D212: Data Mining II

Task 1

Lucas Benish

Western Governors University

# Part I: Research Question

## Describe the purpose of this data mining report by doing the following:

### Propose **one** question relevant to a real-world organizational situation that you will answer using **one** of the following clustering techniques:

### *k*-means

### hierarchical

The question that I will be seeking to answer is as follows: based on the continuous, patient demographic variables present within the dataset, can we accurately cluster patients into readmission and non-readmission groups?

### Define **one** goal of the data analysis. Ensure that your goal is reasonable within the scope of the scenario and is represented in the available data.

One goal of my data analysis is to determine the accuracy of my k-means clustering model on predicting patient readmission, when the model is given only patient demographic information.

# Part II: Technique Justification

## Explain the reasons for your chosen clustering technique from part A1 by doing the following:

### Explain how the clustering technique you chose analyzes the selected dataset. Include expected outcomes.

The clustering technique I have chosen to use is k-means clustering. This technique utilizes continuous variables within the dataset and groups observations based on the distance between them and a cluster centroid. The expected outcome is that patients will be grouped into clusters with observations containing similar characteristics. The appropriate number of clusters is often determined based on domain knowledge or determined based on a method such as the elbow plot or silhouette plot.

### Summarize **one** assumption of the clustering technique.

One assumption of k-means clustering is that the created clusters should contain approximately the same number of observations.

### List the packages or libraries you have chosen for Python or R, and justify how *each* item on the list supports the analysis.

I will once again be utilizing R Studio to perform my analysis. The packages that I will be utilizing are naniar, visdat, readr, cluster, and factoextra. Cluster will be utilized for its clustering models and factoextra will also be utilized for clustering models, as well as visualization of the clusters. Naniar, visdat, and readr will be utilized for data exploration and importing.

# Part III: Data Preparation

## Perform data preparation for the chosen dataset by doing the following:

### Describe **one** data preprocessing goal relevant to the clustering technique from part A1.

Prior to performing k-means clustering, I need to scale my continuous variables. Scaling involves transforming the continuous variable so that the mean equals 0, and the remaining values range from -1 standard deviation to 1 standard deviation from the mean. This will allow for equal weighting of variables so that a variable that is inherently a larger value, such as income, will not bias the clustering.

### Identify the initial dataset variables that you will use to perform the analysis for the clustering question from part A1, and label *each* as continuous or categorical.

The initial variables from my dataset that will be used in the analysis are noted below. These were chosen because they are continuous patient variables that are present prior to a patient being admitted to the hospital.

Age – continuous

Income – continuous

Population – continuous

Children - continuous

### Explain *each* of the steps used to prepare the data for the analysis. Identify the code segment for *each* step.

The first step of any analysis that I perform is setting up the R studio environment by loading in the appropriate libraries. After the libraries are loaded, the data is imported and copied to have an extra dataframe. This was done with the following code:

#setting up environment

library(readr)

library(naniar)

library(visdat)

library(cluster)

library(factoextra)

#importing and copying dataset

medical\_clean <- read\_csv("C:/Users/lgben/OneDrive/Desktop/MSDA/D212 - Data Mining II/medical\_clean.csv")

medical\_clean\_copy <- medical\_clean

After the data is imported, any cleaning necessary will be performed. To begin, I always look at the general structure of the dataframe to familiarize myself with the data types. Following this, I visualize the missingness of the data and determine the overall number of missing values and duplicates. A step that I do not always do at this point, but decided to here, is reduced the number of variables to only those of interest for the current analysis. This was accomplished by the following code:

#general view/exploration

str(medical\_clean)

vis\_miss(medical\_clean)

sum(is.na(medical\_clean))

sum(duplicated((medical\_clean)))

#keeping only pertinent variables - continuous variables that patient has prior to stay, except ZIP/lat/long

#because looking pre-stay, initial days, total charge are excluded

#keeping re-admis for later

medical\_clean\_continuous\_vars <- medical\_clean[, c('Customer\_id', 'Population', 'Children', 'Age', 'Income')]

medical\_clean\_continuous\_vars\_with\_readmit <- medical\_clean[, c('Customer\_id', 'ReAdmis', 'Population', 'Children', 'Age', 'Income')]

medical\_clean\_continuous\_vars\_with\_readmit$ReAdmis\_NBR <- as.numeric(as.factor(medical\_clean\_continuous\_vars\_with\_readmit$ReAdmis))-1

summary(medical\_clean\_continuous\_vars\_with\_readmit$ReAdmis\_NBR)

After narrowing my dataframe down to only the needed variables, I began to look for the presence of outliers. This was performed after reducing the number of columns, as outliers may have been present in other columns, while not in the ones of interest. Outlier presence was determined by creating z-score columns and determining presence of z-scores greater than 3 or less than -3. Vectors of the observations containing outliers were created and combined into a single vector. A new dataframe, without the observations containing outliers, was created. Outliers were removed, rather than imputed, due to the low number of outliers present. The number of outliers totaled 546, or approximately 5% of the dataset. This was accomplished by the following code:

#further exploring/cleaning - outliers?

#z-score columns

medical\_clean\_continuous\_vars$population\_z <- scale(x=medical\_clean\_continuous\_vars$Population)

medical\_clean\_continuous\_vars$children\_z <- scale(x=medical\_clean\_continuous\_vars$Children)

medical\_clean\_continuous\_vars$age\_z <- scale(x=medical\_clean\_continuous\_vars$Age)

medical\_clean\_continuous\_vars$income\_z <- scale(x=medical\_clean\_continuous\_vars$Income)

#outlier vectors

population\_outliers <- which(medical\_clean\_continuous\_vars$population\_z >3 | medical\_clean\_continuous\_vars$population\_z < -3)

children\_outliers <- which(medical\_clean\_continuous\_vars$children\_z >3 | medical\_clean\_continuous\_vars$children\_z < -3)

age\_outliers <- which(medical\_clean\_continuous\_vars$age\_z >3 | medical\_clean\_continuous\_vars$age\_z < -3)

income\_outliers <- which(medical\_clean\_continuous\_vars$income\_z >3 | medical\_clean\_continuous\_vars$income\_z < -3)

#treating outliers

unique\_outliers <- unique(c(children\_outliers, income\_outliers, population\_outliers))

medical\_clean\_continuous\_vars <- medical\_clean\_continuous\_vars[-unique\_outliers, ]

medical\_clean\_continuous\_vars\_with\_readmit <- medical\_clean\_continuous\_vars\_with\_readmit[-unique\_outliers, ]

medical\_clean\_continuous\_vars <- subset(medical\_clean\_continuous\_vars, select = -c(age\_z, income\_z, population\_z, children\_z))

#removing outlier vectors/objects from environment

remove(age\_outliers, children\_outliers, income\_outliers , unique\_outliers, population\_outliers)

After the data was cleaned and extra objects removed from the environment, my data was almost ready for analysis. The final preparation step to complete was scaling my data and writing my new dataset to my personal computer for submission. Scaling involves transforming a continuous variable so that the mean is 0, and all other observations are within 1 to -1 standard deviation from the mean. This was accomplished by the following code:

#scaling data

medical\_cont\_scaled\_ID <- medical\_clean\_continuous\_vars[, 1]

medical\_cont\_scaled <- scale(medical\_clean\_continuous\_vars[, c('Population', 'Children', 'Age', 'Income')])

medical\_cont\_scaled\_with\_ID <- cbind(medical\_cont\_scaled\_ID, medical\_cont\_scaled)

str(medical\_clean\_cont\_scaled)

summary(medical\_cont\_scaled) #verifies that mean for all continuous variables are 0

#writing prepped data to folder

write.csv(medical\_cont\_scaled\_with\_ID, 'C:\\Users\\lgben\\OneDrive\\Desktop\\MSDA\\D212 - Data Mining II\\Task 1\\medical\_cont\_scaled\_with\_ID.csv')

### Provide a copy of the cleaned dataset.

A copy of medical\_cont\_scaled\_with\_ID will be provided with my submission.

# Part IV: Analysis

## Perform the data analysis and report on the results by doing the following:

### Describe the analysis technique you used to appropriately analyze the data. Include screenshots of the intermediate calculations you performed.

The analysis technique that I utilized was k-means clustering. This was done using only continuous variables that were present for patients prior to their stay at the hospital. 3 of these variables were specific to the patient: age, income, children. The final one was related to their area of residence: population.

I decided to utilize an elbow plot and silhouette plot to determine what value of k would best limit the within-cluster sum of squares. Both the elbow plot and silhouette plot determined that k=5 was the best value for clustering observations with like observations based on the data. These will be noted below. However, given the domain knowledge that patients either re-admit to the hospital, or don’t readmit to the hospital, k=2 was selected as the final value of k.

A picture containing text, line, diagram, plot

Description automatically generated

A picture containing text, line, plot, diagram

Description automatically generated

After performing the clustering technique, my clusters were visualized. This visualization will be provided below. After visualizing, I added a column to my original dataframe that contained the case ID and continuous variables prior to scaling. This allowed me to create a confusion matrix to calculate accuracy of the clustering model.

A picture containing text, map, screenshot

Description automatically generated

The confusion matrix, mentioned above, and the accuracy measures of my clustering, were as follows:

A picture containing text, font, screenshot, white

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

### Provide the code used to perform the clustering analysis from Part 2.

The following is the R code utilized to perform the visualization of elbow and silhouette plots, clustering, visualization of clusters, and creation of confusion matrix.

#determining appropriate k via elbow plot

set.seed(4206)

fviz\_nbclust(medical\_cont\_scaled, kmeans, method = 'wss')

fviz\_nbclust(medical\_cont\_scaled, kmeans, method = 'silhouette')

#5 appears to be optimal number of clusters based on both within sum of squares and silhouette

#we know however that patients are either readmitted or not readmitted, so both will be run and compared

#kmeans clustering

#clustering5 <- kmeans(medical\_cont\_scaled, 5, nstart = 25)

clustering2 <- kmeans(medical\_cont\_scaled, 2, nstart = 25)

#fviz\_cluster(clustering5, data=medical\_cont\_scaled)

fviz\_cluster(clustering2, data=medical\_cont\_scaled)

#adding cluster to original df with non-scaled values

medical\_cont\_vars\_with\_readmit2 <- medical\_clean\_continuous\_vars\_with\_readmit

medical\_cont\_vars\_with\_readmit2 <- cbind(medical\_cont\_vars\_with\_readmit2, cluster = clustering2$cluster)

medical\_cont\_vars\_with\_readmit2 %>% group\_by(cluster) %>% summarise\_all("mean")

#medical\_cont\_vars\_with\_readmit5 <- medical\_clean\_continuous\_vars\_with\_readmit

#medical\_cont\_vars\_with\_readmit5 <- cbind(medical\_cont\_vars\_with\_readmit5, cluster = clustering5$cluster)

#medical\_cont\_vars\_with\_readmit5 %>% group\_by(cluster) %>% summarise\_all("mean")

#accuracy

#will be using k=2 for confusion matrix

conf\_mat <- table(medical\_cont\_vars\_with\_readmit2$ReAdmis\_NBR, medical\_cont\_vars\_with\_readmit2$cluster)

print(conf\_mat)

true\_neg <- conf\_mat[1, 1]

true\_pos <- conf\_mat[2, 2]

false\_neg <- conf\_mat[2, 1]

false\_pos <- conf\_mat[1, 2]

specificity <- (true\_neg)/(true\_neg+false\_pos)

sensitivity <- (true\_pos)/(true\_pos+false\_neg)

accuracy <- (true\_neg + true\_pos)/(true\_neg + true\_pos + false\_neg + false\_pos)

# Part V: Data Summary and Implications

## Summarize your data analysis by doing the following:

### Explain the accuracy of your clustering technique.

The overall accuracy of my data model was not great. The accuracy was determined via confusion matrix and noted to be 41% accurate. This means that based solely on continuous variables present to the patient prior to hospital admission, the model accurately assigned a patient to readmission or non-readmission approximately 4 out of 10 times.

### Discuss the results and implications of your clustering technique.

The results of this clustering technique were all my observations within my dataset being assigned one of two clusters, based on the similarities they had to each other. The cluster plot, provided in an above section, demonstrates the amount of overlap present between the clusters. What this then implies is that there may not be enough distinctions in the continuous variables present prior to patient admission to stratify them into just two clusters accurately. This may benefit from revisiting with the recommended k=5 for comparison.

### Discuss **one** limitation of your data analysis.

One limitation of this analysis is that kmeans clustering requires use of continuous variables. There were not many continuous, pre-admission, variables present within the dataset. This limits the input for the clusters, and allows for greater similarity of observations, and therefore overlap of clusters.

### Recommend a course of action for the real-world organizational situation from Part A1 based on your results and implications discussed from Part E2.

I would recommend one of the two following options, based on the creation and performance of my kmeans clustering model:

* If kmeans clustering is the desired model, I believe utilizing more continuous variables that are present pre-admission may allow for more distinct clustering.
* If kmeans clustering does not have to be the model utilized for clustering patient re-admission, other variables within the dataset could be used to form a more robust and accurate model.

# Part VI: Demonstration

## Provide a Panopto video recording that includes a demonstration of the functionality of the code used for the analysis and summary of the programming environment.

## Record the web sources used to acquire data or segments of third-party code to support the analysis. Ensure the web sources are reliable.

No outside sources beyond the datacamp modules were utilized for performance of this performance assessment.

## Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.

## Demonstrate professional communication in the content and presentation of your submission.