DATA MINING PROJECT 2

Cluster Analysis

Dhruvil Patel 49375423
Juhi Shah 49351308
Vishakha Satpute 49258111

ABSTRACT

This report focuses on examining the impact of COVID-19 in California in relation to various demographic and socio-economic factors. The project utilizes clustering techniques, specifically K-means and Hierarchical Clustering, to identify patterns and groupings within a dataset that combines COVID-19 case data with census information. It emphasizes the importance of data preparation, including handling outliers and normalizing data, to ensure accurate and meaningful analysis. By incorporating methods such as outlier management and data normalization, we ensured the accuracy and comparability of our findings across various population sizes.

The analyses reveal distinct demographic and socioeconomic impacts of the pandemic, emphasizing the necessity for tailored public health responses and intervention strategies. The clustering approach autonomously identifies similarities and differences among the data, providing critical insights without predefined criteria. These insights are essential for stakeholders, including public health officials and policymakers, who rely on precise and actionable data to allocate resources efficiently and design effective interventions.

The findings underscore the value of rigorous data analysis in public health contexts, offering a clearer understanding of the pandemic's impact and guiding targeted actions to mitigate its effects.

Contents

LIST OF FIGURES

LIST OF TABLES

DATA PREPARATION

Preparing our data meticulously is crucial for extracting meaningful insights, when exploring critical issues like COVID-19, encompassing aspects of individuals' lifestyles, mobility, and financial status. In our preceding project, we diligently ensured the integrity of our data, eliminating any gaps or redundancies. This rigorous preparation affords us a high degree of confidence in our data's reliability for identifying trends or clusters within it.

When discussing the identification of patterns and groupings, we're referring to the concept of clustering. This method enables us to spot similarities or differences among items autonomously, without predefined criteria set by us. The data attributes we select, such as residential locations, ages, modes of transportation, and income levels, are ideally suited for clustering. This approach facilitates our comprehension of the diverse impacts of COVID-19 across various demographics and regions. For instance, insights into areas predominantly inhabited by the elderly or households with children can inform tailored assistance strategies.

Financial details are crucial for understanding how the pandemic disproportionately impacts various communities. For instance, areas with high unemployment rates or where a substantial portion of income goes towards rent might experience heightened difficulties. Observing living conditions, such as whether people reside in spacious homes or cramped apartments, and their commuting habits can indicate potential virus transmission routes.

Furthermore, demographic information, including population counts, age groups, and family structures, provides valuable insights. It helps identify populations in need of more extensive health education and support. The modes of transportation to work or school are also pivotal in tracing the virus's spread, particularly in regions with prevalent public transit usage.

In essence, these pieces of information contribute to a comprehensive understanding of COVID-19's varied impacts. By integrating these factors, we aim to gain a clearer picture of the current situation and identify effective intervention strategies. This process resembles assembling a complex puzzle, where each piece clarifies a segment of the broader strategy required to safeguard public health.

An equally important aspect of our data preparation involves managing outliers—data points that significantly deviate from the norm. Identifying anomalies, such as a small town with disproportionately high COVID-19 cases, is crucial because they can distort analytical outcomes. For effective clustering, recognizing, and addressing these outliers is imperative to prevent them from misleading our analysis.

Additionally, data normalization is a key step in our process. This adjustment ensures all data are on a comparable scale, facilitating accurate cross-comparisons, such as adjusting illness rates to a per capita basis across different population sizes. This standardization unveils genuine patterns and correlations, devoid of misconceptions due to size or scale discrepancies.

In summary, by meticulously addressing outliers and normalizing data, our analysis becomes more robust, offering a truer depiction of the pandemic's progression across communities. This detailed preparation lays the groundwork for uncovering actionable insights, navigating the complex landscape of the pandemic with greater precision, and understanding.

Table 1: Covid Cases and Census Important Features

Feature	Scale	Information
Confirmed_cases	Ratio	Number of confirmed COVID-19 cases.
Deaths	Ratio	Number of COVID-19 related deaths.
Median_income	Ratio	The median household income.
Income_per_capita	Ratio	Average income earned per person.
Median_age	Ratio	Median age of the population.
Total_pop	Ratio	Total population count.
Male_pop	Ratio	Total male population count.
Female_pop	Ratio	Total female population count.
White_pop	Ratio	Total population identifying as White.
Black_pop	Ratio	Total population identifying as Black or African American.
Asian_pop	Ratio	Total population identifying as Asian.
Hispanic_pop	Ratio	Total population identifying as Hispanic or Latino.
Percent_income_spent_on_rent	Ratio	Percentage of income spent on rent.
Vacant_housing_units	Ratio	Number of unoccupied housing units.
Housing_units	Ratio	Total number of housing units.
Median_rent	Ratio	Median monthly rent cost.
Owner_occupied_housing_units	Ratio	Number of housing units occupied by the owner.
Renter_occupied_housing_units_paying_cash_median_gross_rent	Ratio	Median gross rent among renter-occupied units paying in cash.
Families_with_young_children	Ratio	Number of families with young children.
Unemployed_pop	Ratio	Number of unemployed individuals.
Civilian_labor_force	Ratio	Number of people in the civilian labor force.
Employed_pop	Ratio	Number of employed individuals.
Not_in_labor_force	Ratio	Number of individuals not in the labor force.
Education levels (bachelors_degree_or_higher_25_64)	Ordinal	Number of individuals aged 25-64 with a bachelor's degree or higher.
Commuters_by_public_transportation	Ratio	Number of individuals commuting by public transportation.

Median_year_structure_built	Ratio	Median year of construction for buildings.
Households	Ratio	Total number of households.
pop_15_and_over	Ratio	Total count of individuals aged 15 years and over, important for assessing the majority of the adult and working-age population.
pop_5_years_over	Ratio	Represents the total number of individuals aged 5 years and over, focusing on the population eligible for schooling and beyond.
pop_25_years_over	Ratio	Total individuals aged 25 years and over, highlighting the adult demographic likely to have completed higher education or engaged in the workforce.
population_3_years_over	Ratio	Count of individuals aged 3 years and over, including preschool-aged children and older, relevant for early childhood development studies.
pop_25_64	Ratio	Total population within the prime working ages of 25 to 64 years, crucial for analyses related to the labor market and economic contributions.

The reason for selecting these feature is mentioned below.

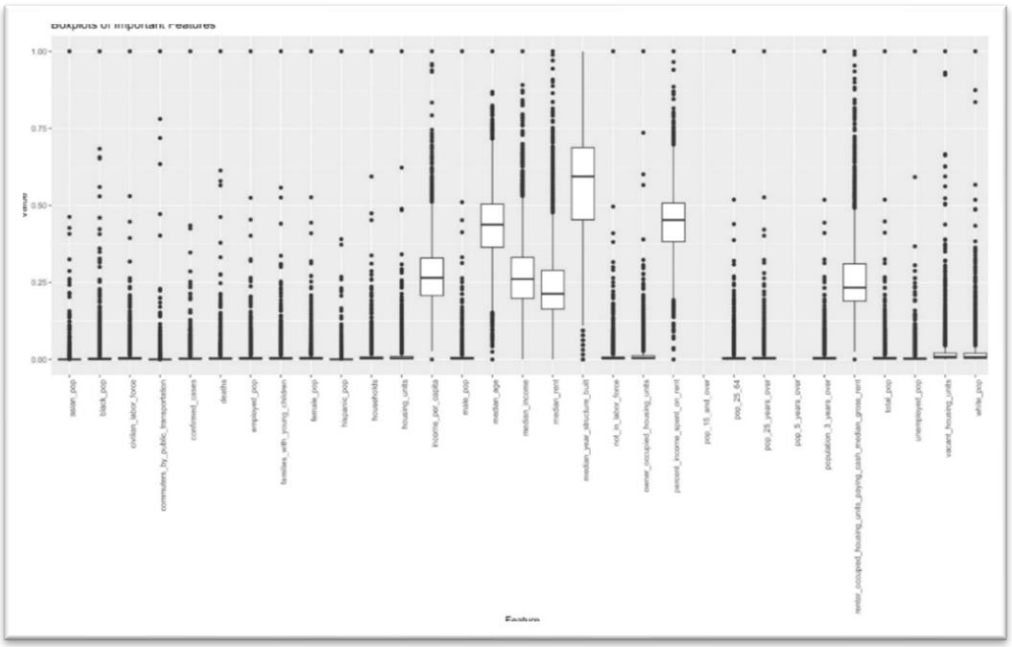

Figure 1: Outliers.

In figure 1, we have identified outliers in our dataset using boxplots. Outliers are observations that lie at an abnormal distance from other values in the data. Boxplots are particularly useful for this purpose because they visually show the spread of the data and the outliers as points outside the whiskers. After identifying these outliers, we have removed them from the dataset. This step is crucial because outliers can skew the results of data analysis, particularly clustering, as they can affect the mean and standard deviation of the data.

In the image above:

- Each boxplot corresponds to a different feature or variable within the dataset.
- The central line in each box represents the median of the data.
- The top and bottom edges of the box indicate the Q3 and Q1, respectively.
- The whiskers extend to the most extreme data points that are not considered outliers.
- Points outside the whiskers are the outliers, which are considered anomalous within the context of this data.

Scale of Measurement and Appropriate Distance Measurement

Statistical Summary of Dataset

Table 2: Statistical Summary of Dataset

Feature	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Confirmed Cases	0	0.000794	0.0019115	0.0075392	0.0049421	1
Deaths	0	0.000861	0.0022245	0.0089571	0.0055253	1
Median Income	0	0.1981	0.2611	0.2764	0.3308	1
Income per Capita	0	0.2073	0.2648	0.2775	0.3288	1
Median Age	0	0.3638	0.4375	0.4365	0.5045	1
Total Population	0	0.001076	0.002535	0.010102	0.006667	1
Male Population	0	0.0011	0.002562	0.010092	0.006716	1
Female Population	0	0.001058	0.002507	0.010113	0.006647	1
White Population	0	0.003016	0.007541	0.023448	0.019978	1
Black Population	0	0.0000775	0.0006182	0.0102389	0.004401	1
Asian Population	0	0.0000215	0.0000957	0.0037483	0.0004939	1
Hispanic Population	0	0.000066	0.0002095	0.0036753	0.0009949	1
Percent Income Spent on Rent	0	0.3825	0.4525	0.4456	0.5075	1
Vacant Housing Units	0	0.004091	0.008854	0.023065	0.020481	1
Housing Units	0	0.001539	0.003521	0.012271	0.008904	1
Median Rent	0	0.1633	0.2131	0.2434	0.2887	1
Owner Occupied Housing Units	0	0.002085	0.004704	0.015957	0.011971	1
Renter Occupied Housing Units Paying Cash Median Gross Rent	0	0.1893	0.2332	0.265	0.3102	1

Families with Young Children	0	0.000983	0.0023811	0.0100577	0.0062983	1

The selected features for Subset 1 focus on demographic data and COVID-19 impact. They are used because:

1. **confirmed_cases and deaths**: These are direct indicators of the impact of COVID-19 and are essential for any analysis related to the pandemic.
2. **median_age**, **total_pop**, **male_pop**, **female_pop**: These provide a snapshot of the population structure, which is vital as COVID-19 affects age groups differently.
3. **white_pop**, **black_pop**, **asian_pop**, **hispanic_pop**: Understanding the pandemic's impact across different racial and ethnic groups is important for identifying disparities and targeting interventions.
4. **households**: This can give insights into living conditions that may affect the spread and impact of COVID-19.
5. **pop_15_and_over, pop_5_years_over, pop_25_years_over, population_3_years_over, pop_25_64:** These age brackets offer detailed insights into potentially economically active groups, dependency ratios, and age-specific vulnerabilities to COVID-19.

These features provide insights into how COVID-19 affects different demographics and can help in crafting targeted public health responses.

Table 3: Subset 1 Demographic Features

Feature	Scale	Description
confirmed_cases	Ratio	Number of confirmed COVID-19 cases.
deaths	Ratio	Number of COVID-19 related deaths.
median_age	Ratio	Median age of the population.
total_pop	Ratio	Total population count.
male_pop	Ratio	Total male population count.
female_pop	Ratio	Total female population count.
white_pop	Ratio	Total population identifying as White.
black_pop	Ratio	Total population identifying as Black or African American.
asian_pop	Ratio	Total population identifying as Asian.
hispanic_pop	Ratio	Total population identifying as Hispanic or Latino.
households	Ratio	Total number of households.
pop_15_and_over	Ratio	Number of individuals aged 15 years and over.
pop_5_years_over	Ratio	Number of individuals aged 5 years and over.

pop_25_years_over	Ratio	Number of individuals aged 25 years and over.
population_3_years_over	Ratio	Number of individuals aged 3 years and over.
pop_25_64	Ratio	Number of individuals within the age range of 25 to 64 years.

Subset 2 features are tailored toward capturing the economic status and living conditions of the population, which are also key factors in understanding the spread and impact of COVID-19. Here's why each feature is important:

1. **median_income** and **income_per_capita**: These are fundamental indicators of the economic well-being of an area's residents, which can influence access to healthcare and resources to cope with the pandemic.
2. **percent_income_spent_on_rent**: This may reflect financial stress in households, which could impact the ability to afford healthcare or to comply with public health measures like staying home when ill.
3. **vacant_housing_units** and **housing_units**: Vacancy rates can be indicative of economic health or migration patterns, while the total housing units give a sense of population density, affecting virus transmission rates.
4. **median_rent**: High rent could be a stressor impacting public health, especially if combined with low income.
5. **owner_occupied_housing_units**: Ownership rates can be connected to economic stability, which might affect how communities experience and respond to COVID-19.
6. **renter_occupied_housing_units_paying_cash_median_gross_rent**: This offers insight into the burden on renters, potentially affecting their financial ability to handle emergencies like a pandemic.
7. **unemployed_pop**, **civilian_labor_force**, **employed_pop**, **not_in_labor_force**: These employment-related statistics can reveal economic vulnerabilities or strengths in the population, influencing both the spread and the effects of COVID-19.
8. **commuters_by_public_transportation**: Higher use of public transportation may correlate with higher risk of virus transmission.
9. **median_year_structure_built**: Older structures may affect residents' health differently compared to newer constructions and may reflect infrastructure investments and living conditions.

Overall, these features can reveal social determinants of health and economic factors that are likely to influence how populations are affected by the pandemic, informing public health interventions and policies.

Table 4: Subset 2 Money-Related Features

Feature	Scale	Description
median_income	Ratio	The median household income.
income_per_capita	Ratio	Average income earned per person.
percent_income_spent_on_rent	Ratio	Percentage of income that households spend on rent.
vacant_housing_units	Ratio	Number of unoccupied housing units.

housing_units	Ratio	Total number of housing units.
median_rent	Ratio	Median monthly rent cost.
owner_occupied_housing_units	Ratio	Number of housing units occupied by the owner.
renter_occupied_housing_units_paying_cash_median_gross_rent	Ratio	Median gross rent among renter-occupied units paying in cash.
unemployed_pop	Ratio	Number of unemployed individuals.
civilian_labor_force	Ratio	Number of people in the civilian labor force.
employed_pop	Ratio	Number of employed individuals.
not_in_labor_force	Ratio	Number of individuals not in the labor force.
commuters_by_public_transportation	Ratio	Number of individuals commuting by public transportation.
median_year_structure_built	Ratio	Median year of construction for buildings.

Distance Measure for Clustering

In clustering analysis, especially hierarchical clustering, understanding the proximity or dissimilarity between data points is crucial. Some of distance measure for clustering are Manhattan Distance, Minkowski Distance, Euclidean distance.

The Euclidean distance measure is a common and intuitive way to quantify this dissimilarity. It calculates the "straight line" distance between two points in multidimensional space, making it a natural choice for many clustering tasks. When we create a dissimilarity matrix using the Euclidean distance method, what we're essentially doing is comparing each data point with every other data point in our dataset based on the features we've selected.

This matrix is a foundational step in hierarchical clustering because it informs the algorithm how closely related (or not) each pair of observations is. Hierarchical clustering then uses this information to start grouping the most similar observations together into clusters, iteratively joining clusters together based on their proximity until all data points are nested within larger and larger clusters.

This process is visualized as a dendrogram, a tree-like diagram that shows the order and distance at which points are merged. The Euclidean distance is particularly useful because it's easy to understand and interpret. For example, in a 2-dimensional space (like plotting two features against each other), the Euclidean distance corresponds to the direct length of the line connecting two points. As more dimensions (or features) are added, this concept extends into higher-dimensional spaces, though it becomes harder to visualize directly.

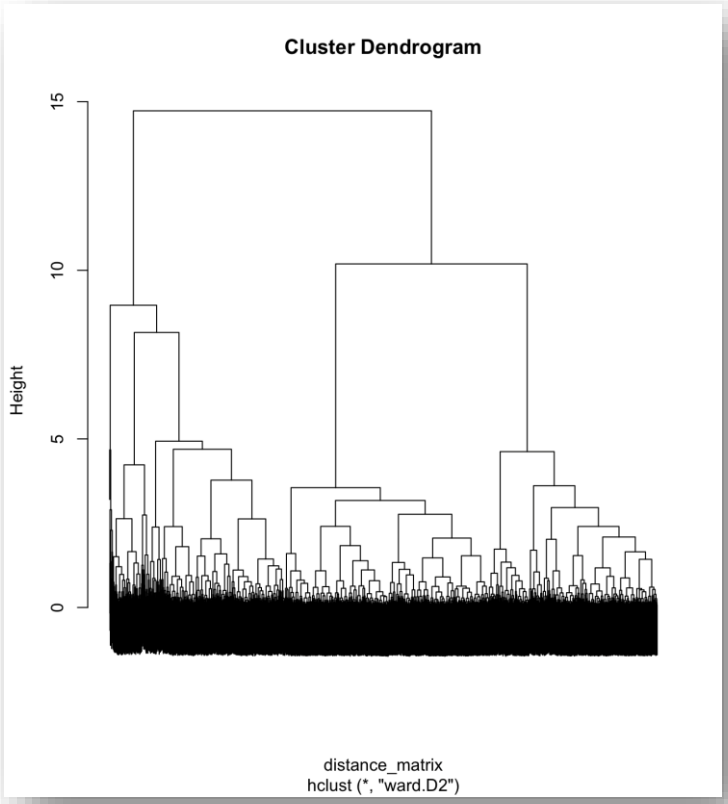

Figure 2: Cluster Dendrogram

MODELING

SUBSET 1

K Means Clustering

In our recent analysis, we explored a range of demographic factors to uncover patterns that could help us understand the spread and impact of COVID-19 across different communities in California. By employing Principal Component Analysis (PCA) and K-means clustering techniques on this data, we've identified distinct groups or 'clusters' that share common characteristics.

PCA is a process that simplifies complex data, making it easier to see overarching trends. Think of it like reducing a multi-coloured image to just a few shades; the nuances are lost, but the main picture is clearer. In our study, PCA condensed numerous demographic factors into two main components. The first component accounted for a vast majority of the variation in the data (over 80%), suggesting that there is one main factor that really sets our clusters apart from each other. The second component, while less influential, still captured an important part of the data's story.

K-means clustering further distilled these insights, grouping the data into three clusters, each representing different demographic profiles with potentially varying levels of COVID-19 cases and mortality. These clusters were visualized on a scatter plot, distinguished by different colours and shapes,

and encased in rectangles to approximate their spread within the two main components identified by PCA.

The striking dominance of the first principal component could imply a significant influence of one or two demographic features, such as population density or median age, on the COVID-19 outcomes. This kind of insight is invaluable. It tells us, in simpler terms, that one major factor could be a driving force behind the trends we see in the virus's spread in different parts of California. For example, areas with older populations might be hit harder and thus might need more resources like hospitals or vaccine centres.

This analysis is more than just numbers and graphs; it's a tool that can help policymakers make informed decisions about where to focus their efforts. By understanding which communities are most at risk, resources can be allocated more effectively, potentially saving lives. The three clusters we've identified are a first step towards a more nuanced strategy in combating the pandemic—each has its own story, with tailored solutions required to address their unique challenges.

The silhouette method and the elbow method are both used to determine the optimal number of clusters in a dataset for clustering algorithms like K-Means and hierarchical clustering.

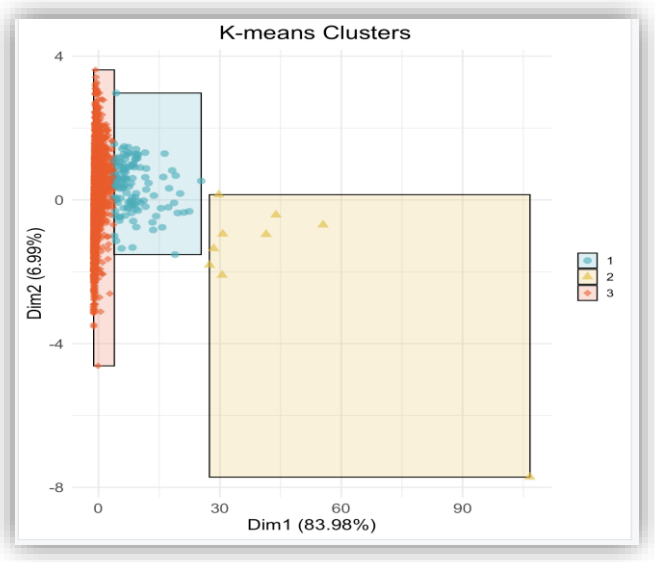

Figure 3: K-means Clusters for Subset 1

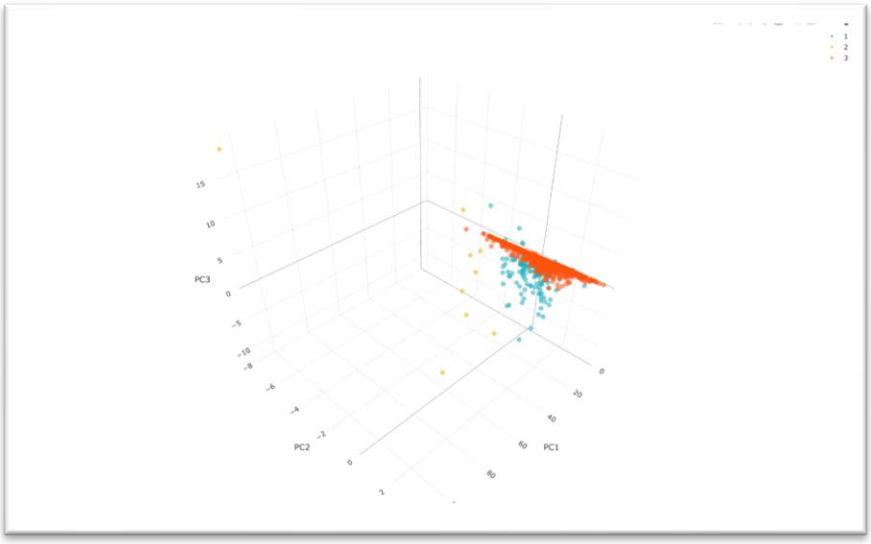

Figure 4: 3D Representation of K-means Clustering for Subset 1

Figure 4 represents a Principal Component Analysis (PCA) of K-means clustering results. In the plot, data points are coloured differently to indicate the clusters they belong to, as determined by the K-means algorithm. PCA is used to reduce the dimensionality of the data, distilling the information into three principal components (PC1, PC2, and PC3) to visualize the clustering in three-dimensional space. This type of visualization helps to understand the separation and grouping of data in a reduced dimensional space after applying clustering techniques.

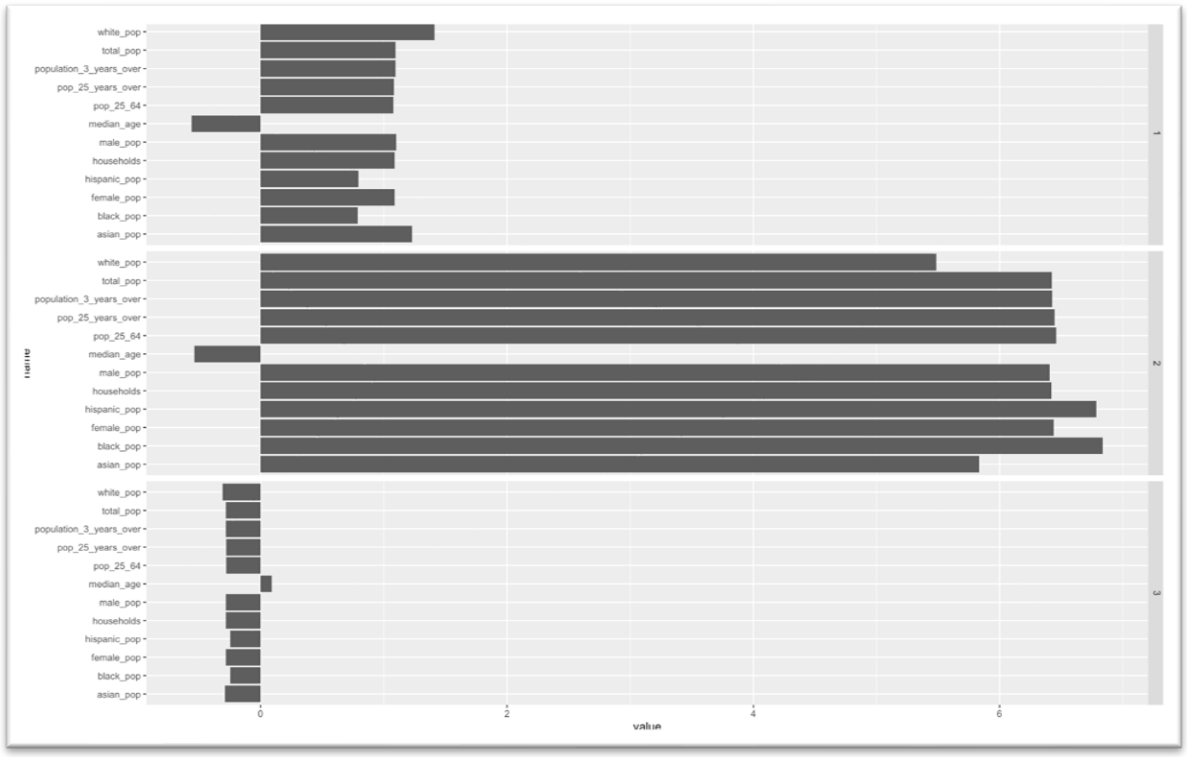

Figure 5: K-means Subset 1 clustering profile.

In figure 5, we are displaying a series of variables (features) on the y-axis and their corresponding values on the x-axis. It seems to show the importance or frequency of various demographic variables in a

dataset, such as the population over 3 years old, population over 25 years old, median age, and populations segmented by gender and ethnicity (e.g., male population, Hispanic population, white population).

Each bar's length represents the value of the associated variable, with longer bars indicating higher values. This type of chart is typically used to compare the magnitude of different variables or to rank them in order of significance or frequency. It appears there are two sets of variables, possibly representing different groups, scenarios, or time periods for comparison.

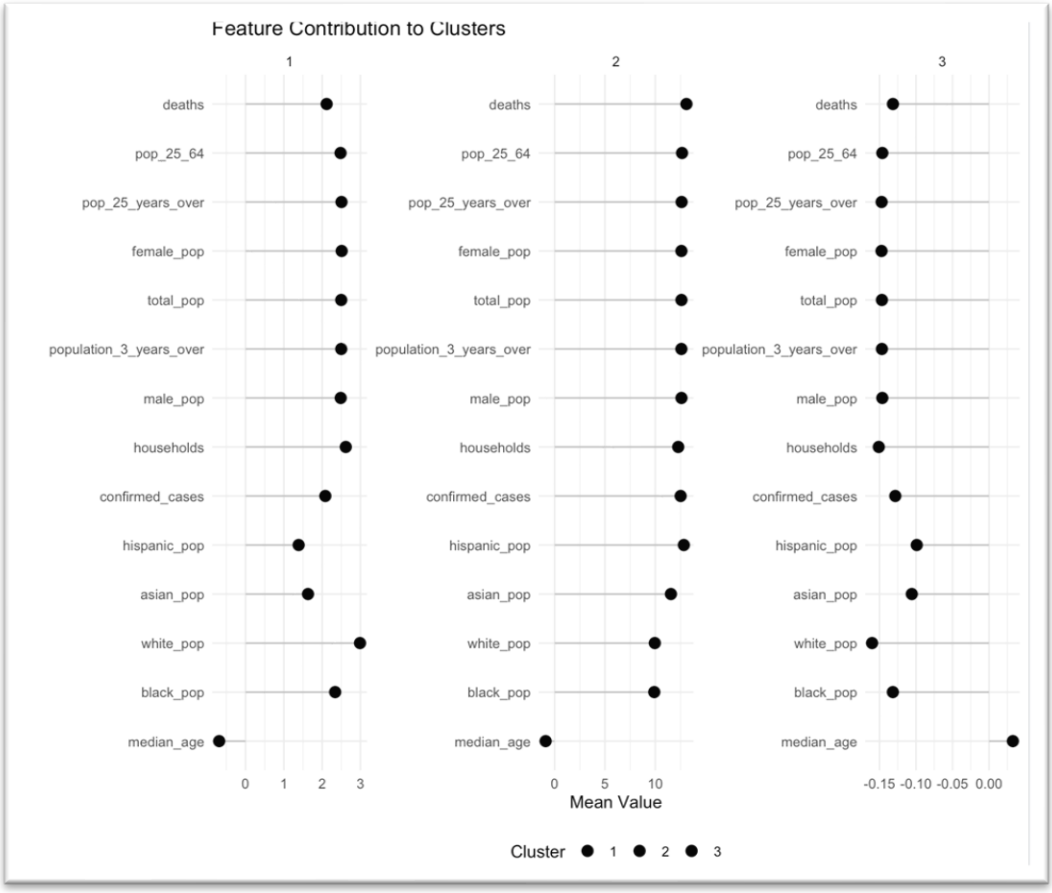

Figure 6: Feature Contribution to Clusters for Subset 1

In the visual analysis of our dataset through clustering, we've uncovered distinct groupings that reveal patterns in the demographics of the population affected by COVID-19. These patterns are depicted in the lollipop charts, where each cluster brings forward a unique demographic footprint defined by the average values of features like age distribution, population size, and confirmed cases, among others.

The charts present a simplified yet informative snapshot of each cluster. For example, one cluster might show a higher average in the 'deaths' feature, which could indicate a group with a higher mortality rate associated with the virus. This cluster may demand more healthcare resources and targeted interventions to manage the impact of the pandemic effectively.

In contrast, another cluster with a higher average in 'confirmed_cases' yet a lower 'deaths' feature suggests a demographic that, while having a higher incidence of COVID-19, may be faring better in terms of survival. This could be due to a younger population, as indicated by lower averages in the 'median_age' feature, or better healthcare systems and practices in place.

Similarly, differences in the 'total_pop' feature across clusters can indicate the density of the population affected, which is a critical factor in the spread of infectious diseases. A cluster with a higher 'total_pop' average could represent urban areas where high population density accelerates the transmission of the virus, necessitating distinct public health strategies compared to less densely populated clusters.

Figure 6 titled ‘Feature Contribution to Clusters for Subset 1’ serve as a basis for strategic planning. Policymakers and public health officials can use these insights to tailor responses and allocate resources where they are most needed. The visual format of the data aids in quickly identifying which demographics are most vulnerable or most affected, ensuring that measures can be as proactive and precise as possible.

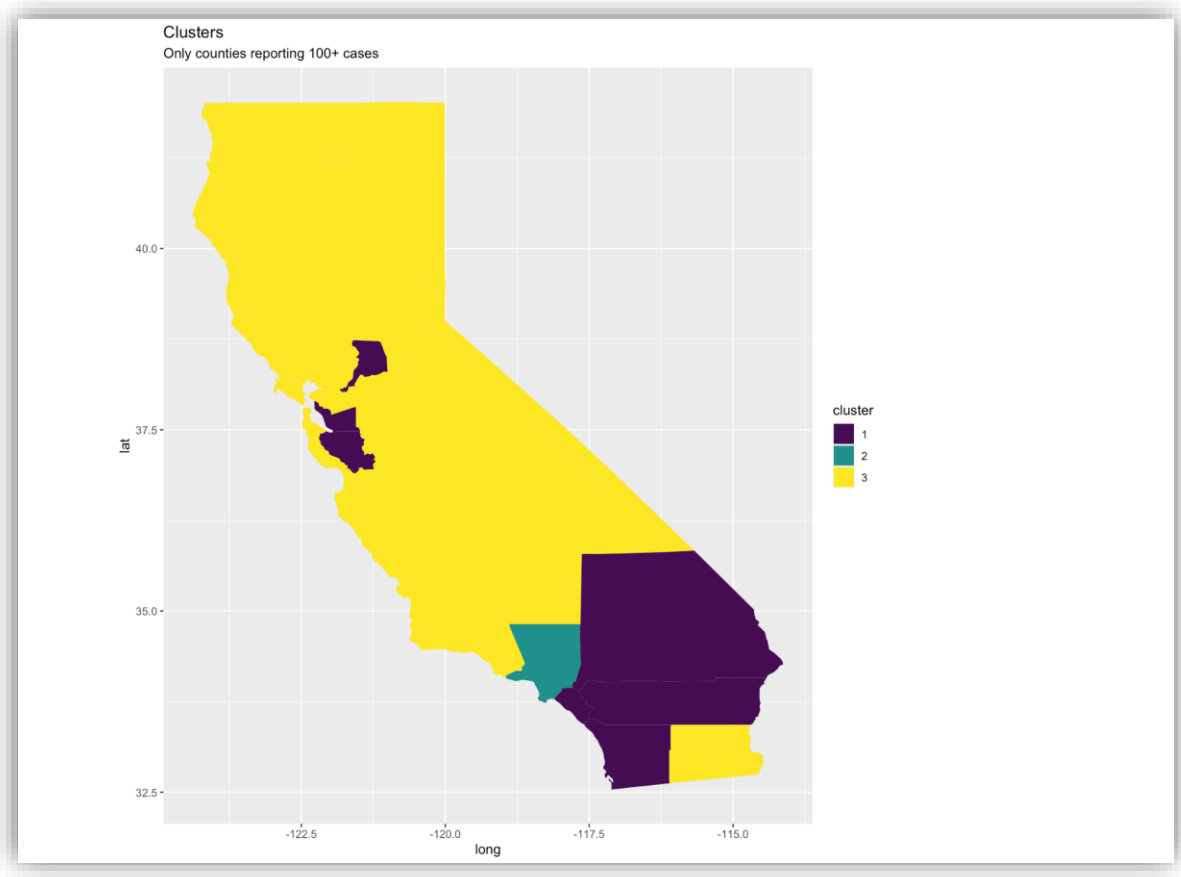

Figure 7: Subset 1 K-Means clustering mapping for subset 1

Figure 7 illustrates the results of a clustering analysis on California county data related to COVID-19 cases. Using geospatial visualization, we can observe the grouping of counties into three distinct clusters, differentiated by colour. The underlying data driving the clustering likely includes various metrics related to the pandemic—such as confirmed cases and deaths—alongside demographic information like age distribution and population metrics. The clusters represent patterns or similarities within the data that the k-means algorithm has identified.

The visualization is effective in conveying how these clusters are geographically distributed across the state of California. The use of colour not only differentiates the clusters but also provides a quick visual cue to the viewer about the concentration of similar characteristics within geographic regions. For instance, we might infer that counties within the same cluster could share demographic similarities or have experienced similar trajectories in the spread of COVID-19. This can be particularly useful for public health officials to tailor responses or resources allocation that are specific to each cluster's needs.

It is important to note that the colour scale, which ranges from dark purple to light yellow, is applied to the counties in such a way that no clear geographic pattern is immediately apparent from north to south or east to west. This suggests that the clustering is not based on geographical proximity alone but is likely a result of the multidimensional nature of the data considered in the analysis. The variance in cluster distribution could indicate that factors such as population density, mobility, healthcare accessibility, or socioeconomic status, which often do not follow simple geographic patterns, are playing significant roles in the clustering result.

The subtitle "Only counties reporting 100+ cases" implies a filtering criterion that has been applied to the dataset before the analysis, focusing the clustering on areas with a significant number of cases. This decision likely serves to concentrate the analysis on regions where the COVID-19 impact was more pronounced and the clustering could yield more meaningful distinctions. By excluding areas with fewer cases, the analysis avoids the noise that low case counts might introduce, thus seeking to ensure that the clusters formed are more representative of the patterns in the data related to the virus's spread and impact.

Overall, this visualization is a snapshot of the analytical power of data mining techniques in understanding complex datasets. Clustering, a fundamental tool in the data science arsenal, has provided a lens to interpret the multi-faceted nature of the pandemic's impact on different regions. Public health strategies can be informed by such analysis, as it reveals hidden structures and relationships within the data that might not be apparent through a simple examination of raw numbers.

Internal Validation

Silhouette Method

The figure ‘Silhouette Method for Determining Optimal Number of Clusters for Subset 1’ below presents the silhouette scores for a range of cluster numbers, offering a quantitative approach to determining the optimal number of clusters for the k-means clustering algorithm. Silhouette scores measure how similar an object is to its own cluster (cohesion) compared to other clusters (separation), with higher values indicating better-defined clusters. The scores range from -1 to +1, where a high value signifies that the object is well matched to its own cluster and poorly matched to neighbouring clusters.

In the depicted silhouette method graph, we observe a sharp decline as the number of clusters increases from 2 to 4, followed by a more gradual decline and some fluctuations in the scores for larger numbers of clusters. The highest silhouette score appears when there are two clusters, suggesting that dividing the dataset into two clusters may provide the most distinct and well-separated grouping according to the k-means clustering algorithm. However, practical application may warrant consideration of additional factors such as domain knowledge and the specific objectives of the clustering to decide if two clusters are indeed the most meaningful division of the data.

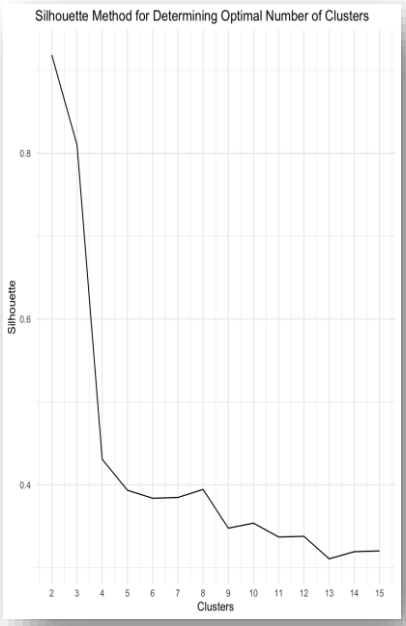

Figure 8: Silhouette Method for Determining Optimal Number of Clusters for Subset 1

Elbow Method

The figure named 'Elbow Method for Determining Optimal Number of Clusters for Subset 1' illustrates the application of the elbow method to determine the optimal number of clusters for k-means clustering. It plots the within-cluster sum of squares (WSS) against the number of clusters, revealing how the total variance within clusters decreases as the number of clusters increases. The WSS is a measure of clustering quality, with lower values indicating that data points are closer to their respective cluster centroids.

In this curve, the WSS rapidly decreases as the number of clusters increases from 1 to around 4 or 5, after which the rate of decrease slows significantly, suggesting an elbow point around this region. The "elbow" represents the point at which adding more clusters does not provide a substantial improvement in the compactness of the clusters. It's at this inflection point where the optimal number of clusters is often considered to be, as it represents a good trade-off between cluster compactness and model complexity.

The visual inspection of this graph suggests that the optimal number of clusters for this dataset might be around 4 or 5, as the elbow appears to be in this vicinity. This implies that beyond this point, the gains in reducing the WSS are marginal compared to the cost of increasing the number of clusters, hence clustering the dataset into 4 or 5 groups would likely be the most efficient use of the k-means algorithm. However, it is also crucial to corroborate this graphical insight with additional analytical methods such as the silhouette score and domain-specific considerations to confirm the most appropriate number of clusters for the dataset.

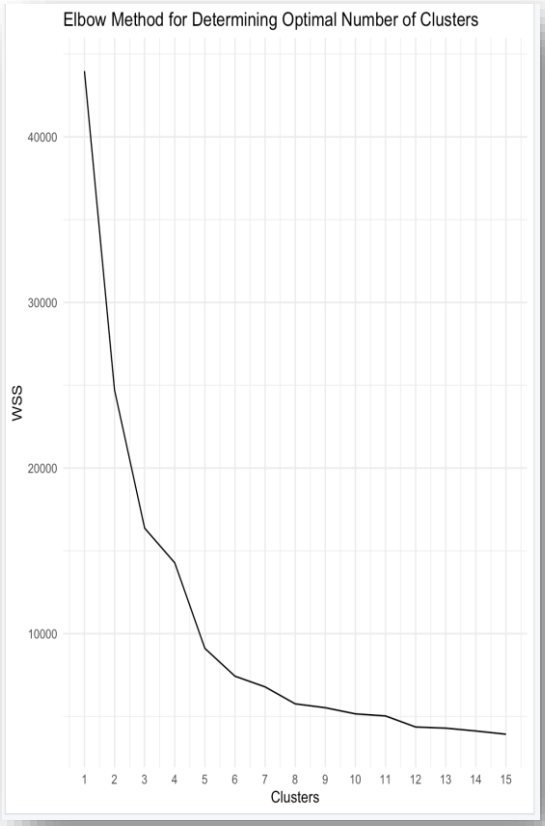

Figure 9: Elbow Method for Determining Optimal Number of Clusters for Subset 1

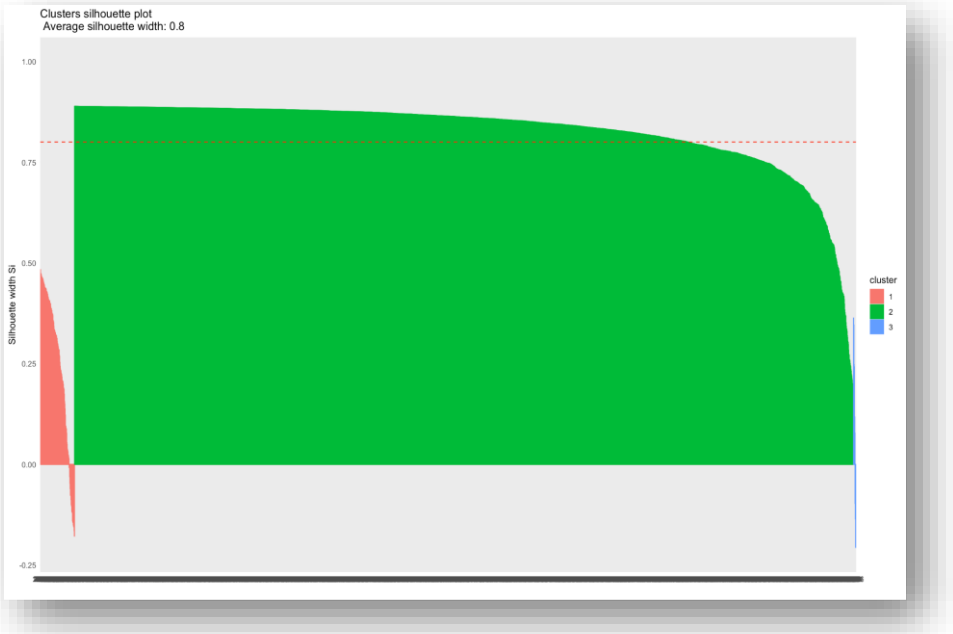

Figure 10: Clusters silhouette plot for Subset 1

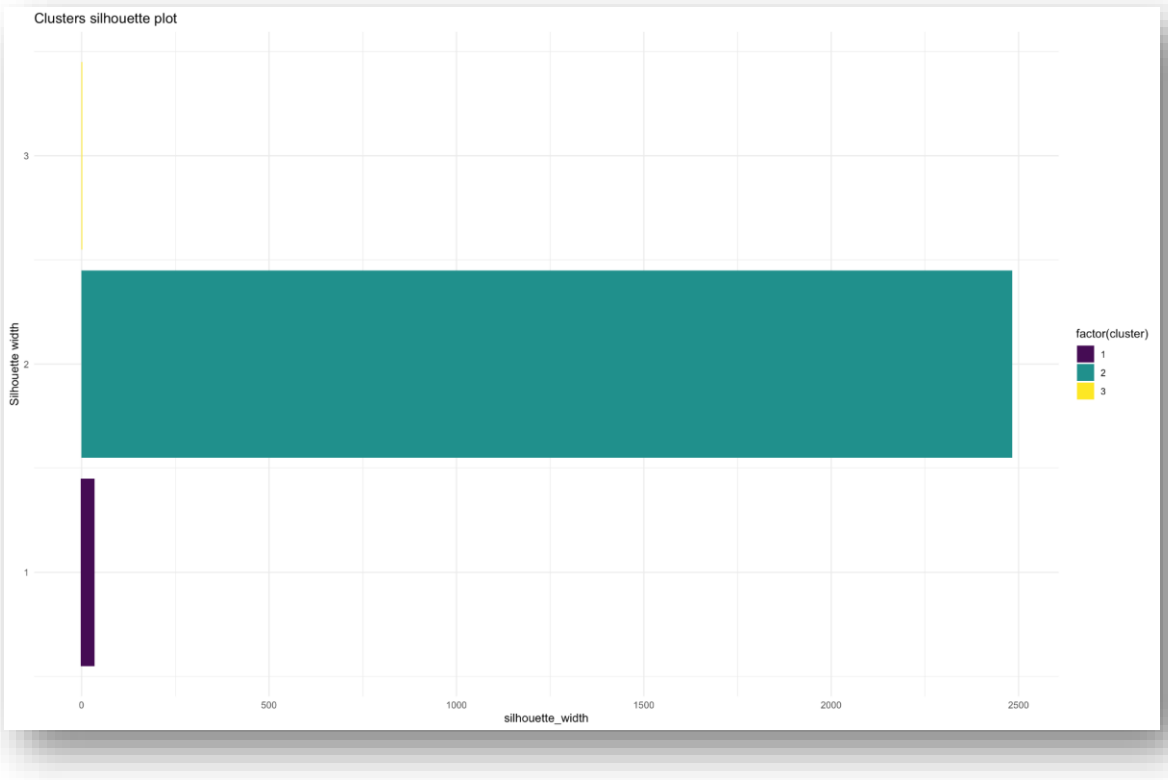

Figure 11: Clusters silhouette plot for Subset 1

The first plot appears to be a bar chart representing the silhouette widths of three clusters. The silhouette width measures how similar an object is to its own cluster compared to other clusters. Values range from -1 to 1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighbouring clusters. In this chart, each bar represents the silhouette width of a cluster, with the x-axis showing the silhouette width and the y-axis representing the cluster number. The varying heights of the bars suggest different levels of cohesion and separation for each cluster, with higher bars indicating clusters that are better defined.

The second plot is a silhouette plot that visualizes how close each point in each cluster is to the points in the neighbouring clusters. This is a more detailed plot where each point on the x-axis represents an instance in the dataset, and its position on the y-axis shows the silhouette width for that instance. Colours differentiate the clusters. We want the silhouette width to be as close to 1 as possible; values close to 0 indicate overlapping clusters, and negative values suggest that samples might have been assigned to the wrong cluster. The dashed line represents the average silhouette score across all the samples, which gives a sense of the overall cluster fit. In this case, the average silhouette width is 0.45, indicating a reasonable structure detected by the clustering process.

External Validation

We will perform External Validation as given steps: -

Define the Rate of Cases: We should decide on the specific metric for the rate of cases. This could be, for example, the number of new cases per 1000 people per week.

Set the Threshold: Determine the threshold for the rate of cases that would differentiate between high risk and low risk. This threshold should be set based on public health guidelines, statistical analysis, or historical data.

Label the Counties: Using the chosen threshold, we would label each county as 'high risk' or 'low risk' based on their rate of cases. This establishes our ground truth.

Perform Clustering: Cluster the counties using appropriate clustering methods (like K-Means or hierarchical clustering) based on other features in the dataset.

Evaluate Clustering: After clustering, we can compare the clusters to our ground truth using evaluation metrics such as Adjusted Rand Index (ARI), Variation of Information (VI), and others.

These metrics will help us to understand how well the clustering matches the actual distribution of high and low-risk counties as defined by the rate of cases. Interpret Results: Based on the values obtained from ARI and VI, we could interpret the quality of our clustering. A higher ARI value would indicate a better match between the clusters and the ground truth, whereas for VI, lower values indicate a better match.

So, in this we ran procedure for analysing COVID-19 case data to identify high-risk and low-risk counties in California. It first filters the dataset for California, then calculates a rate of COVID-19 cases per 1,000 people. Counties are classified as 'high risk' or 'low risk' based on whether their case rate exceeds a specified threshold. Next, K-Means clustering is applied to group counties based on socio-economic features like median income, percent income spent on rent, and median age, which are normalized to ensure fair contribution of each feature. The quality of clustering is evaluated using the Adjusted Rand Index (ARI) and Variation of Information (VI), metrics that measure the similarity between the clustering results and the pre-determined risk categories. For stakeholders, such as public health officials or policymakers, this analysis can pinpoint regions that require more resources or interventions and help assess whether socio-economic factors correlate with COVID-19 risk levels, aiding in informed decision-making and targeted responses.

Why confirmed cases as ground truth?

Think of "confirmed cases" as the actual number of sick people that we know about in different places. By using these numbers as the "truth," we're trying to see if the way we group places based on things like income or age matches up with where the sickest people are. If our groups do match up, it could help us figure out why some places have more sick people than others and what we might do about it.

What is role of threshold?

The threshold defines the cut-off value for the rate of COVID-19 cases per 1,000 people that distinguishes between 'high risk' and 'low risk' counties. Counties with case rates above this threshold are labelled 'high risk', implying they have a higher prevalence of COVID-19 and may require more attention or resources. Conversely, those with case rates below the threshold are labeled 'low risk', indicating a lower prevalence of the disease. The choice of this threshold is crucial as it directly influences the classification of counties and, subsequently, any public health actions that might be taken based on this analysis.

Validation Metrics Explanation

Adjusted Rand Index (ARI): This metric evaluates the similarity between two clustering by considering all pairs of samples and counting pairs that are assigned in the same or different clusters in the predicted and true clustering. The ARI corrects for chance, making it a reliable metric even for randomly assigned clustering. A value close to 1 indicates a high similarity between the clustering and the ground truth.

Mean Silhouette Width: This measure assesses how close each point in one cluster is to points in the neighbouring clusters. The silhouette width for a single data point is a ratio that subtracts the average distance to points in the nearest cluster from the average distance to points in the same cluster and divides it by the maximum of these two values. The overall mean silhouette width provides an insight into the separation distance between the resulting clusters, where higher values indicate better-defined clusters.

Purity: Purity is a simple and transparent evaluation measure. After clustering, each cluster is assigned to the most frequent class of its items (from the ground truth), and then the accuracy of this assignment is measured by counting the correctly assigned items and dividing by the total number of items. Purity thus reflects the degree to which each cluster contains data points from primarily one class.

- Adjusted Rand Index (ARI) of 0.0192: This value is close to 0, suggesting that the clustering results have little to no agreement with the ground truth labels. The clusters derived from features do not correspond well with the actual distribution of confirmed cases.
- Mean Silhouette Width of 0.449: A silhouette score ranges from -1 to 1. A score of 0.449 suggests moderate cohesion and separation; that is, on average, clusters are not too close to each other, and points within a cluster are reasonably similar. However, this value doesn't necessarily align with the ground truth labels we have for confirmed cases.
- Purity of 0.4036: Purity is a measure of the extent to which each cluster contains data points from primarily one class. A purity score of 0.4036 indicates moderate purity, meaning that while there is some homogeneity in the clusters, they are not highly pure. Many points from different classes (in this case, different levels of confirmed cases) are being grouped together.

For stakeholders, such as public health officials, the low ARI indicates that the chosen features for clustering may not be effective indicators of COVID-19 case rates, and the model may need to be reassessed or improved by exploring additional or different features. The moderate silhouette score indicates that the model has some underlying structure, but it does not necessarily align with case rates. The purity score further suggests that while there is some consistency within clusters, there is significant overlap and clusters are not distinctly defined by case rates. Thus, the clustering model as it stands may have limited usefulness for making public health decisions related to COVID-19 case distributions.

Table 5: Validation Metrics for Subset 1

Metric	Value
Adjusted Rand Index (ARI)	0.0192
Mean Silhouette Width	0.449
Purity	0.4036

SUBSET 2

K Means Clustering

Determining Optimal Cluster Count

Elbow Method

The figure titled ‘Elbow Method for Determining Optimal Number of Clusters for Subset 2’ visualizes the relationship between the number of clusters and the within-cluster sum of squares (WSS) to identify

the most suitable cluster count for our data. The WSS is a measure of variance within each cluster; lower values generally indicate that the data points are closer to their respective centroids, hence a better fit.

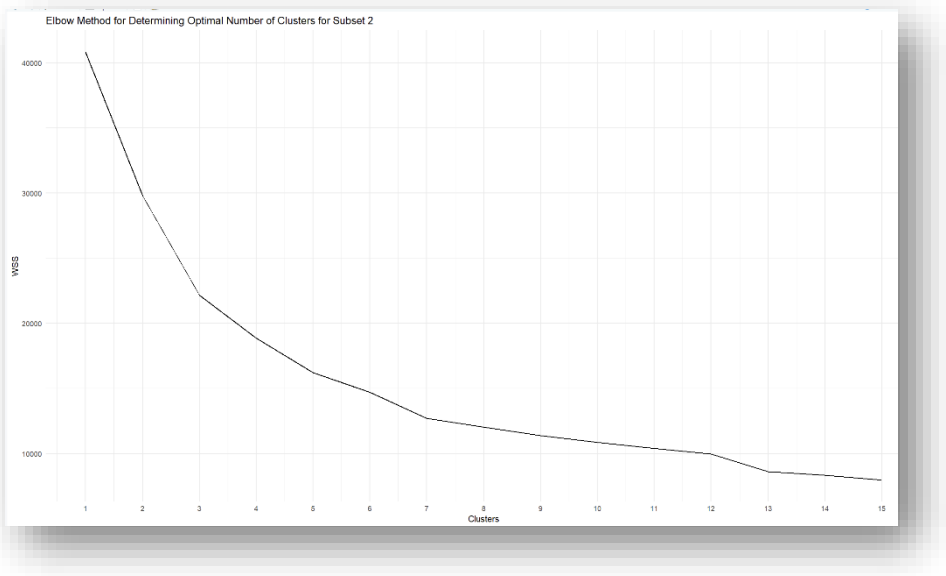

Figure 12: Elbow Method for Determining Optimal Number of Clusters for Subset 2

The graph begins with a steep downward trajectory from one cluster to two, indicating a significant gain in data fitting from the initial single cluster to two clusters. As we progress from two to fifteen clusters, the WSS decreases at a diminishing rate, which is characteristic as the clusters begin to finely tune to the variance in the data.

While the graph depicts a gradual descent without a pronounced elbow, we might observe a subtle inflection around the cluster count of 3 to 5, where the curve starts to plateau. This suggests that increasing the number of clusters beyond this range yields minimal improvement to the model. In the absence of a distinct elbow, the decision on the optimal number of clusters may also be influenced by additional considerations such as interpretability, domain-specific knowledge, or the application of alternative validation metrics.

Silhouette Method

Our examination involves silhouette scores which span from 2 to 15 clusters. Each score quantifies the average similarity of objects within their own cluster compared to other clusters, thus serving as an indicator of cluster separation quality.

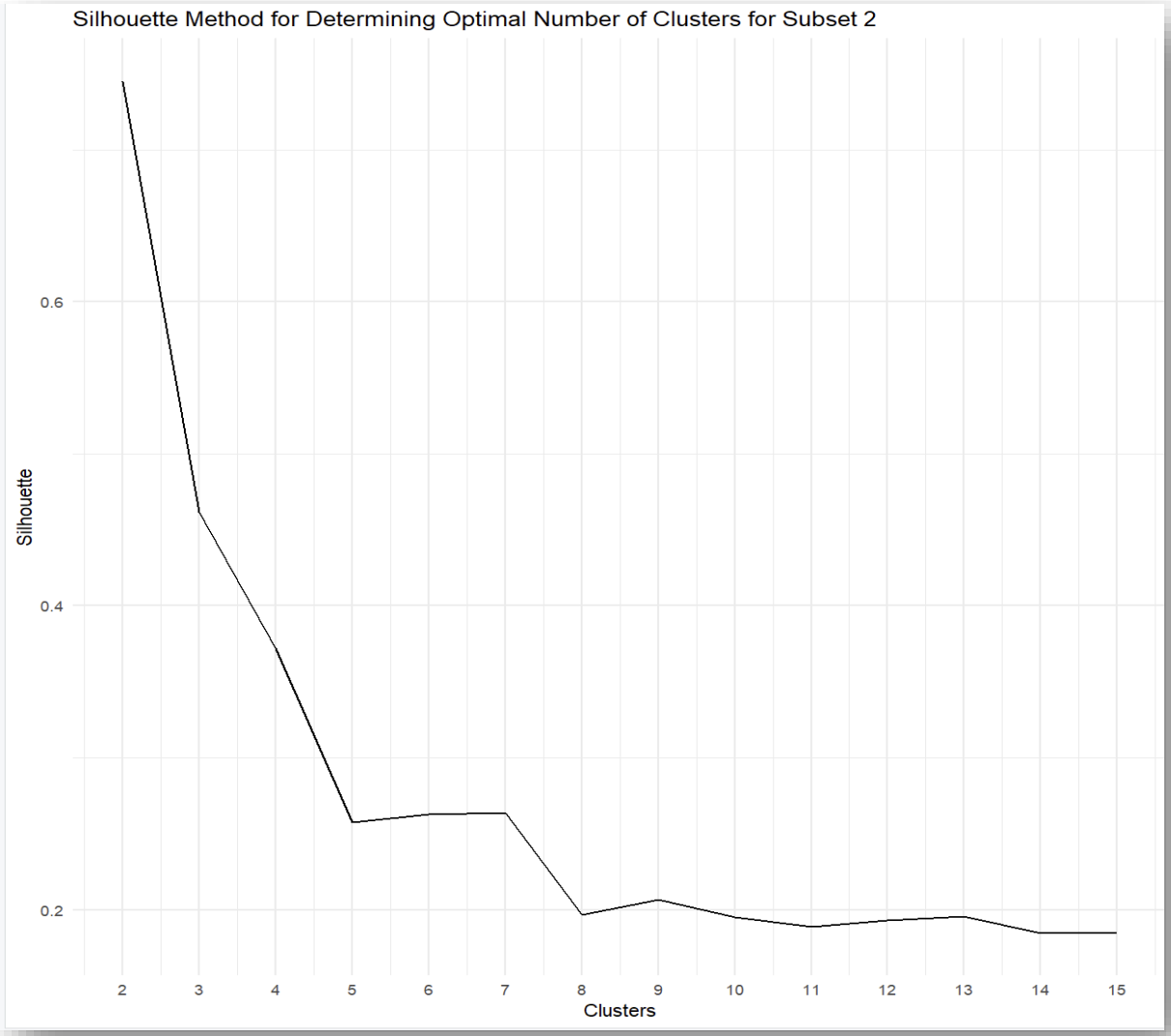

Figure 13: Silhouette Method for Determining Optimal Number of Clusters for Subset 2

Based on the Silhouette Graph, we can see that initially high silhouette scores are seen when the cluster count is low. As we introduce more clusters, there's a general trend of decreasing silhouette scores. This reflects an expectation that more clusters may yield a more nuanced but potentially less meaningful subdivision of data. Post the formation of four clusters, the scores dip, then slightly rise at a five-cluster solution before reaching a steady state. This pattern suggests limited improvement in cluster separation past this point.

The peak silhouette score, evident with two clusters, implies the clearest separation at this juncture. However, such clear separation could also point towards a potentially overly simplistic bifurcation of the dataset.

Hence, based on the silhouette analysis, the graph suggests an inclination towards fewer clusters. Nevertheless, in determining the most suitable cluster count for Subset 2, it's imperative to incorporate dataset context and the specific clustering objectives. Considering the levelling of scores beyond the 6-cluster mark, selecting between 4 to 6 clusters appears to be a judicious balance of silhouette score optimization and cluster interpretability.

The same analytic steps applied to subset 1 will be executed for subset 2. This preliminary visualization of clustering acts as a foundational stage for deeper examination, laying the initial groundwork for identifying each cluster's distinct characteristics. Such insights can drive further modeling techniques or contribute to informed decision-making in line with the dataset's wider relevance.

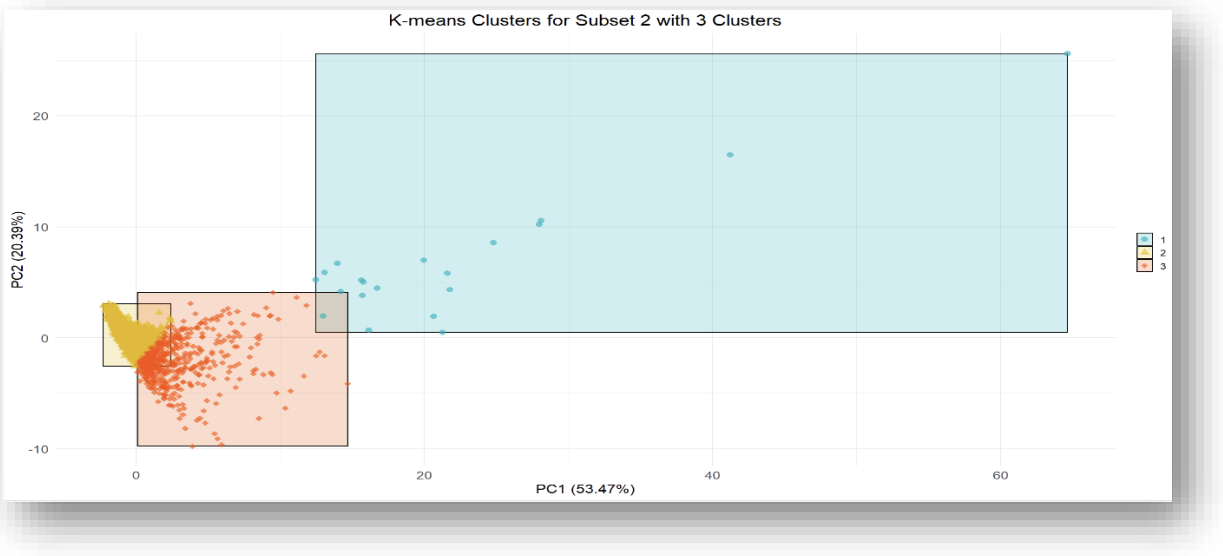

Figure 14: K-means Clusters for subset 2 with 3 clusters.

Summary of the findings in figure named ‘K-means Clusters for subset 2 with 3 clusters’:

- Cluster 1, in light blue, is more dispersed, primarily above the plot's midpoint, suggesting diverse attributes within this group.
- Cluster 2, in yellow, is tightly grouped in the lower-left, indicating strong similarity in features among its data points.
- Cluster 3, in salmon, overlaps with Cluster 2 but is more spread along PC2, hinting at variance in features that are distinct from those driving Cluster 2.

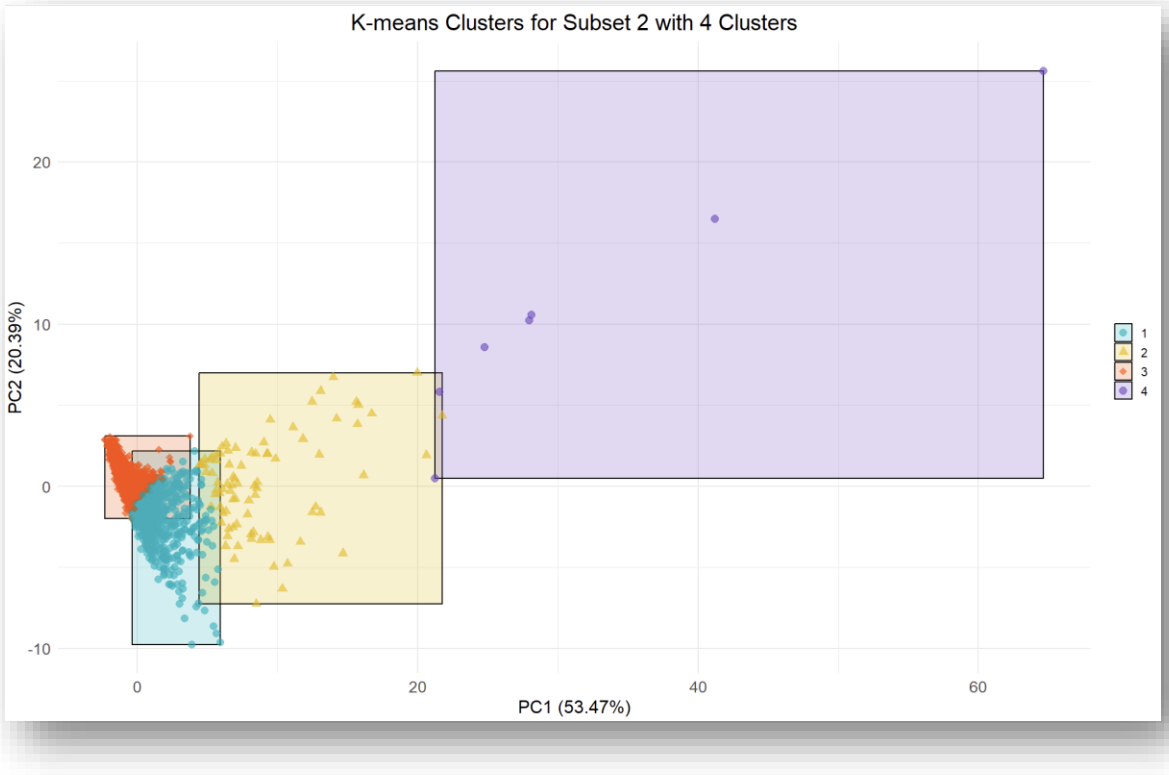

Figure 15: K-means Clusters for subset 2 with 4 clusters.

Summary of the findings from the visualization for Subset 2 with four clusters:

- Cluster 2, represented in yellow with triangle shapes, appears in the lower section of the plot, mostly concentrated around the origin, which may indicate a grouping with average or median values of the dataset's features.
- Cluster 1, in light blue with circle shapes, is tightly grouped near the origin on the PC1 axis but has a lower spread along the PC2 axis, suggesting similar characteristics within this group that are distinct from Cluster 2.
- Cluster 3, marked in salmon with diamond shapes, shows a broad dispersion across the PC1 axis. This suggests a wider range of variation within the attributes that contribute to PC1 for this cluster.
- Cluster 4, in purple with square shapes, is situated further along the PC1 axis, though with few data points, indicating potentially outlier characteristics or a smaller group of data points with high similarity.

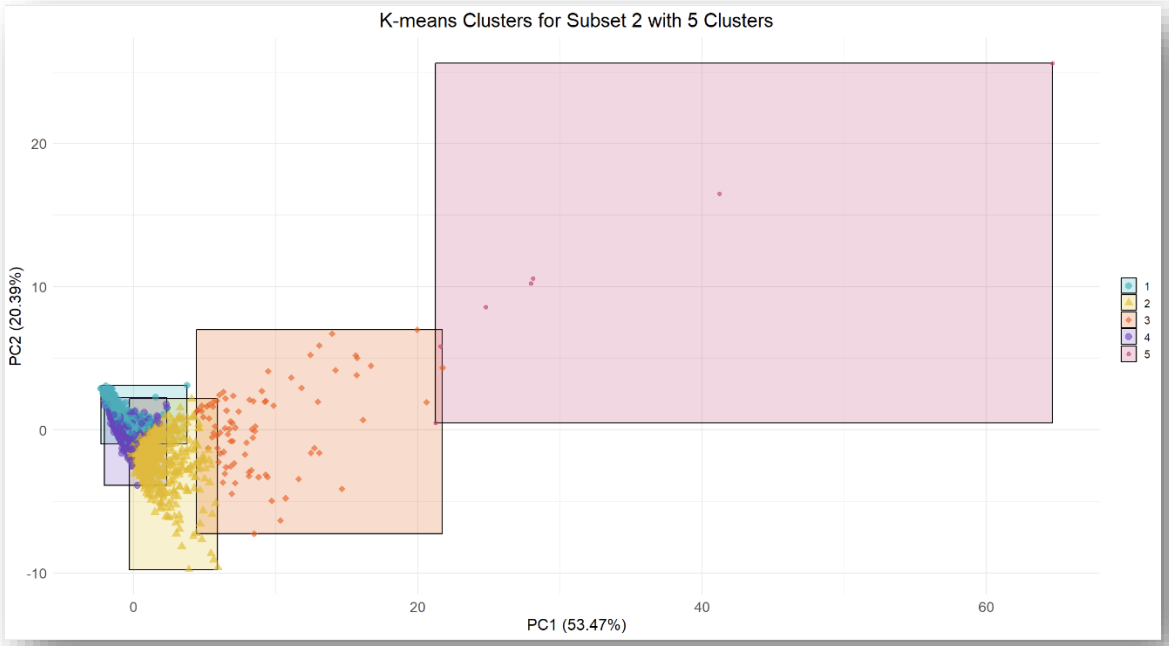

Figure 16: K-means Clusters for subset 2 with 5 clusters.

The visualization of K-means clustering for Subset 2 with five clusters provides a nuanced breakdown of the data points. Here's a synthesis of observations from the five-cluster plot:

- Cluster 1 (Blue): Positioned similarly to Cluster 4 but with fewer data points, this cluster might represent a subset with characteristics slightly distinct from Cluster 1 yet not significantly different in terms of PCA values.
- Cluster 2 (Yellow): Located primarily around the origin in the PCA plot, Cluster 1 represents a group with moderate values on both principal components, suggesting average feature attributes.
- Cluster 3 (Salmon): Shows a larger spread along the PC1 axis, indicating a range in the data that aligns with the variance described by PC1. This suggests a diverse set of attributes within Cluster 3 that have a strong influence on the first principal component.
- Cluster 4 (Purple): This cluster has very few points, mostly located at higher PC1 values, possibly representing outlier data points or a specific subgroup with unique characteristics that strongly influence PC1.
- Cluster 5 (Pink): Like Cluster 4, this cluster includes a minimal number of data points, found at various points on the plot. The data points in this cluster could represent unique or outlier characteristics when compared to the other clusters.

Finalizing total number of clusters:

Upon reviewing the cluster analysis conducted for Subset 2, the decision to proceed with a model comprising three clusters has been made. This model presents a straightforward and easily interpretable structure, delineating distinct data groupings effectively. Opting for three clusters enhances clarity in distinguishing between the groups and prevents over-complication of the model, aligning with the insights gained from both the elbow and silhouette methods. Such a configuration not only streamlines

the model but also supports practical application, yielding actionable insights that are beneficial for strategic decisions tailored to the specific characteristics of each identified data segment.

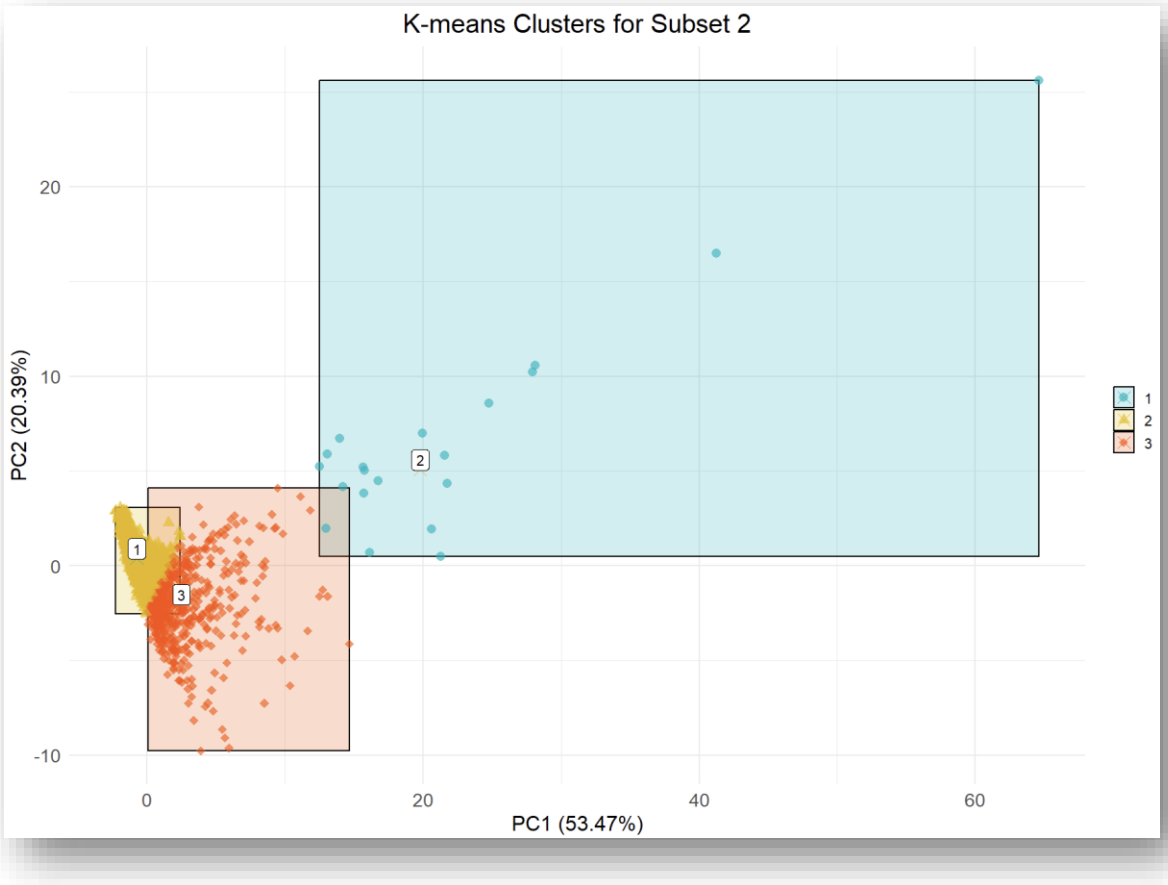

Figure 17: K-means Clusters for Subset 2

The scatter plot titled 'K-means Clusters for Subset 2' presents a two-dimensional simplification of the dataset via the primary principal components, PC1 and PC2. PC1, plotted on the x-axis, captures 53.47% of the variance, suggesting that observations along this axis are likely aligned with economic and housing factors, which often have substantial impact on variability in such datasets. PC2, on the y-axis, contributes 20.03% to the total variance and may indicate variability in demographic or employment features, differing from those denoted by PC1. Combined, these components encompass over 73% of the total variance, thus providing a significant abstraction of the dataset's structure.

K-means clustering segregates the data into three clearly marked clusters, each visually distinguished by distinct colours. They are numerically identified as clusters 1, 2, and 3, with the cluster centroids—emphasized by large squares marked with the corresponding cluster numbers—representing the average position of the data points in the PCA-reduced domain.

Although overlapping with Cluster 1, Cluster 3 spans upwards along the PC2 axis. The positioning of its centroid, apart from Cluster 1, signifies differing characteristics that distinguish this grouping.

Centroids serve as indicators of each cluster's average location within the PCA-constricted space. These centroids' interpretation should be approached with caution due to potential discrepancies in representation between the reduced dimensions and the original data space.

The discernible gap between the centroids of Clusters 3 and 1, despite their intersecting regions, emphasizes the underlying structural diversity within the dataset.

Now, we need to comprehend the feature distribution within our dataset for subset 2, for that we utilize a lollipop chart. This visualization is particularly effective in assessing the magnitude and direction of feature contributions as well as their distinctiveness and uniformity across clusters.

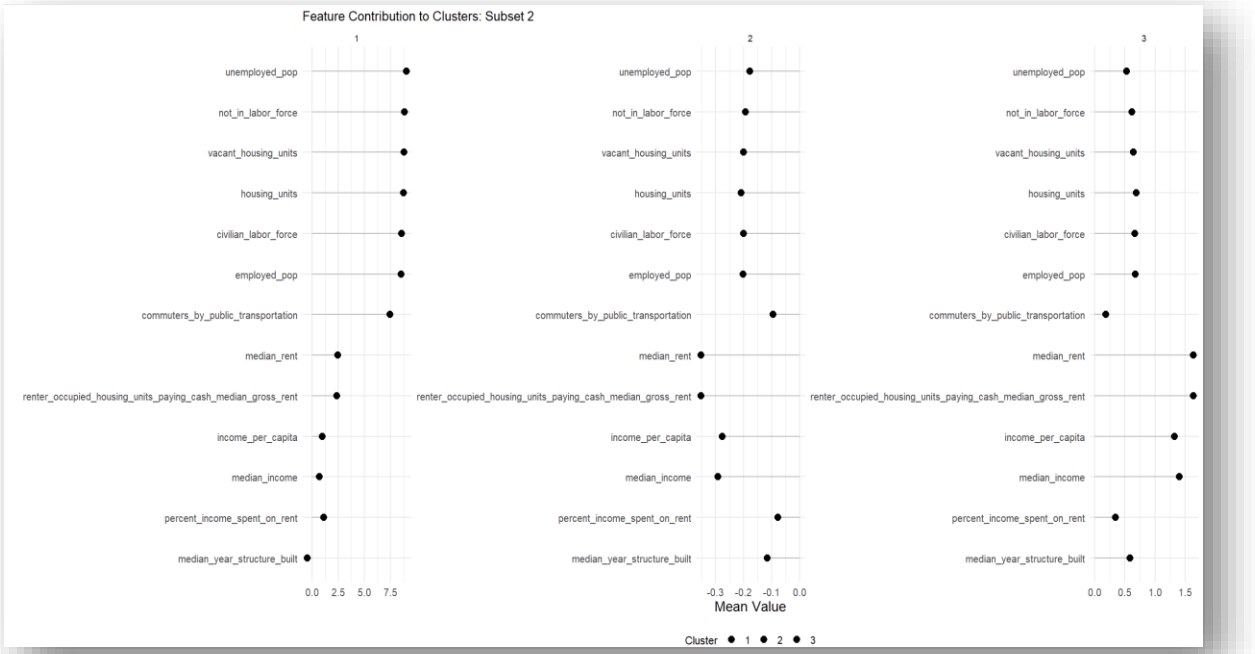

Figure 18: Feature contribution to clusters for subset 2.

The chart provides an insightful depiction of how each feature contributes, on average, to the formation of the three K-means clusters. It's instrumental for unravelling the intricate patterns within the dataset and pinpointing key attributes that are central to each cluster. For instance, a pronounced extension towards the right for 'median_income' in Cluster 1 suggests that this cluster is characterized by relatively higher median incomes when contrasted with Clusters 2 and 3. On the flip side, Cluster 2's 'median_income' feature, which leans towards the left or is abbreviated, indicates a comparatively lower median income in that segment.

Features that maintain a consistent length across clusters, like 'commuters_by_public_transportation', hint at a uniform influence across the groups, while features such as 'housing_units' that show a marked discrepancy in length, underscore their divergent influence on each cluster.

Cluster 3, distinguished by lollipops significantly projecting to the right, reveals a collective profile shaped by the contributing features. Notably, 'not_in_labor_force' and 'unemployed_pop' have a pronounced presence in Cluster 3, implying a cluster demographic with less participation in the labor market. Conversely, features with lollipops stretching to the left in a cluster, such as the 'percent_income_spent_on_rent' in Cluster 2, reveal these features are less representative of that cluster's character, possibly denoting areas where a lower portion of income goes to rent.

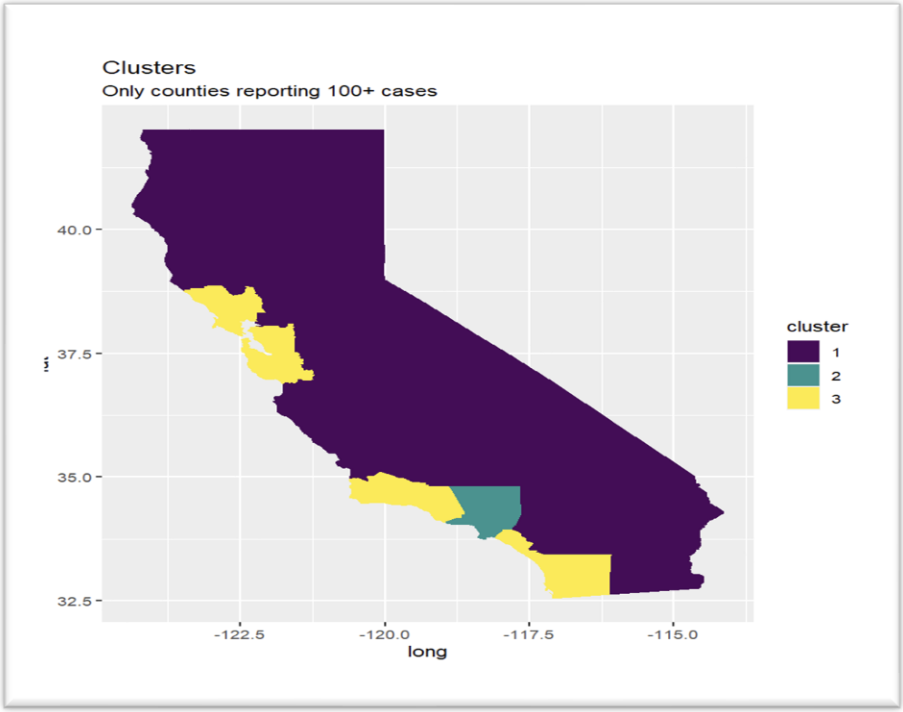

Figure 19: Map representation of the clusters for Subset 2

The 'Map representation of the clusters' of California's counties suggests that based on socio-economic and housing data related to COVID-19 cases, three distinct clusters have emerged. Most counties fall into one extensive cluster, possibly indicating common characteristics across these regions. Two smaller clusters are isolated in the southern part of the state, hinting at unique factors or conditions in these areas.

From this analysis, it is conceivable that regional differences in economic conditions, population density, or public health resources may correlate with the pandemic's impact. This clustering could inform public health strategies, emphasizing tailored approaches for different regions. However, since the map is limited to counties with more than 100 cases, the clusters might not encapsulate the full scope of California's situation and should be considered within this context for policymaking or further research.

Internal Validation:

Now, we used the silhouette method to look at how well our cluster analysis worked for a part of our data called Subset 2. Imagine each piece of data is a point, and we're trying to see if these points are grouped well into clusters. When we use the silhouette method, we get scores that tell us how good our clusters are. If a score is close to +1, that means the points are well placed in their clusters. A score closer to 0 means some points might not belong where we put them, and a negative score would mean we've probably put some points in the wrong clusters. We made two graphs from our silhouette scores.

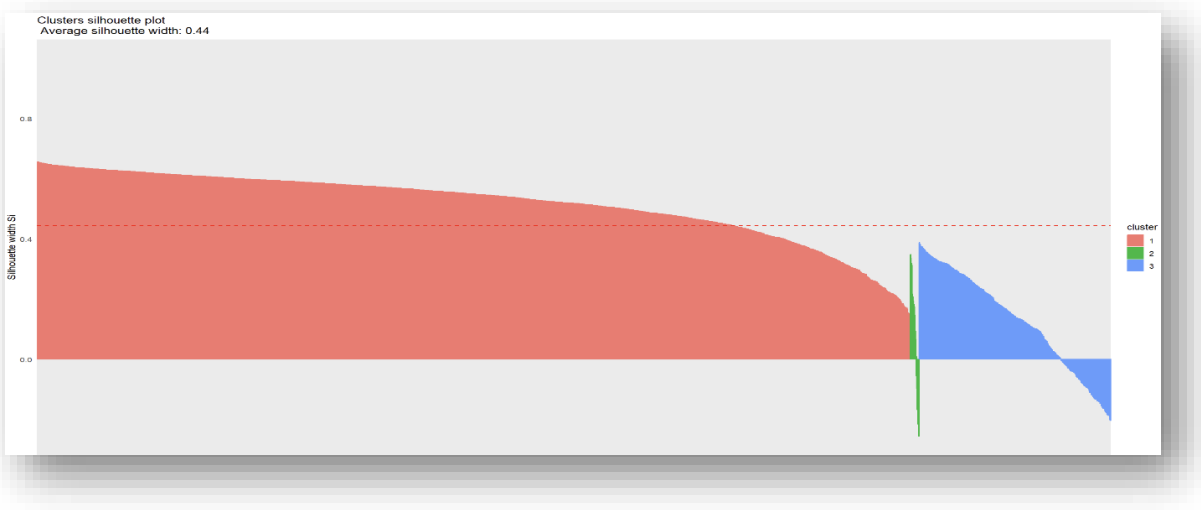

Figure 20: Clusters silhouette plot for Subset 2

The above figure shows us how the scores are spread out for each cluster. We're looking for most of our graph to be closer to the +1 side, which means better clusters.

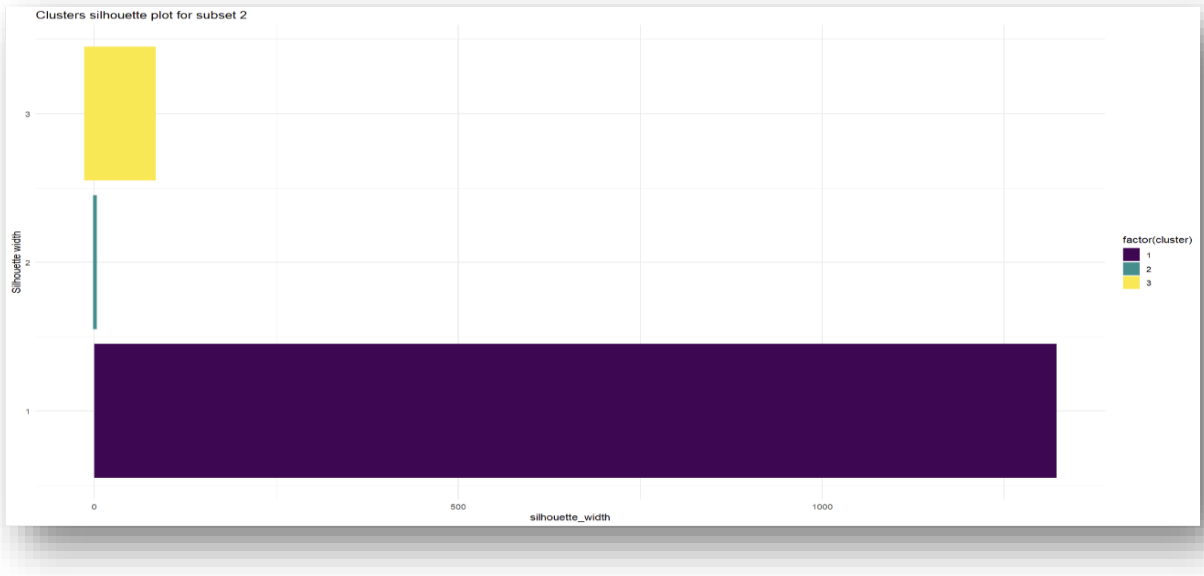

Figure 21: Average score for each cluster for Subset 2

The second graph we made shows the average score for each cluster with bars. Taller bars mean better average scores, which is what we want. This graph gives us a quick way to see which clusters are best.

After looking at our graphs, we think our clusters for Subset 2 are pretty good because most of our silhouette scores are on the positive side. This tells us that our points or data are mostly in the right groups.

External validation for subset 2

In our external validation process for subset 2, we have employed confirmed COVID-19 case counts as a definitive reference point or 'ground truth.' This approach enables us to benchmark the clusters derived from socio-economic and housing data against the actual progression of the pandemic. Through this

comparative analysis, we assess the congruence between the clusters we have identified and the real-world distribution of COVID-19 cases.

The alignment—or lack thereof—between these clusters and case counts can illuminate underlying patterns. Specifically, it allows us to explore the relationship between the infection rates and variables such as socio-economic status, housing quality, employment rates, and reliance on public transportation. Understanding these correlations is pivotal for designing and implementing public health measures that are precisely targeted and more likely to be effective.

Ultimately, using confirmed cases as a ground truth anchors our cluster analysis in the concrete reality of the pandemic's spread. It serves as a crucial tool for decision-makers, aiding in the strategic planning of interventions that address both the health crisis and its socio-economic ramifications. The evaluation metrics obtained are as follows:

Table 6: Validation Metrics for Subset 2

Metric	Value
Adjusted Rand Index (ARI)	0.06743287
Mean Silhouette Width	0.444162
Purity	0.8134946

HIERARCHICAL CLUSTERING

Hierarchical clustering is a method where we start by treating each data point as a separate cluster. Gradually, we merge the closest clusters until we have one large cluster containing all data points. This process is like constructing a family tree for the data, with each merge representing a common ancestor for a set of clusters.

There are two primary approaches to hierarchical clustering. The first is the agglomerative approach, which begins with individual points and merges them upwards to form the tree, akin to a bottom-up approach. The second is the divisive approach, which starts with the entire dataset as one cluster and splits it down into individual points, resembling a top-down approach.

A notable feature of hierarchical clustering is its flexibility in determining the number of clusters. Instead of specifying this number upfront, we can examine the resulting dendrogram, a tree-like diagram, and decide where to 'cut' it to achieve a meaningful grouping of the data.

However, hierarchical clustering has its drawbacks. It can be time-consuming with large datasets, and the results can vary significantly based on the chosen distance measures and merging criteria.

In our cluster analysis, we'll begin by determining the optimal number of clusters using the NbClust method. Based on this, we'll decide on the total number of clusters for each subset. Following this, we'll visualize the dendrograms for these clusters and proceed with internal and external validation methods to assess the quality of our clustering.

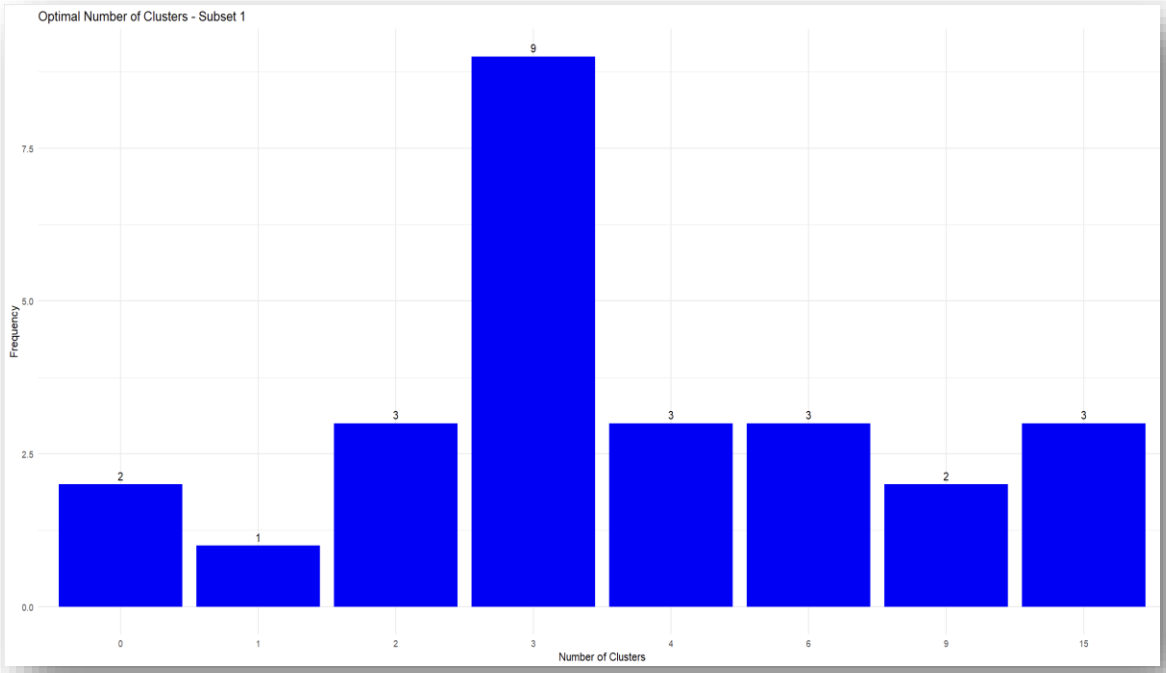

Figure 22: Optimal Number of Clusters for Subset 1

The above bar chart 'Optimal Number of Clusters for Subset 1' presents the results of applying the NbClust method to identify the optimal number of clusters for Subset 1. According to the method's indicators, the most frequent suggestion is that 3 clusters may best represent the inherent groupings within the data. The bars correspond to different numbers of clusters that the NbClust method evaluated, with their heights representing how often each cluster count was recommended by various indices used within the method. Lesser counts, such as 1, 2, 4, 6, 9, 15 clusters, were also suggested but less frequently, indicated by the lower height of those bars. It's evident from the chart that after 3 clusters, the frequency of suggestions drops, and higher numbers of clusters, like 15, are least recommended, suggesting that a more refined clustering might not be as meaningful or could lead to overfitting. This visualization aids in making an informed decision about the number of clusters to use for further analysis with hierarchical clustering techniques.

After evaluating the statistics from our analysis, we've determined that three clusters will provide the most clarity for subset 1 in our hierarchical clustering. This decision ensures a focused yet comprehensive examination of the data.

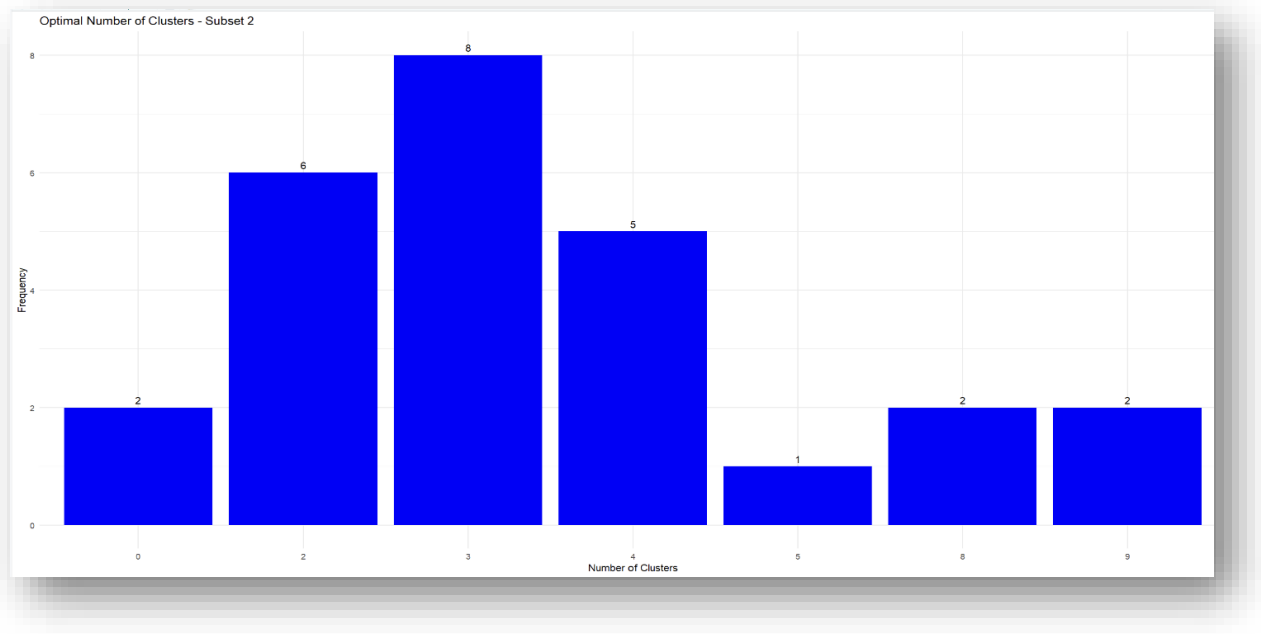

Figure 23: Optimal Number of Clusters for Subset 2

The bar graph titled ‘Optimal Number of Clusters for Subset 2’ represents the suggested optimal number of clusters for Subset 2, determined by the NbClust method across various indices. The method indicates that 2 clusters were having frequency of six times, 3 clusters eight times, and 4 clusters five times, with 8 and 9 clusters each being suggested twice. The tallest bar at 3 clusters suggests that, according to most indices, dividing the dataset into three groups might capture the inherent structure effectively. However, 4 clusters also have a substantial number of recommendations, indicating that this might also be a viable option. Lower recommendations for 2, 8, and 9 clusters suggest these might not delineate the data as effectively according to the indices used.

We decided to opt for 3 clusters in Subset 2 because it aligns with most statistical indicators from the NbClust method, suggesting a natural grouping within the data. This choice balances detail with simplicity, making the clusters distinct and interpretable while avoiding overfitting. It's also a manageable approach for analysis and aligns with practical applications that typically categorize data into low, medium, and high segments. Overall, 3 clusters seem a statistically solid and practically efficient way to understand and utilize the dataset.

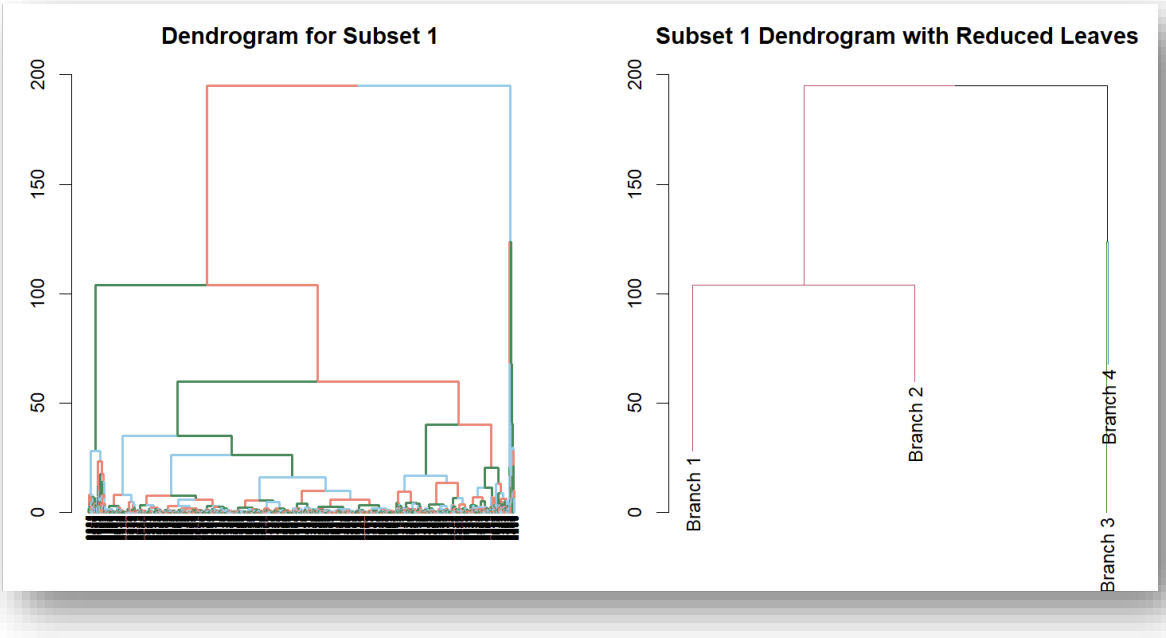

Figure 24: Dendrogram for Subset 1

The dendrogram created from the COVID-19 dataset subset 1, encompassing various demographic features like age, total population, gender distribution, race, household information, and population segments by age, illustrates a structured hierarchy. By applying hierarchical clustering to these attributes, we can discern natural groupings based on demographic similarities. Tight clusters indicate subgroups with shared characteristics, such as similar median ages or household compositions. In contrast, branches that merge at higher points suggest clusters with more significant differences—perhaps contrasting younger and older populations or varying household sizes. This graphical representation aids in understanding the relationships between diverse demographic groups within the data, which can be crucial for targeted public health strategies or resource allocation during the pandemic.

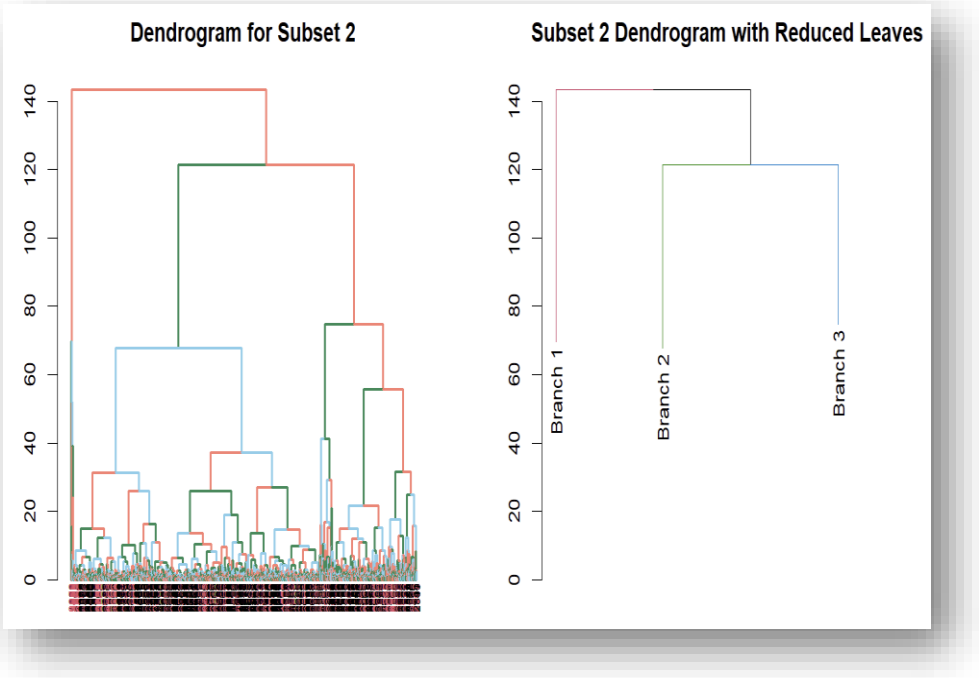

Figure 25: Dendrograms for Subset 2

The above dendrogram derived from Subset 2's analysis—a set that includes a range of socioeconomic and housing features identifies three main clusters. These likely correspond to different local conditions which could be crucial in tailoring the COVID-19 response. Such detailed clustering is an asset for government bodies, enabling them to customize policies and direct resources efficiently to meet the distinctive needs highlighted by each cluster. This can range from financial assistance, housing aid, and health measures to infrastructural enhancements, essentially allowing for a more nuanced approach to pandemic management.

With this dendrogram, agencies gain a bird's-eye view of varying economic and housing conditions, such as income levels and employment rates. A cluster marked by lower incomes and high unemployment rates, for example, might be prioritized for economic stimulus and job programs. Conversely, areas where rent consumes a large income portion might require subsidies or rent control measures. The dendrogram's architecture offers insights into the socioeconomic parallels and distinctions among different groups, paving the way for targeted, impactful community support in the ongoing pandemic scenario.

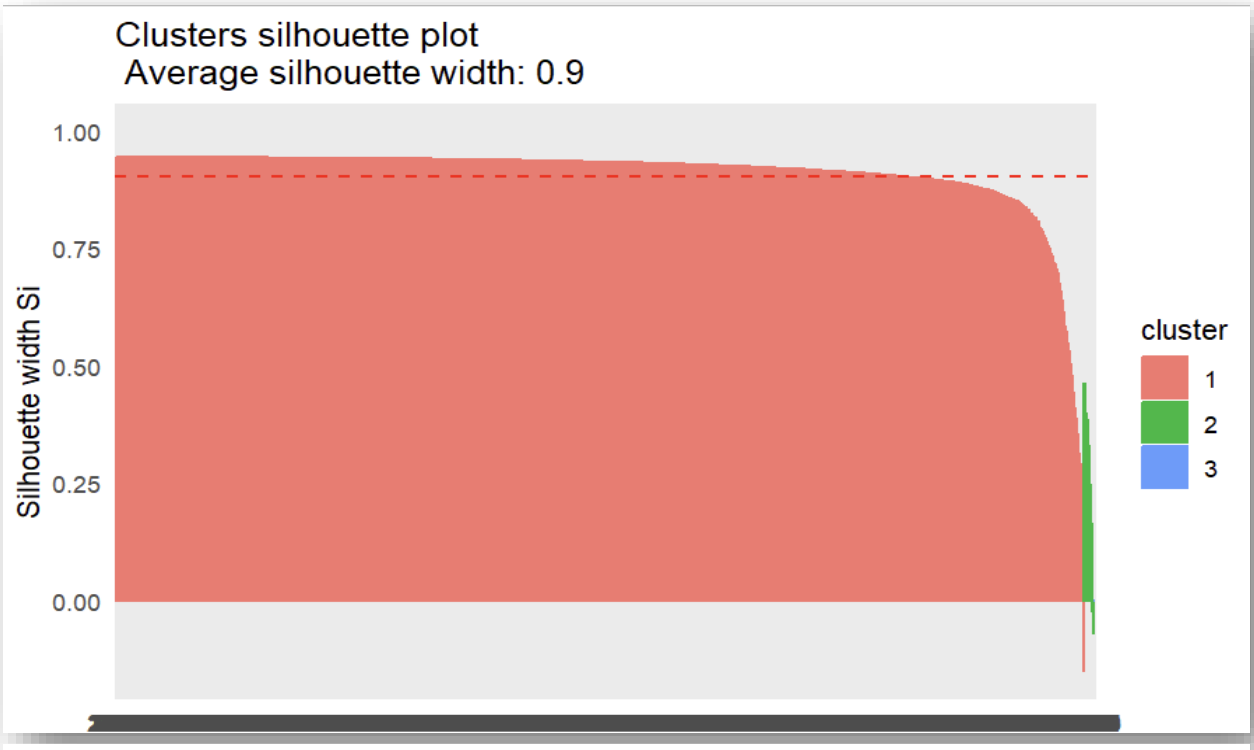

Figure 26: Cluster silhouette plot for subset 1.

In our hierarchical cluster analysis, Cluster 1 stands out with a high average silhouette width of 0.91, indicating a tight-knit and distinct grouping of most of our dataset. Cluster 2, while significantly smaller and with a lower silhouette width of 0.25, points to a less defined and possibly more diverse subset. The single data point in Cluster 3, having no silhouette width, may be an outlier. Essentially, our clustering is highly effective for the largest group, with less certainty for the smaller, and potentially more complex, clusters.

Table 7: Average Silhouette Width of each cluster for subset 1

Cluster	Size	Avg Silhouette Width
1	3108	0.91
2	33	0.25
3	1	0.00

When considering the practical applications of our findings we can conclude that:

- The dominant Cluster 1 can be focused on for any general policies or interventions, as it represents the largest segment of our data.
- Cluster 2 may require more nuanced approaches or targeted strategies since it seems to represent a distinct subgroup within the population.
- Cluster 3 being an outlier could be subjected to further individual analysis to understand its unique characteristics.

Internal Validation Hierarchical Clustering for Subset 2

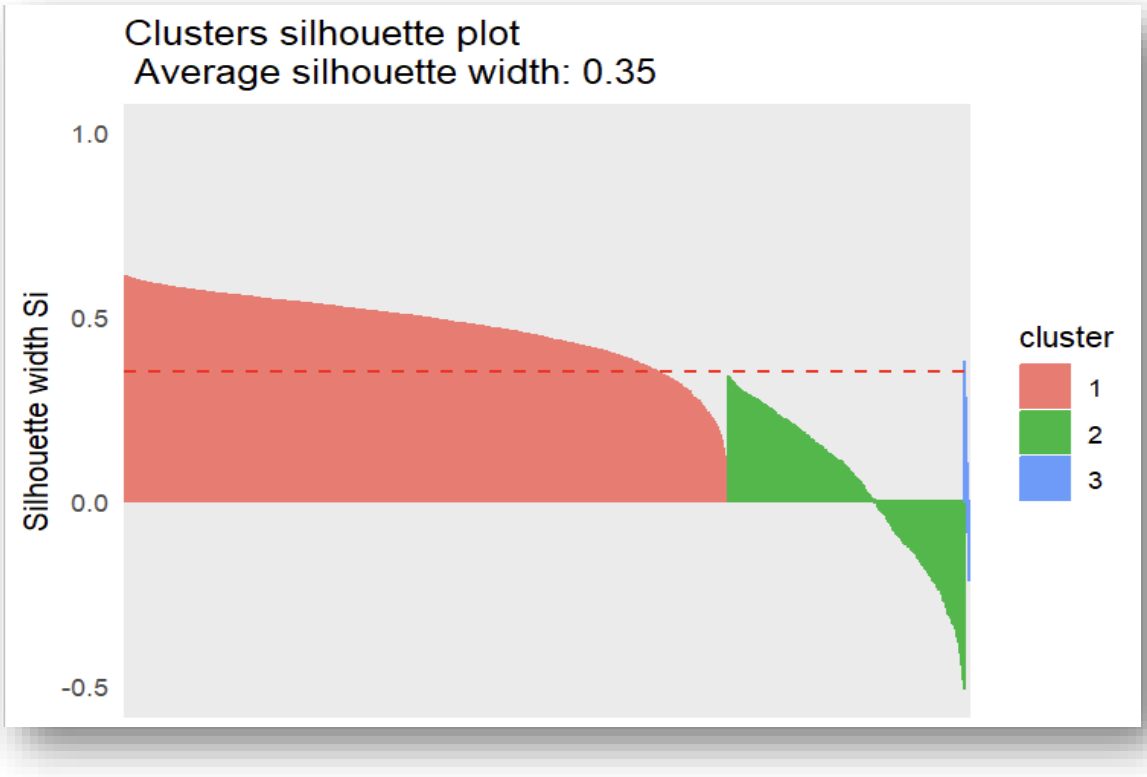

Figure 27: Figure 26: Cluster silhouette plot for subset 2

Our internal analysis for Subset 2's hierarchical clustering reveals three clusters of varying clarity:

- Cluster 1 stands out with a silhouette width of 0.47, hinting at a well-separated and similar group within the data. This is the primary cluster, with most data points, and it represents a distinct grouping.
- Cluster 2 shows a minimal silhouette width of 0.05, indicating that this cluster's data points are not clearly distinguishable from others, suggesting they may have overlapping characteristics with other clusters. This cluster is also sizeable, but its boundaries are less defined.
- Cluster 3, though small with only 17 data points, has a silhouette width of 0.15. This suggests it is a unique, albeit not strongly distinct, cluster that could represent a niche or outlier characteristic in the dataset.

Table 8: Average Silhouette Width of each cluster subset 2

Cluster	Size	Avg Silhouette Width
1	2247	0.47
2	878	0.05
3	17	0.15

For government agencies or stakeholders using this data for COVID-19 responses, these insights can help tailor strategies to the identified groups' needs. For example, Cluster 1 could represent areas with similar socioeconomic conditions that might require similar public health measures or economic assistance. Cluster 2 might need more nuanced strategies, considering the less distinct nature of its data

points. Finally, the unique nature of Cluster 3, albeit small, should not be ignored as it might include special cases that require individual attention.

External validation for HC subset 1:

When we look at the 'confirmed_cases' numbers, it's like getting a snapshot of how deeply COVID-19 is affecting our communities. It makes sense to use this data as a benchmark for our hierarchical clustering because it's a straightforward count of the impact. It's as if we're holding up a mirror to our sociodemographic clusters and asking, "How well do these reflect the real situation out there?" The answer to that question is super important for our partners and decision-makers. They need to know which communities are hit hardest and why, so they can direct health resources and support exactly where it's needed most. By aligning our clusters with the confirmed cases, we are thinking that we will be able to provide a map that can guide a targeted, and hopefully effective, response to this pandemic.

Table 9: External validation of hierarchical clustering for Subset 1

Metric	Value
Adjusted Rand Index (ARI)	0.0004531316
Mean Silhouette Width	0.9019997
Purity	0.3443666

The results from the external validation of hierarchical clustering for Subset 1 of the COVID-19 census dataset indicate that while the clusters formed are internally coherent, they do not correspond effectively to the actual confirmed case numbers. The low Adjusted Rand Index (ARI) shows a negligible correspondence between the clusters and confirmed case groups, the high mean silhouette width reflects that the clusters are well-defined but may not be meaningful with respect to the confirmed cases, and the low purity score suggests that the clusters are not homogenous regarding the distribution of confirmed cases. This means that while the data may be grouped neatly, these groupings do not reflect patterns in the spread or concentration of COVID-19 cases.

External validation for HC subset 2:

By choosing 'confirmed_cases' as our ground truth for subset 2 in our hierarchical clustering analysis, we're really digging into the nitty-gritty of what the numbers can tell us. It's like we're testing the waters to see if our clusters based on economic status and living conditions have anything to do with COVID-19's spread. If these clusters can give us a clear picture of the outbreak's patterns, that's a win for everyone. It means our health experts and policymakers have a solid starting point to tailor their strategies and support to the areas that really need it, focusing on where the economic and housing stresses are making the pandemic hit even harder.

Table 10: External validation of hierarchical clustering for Subset 2

Metric	Value
Adjusted Rand Index (ARI)	0.1152692
Mean Silhouette Width	0.3513174
Purity	0.4894971

The external validation results for hierarchical clustering of Subset 2 indicate moderate correspondence between the formed clusters and the actual confirmed COVID-19 case distributions, with an ARI score of 0.115. However, this score also signifies that the clusters do not precisely reflect the distribution of confirmed cases. The silhouette width of 0.35 suggests that the clusters are not very distinct, with data

points within each cluster not being highly like one another. Furthermore, the purity score of around 0.49 indicates that the clusters are not very pure, meaning they do not consist predominantly of data points with the same confirmed case levels. Therefore, while the clusters offer some insights into the data, they do not provide a definitive grouping based on the confirmed case numbers, suggesting that other unconsidered variables may be impacting the spread or detection of COVID-19 cases within these clusters.

DISCUSSION FOR EFFICIENT CLUSTERING METHOD FOR SUBSET 1

Let's decide which method is efficient for K-means and hierarchical respectively: -

1. Silhouette Method: The silhouette method measures how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette value ranges from -1 to +1, as shown in figure 8 where a high value indicates that the object is well matched to its own cluster and poorly matched to neighbouring clusters. If most objects have a high value, then the clustering configuration is appropriate. In the silhouette plot; the silhouette score seems to drop significantly as the number of clusters increases. The plot indicates that the optimal number of clusters is around 2, as that's where the silhouette score peaks.
2. Elbow Method: The elbow method involves plotting the explained variation as a function of the number of clusters and picking the elbow of the curve as the number of clusters to use. This point is where the rate of decrease sharply changes, indicating that additional clusters do not explain much more variance in the data. In the elbow plot we have provided, there seems to be an "elbow" at around 3 or 4 clusters where the within-cluster sum of squares (WSS) begins to decrease at a slower rate as mentioned in figure 9.
3. For K-Means Clustering: This algorithm is sensitive to the initial placement of the centroids. The elbow method is more popular for determining the number of clusters for K-Means because it directly relates to the sum of squared distances, which K-Means aims to minimize. However, if the silhouette score at this elbow point is reasonably high, it would also suggest that the clusters are well separated and compact. Given the data, an argument could be made for either 2 clusters, as suggested by the silhouette method, or 3-4 clusters, as suggested by the elbow method.
4. For Hierarchical Clustering: This method does not require the number of clusters to be specified in advance. The silhouette method can be particularly useful here since it provides more insight into how well each object lies within its cluster. If the goal is to ensure that each cluster is very distinct from others, the silhouette method might be the better choice. Hence, for hierarchical clustering, 2 clusters could be a better choice if cluster separation and cohesion is prioritized.

Conclusion: The choice between the silhouette and the elbow method may also depend on the specific application and the importance of cluster cohesion versus variance explanation. For K-Means, the elbow method is typically preferred, while for hierarchical clustering, both methods are useful, but the silhouette method might be favoured for its direct measure of cluster quality. In practice, it would be beneficial to consider both methods and perhaps run the clustering with different numbers of clusters to see which results are more meaningful for the application or analysis.

DISCUSSION FOR EFFICIENT CLUSTERING METHOD FOR SUBSET 2

When determining the best number of clusters for Subset 2, we initially considered both the elbow and silhouette methods. After closely analysing the results, we found the silhouette method's visual cues very telling. It hinted at possible cluster counts: 3, 4, and 5. However, upon further inspection and visualization, clusters 4 and 5 presented more overlapping, complicating the clarity of our analysis. In contrast, the delineation of three clusters was significantly more distinct and interpretable, aligning well with the indications from the elbow method.

Consequently, the elbow method, with its pronounced 'elbow,' steered us toward a choice of three clusters. This number showed a stark decrease in within-cluster variation without unnecessarily complicating the model with additional clusters that offered minimal benefit.

Ultimately, this choice harmonizes cluster compactness with simplicity, enabling a structured analysis that is both insightful and actionable. We anticipate that this careful balance will improve the clarity of our clustering analysis, an important factor in making informed strategic decisions where clear direction and applicability are key.

EVALUATION

Table 11: External Validation metrics from All subset's vs the Ground Truth Covid Impact

	Adjusted Rand Index	VI	Entropy	Purity
Kmeans Sub1	0.0192	1.471	0.404	0.40
Kmeans Sub2	0.067	0.393	0.469	0.81
HC Sub1	0.00045	1.1194	0.056	0.34
Hc Sub2	0.11	0.281	0.357	0.48

1. Data Normalization: We started by normalizing the dataset to ensure that all features are on the same scale. This step is crucial because it prevents features with larger scales from dominating the analysis.
2. Handling Missing Values: We checked for missing values in the dataset and imputed them using appropriate techniques. This step ensures that the data is complete and ready for analysis.
3. Subset Creation: We divided the normalized dataset into two subsets based on different criteria. One subset focused on economic factors, while the other subset focused on racial demographics. This step allows for more targeted analysis of specific aspects of the data.
4. Outlier Detection and Removal: We identified and removed outliers from the subsets. Outliers can skew the results of clustering algorithms, so it's essential to address them before proceeding with analysis.
5. Data Scaling: After outlier removal, we scaled the data to further ensure that all features have a similar influence on the clustering process. Scaling helps improve the performance of clustering algorithms by making them less sensitive to the scale of the features.
6. Clustering Algorithms: We applied two clustering algorithms, K-means, and Hierarchical Clustering, to each subset of the data. These algorithms group similar data points together to form clusters based on the features provided.

7. Dendrogram Visualization: For Hierarchical Clustering, we visualized the dendrogram to understand the hierarchical structure of the clusters. This visualization helps in determining the optimal number of clusters for the analysis.
8. Evaluation Metrics: Finally, we evaluated the performance of each clustering algorithm using four evaluation metrics:
    - Adjusted Rand Index (ARI): Measures the similarity between two clustering, adjusted for chance. A higher ARI indicates better agreement between the true labels and the clustering results.
    - Variation of Information (VI): Measures the amount of information lost and gained in the clustering process. Lower VI values indicate better clustering quality.
    - Entropy: Measures the impurity or disorder within clusters. Lower entropy values indicate more homogeneous clusters.
    - Purity: Measures the extent to which clusters contain instances of a single class. Higher purity values indicate better cluster homogeneity.

These evaluation metrics provide insights into the quality of the clustering results and help our **stakeholders** to assess the effectiveness of the clustering algorithms in capturing the underlying structure of the data.

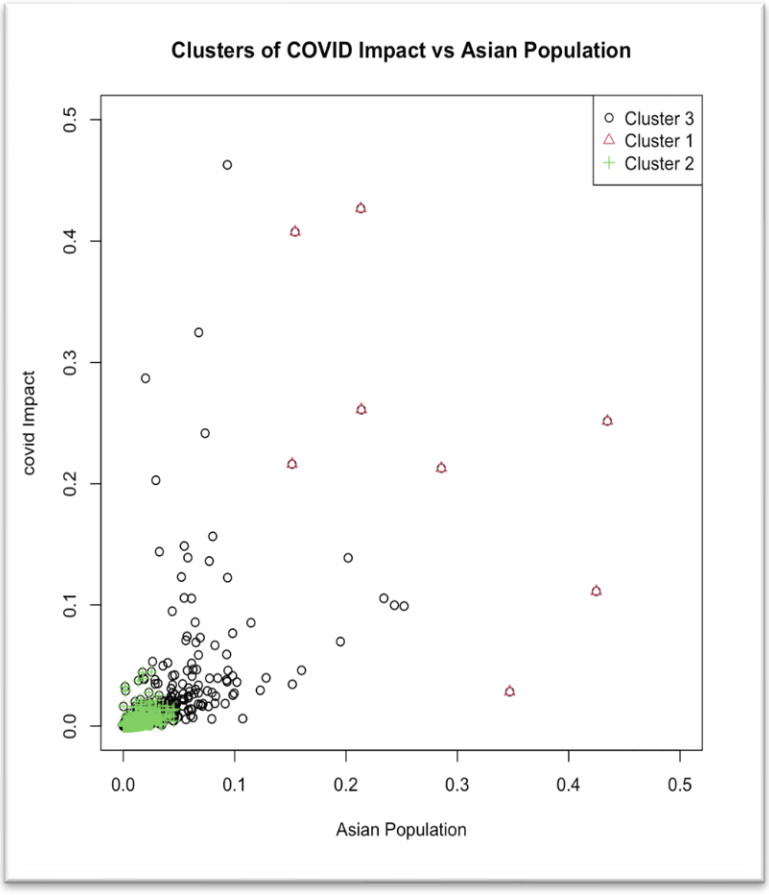

Figure 28: Clusters of COVID Impact vs Asian Population

This scatter plot titled "Clusters of COVID Impact vs Asian Population" appears to show a clustering analysis of data points based on two variables: the impact of COVID-19 (on the y-axis) and the proportion of the Asian population (on the x-axis). Three distinct clusters are indicated using different markers:

1. Cluster 1 (Red Triangles): This cluster contains points that are generally higher on the y-axis, indicating a higher impact of COVID. However, these points are spread across a range of values on the x-axis, showing that the Asian population proportion varies within this cluster. This could suggest that regions with a higher COVID impact have varying proportions of Asian populations.
2. Cluster 2 (Green Pluses): The data points in this cluster are tightly grouped at the lower end of both axes. This indicates that the regions represented here have both a low impact of COVID and a low proportion of the Asian population. The tight clustering suggests a strong correlation between these two variables within this cluster.
3. Cluster 3 (Black Circles): Like Cluster 2, these points are also low on the y-axis but are more spread out on the x-axis. This indicates a generally low COVID impact across regions with varying proportions of the Asian population. There's less tightness in this cluster compared to Cluster 2, which suggests a bit more variation in one or both variables within this group.

Overall Observations:

- Most of the data points, particularly in Clusters 2 and 3, suggest a lower impact of COVID-19 among regions with smaller Asian populations.
- Cluster 1 indicates that a higher COVID impact does not necessarily correlate with a higher proportion of the Asian population.
- The distribution of points suggests that there might not be a simple or direct relationship between the proportion of the Asian population and the impact of COVID-19, especially given the spread of data points in Cluster 1.
- The use of colour and markers to distinguish between clusters allows for easy visualization of how the data points are grouped according to the two variables of interest. However, for more detailed analysis, it would be helpful to have more context on how the "COVID impact" is measured and what the units are for both axes.

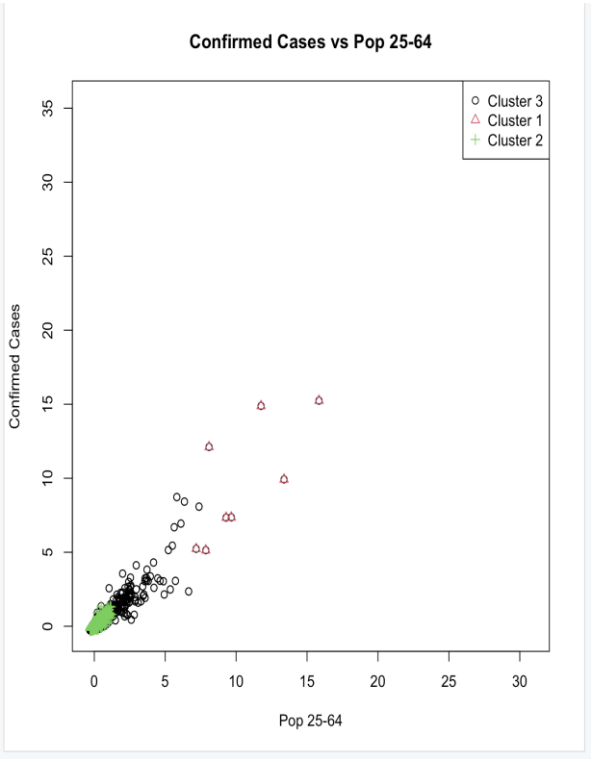

Figure 29: Confirmed Cases vs Pop 25-64

This scatter plot titled "Confirmed Cases vs Pop 25-64" suggests an analysis of the relationship between the number of confirmed cases (presumably of COVID-19, given the common context of such analyses) and the population aged 25-64 in various regions, again indicated by the clustering of data points:

1. Cluster 1 (Red Triangles): This cluster seems to consist of regions with a relatively high number of confirmed cases, spread across a range of population sizes for the 25-64 age group. The data points are more spread out, which could indicate that the confirmed cases do not correlate strongly with the population size in this age range or that there are other factors influencing the number of confirmed cases.
2. Cluster 2 (Green Pluses): These points are tightly grouped near the origin, suggesting regions with both a low number of confirmed cases and a smaller population in the 25-64 age group. This tight grouping could indicate a stronger relationship between the two variables in these regions, potentially suggesting that regions with a smaller population in this age group might have fewer cases.
3. Cluster 3 (Black Circles): Like Cluster 2, these points indicate regions with a low number of confirmed cases but are spread across a range of population sizes. This cluster suggests that the number of confirmed cases is low regardless of the population size in the 25-64 age group in these regions.

Potential issues to consider:

- Clustering Validity: It's unclear what clustering algorithm was used and how well it has identified the natural groupings in the data. The spread of Cluster 1 points might suggest that the clustering could be refined to better capture the underlying patterns.
- Axis Labels: While the axes are labelled, they lack specific context such as whether the population figures are in thousands, millions, or as a percentage of the total population.

EXCEPTIONAL CREDIT

Perform C means Clustering on Subset 1 features and calculate average silhouette widths.

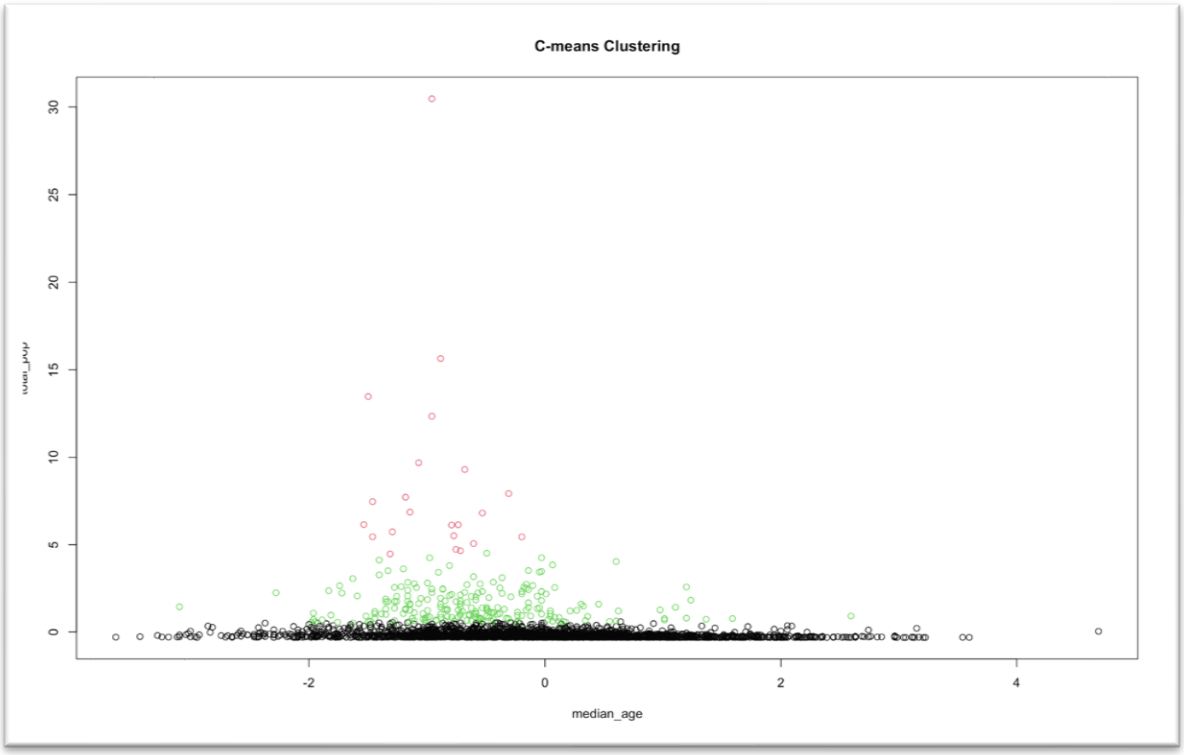

Figure 30: C-means Clustering.

Here Feature 1 and Feature 2 are dimensions.

C-means clustering is a method of clustering that allows one piece of data to belong to two or more clusters. This method is frequently used in situations where data points belong to overlapping groups or categories rather than distinctly separate groups. The algorithm works as follows:

- Initialization: Choose the number of clusters, and then assign each data point a degree of belonging to each cluster. This usually starts with random assignment.
- Cluster Centres: Compute the centroid for each cluster based on the degrees of belonging of all the data points.
- Update Membership: For each data point, compute its degree of belonging to each cluster by considering the distance from the data point to the cluster centre and the fuzziness factor.
- Iterate: Repeat the calculation of the centroids and the updating of the degree of belonging until the coefficients change by an amount below a certain threshold.

In contrast to hard clustering methods like K-means where data points are rigidly assigned to a single cluster, C-means clustering results in a softer, probabilistic clustering where data points can be shared between clusters, with the degree of belonging indicating the strength of the association.

This scatter plot is titled "C-means Clustering" and it shows data points plotted against two variables: median age on the x-axis and total population on the y-axis. It seems that the plot is the result of a clustering analysis using the c-means algorithm, also known as fuzzy c-means clustering.

In c-means clustering, each point has a degree of belonging to clusters, as opposed to k-means where each point belongs to a cluster completely. This might be reflected in the plot by the different colours or shades of the data points; however, this aspect is not clearly visible in the image.

Key observations:

- Many points are clustered near the origin, indicating a large proportion of data with lower median ages and smaller total populations.
- There are some scattered points away from the origin, which may represent areas with higher median ages and/or larger total populations.

For stakeholders, such a visualization can help identify demographic patterns, such as whether certain age demographics are associated with higher population areas. However, the effectiveness of this plot in conveying the nuances of c-means clustering is limited without a clear indication of the fuzzy cluster membership of each point. Additional context or data, such as colour coding or size variation to reflect the degree of membership, would enhance the interpretability of the clustering results.

Based on the silhouette scores for the three clusters, here's a summary of how each cluster is characterized in terms of cohesion and separation:

- Cluster 3 (Silhouette Score: 0.1793): This score is considered good, indicating that the data points within Cluster 3 are quite like each other and well-differentiated from the points in other clusters. The items in this cluster are well matched to their cluster.
- Cluster 2 (Silhouette Score: 0.0569): This is a low score, suggesting that while the data points in Cluster 2 are, on average, more like each other than to data points in other clusters, the degree of similarity is not strong. The cluster has relatively low cohesion and separation from others, and there might be some overlap with other clusters.
- Cluster 1 (Silhouette Score: 0.7718): This is a high score, which is typically considered very good. It implies that the data points within Cluster 1 are highly cohesive and distinct from the points in other clusters. However, given that silhouette scores are typically in the range of -1 to 1, this score should be reviewed to ensure it is not a typographical error. If the score is correct, it would mean that Cluster 1 is exceptionally well defined.

In clustering analysis, a high silhouette score for a cluster indicates that the clustering process has effectively distinguished that cluster from others. Conversely, a low silhouette score may indicate that the clustering is not as definitive, and some data points might not fit well within the assigned cluster. It's important to consider these scores in the context of the specific domain and the clustering goals to determine the most appropriate actions to take, such as potentially adjusting the number of clusters or the algorithm's parameters.

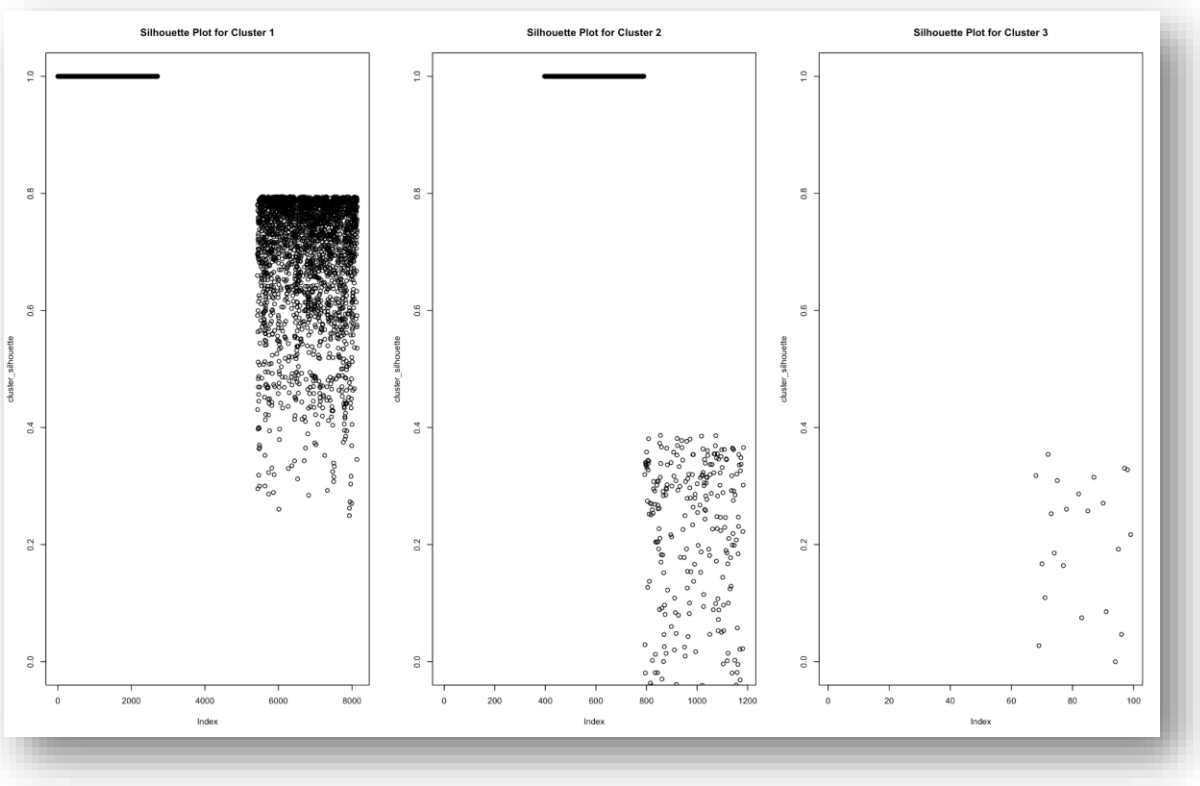

Figure 31: Silhouette plots.

The image shows three separate silhouette plots for what appears to be three different clusters from a clustering algorithm. Each plot corresponds to one cluster and shows the silhouette scores of the individual data points within that cluster.

The silhouette score is a measure of how similar an object is to its own cluster compared to other clusters. The value ranges from -1 to +1, where a high positive score indicates that the object is well matched to its own cluster and poorly matched to neighbouring clusters, and a score close to 0 indicates overlapping clusters.

Key observations from the plots:

- Silhouette Plot for Cluster 1: The plot indicates many data points with a broad range of silhouette scores. Most of the scores are positive and many are clustered around higher values, suggesting good cohesion and separation for this cluster. However, the sheer number of points and the spread of scores hint at some variability within the cluster.
- Silhouette Plot for Cluster 2: This plot shows fewer points than Cluster 1, and the scores are generally positive but lower. This suggests that while data points in Cluster 2 are, on average, closer to their own cluster than to other clusters, the fit is not as good as in Cluster 1. The cluster may be less cohesive or have more overlap with other clusters.
- Silhouette Plot for Cluster 3: This cluster has even fewer points, and the silhouette scores are mostly positive and higher, indicating that these points are very well matched to their own cluster and distinctly separated from other clusters. The good separation is shown by the fact that many points have high silhouette values.

The indexes on the x-axis are likely identifiers for the individual data points. These plots are helpful for evaluating the quality of the clustering. The consistency of the silhouette scores within each cluster can indicate how well the clustering algorithm performed. For example, if most points in a cluster have a high silhouette score, it suggests that the cluster is compact and well-separated from others. Conversely, a wide range of scores or many scores near zero could indicate that the clustering is not as clear-cut.

CONCLUSION

The comprehensive analysis of COVID-19 data, focusing on demographic and economic-housing factors, reveals the pandemic's complex impact on various population segments. The demographic analysis (Subset 1) indicates that age, socioeconomic status, and mobility patterns significantly influence COVID-19's spread and severity, with older populations and those in lower socioeconomic brackets being particularly vulnerable. Similarly, the economic and housing data analysis (Subset 2) underscores the heightened vulnerability of communities facing economic hardships and challenging housing conditions. These insights emphasize the need for integrated public health policies that not only address immediate health concerns but also tackle underlying economic and social vulnerabilities. The findings advocate for targeted interventions aimed at protecting the most at-risk groups and call for expanded research that includes health-related variables for a more comprehensive pandemic response strategy. Ultimately, this analysis highlights the importance of a multi-faceted approach in mitigating the pandemic's effects, combining public health measures with economic support and housing stability initiatives to ensure a resilient recovery for all communities.

DISTRIBUTION OF WORK

Task	Contribution
Document Writing/Formatting	Dhruvil, Juhi, Vishakha
Abstract	Vishakha
Data Preparation	Dhruvil
Modeling	Dhruvil, Juhi, Vishakha
Evaluation	Dhruvil, Juhi
Exceptional Work	Dhruvil, Juhi

All members contributed equally across the project.

REFERENCES

[1] https://console.cloud.google.com/marketplace/browse?filter=solution-%20type:dataset&filter=category:covid19

[2] https://en.wikipedia.org/wiki/COVID-19_pandemic_in_California