PACDII: Quiz IV

7 - João Francisco Botas

14 de outubro de 2023

Elementos do grupo:

- Allan Kardec da Silva Rodrigues, nº 103380
- André Plancha Fernandes, nº 105289
- Diogo Alexandre Alonso de Freitas, nº 104841
- João Francisco Marques Gonçalves da Silva Botas, nº 104782
- Marco Delgado Esperança, nº 110451


```
# Remover tudo!
rm(list = ls())
# Incluir as libraries de que necessita
library(tidyverse)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.1
```

```
## Warning: package 'readr' was built under R version 4.3.1
```

```
## Warning: package 'purrr' was built under R version 4.3.1
```

```
## Warning: package 'dplyr' was built under R version 4.3.1
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.3     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ ggplot2   3.4.3     ✓ tibble    3.2.1
## ✓ lubridate 1.9.2     ✓ tidyr     1.3.0
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts
to become errors
```

```
library(conflicted)
conflicts_prefer(dplyr::filter)
```

```
## [conflicted] Will prefer dplyr::filter over any other package.
```

```
library(here)
```

```
## Warning: package 'here' was built under R version 4.3.1
```

```
## here() starts at C:/Users/35196/Downloads
```

```
library(knitr)
```

```
## Warning: package 'knitr' was built under R version 4.3.1
```

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(here("acidentes.csv"), na.strings = c("NÃO DEFINIDO")) %>%
  mutate(Datahora = ymd_hms(Datahora)) %>%
  mutate(Longitude.GPS = ifelse(Longitude.GPS > 0, Longitude.GPS * -1,
Longitude.GPS),
         Latitude.GPS = ifelse(Latitude.GPS < 0, Latitude.GPS * -1,
Latitude.GPS)) %>%
  filter(Longitude.GPS > -10 & Longitude.GPS < -6 & Latitude.GPS > 37 &
Latitude.GPS < 42)
summary(df)
```

```
##   Id..Acidente          Datahora                      Entidades.Fiscalizadoras
##  Min.   :2.010e+09   Min.   :2010-01-01 00:05:00.00   Length:10908
##  1st Qu.:2.010e+09   1st Qu.:2010-04-10 20:26:15.00   Class :character
##  Median :2.010e+09   Median :2010-07-15 21:12:30.00   Mode  :character
##  Mean   :2.011e+09   Mean   :2010-07-09 17:07:34.68
##  3rd Qu.:2.010e+09   3rd Qu.:2010-10-08 11:00:00.00
##  Max.   :2.012e+09   Max.   :2010-12-31 21:40:00.00
##
##  Velocidade.local Velocidade.geral Dia.da.Semana       Latitude.GPS
##  Min.   : 10.00   Min.   : 20.00   Length:10908       Min.   :37.02
##  1st Qu.: 50.00   1st Qu.: 50.00   Class :character   1st Qu.:38.82
##  Median : 70.00   Median : 90.00   Mode  :character   Median :39.66
##  Mean   : 75.54   Mean   : 81.09                      Mean   :39.82
##  3rd Qu.: 90.00   3rd Qu.:100.00                      3rd Qu.:41.11
##  Max.   :120.00   Max.   :120.00                      Max.   :42.00
##  NA's   :353      NA's   :18
##  Longitude.GPS    Num..Mortos.a.30.dias Num..Feridos.graves.a.30.dias
##  Min.   :-9.944   Min.   :0.00000       Min.   : 0.00000
##  1st Qu.:-8.974   1st Qu.:0.00000       1st Qu.: 0.00000
##  Median :-8.572   Median :0.00000       Median : 0.00000
##  Mean   :-8.545   Mean   :0.04501       Mean   : 0.08865
##  3rd Qu.:-8.279   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:10908              Length:10908
##  1st Qu.: 1.000                  Class :character          Class :character
##  Median : 1.000                  Mode  :character          Mode  :character
##  Mean   : 1.328
##  3rd Qu.: 2.000
##  Max.   :36.000
##
##    Distrito           Concelho          Freguesia         Pov..Proxima
##  Length:10908       Length:10908       Length:10908       Length:10908
##  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:10908       Length:10908       Length:10908       Length:10908
##  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:10908          Length:10908       Length:10908
##  1st Qu.: 13.10   Class :character      Class :character   Class :character
##  Median : 36.75   Mode  :character      Mode  :character   Mode  :character
##  Mean   : 66.94
##  3rd Qu.: 80.12
##  Max.   :737.00
##  NA's   :1680
##  Localizações       Luminosidade        Marca.Via           Natureza
##  Length:10908       Length:10908       Length:10908       Length:10908
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   Obras.Arte         Obstáculos          Sentidos            Sinais
##  Length:10908       Length:10908       Length:10908       Length:10908
##  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:10908       Length:10908       Length:10908       Length:10908
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   Traçado.3          Traçado.4         Via.Trânsito
##  Length:10908       Length:10908       Length:10908
##  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 %>%
  select(2:14) %>%
  filter(Distrito %in% c("Leiria", "Lisboa", "Santarém", "Setúbal")) %>%
  mutate(Distrito = factor(Distrito)) %>%
  drop_na() -> df2
df2 %>%
  count(Distrito)
```

```
##   Distrito    n
## 1   Leiria 1192
## 2   Lisboa 2050
## 3 Santarém  609
## 4  Setúbal  671
```

```
df2 %>%
  count(Distrito) %>%
  kable()
```

Distrito	n
Leiria	1192
Lisboa	2050
Santarém	609
Setúbal	671

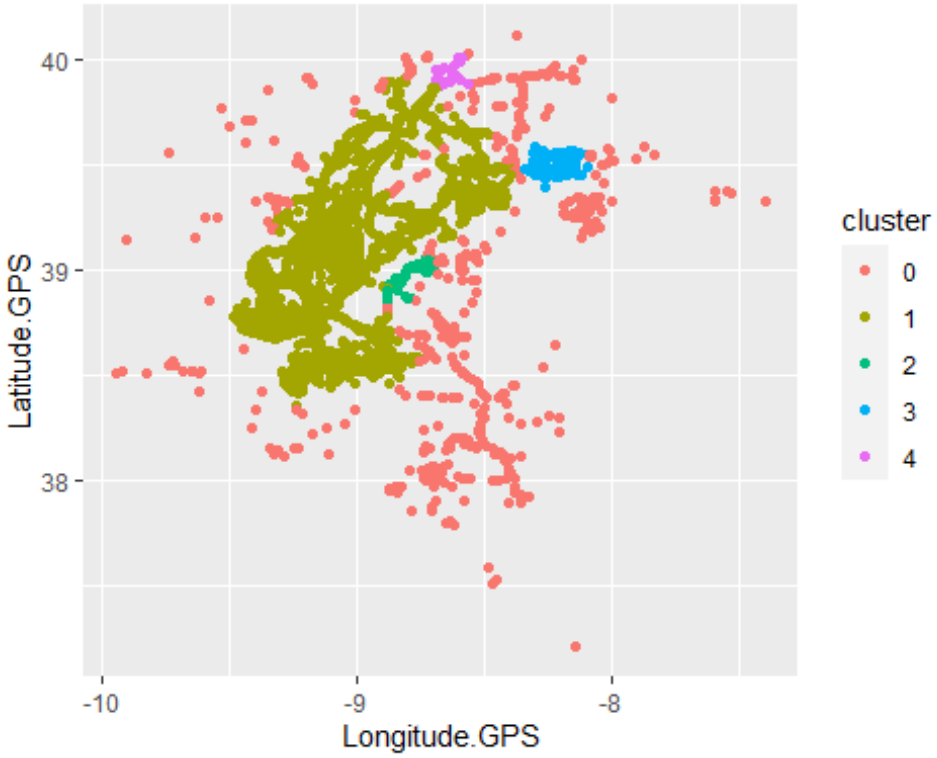
IV.3) [6 valores] Efectue o agrupamento dos dados obtidos em IV.2) com o algoritmo DBSCAN, baseado na Latitude.GPS e Longitude.GPS, utilizando eps=0.07 e minPts=40. Apresente um mapa dos clusters obtidos.No Quiz faça upload, em formato pdf, deste mapa.

```
set.seed(123)
library(dbscan)
```

```
## Warning: package 'dbscan' was built under R version 4.3.1
```

```
df2 %>%
  select(Latitude.GPS, Longitude.GPS) %>%
  dbscan(eps = 0.07, minPts = 40) -> clusters
```

```
df2 %>%
  mutate(cluster = clusters$cluster %>% as.factor) -> df.clusters
df.clusters %>%
  ggplot(aes(x = Longitude.GPS, y = Latitude.GPS, color = cluster)) +
  geom_point()
```




```
clusters
```

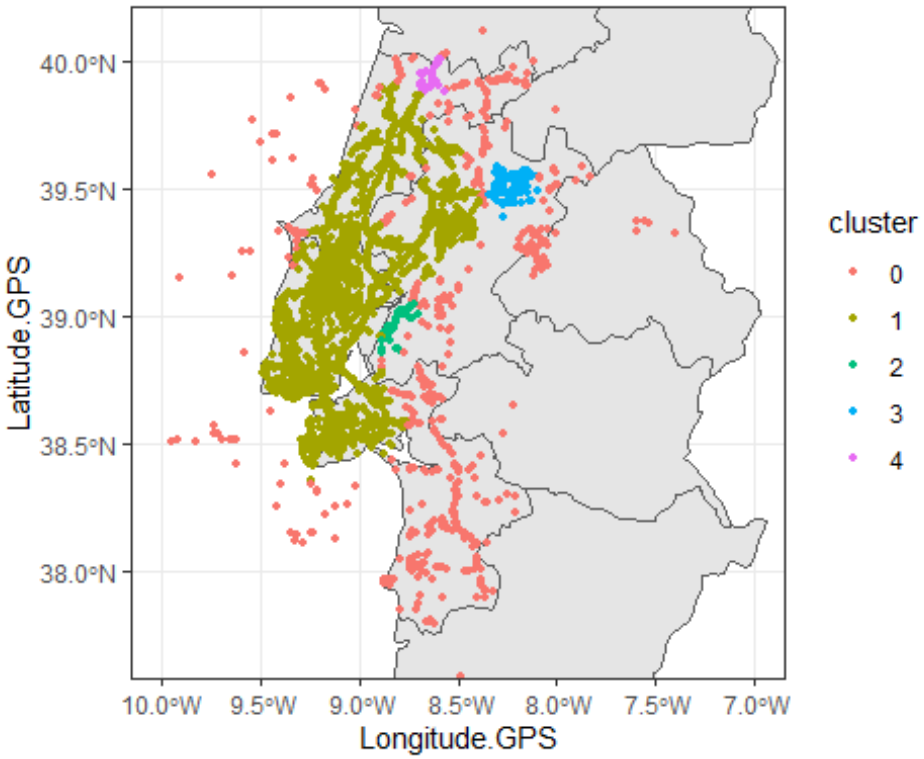
```
## DBSCAN clustering for 4522 objects.
## Parameters: eps = 0.07, minPts = 40
## Using euclidean distances and borderpoints = TRUE
## The clustering contains 4 cluster(s) and 607 noise points.
##
##    0    1    2    3    4
##  607 3673   73  112   57
##
## Available fields: cluster, eps, minPts, dist, borderPoints
```

```
# mapa
pak::pak("ropensci/rnaturalearthhires")
```

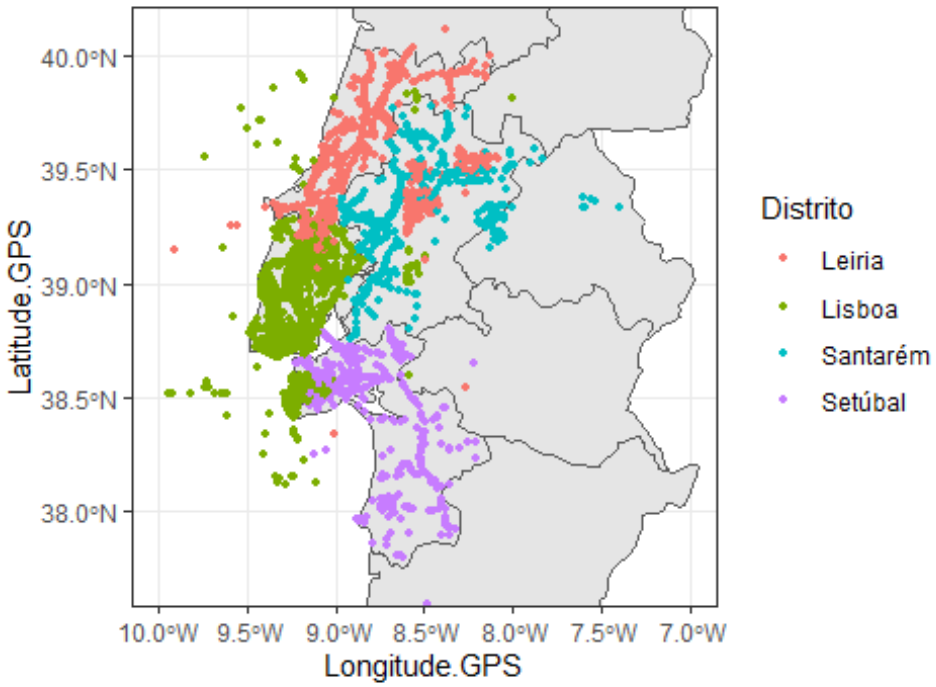
```
## ℹ Loading metadata database✓ Loading metadata database ... done
##
## ℹ No downloads are needed
## ✓ 1 pkg + 2 deps: kept 1 [7.5s]
```

```
library(rnaturalearth)
```

```
portugal <- ne_states(country = "portugal", returnclass = "sf")
portugal %>%
  ggplot() +
    # color districts
  geom_sf() +
  geom_point(data = df.clusters, alpha = 1, size = 1, aes(x =
Longitude.GPS, y = Latitude.GPS, color = cluster)) +
  coord_sf(xlim = c(-10, -7), ylim = c(37.7, 40.1)) +
  theme_bw()
```

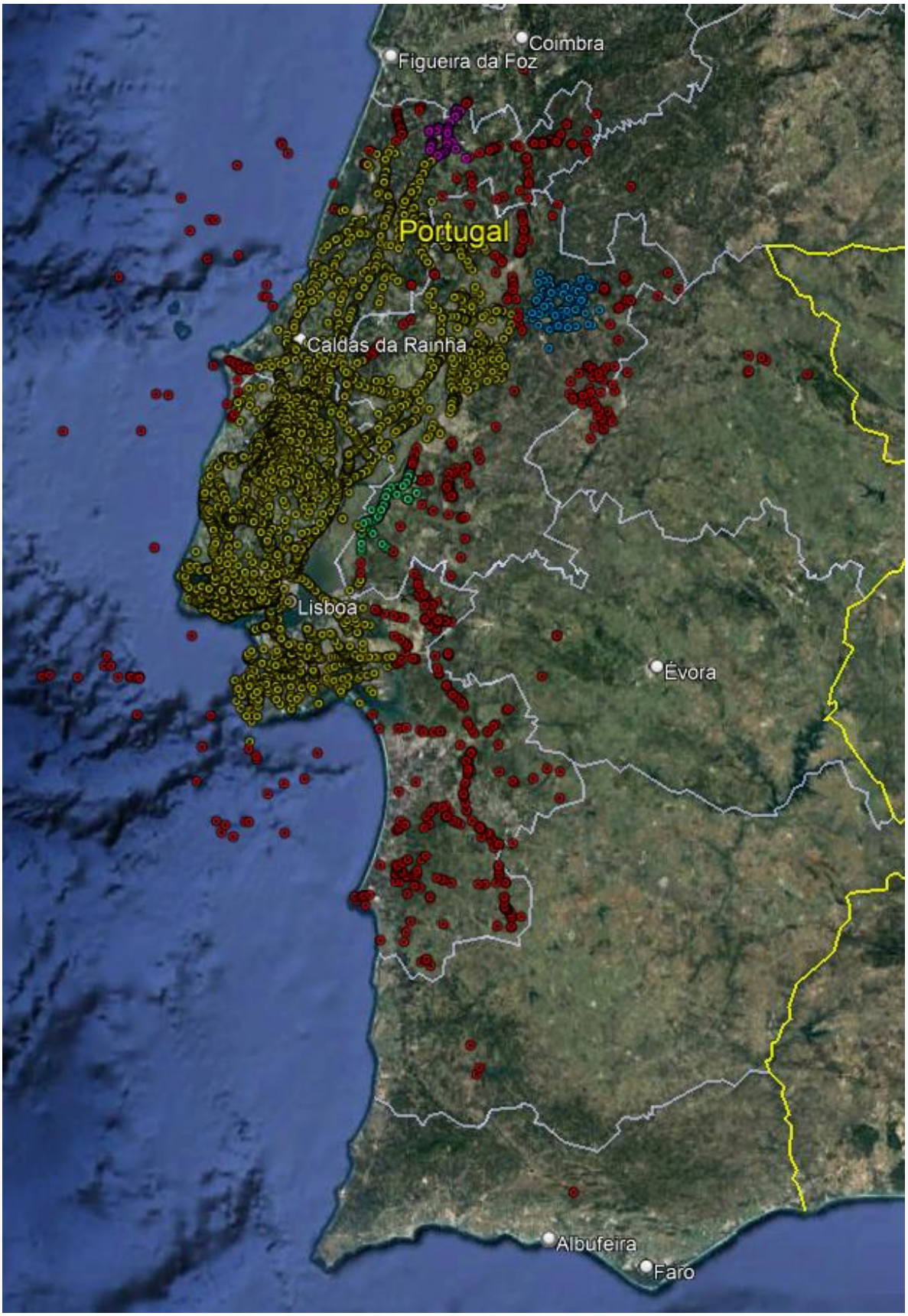



```
# portugal distribuição pelos 4 distritos (Leiria, Lisboa, Santarém,
Setúbal)
portugal %>%
  ggplot() +
  geom_sf() +
  geom_point(data = df.clusters, alpha = 1, size = 1, aes(x =
Longitude.GPS, y = Latitude.GPS, color = Distrito)) +
  coord_sf(xlim = c(-10, -7), ylim = c(37.7, 40.1)) +
  theme_bw()
```




```r
if (F) {
  db_df <- data.frame(clusters$cluster)
  asd <- merge(df2, db_df, by.x = 0, by.y=0, sort=FALSE)

  cluster_0 <- asd[which(asd$clusters.cluster==0),]
  cluster_1 <- asd[which(asd$clusters.cluster==1),]
  cluster_2 <- asd[which(asd$clusters.cluster==2),]
  cluster_3 <- asd[which(asd$clusters.cluster==3),]
  cluster_4 <- asd[which(asd$clusters.cluster==4),]
  write.csv(cluster_0, "cluster_0.csv")
  write.csv(cluster_1, "cluster_1.csv")
  write.csv(cluster_2, "cluster_2.csv")
  write.csv(cluster_3, "cluster_3.csv")
  write.csv(cluster_4, "cluster_4.csv")
}
```

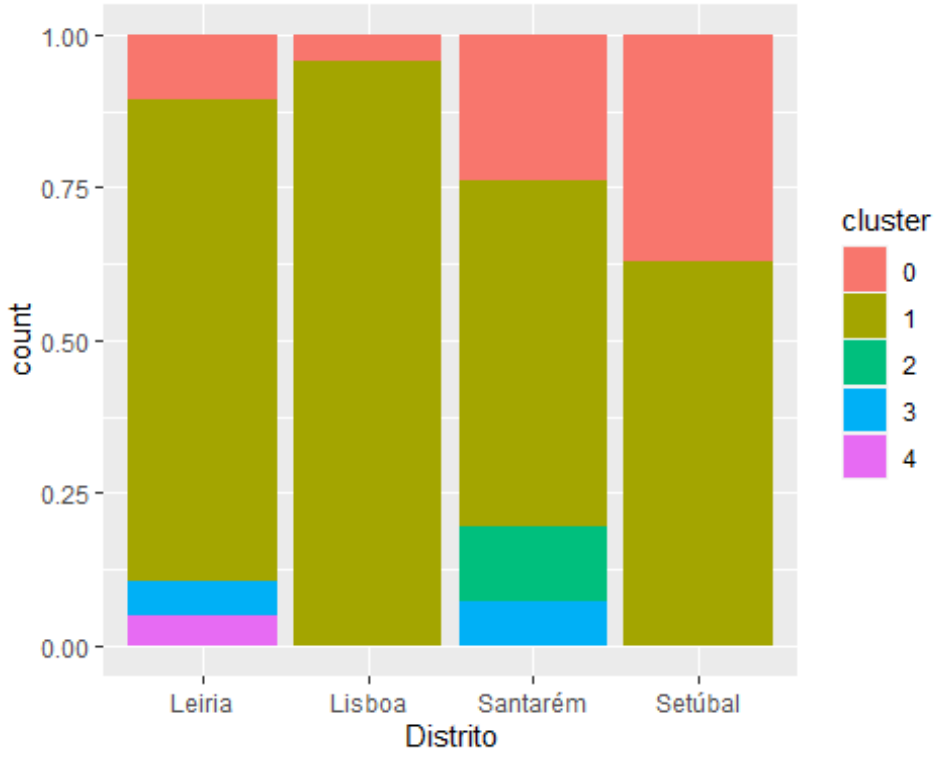

Coimbra
Figueira da Foz
Portugal
Caldas da Rainha
Lisboa
Évora
Albufeira
Faro

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.

```
df.clusters %>%
  select(Distrito, cluster) %>%
  table()
```

```
##           cluster
## Distrito      0    1    2    3    4
##   Leiria    126  942    0   67   57
##   Lisboa     87 1963    0    0    0
##   Santarém  146  345   73   45    0
##   Setúbal   248  423    0    0    0
```

```
# grafico de barras
df.clusters %>%
  select(Distrito, cluster) %>%
  ggplot(aes(x = Distrito, fill = cluster)) +
  geom_bar(position = "fill")
```



A associação entre os Clusters obtidos e Distrito, medida por **V de Cramer** tem o valor **0.32** (com 2 c.d.).

```
# Distrito
lsr::cramersV(df.clusters$Distrito, df.clusters$cluster) %>% round(3)
```

```
## [1] 0.32
```

```
lsr::cramersV(df.clusters$Entidades.Fiscalizadoras, df.clusters$cluster,
simulate.p.value=TRUE) %>% round(3)
```

```
## [1] 0.136
```

```
lsr::cramersV(df.clusters$Dia.da.Semana, df.clusters$cluster) %>%
round(3)
```

```
## [1] 0.043
```

```
lsr::cramersV(df.clusters$Características.Tecnicas1, df.clusters$cluster)
%>% round(3)
```

```
## [1] 0.113
```

```
lsr::cramersV(df.clusters$Cond.Aderência, df.clusters$cluster,
simulate.p.value=TRUE) %>% round(3)
```

```
## [1] 0.045
```

De acordo com a medida de associação **Eta**(R de pearson | R de Spearman | V de Cramer | Eta), **nenhuma das** (todas, algumas, nenhuma das) variáveis métricas são relevante para a caracterização dos clusters obtidos.

```
options(scipen = 999)
df.clusters %>%
  # select if numeric or cluster
  select(where(is.numeric)) %>%
  aov(df.clusters$cluster %>% as.numeric ~ ., data = .) %>%
  lsr::etaSquared() %>%
  as.data.frame() %>%
  select(eta.sq) %>%
  mutate(eta = sqrt(eta.sq)) %>%
  arrange(-eta) %>%
  round(3)
```

```
##                                eta.sq   eta
## Latitude.GPS                    0.078 0.278
## Longitude.GPS                   0.010 0.099
## Num..Feridos.graves.a.30.dias   0.002 0.044
## Velocidade.geral                0.000 0.018
## Num..Mortos.a.30.dias           0.000 0.010
## Num..Feridos.ligeiros.a.30.dias 0.000 0.010
## Velocidade.local                0.000 0.008
```

Dentro do cluster 3 o observa-se a percentagem **59.8**% (com 1 c.d.) de acidentes ocorridos no Distrito de Leiria;e no cluster **1** os acidentes neste distrito representam 25.6%.

```
table(df.clusters$cluster, df.clusters$Distrito) %>% t() %>%
prop.table(.,2) %>% round(3)*100
```

```
##
##                0     1     2     3     4
##   Leiria    20.8  25.6   0.0  59.8 100.0
##   Lisboa    14.3  53.4   0.0   0.0   0.0
##   Santarém  24.1   9.4 100.0  40.2   0.0
##   Setúbal   40.9  11.5   0.0   0.0   0.0
```

A localidade de Abrantes, com Longitude.GPS = **-8.2** (graus decimais com 1 c.d.) e Latitude.GPS = **39.5** (graus decimais com 1 c.d.) classifica-se no cluster **3** ((número do cluster)

Para o concelho de Abrantes - previsão

```
df %>% filter(Concelho == "Abrantes") %>% select(Latitude.GPS,
Longitude.GPS) -> abrantes
lat_abrantes.min <- min(abrantes$Latitude.GPS)
lat_abrantes.max <- max(abrantes$Latitude.GPS)
long_abrantes.min <- min(abrantes$Longitude.GPS)
long_abrantes.max <- max(abrantes$Longitude.GPS)
lat_abrantes.mean <- mean(abrantes$Latitude.GPS)
long_abrantes.mean <- mean(abrantes$Longitude.GPS)
# centro oficial: 39,464294, -8,197861
centroide_abrantes_lat <- 39.464294
centroide_abrantes_long <- -8.197861
tibble(lat_abrantes.min, lat_abrantes.max, lat_abrantes.mean,
centroide_abrantes_lat, long_abrantes.min, long_abrantes.max,
long_abrantes.mean, centroide_abrantes_long) %>% t() %>% round(1)
```

```
##                          [,1]
## lat_abrantes.min         39.2
## lat_abrantes.max         39.5
## lat_abrantes.mean        39.4
## centroide_abrantes_lat   39.5
## long_abrantes.min        -9.1
## long_abrantes.max        -8.0
## long_abrantes.mean       -8.2
## centroide_abrantes_long -8.2
```

```
# df.clusters %>% filter(Latitude.GPS > 39.4 & Latitude.GPS < 39.5 &
Longitude.GPS > -8.2 & Longitude.GPS < -8.1) %>% select(Distrito,
cluster) %>% table()
```

```
abrantes<-matrix (NA ,1,2)
abrantes[1,]<-c(39.5, -8.2)
```

```
(pred_abrantes <-
predict(clusters,as.data.frame(abrantes),df.clusters[,6:7]))

## [1] 3
```

Para o concelho de Óbidos - previsão

```
df %>% filter(Concelho == "Obidos") %>% select(Latitude.GPS,
Longitude.GPS) -> obidos
lat_obidos.min <- min(obidos$Latitude.GPS)
lat_obidos.max <- max(obidos$Latitude.GPS)
long_obidos.min <- min(obidos$Longitude.GPS)
long_obidos.max <- max(obidos$Longitude.GPS)
(lat_obidos.mean <- mean(obidos$Latitude.GPS))

## [1] 39.30474

(long_obidos.mean <- mean(obidos$Longitude.GP))

## [1] -9.111618

obidos <- matrix(NA, 1, 2)
obidos[1,] <- c(39.30, -9.11)
(pred_obidos <- predict(clusters, as.data.frame(obidos),
df.clusters[,6:7]))

## [1] 1
```

Respostas finais

A associação entre os Clusters obtidos e Distrito, medida por **V de Cramer** tem o valor **0.32** (com 2 c.d.). De acordo com a medida de associação **Eta**(R de pearson | R de Spearman | V de Cramer | Eta), **nenhumas das** (todas, algumas, nenhuma das) variáveis métricas são relevante para a caracterização dos clusters obtidos. Dentro do cluster 3 o observa-se a percentagem **59.8**% (com 1 c.d.) de acidentes ocorridos no Distrito de Leiria;e no cluster **1** os acidentes neste distrito representam 25.6%. A localidade de Abrantes, com Longitude.GPS = **-8.2**(graus decimais com 1 c.d.) e Latitude.GPS = **39.5** (graus decimais com 1 c.d.) classifica-se no cluster **3** ((número do cluster).