**Clustering-20237**

Jitesh Rawat

4/9/2021

**library**(palmerpenguins)  
data <-penguins  
**head**(data)

## # A tibble: 6 x 8  
## species island bill\_length\_mm bill\_depth\_mm flipper\_length\_~ body\_mass\_g sex   
## <fct> <fct> <dbl> <dbl> <int> <int> <fct>  
## 1 Adelie Torge~ 39.1 18.7 181 3750 male   
## 2 Adelie Torge~ 39.5 17.4 186 3800 fema~  
## 3 Adelie Torge~ 40.3 18 195 3250 fema~  
## 4 Adelie Torge~ NA NA NA NA <NA>   
## 5 Adelie Torge~ 36.7 19.3 193 3450 fema~  
## 6 Adelie Torge~ 39.3 20.6 190 3650 male   
## # ... with 1 more variable: year <int>

**set.seed**(123)  
data = **na.omit**(data)  
data =data[,**-c**(1,2,7,8)]*#Remove the char and categorical data*  
**head**(data)

## # A tibble: 6 x 4  
## bill\_length\_mm bill\_depth\_mm flipper\_length\_mm body\_mass\_g  
## <dbl> <dbl> <int> <int>  
## 1 39.1 18.7 181 3750  
## 2 39.5 17.4 186 3800  
## 3 40.3 18 195 3250  
## 4 36.7 19.3 193 3450  
## 5 39.3 20.6 190 3650  
## 6 38.9 17.8 181 3625

**dim**(data)

## [1] 333 4

wss = (**nrow**(data)**-**1)**\*sum**(**apply**(data,2,var))  
**for** (i **in** 2**:**5)  
 {wss[i] = **sum**(**kmeans**(data,centers = i)**$**withinss)}  
**plot**(1**:**5, wss, type="b", xlab="Number of Clusters",  
 ylab="Within groups sum of squares")



kmeans3 <- **kmeans**(data, 3)  
kmeans3

## K-means clustering with 3 clusters of sizes 80, 140, 113  
##   
## Cluster means:  
## bill\_length\_mm bill\_depth\_mm flipper\_length\_mm body\_mass\_g  
## 1 48.66250 15.39750 219.9875 5365.938  
## 2 41.12214 17.94643 189.6286 3461.250  
## 3 44.24336 17.44779 201.5487 4310.619  
##   
## Clustering vector:  
## [1] 2 2 2 2 2 2 3 2 2 3 2 2 3 2 3 2 2 2 3 2 2 2 2 2 3 2 3 2 3 2 3 3 2 2 3 2 3  
## [38] 2 3 2 3 2 2 3 2 3 2 3 2 2 2 2 2 2 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3  
## [75] 2 3 2 3 2 2 2 2 3 2 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 2 2 3 2 3 2 3 2 3 3 3 2  
## [112] 2 2 2 2 2 2 2 2 3 2 3 2 3 2 2 2 3 2 3 2 3 2 3 2 2 2 2 2 2 3 2 2 2 2 3 3 1  
## [149] 3 1 1 3 3 1 3 1 3 1 3 1 3 1 3 1 3 1 1 1 3 1 1 1 1 3 1 1 3 1 1 1 1 1 1 3 1  
## [186] 3 1 3 3 1 1 3 1 1 1 1 1 3 1 1 1 3 1 3 1 3 1 3 1 3 1 1 3 1 3 1 1 1 3 1 3 1  
## [223] 3 1 3 1 3 1 3 1 3 1 1 1 1 1 3 1 1 1 1 1 3 1 1 1 1 1 1 3 1 3 1 1 1 3 1 3 1  
## [260] 1 1 1 1 1 1 2 3 2 2 2 3 2 2 3 2 2 2 2 3 2 3 2 2 2 3 2 2 2 2 2 3 2 2 2 3 2  
## [297] 3 2 3 2 3 2 3 2 3 3 2 2 2 2 3 2 3 2 2 2 3 2 3 2 2 2 3 2 2 3 2 2 3 2 2 3 2  
##   
## Within cluster sum of squares by cluster:  
## [1] 9718829 9724809 9318036  
## (between\_SS / total\_SS = 86.6 %)  
##   
## Available components:  
##   
## [1] "cluster" "centers" "totss" "withinss" "tot.withinss"  
## [6] "betweenss" "size" "iter" "ifault"

kmeans2 <- **kmeans**(data, 2)  
kmeans2

## K-means clustering with 2 clusters of sizes 203, 130  
##   
## Cluster means:  
## bill\_length\_mm bill\_depth\_mm flipper\_length\_mm body\_mass\_g  
## 1 41.93300 18.08719 192.1281 3655.665  
## 2 47.20923 15.72462 214.7692 5068.077  
##   
## Clustering vector:  
## [1] 1 1 1 1 1 1 2 1 1 2 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1  
## [38] 1 2 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1  
## [75] 1 2 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 2 1 2 1 1 1 1 1  
## [112] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2  
## [149] 2 2 2 2 2 2 2 2 2 2 2 2 1 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2  
## [186] 1 2 1 2 2 2 1 2 2 2 2 2 2 2 2 2 1 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2  
## [223] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2  
## [260] 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 2 1  
## [297] 1 1 1 1 1 1 2 1 2 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
##   
## Within cluster sum of squares by cluster:  
## [1] 28111986 29085930  
## (between\_SS / total\_SS = 73.4 %)  
##   
## Available components:  
##   
## [1] "cluster" "centers" "totss" "withinss" "tot.withinss"  
## [6] "betweenss" "size" "iter" "ifault"

testsp <- penguins  
**head**(testsp)

## # A tibble: 6 x 8  
## species island bill\_length\_mm bill\_depth\_mm flipper\_length\_~ body\_mass\_g sex   
## <fct> <fct> <dbl> <dbl> <int> <int> <fct>  
## 1 Adelie Torge~ 39.1 18.7 181 3750 male   
## 2 Adelie Torge~ 39.5 17.4 186 3800 fema~  
## 3 Adelie Torge~ 40.3 18 195 3250 fema~  
## 4 Adelie Torge~ NA NA NA NA <NA>   
## 5 Adelie Torge~ 36.7 19.3 193 3450 fema~  
## 6 Adelie Torge~ 39.3 20.6 190 3650 male   
## # ... with 1 more variable: year <int>

testsp <- **na.omit**(testsp)  
**dim**(testsp)

## [1] 333 8

**length**(kmeans3**$**cluster)

## [1] 333

*#the length of kmeans clusters list should be same as species list observations*

**table**(testsp**$**species, kmeans3**$**cluster)

##   
## 1 2 3  
## Adelie 0 94 52  
## Chinstrap 0 46 22  
## Gentoo 80 0 39

**table**(testsp**$**species, kmeans2**$**cluster)

##   
## 1 2  
## Adelie 132 14  
## Chinstrap 63 5  
## Gentoo 8 111

**plot** (data , col =(kmeans3**$**cluster **+**1) , main =" K - Means Clustering  
Results with K = 3" ,pch =20 , cex =2)



idx <- **sample**(1**:dim**(data)[1], 25)  
penguinsSample <- data[idx,]  
penguinsSample**$**species <- NULL

d <- **dist**(penguinsSample, method = "euclidean") *# distance matrix*  
fit <- **hclust**(d, method="ward")

## The "ward" method has been renamed to "ward.D"; note new "ward.D2"

**plot**(fit, hang = -1, labels= penguins**$**species[idx]) *#Just plotting it*  
**rect.hclust**(fit, k=3)



hc.complete = **hclust**(**dist**(penguinsSample), method ="complete")  
hc.average = **hclust**(**dist**(penguinsSample), method ="average")  
hc.single = **hclust**(**dist**(penguinsSample), method ="single")

**par**(mfrow = **c**(1,3))*#To print all the plots in to one*  
**plot**(hc.complete, main =" Complete Linkage " , cex =.9,labels= penguins**$**species[idx])  
**rect.hclust**(hc.complete, k=3)  
groups <- **cutree**(hc.complete, k=3)  
  
**plot**(hc.average, main =" Average Linkage " , cex =.9,labels=penguins**$**species[idx])  
**rect.hclust**(hc.average, k=3)  
groups <- **cutree**(hc.average, k=3)  
  
**plot**(hc.single, main ="Single Linkage " , cex =.9,labels=penguins**$**species[idx])  
**rect.hclust**(hc.single, k=3)

