**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

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.

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?**

De las 27 variables que contiene el dataset

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

Los campos más importantes para resolver el problema serían

1.Age

2. treatment

3. mental\_vs\_physical

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 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 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?**

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, luego 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 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

## **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.

Para este análisis inicialmente podemos realizar unos estudios de manera gráfica.

Comencemos con el análisis de la normalidad para el conjunto de valores de la variable Age que consideran que la organización da mayor importancia a la salud mental vs salud física.

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 pvalor 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 pvalor, 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 influira 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.

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

Vamos a realizar el estudio ANOVA de un factor (one-way ANOVA o independent samples ANOVA) para investigar si existen diferencias en la edad entre los individuos que han tenido tratamiento o no de salud mental.

Nos hubiera gustado realizar un modelo de regresión logística pero finalmente este no ha sido posible.

## **Ejercicio 4**

**Representación de los resultados a partir de tablas y gráficas.**

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

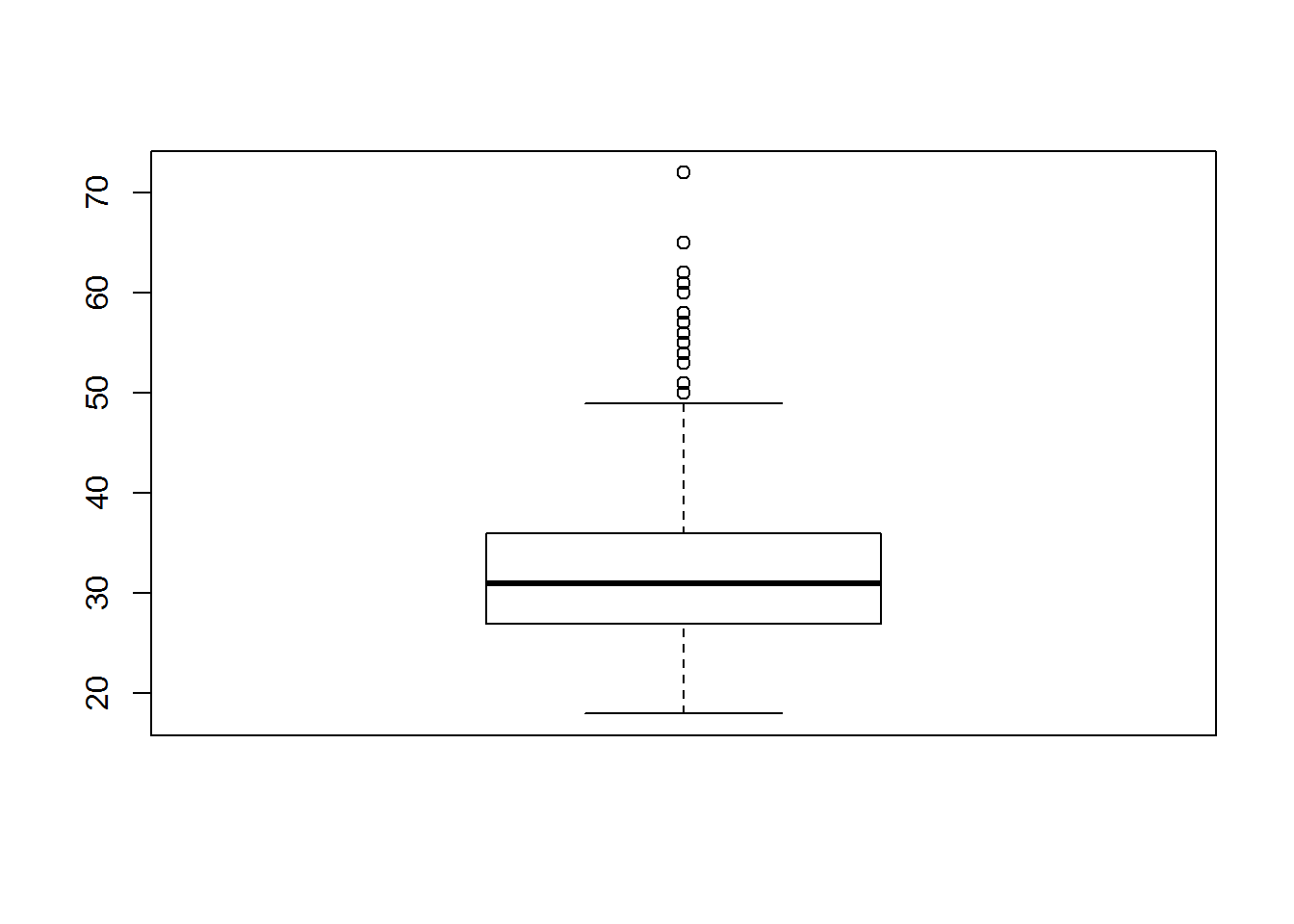
Media\_Recortada\_Age31.655723

Desviacion\_estandar\_Age7.288272

RIC\_Age9.000000

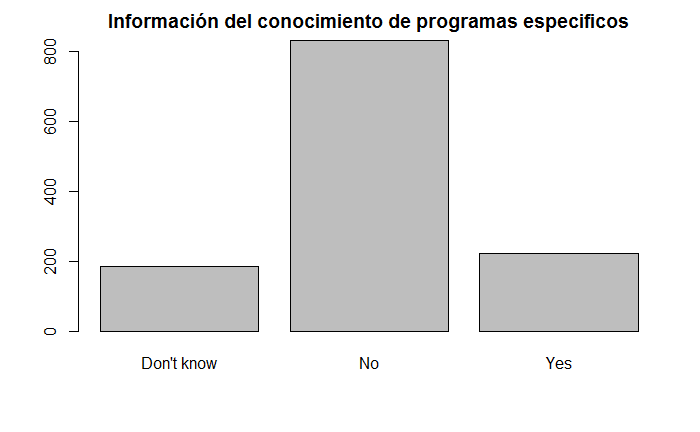
Desviacion\_Absoluta\_Mediana\_Age5.930400

De igual forma realizamos la representación del Box-Plot.

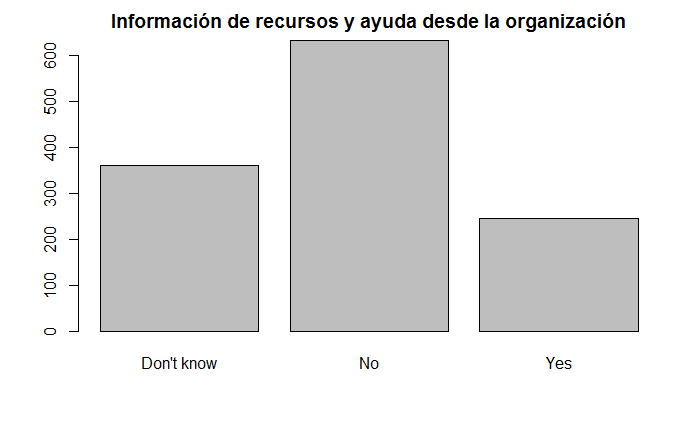


Para el resto que son variables cualitativas se realiza un análisis gráfico mediante gráficos de barras, en los cuales podemos observar los niveles de cada variable y el número de observaciones que pertenece a cada uno de ellos.

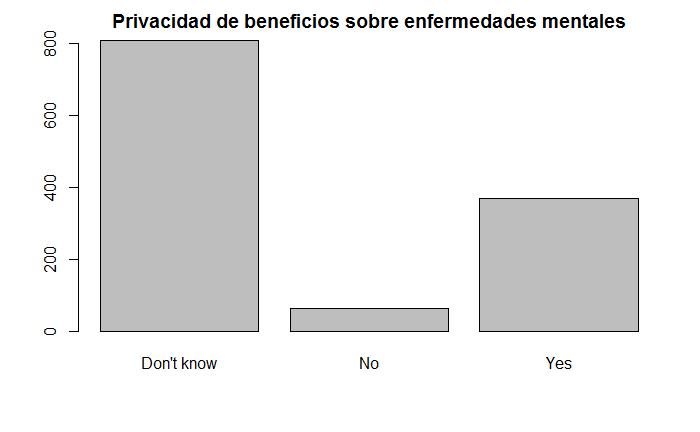
Wellness-program



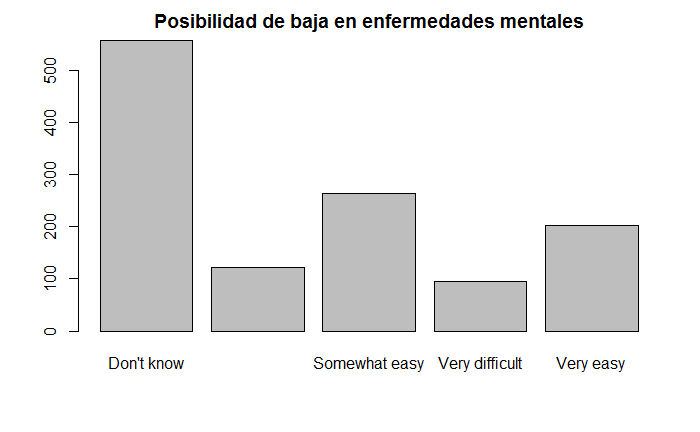
Seek\_help



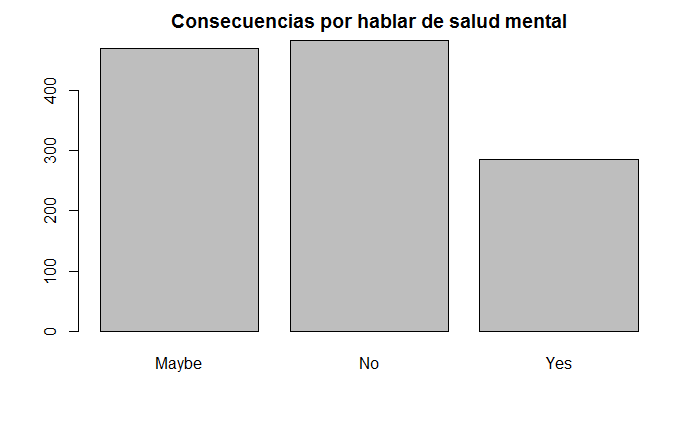
Anonymity



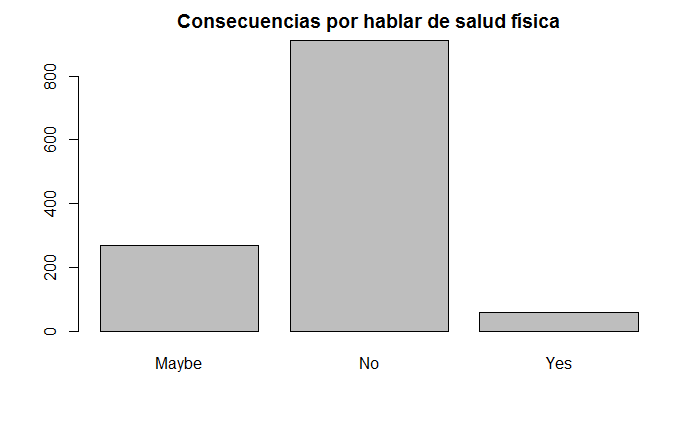
Leave



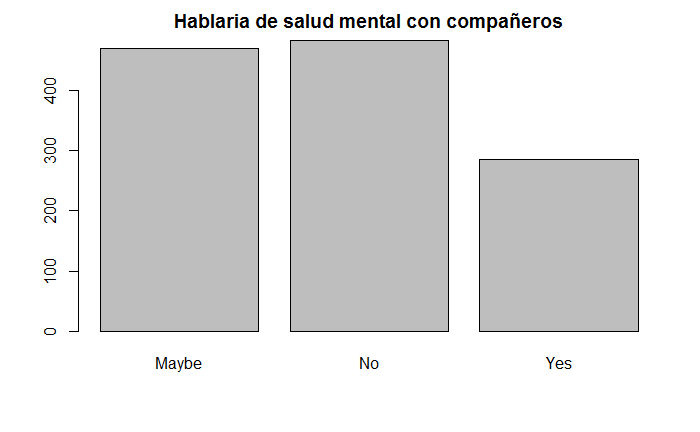
Mental\_health\_consequence



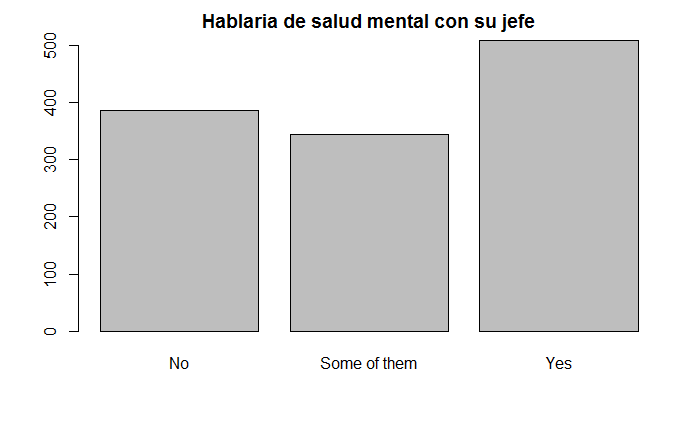
Phys\_health\_consequence



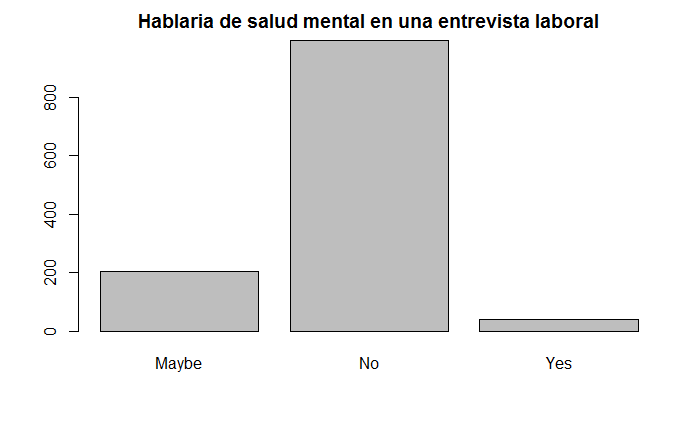
coworkers



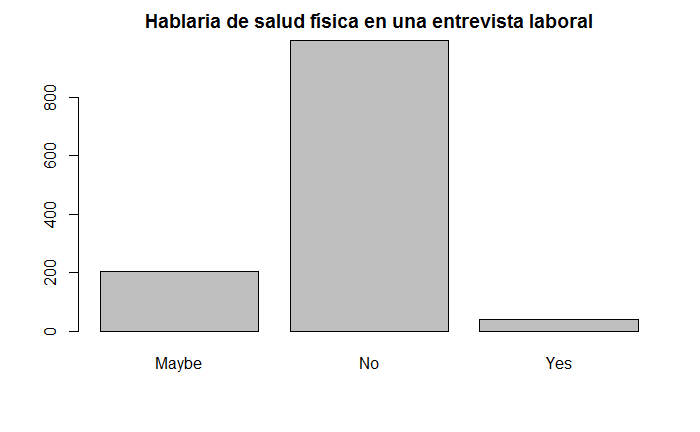
supervisor



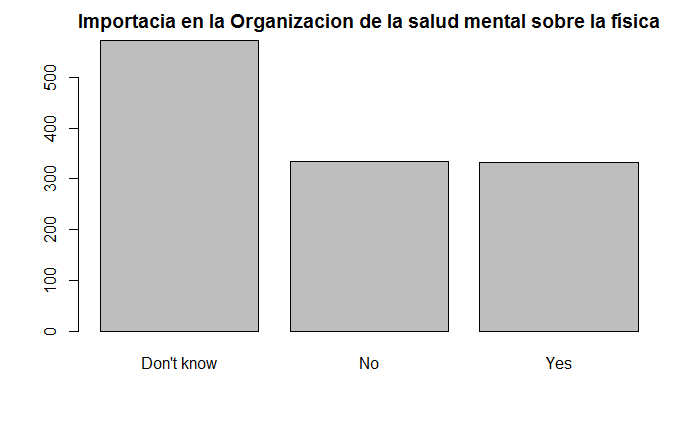
mental\_health\_interview



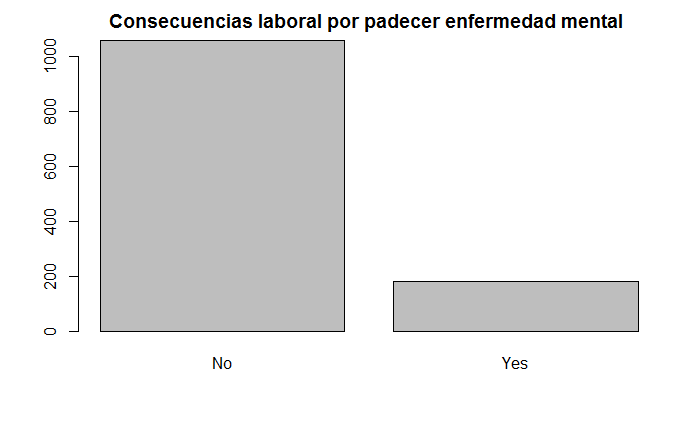
Phys\_health\_interview



mental\_vs\_physical



obs\_consequence



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

Calculo de la media

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

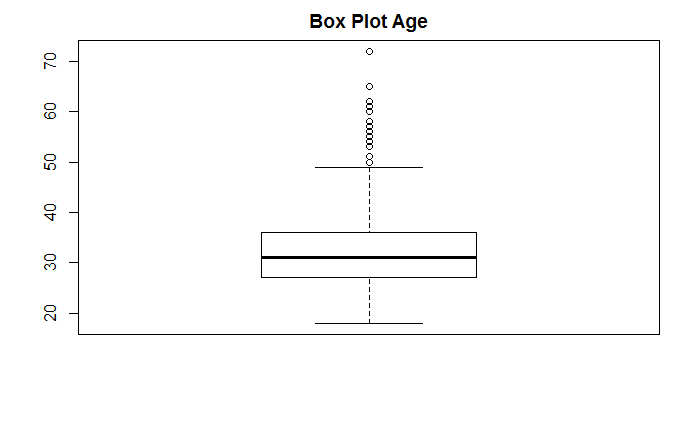
Q1 27

Mediana 31

Q3 36

Máximo 72

Gráfico boxplot para Age



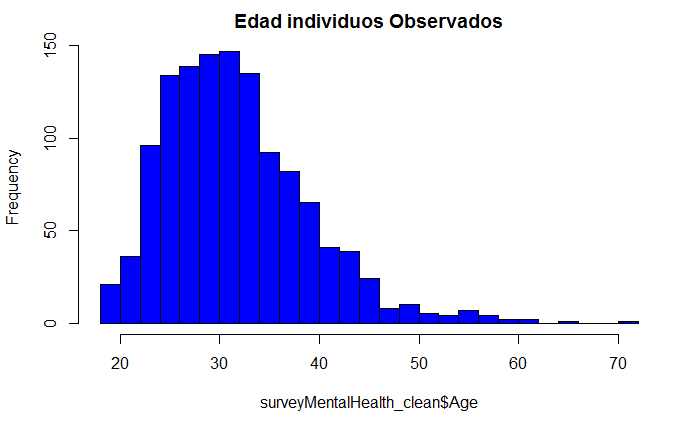
En un gráfico de Boxplot podemos estudiar la simetría, detectar outlier e incluso contrastar algunas hipótesis de la distribución. El gráfico fracciona los datos en 4 partes de igual frecuencia, es decir, cada grupo contiene mas o menos el mismo número de observaciones. Pero la ocupación de estos es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

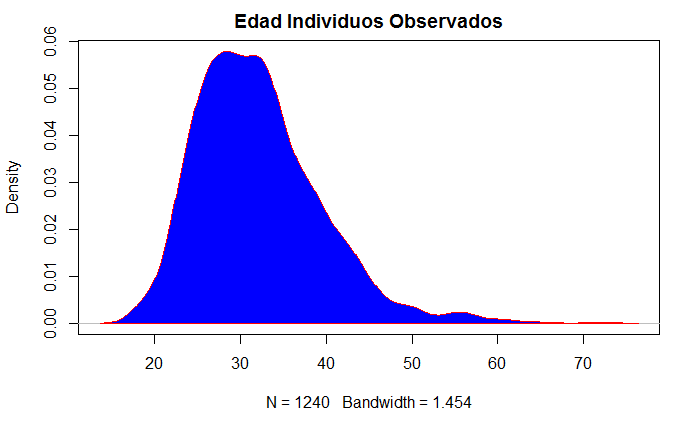
Representamos el histograma de la variable Age de la muestra. Para calcular el número de clases

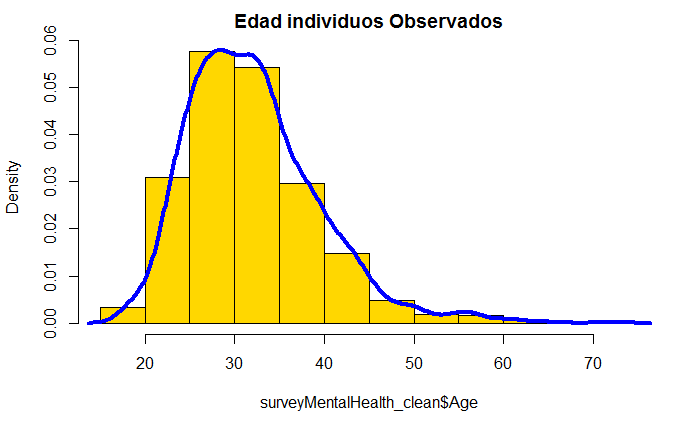
que necesitamos realizamos el siguiente cálculo k=1+3,3\*log(n) ó k=\sqrt(n).

El número de intervalos correspondientes Age es 35.

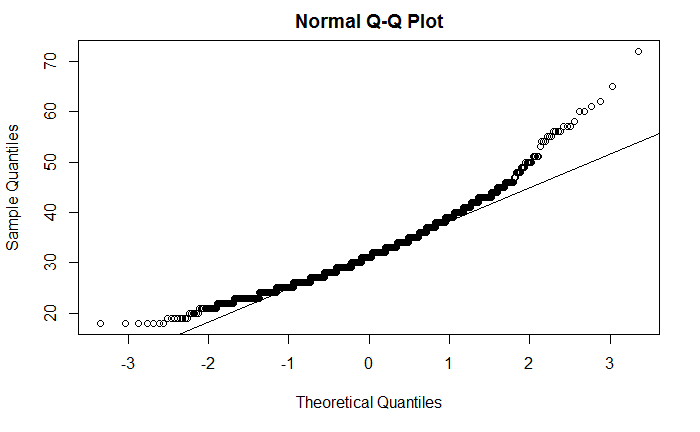


Calculamos la función de densidad de para finalmente superponerla sobre el histograma anteriormente representado





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

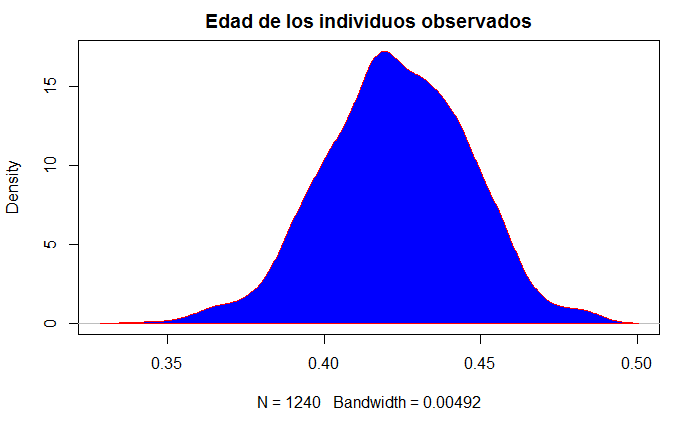
Jarque-Bera test for normality

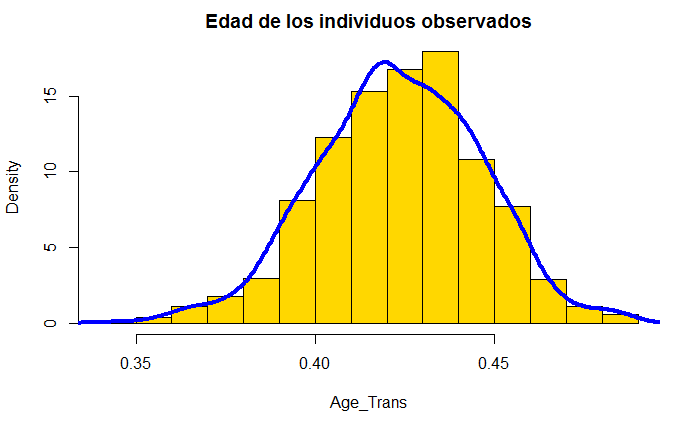
data: surveyMentalHealth\_clean$Age

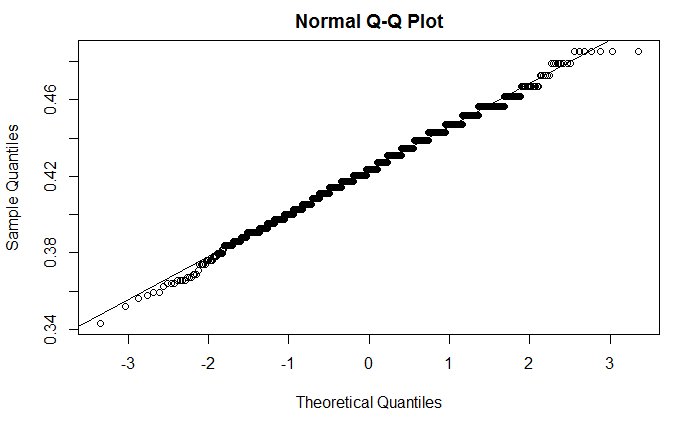
JB = 392.78, p-value < 2.2e-16

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 √(√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.

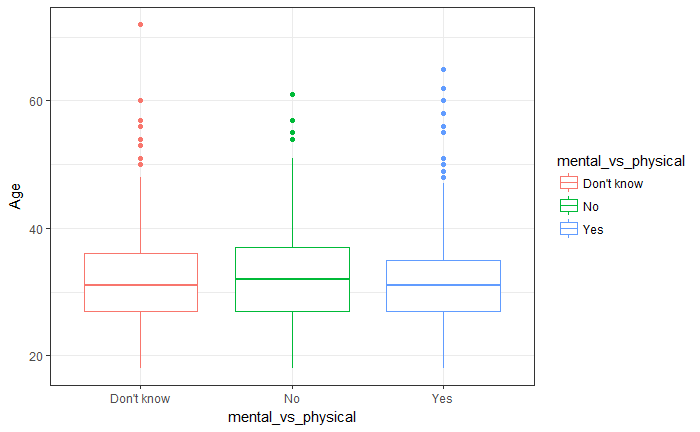
Bartlett test of homogeneity of variances

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.

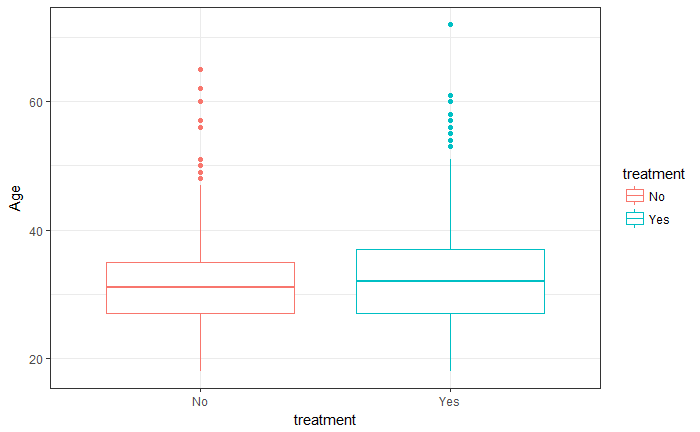
Bartlett test of homogeneity of variances

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.