Data Mining Project 2:

Clustering Analysis of COVID-19 Spread in Texas: Demographics, Mobility, and Policy Implications

Sreshta Ghosh, Muskaan Mahes, and Ridhi Ralli

Executive Summary

Due to the COVID-19 pandemic, this analysis was conducted to support the Texas State Health Services Commissioner by using COVID-19 data to guide future public health funding and policy decisions. To address key questions, unsupervised clustering methods were applied to 12 key demographic, socioeconomic, mobility, and health indicators across Texas counties. By using K-means, Hierarchical, Hierarchical CURE, and DBSCAN clustering, we identified patterns in community vulnerability and COVID-19 severity. Clustering was used to group counties with similar characteristics, allowing us to uncover high-risk areas and inform targets interventions. Among the methods tested, K-means consistently produced the most meaningful and interpretable results, especially in aligning with severity outcomes. These insights offer a data-driven framework to help the state allocate vaccines, staffing, and resource more effectively, while tailoring interventions to the needs of vulnerable populations

Contents

1) Problem Statement (Business Understanding)

The following information is from Project 1 Report (Project1_SG_RR_MM):

Our main question is:

- How can COVID data inform future public health funding and policymaking for the state of Texas?

Additional questions:

- Where are the most outbreaks occurring, and what factors contribute to their spread?
- Which demographics are most affected, and how should interventions be tailored?
- Where should resources like vaccines and medical staff be distributed?

Understanding how key factors affect the spread and severity of the disease is essential for shaping future public health funding and policy decisions. Using these relationships and questions, we can identify the areas of greatest need and allocate resources effectively.

2) Data Preparation

More information can be found in Project 1 Report (Project1_SG_RR_MM).

What are the objects you want to cluster?

In this clustering analysis, the objects of interest are the counties in Texas. This focus directly aligns with our stakeholder—the Commissioner of the Texas Department of State Health Services—who requires actionable insights at the county level to inform public health funding and policymaking. By clustering counties based on these factors, we aim to identify groups of counties with similar public health profiles, ultimately supporting targeted interventions and resource allocation decisions.

Describe which features you want to use for clustering and why.

As shown in Table 1 below, we will use the following features to characterize each Texas county: confirmed_cases, deaths, retail_and_recreation_percent_change_from_baseline, grocery_and_pharmacy_percent_change_from_baseline, workplaces_percent_change_from_baseline, residential_percent_chance_from_baseline, median_income, median_age, nonfamily_households, family_households, and poverty. These features were selected because they provide a comprehensive view of each county's public health situation, socioeconomic status, and behavioral changes during the pandemic. Confirmed_cases and deaths directly capture the COVID-19 impact, while the mobility metrics (retail and recreation, grocery and pharmacy, workplaces, and residential percent changes) offer insight into how residents altered their movements relative to a pre-pandemic baseline. Demographic and economic variables such as total population, median income, median age, and poverty help define the underlying context of each county, and household composition (nonfamily and family households) further enriches this picture by indicating community structure. We decided not to include the transit_stations_percent_change_from_baseline column in our analysis because transit usage is not as prevalent in Texas as it is in other states, making it a less informative indicator for our specific context.

Overall, these chosen features enable us to cluster counties in a way that highlights both the direct effects of COVID-19 and the broader social and economic factors that can influence public health outcomes, thereby providing actionable insights for policy and resource allocation decisions.

Table 1: Important Features' Descriptions

Features	Scale of Measurement	Description
cases_per_10000	Ratio	The number of confirmed COVID-19 cases in a county, scaled per 10,000 residents.
deaths_per_10000	Ratio	The number of COVID-19-related deaths in a county, scaled per 10,000 residents.
retail_and_recreation_percent_change_from_baseline	Interval	Percent change in visits to retail stores and recreation areas indicating shifts in consumer and leisure activities.
grocery_and_pharmacy_percent_change_from_baseline	Interval	Percent change in visits to grocery stores and pharmacies reflecting shifts in shopping behavior.
workplaces_percent_change_from_baseline	Interval	Percent change in visits to workplaces indicating shifts in work-related mobility.
residential_percent_chance_from_baseline	Interval	Percent changes in visits to residential areas reflecting how much more/less people are staying at home relative to normal conditions.
median_income	Ratio	The average household income in a county.
median_age	Ratio	The average age of the population in the county.
nonfamily_households_10000	Ratio	The number of nonfamily households in a county, scaled per 10,000 total households. These typically include individuals living alone or with non-relatives.
family_households_10000	Ratio	The number of family households in a county, scaled per 10,000 total households. These include households with related individuals.
poverty	Ratio	The percentage of the county population living below the federal poverty line.
death_per_case	Ratio	The proportion of confirmed COVID-19 cases that resulted in death. This can be interpreted as a case fatality ratio.

Statistics for the Most Important Features

Table 2: Statistical Summary of Important Variables in composite_normalized

Variables	Data Type	Mean	Median	Mode	St. Dev	Variance	Min	Max	Range (Max-Min)
cases_per_10000	Integer (Numeric)	3545.80	850.78	7255.102	13787.74	190101768	11.79	131783.63	131771.8
deaths_per_10000	Integer (Numeric)	80.856	17.900	28.24859	388.3149	150788.5	0.309	4608.637	4608.328
retail_and_recreation_percent_change_from_baseline	Double (Numeric)	-17.78	-17.00	-14	6.731706	45.31587	-42.00	-2.00	40
grocery_and_pharmacy_percent_change_from_baseline	Double (Numeric)	-12.65	-12.00	-9	8.252246	68.09956	-44.00	4.00	48
workplaces_percent_change_from_baseline	Double (Numeric)	-19.95	-18.00	-14	7.723571	59.65355	-43.00	-1.00	42
residential_percent_chance_from_baseline	Double (Numeric)	9.036	9.000	9	2.739207	7.503253	1.000	17.000	16
median_income	Double (Numeric)	49231	48139	44601	12725.68	161942961	23558	93645	70087
median_age	Double (Numeric)	39.39	39.25	35.6	5.882809	34.60744	25.80	57.10	31.3
nonfamily_households_10000	Double (Numeric)	3082	3059	3574.548	563.6086	317654.6	1790	5188	3397.966
family_households_10000	Double (Numeric)	6918	6941	6425.452	563.6086	317654.6	4812	8210	6425.452
poverty	Integer (Numeric)	13764	3848	1049	34590.85	1196527041	42	304792	1049
death_per_case	Double (Numeric)	0.023651	0.022147	0.01424051	0.011259 34	0.000126772 7	0.003632	0.091667	0.01424051

To ensure meaningful comparisons across Texas counties with varying population sizes, we normalized key variables using each county's total population (total_pop) or total number of households (households). These transformations converted the raw counts into per capita or per household rates. Using the results, we were able to compare counties of different sized on an equitable scale.

Below we have included the formulas we used for normalization:

Normalization using total_pop:

cases_per_10000 = (confirmed_cases/total_pop)*10,000

deaths_per_10000 = (deaths/total_pop)*10,000

Normalization using Household Counts:

nonfamily_households_10000 = (nonfamily_households/households) * 10,000

family_households_10000 = (family_households/households) * 10,0000

Mortality Ratio:

death_per_case = deaths/confirmed_cases

Under each clustering method in the Modeling section, we will include the appropriate measures for similarity and distance to ensure our clusters accurately reflect the relationships among the counties based on the chosen features.

3) Modeling

Cluster Analysis using Different Methods

1. K-Means Clustering

In this analysis, we applied K-Means clustering using four different feature subsets to explore varying aspects of the data, including demographic variables, household composition, poverty indicators, mobility patterns, and COVID-19 impact measures. This multi-feature approach enables a comprehensive understanding of how different dimensions of the data influence clustering outcomes. All variables were standardized prior to clustering to ensure equal contribution to distance calculations, and **Euclidean distance** was used as the similarity measure.

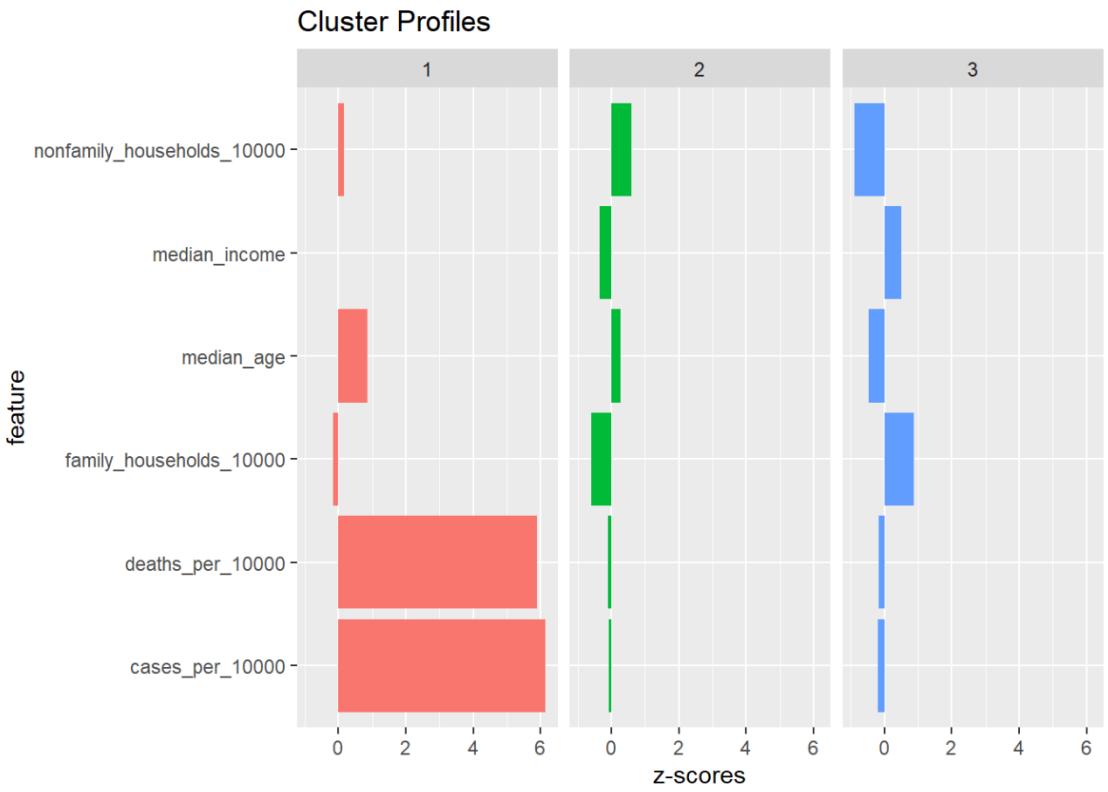


Figure 1: Cluster Profile using Demographic and Household Features

Figure 1, which combines both demographic and household variables, reveals distinct patterns in COVID-19 impact across the clusters. Importantly, all variables were normalized to account for differences in county sizes, allowing for meaningful comparisons. Specifically, cases and deaths were scaled per 10,000 residents (cases_per_10000, deaths_per_10000), and household variables were adjusted relative to total households (nonfamily_households_10000, family_households_10000). Cluster 1 is characterized by significantly higher COVID-19 burden, with elevated levels of both cases and deaths per 10,000 residents, while also displaying higher family household concentration and lower median income and age, potentially indicating areas with larger family households and socioeconomic vulnerabilities. Cluster 2 shows moderate values across most variables but slightly higher non-family households and a balanced demographic profile, suggesting more urbanized or mixed household settings with moderate COVID-19 impact. Cluster 3 exhibits the lowest levels of cases and deaths per 10,000 residents, alongside higher median income and age, and moderate household compositions, which may reflect more affluent or older populations with lower pandemic impact. These findings highlight how demographic and household structures, combined with normalized COVID-19 indicators, can reveal meaningful patterns in how different regions experienced the pandemic, guiding targeted public health strategies.

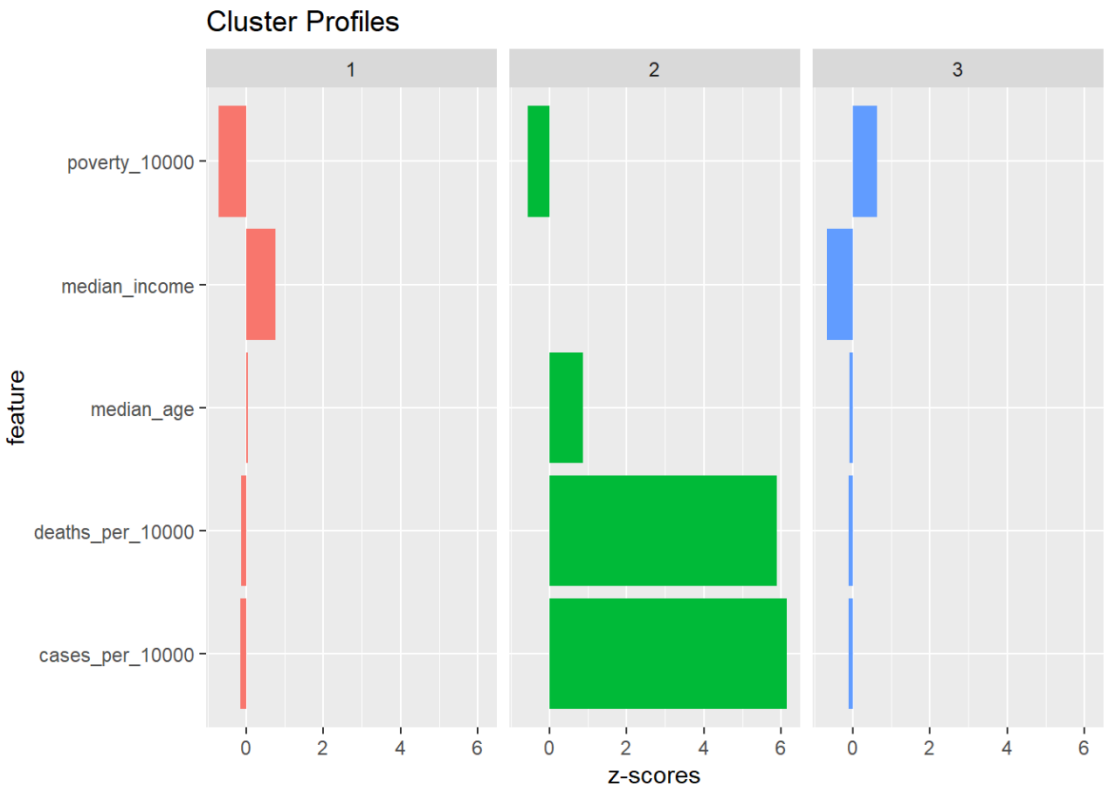


Figure 2: Cluster Profile using Demographic and Poverty Features

Figure 2 focuses on demographic and poverty-related features, alongside normalized COVID-19 indicators, to examine how socio-economic factors correlate with pandemic outcomes across counties. Variables such as cases and deaths were standardized per 10,000 residents (cases_per_10000, deaths_per_10000), and poverty rates were normalized (poverty_10000) to account for population size differences, ensuring a meaningful comparison across regions. Cluster 1 is distinguished by higher poverty rates and lower median income and age, yet it shows lower levels of COVID-19 cases and deaths, which may indicate underreporting or limited testing access in economically disadvantaged areas. Cluster 2 stands out with the highest levels of COVID-19 cases and deaths per 10,000 residents, combined with moderate poverty and lower median income and age, suggesting that economically vulnerable communities experienced disproportionate pandemic burdens. Cluster 3 exhibits higher median income and age with lower poverty rates, and correspondingly, the lowest COVID-19 cases and deaths per 10,000 residents, pointing to better resource availability and potential protection against severe pandemic impact. Overall, this clustering highlights a clear socio-economic gradient in COVID-19 outcomes, reinforcing how poverty and demographic factors intertwine with pandemic vulnerability.

Figure 3: Cluster Profile using Demographic and Mobility Features

Figure 3 focuses on the intersection of demographic characteristics and mobility patterns, alongside normalized COVID-19 impact indicators, to understand how regional behaviors and population traits relate to pandemic outcomes. COVID-19 cases and deaths were normalized per 10,000 residents to ensure comparability across counties of varying sizes. Mobility variables reflect percent changes from pre-pandemic baselines across workplaces, retail and recreation venues, grocery and pharmacy locations, and residential areas, providing insights into behavioral shifts during the pandemic.

Cluster 1 shows moderate decreases in mobility to workplaces and retail locations, and slight increases in residential mobility, suggesting cautious behavioral adaptations during the pandemic. This cluster also displays lower median income and age, but notably lower COVID-19 cases and deaths per 10,000 residents, which could indicate either effective behavioral mitigation or potential underreporting in lower-income areas. Cluster 2 is relatively balanced across all features, with minimal deviations in mobility patterns, demographic variables, and COVID-19 impact. This suggests a more stable or average profile, potentially representing regions with moderate restrictions and moderate pandemic impact. Cluster 3 is the most distinct, characterized by sharp increases in residential mobility (indicating people staying at home more) and significant declines in workplace and retail mobility. This cluster also records the highest COVID-19 case and death rates per 10,000 residents, coupled with higher median income and age. These patterns may reflect urban or higher-income areas where residents were more responsive to mobility restrictions but still experienced substantial pandemic burden, possibly due to higher population density or more widespread testing.

Overall, this clustering highlights how demographic factors and changes in population mobility collectively shaped the spread and impact of COVID-19 across counties, emphasizing the role of behavior adaptation in pandemic dynamics.

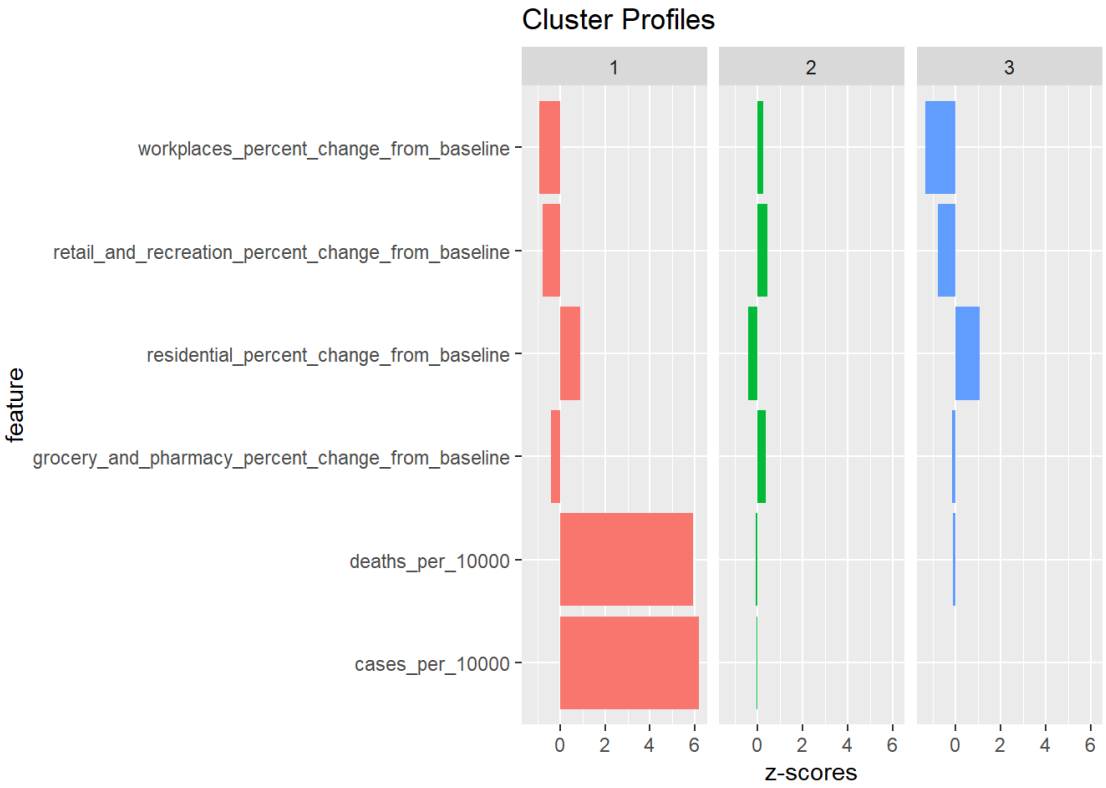


Figure 4: Cluster Profile using Mobility Features

Figure 4 focuses exclusively on mobility behaviors and their relationship with COVID-19 outcomes, using normalized case and death rates (per 10,000 residents) to enable fair comparisons across counties of different population sizes. Mobility variables capture percent changes from pre-pandemic baselines across key categories, including workplaces, retail and recreation, grocery and pharmacy, and residential areas, providing insight into regional behavioral responses during the pandemic.

Cluster 1 is defined by the highest COVID-19 burden, with elevated cases and deaths per 10,000 residents. Interestingly, this cluster shows relatively modest reductions in workplace and retail mobility, alongside limited increases in residential mobility, suggesting that counties in this group may have experienced less behavioral adaptation, potentially contributing to higher transmission rates. Cluster 2 demonstrates moderate reductions in workplace and retail mobility, coupled with small increases in residential mobility, while maintaining lower levels of COVID-19 cases and deaths. This indicates a more balanced response, where moderate adherence to mobility restrictions may have contributed to mitigating viral spread. Cluster 3 is the most behaviorally distinct, with the most substantial reductions in workplace and retail mobility and the largest increases in residential mobility, signaling strong behavioral adaptation to pandemic conditions. Correspondingly, this cluster exhibits the lowest COVID-19 case and death rates, suggesting that higher mobility reductions were effective in limiting viral transmission.

Overall, this clustering highlights the critical role that mobility patterns played in shaping the trajectory of COVID-19 across regions. Counties that demonstrated stronger reductions in mobility tended to experience better health outcomes, reinforcing the importance of behavioral public health measures in pandemic management.

2. Hierarchical Clustering

In this analysis, we applied Hierarchical clustering to the same feature subsets explored in the K-Means analysis, including demographic variables, household composition, poverty indicators, mobility patterns, and normalized COVID-19 impact measures, to determine whether this method provides improved cluster separation and interpretability. We used **Euclidean distance** as the similarity measure to compute pairwise dissimilarities between data points, and employed **Ward's linkage method** to minimize the total within-cluster variance during hierarchical agglomeration. Dendrograms are included in the Appendix.

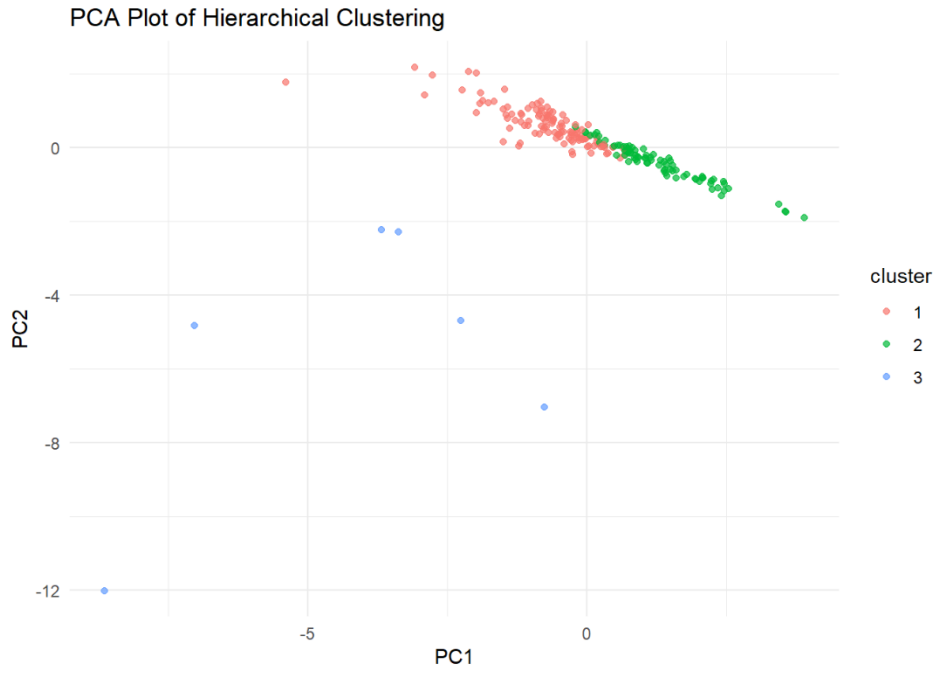


Figure 5: Clustering using Demographic and Household Features

Figure 5 focuses on demographic and household characteristics to explore regional patterns, with COVID-19 cases and deaths per 10,000 residents included for completeness and accuracy of the dataset, but not as the primary drivers of clustering. The clustering reveals three distinct groupings of counties based on median income, median age, and household composition.

Cluster 1 consists predominantly of counties with lower median income and younger populations, combined with higher proportions of family households. This profile may reflect more rural or lower-income suburban areas, where family-centric living arrangements are more common. While not the focus, these areas may also face challenges in health outcomes due to socioeconomic vulnerabilities. Cluster 2 represents counties with moderate median income and age profiles, along with a more balanced mix of family and non-family households. This cluster suggests a middle-ground demographic, potentially suburban regions with a stable household composition. In the context of COVID-19, this group likely experienced average exposure and outcomes, aligning with their demographic and household structures. Cluster 3 is characterized by higher median income and older populations, alongside a greater prevalence of non-family households. These areas likely correspond to more urban or affluent regions with a higher share of individuals living alone or in non-family arrangements. Notably, the points in Cluster 3 appear more dispersed in the PCA plot, which suggests greater internal diversity within this group. This spread likely reflects variations in urban household dynamics and income distribution, as affluent areas can still vary widely in household size, living arrangements, and demographic makeup. For example, this cluster may include both high-density urban centers with a concentration of single-person households and wealthier suburban regions with older populations but differing household compositions. This heterogeneity contributes to the broader spatial distribution observed in Cluster 3.

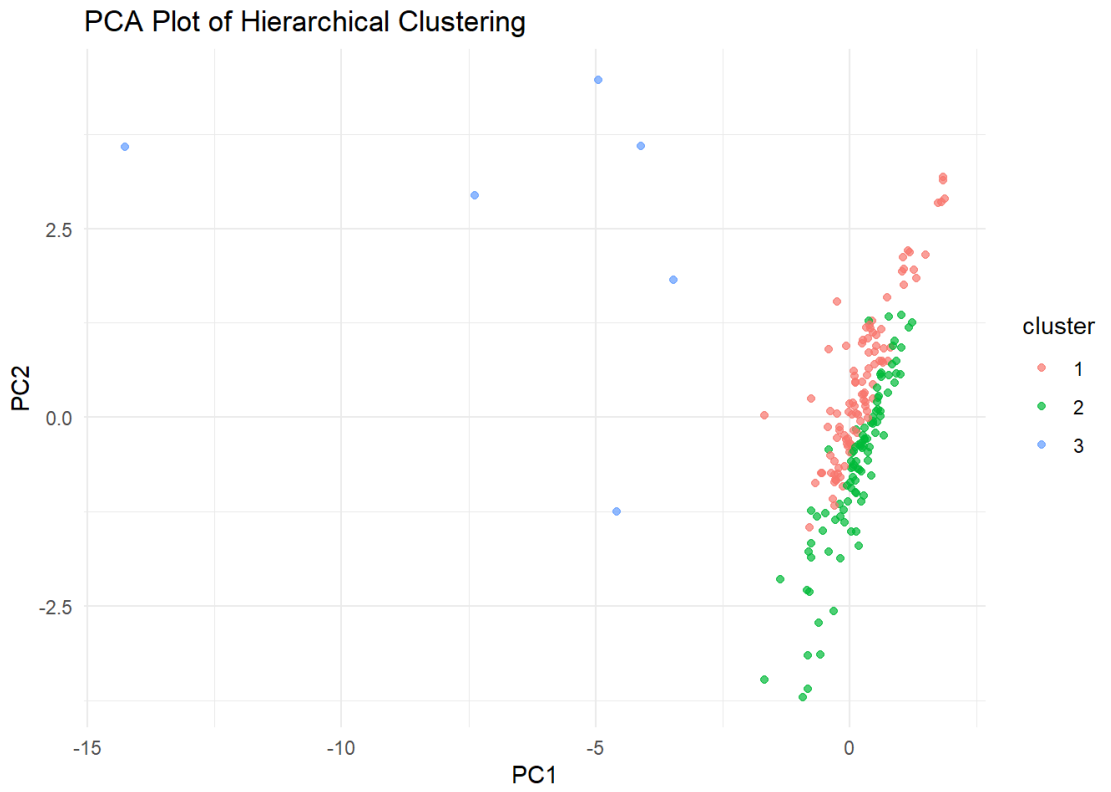


Figure 6: Clustering using Demographic and Poverty Features

Figure 6 centers on demographic and poverty-related features to explore socio-economic patterns across counties, with normalized COVID-19 cases and deaths per 10,000 residents included to ensure data completeness and comparability, but not as primary clustering drivers. The analysis produced three distinct clusters based on median income, median age, and poverty levels.

Cluster 1 comprises counties with higher poverty rates and lower median income and age, potentially reflecting economically disadvantaged regions with younger populations. Although COVID-19 data were not the focus, these areas might face compounded vulnerabilities during public health crises due to limited resources and higher socioeconomic risks. Cluster 2 includes counties with moderate poverty rates and middle-range median income and age profiles, indicating communities with relatively balanced socio-economic conditions. This cluster likely represents a socio-economic middle ground, providing a useful reference point between the extremes represented by other clusters. Cluster 3 is characterized by higher median income and age, along with lower poverty rates, suggesting more affluent and potentially older populations. While these regions may benefit from better access to healthcare and resources, their higher age profiles could still contribute to increased vulnerability in the context of COVID-19, though this was not the primary focus of the clustering. Notably, in the PCA plot, Cluster 3 appears more dispersed compared to the other clusters. This spread suggests greater internal variability within affluent counties, which may arise from differing socio-economic dynamics such as variations in poverty pockets within otherwise wealthier areas, or disparities in population age distributions. Even within generally affluent regions, differences in economic inequality or demographic composition can lead to greater diversity in the data, resulting in a looser cluster formation.

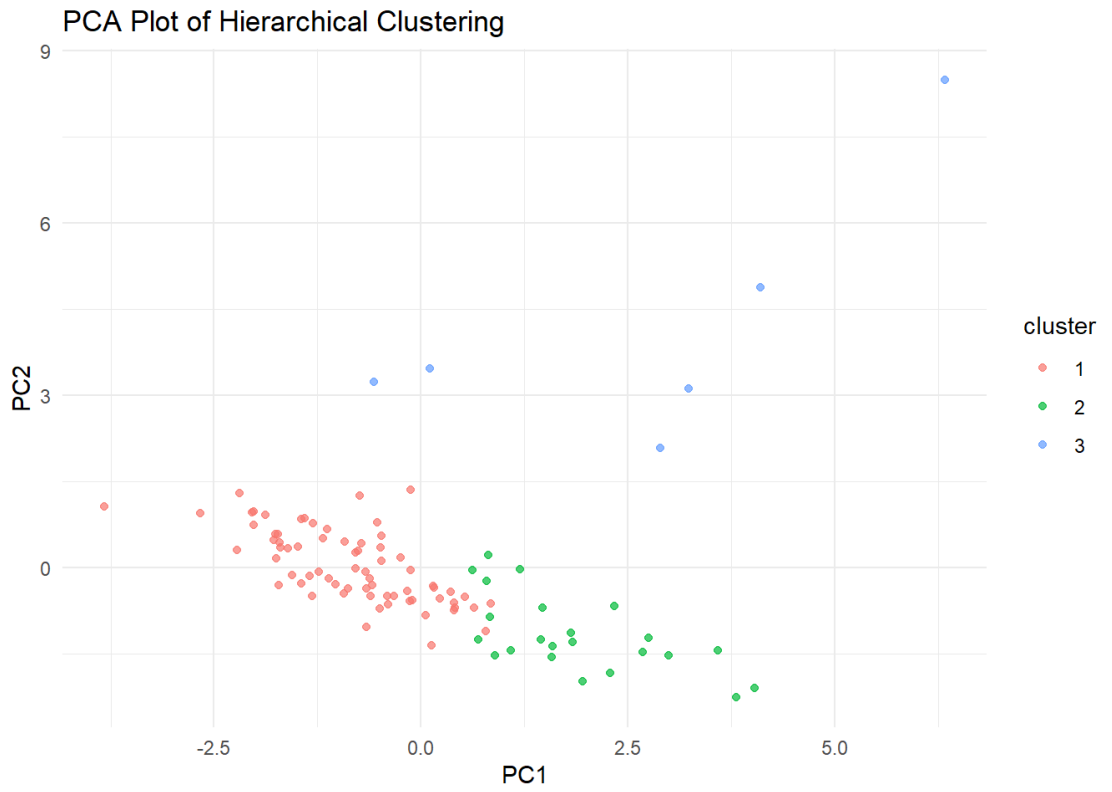


Figure 7: Clustering using Demographic and Mobility Features

Figure 7 focuses on understanding how demographic and mobility patterns shape regional profiles, with COVID-19 cases and deaths per 10,000 residents included for accuracy and completeness but not as primary clustering variables. The analysis resulted in three distinct clusters based on mobility behaviors and demographic characteristics.

Cluster 1 comprises counties exhibiting moderate declines in workplace and retail mobility, paired with increases in residential mobility, indicating behavioral adjustments in response to the pandemic. These counties generally have lower median income and younger populations, suggesting that economic constraints may have influenced the extent of mobility reductions and behavioral adaptation. Cluster 2 displays somewhat greater reductions in workplace and retail mobility, along with notable increases in residential mobility, reflecting regions that more actively reduced movement during the pandemic. This cluster aligns with moderate median incomes and ages, suggesting a more balanced demographic profile with effective behavioral adaptations. Cluster 3 is characterized by the most significant decreases in workplace and retail mobility and substantial increases in residential mobility, indicating strong adherence to mobility restrictions. This cluster also corresponds to higher median incomes and older populations, potentially representing more affluent regions with greater capacity for remote work and mobility reduction. Notably, the points in Cluster 3 appear more dispersed in the PCA plot, suggesting greater internal variability in mobility responses. This spread could be due to differences within higher-income regions, where some counties may have implemented strict mobility reductions, while others maintained moderate activity due to essential workforce demands or varying local policies, leading to less uniformity within the cluster.

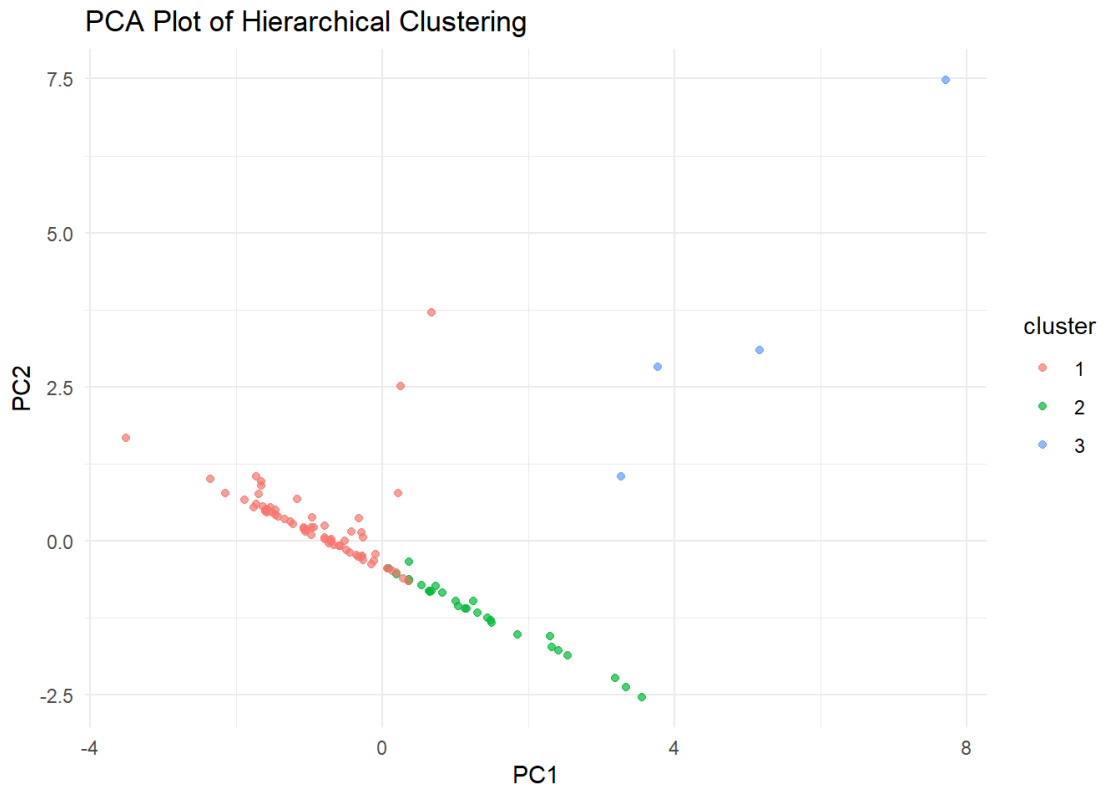


Figure 8: Clustering using Mobility Features

Figure 8 explores patterns in regional mobility behaviors, using COVID-19 cases and deaths per 10,000 residents for contextual accuracy but not as primary clustering variables. Mobility features reflect percent changes from pre-pandemic baselines across workplaces, retail and recreation venues, grocery and pharmacy locations, and residential areas. We applied Euclidean distance to measure dissimilarity and Ward's linkage method to minimize within-cluster variance during clustering, resulting in three distinct mobility behavior clusters.

Cluster 1 is characterized by moderate reductions in workplace and retail mobility and modest increases in residential mobility, reflecting regions that displayed some behavioral adaptation to pandemic conditions but not extensive changes. This cluster likely includes areas with fewer restrictions or more essential workforce presence. Cluster 2 exhibits stronger mobility shifts, with substantial declines in workplace and retail mobility and higher increases in residential mobility, indicating regions that may have implemented stricter mobility restrictions or whose populations were more responsive to public health guidelines. Cluster 3, however, stands out for its greater internal variability. While the cluster generally reflects lower or inconsistent mobility changes across the selected categories, the points are more dispersed in the PCA space. This dispersion suggests that Cluster 3 includes a mix of counties with diverse mobility patterns — potentially combining both rural areas with inherently lower baseline mobility changes and urban areas with variable compliance or behavioral responses. The heterogeneity in mobility responses within this cluster could stem from differences in policy enforcement, population density, or essential workforce composition, leading to less cohesion in movement patterns.

3. Hierarchical Clustering via CURE

In this analysis, we applied a hierarchical clustering approach inspired by the CURE (Clustering Using REpresentatives) method to evaluate whether it improves upon previous clustering techniques. Using the same four feature subsets as in our earlier analyses— demographics, household composition, poverty indicators, and mobility patterns— we aimed to compare the clustering performance of this method against K-Means and standard Hierarchical clustering. For consistency and comparability, we used **Euclidean distance** as the similarity measure to compute dissimilarities between counties, and employed **Ward's linkage method** to minimize total within-cluster variance during agglomeration. Dendrograms are included in the Appendix.

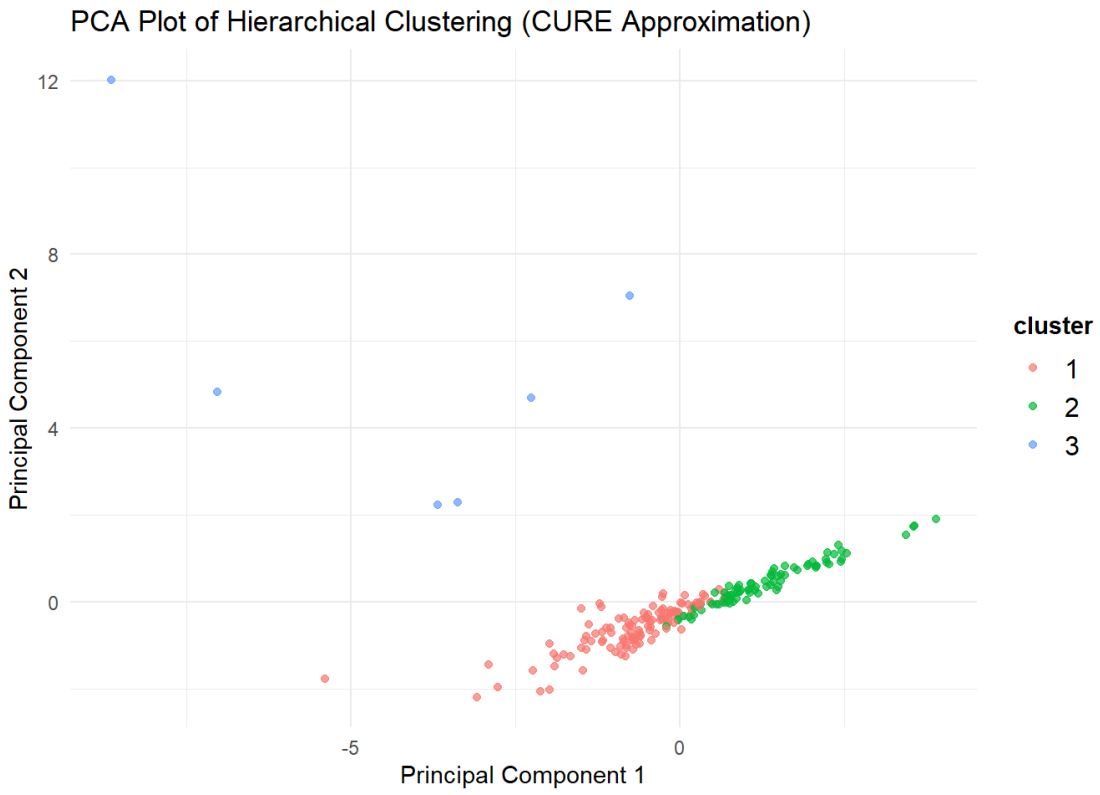


Figure 9: Clustering (CURE Approximation) using Demographic and Household Features

Figure 9 presents the results of hierarchical clustering using a CURE approximation method, focusing on demographic and household composition variables to explore regional patterns. While COVID-19 cases and deaths per 10,000 residents were included to improve data accuracy and completeness, the primary emphasis of the clustering remains on demographic factors such as median income, median age, and household structure (family and non-family households normalized per 10,000 residents). The analysis produced three distinct clusters.

Cluster 1 is composed of counties with lower median incomes, younger populations, and higher proportions of family households. This profile suggests more rural or economically constrained regions, where larger family units are more common and socio-economic vulnerabilities may be present. Cluster 2 features counties with moderate median incomes and age profiles, alongside a balanced household composition. These counties likely represent more typical suburban or semi-urban areas with diverse household structures and relatively stable socio-economic conditions. Cluster 3 is characterized by higher median income and age, along with a higher concentration of non-family households. This cluster likely captures wealthier, possibly urban counties with more individual or non-family living arrangements. However, the PCA plot reveals that points within Cluster 3 are notably more dispersed compared to the other clusters. This spread indicates greater internal heterogeneity

within this cluster, which may stem from a combination of urban areas with varying population densities, differences in household structures within affluent regions, or differing policy responses and socio-economic dynamics. While these counties generally share higher income and age profiles, local variations such as differences in housing affordability, employment sectors (e.g., essential vs. remote-capable industries), and population mobility likely contribute to the observed dispersion.

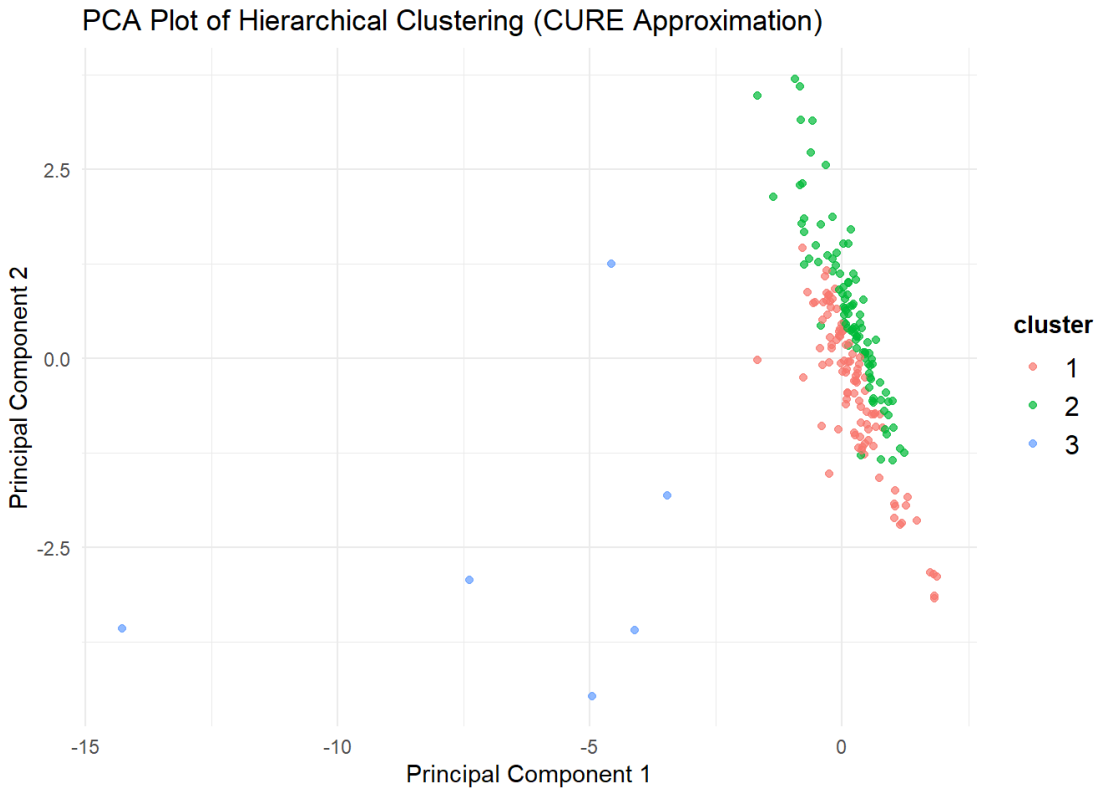


Figure 10: Clustering (CURE Approximation) using Demographic and Poverty Features

Figure 10 displays the results of hierarchical clustering using a CURE approximation method, focusing on demographic and poverty indicators to explore socio-economic patterns across counties. While normalized COVID-19 cases and deaths per 10,000 residents were included to improve data completeness, the primary emphasis remains on demographic factors such as median income, median age, and poverty levels (normalized per 10,000 residents). The clustering resulted in three distinct clusters.

Cluster 1 comprises counties with higher poverty rates, lower median incomes, and younger populations. This grouping likely reflects economically disadvantaged areas with heightened vulnerabilities, which could compound challenges during public health crises, even though COVID-19 data were not the main drivers of this clustering. Cluster 2 includes counties with moderate poverty levels, median incomes, and age profiles, representing regions with relatively balanced socio-economic conditions. These counties might serve as a middle ground, not exhibiting the extremes seen in the other clusters. Cluster 3 is characterized by lower poverty rates and higher median incomes and ages, suggesting more affluent counties with older populations. However, the PCA plot reveals that Cluster 3 is more dispersed compared to the other clusters. This spread indicates greater internal diversity within this group, which could arise from differences among affluent counties — for example, urban centers with varying levels of income inequality, or rural wealthier areas with distinct demographic structures. Even though these counties share general socio-economic advantages, local variations such as differences in aging populations, housing affordability, or regional economic compositions (e.g., tourism, services, or resource-based economies) contribute to this wider spread. This diversity within Cluster 3 suggests

that while income and poverty levels group these counties together broadly, internal heterogeneity remains significant.

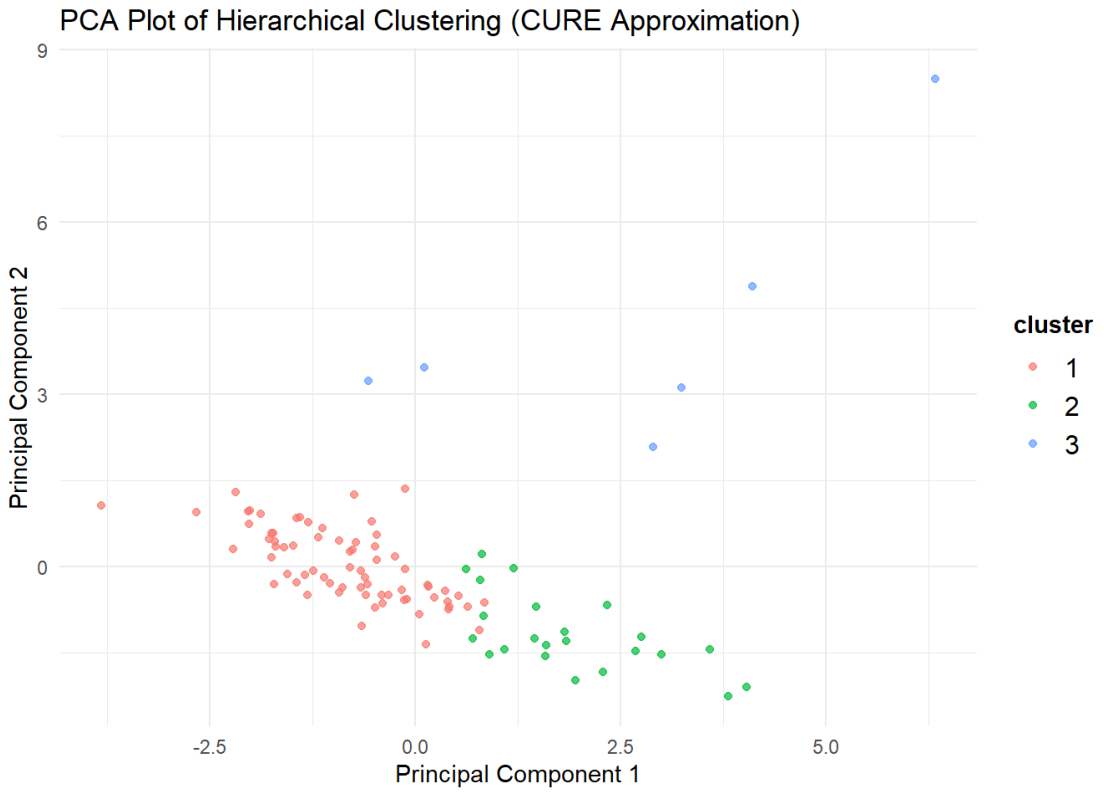


Figure 11: Clustering (CURE Approximation) using Demographic and Mobility Features

Figure 11 presents the results of hierarchical clustering using a CURE approximation method, focusing on demographic and mobility variables to explore regional patterns in behavioral adaptation to the pandemic. While normalized COVID-19 cases and deaths per 10,000 residents were included for completeness and accuracy, they were not primary drivers of clustering. Instead, the analysis prioritized demographic factors like median income and age, alongside mobility patterns capturing changes in workplace, retail and recreation, grocery, and residential activities compared to pre-pandemic baselines. The analysis identified three distinct clusters.

Cluster 1 includes counties with lower median incomes and younger populations, displaying moderate reductions in workplace and retail mobility and slight increases in residential mobility. This suggests limited behavioral shifts, potentially due to economic constraints or higher proportions of essential workers unable to reduce mobility substantially. Cluster 2 consists of counties with moderate median incomes and ages, showing more substantial reductions in workplace and retail mobility, and greater increases in residential mobility. This profile points to regions with more capacity for behavioral adjustment, likely reflecting a balance between public health response and socio-economic factors. Cluster 3 is defined by higher median incomes and older populations, coupled with the most pronounced mobility changes — significant reductions in workplace and retail activity, and notable increases in residential mobility. However, the PCA plot reveals that Cluster 3's points are more widely dispersed compared to the other clusters. This spread indicates greater internal heterogeneity within this group. Affluent regions often encompass a diverse range of community profiles: urban centers with variable population densities and occupational structures, suburban areas with differing remote work capabilities, and retirement communities with naturally reduced mobility. Variations in local policy responses, public health messaging, and levels of essential workforce engagement could all contribute to the spread observed in this cluster. Despite a shared tendency toward higher socio-economic status and strong mobility reduction, the underlying local differences manifest as dispersion in the PCA space.

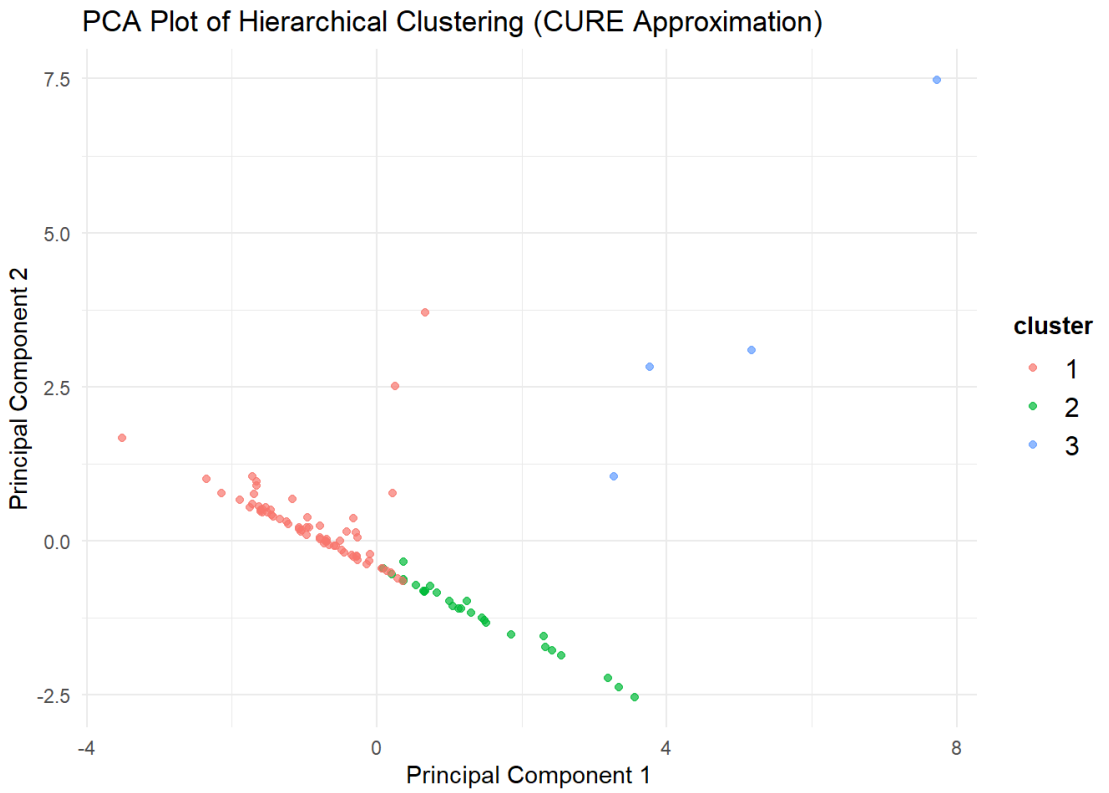


Figure 12: Clustering (CURE Approximation) using Mobility Features

Figure 12 presents the results of hierarchical clustering using a CURE approximation method, focusing exclusively on mobility variables to understand regional behavioral patterns during the pandemic. While COVID-19 cases and deaths per 10,000 residents were included to maintain accuracy and context, the clustering primarily reflects differences in mobility behaviors, including changes in workplace, retail and recreation, grocery and pharmacy, and residential movement compared to pre-pandemic baselines. The analysis identified three clusters.

Cluster 1 includes counties with moderate decreases in workplace and retail mobility and slight increases in residential mobility. The points in this cluster appear somewhat spread out, indicating internal variation in mobility changes among counties with similar behavioral profiles. This dispersion likely reflects regional differences in economic reliance on in-person workforces or varied policy enforcement — for example, some counties within this cluster may have essential industries requiring sustained mobility, while others reduced movement to a greater extent. Cluster 2 forms a tighter grouping, characterized by substantial reductions in workplace and retail mobility and pronounced increases in residential mobility. This suggests consistent behavioral adaptation across these counties, possibly due to more uniform policy measures, higher remote work capabilities, or stronger public health messaging. Cluster 3 displays more extreme shifts in mobility — significant declines in workplace and retail activity and large increases in residential mobility — but like Cluster 1, the points are spread out in PCA space. This dispersion indicates internal heterogeneity within the cluster. It may capture a mix of highly urbanized counties with strict lockdowns and rural or suburban areas where mobility reductions varied based on local socio-economic conditions, availability of essential services, or differences in adherence to public health guidelines. The variation also reflects that even when overall trends are similar (i.e., strong mobility reduction), the degree and nature of those changes can differ widely across regions.

4. DBSCAN Clustering

In this analysis, we applied DBSCAN (Density-Based Spatial Clustering of Applications with Noise) to the same four feature subsets used in the other clustering methods — demographics, household composition, poverty indicators, and mobility patterns — to enable a direct comparison across all techniques. DBSCAN does not require pre-specifying the number of clusters and is particularly well-suited for identifying clusters of varying shapes and detecting noise or outliers in the data, which makes it a valuable complement to methods like K-Means and Hierarchical clustering. For measuring similarity, we used **Euclidean distance**, which is appropriate given that all variables were normalized and scaled to ensure comparability. Euclidean distance allows DBSCAN to effectively compute point densities and define neighborhood thresholds for cluster formation. By applying DBSCAN across these standardized feature sets, we aimed to evaluate whether a density-based approach could uncover more nuanced or irregular cluster structures compared to the partitioning and hierarchical methods, ultimately helping to identify which method provided the most accurate and meaningful segmentation of counties.

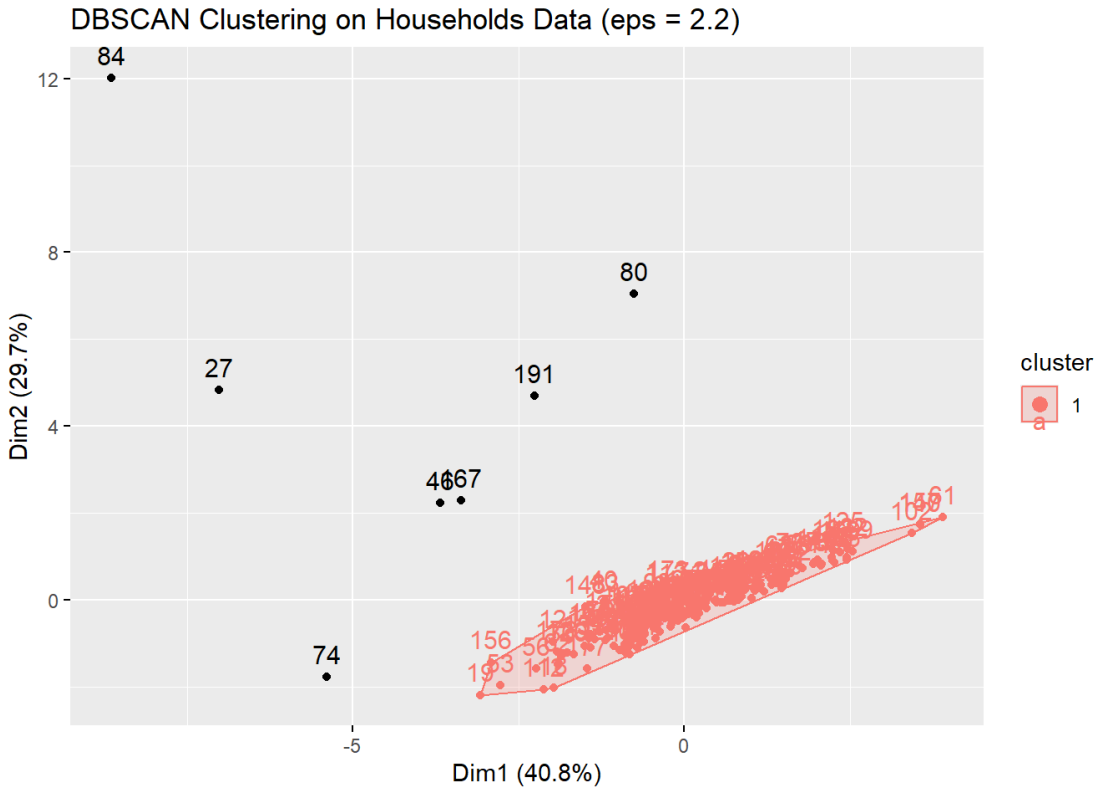


Figure 13: DBSCAN Clustering using Demographic and Household Features

Figure 13 presents the results of DBSCAN clustering applied to demographic and household variables, with normalized COVID-19 cases and deaths per 10,000 residents included to ensure data completeness and comparability. The feature set focuses primarily on understanding socio-economic patterns at the county level, considering factors like median income, median age, and household composition normalized to adjust for varying county sizes. DBSCAN was selected for its ability to detect clusters of varying shapes and densities without the need to pre-define the number of clusters. We used Euclidean distance for similarity measurement, appropriate for our scaled data, and set the parameters to eps = 2.2 and minPts = 7.

The resulting output indicates that DBSCAN primarily formed a single dense cluster, with a small number of points identified as noise or outliers (shown in black). In this case, demographic and household variables, along with normalized COVID-19 rates, do not display sharp density contrasts across counties. Instead, the variables form a continuous gradient, especially after scaling. Because of this, DBSCAN does not detect multiple distinct dense areas but rather treats the data as one broadly connected cluster, with only sparse regions falling outside the density threshold and being labeled as noise. To potentially reveal more structure, a smaller eps value could be tested to tighten the neighborhood definition and expose finer-grained density variations. However, it's also important to note that with socio-economic variables, which often form smooth distributions rather than tight groupings, density-based methods like DBSCAN may have limited effectiveness compared to partitioning (K-Means) or Hierarchical methods.

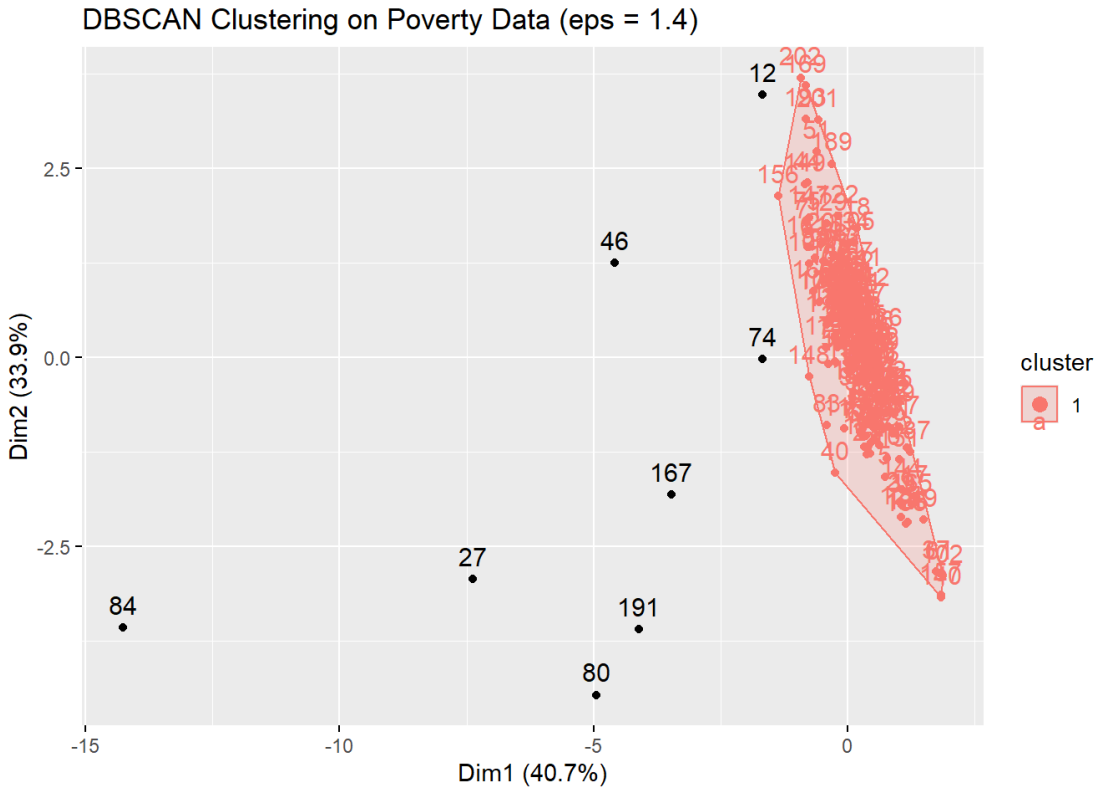


Figure 14: DBSCAN Clustering using Demographic and Poverty Features

Figure 14 presents the results of DBSCAN clustering applied to demographic and poverty-related variables, with normalized COVID-19 cases and deaths per 10,000 residents included to ensure data completeness and comparability. The analysis primarily focuses on understanding socio-economic patterns across counties, using variables such as median income, median age, and poverty rates per 10,000 residents. DBSCAN, which is well-suited for detecting clusters of varying densities and identifying outliers, was used with an eps value of 1.4 and minPts of 7.

The clustering output reveals a single dominant cluster encompassing most counties, accompanied by a few points marked as noise or outliers (shown in black). This result suggests that the distribution of demographic and poverty variables across counties is largely continuous, with no sharp density breaks that DBSCAN could use to form distinct clusters. While eps controls the neighborhood radius and influences the density threshold for cluster formation, the primary factor at play here is the smooth and gradual nature of socio-economic variation across regions. Even though a moderate eps value of 1.4 was used (lower than in some previous runs), the data does not exhibit distinct high-density regions separated by sparse areas— a condition that DBSCAN relies on to form multiple clusters.

As a result, DBSCAN effectively treats the data as part of one broadly connected socio-economic landscape, with only a handful of counties falling outside the density threshold and being labeled as noise. These outliers likely represent counties with extreme demographic or poverty characteristics, such as very low or very high poverty rates relative to the national distribution. While DBSCAN remains valuable for detecting irregular cluster shapes and noise, alternative approaches such as K-Means or Hierarchical clustering may provide clearer insights for datasets with gradual socio-economic transitions like this one.

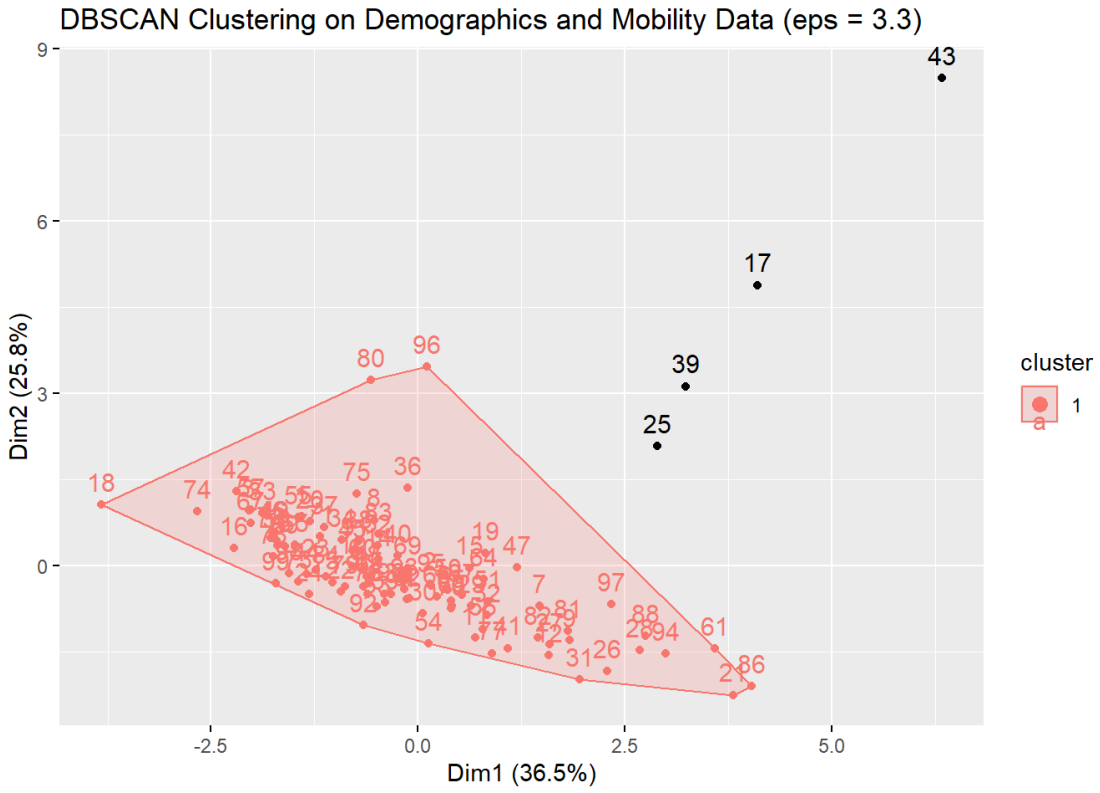


Figure 15: DBSCAN Clustering using Demographic and Mobility Features

Figure 15 presents the results of DBSCAN clustering applied to a combination of demographic and mobility variables, with normalized COVID-19 cases and deaths per 10,000 residents included for data completeness. The feature set focuses on exploring how demographic characteristics, such as median income and median age, intersect with mobility behaviors during the pandemic, including changes in workplace, retail, grocery, and residential movement patterns. DBSCAN was selected for its ability to detect clusters of varying shapes and identify noise points without needing to pre-define the number of clusters. With an eps value of 3.3 and minPts set at 7, Euclidean distance was used as the similarity measure, well-suited for scaled and normalized data.

The clustering result shows that DBSCAN identified a single dominant cluster, encompassing most counties, along with a few scattered points classified as noise or outliers (displayed in black). This pattern indicates that, even after combining mobility behaviors and demographic characteristics, the data distribution remains relatively continuous without distinct high-density separations. While the chosen eps value contributes to this outcome by setting a broad neighborhood radius, the primary factor influencing the result is the inherent structure of the data itself. In this case, counties across the United States exhibit gradual, continuous variation in both demographic and mobility variables, rather than forming sharply defined groups based on density.

The few noise points observed likely correspond to counties with extreme demographic or mobility patterns, such as exceptionally high changes in residential mobility or outlier demographic profiles. These counties fall outside the typical density threshold established by DBSCAN and are appropriately labeled as noise. Overall, this result highlights an important insight: when working with naturally continuous data like demographics and mobility patterns, density-based clustering methods like DBSCAN may struggle to detect multiple distinct clusters. The analysis reinforces the notion that while DBSCAN is highly effective for datasets with clear density separations, partitioning methods like K-Means or hierarchical approaches may offer clearer segmentation when analyzing socio-economic and behavioral data with smooth gradients. Including COVID-19 metrics helped to maintain data accuracy but did not substantially affect the clustering outcome.

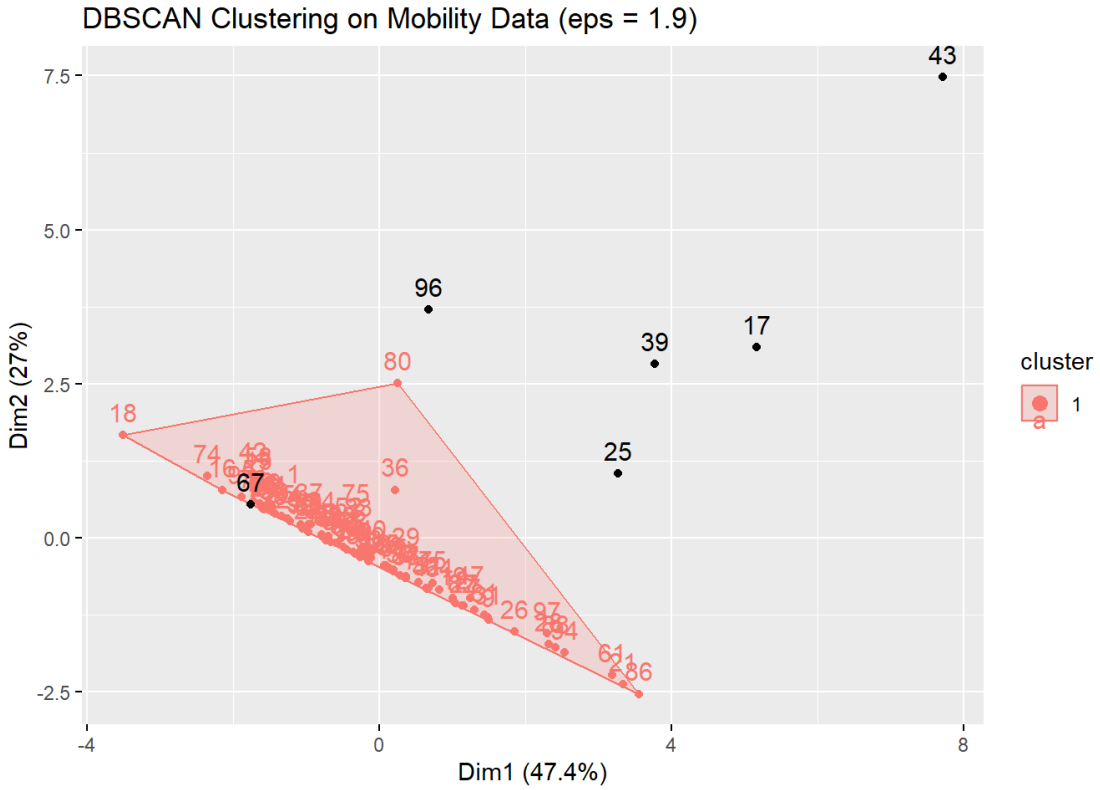


Figure 16: DBSCAN Clustering using Mobility Features

Figure 16 displays the results of DBSCAN clustering focused exclusively on mobility variables, with normalized COVID-19 cases and deaths per 10,000 residents included for completeness and accuracy but not as primary clustering drivers. The mobility features reflect behavioral adaptations during the pandemic, capturing shifts in workplace, retail and recreation, grocery, and residential activities relative to pre-pandemic baselines. Using Euclidean distance as the similarity measure — appropriate for the scaled and normalized dataset — DBSCAN was applied with an eps value of 1.9 and minPts of 7 to explore density-based patterns in mobility behaviors.

The clustering outcome shows that DBSCAN identified a single dominant cluster, encompassing most of the counties, along with a few scattered points labeled as noise (displayed in black). This result suggests that across U.S. counties, mobility behavior changes formed a relatively continuous spectrum, without distinct density breaks that would lead to the formation of multiple clusters. While the eps parameter controls the neighborhood radius, and a moderately large value of 1.9 was used here, the primary influence on this outcome is the nature of the data itself. Regional mobility patterns, shaped by varying local restrictions, behavioral adaptations, and economic necessities, tend to vary along smooth gradients rather than forming tightly separated groups.

The few noise points likely correspond to counties with extreme mobility shifts, either due to stringent lockdowns, unique socio-economic factors, or minimal behavioral change despite broader trends. These counties fall outside the density threshold defined by the eps value and are appropriately treated as outliers. Overall, this clustering reinforces the observation that DBSCAN may be less effective in distinguishing clear groupings within mobility data, which is inherently continuous and lacks sharp density-based separations. Although the method remains valuable for detecting irregularly shaped clusters and identifying noise, alternative clustering methods such as K-Means or Hierarchical clustering may offer clearer insights for analyzing mobility behaviors.

Determining a Suitable Number of Clusters

To determine the appropriate number of clusters for K-Means clustering, we utilized the elbow method, a widely used heuristic that evaluates the total within-cluster sum of squares (WCSS) across different values of k. By plotting WCSS against increasing values of k, we observed that the graph (refer to Appendix) showed a distinct "elbow" at k = 3, indicating a point of diminishing returns. At this point, adding additional clusters yielded only marginal improvements in compactness, suggesting that three clusters effectively capture the underlying structure of the data without overfitting. This visual inflection point guided our selection of three clusters for all K-Means analyses across the different feature subsets.

For hierarchical clustering and the hierarchical CURE approximation, the number of clusters was determined through a combination of dendrogram inspection and domain knowledge about the data's socio-economic patterns. In hierarchical clustering, we applied Ward's method, which focuses on minimizing the total within-cluster variance as clusters are merged. By examining the dendrograms (included in the Appendix), we identified natural "cuts" in the tree at three clusters, where significant linkage distances occurred between merging steps. This indicated that three clusters were an appropriate balance between over-fragmentation and under-segmentation. Using the same approach for the hierarchical CURE approximation, we also selected three clusters for consistency and comparability, as the visual structure of the dendrogram and the cluster density patterns supported this choice. This allowed for a fair cross-method comparison while maintaining methodological rigor.

For DBSCAN, the process of determining the number of clusters differs fundamentally, as DBSCAN does not require the user to predefine the number of clusters. Instead, DBSCAN relies on two key parameters: eps (the neighborhood radius) and minPts (minimum points to form a dense region). We used a k-nearest neighbors (k-NN) distance plot to inform the selection of eps. By plotting the distance to the k-th nearest neighbor (where k = number of dimensions + 1), we identified an inflection point where distances rapidly increased, suggesting a suitable threshold for density separation. These selected eps values varied slightly across feature sets (e.g., 1.9 for mobility data, 2.2 for household data) but consistently aimed to balance between excessive fragmentation and overly coarse clusters. Notably, however, given the continuous nature of our data, DBSCAN often resulted in a single dominant cluster with a few outliers, illustrating its sensitivity to data density and its different conceptual approach to defining clusters compared to partitioning or hierarchical methods.

Unsupervised Evaluation (Silhouette Plots)

To compare the effectiveness of our clustering methods, we used silhouette plots and average silhouette widths as our unsupervised evaluation tools. These metrics allowed us to assess both the cohesion within clusters and the separation between them, providing a clear basis for comparing how well each method captured the structure of the data across different feature subsets.

During our evaluation, we identified a few important considerations. DBSCAN was not suitable for silhouette analysis, as it consistently produced a single large cluster with scattered noise points across all feature sets. Since silhouette plots require at least two clusters, we excluded DBSCAN from this comparison. This limitation further confirmed our earlier finding that the continuous nature of our socio-economic and mobility data limited DBSCAN's ability to form distinct clusters. As a result, our final analysis will focus on K-Means, Hierarchical, and Hierarchical CURE clustering.

Silhouette Plots for Demographic and Household Features

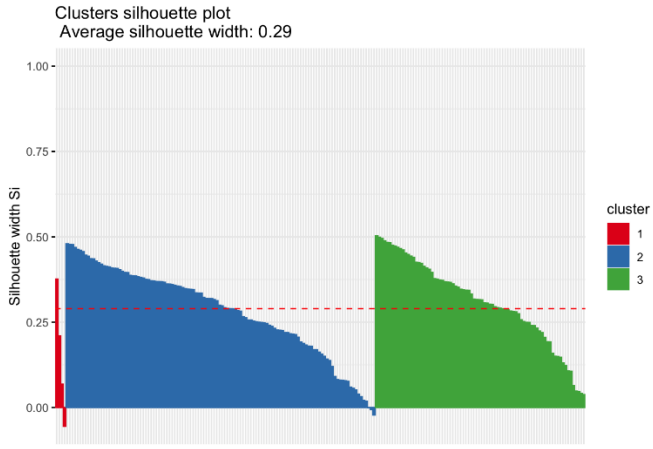


Figure 17: Demographic and Household Features (K-Means)

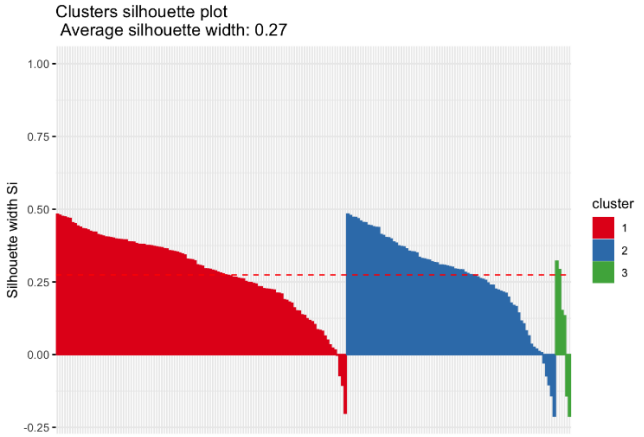


Figure 18: Demographic and Household Features (Hierarchical/CURE)

Figure 17 (K-Means) and Figure 18 (Hierarchical and Hierarchical CURE) present the silhouette analyses for the demographic and household feature subset. Both methods were evaluated using average silhouette width, which reflects the cohesion and separation of the clusters—higher values indicate tighter, better-separated clusters. Notably, since both Hierarchical clustering and the Hierarchical CURE approximation produced identical clustering structures, we represented both methods with a single silhouette plot in Figure 18 to avoid redundancy.

In this case, K-Means achieved an average silhouette width of 0.29, slightly outperforming hierarchical clustering, which recorded 0.27. Although the difference is modest, it suggests that K-Means produced clusters that are marginally more cohesive and better separated. Visually, the K-Means silhouette plot shows that the majority of points have positive silhouette widths across clusters 2 and 3, indicating reasonable assignment of data points to their respective clusters. Cluster 1 in K-Means is smaller but still maintains acceptable silhouette scores. Conversely, the Hierarchical clustering silhouette plot shows a wider spread of silhouette values, including a more noticeable portion of points with negative silhouette widths in cluster 1. Negative silhouette values indicate points that may have been misclassified or are closer to neighboring clusters than to their assigned cluster, reducing overall cluster quality. This points to less clear boundaries between clusters in the hierarchical approach, especially for the smallest cluster.

In the context of COVID-19 analysis, where we aim to understand how demographic and household structures (normalized by population size) may have influenced pandemic dynamics, **K-Means** provides a slightly clearer representation of the underlying patterns. The improved cohesion and separation seen in the K-Means clustering suggest that it better captures the natural groupings in socio-demographic characteristics and household compositions across counties. This, in turn, allows for more reliable interpretation of how factors like median income, age, and household types may relate to variations in COVID-19 impacts across regions.

Silhouette Plots for Demographic and Poverty Features

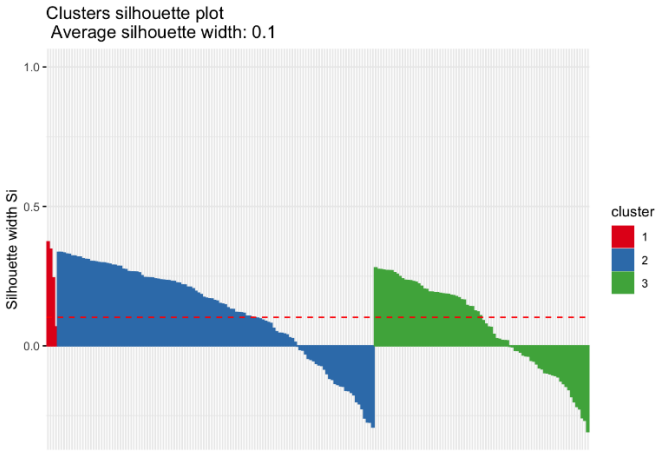


Figure 19: Demographic and Poverty Features (K-Means)

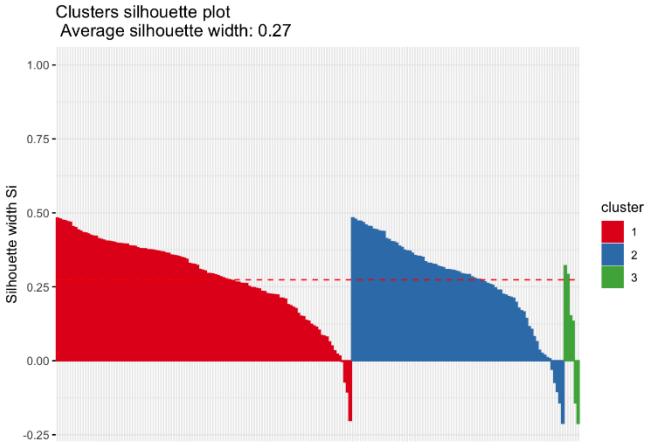


Figure 20: Demographic and Poverty Features (Hierarchical)

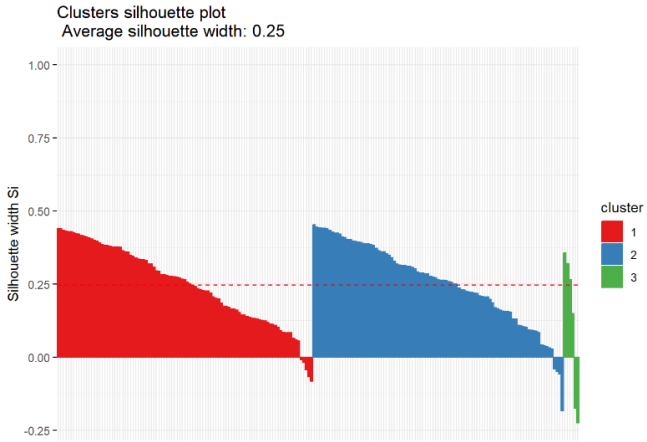


Figure 21: Demographic and Poverty Features (CURE)

Figures 19, 20, and 21 present silhouette plots for three clustering methods—K-Means, Hierarchical, and Hierarchical CURE—applied to the demographic and poverty feature subset. This subset includes variables such as median income, median age, and poverty per 10,000 residents, with COVID-19 cases and deaths included only to enhance accuracy, not to drive the clustering itself.

Among the three methods, Hierarchical Clustering (Figure 20) performed the best, achieving an average silhouette width of 0.27. This indicates well-defined and more cohesive clusters with relatively few negative silhouette values. Visually, Clusters 2 and 3 in the plot show strong internal consistency, suggesting that the hierarchical method effectively captured the underlying structure of the data. Hierarchical CURE (Figure 21) followed closely, with an average silhouette width of 0.25, also demonstrating good cohesion and separation. Although slightly less compact than standard hierarchical clustering, CURE still performed well in distinguishing socio-economic patterns. In contrast, K-Means Clustering (Figure 19) showed significantly poorer performance, with an average silhouette width of just 0.10. The plot reveals many negative silhouette values, particularly in Cluster 1, suggesting substantial overlap between clusters and possible misclassification. This result aligns with expectations, as K-Means tends to struggle with non-spherical or overlapping data distributions—characteristics common in socio-economic data.

In the context of COVID-19, where understanding structural differences across counties is critical, **Hierarchical** clustering proves to be the most suitable method for this subset. It captures nuanced distinctions in demographics and poverty that may influence public health outcomes, offering a more reliable foundation for further interpretation and policy analysis.

Silhouette Plots for Demographic and Mobility Features

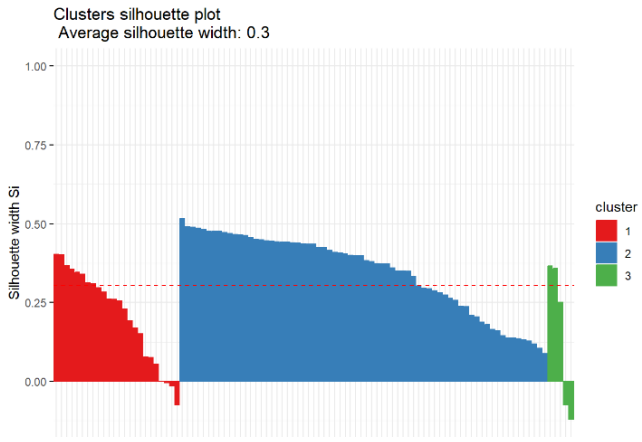


Figure 22: Demographic and Mobility Features (K-Means)

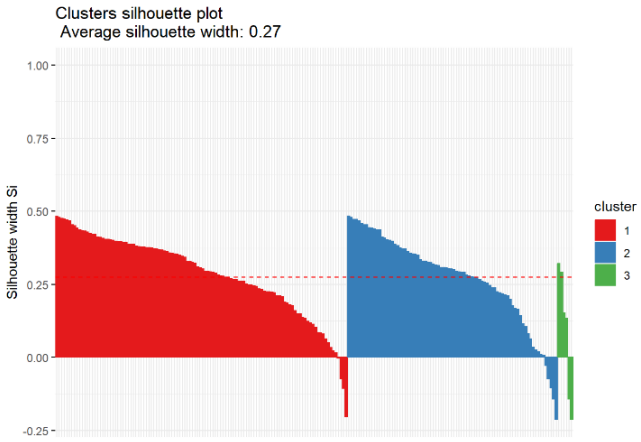


Figure 23: Demographic and Mobility Features (Hierarchical)

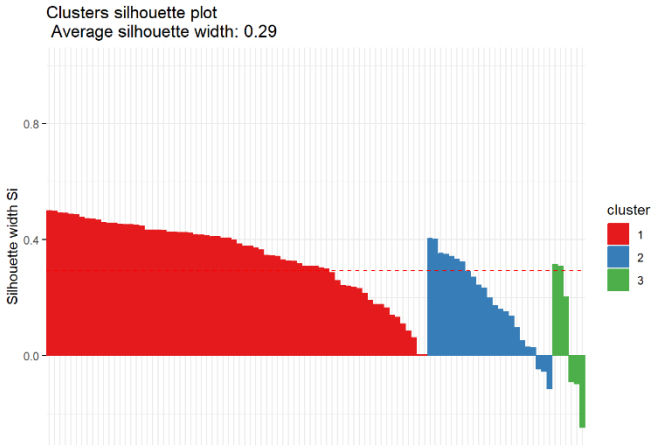


Figure 24: Demographic and Mobility Features (CURE)

Figures 22, 23, and 24 compare the silhouette plots for K-Means, Hierarchical, and Hierarchical CURE clustering methods applied to the demographic and mobility feature subset. This subset includes demographic variables such as median income and median age, alongside mobility indicators like workplace, retail and recreation, grocery and pharmacy, and residential mobility changes. While COVID-19 cases and deaths per 10,000 residents are included to improve normalization across counties, they are not the primary focus of clustering.

Among the three methods, K-Means clustering (Figure 22) emerges as the best performer, achieving the highest average silhouette width of 0.30. The silhouette plot shows a dominant presence of positive silhouette values

across clusters, particularly for Cluster 2, indicating strong internal cohesion and clear separation from neighboring clusters. This suggests that K-Means effectively captured meaningful patterns in the mobility behaviors and demographic characteristics, especially important for understanding how population segments adapted their movements during the pandemic. Hierarchical CURE clustering (Figure 24) follows closely with an average silhouette width of 0.29, which also reflects good clustering quality. The plot demonstrates fairly consistent positive silhouette values across clusters, particularly for Clusters 1 and 2, indicating reliable groupings. Although slightly behind K-Means, CURE's performance confirms its strength in handling more complex data shapes and producing well-defined clusters in this context. On the other hand, Hierarchical clustering (Figure 23) lags behind with an average silhouette width of 0.27. While this is still respectable, the plot reveals a broader distribution of silhouette values and a slightly higher frequency of negative values in Cluster 1, suggesting potential misclassifications and less distinct cluster boundaries.

Overall, for this demographic and mobility feature subset, **K-Means** offers the strongest clustering performance, with **Hierarchical CURE** providing a competitive alternative. Both methods outperform standard hierarchical clustering, highlighting the value of centroid-based and hybrid approaches when capturing complex interactions between socio-demographics and mobility changes during COVID-19.

Silhouette Plots for Mobility Features

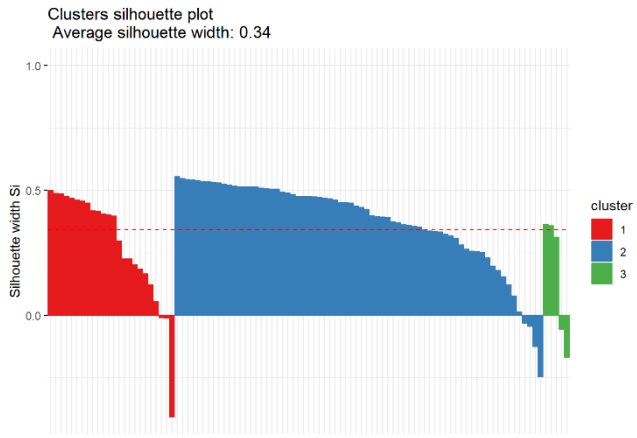


Figure 25: Mobility Features (K-Means)

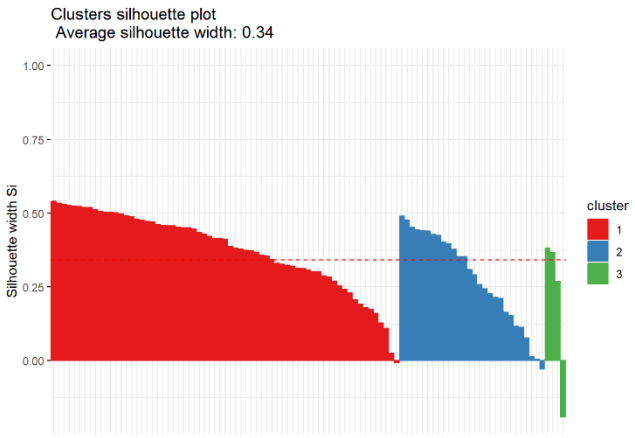


Figure 26: Mobility Features (Hierarchical/CURE)

Figures 25 and 26 display the silhouette plots for K-Means and Hierarchical/CURE clustering methods, applied to the mobility feature subset. This subset included workplace, retail and recreation, grocery and pharmacy, and residential mobility changes, with cases and deaths per 10,000 residents included for normalization but not as primary drivers of clustering. Notably, both K-Means and Hierarchical/CURE produced identical average silhouette widths of 0.34, indicating comparably strong clustering performance.

However, a closer inspection reveals subtle differences that help determine which method performs better. In Figure 25 (K-Means), the silhouette plot exhibits tightly clustered positive values for Clusters 2 and 3, suggesting strong internal cohesion and clear separation between mobility behavior patterns. Even Cluster 1, while smaller and with some lower scores, maintains positive silhouette values for most points, reinforcing the overall robustness of the K-Means clustering for this dataset. In Figure 26 (Hierarchical/CURE), while the average silhouette width matches K-Means numerically, the plot displays a slightly broader distribution of silhouette scores across clusters, with more variability especially in Cluster 1. There is a higher frequency of lower silhouette values compared to K-Means, indicating that while the method captured the overall structure well, it struggled slightly more with cluster compactness and separation.

Given these observations, **K-Means** emerges as the slightly superior method for clustering the mobility feature subset. Its clearer separation and tighter cohesion across clusters make it better suited for understanding mobility

patterns during the pandemic, especially when considering the nuances of human behavior changes such as reduced workplace and retail visits and increased residential stays. These insights are crucial for interpreting how public health measures and personal behavior adaptations varied across regions.

Supervised Evaluation

The supervised evaluation was conducted to assess the quality of clustering results across four feature subsets: households, poverty, demographic and mobility, and mobility. Each subset was clustered using four different methods– K-means, Hierarchical, Hierarchical CURE, and DBSCAN. The resulting cluster assignments were compared to a ground truth which was derived from binned death_per_case values which represented COVID-19 severity levels. To evaluate alignment between clustering results and the ground truth, three metrics were computed: Adjusted Rand Index (ARI) which measures agreement adjusted for chance; purity, which quantifies the homogeneity of clusters; and entropy, which reflects the degree of class mixing within clusters. The supervised evaluation revealed notable differences in performance across methods and subsets. Additionally, it provided a rigorous and quantitative framework for comparing clustering outcomes, which supported the identification of meaningful patterns in COVID-19 severity across different county- level characteristics.

Table 3: Supervised Clustering Evaluation

Method	ARI	Purity	Entropy	Subset
KMeans	0.120	0.568	1.204	Households
Hierarchical	-0.020	0.432	1.370	Households
CURE	-0.020	0.432	1.370	Households
DBSCAN	NA	NA	NA	Households
KMeans	0.085	0.568	1.270	Poverty
Hierarchical	0.027	0.459	1.032	Poverty
CURE	0.027	0.459	1.032	Poverty
DBSCAN	NA	NA	NA	Poverty
KMeans	0.077	0.514	1.224	Demomobility
Hierarchical	0.044	0.486	1.207	Demomobility
CURE	0.044	0.486	1.207	Demomobility
DBSCAN	NA	NA	NA	Demomobility
KMeans	0.065	0.514	1.211	Mobility
Hierarchical	0.004	0.459	1.119	Mobility
CURE	0.004	0.459	1.119	Mobility
DBSCAN	NA	NA	NA	Mobility

Demographic and Household Features

From the household feature subset, the clustering method showed limited alignment with the death_per_case severity ground truth. From Table 3, K-means had the highest ARI of 0.120, a moderate purity score of 0.568, and a relatively high entropy of 1.204, thus indicating a mixed cluster composition. Both the hierarchical and CURE clustering methods had a negative ARI score of –0.020, which reflected poor agreement with the true labels, and achieved lower purity of 0.432 with the highest entropy of 1.370, suggesting substantial class overlap within clusters. The DBSCAN did not produce valid clustering results for this subset, perhaps due to insufficient

density-based separation in the data. Therefore, overall, K-means marginally outperformed the other methods, but all approached demonstrated relatively weak supervised alignment for house and demographic related variables.

Demographic and Poverty Features

The poverty related subset revealed a slightly improved but still modest supervised performance. From Table 3. the k-means produced an ARI of 0.085, and the highest purity of 0.568, though the entropy remained high at 1.270, indicating overlapping classes. The hierarchical and CURE methods performed similarly with ARI's 0.027, lower purity of 0.0459, and slightly better entropy of 1.032, reflecting slightly more compact cluster structures. The DBSCAN again failed to return usable results for this subset suggesting that the poverty-related features may not exhibit strong density-based clustering. Whole overall performance remained weak; K-means provided the most consistent alignment with ground truth.

Demographic and Mobility Features

The demographic and mobility subset integrates both variables demonstrating the most balanced clustering evaluation results among all subsets. As seen in Table 3, K-means attain an ARI of 0.077 with a purity of 0.514 and entropy of 1.224. Hierarchical and CURE both returned slightly lower ARI's of 0.044 and purity of 0.486, but their entropy of 1.207 was marginally better, indicated less class mixing. The DBSCAN failed again as it did not identify meaningful clusters, likely due to the combined complexity and noise in mobility data. Although clustering performance was modest, k-means provided the most consistent alignment with the ground truth.

Mobility Features

The mobility subset, composed of percent changed in movement patterns and COVID-19 metrics, exhibited the weakest supervised clustering performance. From Table 3, the K-means had the best relative performance attaining an ARI of 0.065, a purity score of 0.514, and a high entropy of 1.211, reflecting significant overlap in class labels within clusters. Hierarchical and CURE clustering performed poorly, each yielding an ARI of 0.004, with lower purity of. 0.459, and a slightly improved entropy of 1.119. The DBSCAN again failed to detect stable cluster structure, reinforcing that it was unable to identify any stable cluster structure in this subset. While none of the methods achieved strong clustering, K-means was the best performing technique for this subset.

Best Method Overall

Based on the supervised evaluation results across all four subsets, K-means clustering consistently outperformed the other methods in aligning with the ground truth classification derived from COVID-19 severity (death_per_case). K-means achieved the highest ARI in every subset and produced relatively higher purity scores compared to hierarchical, CURE, and DBSCAN. Although, entropy values remained moderate to high across all methods-indicating some class mixing- K-means consistently yielded the most stable cluster structures. Hierarchical and CURE clustering demonstrated comparable but generally lower performance, while DBSCAN struggled to identity meaningful groupings in any subsets. Overall, these results suggest K-means is the most robust and reliable clustering method for identifying severity-related patterns across diverse county-level demographic, socioeconomic, and mobility indicators.

4) Recommendations

Describing our Results

Where are the most outbreaks occurring, and what factors contribute to their spread?

Based on our analysis focusing solely on mobility features, we identified several important patterns about the spread of COVID-19 and its relationship to movement behaviors across counties. Although clustering performance for this subset was weaker overall, K-Means emerged as the most effective technique, consistently outperforming Hierarchical, CURE, and DBSCAN methods in both unsupervised and supervised evaluations. Specifically, in our supervised evaluation, K-Means achieved the highest Adjusted Rand Index (ARI) of 0.065 and a purity score of 0.514, indicating moderate success in grouping counties with similar outbreak dynamics based on mobility changes.

As seen in Figures 4 and 25 (the clustering profiles and silhouette analyses), we observed that the counties experiencing the most severe outbreaks were generally those associated with Cluster 3 in our K-Means mobility analysis. This cluster was characterized by sharp **declines in workplace and retail mobility** but notable **increases in residential mobility**, suggesting that residents in these areas significantly reduced external movement and stayed home more. Importantly, these areas also recorded higher rates of cases and deaths per 10,000 residents, indicating that despite behavioral adaptations, outbreaks were already widespread — likely due to earlier transmission before mobility restrictions were widely adopted or due to essential workers maintaining mobility despite restrictions. Interestingly, Cluster 1 counties, which exhibited moderate changes in mobility, had comparatively lower COVID-19 metrics. This suggests that regions with less drastic mobility shifts, potentially rural or less densely populated areas, may have experienced fewer outbreaks — either due to lower initial exposure or inherently reduced person-to-person contact.

Our findings point to **mobility as a reactive**, rather than proactive, **indicator** of outbreak severity. The most affected areas showed greater shifts in mobility, likely responding to growing case counts rather than preventing them. This reinforces for our stakeholders the critical importance of early mobility interventions: encouraging reductions in non-essential movement before outbreaks escalate could be more effective in controlling spread. While overall cluster separation was modest, these insights still provide valuable direction for public health policy. Areas displaying high residential mobility increase and steep workplace mobility declines should be prioritized for targeted public health messaging and support, particularly early in outbreak detection phases.

Which demographics are most affected, and how should interventions be tailored?
Our analysis of demographic and household, as well as demographic and poverty features, reveals important insights into which communities were most affected by COVID-19, and how interventions should be designed moving forward. Although clustering performance across both feature sets was moderate at best, K-Means consistently emerged as the strongest performer, particularly when evaluated with supervised metrics aligned to COVID-19 severity outcomes.

For the Demographic and Household feature set, K-Means achieved the best supervised performance with an Adjusted Rand Index (ARI) of 0.120 and a purity score of 0.568, outperforming hierarchical methods which demonstrated poor alignment with ground truth (negative ARI and lower purity). Referencing Figures 1 and 17, clusters generated from this analysis pointed to **higher-risk communities being those with lower median incomes, younger populations, and a higher proportion of family households**. These characteristics likely reflect regions with multi-generational households or essential workers, where living arrangements and job requirements increase potential exposure. Cluster profiles also showed that these counties experienced higher COVID-19 case and death rates, reinforcing that household density and economic vulnerability contribute to increased risk.

For the Demographic and Poverty feature set, supervised evaluation also favored K-Means, with an ARI of 0.085 and the highest purity score of 0.568, indicating that this method best captured relationships between poverty-related demographics and COVID-19 impacts. Referencing Figures 2 and 19, counties within higher-risk clusters in this set typically exhibited **higher poverty rates, lower median incomes, and slightly older populations**. These socio-economic factors likely intensified vulnerabilities through barriers to healthcare access, underlying health conditions, and economic constraints that limited individuals' ability to engage in preventative behaviors such as social distancing or remote work. Interestingly, hierarchical clustering in this subset showed moderately

compact cluster structures (slightly improved entropy), but K-Means still provided better alignment to actual severity outcomes.

From both analyses, a clear pattern emerges: economic disadvantage consistently overlaps with higher COVID-19 burden. Lower income and higher poverty counties, particularly those with dense family households, experienced worse outcomes. For stakeholders, this highlights the need for tailored interventions that prioritize economically vulnerable communities. Specifically, public health strategies should focus on:

- Providing accessible testing and vaccination programs in low-income, high-density household areas.
- Ensuring paid sick leave and financial support for essential workers and multi-generational households to reduce exposure risks.
- Targeted communication campaigns that account for demographic factors like younger population profiles, to encourage preventive behaviors.

While clustering performance suggests the complexity of these demographic relationships, the consistent patterns across both feature sets underscore the importance of proactively supporting socio-economically disadvantaged communities to mitigate COVID-19 impacts both in ongoing recovery efforts and in future public health crises.

Where should resources like vaccines and medical staff be distributed?
Based on the comprehensive COVID-19 data analyses, I recommend that vaccine and medical staff resources should be prioritized for counties belonging to clusters identified by K-means clustering's- particularly Cluster 2 as seen in Figures 22 to Figure 24, which demonstrated strong internal cohesion and distinct separation from other groups. Cluster 2 is likely to represent counties with similar socio-demographic risk profiles and mobility patterns that may have contributed to higher vulnerability during the COVID-19 pandemic. The figures demonstrate a strong silhouette score with and average width of 0.20, and a relatively high alignment with the ground truth outcomes, therefore, suggesting that K-means effectively captured meaningful patterns of risk. These counties may include populations with greater difficulty reducing mobility during outbreaks (essential workers or underdeveloped communities), making them key targets for proactive public health support. Allocating vaccines, medical personnel, and outreach efforts to these well-defined high-risk clusters can help reduce transmission rates and improve equity in healthcare delivery across Texas.

What recommendations can you formulate based on the clustering results?

How can COVID data inform future public health funding and policymaking for the state of Texas?
Building on the initial findings from our first report, our deeper clustering analysis provides more nuanced insights to help inform future public health funding and policy decisions in Texas. Previously, we recommended prioritizing urban centers like Harris, Dallas, Tarrant, Bexar, and El Paso counties — areas with high case counts, dense households, and significant essential worker populations. Our clustering work not only reinforces these earlier observations but also sharpens them, helping to clarify which demographic and behavioral patterns most consistently align with elevated COVID-19 impact.

First, our clustering analyses consistently identified that **economic vulnerability is a key driver of COVID-19 burden across Texas counties**. Both the demographic and household, and demographic and poverty clusters, revealed that lower-income counties with higher poverty rates and dense family households experienced disproportionately high cases and deaths per 10,000 residents. These findings emphasize that public health strategies should prioritize not just urban centers, but **specifically target low-income, densely housed communities within these urban areas**. This adds precision to our first recommendation, confirming that interventions should focus at the neighborhood level, where socioeconomic vulnerability intersects with population density.

Second, our mobility clustering showed that mobility patterns tend to be reactive rather than preventive — counties with the most severe outbreaks also showed the greatest mobility reductions, likely after case counts began to rise. This suggests a lag in behavior change and highlights the critical need for **proactive public health measures**. Early and aggressive outreach in areas showing high baseline mobility could help slow transmission

before outbreaks escalate. Importantly, this aligns with our initial recommendation to enhance protective measures for essential workers, as high mobility persists in populations with limited remote work opportunities.

Third, while DBSCAN and hierarchical clustering struggled to reveal distinct groupings in several feature sets, K-Means emerged as the most consistently effective method, especially in identifying high-risk clusters based on household density and poverty. This reinforces the value of using targeted, data-driven approaches to guide resource allocation, rather than one-size-fits-all strategies.

Based on these insights, we recommend the following actions for Texas public health leadership:

- Continue prioritizing urban centers, with a sharpened focus on economically disadvantaged neighborhoods with dense family households, which consistently align with high-risk clusters.
- Invest in proactive mobility management through early communication campaigns and policies that reduce non-essential movement before case surges.
- Expand support for essential workers, including sustained PPE distribution, accessible testing, and vaccination programs, especially in lower-income, high-mobility communities.
- Enhance telehealth and remote work infrastructure, particularly in socioeconomically vulnerable counties, to reduce exposure risks and support pandemic resilience.
- Deploy hyper-localized public health messaging that resonates with younger populations and essential workforce demographics identified in high-risk clusters.

In comparison to our initial report, these recommendations refine and deepen our earlier guidance. Rather than focusing solely on urban vs. rural distinctions, our clustering analysis highlights the importance of addressing the socio-economic and behavioral patterns within high-risk urban counties. This approach ensures that public health funding and interventions are not only geographically targeted but also socio-demographically precise, leading to more effective allocation of resources and stronger mitigation of future outbreaks.

5) List of References

Centers for Disease Control and Prevention. (n.d.). CDC Covid Data tracker. Centers for Disease Control and Prevention. https://covid.cdc.gov/covid-data-tracker

Centers for Disease Control and Prevention. (2025, February 13). Home page for MMWR. Centers for Disease Control and Prevention. https://www.cdc.gov/mmwr/index.html

Covid-19. Johns Hopkins Medicine. (n.d.). https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus

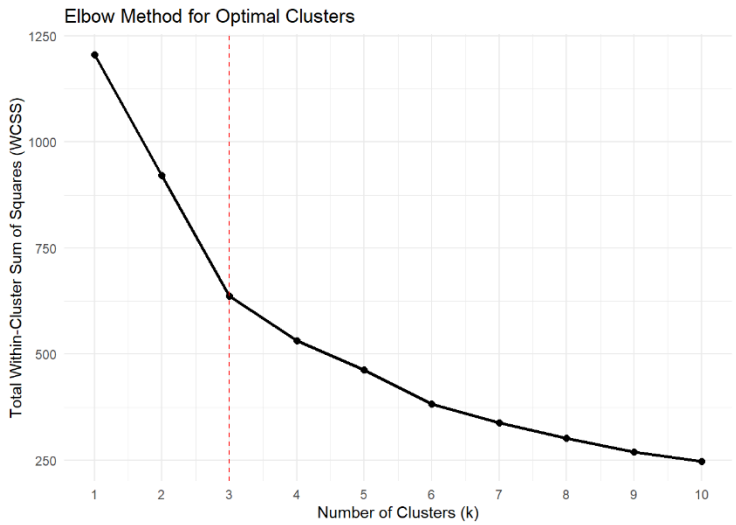
U.S. Department of Health and Human Services. (n.d.). *Coronaviruses*. National Institute of Allergy and Infectious Diseases. https://www.niaid.nih.gov/diseases-conditions/coronaviruses

World Health Organization. (n.d.). Covid-19 cases | WHO COVID-19 Dashboard. World Health Organization. https://covid19.who.int/

"CDC Museum Covid-19 Timeline," Centers for Disease Control and Prevention, https://www.cdc.gov/museum/timeline/covid19.html (Accessed Mar. 12, 2025).

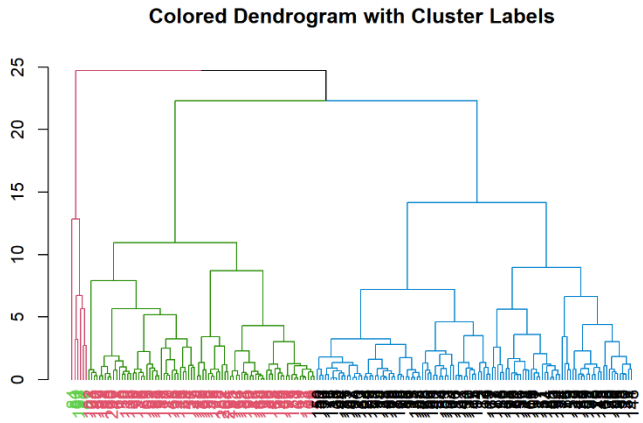
6) Appendix

Elbow Method for Optimal No. Of Clusters (K-Means)

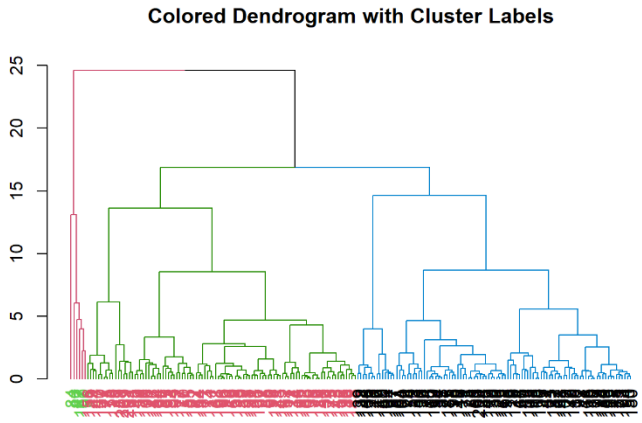


Hierarchical Dendrograms

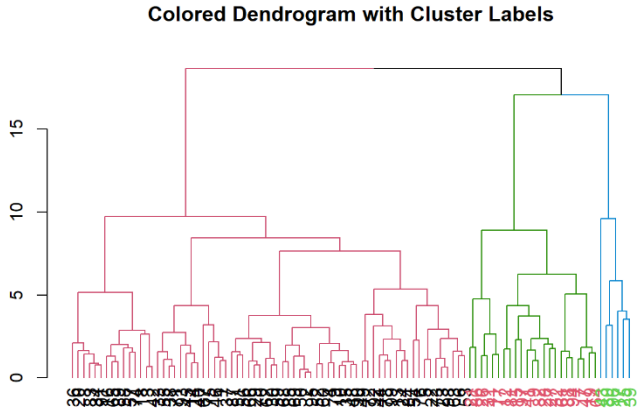
Demographic and Household Features



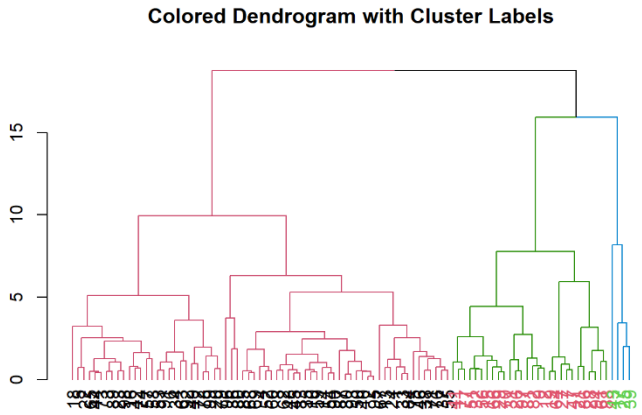
Demographic and Poverty Features



Demographic and Mobility Features

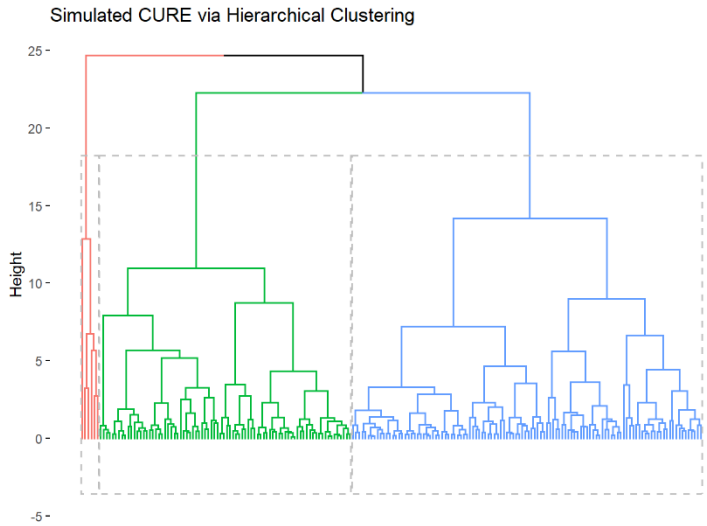


Mobility Features

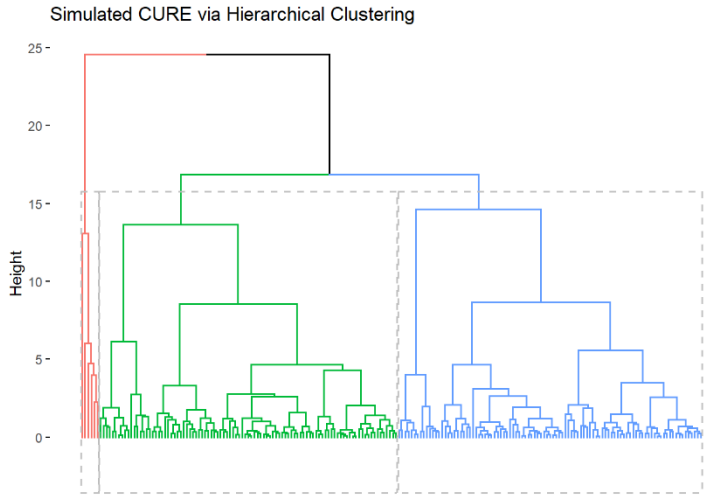


Hierarchical via CURE Dendrograms

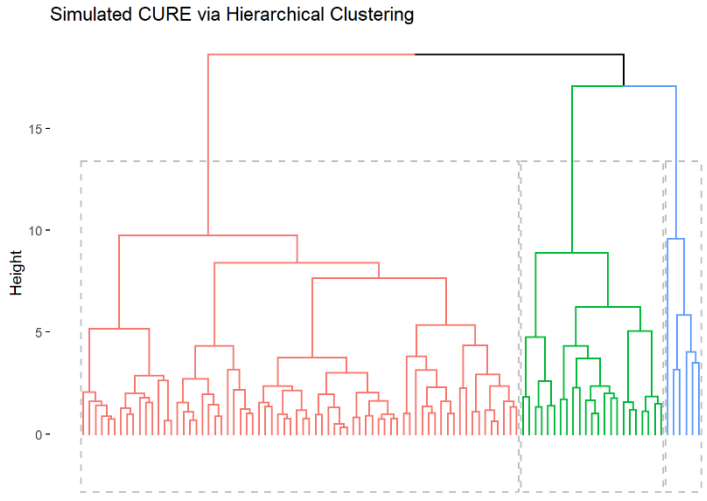
Demographic and Household Features



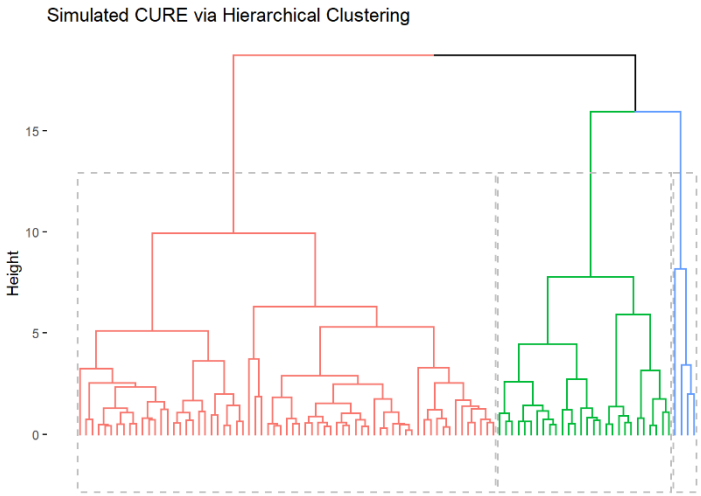
Demographic and Poverty Features



Demographic and Mobility Features



Mobility Features



Student Contributions

All three group members (Sreshta, Muskaan, and Ridhi) contributed equally to the project. We all worked equally on research and code for the report. All three members also worked on report analysis and analysis of graphs. Each person played a crucial role in ensuring the overall success of the project, with the workload distributed evenly among the team.