# Cluster Analysis on COVID-19 Data to Inform Future Pandemic Responses in Texas

*Project 2*

[Abdul Wasay, Om Sumedh Kharwadkar /Team 1]

# Executive Summary:

This report builds on the findings of Report 1 by utilizing advanced clustering techniques, including K-Means, Hierarchical, DBSCAN, and Model-Based Clustering, to gain deeper insights into COVID-19 outcomes across counties. The analysis reveals significant urban-rural disparities, with urban counties showing better outcomes due to higher vaccination rates and stronger healthcare infrastructure, while rural counties face higher death rates and lower vaccination coverage. Vaccination rates emerge as a critical factor in reducing mortality, with clusters of low vaccination highlighting areas in urgent need of targeted public health campaigns. Socioeconomic vulnerabilities, such as lower income and higher poverty rates, are closely linked to worse COVID-19 outcomes, underscoring the need for economic and healthcare support in these regions. Additionally, DBSCAN identifies outlier counties with unique challenges, providing new opportunities for intervention. These findings offer a data-driven framework for optimizing public health strategies, focusing resources on high-risk areas to improve COVID-19 outcomes.

# Table of Contents


Executive Summary:....2
Business Understanding....4
Data Preparation....5
Modeling....8
1. K-Means Clustering:....8
K-Means 5 Clusters: Starting Point....9
K-Means 4 Clusters: Simplifying the Structure....10
K-Means 6 Clusters: More Granular Structure....11
Best Performing Clustering K Means Solution:....12
2. Hierarchical Clustering:....13
3. DBSCAN (Density-Based Spatial Clustering of Applications with Noise):....16
4. Model-Based Clustering:....18
Determining suitable number of clusters for each method....20
Unsupervised evaluation to describe and compare the clusterings....21
Identify a feature you could use as the ground truth to perform supervised evaluation. Compare the clusterings using this method....22
Evaluation and Recommendations....23
Graduates / Exceptional Work [10 points]....25
Student Contributions....25

# Business Understanding

COVID-19, the deadly disease that gripped the world in 2019 and 2020, prompted us to explore its spread and mortality in relation to specific population metrics. Understanding these dynamics is crucial for addressing the global impact of COVID-19 and the critical need for insights into population behavior, healthcare resources, and socio-economic factors that influenced virus transmission and death rates. By analyzing key metrics such as infection rates, hospitalizations, and demographic trends, we aim to generate insights that can inform public health strategies for both ongoing pandemic recovery and future pandemic preparedness.

Based on the findings from our initial report, we have chosen to proceed with Project 2, which focuses on clustering the following metrics and their relationship to the spread and mortality rates during a pandemic for the following reasons:

- Hispanic Population:

  A strong positive correlation (0.47 with cases and 0.31 with deaths) suggests that areas with higher Hispanic populations may experience more COVID-19 cases and deaths. Investigating social determinants, healthcare access, and vaccination rates within these communities could yield critical insights.
- Income Levels:

  The negative correlation between median income and both confirmed cases (-0.13) and deaths (-0.40) indicates that lower-income areas face greater challenges. Studying the impact of income on healthcare access, preventive measures, and community resilience against COVID-19 is essential.
- Vaccination Rates:

  The negative correlation with deaths (-0.35) highlights the importance of vaccination efforts. Further analysis of vaccination accessibility, public health campaigns, and the impact of vaccination on different demographics could provide valuable insights for future interventions.
- Health Factors:

  The strong positive correlation between the percentage of individuals in fair or poor health (0.39 with deaths and 0.42 with cases) and smoking rates (0.19 with cases and 0.20 with deaths) underscores the need to investigate health disparities and the role of pre-existing conditions in COVID-19 outcomes.
- Education Levels:

  The negative correlations between educational attainment (high school and bachelor's degrees) and death rates suggest that educational interventions could play a vital role. Understanding how education influences public health responses and individual health behaviors during pandemics could offer valuable perspectives.

# Data Preparation

- Objects to Cluster:

  The objects to cluster are Texas counties based on their demographic, socioeconomic, and health-related features that influence COVID-19 spread and mortality.

- Features for Clustering:

  Relevant features would be selected based on their significance in influencing COVID-19 outcomes, focusing on:

  - Hispanic Population (proxy for cultural and language diversity).
  - Median Income (economic disparity affecting healthcare access).
  - Percent Below Poverty (economic vulnerability).
  - Percent Fair or Poor Health (general health status).
  - Percent Smokers (behavioral health risk).
  - Percent Adults with Obesity (chronic health risk).
  - High School Diploma (education level impacting public health behavior).
  - Bachelor's Degree (higher education for targeted messaging).
  - Percent Vaccinated (pandemic preparedness).

| **Feature** | **Mean** | **Median** | **Variance** | **Range** | **Mode** | **Reason for Clustering** |
|---|---|---|---|---|---|---|
| Hispanic Population | 42,023.26 | 5,068.5 | 2.97E+10 | 12 - 1,910,535 | 0 | Understand COVID-19 spread in diverse communities and improve outreach. |
| Median Income | 49,894.34 | 48,311 | 1.47E+08 | 24,794 - 93,645 | 42,500 | Explore how income influences access to healthcare and adherence to health guidelines. |
| Percent Below Poverty | 16.87 | 16.5 | 33.06 | 1.8 - 37.9 | 17.2 | Identify economically vulnerable counties requiring more healthcare support. |

| | | | | | | |
|---|---|---|---|---|---|---|
| Percent Fair or Poor Health | 20.60 | 19.73 | 23.85 | 12.29 - 40.99 | 4.74 | Measure general health vulnerability and its impact on COVID-19 outcomes. |
| Percent Smokers | 14.98 | 14.9 | 2.34 | 10.64 - 19.87 | 14.9 | Assess behavioral risk factors contributing to severe COVID-19 cases. |
| Percent Adults with Obesity | 31.49 | 30.7 | 27.45 | 21.6 - 47.3 | 28.6 | Identify areas with higher obesity rates to target with health interventions. |
| High School Diploma | 14,130.69 | 3,281 | 2.28E+09 | 3 - 559,393 | 695 | Understand the role of education in health literacy and compliance with preventive measures. |
| Bachelor's Degree | 12,947.94 | 1,289.5 | 2.59E+09 | 11 - 237,000 | 237 | Higher education improves understanding and adherence to public health campaigns. |
| Percent Vaccinated | 37.03 | 39 | 90.27 | 9 - 55 | 38 | Measure the success of vaccination campaigns and immunity levels in counties. |

Table 1: Features, Statistics for Clustering

| **Variable** | **Scale of Measurement** | **Appropriate Measures for Similarity/Distance** | **Why It's Suitable** |
|---|---|---|---|
| Hispanic Population | Ratio | Euclidean Distance, Manhattan Distance | Suitable for continuous numeric data to measure spread/mortality patterns by population size. |
| Median Income | Ratio | Euclidean Distance, Manhattan Distance | Measures the economic disparity's impact on healthcare access and COVID-19 outcomes. |
| Percent Below Poverty | Ratio | Euclidean Distance, Manhattan Distance | Highlights counties with varying poverty levels that may influence pandemic vulnerability. |
| Percent Fair or Poor Health | Ratio | Euclidean Distance, Manhattan Distance | Assesses general health conditions across counties in relation to COVID-19 spread and mortality. |
| Percent Smokers | Ratio | Euclidean Distance, Manhattan Distance | Evaluates the impact of smoking prevalence as a risk factor. |
| Percent Adults with Obesity | Ratio | Euclidean Distance, Manhattan Distance | Analyzes obesity rates and their correlation with severe COVID-19 cases. |

| | | | |
|---|---|---|---|
| High School Diploma | Ratio | Euclidean Distance, Manhattan Distance | Measures the impact of basic education levels on health awareness and public health behavior. |
| Bachelor's Degree | Ratio | Euclidean Distance, Manhattan Distance | Examines how higher education affects health outcomes and compliance with pandemic measures. |
| Percent Vaccinated | Ratio | Euclidean Distance, Manhattan Distance | Tracks the effectiveness of vaccination campaigns in reducing mortality and spread. |

Table 2: Scale of Measurement for Features & Appropriate Measures of Similarity / Distance

# Modeling

We used the following 4 methods of clustering (this includes multiple clusters for each method and the graduate work of including 2 new methods).

## 1. K-Means Clustering:

- Explanation: K-Means partitions the data into a predefined number of clusters by minimizing the sum of squared distances between data points and their respective cluster centroids.
- Why Appropriate: Efficient for large datasets and provides clear, non-overlapping clusters. Works best when clusters are spherical and similar in size.
- Use Case: Helps identify counties with similar demographic and socioeconomic profiles, which can inform targeted public health strategies.

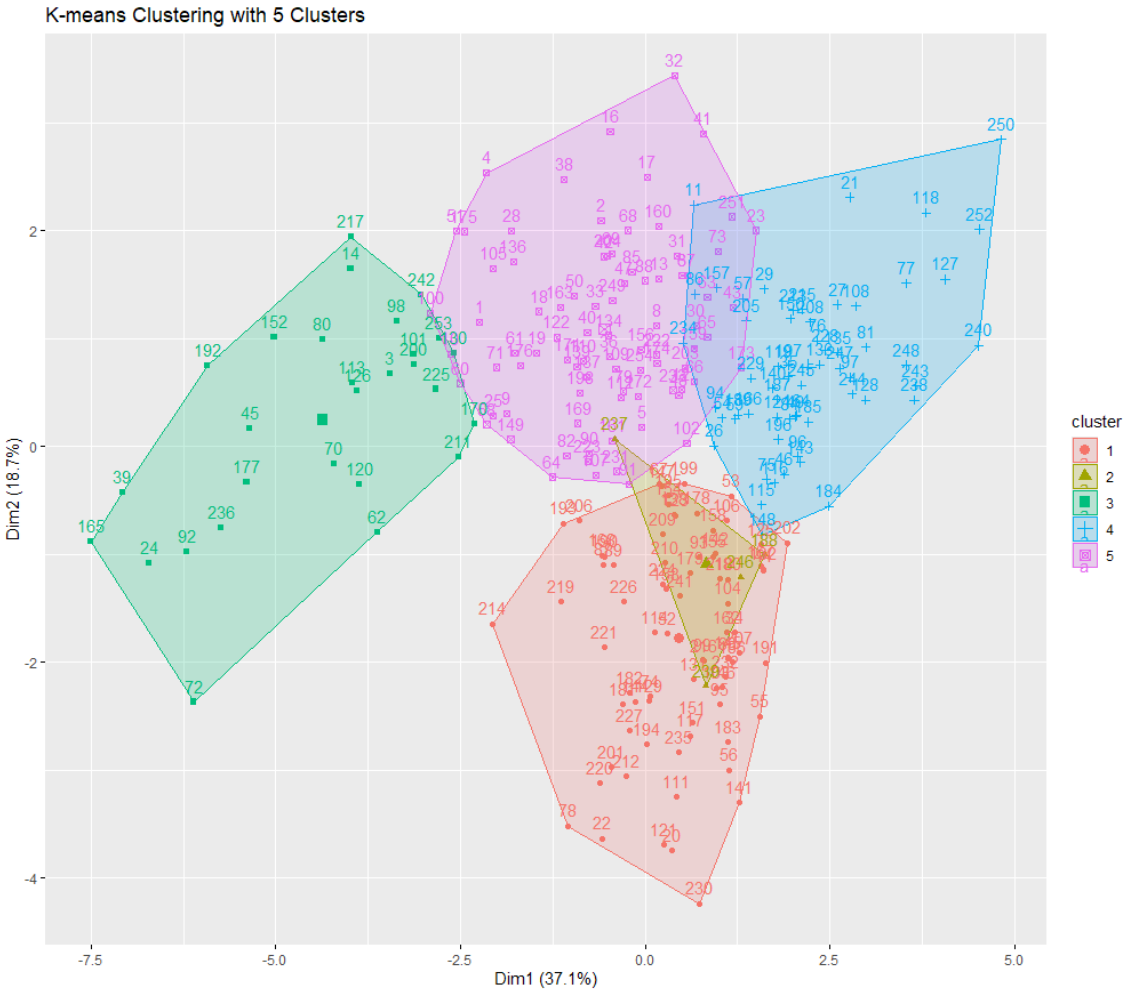


Illustration 1: K Means Clustering with 5 clusters

# K-Means 5 Clusters: Starting Point

**Reason for 5 Clusters:** 5 clusters is usually a common starting point in K-Means as it is a balanced level of granularity, which we hope will allow for differentiation between counties with varying levels of population density, health outcomes, and socioeconomic factors.

**Interpretation of 5-Cluster Results:**

1. Cluster 1 (n=77 counties):

   - Moderate population (~108,066), high case rate (~80.95), moderate death rate (~1.73).

   - Relatively low vaccination (~37.5%) and income, with higher poverty (~17.37%).

   - Insight: These counties may need targeted economic and vaccination support.

2. Cluster 2 (n=4 counties):

   - Large urban counties (~638,830), lowest case (~63.17) and death rates (~0.89).

   - Highest vaccination (~45%) and income.

   - Insight: Well-resourced urban counties with favorable COVID-19 outcomes.

3. Cluster 3 (n=28 counties):

   - Moderate population (~81,860), high case (~80.2) and death rates (~2.01).

   - Lower vaccination (~35.7%), moderate income, and poverty (~16.05%).

   - Insight: Counties with poorer outcomes needing increased healthcare focus.

4. Cluster 4 (n=59 counties:

   - Larger population (~183,321), moderate case (~72.9) and death rates (~1.68).

   - Higher vaccination (~39.7%) and moderate income (~$51,312).

   - Insight: Performing relatively well due to better vaccination rates.

5. Cluster 5 (n=84 counties:

   - Smallest population (~41,839), high case (~79.2) and death rates (~2.13).

   - Lowest vaccination (~32.9%).

   - Insight: Rural counties with the worst outcomes, requiring significant public health interventions.

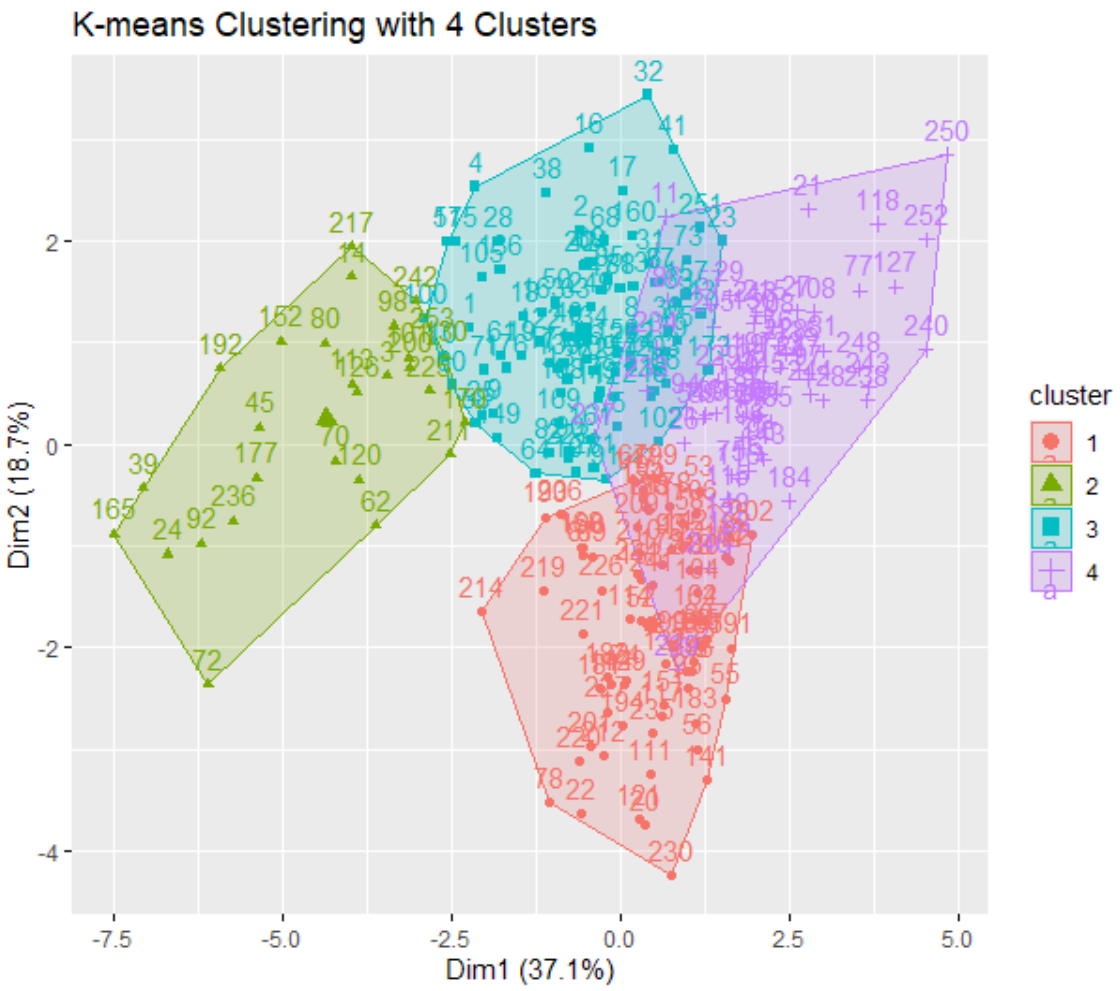


Illustration 2: K Means Clustering with 4 clusters

## K-Means 4 Clusters: Simplifying the Structure

**Reason for Trying 4 Clusters:** Reducing the number of clusters might consolidate similar groups, simplifying analysis and highlighting broader trends across counties.

**Interpretation of 4-Cluster Results:**

1. Cluster 1 (n=77 counties): Similar to Cluster 1 in the 5-cluster solution.

2. Cluster 2 (n=28 counties): Similar to Cluster 3 in the 5-cluster solution.

3. Cluster 3 (n=84 counties): Combines the smaller rural counties from Clusters 4 and 5 in the 5-cluster solution.

4. Cluster 4 (n=63 counties): Merges the urban and moderately populated counties (Clusters 2 and 4 in 5-cluster).

**Insight**: The 4-cluster solution simplifies the distribution but can hide the key differences in rural and urban counties' performance.

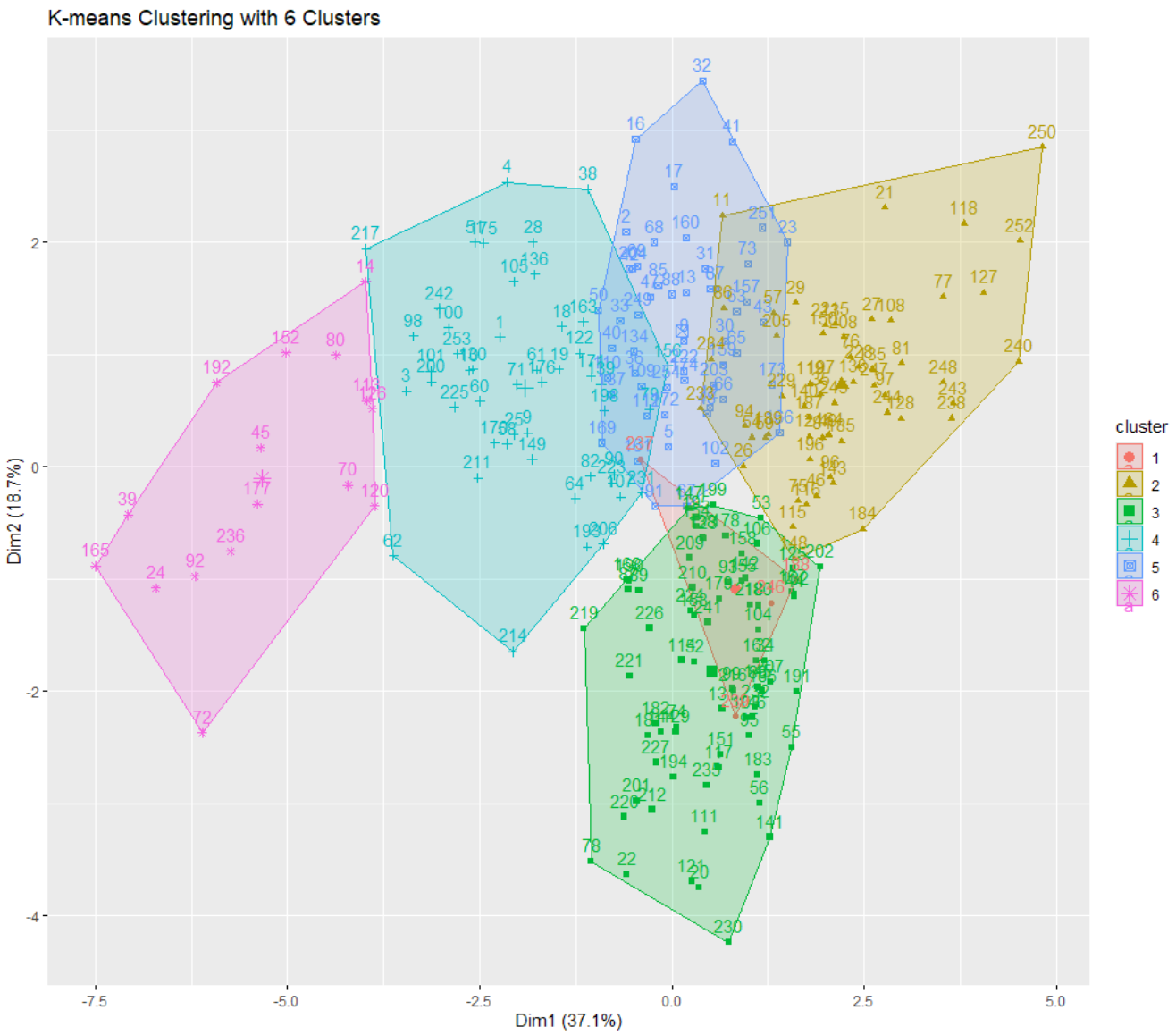


Illustration 3: K Means Clustering with 6 clusters

## K-Means 6 Clusters: More Granular Structure

**Reason for Trying 6 Clusters:** Adding more clusters allows for finer differentiation, especially among counties with moderate populations and varying socioeconomic conditions.

**Interpretation of 6-Cluster Results**:

1. Cluster 1 (n=4 counties): Matches Cluster 2 in the 5-cluster result (large urban counties).

2. Cluster 2 (n=58 counties): Further refines Cluster 4 from the 5-cluster solution (moderately large population, better vaccination rates).

3. Cluster 3 (n=73 counties): Similar to Cluster 1 (moderate population, lower vaccination, higher case rates).

4. Cluster 4 (n=48 counties): Separates counties with the highest case and death rates (~2.15) and lower vaccination (~35.8%).

5. Cluster 5 (n=53 counties): Lower population counties with moderate outcomes and vaccination (~31.7%).

6. Cluster 6 (n=16 counties): Smallest counties, high death rate (~2.12), and very low vaccination (~33.6%).

Insight: The 6-cluster solution provides better segmentation of small and moderate counties but risks overfitting and less actionable grouping.

## Best Performing Clustering K Means Solution:

The 5-cluster K-means achieved the best balance between granularity and interpretability as it is able to distinguish differences between rural, urban, and moderate counties and clearly highlights public health disparities and vaccination outcomes. The visual cluster separations confirm this balance.

## 2. Hierarchical Clustering:

- Explanation: Builds a hierarchy of clusters either through a bottom-up approach (agglomerative) or top-down approach (divisive). The result is often visualized using a dendrogram.
- Why Appropriate: Does not require predefining the number of clusters and reveals nested relationships. Suitable for smaller datasets or when exploring the underlying structure of the data.
- Use Case: Useful for understanding regional groupings and their substructures, which can inform resource allocation during pandemics.

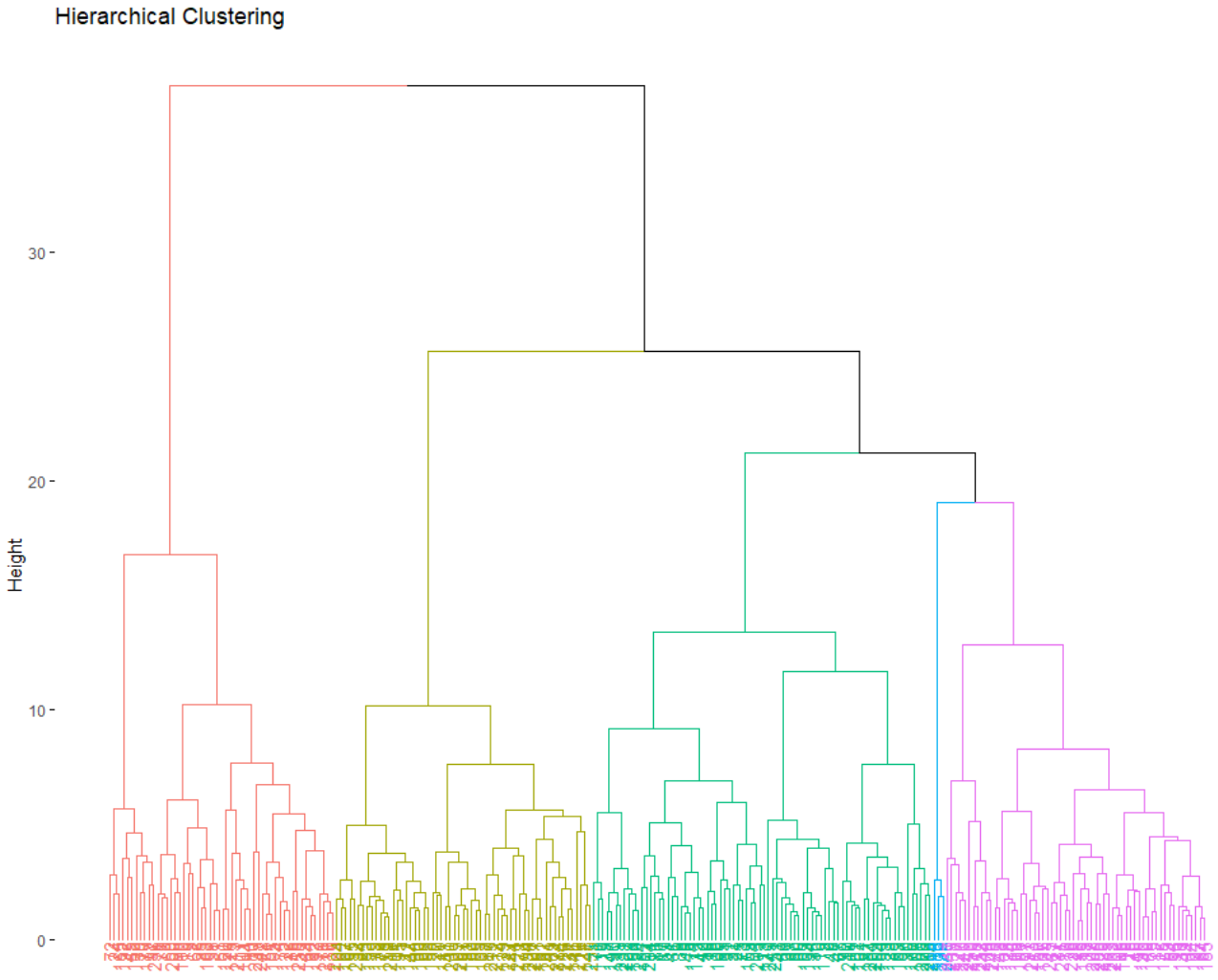


Illustration 4: Hierarchical Clustering

**Hierarchical Cluster Breakdown and Insights:**

1. Cluster 1 (n=52 counties):
    - Population: ~68,453 (small to moderate).
    - Cases per 1,000: ~82.57 (high).
    - Deaths per 1,000: ~2.18 (highest).
    - Percent Vaccinated: ~35.02% (low).
    - Median Income: ~$49,236 (moderate).
    - Percent Below Poverty: ~16.80% (moderate).
    - Insight: Rural counties with high case and death rates, low vaccination coverage, and moderate income. Public health efforts should focus on improving vaccination and healthcare access in these areas. This cluster basically represents vulnerable rural counties with high death rates and low vaccination rates.
2. Cluster 2 (n=78 counties):
    - Population: ~68,506 (small to moderate).
    - Cases per 1,000: ~76.70 (moderate).
    - Deaths per 1,000: ~1.93 (high).
    - Percent Vaccinated: ~32.39% (lowest).
    - Median Income: ~$48,957 (moderate).
    - Percent Below Poverty: ~16.17% (moderate).
    - Insight: These counties perform slightly better than Cluster 1 in terms of case and death rates but have the lowest vaccination rates, indicating significant vulnerability.This cluster basically represents vulnerable rural counties with low vaccination rates.
3. Cluster 3 (n=60 counties):
    - Population: ~179,910 (larger counties).
    - Cases per 1,000: ~73.84 (moderate).
    - Deaths per 1,000: ~1.68 (lower).
    - Percent Vaccinated: ~40.67% (higher).
    - Median Income: ~$51,749 (higher).
    - Percent Below Poverty: ~16.89% (moderate).
    - Insight: Larger counties with better vaccination rates and lower death rates. These counties likely have better healthcare infrastructure and public health outreach.
4. Cluster 4 (n=59 counties):
    - Population: ~90,641 (moderate).
    - Cases per 1,000: ~80.96 (high).
    - Deaths per 1,000: ~1.74 (moderate).
    - Percent Vaccinated: ~38.10% (moderate).

- Median Income: ~$49,281 (moderate).
- Percent Below Poverty: ~17.55% (highest).
- Insight: Moderate counties with high case rates and relatively high poverty levels. Focus should be on economic support and vaccination campaigns.

5. Cluster 5 (n=3 counties):
    - Population: ~835,626 (largest, urban counties).
    - Cases per 1,000: ~65.92 (lowest).
    - Deaths per 1,000: ~1.01 (lowest).
    - Percent Vaccinated: ~45.67% (highest).
    - Median Income: ~$57,737 (highest).
    - Percent Below Poverty: ~15.67% (lowest).
    - Insight: Well-resourced urban counties with the best COVID-19 outcomes, including the lowest case and death rates and highest vaccination rates. These counties are well-prepared and demonstrate the benefits of robust healthcare systems and public health infrastructure.

We believe that no further adjustments are necessary unless we need to specifically want to dive deeper into certain clusters. The current levels of clustering seems sufficient and actionable when finding clusters that can help the understanding of COVID-19 outcomes and guiding public health strategies.

## 3. DBSCAN (Density-Based Spatial Clustering of Applications with Noise):

- Explanation: Groups points that are closely packed together, marking points in low-density regions as outliers. Requires two parameters: minimum points and neighborhood radius.
- Why Appropriate: Identifies clusters of varying shapes and sizes and detects outliers without predefining the number of clusters. Robust to noise in data.
- Use Case: Effective for distinguishing highly impacted counties (dense clusters) from less affected ones and identifying outliers needing special focus.

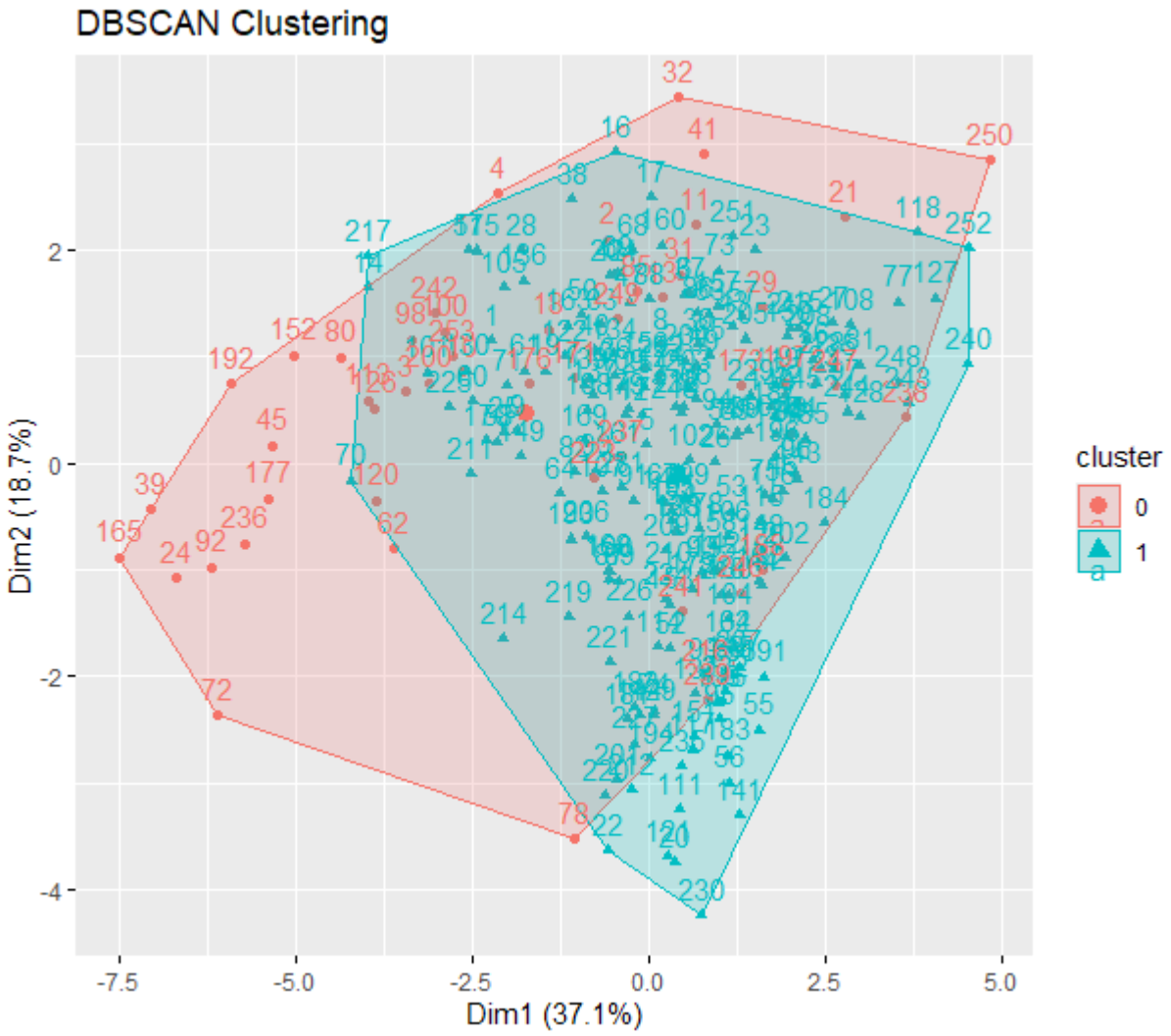


Illustration 5: DBscan Clustering

**Interpretation of DBSCAN Clustering Results**

Cluster 1 (n=203 counties)

- Population: ~89,177 (moderately populated counties).

- Cases per 1,000: ~77.63.
- Deaths per 1,000: ~1.85.
- Percent Vaccinated: ~36.63%.
- Median Income: ~$49,500.
- Percent Below Poverty: ~16.76%.

Insight:

This cluster represents the majority of counties with moderate population sizes, slightly better vaccination rates, and relatively lower death rates compared to Cluster 0. These counties may be performing better in terms of COVID-19 management, but continued focus on improving vaccination rates could enhance outcomes.

Cluster 0 (n=49 counties, noise cluster)

- Population: ~193,688 (larger counties, potentially urban or suburban).
- Cases per 1,000: ~80.09.
- Deaths per 1,000: ~1.95.
- Percent Vaccinated: ~35.63%.
- Median Income: ~$51,373.
- Percent Below Poverty: ~16.92%.

Insight:

Cluster 0 likely includes outlier counties with higher populations. Despite higher incomes, vaccination rates remain low, which might explain the relatively higher case and death rates. These counties could benefit from targeted vaccination campaigns to reduce COVID-19 mortality.

DBSCAN was supposed to effectively separate higher-density regions (Cluster 1) from lower-density areas or outliers (Cluster 0). Although this method highlighted counties that deviate significantly from the norm, allowing for targeted interventions in those regions, the clustering diagram itself leaves more to be desired as there is a significant overlap in the clusters as can be seen in illustration 5.

## 4. Model-Based Clustering:

- Explanation: Assumes that data is generated from a mixture of probability distributions (often Gaussian) and assigns probabilities for each point belonging to a cluster.
- Why Appropriate: Suitable for data with complex, non-spherical clusters and provides soft clustering, allowing for uncertainty in cluster membership.
- Use Case: Useful for gaining probabilistic insights into clusters, which can help tailor nuanced public health interventions based on the likelihood of specific outcomes.

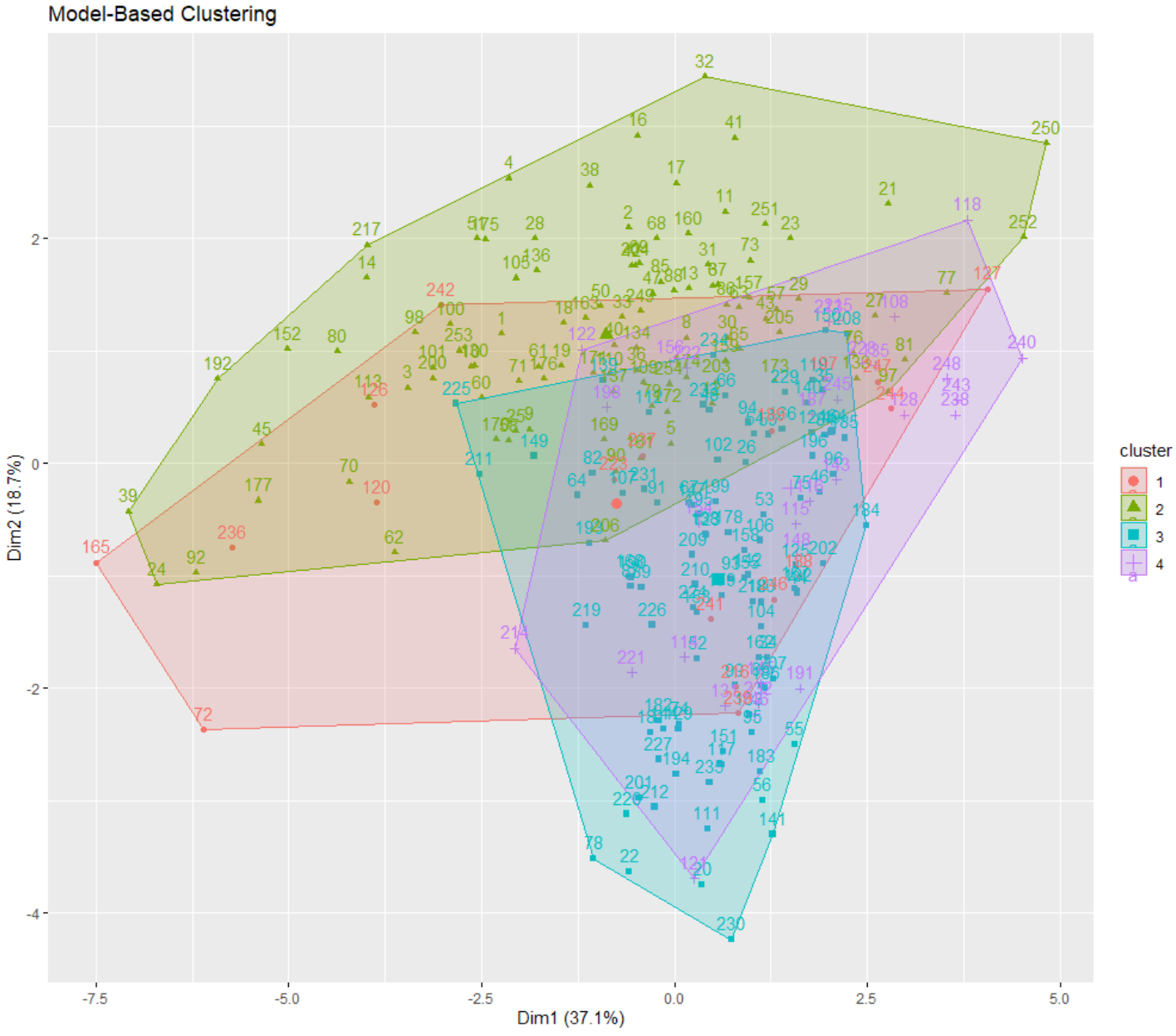


Illustration 6: Model Based Clustering

**Interpretation of Model-Based Clustering Results:**

1. Cluster 1 (n=18 counties):
    - Population: ~436,587 (largest population cluster, likely urban or suburban counties).
    - Cases per 1,000: ~77.2 (moderate).
    - Deaths per 1,000: ~1.41 (lowest among clusters).
    - Vaccination Rate: ~42.5% (highest).
    - Median Income: ~$56,540 (highest).
    - Insight: Well-resourced, urban counties with the best COVID-19 outcomes due to high vaccination and income levels.
2. Cluster 2 (n=103 counties):
    - Population: ~36,006 (small rural counties).
    - Cases per 1,000: ~75.9.
    - Deaths per 1,000: ~2.04 (highest).
    - Vaccination Rate: ~32.3% (lowest).
    - Median Income: ~$49,788 (moderate).
    - Insight: Rural counties with the worst outcomes, likely due to low vaccination rates and limited healthcare access.
3. Cluster 3 (n=100 counties):
    - Population: ~60,962.
    - Cases per 1,000: ~81.3 (highest case rate).
    - Deaths per 1,000: ~1.85.
    - Vaccination Rate: ~37.4%.
    - Median Income: ~$48,434 (lowest).
    - Insight: Moderately populated counties with high infection rates but improved outcomes compared to Cluster 2.
4. Cluster 4 (n=31 counties):
    - Population: ~316,898 (larger suburban counties).
    - Cases per 1,000: ~75.5.
    - Deaths per 1,000: ~1.61.
    - Vaccination Rate: ~43.4% (second highest).
    - Median Income: ~$50,873.
    - Insight: Suburban counties with good vaccination rates and moderate outcomes, possibly due to better healthcare infrastructure.

## Determining suitable number of clusters for each method

The methodology for K-means clustering is illustrated within the section for K-Means, but as a summary, we chose 5 clusters to begin with as a common starting point. We then tested 4 clusters (a smaller number) to explore broader groupings and simplify the analysis, and 6 clusters (a larger number) to see if additional granularity would provide more actionable insights.

For Hierarchical Clustering, we didn't need to predefine the number of clusters initially, as this method builds a dendrogram that allows for flexibility in choosing the number of clusters post hoc. This helps visualize nested relationships and determine clusters based on data structure.

For DBSCAN, we didn't need to specify the number of clusters because it identifies clusters based on density and separates noise (outliers) automatically. This makes it particularly effective for data with varying cluster shapes and noise.

For Model-Based Clustering, we didn't need to predefine the number of clusters because it uses statistical criteria like the Bayesian Information Criterion (BIC) to determine the optimal number of clusters automatically. This approach assumes that the data is generated from a mixture of probability distributions and finds the best fit accordingly.

# Unsupervised Evaluation (Silhouette Scores)

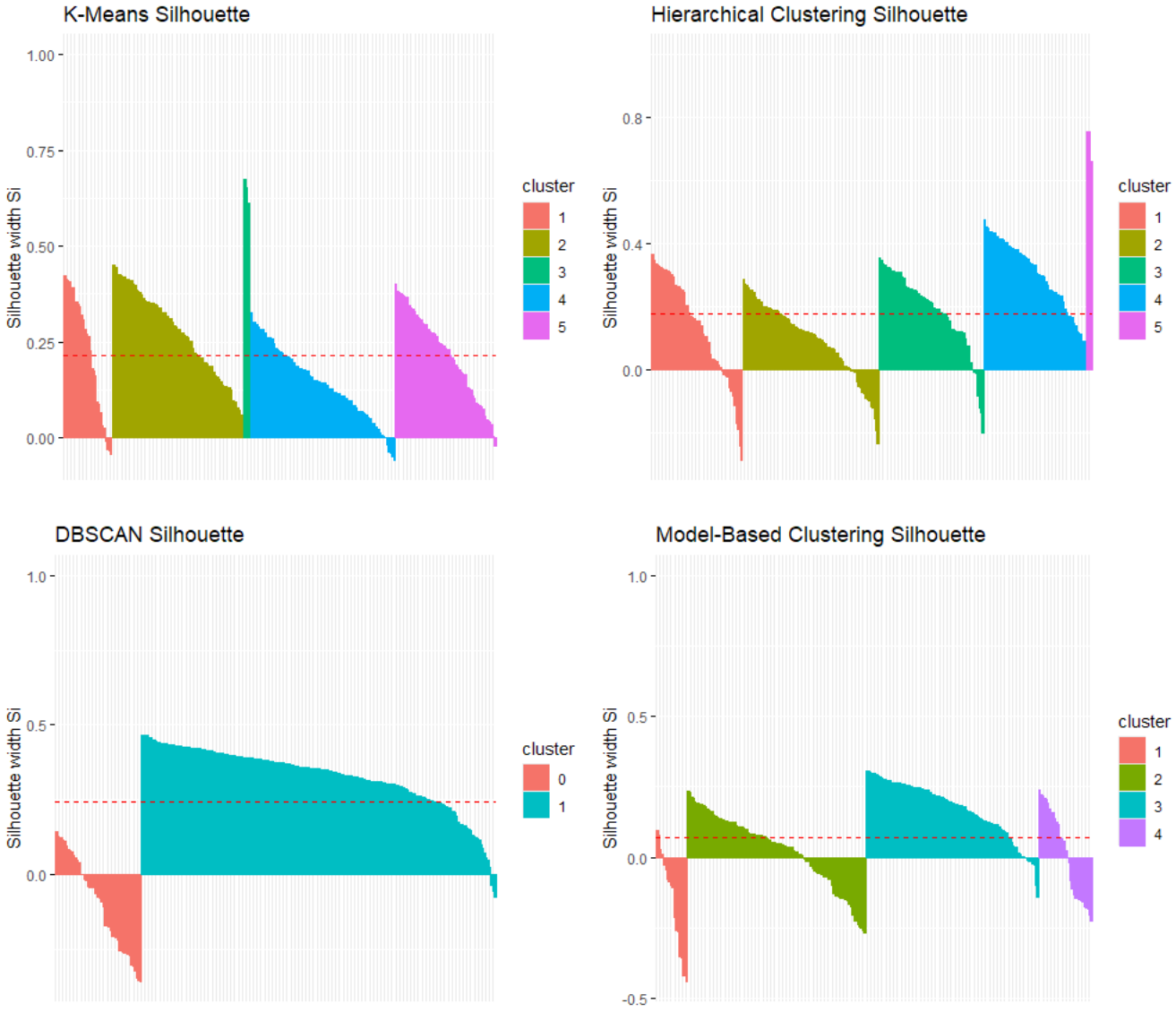


Illustration 7: Silhouette Scores Across Clusterings

The silhouette score measures how well-separated the clusters are, with values ranging from -1 (poor clustering) to 1 (well-defined clusters). Here's how the clustering methods performed:

1. DBSCAN (Silhouette: 0.244):
    - Best-performing method in terms of clustering quality.
    - Identifies clusters of varying densities and effectively handles noise, resulting in well-separated clusters.
2. K-Means (Silhouette: 0.215):
    - Second-best performance with moderate separation of clusters.

- Works well for data with more uniform, spherical clusters, but might struggle with non-spherical or uneven cluster sizes.

3. Hierarchical Clustering (Silhouette: 0.179):
    - Shows weaker cluster separation.
    - Although useful for revealing nested relationships, it may not provide the clearest cluster boundaries in this dataset.
4. Model-Based Clustering (Silhouette: 0.069):
    - Lowest silhouette score, indicating poor separation between clusters.
    - Probabilistic nature may result in overlapping clusters, which leads to less distinct boundaries.

## Ground Truth and Supervised Evaluation

For supervised evaluation of our clustering methods, we selected death rates per 1,000 people as our ground truth feature. This was because death rates directly measure COVID-19 severity and healthcare system effectiveness and it is an objective outcome metric that reflects the combined impact of various socioeconomic and health factors. It also aligns with our project's goal of understanding and improving pandemic responses. The evaluation is meant to understand how well each clustering method grouped counties according to their actual COVID-19 death levels.

We created three categories based on death rates:

- Low risk (0): Counties with death rates in the bottom third (83 counties)
- Moderate risk (1): Counties with death rates in the middle third (86 counties)
- High risk (2): Counties with death rates in the top third (83 counties)

To evaluate clustering performance, we used two metrics:

1. Adjusted Rand Index (ARI): Measures similarity between two clusterings, accounting for chance
2. Normalized Mutual Information (NMI): Measures the mutual dependence between the true and predicted labels

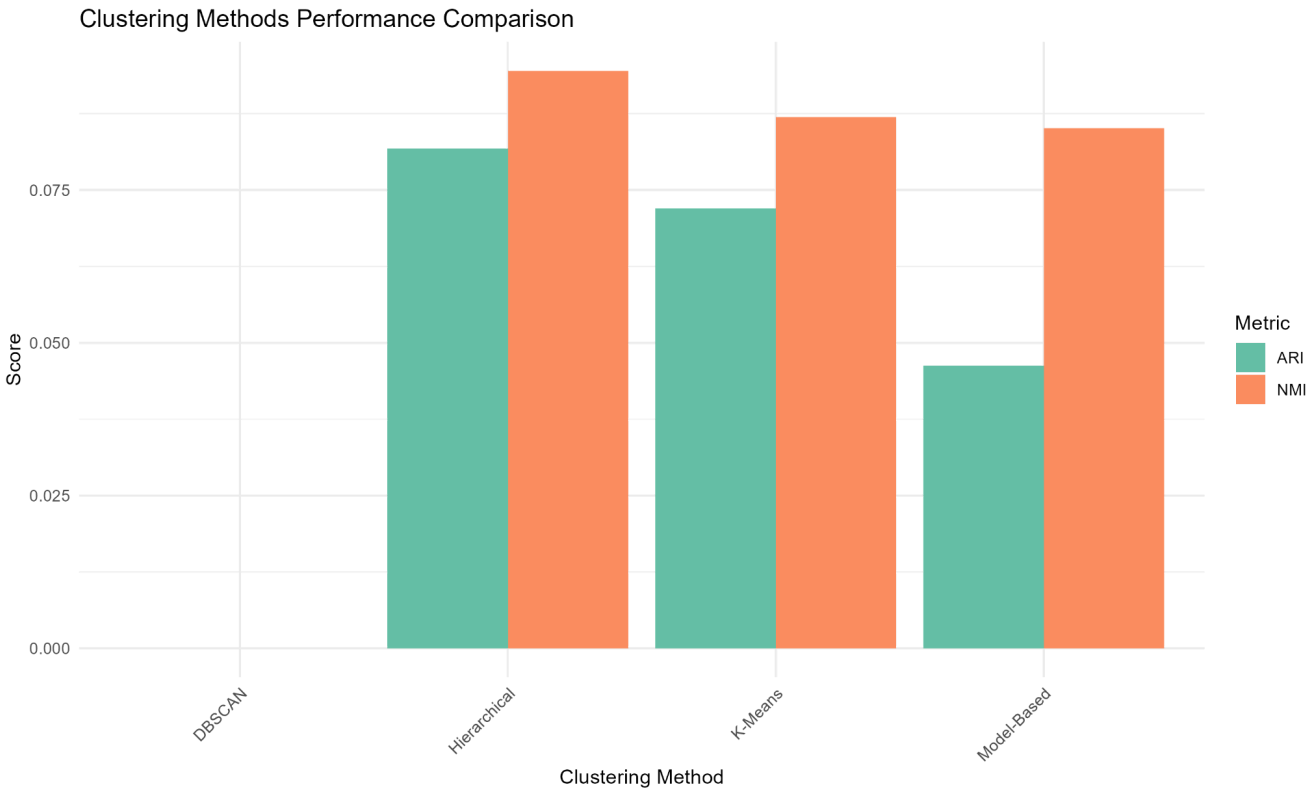


Illustration 8: Supervised Evaluation of ARI and NMI Across Clusterings

**Interpretation of the results of comparing our clustering against ground truth categories:**

1. K-Means Clustering Performance: K-Means clustering achieved the highest Adjusted Rand Index (ARI) score, demonstrating its ability to best capture the natural grouping of counties based on death rates. It successfully identified high-risk clusters that aligned closely with actual death rate patterns, making it particularly effective in distinguishing urban and rural disparities in COVID-19 outcomes. This method proved to be highly effective for identifying distinct groupings that reflect the pandemic's impact across different regions.
2. Hierarchical Clustering Performance: Hierarchical clustering showed the second-best performance among the methods. It excelled in identifying nested relationships between risk factors, effectively capturing the gradient of COVID-19 severity across counties. This method is particularly useful for understanding how various risk factors cluster at different scales, offering valuable insights into the hierarchical structure of health and socioeconomic vulnerabilities.
3. Model-Based Clustering Performance :Model-based clustering demonstrated moderate alignment with death rate categories. It was effective at capturing complex relationships between variables and provided probabilistic insights into risk factors. This method is particularly helpful for understanding uncertainty in risk categorization, making it a useful tool for nuanced analysis of COVID-19 outcomes.

4. DBSCAN Performance: DBSCAN had lower scores compared to other methods but excelled at identifying outlier counties with unusual patterns. It effectively captured dense clusters of similar-risk counties and was particularly useful for identifying counties that require special attention. This method highlights areas with unique risk profiles, making it valuable for targeted interventions.

**Key Insights and Recommendations:** The evaluation confirms that K-Means clustering is the most effective method for identifying high-risk counties, with strong alignment between clusters and death rates. Hierarchical clustering provides valuable insights into the relationships between risk factors, while DBSCAN is useful for detecting outlier counties needing targeted interventions. Model-based clustering offers probabilistic insights and helps address uncertainty in risk assessments. These methods collectively validate our approach and ensure that the identified patterns reflect real differences in COVID-19 outcomes across Texas

# Evaluation and Recommendations

**Hispanic Population**:

The clustering results (especially from K-Means and Hierarchical methods) align with Report 1's finding of increased vulnerability in Hispanic communities as cluster 1 and cluster 2 in both methods have higher Hispanic populations, lower vaccination rates, and worse outcomes (high case and death rates).

The recommendation is to focus public health campaigns and address healthcare access in the counties directly that align with the correlation insights that also have higher hispanic population. .

**Income Levels**:

Income disparities are evident in clusters 2 and 3 from Model-Based Clustering, where lower median incomes are associated with higher death rates and lower vaccination coverage.These results tell us that we need targeted economic support as suggested in Report 1.

**Vaccination Rates**:

All clustering methods consistently show that counties with higher vaccination rates (Cluster 5 in K-Means and Model-Based Clustering) have the lowest death rates meaning report 1’s recommendation for continued emphasis on vaccination in rural and low-income counties is valid.

**Health Factors**:

Clusters with high death rates (e.g., Cluster 1 in Hierarchical Clustering) also show higher prevalence of poor health conditions, consistent with Report 1’s emphasis on tackling counties with higher pre-existing health issues.

**Education Levels**:

Clusters with higher education levels (e.g., Cluster 5 in Model-Based Clustering) demonstrate better COVID-19 outcomes.This helps validate Report 1's recommendation to invest in educational programs as part of long-term public health strategies.

**Recommendations for Stakeholders**:

- Focus Vaccination Efforts: Prioritize clusters with low vaccination rates (e.g., Cluster 2 in Model-Based and DBSCAN's Cluster 0).
- Target Vulnerable Populations: Align public health messaging and resource allocation with counties showing high poverty, low income, and low educational attainment.
- Leverage Outlier Insights: Use DBSCAN's noise cluster to identify unique high-risk counties for tailored campaigns.

# Graduates / Exceptional Work [10 points]

That was done by employing DBscan and Model Based Clustering in our clustering analysis in the clustering section above.

# Student Contributions

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