PACDII: Quiz IIVI

ID Grupo e nome do representante

2023-10-12

# Remover tudo!  
  
# Incluir as libraries de que necessita  
  
library(lubridate)

##   
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':  
##   
## date, intersect, setdiff, union

library(dbscan)

##   
## Attaching package: 'dbscan'

## The following object is masked from 'package:stats':  
##   
## as.dendrogram

library(factoextra)

## Loading required package: ggplot2

## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa

library(RColorBrewer)  
library(lsr)

# IV.1) [4 valores] Leia “Acidentes.csv” e obtenha um resumo (summary) adequado para todas as variáveis do ficheiro (após corrigir Longitude.GPS e Latitude.GPS (\*) e formatar corretamente acidentes$Datahora). No Quiz faça upload – em formato pdf - do referido sumário.

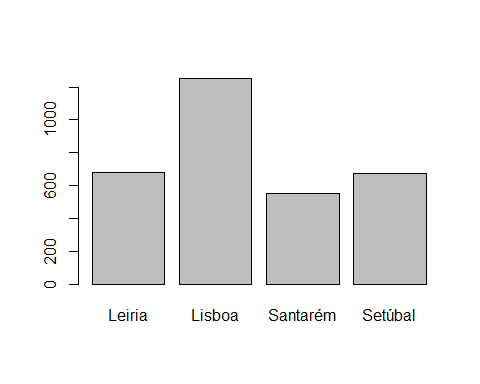
(\*) CONSIDERE QUE: Portugal Continental tem latitude mínima de 37N e máxima de 42N Portugal Continental tem longitude mínima 10W (-10) e máxima de 6W (-6)

df<- read.csv("acidentes.csv")  
df<-subset(df, Latitude.GPS >= 37 & Latitude.GPS <= 42 & Longitude.GPS >= -10 & Longitude.GPS <= -6)  
df$Datahora<- ymd\_hms(df$Datahora)  
  
summary(df)

## Id..Acidente Datahora Entidades.Fiscalizadoras  
## Min. :2.010e+09 Min. :2010-01-01 00:30:00.0 Length:9135   
## 1st Qu.:2.010e+09 1st Qu.:2010-04-06 12:32:30.0 Class :character   
## Median :2.010e+09 Median :2010-07-13 17:00:00.0 Mode :character   
## Mean :2.011e+09 Mean :2010-07-07 23:51:28.2   
## 3rd Qu.:2.010e+09 3rd Qu.:2010-10-08 12:50:00.0   
## Max. :2.012e+09 Max. :2010-12-31 21:40:00.0   
##   
## Velocidade.local Velocidade.geral Dia.da.Semana Latitude.GPS   
## Min. : 10.0 Min. : 30.00 Length:9135 Min. :37.02   
## 1st Qu.: 50.0 1st Qu.: 50.00 Class :character 1st Qu.:38.83   
## Median : 90.0 Median : 90.00 Mode :character Median :40.13   
## Mean : 79.2 Mean : 85.46 Mean :39.94   
## 3rd Qu.:100.0 3rd Qu.:120.00 3rd Qu.:41.20   
## Max. :120.0 Max. :120.00 Max. :42.00   
## NA's :282 NA's :13   
## Longitude.GPS Num..Mortos.a.30.dias Num..Feridos.graves.a.30.dias  
## Min. :-9.474 Min. :0.00000 Min. : 0.00000   
## 1st Qu.:-8.828 1st Qu.:0.00000 1st Qu.: 0.00000   
## Median :-8.547 Median :0.00000 Median : 0.00000   
## Mean :-8.483 Mean :0.04696 Mean : 0.09294   
## 3rd Qu.:-8.229 3rd Qu.:0.00000 3rd Qu.: 0.00000   
## Max. :-6.285 Max. :5.00000 Max. :12.00000   
##   
## Num..Feridos.ligeiros.a.30.dias Características.Tecnicas1 Cond.Aderência   
## Min. : 0.000 Length:9135 Length:9135   
## 1st Qu.: 1.000 Class :character Class :character   
## Median : 1.000 Mode :character Mode :character   
## Mean : 1.349   
## 3rd Qu.: 2.000   
## Max. :36.000   
##   
## Distrito Concelho Freguesia Pov..Proxima   
## Length:9135 Length:9135 Length:9135 Length:9135   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## Nome.arruamento Tipos.Vias Cod.Via Estado.Conservação  
## Length:9135 Length:9135 Length:9135 Length:9135   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## Km Factores.Atmosféricos Reg.Circulação1 Intersecção.Vias   
## Min. : 0.00 Length:9135 Length:9135 Length:9135   
## 1st Qu.: 13.44 Class :character Class :character Class :character   
## Median : 37.55 Mode :character Mode :character Mode :character   
## Mean : 67.48   
## 3rd Qu.: 81.16   
## Max. :737.00   
## NA's :159   
## Localizações Luminosidade Marca.Via Natureza   
## Length:9135 Length:9135 Length:9135 Length:9135   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## Obras.Arte Obstáculos Sentidos Sinais   
## Length:9135 Length:9135 Length:9135 Length:9135   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## Sinais.Luminosos Tipo.Piso Traçado.1 Traçado.2   
## Length:9135 Length:9135 Length:9135 Length:9135   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## Traçado.3 Traçado.4 Via.Trânsito   
## Length:9135 Length:9135 Length:9135   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
##

# IV.2) [3.5 valores] Selecione apenas as variáveis 2 a 14, restrinja o conjunto de dados ao Distrito “Leiria”, “Lisboa”, “Santarém” e “Setúbal”, e, seguidamente, descarte todos os dados omissos.Obtenha uma tabela de frequências absolutas da variável Distrito e, no Quiz, faça upload – em formato pdf - da mesma tabela.

df<-df[,2:14]  
df <- subset(df, Distrito %in% c("Leiria", "Lisboa", "Santarém", "Setúbal"))  
df<- na.omit(df)  
  
contagem<-table(df$Distrito)  
barplot(contagem)



table(df$Distrito)

##   
## Leiria Lisboa Santarém Setúbal   
## 679 1251 552 672

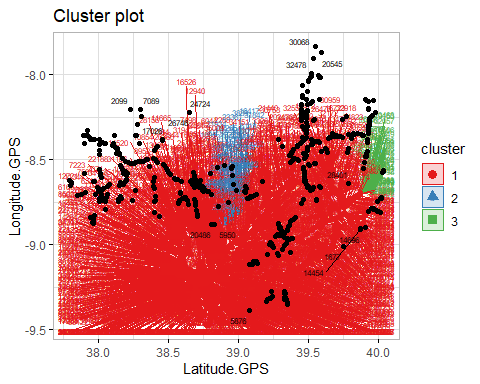
# IV.3) [6 valores] Efectue o agrupamento dos dados obtidos em IV.2) com o algoritmo DBSCAN, baseado na Latitude.GPS e Longitude.GPS, utilizando eps=0.07 e minPts=40. Apresente um mapa dos clusters obtidos.No Quiz faça upload, em formato pdf, deste mapa.

(db<-dbscan(df[,6:7],eps=0.07,minPts = 40))

## DBSCAN clustering for 3154 objects.  
## Parameters: eps = 0.07, minPts = 40  
## Using euclidean distances and borderpoints = TRUE  
## The clustering contains 3 cluster(s) and 622 noise points.  
##   
## 0 1 2 3   
## 622 2402 73 57   
##   
## Available fields: cluster, eps, minPts, dist, borderPoints

fviz\_cluster(db,df[,6:7],stand = F,  
ellipse = T,  
geom = c("point", "text"),  
labelsize = 6,  
repel = TRUE,   
palette="Set1",  
ggtheme = theme\_light())

## Warning: ggrepel: 607 unlabeled data points (too many overlaps). Consider  
## increasing max.overlaps



# IV.4) [6.5 valores] Analise as associações entre as variáveis nominais e métricas disponíveis e os clusters obtidos; use o Distrito para traçar o perfil dos clusters obtendo uma tabela cruzada; determine em que cluster se situam as localidades de Abrantes e Óbidos. Na sequência destas análises complete as frases que se apresentam no Quiz.

Eta<-function(y,x){  
 freqk<-as.vector(table(x))  
 l<-nlevels(x)  
 m<-rep(NA, l)  
 qual<-as.numeric(x)  
 for (k in 1:l) {m[k]<-mean(y[qual == k])}  
   
 return(sqrt(sum(freqk\*(m-mean(y))^2)/sum((y-mean(y))^2)))}

clusters <- db$cluster  
cramersV(df$Entidades.Fiscalizadoras, clusters)

## [1] 0.1905418

cramersV(df$Dia.da.Semana, clusters)

## [1] 0.04420985

cramersV(df$Características.Tecnicas1, clusters)

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## [1] 0.136834

cramersV(df$Cond.Aderência, clusters)

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## [1] 0.0646877

cramersV(df$Distrito, clusters) # alguma correlação

## [1] 0.3534314

cramersV(df$Datahora, clusters) # muita correlação

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## [1] 0.9862514

Eta(df$Velocidade.geral, clusters)

## [1] 0.01962739

Eta(df$Velocidade.local, clusters)

## [1] 0.006719158

Eta(df$Latitude.GPS, clusters)

## [1] 0.04113599

Eta(df$Longitude.GPS, clusters) # alguma correlação

## [1] 0.3647136

Eta(df$Num..Feridos.graves.a.30.dias, clusters)

## [1] 0.04039417

Eta(df$Num..Feridos.ligeiros.a.30.dias, clusters)

## [1] 0.006208389

table(df$Distrito, clusters)

## clusters  
## 0 1 2 3  
## Leiria 115 507 0 57  
## Lisboa 24 1227 0 0  
## Santarém 244 235 73 0  
## Setúbal 239 433 0 0

prop.table(table(df$Distrito, clusters), 2)

## clusters  
## 0 1 2 3  
## Leiria 0.18488746 0.21107410 0.00000000 1.00000000  
## Lisboa 0.03858521 0.51082431 0.00000000 0.00000000  
## Santarém 0.39228296 0.09783514 1.00000000 0.00000000  
## Setúbal 0.38424437 0.18026644 0.00000000 0.00000000

As variáveis que mais influenciam os clusters são o “Distrito”, a “Datahora” e a “Longitude.GPS”.

Ao analisar a tabela de frequências acima, concluimos que Abrantes ficaria no cluster 2 enquanto que Óbidos ficaria no cluster 3.

O cluster 2 e 3 parecem representar acidentes no distríto de Santarém e Leiria respetivamente. O cluster 0 parece ter acidentes mais para fora de Lisboa, enquanto que o 1 parece ter foco em Lisboa e nos distritos do litorais do país (que não é o caso de Santarém porque não se situa perto da costa).