PRACTICA 2

Contents

```
library(rockchalk)
```

```
## Warning: package 'rockchalk' was built under R version 3.4.3
```

```
library(nortest)
library(normtest)
library(moments)
```

```
##
## Attaching package: 'moments'
```

```
## The following objects are masked from 'package:rockchalk':
##
##     kurtosis, skewness
```

```
library(car)
```

```
## Warning: package 'car' was built under R version 3.4.3
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.4.3
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.4.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:car':
##
##     recode
```

```
## The following object is masked from 'package:rockchalk':
##
##     summarize
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Este dataset está formado 27 variables y 1259 observaciones. Estas variables son:

1.Timestamp: momento de presentación de respuestas

2. Age: edad

3. Gender: género

4. Country: país

5. state: estado. ¿Si vives en los Estados Unidos, cual es el estado o el territorio donde vives?

6. **self_employed**: auto-empleado. ¿Es autónomo (auto-empleado)?

7. **family_history**: historia familiar. ¿Tiene antecedentes de enfermedad mental en la familia?

8. **treatment**: tratamiento. ¿Ha sido tratado por una enfermedad mental?

9. **work_interfere**: ¿Si tiene una enfermedad mental, siente que interfiere con su trabajo?

10. **no_employees**: número de empleados. ¿Cuántos empleados tiene su compañía u organización?

11. **remote_work**: ¿realiza teletrabajo (fuera de la oficina) al menos el 50% del tiempo?

12. **tech_company**: ¿su empleador primaro es una organización o empresa de tecnología?

13. **benefits**: ¿su empleador provee beneficios de salud mental?

14. **care_options**: ¿conoce las opciones de cuidado mental de la compañía médica que el empleador provee?

15. **wellness_program**: ¿Su empleador ha mencionado alguna vez que tiene un programa de bienestar mental para sus empleados?

16. **seek_help**: ¿Su empleador proporciona recursos para saber más sobre aspectos de salud mental y cómo encontrar ayuda?

17. **anonymity**: ¿Está protegida su privacidad si elige acogerse a ventajas de salud mental o recursos de tratamiento de abusos de substancias?

18. **leave**: ¿Le sería fácil, acogerse a una baja por situación de salud mental?

19. **mental_health_consequence**: ¿Cree que hablar de un aspecto de salud mental con su empleador, tendría consecuencias negativas?

20. **phys_health_consequence**: ¿Cree que hablar de un aspecto de salud física con su empleador, tendría consecuencias negativas?

21. **coworkers**: ¿Estaría dispuesto a hablar con sus compañeros de un aspecto de salud mental?

22. **supervisor**: ¿Estaría dispuesto a hablar con sus supervisores de un aspecto de salud mental?

23. **mental_health_interview**: ¿Mencionaría un aspecto de salud mental con un potencial empleador en una entrevista?

24. **phys_health_interview**: ¿Mencionaría un aspecto de salud física con un potencial empleador en una entrevista?

25. **mental_vs_physical**: ¿Siente que su empleador se toma la salud mental como un aspecto importante de la salud?

26. **obs_consequence**: ¿Ha oido u observado consecuencias negativas para sus compañeros de trabajo que se encuentren en situación de enfermedad mental en su puesto de trabajo?

27. **comments**: comentarios adicionales

El dataset, de 2014 (facilitado por Open Sourcing Mental Illness), procede de una encuesta que mide las actitudes sobre salud mental y la frecuencia de desórdenes mentales en puestos de trabajo extraido en un contexto de Tecnologías de Información. Es de especial interés dado que aspectos como el uso intensivo de las tecnologías de la información está dando lugar también a nuevas enfermedades, también de tipo mental, como así se pone de manifiesto en la literatura (ver, por ejemplo: Gentile, D., Coyne, S., & Bricolo, F.(2012-12-31). Pathological Technology Addictions: What Is Scientifically Known and What Remains to Be Learned. In The Oxford Handbook of Media Psychology: Oxford University Press).

Para realizar un trabajo de forma correcta, el trabajador debe estar en situación de condiciones mentales normales.

Teniendo en cuenta que la Organización Mundial de la Salud informa que la salud mental no más que una actitud de bienestar para que la persona sea capaz de desarrollar sus capacidades, de afrontar el estrés del

día a día, que en su trabajo se observe una productividad y que sea capaz de aportar a la comunidad. Luego mirándolo de forma positiva, la salud mental es el pilar de un funcionamiento correcto tanto a nivel individual como a nivel comunidad.

No hay que olvidar que durante nuestro día a día nos encontramos con diferentes situaciones tanto a nivel personal como laboral que nos provocan estrés, esto está dentro de unos baremos de la normalidad y en ningún caso debe considerarse como un problema a tratar.

El hecho de sentir estrés no es malo, siempre y cuando sea en unas cantidades que nos permitan en todo momento tener un nivel de sensatez mental adecuado y un positivo rendimiento a nivel de conducta como cognitivo. Se afirma que el estrés agudo, de poca duración, pone en predisposición el cerebro para un mejor rendimiento.

Si lugar a dudas el estrés lleva a las personas a tener problemas de salud, relaciones insuficientes y una baja productividad laboral. Con lo que conlleva aspectos negativos tanto personalmente como profesionalmente. Visiblemente esto se observa con facilidad ya que el individuo se enfada constantemente con los que están más cerca.

Solamente, en la Unión Europea, las enfermedades relacionadas con los músculos del esqueleto superan al estrés laboral.

Una persona con estrés tiene los siguientes síntomas fatiga, tensión muscular, variación en el apetito, bruxismo, cambios en el estímulo sexual, mareos y dolores de cabeza. Psicológicamente estos factores pueden ser la irritabilidad, nerviosismo, falta de energía y ganas de llorar.

La cuestión que podemos llegar a responder es si el trato es el mismo laboralmente en la enfermedad física que en la enfermedad mental.

Pretendemos por tanto con ello responder a la siguiente pregunta/problema: ¿se trata de igual modo en el contexto laboral a las enfermedades físicas y mentales? Por las variables existentes en el conjunto de datos y a partir de estas preguntas previas, deducimos que hay dos aspectos que se podrían tratar: la existencia de enfermedad mental, y las actitudes hacia ésta por las personas en el puesto de trabajo. Nosotros nos centraremos en el segundo aspecto, buscando respuestas en cuanto al trato (o consideración) de igualdad (o no) entre enfermedades físicas y enfermedades mentales.

No debemos olvidar que todo proyecto analítico en ciencia de datos tiene las siguientes fases:

1. Se trata de encontrar la cuestión que deseamos resolver.
2. Consiste en la recodia y almacenamiento de los datos. Conocer de dónde se han extraído los datos y el formato de almacenamiento.
3. Limpieza de datos. Los datos son preparados para el análisis. Para ello es muy posible que se produzca eliminaciones, transformaciones, etc.
4. En esta etapa se produce el estudio de los datos y un aprendizaje de forma automática.
5. Aqui nos encontramos con el estudio de establecer la forma visual más eficiente para la represantación de los datos.
6. Resolvemos la cuestión que se planeó en la primera fase del proyecto.

Sin olvidar la peculiaridad y necesidades de cada proyecto, no todos tienen que llevar a cabo las 6 fases anteriormente nombradas de manera estricta y única. A veces es necesario que alguna fase se repita de manera iterativa.

Cargamos los datos getwd() setwd("C:/Users/David&Sonix/Downloads/Tipologia de Datos/Practica 2/Resolucion Final")

```
surveyMentalHealth<-read.csv("survey.csv", sep=",",na.strings = "NA")
#Mostramos las primeras filas
head(surveyMentalHealth)
```

```
##             Timestamp Age Gender        Country state self_employed
## 1 2014-08-27 11:29:31  37 Female  United States    IL          <NA>
## 2 2014-08-27 11:29:37  44      M  United States    IN          <NA>
## 3 2014-08-27 11:29:44  32   Male         Canada  <NA>          <NA>
## 4 2014-08-27 11:29:46  31   Male United Kingdom  <NA>          <NA>
## 5 2014-08-27 11:30:22  31   Male  United States    TX          <NA>
## 6 2014-08-27 11:31:22  33   Male  United States    TN          <NA>
##   family_history treatment work_interfere   no_employees remote_work
## 1             No       Yes          Often           6-25          No
## 2             No        No         Rarely More than 1000          No
## 3             No        No         Rarely           6-25          No
## 4            Yes       Yes          Often         26-100          No
## 5             No        No          Never        100-500         Yes
## 6            Yes        No      Sometimes           6-25          No
##   tech_company   benefits care_options wellness_program  seek_help
## 1          Yes        Yes     Not sure               No        Yes
## 2           No Don't know           No       Don't know Don't know
## 3          Yes         No           No               No         No
## 4          Yes         No          Yes               No         No
## 5          Yes        Yes           No       Don't know Don't know
## 6          Yes        Yes     Not sure               No Don't know
##    anonymity              leave mental_health_consequence
## 1        Yes      Somewhat easy                         No
## 2 Don't know         Don't know                      Maybe
## 3 Don't know Somewhat difficult                         No
## 4         No Somewhat difficult                        Yes
## 5 Don't know         Don't know                         No
## 6 Don't know         Don't know                         No
##   phys_health_consequence    coworkers supervisor mental_health_interview
## 1                      No Some of them        Yes                      No
## 2                      No           No         No                      No
## 3                      No          Yes        Yes                     Yes
## 4                     Yes Some of them         No                   Maybe
## 5                      No Some of them        Yes                     Yes
## 6                      No          Yes        Yes                      No
##   phys_health_interview mental_vs_physical obs_consequence comments
## 1                 Maybe               Yes              No     <NA>
## 2                    No         Don't know              No     <NA>
## 3                   Yes                No              No     <NA>
## 4                 Maybe                No             Yes     <NA>
## 5                   Yes         Don't know              No     <NA>
## 6                 Maybe         Don't know              No     <NA>
```

Procedemos a leer el fichero *survey.csv*. La carga la realizamos mediante *read.csv*, debido a que ,(coma) es el separador de las variables. Los valores perdidos los designamos por *NA*

De las 27 variables que contiene el data set

1.Timestamp

2. Age

3. Gender

4. Country

5. state

6. self_employed

7. family_history

8. treatment

9. work_interfere

10.no_employees

11.remote_work

12.tech_company

13.benefits

14.care_options

15.wellness_program

16.seek_help

17.anonymity

18.leave

19.mental_health_consequence

20.phys_health_consequence

21.coworkers

22.supervisor

23.mental_health_interview

24.phys_health_interview

25.mental_vs_physical

26.obs_consequence

27.comments

De estas variables, dado que algunas de ellas no son directamente asociadas al objetivo de nuestro trabajo, debido a las razones previamente expuestas, prescindimos de las siguientes 3 variables.

1.Timestamp

5. state

27.comments

Así pues, nos quedamos con 24 variables que, a priori, podrían ser útiles para nosotros.

```
#Eliminación de las variables
surveyMentalHealth$Timestamp<-NULL
surveyMentalHealth$state<-NULL
surveyMentalHealth$comments<-NULL
#summary(surveyMentalHealth)
```

```
colnames(surveyMentalHealth)
```

```
##  [1] "Age"                       "Gender"
##  [3] "Country"                   "self_employed"
##  [5] "family_history"            "treatment"
##  [7] "work_interfere"            "no_employees"
##  [9] "remote_work"               "tech_company"
```

```
## [11] "benefits"                  "care_options"
## [13] "wellness_program"          "seek_help"
## [15] "anonymity"                 "leave"
## [17] "mental_health_consequence" "phys_health_consequence"
## [19] "coworkers"                 "supervisor"
## [21] "mental_health_interview"   "phys_health_interview"
## [23] "mental_vs_physical"        "obs_consequence"
```

Mostramos algunos detalles de los objetos

```
str(surveyMentalHealth)
```

```
## 'data.frame':    1259 obs. of  24 variables:
##  $ Age                      : num  37 44 32 31 31 33 35 39 42 23 ...
##  $ Gender                   : Factor w/ 49 levels "A little about you",..: 16 24 30 30 30 30 16 24 16
##  $ Country                  : Factor w/ 48 levels "Australia","Austria",..: 46 46 8 45 46 46 46 8 46
##  $ self_employed            : Factor w/ 2 levels "No","Yes": NA NA NA NA NA NA NA NA NA NA ...
##  $ family_history           : Factor w/ 2 levels "No","Yes": 1 1 1 2 1 2 2 1 2 1 ...
##  $ treatment                : Factor w/ 2 levels "No","Yes": 2 1 1 2 1 1 2 1 2 1 ...
##  $ work_interfere           : Factor w/ 4 levels "Never","Often",..: 2 3 3 2 1 4 4 1 4 1 ...
##  $ no_employees             : Factor w/ 6 levels "1-5","100-500",..: 5 6 5 3 2 5 1 1 2 3 ...
##  $ remote_work              : Factor w/ 2 levels "No","Yes": 1 1 1 1 2 1 2 2 1 1 ...
##  $ tech_company             : Factor w/ 2 levels "No","Yes": 2 1 2 2 2 2 2 2 2 2 ...
##  $ benefits                 : Factor w/ 3 levels "Don't know","No",..: 3 1 2 2 3 3 2 2 3 1 ...
##  $ care_options             : Factor w/ 3 levels "No","Not sure",..: 2 1 1 3 1 2 1 3 3 1 ...
##  $ wellness_program         : Factor w/ 3 levels "Don't know","No",..: 2 1 2 2 1 2 2 2 2 1 ...
##  $ seek_help                : Factor w/ 3 levels "Don't know","No",..: 3 1 2 2 1 1 2 2 2 1 ...
##  $ anonymity                : Factor w/ 3 levels "Don't know","No",..: 3 1 1 2 1 1 2 3 2 1 ...
##  $ leave                    : Factor w/ 5 levels "Don't know","Somewhat difficult",..: 3 1 2 2 1 1 2
##  $ mental_health_consequence: Factor w/ 3 levels "Maybe","No","Yes": 2 1 2 3 2 2 1 2 1 2 ...
##  $ phys_health_consequence  : Factor w/ 3 levels "Maybe","No","Yes": 2 2 2 3 2 2 1 2 2 2 ...
##  $ coworkers                : Factor w/ 3 levels "No","Some of them",..: 2 1 3 2 2 3 2 1 3 3 ...
##  $ supervisor               : Factor w/ 3 levels "No","Some of them",..: 3 1 3 1 3 3 1 1 3 3 ...
##  $ mental_health_interview  : Factor w/ 3 levels "Maybe","No","Yes": 2 2 3 1 3 2 2 2 2 1 ...
##  $ phys_health_interview    : Factor w/ 3 levels "Maybe","No","Yes": 1 2 3 1 3 1 2 2 1 1 ...
##  $ mental_vs_physical       : Factor w/ 3 levels "Don't know","No",..: 3 1 2 2 1 1 1 2 2 3 ...
##  $ obs_consequence          : Factor w/ 2 levels "No","Yes": 1 1 1 2 1 1 1 1 1 1 ...
```

Comprobamos que el número de filas o registros cargados son correctos.

```
#Número de fila del fichero.
nrow(surveyMentalHealth)
```

```
## [1] 1259
```

Es bueno observar cada una de las características anteriores para saber si el fichero se ha subido adecuadamente.

Observamos el tipo de variable estadística para cada variable.

```
sapply(surveyMentalHealth,class)
```

```
##                       Age                    Gender
##                 "numeric"                  "factor"
##                   Country             self_employed
##                  "factor"                  "factor"
##            family_history                 treatment
##                  "factor"                  "factor"
##            work_interfere              no_employees
```

```
##                  "factor"                  "factor"
##               remote_work              tech_company
##                  "factor"                  "factor"
##                  benefits              care_options
##                  "factor"                  "factor"
##          wellness_program                 seek_help
##                  "factor"                  "factor"
##                 anonymity                     leave
##                  "factor"                  "factor"
## mental_health_consequence   phys_health_consequence
##                  "factor"                  "factor"
##                 coworkers                supervisor
##                  "factor"                  "factor"
##   mental_health_interview     phys_health_interview
##                  "factor"                  "factor"
##        mental_vs_physical           obs_consequence
##                  "factor"                  "factor"
```

Es habitual que R no asigne adecuadamente el tipo de variable estadística a las diferentes variables en estudio. Pero este no es el caso. El tipo de variable estadística esta adecuadamente definido.

No debemos olvidar que la transformación entre los diferentes tipos de datos es una labor ineludible en la limpieza de datos. Hay que tener siempre en mente que esta transformaciones conllevan un riesgo principal, que no es otro que la perdida de datos al transformar un tipo de dato en otro. Recordemos que los principales factores que dan lugar a esta situación son:

-Mismo tipo de dato con transformación en diferente tamaño.

-Transformación con cota de exactitud diferente.

En el caso que nos ocupa todas las variables están definidas de forma correcta.

En este dataset nos encontramos con un conjunto de variables que son cuantitativas y cualitativas. Las cualitativas son las que tienen su origen en caracterísitcas o categorías. Mientras que la variable cuantitativa hace referencia a un valor de naturaleza numérica, estas pueden ser discretas (corresponden a un valor numérico entero) y continuos (toman cualquier valor existente en un intervalo). La forma de analizar estos datos es diferente, la primera de ella es la ordenación, un dato cualitativo no puede ordenarse de manera numérica. Para obtener información de datos cualitativos partimos de distribuciones de frecuencias, en la cual podemos observar el número de veces que sucede una caterogoría o nivel de la variable cualitativa. En variables cuantitativas la distribución de frecuencia nos proporciona una zona visible más espesa donde se establecen el mayor número de observaciones y una zona mas liviana donde nos encontramos con muy pocas observaciones.

En el dataset que nos ocupa la unica variable cuantitativa discreta es *Age* el resto son variables cualitativas.

Cuando hablamos de un dato cero tenemos siempre en mente una asociación a un valor numérico. No hay que olvidar que si el dato es de carácter numérico el valor cero es el que mejor se adapta.

Un dato vacío existe cuando se carece de observación. Este es de utilidad cuando nos encontramos con cadena de caracteres, si añadimos un espacio en blanco el dato pierde el caracter de vacio.

En el momento de la lectura del fichero hemos especificado *na.strings = "NA"* con lo cual cualquier elemento vacio ha sido rellenado con "NA".

Comprobamos que variables tienen datos perdidos

```
#En el momento de la lectura del fichero establecemos
#que si se encuenta un valor perdido los asigne por NA (na.strings = "NA")
table(is.na(surveyMentalHealth$Age))
```

```
##
## FALSE
##  1259
```

```
table(is.na(surveyMentalHealth$Gender))
```

```
##
## FALSE
##  1259
```

```
table(is.na(surveyMentalHealth$Country ))
```

```
##
## FALSE
##  1259
```

```
table(is.na(surveyMentalHealth$self_employed))
```

```
##
## FALSE  TRUE
##  1241    18
```

```
table(is.na(surveyMentalHealth$family_history))
```

```
##
## FALSE
##  1259
```

```
table(is.na(surveyMentalHealth$treatment))
```

```
##
## FALSE
##  1259
```

```
table(is.na(surveyMentalHealth$work_interfere))
```

```
##
## FALSE  TRUE
##   995   264
```

```
table(is.na(surveyMentalHealth$no_employees))
```

```
##
## FALSE
##  1259
```

```
table(is.na(surveyMentalHealth$remote_work))
```

```
##
## FALSE
##  1259
```

```
table(is.na(surveyMentalHealth$tech_company))
```

```
##
## FALSE
##  1259
```

```
table(is.na(surveyMentalHealth$benefits))
```

```
##
```

```
## FALSE
##  1259
```

```
table(is.na(surveyMentalHealth$care_options))
```

```
##
## FALSE
##  1259
```

```
table(is.na(surveyMentalHealth$ wellness_program))
```

```
##
## FALSE
##  1259
```

```
table(is.na(surveyMentalHealth$seek_help))
```

```
##
## FALSE
##  1259
```

```
table(is.na(surveyMentalHealth$anonymity))
```

```
##
## FALSE
##  1259
```

```
table(is.na(surveyMentalHealth$leave))
```

```
##
## FALSE
##  1259
```

```
table(is.na(surveyMentalHealth$mental_health_consequence))
```

```
##
## FALSE
##  1259
```

```
table(is.na(surveyMentalHealth$phys_health_consequence))
```

```
##
## FALSE
##  1259
```

```
table(is.na(surveyMentalHealth$coworkers))
```

```
##
## FALSE
##  1259
```

```
table(is.na(surveyMentalHealth$supervisor))
```

```
##
## FALSE
##  1259
```

```
table(is.na(surveyMentalHealth$mental_health_interview))
```

```
##
## FALSE
```

```
##  1259
```

```
table(is.na(surveyMentalHealth$phys_health_interview))
```

```
##
## FALSE
##  1259
```

```
table(is.na(surveyMentalHealth$mental_vs_physical))
```

```
##
## FALSE
##  1259
```

```
table(is.na(surveyMentalHealth$obs_consequence))
```

```
##
## FALSE
##  1259
```

Podemos concluir que las únicas variables que contienen los valores vacios son *self_employed* con un total de 18 (valor TRUE) y *work_interfere* con 264 (Valor TRUE).

Comprobamos si la unica variable cuantitativa, *Age* posee valor cero

```
surveyMentalHealth$Age[which(surveyMentalHealth$Age==0) ]
```

```
## numeric(0)
```

Por tanto podemos concluir que la variable *Age* no posee datos cero.

Se entiende por dato atípico como una observación fuera de la normalidad de la variable, una observación con una desviación tan grande de las otras observaciones que incluso podemos poner en duda si ha sido producido por los mismos mecanismos que las anteriores. El punto en común es lo alejado que esta del resto de las observaciones de la variable.

Los motivos por los cuales aparecen los datos atípicos pueden ser:

1.Outliers o datos atípicos cuyo origen está en la equivocación de los datos.

2.Valores atípicos u outliers con un propósito.

3.Valores atípicos u outliers cuyo origen son errores del muestreo.

4.Valores atípicos u outliers de errores en la estandarización.

5.Valores atípicos u outliers por asumir distribuciones erróneas.

6.Valor atípico u outliers cuyo origen es el muestreo correcto de la población.

7.Outliers o datos atípicos que proporcionan orígenes de nuevas investigaciones.

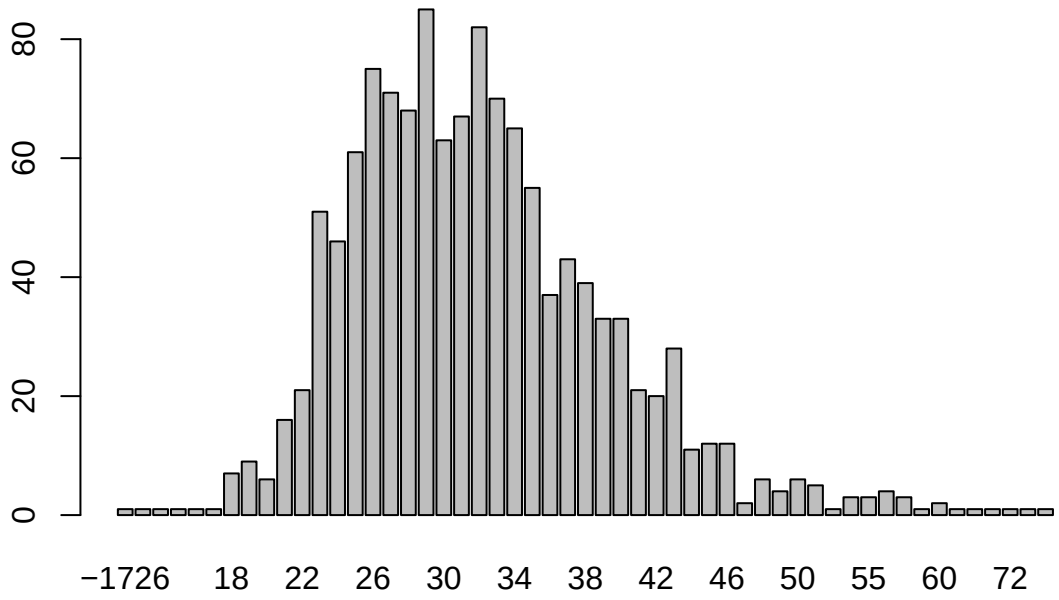
Los datos atípicos pueden tener efectos peligrosos en los diferentes análisis estadísticos que realicemos, con ellos presentes se puede llegar aumentar el error de la varianza y hacer disminuir los resultados de las pruebas estadísticas.

A continuación realizamos un analisis exhaustivo de cada variable. Dentro de este análisis obviamente esta la posibilidad de existencia de datos atipicos y se explicara la forma de gestionarlos.

Age

Detectamos gráficamente la posibilidad de tener datos atípicos. Mediante una grafica de frecuencia

```
barplot(table(surveyMentalHealth$Age))
```



Como podemos observar en este gráfico tenemos datos atípicos ya que tenemos individuos con edades menores de 16 años e incluso negativas, edad legal desde cuando se puede comenzar a trabajar, mayores de 75 años que corresponde a la edad legal a partir de la cual no se puede trabajar.

```
datos_atipicos<-subset(surveyMentalHealth[1:1], surveyMentalHealth$Age<16 | surveyMentalHealth$Age>75)
datos_atipicos
```

```
##             Age
## 144  -2.900e+01
## 365   3.290e+02
## 391   1.000e+11
## 716  -1.726e+03
## 735   5.000e+00
## 990   8.000e+00
## 1091  1.100e+01
## 1128 -1.000e+00
```

Entendemos que la generación de estos datos atípicos no es otra que una mal grabación de los datos o un error en el momento de la recogida de ellos. Si no tenemos posibilidad de rectificar el dato lo mejor es despreciar estas observaciones.

```
surveyMentalHealth_clean<-subset(surveyMentalHealth, surveyMentalHealth$Age>16 & surveyMentalHealth$Age<
nrow(surveyMentalHealth_clean)
```

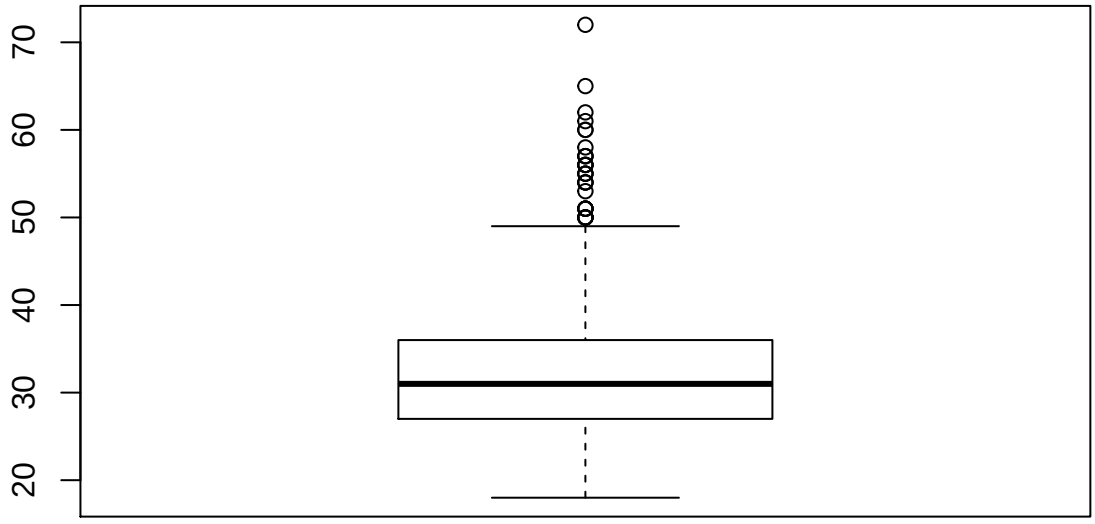
```
## [1] 1251
```

Una vez eliminados los datos atípicos de la variable Age y teniendo en cuenta que esta es cuantitativa procedemos a obtener los siguientes datos:

```r
summary(surveyMentalHealth_clean$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   27.00   31.00   32.08   36.00   72.00
```

```r
boxplot(surveyMentalHealth_clean$Age)
```




```r
#Media Aritmetica
Media_Age<-mean(surveyMentalHealth_clean$Age)
#Mediana
Mediana_Age<-median(surveyMentalHealth_clean$Age)
#Media Recortada
Media_Recortada_Age<-mean(surveyMentalHealth_clean$Age, trim=0.05)
#Desviación estándar
Desviacion_estandar_Age<-sd(surveyMentalHealth_clean$Age)
#Rango Intercuartilico (RIC)
RIC_Age<-IQR(surveyMentalHealth_clean$Age)
#Desviación Absoluta Respecto de la Mediana
Desviacion_Absoluta_Mediana_Age<-mad(surveyMentalHealth_clean$Age)
#Tabla
kable(rbind(Media_Age,Mediana_Age,Media_Recortada_Age,
            Desviacion_estandar_Age,RIC_Age,
            Desviacion_Absoluta_Mediana_Age))
```

Media_Age	32.076739
Mediana_Age	31.000000
Media_Recortada_Age	31.655723

Desviacion_estandar_Age	7.288272
RIC_Age	9.000000
Desviacion_Absoluta_Mediana_Age	5.930400

Gender

Esta es una variable cualitativa. Asi que por definición de la variable debería haber 2 niveles (Male y Female).

```
levels(surveyMentalHealth_clean$Gender)
```

```
##  [1] "A little about you"
##  [2] "Agender"
##  [3] "All"
##  [4] "Androgyne"
##  [5] "cis-female/femme"
##  [6] "Cis Female"
##  [7] "cis male"
##  [8] "Cis Male"
##  [9] "Cis Man"
## [10] "Enby"
## [11] "f"
## [12] "F"
## [13] "femail"
## [14] "Femake"
## [15] "female"
## [16] "Female"
## [17] "Female "
## [18] "Female (cis)"
## [19] "Female (trans)"
## [20] "fluid"
## [21] "Genderqueer"
## [22] "Guy (-ish) ^_^"
## [23] "m"
## [24] "M"
## [25] "Mail"
## [26] "maile"
## [27] "Make"
## [28] "Mal"
## [29] "male"
## [30] "Male"
## [31] "Male-ish"
## [32] "Male "
## [33] "Male (CIS)"
## [34] "male leaning androgynous"
## [35] "Malr"
## [36] "Man"
## [37] "msle"
## [38] "Nah"
## [39] "Neuter"
## [40] "non-binary"
## [41] "ostensibly male, unsure what that really means"
## [42] "p"
## [43] "queer"
## [44] "queer/she/they"
```

```
## [45] "something kinda male?"
## [46] "Trans-female"
## [47] "Trans woman"
## [48] "woman"
## [49] "Woman"
```

Sin embargo, nos encontramos con 49.

Cambiamos sus valores correspondientes por M y F respectivamente. Para aquellos que no es posible determinar, dad la inconcreción de la respuesta, le asignamos NA ("A little about you","Agender","All","Enby","fluid","Genderqueer","Nah","Neuter","non-binary","p","queer","queer/she/they","Trans woman","Trans-female").

```
surveyMentalHealth_clean$Gender<-as.character(surveyMentalHealth_clean$Gender)
surveyMentalHealth_clean$Gender<-replace(surveyMentalHealth_clean$Gender,surveyMentalHealth_clean$Gende

surveyMentalHealth_clean$Gender<-replace(surveyMentalHealth_clean$Gender,surveyMentalHealth_clean$Gende

surveyMentalHealth_clean$Gender<-replace(surveyMentalHealth_clean$Gender,surveyMentalHealth_clean$Gende

surveyMentalHealth_clean$Gender<-as.factor(surveyMentalHealth_clean$Gender)
```

Ahora comprobamos los niveles de la variable cualitativa Gender

```
levels(surveyMentalHealth_clean$Gender)
```

```
## [1] "F" "M"
```

Recordemos que inicialmente hemos comprobado que esta variable no poseia ningún dato vacio, especificado en el momento de la lectura como NA. Seguidamente procederia a eleminar estas valores anteriormente clasificados como NA.

```
surveyMentalHealth_clean<-subset(surveyMentalHealth_clean, surveyMentalHealth_clean$Gender!="NA")
nrow(surveyMentalHealth_clean)
```

```
## [1] 1240
```

Country

```
levels(surveyMentalHealth_clean$Country)
```

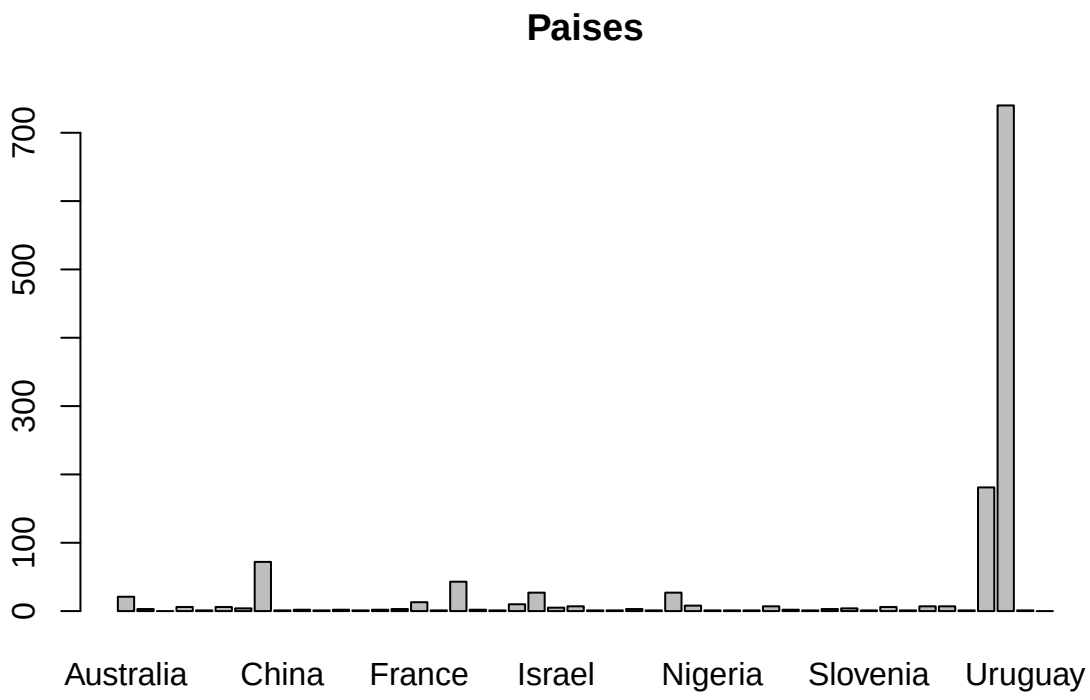
```
##  [1] "Australia"              "Austria"
##  [3] "Bahamas, The"           "Belgium"
##  [5] "Bosnia and Herzegovina" "Brazil"
##  [7] "Bulgaria"               "Canada"
##  [9] "China"                  "Colombia"
## [11] "Costa Rica"             "Croatia"
## [13] "Czech Republic"         "Denmark"
## [15] "Finland"                "France"
## [17] "Georgia"                "Germany"
## [19] "Greece"                 "Hungary"
## [21] "India"                  "Ireland"
## [23] "Israel"                 "Italy"
## [25] "Japan"                  "Latvia"
## [27] "Mexico"                 "Moldova"
## [29] "Netherlands"            "New Zealand"
## [31] "Nigeria"                "Norway"
## [33] "Philippines"            "Poland"
## [35] "Portugal"               "Romania"
```

```
## [37] "Russia"                "Singapore"
## [39] "Slovenia"              "South Africa"
## [41] "Spain"                 "Sweden"
## [43] "Switzerland"           "Thailand"
## [45] "United Kingdom"        "United States"
## [47] "Uruguay"               "Zimbabwe"
```

```
summary(surveyMentalHealth_clean$Country)
```

```
##              Australia                Austria           Bahamas, The
##                     21                      3                      0
##                Belgium Bosnia and Herzegovina                 Brazil
##                      6                      1                      6
##               Bulgaria                 Canada                  China
##                      4                     72                      1
##               Colombia             Costa Rica                Croatia
##                      2                      1                      2
##         Czech Republic                Denmark                Finland
##                      1                      2                      3
##                 France                Georgia                Germany
##                     13                      1                     43
##                 Greece                Hungary                  India
##                      2                      1                     10
##                Ireland                 Israel                  Italy
##                     27                      5                      7
##                  Japan                 Latvia                 Mexico
##                      1                      1                      3
##                Moldova            Netherlands            New Zealand
##                      1                     27                      8
##                Nigeria                 Norway            Philippines
##                      1                      1                      1
##                 Poland               Portugal                Romania
##                      7                      2                      1
##                 Russia              Singapore               Slovenia
##                      3                      4                      1
##           South Africa                  Spain                 Sweden
##                      6                      1                      7
##            Switzerland               Thailand         United Kingdom
##                      7                      1                    181
##          United States                Uruguay               Zimbabwe
##                    740                      1                      0
```

```
barplot(table(surveyMentalHealth_clean$Country),
        main="Paises")
```



family_history

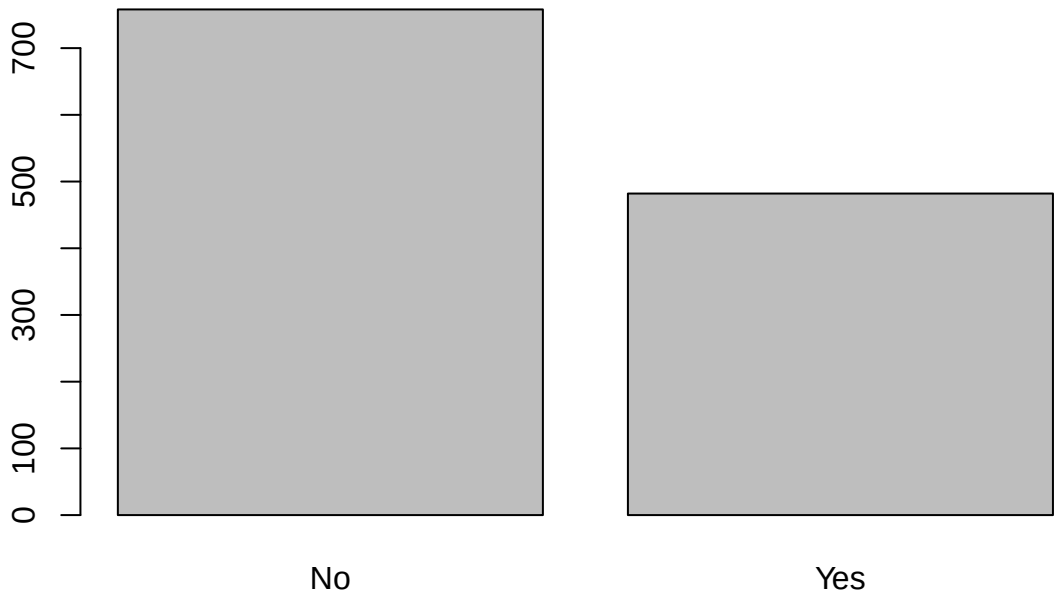
```
levels(surveyMentalHealth_clean$family_history)
```

```
## [1] "No"  "Yes"
```

```
summary(surveyMentalHealth_clean$family_history)
```

```
##  No Yes
## 758 482
```

```
barplot(table(surveyMentalHealth_clean$family_history),
        main="Antecedentes familiares")
```



treatment

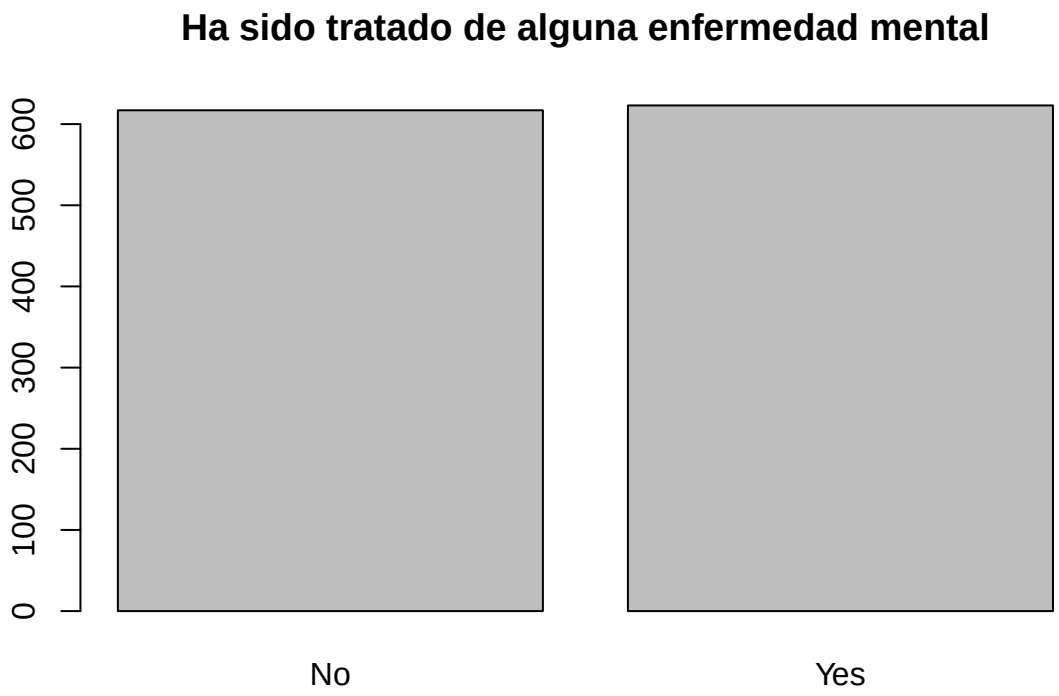
```
levels(surveyMentalHealth_clean$treatment)
```

```
## [1] "No"  "Yes"
```

```
summary(surveyMentalHealth_clean$treatment)
```

```
##  No Yes
## 617 623
```

```
barplot(table(surveyMentalHealth_clean$treatment),
        main="Ha sido tratado de alguna enfermedad mental")
```



work_interfere

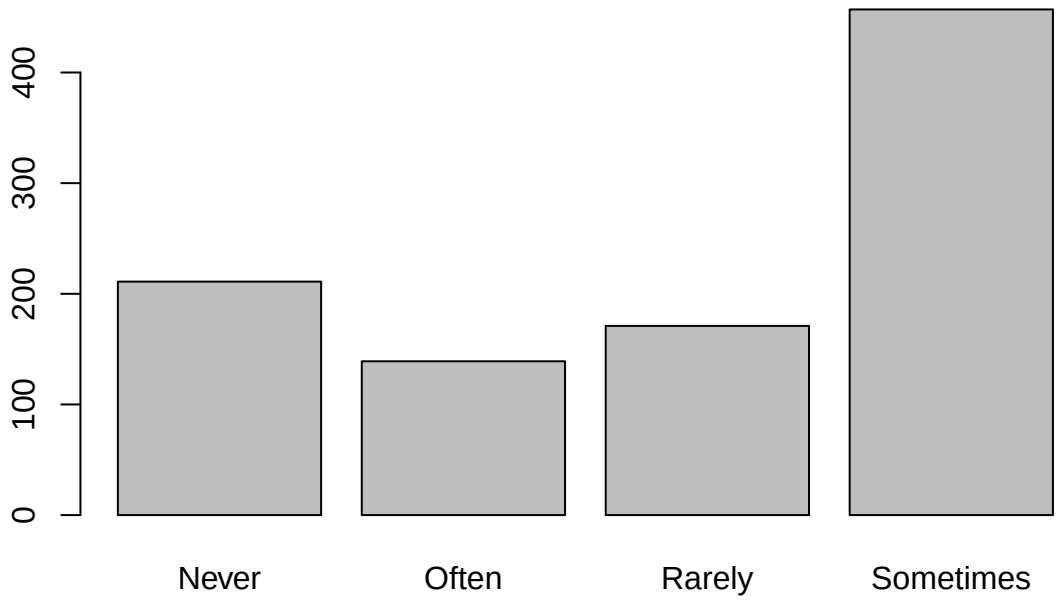
```
levels(surveyMentalHealth_clean$work_interfere)
```

```
## [1] "Never"     "Often"     "Rarely"    "Sometimes"
```

```
summary(surveyMentalHealth_clean$work_interfere)
```

```
##     Never     Often    Rarely Sometimes      NA's
##       211       139       171       457       262
```

```
barplot(table(surveyMentalHealth_clean$work_interfere),
        main="La enfermendad mental interfiere en su trabajo")
```



no_employees

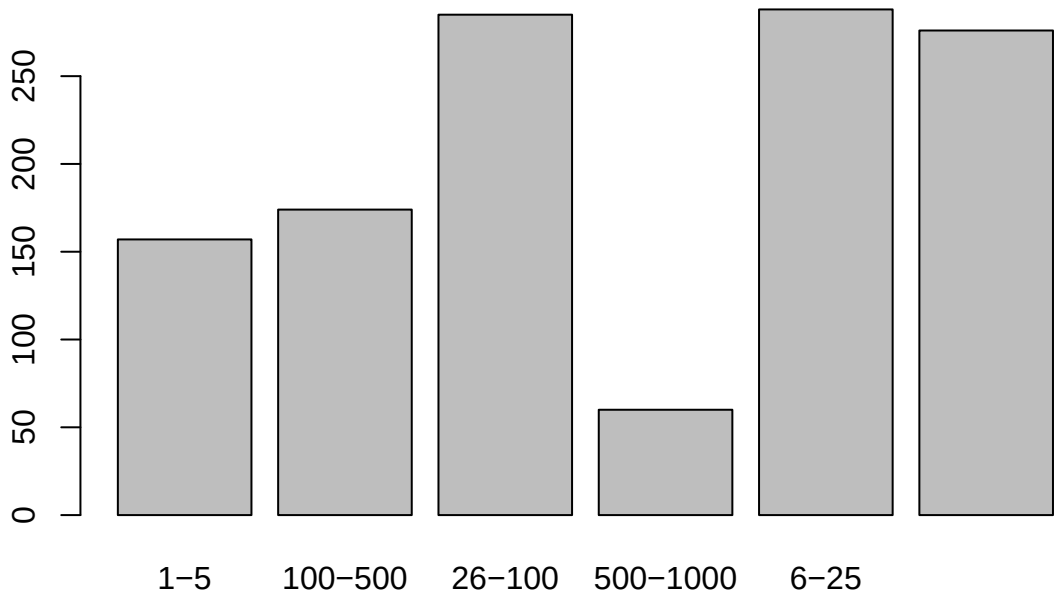
```
levels(surveyMentalHealth_clean$no_employees)
```

```
## [1] "1-5"            "100-500"        "26-100"         "500-1000"
## [5] "6-25"           "More than 1000"
```

```
summary(surveyMentalHealth_clean$no_employees)
```

```
##            1-5        100-500         26-100       500-1000           6-25
##            157            174            285             60            288
## More than 1000
##            276
```

```
barplot(table(surveyMentalHealth_clean$no_employees),
        main="Número de empleados de la compañia u organizacion")
```



remote_work

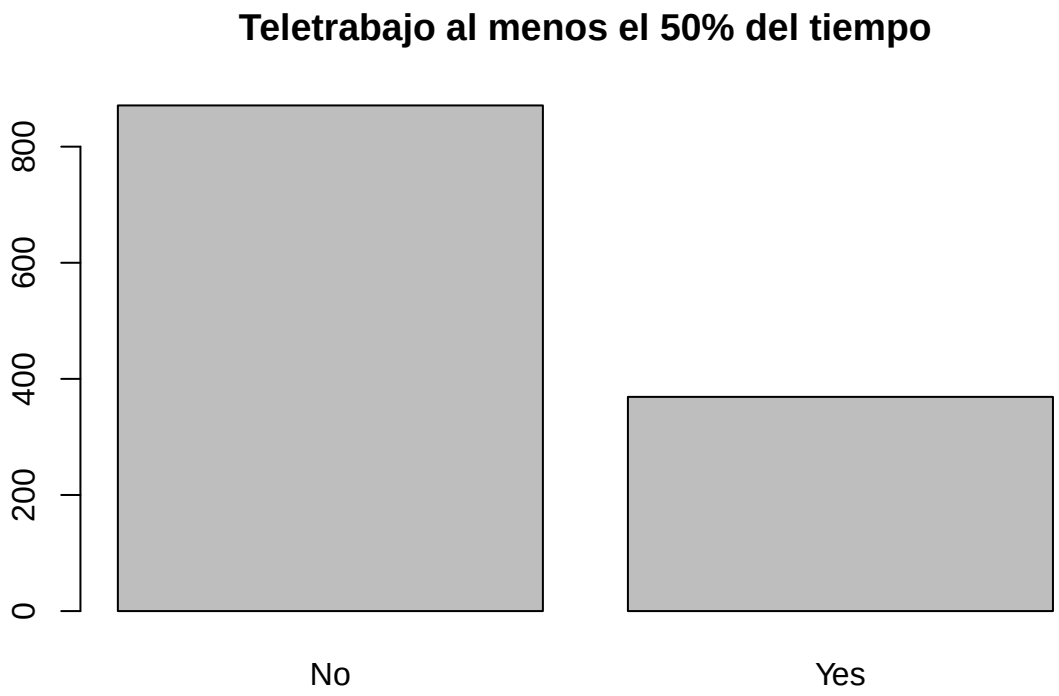
```
levels(surveyMentalHealth_clean$remote_work)
```

```
## [1] "No"  "Yes"
```

```
summary(surveyMentalHealth_clean$remote_work)
```

```
##  No Yes
## 871 369
```

```
barplot(table(surveyMentalHealth_clean$remote_work),
        main="Teletrabajo al menos el 50% del tiempo")
```



tech_company

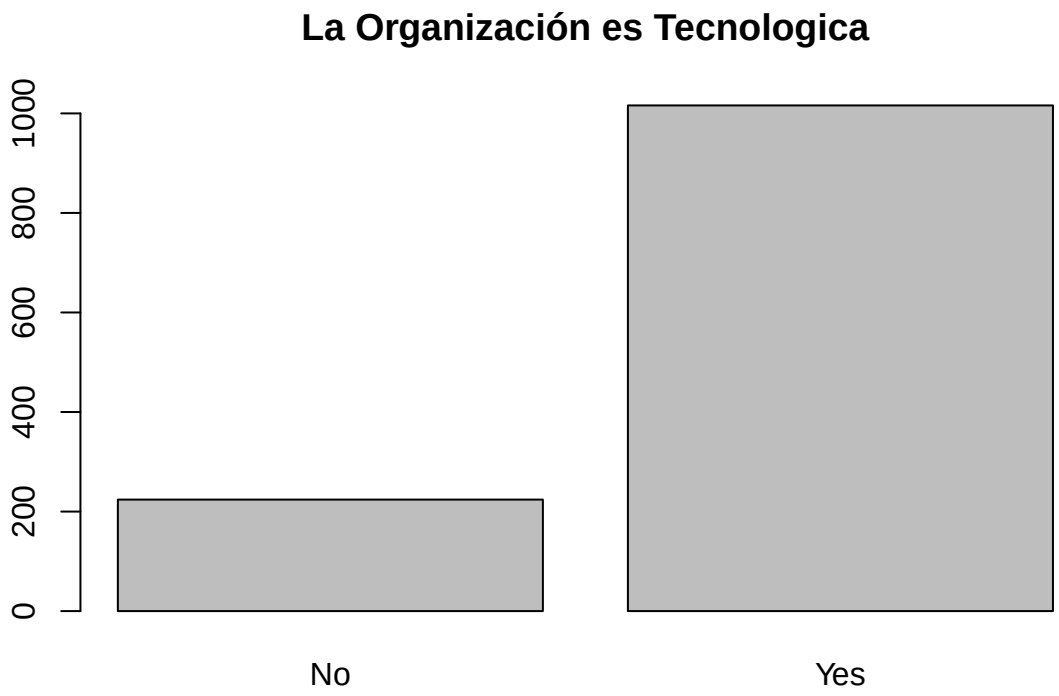
```
levels(surveyMentalHealth_clean$tech_company)
```

```
## [1] "No"  "Yes"
```

```
summary(surveyMentalHealth_clean$tech_company)
```

```
##   No  Yes
##  224 1016
```

```
barplot(table(surveyMentalHealth_clean$tech_company),
        main="La Organización es Tecnologica")
```



benefits

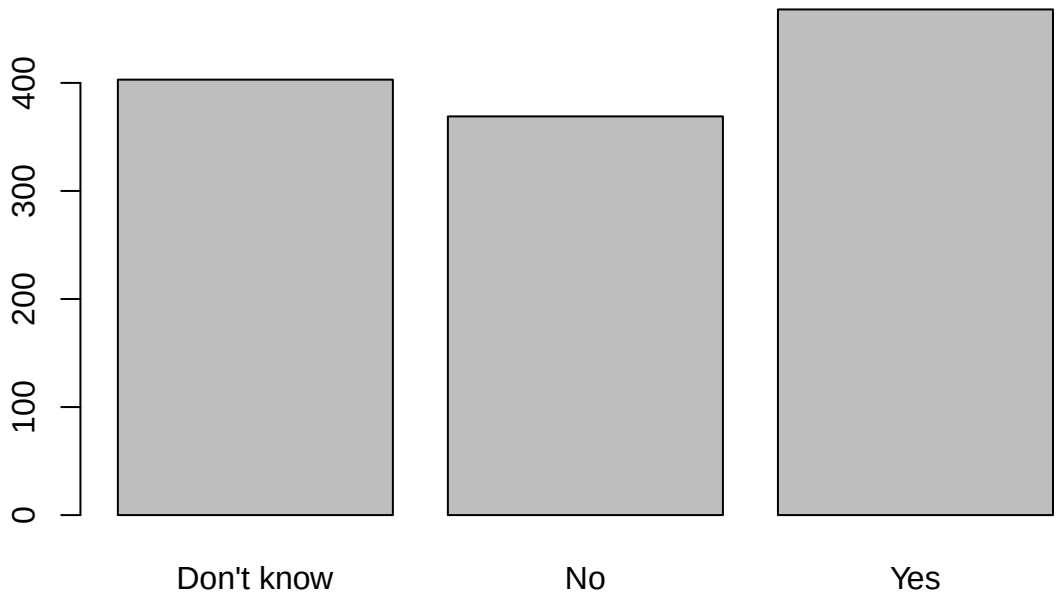
```
levels(surveyMentalHealth_clean$benefits)
```

```
## [1] "Don't know" "No"         "Yes"
```

```
summary(surveyMentalHealth_clean$benefits)
```

```
## Don't know         No        Yes
##        403        369        468
```

```
barplot(table(surveyMentalHealth_clean$benefits),
        main="La Organizacion provee de beneficios de salud Mental")
```



care_options

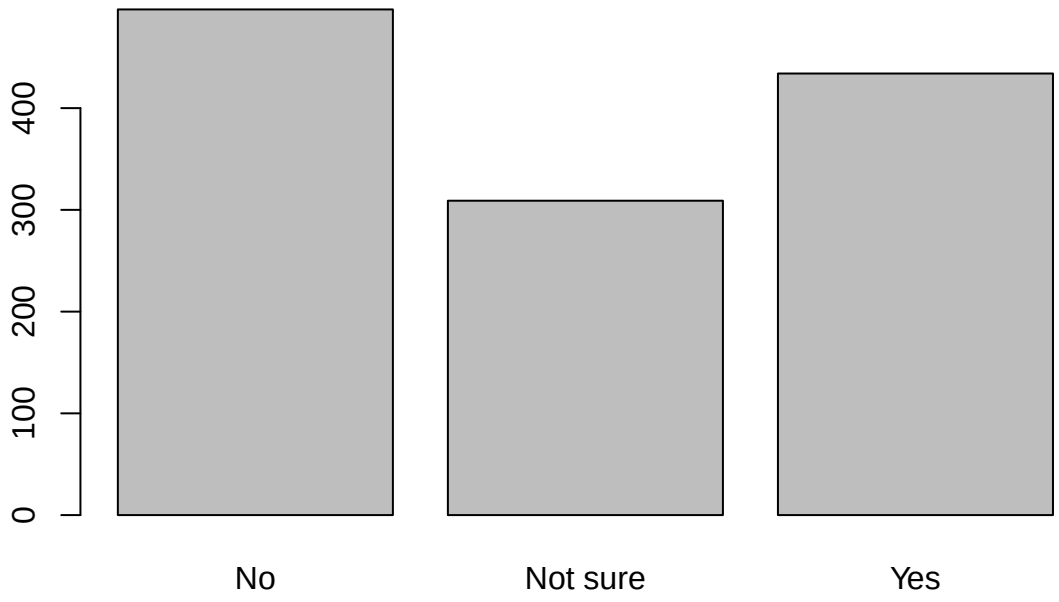
```
levels(surveyMentalHealth_clean$care_options)
```

```
## [1] "No"       "Not sure" "Yes"
```

```
summary(surveyMentalHealth_clean$care_options)
```

```
##        No Not sure       Yes
##       497       309       434
```

```
barplot(table(surveyMentalHealth_clean$care_options),
        main="Conoce Opciones de cuidado mental de su compañia médica")
```



wellness_program

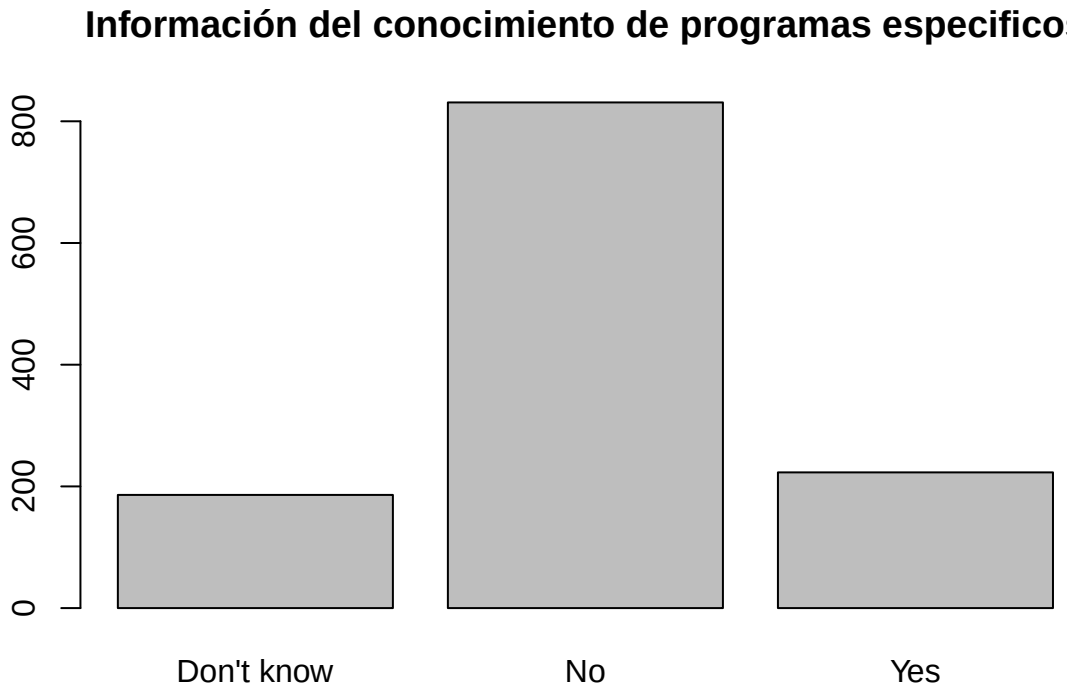
```
levels(surveyMentalHealth_clean$wellness_program)
```

```
## [1] "Don't know" "No"         "Yes"
```

```
summary (surveyMentalHealth_clean$wellness_program)
```

```
## Don't know         No        Yes
##        186        831        223
```

```
barplot(table(surveyMentalHealth_clean$wellness_program),
        main="Información del conocimiento de programas especificos")
```



seek_help

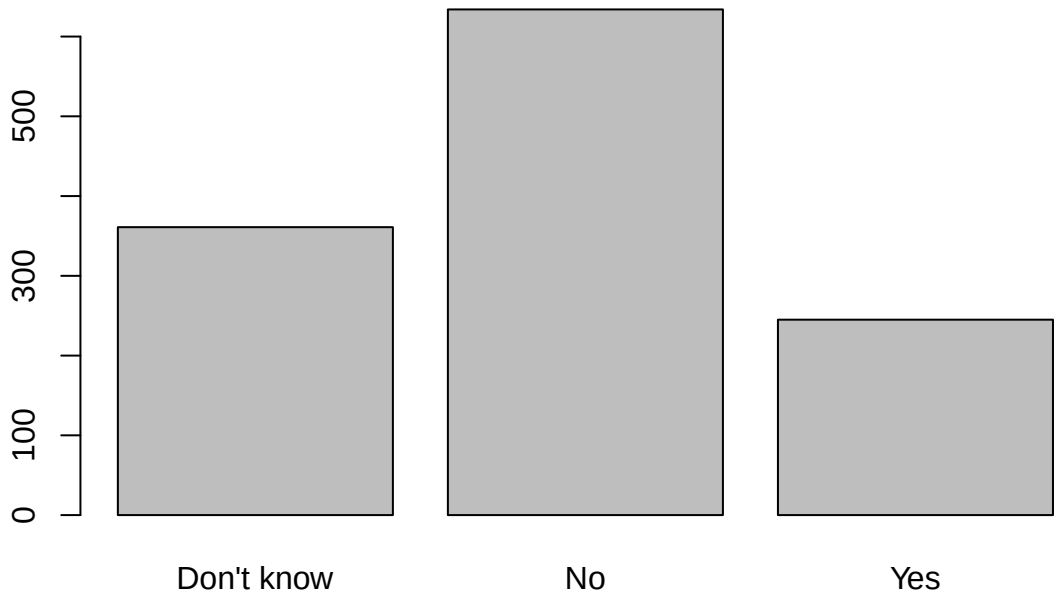
```
levels(surveyMentalHealth_clean$seek_help)
```

```
## [1] "Don't know" "No"         "Yes"
```

```
summary (surveyMentalHealth_clean$seek_help)
```

```
## Don't know         No        Yes
##        361        634        245
```

```
barplot(table(surveyMentalHealth_clean$seek_help),
        main="Información de recursos y ayuda desde la organización")
```



anonymity

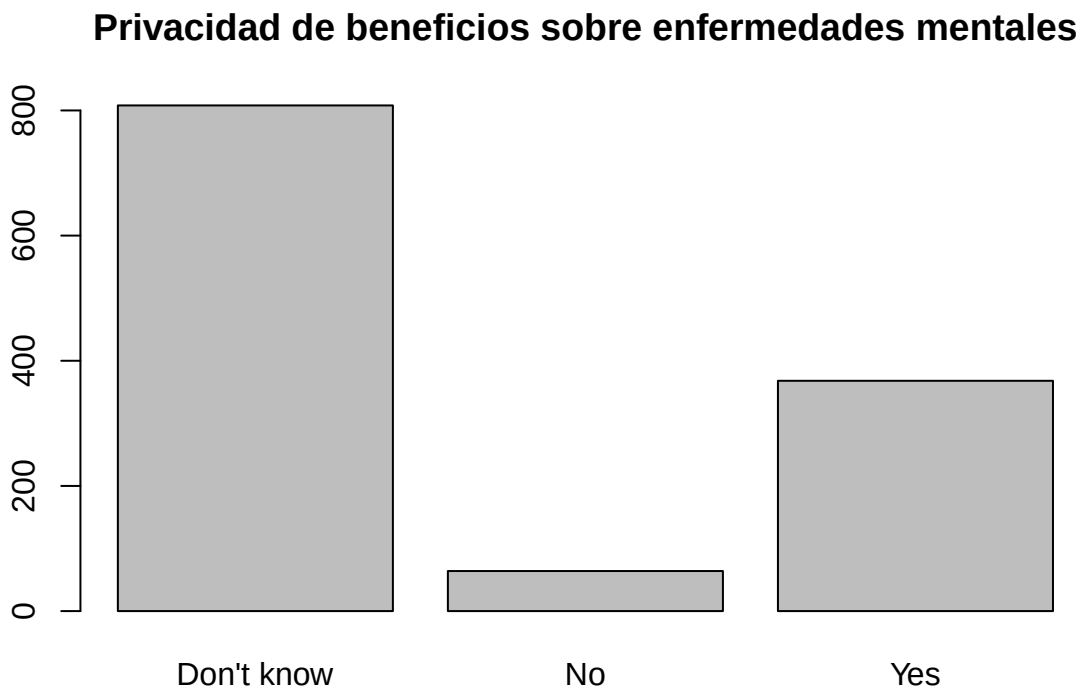
```
levels(surveyMentalHealth_clean$anonymity)
```

```
## [1] "Don't know" "No"         "Yes"
```

```
summary (surveyMentalHealth_clean$anonymity)
```

```
## Don't know         No        Yes
##        808         64        368
```

```
barplot(table(surveyMentalHealth_clean$anonymity),
        main="Privacidad de beneficios sobre enfermedades mentales")
```



leave

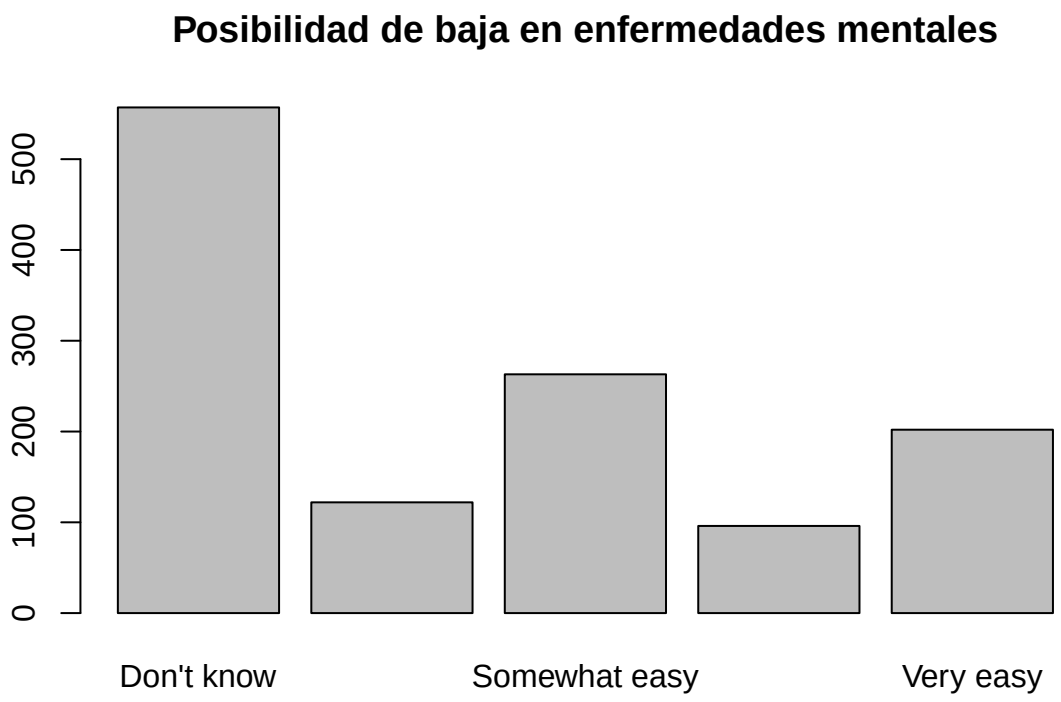
```
levels(surveyMentalHealth_clean$leave)
```

```
## [1] "Don't know"         "Somewhat difficult" "Somewhat easy"
## [4] "Very difficult"     "Very easy"
```

```
summary (surveyMentalHealth_clean$leave)
```

```
##         Don't know Somewhat difficult      Somewhat easy
##                557                122                263
##     Very difficult          Very easy
##                 96                202
```

```
barplot(table(surveyMentalHealth_clean$leave),
        main="Posibilidad de baja en enfermedades mentales")
```



mental_health_consequence

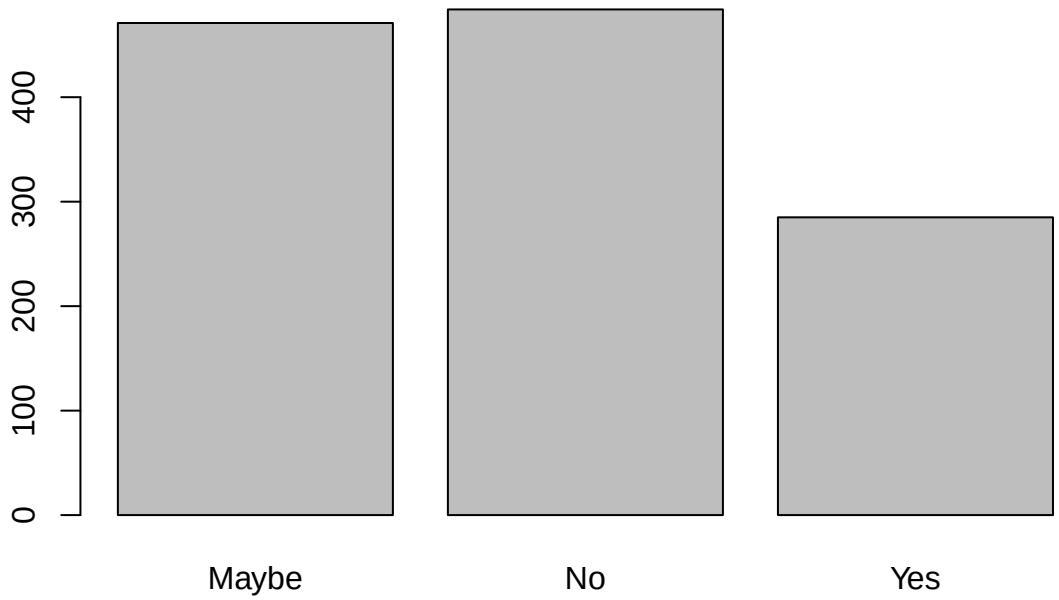
```
levels(surveyMentalHealth_clean$mental_health_consequence)
```

```
## [1] "Maybe" "No"    "Yes"
```

```
summary (surveyMentalHealth_clean$mental_health_consequence)
```

```
## Maybe    No   Yes
##   471   484   285
```

```
barplot(table(surveyMentalHealth_clean$mental_health_consequence),
        main="Consecuencias por hablar de salud mental")
```



phys_health_consequence

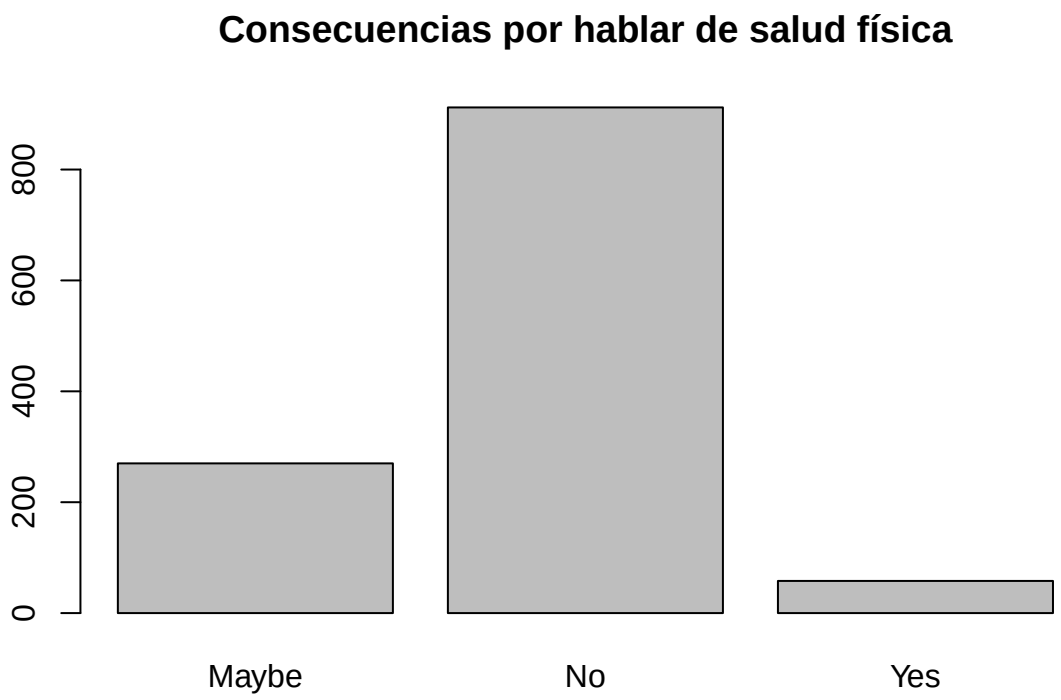
```
levels(surveyMentalHealth_clean$phys_health_consequence)
```

```
## [1] "Maybe" "No"    "Yes"
```

```
summary (surveyMentalHealth_clean$phys_health_consequence)
```

```
## Maybe    No   Yes
##   270   912    58
```

```
barplot(table(surveyMentalHealth_clean$phys_health_consequence),
        main="Consecuencias por hablar de salud física")
```



coworkers

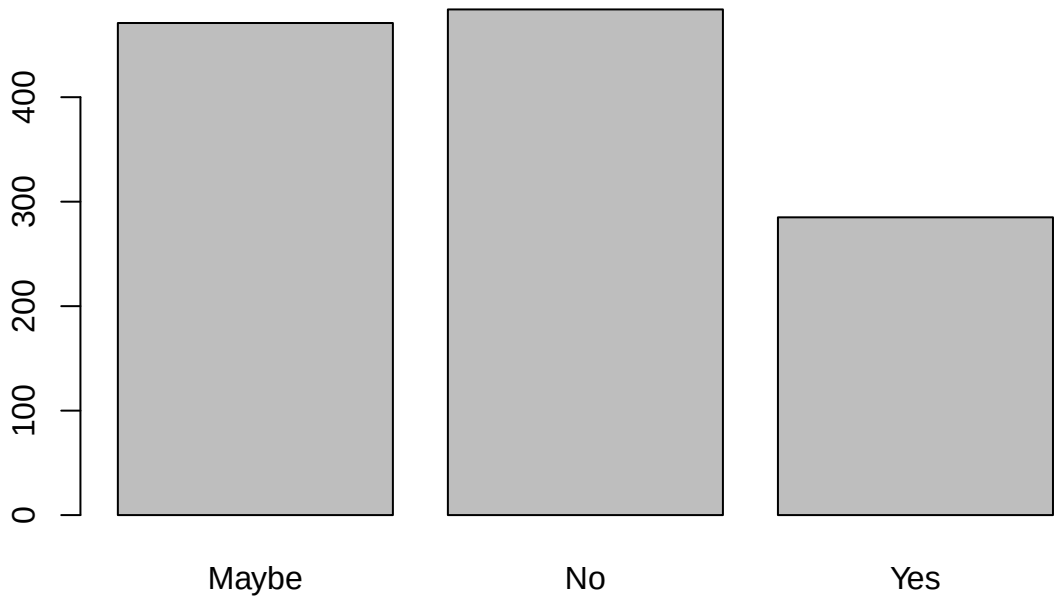
```
levels(surveyMentalHealth_clean$coworkers)
```

```
## [1] "No"           "Some of them" "Yes"
```

```
summary (surveyMentalHealth_clean$coworkers)
```

```
##           No Some of them          Yes
##          255          766          219
```

```
barplot(table(surveyMentalHealth_clean$mental_health_consequence),
        main="Hablaria de salud mental con compañeros")
```



supervisor

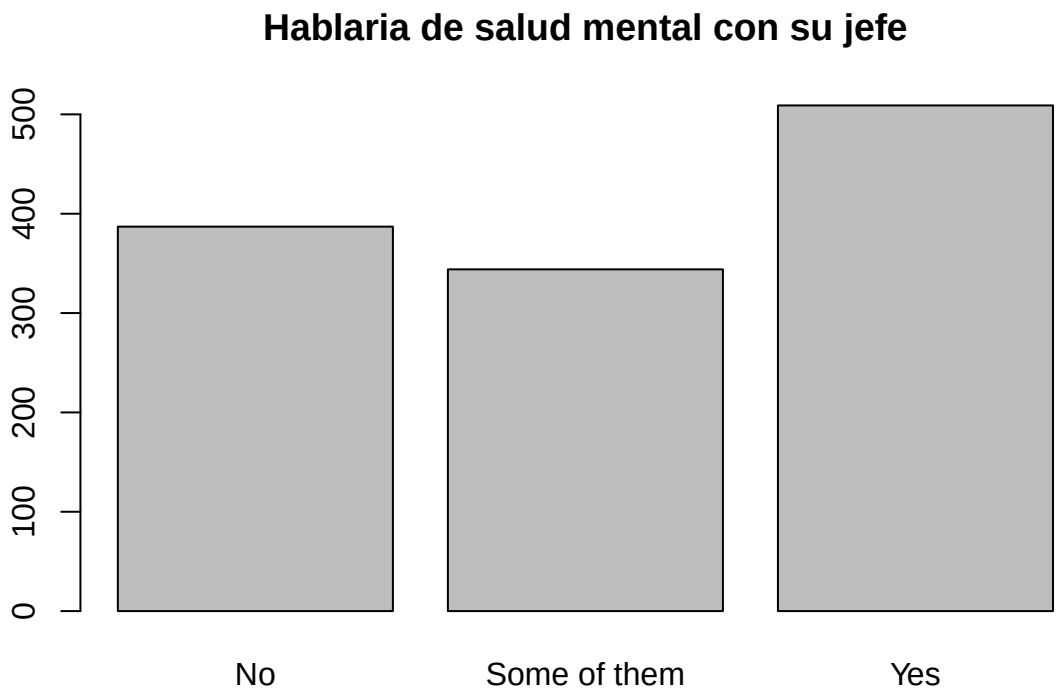
```
levels(surveyMentalHealth_clean$supervisor)
```

```
## [1] "No"           "Some of them" "Yes"
```

```
summary (surveyMentalHealth_clean$supervisor)
```

```
##           No Some of them          Yes
##          387          344          509
```

```
barplot(table(surveyMentalHealth_clean$supervisor),
        main="Hablaria de salud mental con su jefe")
```



mental_health_interview

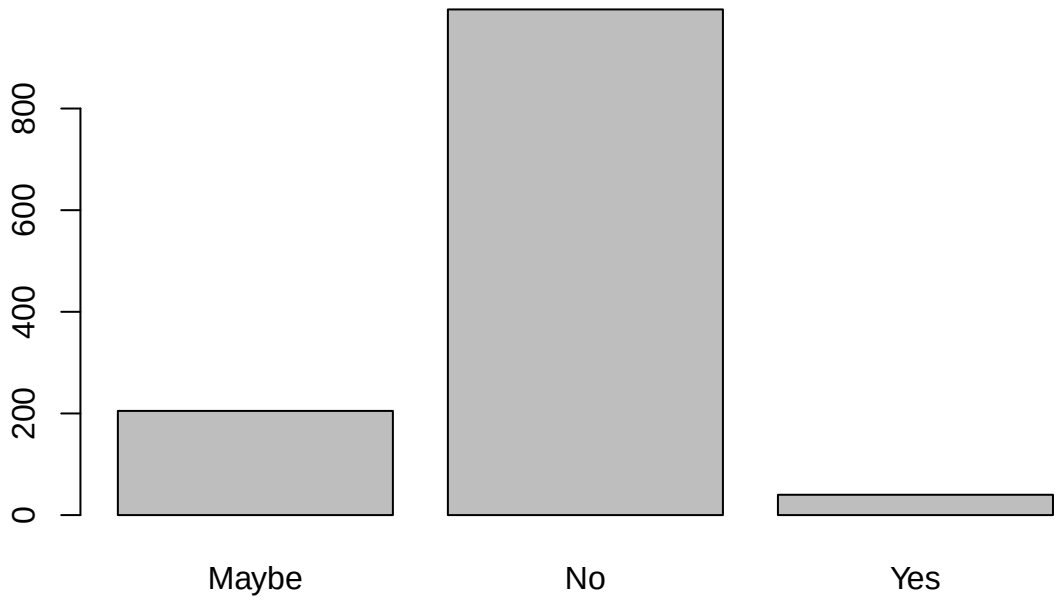
```
levels(surveyMentalHealth_clean$mental_health_interview)
```

```
## [1] "Maybe" "No"    "Yes"
```

```
summary (surveyMentalHealth_clean$mental_health_interview)
```

```
## Maybe    No   Yes
##   205   995    40
```

```
barplot(table(surveyMentalHealth_clean$mental_health_interview),
        main="Hablaria de salud mental en una entrevista laboral")
```



phys_health_interview

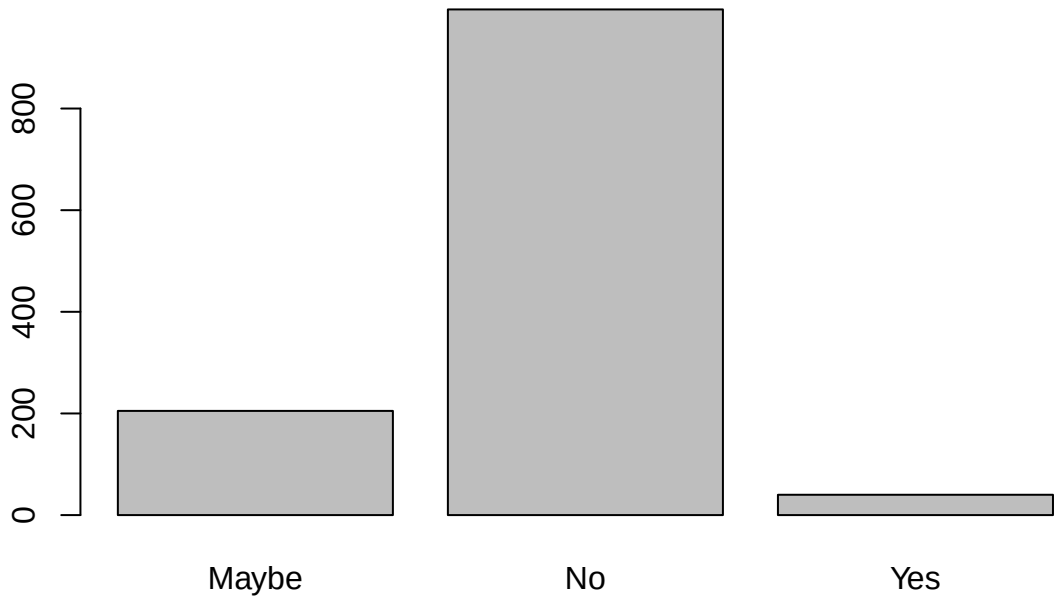
```
levels(surveyMentalHealth_clean$phys_health_interview)
```

```
## [1] "Maybe" "No"    "Yes"
```

```
summary (surveyMentalHealth_clean$phys_health_interview)
```

```
## Maybe    No   Yes
##   550   492   198
```

```
barplot(table(surveyMentalHealth_clean$mental_health_interview),
        main="Hablaria de salud física en una entrevista laboral")
```



mental_vs_physical

```
levels(surveyMentalHealth_clean$mental_vs_physical)
```

```
## [1] "Don't know" "No"         "Yes"
```

```
summary (surveyMentalHealth_clean$mental_vs_physical)
```

```
## Don't know         No        Yes
##        573        334        333
```

```
barplot(table(surveyMentalHealth_clean$mental_vs_physical),
        main="Importacia en la Organizacion de la salud mental sobre la física")
```

obs_consequence

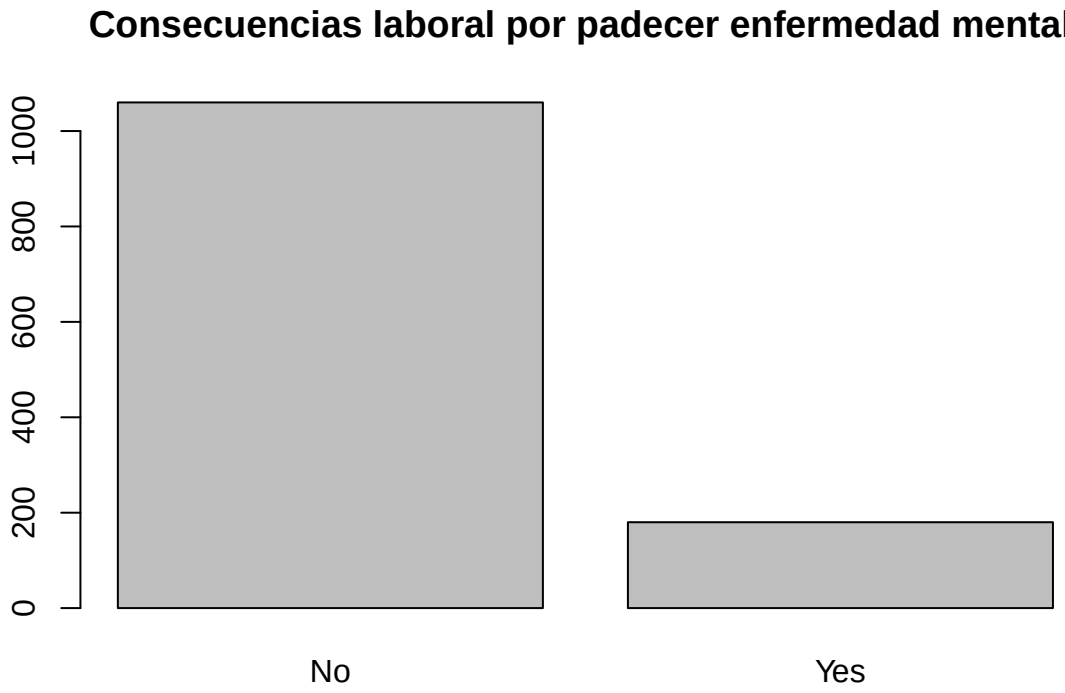
```
levels(surveyMentalHealth_clean$obs_consequence)
```

```
## [1] "No"  "Yes"
```

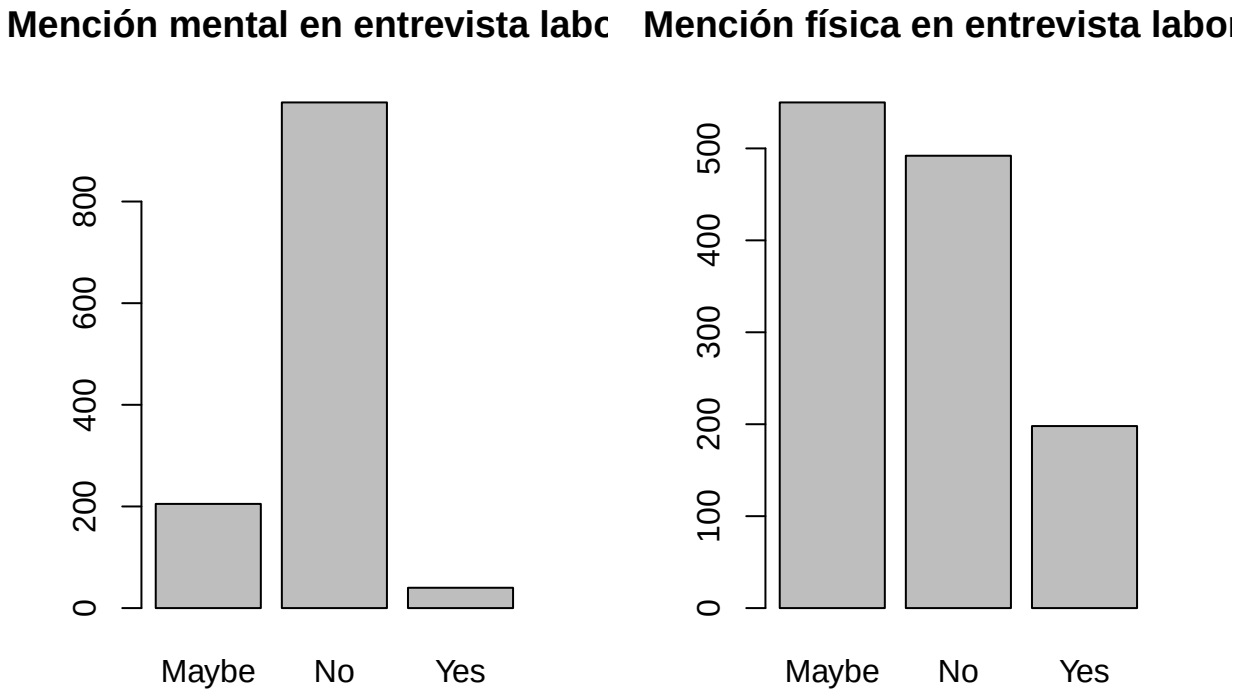
```
summary (surveyMentalHealth_clean$obs_consequence)
```

```
##   No  Yes
## 1060  180
```

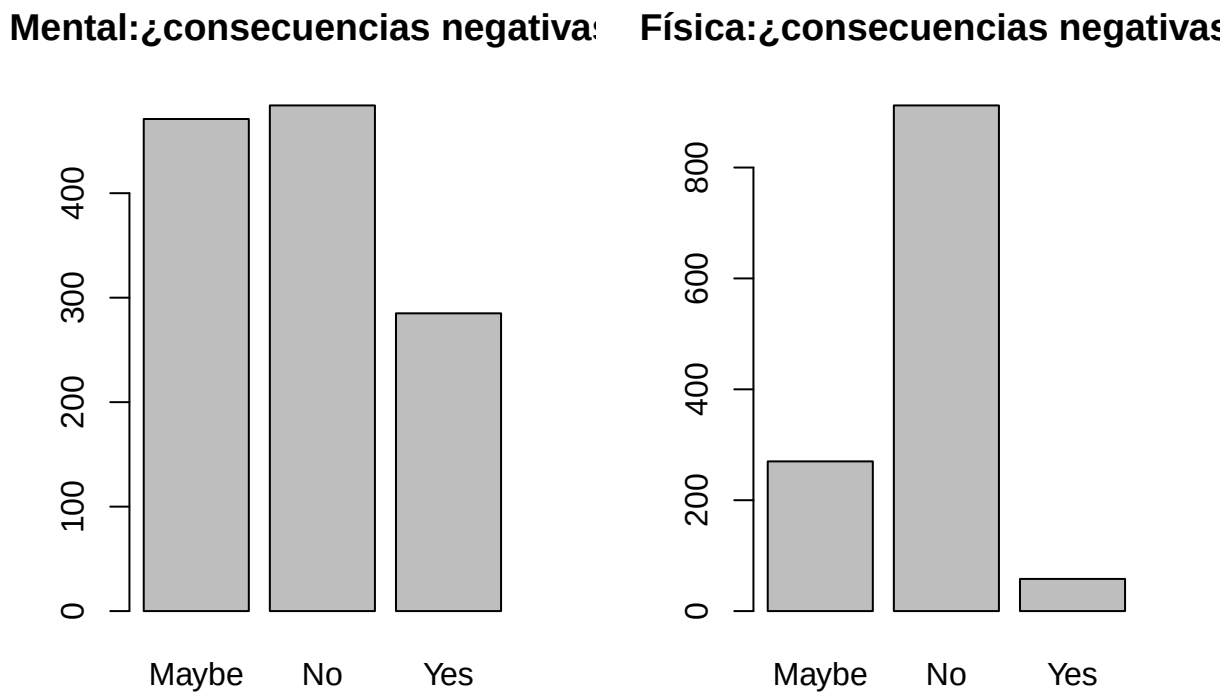
```
barplot(table(surveyMentalHealth_clean$obs_consequence),
        main="Consecuencias laboral por padecer enfermedad mental ")
```




```
#Gráfica comparativa 1
    par(mfrow=c(1,2))
    barplot(table(surveyMentalHealth_clean$mental_health_interview),
            main="Mención mental en entrevista laboral")
    barplot(table(surveyMentalHealth_clean$phys_health_interview),
            main="Mención física en entrevista laboral")
```




```
#Gráfica comparativa 2
    par(mfrow=c(1,2))
    barplot(table(surveyMentalHealth_clean$mental_health_consequence),
            main="Mental:¿consecuencias negativas?")
    barplot(table(surveyMentalHealth_clean$phys_health_consequence),
            main="Física:¿consecuencias negativas?")
```



Vamos a investigar:

-El hecho de recibir tratamiento tiene algo que ver con la edad, es decir, si existen diferencias en la variable Age según la variable treatment (Tratamiento).

-Dependiendo de la edad del individuo como percibe este el hecho de que la organización de igual importancia a la salud mental vs salud física, es decir, si existen diferencias en la variable Age según la variable mental_vs_physical

El analisis de la normalidad o contrastes de normalidad, investigan cuanto de lejos esta la distribución de los valores observados con respecto a una distribución normal con la misma media y desviación típica. Para este analisis inicialmente podemos realizar unos estudios de manera gráfica. Vamos a comenzar observando si existe diferencias significativa según la edad del individuo para pensar que la organización da igual importancia a la salud mental o a la salud fisica.

Cálculo de la media

```
#Calculo Media
mean(surveyMentalHealth_clean$Age)
```

```
## [1] 32.11452
```

Cálculo de la mediana

```
#Calculo Mediana
median(surveyMentalHealth_clean$Age)
```

```
## [1] 31
```

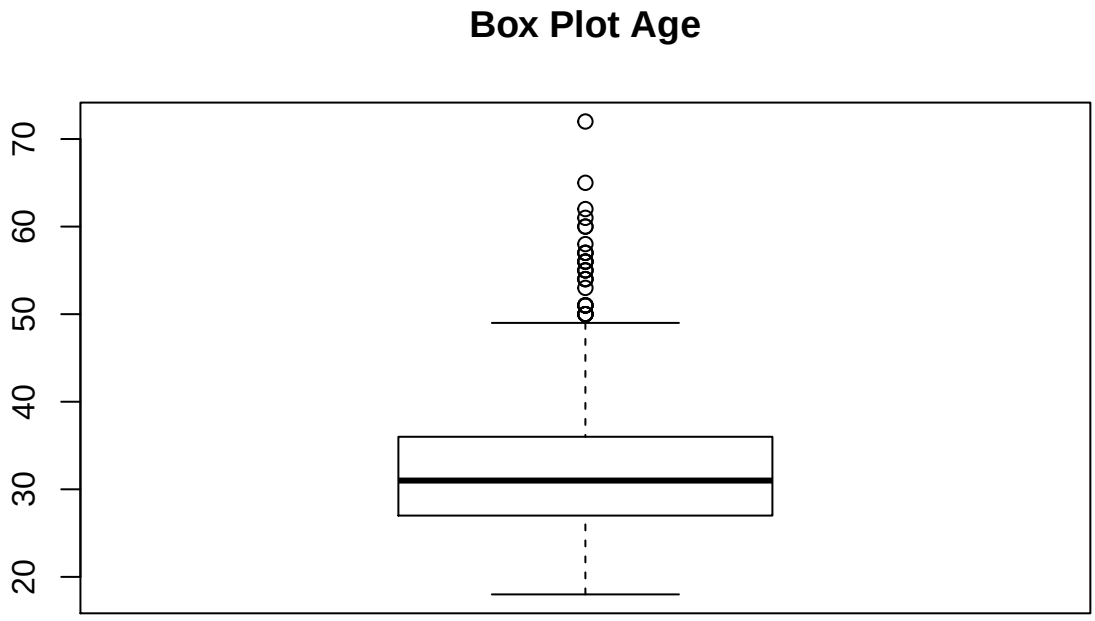
Cálculo de los cinco números de Tukey (Mínimo, Q1, Mediana, Q3 y Máximo)

```r
#Sumario de los cinco números (Mínimo, Q1, Mediana, Q3, Maximo)
fivenum(surveyMentalHealth_clean$Age)
```

```
## [1] 18 27 31 36 72
```

Gráfico de Boxplot

```r
#Diagrama de caja (Boxplot)
boxplot(surveyMentalHealth_clean$Age, main="Box Plot Age")
```



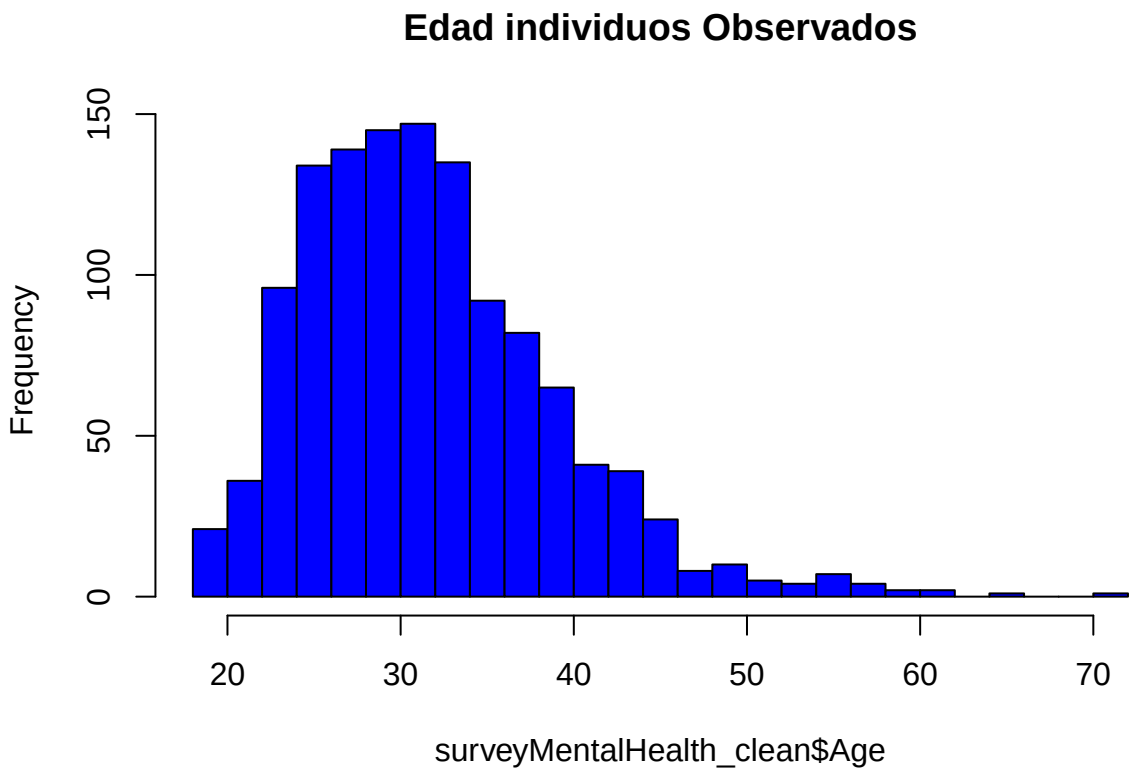
En un gráfico de Boxplot podemos estudiar la simetría, detectar outlier e incluso contrastar algunas hipótesis de la distribución. El gráfico fracciona los datos en 4 partes de igual frecuencia, es decir, cada grupo contiene mas o menos el mismo número de observaciones. Pero la ocupación de estos es diferente.El primer grupo(desde el valor mas pequeño hasta Q1) los valores de la variable Age va desde 18 hasta 27. El último grupo va (desde Q3 hasta el máximo valor) desde 36 hasta 72. Podemos observar que la longitud desde el mínimo hasta Q1 es diferente a la de Q3 al máximo, por lo que podemos decir que no existe simetria con respecto a la mediana, por tanto podemos hablar de asimetria. El 50% de los individuos observados tienen Age entre Q1 y Q3.

Realizamos una representación de un histograma y superponemos una curva normal o función de densidad estimada para que se pueda ver la forma de la gráfica. Representamos el histograma de la variable Age de la muestra. Para calcular el número de clases que necesitamos realizamos el siguiente cálculo $k = 1 + 3,3 * log(n)$ ó $k = \sqrt{(n)}$.

```r
#Calculamos el numero de intervalor
k_Age<- round(sqrt(length(surveyMentalHealth_clean$Age)))
k_Age
```

```
## [1] 35
```

```
hist(surveyMentalHealth_clean$Age ,main="Edad individuos Observados",
     breaks=k_Age, col="blue")
```



Obtenemos de los datos observados en el histograma

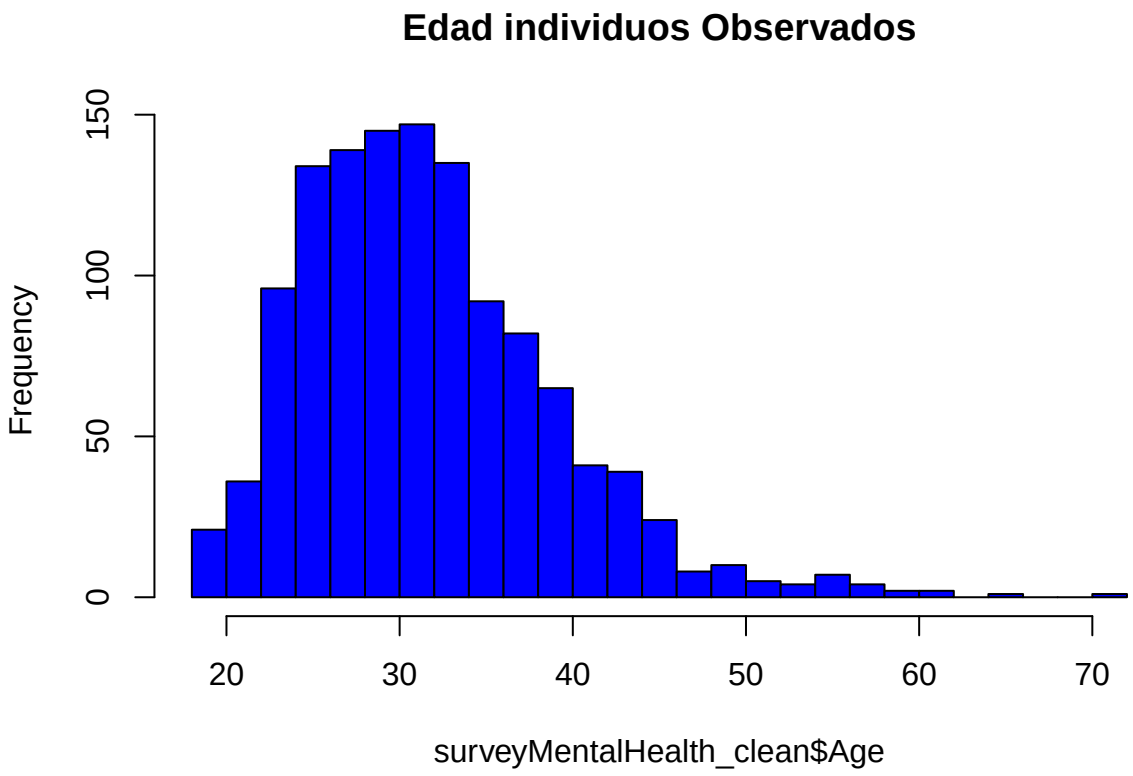
Los valores utilizados para dibujar el histograma son:

$breaks Información de los extremos de los intervalos

$counts Información de las frecuencias absolutas

$mids Información de los valores de las marcas de las clases (puntos medios)

```
hh_Age<-hist(surveyMentalHealth_clean$Age ,main="Edad individuos Observados",
     breaks=k_Age, col="blue")
```



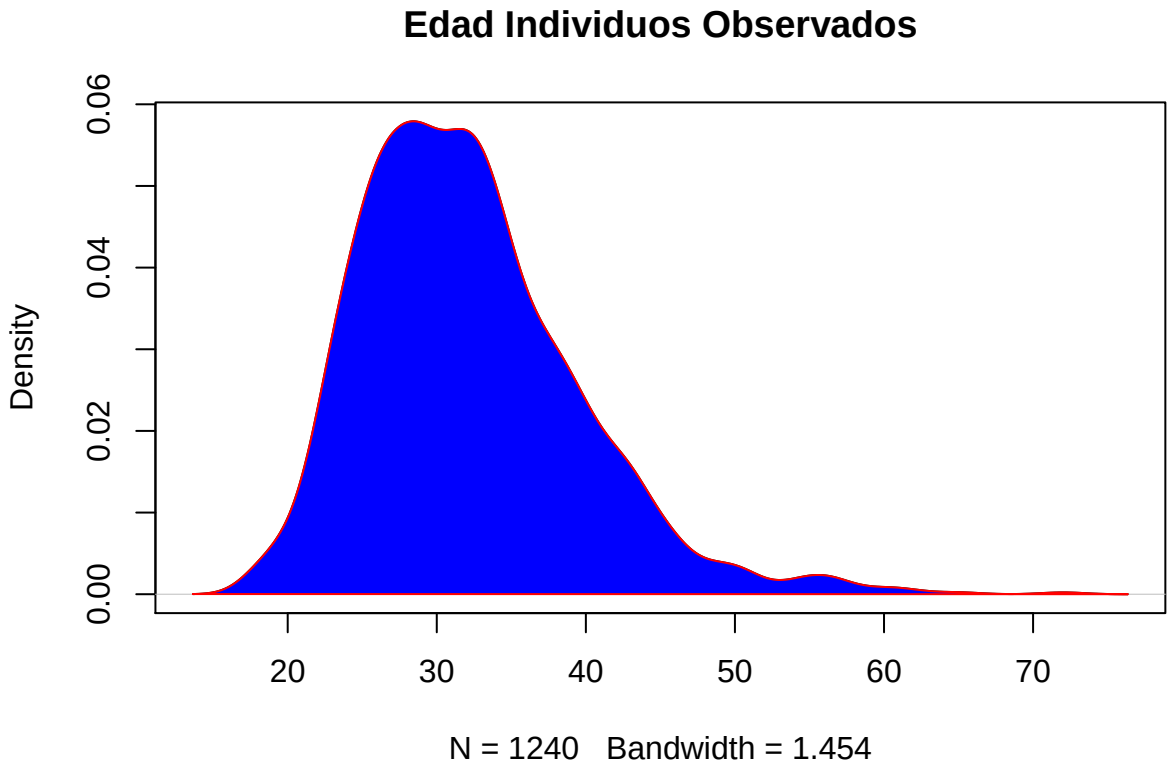

```
hh_Age
```

```
## $breaks
##  [1] 18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48 50 52 54 56 58 60 62
## [24] 64 66 68 70 72
##
## $counts
##  [1]  21  36  96 134 139 145 147 135  92  82  65  41  39  24   8  10   5
## [18]   4   7   4   2   2   0   1   0   0   1
##
## $density
##  [1] 0.0084677419 0.0145161290 0.0387096774 0.0540322581 0.0560483871
##  [6] 0.0584677419 0.0592741935 0.0544354839 0.0370967742 0.0330645161
## [11] 0.0262096774 0.0165322581 0.0157258065 0.0096774194 0.0032258065
## [16] 0.0040322581 0.0020161290 0.0016129032 0.0028225806 0.0016129032
## [21] 0.0008064516 0.0008064516 0.0000000000 0.0004032258 0.0000000000
## [26] 0.0000000000 0.0004032258
##
## $mids
##  [1] 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51 53 55 57 59 61 63
## [24] 65 67 69 71
##
## $xname
## [1] "surveyMentalHealth_clean$Age"
##
## $equidist
```

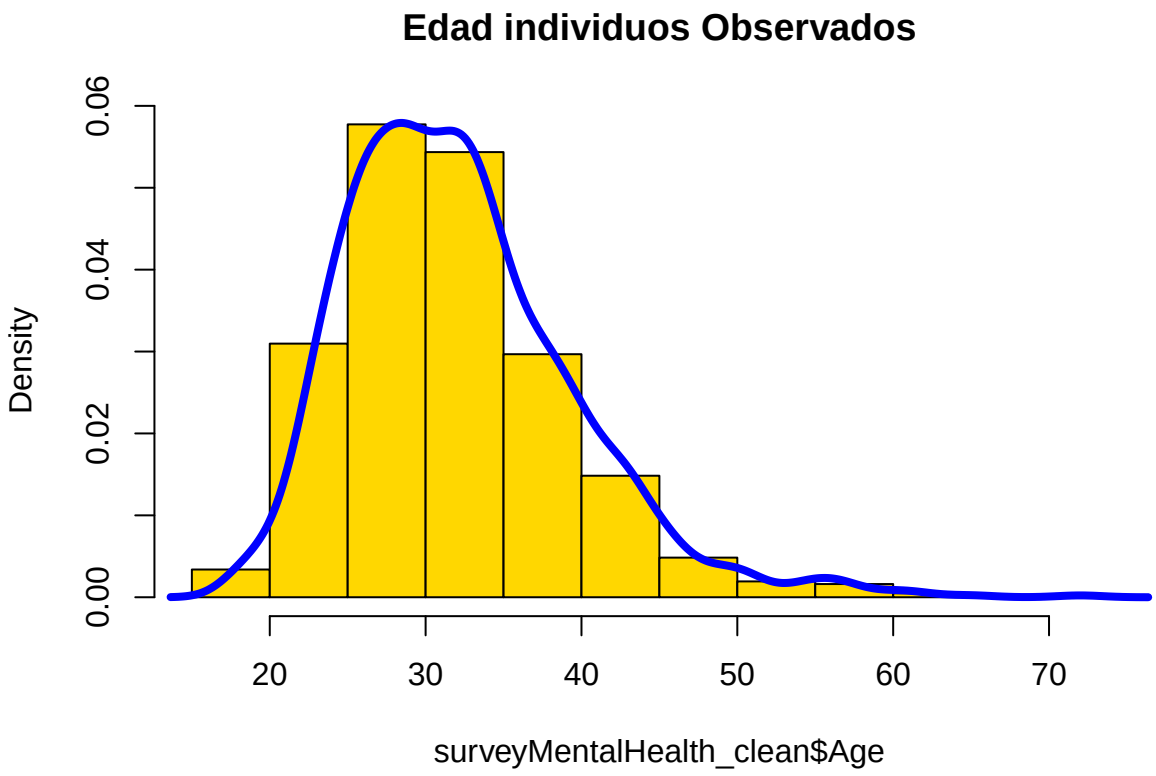
```
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
```

Superponemos la gráfica de la función de la densidad

```
#Calculo de la función de densidad
den_Age<- density(surveyMentalHealth_clean$Age)
plot(den_Age ,main="Edad Individuos Observados")
polygon(den_Age , col="blue", border="red")
```

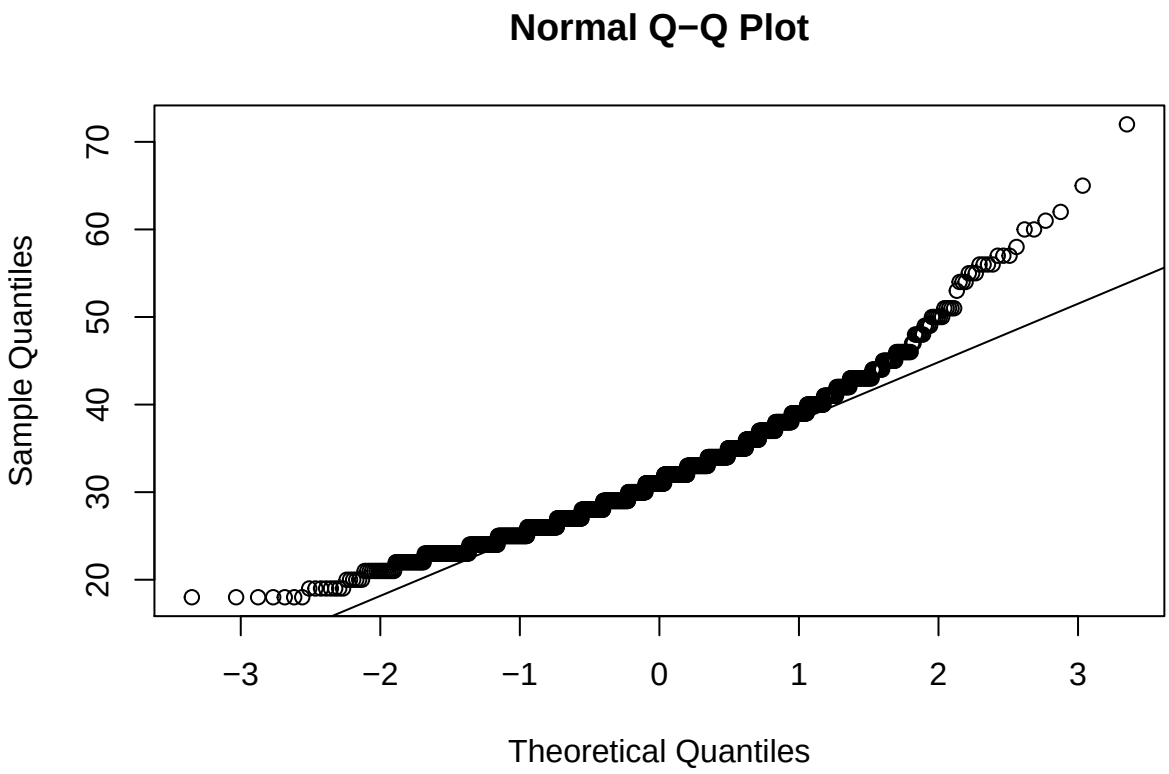



```
#Superposición de las gráficas
hist(surveyMentalHealth_clean$Age ,main="Edad individuos Observados",
     col="gold",freq=FALSE)
lines(den_Age,col="blue",lwd=4)
```



Vamos a utilizar tambien el gráfico de los cuantiles teóricos (Graficos Q-Q). Estos consisten en la comparación de los cuantiles de la distribución observada con los cuartiles teóricos de la distribución normal. Cuanto más se asemejen a una normal, mas alineados están los puntos a una recta.

```
qqnorm(surveyMentalHealth_clean$Age)
qqline(surveyMentalHealth_clean$Age)
```



Realizamos el estudio de la normalidad mediante los contraste de hipótesis. Tenemos diferentes test de hipótesis:

-Test de Shapiro-Wilk: Para muestras de tamaño menor de 50

-Test de Kolmogorov-Smirnov

-Lillefors:Da por hecho que la media y la varianza son desconocidas. Se considera que cuando tenemos muestras con tamaño superior a 50 es la alternativa de Shapiro-Wilk

-Test Jarque-Bera: Esta da valor a la alejania que existe entre los coeficientes de asimetria y curtosis de los esperados por una distribución normal.

Todos estos test tenemos como hipotesis nula que los datos proceden de una distribución normal y la hipótesis alternativa que no lo hacen. El p_{value} nos da la probabilidad de tener una distribución como la observada siempre y cuando los datos proceden de una población con distribución normal. Al estar hablando de p_{value}, hay que tener en cuenta que a mayor tamaño de la muestra más finos son los test y es mas sencillo encontrar evidencias en contra de H_o. De igual manera, a mayor tamaño de la muestra menos sensibles son los test paramétricos en falta de normalidad. No realizamos el test de Shapiro-Wilk ya que nuestra muestra tiene un tamaño mayor a 50. Vamos a utilizar el test de Kolmogorov-Smirnov, para estudiar si una muestra proviene de una población con una distribución de media y desviación típica especifica.

```
ks.test(x=surveyMentalHealth_clean$Age,"pnorm", mean(surveyMentalHealth_clean$Age), sd(surveyMentalHeal
```

```
## Warning in ks.test(x = surveyMentalHealth_clean$Age, "pnorm",
## mean(surveyMentalHealth_clean$Age), : ties should not be present for the
## Kolmogorov-Smirnov test
```

```
##
##  One-sample Kolmogorov-Smirnov test
```

```
##
## data:  surveyMentalHealth_clean$Age
## D = 0.087147, p-value = 1.322e-08
## alternative hypothesis: two-sided
```

Como ya hemos dicho anteriormente el test de Kolmogorov-Smirnov acepta que conoce la media y varianza poblacional, lo que hace que dicho test sea conservador y poco potente. Asi tenemos el test de Lilliefors, en este caso se acepta que la media y la varianza son desconocidas.

```
lillie.test((x=surveyMentalHealth_clean$Age))
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  (x = surveyMentalHealth_clean$Age)
## D = 0.087147, p-value < 2.2e-16
```

Podemos tener en cuenta tambien el test de normalidad de Jarque-Bera, este no pide estimación de los parámetros con los que podemos caracterizar una normal. Este lo que hace es saber lo que se alejan los coeficientes de asimetria y curtosis de una distribución normal.

```
jb.norm.test(x=surveyMentalHealth_clean$Age)
```

```
##
##  Jarque-Bera test for normality
##
## data:  surveyMentalHealth_clean$Age
## JB = 392.78, p-value < 2.2e-16
```

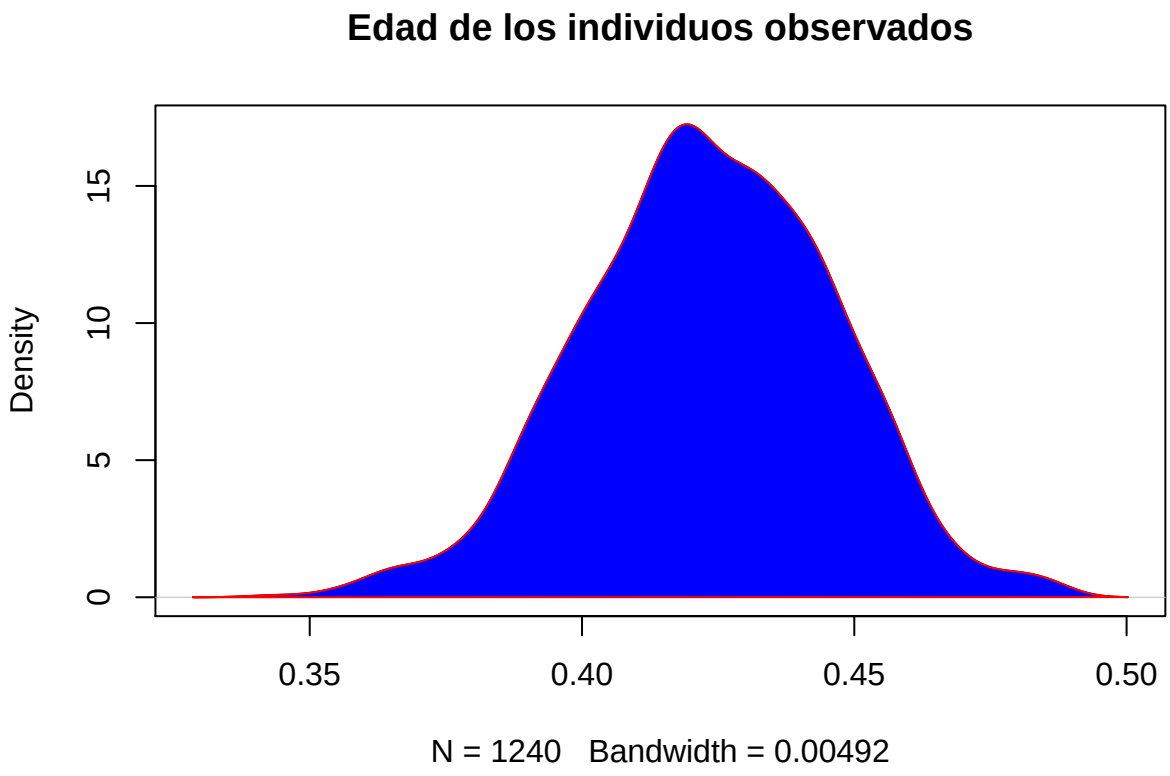
Si no podemos asumir normalidad este hecho nos influira en los test de hipótesis paramétricos y en los modelos de regresion luego los estimadores calculados por mínimos cuadrados no seran eficientes y tanto los intervalosde confianza de los parámetros del modelo como constrantes significativos serán unicamente aproximados y nos exactos. Si tenemos en cuenta el teorema del límite central el cual necesita que las poblacion de las que procede la muestra sea una normal, no las muestras. Si la muestra se distribuye según una normal esta claro que la población tambien lo hara. Puede ocurrir que la muestra no se distribuye segun una norma pero si conocemos que la población se distribuye segun una normal, entonces los contrastes parámetricos si son válidos. El Teorema del Limite Central permite simplificar los requisitos de normalidad cuando las muestras son grandes.

Procedemos a rechazar la hipotesis nula de normalidad ya que en todos los test obtenemos un p_{value}<0.05.El hecho de no cumplir la condición no tiene un efecto grave si el tamaño de la muestra es suficientemente grande. Como n>30 (que es nuestro caso) por el Teorema Central del límite, garantizamos robustez del análisis.

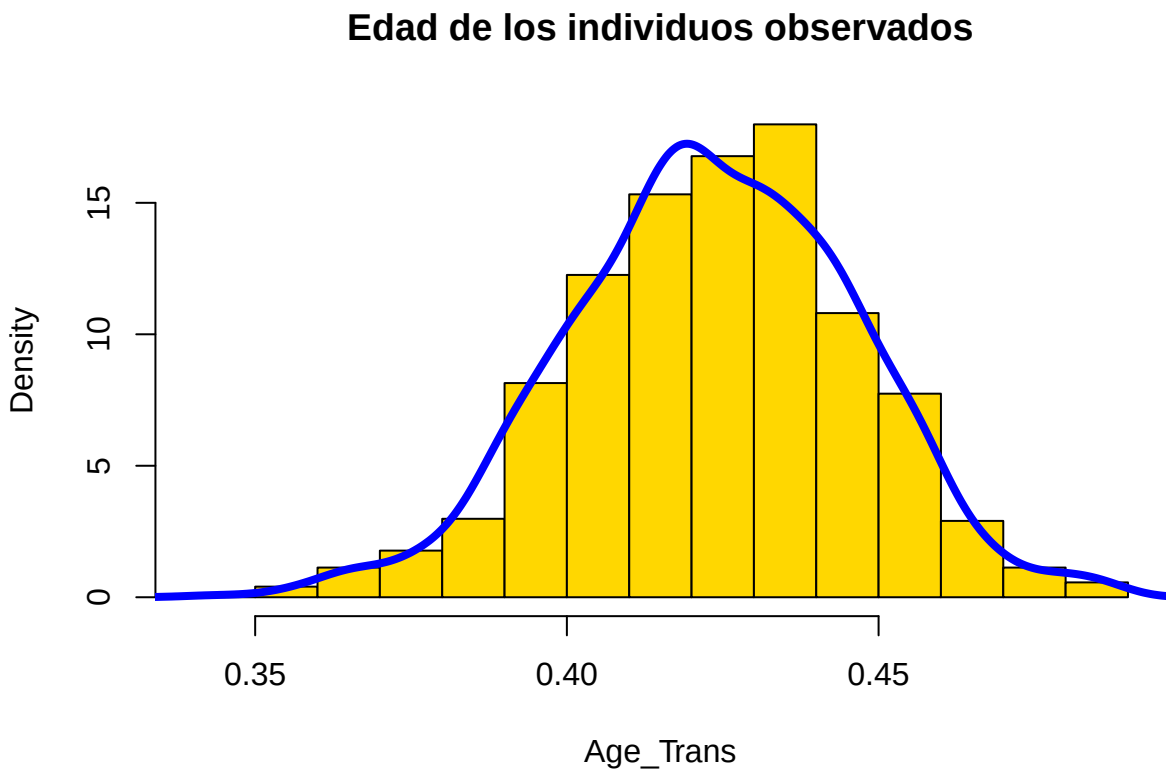
Vamos a realizar la transformación $y = \sqrt{(}\sqrt{(}1/x)$.

```
Age_Trans<-(sqrt(sqrt(1/surveyMentalHealth_clean$Age)))
```

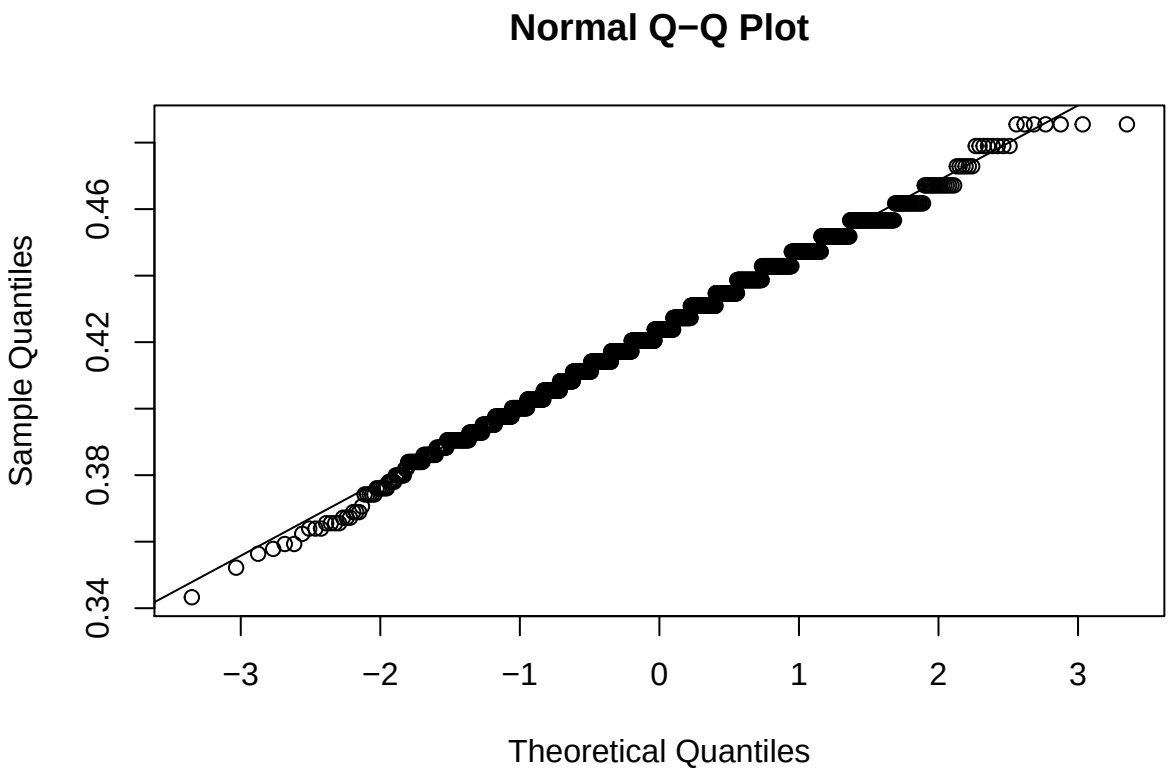
```
#Calculo de la función de densidad
den_Age_Trans<- density(Age_Trans)
plot(den_Age_Trans ,main="Edad de los individuos observados")
polygon(den_Age_Trans , col="blue", border="red")
```




```
#Superposición de las gráficas
hist(Age_Trans ,main="Edad de los individuos observados",
     col="gold",freq=FALSE)
lines(den_Age_Trans ,col="blue",lwd=4)
```




```
qqnorm(Age_Trans)
qqline(Age_Trans)
```




```
lillie.test((x=Age_Trans))
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  (x = Age_Trans)
## D = 0.043067, p-value = 1.264e-05
```

Aun realizando la transformación no conseguimos tener normalidad. Luego continuaremos con los datos origenes sin tener en cuenta la transformada.

A continuación estudiamos la homogeneidad de la varianza u homocedasticidad, se esta considerando que la varianza es constante en los diferentes niveles. Tenemos diferentes test para evaluar la distribución de la varianza. En todos ellos estamos considerando como hipotesis nula que la varianza es la misma en todos los grupos y como hipótesis alternativa que no lo es.

-F-Test. Razón de varianzas: Es recomendado siempres y cuando se tenga la certeza de que las poblaciones se distribuyen con normalidad. Luego es muy sensible en caso de no cumplir normalidad

-Test de Levene: Se puede utilizar en el caso de tener mas de dos poblaciones. Permite elegir entre diferentes estadisticos de centralidad. Lo cual tiene relevancia a la hora de realizar el contraste de homocedasticidad segun se tenga distribuciones normales o no.

-Test de Bartlett: Es muy sensible si no existe normalidad. Permite realizar el contraste para muestras de diferente tamaño.

-Test de Brown-Forsyth: Se basa en el test de Levene pero unicamente se utiliza la mediana como medida de centralidad.

-Test de Fligner-Killeen: Es el idoneo cuando no se cumple la condición de normalidad en las poblaciones. Es

un test no paramétrico donde la comparativa de las varianzas se realizan basandonos en la mediana.

Al tener muestras de diferentes tamaño utilizaremos el test de Bartlett, aunque teniendo en cuenta los resultados anteriormente no seria el mas idoneo ya que este es muy sensible si no existe normalidad.

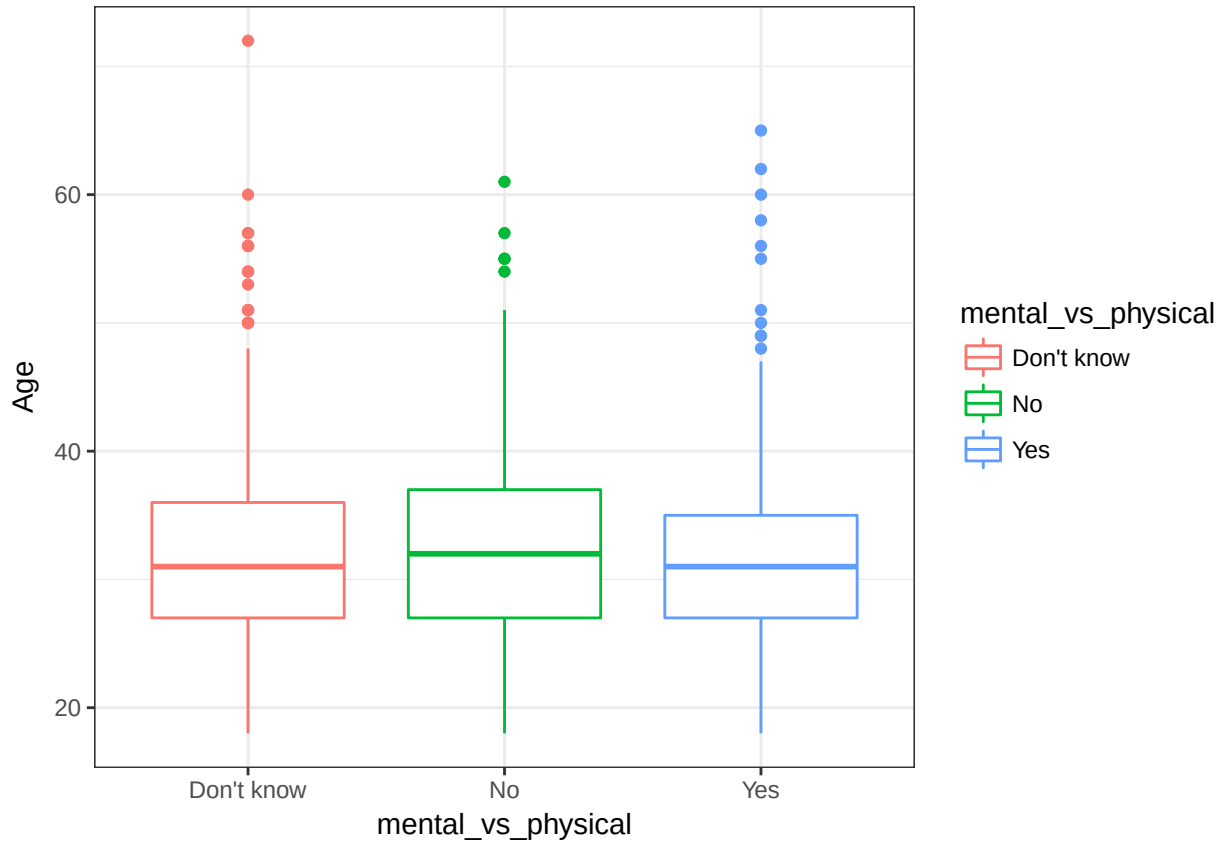
```
Age_Mental<-subset(surveyMentalHealth_clean$Age, surveyMentalHealth_clean$mental_vs_physical=="Yes")
Age_Fisica<-subset(surveyMentalHealth_clean$Age, surveyMentalHealth_clean$mental_vs_physical=="No")
Age_Desconoce<-subset(surveyMentalHealth_clean$Age, surveyMentalHealth_clean$mental_vs_physical=="Don't t
```

```
bartlett.test(list(Age_Mental,Age_Fisica,Age_Desconoce))
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  list(Age_Mental, Age_Fisica, Age_Desconoce)
## Bartlett's K-squared = 0.00013167, df = 2, p-value = 0.9999
```

Podemos concluir que el test no haya diferencias significativas entre las varianzas de los tres grupos.

```
ggplot(surveyMentalHealth_clean, aes(x = mental_vs_physical, y = Age, colour =mental_vs_physical)) + ge
```



Vamos a continuar observando si existe diferencias significativa según la edad del individuo de haber recibido o no tratamiento.

Omitimos el analisis de normalidad para la varible Age ya que se ha realizado anteriormente. A continuación estudiamos la homogeneidad de la varianza u homocedasticidad.

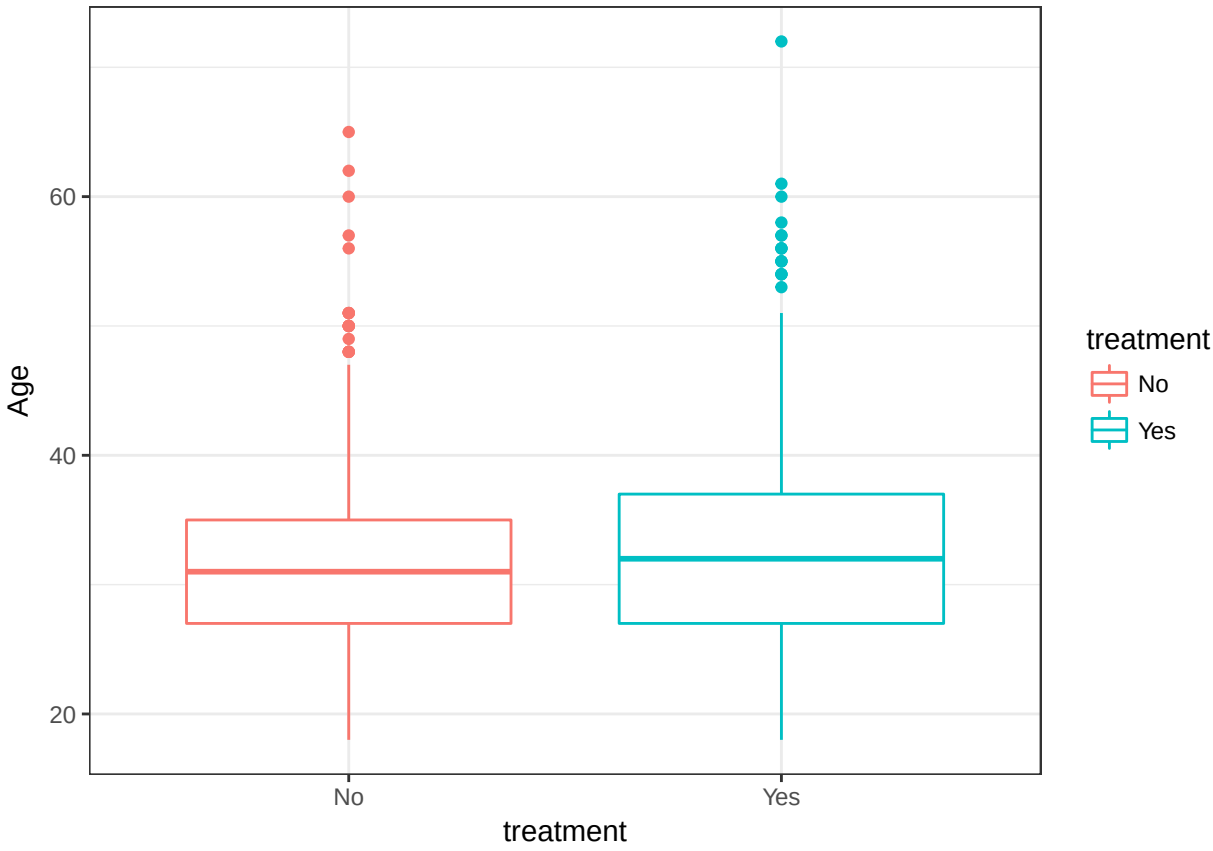
```
Age_Tratamiento<-subset(surveyMentalHealth_clean$Age, surveyMentalHealth_clean$treatment=="Yes")
Age_NTratamiento<-subset(surveyMentalHealth_clean$Age, surveyMentalHealth_clean$treatment=="No")
```

```r
bartlett.test(list(Age_Tratamiento,Age_NTratamiento))
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  list(Age_Tratamiento, Age_NTratamiento)
## Bartlett's K-squared = 4.2709, df = 1, p-value = 0.03877
```

Podemos concluir que si hay diferencias significativas entre las varianzas de los dos grupos.

```r
ggplot(surveyMentalHealth_clean, aes(x = treatment, y = Age, colour =treatment)) + geom_boxplot() + ther
```



El test ANOVA requiere que los datos de la muestra cumplan dos asunciones básicas: normalidad e igualdad de varianzas (homocedasticidad). Si las asunciones de ANOVA no se terminan cumpliendo, se aplica el equivalente no paramétrico de ANOVA, la prueba de Kruskal-Wallis. En el caso de la edad de los individuos teniendo en cuenta si han recibido tratamiento o no relacionado con la salud mental, no se cumple la asunción de homocedasticidad por ello aplicamos el test de Kruskal-Wallis.

```r
kruskal.test(Age~ treatment, data=surveyMentalHealth_clean)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Age by treatment
## Kruskal-Wallis chi-squared = 6.9616, df = 1, p-value = 0.008328
```

Como el p-value es menor de 0.05 podemos concluir que hay diferencias significativas entre el hecho de haber tenido o no tratamiento relacionado con la salud mental.

En el caso de la edad teniendo en cuenta si consideran que la organización da mayor importancia a la salud mental que a la física. Se cumple la asunción de homocedasticidad por este motivo continuamos con el estudio ANOVA de un factor (one-way ANOVA o independent samples ANOVA).

```
fit2=lm(Age~ mental_vs_physical, surveyMentalHealth_clean)
aov(fit2)
```

```
## Call:
##    aov(formula = fit2)
##
## Terms:
##                 mental_vs_physical Residuals
## Sum of Squares               61.69  65790.05
## Deg. of Freedom                  2      1237
##
## Residual standard error: 7.292816
## Estimated effects may be unbalanced
```

```
summary(aov(fit2))
```

```
##                      Df Sum Sq Mean Sq F value Pr(>F)
## mental_vs_physical    2     62   30.85    0.58   0.56
## Residuals          1237  65790   53.19
```

En este caso no hemos encontrado ningún cambio significativa de la variable mental_vs_physical ya que el p_{value} ha sido mayor que 0.05

Se guarda todos los cambios en un fichero.

```
#Se guardan los cambios realizados
write.csv(surveyMentalHealth_clean, file="survey_clean.csv")
```