Análisis Canónico de correspondencias

Omar Sanchez Hernandez

4/6/2022

Introducción

El análisis canónico es una herramienta multivariante la cual estudia la relación entre dos variables que están divididas en grupos.

Este resulta como una generalización del modelo de regresión múltiple, cuyo objetivo es establecer la relación entre un conjunto de variables predictoras y un conjunto de variables respuesta.

Prepararción de la matriz de datos

Para llevar a cabo el desarrollo de la tecnica se utilizara la matriz de datos de los palmer penguins que contiene informacion sobre las caracteristicas medidas a 3 especies de pingüinos.

```
# Instalar paqueterias
library(tidyverse)
library(readxl)
library(palmerpenguins)
library(knitr )
penguins=as.data.frame(read_excel("penguins.xlsx"))
kable(head(penguins))
```

ID	especie	isla	largo_pico_mm	grosor_pico_mm	largo_aleta_mm	masa_corporal_g	genero	año
i1	Adelie	Torgersen	39.1	18.7	181	3750	male	2007
i2	Adelie	Torgersen	39.5	17.4	186	3800	female	2007
i3	Adelie	Torgersen	40.3	18.0	195	3250	female	2007
i4	Adelie	Torgersen	37.8	18.1	190	3700	female	2007
i5	Adelie	Torgersen	36.7	19.3	193	3450	female	2007
i6	Adelie	Torgersen	39.3	20.6	190	3650	male	2007

Exploracion de la matriz

```
dim(penguins)
```

```
## [1] 344   9
```

```
colnames(penguins)
```

```
## [1] "ID"              "especie"         "isla"            "largo_pico_mm"
## [5] "grosor_pico_mm"  "largo_aleta_mm"  "masa_corporal_g" "genero"
## [9] "año"
```

```
str(penguins)
```

```
## 'data.frame':    344 obs. of  9 variables:
##  $ ID             : chr  "i1" "i2" "i3" "i4" ...
##  $ especie        : chr  "Adelie" "Adelie" "Adelie" "Adelie" ...
##  $ isla           : chr  "Torgersen" "Torgersen" "Torgersen" "Torgersen" ...
##  $ largo_pico_mm  : num  39.1 39.5 40.3 37.8 36.7 39.3 38.9 39.2 34.1 42 ...
##  $ grosor_pico_mm : num  18.7 17.4 18 18.1 19.3 20.6 17.8 19.6 18.1 20.2 ...
##  $ largo_aleta_mm : num  181 186 195 190 193 190 181 195 193 190 ...
##  $ masa_corporal_g: num  3750 3800 3250 3700 3450 ...
##  $ genero         : chr  "male" "female" "female" "female" ...
##  $ año            : num  2007 2007 2007 2007 2007 ...
```

```
anyNA(penguins)
```

```
## [1] FALSE
```

La matriz de datos contiene 344 observaciones con 9 variables en donde sus variables son numericas continuas, categoricas y de texto. La matriz de datos se encuentra limpia de valores faltantes.

Generacion de variables X

```
X <- penguins %>%
  select(grosor_pico_mm, largo_pico_mm) %>%
  scale()
head(X)
```

```
##      grosor_pico_mm largo_pico_mm
## [1,]      0.7863145    -0.8825216
## [2,]      0.1267012    -0.8093460
## [3,]      0.4311381    -0.6629947
## [4,]      0.4818776    -1.1203424
## [5,]      1.0907514    -1.3215754
## [6,]      1.7503647    -0.8459338
```

Generacion de variables Y

```
Y <- penguins %>%
  select(largo_aleta_mm,masa_corporal_g) %>%
  scale()
head(Y)
```

```
##      largo_aleta_mm masa_corporal_g
## [1,]     -1.4166210      -0.5646829
## [2,]     -1.0614850      -0.5022529
## [3,]     -0.4222402      -1.1889828
## [4,]     -0.7773762      -0.6271129
## [5,]     -0.5642946      -0.9392628
## [6,]     -0.7773762      -0.6895429
```

Analisis canonico con un par de variables

Visualizacion de la matriz X

```
# Libreria
library(CCA)
# Analisis
ac<-cancor(X,Y)

# Visualizacion de la matriz X
ac$xcoef
```

```
##                       [,1]       [,2]
## grosor_pico_mm  0.03098538 0.04615243
## largo_pico_mm  -0.03746177 0.04107014
```

Visualizacion de la matriz Y

```
# Visualizacion de la matriz Y
ac$ycoef
```

```
##                         [,1]       [,2]
## largo_aleta_mm  -0.055220261 -0.0951545
## masa_corporal_g  0.001411466  0.1100076
```

Visualizacion de la correlacion canonica

```
ac$cor
```

```
## [1] 0.79268475 0.09867305
```

Obtencion de la matriz de variables canonicas

```
# Se obtiene multiplicando los coeficientes por
# cada una de las variables (X1 y Y1)
ac1_X <- as.matrix(X) %*% ac$xcoef[, 1]
ac1_Y <- as.matrix(Y) %*% ac$ycoef[, 1]
```

#Visualizacion de los primeros 20 datos

```
ac1_X[1:20,]
```

```
##  [1] 0.05742508 0.03424542 0.03819593 0.05690117 0.08330590 0.08592589
##  [7] 0.04464608 0.07088939 0.08225809 0.06113346 0.04117935 0.04432371
## [13] 0.02642463 0.10015624 0.12599695 0.06040849 0.06488291 0.06556776
## [19] 0.08491867 0.05415894
```

```
ac1_Y[1:20,]
```

```
##  [1] 0.07742915 0.05790657 0.02163800 0.04204177 0.02983476 0.04195365
##  [7] 0.07720886 0.02414936 0.02987882 0.04301106 0.05702539 0.08126317
## [13] 0.07253771 0.03829586 0.01189829 0.06165247 0.02199048 0.01599667
## [19] 0.06491373 0.02723438
```

Correlacion canonica entre variable X1 y Y1

```
cor(ac1_X,ac1_Y)
```

```
##           [,1]
## [1,] 0.7926848
```

Verificacion de la correlacion canonica

```
assertthat::are_equal(ac$cor[1],
                      cor(ac1_X,ac1_Y)[1])
```

```
## [1] TRUE
```

Analisis canonico con dos pares de variables

Calculo de las variables X2 y Y2

```
ac2_X <- as.matrix(X) %*% ac$xcoef[, 2]
ac2_Y <- as.matrix(Y) %*% ac$ycoef[, 2]
```

Agregamos las variables generadas a la matriz original de penguins

```
ac_df <- penguins %>%
  mutate(ac1_X=ac1_X,
         ac1_Y=ac1_Y,
         ac2_X=ac2_X,
         ac2_Y=ac2_Y)
```

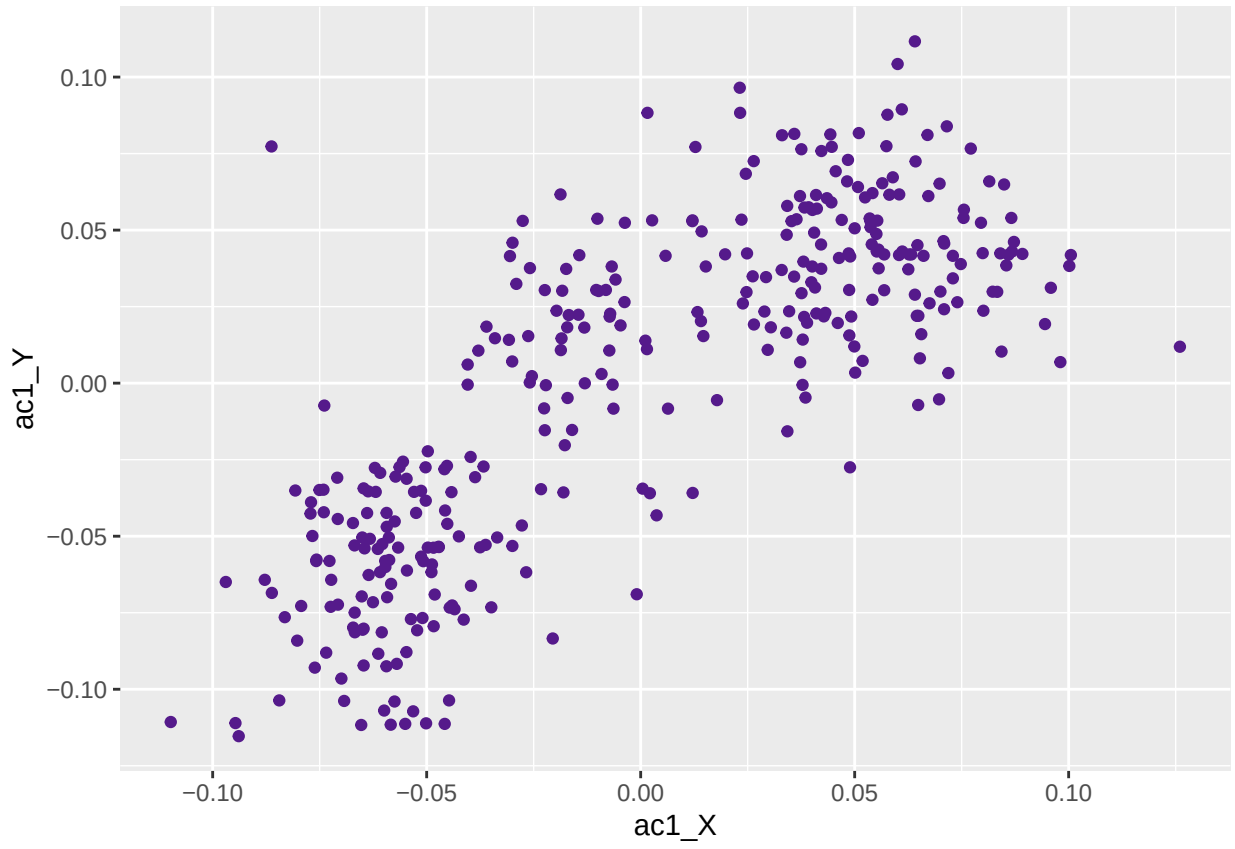
Visualizacion de los nombres de las variables

```
colnames(ac_df)
```

```
##  [1] "ID"              "especie"         "isla"            "largo_pico_mm"
##  [5] "grosor_pico_mm"  "largo_aleta_mm"  "masa_corporal_g" "genero"
##  [9] "año"             "ac1_X"           "ac1_Y"           "ac2_X"
## [13] "ac2_Y"
```

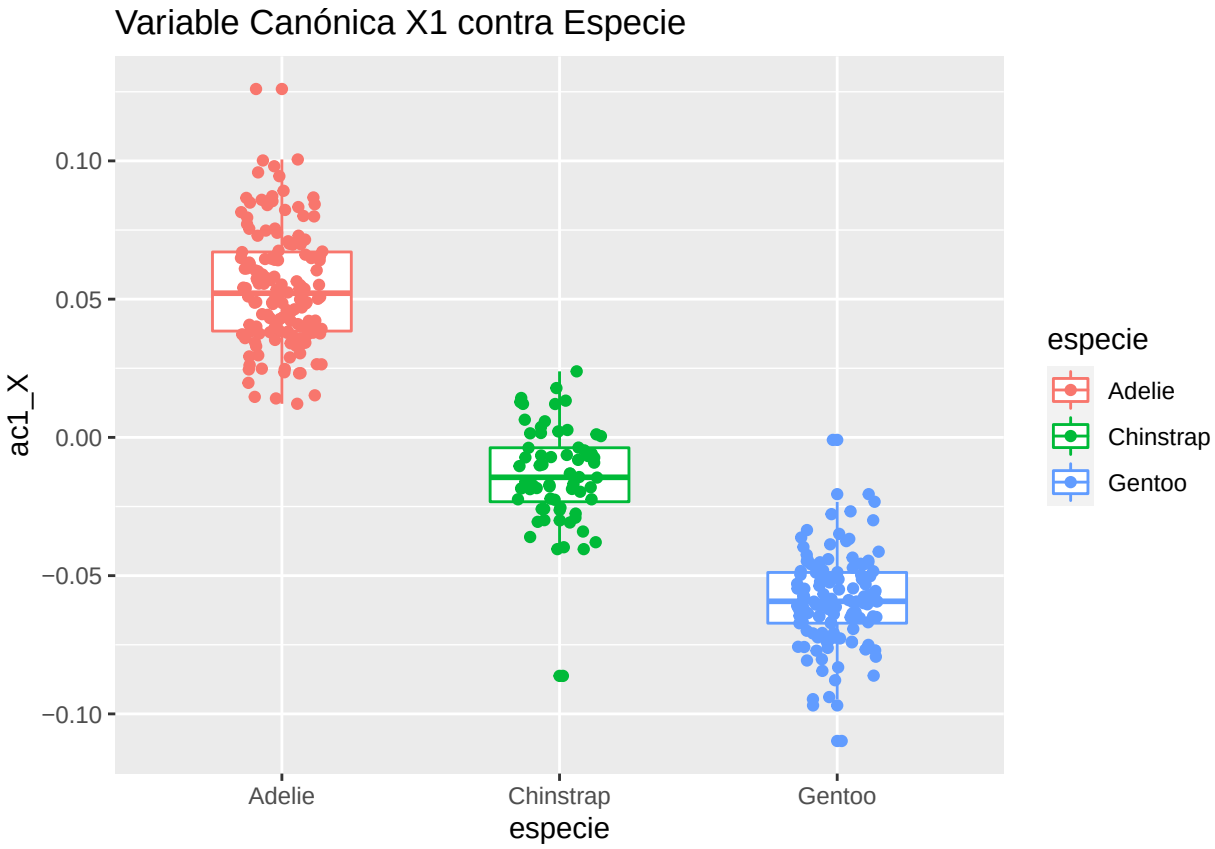
Generacion del grafico scater plot para la visualizacion de X1 y Y1

```
ac_df %>%
  ggplot(aes(x=ac1_X,y=ac1_Y))+
  geom_point(color="#551A8B")
```



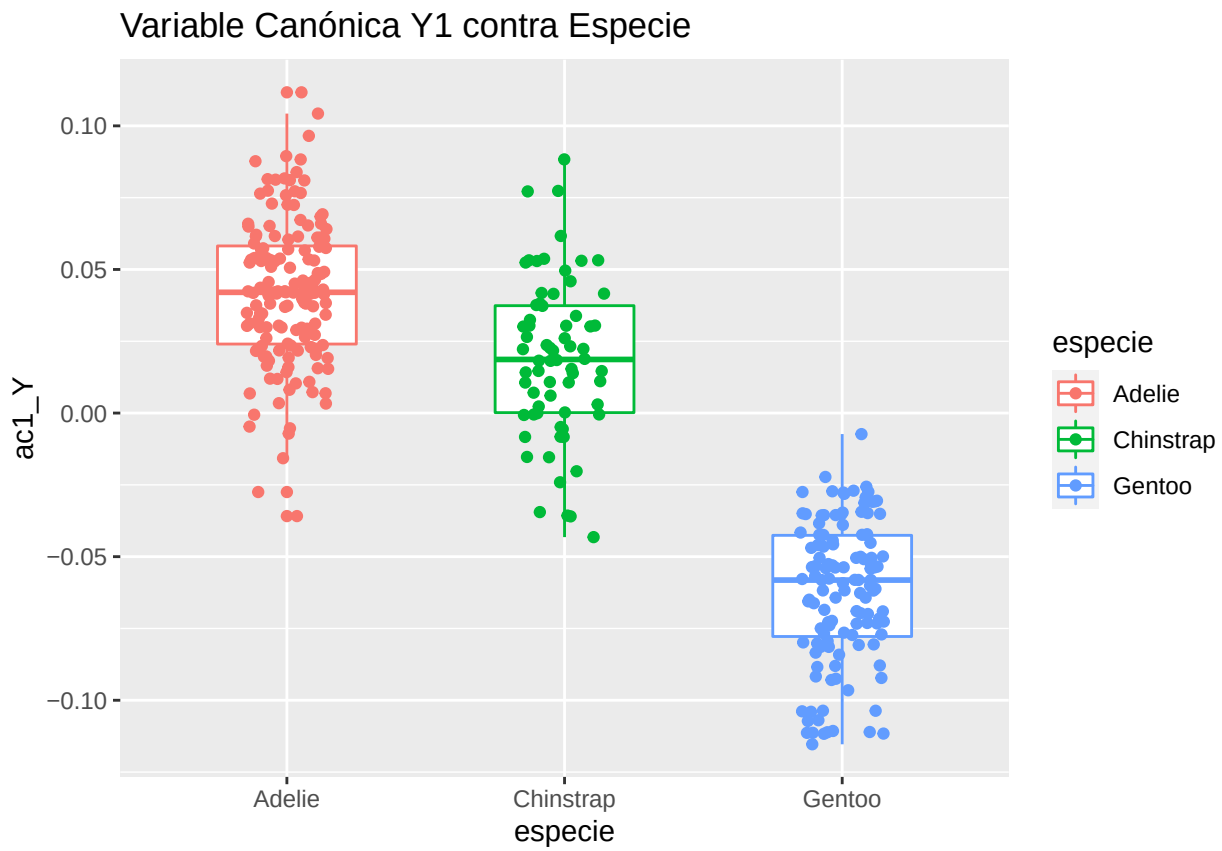
Generacion de un boxplot

```
ac_df %>%
  ggplot(aes(x=especie,y=ac1_X, color=especie))+
  geom_boxplot(width=0.5)+
  geom_jitter(width=0.15)+
  ggtitle("Variable Canónica X1 contra Especie")
```

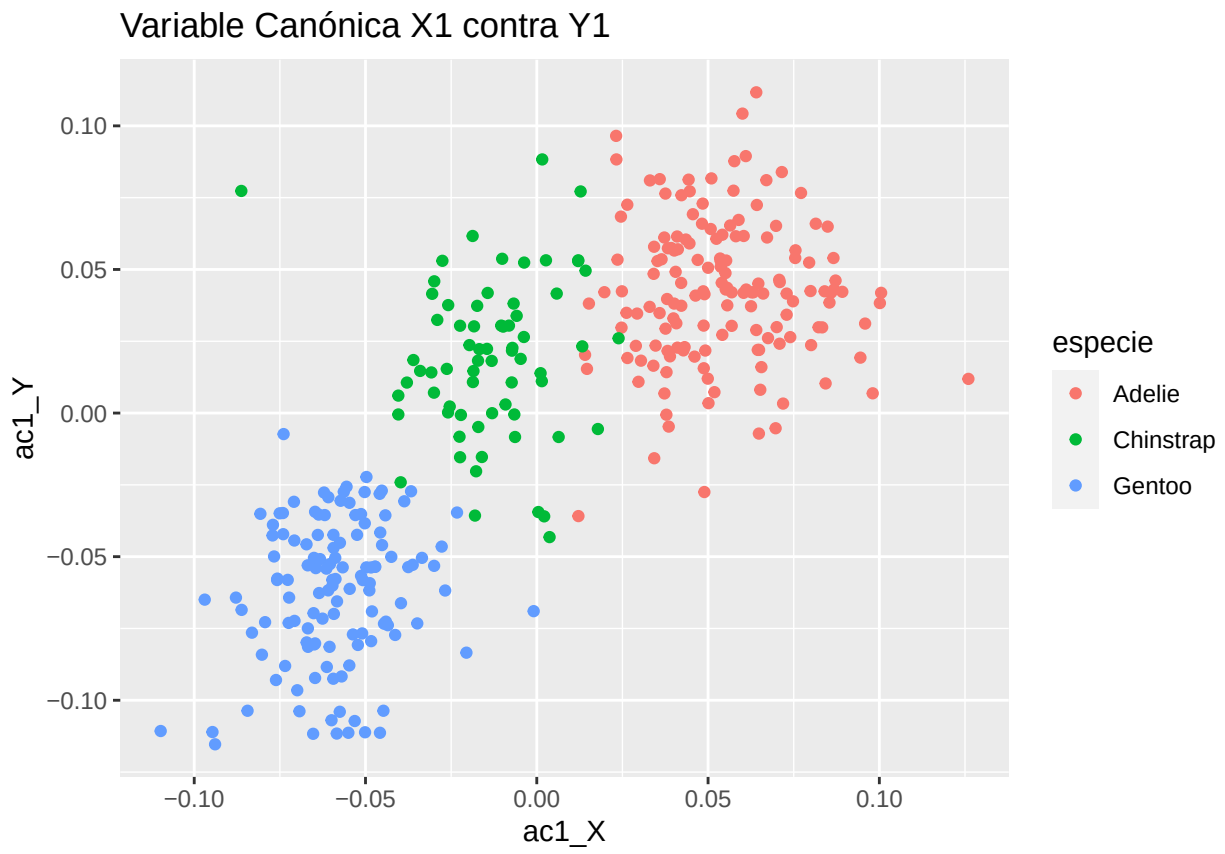


Interpretación: se observa una correlacion entre la variable canónica X1 y la variable latente Especie.

```
ac_df %>%
  ggplot(aes(x=especie,y=ac1_Y, color=especie))+
  geom_boxplot(width=0.5)+
  geom_jitter(width=0.15)+
  ggtitle("Variable Canónica Y1 contra Especie")
```

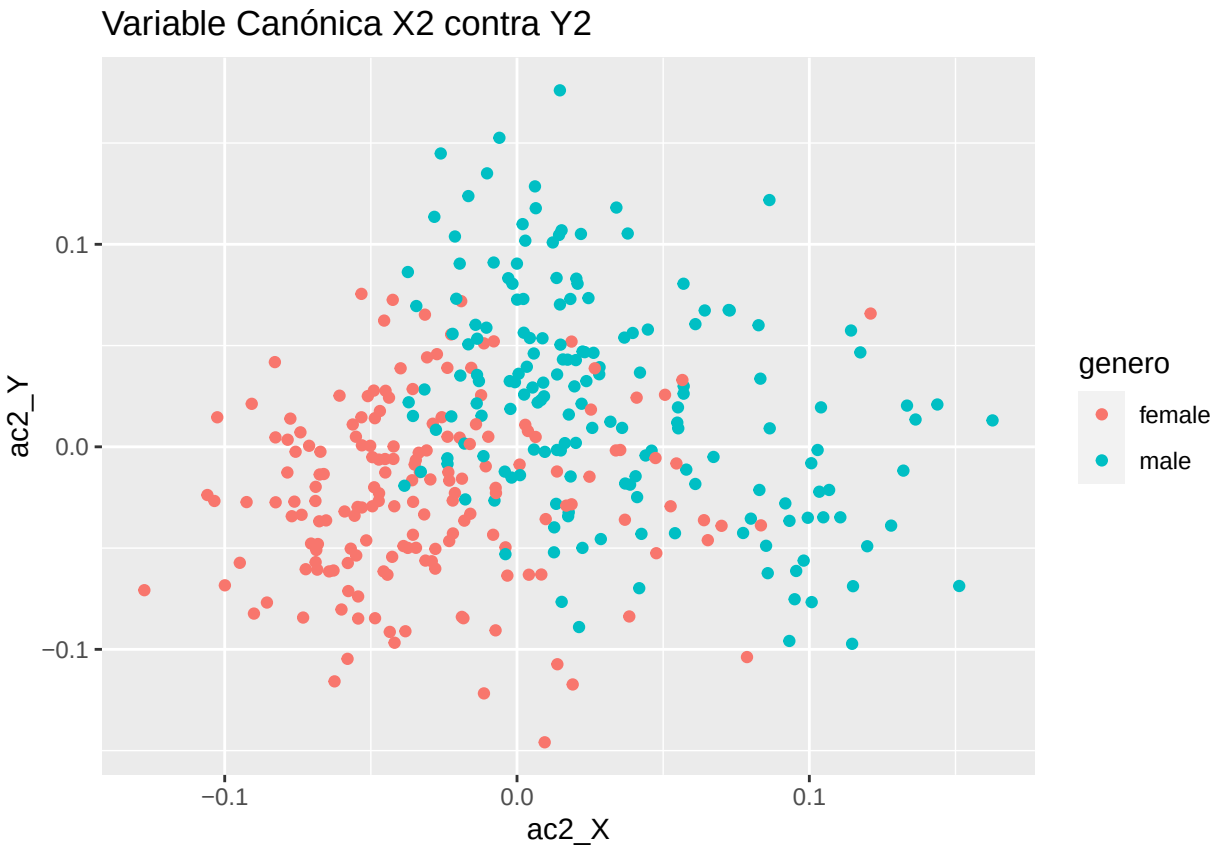



```
ac_df %>%
  ggplot(aes(x=ac1_X,y=ac1_Y, color=especie))+
  geom_point()+
  ggtitle("Variable Canónica X1 contra Y1")
```



Scatterplot con las variables canonicas X2 y Y2 separadas por genero.

```
ac_df %>%
  ggplot(aes(x=ac2_X,y=ac2_Y, color=genero))+
  geom_point()+
  ggtitle("Variable Canónica X2 contra Y2")
```



Interpretacion: No de identifica correlacion entre el conjunto de variables X2 y Y2 separadas por genero.