Análisis conjunto de datos

Nicolas Jadan

Analisis conjunto de datos

```
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.switzerland.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   54    1      1        145        233    1      2       180      0 2.3     3
2   51    1      3        100        222    0      0       143      1 1.2     2
3   55    1      4        140        217    0      0       111      1 5.6     3
4   65    1      1        138        282    1      2       174      0 1.4     2
5   45    0      2        130        234    0      2       175      0 0.6     2
6   56    0      4        200        288    1      2       133      1 4.0     3
```

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

Definimos los datos por sus categorias.

```
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"
```

```r
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
54	Masculino	Tipo 1	145	233	Verdadero	Nivel 2
51	Masculino	Tipo 3	100	222	Falso	Nivel 0
55	Masculino	Tipo 4	140	217	Falso	Nivel 0
65	Masculino	Tipo 1	138	282	Verdadero	Nivel 2
45	Femenino	Tipo 2	130	234	Falso	Nivel 2
56	Femenino	Tipo 4	200	288	Verdadero	Nivel 2

Table: Table continues below

FCardiaca	Angina	ST	PenST	Vasos	Estado	Enfermedad	Corazon
180	No	2.3	Valor 3	0	DF	0	No
143	Si	1.2	Valor 2	0	N	0	No
111	Si	5.6	Valor 3	0	DR	3	Si
174	No	1.4	Valor 2	1	N	1	Si
175	No	0.6	Valor 2	0	N	0	No

```
    133        Si       4    Valor 3     2       DR         3          Si
-------------------------------------------------------------------------------
```

Analisis PCA

Tecnica útil para resumir y explorar datos complejos, reducir la dimensionalidad y encontrar las variables y combinaciones lineales más relevantes en un conjunto de datos.

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

PCA
```

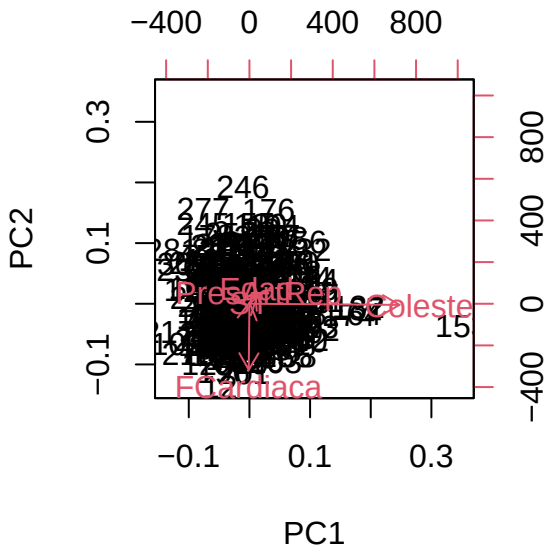
```
Standard deviations (1, .., p=5):
[1] 51.847303 23.282294 17.519092  7.608652  1.073141

Rotation (n x k) = (5 x 5):
                    PC1         PC2          PC3          PC4           PC5
Edad        0.038594890  0.17942893 -0.127647110 -0.974686396 -0.0028877146
PresArtRep  0.050632586  0.09853641 -0.982571564  0.148856819 -0.0110311912
Colesterol  0.997959826 -0.01651227  0.054225029  0.029376598 -0.0004385629
FCardiaca  -0.004643303 -0.97852682 -0.123429549 -0.164203487  0.0163115845
ST          0.001183631  0.01756267 -0.009172248  0.001519055  0.9998018373
```

```
summary(PCA)
```

```
Importance of components:
                          PC1     PC2      PC3    PC4     PC5
Standard deviation     51.8473 23.2823 17.51909 7.6087 1.07314
Proportion of Variance  0.7475  0.1507  0.08535 0.0161 0.00032
Cumulative Proportion   0.7475  0.8982  0.98358 0.9997 1.00000
```

```
biplot(PCA)
```



Analisis Univariante

Se obtienen las medidas de tendencia central, valores minimos y maximos y los quartiles 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)
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.:211.5   1st Qu.:133.5   1st Qu.:0.00
 Median :55.00   Median :130.0   Median :241.0   Median :153.0   Median :0.80
 Mean   :54.41   Mean   :131.7   Mean   :246.7   Mean   :149.7   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
```

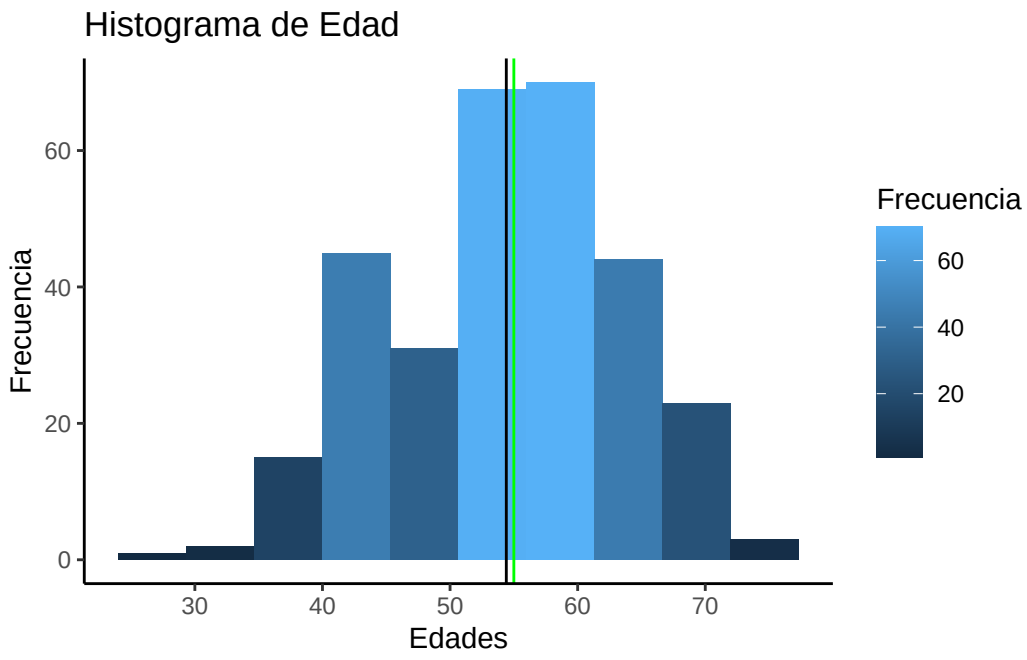
Se calcula la desviacion estandar de las variables especificadas.

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

Variables	Desv.Est
Edad	9.025214
PresArtRep	17.599748
Colesterol	51.752156
FCardiaca	22.920089
ST	1.161075

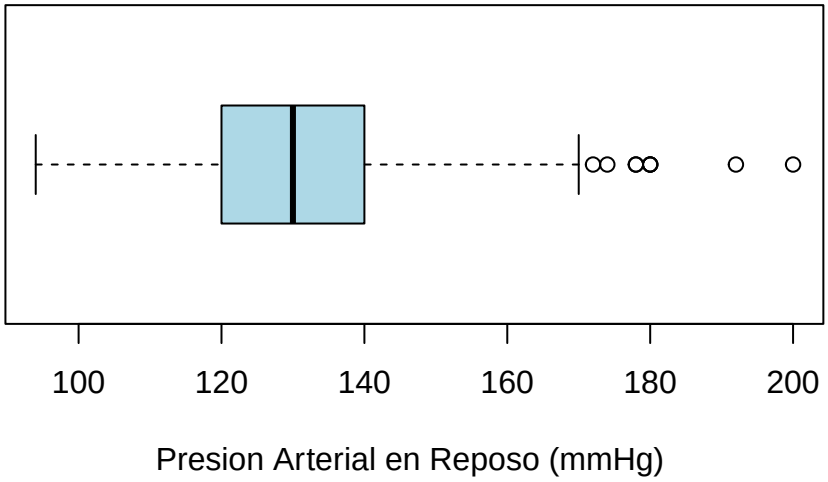
Se realiza un histograma para una de las variables cuantitativas.

```r
ggplot(V.Cuantitativas, aes(Edad)) +
  geom_histogram(aes(fill=after_stat(count)), bins=10) +
  geom_vline(aes(xintercept=mean(Edad)), color="black") +
  geom_vline(aes(xintercept=median(Edad)), color= "green") +
  labs(title = "Histograma de Edad",
       x = "Edades",
       y = "Frecuencia") +
  scale_fill_continuous(name="Frecuencia") +
  theme_classic()
```



- El promedio de edad entre pacientes resulto de 54 años, con una mediana de 56 años (cercana al valor de la media, pero mayor), la línea roja refleja el valor de la media y la amarilla el valor de la mediana, también se puede observar que los datos tienen más concentración en las edades entre 50 a 60 años.

```
boxplot(V.Cuantitativas$PresArtRep, xlab = "Presion Arterial en Reposo (mmHg) ",col =
```

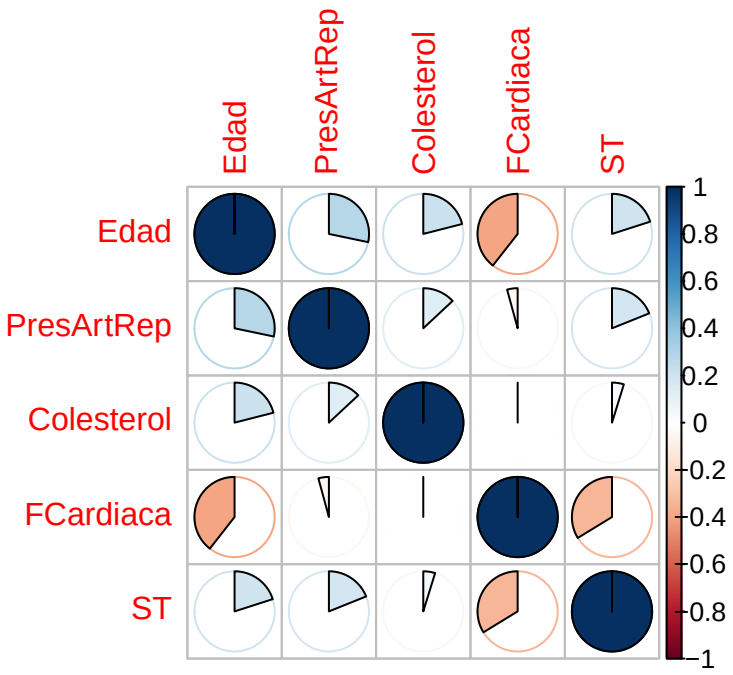


- El 25% de los pacientes presentaron una frecuencia menor o igual a 133.5, el 50% de ellos una frecuencia menor o igual a 153, y el 75% una frecuencia menor o igual a 166. Los datos tienen un comportamiento asimétrico positivo, y hay existencia de valores atípicos.

Analisis Bivariante

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

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



- En la grafica se oberva 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	81.454484	44.932015	98.148897	-81.554445	2.101288
PresArtRep	44.932015	309.751120	118.848513	-17.090857	3.865638
Colesterol	98.148897	118.848513	2678.285642	-6.031331	2.867507
FCardiaca	-81.554445	-17.090857	-6.031331	525.330492	-8.978651
ST	2.101288	3.865638	2.867507	-8.978651	1.348095

Se realiza la comprobacion de que las variables sean independientes.

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

Se realiza la matriz de diagramas de dispersion.

```
pairs(V.Cuantitativas[,1:5], pch = 19,  cex = 0.5,
      col = "lightblue",
            lower.panel=NULL)
```