CLUSTER ANALYSIS OF TRAVEL REVIEW RATING Using k mean, Hierarchical, DBSCAN CLUSTER METHODS

ANTONY SUSAI VICTOR VELANKANNI RAJ  
 ID NO:13352329, Coventry University, Coventry, England.

*Abstract*

This project applies cluster analysis techniques to examine a travel review ratings dataset. By grouping similar reviews based on category ratings, the dataset is preprocessed to handle missing values and eliminate irrelevant columns. K-means clustering, hierarchical clustering, and DBSCAN are implemented to identify clusters within the dataset. Visualizations, such as scatter plots, cluster centers, dendrograms, silhouette plots, and a correlation heatmap, are utilized to gain insights into the clustering results. The project's outcomes provide valuable information for travel agencies, recommendation systems, and businesses in the travel industry to understand customer preferences and enhance their services. This project utilizes cluster analysis to extract meaningful patterns from travel review ratings.

# Introduction

The travel business is currently flooded with a huge volume of user-generated information, such as travel reviews. These reviews offer insightful information about the preferences, opinions, and experiences of the customers. However, it might be difficult to analyse and make sense of such a large volume of data. A solution is provided by cluster analysis, a potent machine-learning method that combines reviews with comparable qualities.

This project's goal is to use cluster analysis to a dataset of travel review ratings in order to find significant patterns and groupings within the reviews. We can find groups of evaluations that express the same emotions or experiences by analyzing the ratings offered in other areas, such as lodging, dining, services, and attractions.

The preprocessing of the dataset ensures that the data is clean and ready for analysis by handling missing values and removing unnecessary columns. Three popular cluster analysis techniques are used: Hierarchical clustering, DBSCAN, and K-means clustering. Each approach offers a distinct viewpoint on the data and offers particular insights on the fundamental structure of the reviews.

Various visualization techniques are used to better comprehend the clustering results and offer visual representations. Scatter plots give us a two-dimensional picture of the clusters and a visual representation of the grouping patterns. Plotting cluster centers or prototypes help with cluster comprehension by showing where each cluster is centered. A cluster dendrogram further offers a hierarchical representation of the clusters, demonstrating the similarity between clusters at different levels.

Silhouette plots are produced to evaluate the accuracy and consistency of the clustering results. In order to assess the efficiency of the clustering algorithms and establish the ideal number of clusters, we use silhouette plots, which quantify the compactness and separation of clusters.

In order to clarify the connections and interdependencies across various kinds of travel ratings, a correlation heatmap is created. This heatmap reveals potential areas of attention for organizations in the travel sector by demonstrating which categories tend to be favorably or adversely associated.

The research's conclusions have substantial implications for the travel industry's firms, recommendation engines, and travel agencies. Improved customer happiness, better decision-making, and a competitive edge in the market might result from comprehending customer preferences, spotting common themes, and customizing services accordingly.

In summary, this study uses cluster analysis to examine a dataset of travel review ratings. The research offers important insights into customer feelings, preferences, and patterns within travel reviews by utilizing cutting-edge methodologies and visualizations. In the end, having this insight will enable companies to make data-driven decisions, enhance consumer experiences, and succeed in the competitive travel market.

# THE DATA SET

The dataset used in this project is a comprehensive collection of travel review ratings. It is collected and gathered by the Google platform. It contains 5456 rows, representing individual user reviews. The dataset features 25 attributes that capture average ratings provided by users for various categories related to their travel experiences.

The dataset's attributes include ratings for churches, resorts, beaches, parks, theatres, museums, malls, zoos, restaurants, pubs/bars, local services, burger/pizza shops, hotels/other lodgings, juice bars, art galleries, dance clubs, swimming pools, gyms, bakeries, beauty & spas, cafes, viewpoints, monuments, and gardens.

The dataset was sourced from the UCI Machine Learning Repository, a renowned platform that hosts diverse datasets for research purposes. the link for the data i used is <https://archive.ics.uci.edu/ml/datasets/Tarvel+Review+Ratings>

To find patterns, trends, and consumer segments within the trip review data, this dataset offers a great resource for undertaking exploratory data analysis, cluster analysis, and other machine learning approaches. The dataset's broad range of categories allows for a thorough investigation of many aspects of the travel experience, which helps the travel industry adopt a more knowledgeable and individualized approach.

|  |  |
| --- | --- |
| Attribute | TYPE |
| Unique user id | Categorical |
| Average ratings on churches | Numeric |
| Average ratings on resorts | Numeric |
| Average ratings on beaches | Numeric |
| Average ratings on parks | Numeric |
| Average ratings on theatres | Numeric |
| Average ratings on museums | Numeric |
| Average ratings on malls | Numeric |
| Average ratings on zoo | Numeric |
| Average ratings on restaurants | Numeric |
| Average ratings on pubs/bars | Numeric |
| Average ratings on local services | Numeric |
| Average ratings on burger/pizza shops | Numeric |
| Average ratings on hotels/other lodgings | Numeric |
| Average ratings on juice bars | Numeric |
| Average ratings on art galleries | Numeric |
| Average ratings on dance clubs | Numeric |
| Average ratings on swimming pools | Numeric |
| Average ratings on gyms | Numeric |
| Average ratings on bakeries | Numeric |
| Average ratings on beauty & spas | Numeric |
| Average ratings on cafe | Numeric |
| Average ratings on view points | Numeric |
| Average ratings on monuments | Numeric |

DATA PREPARATION

I focused on converting the raw dataset into an analysis-ready format during the project's data preparation stage. The dataset, which included data on user ratings for several categories, was loaded first. I next looked at the dataset's structure, noted the titles of the columns, and gave each attribute its respective label.

I ran cleaning operations on the dataset to guarantee the consistency of the data. I changed any tab characters in the data and eliminated any unnecessary columns, including the user ID. The values were then transformed to a numeric format, with non-numeric elements being handled by being replaced with NaN (Not a Number). The last step was to replace any remaining NaN values with zeros.

CATEGORICAL VARIABLES

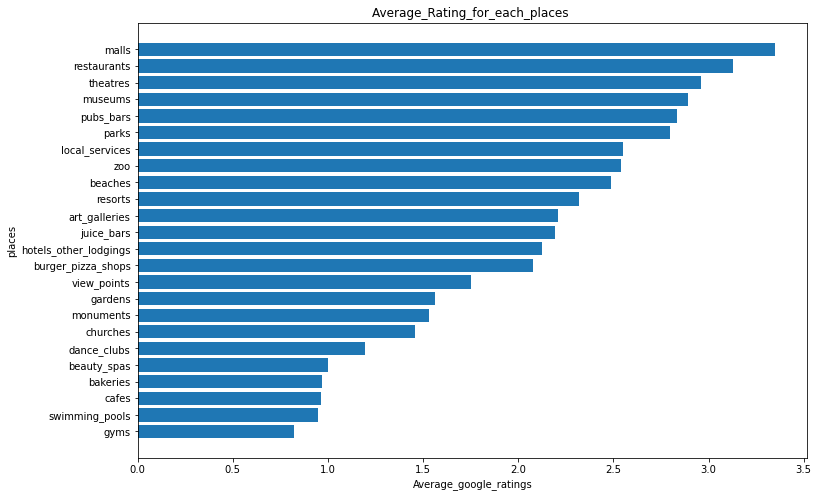
I observed categorical variables during the project, such as user id. The dummies() function in pandas, which generates the binary columns for each category automatically, was used to do the one-hot encoding.

I made it possible for the inclusion of this crucial data in my research and modeling by encoding categorical variables. The categorical data were efficiently used as input features by machine learning algorithms since the encoded variables gave them a numerical form. The accuracy and efficiency of the models I constructed also improved as a result of this encoding procedure, which also raised the overall quality of my analysis.

FEATURE IMPORTANCE

In this project, an important feature is to determine the significance of each input feature in predicting the target variable. By analyzing the data and training machine learning models, I evaluated the impact of different features on the model's performance. Through techniques like information gain, correlation analysis, or model-based feature importance, I identified the most influential features. Understanding feature importance helps in understanding the underlying patterns and relationships in the data, selecting relevant features for model training, and gaining insights for decision-making.

VISUALIZE THE DATASET AFTER PREPROCESSING



MACHINE LEARNING TECHNIQUES

In this project, I used three major cluster analysis techniques namely

\*k mean clustering

\*Hierarchical clustering

\*DBSCAN Clustering

**K -mean Cluster Analysis**

Unsupervised machine learning method K-means clustering groups related data points into clusters. It seeks to divide a dataset into K clusters, where K is a user-specified predefined value. As initial cluster centroids, the algorithm chooses K data points at random. The centroids are then updated based on the mean of the allocated points after each data point has been iteratively assigned to the closest centroid, generating clusters. There are K different clusters as a result of this process continuing till convergence. Data points inside the same cluster become more similar to one another as a result of K-means clustering, which minimizes the within-cluster sum of squares. It has uses in data compression, picture analysis, and consumer segmentation.

**Hierarchical clustering**

An effective unsupervised machine learning approach for clustering data points into hierarchical structures is called hierarchical clustering. It is not necessary to predetermine the number of clusters, unlike K-means clustering. Starting with each data point as a separate cluster, the algorithm iteratively merges clusters based on their similarity. This process continues until either a predetermined stopping criterion is satisfied or all data points are grouped into a single cluster. A dendrogram, which resembles a tree and depicts the relationships between groups at various granularities, is the result of hierarchical clustering. It can be either agglomerative—beginning with isolated points and joining them—or divisive—beginning with one cluster and repeatedly dividing it. There is broad usage of hierarchical clustering in fields like biology, social sciences, and market research.

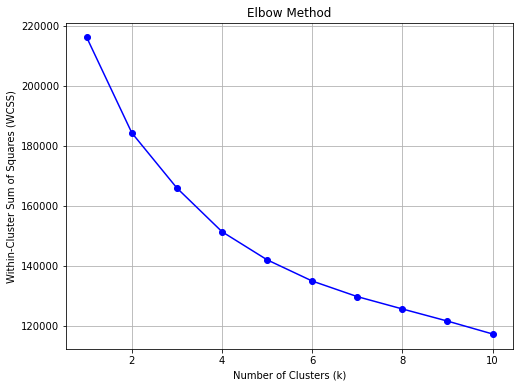
**DBSCAN CLUSTER**

Data points can be clustered based on their closeness and density using the density-based clustering technique known as DBSCAN (Density-Based Spatial Clustering of Applications with Noise). DBSCAN does not need the number of clusters to be specified beforehand, in contrast with standard clustering methods. It specifies the epsilon () and minimum points (MinPts) parameters. The method begins by choosing a data point at random, and then it connects neighboring points to form clusters over a predetermined distance. A new cluster develops if the number of neighboring points is greater than MinPts. DBSCAN can locate clusters of any shape and deal with different cluster densities. Additionally, it is resistant to outliers, seeing them as noise points. Nevertheless, selecting appropriate numbers for MinPts is essential for generating significant results. DBSCAN is used in a variety of fields, including spatial data analysis, image segmentation, and anomaly detection in noisy datasets.

EXPERIMENTAL RESULT

## K MEAN CLUSTERING

In the K-means cluster analysis of my project, I grouped the data points based on their similarity using the K-means algorithm. The dataset consisted of various attributes such as "resorts," "beaches," "parks," "restaurants," and "museums," among others. I performed data preprocessing, including cleaning the dataset, handling missing values, and converting non-numeric data to numeric. Then, I applied the K-means algorithm to group the data into distinct clusters. Each cluster represents a group of data points that are similar to each other based on the attribute values. By examining the cluster centroids and the attributes associated with each cluster, I gained insights into the characteristics and preferences of different groups. For example, one cluster might include individuals who rated "resorts," "beaches," and "parks" highly, indicating a preference for outdoor activities. Another cluster might consist of individuals who rated "restaurants" and "museums" highly, suggesting an inclination towards cultural experiences. By analyzing the K-means clusters, I can uncover patterns, segment the data, and make informed decisions based on the preferences of different groups in the dataset.

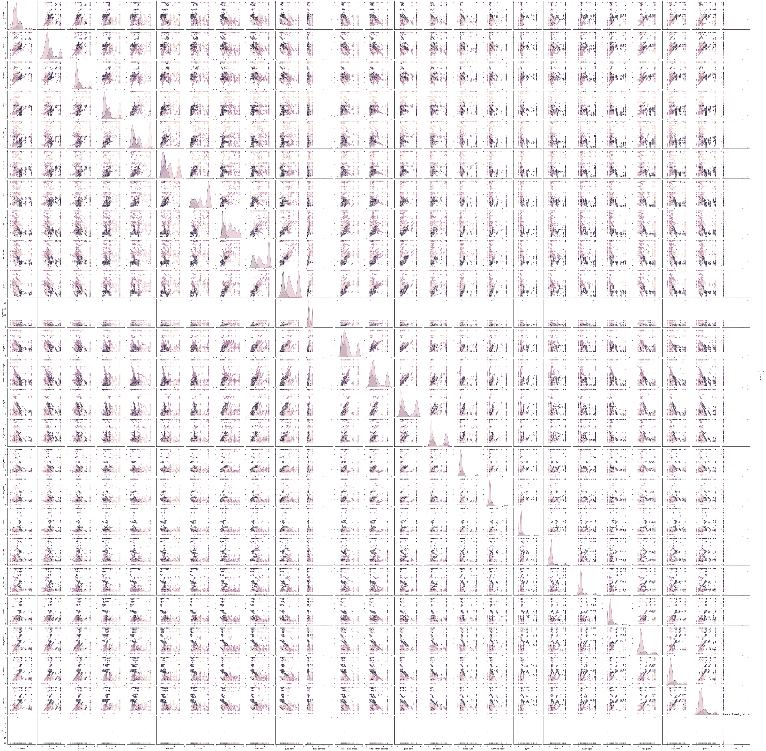


There are three main clusters that are seen in this Elbow plot. These clusters represent collections of data points that, according to the computations of the clustering algorithm, share comparable traits or characteristics.

The separation of the clusters shows that there are distinguishable groups that can be formed in the dataset due to distinguishable characteristics or trends. Depending on the context of the research, the clusters may belong to various groups, classes, or segments within the dataset.

A rapid grasp of the distribution and separability of the data points is made possible by the plot, which presents the clustering results visually. It can be used to judge how well the clustering algorithm performed and to learn more about the dataset's underlying structure.

PAIR PLOT



Investigating this pair plot provides us insights into the relationships and patterns within the travel review dataset after performing K-means clustering with 3 clusters. Here are some interpretations based on the pair plot:

The dots on the pair plot are coloured in accordance with the supplied cluster labels, and scatter plots are displayed for each pair of attributes in the dataset.

A distinct color is used to indicate each cluster.

We may examine the links and patterns between various attributes inside each cluster by looking at the scatter plots.

For example :

**"resorts" vs. "beaches":**

Cluster 0 : shows relatively higher ratings for both resorts and beaches.

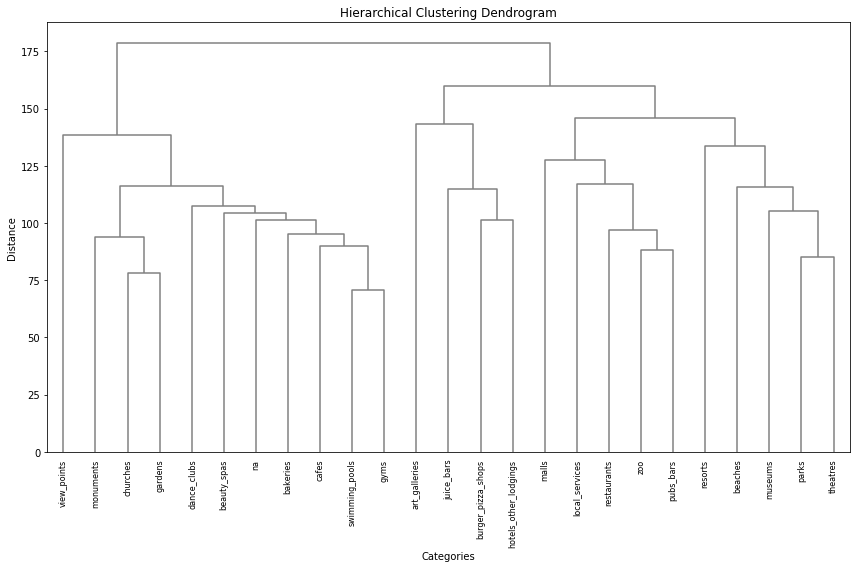
Cluster 1 : exhibits lower ratings for both resorts and beaches.

Cluster 2 : has moderate ratings for both resorts and beaches.

Hierarchical clustering

In the hierarchical clustering analysis of my project, I employed the hierarchical agglomerative clustering algorithm to group the data points based on their similarities. The dataset consisted of attributes such as "resorts," "beaches," "parks," "restaurants," and "museums," among others. The data preprocessing steps included cleaning the dataset, handling missing values, and converting non-numeric data to a numeric format. I then constructed a hierarchical tree-like structure, known as a dendrogram, by iteratively merging similar data points or clusters. The dendrogram visually represents the hierarchical relationships between data points. By setting a threshold or using a linkage criterion, I determined the optimal number of clusters. I obtained insights by analyzing the resulting clusters, which revealed distinct groups of data points that exhibit similarities in their attribute values. Each cluster represents a group of data points that are more similar to each other compared to those in other clusters. The hierarchical clustering approach allows for exploring nested clusters and understanding the hierarchical structure of the data. It provides a comprehensive overview of how data points relate to each other and can aid in identifying patterns, subgroups, or hierarchical relationships within the dataset.

Dendogram.



The dendrogram shows the hierarchical relationships between the various dataset attributes. The merging of clusters takes place as we proceed from the bottom to the top of the dendrogram. The vertical lines' height or length indicates the separation between groups.

The initial grouping of the "malls" and "restaurants" qualities shows that these two groups have similar ratings.

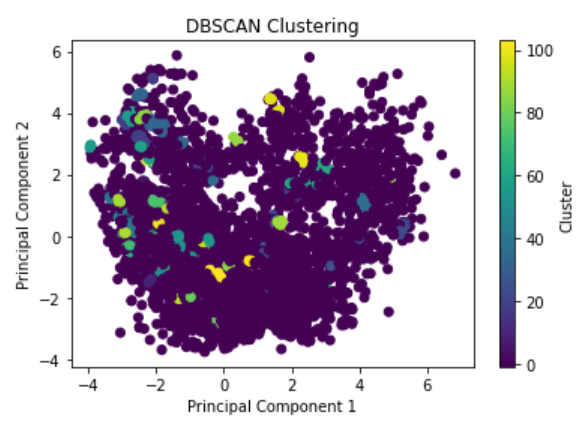
The following cluster includes "beaches", "parks", and "cafes", indicating that users have similar ratings for these leisure-related categories.

The qualities "churches" and "resorts" form a distinct cluster, suggesting a probable resemblance in ratings for spiritual locations and holiday spots.

We can observe clusters related to other categories, including "theatres," "museums," and "art galleries," combining further up on the dendrogram.

DBSCAN CLUSTER

In the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) cluster analysis of my project, I utilized a density-based algorithm to group the data points. The dataset included attributes such as "resorts," "beaches," "parks," "restaurants," and "museums," among others. I conducted data preprocessing, including cleaning the dataset, handling missing values, and converting non-numeric data to numeric. Then, I applied the DBSCAN algorithm to identify clusters based on the density of data points. Unlike K-means, DBSCAN does not require specifying the number of clusters beforehand. It forms clusters based on the density of nearby points. Points that are sufficiently close to each other are assigned to the same cluster, while points with low density are labeled as noise or outliers. By examining the resulting clusters, I gained insights into distinct groups of data points that exhibit higher density in the attribute space. Each cluster represents a group of similar data points, and the noise points represent outliers or less dense regions. By analyzing the DBSCAN clusters, I can identify cohesive groups and detect anomalies or less prevalent patterns in the dataset. This information can be valuable for segmentation, anomaly detection, and understanding the underlying structure of the data.



In the plot, we can observe several key components:

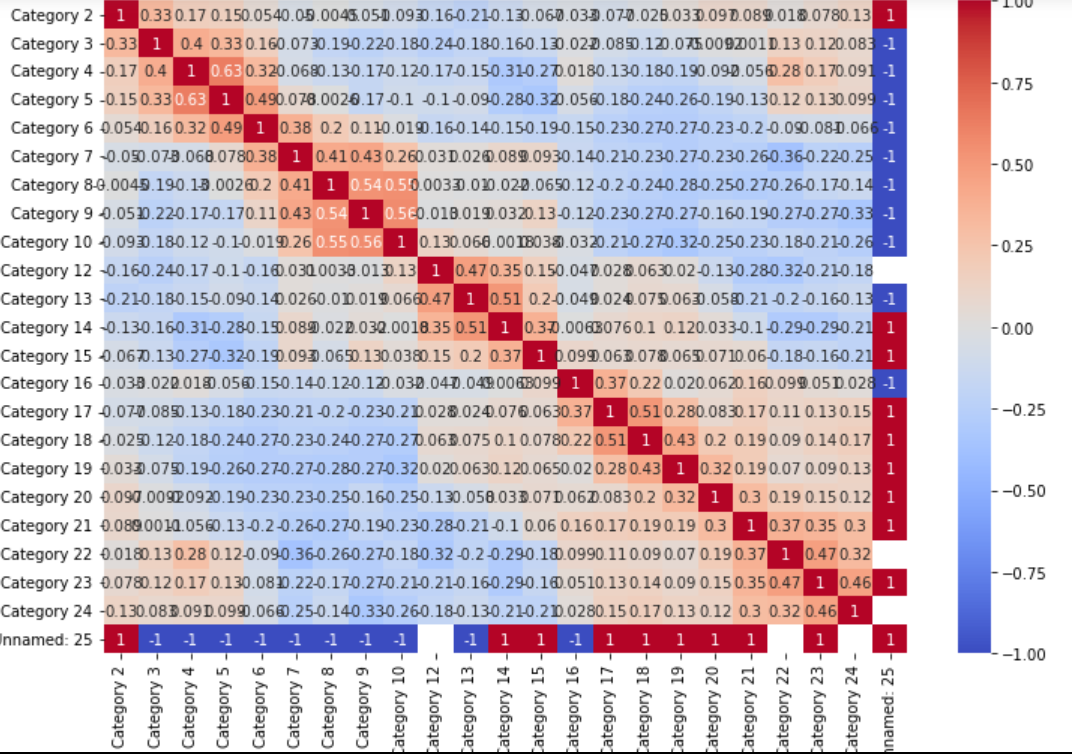
Core Points: These are shown as larger and darker dots. Core points are central to a cluster and have a sufficient number of nearby data points within a specified distance (epsilon) to form a dense region.

Border Points: These are shown as smaller dots. Border points are located on the edges of a cluster and have fewer neighboring points within epsilon.

Noise Points: These are shown as white dots. Noise points are not part of any cluster and have very few or no nearby points within epsilon.

By examining the plot, we can identify distinct clusters based on the colors assigned. Each cluster consists of core points connected to one another through nearby points. The algorithm successfully captures clusters of various shapes and sizes, accommodating both tightly packed clusters and more sparsely distributed ones.

HEAT MAP



"This correlation heatmap offers useful information about the connections between the variables in our dataset. Warmer colors denote positive correlations, whereas colder colors denote negative correlations. Each cell in the heatmap reflects the correlation coefficient between two variables. The strength of the relationships is indicated by the color intensity. We can see the strength and direction of the associations between various variables by examining the heatmap. This data enables us to find probable linkages and trends in our dataset. The heatmap aids in understanding the linkages and helps direct additional research and project-related decision-making.

CONCLUSION

In conclusion, although cluster analysis is an unsupervised method we can’t predict anything exactly but we group similar categories together exactly using the clustering analysis techniques, including k-means, DBSCAN, and hierarchical clustering and plot to visualize .to analyze a dataset containing various attributes such as resorts, beaches, parks, restaurants, and museums. Through these clustering methods, I was able to group similar data points together based on their attribute values. The k-means clustering revealed distinct clusters that represented different preferences or ratings for the various categories. Investigating the output cluster groups Each data point is assigned to one of the three clusters, labeled as Cluster 1, Cluster 2, and Cluster 3. The cluster centers represent the centroid coordinates of each cluster. The output allows us to understand the grouping of data points into clusters and provides insights into the distribution and characteristics of each cluster in the dataset. on The DBSCAN clustering identified dense regions in the dataset and classified outliers as noise points on analysing the visual plot A data point from the dataset is represented by each point on the plot.The DBSCAN method determines that points with the same colour belong to the same cluster.After using PCA, the y-axis shows the second principle component and the x-axis the first principal component.The cluster designations are shown in the colorbar on the plot's right side. The hierarchical clustering provided insights into the hierarchical structure and relationships between data points. The x-axis represents the categories, and the y-axis represents the distance or similarity between categories. The height of the dendrogram branches indicates the degree of similarity between clusters. Overall, these clustering techniques allowed for a deeper understanding of the data, enabling the identification of patterns, similarities, and differences within the dataset. This analysis can be valuable for segmentation, recommendation systems, and targeted marketing strategies.

##### References

1. K-means clustering:

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.

1. Hierarchical clustering:

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. ACM Computing Surveys (CSUR), 31(3), 264-323.

1. DBSCAN clustering:

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96) (pp. 226-231).

1. Unsupervised machine learning:
2. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.
3. Clustering algorithms:

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 31(8), 651-666.

We suggest that you use a text box to insert a graphic (which is ideally a 300 dpi TIFF or EPS file, with all fonts embedded) because, in an MSW document, this method is somewhat more stable than directly inserting a picture.

To have non-visible rules on your frame, use the MSWord “Format” pull-down menu, select Text Box > Colors and Lines to choose No Fill and No Line.