Practica 2 - Preprocesado de datos

1. Dataset

El objetivo de esta práctica es identificar los datos relevantes en un proyecto analítico y usar las herramientas de integración, limpieza, validación y análisis de los datos. Para ello vamos a utilizar un dataset que hemos obtenido de la plataforma Kagle. Los datos que hemos elegido son los siguiente: https://www.kaggle.com/datasets/arashnic/fitbit

Este dataset nos presenta un conjunto de datos recolectados desde un smartwatch que recoge la información de salud de su usuario. Tenemos información relativa al ritmo cardíaco, al sueño, nivel de actividad física, calorías consumidas, el peso y los pasos realizados.

Este dataset contiene tres archivos .csv que explicamos a continuación:

- El primer archivo, *dailyActivity_merged.csv,* trata sobre la actividad diaria de un usuario y, contiene los datos sobre el número de pasos realizados, la distancia total recorrida, y diversos datos que categorizan el tipo de actividad que se ha realizado midiendolas en tiempo y distancia, además de las calorías que ha quemado. Un ejemplo, es un usuario que a lo largo de un día ha realizado 13.162 pasos, con una distancia total recorrida de 8.5 Km. De estos 8.5 Km, 1.87 Km se han realizado en una alta intensidad, 0.5 Km en una intensidad moderada y el resto, 6.05 Km en un paso tranquilo, caminando. Se cuantifica que el usuario ha estado 25 minutos muy activo, 13 min. bastante activo, 328 min. algo activo y 728 en modo sedentario. En total en este periodo de actividad ha consumido 1985 calorías.

- El segundo archivo, *sleepDay_merged.csv,* trata sobre el sueño durante un día de un usuario. Los datos que se recogen en este dataset son el día que grabó el sueño, cuántas veces se grabó, cuántos minutos ha dormido y cuánto tiempo ha estado en la cama. Por ejemplo este usuario el día 12 de abril del 2016, grabó 1 etapa de sueño donde estuvo en la cama durante 346 minutos de los cuales 327 los pasó durmiendo.

- El tercer archivo, *heartrate_second_merged.csv*, trata sobre el registro del pulso cada 5 segundos durante el transcurso de varios días. La primera muestra de este dataset muestra que el usuario el día 12 de abril del 2016 tenía un pulso cardíaco de 97 pulsaciones por minuto a las 7:21:00 de la mañana.

El objetivo de analizar estos tres datasets es comprobar la correlación que existe entre la actividad y el pulso que un usuario registra a lo largo del día y cómo afecta a la calidad del

sueño. Para ello vamos a utilizar estos tres datasets que nos brindan la oportunidad de tratar con datos reales de un usuario.

2. Integración y selección

Los datos obtenidos de Kagle están formados por varios archivos en nuestro caso concreto vamos a utilizar tres de ellos:

- **dailyActivity_merged.csv:** aquí encontramos los datos de actividad de una persona identificada mediante un ID. También la fecha del registro. A parte de esta información se recogen datos como la distancia recorrida, el tiempo de ejercicio, las calorías quemadas…

- **Heartrate_seconds_merged.csv:** en este dataset nos encontramos con el ritmo cardíaco, esté datos también se identifica con un ID y una fecha pero en este caso se recogen datos de varias horas diferentes en un día. En vez de una única observación por día con es el caso del dataset anterior.

- **sleepDay_merged.csv:** por último en este dataset nos encontramos con los datos de sueño de un usuario, es decir, los minutos totales que ha dormido y el rato total que se ha detectado que ha estado en la cama, aunque no necesariamente dormido. También tenemos una fecha pero solo hay un registro por día, como en el primer dataset, aunque se añade la hora.

Analizando los ID que encontramos en los tres datasets, identificamos 12 ID que coinciden en los 3 archivos. Por lo que unificamos los tres datasets en base a ese ID y la fecha del registro de datos. En el caso de los datos del ritmo cardiaco, donde tenemos varias observaciones en un mismo día calcularemos la media de cada día y será esa información la que utilizaremos.

De esta forma evitamos tener un número elevado de datos vacíos o nulos ya que no aportaran información a nuestro estudio.

Por último también debemos adecuar el formato de la fecha a que sea uniforme en los tres datasets para poder generar un único dataset con toda la información. Una vez realizados todos estos pasos y generado el dataset ya podemos pasar al siguiente proceso.

3. Limpieza de los datos

Una vez que tenemos el dataset unificado ya podemos pasar al proceso de limpieza de datos. Lo primero que debemos comprobar es la existencia de datos nulos. Teniendo en cuenta la forma en la que hemos unido los dataset no hay ningún dato nulo pero sí que observamos la existencia de tres columnas donde la mayoría de la información es 0 o apenas varía a lo largo de todas sus observaciones.

Se trata de las siguientes columnas:

```
##  LoggedActivitiesDistance SedentaryActiveDistance TotalSleepRecords
##  Min.   :0.00000          Min.   :0.000000        Min.   :1.000
##  1st Qu.:0.00000          1st Qu.:0.000000        1st Qu.:1.000
##  Median :0.00000          Median :0.000000        Median :1.000
##  Mean   :0.05514          Mean   :0.002088        Mean   :1.121
##  3rd Qu.:0.00000          3rd Qu.:0.000000        3rd Qu.:1.000
##  Max.   :4.08169          Max.   :0.110000        Max.   :3.000
```

Comprobando más en detalle las columnas de “LoggedActivitiesDistance” y “SedentaryActiveDistance”, observamos que solo hay 3 y 6 valores diferentes de 0 respectivamente. Dada la poca información que nos aportan estos datos decidimos quitarlos.

Por otro lado tenemos la columna de “TotalSleepRecords” que se distribuye de la siguiente manera:

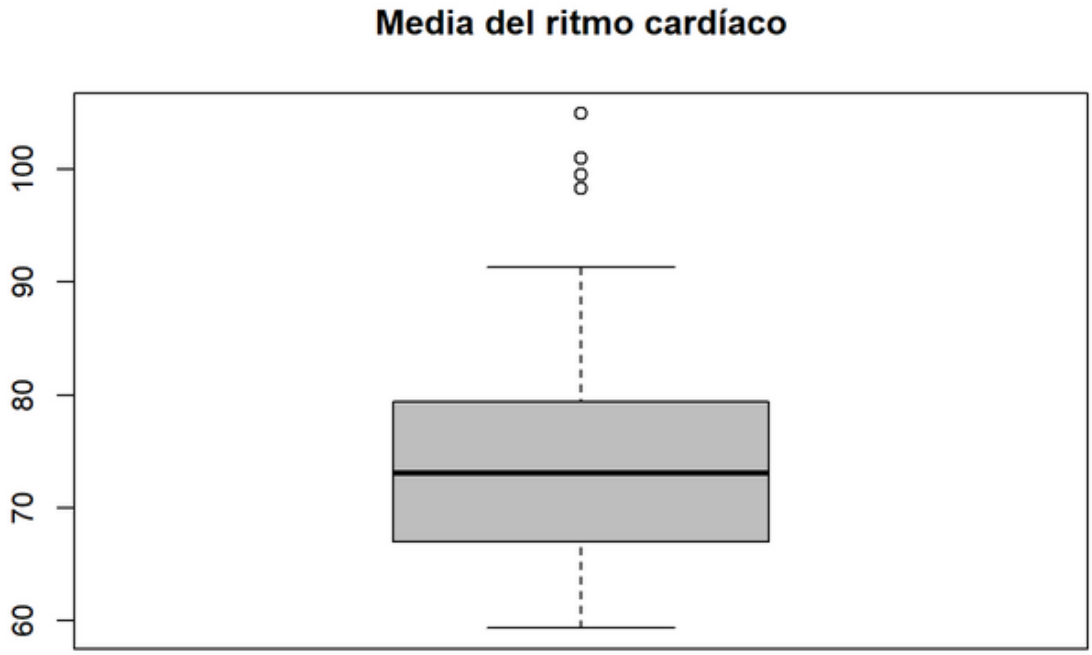
```
##
##   1   2   3
## 162  18   2
```

También debemos mencionar que este dato representa el número de registros de sueño detectados en un día, por otro lado disponemos del tiempo total dormido por lo que consideramos que esta variable no aporta información relevante a nuestro dataset. Además de que apenas varía ya que la mayoría de las observaciones tienen valor uno. Por lo que también decidimos quitarla.

Lo siguiente que debemos comprobar es la existencia de datos atípicos y extremos. Observando el resumen genérico de todos los datos hay dos que nos llaman la atención:

- HeartrateMean
- TotalMinutesAsleep

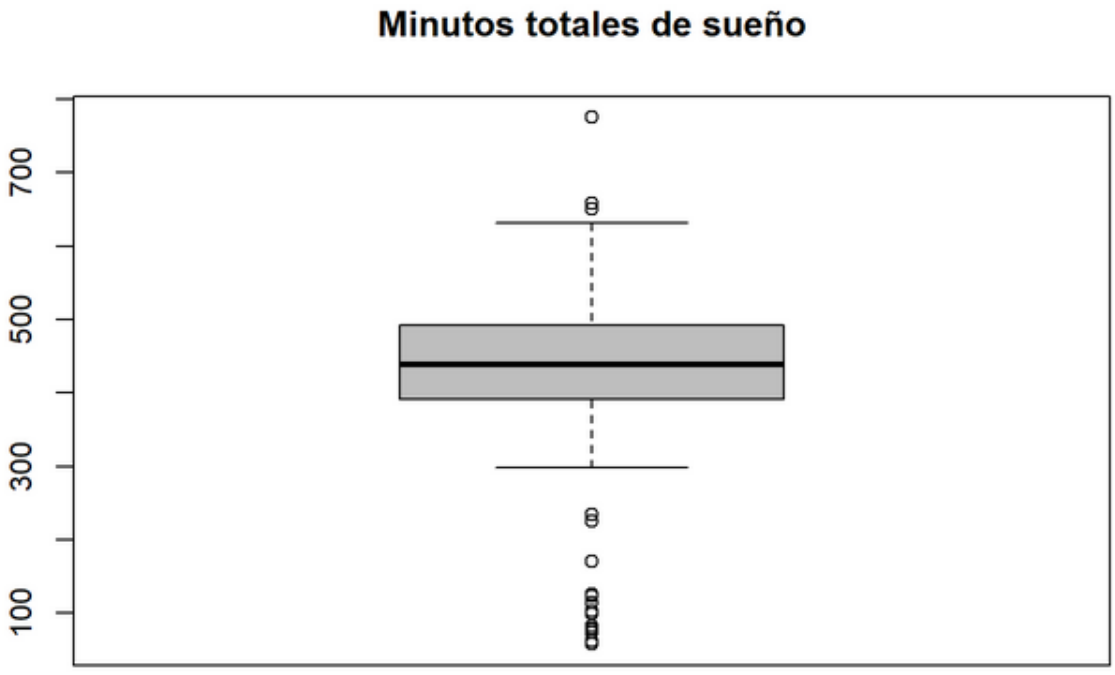
Lo que hacemos es comprobar visualmente cuantos valores extremos existen y si existe la posibilidad de que sean algún error. Comenzamos calculando el boxplot o diagrama de cajas y bigotes de la variable HeartrateMean:



Este tipo de gráficos nos permite visualizar y comparar la distribución y la tendencia central de valores numéricos mediante sus cuartiles. Todo lo que queda fuera de los bigotes lo podemos considerar datos atípicos. En este caso las tres observaciones que podemos ver hacen referencia a medias de 99, 98, 100 y 104 de ritmo cardiaco. Es un ritmo muy elevado pero es posible por lo que no vamos a quitarlo de nuestro dataset.

.

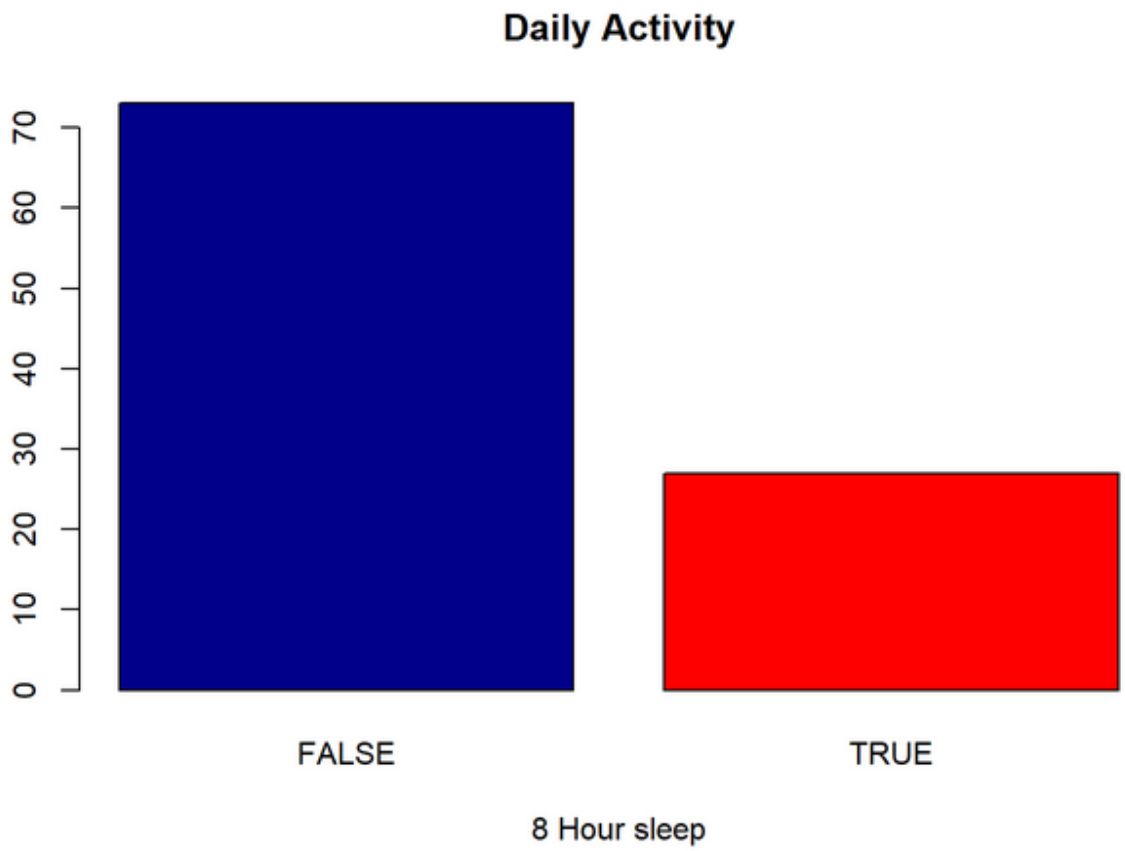
El otro valor que nos ha llamado la atención es de “TotalMinutesAsleep”, vamos a generar el mismo gráfico:



En este caso observamos que hay muchos más valores por fuera de los bigotes. Mostrando los resultados de este cálculo hemos podido comprobar que todos son casos de haber dormido poco más de 1 hora o de dormir entre 12 y 13 horas. Como en el caso anterior son casos extremos pero posibles, por lo que también vamos a mantenerlos en nuestro dataset.

Por otro lado vamos a crear dos variables booleanas que nos ayuden a realizar el estudio de los datos.

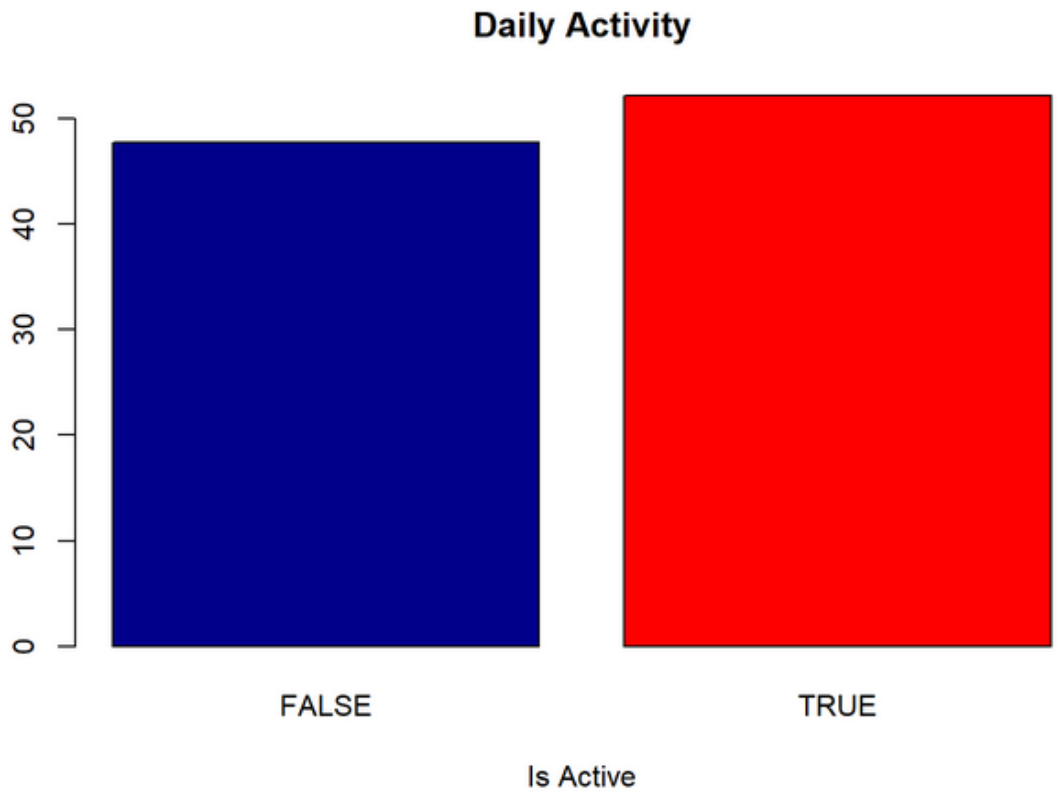
La primera muestra si una persona ha podido dormir más de ocho horas o no, ya que son las recomendadas. Vamos a graficar esta nueva variable para ver cómo se distribuye:



Podemos observar que la mayoría de las personas que han registrado sus valores han dormido menos de 8 horas

.

La otra variable booleana que vamos a crear representa si una persona es activa o no. Para ello la condición que deben cumplir es que hayan hecho 30 minutos o más de ejercicio activo o moderado. Es decir, que la suma de esos dos tiempos sea mayor o igual a treinta. En concreto hace referencia a las variables “VeryActiveMinutes” y “FairlyActiveMinutes”. También vamos a mostrar cómo se distribuye:



En este caso se distribuye de una forma más equitativa aunque sí que hay más casos de gente activa que de no activa.

4. Análisis de los datos

Una vez que ya hemos revisado los datos, hemos podido unificar los tres datasets en uno solo, nos hemos asegurado de que no hay nulos ni datos erróneos. También hemos generado nuevas variables que nos pueden facilitar realizar el análisis de los datos por lo que ya podemos comenzar con dicho proceso.

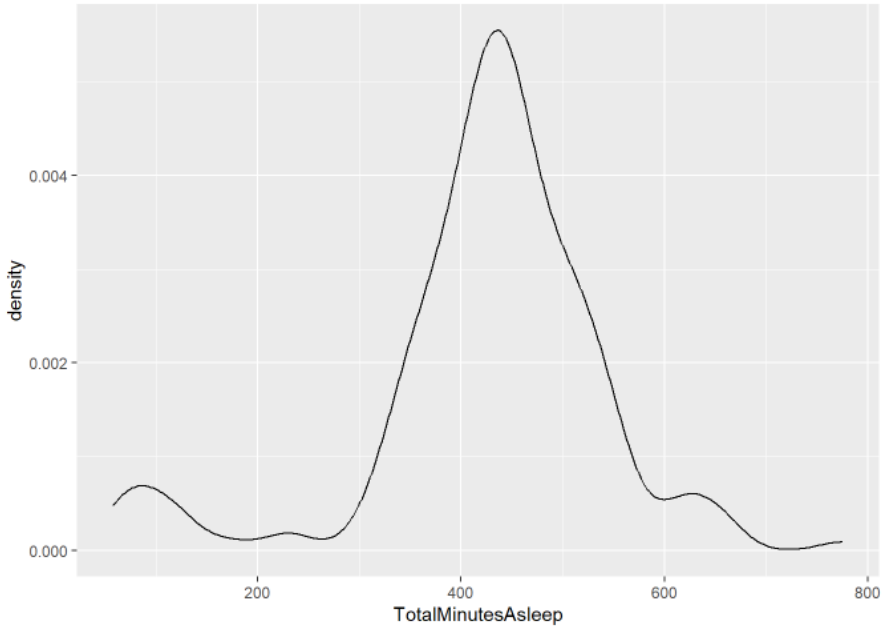
El primer paso es seleccionar los grupos de datos que vamos a utilizar para nuestro estudio. En nuestro caso concreto queremos comprobar cómo afecta o si afecta o no el nivel y tiempo de actividad física a nuestras horas de sueño.

Por ello las variables que vamos a utilizar para nuestro estudio son las siguientes:

- VeryActiveMinutes → minutos de actividad física de intensidad elevada
- FairlyActiveMinutes → minutos de actividad física de intensidad moderada
- LightlyActiveMinutos → minutos de actividad física suave
- SedentaryMinutos → minutos sin actividad física
- TotalMinutesAsleep → minutos totales de sueño
- Sleep8h → boolean que representa si una persona ha dormido 8 horas o más
- IsActive → boolean que representa si una persona ha sido activa a lo largo del día o no

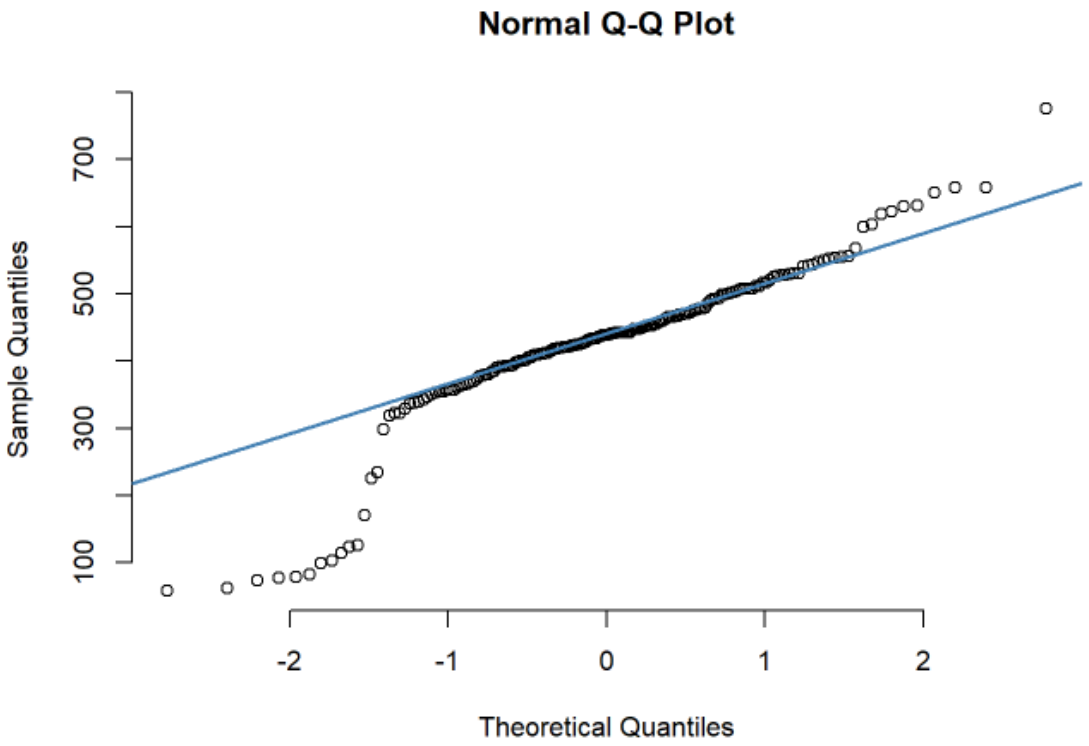
Una vez que ya sabemos sobre qué variables vamos a realizar el estudio debemos comprobar su normalidad y la homocedasticidad. Es decir, si la variable sigue una distribución normal y si la varianza de las variables difiere mucho entre ellas.

Dado que la principal variable que vamos a estudiar son los minutos totales de sueño vamos a comprobar la normalidad de dicha variable. Primero vamos a mostrar el gráfico de su distribución para ver si visualmente podemos ver si se acerca a una distribución normal.



Si solo nos centramos en la parte central de la gráfica sí que podríamos decir que sigue una distribución normal pero en la parte izquierda de la gráfica hay un pico de casos por lo que no podemos afirmarlo.

Otra forma en la que podemos verificar si se trata de una distribución normal es calculando del gráfico QQ plot:



Para que los datos sigan una distribución normal los puntos deben ajustarse a la distribución teórica, que sería la línea diagonal. Como podemos observar en el gráfico no es el caso.

Igualmente como se trata de una muestra mayor a 30 observaciones, por lo que podemos aplicar el teorema central del límite y decir que la media sigue una distribución normal.

Para comprobar la homogeneidad de la varianza también llamada homocedasticidad vamos a aplicar el test de Levene. Un supuesto que hay que cumplir para poder aplicar este test es que la variable siga una distribución normal. Como hemos explicado en el párrafo anterior utilizando el teorema central del límite podemos afirmar que la muestra sigue una distribución normal

.

Lo siguiente que vamos a comprobar es si la varianza estadística de la variable TotalMinutesAsleep es igual o no a la varianza de la variable IsActive, que representa si una persona ha sido activa a lo largo del día o no. Consideremos que es activa si ha realizado 30 minutos o más de actividad intensa o moderada. Una vez que realizamos el cálculo obtenemos el siguiente resultado:

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value Pr(>F)
## group    1  1.3007 0.2556
##        180
```

Obtenemos un p-valor mayor a 0.05, por lo que aceptamos la hipótesis nula de homocedasticidad y concluimos que la variable TotalMinutesAsleep tiene un varianza estadísticamente similar a si una persona es activa o no.

Una vez que ya hemos probado la normalidad y homocedasticidad de la variable podemos pasar al contraste de hipótesis. En el que vamos a comprobar si existen diferencias estadísticamente significativas entre las variables de TotalMinutesAsleep y IsActive. Para ello vamos a aplicar el test de T de Student. Con el que obtenemos el siguiente resultado:

```
##
##  Welch Two Sample t-test
##
## data:  TotalMinutesAsleep by IsActive
## t = 2.7172, df = 171.22, p-value = 0.007259
## alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
## 95 percent confidence interval:
##  12.74652 80.44356
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            450.8161            404.2211
```

El p-valor obtenido es menor a 0.05, es decir es menor al nivel de significancia, por lo que podemos decir que se observan diferencias estadísticamente significativas entre los grupos de datos IsActive y TotalMinutesAsleep.

También vamos a calcular el coeficiente de correlación. Este coeficiente es una medida de la asociación entre dos variables. Este puede tomar valores entre -1 y 1, donde los extremos indican una correlación perfecta y el 0 indica la ausencia de correlación. El signo es negativo cuando valores elevados de una variable se asocian con valores pequeños de la otra, y el signo es positivo cuando ambas variables tienden a incrementar o disminuir simultáneamente.

Vamos a comprobar si hay correlación entre la variable de minutos totales que una persona ha dormido y si es activa o no. Este coeficiente sólo se puede calcular con variables numéricas. Por lo que como hemos utilizado la suma de los minutos de actividad para crear la variable boolena IsActive vamos a calcular el coeficiente de correlación con ese mismo valor.

Para ello vamos a utilizar el método cor.test de r que nos permite obtener este resultado de una forma sencilla:

```
##
##  Pearson's product-moment correlation
##
## data:  df$TotalMinutesAsleep and (df$VeryActiveMinutes + df$FairlyActiveMinutes)
## t = -1.1717, df = 180, p-value = 0.2429
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.22955202  0.05920334
## sample estimates:
##         cor
## -0.08700142
```

Obtenemos un nivel de correlación de -0.087, este valor está bastante alejado de -1 por lo que el nivel de correlación es muy bajo.

Por último vamos a intentar obtener un modelo de regresión lineal que pueda explicar la variables de TotalMinutesAsleep, ya que el objetivo de nuestro estudio es comprobar cómo afecta la actividad física a nuestra calidad del sueño.

Un modelo de regresión lineal es un tipo de modelo matemático que tiene como objetivo aproximar la relación entre una variable dependiente y una o varias variables independientes.

Para ello vamos a utilizar el método lm de r que nos permite realizar este cálculo. Obtenemos el siguiente resultado:

```
##
## Call:
## lm(formula = TotalMinutesAsleep ~ VeryActiveMinutes + FairlyActiveMinutes +
##     LightlyActiveMinutes + SedentaryMinutes, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -392.21  -35.77   12.63   50.44  208.25
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          805.01917   37.14386  21.673  < 2e-16 ***
## VeryActiveMinutes     -0.20316    0.21734  -0.935  0.35119
## FairlyActiveMinutes   -0.92014    0.51061  -1.802  0.07324 .
## LightlyActiveMinutes  -0.23632    0.07547  -3.131  0.00204 **
## SedentaryMinutes      -0.43166    0.03769 -11.452  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 88.32 on 177 degrees of freedom
## Multiple R-squared:  0.4417, Adjusted R-squared:  0.4291
## F-statistic: 35.01 on 4 and 177 DF,  p-value: < 2.2e-16
```

La medida de calidad del modelo es R-squared, tomando valores entre 0 y 1. En nuestro caso toma un valor de 0.4417, es decir, nuestro modelo puede explicar el 44.17% de la variabilidad observada.

5. Resolución del problema - Conclusiones

Una vez realizado el estudio, podemos concluir que a pesar de que pensábamos el tiempo del sueño iba completamente relacionado con la actividad física del usuario, esto no es así ya que en el estudio del coeficiente de correlación nos da un índice de -0,087, un valor muy alejado del -1 esperado, y en el test de T de Student el valor p-valor obtenido es menor de 0.05, por lo tanto nos contradice con nuestra hipótesis.

En un estudio futuro, deberíamos realizar el estudio con la calidad del sueño, teniendo en cuenta cuánto de ese tiempo dormido pertenece a los diferentes estadios del sueño, fase del sueño ligero, etapa de transición, sueño profundo y fase REM. También podríamos considerar la relación entre el tiempo que el usuario pasa en la cama respecto al tiempo dormido e incorporar el registro del pulso durante los diferentes estadios del sueño para determinar el nivel de estrés.

6. Contribuciones

Contribuciones	**Firma**
Investigación previa	DC, PM
Redacción de las respuestas	DC, PM
Desarrollo del código	DC, PM
Participación en el vídeo	DC, PM