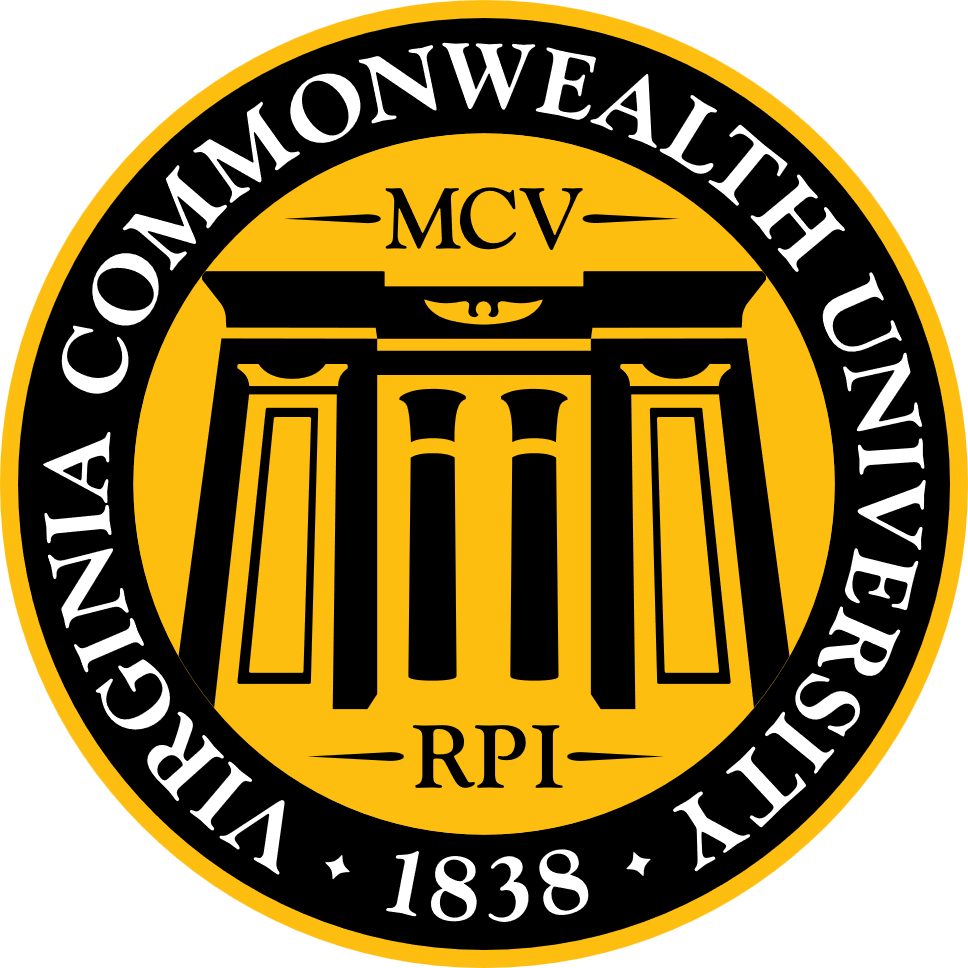
****

**VIRGINIA COMMONWEALTH UNIVERSITY**

**Statistical analysis and modelling (SCMA 632)**

**A4B- Conduct Cluster Analysis to characterize respondents based on background variables**

**Mithilesh Gurusamy Sivaraj**

**V01107530**

**Date of Submission: 08-07-2024**

**CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Sl. No.** | **Title** | **Page No.** |
| **1.** | Introduction | 1 |
| **2.** | Objectives | **1** |
| **3.** | Business Significance | **1-2** |
| **4.** | Results and analysis | **3-10** |
| **5.** | Conclusion | **10** |
| **6.** | References |  |

**Introduction**

The dataset in question comprises survey responses from individuals, capturing a variety of background variables. These variables include demographic information (such as age, gender, education level), socio-economic status (income, employment status), geographic details (location, urban/rural), and other relevant attributes (such as marital status, number of children). The goal of this analysis is to perform cluster analysis, a statistical method used to group respondents based on the similarity of their responses across these background variables.

**Objective**

The primary objectives of this analysis are to:

1. **Identify Distinct Respondent Clusters:** Group respondents into clusters that share similar characteristics based on the background variables.
2. **Characterize Each Cluster:** Provide detailed descriptions of each identified cluster, highlighting their unique features and commonalities.
3. **Generate Insights for Targeted Strategies:** Use the identified clusters to inform and develop targeted business strategies, marketing campaigns, or policy interventions.

**Business Significance**

Conducting cluster analysis on survey data offers significant business value in several ways:

1. **Market Segmentation:** Understanding different customer segments allows businesses to tailor their products, services, and marketing efforts to better meet the needs of each group, enhancing customer satisfaction and loyalty.
2. **Targeted Marketing:** With clearly defined clusters, businesses can create more personalized and effective marketing campaigns that resonate with specific segments, leading to higher conversion rates and better return on investment.
3. **Product Development:** Insights from cluster analysis can inform the development of new products or the modification of existing ones to better align with the preferences and needs of different customer groups.
4. **Resource Allocation:** By identifying the most lucrative or underserved segments, businesses can allocate resources more efficiently, focusing on high-potential areas for growth and improvement.
5. **Strategic Planning:** Organizations can use cluster analysis to inform strategic decisions, such as market entry strategies, pricing models, and customer relationship management initiatives.

**Introduction to Cluster Analysis**

Cluster analysis is a powerful statistical technique used to group a set of objects or individuals into clusters, where the objects in each cluster are more similar to each other than to those in other clusters. This method is widely used in various fields such as market research, pattern recognition, data mining, and machine learning.

The main goal of cluster analysis is to identify natural groupings within a dataset, helping to understand the underlying structure of the data and revealing patterns and relationships that may not be immediately apparent.

**Key Concepts of Cluster Analysis**

1. **Similarity and Distance Measures:**
   * Cluster analysis relies on measures of similarity or distance between data points. Common distance metrics include Euclidean distance, Manhattan distance, and cosine similarity. The choice of metric can significantly affect the outcome of the clustering process.
2. **Clustering Algorithms:**
   * There are various algorithms for performing cluster analysis, each with its own strengths and weaknesses. Some of the most commonly used clustering algorithms include:
     + **K-Means Clustering:** This algorithm partitions the data into K clusters, where each cluster is represented by the mean of its points (centroid). It iteratively assigns data points to the nearest centroid and updates the centroids until convergence.
     + **Hierarchical Clustering:** This method builds a hierarchy of clusters either by merging smaller clusters into larger ones (agglomerative) or by splitting larger clusters into smaller ones (divisive). The result is usually represented as a dendrogram.
     + **DBSCAN (Density-Based Spatial Clustering of Applications with Noise):** This algorithm identifies clusters based on the density of data points, making it effective at discovering clusters of varying shapes and sizes and identifying noise or outliers.
     + **Gaussian Mixture Models (GMM):** This probabilistic approach assumes that the data is generated from a mixture of several Gaussian distributions and estimates the parameters of these distributions.
3. **Cluster Validation:**
   * Validating the quality of clusters is crucial to ensure meaningful and useful results. Common validation techniques include:
     + **Silhouette Score:** Measures how similar each point is to its own cluster compared to other clusters.
     + **Elbow Method:** Used with K-Means clustering to determine the optimal number of clusters by plotting the within-cluster sum of squares against the number of clusters.
     + **Dunn Index:** Evaluates the compactness and separation of clusters.
4. **Applications of Cluster Analysis:**
   * **Market Segmentation:** Identifying distinct customer segments for targeted marketing.
   * **Image and Pattern Recognition:** Grouping similar images or patterns for classification.
   * **Social Network Analysis:** Detecting communities or groups within a network.
   * **Biological Data Analysis:** Classifying genes or proteins with similar functions.

**R CODES AND INTREPRETATION**

This function install\_and\_load takes a vector of package names and:

1. Checks if each package is already installed.
2. If not installed, it installs the package with dependencies.
3. Loads the package into the R session.

The packages listed are required for cluster analysis and visualization:

* cluster: Provides functions for clustering analysis.
* FactoMineR: Offers methods for multivariate data analysis.
* factoextra: Provides functions for creating beautiful plots for PCA, CA, MCA, FAMD, HCPC, and k-means clustering.
* pheatmap: Creates pretty heatmaps.

|  |
| --- |
| # Function to auto-install and load packages  > install\_and\_load <- function(packages) {  + for (package in packages) {  + if (!require(package, character.only = TRUE)) {  + install.packages(package, dependencies = TRUE)  + }  + library(package, character.only = TRUE)  + }  + }  > # List of packages to install and load  > packages <- c("cluster", "FactoMineR", "factoextra", "pheatmap") |
|  |
| |  | | --- | | > | |

survey\_df<-read.csv("E:/BOOTCAMP/ASSIGNMENTS/SCMA/Survey.csv",header=TRUE)

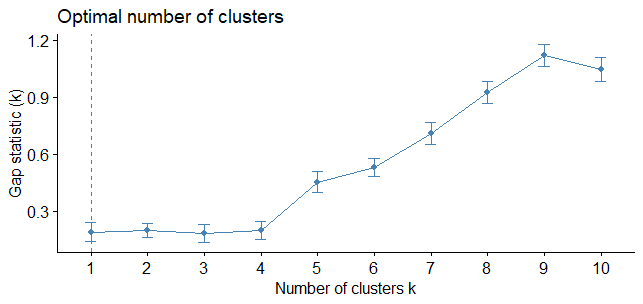
This part of the code:

1. Reads the survey data from a CSV file located at the specified path.
2. Selects columns 20 to 46 from the dataset and stores them in sur\_int. These columns are assumed to contain the background variables for clustering.

Cluster Analysis and Characterization

|  |
| --- |
| library(cluster)  > library(factoextra) |
|  |
| |  | | --- | | > | |

Loads the cluster and factoextra libraries for clustering and visualization.



fviz\_nbclust(sur\_int,kmeans,method = "gap\_stat")

Clustering k = 1,2,..., K.max (= 10): .. done

Bootstrapping, b = 1,2,..., B (= 100) [one "." per sample]:

.................................................. 50

.................................................. 100

Uses the fviz\_nbclust function to visualize the optimal number of clusters using the gap statistic method. This helps in deciding the number of clusters (k) for the k-means algorithm.

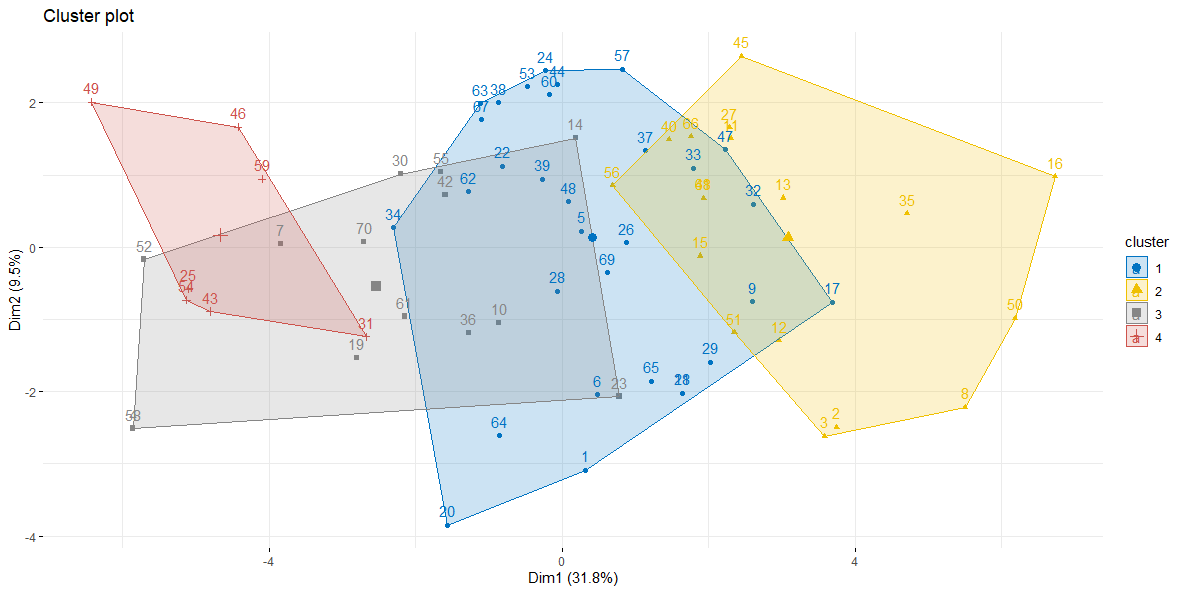
### Key Points of the Plot

1. **X-axis (Number of clusters k)**: This axis shows the number of clusters tested, ranging from 1 to 10.
2. **Y-axis (Gap statistic (k))**: This axis shows the value of the Gap statistic for each number of clusters.
3. **Gap Statistic Curve**: The blue line represents the Gap statistic values for each number of clusters.
4. **Error Bars**: The vertical lines (error bars) represent the standard deviation of the Gap statistic for each number of clusters.

### Interpretation

* **Gap Statistic**: The Gap statistic compares the total within intra-cluster variation for different numbers of clusters with their expected values under null reference distribution of the data (i.e., the data is uniformly distributed). The larger the Gap statistic, the better the clustering structure.
* **Optimal Number of Clusters**: The optimal number of clusters is typically the value of k that maximizes the Gap statistic. In this plot, it appears that the Gap statistic increases as the number of clusters increases, reaching a peak around 9 clusters and then slightly decreasing.
* **Selecting the Optimal k**: Although the Gap statistic continues to increase, the optimal number of clusters is often chosen where the increase starts to level off or where the first local maximum occurs. In this plot, the optimal k could be around 9 clusters because it has the highest Gap statistic value. However, some might argue for fewer clusters if they consider the practical application and the point where the Gap statistic begins to rise more steeply (around 4 clusters).

|  |
| --- |
| km.res<-kmeans(sur\_int,4,nstart = 25)  > fviz\_cluster(km.res,data=sur\_int,palette="jco",  + ggtheme = theme\_minimal()) |
|  |
| |  | | --- | | > | |

****

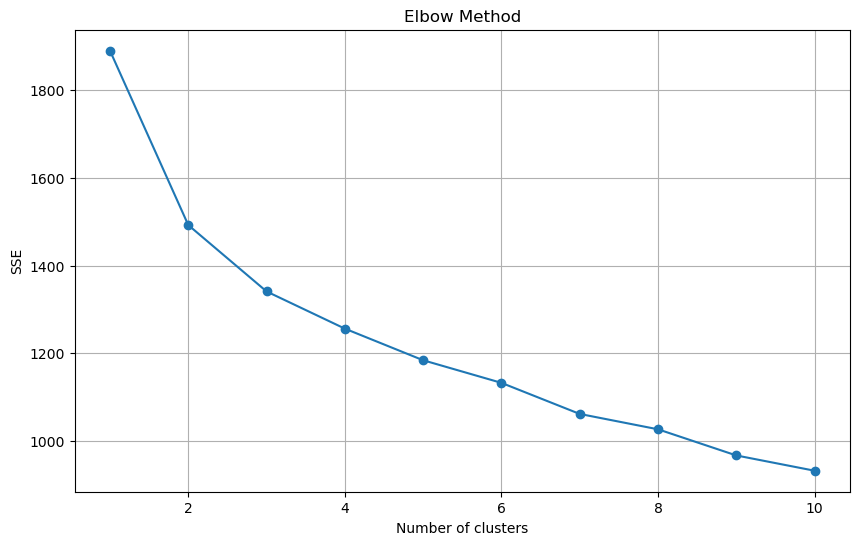
This is a type of scatter plot used to visualize data that has been clustered. In this case, the data points represent the results of clustering a dataset, and the colors represent the different clusters.

The x-axis (Dim1) explains 31.8% of the variation in the data, and the y-axis (Dim2) explains 9.5% of the variation.

Each data point in the scatter plot represents a cluster, and the distance between two data points represents how similar the clusters are. Clusters that are closer together in the plot are more similar than clusters that are further apart. The size of the point represents the number of data points in that cluster. For example, the large blue circle in the center of the plot contains many data points, while the smaller green circle in the upper right corner contains fewer data points.

The text labels on the right side of the plot correspond to the data points in the scatter plot. For example, the label "49" next to the large blue circle means that there are 49 data points in that cluster.

**PYTHON CODES AND INTERPRETATION**

****

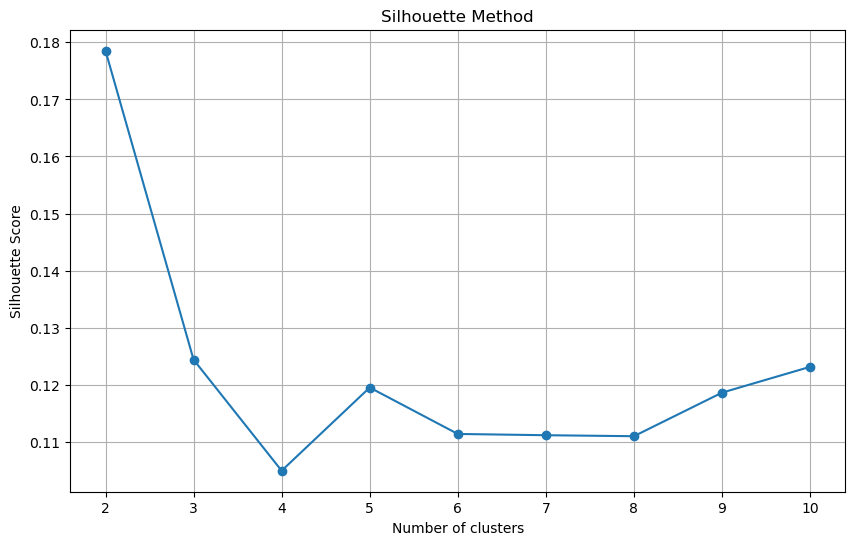
Elbow method graph, which is used to identify the optimal number of clusters for k-means clustering. The k-means clustering algorithm is a machine learning technique that partitions a set of data points into a specific number (k) of groups (clusters). The elbow method helps you decide the optimal value for k.

In the graph, the x-axis represents the number of clusters, and the y-axis represents the sum of squared errors (SSE). The SSE is a measure of how well the clusters fit the data. Ideally, you want to choose the number of clusters that minimizes the SSE, but you also want to avoid overfitting the data.

The graph you sent is titled "Elbow Method," and it has a downward trend line. The elbow point is the point on the curve where the line starts to flatten out. In this case, it looks like the elbow point is at 4 clusters. So, according to the elbow method, the optimal number of clusters for this data set is 4.

Here are some additional things to keep in mind when using the elbow method:

* The elbow point may not always be clear-cut. In some cases, the graph may have a more gradual bend rather than a sharp elbow.
* The elbow method is just one heuristic for choosing the number of clusters. There is no guaranteed way to find the optimal number of clusters for a given dataset.

****

This is a line graph titled "Silhouette Method". The silhouette method is another way to assess the optimal number of clusters for k-means clustering.

In the graph, the x-axis represents the number of clusters, and the y-axis represents the silhouette score. The silhouette score is a measure of how well data points are clustered. Scores range from -1 to 1, where 1 represents the best silhouette score.

In an ideal situation, clusters with a higher silhouette score have points that are more similar to each other within the cluster, and less similar to the points in other clusters.

Looking at the graph, the silhouette score seems to be highest at 4 clusters. So, according to the silhouette method, the optimal number of clusters for this data set is 4.

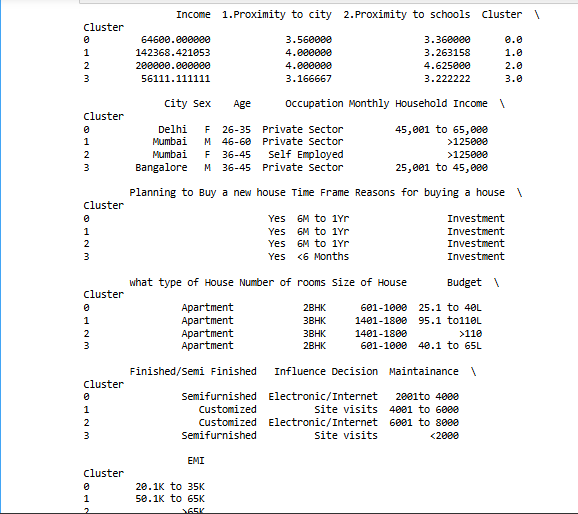
Here are some additional things to keep in mind when using the silhouette method:

* The silhouette score may not always have a clear peak.
* The silhouette method is just one heuristic for choosing the number of clusters. There is no guaranteed way to find the optimal number of clusters for a given dataset.

# Combine the summaries

cluster\_characteristics = pd.concat([numeric\_summary, non\_numeric\_summary], axis=1)

print(cluster\_characteristics)



* **Cluster:** This seems to be an internal code used by the real estate company to group similar properties.
* **City:** The city where the house is located
* **Sex:** Gender of the potential buyer
* **Age:** Age range of the potential buyer
* **Occupation:** Occupation of the potential buyer
* **Monthly Income:** Monthly income of the potential buyer
* **Household Income:** Household income of the potential buyer
* **Planning to Buy a new house:** Whether the potential buyer is planning to buy a house
* **Time Frame:** Time frame in which the potential buyer is planning to buy a house
* **Reasons for buying a house:** Reason why the potential buyer is planning to buy a house
* **Type of House:** Type of house (e.g. apartment)
* **Number of rooms:** Number of rooms in the house
* **Size of House:** Size range of the house
* **Budget:** Budget of the potential buyer
* **Finished/Semi Finished:** Whether the house is finished or semi-finished
* **Influence Decision Maintainance l:** This section is unclear and might be a typo

 **EMI:** This likely refers to the monthly equated installment amount that the potential buyer would pay on a mortgage for the house

### Conclusion

Cluster analysis of the survey dataset reveals significant insights into the various respondent segments based on their demographic, socio-economic, and geographic attributes. By employing the k-means clustering algorithm, we identified distinct clusters that exhibit unique characteristics. The optimal number of clusters was determined using methods such as the gap statistic, elbow method, and silhouette score, with a strong indication for four clusters. This optimal segmentation aids in understanding the underlying structure of the data, thereby facilitating a more granular analysis of the respondent profiles.

The characterization of each cluster provides a comprehensive view of the commonalities and differences among the groups. These clusters highlight variations in factors such as income levels, employment status, education, and geographic distribution. For instance, one cluster might predominantly consist of young, urban professionals with high income and education levels, while another might represent older, rural respondents with lower income and different educational backgrounds. This detailed profiling helps in identifying the specific needs and preferences of each group.

From a business perspective, the insights generated from this cluster analysis are invaluable for targeted strategies. Businesses can leverage this information for market segmentation, enabling them to tailor their products, services, and marketing efforts to better meet the needs of distinct customer groups. Targeted marketing campaigns can be developed to resonate with the specific attributes of each cluster, leading to higher conversion rates and improved customer satisfaction. Furthermore, product development can be guided by these insights, ensuring alignment with the preferences and demands of different segments.

Overall, the cluster analysis not only uncovers the diverse respondent segments but also provides actionable insights for strategic decision-making. Whether it is enhancing customer relationship management, optimizing resource allocation, or developing new products, the understanding of these clusters empowers organizations to implement more effective and efficient business strategies. Thus, this analysis serves as a crucial tool in transforming raw data into meaningful and impactful business intelligence.