ANA 515 Assignment 4 CK

Clarentia Kurnia

4/27/2022

## 1. Discuss the business problem/goal

This data analytics project aims to segment mall customers based on their various characteristics, such as age, gender and interest. Customer segmentation is an important technique that helps businesses personalize their marketing strategies. Customer segmentation divides customers into groups that share similar characteristics. We will be using the Mall Customers dataset that has customers’ gender, customer ID, age, annual income, and spending score.

## 2. Identify where the dataset was retrieved from

The Mall Customers dataset was retrieved from Kaggle, an online data science and machine learning community. It was uploaded to the website by a user 4 years ago and has a public domain license.

## 3. Identify the code that imported and saved your dataset in R

library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──

## ✓ ggplot2 3.3.5 ✓ purrr 0.3.4  
## ✓ tibble 3.1.3 ✓ dplyr 1.0.7  
## ✓ tidyr 1.2.0 ✓ stringr 1.4.0  
## ✓ readr 2.1.2 ✓ forcats 0.5.1

## Warning: package 'tidyr' was built under R version 4.1.2

## Warning: package 'readr' was built under R version 4.1.2

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(readr)  
file <- "Mall\_Customers.csv"  
mall <- read\_csv(file)

## Rows: 200 Columns: 5

## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (2): CustomerID, Genre  
## dbl (3): Age, Annual Income (k$), Spending Score (1-100)

##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

I installed the libraries tidyverse and readr to read and save the data in R. Then I assigned a named dataframe (mall) to the data.

## 4. Describe your data set (using the common attributes such as #rows, #columns, variable names, types, means, SD, min/max, NAs, etc…)

a <- nrow(mall)  
b <- ncol(mall)

This dataframe has 200 rows and 5 columns. The names of the columns and a brief description of each are in the table below:

|  |  |
| --- | --- |
| Column Name | Description |
| CustomerID | Unique customer identifier |
| Genre | Each customer gender |
| Age | Number of measures of spirit consumed |
| Annual Income (k$) | Number of glasses of wine consumed |
| Spending Score (1-100) | Number of litres of hard liquor consumed |

summary(mall)

## CustomerID Genre Age Annual Income (k$)  
## Length:200 Length:200 Min. :18.00 Min. : 15.00   
## Class :character Class :character 1st Qu.:28.75 1st Qu.: 41.50   
## Mode :character Mode :character Median :36.00 Median : 61.50   
## Mean :38.85 Mean : 60.56   
## 3rd Qu.:49.00 3rd Qu.: 78.00   
## Max. :70.00 Max. :137.00   
## Spending Score (1-100)  
## Min. : 1.00   
## 1st Qu.:34.75   
## Median :50.00   
## Mean :50.20   
## 3rd Qu.:73.00   
## Max. :99.00

## 5. Discuss any data preparation, missing values and errors (if the dataset was clean and there is no prep in the code, include a comment that explains what likely data preparation was done. What are the common issues with raw data?)

The mall dataset is clean and there is no prep needed on my end. Some potential data preparation that was done: Padding a “0” at the beginning of the Customer IDs so they can be easily sorted and have the same length of numbers on every row. Changing the annual income from full amounts to thousand of dollars (k$) Adjusting the spending score to only fall within the range of 1-100. They may have calcultated this from a collection of surveys, interviews, or other data collection methods.

## 6. Discuss the modeling

The modeling we are doing is hierarchical clustering, also known as hierarchical cluster analysis, an algorithm that groups similar items into clusters. The result is clusters which are distinct from one another, while each cluster has similarities within.

The modeling involves using a dendrogram to find the optimal number of clusters, fitting hierarchical clustering to the dataset, and finally visualizing the clusters.

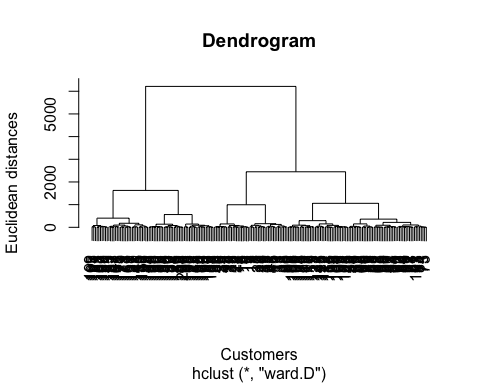
## 7. Produce and discuss the output

#Using a dendrogram to find the optimal number of clusters

dendrogram = hclust(d = dist(mall, method = 'euclidean'), method = 'ward.D')

## Warning in dist(mall, method = "euclidean"): NAs introduced by coercion

plot(dendrogram,  
 main = paste('Dendrogram'),  
 xlab = 'Customers',  
 ylab = 'Euclidean distances')



From this dendrogram, we want to locate the largest vertical difference between nodes, and in the middle form a horizontal line. The number of vertical lines intersecting it is the optimal number of clusters.

#Fitting hierarchical clustering to the dataset

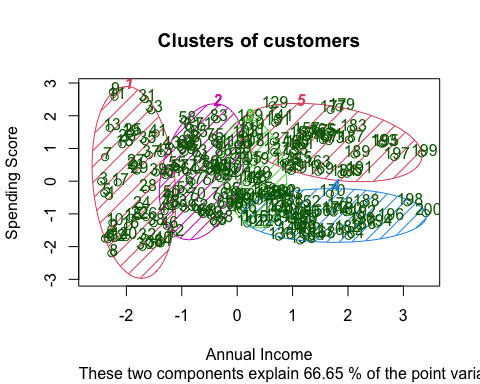
hc = hclust(d = dist(mall, method = 'euclidean'), method = 'ward.D')

## Warning in dist(mall, method = "euclidean"): NAs introduced by coercion

y\_hc = cutree(hc, 5)

## 8. Provide explanation with any visuals

library(cluster)  
clusplot(mall,  
 y\_hc,  
 lines = 0,  
 shade = TRUE,  
 color = TRUE,  
 labels= 2,  
 plotchar = FALSE,  
 span = TRUE,  
 main = paste('Clusters of customers'),  
 xlab = 'Annual Income',  
 ylab = 'Spending Score')



This visualization displays clusters of customers based on their annual income and spending score. As we can see, there are 5 clusters of customers. Each cluster is positioned differently on the diagram. Cluster 1 consists of customers with lower annual income with various spending score. These are mostly the lower numbers on the variable CustomerID. Cluster 2 customers have slightly higher annual income and slightly smaller range of spending score - more are on the higher side. Cluster 3 is like the median, they are in the middle in terms of annual income and spending score. Cluster 4 customers earn more annually but have low spending score. Cluster 5 customers earn more annually and have high spending score.

## 9. Your document should clearly and concisely communicate the project

In conclusion, this data clustering project aims to place customers into groups, so the stores and businesses in the mall can target their goods and/or services accordingly. As we know, it is a lot more efficient and profitable to target your customers rather than to satisfy everyone’s needs. By having these clusters of customers visualized, the mall can differentiate their marketing and advertising strategies depending on the type of customers they have.

This was a fun project to analyze and it was very interesting to run the codes for creating the dendrogram and the cluster!