**ATIVIDADE INDIVIDUAL 3**

**Disciplina: Métodos Matriciais e Análise de Clusters.**

**Professor responsável: Rodrigo Togneri.**

**Versão: 2019.10**

**Sempre que aplicável, utilize software para fazer os cálculos e coloque aqui somente os resultados finais, entregando a solução por software em arquivo à parte.**

|  |  |
| --- | --- |
| **Matrícula** | **Nome Completo** |
| A57622988 | Leandro de Oliveira Daniel |

**Tema: Clusters.**

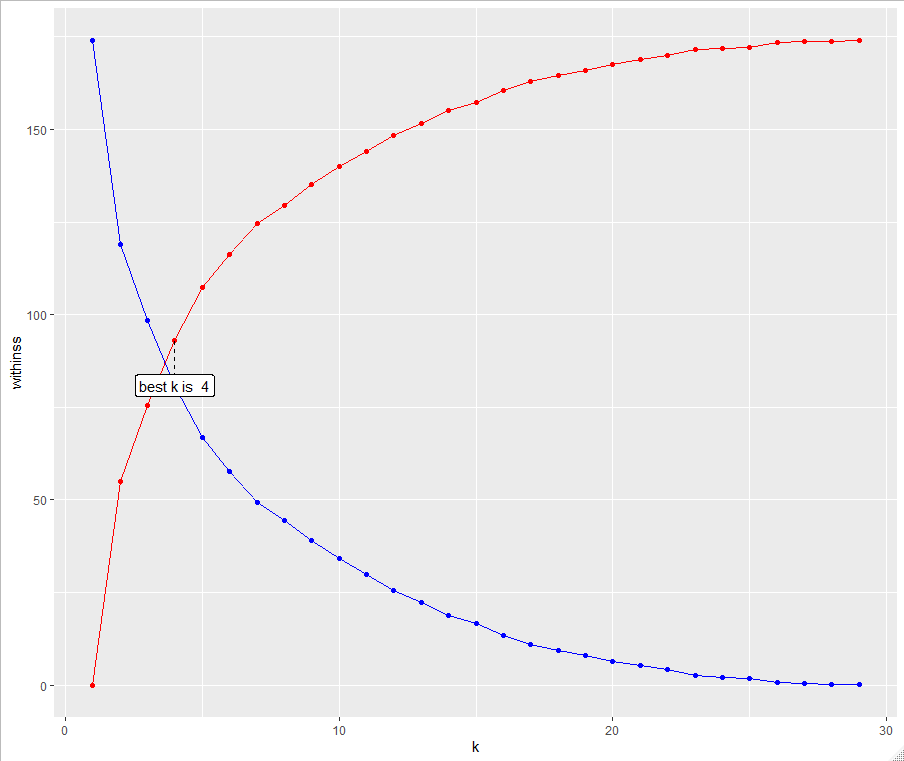
Utilize a base de dados ATIBAIA e os procedimentos de K-Means descritos no arquivo AnClusters\_primeirosPassos.rmd, e:

1. **Crie um programa que, além de realizar o pré-processamento de dados corretamente, forneça os gráficos de withinss e betweenss variáveis segundo K.**

|  |
| --- |
| # loading required libraries --------------------------------------------------  #libraries for file and data manipulation  library(readxl)  library(dplyr)  # cluster analysis  library(fpc)  library(factoextra)  library(cluster)  library(ggcorrplot)  # data injestion --------------------------------------------------------------  ATIBAIA <- read\_xlsx("./data/raw/ATIBAIA.xlsx",  sheet = "ATIBAIA",  col\_names = TRUE)  # checking the class for each column and get a summary  sapply(ATIBAIA, class)  summary(ATIBAIA)  # dataset preparation ---------------------------------------------------------  # changing vars biling, estac and ti from numeric to factor  ATIBAIA$biling = as.factor(ATIBAIA$biling)  ATIBAIA$estac = as.factor(ATIBAIA$estac)  ATIBAIA$ti = as.factor(ATIBAIA$ti)  # getting only meaningful vars as clustering drivers:  ATIBAIA\_drivers = ATIBAIA[, -c(1,2,9)]  # transforming and scaling relevant vars to get the final dataset  ATIBAIA\_drivers\_num <- ATIBAIA\_drivers  ATIBAIA\_drivers\_num$biling = as.numeric(ATIBAIA\_drivers\_num$biling)  ATIBAIA\_drivers\_num$estac = as.numeric(ATIBAIA\_drivers\_num$estac)  ATIBAIA\_drivers\_num$ti = as.numeric(ATIBAIA\_drivers\_num$ti)  ATIBAIA\_drivers\_num\_z <- as.data.frame(lapply(ATIBAIA\_drivers\_num, scale))  # calculating vars correlation  ggcorrplot(cor(ATIBAIA\_drivers\_num\_z),  hc.order = TRUE,  lab = TRUE,  lab\_size = 3,  method = "square",  colors = c("tomato2", "white", "springgreen3"),  title = "Correlation Matrix")  # withinss and betweenss according to its K ----------------------------------  max\_number\_of\_clusters <- nrow(ATIBAIA\_drivers\_num\_z) - 1  results <- tibble(k = numeric(),  withinss = numeric(),  betweenss = numeric(),  best\_fit = numeric())  for (count in seq(1, max\_number\_of\_clusters, by = 1)) {  number\_of\_clusters <- as.numeric(count)  KMeans\_clustering <- kmeans(ATIBAIA\_drivers\_num\_z, number\_of\_clusters, nstart = 20)  results <- add\_row(results,  k = number\_of\_clusters,  withinss = sum(KMeans\_clustering$withinss),  betweenss = KMeans\_clustering$betweenss,  best\_fit = abs(sum(KMeans\_clustering$withinss) - KMeans\_clustering$betweenss))  }  ggplot(data = results, aes(x = k)) +  geom\_point(aes(y = withinss), color = 'blue') +  geom\_point(aes(y = betweenss), color = 'red') +  geom\_line(aes(y = withinss), color = 'blue') +  geom\_line(aes(y = betweenss), color = 'red') +  geom\_segment(aes(x = which.min(results$best\_fit),  y = results$withinss[which.min(results$best\_fit)],  xend = which.min(results$best\_fit),  yend = results$betweenss[which.min(results$best\_fit)]),  linetype = "dashed", color = "black") +  geom\_label(aes(x = which.min(results$best\_fit),  y = results$withinss[which.min(results$best\_fit)],  label = paste('best k is ', which.min(results$best\_fit))))  # choosing the K that best fits -----------------------------------------------  KMeans\_clustering <- kmeans(ATIBAIA\_drivers\_num\_z, which.min(results$best\_fit), nstart = 20)  # Visualizing clusters  plotcluster(ATIBAIA\_drivers\_num\_z, KMeans\_clustering$cluster)  fviz\_cluster(list(data = ATIBAIA\_drivers\_num\_z,  cluster = KMeans\_clustering$cluster),  show.clust.cent = T) |

1. **Com base no resultado de a), faça a escolha do melhor K e justifique sua escolha.**

Com base nos resultados do programa em R descrito no item a), foi considerado como parâmetro de escolha o número de cluster onde tivemos, simultaneamente, um bom balanço entre withinss e betweenss. No caso, com 4 clusters obtivemos o melhor resultado.



1. **As métricas coletadas indicam que os clusters encontrados são adequados? Justifique.**

Observando o gráfico com os 4 clusters gerados, conforme explicado no item b), vemos uma classificação adequada, dado o dataset proposto.

