

# Examen

J. Patricio Parada G.

28/08/2020

## Índice

<b>1. El Paro cardíaco</b>	<b>1</b>
<b>2. El Dataset</b>	<b>1</b>
2.1. Columnas . . . . .	1
2.2. Estructura de los datos . . . . .	2
2.3. Tipo de variables . . . . .	2
2.4. Resumen de datos . . . . .	2
2.5. Filtrado . . . . .	5
2.6. Visualización . . . . .	6
2.7. Análisis . . . . .	11
2.8. Probabilidades . . . . .	13

## 1. El Paro cardíaco

Comúnmente llamado ataque cardíaco, el paro cardíaco es una condición riesgosa y virtualmente mortífera que pone fin a millones de vidas al año. Es una de las causas de muerte más frecuentes en humanos y se debe a variados factores; puede ser a consecuencia del estilo de vida llevado o debido a otras afecciones o enfermedades.

El conjunto de datos anexo presenta 12 factores que eventualmente proporcionan información respecto a si un paciente es candidato a sufrir un ataque cardíaco.

## 2. El Dataset

El conjunto de datos adjunto corresponde a

```
df <- read.csv("CRP_dataset_clean.csv", stringsAsFactors = FALSE)
```

### 2.1. Columnas

Las variables incluídas en el conjunto de datos corresponden a

```
colnames(df)
```

```
## [1] "Age" "Gender"
## [3] "Chain_smoker" "Consumes_other_tobacco_products"
## [5] "HighBP" "Obese"
## [7] "Diabetes" "Metabolic_syndrome"
## [9] "Use_of_stimulant_drugs" "Family_history"
## [11] "History_of_preeclampsia" "CABG_history"
## [13] "Respiratory_illness" "UnderRisk"
```

en donde:

- Age: Edad.
- Gender: género, 1 para masculino y 2 femenino.

- **Chain\_smoker**: fumador, 0 para no fumador y 1 en caso contrario.
- **Consumes\_other\_tobacco\_products**: consumo de otros productos derivados del tabaco. 1 para consumidor, 0 para no consumidor.
- **HighBP**: hipertensión. 0 persona sin hipertensión, 0 indica afección.
- **Obese**: obesidad. 0 para rangos de peso normales, 1 para obesidad.
- **Diabetes**: diabetes. 1 para diabético, 0 para no diabético.
- **Metabolic\_syndrome**: hídrome metabólico. 0 para ausencia, 1 indica presencia.
- **Use\_of\_stimulant\_drugs**: hso de drogas estimulantes. 0 para no consumidor, 1 para consumidor.
- **Family\_history**: historial familiar de ataques cardíacos. 1 para antecedentes, 0 en caso contrario.
- **History\_of\_preeclampsia**: historial de preeclampsia. 1 casi afirmativo, 0 negativo.
- **CABG\_history**: historial de cirugía de bypass de la arteria coronaria. 1 para operado, 0 en caso adverso.
- **Respiratory\_illness**: enfermedades respiratorias. 0 no presenta, 1 presenta.
- **UnderRisk**: riesgoso. yes para sí, no para no.

## 2.2. Estructura de los datos

Las observaciones se encuentran estructuradas como

```
str(df)
```

```
## 'data.frame':    889 obs. of  14 variables:
## $ Age                : int   84 55 80 40 45 78 77 56 999 75 ...
## $ Gender              : int   1 1 1 1 1 2 1 2 1 1 ...
## $ Chain_smoker        : int   1 0 0 0 0 0 0 0 0 0 ...
## $ Consumes_other_tobacco_products: int  1 1 1 1 0 1 1 1 1 1 ...
## $ HighBP              : int   0 0 0 0 0 0 0 0 0 0 ...
## $ Obese               : int   1 1 1 1 0 1 0 1 1 0 ...
## $ Diabetes            : int   0 0 0 0 0 0 0 0 0 0 ...
## $ Metabolic_syndrome  : int   0 0 0 0 1 0 0 0 0 0 ...
## $ Use_of_stimulant_drugs : int   0 0 0 0 1 0 1 0 0 1 ...
## $ Family_history      : int   1 1 1 1 0 1 1 1 1 1 ...
## $ History_of_preeclampsia : int   0 0 0 0 0 0 0 0 0 0 ...
## $ CABG_history        : int   0 0 0 0 0 0 0 0 0 0 ...
## $ Respiratory_illness  : int   0 0 0 0 0 0 0 0 0 0 ...
## $ UnderRisk           : chr  "no" "no" "no" "no" ...
```

Se puede observar el tipo de dato de cada parámetro observado, en donde es posible ver que casi todos son de tipo **int**, a excepción del parámetro **UnderRisk**, el cual es de tipo **char**, lo cual es esperable a partir de la descripción de las variables dada anteriormente.

También se indica el número de observaciones, que corresponden a 889.

## 2.3. Tipo de variables

Para efectos prácticos, serán consideradas todas las variables como variables cualitativas, a excepción de la variable **Age**. Durante el desarrollo del presente texto, el tipo de dato presente en el dataframe será ajustado para facilitar operaciones.

## 2.4. Resumen de datos

Antes de realizar cualquier tipo de limpieza de los datos, se procede a hacer un resumen estadístico de los datos en bruto:

```
summary(df)
```

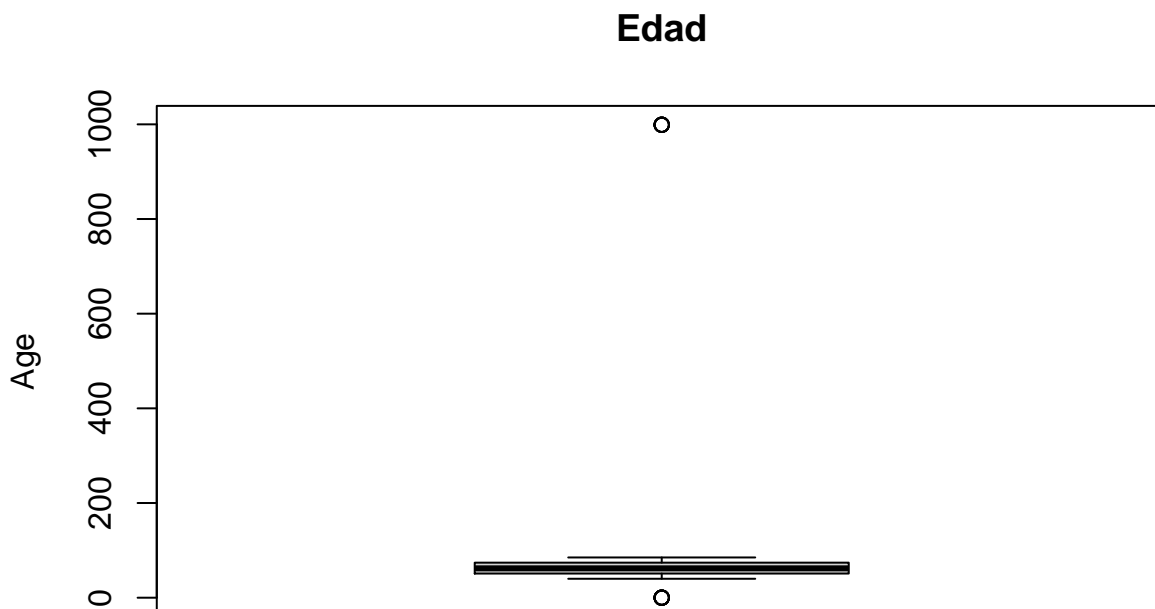
```
##      Age                Gender                Chain_smoker
## Min.   : 0.0000000  Min.   :1.00000000  Min.   :0.000000000
## 1st Qu.: 51.0000000  1st Qu.:1.00000000  1st Qu.:0.000000000
## Median : 62.0000000  Median :1.00000000  Median :0.000000000
## Mean   : 67.1023622  Mean   :1.30033746  Mean   :0.120359955
```

```
## 3rd Qu.: 74.0000000 3rd Qu.:2.00000000 3rd Qu.:0.000000000
## Max. :999.0000000 Max. :2.00000000 Max. :1.000000000
## Consumes_other_tobacco_products HighBP Obese
## Min. :0.000000000 Min. :0.000000000 Min. :0.000000000
## 1st Qu.:1.000000000 1st Qu.:0.000000000 1st Qu.:1.000000000
## Median :1.000000000 Median :0.000000000 Median :1.000000000
## Mean :0.838020247 Mean :0.0866141732 Mean :0.919010124
## 3rd Qu.:1.000000000 3rd Qu.:0.000000000 3rd Qu.:1.000000000
## Max. :1.000000000 Max. :1.000000000 Max. :1.000000000
## Diabetes Metabolic_syndrome Use_of_stimulant_drugs
## Min. :0.000000000 Min. :0.000000000 Min. :0.000000000
## 1st Qu.:0.000000000 1st Qu.:0.000000000 1st Qu.:0.000000000
## Median :0.000000000 Median :0.000000000 Median :0.000000000
## Mean :0.0551181102 Mean :0.0427446569 Mean :0.0821147357
## 3rd Qu.:0.000000000 3rd Qu.:0.000000000 3rd Qu.:0.000000000
## Max. :1.000000000 Max. :1.000000000 Max. :1.000000000
## Family_history History_of_preeclampsia CABG_history
## Min. :0.000000000 Min. :0.000000000 Min. :0.000000000
## 1st Qu.:1.000000000 1st Qu.:0.000000000 1st Qu.:0.000000000
## Median :1.000000000 Median :0.000000000 Median :0.000000000
## Mean :0.92575928 Mean :0.0179977503 Mean :0.0213723285
## 3rd Qu.:1.000000000 3rd Qu.:0.000000000 3rd Qu.:0.000000000
## Max. :1.000000000 Max. :1.000000000 Max. :1.000000000
## Respiratory_illness UnderRisk
## Min. :0.000000000 Length:889
## 1st Qu.:0.000000000 Class :character
## Median :0.000000000 Mode :character
## Mean :0.0326209224
## 3rd Qu.:0.000000000
## Max. :1.000000000
```

en donde se puede apreciar que las variable están dentro de los valores esperados. La única variable que presenta un valores atípicos sería Age.

Realizando un boxplot a la variable

```
boxplot(df$Age, main = "Edad", "ylab" = "Age")
```



en donde se observan valores extremos demasiado lejanos. Ordenando los datos de dicha columna de manera ordenada

```
sort(df$Age)
```

```
## [1] 0 0 0 0 0 0 0 40 40 40 40 40 40 40 40 40 40 40
## [19] 40 40 40 40 40 40 40 40 41 41 41 41 41 41 41 41 41 41
## [37] 41 41 41 41 41 41 41 41 41 41 41 41 41 41 42 42 42 42
## [55] 42 42 42 42 42 42 42 42 42 42 42 43 43 43 43 43 43 43
## [73] 43 43 43 43 43 43 43 43 43 43 43 43 44 44 44 44 44 44
## [91] 44 44 44 44 44 44 44 44 44 44 44 44 44 44 44 44 44 44
## [109] 44 45 45 45 45 45 45 45 45 45 45 45 45 45 45 45 45 45
## [127] 45 45 45 45 45 45 46 46 46 46 46 46 46 46 46 46 46 46
## [145] 46 46 46 46 46 46 47 47 47 47 47 47 47 47 47 47 47 47
## [163] 47 47 47 47 47 48 48 48 48 48 48 48 48 48 48 48 48 48
## [181] 48 48 48 49 49 49 49 49 49 49 49 49 49 49 49 49 49 49
## [199] 49 49 49 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50
## [217] 50 50 50 51 51 51 51 51 51 51 51 51 51 51 51 51 51 52
## [235] 52 52 52 52 52 52 52 52 52 52 52 52 52 52 52 52 52 52
## [253] 52 52 52 52 52 52 52 52 52 52 53 53 53 53 53 53 53 53
## [271] 53 53 53 53 53 53 53 53 53 53 53 53 53 53 53 53 54 54
## [289] 54 54 54 54 54 54 54 54 54 54 54 54 54 54 54 54 54 54
## [307] 55 55 55 55 55 55 55 55 55 55 55 55 55 55 55 55 55 55
## [325] 55 55 56 56 56 56 56 56 56 56 56 56 56 56 56 56 57 57
## [343] 57 57 57 57 57 57 57 57 57 57 57 57 57 57 57 57 57 57
## [361] 57 57 57 57 57 58 58 58 58 58 58 58 58 58 58 59 59 59
## [379] 59 59 59 59 59 59 59 59 59 59 59 59 59 59 59 59 60 60
## [397] 60 60 60 60 60 60 60 60 60 60 60 60 60 60 60 60 60 61
## [415] 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61 61
## [433] 61 61 62 62 62 62 62 62 62 62 62 62 62 62 62 62 62 62
## [451] 62 62 62 62 63 63 63 63 63 63 63 63 63 63 63 63 63 63
## [469] 63 63 63 63 63 63 64 64 64 64 64 64 64 64 64 64 64 64
## [487] 64 64 64 64 64 64 64 64 64 64 64 64 64 64 65 65 65 65
## [505] 65 65 65 65 65 65 65 65 65 65 65 65 65 65 66 66 66 66
## [523] 66 66 66 66 66 66 66 66 66 66 66 66 66 66 67 67 67 67
## [541] 67 67 67 67 67 67 67 67 67 68 68 68 68 68 68 68 68 68
## [559] 68 68 68 68 68 69 69 69 69 69 69 69 69 69 69 69 69 69
## [577] 69 69 69 70 70 70 70 70 70 70 70 70 70 70 70 70 70 70
## [595] 70 70 70 70 70 70 70 71 71 71 71 71 71 71 71 71 71 71
## [613] 71 71 71 71 71 71 71 71 71 72 72 72 72 72 72 72 72 72
## [631] 72 72 72 72 72 72 72 72 72 72 73 73 73 73 73 73 73 73
## [649] 73 73 73 73 73 73 73 73 73 74 74 74 74 74 74 74 74 74
## [667] 74 74 74 74 74 74 74 74 74 74 75 75 75 75 75 75 75 75
## [685] 75 75 75 75 75 75 75 75 75 75 75 75 75 75 75 75 76 76
## [703] 76 76 76 76 76 76 76 76 76 76 76 76 76 76 77 77 77 77
## [721] 77 77 77 77 77 77 77 77 77 77 78 78 78 78 78 78 78 78
## [739] 78 79 79 79 79 79 79 79 79 79 79 79 79 79 79 79 79 79
## [757] 79 79 79 79 80 80 80 80 80 80 80 80 80 80 80 80 80 80
## [775] 80 80 80 80 80 80 80 81 81 81 81 81 81 81 81 81 81 81
## [793] 81 81 81 81 81 81 81 81 81 81 81 81 82 82 82 82 82 82
## [811] 82 82 82 82 82 82 82 82 82 82 82 82 82 83 83 83 83 83
## [829] 83 83 83 83 83 83 83 83 83 83 83 83 83 83 84 84 84 84
## [847] 84 84 84 84 84 84 84 84 84 84 84 84 84 84 84 84 84 84
## [865] 84 85 85 85 85 85 85 85 85 85 85 85 85 85 85 85 85 85
## [883] 85 85 999 999 999 999 999
```

se puede ver que el valor extremo 999 no tiene sentido, en consecuencia inválido, ya que es imposible que un humano viva dicha cantidad de años. El otro valor extremo, 0, es técnicamente válido, ya que podría representar la edad de neonatos menores a 1 año, no será considerado como tal por encontrarse demasiado alejado del grueso de los datos.

```

outliers <- boxplot.stats(df$Age)$out
df$Age <- ifelse(df$Age %in%
                outliers,
                NA, df$Age)

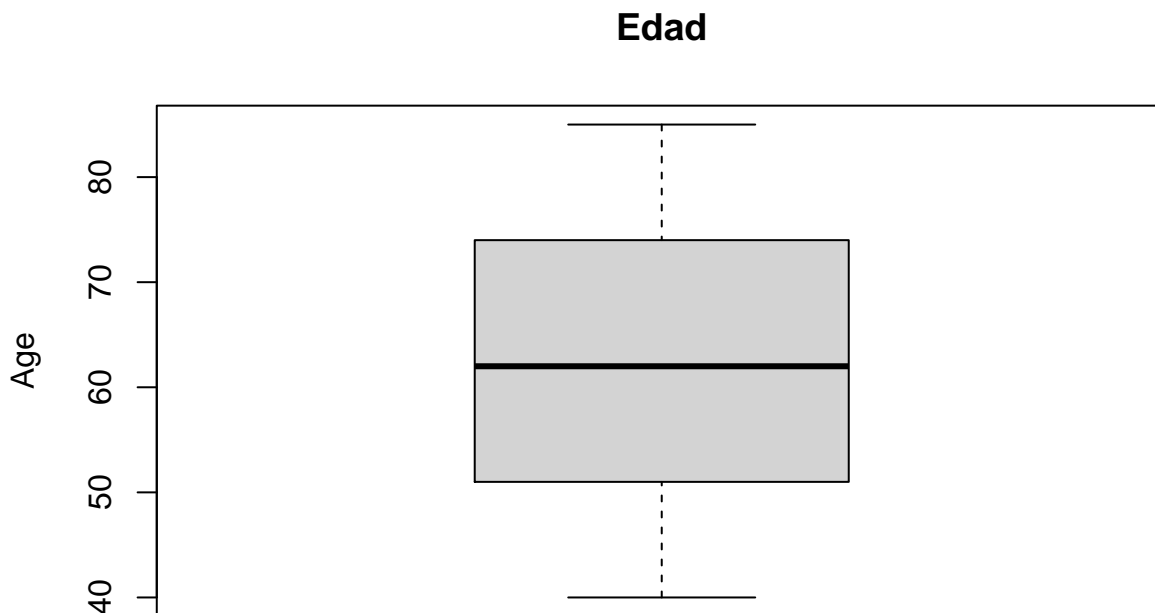
```

Nuevamente se presenta un diagrama de caja para Age

```

boxplot(df$Age, main = "Edad", "ylab" = "Age")

```



además del correspondiente resumen

```

summary(df$Age)

```

```

##      Min.   1st Qu.   Median     Mean   3rd Qu.    Max.     NA's
## 40.000000  51.000000  62.000000  62.2539863  74.000000  85.000000     11

```

En donde el rango de edades ahora va de 40 a 85 años.

## 2.5. Filtrado

Para efectos prácticos, ahora se trabajará con un conjunto reducido en que se eliminarán los valores atípicos del parámetro Age, manteniendo sólo las columnas Gender, Chain\_smoker, Obese, Diabetes, Use\_of\_stimulant\_drugs, Family\_history y UnderRisk.

```

cols <- c(
  "Age", "Gender", "Chain_smoker", "Obese",
  "Diabetes", "Use_of_stimulant_drugs",
  "Family_history", "UnderRisk")
new_df <- na.omit(df[, cols])

```

También se cambiará el tipo de dato de las variables cualitativas, las que serán transformadas a lógicas, con excepción de género, que está como factor.

```

new_df$Gender <- as.factor(new_df$Gender)
new_df$Chain_smoker <- as.logical(new_df$Chain_smoker)
new_df$Obese <- as.logical(new_df$Obese)
new_df$Diabetes <- as.logical(new_df$Diabetes)
new_df$Use_of_stimulant_drugs <- as.logical(new_df$Use_of_stimulant_drugs)
new_df$Family_history <- as.logical(new_df$Family_history)

```

Para el caso de UnderRisk, se hará uso del paquete batman.

```

if (!("batman" %in% rownames(installed.packages()))) {
  install.packages("batman")
}
library("batman")
new_df$UnderRisk <- to_logical(new_df$UnderRisk)

```

Exponiendo nuevamente la estructura y su resumen estadístico

```
str(new_df)
```

```

## 'data.frame':    878 obs. of  8 variables:
## $ Age                : int  84 55 80 40 45 78 77 56 75 47 ...
## $ Gender              : Factor w/ 2 levels "1","2": 1 1 1 1 1 2 1 2 1 1 ...
## $ Chain_smoker        : logi  TRUE FALSE FALSE FALSE FALSE FALSE ...
## $ Obese               : logi  TRUE TRUE TRUE TRUE FALSE TRUE ...
## $ Diabetes            : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
## $ Use_of_stimulant_drugs: logi  FALSE FALSE FALSE FALSE TRUE FALSE ...
## $ Family_history       : logi  TRUE TRUE TRUE TRUE FALSE TRUE ...
## $ UnderRisk           : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
## - attr(*, "na.action")= 'omit' Named int [1:11] 9 150 311 369 378 412 512 580 594 723 ...
## .. attr(*, "names")= chr [1:11] "9" "150" "311" "369" ...

```

```
summary(new_df)
```

```

##      Age                Gender Chain_smoker      Obese      Diabetes
## Min. :40.0000000    1:613   Mode :logical   Mode :logical   Mode :logical
## 1st Qu.:51.0000000    2:265   FALSE:772   FALSE:70      FALSE:830
## Median :62.0000000          TRUE :106      TRUE :808      TRUE :48
## Mean   :62.2539863
## 3rd Qu.:74.0000000
## Max.   :85.0000000
## Use_of_stimulant_drugs Family_history UnderRisk
## Mode :logical          Mode :logical   Mode :logical
## FALSE:807              FALSE:65        FALSE:690
## TRUE :71               TRUE :813        TRUE :188
##
##
##

```

## 2.6. Visualización

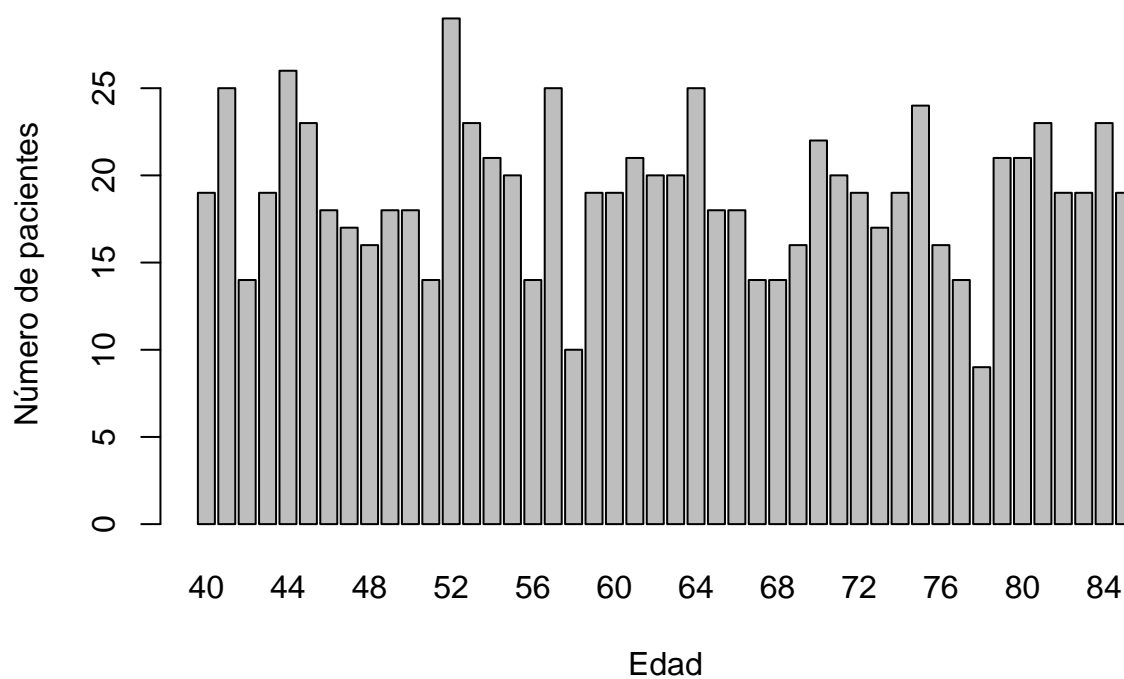
Ahora se verán algunas gráficas del nuevo subconjunto de datos

```

barplot(
  table(new_df$Age),
  main = "Edad",
  xlab = "Edad",
  ylab = "Número de pacientes"
)

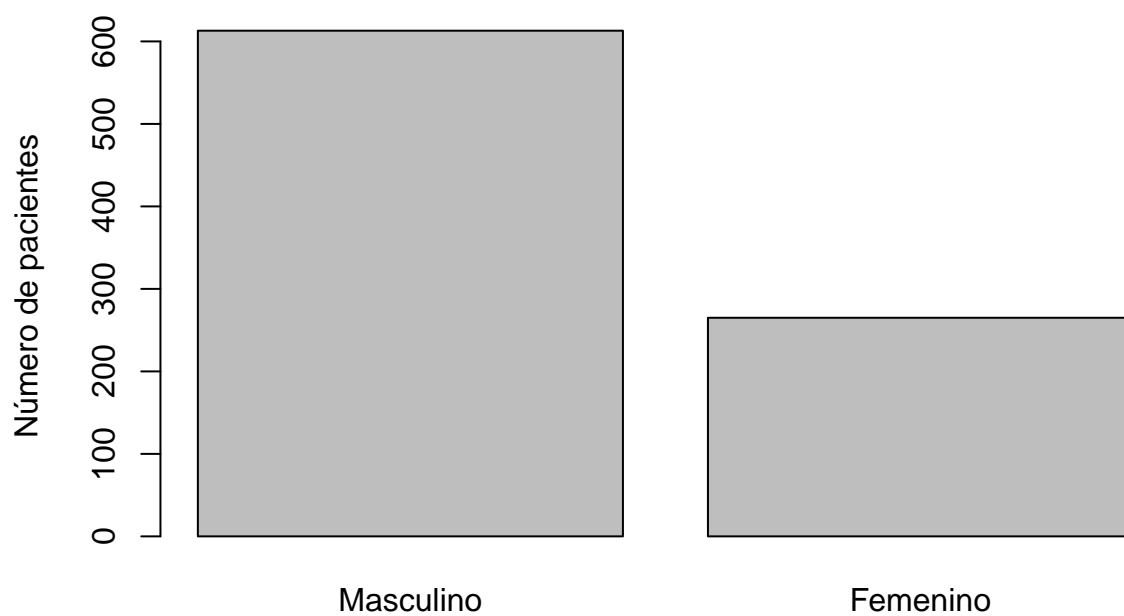
```

## Edad



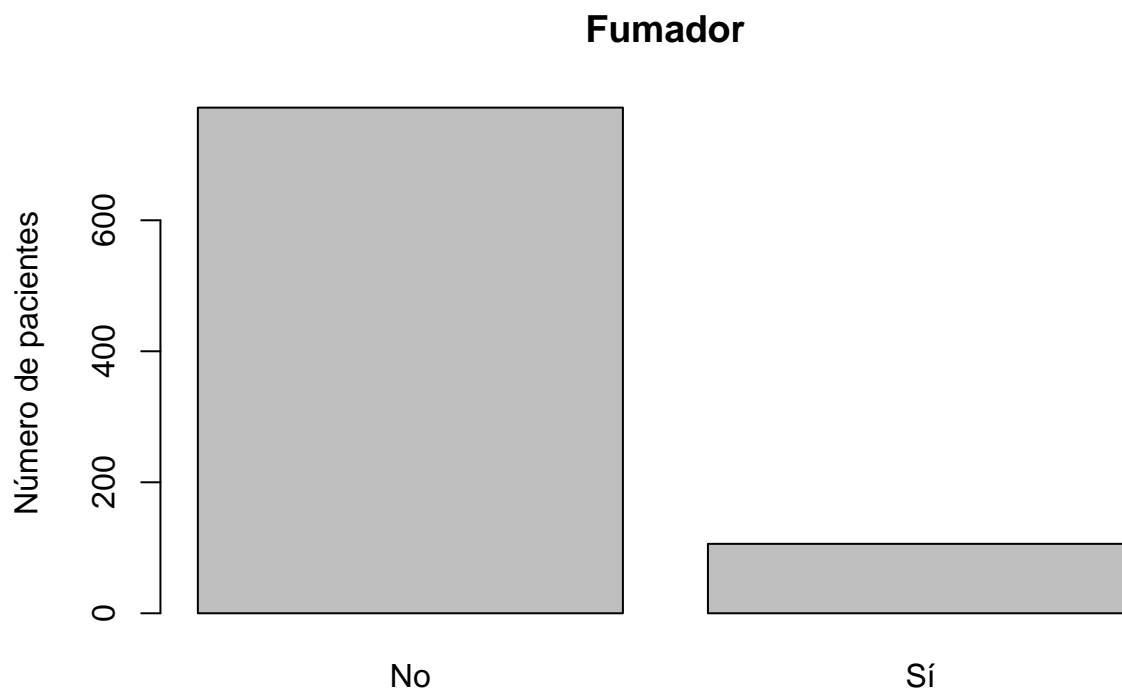
```
barplot(
  table(new_df$Gender),
  main = "Género",
  names.arg = c("Masculino", "Femenino"),
  ylab = "Número de pacientes"
)
```

## Género

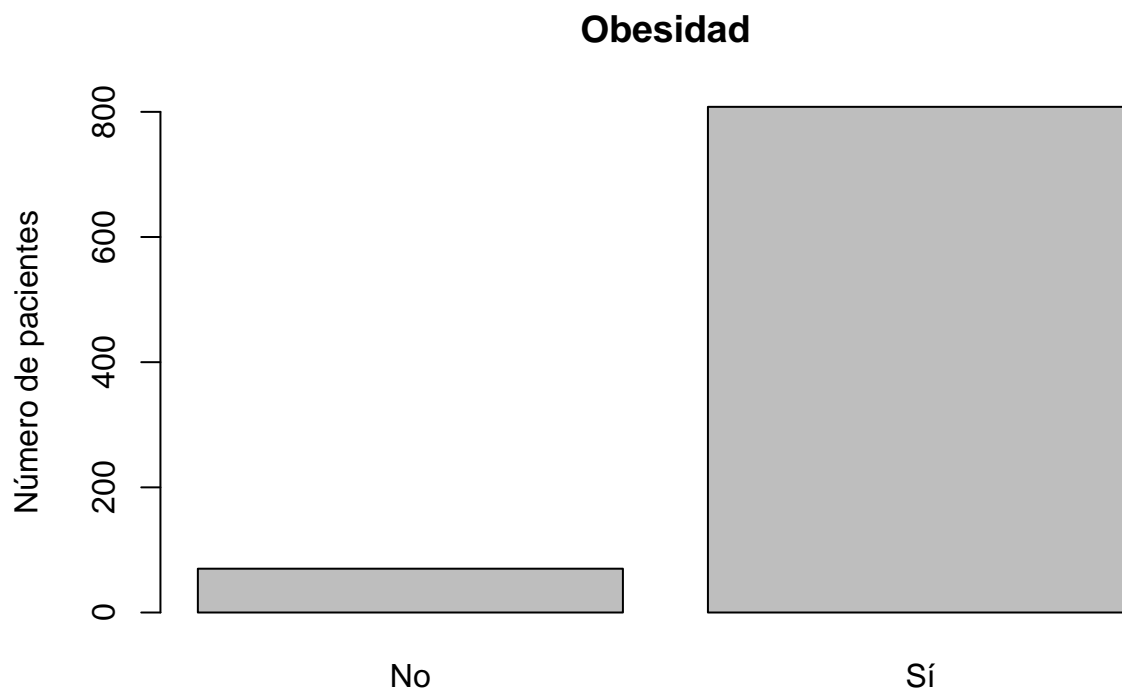


```
barplot(
  table(new_df$Chain_smoker),
  main = "Fumador",
  names.arg = c("No", "Sí"),
  ylab = "Número de pacientes"
)
```

```
) ylab = "Número de pacientes"
```



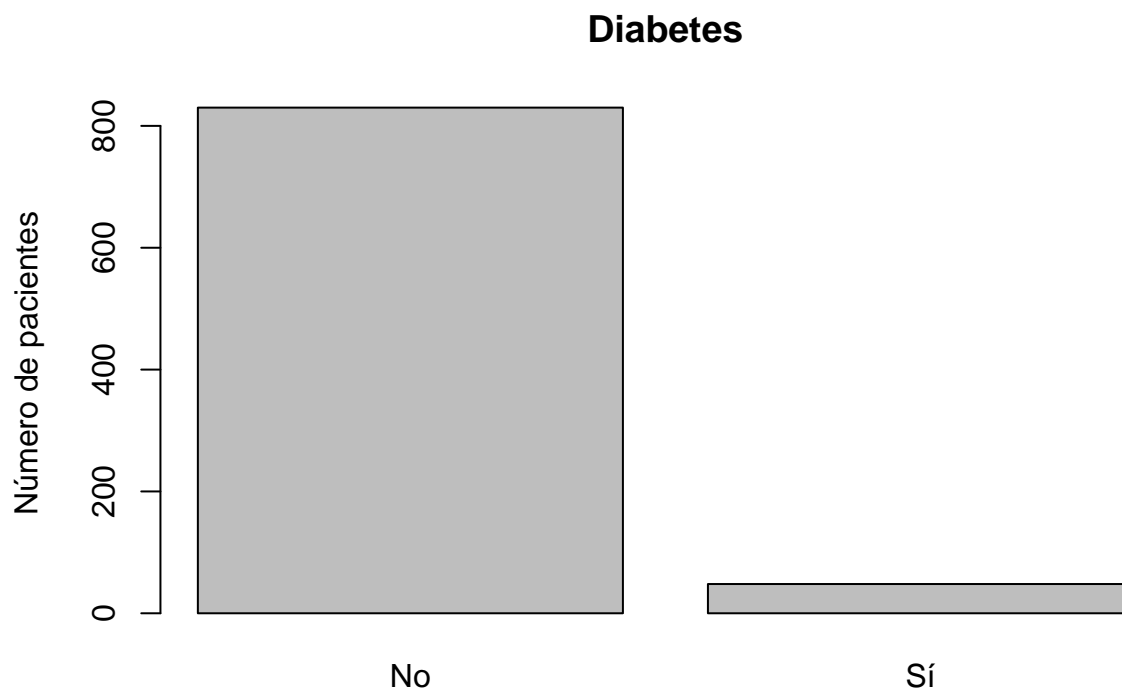
```
barplot(  
  table(new_df$Obese),  
  main = "Obesidad",  
  names.arg = c("No", "Sí"),  
  ylab = "Número de pacientes"  
)
```



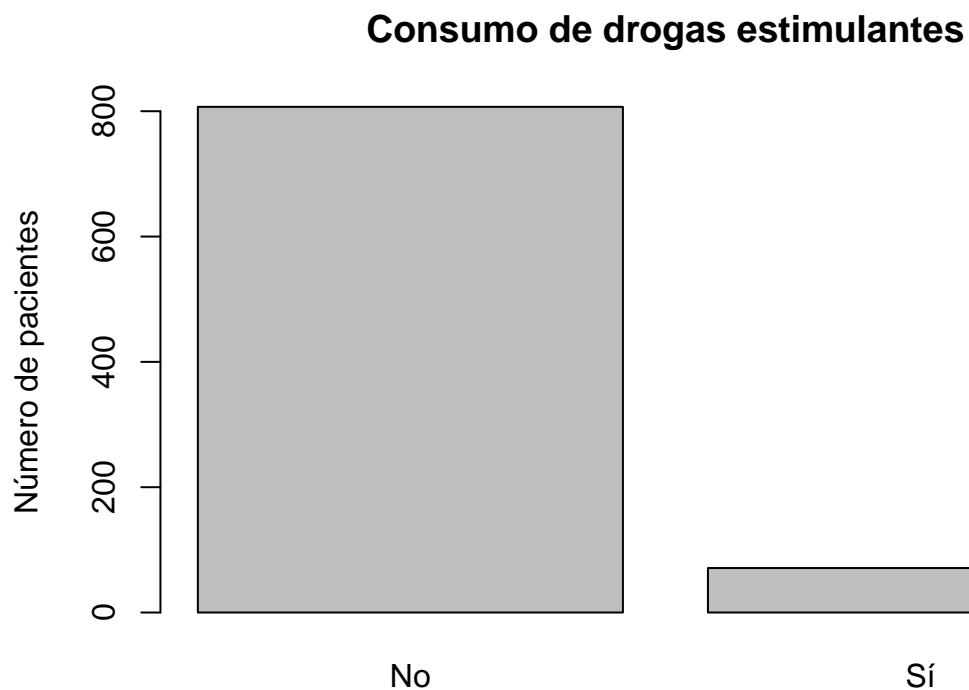
```
barplot(  
  table(new_df$Diabetes),  
  main = "Diabetes",  
  names.arg = c("No", "Sí"),
```



```
)
  ylab = "Número de pacientes"
)
```



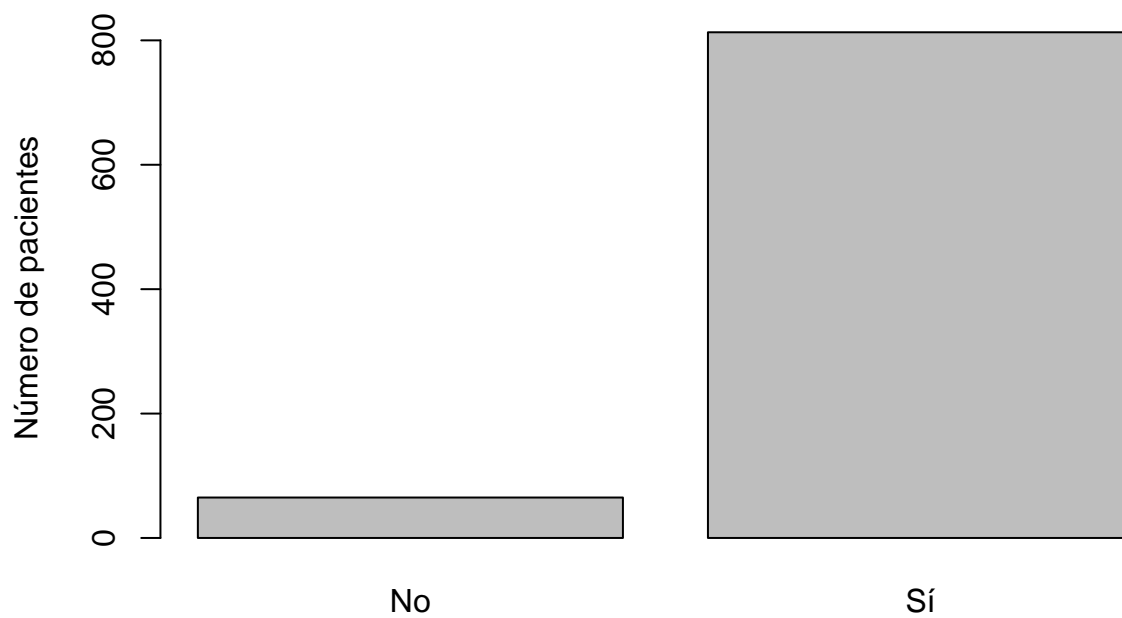
```
barplot(
  table(new_df$Use_of_stimulant_drugs),
  main = "Consumo de drogas estimulantes",
  names.arg = c("No", "Sí"),
  ylab = "Número de pacientes"
)
```



```
barplot(
  table(new_df$Family_history),
  main = "Historial familiar de ataque cardíaco",
  names.arg = c("No", "Sí"),
  ylab = "Número de pacientes"
)
```

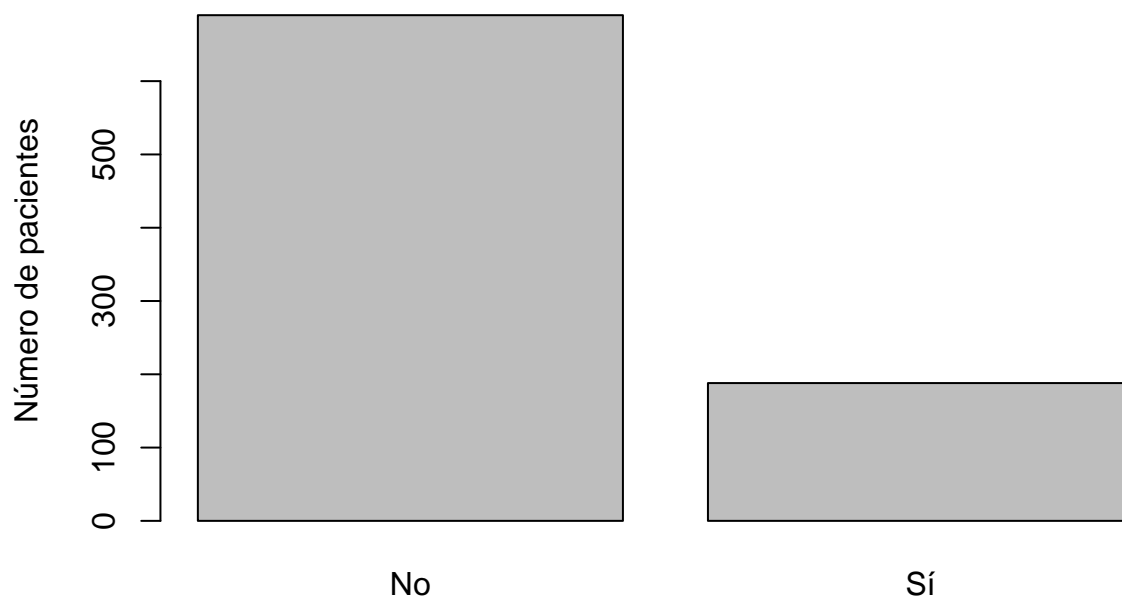
```
) ylab = "Número de pacientes"
```

### Historial familiar de ataque cardíaco



```
barplot(  
  table(new_df$UnderRisk),  
  main = "Riesgo de ataque cardíaco",  
  names.arg = c("No", "Sí"),  
  ylab = "Número de pacientes"  
)
```

### Riesgo de ataque cardíaco



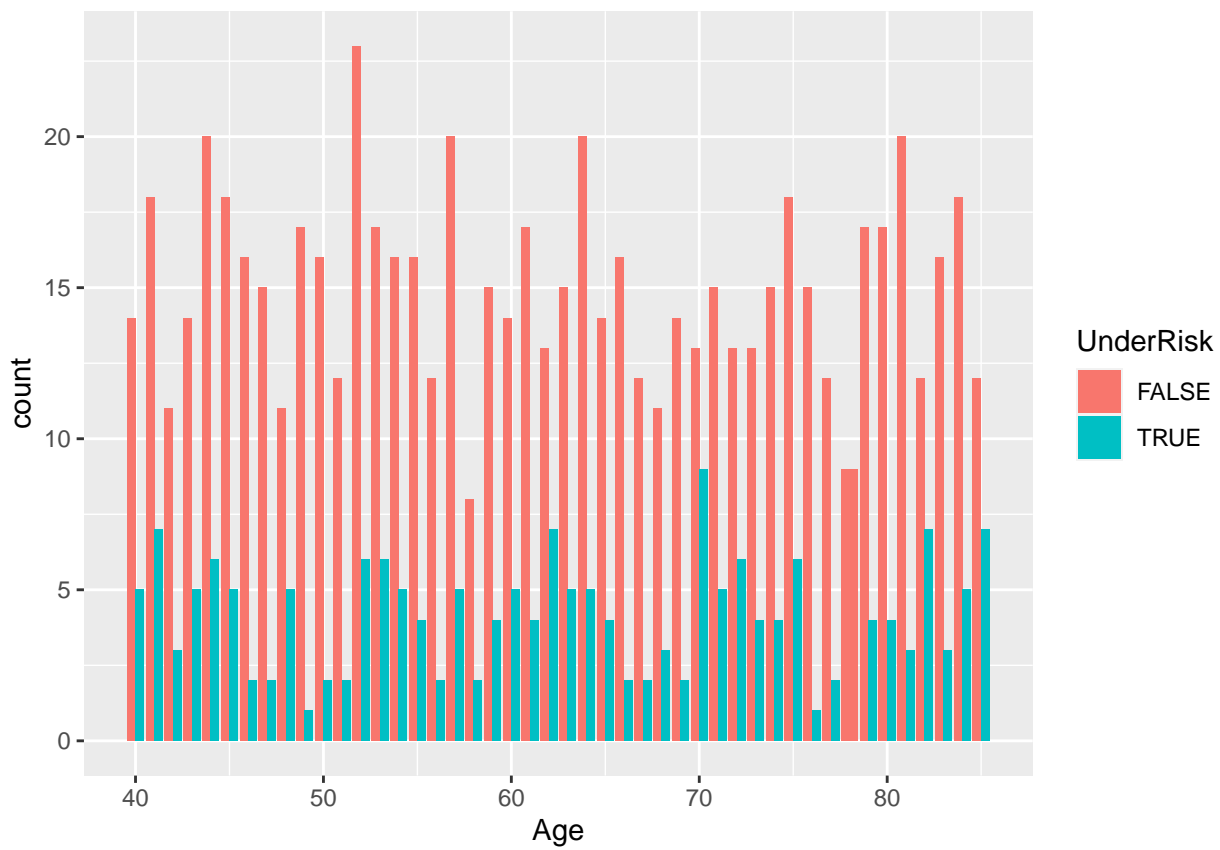
## 2.7. Análisis

Ahora se hará un filtrado para dejar sólo las personas que están bajo riesgo de un infarto y se mostrarán distintos gráficos

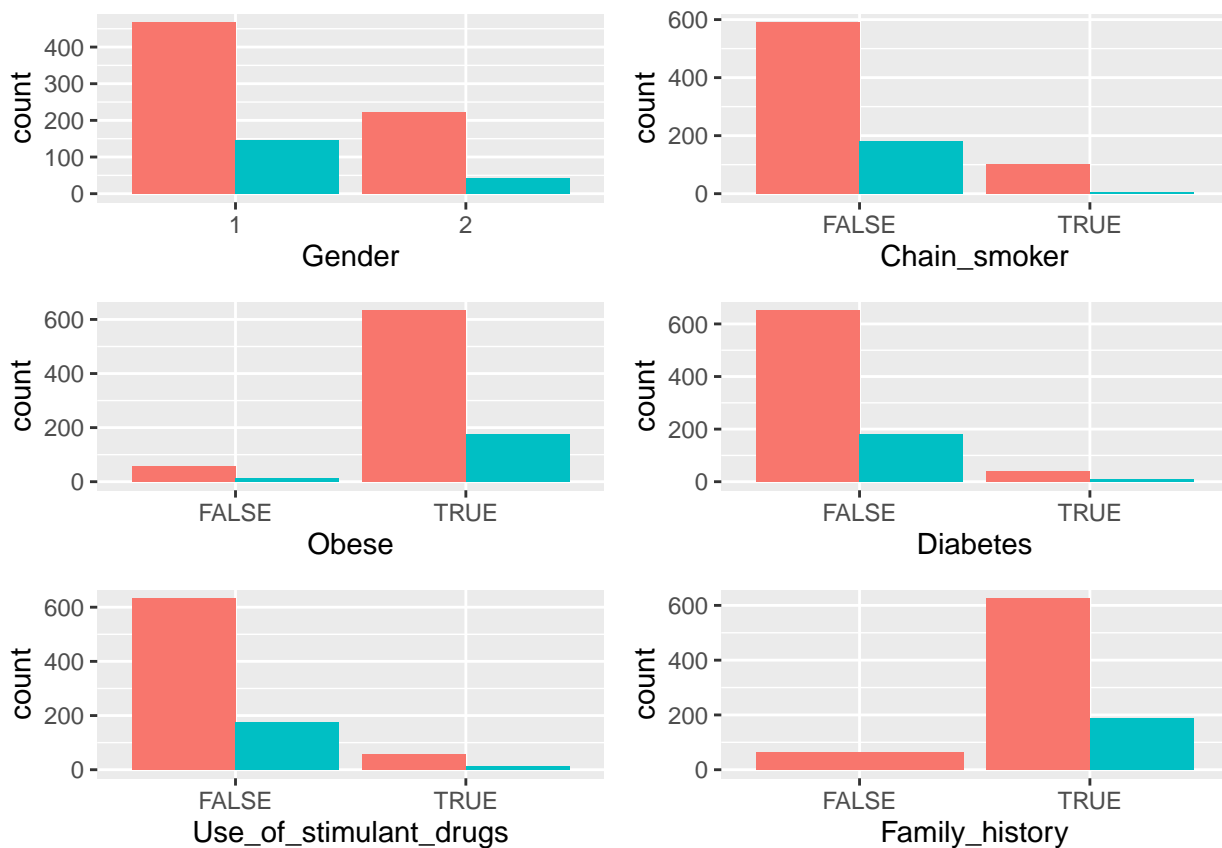
```
library("tidyverse")
library("gridExtra")
filtered <- filter(new_df, UnderRisk == TRUE)

bar_age <- ggplot(new_df, aes(x = Age, fill = UnderRisk)) +
  geom_bar(position = position_dodge()) +
  theme(legend.position = "right")
bar_gender <- ggplot(new_df, aes(x = Gender, fill = UnderRisk)) +
  geom_bar(position = position_dodge()) +
  theme(legend.position = "none")
bar_smoker <- ggplot(new_df, aes(x = Chain_smoker, fill = UnderRisk)) +
  geom_bar(position = position_dodge()) +
  theme(legend.position = "none")
bar_obese <- ggplot(new_df, aes(x = Obese, fill = UnderRisk)) +
  geom_bar(position = position_dodge()) +
  theme(legend.position = "none")
bar_diabetes <- ggplot(new_df, aes(x = Diabetes, fill = UnderRisk)) +
  geom_bar(position = position_dodge()) +
  theme(legend.position = "none")
bar_drugs <- ggplot(new_df, aes(x = Use_of_stimulant_drugs, fill = UnderRisk)) +
  geom_bar(position = position_dodge()) +
  theme(legend.position = "none")
bar_history <- ggplot(new_df, aes(x = Family_history, fill = UnderRisk)) +
  geom_bar(position = position_dodge()) +
  theme(legend.position = "none")

bar_age
```



```
grid.arrange(  
  bar_gender,  
  bar_smoker,  
  bar_obese,  
  bar_diabetes,  
  bar_drugs,  
  bar_history)
```



De los gráficos se puede concluir, grosso modo, que dentro de las personas que presentan riesgos de sufrir paros cardíacos los factores que más influyen son obesidad y la presencia de antecedentes familiares de ataques cardíacos.

## 2.8. Probabilidades

Fallé :c