

Analisis de Datos

Maria Jose Bustamante - Paola Peralta Flores

Analisis Descriptivo e Inferencial

Librerias que vamos a utilizar.

```
library(ggplot2)
library(ggpubr)
library(dplyr)
```

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

filter, lag

The following objects are masked from 'package:base':

intersect, setdiff, setequal, union

```
library(glmnet)
```

Loading required package: Matrix

Loaded glmnet 4.1-7

```
library(caret)
```

Loading required package: lattice

```
library(e1071)
library(ggstatsplot)
```

You can cite this package as:

Patil, I. (2021). Visualizations with statistical details: The 'ggstatsplot' approach. Journal of Open Source Software, 6(61), 3167, doi:10.21105/joss.03167

```
library(corrplot)
```

corrplot 0.92 loaded

```
library(lavaan)
```

This is lavaan 0.6-15

lavaan is FREE software! Please report any bugs.

Leer los datos y renombrar las variables.

```
Datos <- read.csv("./processed.cleveland.data",header=FALSE,sep=" ",
                 na.strings = '?')
names(Datos) <- c( "Edad", "Sexo", "DPecho", "PresArtRep", "Colesterol", "AzuA",
                  "ECGRep", "FCardiaca", "Angina", "ST", "PenST", "Vasos",
                  "Estado", "Enfermedad")
```

Exploracion de los datos. Con la funcion head se muestran las primeras seis filas del conjunto de datos.

```
head(Datos)
```

	Edad	Sexo	DPecho	PresArtRep	Colesterol	AzuA	ECGRep	FCardiaca	Angina	ST	PenST
1	63	1	1	145	233	1	2	150	0 2.3	3	
2	67	1	4	160	286	0	2	108	1 1.5	2	
3	67	1	4	120	229	0	2	129	1 2.6	2	
4	37	1	3	130	250	0	0	187	0 3.5	3	
5	41	0	2	130	204	0	2	172	0 1.4	1	
6	56	1	2	120	236	0	0	178	0 0.8	1	

	Vasos	Estado	Enfermedad
1	0	6	0
2	3	3	2
3	2	7	1
4	0	3	0
5	0	3	0
6	0	3	0

Definimos los datos por sus categorías.

```
Datos$Sexo[Datos$Sexo==1] <- "Masculino"
Datos$Sexo[Datos$Sexo==0] <- "Femenino"
Datos$DPecho[Datos$DPecho==1] <- "Tipo 1"
Datos$DPecho[Datos$DPecho==2] <- "Tipo 2"
Datos$DPecho[Datos$DPecho==3] <- "Tipo 3"
Datos$DPecho[Datos$DPecho==4] <- "Tipo 4"
Datos$Azua[Datos$Azua==1] <- "Verdadero"
Datos$Azua[Datos$Azua==0] <- "Falso"
Datos$ECGRep[Datos$ECGRep==0] <- "Nivel 0"
Datos$ECGRep[Datos$ECGRep==1] <- "Nivel 1"
Datos$ECGRep[Datos$ECGRep==2] <- "Nivel 2"
Datos$Angina[Datos$Angina==1] <- "Si"
Datos$Angina[Datos$Angina==0] <- "No"
Datos$PenST[Datos$PenST==1] <- "Valor 1"
Datos$PenST[Datos$PenST==2] <- "Valor 2"
Datos$PenST[Datos$PenST==3] <- "Valor 3"
Datos$Estado[Datos$Estado==3] <- "N"
Datos$Estado[Datos$Estado==6] <- "DF"
Datos$Estado[Datos$Estado==7] <- "DR"
```

Agregamos una columna, modificando las etapas de “Enfermedad”:

- Saludable (0 - No)
- Enfermo (1,2,3,4 - Si).

```
c <- Datos$Enfermedad
Corazon <- data.frame("Corazon"=c(c))
Corazon$Corazon[Corazon$Corazon==0] <- "No"
Corazon$Corazon[Corazon$Corazon==1] <- "Si"
Corazon$Corazon[Corazon$Corazon==2] <- "Si"
Corazon$Corazon[Corazon$Corazon==3] <- "Si"
```

```

Corazon$Corazon[Corazon$Corazon==4] <- "Si"

Datos <- cbind(Datos, Corazon)

# Permite visualizar los datos de la tabla de mejor manera
pander::pandoc.table(
  head(Datos))

```

Edad	Sexo	DPecho	PresArtRep	Colesterol	AzuA	ECGRep
63	Masculino	Tipo 1	145	233	Verdadero	Nivel 2
67	Masculino	Tipo 4	160	286	Falso	Nivel 2
67	Masculino	Tipo 4	120	229	Falso	Nivel 2
37	Masculino	Tipo 3	130	250	Falso	Nivel 0
41	Femenino	Tipo 2	130	204	Falso	Nivel 2
56	Masculino	Tipo 2	120	236	Falso	Nivel 0

Table: Table continues below

FCardiaca	Angina	ST	PenST	Vasos	Estado	Enfermedad	Corazon
150	No	2.3	Valor 3	0	DF	0	No
108	Si	1.5	Valor 2	3	N	2	Si
129	Si	2.6	Valor 2	2	DR	1	Si
187	No	3.5	Valor 3	0	N	0	No
172	No	1.4	Valor 1	0	N	0	No

178	No	0.8	Valor 1	0	N	0	No
-----	----	-----	---------	---	---	---	----

Analisis Univariante

Se obtienen las medidas de tendencia central, valores minimos y maximos y los cuantiles de las variables especificadas. Se visualizan las variables de manera independiente, utilizando variables categoricas y variables cuantitativas.

```
V.Cuantitativas <- data.frame("Edad" = Datos$Edad,"PresArtRep" = Datos$PresArtRep,
                              "Colesterol" = Datos$Colesterol,
                              "FCardiaca" = Datos$FCardiaca,"ST" = Datos$ST)
# Se realiza una tabla con todas estas variables
knitr::kable(summary(V.Cuantitativas))
```

Edad	PresArtRep	Colesterol	FCardiaca	ST
Min. :29.00	Min. : 94.0	Min. :126.0	Min. : 71.0	Min. :0.00
1st Qu.:48.00	1st Qu.:120.0	1st Qu.:211.0	1st Qu.:133.5	1st Qu.:0.00
Median :56.00	Median :130.0	Median :241.0	Median :153.0	Median :0.80
Mean :54.44	Mean :131.7	Mean :246.7	Mean :149.6	Mean :1.04
3rd Qu.:61.00	3rd Qu.:140.0	3rd Qu.:275.0	3rd Qu.:166.0	3rd Qu.:1.60
Max. :77.00	Max. :200.0	Max. :564.0	Max. :202.0	Max. :6.20

Se calcula la desviacion estandar de las variables especificadas.

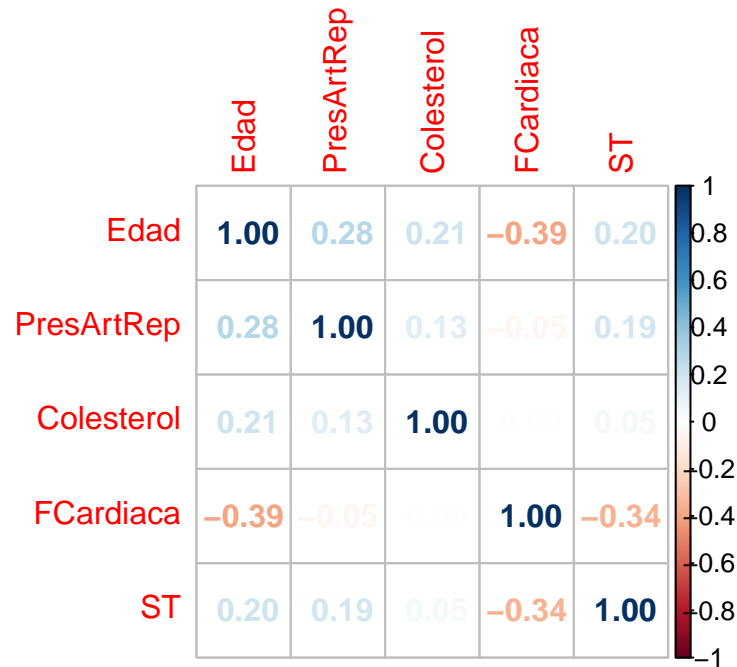
```
z <- data.frame("Variables"= c("Edad","PresArtRep","Colesterol","FCardiaca","ST"),"Desv Est"=
knitr::kable(z)
```

Variables	Desv.Est
Edad	9.038662
PresArtRep	17.599748
Colesterol	51.776918
FCardiaca	22.875003
ST	1.161075

Analisis Bivariante

Se realiza una correlacion de las variables almacenadas en “V.Cuantitativas”

```
Cor <- cor(V.Cuantitativas)
corrplot(Cor, method="number")
```



- En la grafica se observa esta correlacion, variando entre los colores rojo y azul, diferenciando de esta manera cuando la correlacion entre variables se hace cada vez más fuerte. En este caso las variables son muy débiles, ninguna supera el 0.5 para concluir que existe al menos una correlacion moderada o fuerte entre las variables.

Se realiza una matriz de varianzas y covarianzas.

```
Covarianza <- cov(V.Cuantitativas)
knitr::kable(Covarianza)
```

	Edad	PresArtRep	Colesterol	FCardiaca	ST
Edad	81.69742	45.328678	97.787489	-81.423065	2.138850
PresArtRep	45.32868	309.751120	118.573340	-18.258005	3.865638
Colesterol	97.78749	118.573340	2680.849190	-4.064652	2.799282
FCardiaca	-81.42307	-18.258005	-4.064652	523.265775	-9.112209
ST	2.13885	3.865638	2.799282	-9.112209	1.348095

Se realiza la comprobacion de que las variables sean independientes.

```
chi <- chisq.test(table(Datos$FCardiaca,
                        Datos$Colesterol))
```

Warning in chisq.test(table(Datos\$FCardiaca, Datos\$Colesterol)): Chi-squared approximation may be incorrect

Analisis PCA

```
PCA <- prcomp(Datos[,c("Edad", "PresArtRep", "Colesterol", "FCardiaca", "ST")])
```

PCA

Standard deviations (1, ..., p=5):

```
[1] 51.871246 23.245850 17.513339 7.619184 1.070944
```

Rotation (n x k) = (5 x 5):

	PC1	PC2	PC3	PC4	PC5
Edad	-0.038400706	-0.18061002	-0.12731753	-0.97451825	0.003136272
PresArtRep	-0.050463490	-0.10499968	-0.98177615	0.14974926	0.010920408
Colesterol	-0.997979694	0.01594758	0.05406571	0.02930745	0.000421610
FCardiaca	0.003744198	0.97763994	-0.13001050	-0.16440401	-0.016574588
ST	-0.001154567	-0.01791396	-0.00894484	0.00131652	-0.999797986

```
summary(PCA)
```

Importance of components:

	PC1	PC2	PC3	PC4	PC5
Standard deviation	51.871	23.2458	17.51334	7.61918	1.07094
Proportion of Variance	0.748	0.1502	0.08527	0.01614	0.00032
Cumulative Proportion	0.748	0.8983	0.98354	0.99968	1.00000

```
biplot(PCA)
```

