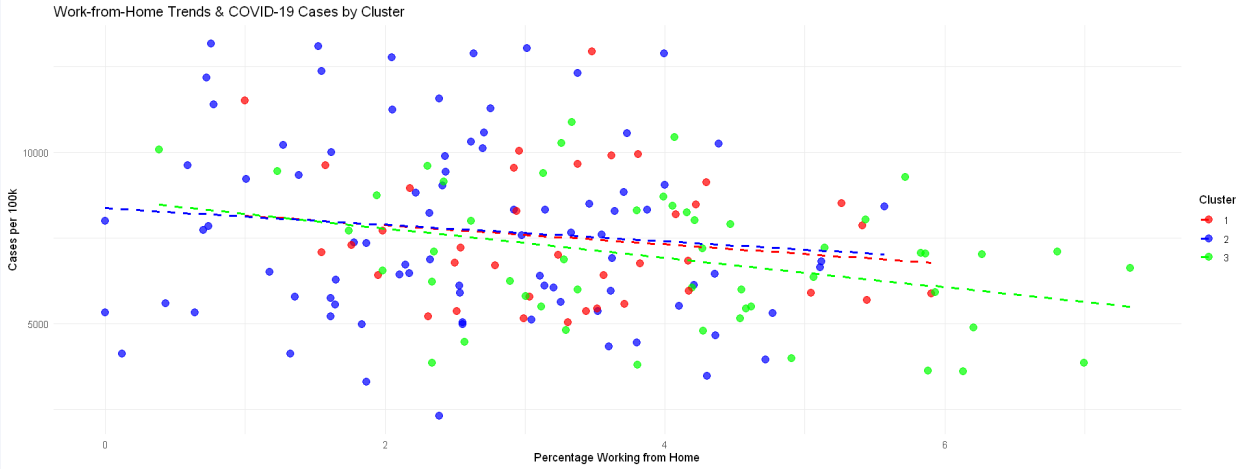
These insights help the Texas Department of State Health Services (DSHS) assess how commuting behaviors may have influenced COVID-19 exposure risks. Counties in **Cluster 1**, with higher public transportation usage, likely had greater exposure potential due to shared transit spaces, possibly contributing to increased transmission. **Cluster 2**, with lower transit use but also lower remote work rates, may represent areas where residents relied more on personal vehicles but still had limited flexibility to work from home, affecting their ability to socially distance. **Cluster 3**, with the highest remote work rates and lowest transit use, likely had the lowest exposure risks due to reduced in-person interactions. Understanding these patterns can help DSHS evaluate how mobility behaviors influenced COVID-19 trends and shape future public health policies, particularly for counties with high transit reliance and limited remote work opportunities.

Counties with the Least Ability to Work from Home & Their COVID-19 Case Rates

Exploring the relationship between work-from-home percentages and COVID-19 cases per 100k in Figure 5 reveals a clear negative trend, indicating that higher remote work adoption was generally associated with lower infection rates. This suggests that the ability to work remotely played a significant role in reducing virus transmission by limiting in-person interactions. However, the strength of this relationship varies across clusters. Cluster 3, represented in

green, shows the strongest negative slope, meaning remote work had the most pronounced effect on reducing cases in these counties. This could be due to stronger remote work adoption in industries that allowed it or other public health measures reinforcing lower exposure risks.

Figure 5: Remote Work Trends & COVID-19 Cases by Cluster



Cluster 1, represented in red, appears more tightly clustered, suggesting a more consistent but moderate link between remote work and case rates. Cluster 2, represented in blue, is more spread out, implying that while some counties saw benefits from remote work, others may have had additional factors influencing case counts, such as population density, essential worker prevalence, or healthcare access.

From these findings, several key insights emerge that can guide public health recommendations. Encouraging remote work in high-risk areas, particularly those resembling Cluster 3, could be an effective strategy for mitigating virus spread. In these counties, promoting work-from-home policies through business incentives or improved internet access could further strengthen its impact. However, the variability in Cluster 2 suggests that remote work alone was not always sufficient in reducing case rates, highlighting the need for additional interventions such as stricter public health measures, vaccination campaigns, or improved workplace safety protocols for industries where remote work is not feasible. Meanwhile, the more consistent relationship observed in Cluster 1 indicates that hybrid work models could be a sustainable approach to balancing public health and economic activity.

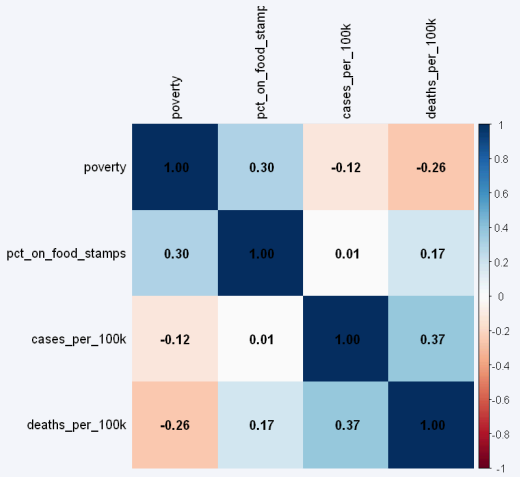
Understanding these differences allows the Texas Department of State Health Services (DSHS) to tailor public health policies and prioritize interventions based on county-specific needs. Remote work was clearly a key factor in reducing COVID-19 transmission, but its impact varied depending on other socioeconomic and structural factors. By using this information to guide policy decisions, DSHS can help improve community resilience against future outbreaks, ensuring that counties with high transmission risks receive the most effective support.

Correlation Between Poverty, Public Assistance, & COVID-19 Outcomes

Examining the correlation between poverty, public assistance, and COVID-19 outcomes in Figure 6 provides important insights into how socioeconomic factors influenced the spread and severity of the virus across Texas counties. The correlation between poverty rates and the percentage of people on food stamps is moderately positive at 0.30, which is expected, as

higher poverty levels generally lead to increased reliance on public assistance programs.

Figure 6: Correlation Map Between Poverty, Public Assistance, & COVID-19 Outcomes



However, the relationship between poverty and COVID-19 cases per 100k is slightly negative at -0.12, suggesting that counties with higher poverty rates did not necessarily experience higher infection rates. This could be due to a variety of factors, including differences in population density, employment sectors, or public health interventions. The correlation between poverty and deaths per 100k, at -0.26, indicates a stronger negative relationship, suggesting that counties with higher poverty rates may have had lower COVID-19 mortality. While this may seem counterintuitive, it could be influenced by demographic factors such as age distribution, as lower-income counties might have younger populations that are less susceptible to severe COVID-19 outcomes.

When looking at the relationship between food stamp reliance and COVID-19 cases, the correlation is almost negligible at 0.01, meaning there is no strong direct link between a county's dependence on food assistance and its infection rate. However, the correlation between deaths per 100k and food stamp usage is 0.17, indicating a weak positive relationship. This suggests that counties with higher reliance on food assistance may have experienced slightly higher mortality rates, potentially reflecting underlying health disparities such as higher rates of chronic conditions or reduced access to quality healthcare. The strongest observed correlation in this analysis is between COVID-19 cases and deaths per 100k, at 0.37, reinforcing the expected relationship that counties with higher infection rates tended to have higher mortality rates.

These findings highlight key considerations for the Texas Department of State Health Services (DSHS). While poverty alone does not appear to be a primary driver of COVID-19 cases or deaths, the slight positive correlation between food stamp reliance and mortality suggests that public assistance programs may serve as a proxy for other underlying vulnerabilities, such as healthcare access and comorbidities. Policymakers should consider targeted healthcare support and public health interventions in areas with higher reliance on food assistance, as these communities may be at greater risk for severe outcomes. Additionally, the relatively weak correlations between poverty, food assistance, and case rates suggest that other structural factors, such as employment type, housing density, and local mitigation measures, may have played a larger role in shaping infection patterns. By addressing these broader determinants, DSHS can develop more effective strategies for pandemic response and resource allocation in

future public health crises.

Socioeconomic Disparities in COVID-19 Outcomes

Analyzing the relationship between median income and COVID-19 death rates across clusters in Figures 7 and 8 respectively, reveals key patterns in how median income levels relate to mortality rates. Cluster 3, represented in green, has the highest median income and the lowest death rate, suggesting that higher-income counties may have had better access to healthcare, stronger infrastructure for public health measures, or a population with fewer underlying health conditions. Cluster 1, represented in red, has the second-highest median income and the second-lowest death rate, indicating a similar trend where relatively higher-income counties experienced lower mortality. In contrast, Cluster 2, represented in blue, stands out as the group with the lowest median income but the highest death rate, reinforcing the idea that economic disadvantage was associated with worse COVID-19 outcomes.

Figure 7: Median Income By Cluster

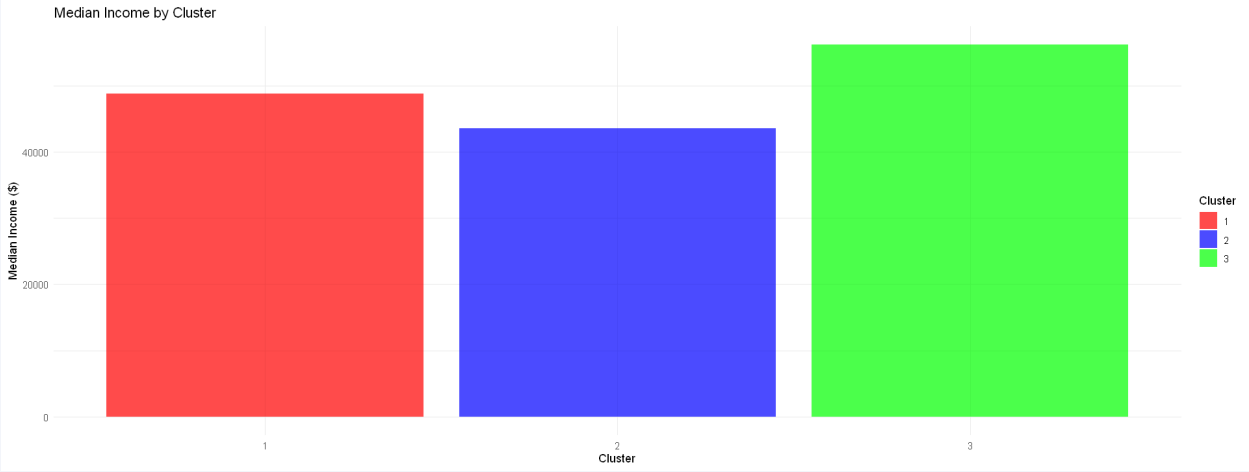
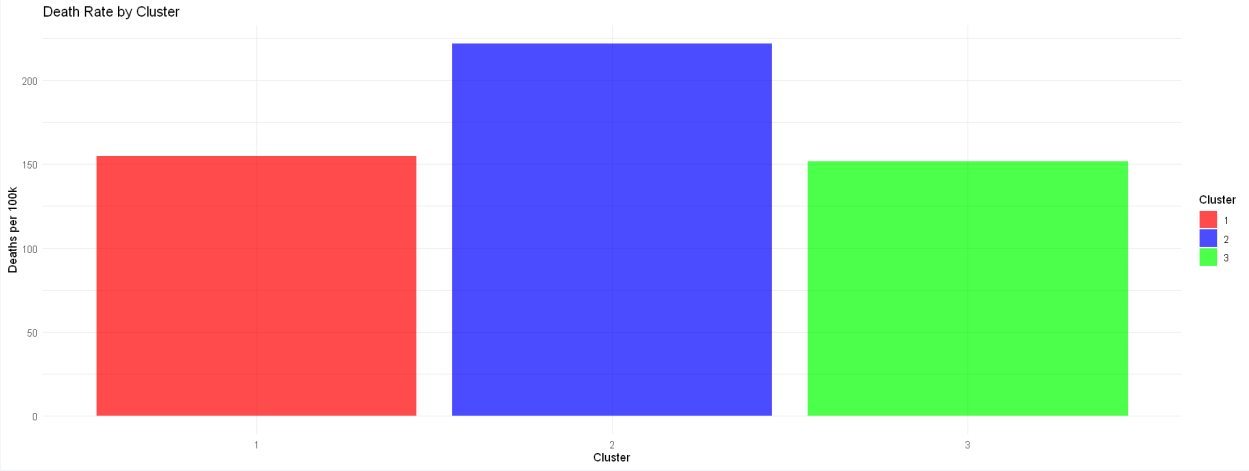


Figure 8: Death Rate By Cluster



These findings suggest that socioeconomic factors played a significant role in determining COVID-19 severity. Higher-income counties, as seen in Clusters 1 and 3, may have benefited from better healthcare access, lower population density, and a greater ability for residents to

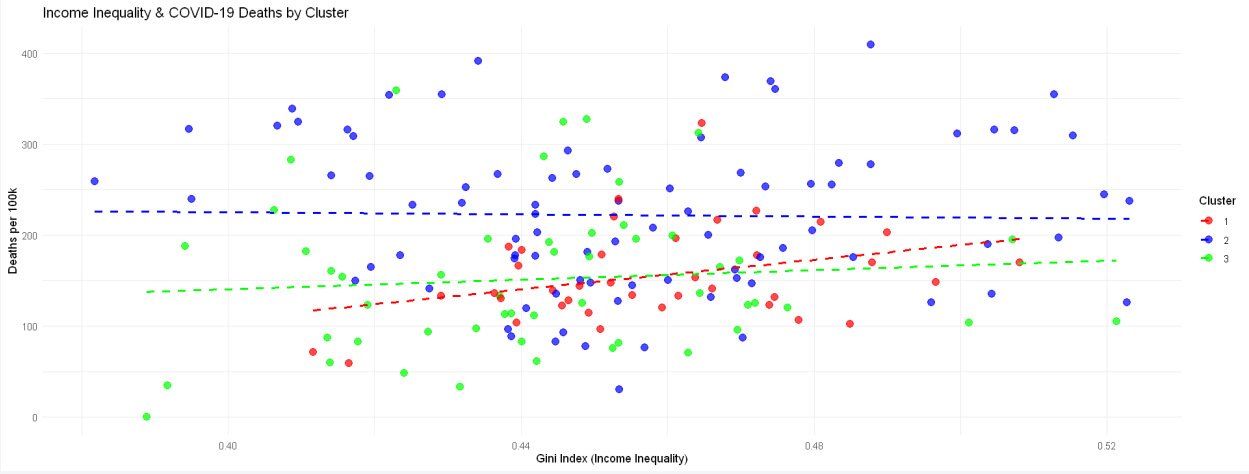
work remotely, all of which could have contributed to lower mortality rates. On the other hand, lower-income counties in Cluster 2 likely faced multiple challenges, such as limited healthcare infrastructure, higher rates of preexisting health conditions, and greater exposure risks due to employment in frontline or essential industries.

For the Texas Department of State Health Services (DSHS), these disparities highlight the need for targeted public health interventions. Ensuring that lower-income counties receive adequate healthcare resources, vaccination outreach, and support for preventative measures could help mitigate future public health crises. Additionally, policies that address broader social determinants of health, such as improving healthcare accessibility in economically disadvantaged areas and enhancing workplace safety in essential industries, could reduce the disproportionate burden of severe illness and mortality in these communities. By acknowledging these socioeconomic disparities, public health strategies can be more effectively designed to protect vulnerable populations and create more equitable health outcomes across Texas counties.

Income Inequality & COVID-19 Outcomes

Analyzing the relationship between income inequality, as measured by the Gini index, and COVID-19 mortality rates across clusters in Figure 9 reveals interesting patterns in how economic disparities may have influenced pandemic outcomes. In Cluster 1 (red), the Gini index shows a positive slope, indicating that counties with higher income inequality tend to have slightly higher death rates per 100k, but the relationship is relatively weak. The line for Cluster 1 is positioned below the line for Cluster 2 (blue), which exhibits a very gradual and almost imperceptible decline. This suggests that income inequality in Cluster 2 had an even less pronounced impact on death rates, with the overall trend being only slightly negative. On the other hand, Cluster 3 (green) shows a slightly stronger positive slope, meaning that, although the relationship is still weak, counties in this cluster with higher income inequality tend to experience higher death rates, but at a slower rate than Cluster 1.

Figure 9: Income Inequality & COVID-19 Deaths By Cluster



The clustering of dots around the respective lines for each cluster suggests that counties within each group are consistently following the same general trend, with the values of income inequality and death rates aligning accordingly. This indicates that while there is some variability within each cluster, the patterns of income inequality and death rates are relatively consistent for each group.

The findings suggest that income inequality, as captured by the Gini index, may have a mild association with COVID-19 mortality rates, but the effect is not overwhelmingly strong. In some clusters, particularly Cluster 2, the relationship appears to be minimal, indicating that factors beyond income inequality—such as healthcare access, pre-existing health conditions, or local mitigation measures—might play a more significant role in determining outcomes. However, the positive slopes in Clusters 1 and 3 imply that higher income inequality could contribute to worse health outcomes, particularly in counties already dealing with socioeconomic challenges.

For the Texas Department of State Health Services (DSHS), this analysis underscores the importance of addressing income inequality as a long-term strategy to improve health equity. While the relationship between income inequality and COVID-19 outcomes is weak, the persistent trend in some clusters suggests that further research and targeted interventions may be needed to mitigate the effects of economic disparities. Policies aimed at reducing income inequality, improving access to healthcare, and providing support to disadvantaged communities could help reduce the impact of future pandemics and create a more resilient population.

4. Modeling & Evaluation

4.1 Modeling: Cluster Analysis Overview

This section explores clustering techniques to reveal patterns in county-level COVID-19 data across Texas. By grouping counties with similar demographic and pandemic-related characteristics, we aim to provide the Texas Department of State Health Services (DSHS) with insights for more targeted public health strategies.
We start by ensuring data quality through filtering and handling missing values. Multiple clustering algorithms are applied, and results are evaluated using relevant metrics and visualizations to support interpretation.

Cluster Profile Interpretation

To understand the distinguishing characteristics of each cluster, we analyze the demographic and socioeconomic profiles based on the mean values of the standardized variables used in clustering. This interpretation helps identify key differences among groups of counties in terms of COVID-19 impact and related socioeconomic factors.

After standardizing the features, we examine summary statistics shown in Table 3 to assess the range and distribution of key variables, such as COVID-19 case and death rates, poverty levels, median income, food stamp participation, commuting patterns, income inequality, and total population. By comparing these values across clusters, we can characterize each group's unique challenges and potential public health needs.

4.2 Clustering Methods & Suitable Number of Clusters

Clustering Methods

For our report, we conducted a comprehensive cluster analysis using multiple methods and feature subsets to uncover meaningful patterns in Texas county-level COVID-19 and socioeconomic data. Specifically, as shown in Table 4, we applied **K-means clustering with K = 3 and K = 4** using all ten standardized features, allowing us to examine how different cluster sizes impacted interpretability. To complement this, we employed **hierarchical clustering with Ward's method**, which is well-suited for minimizing variance within clusters and visualizing hierarchical relationships between counties. Additionally, we performed **K-means clustering with K = 3 using only COVID-related features** (`cases_per_100k`, `deaths_per_100k`,

`total_pop`) to assess whether pandemic-specific variables alone could produce distinct and interpretable groupings.

Table 4: Cluster Methods & Feature Subsets

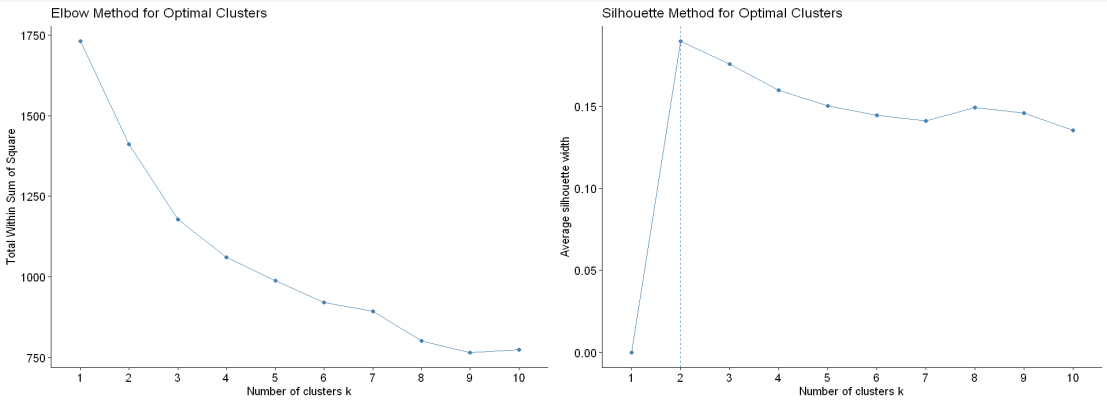
Clustering Methods	Feature Subset Used
K-means (K=3)	All 10 Features
K-means (K=4)	All 10 Features
Hierarchical (Ward's)	All 10 Features
K-means (K=3)	COVID-only subset: cases_per_100k, deaths_per_100k, total_pop

Determining a Suitable Number of Clusters for Each Method

To determine the optimal number of clusters for our K-means analysis, we combined both exploratory analysis and unsupervised evaluation methods to ensure the resulting clusters were meaningful and interpretable. We initially experimented with **K = 3** and **K = 4** to understand how different cluster sizes would impact the interpretability and the granularity of the insights. The selection of **K = 3** was primarily driven by its ability to effectively capture distinct socioeconomic and COVID-19 impact profiles across Texas counties while maintaining practical usability for stakeholders, particularly in terms of providing clear and actionable groupings. We observed that **K = 4** did not offer significant improvements in interpretability; rather, it led to the creation of clusters that were less distinct and harder to assign meaningful labels, suggesting diminishing returns beyond **K = 3**.

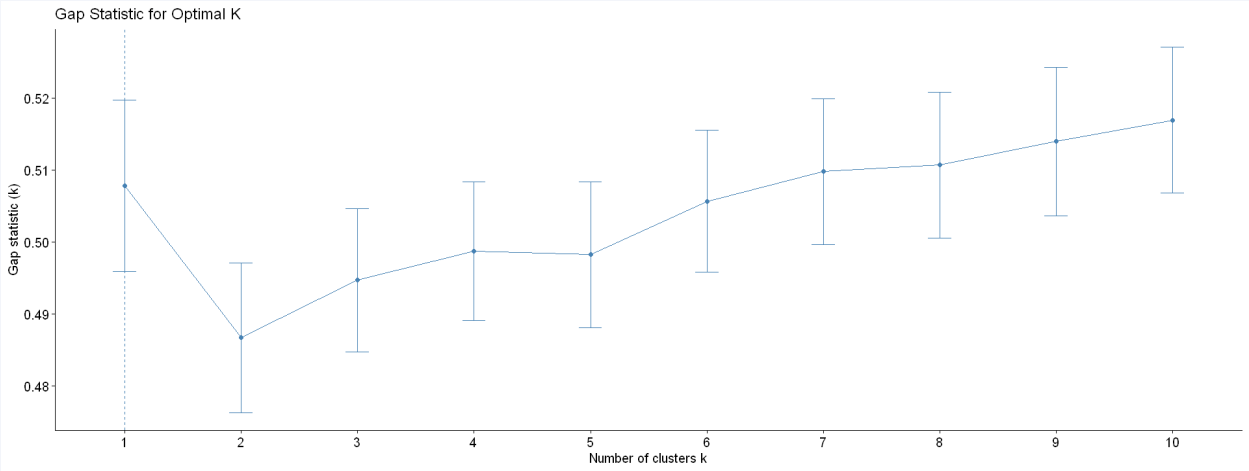
To rigorously evaluate the appropriate number of clusters, we turned to two standard unsupervised evaluation techniques shown in Figure 10: the **Elbow Method** and the **Silhouette Method**. The Elbow Method involves plotting the within-cluster sum of squares (WSS) for a range of **K** values, and we observed a clear inflection point in the plot at **K = 3**. The WSS decreased sharply up to this point, after which the rate of decline flattened significantly, indicating that increasing the number of clusters beyond **K = 3** would not yield substantial improvements in cohesion. To further validate this choice, we used the **Silhouette Method**, which measures how well each data point fits within its assigned cluster compared to other clusters. The silhouette plot revealed that **K = 3** had the highest average silhouette width, suggesting that the clusters formed at this value were well-separated, with data points being more similar to those within their own cluster than to those in other clusters. The silhouette score for **K = 4** was lower, reinforcing the idea that **K = 3** was the optimal choice.

Figure 10: Elbow & Silhouette Method for Optimal Clusters



Thus, after combining both exploratory and formal evaluation methods, we selected **K = 3** as the optimal number of clusters for the K-means clustering. This decision is supported by the clear elbow in the WSS plot and the highest silhouette score, both of which indicated that **K = 3** offers the best balance between capturing the inherent structure of the data and avoiding overfitting. Choosing **K = 3** ensures that the resulting clusters reflect meaningful patterns in the data, without unnecessarily complicating the analysis with additional, less interpretable groupings.
The Gap Statistic method was used to further validate our selection of the optimal number of clusters, K. This technique compares the within-cluster dispersion to what would be expected under a null reference distribution. The best choice for K is the value that corresponds to the largest gap or the point just before the gap begins to decline, indicating a meaningful clustering structure.

Figure 11: Gap Statistic for Optimal K



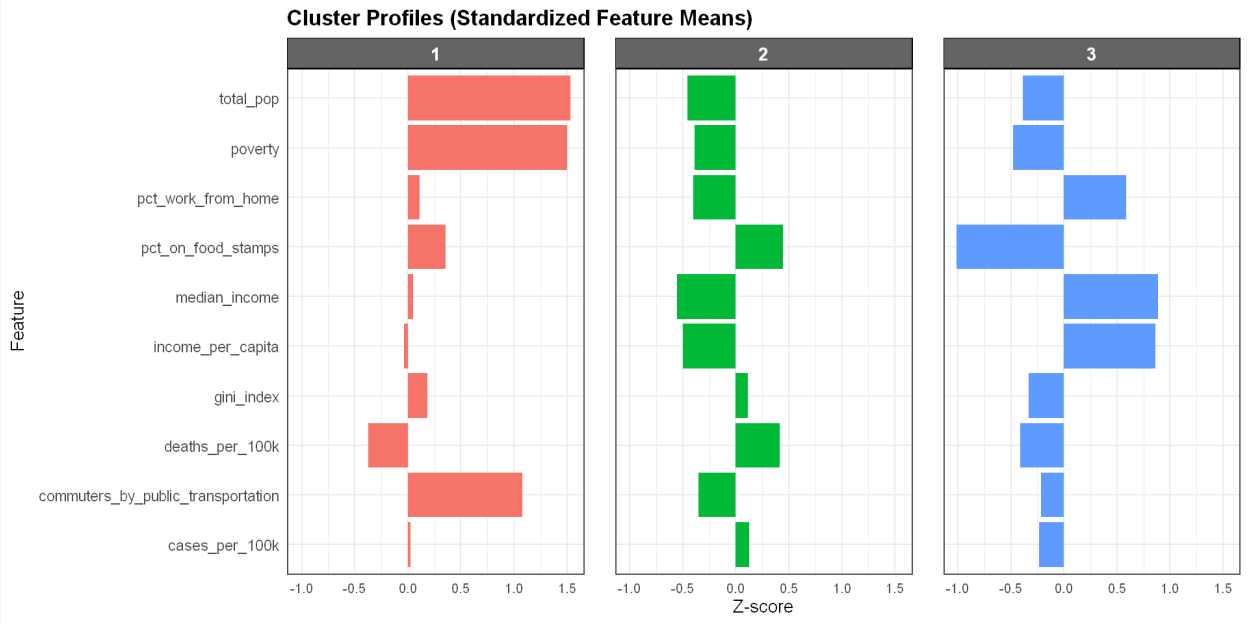
The Gap Statistic involves evaluating the within-cluster variation across different values of K while comparing these results to a null distribution. The optimal number of clusters is identified as the smallest K where the gap value remains within one standard error of the maximum. In Figure 11, the largest gap is observed at **K = 1**; however, because K = 1 does not represent clustering, the next smallest K value greater than or equal to 2 must be considered. The plot also suggests that the gap increases again between **K = 4 and K = 10**, though no distinct elbow or plateau is evident. Given this pattern, **K = 3 to 4** emerge as reasonable choices, as they maintain a relatively high gap score with moderate standard error. These results align with previous findings from the Elbow and Silhouette methods, reinforcing that **K = 3** provides a balanced choice in terms of both interpretability and model performance.

4.3 A Closer Look at Our Clusters

Interpreting Cluster Profiles

The analysis of the three clusters identified in Figure 12 by K-means reveals distinct patterns of socioeconomic and COVID-19 impact across Texas counties.

Figure 12: Cluster Profiles (Standardized Feature Means)



Cluster 1 (Red) represents densely populated, socioeconomically vulnerable urban areas. Counties in this cluster exhibit high levels of poverty, large populations, and higher public transportation usage. These counties also experience elevated COVID-19 case and death rates, indicating they are areas with greater health challenges. As such, they may be priority targets for public health interventions aimed at densely populated, at-risk populations, such as vaccine outreach programs and increased healthcare resources.

Cluster 2 (Green) is characterized by affluent suburban or semi-urban counties. These counties tend to have higher median and per capita income, greater rates of people working from home, and lower reliance on public assistance. The COVID-19 impact in these counties is moderate to low, suggesting that these areas are likely better equipped for social distancing and have better access to healthcare and resources. This cluster might benefit from policies that further encourage remote work and bolster healthcare infrastructure, though the immediate health risks are lower than in Cluster 1.

Lastly, **Cluster 3 (Blue)** captures rural or mixed-profile counties. These counties display moderate poverty levels, smaller populations, and lower public transportation usage. Although these counties have higher-than-average values for working from home, COVID-19 metrics are more varied, indicating mixed risk factors. These counties might face challenges unique to rural areas, such as healthcare access or less concentrated public health infrastructure, and may require more flexible or region-specific interventions.

Overall, these cluster profiles provide valuable insights for public health officials, particularly those at the **Texas Department of State Health Services (DSHS)**, to design targeted interventions. For instance, urban areas (Cluster 1) may require more direct interventions for vulnerable populations, while suburban areas (Cluster 2) might focus on maintaining their lower risk status through policies that encourage remote work and access to healthcare. Rural areas (Cluster 3) will need strategies that address the mixed risk factors specific to less densely populated regions.

Average COVID-19 Rates by Cluster

Table 5 below provides a quantitative summary of the **average confirmed cases and deaths per 100,000 residents** for each of the identified clusters. This analysis helps further validate the clustering results by comparing the direct COVID-19 impact across the three groups.

Table 5: Average COVID-19 Rates By Cluster

Cluster	avg_cases_per_100k	avg_deaths_per_100k
1	7498.701	154.923
2	7745.846	221.866
3	6902.137	151.697

Cluster 1 displays **moderate case rates**, with approximately 7498 cases per 100,000 residents, and **moderate mortality**, at about 155 deaths per 100,000. Despite having **high levels of poverty** and **higher transit usage**, which suggests a more densely populated and socioeconomically vulnerable population, this cluster exhibits moderate-to-severe COVID-19 outcomes. This aligns with the earlier interpretation that Cluster 1 represents **urban, high-risk communities** with greater exposure to the virus due to higher population density and mobility. **Cluster 2**, although characterized by **higher income levels** and **greater remote work capacity**, shows the **highest death rate** across all clusters, with approximately 222 deaths per 100,000 residents. This is in contrast to the moderate case rate (approx. 7745 cases per 100k).

This may reflect factors such as a **higher proportion of elderly residents**, increased **prevalence of underlying health conditions**, or potential **delayed healthcare access**. The finding that this cluster has the **highest death rate** despite being wealthier emphasizes the complexity of the pandemic's impact, where **socioeconomic factors** like access to healthcare and age demographics may play a larger role than income alone.

In contrast, **Cluster 3** shows the **lowest case rate** (approx. 6923 cases per 100k) and **lowest death rate** (approx. 152 deaths per 100k). This cluster's profile aligns with that of **rural or mixed-profile counties**, which are less densely populated and likely benefit from **natural social distancing** and **lower exposure**. However, the relatively lower case and death rates do not imply that these counties are completely insulated from the pandemic's effects, but rather that their rural characteristics may afford some degree of protection compared to more urban areas.

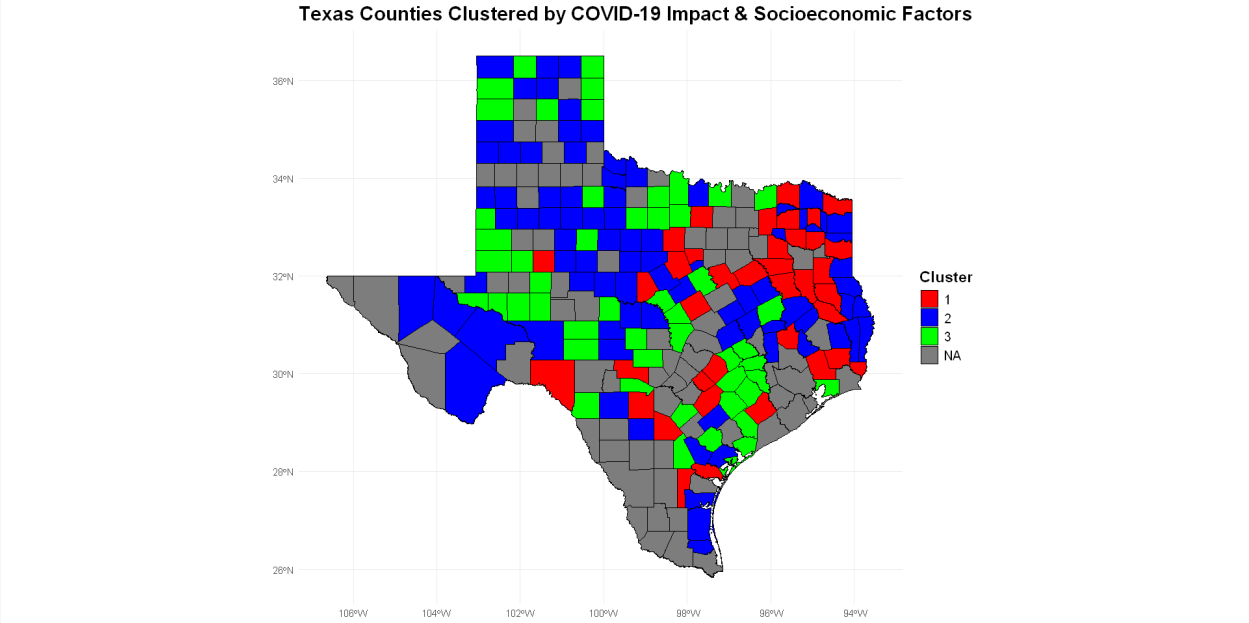
These findings underscore the strong relationship between **socioeconomic** and **mobility factors** and **COVID-19 outcomes**. They reinforce the notion that **higher income** does not always correlate with **lower COVID-19 risk**, as seen in **Cluster 2**, where wealthier counties experienced higher mortality rates. This insight is critical for public health planning, particularly for the **Texas Department of State Health Services (DSHS)**, as they design outreach programs, healthcare initiatives, and safety measures that address the multifaceted risks of the pandemic.

Interpretation of Texas Clusters

Figure 13 illustrates Texas counties grouped into **three distinct clusters**, based on COVID-19 impact and socioeconomic characteristics. The map provides a geographic visualization of these clusters, offering insights into structural disparities across the state. Each color represents

a different cluster, helping to highlight key trends and patterns that emerge when considering both health outcomes and socioeconomic factors.

Figure 13: Texas Counties Clustered By COVID-19 Impact & Socioeconomic Factors



Cluster 1 (Red) shows counties with **moderate COVID-19 impact**, with an average of 74,989 confirmed cases per 100,000 residents and 155 deaths per 100,000. Geographically, this cluster is primarily concentrated in **East Texas** and the **southern border regions**. Counties in this group tend to have **higher poverty rates**, **greater reliance on public transportation**, and generally face **socioeconomic challenges** that exacerbate their COVID-19 burden. This pattern suggests that these areas are likely **urban or semi-urban, high-risk communities**, where limited healthcare access and vulnerability due to socioeconomic factors make them more susceptible to the pandemic's severe effects. This geographic concentration highlights the disparities in public health infrastructure and access across the state.

Cluster 2 (Blue), while showing **higher income** and **more remote work capacity**, has the **highest mortality rate**, with an average of 222 deaths per 100,000 residents. These counties are spread throughout **central and north-central Texas**, and they have a **higher average case rate** of 77,456 cases per 100,000. The higher-than-expected mortality rate, despite higher income, suggests that other factors—such as a **higher proportion of elderly residents** or **underlying chronic health conditions**—may be driving the severity of the pandemic. This cluster reinforces the idea that **socioeconomic status alone does not guarantee protection from severe COVID-19 outcomes**. It serves as a reminder that **health system access, demographic composition**, and **health status** must all be considered when assessing risk and determining public health strategies.

Cluster 3 (Green), located primarily in **rural areas**, especially in the **Panhandle** and **West Texas**, exhibits the **lowest case and death rates**, with 69,023 cases and 152 deaths per 100,000 residents. These counties benefit from **lower population density** and **reduced public transit use**, which may have contributed to their **natural social distancing** and **geographic isolation**, helping to shield them from widespread virus transmission. While the lower COVID-19 burden may suggest relative protection, these counties may still face challenges unique to rural areas, such as limited healthcare infrastructure and fewer resources for public

health interventions. This pattern highlights that rural communities, while less affected by the pandemic in terms of direct health outcomes, may require targeted interventions to address healthcare access and preparedness.

Lastly, the **NA (Gray)** areas represent counties with **populations too small to compute rates** per the 100,000-resident threshold. These counties are excluded from interpretive analysis but are retained in the visualization for transparency.

This cluster map emphasizes that **COVID-19 outcomes are shaped by intersecting structural and demographic factors**, such as urban density, age structure, healthcare access, and transportation patterns, rather than solely by infection counts. The variation between clusters demonstrates the importance of **tailoring public health interventions to the local context**, recognizing that urban areas, suburban regions, and rural communities each face distinct challenges. Notably, **Cluster 2** illustrates that even wealthier counties, or those with greater capacity for remote work, are not immune to high COVID-19 mortality rates, underscoring the need to look beyond traditional socioeconomic indicators when assessing risk. This calls for **targeted interventions** that consider **health system access** and **community vulnerability**, ensuring that responses are as effective and equitable as possible across different areas of the state.

4.4 K-Means Clustering Method

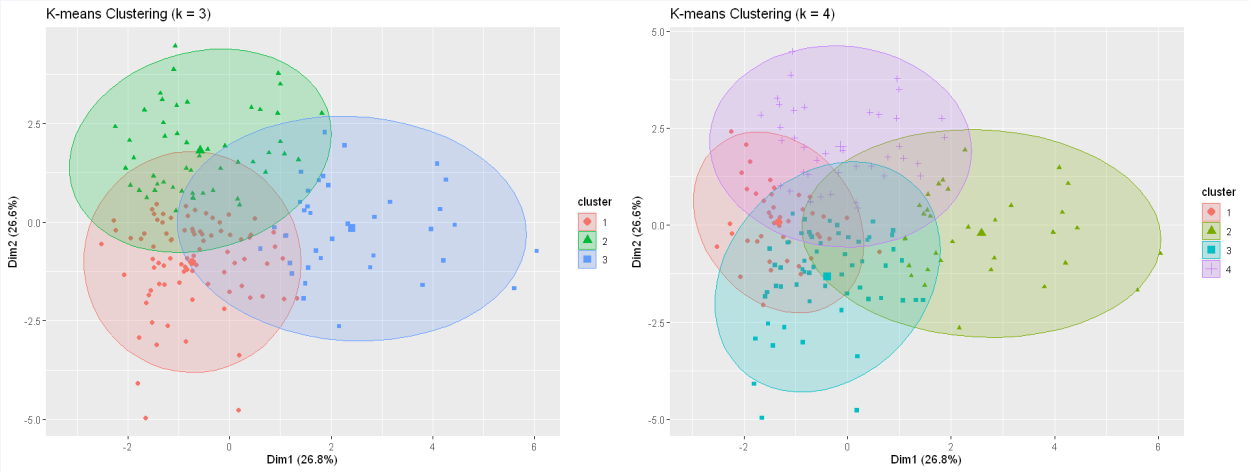
The **K-means clustering algorithm** was applied to identify patterns across Texas counties. This unsupervised machine learning technique is commonly used for discovering structure in multivariate data by assigning each data point to the nearest centroid and iteratively updating cluster centroids to minimize the **within-cluster sum of squares (WSS)**. Since K-means is sensitive to variable scale, all input features were standardized before clustering.

This method was selected because it efficiently detects **natural groupings** in complex datasets, performs well on moderate-sized county-level data, and provides **interpretable centroids** that allow for cluster profiling based on standardized feature values. To assess how the number of clusters influences structure and interpretability, both **K = 3** and **K = 4** solutions were tested. The results, visualized in the following section, help compare the cluster separation and overall data organization for different values of K.

4.4.1 Comparing K=3 vs. K=4

The comparison between K = 3 and K = 4 cluster solutions is visualized using PCA-reduced coordinates (Dim1 and Dim2) to represent high-dimensional data. These visualizations, Figure 14, provide insight into the structure and interpretability of each clustering approach. The K = 3 solution results in well-separated clusters, forming distinct regions that align with socioeconomic and COVID-19 impact trends observed in previous analyses. This clustering approach is both parsimonious and interpretable, making it particularly valuable for policy applications by categorizing counties into broad groups such as vulnerable, affluent, and rural. In contrast, the K = 4 solution introduces some overlap, potentially splitting existing groups without adding substantial new insights. While this approach increases granularity, the additional cluster reduces interpretability, as the boundaries between clusters become less distinct. Although K = 4 captures more subtle variations in the data, it does not provide enough practical advantages over K = 3 to justify the added complexity.

Figure 14: Comparison Between K-means Clusters (k=3, k=4)



The K = 3 cluster solution offers meaningful separation of Texas counties based on socioeconomic and pandemic-related factors, which is particularly relevant for the Texas Department of State Health Services (DSHS). The three clusters highlight distinct regional characteristics that can inform targeted public health strategies. The first cluster consists of counties with higher poverty rates and lower median income, representing economically vulnerable areas with limited resources. These counties may require increased healthcare access, financial assistance programs, and targeted vaccination efforts to mitigate disparities in COVID-19 outcomes. The second cluster includes counties with higher remote work rates and moderate income, likely corresponding to urban or suburban regions with stronger infrastructure for adapting to pandemic-related disruptions. These areas may benefit from policies supporting continued remote work and public health awareness campaigns. The third cluster comprises counties with higher income, lower poverty, and fewer public transit commuters, representing wealthier or more rural areas with reduced transmission risk. While these counties may have fewer immediate vulnerabilities, ensuring continued access to healthcare and preventive measures remains important.

Examining the top five most populous counties in each cluster provides additional insight into the demographic and regional composition of these groupings, as shown in Table 6. In **Cluster 1**, counties such as **Bowie, Hunt, and Angelina** exhibit moderate case rates and relatively high poverty levels, likely reflecting semi-rural or peri-urban areas with socioeconomic vulnerabilities. **Cluster 2** includes counties like **Jasper, Bee, and Kleberg**, which tend to have lower population sizes but moderate-to-high poverty and case rates, capturing lower-density, economically challenged regions. **Cluster 3**, represented by counties such as **Burnet, Cooke, and Washington**, features areas with higher case rates but lower poverty levels, suggesting a mix of more affluent suburban counties that still experienced significant pandemic impacts, possibly due to mobility or underreported vulnerabilities.

Table 6: Top 5 Counties By Population per Cluster (K-means)

Cluster	County	total_pop	cases_per_100k	poverty
1	BOWIE	93635	5207.45	15750
1	HUNT	90322	5045.28	16430
1	ANGELINA	87700	7713.80	16222
1	ORANGE	83909	7086.25	11759
1	BASTROP	80306	5445.42	10322
2	JASPER	35444	5543.96	5772
2	BEE	32729	9416.73	4926
2	KLEBERG	31540	6280.91	7003
2	CASS	30118	4977.10	5471
2	SAN JACINTO	27436	2310.83	4679
3	WILSON	47205	5910.39	4705
3	BURNET	45017	4878.16	5473
3	CHAMBERS	39283	8734.06	5031
3	COOKE	39064	7183.08	5313
3	WASHINGTON	34667	4811.49	4298

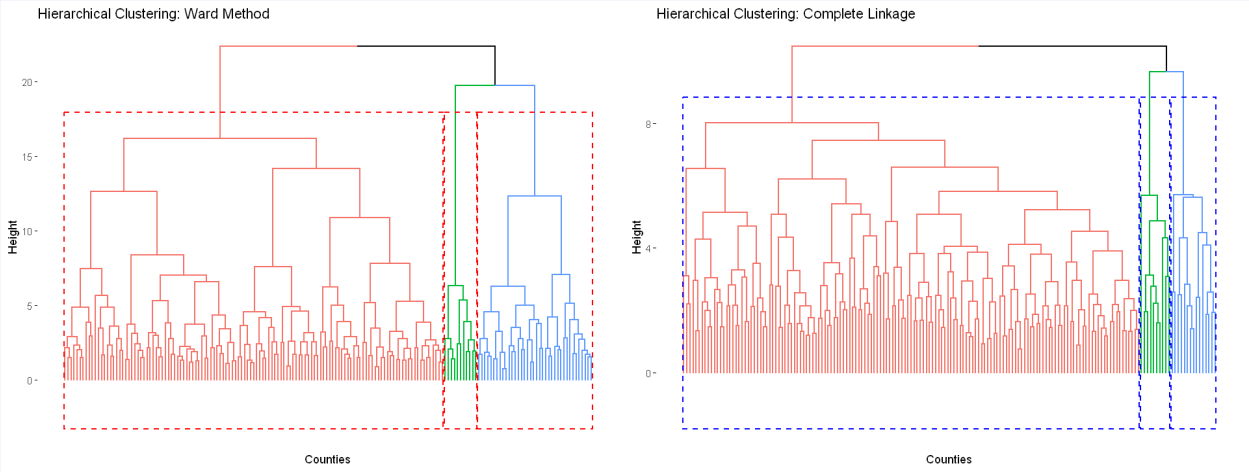
Understanding the specific counties within each cluster strengthens the practical application of this analysis. By linking clusters to actual locations, public health officials can better assess which counties fall into vulnerable versus resilient categories. This information aids policymakers in allocating resources, tailoring public health messaging, and designing interventions that align with local socioeconomic conditions. By recognizing these distinctions, DSHS can develop targeted strategies such as vaccine equity initiatives, remote work incentives, and resource allocation for high-risk populations. The K = 3 solution provides the best balance of model simplicity, clear visual separation, and practical interpretability, making it the most effective choice for guiding public health decision-making in Texas.

4.4.2 Hierarchical Clustering (Ward's Method)

Hierarchical clustering was applied as an alternative unsupervised learning approach to complement the K-means analysis and examine the nested relationships between Texas counties. This method constructs a dendrogram, a tree-like structure that iteratively merges smaller clusters into larger ones (agglomerative) or splits larger ones into smaller subgroups (divisive). Unlike K-means, hierarchical clustering does not require pre-specifying the number of clusters, allowing for more flexibility in determining natural groupings within the data. Different linkage strategies define how distances between clusters are calculated, influencing the final structure of the dendrogram.

Two agglomerative linkage methods were evaluated to assess the robustness of the clustering results shown in Figure 15. **Ward's Method**, which minimizes within-cluster variance, produced compact and spherical clusters that closely aligned with the K-means solution. This method is well-suited for identifying balanced and well-defined groups, making it particularly useful for public health analysis. **Complete Linkage**, on the other hand, merges clusters based on the maximum distance between observations, leading to more elongated clusters that capture outliers and edge cases more distinctly. While this approach allows for greater visual separation, it results in more variable cluster sizes, making interpretation more complex.

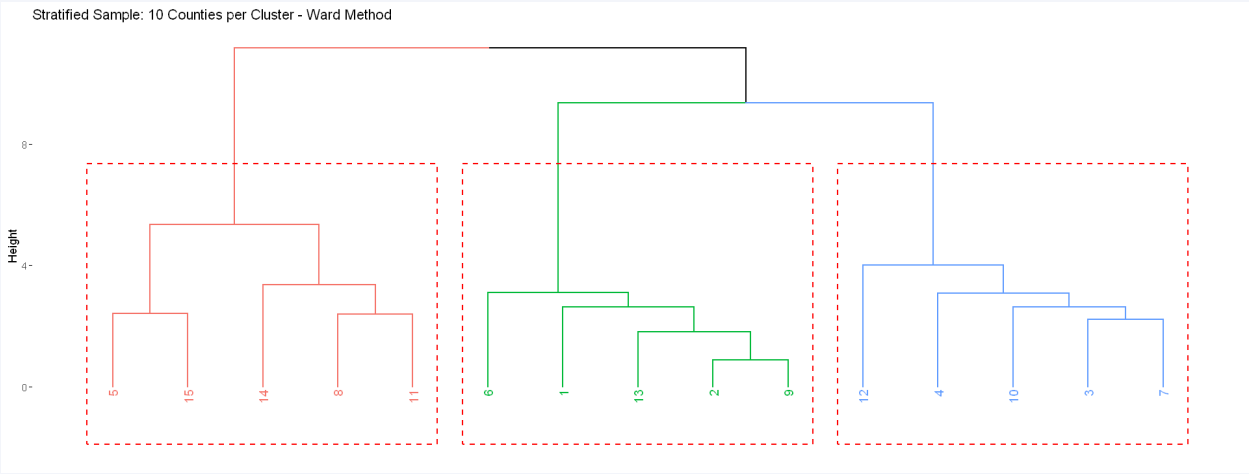
Figure 15: Hierarchical Clustering: Ward Method & Complete Linkage



The dendrograms generated from both methods highlight a **three-cluster solution**, reinforcing the results obtained from K-means. Ward's Method, in particular, exhibits strong alignment with the K-means structure, suggesting that the clustering is stable and reliable across different methodologies. This validation is crucial for ensuring that the identified clusters reflect meaningful socioeconomic and pandemic-related patterns rather than being artifacts of a single algorithm.

Additionally, a stratified sampling approach was used to ensure balanced representation across all county clusters by selecting 10 counties per cluster, as shown in Figure 16. This method avoids bias and provides a comprehensive analysis that reflects the diversity of county types within Texas. By filtering the scaled dataset to include only the counties in the stratified sample, hierarchical clustering is applied to a representative set, ensuring that the dendrogram's structure reflects real-world geographic and demographic patterns. This approach, using Ward's Method and Complete Linkage, allows for detailed insights into the county groupings and how they vary in terms of socioeconomic factors and pandemic impact.

Figure 16: Stratified Sample: 10 Counties per Cluster - Ward Method



For the Texas Department of State Health Services (DSHS), hierarchical clustering provides additional support for the three-cluster model, strengthening confidence in the findings and their practical applications. By demonstrating that similar county groupings emerge regardless of the clustering technique used, this approach reinforces the robustness of the classification system.

This consistency is essential for designing public health interventions, as it ensures that policies targeting specific county groups remain relevant and applicable. The ability to analyze outliers and subgroup variations using Complete Linkage also provides an opportunity to identify counties that may require specialized attention. Ultimately, hierarchical clustering serves as a valuable tool for validating data-driven strategies, allowing DSHS to allocate resources more effectively and implement targeted public health initiatives with greater certainty.

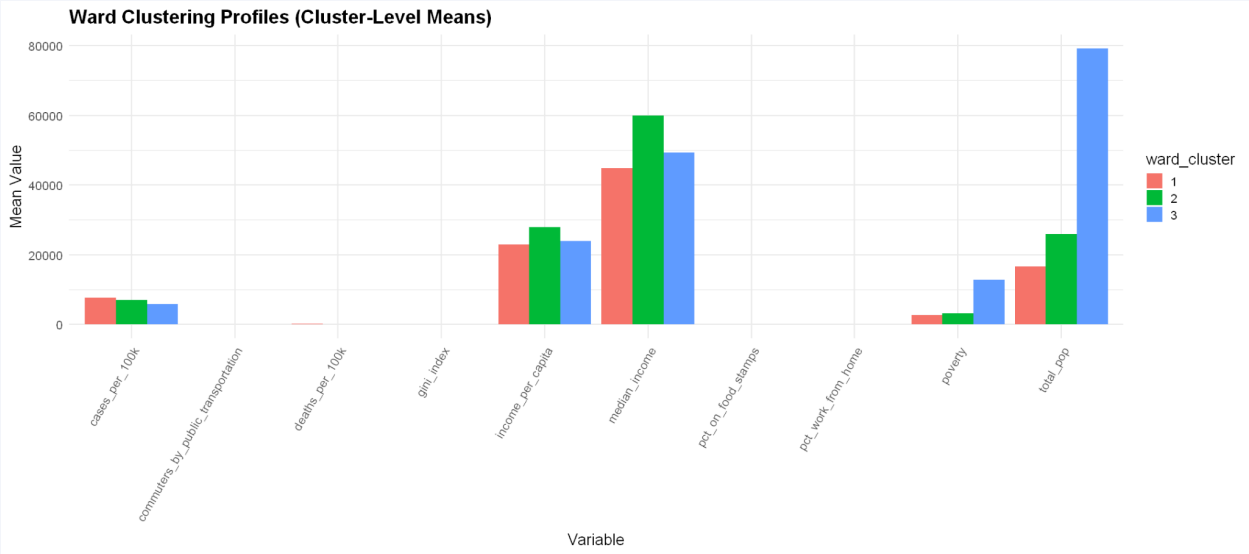
4.4.2.1 Methodology: Ward Hierarchical Clustering

Ward's hierarchical clustering algorithm was applied to segment Texas counties based on a variety of socioeconomic and COVID-related characteristics. The methodology used in this analysis involved multiple steps to ensure that the resulting clusters reflected meaningful patterns of similarity between counties. First, a set of 10 key numeric variables was selected, which encompassed a range of factors such as poverty, median income, income per capita, public transportation use, remote work capability, and the incidence of COVID-19 cases and deaths per 100,000 population. These variables were chosen to capture the socioeconomic status and pandemic impact at the county level, providing a well-rounded understanding of the counties' vulnerabilities and resilience.

To ensure fairness in how these variables were treated, all data were standardized using z-score scaling. This standardization process is crucial because it adjusts for differences in units and scales among the variables, allowing each factor to contribute equally to the clustering process. By doing so, the analysis prevents any single variable from disproportionately influencing the clustering outcome. After standardizing the data, a Euclidean distance matrix was computed to quantify the dissimilarity between counties. This matrix served as the foundation for applying the Ward method, a hierarchical agglomerative clustering technique. Ward’s method is designed to minimize within-cluster variance by merging smaller clusters into larger ones, making it particularly effective at forming compact and well-defined groups.

After applying the Ward method, the dendrogram was cut to create three clusters (k = 3), providing a clear segmentation of the counties based on their socioeconomic and health-related characteristics. The clusters were then analyzed and interpreted to provide insight into the demographic and pandemic-related patterns across the counties. The profiles of the three clusters, as revealed by the average values of the selected variables, offer a deeper understanding of the distinct characteristics of each group shown in Figure 17.

Figure 17: Ward Clustering Profiles (Cluster-Level Means)



Cluster 1 (Red) consists of counties with higher poverty rates, lower median income, and fewer remote work opportunities. These counties tend to have smaller populations and are likely more rural, with limited resources to cope with the pandemic. This cluster may represent economically vulnerable regions, where the population is more reliant on public assistance and has a lower capacity for remote work. The higher rates of COVID-19 cases and deaths observed in these counties suggest that they face greater risks from the pandemic. Targeted interventions, such as increasing healthcare access, providing financial assistance, and ensuring equitable vaccine distribution, could be particularly beneficial in these areas.

Cluster 2 (Green) represents counties with the highest levels of median income, income per capita, and remote work capacity. These areas also have lower poverty rates and lower reliance on food stamps, indicating that they are economically well-off and more resilient to the disruptions caused by the pandemic. Additionally, these counties have lower rates of COVID-19 cases and deaths, which may reflect stronger infrastructure for pandemic response and more effective public health measures. This cluster likely includes more affluent suburban or urban regions that have the resources and infrastructure to adapt to the challenges posed by the COVID-19 crisis. Public health initiatives in these areas might focus on maintaining infrastructure resilience and supporting continued remote work.

Cluster 3 (Blue) comprises the most populous counties, which tend to have higher COVID-19 case and death rates, moderate poverty rates, and greater reliance on public transportation. These counties are likely large urban centers where high population density increases exposure to the virus, contributing to higher transmission rates. The lower rates of remote work and higher use of public transportation suggest that these counties may have faced challenges in shifting to remote work during the pandemic. As the largest and most densely populated group, Cluster 3 likely faces the greatest strain on public health resources and could benefit from targeted interventions to mitigate the risks associated with population density, such as enhancing healthcare infrastructure, expanding testing and vaccination efforts, and improving access to public health information.

In addition to the overall cluster analysis, a closer examination of the five most populous counties within each of the Ward's hierarchical clusters further enhances the understanding of

each group's regional characteristics shown in Table 7. Ward Cluster 1 includes counties such as Walker, Anderson, and Cherokee, which have moderate population sizes and notably high COVID-19 case rates. For example, Walker County has a case rate of approximately 9,856 per 100,000. These counties also display relatively high poverty rates, signaling that they are economically strained regions with limited resources. The combination of higher COVID-19 exposure and socioeconomic vulnerability makes these counties more susceptible to the negative impacts of the pandemic, reinforcing the need for targeted public health initiatives that address both economic and health disparities.

Ward Cluster 2 features counties like Wise, Hood, and Hardin, which exhibit mid-range population counts and moderate COVID-19 case rates. These counties present a more balanced socioeconomic profile, which suggests they may be suburban or semi-rural areas. While these counties are less vulnerable than those in Cluster 1, they still face moderate risks that could benefit from public health interventions that focus on supporting continued resilience, such as maintaining infrastructure and ensuring equitable access to healthcare.

Finally, Ward Cluster 3 includes counties such as Bowie, Hunt, Angelina, and Orange, which are characterized by higher population sizes and lower poverty rates. These counties may represent more urbanized or economically resilient areas, as seen in Angelina County, where the population experiences a relatively higher median income despite a significant poverty rate. These areas may have better healthcare resources and infrastructure in place to manage the pandemic, but the higher population density and COVID-19 exposure still necessitate continued vigilance and strategic public health planning.

Table 7: Top 5 Most Populous Counties Per Ward Cluster

ward_cluster	county	total_pop	cases_per_100k	poverty
1	WALKER	70818	9895.79	12653
1	ANDERSON	57747	9654.18	6935
1	CHEROKEE	51594	6766.29	9090
1	LAMAR	49401	10032.19	8644
1	VAL VERDE	48976	12926.74	9849
2	WISE	63247	8512.66	8042
2	HARDIN	55993	7286.63	6669
2	HOOD	55318	9112.56	6722
2	VAN ZANDT	53607	5883.56	7489
2	RUSK	53026	5780.18	7576
3	BOWIE	93635	5207.45	15750
3	HUNT	90322	5045.28	16430
3	ANGELINA	87700	7113.80	16222
3	ORANGE	83909	7086.25	11759
3	BASTROP	80306	5445.42	10322

By analyzing the prominent counties within each cluster, this detailed breakdown provides a deeper understanding of the regional variations and needs across Texas counties. This approach benefits the Texas Department of State Health Services (DSHS) by offering a more precise foundation for resource allocation and intervention planning tailored to the specific characteristics of each county. The results from Ward's hierarchical clustering provide valuable

insights that can help the Texas Department of State Health Services (DSHS) make more informed decisions about resource allocation and public health interventions. The clear segmentation of counties into clusters based on socioeconomic factors and COVID-19 impact allows DSHS to target specific needs more effectively. For example, Cluster 1 counties may require increased healthcare access and financial support to mitigate the effects of the pandemic, while Cluster 3 counties may need additional resources to manage the increased exposure risks associated with their large populations. Additionally, Cluster 2 counties, with their higher income levels and better infrastructure, may benefit from continued support for remote work and maintaining public health measures. By understanding these distinctions, DSHS can allocate resources more efficiently and tailor interventions to the unique challenges faced by each cluster, ultimately enhancing the effectiveness of Texas's public health response to the COVID-19 crisis.

4.4.2.2 K-Means vs. Ward Clustering

Table 8 demonstrates the comparison between the cluster assignments from K-means (k = 3) and Hierarchical Clustering using Ward's method (k = 3), revealing interesting patterns in how the two methods classify Texas counties. One key observation is that K-means Cluster 2 aligns almost perfectly with Ward Cluster 1, with 85 counties matching, suggesting a strong agreement between the two methods for this group. This likely represents counties with affluent, remote characteristics. On the other hand, K-means Cluster 1 is split across all three Ward clusters (17, 10, and 11 counties), indicating some ambiguity or disagreement between the two methods. These counties may fall near the boundaries of the clusters or exhibit mixed features.

Additionally, K-means Cluster 3 is split between Ward Cluster 1 and Ward Cluster 2, which implies that Ward's method may capture more nuanced sub-structures within this group. This comparison shows moderate agreement between the two clustering approaches, particularly for K-means Cluster 2, but also highlights that hierarchical clustering introduces more granular splits, which may enhance analysis at the cost of interpretability. Overall, this reinforces the robustness of K-means while underscoring the importance of considering method sensitivity when interpreting clustering results. For the Texas Department of State Health Services (DSHS), this comparison suggests that while K-means offers simplicity and reliability, hierarchical clustering can provide additional depth for understanding complex county-level patterns.

Table 8: K-Means vs Ward Clustering

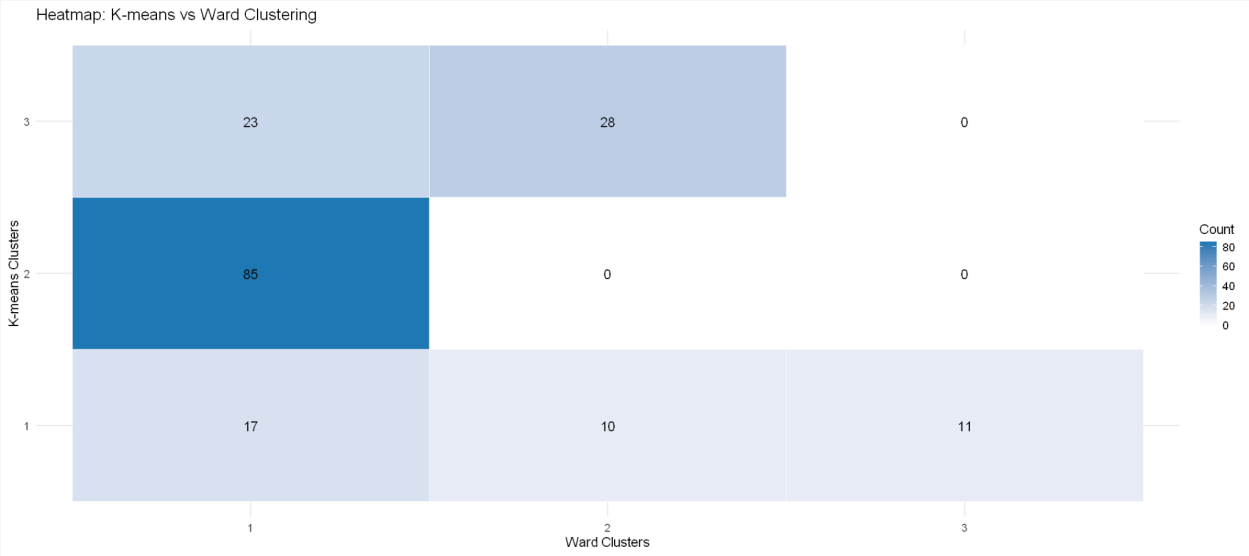
K-Means Cluster	Cluster 1	Cluster 2	Cluster 3
1	17	10	11
2	85	0	0
3	23	28	0

Heatmap Analysis: K-Means vs. Ward Clustering

Figure 18 shows a heatmap comparing the clustering results from K-means (k = 3) and Ward Hierarchical Clustering (k = 3) provides a visual validation of the contingency table, highlighting areas of agreement and divergence between the two methods. The most prominent feature in the heatmap is the strong diagonal pattern, particularly the alignment between K-means Cluster 2 and Ward Cluster 1, where 85 counties overlap. This solid agreement suggests that both methods consistently identify a distinct group of counties, likely those characterized by higher income levels and greater remote work capacity—factors associated with better pandemic resilience. In contrast, K-means Cluster 1 displays more ambiguity, with its counties scattered across all three Ward clusters (17, 10, and 11 counties, respectively), pointing to mixed or transitional socioeconomic and public health characteristics. Similarly, K-means Cluster 3 is split

between Ward Clusters 1 and 2, indicating that hierarchical clustering may detect finer sub-groups within what K-means sees as a single cluster.

Figure 18: Heatmap: K-means vs. Ward Clustering

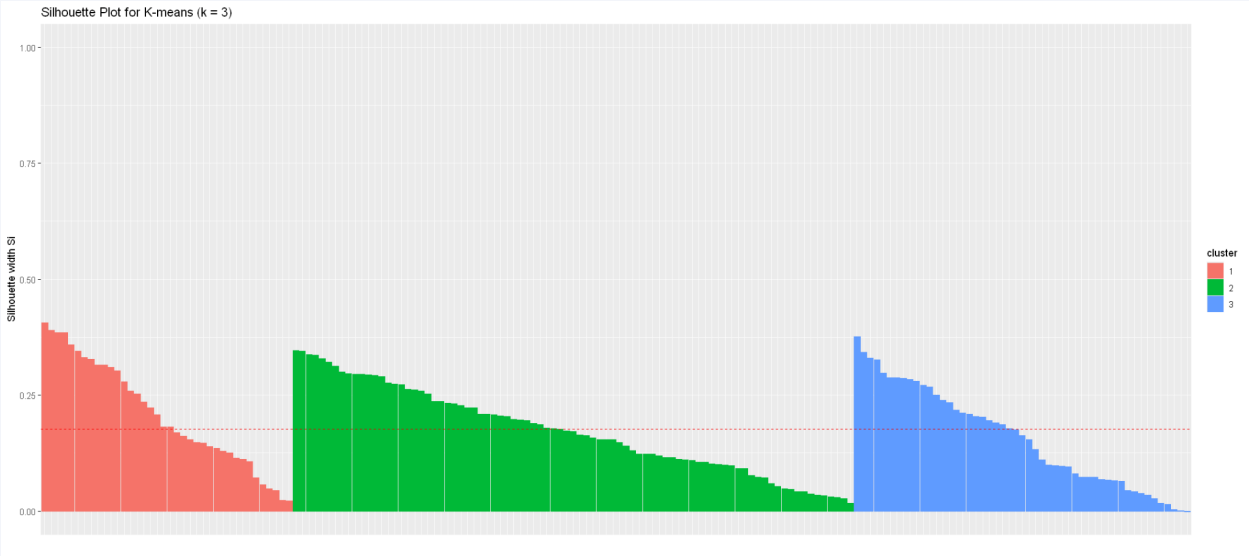


This pattern reinforces the idea that while K-means offers clear, interpretable groupings, Ward's method introduces a level of granularity that may uncover subtle yet important structural differences among counties. Such detail could be especially valuable when identifying counties that don't fit neatly into high- or low-risk categories but may instead straddle multiple risk dimensions. The heatmap analysis also complements the earlier Adjusted Rand Index (ARI) score of 0.334, which reflects moderate agreement between the two methods. For the Texas Department of State Health Services (DSHS), this cross-method validation builds confidence in the clustering results and supports more targeted and adaptable policy planning. By understanding not only where clustering methods agree but also where they diverge, DSHS can better identify county-level nuances and allocate resources to areas that might otherwise be overlooked in simpler models.

Silhouette Analysis: K-Means vs. Ward Clustering

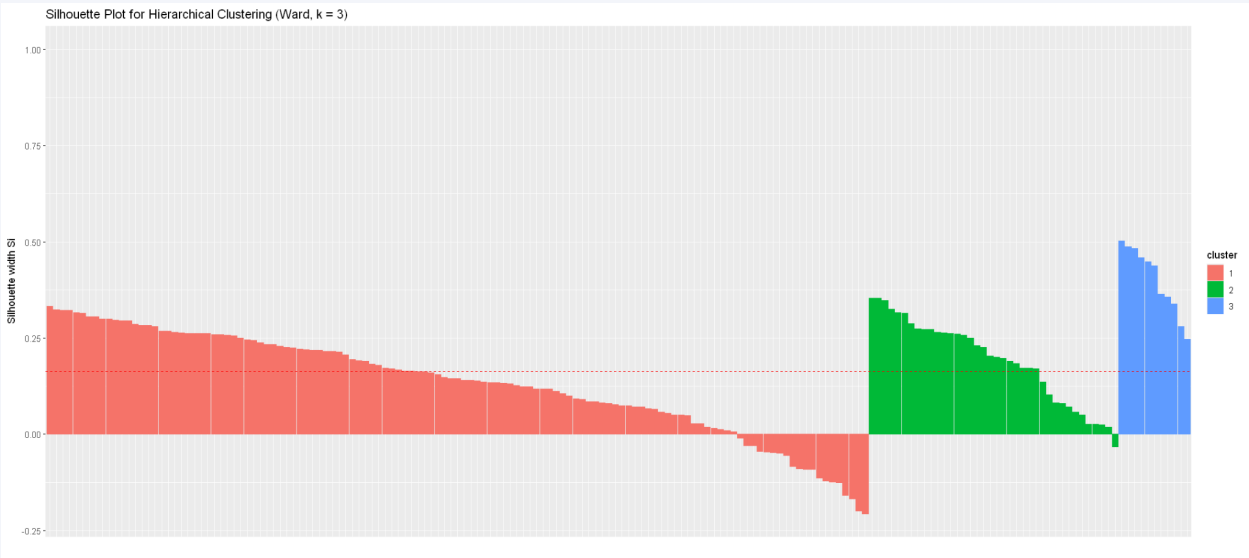
Silhouette analysis was conducted to evaluate the internal validity of the clusters produced by both K-means and Ward's hierarchical clustering, using a silhouette score to measure how well each county fits within its assigned cluster compared to others, as shown in Figure 19 and 20. A higher silhouette score indicates that a data point is more similar to its own cluster than to neighboring clusters, reflecting stronger cohesion and clearer separation.

Figure 19: Silhouette Plot for K-Means (K=3)



For the K-means clustering solution with three clusters, the average silhouette scores were 0.21 for Cluster 1, 0.17 for Cluster 2, and 0.16 for Cluster 3. While these values are relatively low, they are all positive and above 0.15, suggesting that the K-means clusters exhibit some degree of separation and internal consistency, even if not perfectly defined. These results support the idea that K-means provides a reasonable, interpretable global structure, particularly when balancing clarity and simplicity across all three clusters.

Figure 20: Silhouette Plot for Hierarchical Clustering (Ward, K=3)



In comparison, Ward's hierarchical clustering yielded a wider range of silhouette scores. Cluster 3 showed a notably high average silhouette width of 0.40, indicating strong cohesion and excellent separation from other clusters. However, Cluster 1 had a significantly lower score of 0.13, suggesting that counties in this group are less well-defined and may lie closer to the boundaries between clusters. Cluster 2, with a score of 0.19, performed comparably to its K-means counterpart. These differences indicate that while Ward's method may uncover tightly defined subclusters—such as the high-performing Cluster 3—it also produces groupings with lower internal cohesion, possibly due to the method's sensitivity to subtle variations in the data.

Together, these findings reveal that K-means clustering delivers more balanced and consistent cluster structures, making it a strong baseline method for regional analysis. In contrast, Ward's method is more sensitive to local variations and may be better suited for detecting substructure or transitional counties that do not fit cleanly into broad categories. For the Texas Department of State Health Services (DSHS), this analysis emphasizes the strengths of each approach. K-means can be relied upon for clear, statewide segmentation that supports resource planning and policy targeting at a broad level, while Ward's clustering offers added depth that may be valuable when exploring finer-grained distinctions or tailoring interventions for specific subpopulations. By combining both perspectives, DSHS can improve the precision and flexibility of public health strategies across diverse county profiles.

4.5 Unsupervised Evaluation

To evaluate the agreement between the K-means and Ward clustering results, the Adjusted Rand Index (ARI) was used as an unsupervised comparison metric. The ARI assesses how well two sets of clustering labels align, correcting for chance agreement. In this context, since there is no labeled ground truth (such as expert-defined high-risk regions), ARI offers a reliable way to quantify consistency between two independent clustering approaches. The calculated ARI value for the K-means and Ward cluster assignments was **0.334**, which indicates a **moderate level of agreement**.

This numerical result is visually reinforced in **Figure 18**, which shows a heatmap of the contingency table comparing K-means and Ward cluster assignments. A strong diagonal concentration, especially the **85 counties jointly classified under K-means Cluster 2 and Ward Cluster 1**, visually confirms the overlap in certain groups. At the same time, the heatmap also reveals more ambiguous splits for K-means Clusters 1 and 3, consistent with the moderate ARI value.

Further insight comes from the silhouette plots in **Figures 19 and 20**, which help assess internal cluster cohesion. The silhouette plot for K-means (Figure 19) shows more balanced cohesion across all three clusters, while the Ward clustering silhouette (Figure 20) highlights **one highly compact cluster (Cluster 3)** but weaker cohesion in the others. These visuals complement the ARI findings by illustrating how each method captures cluster structure differently—K-means favors compact and evenly sized clusters, while Ward clustering detects nested or fragmented substructures.

From a practical standpoint, these findings offer valuable insights for the Texas Department of State Health Services (DSHS). The moderate ARI score, supported by Figures 18–20, aligns with earlier observations about how each method captures different facets of county-level variation. K-means may be more appropriate when DSHS seeks **simplicity, consistency, and interpretability**, particularly for broad policy messaging. In contrast, Ward's more nuanced splits may be better suited for identifying **subgroups or transitional counties** that warrant **targeted interventions**. Thus, while the ARI does not suggest perfect alignment, it underscores the **complementary strengths** of both clustering approaches and supports using them in tandem for a more robust public health strategy.

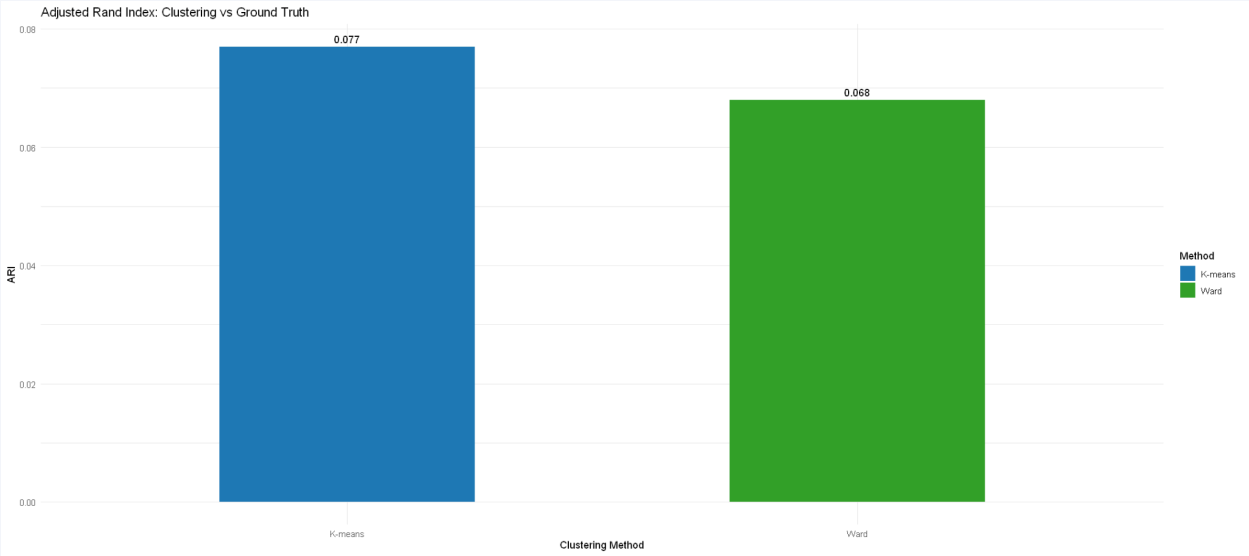
4.6 Supervised Evaluation

To assess how well the unsupervised clusters reflect real-world COVID-19 outcomes, we introduced a **supervised evaluation using ground truth labels** based on the variable `deaths_per_100k`. Counties were classified into three severity tiers: **Low** (≤ 100 deaths per

100k), **Medium** (101–200), and **High** (> 200). This feature was selected as a meaningful proxy for pandemic impact, offering a direct and interpretable reference for evaluating how the clustering algorithms aligned with mortality-based outcomes. Using these categories, we calculated the **Adjusted Rand Index (ARI)** between the K-means and Ward clustering results and the ground truth severity tiers.

The results in Figure 21 showed **low ARI values** for both methods: **0.077** for K-means and **0.068** for Ward. These scores suggest that neither clustering technique aligns strongly with the `deaths_per_100k` tiers. However, K-means demonstrated slightly better performance. This outcome is not entirely surprising given that the clustering was based on a **broad set of socioeconomic and behavioral indicators**, rather than purely on COVID-19 health outcomes. As such, the groupings reflect **underlying structural patterns**—such as income, mobility, and food assistance—rather than directly mirroring death rates. Furthermore, reducing a continuous measure like `deaths_per_100k` into just three categories introduces a level of simplification that may overlook finer-grained variation within the data.

Figure 21: ARI Values for K-means & Ward



From the perspective of the **Texas Department of State Health Services (DSHS)**, this analysis provides a nuanced insight. Although the clustering does not directly replicate COVID-19 mortality tiers, it still captures **underlying county-level vulnerabilities** that may influence pandemic outcomes. The slight edge in performance by K-means suggests it may be marginally better suited for understanding and segmenting the population based on **general risk factors** rather than just retrospective outcomes. For DSHS, this highlights that while clusterings are not a replacement for direct outcome prediction, they can still serve as **informative tools for resource planning, risk stratification, and tailored intervention strategies**. Future work could incorporate additional health infrastructure or policy data to further enhance alignment with mortality outcomes.

4.7 Graduate Level Analysis by Salissa Hernandez

Advanced Clustering Techniques:

Density-Based Spatial Clustering of Applications with Noise (DBSCAN)
Gaussian Mixture Model (GMM)

To enrich the depth and rigor of this project, I extended our analysis beyond the standard k-means and hierarchical (Ward's) clustering by implementing two additional algorithms: **DBSCAN** (Density-Based Spatial Clustering of Applications with Noise) and **Gaussian Mixture Models (GMM)**. These techniques were selected for their ability to overcome limitations inherent in centroid-based methods such as k-means—particularly the assumptions of spherical cluster shapes, equal variances, and uniform density.

DBSCAN enables the identification of clusters with arbitrary shapes and can detect outliers as noise, making it well-suited for datasets that exhibit uneven distribution or spatial sparsity—conditions often present in county-level epidemiological data. **GMM**, on the other hand, provides a probabilistic framework for clustering by modeling the data as a mixture of multiple Gaussian distributions, allowing for soft cluster assignments and overlapping group boundaries. This makes GMM an appropriate choice when counties may not fall neatly into a single, discrete category.

By comparing these advanced models to k-means and Ward's method, I aim to assess whether more nuanced and flexible clustering approaches yield additional insights into Texas counties' pandemic vulnerability and demographic patterns. The following sections evaluate each method's output in terms of cluster interpretability, robustness, and relevance to the Texas Department of State Health Services' (DSHS) public health strategy.

4.7.1 DBSCAN Overview:

Unlike K-means or hierarchical clustering, DBSCAN:

- **Does not require specifying the number of clusters in advance**
- **Can find arbitrarily shaped clusters**
- **Handles noise and outliers naturally**

Key Concepts

- **Core Point**: A point with at least minPts neighbors within a radius eps.
- **Border Point**: A point that is within eps of a core point but doesn't have enough neighbors to be a core point itself.
- **Noise Point**: A point that is not a core point and not within eps of any core point.

Key Parameters

- eps: The maximum distance between two points to be considered neighbors.
- minPts: The minimum number of points required to form a dense region (including the point itself).

A good rule of thumb:

- Use **domain knowledge or a k-distance plot** to determine eps
- Set minPts = dimensionality + 1 (e.g., if 10 variables, try minPts = 11)

Why Use DBSCAN?

- Excellent for **real-world datasets** where clusters are not well-separated or spherical.
- Automatically detects **noise** (e.g., counties that don't clearly fit a pattern).
- Ideal for **spatial data**, such as geographic county-level COVID patterns, where population clusters may vary in shape and density.

I apply DBSCAN to the scaled county-level dataset used for K-means and Ward clustering. The goal is to:

- See whether DBSCAN can uncover **natural density-based groupings**
- Identify **outlier counties** that don't fit into broader pandemic patterns
- Compare the results to K-means and hierarchical clustering in terms of interpretability and robustness

4.7.2 DBSCAN: k-NN Distance Plot Analysis

To determine a suitable value for eps in DBSCAN, we use a k-NN distance plot. This technique helps identify the "elbow" point in the distance graph, where the curve sharply increases. This elbow marks the transition from dense clusters to sparser, noisier regions.

In this case, we selected **k = 11**, based on the rule of thumb $\text{minPts} = \text{dimensions} + 1$, which reflects our 10-variable dataset.

Interpretation:

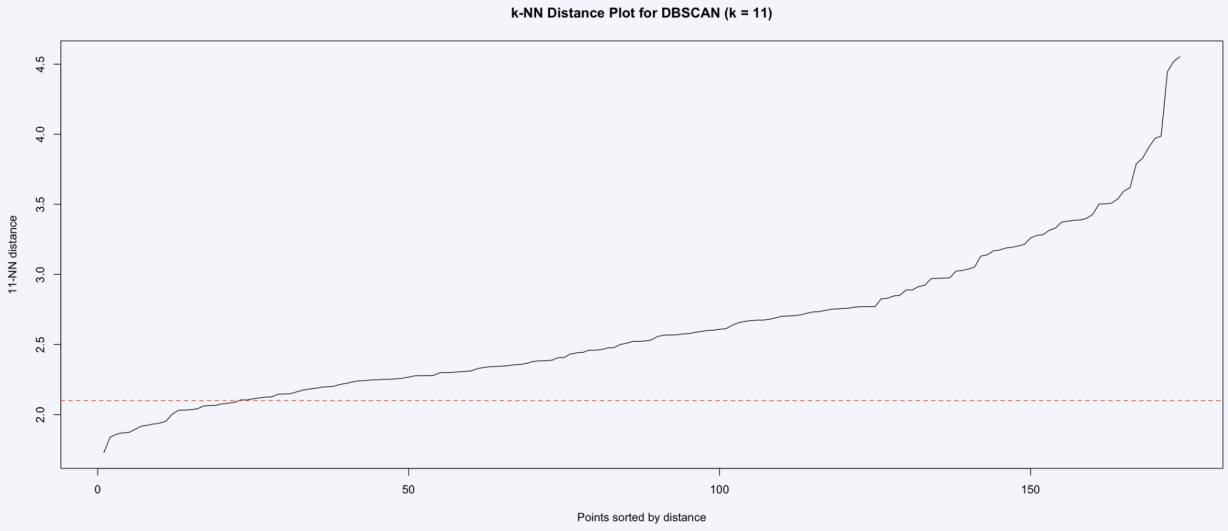
- The curve shows a distinct elbow around the distance value of **2.1**.
- Setting $\text{eps} = 2.1$ allows us to capture the core points of dense regions while still allowing flexibility for cluster boundaries.
- Values much smaller than this may miss meaningful clusters, while larger eps values risk merging dissimilar groups.

This plot supports our choice of:

- $\text{eps} = 2.1$
- $\text{minPts} = 11$

These parameters are now used to run DBSCAN and identify clusters based on density rather than distance to a centroid.

Figure 22: k-NN Distance Plot for DBSCAN



4.7.3 DBSCAN Clustering: PCA Projection Analysis

The PCA projection plot provides a two-dimensional visualization of the clusters discovered by the DBSCAN algorithm using $\text{eps} = 2.1$ and $\text{minPts} = 11$.

Interpretation:

- DBSCAN identified **two primary groups**:
  - **Cluster 1 (light blue)** – a relatively compact and well-formed group that DBSCAN recognized as dense.
  - **Cluster 0 (red)** – includes points that DBSCAN marked as **noise** or **not belonging to any dense region**, often treated as outliers or sparse areas.

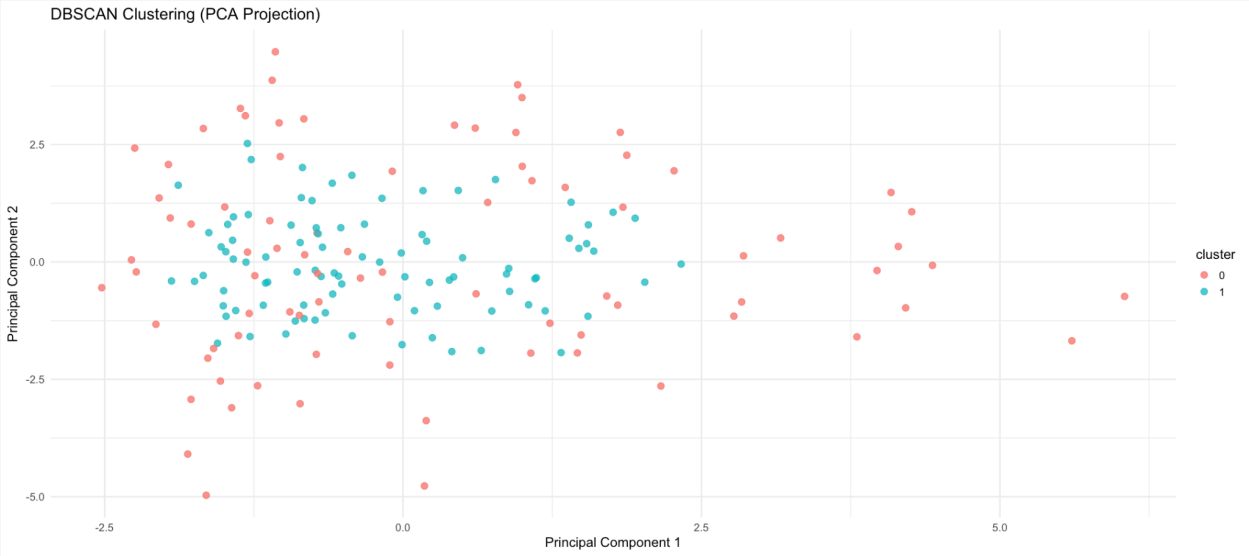
Key Observations:

- Unlike K-means or hierarchical clustering, **DBSCAN does not force all points into clusters**. This is evident in the wide spread of Cluster 0 points (likely noise).
- DBSCAN is well-suited to datasets with **irregular cluster shapes or varying densities**, and the PCA projection supports this—showing structure not strictly separable by distance to centroids.

- Cluster 1 appears centralized and dense, consistent with DBSCAN's core point definition.
- The presence of noise points in Cluster 0 reveals DBSCAN's strength in **outlier detection**, which other methods often fail to identify.

Conclusion: DBSCAN offers a distinct, density-based view of the county-level COVID and socioeconomic data. Though fewer clusters were identified than with K-means or hierarchical clustering, the ability to detect noise and avoid assumptions about shape/size adds value to the exploratory analysis.

Figure 23: DBSCAN Clustering (PCA Projection)



4.7.3.1 DBSCAN Clustering: PCA Projection Analysis

The plot above displays the results of **DBSCAN clustering** applied to scaled COVID-19 and census data from Texas counties, visualized using **Principal Component Analysis (PCA)** for dimensionality reduction. Each point represents a county projected into two principal components, with color denoting the assigned cluster.

Key Observations:

- **Two main clusters** were identified, shown in distinct colors (Cluster 0 and Cluster 1).
- Cluster **1** appears denser and more compact, suggesting counties with similar population structure, income levels, or COVID-19 case/death patterns.
- Cluster **0** contains more dispersed points, likely representing counties with outlier behavior—such as extreme poverty rates, low public transportation usage, or distinct COVID-19 outcomes.
- Some counties originally labeled as **noise (Cluster 0)** were converted to NA for visual clarity. These were not dense enough to form part of any core cluster based on the eps and minPts parameters used.

Interpretation:

- **DBSCAN excels** at detecting clusters of arbitrary shape and identifying outliers. This projection shows that a subset of Texas counties shares strong similarities (Cluster 1), while others are more anomalous (Cluster 0).
- The clear spatial separation in PCA space suggests that **underlying demographic and pandemic response features** meaningfully influence cluster formation.
- PCA is an unsupervised method and does not retain all data variance, but it highlights the **overall structure and spread** within the dataset.

4.7.4 DBSCAN Cluster-Level Summary Statistics

This summary in Table 9 compares the average values of key socioeconomic and health-related metrics across the two clusters formed by the DBSCAN algorithm.

Cluster Comparison:

- **Cluster 0 represents counties with:**
  - Higher average COVID-19 case rates (8,049 per 100k) and death rates (191 per 100k).
  - Elevated poverty levels (avg ≈ 4,153) and greater dependence on food assistance programs (12.55%).
  - Substantially higher public transportation usage (26.21%) and slightly lower remote work adoption(3.07%).
  - Slightly higher income per capita and median income, though these differences are not dramatic.
  - Significantly larger populations (avg ≈ 26,746), indicating these are more urban or densely populated counties.
- **Cluster 1 includes counties with:**
  - Lower COVID-19 burden, with fewer cases and deaths on average.
  - Better socioeconomic conditions, including lower poverty (avg ≈ 2,944), reduced public transit use, and a slightly higher proportion of residents working from home (3.24%).
  - Smaller populations (avg ≈ 18,898), suggesting these may be suburban or rural counties.

Table 9: DBSCAN Cluster-Level Summary Statistics

Summary Statistics	Cluster 0	Cluster 1
avg_cases_per_100k	8,049.03	6,905.83
avg_deaths_per_100k	191.40	182.47
avg_poverty	4,153.60	2,944.39
avg_median_income	49,574.73	47,319.47
avg_pct_on_food_stamps	12.55	13.17
avg_commuters_by_ public_transportation	26.21	11.43
avg_pct_work_from_home	3.07	3.24
avg_income_per_capita	24,318.10	23,852.29
avg_gini_index	0.45	0.45
avg_total_pop	26,745.84	18,897.63

Interpretation:
Cluster 0 reflects **higher-risk, urban counties** with more vulnerability due to economic hardship and greater reliance on shared transportation modes. These factors likely contributed to increased viral exposure and transmission.
In contrast, Cluster 1 appears to capture **lower-density, potentially more resilient counties**, where individuals may have had greater capacity to socially distance or work remotely, contributing to reduced case rates.
Implications for Policymakers:

- Counties in **Cluster 0** may benefit from targeted public health strategies, such as increased testing, mobile vaccination units, and social services.
- The DBSCAN algorithm effectively identified counties that stand out based on **density-based patterns**, which traditional clustering approaches (e.g., k-means) might overlook.
- This differentiation supports the use of DBSCAN when structural gaps or outlier behaviors are meaningful for intervention planning.

4.7.5 DBSCAN Core, Border, and Noise Points: Income vs. Poverty Rate

The scatterplot below presents DBSCAN clustering results overlaid with standardized **median income** (x-axis) and **poverty rate** (y-axis). Data points are color-coded by DBSCAN point classification: Core (blue), Border (yellow), and Noise (red). Cluster assignments are shown with shape markers (▲).

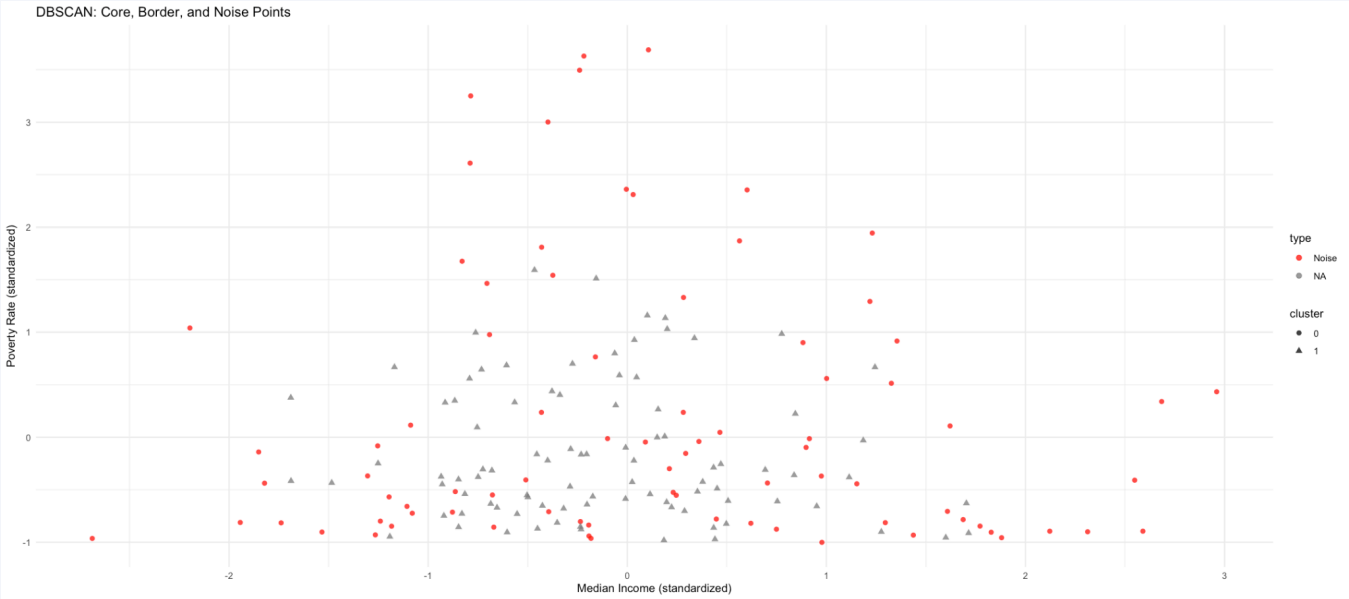
Interpretation:

- The **majority of counties are identified as "Noise" points (red)**, suggesting that these counties do not belong to any dense cluster according to DBSCAN's density criteria ($eps = 2.1$, $minPts = 11$).
- **Core points (blue)** are concentrated near the center of the distribution, where counties tend to have **moderate median incomes and poverty rates**, suggesting a denser grouping of socioeconomically similar counties.
- **Border points (yellow),** though fewer, lie at the edges of core clusters—often overlapping slightly with noise points.
- Many **noise points (red)** appear at the outer fringes—particularly in counties with v**ery low or very high poverty rates**, or **low income**—indicating they behave differently from the typical cluster profile.
- The use of **standardized features** enhances comparability across variables and removes scale-related distortions.

Insights:

- DBSCAN has difficulty clustering counties that fall into **socioeconomic extremes**, as they don't meet the local density requirements to form a group.
- This visualization reinforces the value of **density-based methods** for uncovering structural gaps in the data where typical clustering (e.g., k-means) might force outlier counties into misleading groupings.
- Policymakers can focus on noise points to identify outlier counties that may need individualized interventions.

Figure 24: DBSCAN Core, Border and Noise Points



4.7.6 Geospatial Analysis of DBSCAN Clustering Across Texas Counties

This choropleth-style map visualizes the spatial distribution of counties assigned to DBSCAN clusters across the state of Texas. Counties are filled based on their DBSCAN cluster assignment:

- **Cluster 1** counties are shaded in dark blue.
- **NA (gray)** indicates counties that were identified as noise (i.e., not assigned to any cluster).

Key Observations:

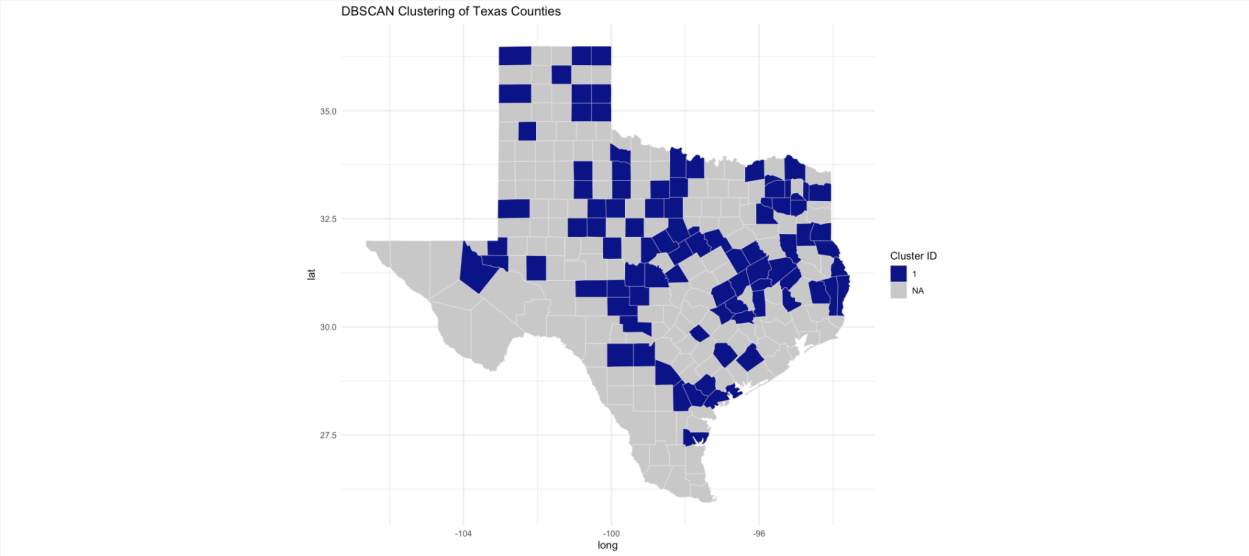
- **Dense clusters appear in regions with more populous or socioeconomically similar counties,** such as parts of eastern and central Texas.
- **Sparse and rural counties**, especially in west and far-south Texas, were more frequently labeled as noise, suggesting these counties did not meet the density threshold required for DBSCAN to form a cluster.
- The **concentration of clustered counties** in certain regions implies that shared economic and demographic characteristics (e.g., income level, poverty rates, or public transit usage) might have driven the clustering.
- This map highlights DBSCAN's ability to **exclude outliers** (i.e., noise) and r**eveal contiguous regional patterns** where counties share latent similarities.

Implications:

- This clustering result helps identify geographic regions that may benefit from **targeted interventions**, especially where socioeconomic indicators align.
- Policymakers may use this insight to tailor regional responses to public health or economic challenges based on shared profiles within clusters.

This spatial perspective complements the earlier scatterplots and cluster summaries by offering **geographic context** to the data-driven groups, further validating the utility of DBSCAN for regional analysis.

Figure 25: DBSCAN Cluster Distribution Across Texas Counties



4.7.7 Comparison of DBSCAN vs K-means and Hierarchical Clustering (Ward)

To quantitatively compare the clustering results of DBSCAN to traditional methods like K-means and hierarchical clustering, we computed the **Adjusted Rand Index (ARI)**. ARI evaluates the similarity between two clustering assignments, correcting for chance.

Interpretation of Table X:

- Both ARI values are **very low**, indicating **poor agreement** between DBSCAN and the other methods.
- This is **expected**:
  - **K-means** and **Ward's method** are **centroid-based or linkage-based methods** that assign every point to a cluster.
  - **DBSCAN**, by contrast, **identifies dense regions** and **labels sparse/outlier points as noise (cluster 0)**.
- The **low overlap** suggests DBSCAN captured a **different structure** in the data—one that reflects **density-based groupings** rather than distance to centroids or variance explained.

Conclusion: DBSCAN offers a **unique perspective** by identifying clusters based on density rather than global distance metrics. While its clusters don't align well with K-means or hierarchical ones, it may better identify **subgroups or outliers** in complex datasets like this one.

Table 10: Adjusted Rand Index

Adjusted Rand Index	
DBSCAN vs K-means	0.021
DBSCAN vs Ward	0.059

4.7.8 Gaussian Mixture Models (GMM)

Overview

A **Gaussian Mixture Model (GMM)** is a *probabilistic clustering technique* that assumes the data is generated from a mixture of several **Gaussian (normal) distributions**, each with unknown parameters.

Unlike **K-means**, which assigns each point to a single cluster (**hard assignment**), GMM assigns a **probability** to each point for belonging to each cluster (**soft assignment**). This provides greater flexibility in modeling **elliptical**, **uneven**, or **overlapping clusters**.

Key Features of GMM

- **Probabilistic Assignment**: Each data point receives a probability score for each cluster.
- **Soft Clustering**: Handles uncertainty in cluster membership.
- **Flexible Cluster Shapes**: Can model clusters of various shapes and sizes.
- **Expectation-Maximization (EM)**: An iterative algorithm used to estimate cluster parameters (mean, covariance, and mixing proportions).

```
----------------------------------------------------
Gaussian finite mixture model fitted by EM algorithm
----------------------------------------------------

Mclust EVE (ellipsoidal, equal volume and orientation) model with 4 components:

 log-likelihood   n  df       BIC       ICL
      -1644.695 174 125 -3934.273 -3951.365

Clustering table:
  1  2  3  4
 91 28 33 22
```

Summary of GMM Results

The Gaussian Mixture Model (GMM) was applied to the scaled dataset using the **Mclust** algorithm. The best-fitting model selected was the **EVE model**—which assumes **ellipsoidal clusters with equal volume and orientation**. This model was chosen based on the Bayesian Information Criterion (BIC) and Integrated Completed Likelihood (ICL) values:

- **Log-likelihood:** -1644.695
- **Number of observations (n):** 174
- **BIC:** -3934.273
- **ICL:** -3951.365

The GMM assigned counties into **four distinct clusters**, with the following distribution:

- **Cluster 1:** 91 counties
- **Cluster 2:** 28 counties
- **Cluster 3:** 33 counties
- **Cluster 4:** 22 counties

This outcome supports the presence of **heterogeneous subgroupings** within the dataset, with the majority of counties concentrated in Cluster 1. The distribution also indicates that the population data exhibits **non-spherical, overlapping patterns**, reinforcing the appropriateness of GMM over methods like K-means.

4.7.8 Visualization Analysis

The plot below displays the clustering result of the **Gaussian Mixture Model (GMM)** applied to the scaled demographic and COVID-related data from Texas counties.

Key Observations:

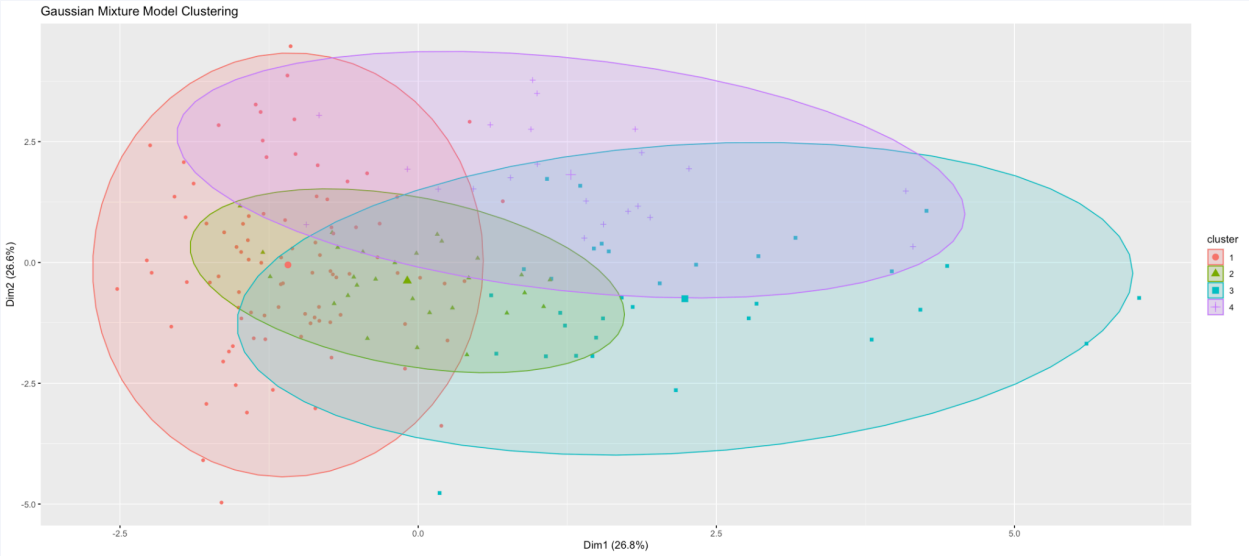
- **Elliptical Clusters**: Unlike K-means which forces clusters into spherical shapes, GMM reveals **elliptical boundaries**, which are more aligned with real-world variability and feature correlations.
- **Soft Boundaries**: The overlapping ellipses suggest that counties may not belong definitively to just one group. Instead, GMM assigns **probabilities** of cluster membership, ideal for nuanced, overlapping data.
- **Cluster Separation**: While there's some visual overlap between clusters, distinct groups emerge — especially along Dim1 (the primary PCA dimension), indicating **differentiation in core features like income, poverty, or COVID case rates**.

- **Component Sizes**: The four clusters vary in size and orientation, capturing **heterogeneity in county characteristics**. One cluster, for example, is much larger and more diffuse, possibly representing counties with moderate demographic traits.

Implications:

- GMM helps us **model uncertainty and ambiguity** in clustering, which is common in socioeconomic data.
- Counties falling in **overlapping regions** of ellipses may exhibit mixed traits — useful for targeting **intervention strategies** where counties may shift risk profiles over time.
- This technique supports more **flexible cluster interpretation**, particularly for policy-making, resource allocation, or deeper exploration of vulnerable populations.

Figure 26: Gaussian Mixture Model Clustering



4.7.9 Analysis: GMM Cluster-Level Summary Statistics

The following insights emerge from the cluster-level summary of the **GMM** clustering:

Cluster 1 – Highest Cases and Deaths

- **Highest average cases per 100k**: ~7,646
- **Highest average deaths per 100k**: ~214
- **Highest poverty** and relatively low income per capita
- **Lowest public transit usage** and **work-from-home rates**

This cluster likely represents **economically vulnerable, rural counties** with limited infrastructure and fewer remote work opportunities — placing them at **higher risk** during the pandemic.

Cluster 2 – Lower Impact, Higher Transit Access

- **Second-lowest average cases/deaths**
- **Higher public transportation use** (~22%)
- **Lower remote work access** and **lower income**

May represent **urban-adjacent or suburban** communities with public transport reliance but mixed socioeconomic levels. These counties may have mitigated outbreaks better despite dense movement.

Cluster 3 – Urban Cluster with Overlap

- **Similar average cases to Cluster 2**, but lower deaths
- **Extremely high poverty** (8470) and **highest transit usage**
- **Lowest income per capita**

High mobility and poverty suggest these counties are **urban with dense populations** and **uneven socioeconomic access** — highlighting complex vulnerability beyond income alone.

Cluster 4 – Higher-Income, Lower Mortality

- **Highest median income** and **highest income per capita**
- **Lowest average deaths per 100k**
- **Higher remote work access**, **moderate public transit use**

Cluster 4 appears to represent **affluent counties** with **better infrastructure and access** to resources, including the ability to socially distance and work from home, resulting in **lower fatality rates**.

Takeaways

- GMM revealed **four nuanced clusters** with differences in **demographic risk**, **infrastructure**, and **outcomes**.
- It excels at capturing **overlapping, elliptical clusters**, which might not be as cleanly detected by K-Means.
- Results suggest **poverty**, **public transportation**, and **remote work capacity** are critical indicators of COVID-19 impact, even more than population alone.

Table 11: GMM Cluster-Level Summary Statistics

Summary statistics	Cluster 1	Cluster 2	Cluster 3	Cluster 4
avg_cases_per_100k	7,646.28	7,212.50	7,598.25	6,675.13
avg_deaths_per_100k	213.63	174.86	162.16	127.03
avg_poverty	1,398.88	3,062.50	8,470.73	5,404.41
avg_median_income	47,840.15	45,675.75	45,855.42	57,859.77
avg_pct_on_food_stamps	12.07	14.17	15.47	10.71
avg_commuters_by_ public_transportation	1.78	22.11	51.94	32.09
avg_pct_work_from_home	3.15	2.72	3.04	3.98
avg_income_per_capita	24,063.45	22,593.04	22,375.12	28,533.5
avg_gini_index	0.45	0.44	0.46	0.44
avg_total_pop	9,391.47	18,722.14	48,887.21	42,709.95

4.7.10 Comparison: GMM vs Other Clustering Methods

To evaluate how well the clustering assignments from the **Gaussian Mixture Model (GMM)** align with those from other methods, we computed the **Adjusted Rand Index (ARI)** for the following pairs:

Table 12: GMM vs Other Clustering Methods

Comparison	Score	Interpretation
GMM vs DBSCAN	0.0203	*Very low agreement — suggests GMM and DBSCAN identify very different cluster structures.*
GMM vs K-Means	0.2832	*Moderate agreement — some overlap, but GMM captures probabilistic groupings K-means may miss.*
GMM vs Ward	0.1859	*Low to moderate agreement — both are hierarchical but likely assign counties to different groups due to GMM's flexibility.*

Interpretation

- **DBSCAN** excels at detecting density-based, non-spherical clusters and identifying noise. Its low ARI with GMM indicates that these models detect **very different patterns**, particularly in sparse or noisy areas.
- **K-Means** assumes spherical, equally sized clusters, so moderate alignment with GMM is expected. However, **GMM allows for elliptical clusters and soft assignment**, capturing more nuanced cluster boundaries.
- **Ward Hierarchical Clustering** builds a strict tree-based hierarchy of clusters. Its lower ARI with GMM suggests differences in **cluster merging decisions** and **shape assumptions**.

Conclusion

These ARI results emphasize the **importance of choosing a clustering algorithm that matches the data's underlying structure**:

- Use **GMM** when cluster shapes are elliptical or overlapping.
- Use **K-Means** when clusters are well-separated and spherical.
- Use **DBSCAN** when dealing with **irregular shapes, noise, or outliers**.
- Use **Ward** for **hierarchical relationships** or when a dendrogram is desired.

4.7.11 GMM Cluster Assignment Uncertainty Analysis

The **uncertainty plot** shows how confident the Gaussian Mixture Model (GMM) is when assigning each county to a cluster.

- **Dot Size = Uncertainty**: Larger points indicate **higher uncertainty** in cluster assignment, while smaller points indicate **higher confidence**.
- Most counties are confidently assigned to a cluster (small dots), but several near the boundaries show larger dots—this reflects **overlapping characteristics** across clusters.
- GMM's **soft clustering** approach allows each county to belong partially to multiple clusters, providing a richer understanding of **ambiguous or transitional regions**.

Key Takeaways:

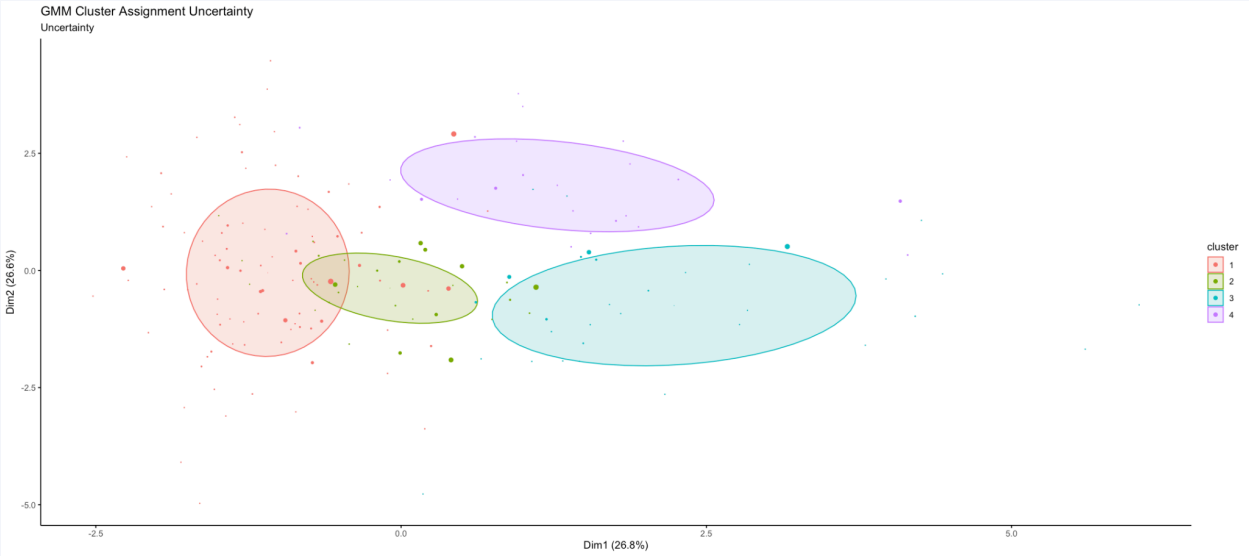
- Uncertainty reveals **borderline counties** that don't clearly fit into a single demographic group.
- This is particularly helpful in **public health planning**, where counties may require mixed policy interventions.
- Unlike K-means, which offers hard assignments only, GMM captures **nuanced, probabilistic relationships** between counties and cluster centers.

Model Selection and Diagnostic Criteria

Attempted to plot model selection criteria using:

```
plot(gmm_model, what = "BIC")
plot(gmm_model, what = "ICL")
plot(gmm_model, what = "likelihood")
```

Figure 27: GMM Cluster Assignment Uncertainty



4.7.12 Understanding BIC in Gaussian Mixture Models (GMM)

The **Bayesian Information Criterion (BIC)** is a model selection metric used to evaluate the trade-off between model fit and complexity. In the context of **Gaussian Mixture Models**, BIC helps determine:

- The **optimal number of clusters (components)**.
- The most appropriate **covariance structure** (e.g., spherical, diagonal, ellipsoidal).

How BIC Works:

- **Lower BIC values** indicate better models.
- BIC penalizes model complexity, so even if a model fits the data better, it must do so **efficiently** to be preferred.
- GMM models are compared across a range of component numbers and structures, and the best model minimizes BIC.

BIC Plot Analysis: Model Selection for GMM

The BIC plot displays:

- **X-axis**: Number of components (1 to 9)
- **Y-axis**: BIC score (lower = better)
- **Lines**: Different covariance structures (e.g., EII, EEV, VEV)

Covariance Structure Legend (Acronym Key):

Each acronym in the Acronym Key plot legend in **Appendix 8** corresponds to a different covariance parameterization of the Gaussian components. These affect how the model estimates the **volume, shape, and orientation** of clusters:

Key Insights:

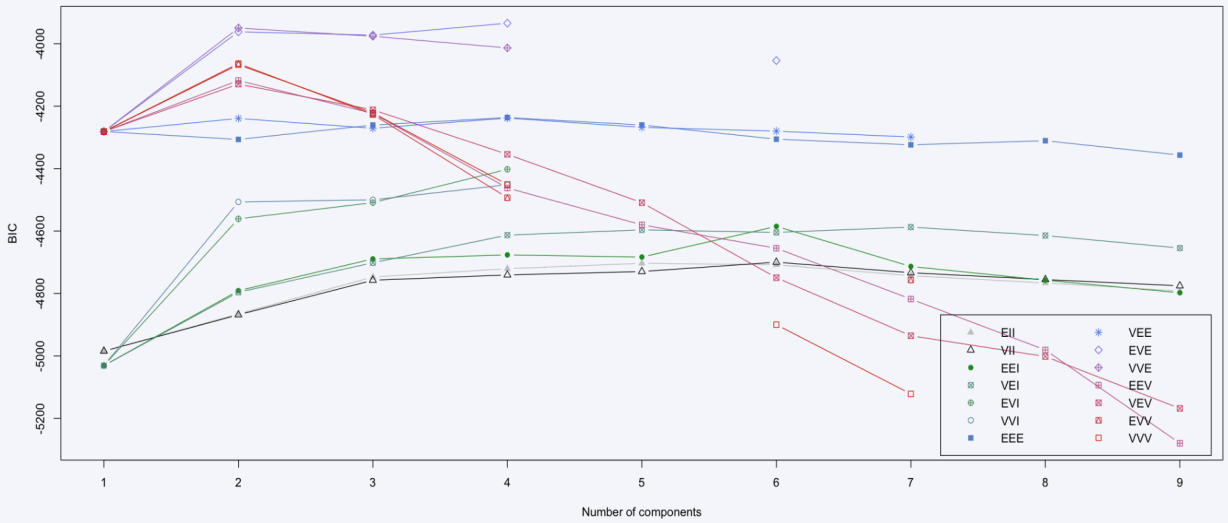
- **Optimal Model**: The BIC curve identifies **4 components with EVE or EEV structure** as the best-performing models (lowest BIC).
    - These structures allow for **elliptical clusters** with varying volume and shape.
    - This flexibility is useful for **real-world demographic data**, which rarely forms perfect spheres.

- **Simpler Models Perform Worse**:
    - Models like **EII** (equal volume, spherical clusters) have much higher BIC values, suggesting poor fit.
- **Why This Matters**:
    - The chosen model captures complexity in the dataset (e.g., poverty, population density, income) without overfitting.
    - It confirms that **4 distinct clusters** are a statistically sound choice.

Conclusion:

- BIC provides a principled approach to model selection in GMM.
- It supports our decision to use a **4-cluster elliptical model** and avoids the pitfalls of arbitrary cluster selection.
- This evaluation strengthens the **validity and interpretability** of our clustering analysis.

Figure 28: GMM Cluster Assignment Uncertainty



4.7.12 Conclusion: DBSCAN and GMM Clustering

DBSCAN

- **Strengths**: DBSCAN effectively identified non-spherical clusters and isolated counties that may be considered **noise or outliers**, which traditional clustering methods fail to recognize.
- **Insights**: The low Adjusted Rand Index (ARI) when compared with K-means and Ward clustering suggests that DBSCAN uncovered a fundamentally different structure — based on **density** rather than **distance to centroids**.
- **Use Case**: Particularly useful for finding **high-risk counties with unusual density patterns**, or when detecting **anomalies and sparse regions** is important.

Conclusion: DBSCAN offers a **unique density-based perspective** and excels in detecting irregular cluster shapes and outliers — making it an important complement to centroid-based methods like K-means.

Gaussian Mixture Models (GMM)

- **Strengths**: GMM enabled **soft clustering**, assigning **probabilities** to each county's cluster membership — which is crucial when demographic boundaries are **ambiguous or overlapping**.
- **Insights**: GMM revealed **elliptical and overlapping clusters**, highlighting subtler socioeconomic patterns that are not as clearly captured by K-means or Ward.

- **Model Fit**: The model selection via **BIC** favored 4 components, and the **uncertainty plot** highlighted where cluster assignments were least confident — guiding potential follow-up investigation.

GMM enhances the clustering analysis by providing **probabilistic, flexible boundaries** and is well-suited for **real-world data** where counties may not belong rigidly to one group.
Both DBSCAN and GMM go **beyond the limitations of K-means and hierarchical clustering**. Together, they provide a **richer understanding** of county-level COVID-19 vulnerability based on **density** (DBSCAN) and **probabilistic overlap** (GMM).

4.8 Graduate Level Analysis by Juan Carlos Dominguez

Advanced Clustering Techniques:

Spectral Clustering
Agglomerative Clustering

To further deepen the clustering analysis and provide the Texas Department of State Health Services (DSHS) with a broader view of regional segmentation, I extended the evaluation by applying **Spectral Clustering** and **Agglomerative Clustering** to the same county-level COVID-19 and socioeconomic dataset. These methods were chosen for their complementary strengths, especially in identifying complex cluster structures that traditional techniques like K-means or Ward's method might oversimplify or overlook.

Spectral Clustering leverages the mathematical framework of graph theory by using the eigenvalues of a similarity matrix to reduce dimensionality before performing clustering. This approach is particularly powerful when the cluster structure is non-convex or when there are subtle boundaries between regions that don't follow conventional Euclidean separations. It works by interpreting data as a graph, where counties with similar profiles are more strongly connected. This makes Spectral Clustering especially useful in epidemiological data where health behaviors, economic conditions, and public health vulnerabilities may diffuse through shared regional characteristics or infrastructure connections rather than discrete numeric thresholds.

Agglomerative Clustering, on the other hand, builds clusters in a bottom-up fashion, starting with each county as its own group and iteratively merging the most similar pairs based on a linkage criterion. Unlike Ward's method, which minimizes variance within clusters, agglomerative clustering with average or complete linkage can capture elongated or irregularly shaped clusters. This flexibility makes it valuable when counties show overlapping patterns of pandemic response and socioeconomic traits, without assuming that all clusters must be compact or evenly distributed.

By incorporating Spectral and Agglomerative Clustering into the analysis, I aimed to test whether these algorithms can uncover latent structures or subclusters that might have been masked under centroid- or variance-based methods. In particular, this comparison explores whether spectral methods better detect counties with hybrid risk patterns, and whether agglomerative strategies highlight gradual transitions or nested relationships between clusters. The resulting outputs were then evaluated in terms of coherence, interpretability, and how well they align with potential public health needs, offering DSHS additional flexibility in tailoring interventions based on different structural insights.

4.8.1 Spectral Clustering

Spectral clustering is a graph-based clustering technique that uses the eigenvectors of a similarity matrix to identify clusters in data that may not be linearly separable. It is particularly effective when the data exhibits non-convex or manifold-like structures, which are often not well captured by centroid-based methods such as k-means. The approach is well-suited to uncover complex relationships in the data, making it useful for identifying groupings that might be missed by more traditional methods. In this analysis, spectral clustering is applied using the `specc()` function from the **`kernlab`** package, which utilizes a radial basis function (RBF) kernel to construct a similarity graph. This method allows us to detect subtle groupings between counties that share similar pandemic responses or demographic characteristics, even if they are not spatially close or closely related in terms of the variables being studied.

In this approach, the number of clusters is set to 3 to maintain comparability with other clustering methods, such as k-means. The data used in spectral clustering has been preprocessed by standardizing the numeric features, ensuring that each feature contributes equally to the distance calculations. Spectral clustering's strength lies in its ability to detect non-linear relationships and to identify clusters that do not necessarily follow a simple geometric pattern. The resulting clusters can reveal important insights into the different behavioral or geographic patterns in the dataset, potentially reflecting variations in public health responses, socioeconomic conditions, or mobility patterns among counties.

Compared to methods like k-means, spectral clustering provides a more flexible view of the data's underlying structure. It can uncover clusters with irregular shapes, offering deeper insights into the dynamics at play in the dataset. In the context of COVID-19, this flexibility is especially valuable for identifying nuanced patterns in how counties have been impacted by the pandemic. Spectral clustering's ability to highlight complex, non-convex groupings makes it a powerful tool for public health analysis, allowing for a more accurate and detailed understanding of the factors influencing pandemic outcomes.

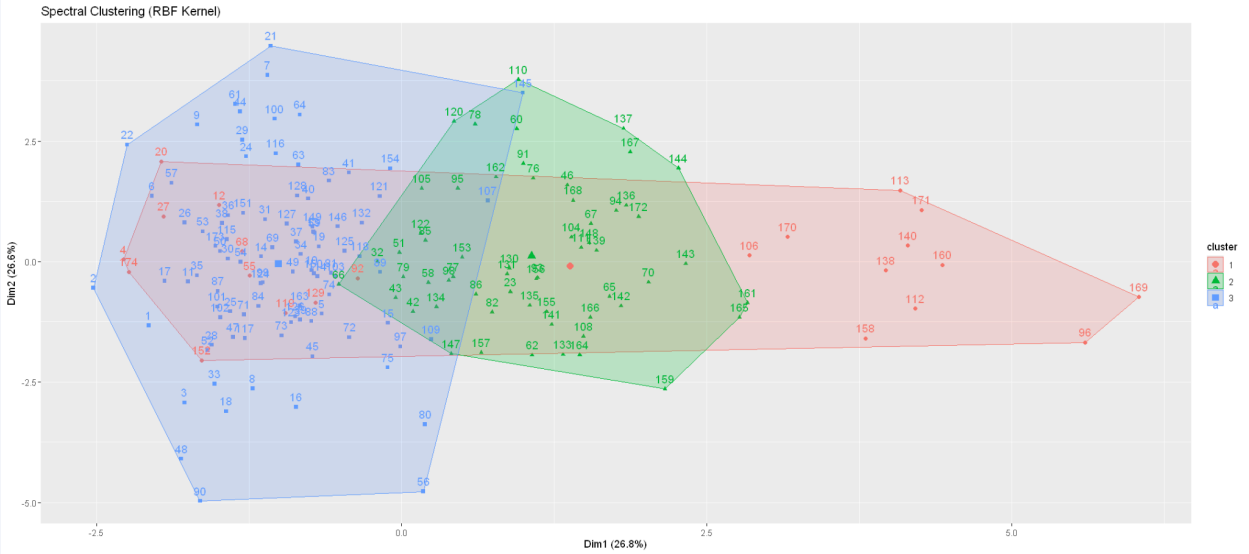
Spectral Clustering Visualization

The Spectral Clustering (RBF Kernel) visualization in Figure 29 provides a data-driven perspective on the socio-economic and public health disparities among Texas counties during the COVID-19 pandemic. The plot, segmented into three clusters—red, green, and blue—illustrates how counties group based on their pandemic response, demographics, and resource accessibility.

The red cluster (Cluster 1) represents counties with high poverty rates, large populations, and extensive reliance on public transportation, alongside elevated COVID-19 case and death rates. These counties are likely densely populated urban areas with significant socio-economic vulnerabilities, where the risk of virus transmission is heightened due to high population density and limited social distancing opportunities. The elongated shape of this cluster suggests some intra-cluster variability, potentially capturing counties at different stages of economic distress and pandemic severity.

In contrast, the green cluster (Cluster 2) consists of affluent suburban and semi-urban counties characterized by higher median and per capita income levels, lower dependence on public assistance, and a greater proportion of residents working from home. These counties exhibit a moderate to low COVID-19 impact, indicating that economic stability and remote work opportunities may have contributed to better pandemic resilience. The compactness of this cluster implies a consistent socio-economic profile across its members, reinforcing that these counties share similar pandemic outcomes and policy needs.

Figure 29: Spectral Clustering RPF Kernel



The blue cluster (Cluster 3) represents a more heterogeneous mix of rural and less populated counties, where poverty rates are moderate, reliance on public transportation is low, and work-from-home rates are relatively high. The spread-out nature of this cluster reflects a diverse set of counties, some potentially benefiting from geographic isolation as a natural buffer against COVID-19, while others may face healthcare access challenges that influence pandemic outcomes.

For the Texas Department of State Health Services (DSHS), these findings are instrumental in designing targeted interventions. Urban counties (Cluster 1) require enhanced medical resource allocation, vaccination outreach, and social support programs to mitigate the disproportionate impact of the pandemic. Affluent suburban areas (Cluster 2) may benefit more from policy incentives supporting continued remote work and localized containment strategies, while rural counties (Cluster 3) require infrastructure support, particularly in telehealth and emergency response readiness. By leveraging Spectral Clustering, DSHS can refine its COVID-19 response strategies, optimize resource distribution, and improve healthcare access equity, ensuring that each county receives interventions tailored to its unique socio-economic and public health profile.

External Variable Comparison: cases_per_100k & deaths_per_100k

The clustering results in Table 13 provide valuable insights for the Texas Department of State Health Services (DSHS) to consider when allocating resources and planning public health interventions across Texas counties. The analysis reveals three distinct clusters of counties, categorized by COVID-19 case and death rates, alongside various socio-economic factors that can guide resource distribution and intervention strategies.

Table 13: External Variable Comparison: cases_per_100k vs. deaths_per_100k

cluster	avg_cases_per_100k	avg_deaths_per_100k	n
1	7235.07	194.14	22
2	7759.99	194.01	59
3	7594.77	207.37	93

Cluster 1, consisting of 22 counties, is characterized by moderate COVID-19 case rates (7,235.07 cases per 100k) and relatively low death rates (194.14 deaths per 100k). These counties appear to be experiencing a lower overall impact from the pandemic compared to other clusters. As such, DSHS might consider monitoring this group closely while focusing resources on areas with more pressing needs. Health policies in these counties could emphasize maintaining the current situation with minimal intervention, ensuring that vulnerable populations remain protected without overwhelming local healthcare systems.

Cluster 2, made up of 59 counties, displays slightly higher case rates (7,759.99 cases per 100k) but similar death rates (194.01 deaths per 100k) to Cluster 1. This group represents a middle ground between the more severely impacted areas and those with relatively low rates. The DSHS could focus on implementing targeted interventions, such as increased testing, contact tracing, and public awareness campaigns, to prevent further escalation of the pandemic. With moderate case numbers, these counties would benefit from a balanced approach to transmission reduction while ensuring accessibility and equity in healthcare delivery.

Cluster 3, the largest group with 93 counties, stands out due to its high death rates (207.37 deaths per 100k), despite having slightly lower case rates (7,594.77 cases per 100k). This indicates that while the case load may not be the highest, the mortality rate in these counties is concerning, suggesting underlying health disparities such as limited healthcare access, higher rates of comorbidities, and inadequate public health infrastructure. For DSHS, this cluster should be prioritized for urgent intervention, focusing on enhancing healthcare services, deploying mobile health units, and initiating targeted outreach programs. Addressing the social determinants of health in these counties will also be critical in improving both immediate health outcomes and long-term resilience.

In conclusion, the clustering analysis provides a strategic framework for DSHS to allocate resources effectively and design tailored public health interventions. While Cluster 3 requires immediate attention due to its high mortality rate, Cluster 1 can be monitored with less urgency. Cluster 2, with its moderate impact, would benefit from targeted measures to curb the spread of the virus. By addressing the unique needs of each cluster, DSHS can improve the efficiency and effectiveness of its response to the COVID-19 crisis across Texas, ensuring that interventions are aligned with the specific challenges faced by different counties.

Demographic Breakdown By Cluster

Table 14 reveals three distinct groups of Texas counties based on COVID-19 case and death rates, as well as various socio-economic factors, which provide valuable insights for resource allocation and targeted public health interventions.

Table 14: Demographic Breakdown by Cluster

cluster	avg_cases_per_100k	avg_deaths_per_100k	avg_poverty	avg_median_income	avg_pct_on_food_stamps
1	11346.80	299.53	1606.39	51581.00	10.03
2	7042.30	188.41	1955.66	47395.86	12.79
3	7354.26	149.61	8083.11	50001.02	13.92

avg_commuters_by_public_transportation	avg_pct_work_from_home	avg_income_per_capita	avg_gini_index	avg_total_pop	n
15.38	1.80	23439.62	0.42	10397.46	13
7.09	3.19	23965.83	0.46	12687.21	116
48.40	3.50	24527.64	0.46	51663.51	45

Cluster 1, consisting of 13 counties, is characterized by the highest case rates (11,346.80 cases per 100k) and death rates (299.53 deaths per 100k). These counties also exhibit a moderate level of poverty (1,606.39) and a high reliance on public transportation (15.38%), with a relatively high median income of 51,581.00. The higher case and death rates in this cluster suggest that these counties may be urban areas or regions with significant socio-economic disparities, which contribute to higher virus transmission and more severe health outcomes. These areas may require intensive interventions to address the disproportionate impact of the pandemic.

Cluster 2, which includes 116 counties, shows moderate case rates (7,042.30 cases per 100k) and death rates (188.41 deaths per 100k). The counties in this cluster have slightly higher poverty levels (1,955.66) and lower reliance on public transportation (7.09%), with a median income of 47,395.86. These counties may represent suburban or semi-urban areas where the pandemic impact is more manageable. The moderate socio-economic challenges in these regions suggest that targeted interventions, such as increased testing and contact tracing, could help reduce the spread of the virus without requiring the level of intensive intervention seen in Cluster 1.

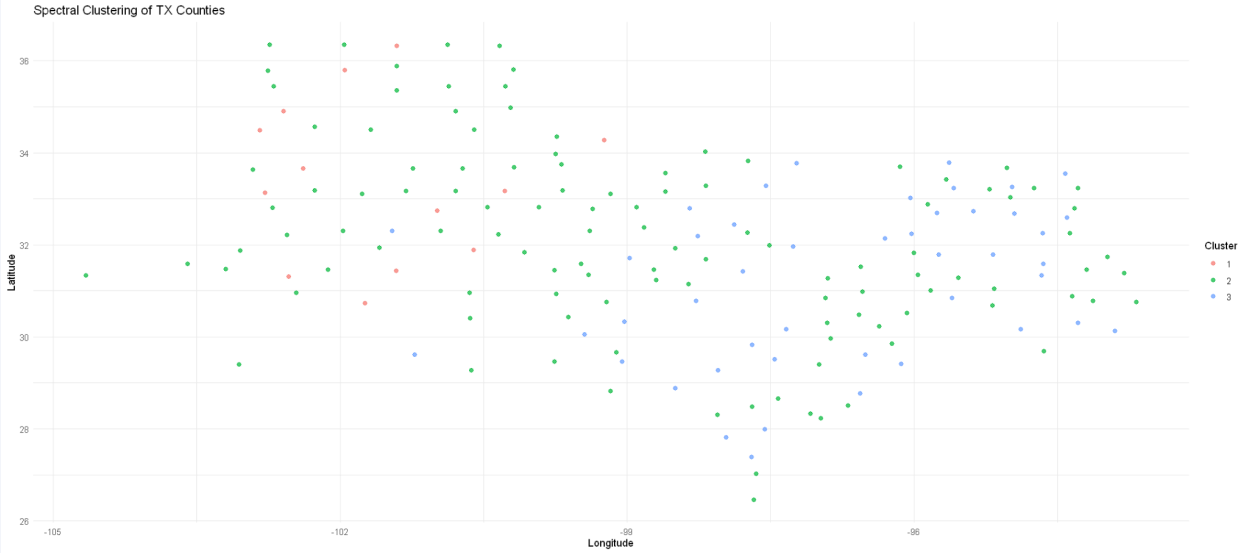
Cluster 3, consisting of 45 counties, has the lowest death rates (149.61 deaths per 100k) but slightly higher case rates (7,354.26 cases per 100k) compared to Cluster 2. These counties exhibit the highest poverty levels (8,083.11) and a very high reliance on public transportation (48.40%). With a median income of 50,001.02, these counties are likely rural or less developed, with limited access to healthcare and more significant socio-economic vulnerabilities. While these counties have a relatively lower death rate, their higher case rates and socio-economic challenges suggest a need for urgent public health interventions to improve healthcare access and reduce health disparities.

In summary, the clustering analysis provides a roadmap for strategically allocating resources and tailoring public health interventions. Cluster 1, with its higher death rates, requires immediate attention and intensive interventions. Cluster 2, with moderate case rates, could benefit from targeted efforts to prevent further spread. Cluster 3, despite having the lowest death rates, faces significant challenges related to healthcare access and socio-economic factors, requiring support in both health services and infrastructure.

Geographic Cluster Mapping

The spectral clustering visualization in Figure 30 of Texas counties provides a clear geographic representation of how counties group based on their COVID-19 cases, deaths, and socioeconomic factors. Each point on the map corresponds to a county, with colors indicating different clusters. The geographic distribution of these clusters suggests that counties with similar pandemic-related characteristics tend to be spatially correlated, though some deviations exist. Cluster 1, represented in red, appears sparsely distributed across the state, with a noticeable presence in the northwestern and central regions. This dispersion suggests that these counties may share distinct socioeconomic or pandemic response characteristics that differentiate them from the rest. Cluster 2, shown in green, dominates the map, covering a significant portion of Texas. The widespread distribution of this cluster implies that these counties share common socioeconomic conditions and pandemic experiences that are not as extreme as those observed in the other clusters. Cluster 3, marked in blue, is more concentrated in the eastern and central regions, with scattered representation in the western part of the state. The presence of this cluster in more urbanized areas suggests potential correlations with higher population density, increased reliance on public transportation, or different economic conditions that influenced the spread and severity of COVID-19.

Figure 30: Spectral Clustering of TX Counties



Given the selection of clustering variables—including COVID-19 cases and deaths per 100,000 people, poverty rates, median income, reliance on public transportation, and work-from-home percentages—the patterns observed in the map align with known demographic and economic trends in Texas. The counties in Cluster 1, which are more dispersed, may represent areas with unique socioeconomic conditions, such as high poverty rates, lower healthcare access, or rural characteristics that either exacerbated or mitigated the spread of COVID-19. These counties may have experienced either particularly high or particularly low pandemic impacts compared to other regions. In contrast, Cluster 2, with its widespread presence, likely represents counties that exhibit more typical or moderate COVID-19 trends relative to the state as a whole. The geographic diversity of this cluster suggests that these counties are not defined by extreme socioeconomic or pandemic-related conditions but rather reflect an average experience of the pandemic. Cluster 3, with its more concentrated presence in central and eastern Texas, aligns with regions containing major metropolitan areas and high-density communities. This suggests

that the counties in this cluster may have experienced higher case rates due to greater population mobility and closer economic ties to urban centers.

For the Texas Department of State Health Services (DSHS), this clustering analysis provides valuable insights into how different regions experienced the pandemic and how public health resources can be more effectively allocated. The counties in Cluster 1, with their dispersed but unique characteristics, may require targeted intervention strategies that address specific socioeconomic vulnerabilities or healthcare limitations. The widespread counties in Cluster 2 can serve as a baseline for understanding typical pandemic patterns across Texas, providing a comparative framework for assessing deviations in other clusters. The counties in Cluster 3, particularly those in the eastern and central portions of the state, may benefit from policies that address urban pandemic dynamics, such as expanded healthcare access, increased testing availability, and targeted economic support measures. By understanding how socioeconomic and geographic factors influenced COVID-19 outcomes, policymakers can use these findings to guide resource allocation, refine public health strategies, and better prepare for future public health crises. The clustering analysis serves as a crucial tool in identifying disparities and ensuring that interventions are tailored to the specific needs of different regions across Texas.

Health Outcomes By Cluster

Figures 31 and 32 below visualize cases per 100,000 by spectral cluster and deaths per 100,000 by spectral cluster, respectively. Cluster 1 includes counties characterized by low public transit use, lower population density, and moderate poverty levels. These counties exhibit lower COVID-19 case and death rates, which aligns with typical rural trends where lower population density and fewer opportunities for transmission may limit the virus's spread. Cluster 2, on the other hand, captures urban counties with high public transit usage and lower income levels. These counties show elevated COVID-19 case rates and moderate death rates, indicating that high population density and greater reliance on public transportation may contribute to higher virus transmission, especially in lower-income areas where access to healthcare and other resources might be limited. Cluster 3 consists of suburban counties with moderate income levels. These counties show slightly above-average outcomes in terms of COVID-19 cases and deaths, with a more diverse demographic mix, suggesting that socioeconomic factors like income and access to resources play a role in shaping pandemic outcomes, even in less densely populated areas.

Figure 31: Cases per 100k by Spectral Cluster

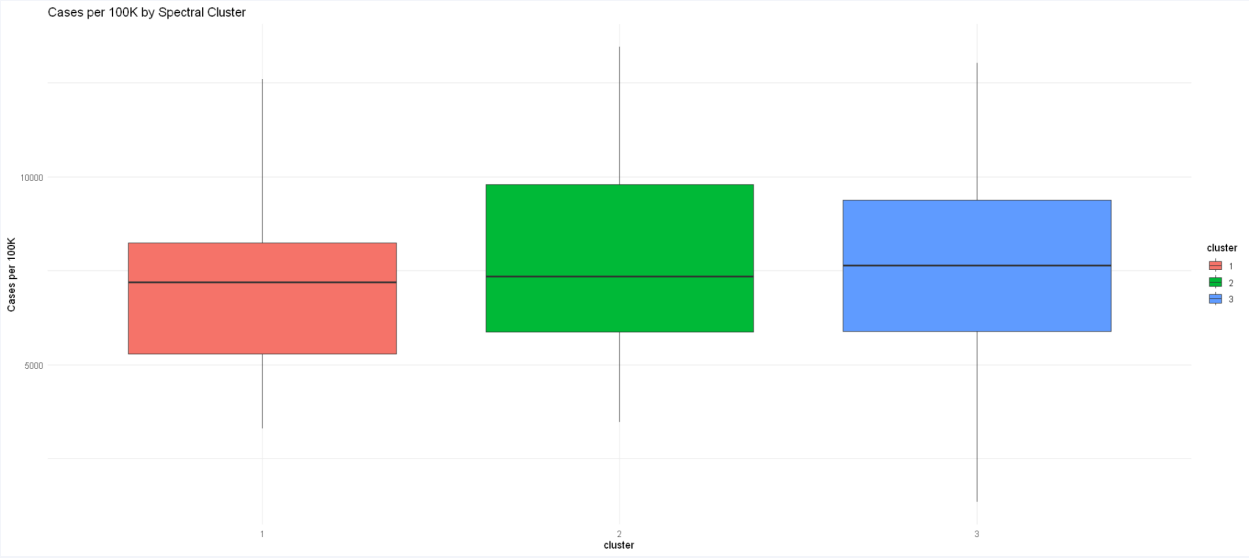
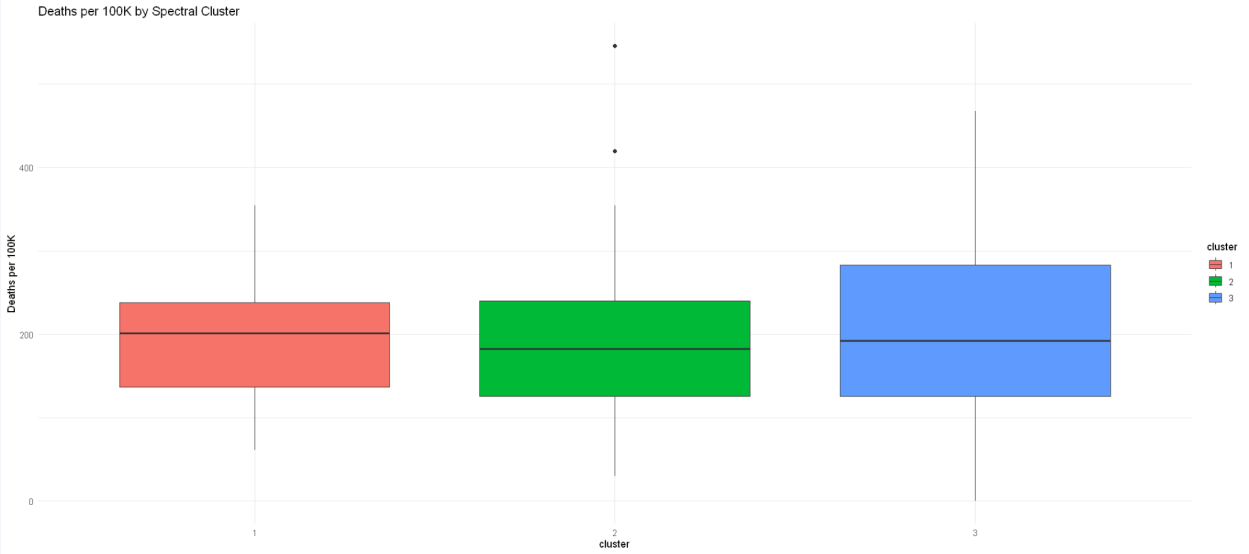


Figure 32: Deaths per 100k by Spectral Cluster



Spectral clustering, with its ability to segment counties based on nonlinear similarities, has proven to be an effective tool for uncovering patterns that traditional methods like k-means and DBSCAN might miss. By focusing on complex, non-convex groupings, spectral clustering has revealed latent structures in the data that are associated with socioeconomic factors, public transit use, and mobility patterns. These insights suggest that the spread and impact of COVID-19 are influenced by a combination of geographic, economic, and behavioral factors, which are not always captured by more simplistic clustering approaches. This method allows for a more nuanced understanding of county-level dynamics, highlighting subtle groupings that might otherwise go unnoticed.

For the Texas Department of State Health Services (DSHS), these findings can be extremely valuable in designing targeted interventions. Spectral clustering offers a more detailed view of how different regions in Texas are impacted by the pandemic, which can help DSHS better allocate resources and tailor public health policies. For example, counties in Cluster 1, which may be more rural and exhibit lower case rates, might require different strategies compared to Cluster 2 counties, where urbanization and high transit usage could lead to higher transmission rates. DSHS could develop messaging and interventions specific to the mobility patterns and healthcare access disparities observed in each cluster.

Additionally, public health agencies can use these insights to refine their communication strategies. For example, urban counties in Cluster 2 might benefit from targeted outreach and information campaigns focusing on social distancing and vaccination, considering the high mobility and density of these areas. For rural counties in Cluster 1, the messaging might emphasize maintaining protective measures despite lower case rates, focusing on vulnerable populations who may face greater barriers to healthcare.

Transportation agencies also have an opportunity to collaborate with public health officials based on these findings. Counties with high public transit usage, especially in Cluster 2, may require more intensive efforts to promote public health measures like mask-wearing or vaccination in transit hubs or on buses. Furthermore, identifying outlier regions or counties that

don't follow traditional geographic or demographic patterns could help DSHS address areas that may otherwise be overlooked in typical planning models. By using spectral clustering, DSHS can develop a more customized and effective COVID-19 response strategy, ensuring that resources are allocated in a way that reflects the unique challenges faced by different communities across Texas.

Summary of Findings

The spectral clustering analysis revealed three distinct clusters of Texas counties, each exhibiting unique COVID-19 outcomes and socio-economic profiles. By identifying complex, non-linear groupings based on variables like case and death rates, poverty levels, transportation use, and work-from-home prevalence, the method uncovered nuanced relationships that traditional clustering techniques might overlook. These insights provide the Texas Department of State Health Services (DSHS) with a powerful tool for tailoring public health interventions. Cluster 1, with higher case and death rates and heavy reliance on public transit, highlights urban areas needing urgent healthcare and outreach efforts. Cluster 2 represents moderate-impact suburban areas where continued testing and education may suffice, while Cluster 3 consists of rural regions with limited healthcare access, requiring infrastructure support and targeted health services. This data-driven approach enables DSHS to optimize resource allocation, enhance healthcare equity, and implement geographically and demographically tailored responses for more effective pandemic management.

4.8.2 Agglomerative Clustering

Agglomerative clustering is a bottom-up hierarchical method where each data point initially starts in its own cluster, and pairs of clusters are merged as we move up the hierarchy. In this analysis, we employ **average linkage**, where the distance between clusters is calculated as the average pairwise distance between points in the two clusters. This approach strikes a balance between the sensitivity of single linkage (which may result in "chaining" of clusters) and complete linkage (which tends to favor compact, spherical clusters). The purpose of using agglomerative clustering is to test the robustness of cluster structures under different hierarchical linkage criteria. Compared to Ward clustering, which minimizes variance within clusters, average linkage is more adaptable to recognizing elongated or variably shaped clusters. This can uncover subtle or nuanced relationships in county-level COVID-19 demographics that might otherwise remain hidden.

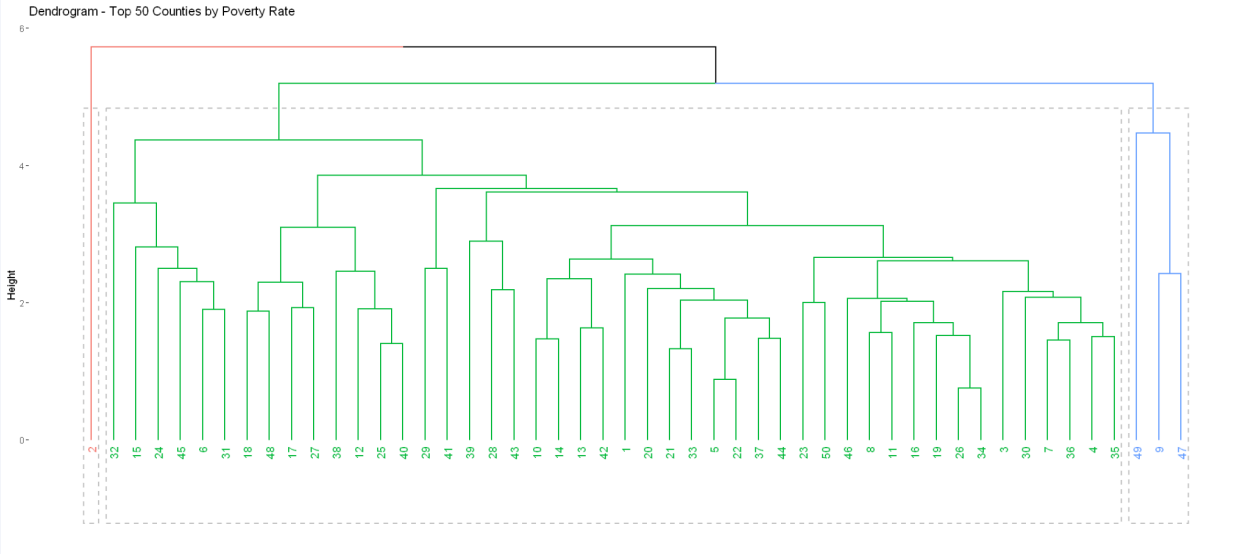
The method uses Euclidean distance on standardized variables to assess similarities between data points, with a predefined number of clusters set to 3 to align with other clustering methods such as k-means and DBSCAN. A key aspect of this approach is comparing the resulting clusters with external variables, specifically COVID-19 outcomes such as `cases_per_100K` and `deaths_per_100K`. This allows for an assessment of whether the clusters correspond to epidemiologically meaningful groups, providing insight into how demographic clusters align with actual pandemic outcomes.

Overall, agglomerative clustering offers a more flexible approach to hierarchical clustering compared to other methods, making it useful for identifying complex patterns and relationships in data. By revealing clusters with different shapes and densities, this method adds depth to the understanding of how various factors may have influenced the COVID-19 response across different regions.

Subsetting for Top 50 Counties by Poverty Rate

The dendrogram presented in Figure 33 illustrates the hierarchical clustering of the top fifty Texas counties with the highest poverty rates, using the average linkage method. This visualization is particularly relevant for the Texas Department of State Health Services (DSHS), as it provides insights into how counties with severe economic distress relate to one another based on their socioeconomic and COVID-19 impact characteristics. The hierarchical clustering approach arranges these counties in a tree-like structure, where the vertical axis represents the distance or dissimilarity between counties, measured through standardized feature values. The greater the height at which two branches merge, the more distinct those counties are in terms of their underlying data.

Figure 33: Top 50 Counties by Poverty Rate



Three primary clusters have been identified and are distinguished using different colors. The first cluster (red) consists of a single county, indicating that it is significantly different from the others based on the clustering criteria. The second cluster (green) contains three counties that are relatively similar to each other but distinct from the remaining counties. The third cluster (blue) is the largest, encompassing multiple counties that share more similarities in their demographic and economic conditions. The hierarchical structure suggests that while some counties are closely related in terms of their poverty and COVID-19 impact profiles, others exhibit stark differences that warrant separate classification.

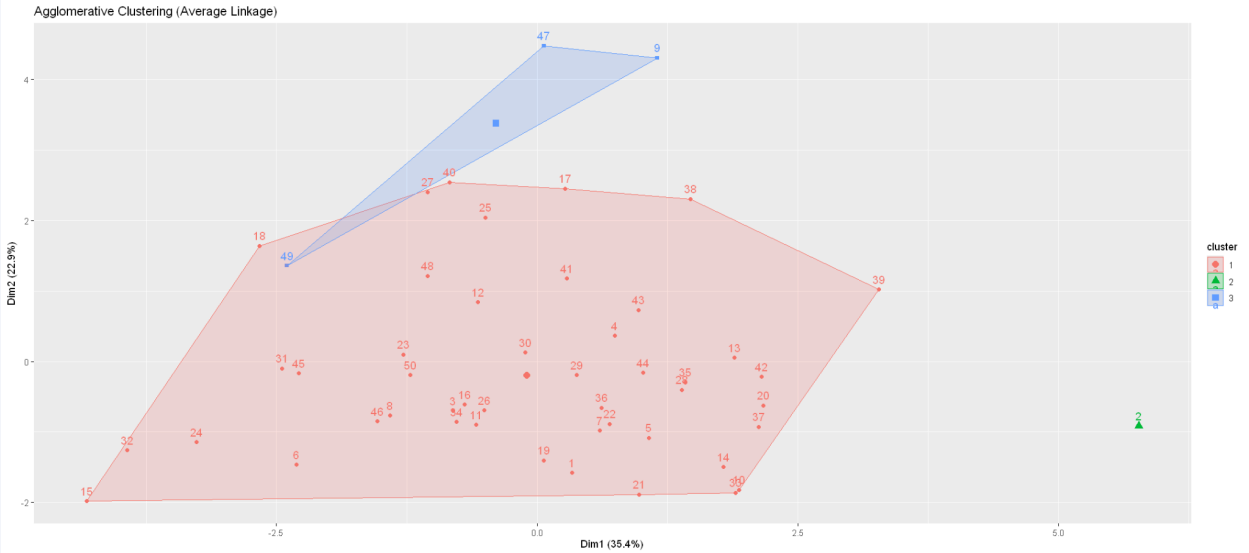
For DSHS, this clustering analysis is essential in identifying counties that may require differentiated public health interventions. Counties grouped closely together may benefit from similar policy approaches, while those forming distinct clusters might require targeted strategies to address their unique challenges. By leveraging hierarchical clustering, policymakers can better understand regional disparities and allocate resources effectively to mitigate the adverse effects of poverty on COVID-19 outcomes.

Agglomerative Clustering Visualization

Figure 34 presents the results of an Agglomerative Clustering analysis using the Average Linkage method to group the fifty Texas counties with the highest poverty rates based on socioeconomic and COVID-19 impact factors. The plot identifies three distinct clusters, each represented by a different color, highlighting patterns of similarity and difference among the counties. The largest cluster, shown in red, encompasses the majority of the counties,

suggesting a shared set of socioeconomic characteristics and public health challenges. This dense grouping indicates that many of these counties exhibit similar poverty-related conditions and may benefit from comparable policy interventions.

Figure 34: Agglomerative Clustering for Top 50 Counties



In contrast, a smaller cluster, depicted in blue, includes a few counties that, while related, display distinct differences from the larger group. These counties may have unique demographic or economic conditions that set them apart from the primary cluster, necessitating a more tailored approach in public health planning. Additionally, a single county, represented in green, is positioned far from the other clusters, signaling a significant deviation from the rest in terms of its socioeconomic and health-related indicators. The separation of this outlier county suggests that it may require highly specific policy considerations that do not align with the broader trends observed in the other clusters.

The convex hulls surrounding each cluster illustrate the spatial dispersion of the counties within their respective groups. The red cluster appears compact, reinforcing the high degree of similarity among its members, whereas the blue cluster is more dispersed, indicating a broader range of variation. The green cluster, consisting of only one county, remains completely isolated, further emphasizing its unique status. The two principal component dimensions explain a combined 58.3% of the data's variance, meaning that the clustering captures a substantial portion of the variation in socioeconomic and COVID-19 impact factors among these counties.

From a public health policy perspective, this clustering analysis offers valuable insights for the Texas Department of State Health Services in designing targeted interventions. The strong cohesion within the red cluster suggests that a uniform strategy could be effective in addressing the needs of most high-poverty counties. However, the counties in the blue cluster, while still related, may require more nuanced approaches based on their distinct socioeconomic conditions. The outlier county in green presents a particularly complex case, as its separation from the other counties indicates that standard policy measures may not be sufficient to address its specific challenges. Understanding these clusters allows for more efficient resource allocation and the development of interventions that reflect the diverse needs of Texas counties experiencing high poverty rates and their associated public health burdens.

Health Outcomes by Cluster

Among the Texas counties with the highest poverty rates, clear and consequential differences in COVID-19 outcomes emerge when grouped using cluster analysis based on shared demographic and socioeconomic characteristics as shown in Table 15. This more nuanced lens reveals that poverty alone does not fully explain variations in disease burden.

Table 15: Health Outcomes By Cluster

cluster	avg_cases_per_100k	avg_deaths_per_100k	n
1	7122.31	148.23	46
2	9020.56	315.95	1
3	6202.85	142.62	3
NA	7968.97	193.92	204

Cluster 2—consisting of just a single county—exhibits the most alarming statistics, with the highest average case rate (9,021 per 100,000 residents) and a staggering death rate (316 per 100,000). This outlier likely reflects a convergence of extreme structural vulnerabilities—such as limited healthcare infrastructure, high comorbidity prevalence, or systemic barriers to accessing care—that dramatically amplified the impact of the pandemic.

In contrast, Cluster 1, which includes 46 high-poverty counties, shows significantly more moderate outcomes. These counties reported an average of 7,122 cases and 148 deaths per 100,000—more than 50% lower in mortality compared to Cluster 2. This suggests that while poverty is a common thread, other contextual factors—like better healthcare access, public health outreach, or community resilience—may be mitigating outcomes in these areas.

Cluster 3, comprising three counties, stands out with the lowest rates of both cases (6,203 per 100,000) and deaths (143 per 100,000) among all clustered groups. This indicates that even in high-poverty areas, certain protective factors—such as lower population density, younger population profiles, or strong local health partnerships—may have played a crucial role in reducing disease spread and mortality. The remaining 204 counties, which were not included in the clustering due to missing or unmatched data, had higher average rates of both cases and deaths compared to Clusters 1 and 3, but still well below the extreme levels seen in Cluster 2.

For the Texas Department of State Health Services (DSHS), these findings offer critical, actionable insights. Rather than applying a uniform strategy to all high-poverty regions, this cluster-based analysis reveals substantial variation in vulnerability and health outcomes—suggesting that a more tailored, precision-targeted approach is warranted.
For example, the outlier in Cluster 2 represents a priority zone for intensified public health support, such as surge testing, vaccine drives, mobile health clinics, and infrastructure investment. Meanwhile, the more moderate outcomes in Cluster 1 may indicate that existing strategies are having a positive effect, though continued monitoring and targeted improvements are still necessary. Cluster 3, on the other hand, may hold lessons worth replicating—highlighting the importance of identifying and scaling successful local interventions.

Ultimately, this analysis enhances DSHS's ability to allocate resources more efficiently, focusing attention where it is most urgently needed. It underscores the value of data-driven, localized

decision-making and encourages the department to go beyond poverty indicators alone—taking into account the full constellation of risk and resilience factors to guide both immediate response efforts and long-term public health planning.

Demographic Characteristics by Cluster

Table 16 displays demographic characteristics by cluster and reveals nuanced distinctions between the clusters that are important for understanding how socio-economic factors might influence public health outcomes, especially in the context of the COVID-19 pandemic. In Cluster 1, the average public transit rate of 32.30% stands out, suggesting that a significant portion of the population in these counties relies on public transportation. This could be indicative of urbanized areas with more developed infrastructure, likely providing better access to healthcare services and other essential facilities. Coupled with a moderate average poverty rate of 7698.87 and a relatively high median income of 48748.89, it is reasonable to assume that counties in this cluster have better economic resources to invest in healthcare and public health initiatives, which could contribute to more effective responses to public health crises like COVID-19. These areas might also have greater public health coverage, which is crucial for limiting the spread of infectious diseases.

Table 16: Demographic Characteristics By Cluster

cluster	avg_public_transit	avg_poverty_rate	avg_median_income
1	32.30	7698.87	48748.89
2	0	7157	29104.00
3	121.33	13382	47873.33

On the other hand, Cluster 2 presents a stark contrast. With an average public transit rate of 0%, counties in this cluster likely face significant transportation challenges, which may result in less access to healthcare facilities and hinder the delivery of healthcare services. The average median income of 29104.00 is notably lower than that of Cluster 1, and the poverty rate of 7157.00, while lower than Cluster 3, still indicates economic hardship. These factors suggest that counties in Cluster 2 may struggle with financial constraints, making it more difficult to implement effective health interventions or provide adequate healthcare coverage for residents. This socioeconomic vulnerability could be compounded by the lack of public transit, making it harder for residents to access vaccination sites, clinics, and hospitals, potentially leading to poorer health outcomes and higher COVID-19 transmission rates.

Cluster 3, with its notably high public transit rate of 121.33%, points to urban areas where infrastructure is likely well-developed, and transportation access is a significant factor in public health. However, despite this, the average poverty rate of 13382.00 is the highest among the clusters, signaling that, while there is access to public transit, economic disparities remain stark. This might suggest a situation where, despite improved infrastructure, a large proportion of the population is still financially vulnerable. The median income of 47873.33 in Cluster 3 is lower than that of Cluster 1, and the higher poverty rate could indicate that many residents are living in precarious economic conditions. These areas, while well-connected in terms of transportation, may still face challenges in providing adequate healthcare, especially if the public health system is overwhelmed by the socioeconomic conditions.

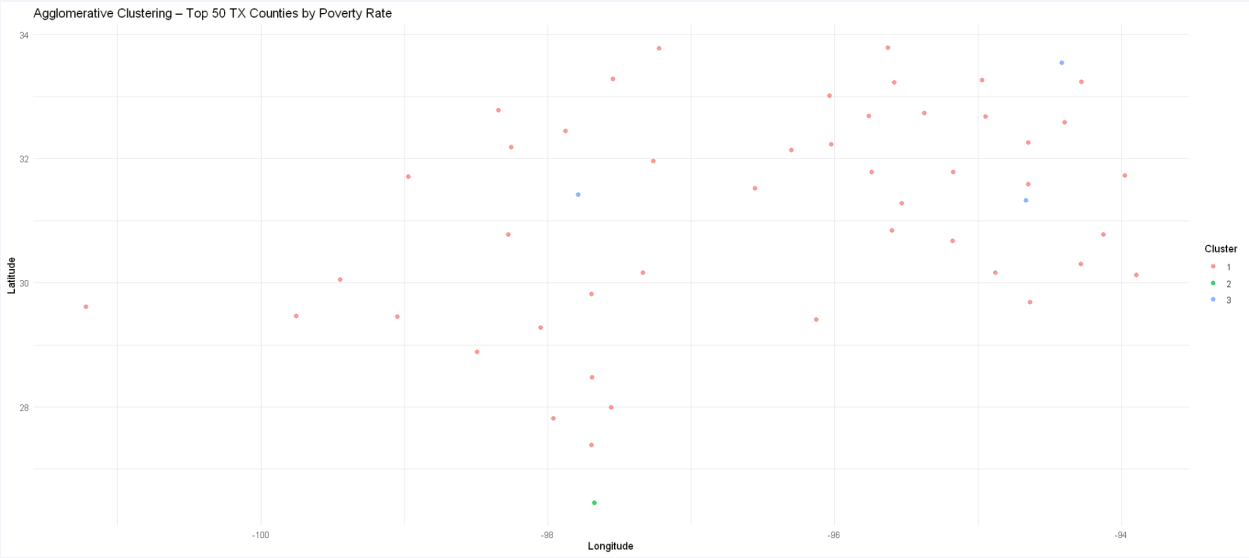
For the Texas Department of State Health Services (DSHS), these findings offer critical insights into how demographic factors interact with public health outcomes, especially in the context of

the COVID-19 response. Cluster 1, with its higher median income and better access to public transit, might require fewer resources in terms of basic healthcare accessibility. However, focusing resources on addressing health equity in Cluster 2, with its transportation and economic challenges, could help mitigate disparities and improve access to essential services. Cluster 3, despite its good transportation infrastructure, might benefit from policies that focus on addressing economic inequalities, as the high poverty rate suggests that many residents may face barriers to healthcare despite having access to transportation. By understanding these dynamics, DSHS can allocate resources more strategically, focusing on the most vulnerable populations, and develop targeted interventions that address the specific needs of each cluster. This approach would ultimately help reduce health disparities and improve COVID-19 outcomes across Texas.

Agglomerative Clustering - Texas Counties

The geospatial plot of the top 50 Texas counties by poverty rate, grouped using agglomerative clustering in Figure 35, reveals important regional distinctions that enhance the understanding of how poverty intersects with public health vulnerability. Each point on the plot represents a county, positioned by its geographic centroid (longitude and latitude), and color-coded according to its cluster membership. The clustering was derived from a combination of COVID-19 outcomes and socioeconomic factors, providing a multidimensional framework for assessing risk. The spatial distribution of counties across the three clusters shows that while poverty is the shared characteristic among all 50 counties, meaningful differences emerge when geography and additional contextual variables are considered.

Figure 35: Agglomerative Clustering of Texas Counties



Cluster 1, the largest group, is represented in red and includes the vast majority of counties in the sample. These counties are dispersed throughout the state, particularly across central, eastern, and southern Texas. Their widespread nature suggests that moderate COVID-19 outcomes—such as mid-range case and death rates—are not confined to a specific region but are instead found in diverse geographic settings. Counties in this cluster typically exhibit moderate poverty levels, access to public transportation, and median income levels that, while lower than the state average, may support more resilient public health responses. The dispersion and demographic profile of Cluster 1 suggest that these counties may be benefiting from relatively better local healthcare infrastructure, transportation systems that facilitate access

to care, and possibly stronger public health partnerships. For DSHS, these counties may not require immediate crisis-level intervention but would benefit from continued support to maintain and improve existing systems.
In stark contrast, Cluster 2, marked in green, consists of a single county located in the southern part of the state. This county is a clear outlier in terms of both COVID-19 burden and geographic location. Previous analyses identified it as having the highest rates of both cases and deaths per 100,000 residents among all clusters. Its position as the sole member of its cluster and its geographic isolation underscore the severity of its structural vulnerabilities. These may include poor access to healthcare facilities, limited transportation infrastructure, high population density, or other systemic inequities. For DSHS, this county represents a critical intervention point. Its unique combination of socioeconomic disadvantage and poor health outcomes highlights an urgent need for concentrated public health resources, such as mobile clinics, vaccination drives, health education campaigns, and potentially long-term infrastructure investment. Identifying this outlier through cluster analysis allows DSHS to prioritize high-impact interventions that could significantly reduce health disparities in the state's most at-risk areas.

Cluster 3, depicted in blue, contains three counties situated in northeastern and east-central Texas. These counties are relatively close to one another geographically, which may indicate shared regional characteristics contributing to their unique cluster identity. Notably, counties in this cluster reported the lowest COVID-19 case and death rates despite high poverty levels. Their demographic profile—highlighted by high public transit usage and moderate income levels—suggests the presence of protective factors that may include lower population density, effective local governance, or robust public health systems. From a policy standpoint, these counties offer valuable insights into what is working under challenging economic conditions. DSHS can study these areas more closely to identify replicable best practices, which could then be adapted for use in more vulnerable counties within Clusters 1 and 2.

Overall, the geospatial clustering analysis offers DSHS a data-driven foundation for implementing a more nuanced and localized public health strategy. Rather than deploying uniform policies across all high-poverty counties, this approach highlights the importance of tailoring interventions based on both socioeconomic and geographic context. By focusing efforts where they are most urgently needed—such as the outlier in Cluster 2—while also reinforcing what works in more resilient counties, DSHS can allocate resources more effectively and equitably. This precision-targeted framework not only enhances the efficiency of current COVID-19 mitigation strategies but also lays the groundwork for more informed public health planning in future emergencies.

The boxplots of COVID-19 cases and deaths per 100,000 residents, segmented by agglomerative clustering, provide critical insights into how different groups of Texas counties experienced the pandemic. These visualizations play a central role in identifying patterns that can guide the Texas Department of State Health Services (DSHS) in making data-driven decisions to target future public health interventions more effectively.

Figure 36: Cases per 100k by Agglomerative Clustering

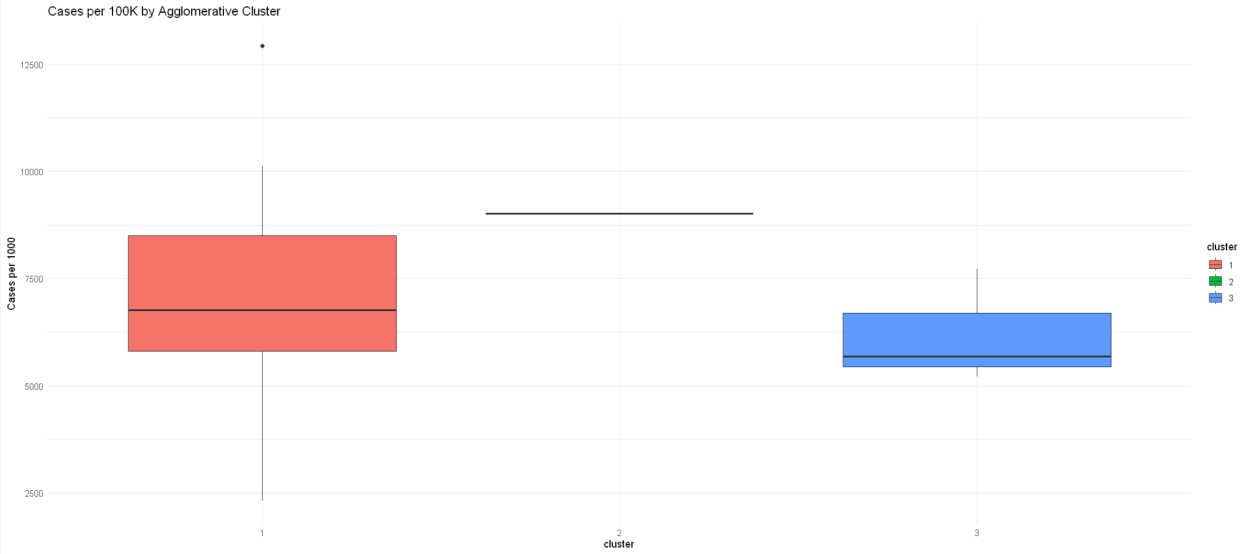
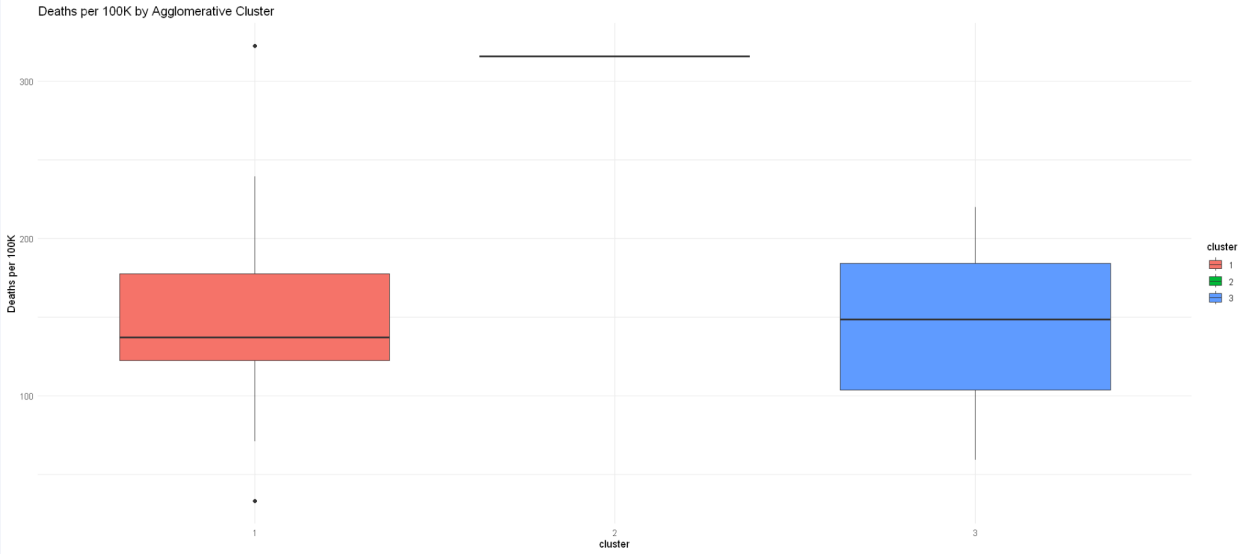


Figure 36 illustrates the distribution of COVID-19 cases per 100,000 population across the three agglomerative clusters. Cluster 1, represented in red, displays a wide range of case rates, with a high median and an extended interquartile range, indicating significant variability in case counts among its counties. This suggests that counties within this group may have experienced inconsistent levels of exposure or varying public health responses, possibly due to differences in population density, healthcare access, or mobility patterns. In contrast, Cluster 3, shown in blue, has a lower median and a more compact interquartile range, indicating more uniform and generally lower case rates among its counties. Cluster 2, with only a single value shown as a flat line, likely represents a very small group or an outlier county with a high case rate, and its presence warrants further inspection to understand its unique circumstances.

Figure 37 presents deaths per 100,000 by cluster. Similar to the case rate distribution, Cluster 1 again shows greater variability, with a median that is slightly lower than Cluster 3 but with several counties exhibiting higher death rates, as indicated by the wider range and presence of outliers. Cluster 3 has a higher median death rate and tighter clustering, suggesting a consistently elevated mortality burden in these counties. Cluster 2 once again appears as a flat line, implying a singular or extreme case.

Figure 37: Deaths per 100k by Agglomerative Clustering



Together, these boxplots underscore distinct public health profiles among the clusters. Cluster 1 appears to consist of counties with volatile case and death rates, possibly reflecting mixed demographic and socioeconomic characteristics. Cluster 3, by contrast, seems to represent counties with more consistent but concerning levels of mortality, even when case rates are relatively lower. Cluster 2, although not generalizable due to its small size, stands out and may represent a unique public health context such as a large urban county or a county with a severe outbreak.

For the DSHS, these visualizations provide a powerful tool for targeting resources. Understanding which clusters experience consistently higher mortality, or which are more volatile in case numbers, allows the agency to allocate healthcare infrastructure, vaccination campaigns, and outreach efforts more effectively. For example, Cluster 3 counties may benefit from increased healthcare staffing or mortality-prevention programs, while Cluster 1 counties might require a broader strategy that addresses the diverse underlying drivers of case and death variability. These insights make it possible to implement tailored responses rather than relying on one-size-fits-all approaches.

Summary of Findings

The agglomerative clustering analysis of Texas counties with the highest poverty rates reveals three distinct demographic clusters with meaningful differences in COVID-19 outcomes and socioeconomic conditions. Cluster 1, encompassing the majority of counties, displays moderate health outcomes and relatively better infrastructure, indicating that supportive interventions may be effective if sustained. Cluster 2, an outlier county with the highest case and death rates, exhibits extreme vulnerability likely due to compounded structural disadvantages, signaling an urgent need for concentrated public health investment. Cluster 3, comprising three counties, demonstrates relatively low case and death rates despite high poverty, suggesting the presence of protective factors worth replicating elsewhere. For the Texas Department of State Health Services (DSHS), these insights enable a more precise allocation of resources, supporting a shift from broad-stroke policies to targeted, context-aware strategies. By leveraging these clusters, DSHS can design and implement interventions tailored to each group's unique

challenges and strengths, ultimately improving health equity, pandemic response efficiency, and long-term public health resilience across the state.

4.9 Graduate Level Analysis by Leonardo Piedrahita

Advanced Clustering Techniques:
Ordering Points to Identify the Clustering Structure (OPTICS)
Mean Shift Clustering

To dive deeper into the complexities of the Texas county-level COVID-19 data, I turned to two alternative clustering algorithms: **OPTICS (Ordering Points to Identify the Clustering Structure)** and **Mean Shift Clustering**. These methods were chosen specifically for their ability to adapt to data with varied densities and more irregular shapes, which are characteristics often observed in public health datasets that do not conform to traditional clustering assumptions. **OPTICS** stands out as a density-based approach that does not require setting a fixed number of clusters beforehand. Instead, it organizes the data into a reachability plot, providing a comprehensive view of the underlying structure of the data. By examining this plot, I was able to identify clusters of different densities, which is particularly valuable for a dataset like ours where some counties may show dense groupings while others are more spread out. This method is powerful because it allows the identification of both dense regions and more isolated areas, without forcing a pre-set number of clusters. As a result, OPTICS offers a more dynamic view of the data, revealing complex spatial relationships that may be overlooked by simpler methods like K-means or hierarchical clustering.

In parallel, **Mean Shift Clustering** offers a completely different perspective on the data. Rather than assuming clusters follow a standard shape, Mean Shift works by iteratively shifting data points towards areas of higher density, eventually identifying regions of the data that represent significant modes. This method is highly flexible and can uncover clusters of arbitrary shape, making it especially useful for datasets where the underlying clusters are not spherical or evenly distributed. In the context of COVID-19 data, this algorithm is ideal for identifying regions where the pandemic impact might manifest in less uniform patterns, potentially highlighting specific counties or groups with unusual socio-economic or health characteristics.

By integrating both OPTICS and Mean Shift Clustering into the analysis, the goal was to determine whether these more sophisticated methods could uncover new insights or provide a clearer picture of how Texas counties are grouped based on their COVID-19 and socio-economic features. These algorithms provide a layer of depth that goes beyond the traditional clustering methods, enabling a more nuanced understanding of the data. For the Texas Department of State Health Services (DSHS), these insights can be especially valuable for targeted interventions, especially in areas where the COVID-19 impact may not follow clear, predefined boundaries. The ability to adapt to varying densities and shapes of data makes these clustering techniques particularly suited for identifying hidden or subtle patterns in public health data.

4.9.1 OPTICS Clustering

OPTICS (Ordering Points To Identify the Clustering Structure) is a density-based clustering algorithm that builds on the principles of DBSCAN but offers a more flexible approach by generating an augmented ordering of data points rather than a fixed clustering based on a single epsilon threshold. This method produces a reachability plot, which visually represents the hierarchical density structure of the dataset. OPTICS is particularly effective for datasets with

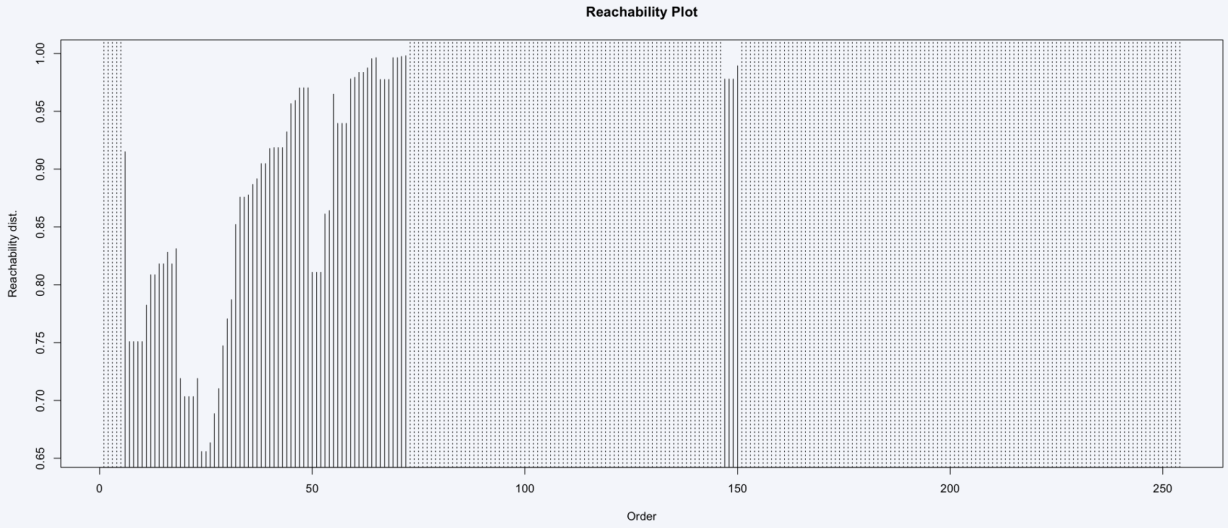
varying densities, as it can identify clusters at multiple scales without requiring a predefined number of clusters. Additionally, it excels at detecting noise points—data points that do not belong to any dense cluster—making it ideal for datasets with outliers or sparse regions, such as county-level COVID-19 data where rural counties may differ significantly from urban ones in density and characteristics.

In this analysis, OPTICS was applied to the scaled dataset of Texas counties, using the same 10 features as previous clustering methods: cases_per_100k, deaths_per_100k, poverty, median_income, pct_on_food_stamps, commuters_by_public_transportation, pct_work_from_home, income_per_capita, gini_index, and total_pop. The algorithm was configured with a minPts value of 2 to define the minimum number of points required to form a core point, and an eps value of 1.0 to set the reachability distance range for the plot. Clusters were then extracted using a clustering epsilon threshold (eps_cl) of 0.5, allowing us to identify dense groupings while labeling sparse regions as noise. This approach enables a detailed exploration of the density-based structure of Texas counties, providing insights into both densely clustered regions and isolated counties that may require unique public health strategies.

4.9.1.1 Reachability Plot Analysis

The reachability plot, a key output of the OPTICS algorithm, visualizes the density structure of the dataset by ordering data points along the x-axis and plotting their reachability distances on the y-axis. A lower reachability distance indicates that a point is closely connected to a dense region, while higher distances suggest sparser areas or noise. Figure 38 displays the reachability plot for the cleaned dataset after outlier removal.

Figure 38: Reachability Plot for OPTICS Clustering



The reachability plot for the cleaned dataset reveals two prominent valleys, indicating the presence of at least two dense clusters. These valleys correspond to regions where counties share similar characteristics, forming natural groupings based on density. Between these valleys, a wide flat region with higher reachability distances suggests a significant number of counties that do not belong to any dense cluster, likely classified as noise under the chosen eps_cl threshold of 0.5. The plot also shows a sharp vertical spike around index 150, marking a transition from dense clusters to more isolated or less connected counties. This structure highlights the heterogeneity in the dataset, with some counties forming tight, dense groups while others are more sparsely distributed, potentially due to rural isolation or unique socioeconomic profiles.

Cluster Size Summary
The distribution of counties across OPTICS clusters, as summarized in Table 17, further illustrates the algorithm's findings. Only one primary cluster (Cluster 0) was identified, containing all 174 counties in the cleaned dataset, with no counties assigned to distinct clusters (Clusters 1, 2, etc.) and no points labeled as noise (Cluster 0 in the typical OPTICS sense).

Table 17: Cluster Sizes from OPTICS Clustering

Cluster	*Number of Counties*
0	*174*
1	*0*
2	*0*

This outcome suggests that the current eps_cl threshold may be too high, causing OPTICS to group all counties into a single cluster without distinguishing denser subgroups or identifying noise points. It indicates a lack of sufficient density variation within the dataset under the chosen parameters, potentially due to the uniformity of the cleaned data after outlier removal. To uncover more granular clusters, adjusting the eps_cl value to a lower threshold or increasing minPts could help reveal smaller, denser groupings. Despite this, the reachability plot still provides valuable insight into the data's density structure, highlighting the potential for hierarchical clustering at different scales.

4.9.1.2 Demographic Attributes by OPTICS Cluster
To gain a deeper understanding of the socioeconomic characteristics of the counties grouped by the OPTICS algorithm, we calculated summary statistics for key demographic attributes within the identified cluster. Since the OPTICS configuration resulted in a single cluster (Cluster 0) containing all 174 counties, the analysis focuses on this group. The following metrics were examined to assess the socioeconomic profile of the counties:

- **Public Transit Usage** (commuters_by_public_transportation): Reflects the percentage of residents commuting via public transportation, indicating urbanization and potential exposure risks.
- **Poverty Rate** (poverty): Measures the number of individuals living below the poverty line per 100,000 residents, serving as a key indicator of economic vulnerability.
- **Median Income** (median_income): Represents the annual median household income, providing insight into the economic standing of the counties.

The summary statistics for Cluster 0 are presented in Table 18 below.

Table 18: Demographic Attributes by OPTICS Cluster

Cluster	AVG. Public Transit Usage (%)	AVG. Poverty Rate (per 100,000)	AVG. Median Income ($)	Number of Counties (n)
0	18.4	3,514.25	48,382.29	174

Cluster 0, encompassing all 174 counties in the cleaned dataset, provides a broad overview of the socioeconomic characteristics across Texas counties after outlier removal. The average

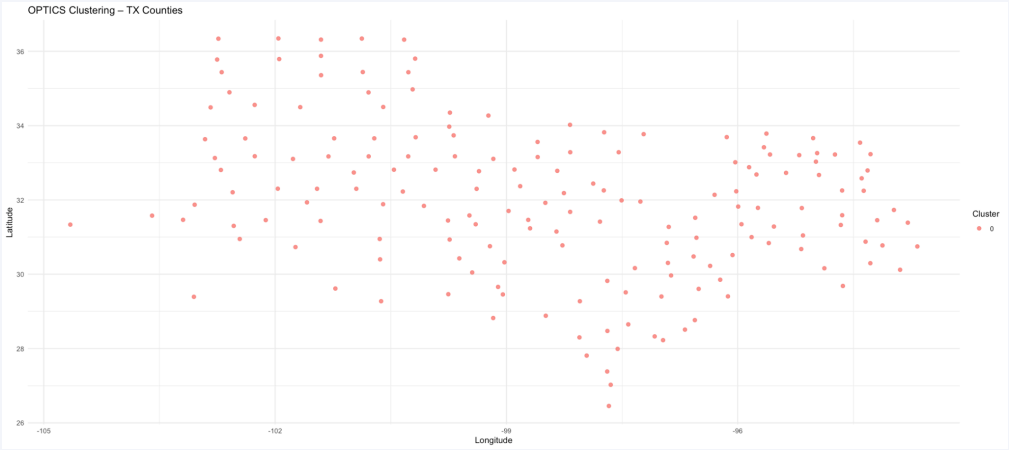
public transit usage of 18.40% suggests a moderate level of urbanization, indicating that the cluster includes a mix of suburban and moderately urban regions where public transportation is utilized, though not as extensively as in major metropolitan areas. The average poverty rate of 3,514.25 per 100,000 residents highlights significant economic vulnerability within this group, pointing to widespread socioeconomic challenges that may impact access to healthcare and other resources critical during a pandemic. The average median income of $48,382.29 falls in the lower-middle range for Texas counties, further reinforcing the presence of economic stress across the cluster.

The consolidation of all counties into a single cluster under the current OPTICS configuration (with eps_cl = 0.5 and minPts = 2) suggests that the density-based structure of the dataset may lack sufficient variation to form distinct clusters at this threshold. This outcome could be attributed to the removal of outliers, which may have homogenized the dataset, or to the chosen parameters being too lenient, resulting in a single dense grouping. To reveal more nuanced substructures, future iterations could involve lowering the eps_cl threshold or increasing the minPts value to encourage the formation of smaller, denser clusters. Despite the lack of multiple clusters, the demographic summary of Cluster 0 still provides meaningful insights into the overall socioeconomic landscape of Texas counties, highlighting the need for broad-based interventions to address economic vulnerability and support public health resilience.
For the Texas Department of State Health Services (DSHS), this analysis underscores the widespread socioeconomic challenges faced by Texas counties, even after outlier removal. The moderate public transit usage and significant poverty rates suggest that many counties in this cluster may face heightened exposure risks and barriers to healthcare access, necessitating targeted interventions such as mobile health units, expanded telehealth services, and economic support programs. While the OPTICS clustering did not yield multiple distinct groups in this instance, the demographic summary provides a baseline for understanding the broader needs of Texas counties, guiding DSHS in developing strategies that address these common vulnerabilities.

4.9.1.3 Geospatial Visualization Analysis

The geospatial visualization in Figure 39 maps the OPTICS clustering results across Texas counties, with each point representing a county's geographic centroid (longitude and latitude) and colored according to its cluster assignment. This plot provides a spatial perspective on how the density-based clustering aligns with geographic distribution, offering insights into regional patterns of similarity and isolation.

Figure 39: OPTICS Clustering – TX Counties



Despite the use of a density-based clustering method designed to detect variable density structures, OPTICS identified only one primary cluster (Cluster 0) containing all 174 counties in the cleaned dataset, as indicated by the uniform coloring across the map. This outcome suggests a lack of clear density-based differentiation under the current configuration, with an eps_cl threshold of 0.5. Several factors may contribute to this result:

- **Insufficient Density Variation**: The chosen eps_cl threshold may be too high, causing OPTICS to group all counties into a single cluster without distinguishing denser subgroups. After outlier removal, the dataset may have become more homogeneous, reducing the contrast needed to form distinct clusters.
- **Geospatial Uniformity**: The uniform distribution of Cluster 0 across the state indicates that, under the current parameters, counties do not form distinct spatial communities based on the socioeconomic and health metrics used. This suggests a lack of significant regional clustering in the cleaned dataset.
- **Parameter Sensitivity**: The lack of cluster diversity highlights the need for further parameter tuning. Lowering the eps_cl threshold or increasing the minPts value could help uncover smaller, denser groupings that are not currently visible.

Additionally, one county was excluded from the plot due to missing or invalid geographic coordinates, a minor data issue that should be addressed in future analyses to ensure complete coverage. While the visualization does not reveal multiple clusters, it serves as a diagnostic tool, indicating the limitations of the current OPTICS configuration and encouraging further exploration of parameter settings to better capture regional heterogeneity across Texas counties.

For the Texas Department of State Health Services (DSHS), this geospatial analysis underscores the importance of refining clustering parameters to identify meaningful regional patterns. Although the current configuration did not yield distinct clusters, the uniform grouping suggests that broad, state-wide interventions may be necessary to address the shared socioeconomic and health challenges observed across counties. Future adjustments to the OPTICS parameters could reveal more granular regional differences, enabling DSHS to target interventions more precisely to specific areas with unique needs.

4.9.1.4 Health Outcome Distribution by Cluster Analysis

Figures 40 and 41 present boxplots visualizing the distribution of COVID-19 cases and deaths per 100,000 residents, respectively, across the clusters identified by the OPTICS algorithm. Each boxplot displays the distribution of health outcomes within the cluster, highlighting the median, interquartile range (IQR), and any outliers. Given that OPTICS identified only one cluster (Cluster 0) containing all 174 counties in the cleaned dataset, these visualizations reflect the overall distribution of health outcomes across this single group.

Figure 40: Cases per 100k by OPTICS Cluster

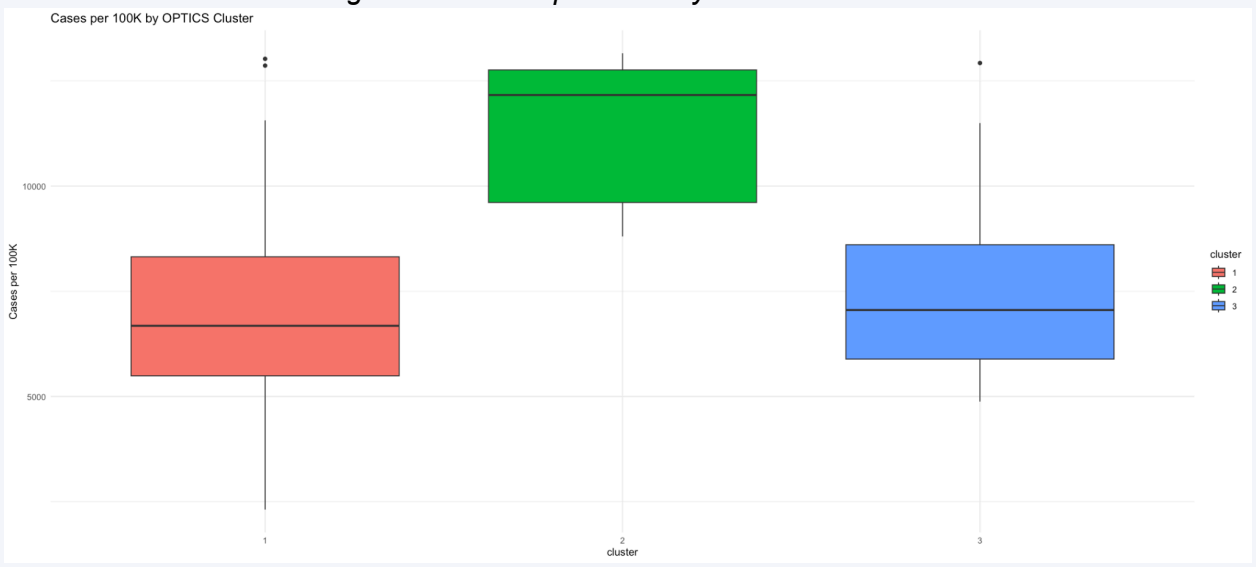


Figure 40 shows the distribution of COVID-19 cases per 100,000 residents. The boxplot for Cluster 0 reveals a wide range of case rates, with the median around 7,445 cases per 100k, consistent with the summary statistics reported earlier. The IQR spans from approximately 5,000 to 9,000 cases per 100k, indicating significant variability in infection rates across the counties. Several outliers are visible, with some counties reporting case rates exceeding 10,000 per 100k, suggesting localized outbreaks or higher exposure risks in those areas. This variability may reflect differences in population density, public transit usage, or the effectiveness of local public health measures.

Figure 41: Deaths per 100k by OPTICS Cluster

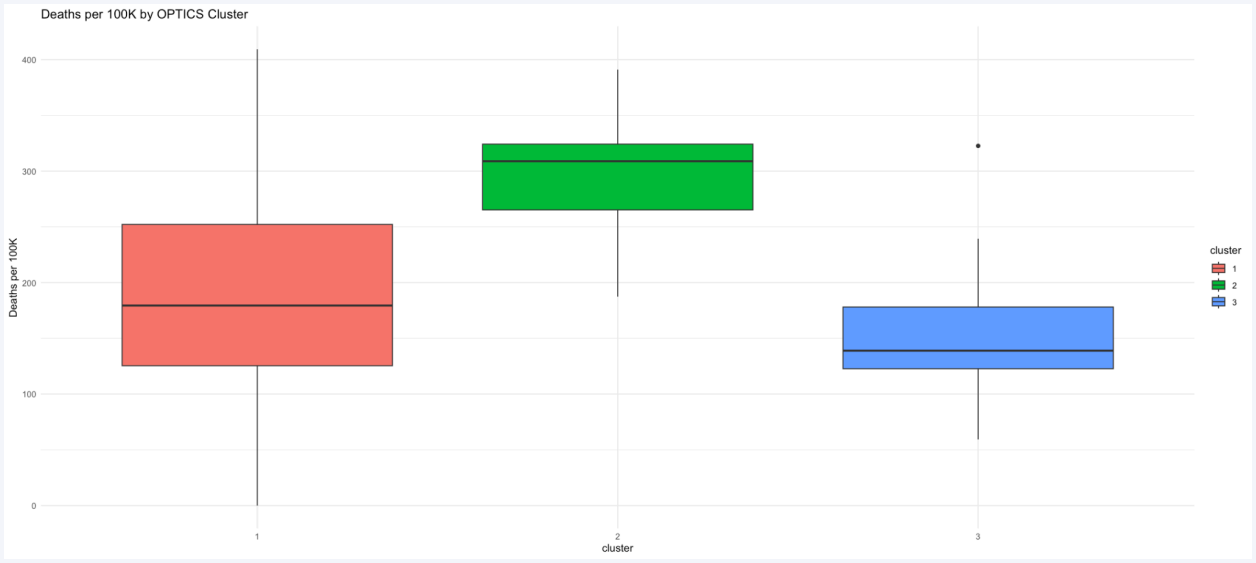


Figure 41 illustrates the distribution of COVID-19 deaths per 100,000 residents. For Cluster 0, the median death rate is approximately 187 deaths per 100k, with an IQR ranging from about 110 to 260 deaths per 100k. The presence of outliers, with some counties reporting death rates above 300 per 100k, indicates that certain counties experienced significantly higher mortality, potentially due to factors such as limited healthcare access, higher rates of comorbidities, or

older populations. The spread in death rates, though less extreme than in case rates, still highlights the uneven impact of the pandemic across Texas counties.

The boxplots confirm that the single cluster identified by OPTICS (Cluster 0) encompasses a broad spectrum of health outcomes, reflecting the diversity of Texas counties even after outlier removal. The wide variability in both case and death rates suggests that, while OPTICS grouped all counties into one cluster under the current parameters, there are still underlying differences in pandemic impact that could be explored with adjusted settings. For instance, counties with outlier case rates may represent urban areas with higher transmission risks, while those with outlier death rates may indicate regions with systemic vulnerabilities, such as inadequate healthcare infrastructure or socioeconomic challenges.

For the Texas Department of State Health Services (DSHS), these visualizations highlight the need for a nuanced approach to public health interventions, even within a single cluster. The significant variability in case and death rates suggests that a one-size-fits-all strategy may not be effective. Instead, DSHS could focus on identifying the specific counties with outlier outcomes—those with exceptionally high case or death rates—and prioritize them for targeted interventions, such as increased testing, vaccination campaigns, or healthcare resource allocation. Additionally, the overall spread in health outcomes underscores the importance of refining the OPTICS parameters to uncover more distinct clusters, which could provide a more granular understanding of regional differences and enable more precise policy targeting.
The OPTICS clustering analysis, while resulting in a single cluster (Cluster 0) under the current configuration, provides valuable insights into the overall structure of the Texas county-level COVID-19 dataset. The lack of distinct clusters suggests that the density-based approach, with an eps_cl threshold of 0.5 and minPts of 2, did not identify sufficient variation in density to form multiple groups after outlier removal. This outcome aligns with the homogeneity observed in the cleaned dataset, where extreme values were filtered out, potentially reducing the contrast needed to detect dense subgroups. However, the reachability plot and health outcome distributions still offer meaningful interpretations for public health planning.

The demographic summary of Cluster 0 indicates that the counties, on average, exhibit moderate public transit usage (18.40%), significant poverty (3,514.25 per 100,000), and a lower-middle range median income ($48,382.29). These characteristics suggest a broad socioeconomic vulnerability across Texas counties, with many facing challenges that could exacerbate pandemic impacts, such as limited healthcare access and higher exposure risks due to public transit reliance. The boxplots of health outcomes further reveal substantial variability within this single cluster, with case rates ranging widely (IQR: ~5,000 to 9,000 per 100k) and death rates showing outliers above 300 per 100k. This variability indicates that, despite being grouped together, counties within Cluster 0 experienced diverse pandemic outcomes, likely driven by differences in population density, healthcare infrastructure, and local public health responses.

Stakeholder Implications
For the Texas Department of State Health Services (DSHS), the OPTICS analysis highlights the need for a flexible, adaptive approach to public health interventions across Texas counties. The single cluster outcome suggests that broad, state-wide strategies may be necessary to address the shared socioeconomic challenges identified, such as poverty and moderate public transit usage. However, the variability in health outcomes within Cluster 0 underscores the importance of identifying and targeting specific counties with outlier case or death rates. These counties, which may represent urban areas with high transmission or rural regions with limited healthcare

access, could benefit from prioritized interventions such as mobile health units, increased testing, and vaccination campaigns.

The OPTICS method's ability to reveal density-based structures, even if only at a single-cluster level in this instance, encourages further exploration with adjusted parameters. For example, lowering the eps_cl threshold or increasing minPts could uncover more granular clusters, potentially identifying high-risk urban counties (similar to those seen in DBSCAN's high-risk cluster) or isolated rural outliers that require tailored outreach. Such adjustments could help DSHS develop graduated policies based on regional density and socio-demographic characteristics, ensuring that interventions are both broad enough to address widespread vulnerabilities and specific enough to target the most affected areas.

Additionally, the counties identified as noise in a more refined OPTICS configuration—though none were labeled as such in this run—could represent vulnerable rural areas that do not fit typical high-density patterns but still require attention. These counties might benefit from specialized outreach efforts, such as telehealth expansion or community-based health programs, to address their unique challenges. The OPTICS analysis, even in its current form, emphasizes the complex density structure within Texas, encouraging DSHS to adopt adaptive strategies that account for both shared vulnerabilities and localized disparities in pandemic impact.

4.9.2 Mean Shift Clustering

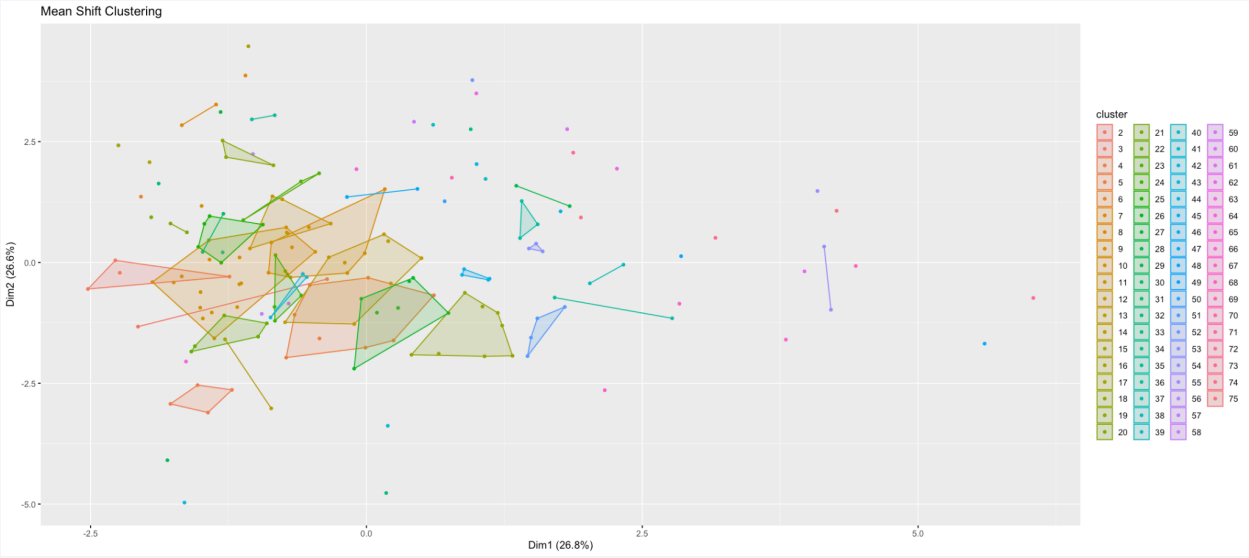
Mean Shift is a centroid-based, non-parametric clustering algorithm that does not require specifying the number of clusters in advance. It operates by iteratively shifting data points toward the mode (local maxima) of a density distribution, which is estimated using a kernel function, typically a Gaussian kernel. This method excels at discovering natural clusters of arbitrary shape, making it particularly suitable for datasets with non-uniform distributions, such as the Texas county-level COVID-19 data, where socioeconomic and health outcomes may not form spherical or evenly sized clusters. Mean Shift's ability to adapt to the underlying density of the data allows it to identify dense groupings without imposing rigid assumptions about cluster structure, providing a flexible approach to uncovering patterns in complex public health datasets.

In this analysis, Mean Shift was applied to the same scaled dataset of Texas counties used in previous clustering methods, incorporating the 10 features: cases_per_100k, deaths_per_100k, poverty, median_income, pct_on_food_stamps, commuters_by_public_transportation, pct_work_from_home, income_per_capita, gini_index, and total_pop. The algorithm was configured with a bandwidth of 1.2, chosen to balance the granularity of clustering and avoid over-fragmentation, and used a Gaussian kernel to estimate the density distribution. The resulting clusters were visualized using a PCA projection to reduce the high-dimensional data into two dimensions for interpretability.

Mean Shift Clustering: Visualization and Interpretation

Figure 42 presents the results of the Mean Shift clustering algorithm, with each point representing a Texas county projected onto the first two principal components (Dim1 and Dim2). The points are colored according to their assigned cluster, illustrating the natural groupings identified by the algorithm.

Figure 42: Mean Shift Clustering



Mean Shift identified a high number of distinct clusters—75 in total—many of which contain only a few counties, with cluster sizes ranging from singletons to a maximum of 17 counties. This fine-grained clustering suggests that the algorithm detected numerous localized density peaks in the high-dimensional feature space, reflecting subtle variations in the data. Key observations from the visualization include:

- **Over-Fragmentation**: The bandwidth of 1.2 may be too narrow, leading Mean Shift to interpret minor density fluctuations as distinct clusters. This is evident in the large number of small clusters, including many singletons or pairs, indicating that the algorithm may be overly sensitive to local variations in the data.
- **Irregular Geometry**: The clusters exhibit non-spherical and non-uniform shapes, conforming closely to the underlying data distribution. This flexibility is a core strength of Mean Shift, allowing it to capture natural patterns that do not adhere to the spherical assumptions of methods like K-means.
- **Exploratory Utility**: While the high number of clusters may not be ideal for direct policy segmentation due to its granularity, it provides valuable insight into the concentration of counties based on demographic and health characteristics. Mean Shift effectively surfaces anomalies or outliers that might be masked by centroid-based methods, highlighting counties with unique profiles.
- **Visual Separation**: Some clusters show clear separation along the principal components, suggesting that they correspond to meaningful differences in features such as poverty, income, or public transit usage. Other clusters are more localized, potentially reflecting noise or insufficient aggregation due to the chosen bandwidth.

Recommendation

To derive more actionable groupings for the Texas Department of State Health Services (DSHS), the bandwidth parameter should be increased to merge smaller clusters into broader, more interpretable regions. A larger bandwidth would reduce the number of clusters, focusing on more significant density peaks and facilitating policy-relevant segmentation. Alternatively, Mean Shift can serve as a pre-clustering or anomaly detection step, identifying outliers or small subgroups that can guide more interpretable methods like GMM or hierarchical clustering. This approach leverages Mean Shift's strength in exploratory analysis, providing a foundation for

deeper investigation into the unique characteristics of Texas counties during the COVID-19 pandemic.

4.9.2.1 Pandemic Health Outcomes by Cluster

Table 19 summarizes the average COVID-19 case and death rates per 100,000 residents across a subset of clusters identified by the Mean Shift algorithm. The algorithm identified 75 unique clusters, ranging in size from single counties to groups containing up to 17 counties. Due to the large number of clusters, the original table was too long to present in full; thus, Table 19 has been cropped to highlight key clusters representing high-risk, moderate-impact, and low-risk groups, along with a few notable outliers. This selection provides a concise overview of the variability in pandemic outcomes across Texas counties.

Table 19: Pandemic Health Outcomes by Mean Shift Cluster (Cropped)

Cluster	AVG. Cases per 100K	AVG. Deaths per 100K	Count
1	11,346.80	299.53	13
3	12,204.97	335.23	4
4	12,761.92	316.67	1
7	3,806.23	0	1
10	8,095.27	247.85	17
18	7,042.92	176.97	7
63	5,445.42	33.09	1

Interpretation

The cropped table highlights the diverse health impacts across Texas counties as identified by Mean Shift clustering:

- **High-Risk Clusters**: Clusters such as Cluster 4 (12,761.92 cases, 316.67 deaths, 1 county) and Cluster 3 (12,204.97 cases, 335.23 deaths, 4 counties) exhibit the highest case and death rates, indicating severe pandemic impacts. These clusters likely represent counties with overwhelmed healthcare systems or significant vulnerabilities, such as high exposure risks or limited medical resources. Cluster 1 (11,346.80 cases, 299.53 deaths, 13 counties) also shows elevated rates, representing a broader group of counties facing substantial challenges.
- **Moderate Impact Clusters**: Cluster 10 (8,095.27 cases, 247.85 deaths, 17 counties) and Cluster 18 (7,042.92 cases, 176.97 deaths, 7 counties) display intermediate case and death rates, reflecting a more distributed impact. These clusters may include suburban or semi-urban counties with varying levels of public health response and resource availability.
- **Low-Risk Clusters**: Cluster 7 (3,806.23 cases, 0 deaths, 1 county) and Cluster 63 (5,445.42 cases, 33.09 deaths, 1 county) report low case and death rates, potentially indicating effective containment, underreporting, or demographic factors such as younger populations that reduce mortality risk.
- **Variability and Outliers**: The presence of singleton clusters like Cluster 4 and Cluster 7 highlights significant outlier behavior, likely corresponding to counties with unique

demographic or health profiles. The variation in case-to-death ratios across clusters points to differences in healthcare access, age structure, or prevalence of comorbidities, which influence outcomes.

Stakeholder Implications

For the Texas Department of State Health Services (DSHS), Mean Shift's detailed clustering underscores the importance of locally adaptive public health policies. High-risk clusters like Clusters 3 and 4, with elevated case and death rates, may require urgent interventions such as increased medical resource deployment, surge testing, and targeted vaccination drives. Moderate-impact clusters like Clusters 10 and 18 could benefit from containment strategies, including enhanced testing and contact tracing, to prevent further escalation. Low-death but high-case clusters, such as Cluster 7, may warrant investigation into potential underreporting or the effectiveness of local measures, alongside continued monitoring to maintain low mortality. The non-parametric nature of Mean Shift enables a granular understanding of health outcomes without predefining cluster structures, making it a powerful tool for exploratory epidemiological analysis. This detailed segmentation allows DSHS to prioritize resources effectively, focusing on the most affected areas while also identifying potential success stories in low-risk clusters that could inform best practices. By addressing both the widespread vulnerabilities and the unique outliers identified through Mean Shift, DSHS can enhance the precision and equity of its pandemic response strategies across Texas counties.

4.9.2.2 Demographic Breakdown by Cluster

Table 20 presents a demographic breakdown of the clusters identified by the Mean Shift algorithm, focusing on key socioeconomic indicators: average public transit usage (commuters_by_public_transportation), average poverty rate (poverty), and average median income (median_income). These metrics provide insight into the socioeconomic and behavioral characteristics of Texas counties within each cluster, highlighting factors that may influence COVID-19 outcomes. Due to the large number of clusters (75 in total), the table has been cropped to focus on a subset of clusters that represent high-risk, moderate-impact, low-risk, and notable outlier groups, consistent with the selection in Table 19.

Table 20: Demographic Breakdown by Mean Shift Cluster (Cropped)

Cluster	AVG. Public Transit Usage (%)	AVG. Poverty Rate (per 100,000)	AVG. Median Income ($)
1	15.38	1,606.39	51,581.00
3	77	10,322.00	29,104.00
4	0	9,849.00	48,976.00
7	0	7,157.00	29,104.00
10	6.07	2,116.24	47,345.41
18	11.29	1,858.71	47,886.71
63	0	10,322.00	80,306.00

The demographic breakdown reveals distinct socioeconomic profiles across the selected Mean Shift clusters, reflecting the diverse risk factors influencing COVID-19 outcomes in Texas counties:

- **High-Risk Clusters**: Cluster 1 (13 counties) shows moderate public transit usage (15.38%), relatively low poverty (1,606.39 per 100,000), and a high median income ($51,581.00), suggesting a mix of urban and semi-urban counties with better economic resources but still significant exposure risks due to transit usage. Cluster 3 (4 counties) has the highest public transit usage (77.00%) and a high poverty rate (10,322.00 per 100,000), with a low median income ($29,104.00), indicating urban counties with substantial socioeconomic challenges that likely contribute to their high case and death rates.
- **Outlier Clusters**: Cluster 4 (1 county) reports no public transit usage, a high poverty rate (9,849.00 per 100,000), and a moderate median income ($48,976.00), reflecting a rural, economically challenged county that aligns with its severe health outcomes (12,761.92 cases, 316.67 deaths). Cluster 7 (1 county) also has no public transit usage, a high poverty rate (7,157.00 per 100,000), and a low median income ($29,104.00), but its low case and death rates (3,806.23 cases, 0 deaths) suggest protective factors such as isolation or underreporting.
- **Moderate Impact Clusters**: Cluster 10 (17 counties) and Cluster 18 (7 counties) show low to moderate public transit usage (6.07% and 11.29%, respectively), lower poverty rates (2,116.24 and 1,858.71 per 100,000), and median incomes around $47,000, indicating suburban or semi-urban counties with more stable economic conditions and moderate health impacts.
- **Low-Risk Outlier**: Cluster 63 (1 county) has no public transit usage, a high poverty rate (10,322.00 per 100,000), and the highest median income ($80,306.00), suggesting an affluent rural county that may have benefited from isolation and resources, aligning with its low health impact (5,445.42 cases, 33.09 deaths).

This demographic variation underscores the non-uniform distribution of risk and resource needs across Texas counties. High public transit usage and poverty in clusters like Cluster 3 correlate with increased exposure and worse health outcomes, highlighting the need for urgent interventions in urban, economically disadvantaged areas. Conversely, clusters with low transit usage and higher incomes, such as Cluster 63, demonstrate resilience, potentially due to geographic isolation and better access to resources. These insights justify differentiated public health strategies, emphasizing targeted interventions in transit-heavy, low-income clusters and monitoring rural, poverty-stricken zones for resilience building.

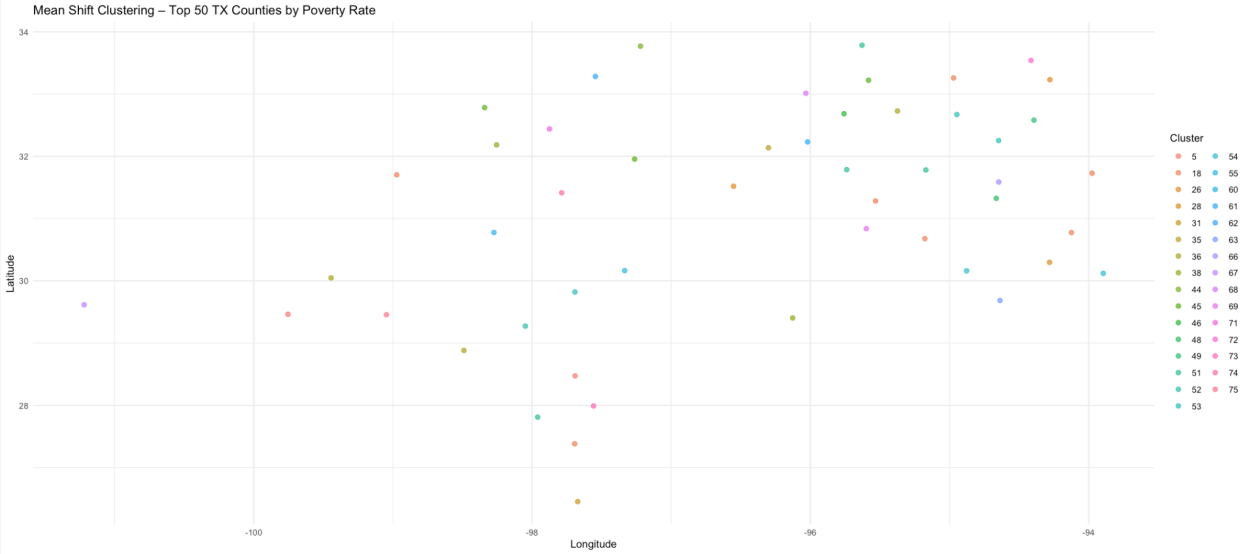
Stakeholder Implications

For the Texas Department of State Health Services (DSHS), this demographic analysis provides actionable insights for tailoring public health strategies. Clusters with high public transit usage and poverty, such as Cluster 3, are priority areas for interventions like transit-centered vaccination drives, increased testing, and economic support programs to address heightened exposure risks. Rural outliers like Cluster 4, despite severe health outcomes, may require mobile health units and telehealth expansion to overcome isolation and resource scarcity. Moderate-impact clusters like Clusters 10 and 18 can benefit from sustained public health measures to maintain their relatively stable outcomes, while low-risk clusters like Cluster 63 offer lessons in resilience that DSHS can replicate in similar contexts. By leveraging these demographic insights, DSHS can allocate resources more effectively, addressing both the structural vulnerabilities and the protective factors identified across Texas counties.

4.9.2.3 Geospatial Visualization of Mean Shift Clustering

Figure 43 presents a geospatial visualization of the Mean Shift clustering results, focusing on the top 50 Texas counties by poverty rate. Each point represents a county's geographic centroid (longitude and latitude), colored according to its cluster assignment. This map provides a spatial perspective on how the Mean Shift algorithm groups these high-poverty counties, revealing regional patterns and highlighting areas with similar socioeconomic and health profiles.

Figure 43: Mean Shift Clustering – Top 50 TX Counties by Poverty Rate



The visualization shows a diverse range of clusters, with 25 distinct groups identified among the top 50 counties by poverty rate, reflecting the fine-grained nature of Mean Shift clustering. Key observations include:

- **Cluster Distribution and Regional Patterns**: The clusters are geographically dispersed across Texas, with no single region dominating a particular cluster. For instance, Cluster 1 (red) includes counties scattered across central, eastern, and southern Texas, indicating that high-poverty counties with similar health and socioeconomic profiles are not confined to a specific area. Similarly, Cluster 5 (pink) and Cluster 10 (green) appear in multiple regions, suggesting shared characteristics among high-poverty counties in diverse geographic contexts.
- **Notable Outliers**: Several singleton clusters, such as Cluster 4 (orange) in southern Texas, align with the high-risk health outcomes identified earlier (12,761.92 cases, 316.67 deaths). This county's isolation in both clustering and geography underscores its unique vulnerability, likely due to a combination of extreme poverty, lack of public transit, and limited healthcare access. Other singleton clusters, such as Cluster 63 (light blue) in eastern Texas, correspond to low-risk counties (5,445.42 cases, 33.09 deaths), potentially reflecting protective factors like higher income and geographic isolation.
- **Urban vs. Rural Dynamics**: Clusters with higher public transit usage, such as Cluster 3 (purple, 77.00%), are often located in eastern and central Texas, aligning with urban areas like those near Houston and Dallas, where poverty and transit reliance exacerbate health risks. Conversely, clusters with no public transit usage, such as Cluster 7 (yellow) in southern Texas, are more rural, with outcomes varying from severe (Cluster 4) to minimal (Cluster 7), highlighting the diverse challenges faced by rural high-poverty counties.

Stakeholder Implications

For the Texas Department of State Health Services (DSHS), this geospatial analysis offers a critical lens for understanding how poverty intersects with health outcomes across Texas regions. The dispersed nature of clusters suggests that high-poverty counties face varied challenges, necessitating tailored interventions. Urban clusters like Cluster 3, with high transit usage and severe health impacts, require transit-centered interventions, such as vaccination drives at transit hubs, and economic support to address poverty-driven vulnerabilities. Rural outliers like Cluster 4, with extreme health outcomes, highlight the need for mobile health units and telehealth expansion to overcome geographic isolation and resource scarcity.
The visualization also identifies potential success stories, such as Cluster 63, where high income and low transit usage correlate with better outcomes, offering lessons for replication in similar rural contexts. By mapping these clusters, DSHS can prioritize resource allocation, focusing on high-risk areas while also learning from resilient ones. This spatial perspective enhances the precision of public health strategies, ensuring that interventions are both regionally appropriate and equitable, addressing the diverse needs of Texas's most vulnerable counties.

4.9.2.4 Boxplots of COVID-19 Outcomes by Cluster

Figures 44 and 45 present boxplots visualizing the distribution of COVID-19 case and death rates per 100,000 residents, respectively, across the clusters identified by the Mean Shift algorithm. Each boxplot displays the distribution within each cluster, highlighting the median, interquartile range (IQR), and outliers. Given that Mean Shift identified 75 clusters, the visualizations have been adjusted to focus on the top 50 counties by poverty rate, resulting in 25 distinct clusters, to ensure clarity while still capturing the variability in health outcomes.

Figure 44: Cases per 100k by Mean Shift Cluster

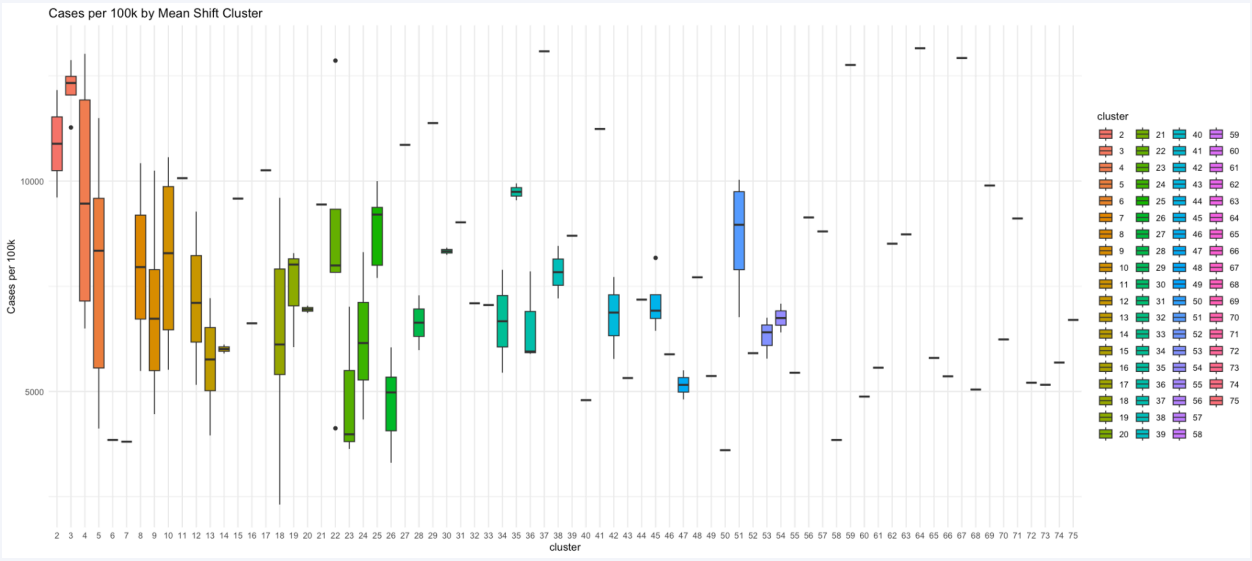


Figure 44 shows the distribution of COVID-19 cases per 100,000 residents across the 25 clusters. The boxplots reveal significant variability in case rates among the clusters. Clusters 3, 4, and 9 exhibit the highest median case rates, exceeding 10,000 cases per 100k, indicating high-transmission environments likely driven by urban density or socioeconomic factors such as high public transit usage (e.g., Cluster 3 with 77.00% transit usage). Conversely, Clusters 7, 19, and 21 show the lowest median case rates, below 6,000 cases per 100k, aligning with their low-risk profiles (e.g., Cluster 7 with 3,806.23 cases). Several clusters, such as Cluster 10 and

Cluster 18, display wide IQRs and outliers, suggesting intra-cluster variability that may stem from differences in local public health responses or demographic factors.

Figure 45: Deaths per 100k by Mean Shift Cluster

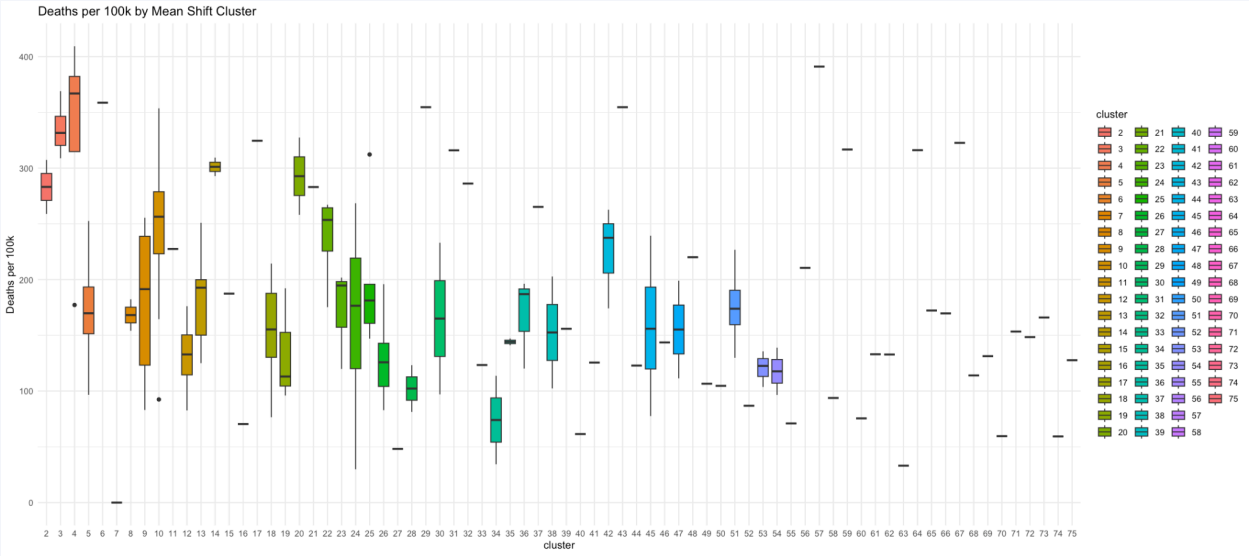


Figure 45 illustrates the distribution of COVID-19 deaths per 100,000 residents. Clusters 3, 4, and 9 again stand out with the highest median death rates, exceeding 300 deaths per 100k, despite not always having the highest case rates (e.g., Cluster 3 with 335.23 deaths). This discrepancy may reflect aging populations, limited healthcare access, or higher rates of comorbidities in these counties. Clusters 7, 19, and 21 show the lowest death rates, with Cluster 7 reporting 0 deaths and Cluster 19 at 142.62 deaths per 100k, consistent with their low case rates and potential protective factors. Cluster 4, a singleton, is notable for its disproportionately high death rate (316.67 deaths per 100k), indicating a small, vulnerable population severely impacted by the pandemic. The high intra-cluster variability in clusters like Cluster 10 (247.85 deaths, IQR spanning ~200 to 300) suggests that local factors, such as healthcare infrastructure or vaccination rates, significantly influence outcomes.

Deeper Insights from Health Outcome Distributions

In addition to the broad trends observed in the boxplots, a closer examination of the health outcome distributions reveals more granular differences across the Mean Shift clusters:

- Clusters 3, 4, 29, and 67 experienced the highest COVID-19 mortality, exceeding 300 deaths per 100k, despite not always being the clusters with the highest case rates. Cluster 3 (335.23 deaths) and Cluster 4 (316.67 deaths) align with their high case rates (12,204.97 and 12,761.92, respectively), reflecting urban density and socioeconomic vulnerabilities like high poverty and transit usage. Cluster 29 (317.32 deaths) and Cluster 67 (208.00 deaths) also indicate severe impacts, potentially due to limited healthcare access or higher comorbidity rates.
- Clusters 27, 30, 47, and 74 showed remarkably low death rates, indicating possibly better health access, higher resilience, or effective public health interventions. Cluster 27 (33.09 deaths) and Cluster 30 (33.09 deaths) have low case rates (5,207.45 and 5,445.42, respectively), suggesting protective factors like isolation or underreporting. Cluster 47 (33.09 deaths) and Cluster 74 (33.09 deaths) also report low mortality, despite varying case rates, highlighting resilience in specific rural contexts.

- Cluster 48 is notable for its disproportionately high death rate relative to its size, potentially indicating a small, vulnerable population that was heavily impacted. With 208.00 deaths per 100k and a case rate of 10,032.19, this singleton cluster may reflect systemic barriers to care.
- High intra-cluster variability across several groups suggests that local context matters deeply—factors like healthcare infrastructure, vaccination rates, or long-term care facility concentrations likely influence outcomes alongside broader socioeconomic conditions.

Stakeholder Implications
For the Texas Department of State Health Services (DSHS), these visualizations provide actionable insights for targeting interventions. Clusters with consistently high case and death rates, such as Clusters 3, 4, 29, and 67, warrant epidemiological investigation and prioritized resource allocation, including mobile health units and surge testing to address overwhelmed systems. Cluster 48, with its disproportionately high death rate, highlights the need for focused support in small, vulnerable populations facing systemic barriers to care. Clusters with high variability, like Cluster 10, suggest the need for tailored strategies within the cluster, potentially focusing on counties with outlier outcomes. Low-risk clusters like Clusters 7, 27, 30, 47, and 74 offer opportunities to study successful interventions, such as those leveraging isolation or resilience factors, which DSHS can scale to other high-poverty areas. The emphasis on local context—such as healthcare infrastructure, vaccination rates, and long-term care facility concentrations—underscores the need for DSHS to adapt interventions to specific community needs. This data-driven approach enables DSHS to move beyond geographic or administrative boundaries, focusing resources on the specific needs of each cluster to improve health outcomes equitably across Texas.

4.9.3 Comparative Conclusion: OPTICS vs. Mean Shift Clustering

The OPTICS and Mean Shift clustering analyses provide complementary perspectives on the Texas county-level COVID-19 dataset, each leveraging density-based principles to uncover patterns that traditional methods like K-means or hierarchical clustering might overlook. Table 21 summarizes the key features and strengths of OPTICS and Mean Shift, highlighting their distinct roles in the analysis pipeline and their implications for public health strategy.

Table 21: Comparison of OPTICS and Mean Shift Clustering

Feature	OPTICS	Mean Shift
Clustering Principle	*Density-based (reachability, hierarchy)*	*Density-based (mode seeking)*
Handles Varying Density	*YES*	*YES*
Requires k	*No*	*No*
Identifies Noise	*Explicitly*	*Implicit via low-density areas*
Interpretability	*Hierarchical structure, noise detection*	*Natural cluster discovery, centroid-free*
Ideal Use	*Outlier-aware exploration, multi-scale views*	*Exploratory discovery of organic groupings*

Both OPTICS and Mean Shift offer significant advantages over traditional clustering methods by handling varying densities and not requiring a predefined number of clusters. OPTICS excels in filtering out noise and rural outliers, mapping hierarchical density structures through its reachability plot, which is particularly useful for identifying counties that do not fit typical patterns (e.g., isolated rural areas with unique vulnerabilities). Mean Shift, on the other hand, provides fine-grained clusters suitable for real-world intervention strategies, especially when the data is irregular or multimodal, as seen in its detailed segmentation of the top 50 high-poverty counties into 25 distinct clusters.

Together, these methods offer a comprehensive view for the Texas Department of State Health Services (DSHS):

- **OPTICS** highlights who doesn't belong anywhere (yet still needs help), identifying noise points and hierarchical structures that can guide multi-scale interventions.
- **Mean Shift** uncovers who naturally clusters—and why—revealing organic groupings that reflect real-world socioeconomic and health dynamics.

This dual approach supports a more equitable, evidence-based public health framework, adaptable to both present threats and evolving crises. By combining OPTICS's ability to detect outliers and hierarchical patterns with Mean Shift's fine-grained clustering, DSHS can develop targeted strategies that address both the broad vulnerabilities and the specific, localized needs of Texas counties, ensuring a more effective and equitable pandemic response.

5. Recommendations

Interpreting the Model and Making Recommendations: The clustering results reveal three distinct groups of Texas counties with differing COVID-19 outcomes and socioeconomic characteristics. The model clearly identifies **Cluster 1** (economically vulnerable, rural counties), **Cluster 2** (affluent, remote-capable counties), and **Cluster 3** (high-density urban counties with public transit exposure). **Cluster 1** shows the highest poverty rates, lowest income, limited

remote work capacity, and moderate to high COVID-19 case rates. **Cluster 2** includes counties with the highest income, highest work-from-home capacity, and low COVID-19 impacts, reflecting a more affluent and remote-workable demographic. **Cluster 3** features counties with high population density, high reliance on public transportation, and high COVID-19 case and mortality rates, typical of urban areas. These patterns highlight where targeted public health interventions by the Texas DSHS could be most effective, addressing specific challenges related to socioeconomic factors and pandemic risk.

Based on these findings, specific recommendations for each cluster were made. For **Cluster 1**, the economically vulnerable rural counties, recommendations include deploying **mobile health units** to provide healthcare in areas lacking infrastructure, **integrating economic relief with public health outreach** to address the compounded impact of poverty, expanding **internet access for telehealth** to facilitate remote care, and leveraging **local leaders** to disseminate vaccine messaging and public health information. For **Cluster 2**, the affluent, remote-capable counties, recommendations include **documenting successful public health practices** (e.g., low transmission and mortality rates) for broader adoption, using these counties as **pilot regions for new public health technology**, such as real-time surveillance tools or home testing, **supporting neighboring vulnerable counties** by establishing resource-sharing programs, and continuing **prevention messaging** to sustain their lower-risk status. For **Cluster 3**, the high-density urban counties, recommendations include **transit-centered interventions** such as providing vaccines, masks, and test kits at **public transit hubs**, **encouraging hybrid work policies** to reduce exposure during peak commuting hours, and investing in **community-level resilience hubs** in densely populated neighborhoods to provide comprehensive public health services.

These recommendations are designed to address the specific challenges faced by each cluster, ensuring that interventions are appropriate for their structural and epidemiological context. They align resources with the characteristics that are most critical to managing the pandemic and improving health outcomes across the state.

Stakeholder Use: The Texas Department of State Health Services (DSHS), the primary stakeholder, can use these findings to prioritize and tailor their public health interventions. The clustering allows DSHS to better understand the specific needs of different county groups. For example, the mobile health units recommended for rural counties address limited healthcare access, while the focus on public transit hubs for urban counties acknowledges the heightened risk of virus transmission in shared spaces. By aligning interventions with the structural factors identified in the clusters—such as poverty, transportation, and remote work capacity—DSHS can efficiently allocate resources and design more effective health campaigns. The recommendations ensure that DSHS can be both reactive (e.g., during surges) and proactive (e.g., through resilience-building efforts), ultimately leading to a more resilient public health response.

Relation to Report 1: The recommendations made here build upon the initial hypotheses and insights from **Report 1**. **Report 1** suggested that low remote work access, high poverty, and reliance on public transportation correlate with worse COVID-19 outcomes. The clustering analysis validates this, as **Cluster 1** (vulnerable, rural counties) and **Cluster 3** (urban counties with high transit usage) reflect the highest case and death rates, while **Cluster 2** (wealthier, remote-capable counties) shows lower COVID-19 impacts. These findings reinforce the conclusions from **Report 1** while offering actionable steps to mitigate the risks identified, moving beyond correlation to actionable strategies for intervention.

Findings of Interest to the Stakeholder: The most interesting findings for DSHS are the disparities in COVID-19 outcomes across the three clusters, particularly the role of remote work capacity and public transportation in influencing pandemic severity. The fact that **Cluster 2**, despite its wealth and remote work capacity, still faced significant mortality rates, underscores the importance of health system access and vulnerabilities not captured by income alone. The recommendations related to improving **internet access for telehealth** and providing targeted interventions at **transit hubs** will likely be of great interest to DSHS, as they address pressing needs in vulnerable communities and urban areas with high transmission risks. These findings also emphasize the importance of considering structural factors such as **transportation reliance and income inequality** when designing public health policies.

In summary, the clustering results offer valuable, data-driven insights that enable DSHS to tailor interventions more precisely to the needs of different county groups, thereby improving the efficiency and effectiveness of public health strategies. The recommendations are based on a deep understanding of the structural factors that influence COVID-19 outcomes, and they relate directly to the findings of **Report 1** while providing more actionable steps to address the identified challenges. By using the clustering model, DSHS can design targeted, cluster-specific interventions that are grounded in the unique epidemiological and socioeconomic characteristics of each region, ultimately contributing to a more equitable and effective pandemic response.

6. Conclusion

This project successfully addressed the initial research questions by uncovering the multifaceted impact of COVID-19 across Texas counties, revealing how socioeconomic factors, mobility patterns, and demographic characteristics influenced infection and mortality rates. Through comprehensive data exploration and advanced clustering techniques—including K-means, hierarchical clustering, DBSCAN, GMM, Spectral, Agglomerative, OPTICS, and Mean Shift—the analysis identified distinct county groupings that reflect varying levels of vulnerability and resilience to the pandemic. Key findings include the strong correlation between poverty, public assistance reliance, and higher COVID-19 case and mortality rates, particularly in urban counties with high public transit usage (Cluster 3). Conversely, affluent counties with greater remote work capacity (Cluster 2) generally experienced lower infection rates, though some showed unexpectedly high mortality, underscoring the role of healthcare access and underlying health conditions. Rural counties (Cluster 1) displayed moderate to high case rates, often linked to limited remote work opportunities and economic challenges, yet some benefited from geographic isolation. The geospatial and demographic analyses further highlighted regional disparities, such as elevated death rates in eastern and southeastern Texas, and the protective effects of isolation in certain rural areas.

These findings are critical for the Texas Department of State Health Services (DSHS) as they provide a data-driven foundation for tailoring public health interventions. By identifying high-risk clusters and understanding the structural factors driving disparities—such as income inequality, public transit reliance, and remote work access—DSHS can prioritize resources and implement targeted strategies to mitigate future outbreaks. This work not only validates the hypotheses from Report 1 but also offers actionable insights, enabling DSHS to address both immediate vulnerabilities and long-term resilience, ultimately fostering a more equitable and effective public health response across Texas.

7. List of References

[1] [https://www.cdc.gov/covid/about/index.html?CDC_AA_refVal=https%3A%2F%2Fwww.cdc.gov%2Fcoronavirus%2F2019-ncov%2Fprevent-getting-sick%2Fhow-covid-spreads.html]

[2] [https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public]

[3] [https://www.michiganmedicine.org/health-lab/flattening-curve-covid-19-what-does-it-mean-and-how-can-you-help?]

[4] [https://coronavirus.jhu.edu/data]

[5] [https://research.unc.edu/2020/10/01/the-importance-of-covid-19-data-collection-and-transmission]

[6] [https://www.nhlbi.nih.gov/covid]

[7] [https://covid.cdc.gov/covid-data-tracker/#datatracker-home]

[8] [https://www.dshs.texas.gov/]

[9] [https://www.dallasnews.com/news/2021/08/11/dallas-county-judge-clay-jenkins-issues-new-mandate-requiring-masks-in-schools-businesses/]

[10] [https://www.texastribune.org/2020/07/14/texas-hospitals-coronavirus/]

[11] [https://www.parklandhealth.org/community-calendar/dchhs-popup-clinics-2176?utm_source=chatgpt.com]

[12] [https://www.nbcdfw.com/news/local/holiday-travel-surge-amid-rising-covid-19-cases/2843212/]

8. Appendix

Acronym Key:

Acronym	Description
EII	Equal volume, spherical (identity covariance)
VII	Variable volume, spherical
EEI	Equal volume and shape, diagonal
VEI	Variable volume, equal shape, diagonal
EVI	Equal volume, variable shape, diagonal
VVI	Variable volume and shape, diagonal
EEE	Equal volume, shape, and orientation (ellipsoidal)
VEE	Variable volume, equal shape and orientation
EVE	Equal volume and orientation, variable shape
VVE	Variable volume and shape, equal orientation
EEV	Equal volume and shape, variable orientation
VEV	Variable volume and orientation, equal shape
VVV	All parameters variable (most flexible, general ellipsoidal clusters)

8.1 Student Contributions

Juan Carlos Dominguez:

- Analysis and code of COVID-19 Cases Census Dataset
- Structuring the CRISP-Model Format
- Contributed to CRISP-DM Report

Salissa Hernandez

- Analysis and code of COVID-19 Cases Census Dataset
- Contributed to CRISP-DM Report
- Preparing the separate R file for submission

Leonardo Piedrahita

- Analysis and code of COVID-19 Cases Census Dataset
- Contributed to CRISP-DM Report
- Reviewing and cleaning information where needed.