Estadística descriptiva e interpretación de variables

Natalia Gordo Herrera

10/03/2023

Creación y descripción de la base de datos.

Lo primero que hay que hacer antes de empezar ningún tipo de análisis es conocer la base de datos que vamos a utilizar. Esto implica saber el contenido y, además, poder clasificar las diferentes variables en función de su clase, es decir, en función de si son variables continuas o discretas (categóricas).

Además, habrá que crear variables en función de nuestros intereses.

Leemos, a continuación, la base de datos.

```
data <- read.csv('Sleep_Efficiency.csv', sep = ',')
head(data)
```

```
##   ID Age Gender             Bedtime         Wakeup.time Sleep.duration
## 1  1  65 Female 2021-03-06 01:00:00 2021-03-06 07:00:00            6.0
## 2  2  69   Male 2021-12-05 02:00:00 2021-12-05 09:00:00            7.0
## 3  3  40 Female 2021-05-25 21:30:00 2021-05-25 05:30:00            8.0
## 4  4  40 Female 2021-11-03 02:30:00 2021-11-03 08:30:00            6.0
## 5  5  57   Male 2021-03-13 01:00:00 2021-03-13 09:00:00            8.0
## 6  6  36 Female 2021-07-01 21:00:00 2021-07-01 04:30:00            7.5
##   Sleep.efficiency REM.sleep.percentage Deep.sleep.percentage
## 1             0.88                   18                    70
## 2             0.66                   19                    28
## 3             0.89                   20                    70
## 4             0.51                   23                    25
## 5             0.76                   27                    55
## 6             0.90                   23                    60
##   Light.sleep.percentage Awakenings Caffeine.consumption Alcohol.consumption
## 1                     12          0                    0                    0
## 2                     53          3                    0                    3
## 3                     10          1                    0                    0
## 4                     52          3                   50                    5
## 5                     18          3                    0                    3
## 6                     17          0                   NA                    0
##   Smoking.status Exercise.frequency
## 1            Yes                 3
## 2            Yes                 3
## 3             No                 3
## 4            Yes                 1
## 5             No                 3
## 6             No                 1
```

Encontramos una base de datos que consta de 452 observaciones y 15 variables.

Algunas de estas variables sirven para identificar a cada una de las personas que se ha hecho seguimiento, como son la edad o el género. Otras, sin embargo, hacen referencia al tipo de sueño y la duración del mismo.

Entendemos que es interesante, para este estudio, comprobar qué factores pueden afectar o noa la calidad del sueño o al tiempo de sueño. Por ello, tendremos que elegir una de estas variables como variable objetivo.

Por último, encontramos variables que hacen referencia al comportamiento de las personas en cuanto al consumo de sustancias como tabaco o cafeína y la rutina de ejercicio que siguen.

A partir de las variables que marcan la hora de acostarse y levantarse, vamos a crear una nueva variable que indique el día de la semana que se trata en cada caso. De esta forma, podremos estudiar si el sueño los días de diario es parecido al sueño durante los fines de semana.

```
# Pasamos ambas variables a formato fecha
data$Bedtime <- as.Date(data$Bedtime)
data$Wakeup.time <- as.Date(data$Wakeup.time)

# Creamos una variable que represente el día de la semana
data$dia <- format(data$Bedtime,"%A")

# Mostramos la base de datos
head(data)
```

```
##   ID Age Gender    Bedtime Wakeup.time Sleep.duration Sleep.efficiency
## 1  1  65 Female 2021-03-06  2021-03-06            6.0             0.88
## 2  2  69   Male 2021-12-05  2021-12-05            7.0             0.66
## 3  3  40 Female 2021-05-25  2021-05-25            8.0             0.89
## 4  4  40 Female 2021-11-03  2021-11-03            6.0             0.51
## 5  5  57   Male 2021-03-13  2021-03-13            8.0             0.76
## 6  6  36 Female 2021-07-01  2021-07-01            7.5             0.90
##   REM.sleep.percentage Deep.sleep.percentage Light.sleep.percentage Awakenings
## 1                   18                    70                     12          0
## 2                   19                    28                     53          3
## 3                   20                    70                     10          1
## 4                   23                    25                     52          3
## 5                   27                    55                     18          3
## 6                   23                    60                     17          0
##   Caffeine.consumption Alcohol.consumption Smoking.status Exercise.frequency
## 1                    0                   0            Yes                  3
## 2                    0                   3            Yes                  3
## 3                    0                   0             No                  3
## 4                   50                   5            Yes                  1
## 5                    0                   3             No                  3
## 6                   NA                   0             No                  1
##         dia
## 1    sábado
## 2   domingo
## 3    martes
## 4 miércoles
## 5    sábado
## 6    jueves
```

Podríamos eliminar aquellas variables de tipo fecha, que no vamos a utilizar y no aportan información al modelo en este caso, por ser de tipo temporal.

Además, eliminamos la variable id, que solo puede crear confusiones (es un identificador).

```
df <- data[,-c(1,4,5)]
```

Limpieza de la base de datos - estadística descriptiva.

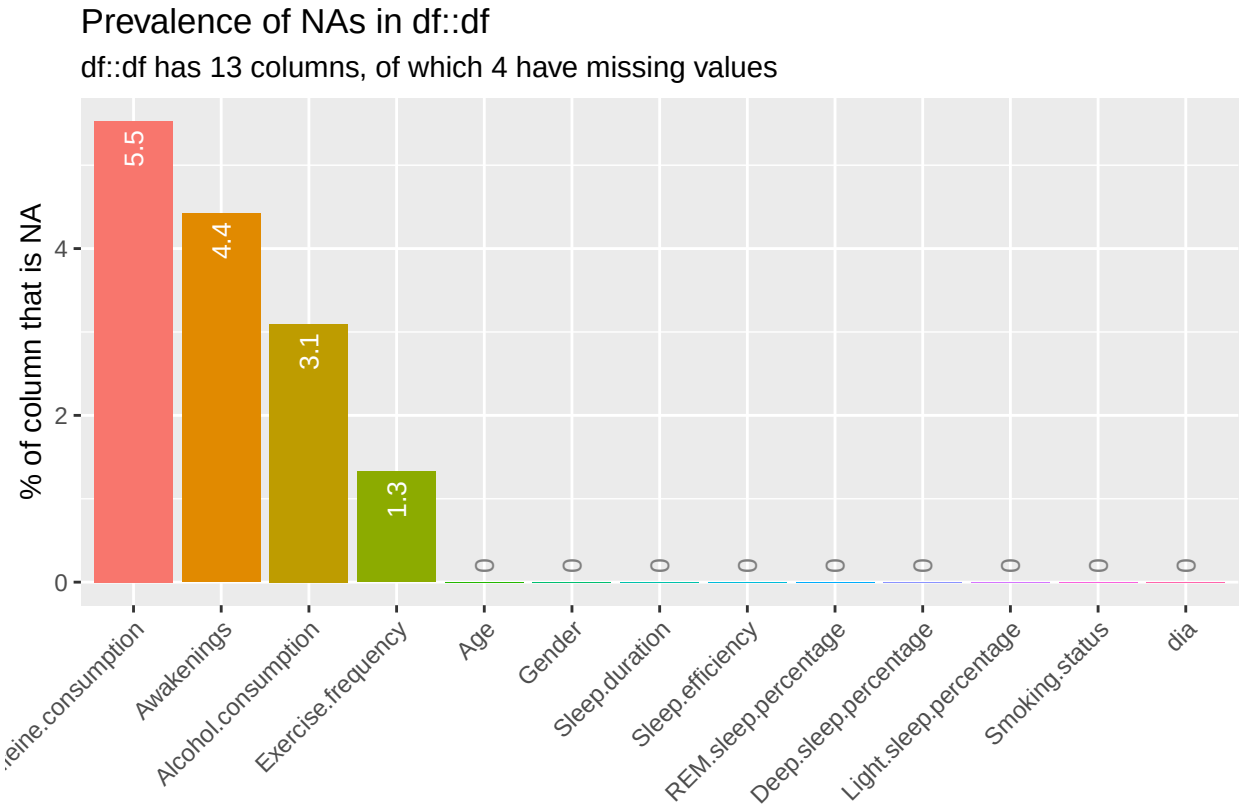
Una vez hemos definido las variables y sabemos lo que significan, tenemos que depurar la base de datos para poder trabajar con ella.

En primer lugar, miramos si hay NA entre las observaciones (datos en blanco).

```
library(inspectdf)
```

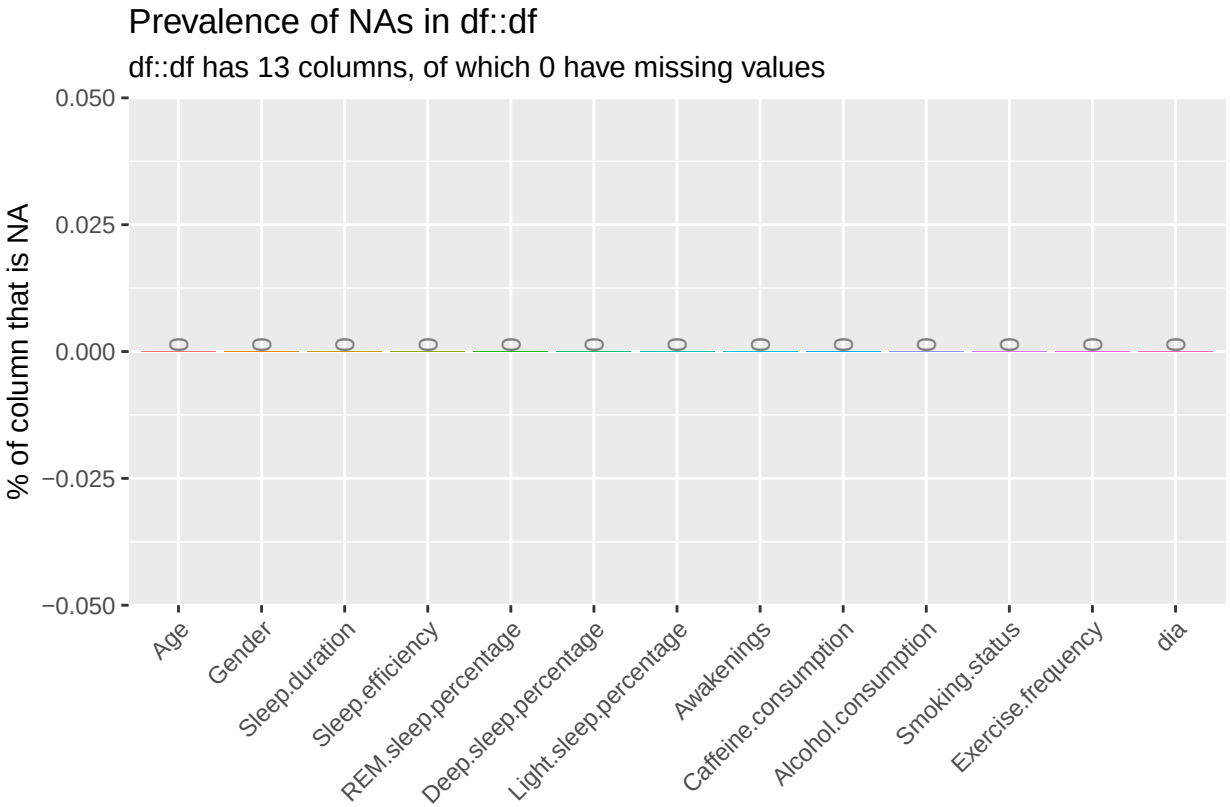
```
## Warning: package 'inspectdf' was built under R version 4.2.2
```

```
show_plot(inspect_na(df))
```



Hay 4 variables con NA y, el porcentaje de estos es menor al 10% en todos los casos, así que eliminaremos LAS OBSERVACIONES que tengan NA en alguna variable.

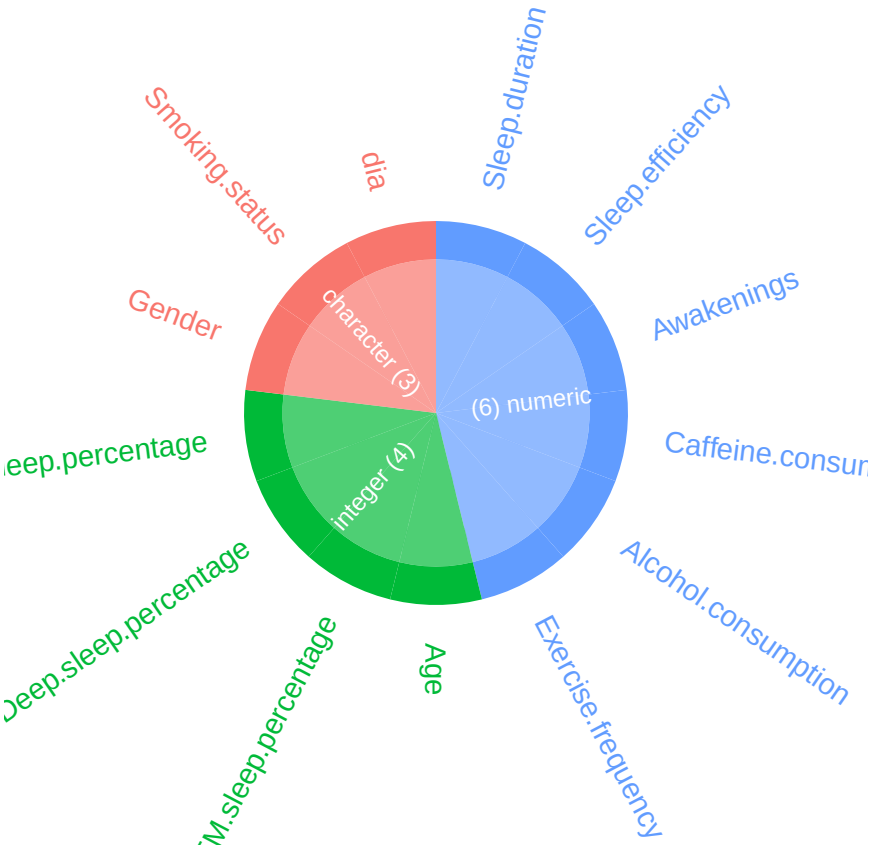
```
df <- na.omit(df)
show_plot(inspect_na(df))
```



Nos quedamos con 388 observaciones.

Lo siguiente que debemos realizar es un estudio de la naturaleza de las variables, es decir, asignar el tipo factor a aquellas que consideremos categóricas y comprobar que las numéricas estén bien identificadas.

```
show_plot(inspect_types(df))
```



Tenemos 3 variables de tipo char que pasaremos a factor. Además, debemos pasar a factor las variables que representan categorías aun siendo variables numéricas, como puede ser la frecuencia con la que se hace ejercicio o el consumo (1) o no (0) de las sustancias mencionadas anteriormente.

```
df[,c(2,8:13)] <- lapply(df[,c(2,8:13)], factor)
summary(df)
```

```
##       Age           Gender    Sleep.duration   Sleep.efficiency
##  Min.   : 9.00   Female:194   Min.   : 5.000   Min.   :0.5000
##  1st Qu.:29.00   Male  :194   1st Qu.: 7.000   1st Qu.:0.7000
##  Median :41.00                Median : 7.500   Median :0.8200
##  Mean   :40.83                Mean   : 7.451   Mean   :0.7893
##  3rd Qu.:52.00                3rd Qu.: 8.000   3rd Qu.:0.9000
##  Max.   :69.00                Max.   :10.000   Max.   :0.9900
##
##  REM.sleep.percentage Deep.sleep.percentage Light.sleep.percentage Awakenings
##  Min.   :15.00        Min.   :18.00        Min.   : 7.0           0: 87
##  1st Qu.:20.00        1st Qu.:51.00        1st Qu.:15.0           1:141
##  Median :22.00        Median :58.00        Median :18.0           2: 49
##  Mean   :22.68        Mean   :52.82        Mean   :24.5           3: 55
##  3rd Qu.:25.00        3rd Qu.:63.00        3rd Qu.:24.0           4: 56
##  Max.   :30.00        Max.   :75.00        Max.   :63.0
##
##  Caffeine.consumption Alcohol.consumption Smoking.status Exercise.frequency
##  0  :195              0:221               No :255        0:110
##  25 : 73              1: 47               Yes:133        1: 78
```

```
##  50 : 97             2: 35                           2: 45
##  75 : 19             3: 39                           3:113
##  100:  1             4: 19                           4: 35
##  200:  3             5: 27                           5:  7
##
##         dia
##  domingo  :50
##  jueves   :55
##  lunes    :49
##  martes   :48
##  miércoles:59
##  sábado   :64
##  viernes  :63
```

Observamos que hay 6 variables continuas, de las que posteriormente estudiaremos las distribuciones y se tomarán las decisiones oportunas:

- **Age:** Edad de cada individuo.
- **Sleep.duration:** Horas que duerme cada individuo cada día de la muestra.
- **Sleep.efficiency:**
- **REM.sleep.percentage, Deep.sleep.percentage, Light.sleep.percentage:** Porcentaje de las horas de sueño que el individuo está en cada fase (REM, profundo o ligero).

Además, observamos las variables categóricas que habíamos mencionado con anterioridad en las que encontramos algunas observaciones que toman ciertos valores no esperados.

Por ejemplo, en la variable que representa el consumo de cafeína tiene 4 valores que parecen 'outliers', al menos no son entendibles pues tomar 100-200 mg de cafeína al día implicaría tomar un total de (aproximadamente) 8 cafés. Eliminaremos Estas tres observaciones.

```
# Elimino las observaciones
df <- df[df$Caffeine.consumption != 200,]
df <- df[df$Caffeine.consumption != 100,]

# Elimino las categorías que se han quedado con representación 0
df$Caffeine.consumption <- droplevels(df$Caffeine.consumption)

summary(df)
```

```
##       Age          Gender    Sleep.duration   Sleep.efficiency
##  Min.   : 9.0   Female:194   Min.   : 5.000   Min.   :0.5000
##  1st Qu.:29.0   Male  :190   1st Qu.: 7.000   1st Qu.:0.6975
##  Median :41.0                Median : 7.500   Median :0.8200
##  Mean   :40.9                Mean   : 7.453   Mean   :0.7880
##  3rd Qu.:52.0                3rd Qu.: 8.000   3rd Qu.:0.9000
##  Max.   :69.0                Max.   :10.000   Max.   :0.9900
##
##  REM.sleep.percentage Deep.sleep.percentage Light.sleep.percentage Awakenings
##  Min.   :15.00        Min.   :18.00        Min.   : 7.00          0: 84
##  1st Qu.:20.00        1st Qu.:47.50        1st Qu.:15.00          1:141
##  Median :22.00        Median :58.00        Median :18.00          2: 48
```

```
##  Mean   :22.68        Mean   :52.74         Mean   :24.58          3: 55
##  3rd Qu.:25.00        3rd Qu.:63.00         3rd Qu.:32.50          4: 56
##  Max.   :30.00        Max.   :75.00         Max.   :63.00
##
##  Caffeine.consumption Alcohol.consumption Smoking.status Exercise.frequency
##  0 :195               0:217               No :251        0:110
##  25: 73               1: 47               Yes:133        1: 78
##  50: 97               2: 35                              2: 44
##  75: 19               3: 39                              3:110
##                       4: 19                              4: 35
##                       5: 27                              5:  7
##
##         dia
##  domingo  :49
##  jueves   :55
##  lunes    :48
##  martes   :47
##  miércoles:59
##  sábado   :63
##  viernes  :63
```

Otra variable categórica que podría dar problemas es la que representa la frecuencia de ejercicio de los individuos. Por ello, crearé una categoría única con las observaciones que tienen frecuencia 4 y 5.

```
library(car)
```

```
## Warning: package 'car' was built under R version 4.2.2
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 4.2.2
```

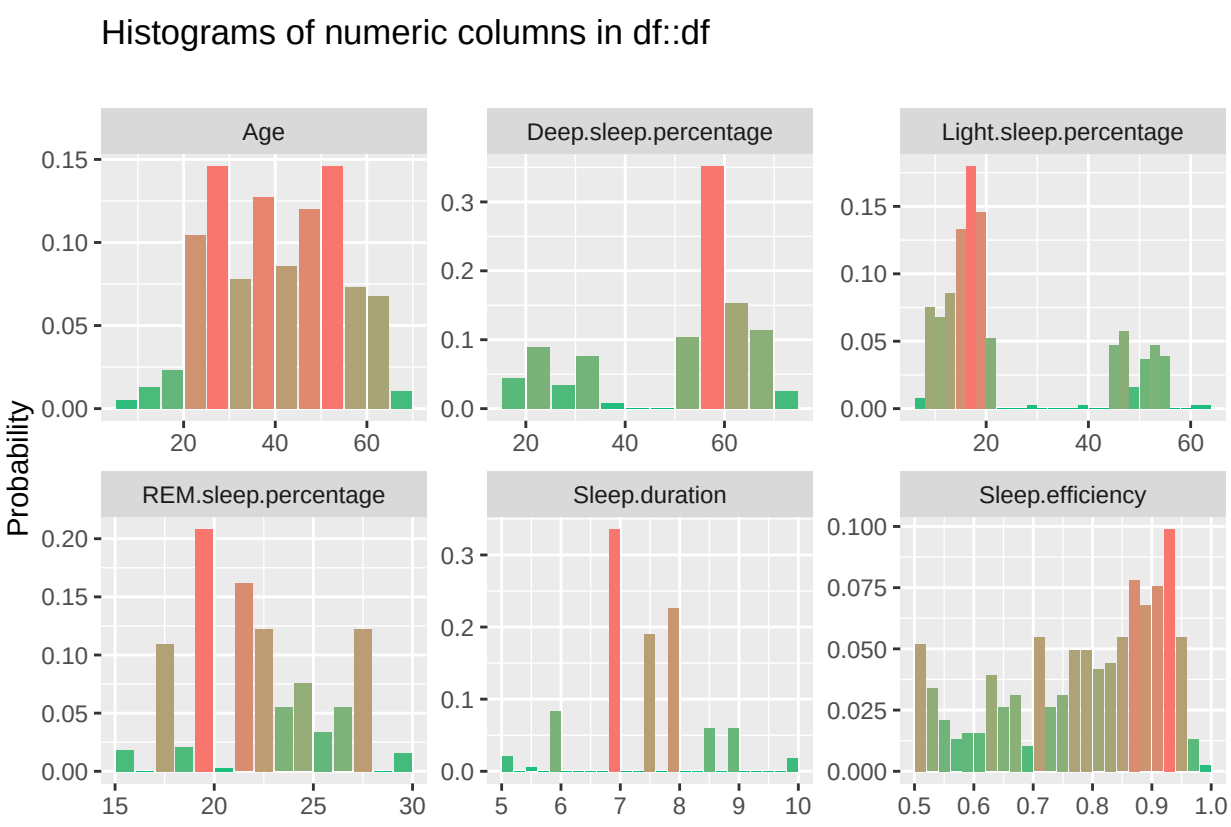
```
df$Exercise.frequency <- recode(df$Exercise.frequency, "'4' = '4-5';'5' = '4-5'")

summary(df)
```

```
##       Age          Gender    Sleep.duration   Sleep.efficiency
##  Min.   : 9.0   Female:194   Min.   : 5.000   Min.   :0.5000
##  1st Qu.:29.0   Male  :190   1st Qu.: 7.000   1st Qu.:0.6975
##  Median :41.0                Median : 7.500   Median :0.8200
##  Mean   :40.9                Mean   : 7.453   Mean   :0.7880
##  3rd Qu.:52.0                3rd Qu.: 8.000   3rd Qu.:0.9000
##  Max.   :69.0                Max.   :10.000   Max.   :0.9900
##
##  REM.sleep.percentage Deep.sleep.percentage Light.sleep.percentage Awakenings
##  Min.   :15.00        Min.   :18.00         Min.   : 7.00          0: 84
##  1st Qu.:20.00        1st Qu.:47.50         1st Qu.:15.00          1:141
##  Median :22.00        Median :58.00         Median :18.00          2: 48
##  Mean   :22.68        Mean   :52.74         Mean   :24.58          3: 55
##  3rd Qu.:25.00        3rd Qu.:63.00         3rd Qu.:32.50          4: 56
##  Max.   :30.00        Max.   :75.00         Max.   :63.00
##
```

```
##  Caffeine.consumption Alcohol.consumption Smoking.status Exercise.frequency
##  0 :195               0:217               No :251        0  :110
##  25: 73               1: 47               Yes:133        1  : 78
##  50: 97               2: 35                              2  : 44
##  75: 19               3: 39                              3  :110
##                       4: 19                              4-5: 42
##                       5: 27
##
##         dia
##  domingo  :49
##  jueves   :55
##  lunes    :48
##  martes   :47
##  miércoles:59
##  sábado   :63
##  viernes  :63
```

Vamos a analizar la distribución de las variables continuas de forma gráfica.

```
show_plot(inspect_num(df))
```



Podemos ver que la variable que representa el tiempo de sueño no parece ser una variable continua, pues toma un número finito de valores de forma discreta. Crearemos una variable categórica.

```r
# Cut marca los límites superiores, incluidos

df$Sleep.duration <- cut(df$Sleep.duration, c(0,7,8,12),
                  labels = c('Poco','Ok','Mucho'))

table(df$Sleep.duration)
```

```
##
##  Poco    Ok Mucho
##   171   160    53
```

Además, observamos que los porcentajes del tiempo que los individuos pasan en cada fase, pudieran ser variables dependientes. ¿Mantienen una relación lineal? La lógica dice que el 100% de tu tiempo durmiendo, tendrá que ser la suma del porcentaje que pasas en cada fase. Vamos a comprobar si esto se cumple.

```r
summary(data.frame(apply(df[,c(5:7)],1,sum)))
```

```
##  apply.df...c.5.7....1..sum.
##  Min.   :100
##  1st Qu.:100
##  Median :100
##  Mean   :100
##  3rd Qu.:100
##  Max.   :100
```

Efectivamente, todas las sumas son 100%, por lo que las variables son dependientes de forma lineal. Para evitar errores, eliminamos una de las variables. En este caso, eliminamos el porcentaje de sueño profundo, pues me parecen más interesantes las variables de REM y sueño ligero.

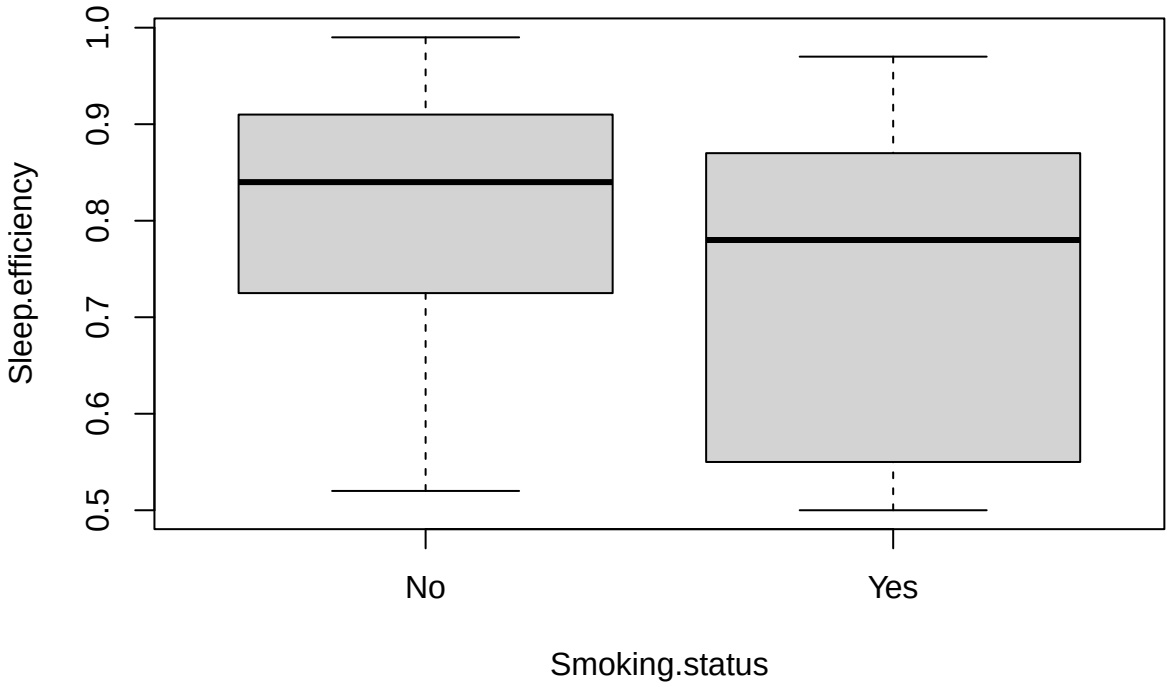
```r
df <- df[,-c(6)]
```

Análisis a priori - influencia de las variables.

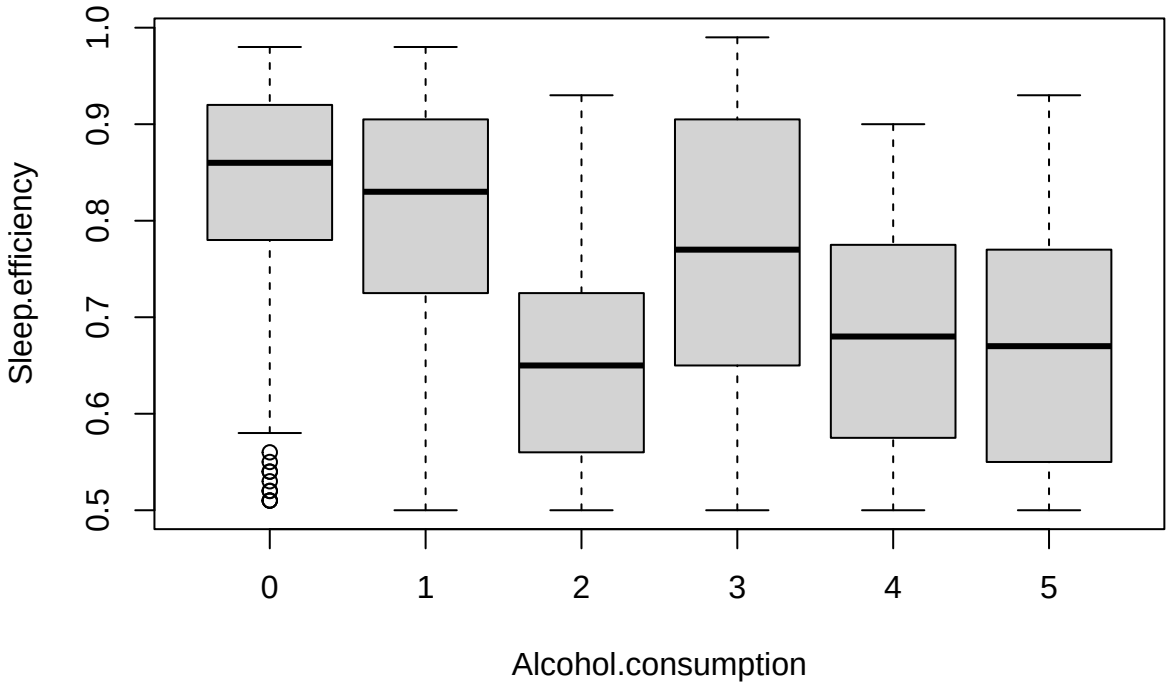
Ahora que concoemos la base de datos y la hemos depurado, vamos a empezar a sacar algunas conclusiones. Por ejemplo, nos interesa ver si la calidad del sueño se ve afectado por el consumo de alcohol o el consumo de tabaco.

Para ello realizo los siguientes gráficos:

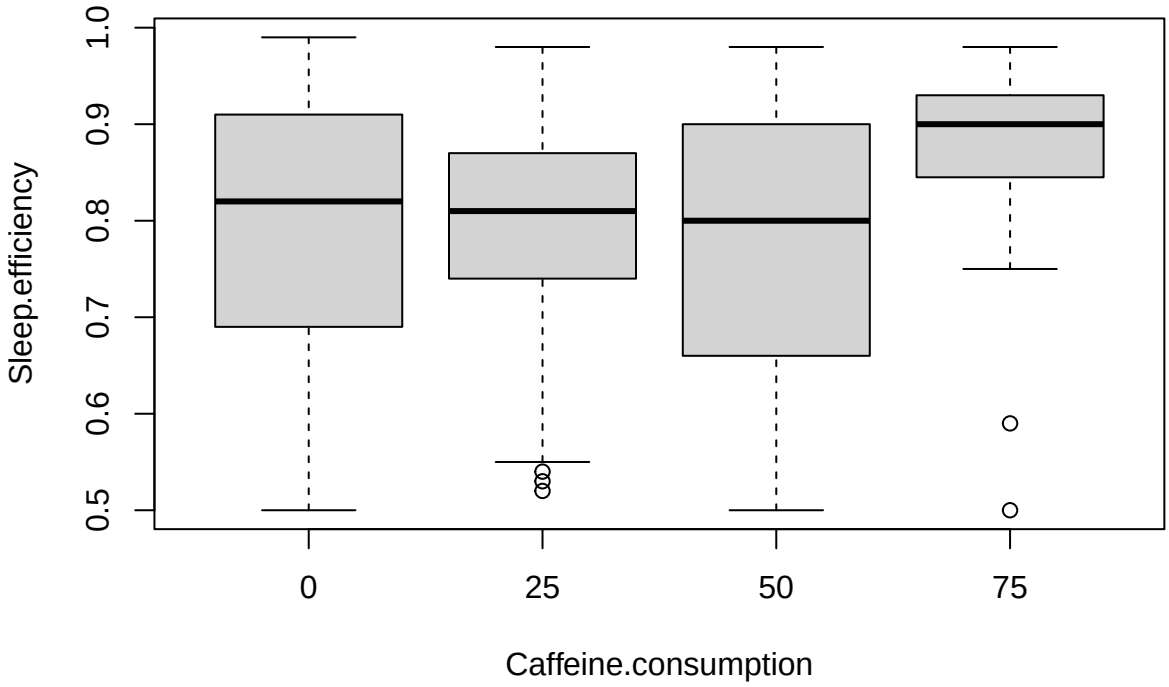
```r
boxplot(Sleep.efficiency ~ Smoking.status, data = df)
```




```
boxplot(Sleep.efficiency ~ Alcohol.consumption, data = df)
```

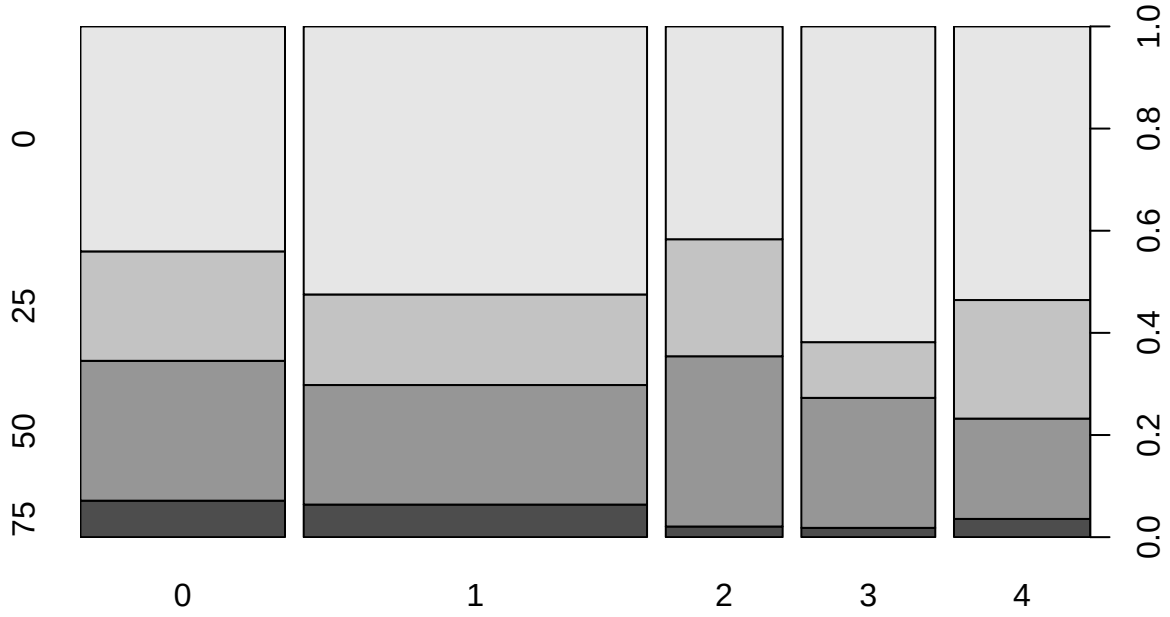



```
boxplot(Sleep.efficiency ~ Caffeine.consumption, data = df)
```



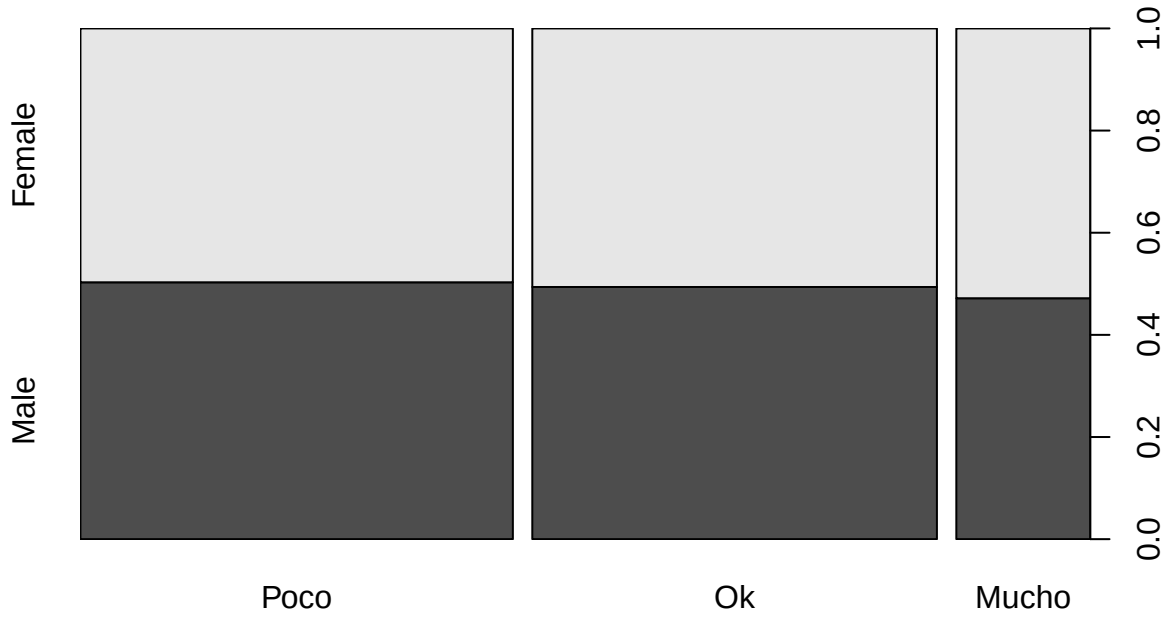
También puede ser interesante cómo es la distribución de las personas en función del consumo de cafeína y el número de veces que se levantan mientras duermen.

```
spineplot(table(df$Awakenings,df$Caffeine.consumption))
```



Vamos a comprobar, de forma gráfica, si el género puede afectar en las horas de sueño. De esta forma, en el futuro, podremos realizar más comprobaciones de forma analítica.

```
spineplot(table(df$Sleep.duration,df$Gender))
```



Intervalo de confianza.

En este apartado crearemos un intervalo de confianza para la calidad de sueño. Hemos visto anteriormente el estimador puntual para este valor, aproximadamente el 79%, pero vamos a crear un intervalo de confianza para este valor y poder estimar cuánto valdría este valor en la población.

Queremos hacer el intervalo de confianza para la calidad del sueño **media**.

```
# Intervalo para la media

conf <- t.test(x =  df$Sleep.efficiency,      # Muestra 1
               y = NULL,                      # Muestra 2
               alternative = c("two.sided"),  # Tipo de intervalo
               paired = FALSE,                # Variables dependientes
               var.equal = FALSE,             # Varianzas iguales
               conf.level = 0.95)             # Nivel de confianza (1-nivel significación)

# Muestro el resultado del intervalo

conf$conf.int
```

```
## [1] 0.7744254 0.8016684
## attr(,"conf.level")
## [1] 0.95
```

Contraste de hipótesis.

Hemos visto de forma gráfica, de dos formas diferentes, que el género no afectaba a la calidad del sueño pero ser fumador sí podía influir.

Crearemos un primer contraste de forma analítica para comprobar que la media de calidad de sueño entre hombres y mujeres es igual. Es decir, queremos plantear el siguiente contraste de hipótesis:

$$\begin{cases} H_0: & \mu_{hombres} = \mu_{mujeres} \\ H_1: & \mu_{hombres} \neq \mu_{mujeres} \end{cases}$$

```
# Contraste para la diferencia de medias

hombres <- df[df$Gender == 'Male', 'Sleep.efficiency']
mujeres <- df[df$Gender == 'Female', 'Sleep.efficiency']

t.test(x = hombres,                    # Muestra 1
       y = mujeres,                    # Muestra 2
       alternative = c("two.sided"),   # Signo hipótesis alternativa
       paired = FALSE,                 # Variables dependientes
       var.equal = FALSE,              # Varianzas iguales
       conf.level = 0.95)              # Nivel de confianza (1-nivel significación)
```

```
##
##  Welch Two Sample t-test
##
## data:  hombres and mujeres
## t = -0.33742, df = 380.32, p-value = 0.736
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.03192836  0.02257513
## sample estimates:
## mean of x mean of y
## 0.7856842 0.7903608
```

Como el p-valor es mayor que el nivel de significación, podemos concluir que no hay evidencias suficientes para rechazar H_0 y por tanto no existen diferencias entre la calidad del sueño de los hombres y las mujeres.

Ahora vamos a comprobar qué pasa con los fumadores. Para ello, vamos a plantear la siguiente hipótesis, pues gráficamente es la que hemos visto que se cumple:

$$\begin{cases} H_0: & \mu_{fumadores} \leq \mu_{nofumadores} \\ H_1: & \mu_{fumadores} > \mu_{nofumadores} \end{cases}$$

```
# Contraste para la diferencia de medias

yes <- df[df$Smoking.status == 'Yes', 'Sleep.efficiency']
no <- df[df$Smoking.status == 'No', 'Sleep.efficiency']

t.test(x = yes,                      # Muestra 1
       y = no,                       # Muestra 2
       alternative = c("greater"),   # Signo hipótesis alternativa
       paired = FALSE,               # Variables dependientes
       var.equal = FALSE,            # Varianzas iguales
       conf.level = 0.99)            # Nivel de confianza (1-nivel significación)
```

```
##
##  Welch Two Sample t-test
##
## data:  yes and no
## t = -5.2905, df = 201.32, p-value = 1
## alternative hypothesis: true difference in means is greater than 0
## 99 percent confidence interval:
##  -0.1187049        Inf
## sample estimates:
## mean of x mean of y
## 0.7342857 0.8165339
```

El p-valor es muy grande, lo que quiere decir que nuestro estadístico de diferencia de medias está muy lejos de la región crítica, por lo que no tenemos evidencias suficientes para rechazar la hipótesis nula y podemos concluir que sí hay diferencia entre la calidad del sueño de los fumadores y no fumadores, siendo la calidad de los primeros menor.

ANOVA.

Por último, vamos a comprobar si el consumo de alcohol es un factor significativo para la calidad del sueño. Gráficamente hemos visto que a medida que aumenta el consumo, la calidad es menor.

Definimos el siguiente contraste:

$$\begin{cases} H_0: & \mu_{c0} = \mu_{c1} = \mu_{c2} = \mu_{c3} = \mu_{c4-5} \\ H_1: & Diferentes \end{cases}$$

```
# Aplicamos ANOVA
anova <- aov(Sleep.efficiency ~ Alcohol.consumption, data = df)
summary(anova)
```

```
##                      Df Sum Sq Mean Sq F value Pr(>F)
## Alcohol.consumption   5  1.448 0.28956   19.51 <2e-16 ***
## Residuals           378  5.611 0.01484
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Como resultado, obtenemos la tabla de ANOVA vista en clase, con la diferencia de encontrar el p-valor en lugar del valor teórico para la distribución de la F.

El criterio de rechazo es el mismo que tenemos en los contrastes de hipótesis, es decir, si el p-valor es mayor que el nivel de significación, rechazamos H_0.

En este caso tenemos un p-valor muy pequeño, bajo un nivel de significación 0,05, rechazamos H_0, es decir, no hay evidencias suficientes para aceptar que las medias de la calidad del sueño en los diferentes grupos según el consumo de alcohol sean iguales. Podemos concluir que estas medias son diferentes y por tanto, el consumo de alcohol sí es un factor influyente en la calidad del sueño.

Gráficamente habíamos comprobado que los factores género y tiempo de sueño no tienen relación entre ellos, pero comprobaremos si son influyentes en la calidad del sueño. Además, comprobaremos si la interacción de ambos factores es influyente.

```r
an <- aov(Sleep.efficiency ~ Sleep.duration*Gender, data = df)
summary(an)
```

```
##                          Df Sum Sq Mean Sq F value Pr(>F)
## Sleep.duration            2  0.076 0.03801   2.071  0.128
## Gender                    1  0.002 0.00233   0.127  0.722
## Sleep.duration:Gender     2  0.042 0.02085   1.136  0.322
## Residuals               378  6.939 0.01836
```

Todos los p-valores son mayores que el nivel de significación (0,05), por lo que, de forma alítica, podemos concluir que ni los factores género y tiempo de sueño ni su interacción afectan a la calidad del sueño, pues la media de todos los tratamientos es la misma (contraste que hemos planteado).