D212

Data Mining II

Task 1: Clustering Techniques

28 March 2024

WGU

I: Research Question

A1. Using k-means clustering, how does the distribution of customers across clusters vary, especially concerning the continous variables: population, age, income, and tenure?

A2. A goal of this analysis is to identify customer segments based on population, age, income, and tenure.

II: Technique Justification

B1. K-means clustering is the chosen technique due to its ability to partition the data into distinct clusters based on similarity of data points. This makes it suitable for exploring patterns and structures within the data. Additionally, the technique is also easily scalable to large datasets, making it applicable for future expansion of the dataset. Expected outcomes include the formation of k clusters represented by centroids, clear separation between clusters, and interpretation of clusters will gain insights into the underlying structure of the date.

B2. An assumption of K-means clustering is the clusters are spherical, implying groups with different sizes and densities may not be captured well by this technique (DataCamp, 2023).

B3.

A screenshot of a computer code

Description automatically generated

III: Data Preparation

C1. Preprocessing data by standardization is important due to the sensitivity of k-means clustering to the scale of variables. This was preformed using StandardScaler.

C2.

|  |  |
| --- | --- |
| Population | Continuous |
| Age | Continuous |
| Income | Continuous |
| Tenure | Continuous |

C3.

A screenshot of a computer code

Description automatically generated

A screenshot of a bar code

Description automatically generated

A screenshot of a computer program

Description automatically generated

A graph of a graph and a graph of a graph

Description automatically generated

A graph of a graph of a number of blue and black lines

Description automatically generated with medium confidence

A screenshot of a graph

Description automatically generated

A screen shot of a graph

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A screen shot of a graph

Description automatically generated

A screen shot of a computer code

Description automatically generated

A screen shot of a graph

Description automatically generated

C4.

A white card with black text

Description automatically generated

IV: Analysis

D1. The optimal number of clusters was determined to be 6, through use of an elbow plot.

A screen shot of a graph

Description automatically generated

D2.

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

V: Data Summary and Implications

E1. When assessing clusters based on homogeneity, it becomes apparent that each cluster exhibits a certain level of similarity, with some variation in the similarity across the clusters. For instance, Cluster 1 displays higher density compared to Cluster 2, indicating a greater homogeneity among its data points. Meanwhile, Cluster 2 demonstrates higher income compared to Cluster 1.

E2. Insights into customer segmentation based on continuous variables: population, age, income, and tenure provide insight into the customer base, aiding in development of targeted marketing strategies for customer retention. For instance, Cluster 3 is characterized by relatively high income and tenure, which may represent a segment for premium services. On the other hand, Cluster 5 exhibits lower income and tenure, suggesting the need for promotional offers to enhance retention rates.

E3. One limitation of this analysis is the assumption of spherical clusters. This assumption may not hold true for datasets with different sizes and densities, leading to suboptimal clustering outcomes due to the high complexity of data (Google for Developers, 2022).

E4. Recommended course of action is to further explore cluster characteristics through dashboard creation or expand visualizations.

VI: Demonstration

F1. Panopto

G. Reference Sources

Wilson, B. (2024). Unsupervised Learning in Python [Online course]. DataCamp. Retrieved from <https://app.datacamp.com/learn/courses/unsupervised-learning-in-python>

Yusuf, F. (2023, June 1). Difference between K-means and Hierarchical Clustering. Medium. <https://medium.com/@waziriphareeyda/difference-between-k-means-and-hierarchical-clustering-edfec55a34f8>

Gross, K. (2020, June 18). Clustering: How It Works (In Plain English!). Data Basics. Retrieved from <https://blog.dataiku.com/clustering-how-it-works-in-plain-english>

H. In-text Sources

DataCamp. (2023). K-Means Clustering in R Tutorial. Retrieved from <https://www.datacamp.com/tutorial/k-means-clustering-r>

Google for Developers. (2022, July 18). k-Means Advantages and Disadvantages. Retrieved from <https://developers.google.com/machine-learning/clustering/algorithm/advantages-disadvantages>

I. Communication