**Performance Assessment: D212- Clustering Techniques**

**A. Research Question**

**1.** For this assessment, the research question is as follows: how many clusters (or value of K) exist in the churn data set amongst the four continuous variables; these variables being income, outage\_sec\_perweek, tenure, and bandwidth\_GB\_year. Subsequently, what do the clusters tell us about the data?

**2.** One goal of this analysis is to determine the number of clusters that are present between these four continuous variables and what they tell us about the data overall.

**B. Technique Justification**

**1.** The k-means algorithm starts by taking the data points from the data source and randomly placing them into different groups, each with a central point. Using this central point, the algorithm determines the distance from that point to the center of that cluster, which is referred to as the Euclidian distance, “and if the distance of a point is higher than to another centroid, the point is reassigned to the ‘other’ centroid” (Oyoga 2023). This process repeats until it forms a predetermined number of clusters (which are inputted by the user) where each cluster is equally sized and well separated. Expected outcomes include the assignment of clusters based upon the predetermined number (or the value of k). We can find the optimal number of K by using the silhouette method, where the algorithm runs various values of k and returns the width of the silhouettes using those values, where the largest width being the best value for k.

**2.** One assumption of k-means clustering is that the clusters “are spherical” (“Exploring Assumptions of K-Means Clustering Using r” 2017). What this means is that the clusters created by the algorithm are of equal size and form a spherical shape, where each point in the cluster is of a closer distance to the point of that particular cluster than the center of all of the clusters, giving the full plot of clusters a spherical shape.

**3.** The following packages were used to perform the analysis in R:

* Factoextra: for clustering and visualization
* Cluster: for clustering and visualization

**C. Data Preparation**

**1.** There are a handful of steps required to prepare the data for analysis. In order, first the columns that are not going to be used in the analysis need to be removed from the data frame, after which I will check the remaining variables for missing data and replace those (if any exist) with the mean. The next step, since K-means is sensitive to outliers, is to remove all outliers (if any exist) from the remaining variables. The final step to prepare the data for analysis is to scale the variables so they are balanced and proportional to each other from 0 to 1.

**2.** The variables that will be analyzed are **ALL** continuous variables and they are as follows: income, outage\_sec\_perweek, tenure, and bandwidth\_GB\_year.

**3.** The following is a list of the steps used to prepare the data for analysis:

* Remove the unnecessary data from the CSV file by using the subset function in R
* Check for missing data using the colSums function; no data was missing
* Use a boxplot to check for outliers. Outliers were present for both Income and Outage\_sec\_perweek. I removed the outliers for Outage\_sec\_perweek but chose to keep the outliers for Income as it is plausible that a company would have customers of significantly varying degrees of income on all sides of a monetary spectrum.
* Last step is to scale the data, so each field is proportional to each other

The code used to prepare the data has been attached as an RScript file alongside this written assessment.

**4.** A copy of the cleaned data set has been submitted alongside this written assessment.

**D. Analysis**

**1.** For this assessment, the optimal number of clusters (or the value for K is 2). This number was determined by using the silhouette method within the fviz\_nbclust function in the factoextra package. Entering this command results in the following chart:

A graph showing the average value of a graph

Description automatically generated with medium confidence

This chart, by looking at the average width of the silhouettes from a silhouette chart with varying numbers for K, clearly states that the optimal number of clusters is 2. This can be seen via the dotted line as by setting k equal to 2 results in an average silhouette width of over 0.4.

**2.** The code used to perform the analysis have been attached as an RScript file alongside this written assessment (Zach 2022).

**E. Data Summary and Implications**

**1.** The resulting analysis resulted in clusters that are represented in the following screenshot:

A screenshot of a graph

Description automatically generated

By observing the cluster plot, you can see that there is a clear separation between the two clusters along the Dim1 axis. Both of the clusters are easy to distinguish from each other and the algorithm was able to partition the data very well to create these clusters. It can also be observed that these two clusters account for 49.8% and 25.2% of the total variance, which gives them a combined 72.4% of the total variance of the data. This is significant as the algorithm was able to create these two clusters in a way that captured the majority of the variance of the data. This means that these clusters have a high quality, although 27.6% of the variance is still not represented. But overall, the k-means algorithm did a good job.

**2.** Upon aggregating the model, the following screenshot displays the means of each cluster:

A close up of numbers

Description automatically generated

Upon observing this summary aggregation, both clusters have very similar means for their income as well as their Outage\_sec\_perweek, however, there is a clear distinction in the means of each clusters tenure as well as their bandwidth\_GB\_year. The first cluster has an average mean for tenure of approximately 9.14 and an average bandwidth\_GB\_year of approximately 1312.364 while the second cluster has averages of 59.94 and 5474.206 respectively. Because of these averages, we can make the conclusion that cluster one is made up over new customers and cluster two consists of long-term users of the service, due to the massive difference in the tenure in each cluster. We can also determine that the users in the first cluster are more likely to be on basic or smaller plans because of the smaller bandwidth while the users in cluster two appear to be on more premium plans that result in higher data usage.

**3.** This particular analysis only focuses on income, outage\_sec\_perweek, tenure, and bandwidth\_GB\_year so the analysis is limited in that it does not include the age, gender, education levels, etc. of the customers which could be used to discover more detailed information in regard to customer preferences for the service.

**4.** Since the clusters show a clear distinction in bandwidth for longer tenured customers vs newer users, a recommended course of actio**n is to** focus on retention strategies in order to keep all of the customers happy, so the new customers eventually turn into longer users.

**F. Sources**

Zach. “K-Means Clustering in R: Step-by-Step Example.” Statology, 8 Sept. 2022, www.statology.org/k-means-clustering-in-r/.

Straw, Eric. “K Means Clustering.” Vimeo, 2 April, 2021, vimeo.com/532083445.

“Exploring Assumptions of K-Means Clustering Using r.” R-Bloggers, 7 Aug. 2017, www.r-bloggers.com/2017/08/exploring-assumptions-of-k-means-clustering-using-r/.

Oyoga, Clinton. “What Is K-Means Clustering and How Does Its Algorithm Work?” Saturn Cloud Blog, 4 April. 2023, saturncloud.io/blog/what-is-k-means-clustering-and-how-does-its-algorithm-work/.