AnalisisExploratorio

DavidVelasco

2022-12-21

1. Introducción con el objetivo del análisis

- Se pretende solo hasta el Exploratorio
- Se debe plantear hipotesis posterior(Clasificación)

2. Carga de los datos
3. Análisis descriptivo

- Comandos de la sesion que estan en el apartado “Summarizing”

4. Analisis Exploratorio

- Graficos Exploratorios y Clustering del Script

5. Conclusiones.

- ¿Pensamos que este dataset puede servir para la finalidad/modelo que habíamos planteado en la Introducción? ¿Tenemos ya alguna conclusión preliminar sobre qué variables pueden resultar más útiles para dicha finalidad/modelo(clasificación)?

Nota: El SCRIPT debe de ser REPRODUCIBLE: No debe depender de las rutas locales (directorios, paths) del equipo del alumno. Se recomienda utilizar setwd(), getwd(), rutas relativas (./, ../) y funciones de modo conveniente.

La entrega será adjuntando el archivo R Markdown al Moodle y el informe generado (PDF, Word, HTML o PPT).

1. Introducción de datos

- El objetivo es intentar agrupar los datos en clusters y ver que variables son mas influyentes en dicha agrupación.

2. Carga de los datos

```
#Comprobación de que están los *.csv en el directorio

currentDir <- getwd()
list.files(path="../datos")
```

```
## [1] "anemonefish.xls"          "beer2.csv"
## [3] "DatasetLimpio.xlsx"       "EXAMPLE_DataToClean.xlsx"
## [5] "output"                   "religions.csv"
## [7] "student-mat.csv"          "student-por.csv"
## [9] "student.zip"
```

```
if (!file.exists("../datos"))
{stop(paste0("Se necesita que el directorio datos esté en: ",currentDir))}


ComprobarInputs <- function(path, dir,file)
{if (!file.exists(paste0(path,"/",dir)))
{stop(paste0("Se necesita que el directorio ", dir, " esté en: ",path))}
  else if (!file.exists(paste0(path, "/",dir,"/", file)))
    {stop(paste0("Se necesita que ", file," esté en: ", path, "/", dir))}}

parentPath <- dirname(currentDir)

try(ComprobarInputs(parentPath,"datos", "student-por.csv"), FALSE)
try(ComprobarInputs(parentPath,"datos", "student-mat.csv"), FALSE)

#Unificacion de los dos ficheros en un dataframe

d1=read.table("../datos/student-mat.csv",sep=";",header=TRUE)
d2=read.table("../datos/student-por.csv",sep=";",header=TRUE)
d3=merge(d1,d2,by=c("school","sex","age","address","famsize",
                    "Pstatus","Medu","Fedu","Mjob","Fjob","reason","nursery","internet"))
class(d3)
```

```
## [1] "data.frame"
```

3. Análisis Descriptivo

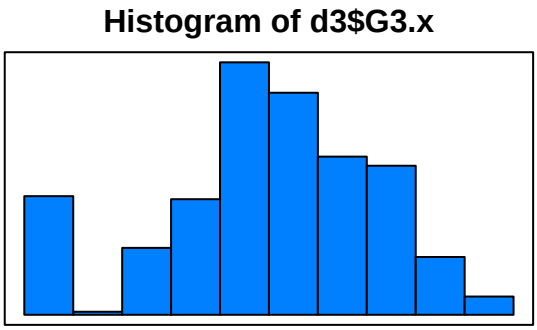
```
#Resumen por trimestres

require(Hmisc)
Hmisc::describe(d3$G1.x)
Hmisc::describe(d3$G2.x)
Hmisc::describe(d3$G3.x)

require(pastecs)
res <-stat.desc(d3)
res <-round(res,2)

#Borro las columnas que no tienen valores numericos
res <- stat.desc(d3[,-c(1,2,4,5,6,9,10,11,12,13,14,18,19,20,21,22,23,34,38,39,40,41,42,43)])

require(psych)
psych::describe(d3$G1.x)
psych::describe(d3$G2.x)
psych::describe(d3$G3.x)
```

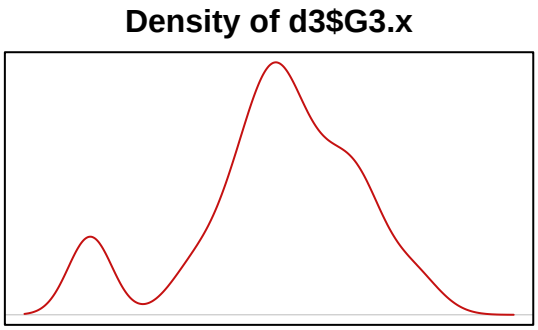
```
psych::describeBy(d3$G1.x, group=d3$sex)
psych::describeBy(d3$G1.x, group=d3$address)

#Media de edades
mean(d3$age)

#Tablas cruzadas
table(d3$sex)
table(d3$sex, d3$age)

tMale<-table(d3$Mjob, d3$Medu)
prop.table(tMale)

tFemale<-table(d3$Fjob, d3$Fedu)
prop.table(tFemale)

##Tablas de frecuencias
prop.table(table(d3$sex, d3$Fedu))
prop.table(table(d3$sex, d3$Medu))
prop.table(table(d3$studytime.x, d3$G3.x))

table(d3$Fedu)
table(d3$Medu)
table(d3$romantic.x)

#Resumen de todas las columnas
summary(d3)
```

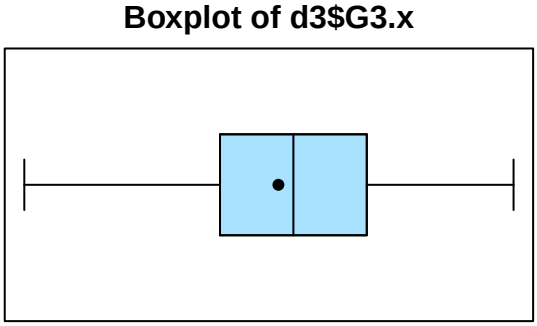
4. Análisis Exploratorio

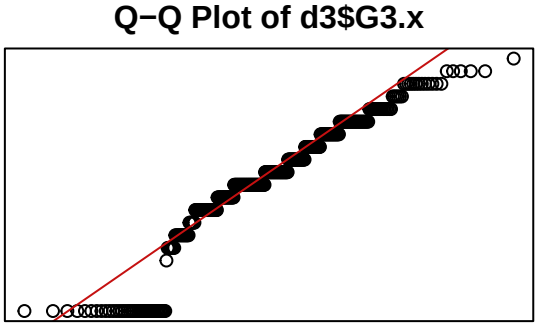
```
#Ejemplo sobre G3
EDA(d3$G3.x)
```

EXPLORATORY DATA ANALYSIS



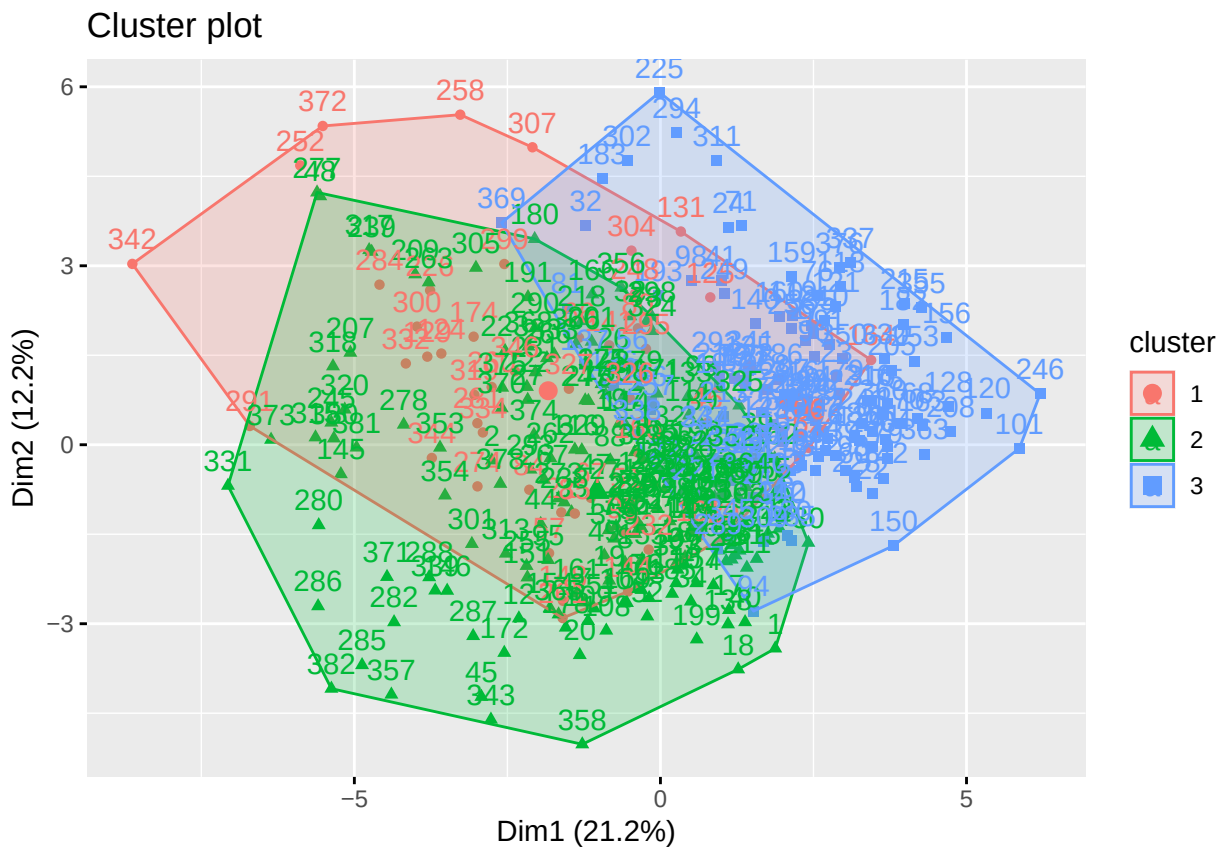






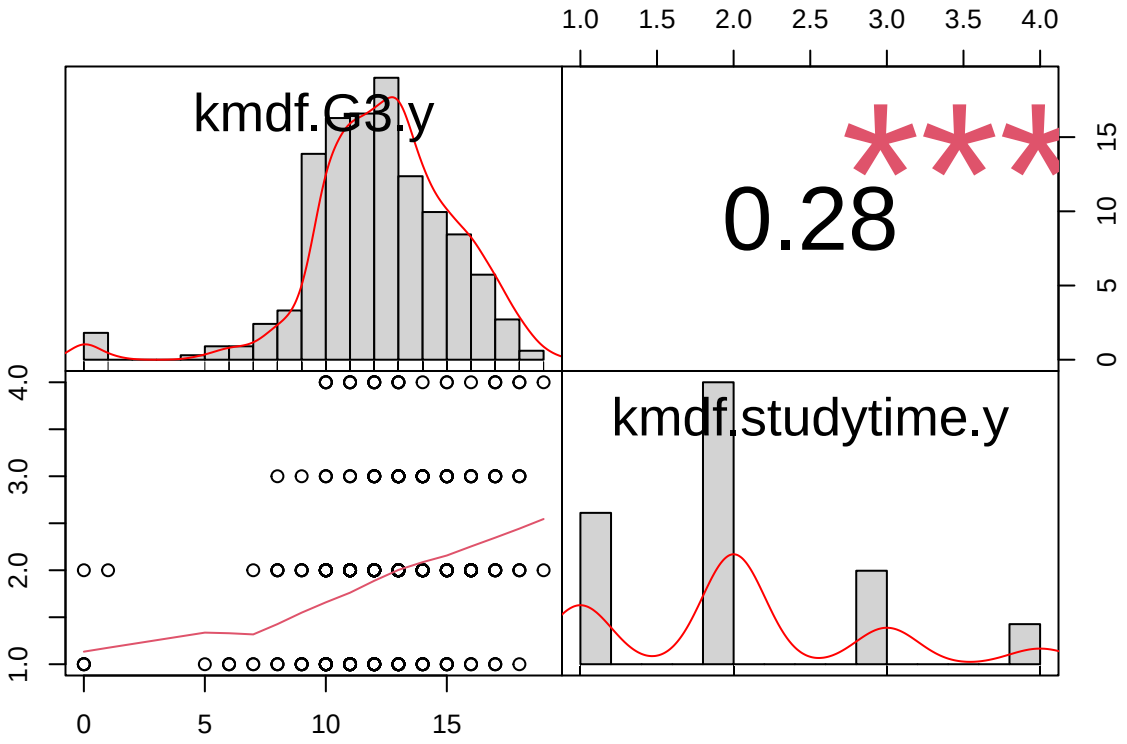

```
## Size (n)  Missing  Minimum   1st Qu     Mean   Median   TrMean   3rd Qu
##  382.000    0.000    0.000    8.000   10.387   11.000   10.520   14.000
##      Max    Stdev      Var  SE Mean   I.Q.R.    Range Kurtosis Skewness
##   20.000    4.687   21.970    0.240    6.000   20.000    0.242   -0.700
## SW p-val
##    0.000
```

```r
#Analisis mediante k-means con k=3

require("factoextra")
#Quito las columnas que no son numericas y las que no me interesen
kmdf<- (d3[,-c(1,2,4,5,6,9,10,11,12,13,14,18,19,20,21,22,23,34,38,39,40,41,42,43)])
km <- kmeans(kmdf, centers = 3, nstart = 25)
fviz_cluster(km, data = kmdf)
```




```
#Correlacion entre dos variables

require(PerformanceAnalytics)
dat1 <- data.frame(kmdf$G3.y, kmdf$studytime.y)
chart.Correlation(dat1)
```

```
## Warning in par(usr): argument 1 does not name a graphical parameter
```



5. Conclusiones

- El objetivo era hacer grupos pero debido a la baja correlación entre todas las variables y observar que el dataset no es el mejor no sirve para la finalidad planteada en la introducción.
- Las variables que mas correlación positiva tienen sobre la nota final son la educación del padre y de la madre y el tiempo de estudio