Heart dataset descriptive analysis

Julián Alejandro Úsuga Ortiz

3/23/2022

Sobre que son estos datos?

Esta base de datos llamada **heart** fue obtenida de la libreria de R **catdata**

Contiene muestras realizadas a hombres en una zona de alto riesgo de enfermedad de corazón (cardiopatía) ubicada en Cabo Occidental, Sudafrica. Este estudio contiene 10 variables, las siguientes:

- **age**: Edad en el momento del estudio (Discreta ordinal pero la trataré como Continua) *
- **alcohol**: Consumo de alcohol (Continua) *
- **tobacco**: Consumo acumulativo de tabaco (Continua) *
- **obesity**: Obesidad (Continua)
- **typea**: Conducta Tipo-A (Continua)
- **famhist**: Historia familiar de enfermedad al corazón (Binaria, 1:Si, 0:No) *
- **adiposity**: Adiposidad / Ganancia de grasa (Continua)
- **ldl**: Colesterol de lipoproteínas de baja densidad (Continua)
- **sbp**: Presión sanguínea sistólica (Continua) *
- **y**: ¿Tiene enfermedad coronaria? (Binaria, 1:Si, 0:No) *

(* = va a ser analizada)

1. Estadistica descriptiva

Como podemos ver en la siguiente tabla, tenemos 4 variables continuas y 2 variables discretas.

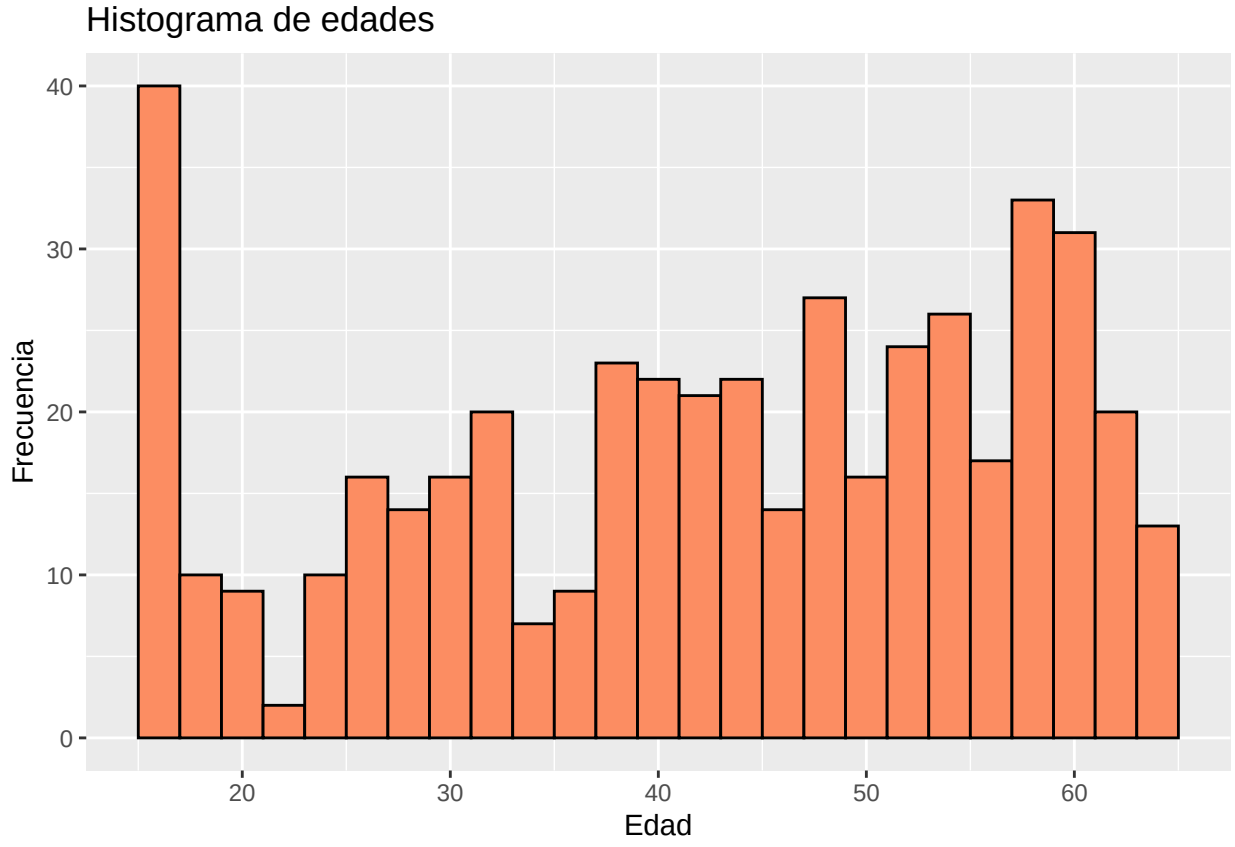
- Las edades observadas van desde los 15 hasta los 64 años
- El consumo de alcohol oscila entre 0 y 147
- El consumo de tabaco oscila entre 0 y 31.2
- La presión sanguínea oscila entre 101 y 218; Hay personas con una presión sanguínea elevada.
- El ~41.55% de las personas tiene historial familiar de enfermedad al corazón.
- El ~34.63% de las personas sufren de enfermedad coronaria al corazón.

```
##       age           alcohol          tobacco            sbp        famhist
##  Min.   :15.00   Min.   :  0.00   Min.   : 0.0000   Min.   :101.0   0:270
##  1st Qu.:31.00   1st Qu.:  0.51   1st Qu.: 0.0525   1st Qu.:124.0   1:192
##  Median :45.00   Median :  7.51   Median : 2.0000   Median :134.0
##  Mean   :42.82   Mean   : 17.04   Mean   : 3.6356   Mean   :138.3
##  3rd Qu.:55.00   3rd Qu.: 23.89   3rd Qu.: 5.5000   3rd Qu.:148.0
##  Max.   :64.00   Max.   :147.19   Max.   :31.2000   Max.   :218.0
##  y
##  0:302
##  1:160
##
##
##
##
```

a) & b)

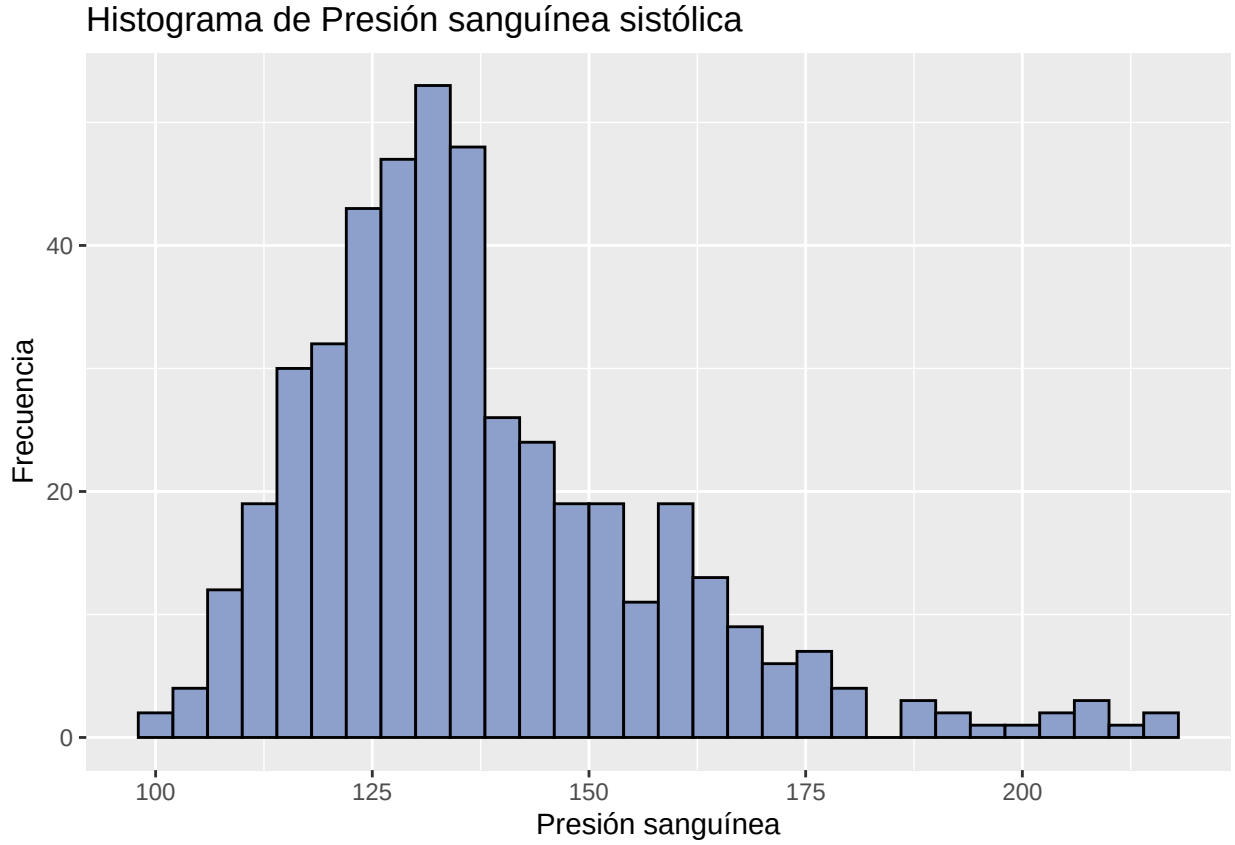
Edad

Como podemos ver en el siguiente gráfico la distribución de las edades es casi plana, pero hay una clara mayoría para el grupo de edades de personas mayores de ~40.



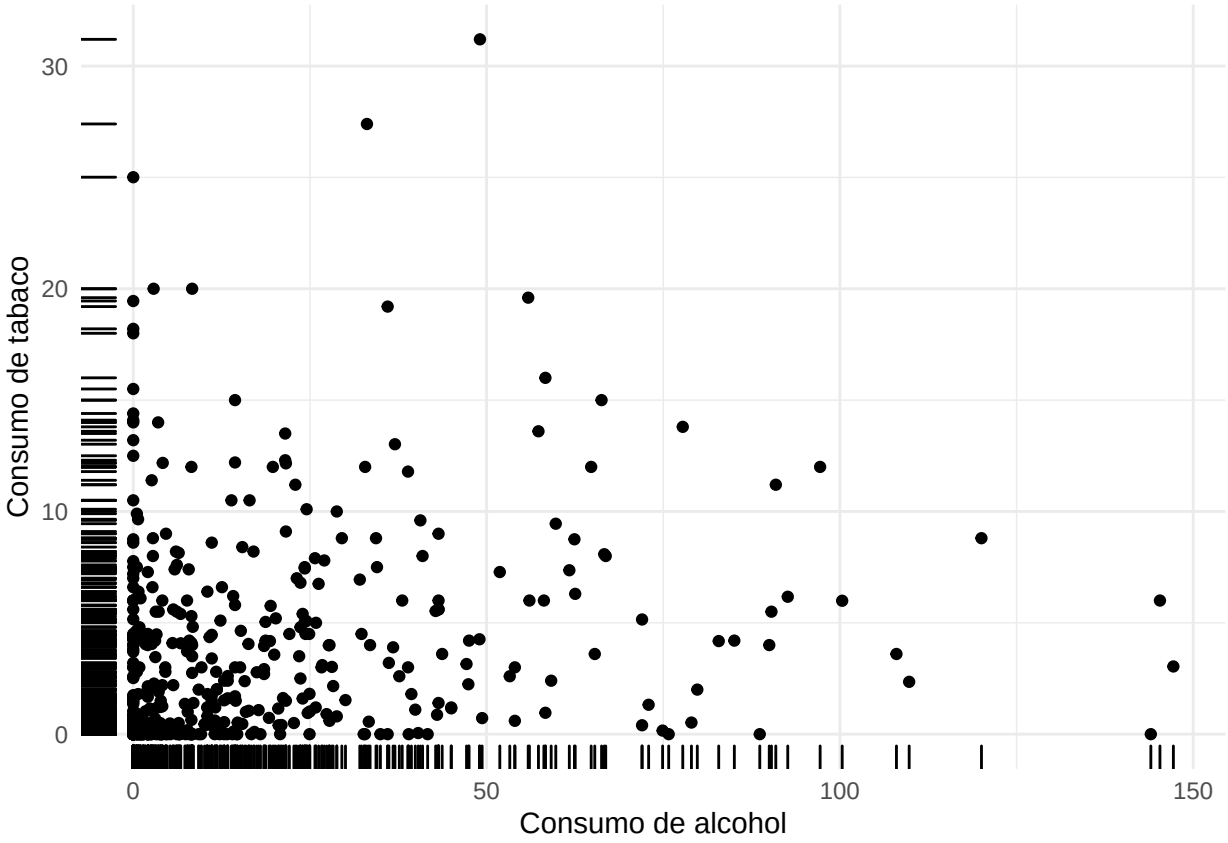
Presión sanguínea sistólica

Como podemos ver en la presión sanguínea, esta distribución parece ser normal pero con una cola larga a la derecha, en presiones sanguíneas altas, lo cual sabemos que es malo. La moda de los datos es ~134.



Alcohol y Tabaco

En este gráfico podemos ver que no hay mucha relación entre el consumo de trabajo y el consumo de alcohol, sin embargo podemos ver que existe un bajo con sumo de estas dos sustancias en los sujetos medidos en nuestra base de datos.

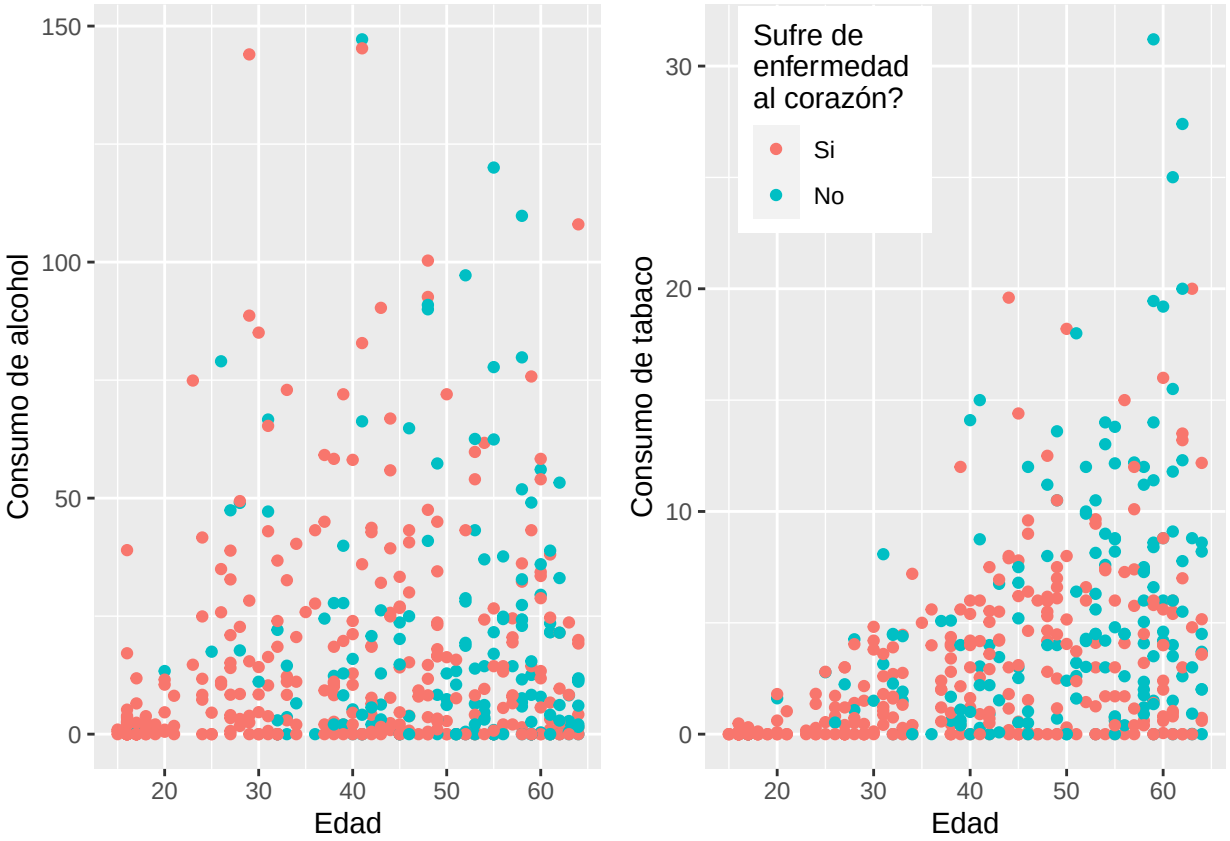


c)

Edad, alcohol y consumo de tabaco

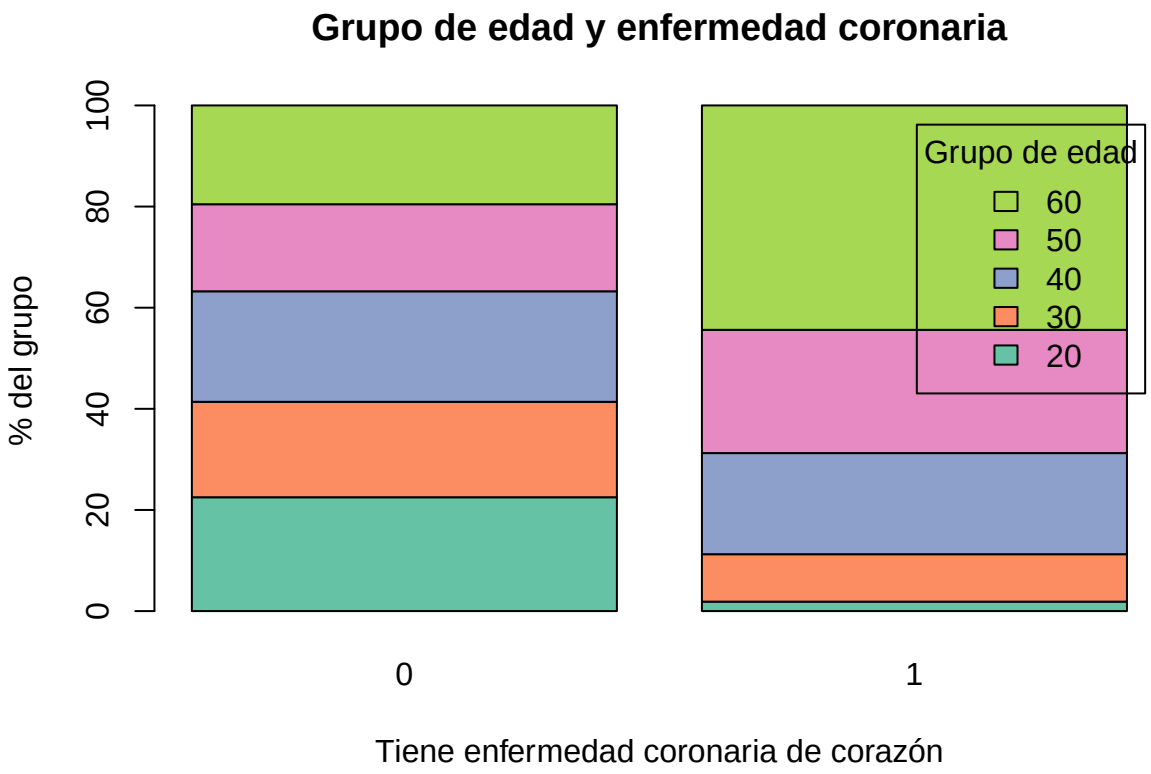
Se sospecha que puede haber una relación positiva entre la edad y estas dos variables por lo que graficamos para confirmar (Los puntos en rojo son las personas que tienen la enfermedad del corazón).

De este gráfico podemos observar varias cosas, que a medida que las personas tienden a envejecer su consumo de tabaco aumenta y que las personas jóvenes no tienden a sufrir de esta enfermedad al corazón.



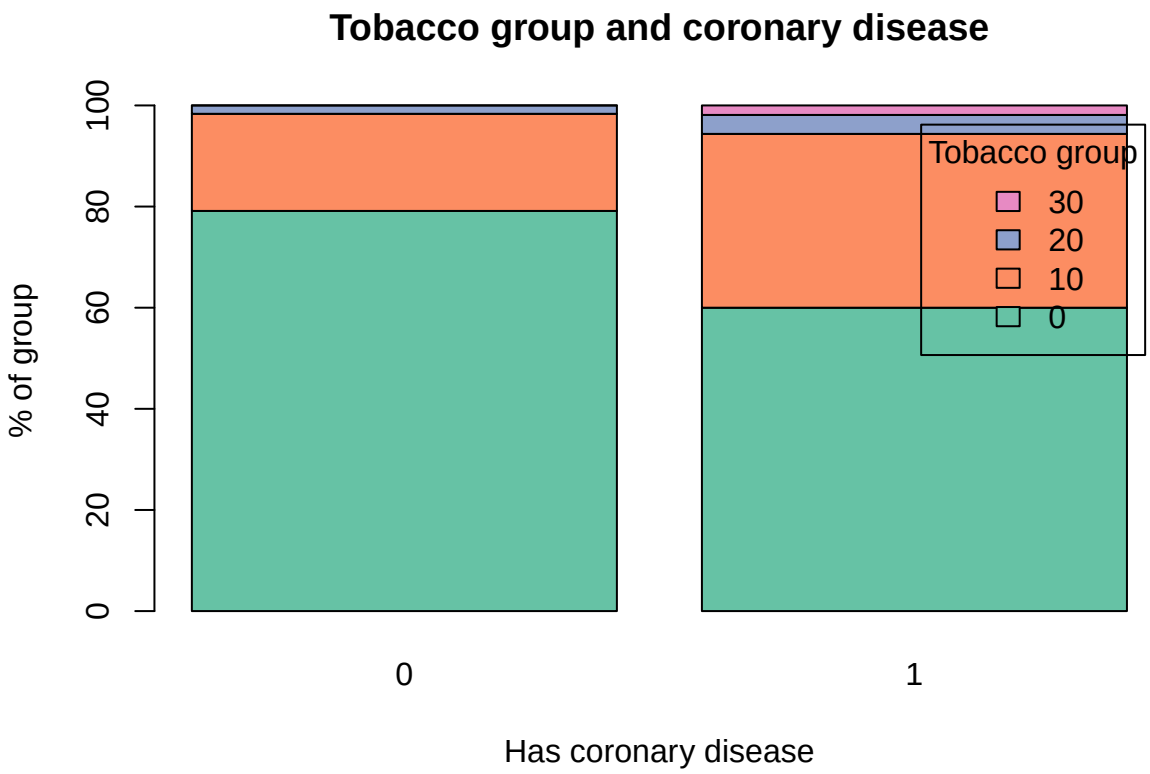
De la siguiente gráfica podemos sacar la conclusión de que el grupo de personas mayores a 45 son la mayoría en el grupo de personas que tienen la enfermedad al corazón.

La agrupación de edades se hace aproximando al numero múltiplo de 10 mas cercano.



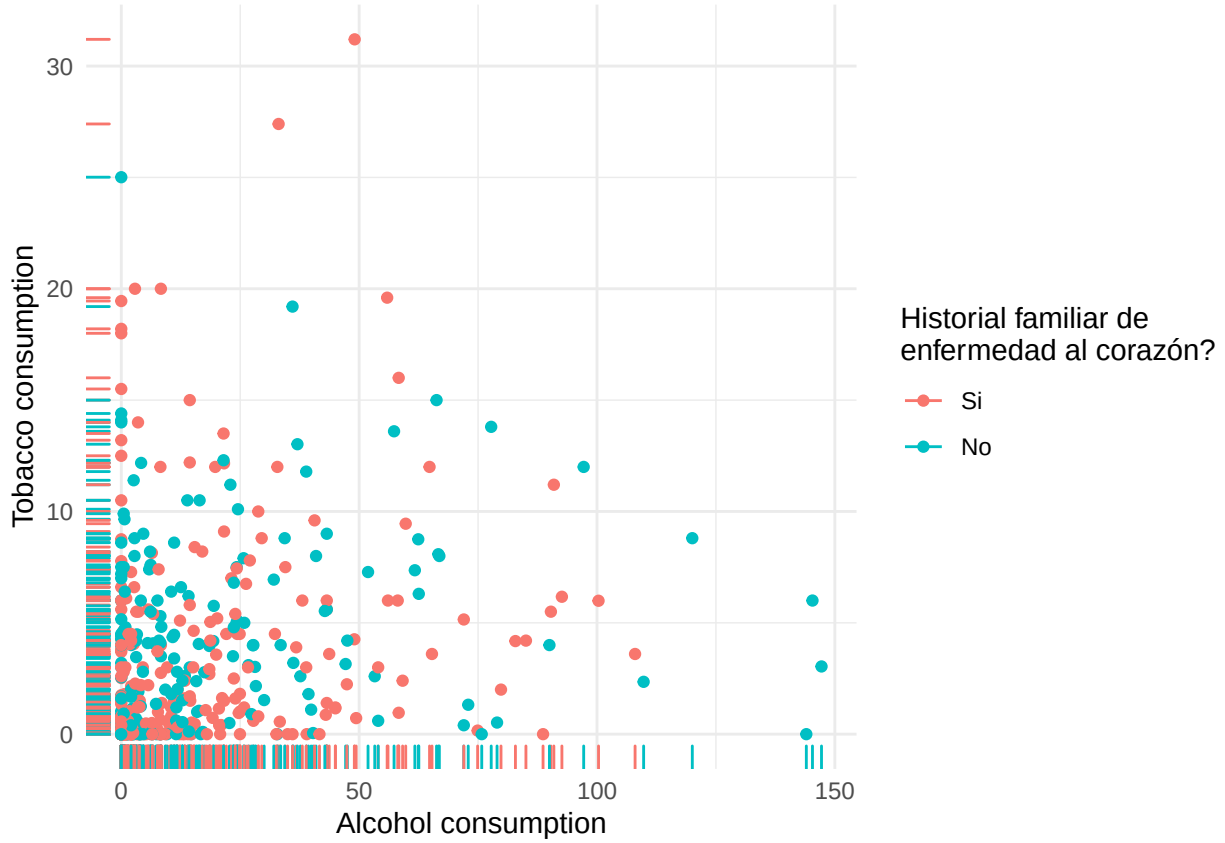
Edad y enfermedad coronaria del corazón

En este gráfico podemos ver que si comparamos a los grupos que tienen o no la enfermedad y los comparamos con su consumo de tabaco, el porcentaje de personas con la enfermedad tienen mayor consumo de tabaco que las personas que no sufren de esta enfermedad al corazón.



Alcohol, tabaco y historial familiar de enfermedad al corazón

Podemos pensar que es posible que haya cierto tipo de relación entre los hábitos de consumo de tabaco y alcohol con el historial familiar de enfermedad al corazón.



2.

Vector de medias $\underline{\bar{x}}$

Nuestro vector de medias nos dice el promedio muestral para cada variable, por ejemplo, podemos observar que la edad promedio de las personas en nuestro estudio es de 42.81 años.

	x
age	42.816
alcohol	17.044
tobacco	3.636
sbp	138.327

Vector de Varianzas-Covarianzas S

En esta tabla podemos ver nuestra matriz de varianzas y covarianzas, podemos interpretarla como: - La diagonal de esta matriz son las varianzas de cada variable, por ejemplo, la varianza de la edad es 213.42

- Los elementos que no estan en la diagonal van a ser nuestras covarianzas entre dos de nuestras variables, por ejemplo, la covarianza entre el consumo de alcohol y la edad es de 36.166, ademas, podemos ver es lo mismo que la covarianza entre la edad y el consumo de alcohol, por propiedades de la covarianza.

	age	alcohol	tobacco	sbp
age	213.422	36.166	30.217	116.410
alcohol	36.166	599.322	22.580	70.296

	age	alcohol	tobacco	sbp
tobacco	30.217	22.580	21.096	19.981
sbp	116.410	70.296	19.981	420.099

Vector de Varianzas-Covarianzas muestrales S_n

	age	alcohol	tobacco	sbp
age	212.960	36.088	30.151	116.158
alcohol	36.088	598.025	22.531	70.144
tobacco	30.151	22.531	21.050	19.938
sbp	116.158	70.144	19.938	419.190

Matriz de correlación muestral

Esta matriz nos dice que tanta correlación lineal existe entre las variables, podemos concluir lo siguiente:

- La correlación va desde -1 a 1
- La correlación de una variable con si misma siempre va a ser igual a 1.
- El par de variables que mas correlación lineal tiene son la edad y el consumo de tabaco, lo cual habíamos graficado anteriormente; esto quiere decir que a medida que aumente la edad se tiende a aumentar el consumo de tabaco.

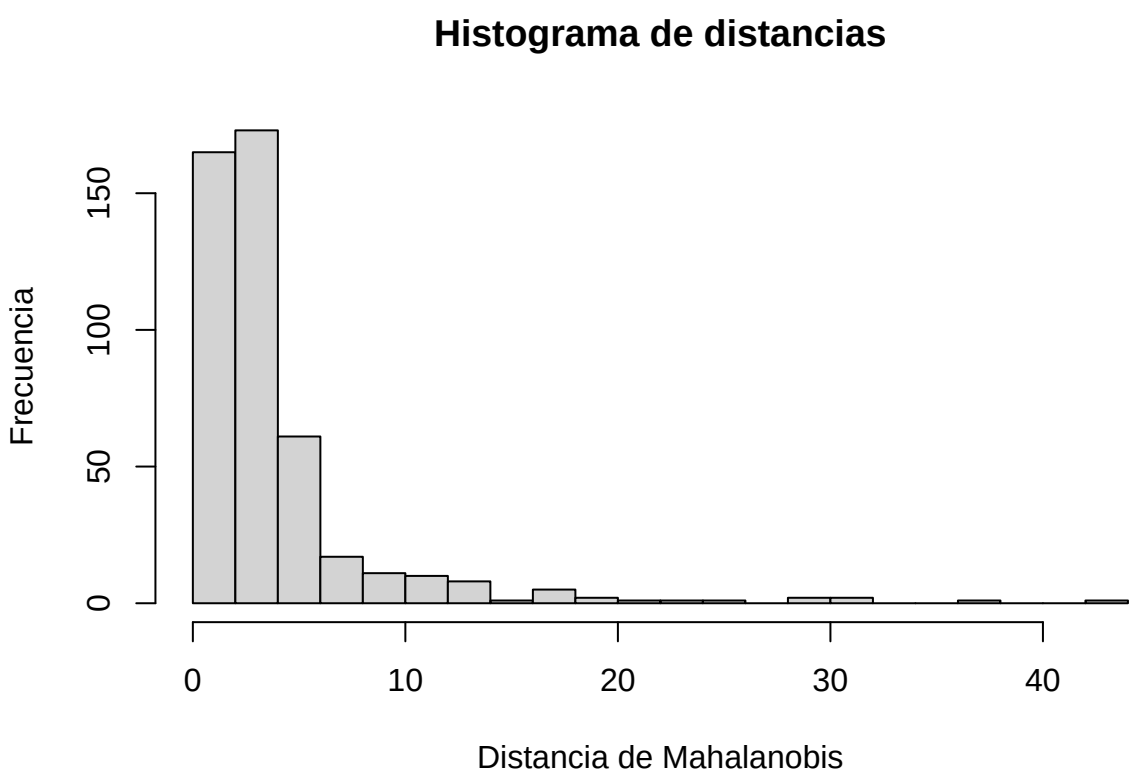
	age	alcohol	tobacco	sbp
age	1.000	0.101	0.450	0.389
alcohol	0.101	1.000	0.201	0.140
tobacco	0.450	0.201	1.000	0.212
sbp	0.389	0.140	0.212	1.000

3.

Podemos ver que la varianza total y la varianza generalizada de nuestros datos es de 1253.939 y 726644866.11 respectivamente.

?

Varianza total	Varianza generalizada
1253.94	726644866.11




```
##        age    alcohol    tobacco        sbp
##  42.816017  17.044394   3.635649 138.326840
```

```
##       241
## 0.1197206
```

```
##     age alcohol tobacco sbp famhist y
## 241  45   20.17     5.2 140       0 1
```

```
##      115
## 43.04026
```

```
##     age alcohol tobacco sbp famhist y
## 115  59   49.06    31.2 116       0 1
```