D212: Data Mining II

Performance Assessment

Task 1

Logan Rosemeyer

Western Governor's University

D212: Data Mining II

Dr. Keiona Middleton

February 22, 2023

Contents

A1: Proposal of Question

My question is: Can we use k-means to cluster patients based on their age and income?

A2: Defined Goal

The goal of the data analysis is to cluster patients based on their age and income. This will allow the organization to better understand the characteristics of its patients.

B1: Explanation of Clustering Techniques

I chose the k-means clustering technique to analyze my dataset. The k-means clustering technique takes the distance between points as a measure of similarity (Al-Masri, 2019). The iterative k-means algorithm tries to iteratively improve the quality of solution of the k-means by removing one cluster, dividing by another one, and applying re-clustering again, in each iteration (Ismkhan, 2018). I expect the algorithm to cluster the data into 9 clusters. This would be broken down into three age clusters and three income clusters. I expect the algorithm to do this because typically we think of income in three clusters: low class, middle class, upper class. We also typically think of age in three clusters: young, middle-age, elderly.

B2: Summary of Technique Assumption

One assumption of k-means is that you know how many clusters you want the data to be broken into. There are ways to find which value of k would be best, but the actual k-means function assumes you know what k is. (Murphy, 2021)

B3: Packages or Libraries List

I used the dplyr package to be able to create my new data set with the variables age and income. I used the ggplot2 package to visualize the clusters of age and income created by k-means. I used the purrr package to use the map_dbl function to be able to build an elbow graph to find the best value for k.

C1: Data Preprocessing

My main data preprocessing goal was to create a data frame that includes only age and income. This is relevant to my clustering technique because I can graph age and income on a coordinate plane and find the distance between the points. This will allow me to cluster points that are similar.

C2: Dataset Variables

I plan to use the age and income variables. They are both continuous.

C3: Steps for Analysis

The first step I took to prepare the data for analysis was to make sure there were no null values or duplicates.

```
sum(duplicated(medical))
```

```
sum(is.na(medical))
```

The next step was to create a data frame that includes the age and income variables. I also scaled the data using the scale function. Without this, I would've just gotten clusters for income, because the distance between different ages would have been very small.

```
ageinc <- select(medical, 16:17)

ageinc <- as.data.frame(scale(ageinc))
```

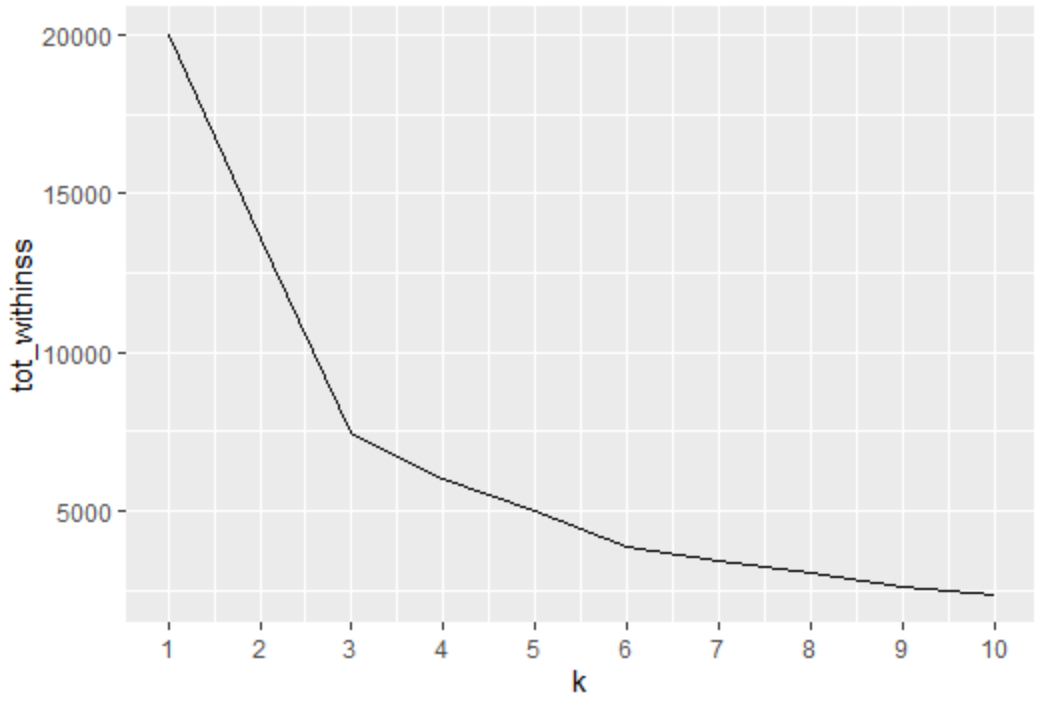
C4: Cleaned Dataset

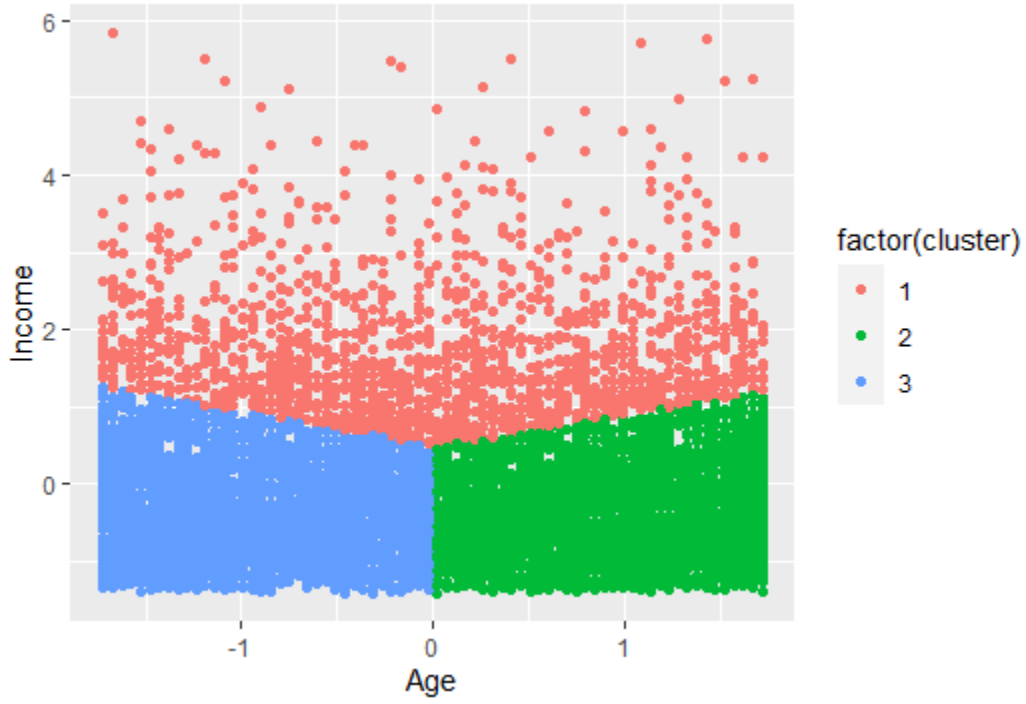
The cleaned dataset in included in my submission.

D1: Output and Intermediate Calculations

I used k-means to analyze my data. I decided to start with 9 clusters. I did this because typically we think of income in three clusters: low class, middle class, upper class. We also typically think of age in three clusters: young, middle-age, elderly. The function was pretty close to doing this as you can see by the graph below.

After this, I decided to create graph with total sum of squares for each k value. The graph starts to level out when k is 3, so I decided to run the k-means function again, this time with 3 clusters.





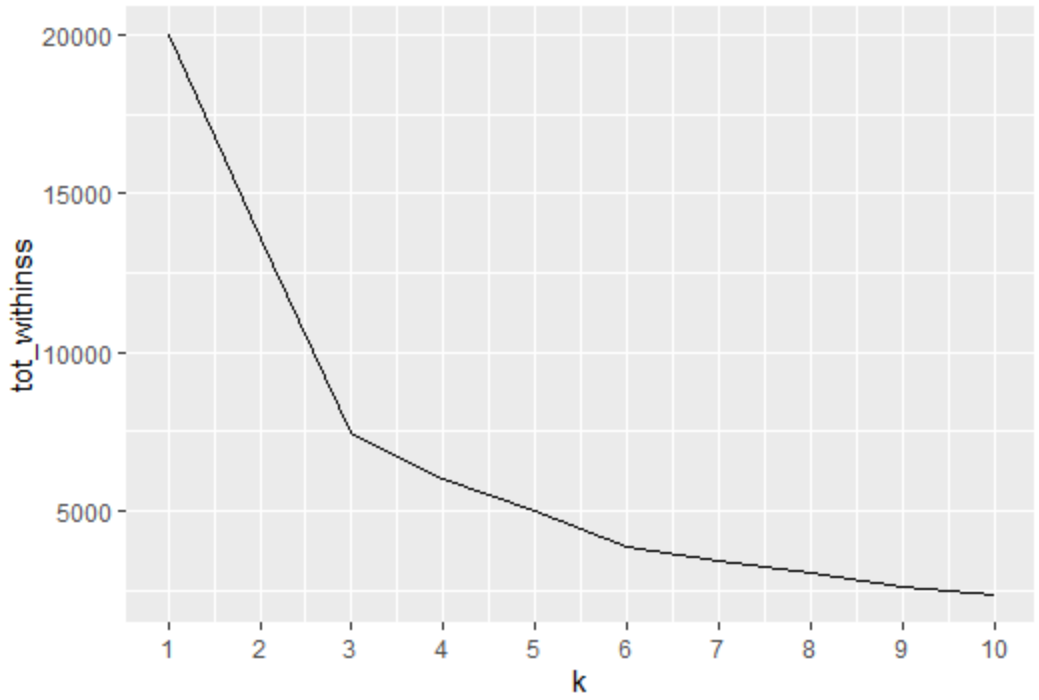
I believe that 3 clusters is the best going forward for this part of my analysis.

D2: Code Execution

```
model_km9 <- kmeans(ageinc, centers=9)

clust_km9 <- model_km9$cluster

ageinc_km9 <- mutate(ageinc, cluster=clust_km9)

head(ageinc_km9)

ggplot(ageinc_km9, aes(x=Age, y=Income, color=factor(cluster))) +
geom_point()


tot_withinss <- map_dbl(1:10, function(k){

  model <- kmeans(x=ageinc, centers=k)

  model$tot.withinss

})


elbow_df <- data.frame(k = 1:10, totwithinss = tot_withinss)

ggplot(elbow_df, aes(x=k, y=tot_withinss)) + geom_line() +
scale_x_continuous(breaks=1:10)


model_km3 <- kmeans(ageinc, centers=3)

clust_km3 <- model_km3$cluster

ageinc_km3 <- mutate(ageinc, cluster=clust_km3)

ggplot(ageinc_km3, aes(x=Age, y=Income, color=factor(cluster))) +
geom_point()
```

E1: Accuracy of Clustering Technique

My k-means technique was fairly accurate. My BSS/TSS ratio was 62.7% and with only three clusters I feel like that is very acceptable. As you can see from the graph below, the total within-cluster sum of squares starts to level off at k=3, which is a good indicator that 3 is the best amount of clusters.




```
K-means clustering with 3 clusters of sizes 4179, 1711, 4110

Cluster means:
         Age     Income
1 -0.8861190 -0.3373937
2  0.0314189  1.7054506
3  0.8879157 -0.3669241

Clustering vector:
   [1] 2 1 1 3 1 2 1 1 1 3 3 3 1 1 3 2 1 3 3 3 3 1 1 1 1 3 1 2 3 3 2 1 1 3 3 3 2 3 3 3 3 3 3 1 1 1 3 3 1 3 2 2 3 3 3 3 1 1 2 3 1
  [62] 3 3 2 2 3 3 1 3 2 1 1 3 1 1 1 1 2 2 2 1 3 3 2 1 1 2 1 2 1 3 1 1 1 3 3 3 3 3 1 1 3 2 3 1 2 1 2 3 1 1 1 1 3 3 1 1 3 1 1 3 3
 [123] 1 3 3 3 2 3 3 1 1 1 3 1 3 2 1 1 3 1 3 3 3 2 3 3 1 1 3 3 3 1 1 3 3 3 1 3 3 3 3 3 1 3 1 3 1 3 1 3 1 1 3 1 3 2 1 2 3 1 3 2 3 1 2 3 1
 [184] 1 3 3 1 2 3 2 2 3 3 2 3 3 1 3 1 2 1 3 1 3 1 1 2 1 1 2 1 3 3 1 1 3 3 3 3 3 2 1 3 2 3 3 1 3 3 3 3 1 1 3 3 3 2 1 1 3 3 2 3 3 1 2
 [245] 3 3 1 3 2 1 1 3 1 1 3 1 2 1 1 2 1 1 3 1 1 1 3 1 1 3 1 3 3 3 3 1 3 1 3 3 3 3 1 1 2 2 1 3 3 3 1 3 3 1 1 2 2 2 1 1 1 1 1 3 1 2 2
 [306] 1 3 3 3 2 2 1 1 1 1 3 1 1 2 1 1 3 2 1 1 1 1 2 1 3 1 2 1 1 2 3 3 1 2 1 1 1 1 1 1 2 3 1 2 1 3 1 1 3 3 3 3 3 3 3 3 1 2 3 3 2 3
 [367] 2 3 3 3 2 1 3 1 3 1 1 2 1 1 3 2 3 3 3 3 1 1 1 3 1 3 2 3 1 3 1 1 3 1 3 3 1 1 3 1 3 1 3 1 2 2 1 3 1 3 3 1 3 3 2 3 1 1 1 3 3 1 1 1
 [428] 3 1 1 3 3 1 3 2 3 3 1 3 3 3 1 2 1 1 1 2 1 3 1 3 3 1 3 1 3 3 1 1 3 1 3 1 3 3 2 2 1 1 1 1 3 1 3 1 1 1 1 1 1 3 1 3 1 1 3 3 3
 [489] 3 3 2 1 1 3 1 3 3 3 1 3 1 3 3 1 2 1 3 3 1 1 3 3 3 1 3 3 1 3 1 1 3 3 1 3 1 1 3 2 2 3 3 3 1 1 1 3 3 3 3 1 3 1 1 1 2 3 2 3 1
 [550] 3 3 3 3 3 3 2 3 1 1 2 1 3 3 1 1 1 3 1 1 3 1 3 2 3 3 3 1 2 1 1 3 1 2 3 3 3 3 2 2 3 2 2 3 3 1 1 2 1 1 3 1 3 1 1 3 3 3 2 3 2
 [611] 1 2 3 1 3 2 2 1 3 3 3 1 2 3 3 3 1 1 3 3 1 1 3 3 1 3 2 1 2 1 3 3 1 2 2 1 3 1 1 1 1 3 1 3 1 1 1 3 1 1 3 2 3 3 1 1 1 2 2 1 1
 [672] 3 3 2 1 2 3 3 3 3 1 1 3 1 1 3 3 3 1 3 1 3 1 1 3 3 1 1 3 3 3 2 3 2 2 1 3 1 3 1 1 1 2 3 3 1 3 3 1 3 1 3 2 3 1 3 1 1 1 3 2 3
 [733] 2 1 2 3 1 3 3 2 1 1 1 1 1 3 3 3 3 3 2 2 3 3 1 1 1 3 3 1 3 1 3 2 3 2 1 3 3 1 3 2 3 1 3 3 3 3 3 2 2 3 1 3 1 1 3 2 1 3 3 3 3
 [794] 1 3 1 3 1 1 3 1 3 2 3 1 1 1 1 1 1 3 3 3 1 2 2 3 3 3 2 1 1 2 3 1 1 1 2 1 1 2 1 1 3 2 3 1 1 3 1 1 2 1 3 1 1 3 2 3 1 3 3 1 2
 [855] 1 1 2 1 3 3 2 3 3 1 1 3 2 2 3 1 2 2 1 3 3 1 3 1 3 1 1 3 1 1 1 3 3 2 2 2 1 2 1 2 2 2 1 3 3 3 2 1 3 2 1 1 3 3 3 1 1 2 2 3 3
 [916] 3 3 2 2 1 3 3 3 2 1 2 3 2 3 3 3 1 3 2 1 2 3 3 3 3 3 1 1 3 1 1 3 1 1 3 1 3 3 2 3 3 3 2 3 1 3 3 2 2 3 1 2 3 1 3 1 1 2 3 1 2
 [977] 3 3 3 3 1 3 1 1 2 1 3 1 3 1 1 3 1 2 3 1 3 3 3 2
 [ reached getOption("max.print") -- omitted 9000 entries ]

Within cluster sum of squares by cluster:
[1] 2442.907 2680.727 2345.391
 (between_SS / total_SS =  62.7 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"         "iter"
[9] "ifault"
```

E2: Results and Implications

My k-means algorithm clustered the data into 3 groups. One group is people who have an income that is about one standard deviation higher than the median. Another group is people who have an income that is less than about one standard deviation higher than the median and an age that is less than the median. The last group is people who have an income that is less than about one standard deviation

higher than the median and an age that is more than the median. This will allow the organization to get a much better look at characteristics of patients based on age and income because there are only 3 groups, as opposed to a very high number of individual values.

E3: Limitation

One limitation of k-means clustering is that I have to choose the number of clusters. Although I used an elbow graph to determine which value of k was best, this value might not be what the organization is looking for. If the number of clusters the organization wants is different than the optimal number of clusters, our analysis might not be ideal.

E4: Course of Action

The next step I would recommend to the organization would be to look at readmission percent based on the three clusters that were created. I think there would be higher readmission in cluster 1 because that cluster is more likely to be able to afford medical bills compared to the other two clusters, but I wouldn't know for sure unless I analyzed that data.

F: Panopto Recording

Panopto recording is included in the submission and can also be found using the link below.

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=aa2016b0-5e63-44dc-b635-afb700e0aeb8

G: Sources for Third-Party Code

Kmeans: K-Means Clustering. RDocumentation. Retrieved February 26, 2023. https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/kmeans

H: Sources

Murphy, Phil. *Clustering Data in R*. RPubs by RStudio. April 25, 2021. Retrieved February 27, 2023. https://rpubs.com/pjmurphy/599072

Al-Masri, Anas. *How Does k-Means Clustering in Machine Learning Work?.* Towards Data Science. May 14, 2019. Retrieved February 26, 2023. https://towardsdatascience.com/how-does-k-means-clustering-in-machine-learning-work-fdaaaf5acfa0

Ismkhan, Hassan, *I-k-means−+: An iterative clustering algorithm based on an enhanced version of the k-means*, Pattern Recognition, Volume 79, 2018, Pages 402-413.