Práctica 2

Ejercicio 1

Descripción del dataset. ¿Por que es importante y que pregunta/problema pretende responder?

Este dataset está formado 27 variables y 1260 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 dónde 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 primario 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 persona 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.

Los datos corresponden a la encuesta realizada durante el 2014. La licencia que tiene toda esta información es Creative Commons Attribution-ShareAlike 3.0 Unported License.

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 recogida 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. Aquí nos encontramos con el estudio de establecer la forma visual más eficiente para la representación 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.

Ejercicio 2

Limpieza de los datos.

2.1 Selección de los datos de interés a analizar. ¿Cuáles son los campos más relevantes para responder al problema?

Dado que el dataset, contiene las siguientes variables y preguntas asociadas a su explicación:

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 o 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 una aspecto de salud mental?

22. supervisor: ¿Estaría dispuesto a hablar con sus supervisores de una 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

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

6. comments

Así pues, tenemos 1259 observaciones y 27 variables de las que nos quedamos con 24 variables que, a priori, podrían ser útiles para nosotros.

Los campos más importantes para resolver el problema serían

1.Age
2. treatment
3. mental_vs_physical

En este dataset nos encontramos con un conjunto de variables que son cuantitativas y cualitativas.

Las cualitativas son las que tienen su origen en características 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 categoría 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 más liviana donde nos encontramos con muy pocas observaciones.

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

2.2 ¿Los datos contienen ceros o elementos vacíos? ¿Y valores extremos? ¿Cómo gestionarás cada uno de estos casos?

Tenemos 1259 observaciones. Vemos que, en primer lugar, hay niveles y valores inadecuados en algunas de las variables. Es necesario estandarizarlos.

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 carácter de vacío.

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

Comprobamos que variables tienen datos perdidos. Las únicas variables que contienen valores vacíos son self_employed con un total de 18 (Valor TRUE) y work_Interfere con 264 (Valor TRUE). Como estas variables no son de interés para nuestro estudio hemos decido no eliminar las observaciones con valor “NA”. En caso de que hubiéramos deseado tener estas variables como parte importante del estudio hubiéramos procedido a la eliminación de los registros u observaciones donde los valores de estas variables fueran “NA”, esta decisión principalmente está fundamentada en el hecho de que son variables cualitativas.

Las variables no_employees, family_history, remote_work, tech_company, benefits, care_options, wellness_program, y treatment no tiene NA.

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 outlets con un propósito.
3.Valores atípicos u outlets cuyo origen son errores del muestreo.
4.Valores atípicos u outlets de errores en la estandarización.
5.Valores atípicos u outlets por asumir distribuciones erróneas.
6.Valor atípico u outlets 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.

Las únicas variables que posee datos atípicos son Age y Gender.

En general, hemos acordado eliminar las observaciones cuyas respuestas no sean adecuadas dentro de los límites de lo aceptable para las preguntas formuladas, dado que, si para Age o Gender el encuestado está optando una actitud que no refleja un comportamiento serio para estas variables, entendemos que tampoco es fiable la respuesta que dé en el resto de la encuesta.

Así pues, eliminaremos las observaciones que contienen respuestas inadecuadas o fuera de rango aceptable.

Para la variable Gender debería haber 2 niveles (Male, Female), cuando hay 49. Así pues, para el variable género, cambiamos sus valores correspondientes por M y F respectivamente.

En aquellas observaciones que no es posible determinar o que la respuesta no es adecuada procedemos a eliminar la observación en lugar de asignar NA (“A little about you”, “Agender”, “All”, “Enby”, “fluid”,”Genderqueer”, “Nah”, “Neuter”, “non-binary”,”p”,”queer”,”queer/she/they”,”Trnas woman”,”Trans-female”). Nos quedamos por tanto con 1245 observaciones.

Exploramos a continuación los valores de la variable Age, variable cuantitativa. Dado que hemos visto que hay respuestas inadecuadas, como edades con signos negativos o valores que no pueden corresponderse con edad del encuestado, entendemos que son inadecuados y es preciso corregirlos (-29000, 329000, 1.000e+11, -1.726e+03, 5.000e+00, 8.000e+00, 1.100e+01, -1.000e+00)

Por tanto, del total de observaciones nos quedamos con 1240.

Ejercicio 3

Análisis de los datos.

3.1 Selección de los grupos de datos que se quieren analizar/comparar.

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

3.2 Comprobación de la normalidad y homogeneidad de la varianza. Si es necesario (y posible), aplicar transformaciones que normalicen los datos.

El análisis de la normalidad o contrastes de normalidad, investigan cuanto de lejos está la distribución de los valores observados con respecto a una distribución normal con la misma media y desviación típica. Este estudio será realizado para la variable Age.

Los estudios gráficos que utilizamos son Box Plot, histogramas, función de densidad, Normal Q-Q Plot, grafico para comprobar la homogeneidad de la varianza.

Realizamos el estudio de la normalidad mediante los contrastes 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 a lejanía que existe entre los coeficientes de asimetria y curtosis de los esperados por una distribución normal.

Todos estos test tenemos como hipótesis nula que los datos proceden de una distribución normal y la hipótesis alternativa que no lo hacen. El p-valor 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-valor, hay que tener en cuenta que a mayor tamaño de la muestra más finos son los test y es más sencillo encontrar evidencias en contra de H0. 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.

Si no podemos asumir normalidad este hecho nos influirá en los test de hipótesis paramétricos y en los modelos de regresión luego los estimadores calculados por mínimos cuadrados no serán eficientes y tanto los intervalos de confianza de los parámetros del modelo como contrastes significativos serán únicamente aproximados y nos exactos.

Si tenemos en cuenta el teorema del límite central el cual necesita que las poblaciones de las que procede la muestra sea una normal, no las muestras. Si la muestra se distribuye según una normal está claro que la población también lo hará. Puede ocurrir que la muestra no se distribuye según una norma, pero si conocemos que la población se distribuye según una normal, entonces los contrastes paramétricos si son válidos. El Teorema del Limite Central permite simplificar los requisitos de normalidad cuando las muestras son grandes.

A continuación, estudiamos la homogeneidad de la varianza u homocedasticidad, se está 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 hipótesis 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 siempre 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 más de dos poblaciones. Permite elegir entre diferentes estadísticos de centralidad. Lo cual tiene relevancia a la hora de realizar el contraste de homocedasticidad según 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 únicamente se utiliza la mediana como medida de centralidad.

-Test de Fligner-Killeen: Es el idóneo cuando no se cumple la condición de normalidad en las poblaciones. Es un test no paramétrico donde la comparativa de las varianzas se realizan basándonos en la mediana.

Al tener muestras de diferentes tamaños utilizaremos el test de Bartlett, aunque teniendo en cuenta los resultados anteriormente no sería el más idóneo ya que este es muy sensible si no existe normalidad.

Comencemos con el análisis de la normalidad para el la variable Age.

Observamos primero los datos de una manera gráfica, con un Boxplot, en este análisis comprobamos que existe cierta asimetría a la derecha, ya que las colas no son de igual longitud. Este 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 donde los individuos consideran que la organización da mayor importancia a la salud mental que a la salud física va desde 18 hasta 27. El último grupo (desde Q3 hasta el máximo valor) desde 36 hasta 72. El 50% de los individuos observados tienen Age entre Q1 y Q3.

Al comprobar para este grupo tanto la función de densidad como su histograma confirmamos que se produce cierta asimetría a la derecha.

Vamos a utilizar también 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, más alineados están los puntos a una recta. Al observar este gráfico vemos como los puntos situados en la parte más alta (derecha) se alejan de la recta por lo tanto ya vamos con disposición de que no existe normalidad.

Si realizamos los test de normalidad de Kolmogorov-Smirnov, Lilliefors y Jarque-Bera procedemos a rechazar la hipótesis nula de normalidad ya que en todos ellos obtenemos un pvalor<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 del Limite Central, garantizamos robustez del análisis.

Realizamos la transformación $\sqrt{(\sqrt{1/Age})}$ para eliminar la asimetría de la derecha y suavizar la gráfica e intentar que esta transformación de la variable acepte un contraste de normalidad. Realizando las mismas pruebas gráficas y de contrates

podemos afirmar que procedemos a rechazar la hipótesis nula de normalidad ya que en todos ellos obtenemos un pvalor<0.05.

Para el análisis de la homogeneidad de la varianza para el conjunto de la variable edad donde consideran que la organización da mayor importancia a la salud mental vs salud física y el conjunto de la variable edad donde consideran que la organización da menor importancia a la salud mental vs salud física. Realizamos el test de Bartlett ya que consideramos que es el más idóneo ya que las muestras tienen diferentes tamaños, aunque sea muy sensible si no existe normalidad. Procedemos a concluir que el test no haya diferencias significativas entre las varianzas de los dos grupos. Ya que el pvalor>0.05. Este resultado también lo confirmamos gráficamente mediante un plot.

Comenzamos con la homogeneidad de la varianza, para su estudio realizamos el test de Bartlett con su resultado podemos concluir que si hay diferencias entre las varianzas de los grupos debido a que el pvalor <0.05. Este hecho también lo podemos ver gráficamente mediante un plot.

3.3 Aplicación de pruebas estadísticas (tantas como sea posible) para comparar los grupos de datos.

El test ANOVA requiere que los datos de la muestra cumplan dos condiciones básicas: normalidad e igualdad de varianzas (homocedasticidad). Si las condiciones 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. Como el p-valor es menor de 0.05 podemos concluir que hay diferencias significativas en la edad 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). En este caso no hemos encontrado ningún cambio significativo de la variable edad según la variable mental_vs_physical ya que el p-valor ha sido mayor que 0.05.

No debemos olvidar el hecho de que en todo momento tenemos una variable dependiente cuantitativa Age y las variables independientes son cualitativas (mental_vs_physical, treatment)

Ejercicio 4

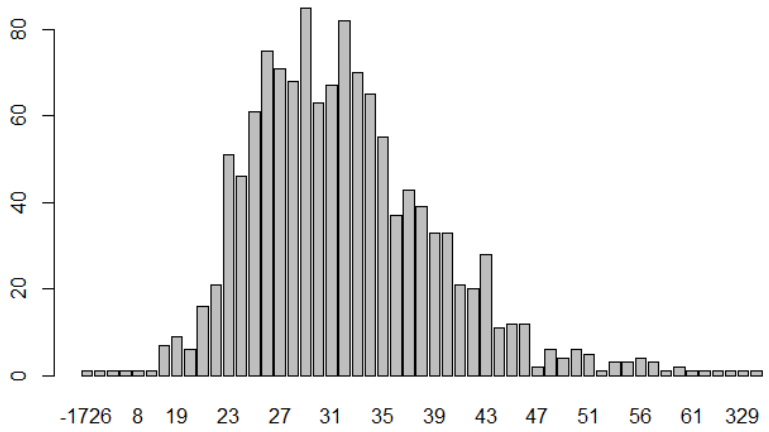
Representación de los resultados a partir de tablas y gráficas.

```
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"
```

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 estas 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 la observación de los datos atípicos de la variable Age comenzamos visualizándolos mediante la gráfica

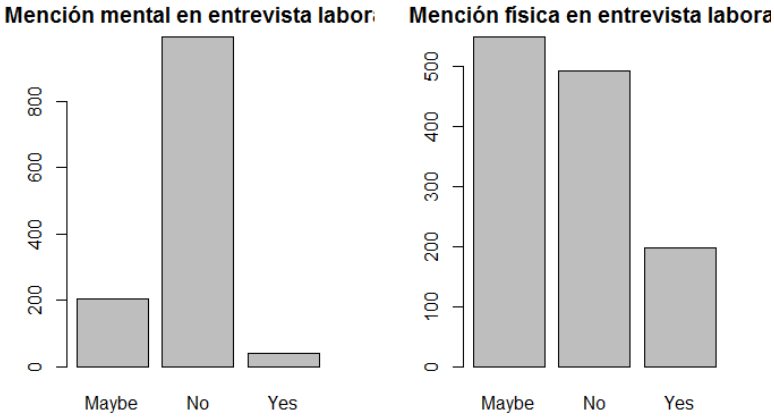


En la variable Gender obtenemos los siguientes niveles que nos incitan a un estudio más en detalle de los datos atípicos para esta variable

```
## [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"

Realizamos una comparativa entre la Mención mental en entrevista laboral y la Mención física en entrevista laboral.



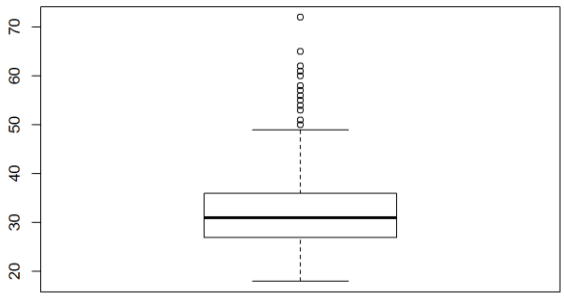
Para la variable cuantitativa Age estudiamos los valores del mínimo, Q1, Mediana, Media, Q3 y máximo.

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

Y de los siguientes indicadores

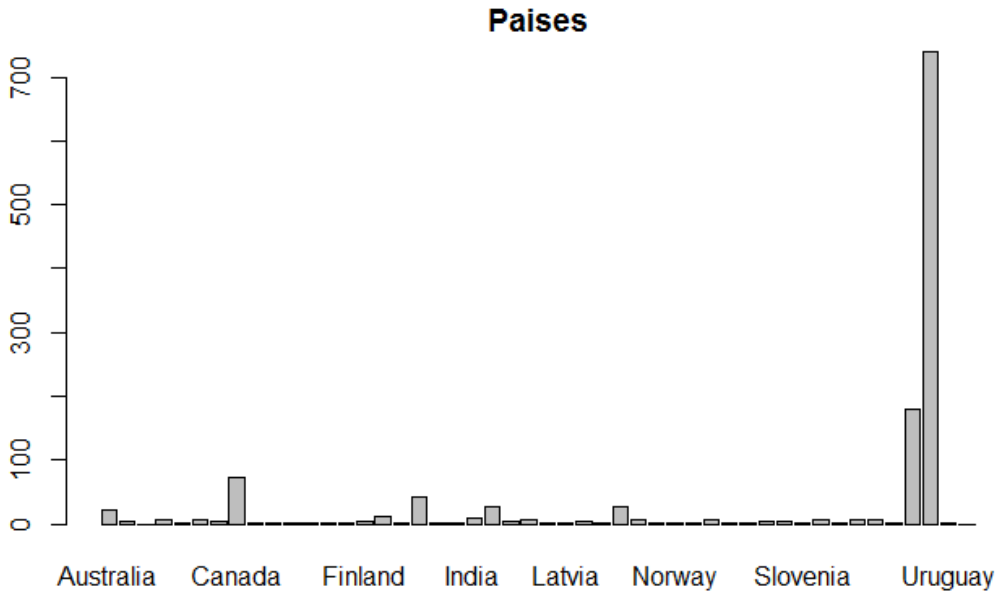
Media_Age32.076739
Mediana_Age31.000000
Media_Recortada_Age31.655723
Desviacion_estandar_Age7.288272
RIC_Age9.000000
Desviacion_Absoluta_Mediana_Age5.930400

De igual forma realizamos la representación del Box-Plot.

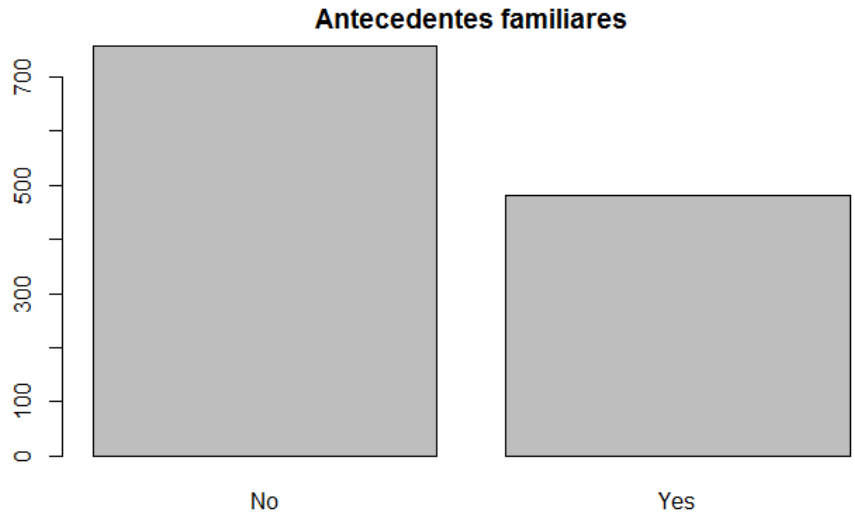


Para el resto que son variables cualitativas se realiza un análisis gráfico mediante gráficos de barras, en los cuales podemos observar los niveles de cada variable y el número de observaciones que pertenece a cada uno de ellos.

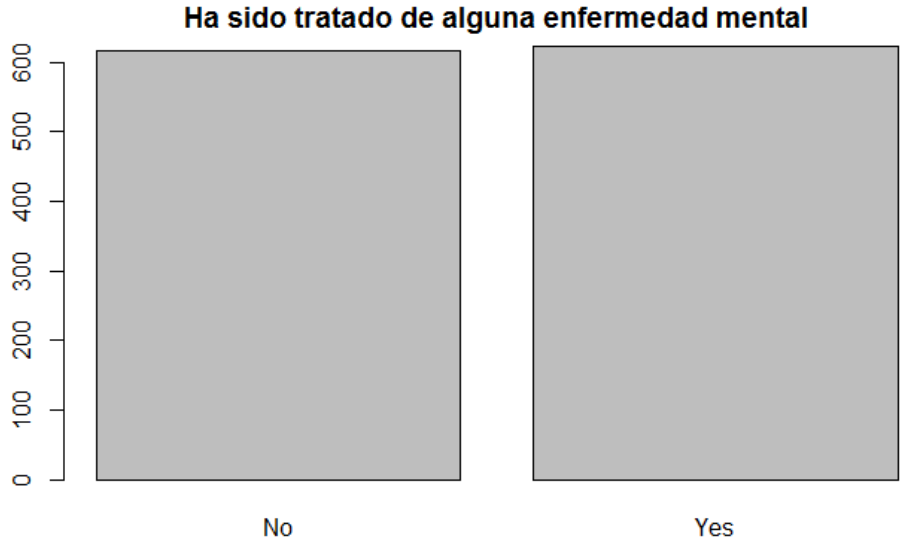
Country



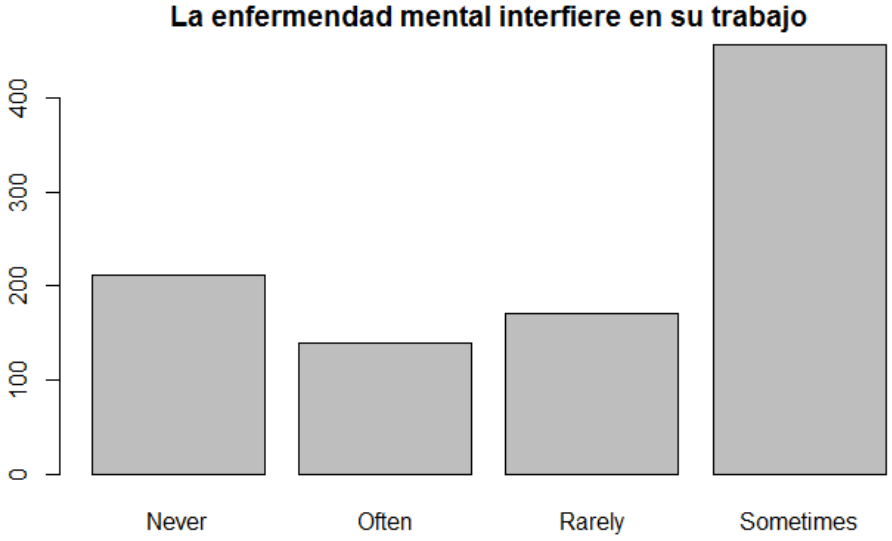
family_history



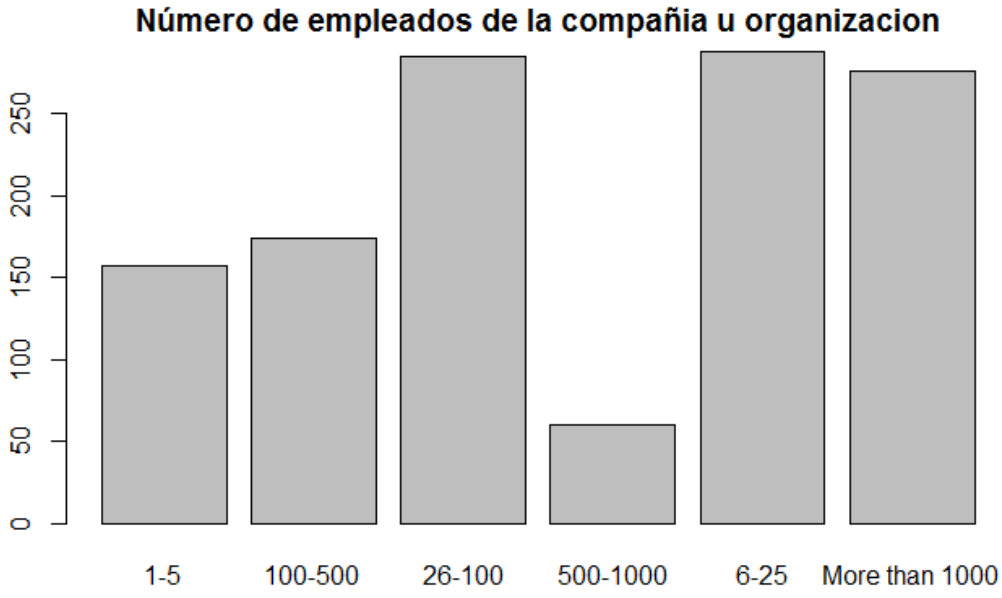
Treatment



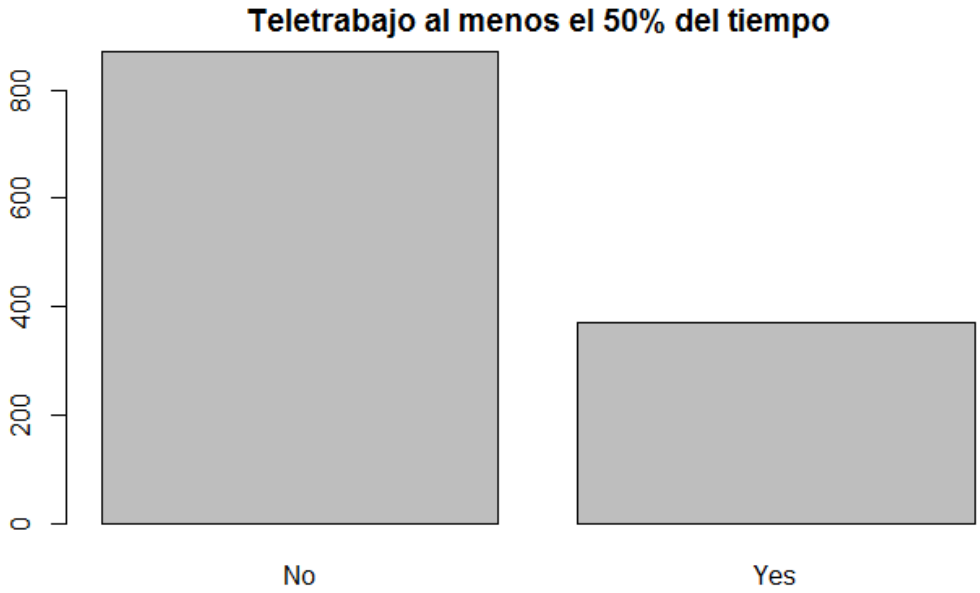
work_interfere



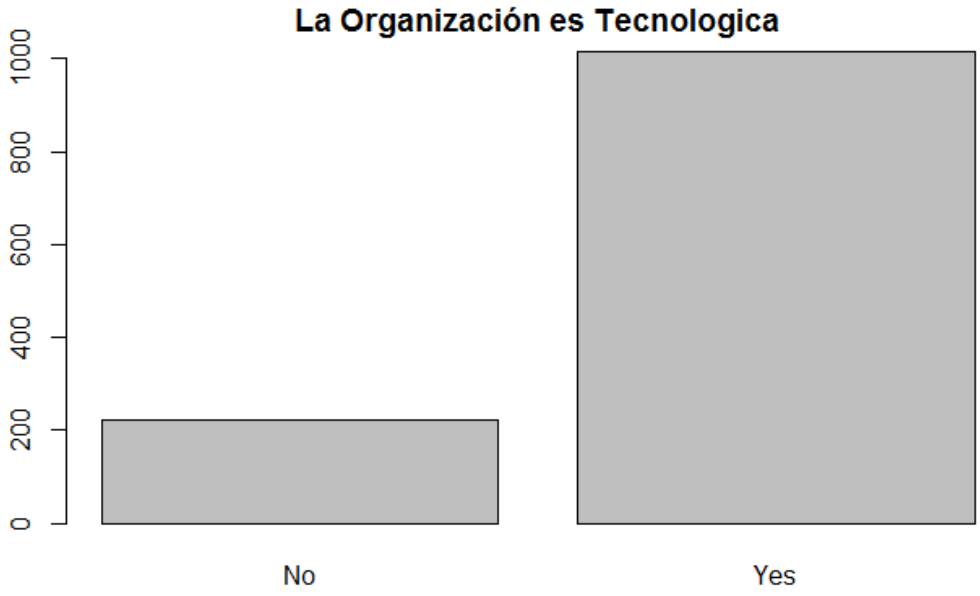
no_employees



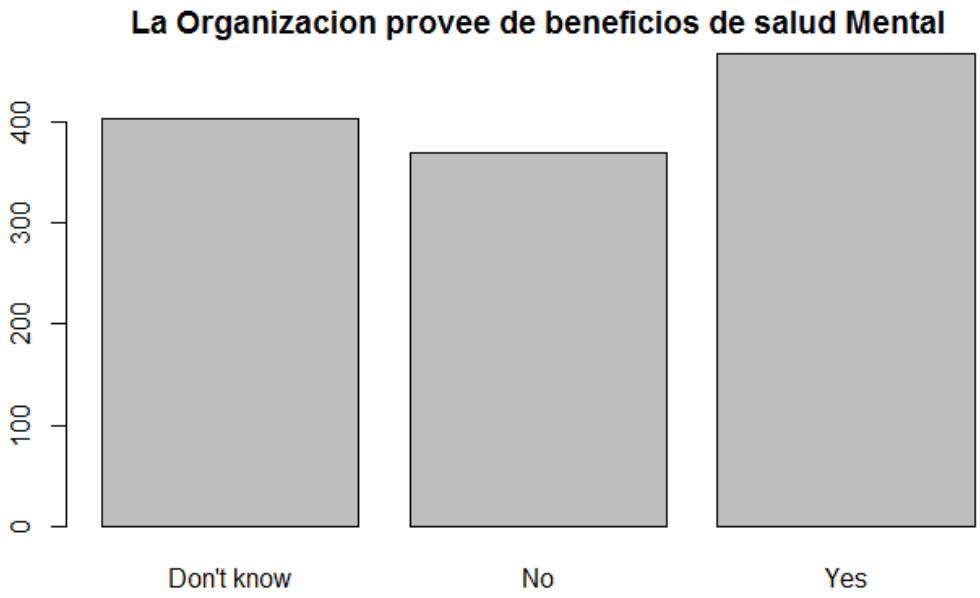
remote_work



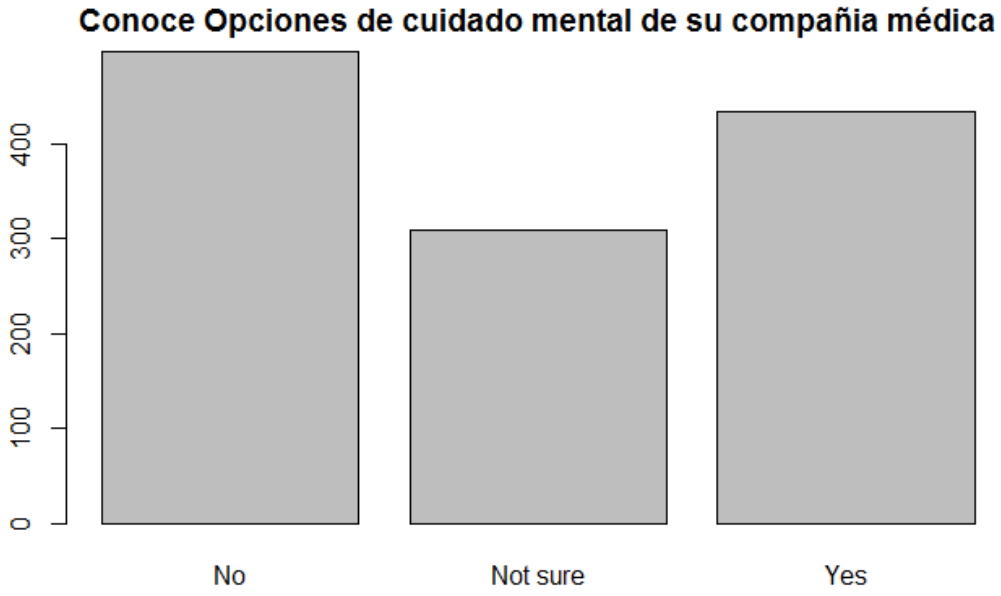
tech_company



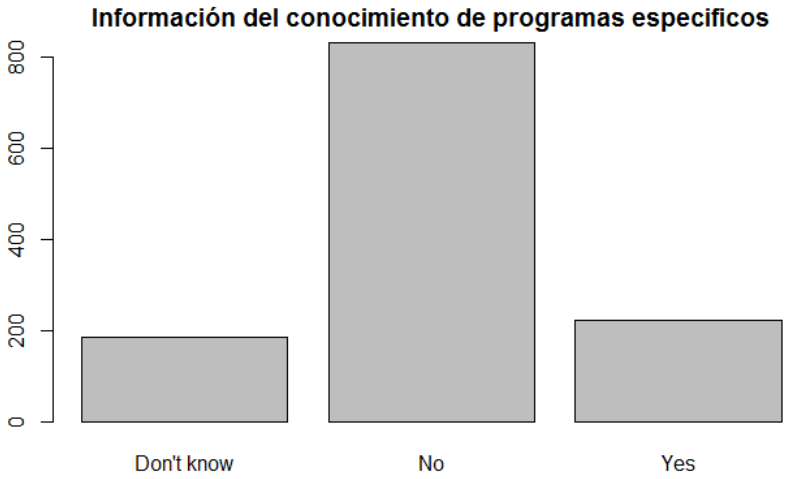
Benefits



care_options

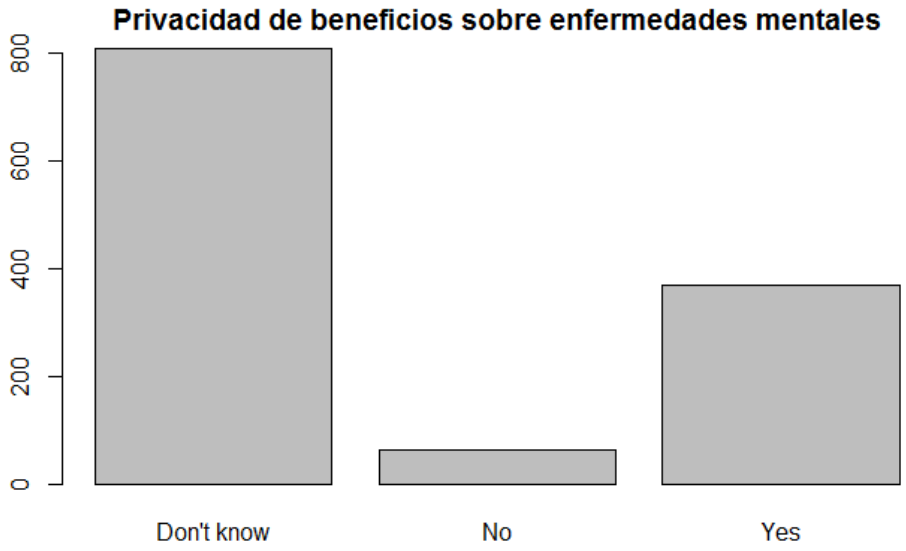


Wellness-program

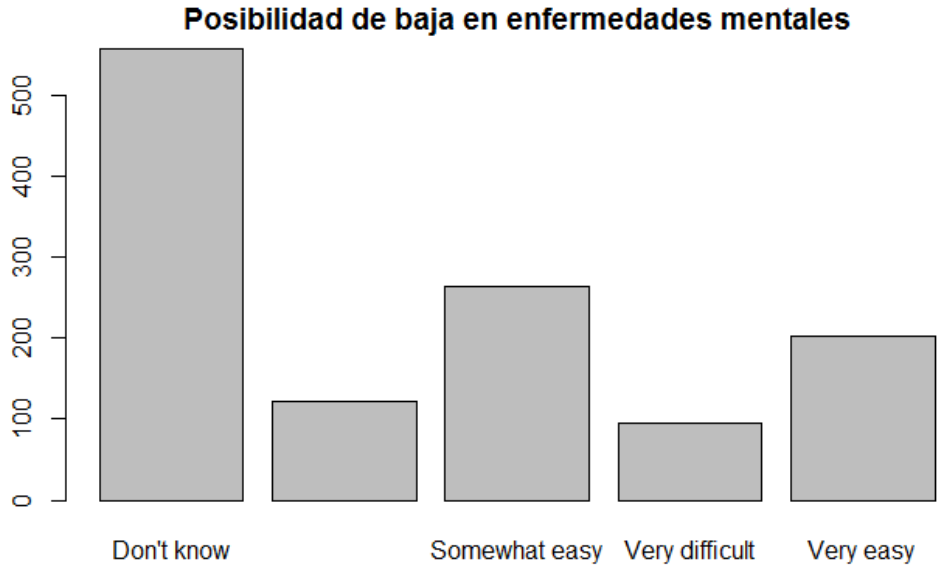


Seek_help

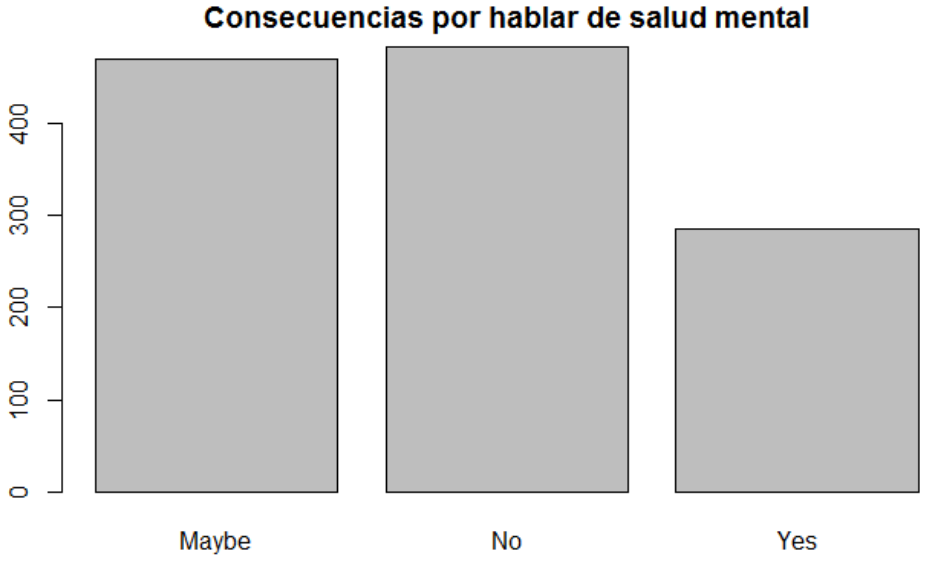
Anonymity



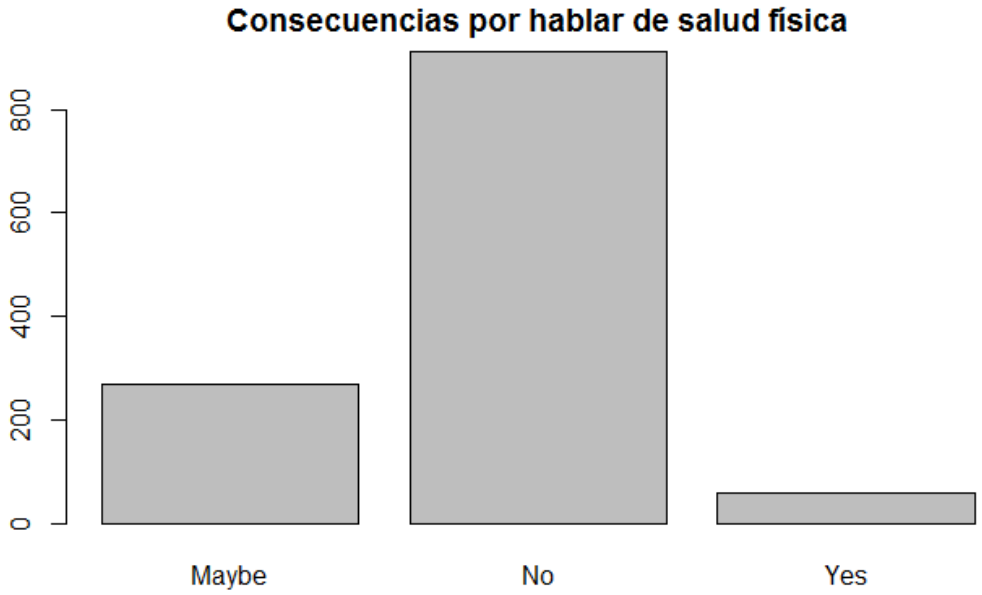
Leave



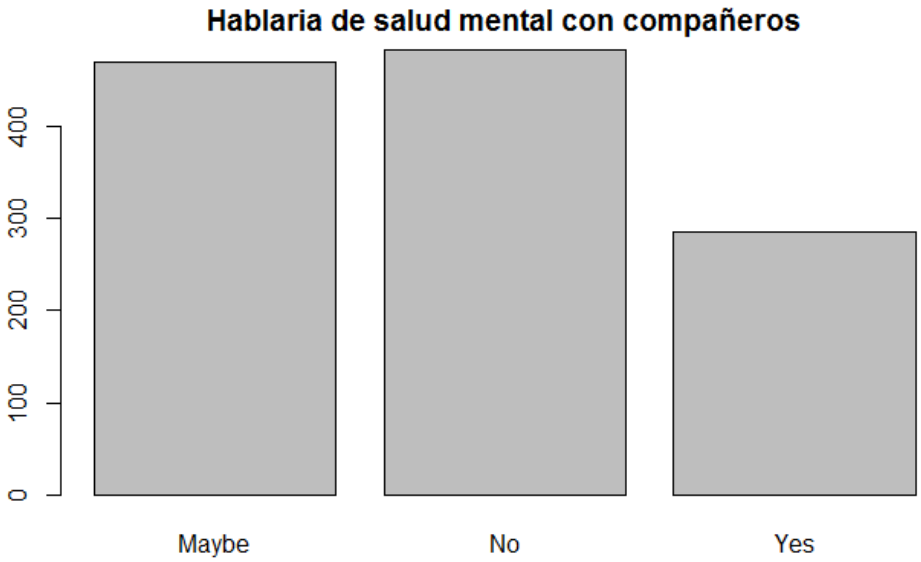
Mental_health_consequence



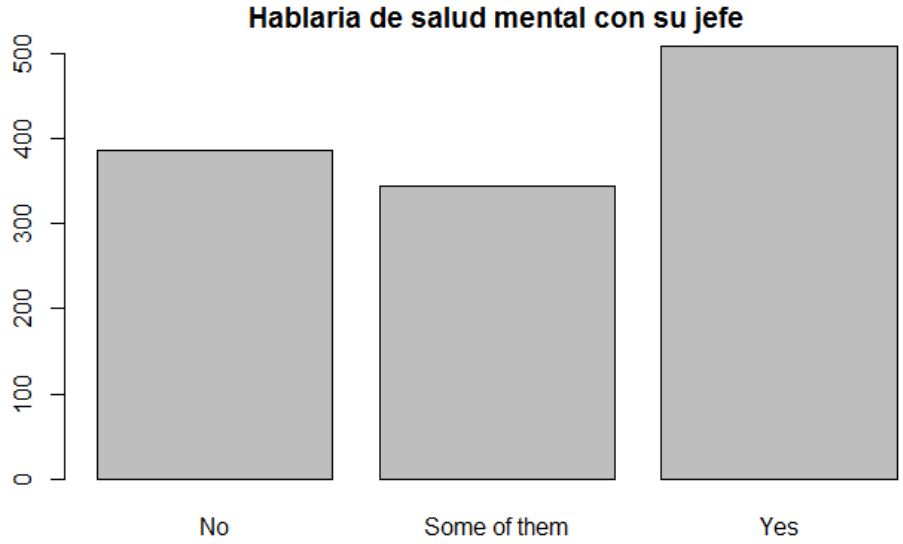
Phys_health_consequence



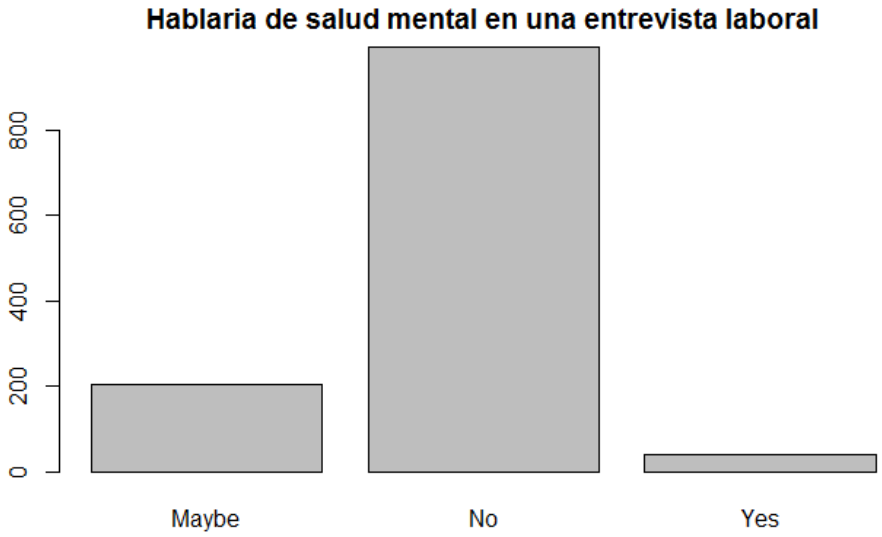
coworkers



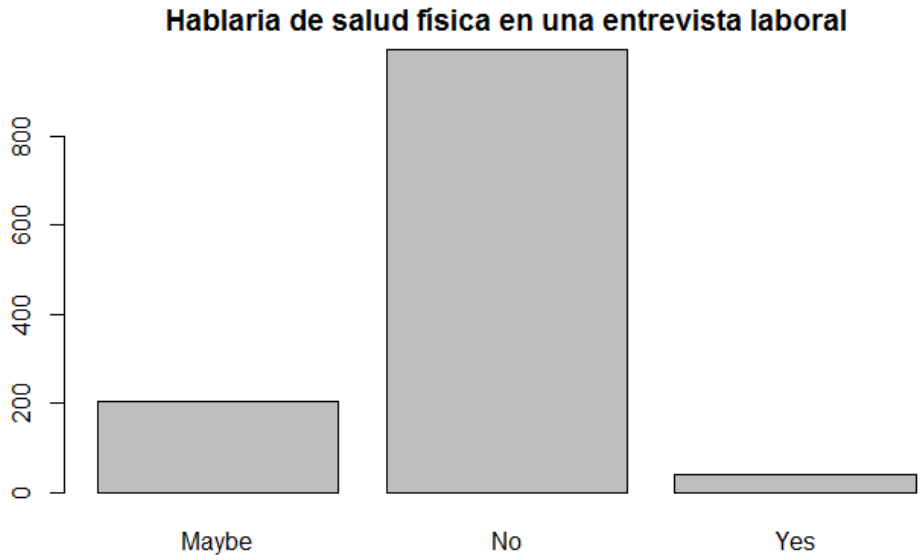
supervisor



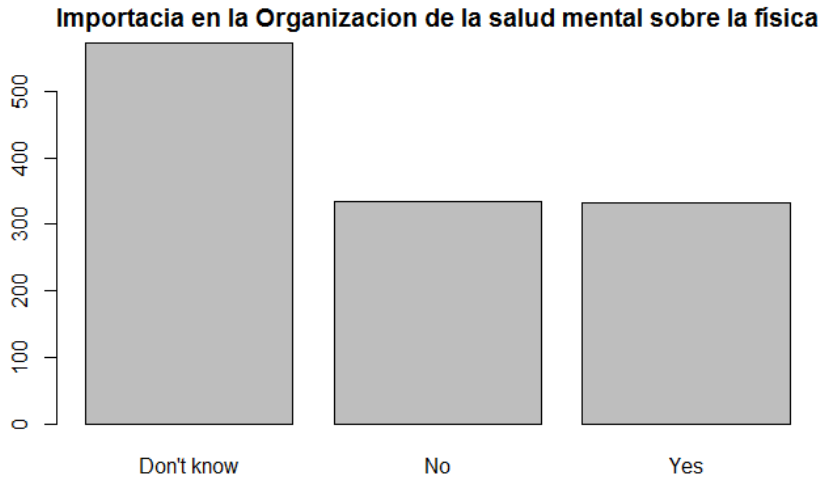
mental_health_interview



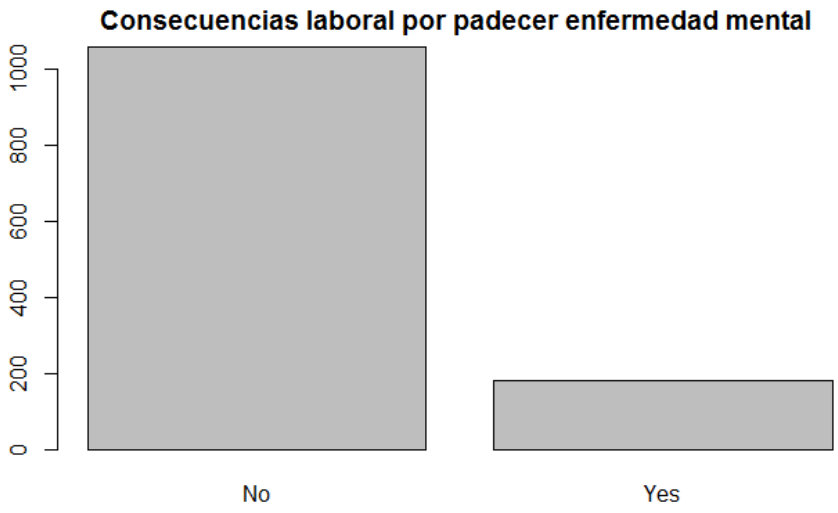
Phys_health_interview



mental_vs_physical



obs_consequence



En varias de las variables hay un porcentaje relevante de observaciones que indican incertidumbre de los propios encuestados, por lo que el análisis debería tomarse con cautela. Ello nos ayudaría a sacar conclusiones a priori sobre los observados, dado que muestra un grado de incertidumbre del sujeto, pero no permite extraer otras conclusiones sobre el entorno que pretendemos analizar.

Teníamos previsto realizar análisis de regresión logística para buscar la probabilidad e influencia entre las diferentes variables. Sin embargo, a la vista del análisis de los datos previo visto gráficamente, entendemos que en los mismo hay importantes carencias que, en aspectos que van desde la adecuada representatividad de la muestra hasta el control de respuestas que se ha observado, podrían implicar la invalidez de cualquier análisis.

Entendemos que es especialmente positivo para esta práctica, en este caso, la comprobación de la importancia del reprocesado y preparación de los datos y comprobación y validación de las respuestas en la muestra en relación con aspectos como la validez de las observaciones, niveles, o representatividad, así como sus valores. Ello nos ha permitido que -antes de entrar a hacer análisis estadísticos más avanzados para extraer conclusiones-, (ya que pueden por si mismos asegurar los posteriores pasos a tratar en el análisis, (como podría ser otras pruebas más avanzadas, como regresión logistica, por ejemplo) podamos comprender la importancia del análisis de los datos en la muestra para garantizar su validez desde la recogida del dato. Lejos de ser un aspecto accesorio, ha resultado ser crítico.

Desafortunadamente esto por otro lado, impide entrar a realizar análisis más profundos para responder a la pregunta que habíamos planteado.

Teniendo en cuenta que los grupos de datos que vamos estudiar corresponde a:

1.Dependiendo de la edad del individuo como percibe el hecho de que la organización de igual importancia a la salud mental vs salud física.

2.El hecho de recibir tratamiento tiene algo que ver con la edad

Comenzamos con el grupo 1 para ello realizamos los siguientes estudios

Un análisis de la normalidad de la variable cuantitativa Age

Calculo de la media

```
mean(surveyMentalHealth_clean$Age)
32.11452
```

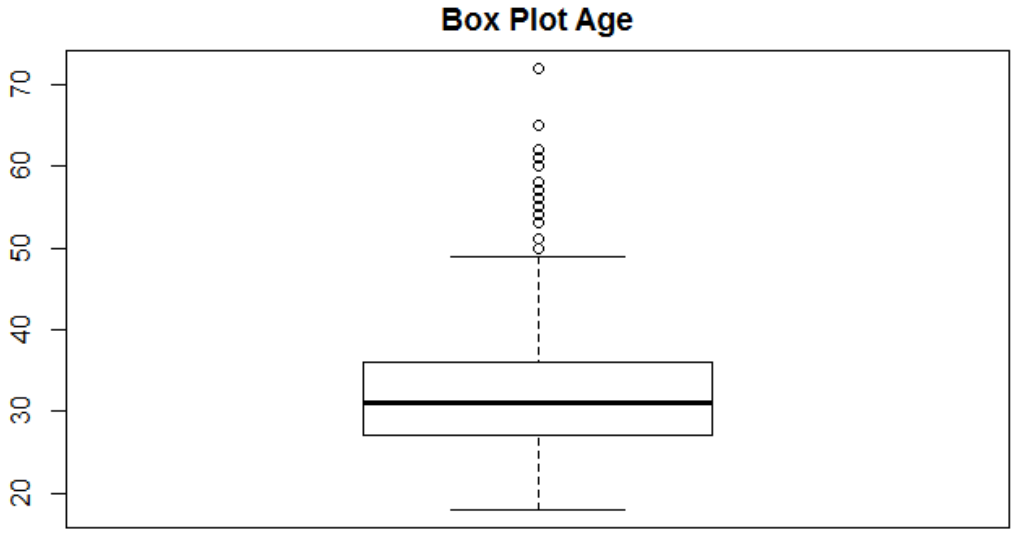
Cálculo de la mediana

```
median(surveyMentalHealth_clean$Age)
31
```

Cálculo de los cinco números de Tukey para Age

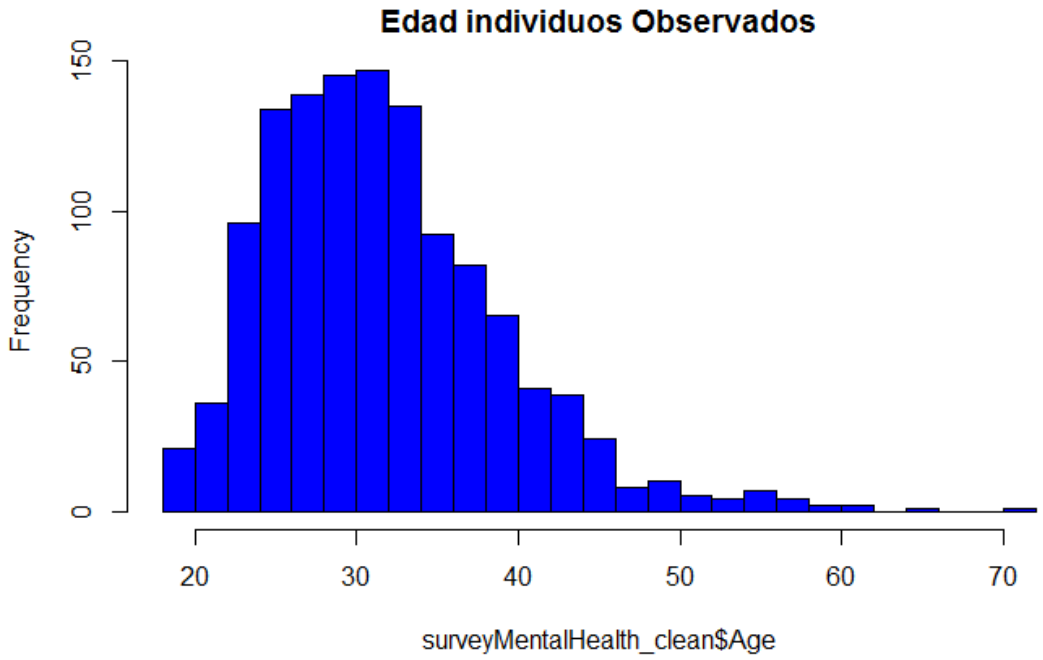
Mínimo	18
Q1	27
Mediana	31
Q3	36
Máximo	72

Gráfico boxplot para Age

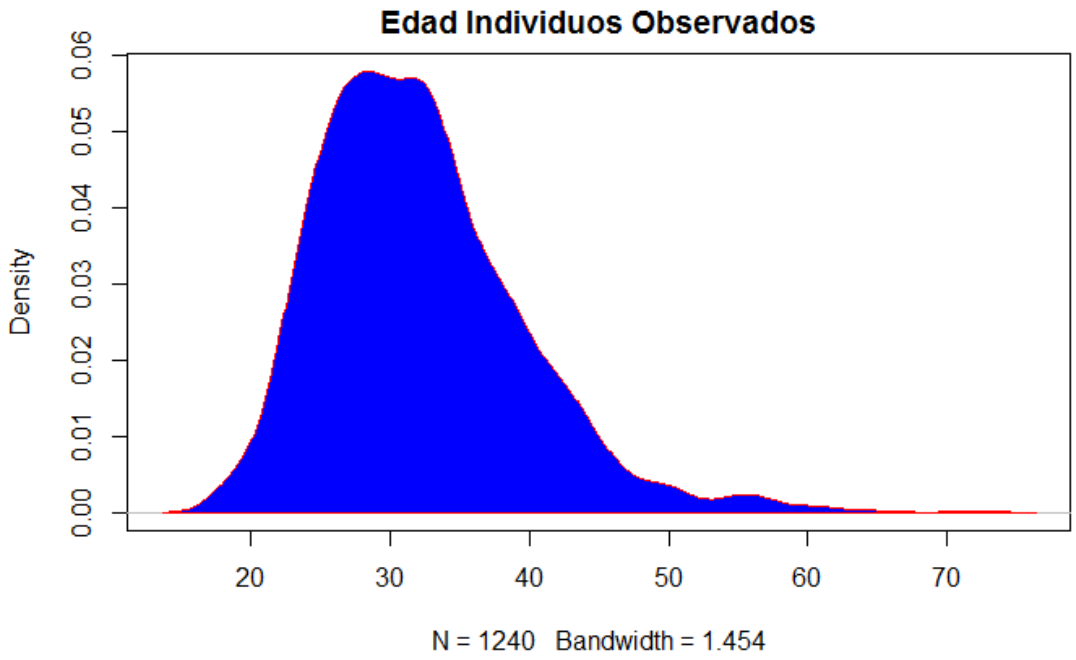


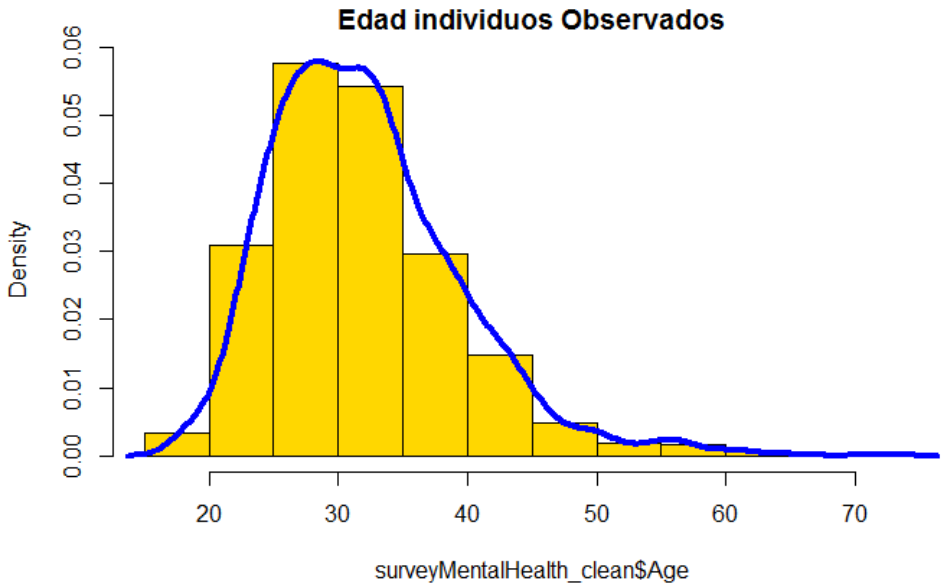
En un gráfico de Boxplot podemos estudiar la simetría, detectar outlier e incluso contrastar algunas hipótesis de la distribución. El gráfico fracciona los datos en 4 partes de igual frecuencia, es decir, cada grupo contiene mas o menos el mismo número de observaciones. Pero la ocupación de estos es 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).
El número de intervalos correspondientes Age es 35.

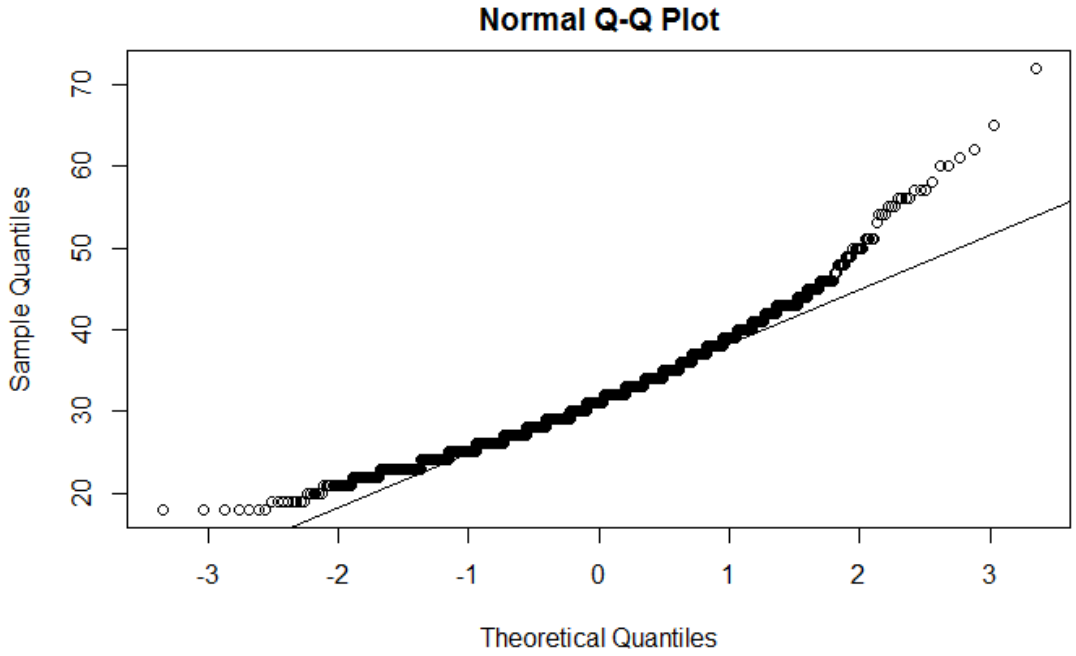


Calculamos la función de densidad de para finalmente superponerla sobre el histograma anteriormente representado





Vamos a utilizar también el gráfico de los cuantiles teóricos (Gráficos 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, más alineados están los puntos a una recta.



Para el estudio de la normalidad utilizamos el test de Kolmogorov-Smirnov

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. Así tenemos el test de Lilliefors, en este caso se acepta que la media y la varianza son desconocidas.

Age

```
Lilliefors (Kolmogorov-Smirnov) normality test

data:  (x = surveyMentalHealth_clean$Age)
D = 0.087147, p-value < 2.2e-16
```

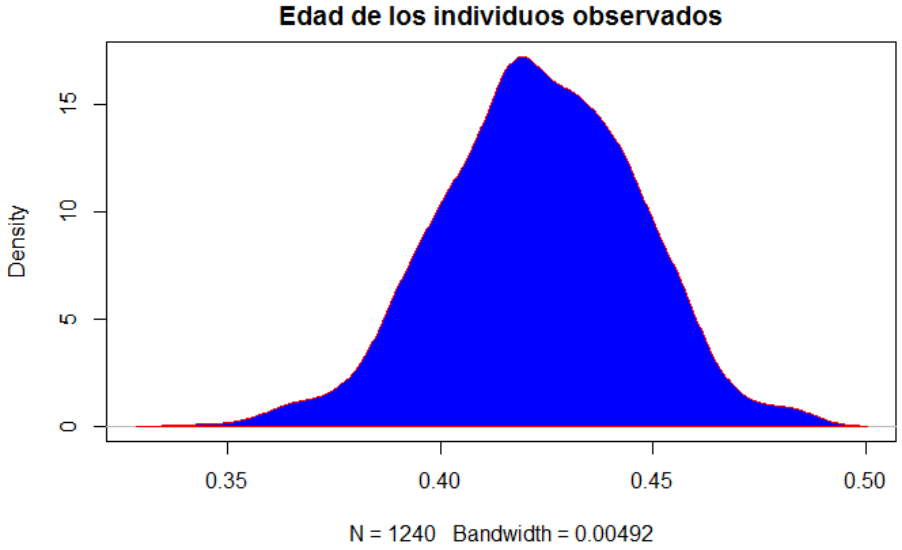
Podemos tener en cuenta también el test de normalidad de Jarque-Bera, este no pide estimación de los parámetros con los que podemos caracterizar una normal. Este lo que hace es saber lo que se alejan los coeficientes de asimetría y curtosis de una distribución normal.

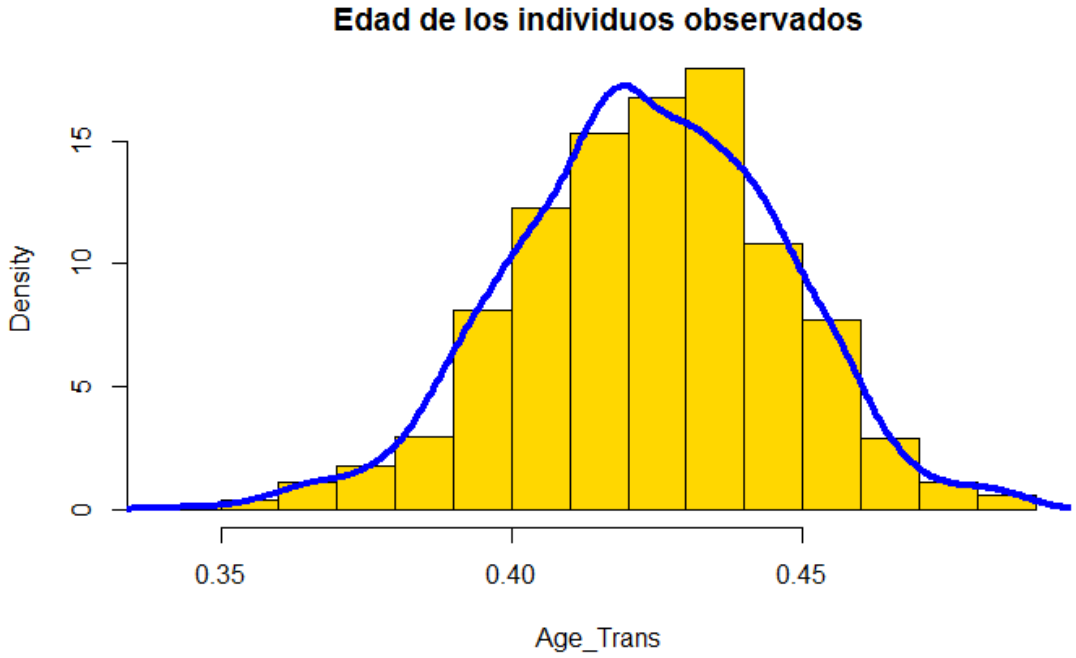
Age

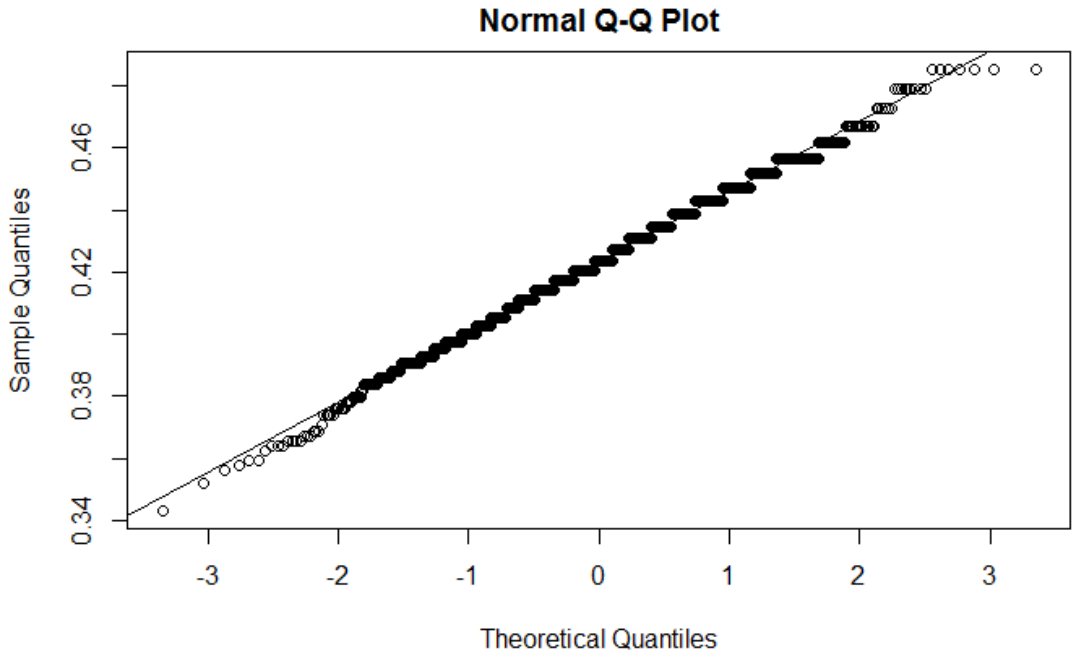
```
Jarque-Bera test for normality

data:  surveyMentalHealth_clean$Age
JB = 392.78, p-value < 2.2e-16
```

Procedemos a rechazar la hipótesis nula de normalidad ya que en todos los test obtenemos un p-valor<0.05. Pero no debemos olvidar que el Teorema del Limite Central permite simplificar los requisitos de normalidad cuando las muestras son grandes, como es el caso.

Realizamos la transformación $\sqrt{(\sqrt{1/Age})}$ y con ella calculamos su correspondiente función de densidad y la superposición de esta sobre su histograma







Observamos los resultados del test de Lillie para dicha transformación

```
        Lilliefors (Kolmogorov-Smirnov) normality test

data:  (x = Age_Trans)
D = 0.043067, p-value = 1.264e-05
```

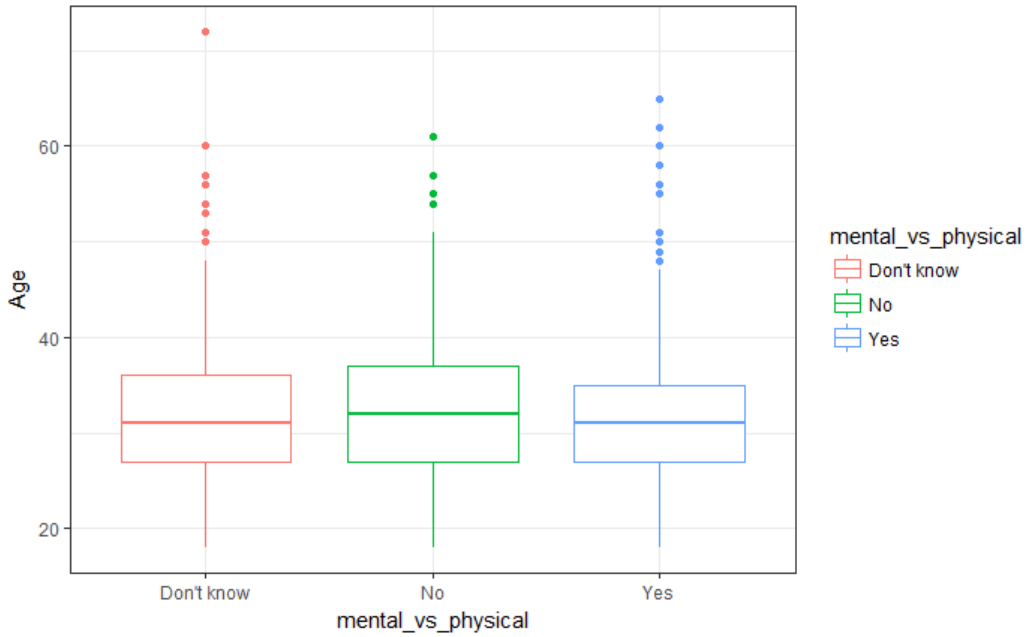
Donde continuamos rechazando la hipótesis nula de normalidad ya que p-valor<0.05
Por tanto, continuaremos con los datos orígenes sin tener en cuenta la transformada.

A continuación, estudiamos la homogeneidad de la varianza u homocedasticidad, se está 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 hipótesis nula que la varianza es la misma en todos los grupos y como hipótesis alternativa que no lo es.

Al tener muestras de diferentes tamaños utilizaremos el test de Bartlett, aunque teniendo en cuenta los resultados anteriormente no sería el más idóneo ya que este es muy sensible si no existe normalidad.

```
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 hay diferencias significativas entre las varianzas de los dos grupos, p-valor>0.05
Gráficamente también lo podemos visualizar



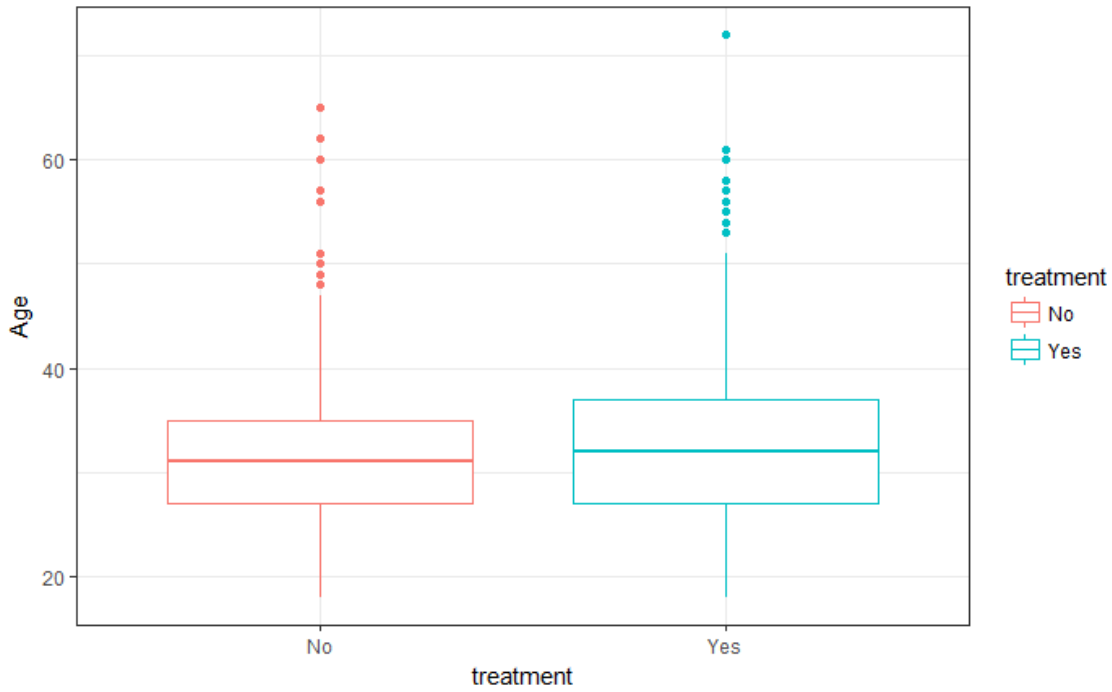
Vamos a continuar observando si existe diferencias significativas según la edad del individuo de haber recibido o no tratamiento.

Omitimos el análisis de normalidad para la variable Age ya que se ha realizado anteriormente.

A continuación, procedemos a estudiar la homogeneidad de la varianza u homocedasticidad.

```
Bartlett test of homogeneity of variances

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

Debido a que el pvalor<0.05 podemos concluir que si hay diferencias significativas entre las varianzas de los dos grupos.
Grafica también lo podemos observar



En el caso de estudio de la edad teniendo en cuenta si han recibido tratamiento o no relacionado con la salud mental aplicamos el equivalente no paramétrico de ANOVA, la prueba de Kruskal-Wallis, ya que no se cumple la condición de homocedasticidad.

```
        Kruskal-Wallis rank sum test

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

Como el p-valor 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).

```
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

                     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 significativo de la variable mental_vs_physical ya que el p-valor ha sido mayor que 0.05

Se guarda todos los cambios en un fichero.

Ejercicio 5

Resolución del problema. A partir de los resultados obtenidos. ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?

Una primera aproximación a la respuesta de nuestra pregunta, podría ser la que viene derivada de la comparación en un contexto de entrevista laboral, de trabajo entre la enfermedad física y mental, así como la comparativa de consecuencias negativas por hablar de salud mental con el empleador.

Si tenemos en cuenta que los encuestados consideran que las respuestas sobre consecuencias de hablar de salud mental vs física con el empleador son (quizá 477, No 490, Sí 292) versus (quizá 273, no 925 y si 61), para el ámbito mental y físico, respectivamente y que asimismo los encuestados en una entrevista de trabajo y en relación con la salud mental vs física son (quizá 207, No 1008, Sí 44) versus (Quizá 557, No 500, Si 202), para el ámbito mental y físico, respectivamente, tal y como se aprecia en los gráficos previos, podríamos inducir cierta condición de estigmatización de la enfermedad mental, versus enfermedad física.

También, podemos concluir que hay diferencia significativa en la edad según en algún momento haya recibido tratamiento relacionado con la salud mental.

De igual modo podemos concluir que no hay diferencia significativa en la edad según considera el individuo que la organización o empresa da mayor importancia a la salud mental frente a la salud física.

Por otra parte, el grado de incertidumbre que los datos presentan en muchas de las respuestas de los encuestados apuntarían a la falta de conocimiento del sujeto; es decir, no solo a aspectos externos derivados del entorno de trabajo o la organización, sino también a aspectos de voluntad o interés asociados al propio sujeto encuestado, lo cual podría ser un indicador más a tener en cuenta.

Sin embargo, es necesario ser sumamente cautelosos al extraer esta conclusión, ya que otros factores que no aparecen en la encuesta podrían ser determinantes, como el tipo de trabajo objetivo, el perfil del encuestado y el tipo de trabajo que desempeña, etc. ya que esto podría condicionar la respuesta. Si, además, tenemos en cuenta las observaciones realizadas a lo largo de la práctica sobre los datos, como la relativa al número de observaciones por país, o la relativa a la variable work_interfere, debemos concluir que los resultados no permiten responder al problema o pregunta planteada, por varios motivos, a los que hemos llegado a partir del análisis de los datos. Entendemos que existen errores importantes en el diseño del proceso para la obtención de los datos, debido a la falta de completitud del mismo. Derivado de ello, y a partir de los datos extraídos, el conjunto de datos presenta un problema de validez de la muestra y no permite responder a la pregunta.

Ejercicio 6

Código: Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos. Si lo preferís, también podéis trabajar en Python.

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

```{r load_libraries, include=FALSE}
library(knitr)

```

```{r}
library(rockchalk)
library(nortest)
library(normtest)
library(moments)
```

```
library(car)
library(ggplot2)
library(dplyr)
```

Cargamos los datos

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

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

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

```{r}
colnames(surveyMentalHealth)
```

```{r str}
str(surveyMentalHealth)
```

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

```{r}
sapply(surveyMentalHealth,class)
```

```{r}
#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))
```

```{r}
table(is.na(surveyMentalHealth$Gender))
```

```{r}
table(is.na(surveyMentalHealth$Country ))
```

```{r}
table(is.na(surveyMentalHealth$self_employed))
```

```{r}
table(is.na(surveyMentalHealth$family_history))
```

```{r}
table(is.na(surveyMentalHealth$treatment))
```

```{r}
table(is.na(surveyMentalHealth$work_interfere))
```

```

```{r}
table(is.na(surveyMentalHealth$no_employees))
```

```{r}
table(is.na(surveyMentalHealth$remote_work))
```

```{r}
table(is.na(surveyMentalHealth$tech_company))
```

```{r}
table(is.na(surveyMentalHealth$benefits))
```

```{r}
table(is.na(surveyMentalHealth$care_options))
```

```{r}
table(is.na(surveyMentalHealth$ wellness_program))
```

```{r}
table(is.na(surveyMentalHealth$seek_help))
```

```{r}
table(is.na(surveyMentalHealth$anonymity))
```

```{r}
table(is.na(surveyMentalHealth$leave))
```

```{r}
table(is.na(surveyMentalHealth$mental_health_consequence))
```
```

```

```{r}
table(is.na(surveyMentalHealth$phys_health_consequence))
```

```{r}
table(is.na(surveyMentalHealth$coworkers))
```

```{r}
table(is.na(surveyMentalHealth$supervisor))
```

```{r}
table(is.na(surveyMentalHealth$mental_health_interview))
```

```{r}
table(is.na(surveyMentalHealth$phys_health_interview))
```

```{r}
table(is.na(surveyMentalHealth$mental_vs_physical))
```

```{r}
table(is.na(surveyMentalHealth$obs_consequence))
```

```{r}
surveyMentalHealth$Age[which(surveyMentalHealth$Age==0) ]

```

```{r}
barplot(table(surveyMentalHealth$Age))
```
```

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

```{r}
surveyMentalHealth_clean<-subset(surveyMentalHealth, surveyMentalHealth$Age>16
& surveyMentalHealth$Age<75)
nrow(surveyMentalHealth_clean)
```

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

```{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))
```

```{r}
levels(surveyMentalHealth_clean$Gender)
```

```{r}
surveyMentalHealth_clean$Gender<-as.character(surveyMentalHealth_clean$Gender)
surveyMentalHealth_clean$Gender<-
replace(surveyMentalHealth_clean$Gender,surveyMentalHealth_clean$Gender=="Cis Female"|surveyMentalHealth_clean$Gender=="cis-female/femme"|surveyMentalHealth_clean$Gender=="f"|surveyMentalHealth_clean$Gender=="femail"|surveyMentalHealth_clean$Gender=="Femake"|surveyMentalHealth_clean$Gender=="female"|surveyMentalHealth_clean$Gender=="Female"|surveyMentalHealth_clean$Gender=="Female "|surveyMentalHealth_clean$Gender=="Female (cis)"|surveyMentalHealth_clean$Gender=="Female (trans)"|surveyMentalHealth_clean$Gender=="Woman"|surveyMentalHealth_clean$Gender=="woman","F")

surveyMentalHealth_clean$Gender<-
replace(surveyMentalHealth_clean$Gender,surveyMentalHealth_clean$Gender=="Male"|surveyMentalHealth_clean$Gender=="male"|surveyMentalHealth_clean$Gender=="Cis Male"|surveyMentalHealth_clean$Gender=="m"|surveyMentalHealth_clean$Gender=="Mail"|surveyMentalHealth_clean$Gender=="Make"|surveyMentalHealth_clean$Gender=="male leaning androgynous"|surveyMentalHealth_clean$Gender=="Malr"|surveyMentalHealth_clean$Gender=="msle"|surveyMentalHealth_clean$Gender=="ostensibly male, unsure what that really means"|surveyMentalHealth_clean$Gender=="something kinda male?"|surveyMentalHealth_clean$Gender=="Androgyne"|surveyMentalHealth_clean$Gender=="cis male"|surveyMentalHealth_clean$Gender=="Cis Man"|surveyMentalHealth_clean$Gender=="Guy (-ish) ^_^"|surveyMentalHealth_clean$Gender=="maile"|surveyMentalHealth_clean$Gender=="Mal"|surveyMentalHealth_clean$Gender=="Male (CIS)"|surveyMentalHealth_clean$Gender=="Male-ish"|surveyMentalHealth_clean$Gender=="Man"|surveyMentalHealth_clean$Gender=="Male ","M")
```

```
surveyMentalHealth_clean$Gender<-
replace(surveyMentalHealth_clean$Gender,surveyMentalHealth_clean$Gender=="A
little about
you"|surveyMentalHealth_clean$Gender=="Agender"|surveyMentalHealth_clean$Gend
er=="All"|surveyMentalHealth_clean$Gender=="Enby"|surveyMentalHealth_clean$Gen
der=="fluid"|surveyMentalHealth_clean$Gender=="Genderqueer"|surveyMentalHealth_
clean$Gender=="Nah"|surveyMentalHealth_clean$Gender=="Neuter"|surveyMentalHe
alth_clean$Gender=="non-
binary"|surveyMentalHealth_clean$Gender=="p"|surveyMentalHealth_clean$Gender==
"queer"|surveyMentalHealth_clean$Gender=="queer/she/they"|surveyMentalHealth_cle
an$Gender=="Trans woman"|surveyMentalHealth_clean$Gender=="Trans-
female",NA)

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

```{r}
levels(surveyMentalHealth_clean$Gender)
```

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

```{r}
levels(surveyMentalHealth_clean$Country)
summary(surveyMentalHealth_clean$Country)
```

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

```{r}
levels(surveyMentalHealth_clean$family_history)
```

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

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

```{r}
levels(surveyMentalHealth_clean$treatment)
summary(surveyMentalHealth_clean$treatment)
```

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

```{r}
levels(surveyMentalHealth_clean$work_interfere)
summary(surveyMentalHealth_clean$work_interfere)
```

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

```{r}
levels(surveyMentalHealth_clean$no_employees)
summary(surveyMentalHealth_clean$no_employees)
```

```{r}
```

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

```{r}
levels(surveyMentalHealth_clean$remote_work)
summary(surveyMentalHealth_clean$remote_work)

```

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

```{r}
levels(surveyMentalHealth_clean$tech_company)
summary(surveyMentalHealth_clean$tech_company)

```

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

```{r}
levels(surveyMentalHealth_clean$benefits)
summary(surveyMentalHealth_clean$benefits)

```

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

```{r}
levels(surveyMentalHealth_clean$care_options)
summary(surveyMentalHealth_clean$care_options)
```

```
```

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

```{r}
levels(surveyMentalHealth_clean$wellness_program)

```

```{r}
summary (surveyMentalHealth_clean$wellness_program)

```

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

```{r}
levels(surveyMentalHealth_clean$seek_help)
```

```{r}
summary (surveyMentalHealth_clean$seek_help)

```

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

```{r}
levels(surveyMentalHealth_clean$anonymity)
```

```{r}
summary (surveyMentalHealth_clean$anonymity)

```

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

```{r}
levels(surveyMentalHealth_clean$leave)
```

```{r}
summary (surveyMentalHealth_clean$leave)

```

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

```{r}
levels(surveyMentalHealth_clean$mental_health_consequence)
```

```{r}
summary (surveyMentalHealth_clean$mental_health_consequence)

```

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

```{r}
```

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

```{r}
summary (surveyMentalHealth_clean$phys_health_consequence)

```

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

```{r}
levels(surveyMentalHealth_clean$coworkers)
```

```{r}
summary (surveyMentalHealth_clean$coworkers)

```

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

```{r}
levels(surveyMentalHealth_clean$supervisor)
```

```{r}
summary (surveyMentalHealth_clean$supervisor)

```

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

```{r}
levels(surveyMentalHealth_clean$mental_health_interview)
```

```{r}
summary (surveyMentalHealth_clean$mental_health_interview)

```

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

```{r}
levels(surveyMentalHealth_clean$phys_health_interview)
```

```{r}
summary (surveyMentalHealth_clean$phys_health_interview)

```

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

```{r}
levels(surveyMentalHealth_clean$mental_vs_physical)
```

```{r}
summary (surveyMentalHealth_clean$mental_vs_physical)

```

```{r }
barplot(table(surveyMentalHealth_clean$mental_vs_physical),
```

```
      main="Importacia en la Organizacion de la salud mental sobre la física")
```

```{r}
levels(surveyMentalHealth_clean$obs_consequence)
```

```{r}
summary (surveyMentalHealth_clean$obs_consequence)

```

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

```{r}
#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")
```

```{r}
#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?")
```

```{r}
#Calculo Media
mean(surveyMentalHealth_clean$Age)
```

```{r}
#Calculo Mediana
median(surveyMentalHealth_clean$Age)
```

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

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

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

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

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

```{r}
#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")

```

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

```{r}
qqnorm(surveyMentalHealth_clean$Age)
qqline(surveyMentalHealth_clean$Age)
```

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

```{r}

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

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

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

```{r}
#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")

```

```{r}
#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)
```

```{r}
qqnorm(Age_Trans)
qqline(Age_Trans)
```

```{r}

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

```{r}
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 know")
```

```{r}
bartlett.test(list(Age_Mental,Age_Fisica,Age_Desconoce))
```

```{r}
ggplot(surveyMentalHealth_clean, aes(x = mental_vs_physical, y = Age, colour
=mental_vs_physical)) + geom_boxplot() + theme_bw()
```

```{r}
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))
```

```
```

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

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

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

```{r}
summary(aov(fit2))
```

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

Realizada por: Carlos E. Jimenez Gomez
Mª Sonia Rodríguez Cepedano