R version 4.1.0 (2021-05-18) -- "Camp Pontanezen"

Copyright (C) 2021 The R Foundation for Statistical Computing

Platform: x86\_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.

You are welcome to redistribute it under certain conditions.

Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.

Type 'contributors()' for more information and

'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or

'help.start()' for an HTML browser interface to help.

Type 'q()' to quit R.

> library(tidyverse)

-- Attaching packages -------------------------------------------------------------------------------- tidyverse 1.3.1 --

v ggplot2 3.3.4 v purrr 0.3.4

v tibble 3.1.2 v dplyr 1.0.6

v tidyr 1.1.3 v stringr 1.4.0

v readr 1.4.0 v forcats 0.5.1

-- Conflicts ----------------------------------------------------------------------------------- tidyverse\_conflicts() --

x dplyr::filter() masks stats::filter()

x dplyr::lag() masks stats::lag()

> library(readr)

> Ecommerce <- read\_csv("Ecommerce.csv")

-- Column specification -------------------------------------------------------------------------------------------------

cols(

InvoiceNo = col\_character(),

StockCode = col\_character(),

Description = col\_character(),

Quantity = col\_double(),

InvoiceDate = col\_character(),

UnitPrice = col\_double(),

CustomerID = col\_double(),

Country = col\_character(),

X9 = col\_logical()

)

Warning message:

Missing column names filled in: 'X9' [9]

> View(Ecommerce)

> Ecommerce <- Ecommerce %>% mutate(TotalPrice = Quantity \* UnitPrice)

> Ecommerce$CustomerID <- as.character(Ecommerce$CustomerID)

> na\_cust <- Ecommerce %>% filter(is.na(CustomerID))

> Ecommerce <- anti\_join(Ecommerce, na\_cust)

Joining, by = c("InvoiceNo", "StockCode", "Description", "Quantity", "InvoiceDate", "UnitPrice", "CustomerID", "Country", "X9", "TotalPrice")

> # Create a new df

> ecommerce\_sales <- Ecommerce %>% filter(TotalPrice > 0) %>% group\_by(CustomerID) %>%

+ summarize(Sales=sum(TotalPrice)) %>% arrange(CustomerID)

> # K-Means Clustering

> KMC <- kmeans(as.matrix(ec\_sales), centers = 4, iter.max = 1000)

Error in as.matrix(ec\_sales) : object 'ec\_sales' not found

> # K-Means Clustering

> KMC <- kmeans(as.matrix(ecommerce\_sales), centers = 4, iter.max = 1000)

> KC1 <- subset(ec\_sales, KMC$cluster==1)

Error in subset(ec\_sales, KMC$cluster == 1) : object 'ec\_sales' not found

>

> KC1 <- subset(ecommerce\_sales, KMC$cluster==1)

>

> mean(kc1$sales)

Error in mean(kc1$sales) : object 'kc1' not found

> mean(kC1$sales)

Error in mean(kC1$sales) : object 'kC1' not found

> mean(KC1$sales)

[1] NA

Warning messages:

1: Unknown or uninitialised column: `sales`.

2: In mean.default(KC1$sales) :

argument is not numeric or logical: returning NA

> mean(KC1$Sales)

[1] 209342.3

> KC2 <- subset(ecommerce\_sales, KMC$cluster==2)

>

> KC3 <- subset(ecommerce\_sales, KMC$cluster==3)

>

> KC4 <- subset(ecommerce\_sales, KMC$cluster==4)

>

> mean(KC1$Sales)

[1] 209342.3

> mean(KC2$Sales)

[1] 862.5538

> mean(KC3$Sales)

[1] 2135.438

> mean(KC4$Sales)

[1] 53366.49

> # Hierarchical Clustering

> distance <- dist(ecommerce\_sales, method='euclidean')

> clusterIntensity <- hclust(distance, method="ward.D")

> plot(clusterIntensity)

> salesClusters <- cutree(clusterIntensity, k = 4)#4 clusters

>

> table(salesClusters)

salesClusters

1 2 3 4

38 390 2439 1471

> tapply(ecommerce\_sales$Sales, salesClusters, mean)

1 2 3 4

71768.1384 6934.7058 821.0115 1004.2396

> #so we have 38 high value customers and hierchical clustering is best clustering algo

> #K-Means is able to pick up the biggest spending customers ($209, 342.3); H-Cluster is unable to do so hence overall Kmean is better

> ecommerce\_sales<-cbind(ecommerce\_sales,Clusternum=KMC$cluster)

> View(ecommerce\_sales)

> table(KMC)

Error in table(KMC) : all arguments must have the same length

> summary(KMC)

Length Class Mode

cluster 4338 -none- numeric

centers 8 -none- numeric

totss 1 -none- numeric

withinss 4 -none- numeric

tot.withinss 1 -none- numeric

betweenss 1 -none- numeric

size 4 -none- numeric

iter 1 -none- numeric

ifault 1 -none- numeric

> sum(KC1)

Error in FUN(X[[i]], ...) :

only defined on a data frame with all numeric-alike variables

> table(ecommerce\_sales$Clusternum)

1 2 3 4

5 2298 2005 30

> #hence we have 5 really high value customers with average spending of $209, 342.3