Raquel Ocasio
D212 – Data Mining II, Task 1
November 19, 2023
Western Governors University

Part A1

Which continuous variables can help us to better understand the similar characteristics of patients who are re-admitted to the hospital?

Part A2

The goal of the data analysis is to use k-means clustering to identify the characteristics of patients who are re-admitted to the hospital.

Part B1

The k-means clustering technique analyzes the dataset by randomly assigning each record to a cluster, then computing the centroid for each cluster and assigning each observation to the cluster whose centroid is the closest as defined by Euclidean distance. The centroid calculation and observation assignment to a cluster are repeated until the cluster assignments stop changing. (Zach, 2022b)

Expected outcomes is a set of clusters where each observation in the dataset is assigned to a cluster.

Part B2

One assumption of k-means clustering is that the clusters have a spherical shape. (Medium, n.d.)

Part B3

The factoextra, cluster, dplyr, and readr libraries were used with R. The factoextra library supports the analysis by helping to enhance the output of clustering techniques. The cluster library supports the analysis by providing functions for cluster analysis. The dplyr library supports the analysis by providing functions for data manipulation, such as filter(). The readr library supports the analysis by providing functions to read files into R.

Part C1

One goal of preprocessing the dataset is to ensure that only continuous variables are used in the analysis.

Part C2

The initial dataset variables and their labels are listed below.

Variable Name	***Variable Description***	***Variable Type***
Income	Annual income of the patient	Continuous
VitD_levels	Patient’s vitamin D levels measured in ng/mL	Continuous

Variable Name	***Variable Description***	***Variable Type***
TotalCharge	Amount charged daily to the patient	Continuous
Additional_charges	Average amount charged to patient for miscellaneous items	Continuous

Part C3

Five steps were taken to prepare the data for analysis:

1. Filtering the data to only those records where the patient was re-admitted. The code segment is on lines 9-13.
2. Saving the required columns to a new dataframe. The code segment is on lines 15-19.
3. Checking the new dataframe for duplicates. The code segment is on line #26.
4. Checking the new dataframe for missing values. The code segment is on line #29.
5. Checking the new dataframe for outlier values. The code segment is on lines 31-63.

Part C4

See attached file.

Part D1

The optimal number of clusters for the dataset is five. The optimal number of clusters was determined by creating two plots: number of clusters vs the total within sum of squares, and number of clusters vs gap statistic. The number of clusters vs the total within sum of squares plot displays a level off point that is typically considered the optimal number of clusters. The number of clusters vs gap statistic plot displays a number for the optimal number of clusters.

Part D2

See attached code.

Part E1

Inertia measures how far the points within a cluster are from the centroid of that cluster. A lower inertia indicates better-defined clusters. The inertia metric was selected as the method for determi nation of cluster quality because it provides it makes it easy to understand how well defined the c lusters are.

Based on the high inertia value of 7,264.324, the quality of the created clusters is low. The high i nertia value indicates that the clusters are not tightly packed or well defined.

Part E2

The k-means cluster analysis produced five clusters of size 438, 798, 823, 961, and 649, with an inertia value of 7,264.324. The low quality of the clusters indicates that the continuous variables used in the analysis are not useful for understanding the similar characteristics of patients who are admitted to the hospital.

The analysis produced a cluster means value for each variable, indicating which variable contributes the most to the formation of that cluster (having the highest mean). A table for the cluster and variable means is below.

Cluster #	***Variable with highest mean***
1	Income
2	TotalCharge
3	VitD_levels
4	VitD_levels
5	Additional_charges

Part E3

One limitation of this data analysis was determining the optimal number of clusters. The two plot methods used gave varying results which made it harder to determine which recommendation for a value for k was best.

Part E4

It is recommended that the stakeholders consider other variables that may help to better categorize the characteristics of those patients who are re-admitted to the hospital.

Parts F and F1

The video demonstration can be viewed at https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=aab2d265-6904-401b-8cbf-b0bf0062d1aa

Part G

Zach. (2022b, September 8). K-Means Clustering in R: Step-by-Step example. Statology. https://www.statology.org/k-means-clustering-in-r/

Medium. (n.d.). Medium. https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks

Part H

None used.