CUSTOMER SEGMENTATION USING KMEANS AND USING PRINCIPAL COMPONENT ANALYSIS FOR DIMENSIONALITY REDUCTION

PLEASE REFER TO THE BELOW LINK FOR COMPLETE ANALYSIS OF MARKET SEGEMENTATION ANALYSIS USING K MEANS https://github.com/snehil943/Market-Segmentation-Analysis-/blob/main/MARKET_SEGMENTATION_KMEANS.ipynb

EXPLORING DATA

- Before starting data analysis, it is important to clean the data by checking for correct values and consistent labels for categorical variables.

- Implausible values should be identified and corrected, indicating errors during data collection or entry.

- Categorical variables should contain only permissible values, and any other values need to be corrected during data cleaning.

- In the Australian travel motives data set, variables like Gender and Age do not require cleaning.

- The variable Income2 in the data set has categories that are not sorted in order due to how data is read into R.

- Categorical variables are stored as factors in R, and their levels are sorted alphabetically by default.

- The income variable categories can be re-ordered by creating a helper variable, storing its levels, finding the correct order, and converting it into an ordered factor in R.

DATA CLEANING

- Understanding the data is crucial to avoid misinterpreting complex analysis results.

- Descriptive numeric and graphic representations provide insights into the data.

- The "summary()" command in R provides a numeric summary for numeric variables and frequency counts for categorical variables, including the number of missing values.

- Histograms, boxplots, scatter plots, and bar plots are useful graphical methods for analyzing data.

- Histograms visualize the distribution of numeric variables

DESCRIPTIVE ANALYSIS AND DATA PREPROCESSING

- Histograms, boxplots, and scatter plots are useful graphical methods for analyzing numeric data.

- Bar plots and mosaic plots are helpful for visualizing categorical variables.

- Histograms visualize the distribution of numeric variables by creating bins and displaying the frequency of observations in each bin.

- Boxplots summarize unimodal distributions by compressing data into minimum, first quartile, median, third quartile, and maximum values.

- Boxplots can reveal distributional properties like skewness and the presence of outliers.

- Graphical methods provide a quick and intuitive overview of the data's structure.

- Dot charts can represent the percentage of agreement with different variables, offering insights into the heterogeneity of responses.

- Graphical inspection confirms the suitability of variables for segmentation analysis in market research.

PRINCIPAL COMPONENT ANALYSIS

- Principal components analysis (PCA) transforms a multivariate data set into a new data set of uncorrelated variables called principal components.

- Principal components are ordered by importance, with the first component containing the most variability.

- PCA preserves the relative positions of observations while changing the perspective of the data.

- PCA is based on the covariance or correlation matrix of numeric variables.

- If variables are measured on the same scale and have similar ranges, the covariance matrix can be used. Otherwise, the correlation matrix should be used.

- PCA is often used to reduce the dimensionality of high-dimensional data for visualization purposes.

- The first few principal components, which capture the most variation, are typically used for plotting.

- Scatter plots and scatter plot matrices can be used to visualize principal components.

INTERPRETING PRINCIPAL COMPONENT ANALYSIS RESULTS

- The output of PCA provides information about each principal component, including the standard deviation, proportion of explained variance, and cumulative proportion of explained variance.

- Principal component 1 explains approximately 18% of the variance in the original data, while principal component 2 explains about 9%.

- Together, the first two principal components explain 27% of the total variation in the original data.

- Principal components 3 to 15 explain smaller proportions of the original variation, ranging from 8% to 3%.

- The low proportion of variance explained by the first few principal components suggests that all the original variables are necessary for segmentation, as they provide valuable information and are not redundant.

- From a projection perspective, this means that it is not easy to represent the data in lower dimensions if only a small number of principal components explain a significant proportion of the variance.

- Using more principal components can provide a better visual representation of the proximity of observations to each other.

GROUPING CONSUMERS USING DISTANCE BASED METHODS (K MEANS AND HIERARCHICAL CLUSTERING)

Market segmentation analysis using data-driven methods can be challenging due to the unstructured nature of consumer data and the lack of clear consumer groups. The results of segmentation methods heavily rely on the assumptions made about the structure of segments and the interaction between the data and the chosen algorithm. Different clustering methods, commonly used for segmentation, can yield different segmentation solutions. For example, k-means cluster analysis may fail to identify complex patterns in the data, while single linkage hierarchical clustering can capture such patterns.

However, there is no single best algorithm for all situations, and the choice of algorithm depends on the characteristics of the data and the desired segment characteristics.

The chapter aims to provide an overview of popular extraction methods for market segmentation. Distance-based methods focus on finding groups of similar observations based on a distance measure, while model-based methods formulate stochastic models for segments. Additionally, some methods incorporate variable selection during segmentation. Comparing alternative segmentation solutions is crucial, considering data characteristics, expected segment characteristics, sample size, and the scale level of variables. Distance measures such as Euclidean and Manhattan are commonly used.

Hierarchical clustering methods mimic how humans would approach dividing observations into groups. Divisive hierarchical methods start with the complete data set and split it into segments iteratively, while agglomerative methods merge the closest segments until one large segment is formed. The hierarchical clustering process results in a sequence of nested partitions.

While deterministic algorithms are commonly used for hierarchical clustering, the specific algorithm choice depends on the requirements of the analysis.

K MEANS AND K CENTROID CLUSTERING

- The goal of partitioning clustering is to divide observations (consumers) into subsets (market segments) that are similar within each segment and dissimilar between segments.

- The centroid represents a market segment and is calculated as the column-wise mean values across all members of the segment.

- The partitioning algorithm involves five steps: specifying the desired number of segments, randomly selecting initial cluster centroids, assigning observations to the closest centroid, recomputing centroids based on cluster membership, and repeating until convergence or a maximum number of iterations.

- The algorithm always converges but may take longer for large data sets or a large number of segments.

- Different random initial centroids lead to different segmentation solutions, so repetition is important for obtaining the best segmentation solution.

- Determining the optimal number of segments is a challenge, and stability analysis or indices can assist in selecting the appropriate number.

- Partitioning clustering requires specifying the number of segments in advance.

- Distance measures, such as squared Euclidean distance, Manhattan distance, or angle difference, significantly impact the resulting segmentation solution, often more than the choice of algorithm.

- The choice of distance measure influences the shape and orientation of cluster borders in the resulting partitions.

PLEASE REFER TO THE BELOW LINK FOR COMPLETE ANALYSIS OF MARKET SEGEMENTATION ANALYSIS USING K MEANS https://github.com/snehil943/Market-Segmentation-Analysis-/blob/main/MARKET_SEGMENTATION_KMEANS.ipynb