**D212 PA**

Jon Isak

D212: Data Mining II

Contents

[**D212 Task 1 3**](#_3xni6yjcaxld)

[**Part I: Research Question 3**](#_83kpmydrb0mx)

[**Part II: Technique Justification 3**](#_bgv1yfgfld31)

[**Part III: Data Preparation 4**](#_3u4qcwvukb05)

[**Part IV: Analysis 5**](#_6i10ucbw6icv)

[**Part V: Data Summary and Implications 5**](#_2zeldq84puwr)

[**Part VI: Demonstration 5**](#_bcnpyxmpx6qr)

## 

## 

## 

## D212 Task 1

## Part I: Research Question

A1. This paper will look into the question, how can the identification of customer clusters through k-means clustering aid in understanding customer behavior and preferences?

A2. The primary goal of the data analysis is to identify distinct customer segments within the company's customer base using continuous variables. This segmentation will enable the company to understand customer behaviors, preferences, and needs, ultimately facilitating targeted strategies for improving customer satisfaction and providing more value to the shareholders. This analysis aims to group customers based on similarities in these continuous attributes, enabling the company to tailor services, marketing, and customer service efforts to specific segments identified through clustering. This goal aligns with the company's objective of leveraging customer data to make data driven decisions, create personalized experiences, and ultimately improve customer loyalty and profitability.

## Part II: Technique Justification

B1. The k-means clustering technique partitions the dataset into clusters based on similarities in the variables. It works by iteratively assigning each data point to the nearest center and then updating the centroids based on the mean of the points assigned to each cluster. This process continues until centroids no longer change significantly, optimizing the clustering solution (Frost, n.d.). I expect to identify clusters representing distinct customer segments based on similarities in continuous variables that will be used. Each cluster will have customers who share similarities in these continuous attributes, allowing the company to understand different customer behavior patterns and preferences.

B2. One assumption of the clustering technique is that clusters are spherical and have a similar variance. The algorithm assumes clusters have a roughly equal number of observations and are evenly distributed (Perceptive Analytics, 2017).

B3. The analysis will require multiple packages and libraries. Scikit-learn will be useful for performing the k\_means clustering algorithm. Pandas will be useful for data manipulation, handling datasets, and preparing data. Matplotlib will be useful for creating visualizations.

## Part III: Data Preparation

C1. One goal of data preprocessing is to make the data suitable for analysis. This will be done by *standardizing* the scale of the variables.

C2. The initial data set variables will be used to perform the analysis include MonthlyCharge, Bandwidth\_GB\_Year, Income, and Outage\_sec\_perweek. All these variables are continuous, as required by the k\_means algorithm.

C3. One of the first steps in preparing the data for the analysis is scaling. I will use StandardScaler to scale the continuous variables identified earlier.

C4. The data set is attached.

## Part IV: Analysis

D1. The optimal number of clusters in the data set is 3. The method used to determine this number is by creating a graph and looking for the elbow point (Saji, 2023).

D2. The code is attached.

## Part V: Data Summary and Implications

E1. The quality of the clusters created is good. A silhouette score of .25 suggests a moderate separation. Each cluster shows distinctiveness with minimal overlapping.

E2. The clustering analysis revealed three distinct customer segments with varying attributes. Each cluster represents customers with different incomes, usage patterns, charges, and outages. For cluster 0 centroid characteristics, the averages were 52.87 for age, $30k for income, $172.59 for monthly charge, 1303 for bandwidth, and 10.05 sec for outage time. Customers in Cluster 0 display moderate income levels, moderate service usage, and moderate outage times. For cluster 1 centroid characteristics, the averages were 52.89for age, $89,934 for income, $170.09 for monthly charge, 3325.80 for bandwidth, and 9.78 sec for outage time. This cluster indicates moderate metrics across all variables. Customers in this cluster have higher income levels, higher service usage, and lower outage time. For cluster 2 centroid characteristics, the averages were 53.36 for age, $ $30,425 for income, $173.63 for monthly charge, 5507.47for bandwidth, and 10.04 sec for outage time. This cluster indicates moderate metrics across all variables. Customers in Cluster 2 show lower income levels, higher service usage, and moderate outage time. Understanding these segments can help tailor marketing strategies, service offerings, and customer service initiatives to suit each group's specific needs and preferences.

E3. The clustering analysis was performed based on a small set of variables from the dataset. Using a more comprehensive set of features might provide deeper insights and potentially improve the quality of segmentation.

E4. The company should utilize the identified customer segments to develop targeted strategies to address the needs of each segment. Cluster 0 has moderate income and usage. In segment, the company should focus on cost effectiveness and reliability. Marketing strategies could highlight reliability and promotions that over a god value. Cluster 1 has high income and usage. For this cluster, the company could offer premium services, more features, and personalized support. These additional offerings could increase company revenue from customer sin these segments. Cluster 2 has lower income and higher usage. In this cluster, the company could offer budget services and pay by use plans. This could improve their satisfaction by tailoring services to their needs.

## Part VI: Demonstration

F. https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=76f665ea-ceed-47af-bb1c-b0d4011e375e

G. Saji, B. (2023, September 21). *Elbow Method for Finding the Optimal Number of Clusters in K-Means*. Analytics Vidhya. Retrieved December 10, 2023, from https://www.analyticsvidhya.com/blog/2021/01/in-depth-intuition-of-k-means-clustering-algorithm-in-machine-learning/

H.

Frost, J. (n.d.). *What is K Means Clustering? With an Example*. Statistics By Jim. Retrieved December 9, 2023, from https://statisticsbyjim.com/basics/k-means-clustering/

Perceptive Analytics. (2017, August 7). *Exploring Assumptions of K-means Clustering using R*. R-bloggers. Retrieved December 9, 2023, from https://www.r-bloggers.com/2017/08/exploring-assumptions-of-k-means-clustering-using-r/

Saji, B. (2023, September 21). *Elbow Method for Finding the Optimal Number of Clusters in K-Means*. Analytics Vidhya. Retrieved December 10, 2023, from https://www.analyticsvidhya.com/blog/2021/01/in-depth-intuition-of-k-means-clustering-algorithm-in-machine-learning/