Estimación de niveles de obesidad con base en los hábitos alimenticios y el estado físico

7 de Enero de 2024

4. **¿Existen diferencias significativas en los hábitos alimenticios (ej. consumo de vegetales `FCVC`, consumo de alimentos entre comidas `CAEC`) entre los diferentes niveles de obesidad?** Buscamos comparar los patrones de alimentación entre los distintos grupos de obesidad, para identificar posibles factores de riesgo o hábitos asociados a cada nivel.

5. **¿Es posible construir un modelo predictivo que clasifique con precisión el nivel de obesidad de los individuos a partir de variables como la edad (`Age`), el género (`Gender`), los hábitos alimenticios y la actividad física?** Nos centraremos en la capacidad de predecir la obesidad utilizando un modelo estadístico o de aprendizaje automático, con base en las variables disponibles en el conjunto de datos.

Visualizamos el conjunto de datos unva vez cargado:

```
##     Gender                Age             Height          Weight
##  Length:2111        Min.   :14.00   Min.   :1.450   Min.   : 39.00
##  Class :character   1st Qu.:19.95   1st Qu.:1.630   1st Qu.: 65.47
##  Mode  :character   Median :22.78   Median :1.700   Median : 83.00
##                     Mean   :24.31   Mean   :1.702   Mean   : 86.59
##                     3rd Qu.:26.00   3rd Qu.:1.768   3rd Qu.:107.43
##                     Max.   :61.00   Max.   :1.980   Max.   :173.00
##  family_history_with_overweight     FAVC                FCVC
##  Length:2111                    Length:2111        Min.   :1.000
##  Class :character               Class :character   1st Qu.:2.000
##  Mode  :character               Mode  :character   Median :2.386
##                                                    Mean   :2.419
##                                                    3rd Qu.:3.000
##                                                    Max.   :3.000
##       NCP            CAEC              SMOKE               CH2O
##  Min.   :1.000   Length:2111        Length:2111        Min.   :1.000
##  1st Qu.:2.659   Class :character   Class :character   1st Qu.:1.585
##  Median :3.000   Mode  :character   Mode  :character   Median :2.000
##  Mean   :2.686                                         Mean   :2.008
##  3rd Qu.:3.000                                         3rd Qu.:2.477
##  Max.   :4.000                                         Max.   :3.000
##      SCC                 FAF              TUE             CALC
##  Length:2111        Min.   :0.0000   Min.   :0.0000   Length:2111
##  Class :character   1st Qu.:0.1245   1st Qu.:0.0000   Class :character
##  Mode  :character   Median :1.0000   Median :0.6253   Mode  :character
##                     Mean   :1.0103   Mean   :0.6579
##                     3rd Qu.:1.6667   3rd Qu.:1.0000
##                     Max.   :3.0000   Max.   :2.0000
##     MTRANS           NObeyesdad
##  Length:2111        Length:2111
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

Y confirmamos que efectivamente, tenemos 2111 instancias y 17 atributos sin valores perdidos. Con el objetivo de simular un escenario real de datos incompletos y poner en práctica diferentes métodos de limpieza, se introducirán valores perdidos de forma aleatoria en el dataset original.
Utilizaremos la función `introduce_na` en R que permitirá especificar el porcentaje de valores perdidos máximo a introducir.
Se optó por introducir un 5% como máximo de valores perdidos de forma aleatoria en todas las variables del dataset **mediante una funcion**. Este proceso de "ensuciamiento" asegura que los valores perdidos se

distribuyan aleatoriamente en el conjunto de datos, simulando un escenario realista donde la ausencia de datos no depende de ningún patrón específico

Introdujimos un máximo de un 5% de valores perdidos con el siguiente código creando un dataset `dataset_dirt.csv`:

Recargamos y revisamos el estado del conjunto de datos:

```
##     Gender               Age            Height          Weight
##  Length:2111        Min.   :14.00   Min.   :1.456   Min.   : 39.00
##  Class :character   1st Qu.:19.91   1st Qu.:1.630   1st Qu.: 65.06
##  Mode  :character   Median :22.77   Median :1.701   Median : 82.95
##                     Mean   :24.29   Mean   :1.702   Mean   : 86.61
##                     3rd Qu.:26.00   3rd Qu.:1.769   3rd Qu.:107.56
##                     Max.   :61.00   Max.   :1.980   Max.   :165.06
##                     NA's   :106     NA's   :106     NA's   :106
##  family_history_with_overweight     FAVC                FCVC
##  Length:2111                    Length:2111        Min.   :1.000
##  Class :character               Class :character   1st Qu.:2.000
##  Mode  :character               Mode  :character   Median :2.375
##                                                    Mean   :2.418
##                                                    3rd Qu.:3.000
##                                                    Max.   :3.000
##                                                    NA's   :42
##       NCP            CAEC              SMOKE                CH2O
##  Min.   :1.000   Length:2111        Length:2111        Min.   :1.000
##  1st Qu.:2.673   Class :character   Class :character   1st Qu.:1.582
##  Median :3.000   Mode  :character   Mode  :character   Median :2.000
##  Mean   :2.689                                         Mean   :2.009
##  3rd Qu.:3.000                                         3rd Qu.:2.482
##  Max.   :4.000                                         Max.   :3.000
##  NA's   :84                                            NA's   :21
##      SCC                 FAF              TUE             CALC
##  Length:2111        Min.   :0.0000   Min.   :0.0000   Length:2111
##  Class :character   1st Qu.:0.1196   1st Qu.:0.0000   Class :character
##  Mode  :character   Median :1.0000   Median :0.6186   Mode  :character
##                     Mean   :1.0057   Mean   :0.6516
##                     3rd Qu.:1.6587   3rd Qu.:1.0000
##                     Max.   :3.0000   Max.   :2.0000
##                     NA's   :42       NA's   :63
##     MTRANS           NObeyesdad
##  Length:2111        Length:2111
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
##
```

Analizamos como ha resultado la perdida de datos inducida:

```
##                                Num_NA Blancos
## Gender                              0      21
## Age                               106      NA
## Height                            106      NA
## Weight                            106      NA
```

```
## family_history_with_overweight      0      63
## FAVC                                 0      21
## FCVC                                42      NA
## NCP                                 84      NA
## CAEC                                 0     106
## SMOKE                                0     106
## CH2O                                21      NA
## SCC                                  0      63
## FAF                                 42      NA
## TUE                                 63      NA
## CALC                                 0      21
## MTRANS                               0      84
## NObeyesdad                           0      63
```

```
##                                 Num_NA Blancos Porcentaje_NA Porcentaje_Blancos
## Gender                               0      21          0.00                0.99
## Age                                106      NA          5.02                  NA
## Height                             106      NA          5.02                  NA
## Weight                             106      NA          5.02                  NA
## family_history_with_overweight       0      63          0.00                2.98
## FAVC                                 0      21          0.00                0.99
## FCVC                                42      NA          1.99                  NA
## NCP                                 84      NA          3.98                  NA
## CAEC                                 0     106          0.00                5.02
## SMOKE                                0     106          0.00                5.02
## CH2O                                21      NA          0.99                  NA
## SCC                                  0      63          0.00                2.98
## FAF                                 42      NA          1.99                  NA
## TUE                                 63      NA          2.98                  NA
## CALC                                 0      21          0.00                0.99
## MTRANS                               0      84          0.00                3.98
## NObeyesdad                           0      63          0.00                2.98
```

Importancia y problema a resolver:

La obesidad se ha convertido en una preocupante epidemia global, con graves consecuencias para la salud individual y un alto costo económico para los sistemas de salud. Este *dataset* nos brinda la oportunidad de explorar la compleja relación entre los hábitos de vida y la obesidad, con el objetivo de identificar factores de riesgo y contribuir al desarrollo de estrategias de prevención y tratamiento más efectivas.

Preguntas de investigación:

En este trabajo buscamos responder a la siguiente pregunta central: ***¿Cómo se relacionan los hábitos alimenticios y la condición física con los diferentes niveles de obesidad?***

Para abordar esta cuestión, nos planteamos las siguientes preguntas específicas:

1. ¿Cuáles son los hábitos alimenticios que mejor predicen el desarrollo de la obesidad?
2. ¿Existe una interacción entre la condición física y los hábitos alimenticios en la determinación del nivel de obesidad?
3. ¿Se observan diferencias significativas en los hábitos alimenticios y la condición física entre los distintos niveles de obesidad?

Variables:

El dataset contiene 17 variables o instancias descriptoras de las características de los individuos, incluyendo hábitos alimenticios, condición física y datos demográficos.

A continuación realizamos una descripción detallada de cada variable:

1. Variables relacionadas con los hábitos alimenticios:

- FAVC (Frequent consumption of high caloric food): Indica si la persona consume frecuentemente alimentos altos en calorías (Sí/No).
- FCVC (Frequency of consumption of vegetables): Frecuencia de consumo de verduras (1: nunca, 2: algunas veces por mes, 3: una vez por semana, 4: 2-4 veces por semana, 5: 5-6 veces por semana, 6: todos los días).
- NCP (Number of main meals): Número de comidas principales al día (1-3).
- CAEC (Consumption of food between meals): Frecuencia de consumo de alimentos entre comidas (1: No, 2: Algunas veces, 3: Frecuentemente, 4: Siempre).
- CH2O (Consumption of water daily): Cantidad de agua consumida diariamente (1: menos de un litro, 2: entre 1 y 2 litros, 3: más de 2 litros).
- SCC (Calories consumption monitoring): Indica si la persona monitorea su consumo de calorías (Sí/No).
- FAF (Physical activity frequency): Frecuencia de actividad física (0: nunca, 1: 1-2 días a la semana, 2: 3-4 días a la semana, 3: 5-7 días a la semana).
- TUE (Time using technology devices): Tiempo diario dedicado al uso de dispositivos tecnológicos (0: 0-2 horas, 1: 3-5 horas, 2: más de 5 horas).

2. Variables relacionadas con la condición física:

- CALC (Consumption of alcohol): Frecuencia de consumo de alcohol (1: No consumo, 2: Algunas veces a la semana, 3: Todos los días).
- MTRANS (Transportation used): Medio de transporte habitual (Automóvil, Motocicleta, Bicicleta, Transporte público, Caminando).

3. Datos demográficos:

- Gender: Género del individuo (Hombre/Mujer).
- Age: Edad del individuo en años.
- Height: Altura del individuo en metros.
- Weight: Peso del individuo en kilogramos.
- family_history_with_overweight: Indica si existen antecedentes familiares de sobrepeso (Sí/No).
- SMOKE: Indica si la persona fuma (Sí/No).

4. Variable objetivo:

- NObeyesdad: Nivel de obesidad del individuo cuyos valores se corresponden al indice de masa corporal, cuya formula se corresponde con:

$$IMC = \frac{Peso}{Altura^2}$$

y que clasificamos consecuentemente como:

- Underweight Less than 18.5
- Normal 18.5 to 24.9
- Overweight 25.0 to 29.9
- Obesity I 30.0 to 34.9
- Obesity II 35.0 to 39.9
- Obesity III Higher than 4

Vamos a factorizar las variables categóricas que están codificadas como números, denotando en este punto que existe un problema con la variable NCP que abordaremos posteriormente:

```
##     Gender               Age            Height          Weight
##  Length:2111        Min.   :14.00   Min.   :1.456   Min.   : 39.00
##  Class :character   1st Qu.:19.91   1st Qu.:1.630   1st Qu.: 65.06
##  Mode  :character   Median :22.77   Median :1.701   Median : 82.95
##                     Mean   :24.29   Mean   :1.702   Mean   : 86.61
##                     3rd Qu.:26.00   3rd Qu.:1.769   3rd Qu.:107.56
```

```
##                    Max.   :61.00   Max.   :1.980   Max.   :165.06
##                    NA's   :106     NA's   :106     NA's   :106
##  family_history_with_overweight     FAVC               FCVC
##  Length:2111                    Length:2111        Nunca  : 198
##  Class :character               Class :character   A veces:1233
##  Mode  :character               Mode  :character   Siempre: 638
##                                                    NA's   :  42
##
##
##
##        NCP           CAEC              SMOKE             CH2O
##  Ninguna : 377   Length:2111        Length:2111        1L  : 762
##  Una-dos : 276   Class :character   Class :character   1-2L:1166
##  Tres    :1307   Mode  :character   Mode  :character   2+L : 162
##  Mas de 3:  67                                         NA's:  21
##  NA's    :  84
##
##
##      SCC                 FAF           TUE           CALC
##  Length:2111        Sin AF  :994   0-2 hrs:1382   Length:2111
##  Class :character   1-2 días:711   3-5 hrs: 566   Class :character
##  Mode  :character   2-4 días:292   5hrs + : 100   Mode  :character
##                     4-5 días: 72   NA's   :  63
##                     NA's    : 42
##
##
##     MTRANS           NObeyesdad
##  Length:2111        Length:2111
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
##
```

Y observamos que los cambios discurrieron sin problema, teniendo en este punto a todas las variables categoricas con orden implícito, factorizadas. Por otro lado, el atributo 'CALC' es ordinal, es decir,las categorías tienen un orden implícito, aspecto que tendremos en cuenta al realizar el análisis exploratorio de datos. En cuanto a otras variables categóricas como *"family_history_with_overweight," "NObeyesdad," "SMOKE," "MTRANS," "CAEC," "FAVC," "SCC" y "Gender"* compuestas por cadenas de texto, las iremos abordando en cuanto a su factorización en cada paso.

Tamaño:

El dataset como ya observamos anteriormente, cuenta con más de 2000 instancias, proporcionando una muestra respetable al objeto de realizar análisis estadísticos robustos y poder obtener resultados significativos.

2. Selección e integración

En esta investigación, emplearemos el conjunto de datos *"Estimation of Obesity Levels Based On Eating Habits and Physical Condition"* de la UCI Machine Learning Repository.
Hemos optado por no realizar una selección inicial de atributos ni integrar datos externos en esta etapa del análisis.
A continuación, detallamos las razones que justifican esta decisión:

Información variada y representativa

El dataset ofrece una muestra representativa con información sobre individuos de ambos géneros, con diversidad en hábitos alimenticios, niveles de actividad física y datos demográficos. Esto permite un análisis completo de la relación entre los factores de riesgo y la obesidad, considerando las posibles diferencias entre hombres y mujeres.

Poder estadístico
El tamaño muestral (N=2111) es suficientemente grande para garantizar un poder estadístico adecuado en los análisis.

A modo de ejemplo, se realizaron cálculos de tamaño muestral para diferentes pruebas estadísticas:

Prueba T de dos muestras independientes:
Para comparar hombres y mujeres, con un nivel de significancia del 1% y una potencia del 95%, se necesitaría un tamaño muestral de 225 por grupo. Nuestro dataset cuenta con más de 1000 individuos por género.

```
##
##      Two-sample t test power calculation
##
##               n = 224.3426
##               d = 0.4
##       sig.level = 0.01
##           power = 0.95
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

ANOVA de un factor:
Para comparar los 6 niveles de obesidad, con un nivel de significancia del 1% y una potencia del 95%, el tamaño muestral requerido por grupo es mucho menor al disponible en nuestro dataset, como se puede observar.

```
##
##      Balanced one-way analysis of variance power calculation
##
##               k = 6
##               n = 53.6784
##               f = 0.25
##       sig.level = 0.05
##           power = 0.95
##
## NOTE: n is number in each group
```

Regresión lineal:

Con un nivel de significancia del 1%, una potencia del 95%, 2 predictores y un tamaño del efecto pequeño, se necesitaría un tamaño muestral de 176. Nuestro dataset supera ampliamente este número.

```
##
##      Multiple regression power calculation
##
##               u = 2
##               v = 175.9183
##              f2 = 0.1
##       sig.level = 0.01
##           power = 0.9
```

Y nuevamente se requiere un tamaño muestral (176) muy inferior al numero de observaciones recogidas en nuestro conjunto de datos. Estos cálculos confirman que el conjunto de datos es lo suficientemente grande para realizar análisis robustos.

Alineación con los objetivos
El objetivo principal de la investigación es analizar la influencia de los hábitos alimenticios y la condición física en los niveles de obesidad. Utilizar la totalidad de los datos disponibles nos permite abordar este objetivo de manera efectiva.

Potencial para análisis de subgrupos
En etapas posteriores de la investigación, realizaremos análisis de subgrupos para profundizar en la comprensión de la obesidad.

Algunos subgrupos de interés son:

- Comparación entre hombres y mujeres: Analizar si existen diferencias en los factores de riesgo y la prevalencia de obesidad entre géneros.
- Grupos de edad: Investigar cómo varían los hábitos alimenticios y la condición física en relación con la obesidad en diferentes grupos de edad.

Integración de datos
Aunque en este estudio no integraremos datos externos, en el futuro podríamos considerar la incorporación de información adicional teniendo en cuenta los contextos socioculturales de los países de procedencia de la muestra - Colombia, Peru y Mexico, como pudiera ser:

- Datos socioeconómicos: Nivel de ingresos, nivel educativo, acceso a servicios de salud.
- Datos geográficos: Lugar de residencia, acceso a espacios verdes, densidad de población.
- Datos climáticos: Temperatura, precipitaciones, horas de sol.

Estos datos podrían enriquecer el análisis y proporcionar una comprensión más completa de los factores que influyen en la obesidad pero, en todo caso no hay ningún atributo que nos permitiera inferir la procedencia de cada una de las muestras. En todo caso y por todo lo detallado en este apartado, podemos considerar el *dataset* apropiado y suficiente a juzgar por nuestras preguntas de investigación.

3. Limpieza

Aunque originalmente el conjunto de datos proporcionado por los creadores del mismo no contenía valores perdidos, hemos alterado el contenido como pudimos ver en el apartado 1.

Comenzamos verificando la estructura del dataset:

```
## [1] "Visual de la estructura con 'STR':"
```

```
## 'data.frame':    2111 obs. of  17 variables:
##  $ Gender                         : chr  "Female" "Female" "Male" "Male" ...
##  $ Age                            : num  21 21 23 27 22 29 23 22 24 22 ...
##  $ Height                         : num  1.62 1.52 1.8 1.8 1.78 1.62 1.5 1.64 1.78 1.72 ...
##  $ Weight                         : num  64 56 77 87 89.8 53 55 53 64 68 ...
##  $ family_history_with_overweight: chr  "yes" "yes" "yes" "no" ...
##  $ FAVC                           : chr  "no" "no" "no" "no" ...
##  $ FCVC                           : Factor w/ 3 levels "Nunca","A veces",..: 2 3 2 3 2 2 3 2 3 2 ...
##  $ NCP                            : Factor w/ 4 levels "Ninguna","Una-dos",..: 3 3 3 3 1 3 3 3 3 3 ..
##  $ CAEC                           : chr  "Sometimes" "Sometimes" "Sometimes" "Sometimes" ...
##  $ SMOKE                          : chr  "no" "yes" "no" "no" ...
##  $ CH2O                           : Factor w/ 3 levels "1L","1-2L","2+L": 2 3 2 2 2 2 2 2 2 2 ...
##  $ SCC                            : chr  "no" "yes" "" "no" ...
##  $ FAF                            : Factor w/ 4 levels "Sin AF","1-2 días",..: 1 4 3 3 1 1 2 4 2 2 ..
##  $ TUE                            : Factor w/ 3 levels "0-2 hrs","3-5 hrs",..: NA 1 2 1 1 1 1 1 2 NA
##  $ CALC                           : chr  "no" "Sometimes" "Frequently" "Frequently" ...
##  $ MTRANS                         : chr  "Public_Transportation" "Public_Transportation" "Public_Trans
##  $ NObeyesdad                     : chr  "Normal_Weight" "Normal_Weight" "Normal_Weight" "Overweight_
```

3.1 Perdida de datos en el dataset

Anteriormente obtuvimos los valores perdidos y valores en blanco:

	Num_NA	Blancos	Porcentaje_NA	Porcentaje_Blancos
Gender	0	21	0.00	0.99
Age	106	NA	5.02	NA
Height	106	NA	5.02	NA
Weight	106	NA	5.02	NA
family_history_with_overweight	0	63	0.00	2.98
FAVC	0	21	0.00	0.99
FCVC	42	NA	1.99	NA
NCP	84	NA	3.98	NA
CAEC	0	106	0.00	5.02
SMOKE	0	106	0.00	5.02
CH2O	21	NA	0.99	NA
SCC	0	63	0.00	2.98
FAF	42	NA	1.99	NA
TUE	63	NA	2.98	NA
CALC	0	21	0.00	0.99
MTRANS	0	84	0.00	3.98
NObeyesdad	0	63	0.00	2.98

> Nota: `colSums(df == "")` no pudo calcular la suma de celdas vacías para algunas variables debido a que esas variables no son de tipo carácter, lo que significa que no pueden contener cadenas vacías (”“) como valores perdidos pero que a los efectos no supone mayor problema. Cabe en este punto hacer la distinción de que los atributos numéricos contienen ‘NA’ y los atributos categóricos con cadenas de texto contendían valores en blanco o””, como puede observarse en el caso de ‘FAVC’ por citar un ejemplo.

Procedemos a trabajar con los valores perdidos variable a variable:

3.2 Imputación de variables

Height La variable ‘Altura’ presenta 63 valores ‘NA.’ Queremos por conocer si la variable tiene una distribución normal, evaluando si la hipótesis nula de que los datos provienen de una distribución normal. Emplearemos la prueba de *Shapiro-Wilk*:

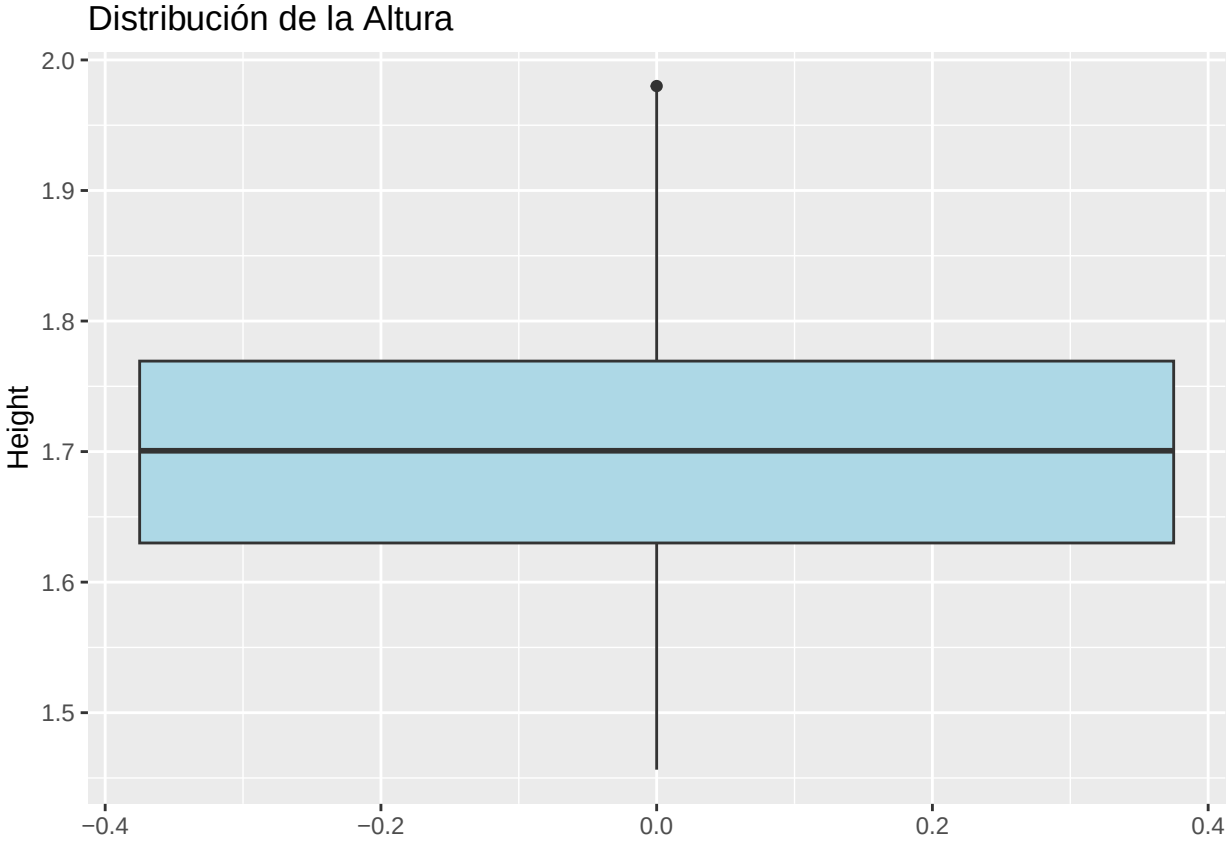
```
##
##  Shapiro-Wilk normality test
##
## data:  df$Height
## W = 0.99278, p-value = 2.275e-08
```

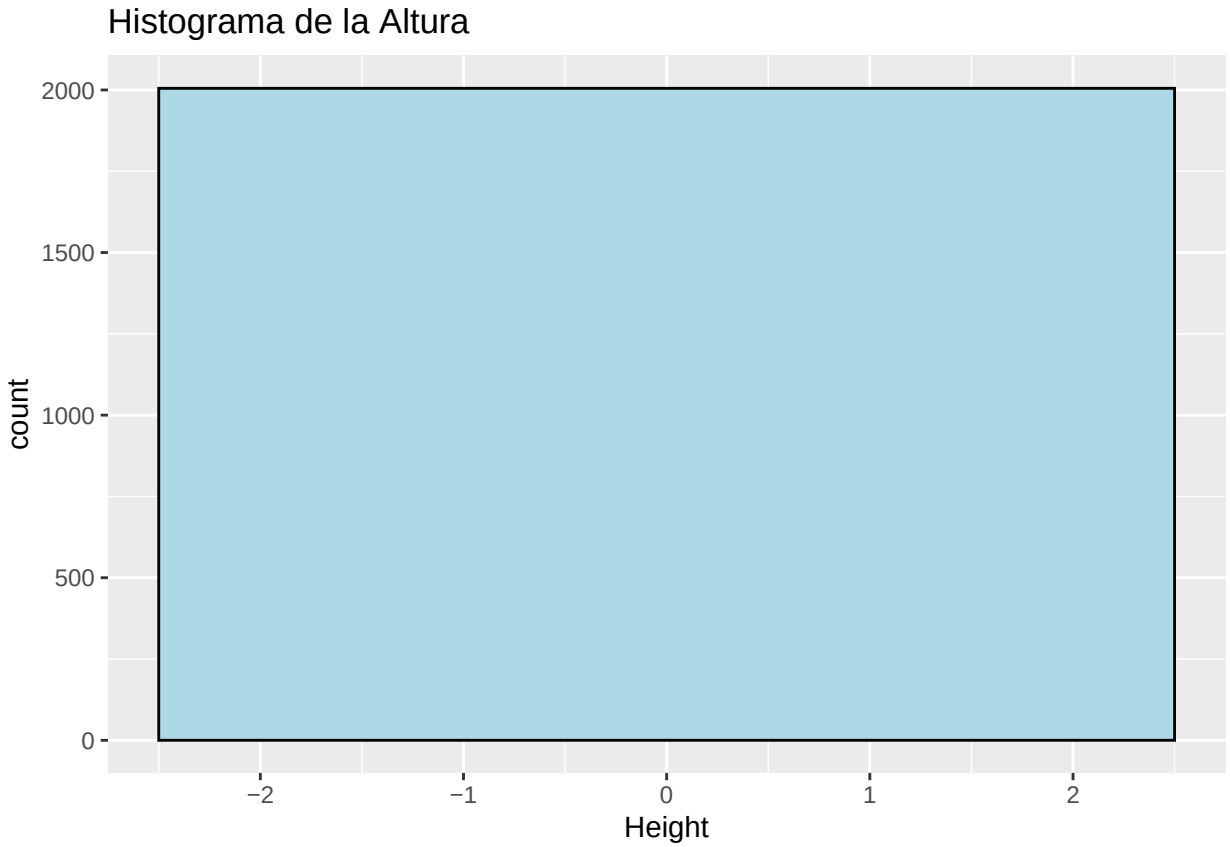
Evaluamos la hipótesis de que la variable “Height” (altura) no siguiera una distribución normal. Los resultados indican que está muy cerca de serlo. Obtuvimos un estadístico $\mathbf{W = 0.99278}$ y un valor $\mathbf{p = 2.275e\text{-}08}$.

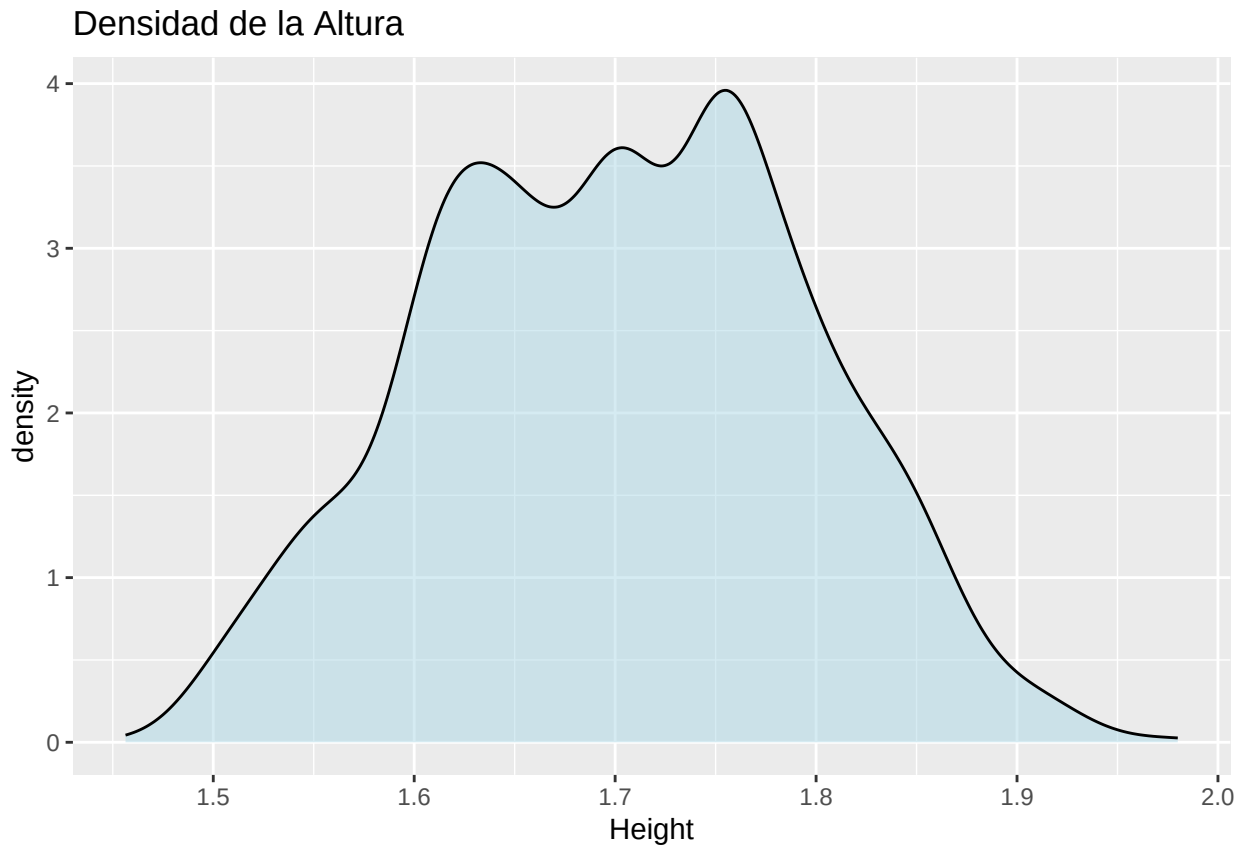
A pesar de que el valor p es bajo, no se puede rechazar la hipótesis nula de normalidad si consideramos un nivel de significancia convencional de 0.05. No hay suficiente evidencia para afirmar que la distribución de la altura en la muestra se desvía significativamente de una distribución normal aunque tampoco matemáticamente es normal

La prueba es sensible al tamaño de la muestra. En muestras grandes, incluso pequeñas desviaciones de la normalidad pueden resultar en un valor p significativo. Por lo tanto, este análisis se complementará con la visualización de la distribución de `Height` mediante histogramas y gráficos de densidad para obtener una comprensión más completa de su forma y características.

En caso de que el análisis visual confirme una distribución aproximadamente normal, la imputación a la media podría ser una estrategia adecuada para los valores faltantes en `Height`. En caso contrario, consideraremos alternativas como la imputación por la mediana o la transformación de la variable. Esto permitirá tomar decisiones informadas sobre el método de imputación más adecuado para la variable, asegurando la integridad y la validez de los datos para los análisis posteriores.




Histograma de la Altura
2000
1500
1000
500
0
count
−2
−1
0
1
2
Height



Las gráficas sobre la variable `Height` nos dicen sobre la distribución:

- **Gráfico de densidad:** Muestra una forma general que se asemeja a una campana, lo que sugiere una posible distribución normal. No obstante, se observan algunos picos que indican cierta multimodalidad, la posible presencia de subpoblaciones con alturas similares.
- **Diagrama de caja:** Confirma la presencia de un valor atípico con una altura considerablemente mayor al resto de los datos. Este valor atípico puede ser la causa de la no normalidad detectada por la prueba de *Shapiro-Wilk*, ya que influye en la media y la desviación estándar de la distribución. Además, la caja del diagrama muestra una distribución bastante simétrica alrededor de la mediana, con una ligera concentración de valores en la parte superior de la caja.

Implicaciones:

- En general la distribución es cercana a la normal, la imputación a la media podría ser una opción a considerar. Sin embargo, la presencia del valor/es atípico/s y los picos en la distribución podrían afectar la precisión de la imputación.
- Consideraremos la imputación por la mediana.
- Analizaremos el valor atípico con mayor detalle para determinar si se trata de un error de medición o de un dato real que requiere un tratamiento especial.

Buscamos valores atipicos:

```
##      Gender      Age   Height   Weight family_history_with_overweight FAVC
## 350    Male 20.00000 1.980000 125.0000                             yes  yes
## 1351   Male 20.49148 1.975663 120.7029                             yes  yes
##         FCVC  NCP      CAEC SMOKE CH2O SCC      FAF     TUE      CALC
## 350  A veces Tres    Always        2+L  no 1-2 días 3-5 hrs Sometimes
## 1351 A veces Tres Sometimes        2+L  no   Sin AF 3-5 hrs Sometimes
```

```
##                       MTRANS     NObeyesdad
## 350  Public_Transportation Obesity_Type_I
## 1351 Public_Transportation Obesity_Type_I
```

Hemos buscado valores extremos o 'outliers' y en este caso observamos que están en un rango que puede considerarse normal. No sería recomendable eliminar los valores atípicos, ya que representan datos reales y no errores de medición.

Dado que la prueba de *Shapiro-Wilk* ha mostrado que la variable no sigue una distribución normal y que además hemos identificado un valor atípico en el diagrama de caja, imputar por la mediana parece la decisión correcta y adecuada.

La mediana es una medida de tendencia central solida que no se ve afectada por valores extremos o asimetrías en la distribución. En este caso, la mediana proporcionará una estimación más representativa de la altura para los valores faltantes en comparación con la media, que podría verse influenciada por el valor atípico.

Procedemos a imputar a la mediana:

Y verificamos:

```
## [1] "No hay valores perdidos en la columna 'Height'"
```

Weight La variable peso es un atributo muy importante dada la naturaleza de este trabajo. Como en las variables que hemos trabajado anteriormente, vamos a proceder al análisis de como se distribuyen los datos:

Emplearemos nuevamente la prueba de *Shapiro-Wilk*:

```
##
##  Shapiro-Wilk normality test
##
## data:  df$Weight
## W = 0.97573, p-value < 2.2e-16
```

Evaluamos la hipótesis de que la variable `Weight` (peso) siguiera una distribución normal. Los resultados de la prueba arrojaron un estadístico $\mathbf{W = 0.97573}$ y un valor $\mathbf{p < 2.2e\text{-}16}$.

El valor p, que representa la probabilidad de obtener un estadístico W igual o menor al observado si la distribución fuera normal, es extremadamente bajo. Rechazamos la hipótesis nula y concluimos que la distribución del peso en la muestra no se ajusta a una distribución normal.

Esta desviación de la normalidad puede influir en la elección de métodos de imputación para los valores faltantes en la variable. La imputación a la media, que asume una distribución normal podría no ser la estrategia más adecuada en este caso. Consideramos alternativas para garantizar una imputación precisa y evitar la introducción de sesgos en los datos.

El análisis se complementará con la visualización de la distribución de `Weight` mediante histogramas y gráficos de densidad para obtener una comprensión más completa de su forma y características lo permitirá tomar decisiones informadas sobre el método de imputación más adecuado para la variable "Weight," asegurando la integridad y la validez de los datos para los análisis posteriores.

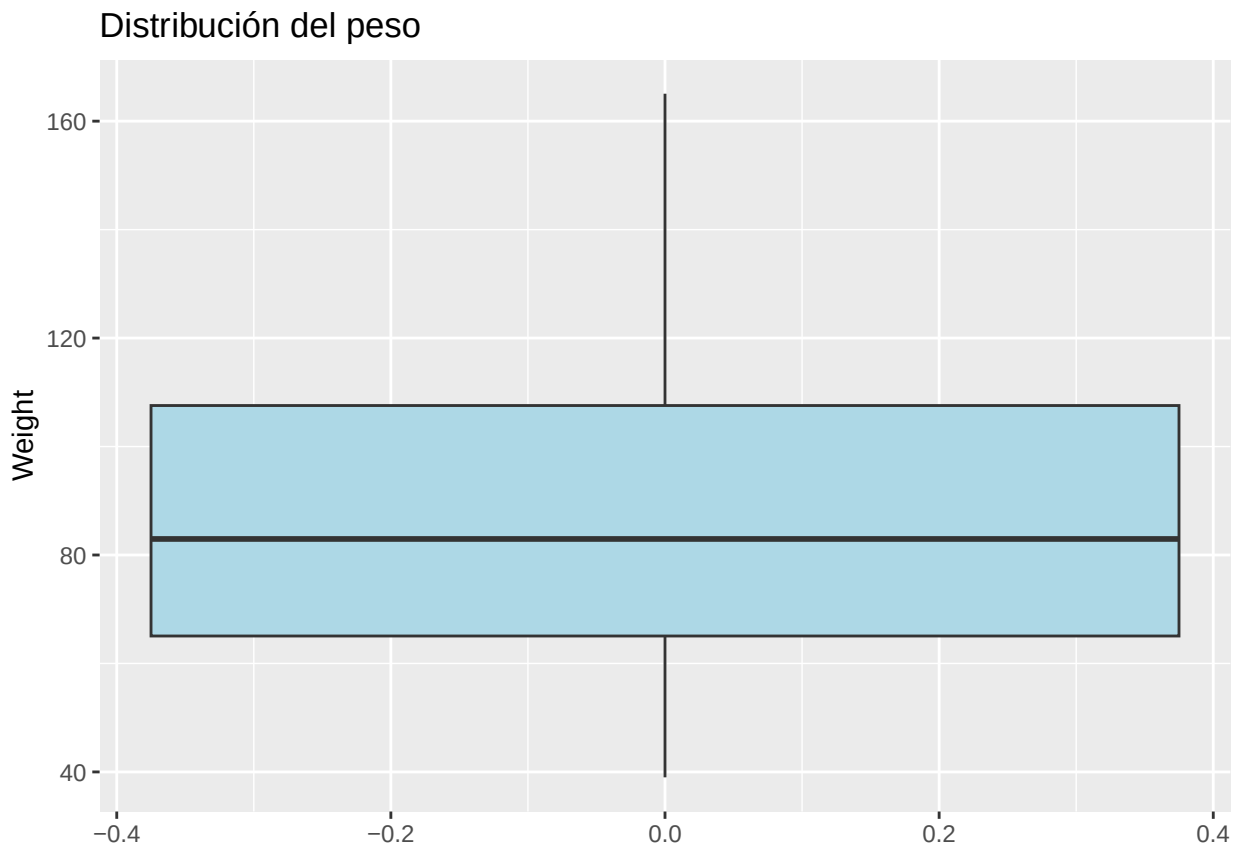
Buscamos outliers como en el caso anterior:

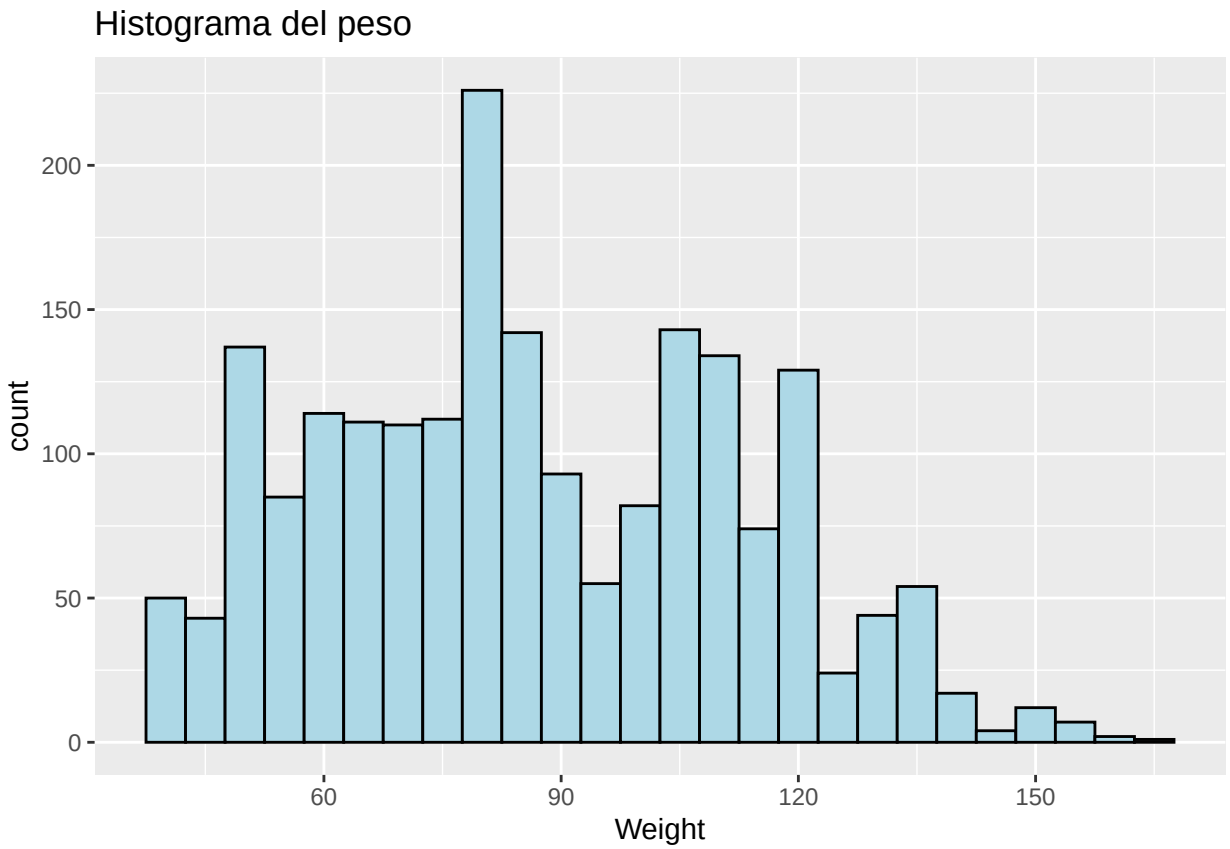
```
##      Gender      Age   Height   Weight family_history_with_overweight FAVC
## 503  Female 21.90012 1.843419 165.0573                              yes  yes
## 1899 Female       NA 1.793824 160.9354                              yes  yes
## 1911 Female 21.52129 1.803677 160.6394                                   yes
##         FCVC  NCP      CAEC SMOKE CH2O SCC      FAF     TUE      CALC
## 503  Siempre <NA> Sometimes    no 1-2L  no   Sin AF 0-2 hrs Sometimes
## 1899 Siempre Tres Sometimes    no 1-2L  no 1-2 días 0-2 hrs Sometimes
## 1911 Siempre Tres Sometimes    no 1-2L  no   Sin AF 0-2 hrs Sometimes
##                      MTRANS       NObeyesdad
```

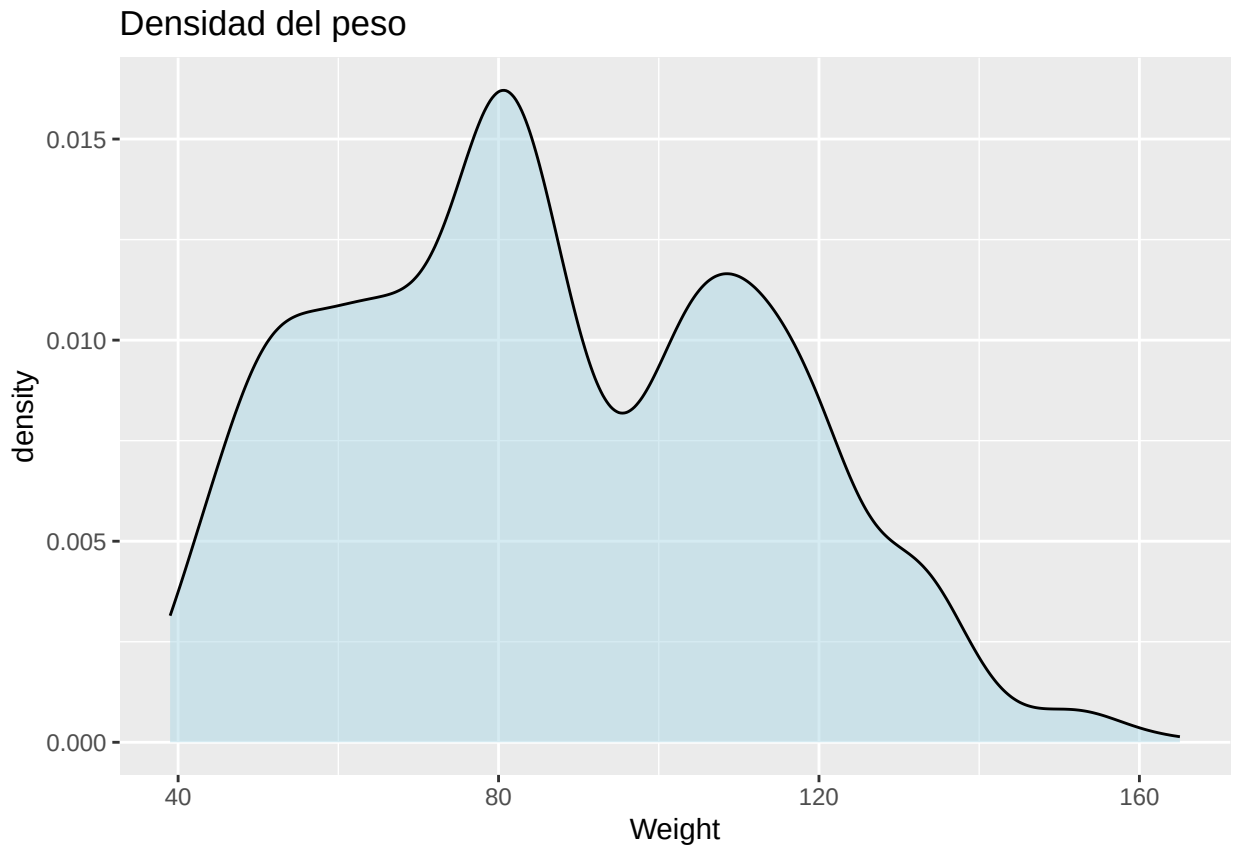
```
## 503  Public_Transportation Obesity_Type_III
## 1899 Public_Transportation Obesity_Type_III
## 1911 Public_Transportation
```

Y nuevamente, observamos que estos pesos iguales o mayores a 160 kilos entran en lo plausible y teniendo en cuenta que el dataset proviene de un estudio sobre el peso y los hábitos alimenticios. No parece proceder eliminar estas instancias pues no asemejan ser casos de valores extremos.

Procedemos a analizar gráficamente:




Histograma del peso
count
200
150
100
50
0
60
90
120
150
Weight



Las gráficas de Weight muestran una distribución asimétrica positiva, hacia la derecha, con la mayor parte de los datos concentrados en pesos menores y una cola que se extiende hacia pesos mayores. Se observa la presencia de algunos picos en el histograma y un valor atípico en el diagrama de caja, que probablemente representa un dato real de una persona con un peso elevado.

Esta información es importante para la imputación de valores faltantes en `Weight`, ya que la asimetría y los valores extremos pueden afectar la precisión de la imputación a la media. Consideraremos alternativas como la imputación por la mediana o la transformación de la variable para obtener más robustez y puesto que la variable peso es muy relevante, vamos a intentar la imputacion con metodos robusto como la imputacion multiple:

```
##
##  iter imp variable
##   1   1  Age  Weight  FCVC  NCP  CH2O  FAF  TUE
##   1   2  Age  Weight  FCVC  NCP  CH2O  FAF  TUE
##   1   3  Age  Weight  FCVC  NCP  CH2O  FAF  TUE
##   1   4  Age  Weight  FCVC  NCP  CH2O  FAF  TUE
##   1   5  Age  Weight  FCVC  NCP  CH2O  FAF  TUE
##   2   1  Age  Weight  FCVC  NCP  CH2O  FAF  TUE
##   2   2  Age  Weight  FCVC  NCP  CH2O  FAF  TUE
##   2   3  Age  Weight  FCVC  NCP  CH2O  FAF  TUE
##   2   4  Age  Weight  FCVC  NCP  CH2O  FAF  TUE
##   2   5  Age  Weight  FCVC  NCP  CH2O  FAF  TUE
##   3   1  Age  Weight  FCVC  NCP  CH2O  FAF  TUE
##   3   2  Age  Weight  FCVC  NCP  CH2O  FAF  TUE
##   3   3  Age  Weight  FCVC  NCP  CH2O  FAF  TUE
##   3   4  Age  Weight  FCVC  NCP  CH2O  FAF  TUE
```

```
##   3   5  Age  Weight  FCVC  NCP  CH2O  FAF  TUE
##   4   1  Age  Weight  FCVC  NCP  CH2O  FAF  TUE
##   4   2  Age  Weight  FCVC  NCP  CH2O  FAF  TUE
##   4   3  Age  Weight  FCVC  NCP  CH2O  FAF  TUE
##   4   4  Age  Weight  FCVC  NCP  CH2O  FAF  TUE
##   4   5  Age  Weight  FCVC  NCP  CH2O  FAF  TUE
##   5   1  Age  Weight  FCVC  NCP  CH2O  FAF  TUE
##   5   2  Age  Weight  FCVC  NCP  CH2O  FAF  TUE
##   5   3  Age  Weight  FCVC  NCP  CH2O  FAF  TUE
##   5   4  Age  Weight  FCVC  NCP  CH2O  FAF  TUE
##   5   5  Age  Weight  FCVC  NCP  CH2O  FAF  TUE
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   39.00   65.10   82.85   86.48  106.78  165.06
```

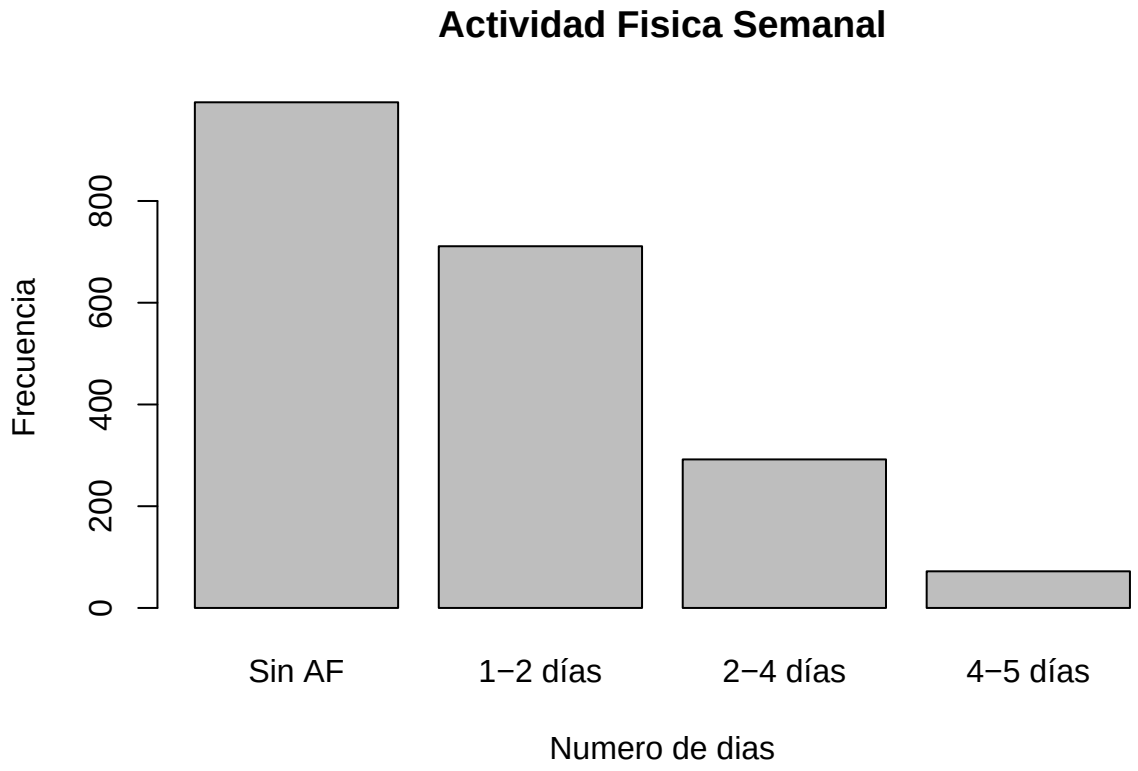
Y verificamos:

```
## [1] "No hay valores perdidos en la columna 'Weight'"
```

FAF Se refiere a la actividad fisica (FAF) que toma valores desde 0 a 3 que se corresponde va desde no quien realiza actividad fisica, 1 o 2 dias de actividad fisica, 2 a 4 dias y 4 o 5 dias, respectivamente. Contiene 42 valores ausentes.

En el mismo sentido que los casos anteriores, realizamos una tabla de frecuencias:

```
##
##   Sin AF 1-2 días 2-4 días 4-5 días
##      994      711      292       72
```



Observamos una concentracion en sobre la ausencia de actividad fisica y que gradualmente va disminuyendo la frecuencia de instancias a medida que aumenta el numero de dias de actividad fisica.

Emplearemos la prueba de *Shapiro-Wilk*:

```
##
##  Shapiro-Wilk normality test
##
## data:  df_base$FAF
## W = 0.91466, p-value < 2.2e-16
```

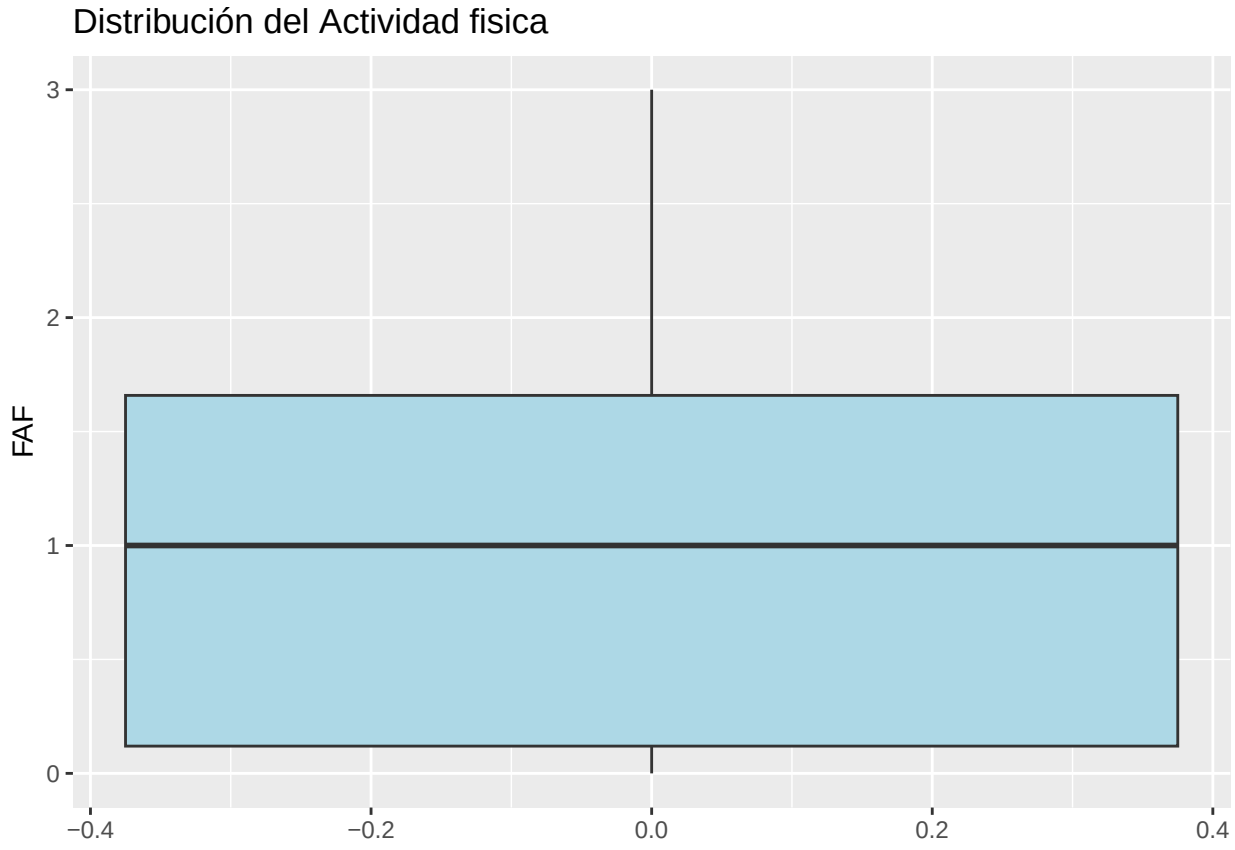
Los resultados de la prueba Shapiro-Wilk arrojaron un estadístico W = 0.91466 y un valor p < 2.2e-16.

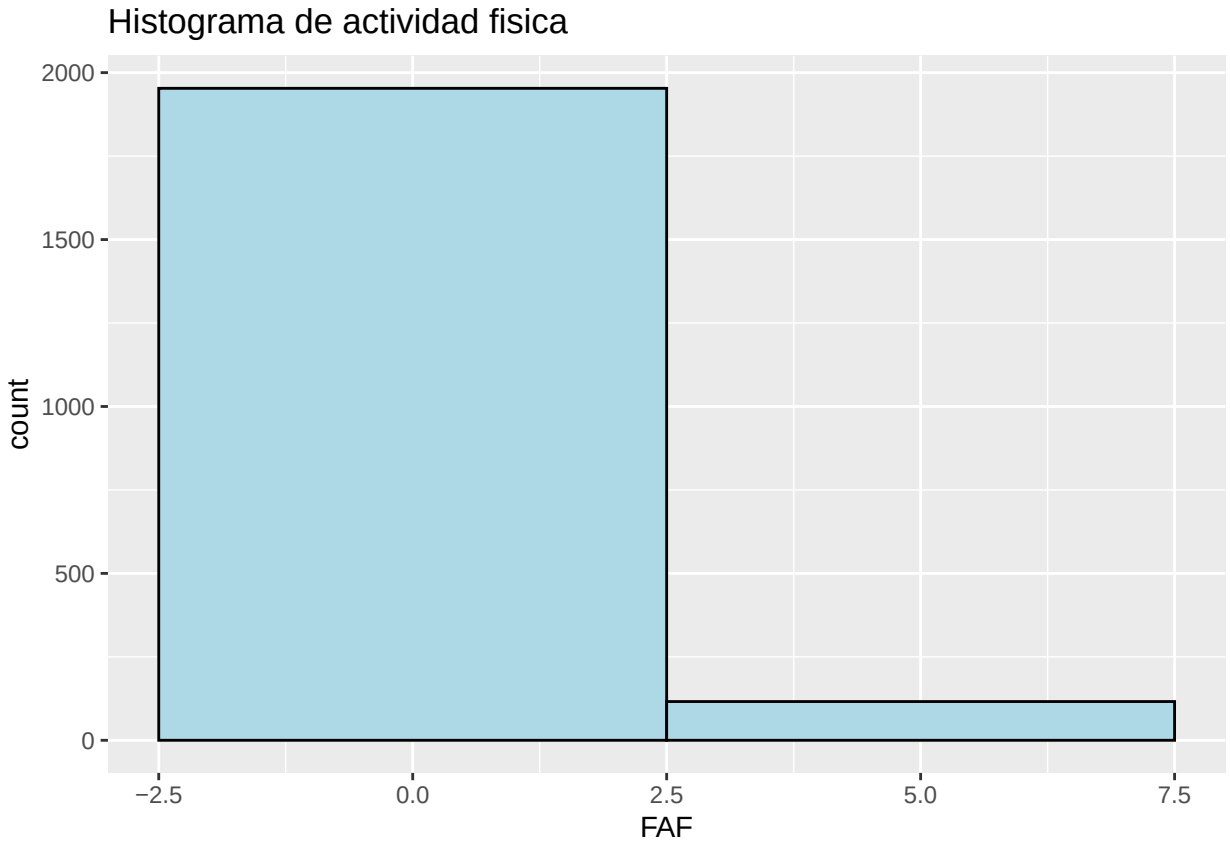
El valor p, nuevamente es extremadamente bajo. Proporciona evidencia contundente para rechazar la hipótesis nula de normalidad. Por lo tanto, se concluye que la distribución de la frecuencia de actividad física en la muestra **no se ajusta a una distribución normal.**

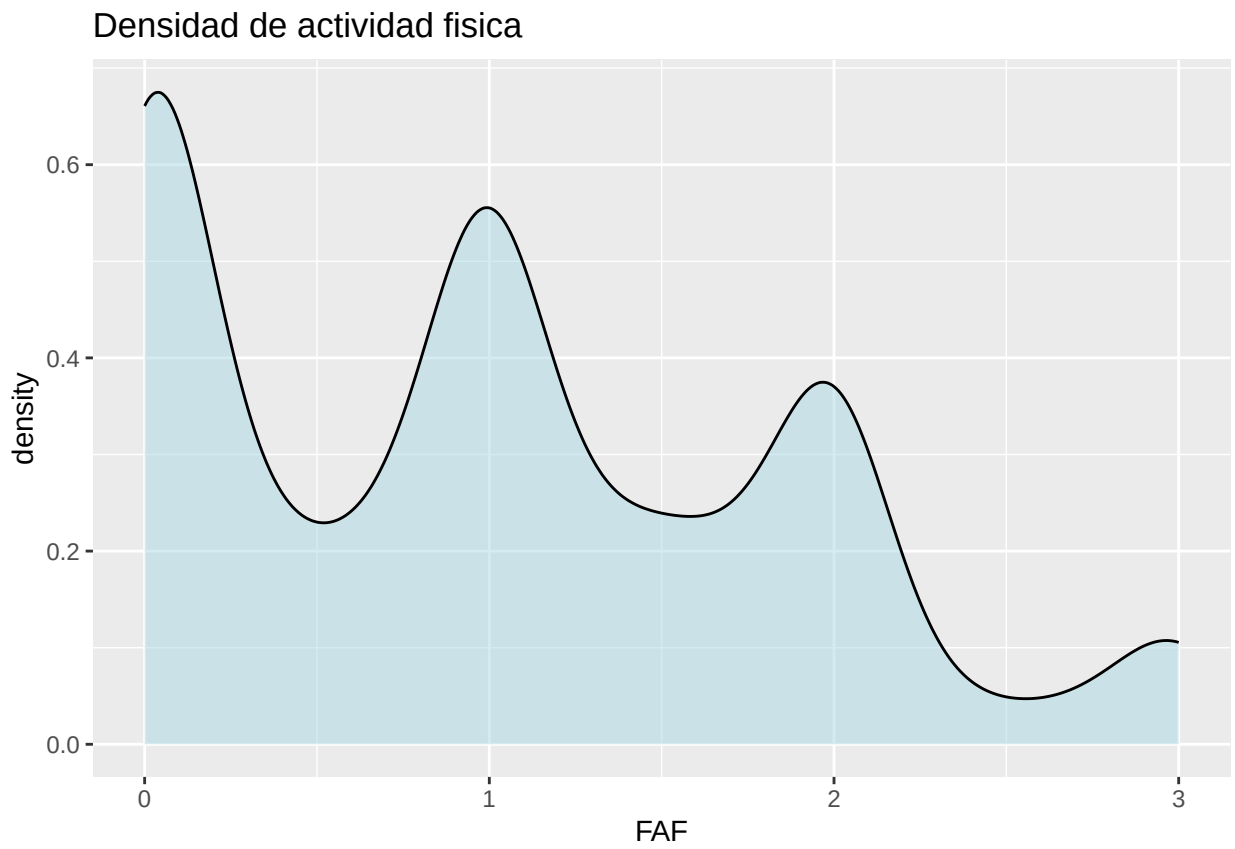
Esta desviación de la normalidad puede ser atribuible a la naturaleza discreta y ordinal de la variable FAF que registra la frecuencia de actividad física en días por semana (0 a 3). Las variables discretas con un número limitado de valores a menudo no se ajustan a una distribución normal.

A pesar de la no normalidad, la variable FAF es relevante para el análisis y se mantendrá en el conjunto de datos. Se utilizarán métodos de análisis apropiados para variables ordinales y se considerarán transformaciones de la variable o métodos no paramétricos si es necesario, para asegurar la validez de los análisis posteriores.

Se complementará este análisis con la visualización de la distribución de "FAF" mediante histogramas y gráficos de densidad para obtener una comprensión más completa de su forma y características.




Histograma de actividad fisica
2000
1500
1000
500
0
count
−2.5
0.0
2.5
5.0
7.5
FAF



El análisis gráfico de la variable `FAF` (frecuencia de actividad física), a través del histograma y el diagrama de caja proporcionados muestran las siguientes características:

Histograma:

- **Claramente discreta:** El histograma muestra que la variable `FAF` es discreta, con valores concentrados en los enteros 0, 1, 2 y 3. Esto es coherente con la naturaleza de la variable, que representa la frecuencia de actividad física en días por semana.
- **Asimetría hacia la derecha:** La mayor parte de los datos se concentra en el valor 0 (sin actividad física), y la frecuencia disminuye a medida que aumenta el valor de `FAF`. No parece haber una asimetría positiva claramente marcada. La gráfica muestra una distribución bimodal (dos picos), pero no hay una cola extendida significativa hacia ninguno de los lados. La densidad cae de forma relativamente equilibrada después de cada pico, lo que sugiere que la asimetría no es un factor predominante en esta distribución.
- **Multimodalidad:** Se observan picos en cada uno de los valores discretos de `FAF`, lo que confirma la multimodalidad de la distribución. Cada pico representa la concentración de individuos en cada nivel de frecuencia de actividad física.

Diagrama de caja:

- **Mediana:** La línea dentro de la caja representa la mediana, que parece estar en 0, indicando que la mitad de los individuos no realiza actividad física.
- **Rango intercuartílico:** La caja representa el rango intercuartílico (IQR), que contiene el 50% central de los datos. En este caso, el IQR abarca desde 0 hasta 1, lo que significa que la mitad de los individuos realiza actividad física entre 0 y 2 días a la semana.
- **Asimetría:** El diagrama de caja también muestra la asimetría hacia la derecha, con un "bigote" superior más largo que el inferior.
- **Valores atípicos:** No se observan valores atípicos (outliers) en el diagrama de caja, lo que indica que

no hay valores de `FAF` que se desvíen significativamente del resto de los datos.

Tenemos por tanto una distribución discreta con tendencia leve no muy marca de asimetria hacia la derecha multimodal.
La mayoría de los individuos no realiza actividad física o la realiza solo 1 o 2 días a la semana. La frecuencia disminuye a medida que aumenta el valor de `FAF`.

Implicaciones para la imputación:

La no normalidad y la naturaleza discreta de `FAF` sugieren que la imputación a la media no sería la mejor opción. La imputación por la moda (0 en este caso) parece ser la estrategia más adecuada, ya que refleja la mayor concentración de datos en la distribución.

```
## [1] Sin AF
## Levels: Sin AF 1-2 días 2-4 días 4-5 días
```

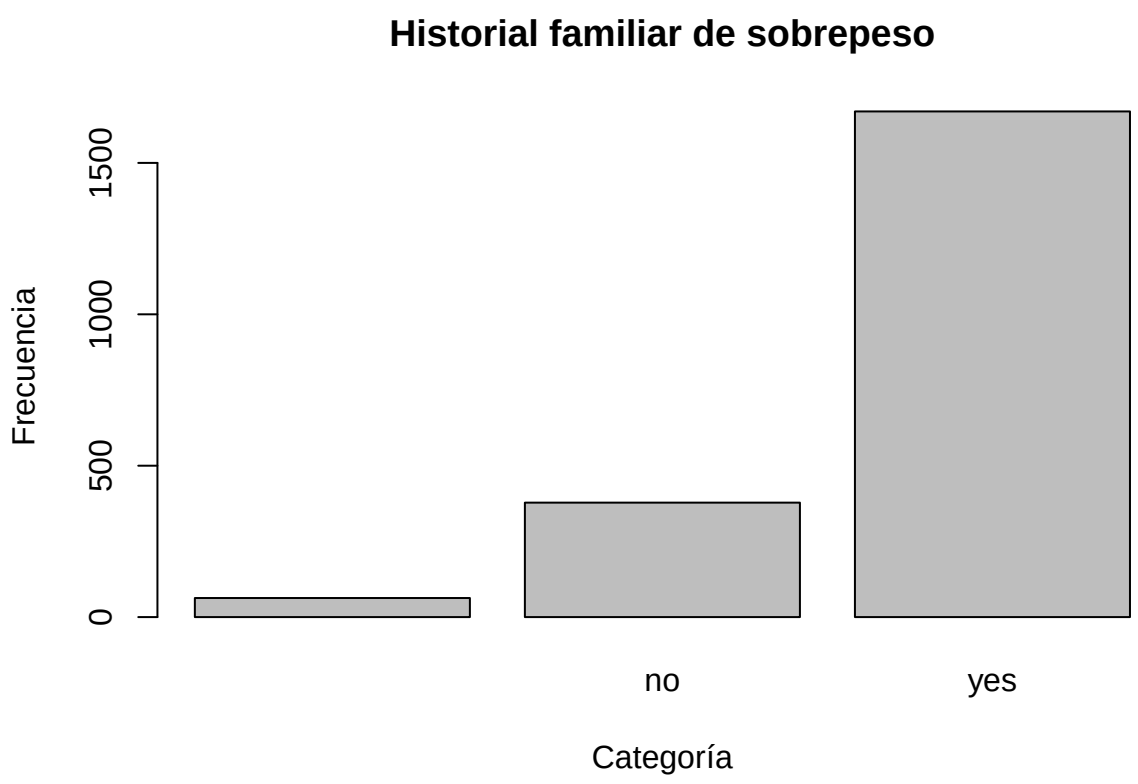
Y verificamos:

```
## [1] "No hay valores perdidos en la columna 'FAF'"
```

family_history_with_overweight Este campo se refiere a si existen antecedentes de sobrepeso entre la familia biologica del sujeto. En el caso del 'historial de familiares con sobrepeso,' aunque los valores se almacenen como cadenas de texto (str), la naturaleza de la variable es binaria, ya que solo hay dos categorías excluyentes.

Tenemos 42 registros 'blank' o de cadenas de texto vacias.
Para analizar la variable `family_history_with_overweight`, una variable categórica binaria (yes/no), utilizaremos métodos apropiados para variables categóricas como tablas de frecuencia, gráficos de barras y pruebas de chi-cuadrado.

```
##
##        no  yes
##   63  378 1670
```



Tanto la tabla de frecuencias como el gráfico de barras indican que hay una mayor proporción de individuos con historial familiar de sobrepeso ("yes") en comparación con aquellos que no lo tienen ("no"). Esto puede tener implicaciones no solo a la hora de elegir un metodo de imputacion (fase en la que estamos) sino que, podria tener implicaciones en el analisis.

Queremos comprender mejor el contexto y la posible influencia de esta variable en el análisis, por lo que valiendonos de una prueba chi cuadrado vamos a analizar como se relaciona esta variable con la variable objetivo 'Nobeyesdad' o nivel de obesidad:

```
##          Insufficient_Weight Normal_Weight Obesity_Type_I Obesity_Type_II Obesity_Type_III Overweigh
##
##        6                   4             4             10              11                9
## no    16                 139           122              7               1                0
## yes   41                 124           146            319             274              306
```

```
##
##  Pearson's Chi-squared test
##
## data:  tabla_contingencia
## X-squared = 590.05, df = 14, p-value < 2.2e-16
```

La prueba evalúa si existe una asociación significativa entre las variables, si la distribución de una variable difiere significativamente entre las categorías de la otra variable.

El resultado principal de la prueba es el *valor p* representa la probabilidad de observar la relación entre las variables en la muestra si no hubiera ninguna asociación real entre ellas en la población. A tenor de estos resultados ($p <$ de 0.05), podemos rechazar la hipótesis nula en la prueba de chi-cuadrado que establece que *no hay asociación entre las variables 'family_history_with_overweight' (historial familiar de sobrepeso) y 'NObeyesdad' (nivel de obesidad).

Concluimos que **sí existe una asociación estadísticamente significativa** entre estas dos variables*. La distribución del nivel de obesidad difiere significativamente entre las personas con y sin historial familiar de sobrepeso. Es importante recalcar que estos resultados no nos dicen la fuerza o la dirección de la asociación, solo que existe una asociación estadistica.

El desequilibrio entre las categorías "no" y "yes" de esta variable puede afectar la precisión de algunos métodos de imputación. Como ejemplo, una imputación por la moda ("yes" en este caso) podría aumentar aún más el desequilibrio. Por otro lado, `family_history_with_overweight` es una variable binaria que representa la presencia o ausencia de antecedentes familiares de sobrepeso. Este tipo de variable puede ser relevante para el análisis de la obesidad, por lo que es importante considerar métodos de imputación que preserven la información y no introduzcan sesgos.

Imputación con missForest para la variable "family_history_with_overweight"

La variable "family_history_with_overweight" (historial familiar de sobrepeso), que registra la presencia o ausencia de antecedentes familiares de obesidad, presenta valores faltantes en el conjunto de datos. Para abordar este problema, se ha seleccionado el método de imputación `missForest`, basado en el algoritmo de Random Forest. Podemos justificar nuestra decision en:

1. La naturaleza de la variable:

`family_history_with_overweight` es una variable categórica binaria. `missForest` es capaz de manejar variables categóricas de forma adecuada a diferencia de otros métodos de imputación centrados en variables numéricas.

2. La relación con otras variables:

Existe una relación entre el historial familiar de sobrepeso y la variable objetivo nivel de obesidad (`NObeyesdad`) y posiblemente entre los hábitos alimenticios. El algoritmo utiliza la información de todas las variables en el conjunto de datos para predecir los valores faltantes, permitiendo capturar relaciones complejas y mejorar la precisión de la imputación.

3. La potencia:

`missForest` es un método no paramétrico que no requiere asumir una distribución específica para los datos faltantes. Esto lo hace potente frente a diferentes patrones de datos faltantes y a la presencia de valores atípicos.

4. La precisión:

`missForest` suele tener en *ratios* un buen rendimiento en la imputación de datos, con una alta precisión en la predicción de valores faltantes.

5. La simplicidad:

A pesar de su sofisticación, `missForest` es relativamente fácil de implementar en R con una función que maneja automáticamente la imputación de variables mixtas (numéricas y categóricas).

Por todo lo anterior, procedemos a la implementacion:

```
##      NRMSE        PFC
## 0.04593892 0.11211911
```

```
##     Gender          Age            Height          Weight
##  Female:1042   Min.   :14.00   Min.   :1.456   Min.   : 39.00
##  Male  :1069   1st Qu.:20.00   1st Qu.:1.632   1st Qu.: 65.10
##                Median :22.77   Median :1.701   Median : 82.85
##                Mean   :24.30   Mean   :1.702   Mean   : 86.48
##                3rd Qu.:26.00   3rd Qu.:1.765   3rd Qu.:106.78
##                Max.   :61.00   Max.   :1.980   Max.   :165.06
##
##  family_history_with_overweight  FAVC           FCVC            NCP
```

```
##  no : 390                       no : 246   Nunca  : 204   Ninguna : 387
##  yes:1721                       yes:1865   A veces:1259   Una-dos : 298
##                                            Siempre: 648   Tres    :1358
##                                                           Mas de 3:  68
##
##
##
##          CAEC      SMOKE        CH2O       SCC             FAF
##  Always    :  49   no :2069   1L  : 770   no :2016   Sin AF  :1036
##  Frequently: 245   yes:  42   1-2L:1179   yes:  95   1-2 días: 711
##  no        :  50              2+L : 162              2-4 días: 292
##  Sometimes :1767                                     4-5 días:  72
##
##
##
##       TUE               CALC                          MTRANS
##  0-2 hrs:1412   Always    :   1   Automobile           : 459
##  3-5 hrs: 591   Frequently:  70   Bike                 :   7
##  5hrs + : 108   no        : 642   Motorbike            :  12
##                 Sometimes :1398   Public_Transportation:1578
##                                   Walking              :  55
##
##
##                NObeyesdad
##  Insufficient_Weight:272
##  Normal_Weight      :287
##  Obesity_Type_I     :351
##  Obesity_Type_II    :296
##  Obesity_Type_III   :324
##  Overweight_Level_I :289
##  Overweight_Level_II:292
```

Los valores que ha devuelto la linea `print(df_imputado$OOBerror)` se corresponden a las medidas de error del algoritmo missForest:

NRMSE (Normalized Root Mean Squared Error):
Medida del error para las variables numéricas. Es la raíz cuadrada del error cuadrático medio, normalizada por el rango de la variable. Un valor de 0 indica un ajuste perfecto, y valores más altos indican un mayor error. En nuestro caso **0.0464599** sugiere un error bajo en la imputación de las variables numéricas.

PFC (Proportion of Falsely Classified):
Medida del error para las variables categóricas. Se trata de la proporción de casos que fueron clasificados incorrectamente por el algoritmo. Un valor de 0 indica una clasificación perfecta, y valores más altos indican un mayor error. En nuestro caso, el valor de **0.1127890** indica que aproximadamente el 11.30% de las categorías en las variables categóricas fueron imputadas incorrectamente.

En general, los resultados de la imputación con `missForest` parecen ser buenos. El error en las variables numéricas es bajo, y el error en las variables categóricas (las que mas problemas no estan suponiendo hasta ahora) es relativamente bajo.

Con el objetivo de transparentar el proceso de imputación y justificar cada decisión metodológica, se ha optado por un enfoque atributo a atributo. Aunque `missForest` ha demostrado ser un método eficiente para la imputación del conjunto de datos completo, este enfoque paso a paso nos permite un análisis más detallado de cada variable y una mejor comprensión del impacto de la imputación en los resultados. Asimismo, habilita la comparación entre diferentes métodos de imputación y la elección de la estrategia más adecuada para cada atributo.

Imputamos los resultados del atribuito `family_history_with_overweight` en el dataset:

```
## 'data.frame':    2111 obs. of  17 variables:
##  $ Gender                        : chr  "Female" "Female" "Male" "Male" ...
##  $ Age                           : num  21 21 23 27 22 29 23 22 24 22 ...
##  $ Height                        : num  1.62 1.52 1.8 1.8 1.78 1.62 1.5 1.64 1.78 1.72 ...
##  $ Weight                        : num  64 56 77 87 89.8 53 55 53 64 68 ...
##  $ family_history_with_overweight: Factor w/ 2 levels "no","yes": 2 2 2 1 1 1 1 1 2 2 ...
##  $ FAVC                          : chr  "no" "no" "no" "no" ...
##  $ FCVC                          : Factor w/ 3 levels "Nunca","A veces",..: 2 3 2 3 2 2 3 2 3 2 ...
##  $ NCP                           : Factor w/ 4 levels "Ninguna","Una-dos",..: 3 3 3 3 1 3 3 3 3 3 ..
##  $ CAEC                          : chr  "Sometimes" "Sometimes" "Sometimes" "Sometimes" ...
##  $ SMOKE                         : chr  "no" "yes" "no" "no" ...
##  $ CH2O                          : Factor w/ 3 levels "1L","1-2L","2+L": 2 3 2 2 2 2 2 2 2 2 ...
##  $ SCC                           : chr  "no" "yes" "" "no" ...
##  $ FAF                           : Factor w/ 4 levels "Sin AF","1-2 días",..: 1 4 3 3 1 1 2 4 2 2 ..
##  $ TUE                           : Factor w/ 3 levels "0-2 hrs","3-5 hrs",..: NA 1 2 1 1 1 1 1 2 NA
##  $ CALC                          : chr  "no" "Sometimes" "Frequently" "Frequently" ...
##  $ MTRANS                        : chr  "Public_Transportation" "Public_Transportation" "Public_Trans
##  $ NObeyesdad                    : chr  "Normal_Weight" "Normal_Weight" "Normal_Weight" "Overweight_L
```

```
##     Gender               Age            Height          Weight
##  Length:2111        Min.   :14.00   Min.   :1.456   Min.   : 39.00
##  Class :character   1st Qu.:19.91   1st Qu.:1.632   1st Qu.: 65.10
##  Mode  :character   Median :22.77   Median :1.701   Median : 82.85
##                     Mean   :24.29   Mean   :1.702   Mean   : 86.48
##                     3rd Qu.:26.00   3rd Qu.:1.765   3rd Qu.:106.78
##                     Max.   :61.00   Max.   :1.980   Max.   :165.06
##                     NA's   :106
##  family_history_with_overweight     FAVC               FCVC
##  no : 390                       Length:2111        Nunca  : 198
##  yes:1721                       Class :character   A veces:1233
##                                 Mode  :character   Siempre: 638
##                                                    NA's   :  42
##
##
##
##        NCP           CAEC              SMOKE             CH2O
##  Ninguna : 377   Length:2111        Length:2111        1L  : 762
##  Una-dos : 276   Class :character   Class :character   1-2L:1166
##  Tres    :1307   Mode  :character   Mode  :character   2+L : 162
##  Mas de 3:  67                                         NA's:  21
##  NA's    :  84
##
##
##      SCC                 FAF            TUE           CALC
##  Length:2111        Sin AF  :1036   0-2 hrs:1382   Length:2111
##  Class :character   1-2 días: 711   3-5 hrs: 566   Class :character
##  Mode  :character   2-4 días: 292   5hrs + : 100   Mode  :character
##                     4-5 días:  72   NA's   :  63
##
##
##
##     MTRANS           NObeyesdad
##  Length:2111        Length:2111
```

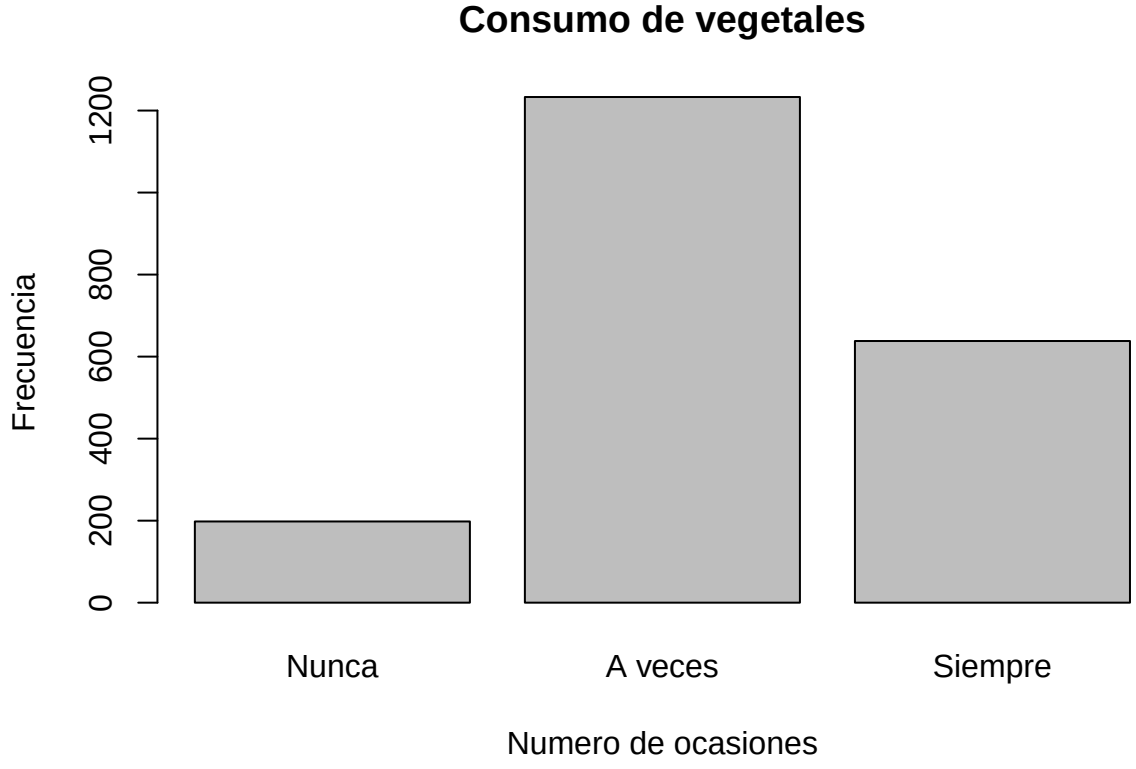
```
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
##
```

```
## [1] "No hay valores perdidos en la columna 'family_history_with_overweight'"
```

FCVC Abordamos el analisis de la variable "FCVC" (frecuencia de consumo de vegetales) para comprender su distribución y características. Se trata de una variable discreta y ordinal que registra la frecuencia con la que los individuos consumen vegetales en una escala de 1 a 3 donde:

- 1 = Nunca
- 2 = A veces
- 3 = Siempre Asimismo sabemos que en ella se contienen 42 valores perdidos. Procedemos con el analisis exploratorio creando una tabla de frecuencias:

```
##
##   Nunca A veces Siempre
##     198    1233     638
```



Observamos nuevamente una asimetria, ya que el grueso recae en la categoria 'a veces.'

A continuacion creamos una tabla de contingencia:

```
##              Insufficient_Weight Normal_Weight Obesity_Type_I Obesity_Type_II Obesity_Type_III Overwe
##
## Nunca      8                  34            16             29              49                0
```

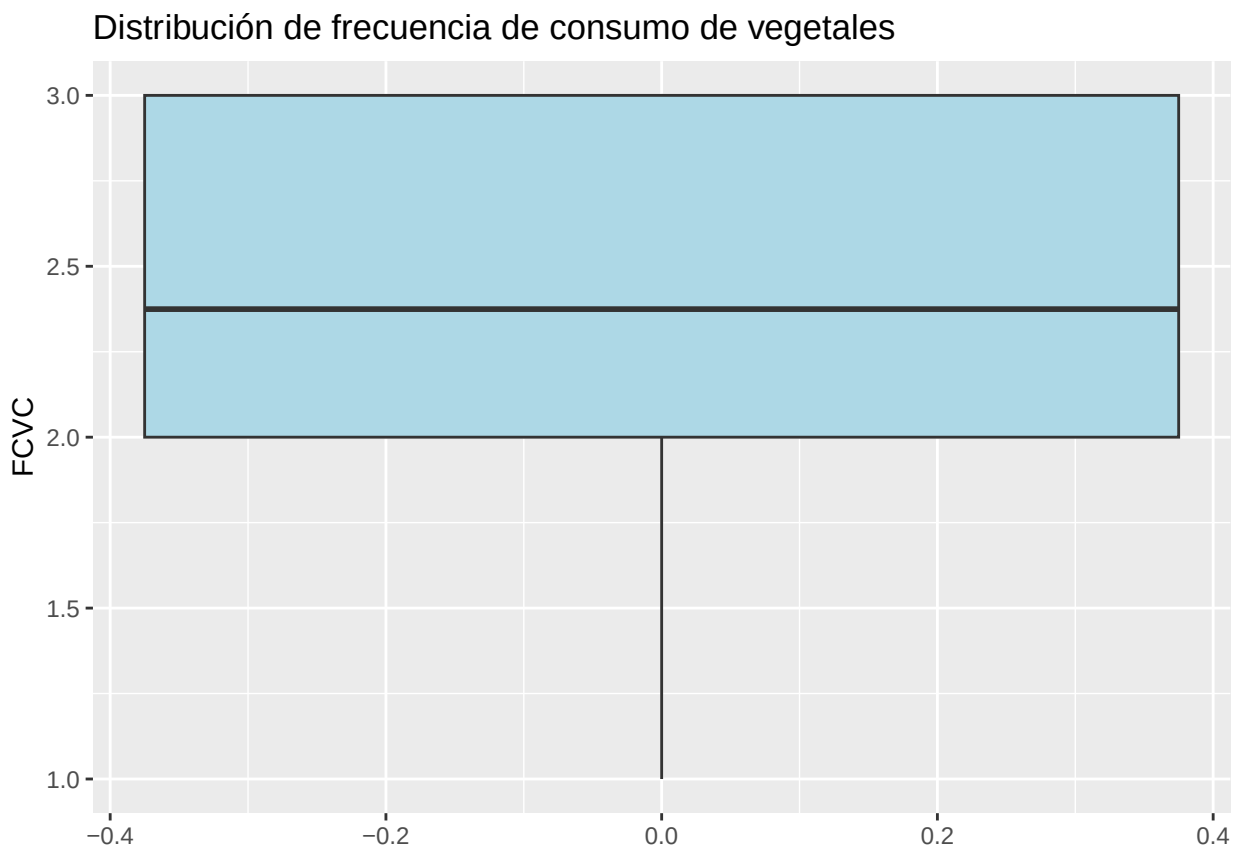
```
## A veces   34                145           145           273           214                0
## Siempre   20                 80           108            26            20              309
```

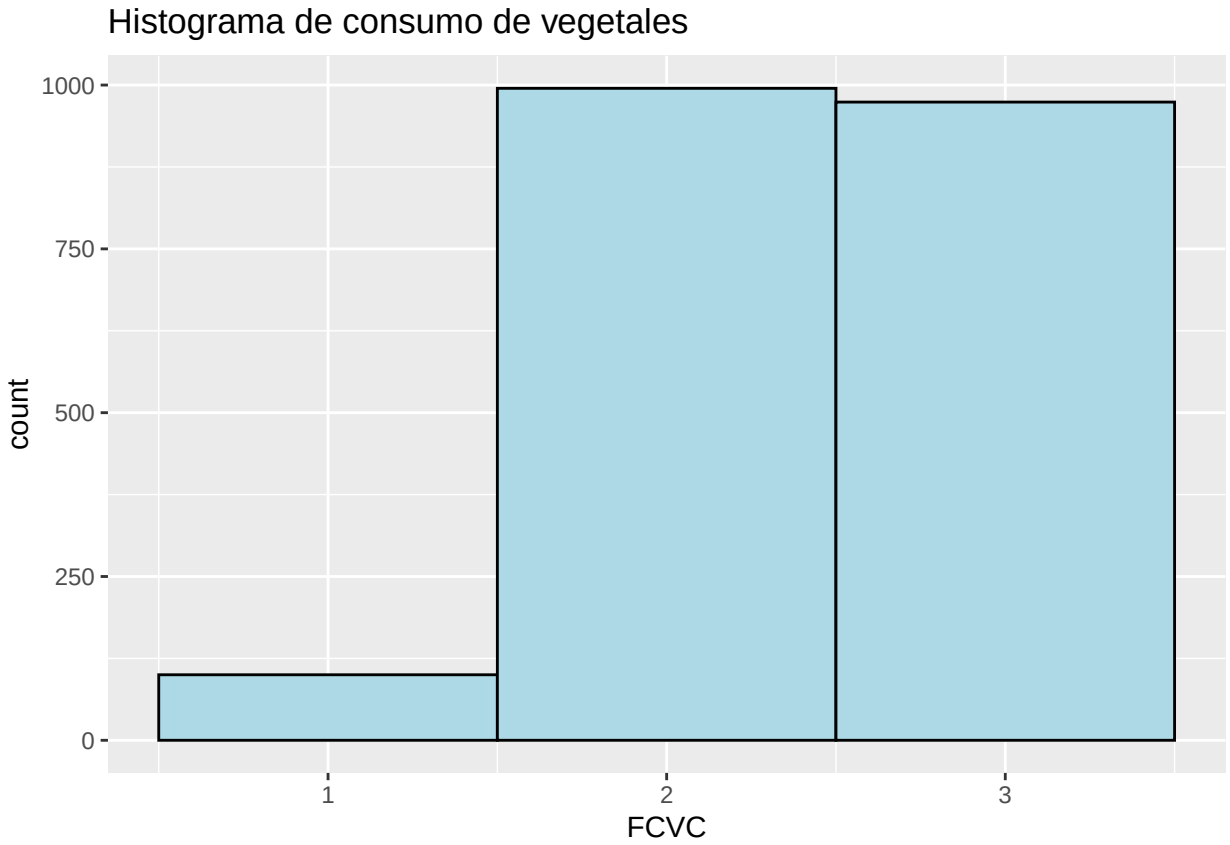
```
##
##  Pearson's Chi-squared test
##
## data:  tabla_contingencia
## X-squared = 961.8, df = 14, p-value < 2.2e-16
```

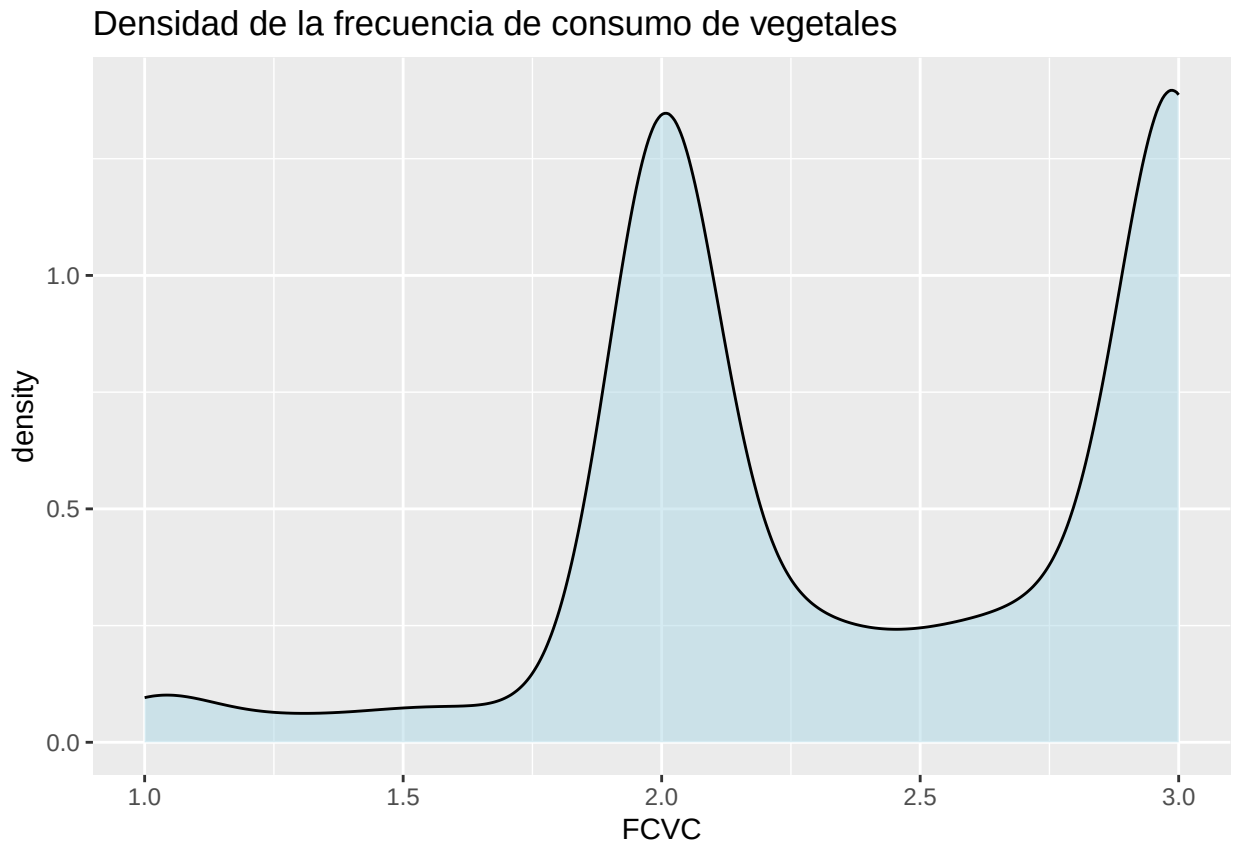
Emplearemos nuevamente la prueba de *Shapiro-Wilk*:

```
##
##  Shapiro-Wilk normality test
##
## data:  df$Weight
## W = 0.97573, p-value < 2.2e-16
```

Los resultados obtenidos para la variable `Weight` **(W = 0.97573, p-value < 2.2e-16)** indican que la distribución no se ajusta a una distribución normal. El valor p se repite e indica una fuerte evidencia para rechazar la hipótesis nula de que los datos provienen de una distribución normal. Asimismo la prueba chi cuadrado indica una relacion con la variable objetivo de los niveles de obesidad. Graficamos:




Histograma de consumo de vegetales
1000
750
500
250
0
count
1
2
3
FCVC



El análisis exploratorio a través de histogramas y diagramas de caja muestra que:

- **Concentración en valores altos:** Se observa una clara concentración de datos en los valores 2 y 3, lo que indica que la mayoría de los individuos consume vegetales con cierta frecuencia (a veces o siempre).
- **Baja frecuencia del valor "Nunca":** El valor 1 ("Nunca") presenta una frecuencia mucho menor, sugiriendo que son pocos los individuos que no consumen vegetales.
- **Asimetría hacia la izquierda:** La distribución muestra una asimetría negativa, con una cola más larga hacia la izquierda (valores menores).

ELos resultados muestran una tendencia hacia un consumo frecuente de vegetales en la población estudiada. La asimetría observada sugiere que la mayoría de los individuos se inclinan hacia un consumo regular de vegetales, mientras que una minoría reporta no consumirlos nunca.

La distribución de `FCVC` proporciona información importante para la comprensión de los hábitos alimenticios de la población y sienta las bases para análisis posteriores como la evaluación de la asociación entre el consumo de vegetales y otras variables de interés - nivel de obesidad -, por lo que es una variable a tener en cuenta.

A la hora de imputar los valores perdidos, asumiremos el contexto de esta variable. La asimetría a la izquierda y la naturaleza discreta de la variable `FCVC` tienen implicaciones a la hora de imputar los valores faltantes. Las medidas de tendencia central como la media **no parecen la mejor opción**.

- **Distorsión por la asimetría:** En una distribución asimétrica, la media se ve "arrastrada" hacia la cola más larga. En este caso, la media se vería influenciada por los pocos casos con valores bajos de `FCVC` (que nunca consumen vegetales), lo que podría llevar a subestimar la frecuencia de consumo al imputar.
- **Variable discreta:** `FCVC` solo toma valores enteros (1, 2 y 3). La media podría resultar en un valor decimal que no tiene sentido en el contexto de la variable.

Por todo ello, consideramos la imputación por la moda, la mediana, regresion multiple y la imputacion multiple. En este caso, la **imputación por la moda** parece ser la opción más simple y adecuada, dado que la moda es clara y la variable es discreta.

```
## [1] A veces
## Levels: Nunca A veces Siempre
```

Verificamos que todo ha resultado bien y continuamos con la siguiente variable a tratar:

```
## [1] "No hay valores perdidos en la columna 'FCVC'"
```

```
##  Factor w/ 3 levels "Nunca","A veces",..: 2 3 2 3 2 2 3 2 3 2 ...
```

NCP Esta variable se refiere al numero de comidas diarias ingeridas o numero de comidas principales toma en el dataset valores entre 1 y 4. A primera vista podría parecer una variable de intervalo continua. Sin embargo, al analizar la descripción de la variable y cómo se recopilaron los datos, se evidencia una incongruencia, como puede observarse al revisar la documentacion:

https://syw.under.jp/img/incongruencia.png

En la documentación por tanto observamos que las opciones proporcionadas a los participantes fueron:

- *Between 1 y 2*
- *Three*
- *More than three*

Esto nos lleva a concluir que **NCP no es una variable continua, sino una variable ordinal discreta**, ya que las categorías representan un orden o jerarquía en la frecuencia de consumo de comidas principales. La incongruencia surge al asignar valores numéricos (1 a 4) a estas categorías. Si 1 correspondiera a "Between 1 y 2," no tendría sentido que 2, 3 y 4 representaran *"Three"* y *"More than three"*. Es probable y asumimos que hay un error en la codificación de la variable en el conjunto de datos, o que la descripción de la variable no sea completamente precisa.

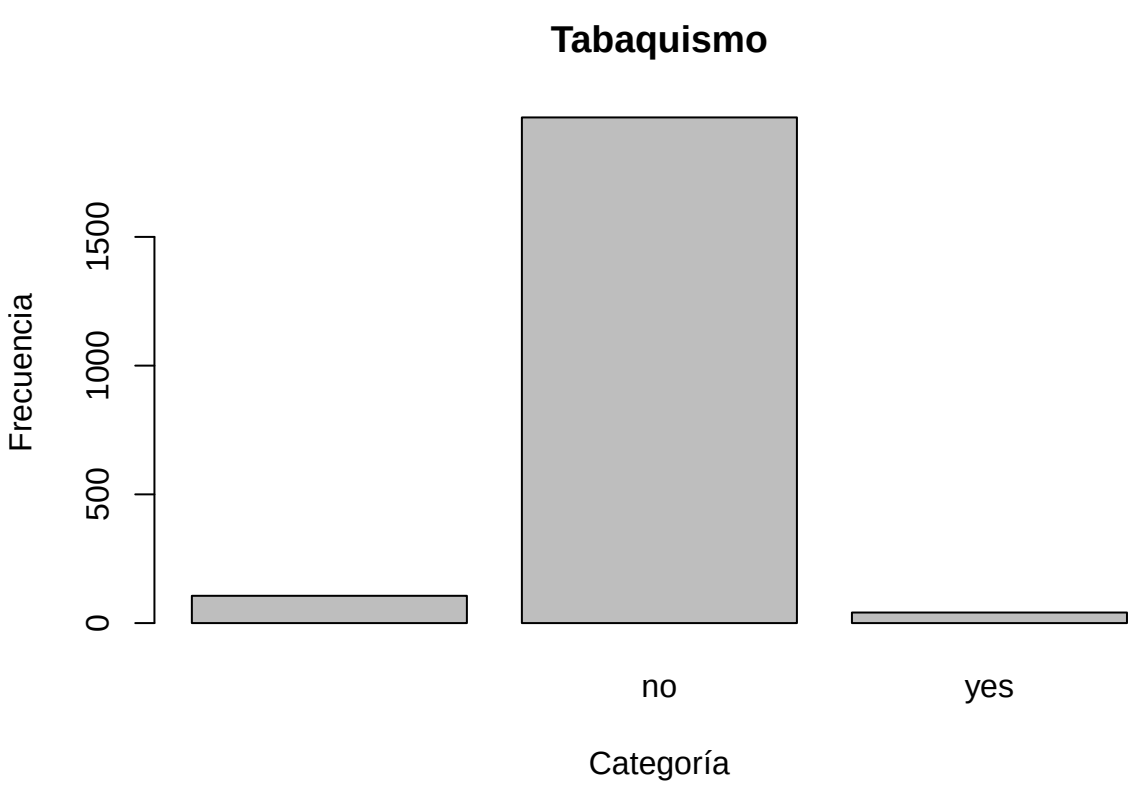
Se ha decidido **excluir la variable NCP (número de comidas principales al día) del análisis** debido a la ambigüedad en su definición y la posible incongruencia en su codificación. La documentación del conjunto de datos no proporciona una descripción clara de cómo se codificaron los valores de la variable, lo que podría llevar a errores de interpretación y sesgos en el modelo. Se considera que la exclusión de NCP es la opción más prudente para asegurar la validez y la precisión del análisis, evitando la introducción de información ambigua o potencialmente errónea.

```
##  [1] "Gender"                          "Age"
##  [3] "Height"                          "Weight"
##  [5] "family_history_with_overweight" "FAVC"
##  [7] "FCVC"                            "CAEC"
##  [9] "SMOKE"                           "CH2O"
## [11] "SCC"                             "FAF"
## [13] "TUE"                             "CALC"
## [15] "MTRANS"                          "NObeyesdad"
```

SMOKE La variable SMOKE registra la presencia o ausencia de adicción al tabaco en cada sujeto experimental. Con un 3% de valores ausentes, esta variable nominal cualitativa dicotómica toma valores *"yes"* o *"no"*.

Para comprender la distribución de la adicción al tabaco y su relación con los niveles de obesidad se procede con el siguiente análisis exploratorio:

```
##
##        no  yes
##  106 1964   41
```



Nuevamente queda patente la desproporcion en las respuestas. Como en los casos anteriores procedemos a observar como se relaciona con la variable objetivo `NObeyesdad`:

```
##          Insufficient_Weight Normal_Weight Obesity_Type_I Obesity_Type_II Obesity_Type_III Overweigh
##
##        1                  15            13             16              18               12
## no    62                 252           246            314             255              302
## yes    0                   0            13              6              13                1
```

```
##
##  Pearson's Chi-squared test
##
## data:  tabla_contingencia
## X-squared = 38.187, df = 14, p-value = 0.0004864
```

La prueba chi-cuadrado de independencia para evaluar la posible asociación entre el tabaquismo (SMOKE) y el nivel de obesidad (NObeyesdad) devolvio un valor de chi-cuadrado **38.187** con 14 grados de libertad y un valor p de 0.0004864.

Con estos datos en la ano **se rechaza la hipótesis nula de independencia y se concluye que existe una asociación estadísticamente significativa entre el tabaquismo y el nivel de obesidad**. La distribución del nivel de obesidad difiere significativamente entre los individuos fumadores y no fumadores.

A pesar del desequilibrio observado en la variable SMOKE, con una mayor proporción de individuos no fumadores, la prueba de chi-cuadrado ha detectado una asociación significativa con el nivel de obesidad. Ello sugiere que **el tabaquismo podría ser un factor a considerar en el estudio de la obesidad** aunque se requiere comprender mejor esta relación.

Aunque la imputación por la moda podría parecer adecuada en este caso debido al desequilibrio en la variable

SMOKE, donde la gran mayoría de los individuos no fuman, optar por un método más sofisticado como la imputación múltiple parece ser lo adecuado por:

- Manejo de la incertidumbre: La imputación múltiple genera varios conjuntos de datos imputados, lo que permite tener en cuenta la incertidumbre asociada a la imputación de los valores faltantes. Esto puede llevar a resultados más robustos y precisos en comparación con la imputación por la moda, que solo genera un conjunto de datos imputado.
- Preservación de la variabilidad: La imputación múltiple tiende a preservar mejor la variabilidad original de la variable SMOKE en comparación con la imputación por la moda, que puede reducir la variabilidad al asignar el mismo valor (la moda) a todos los valores faltantes.
- Flexibilidad: La imputación múltiple permite especificar diferentes modelos de imputación para cada variable, lo que puede ser útil ante variables con diferentes características o patrones de datos perdidos.
- Rigor académico: La imputación múltiple es un método ampliamente reconocido y aceptado en la comunidad científica, lo que puede fortalecer la validez de este trabajo.

Anteriormente empleamos este mismo metodo de imputacion mediante `pmm` (*Predictive Mean Matching*), pero en este caso contamos con mas variables a las que hemos imputados los valores perdidos por lo que volvemos a reconstruir el modelo

```
##
##  iter imp variable
##   1   1  Gender  Age  FAVC  CAEC  SMOKE  CH2O  SCC  TUE  CALC  MTRANS  NObeyesdad
##   1   2  Gender  Age  FAVC  CAEC  SMOKE  CH2O  SCC  TUE  CALC  MTRANS  NObeyesdad
##   1   3  Gender  Age  FAVC  CAEC  SMOKE  CH2O  SCC  TUE  CALC  MTRANS  NObeyesdad
##   1   4  Gender  Age  FAVC  CAEC  SMOKE  CH2O  SCC  TUE  CALC  MTRANS  NObeyesdad
##   1   5  Gender  Age  FAVC  CAEC  SMOKE  CH2O  SCC  TUE  CALC  MTRANS  NObeyesdad
##   2   1  Gender  Age  FAVC  CAEC  SMOKE  CH2O  SCC  TUE  CALC  MTRANS  NObeyesdad
##   2   2  Gender  Age  FAVC  CAEC  SMOKE  CH2O  SCC  TUE  CALC  MTRANS  NObeyesdad
##   2   3  Gender  Age  FAVC  CAEC  SMOKE  CH2O  SCC  TUE  CALC  MTRANS  NObeyesdad
##   2   4  Gender  Age  FAVC  CAEC  SMOKE  CH2O  SCC  TUE  CALC  MTRANS  NObeyesdad
##   2   5  Gender  Age  FAVC  CAEC  SMOKE  CH2O  SCC  TUE  CALC  MTRANS  NObeyesdad
##   3   1  Gender  Age  FAVC  CAEC  SMOKE  CH2O  SCC  TUE  CALC  MTRANS  NObeyesdad
##   3   2  Gender  Age  FAVC  CAEC  SMOKE  CH2O  SCC  TUE  CALC  MTRANS  NObeyesdad
##   3   3  Gender  Age  FAVC  CAEC  SMOKE  CH2O  SCC  TUE  CALC  MTRANS  NObeyesdad
##   3   4  Gender  Age  FAVC  CAEC  SMOKE  CH2O  SCC  TUE  CALC  MTRANS  NObeyesdad
##   3   5  Gender  Age  FAVC  CAEC  SMOKE  CH2O  SCC  TUE  CALC  MTRANS  NObeyesdad
##   4   1  Gender  Age  FAVC  CAEC  SMOKE  CH2O  SCC  TUE  CALC  MTRANS  NObeyesdad
##   4   2  Gender  Age  FAVC  CAEC  SMOKE  CH2O  SCC  TUE  CALC  MTRANS  NObeyesdad
##   4   3  Gender  Age  FAVC  CAEC  SMOKE  CH2O  SCC  TUE  CALC  MTRANS  NObeyesdad
##   4   4  Gender  Age  FAVC  CAEC  SMOKE  CH2O  SCC  TUE  CALC  MTRANS  NObeyesdad
##   4   5  Gender  Age  FAVC  CAEC  SMOKE  CH2O  SCC  TUE  CALC  MTRANS  NObeyesdad
##   5   1  Gender  Age  FAVC  CAEC  SMOKE  CH2O  SCC  TUE  CALC  MTRANS  NObeyesdad
##   5   2  Gender  Age  FAVC  CAEC  SMOKE  CH2O  SCC  TUE  CALC  MTRANS  NObeyesdad
##   5   3  Gender  Age  FAVC  CAEC  SMOKE  CH2O  SCC  TUE  CALC  MTRANS  NObeyesdad
##   5   4  Gender  Age  FAVC  CAEC  SMOKE  CH2O  SCC  TUE  CALC  MTRANS  NObeyesdad
##   5   5  Gender  Age  FAVC  CAEC  SMOKE  CH2O  SCC  TUE  CALC  MTRANS  NObeyesdad
```

```
##   no  yes
## 2066   45
```

Y comprobamos nuevamente que no existan valores perdidos en el dataframe, aunque los resultados de summary ya lo muestran claramente:

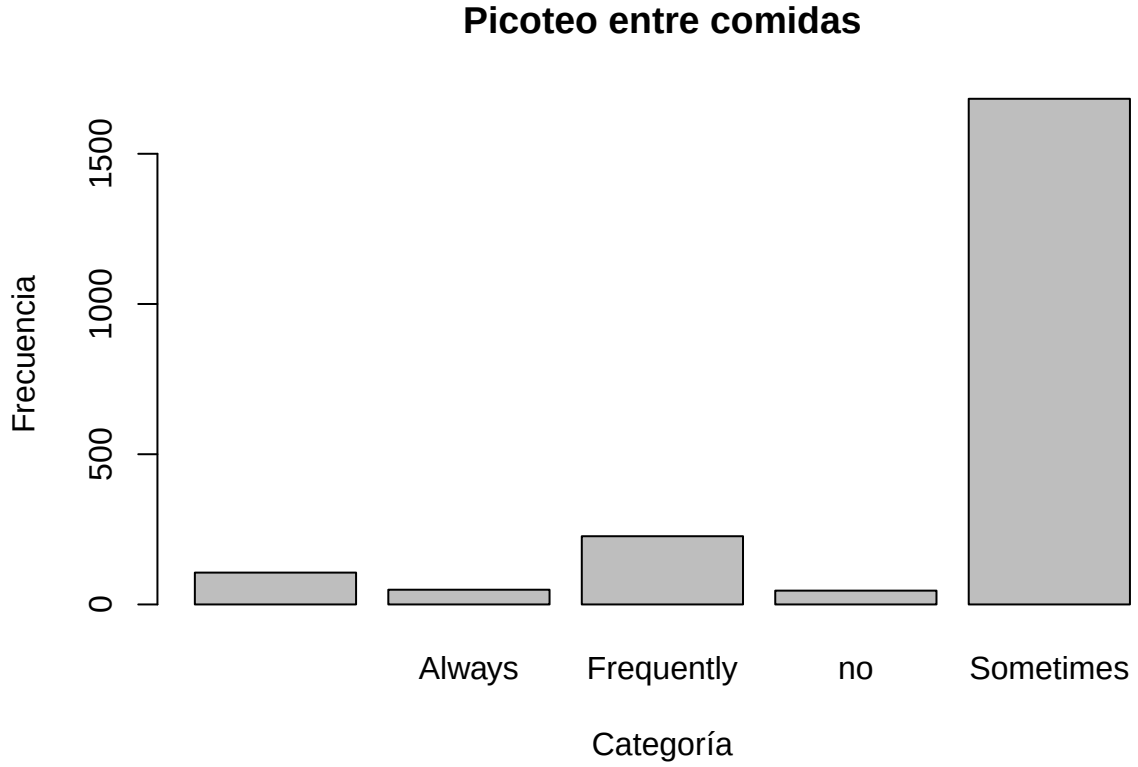
```
## [1] "No hay valores perdidos en la columna 'SMOKE'"
```

Nuevamente y como en el caso del modelo `missforest` empleado anteriormente, tenemos que se han imputado todas variables. En nuestro caso volveremos a ir paso a paso con cada variable empleando varios metodos de imputacion

CAEC Esta variable se refiere a si el sujeto picotea o ingiere alimentos entre las comidas pricipales. `CAEC` por tanto se trata una variable nominal cualitativa politómica,describe una cualidad (el hábito de comer entre comidas) con varias categorías nominales (*"No," "Sometimes," "Frequently," "Always"*) sin un orden intrínseco entre ellas.

Vamos a proceder al analisis exploratorio:

```
##
##               Always Frequently         no  Sometimes
##        106         49        227         46       1683
```



Observamos 106 valores perdidos, lo que supone un 5% del total. Asimismo, tenemos una gran disparidad o desproporcion en las respuestas. Procedemos a continuacion con una tabla de contigencia:

```
##                Insufficient_Weight Normal_Weight Obesity_Type_I Obesity_Type_II Obesity_Type_III Ov
##
##              5                  18            18             12               9              13
## Always       0                   1            32              6               2               0
## Frequently   6                 117            69              5               1               1
## no           1                   3             9              1               1               0
## Sometimes   51                 128           144            312             273             301
```

```
##
##  Pearson's Chi-squared test
##
```

```
## data:  tabla_contingencia
## X-squared = 770.24, df = 28, p-value < 2.2e-16
```

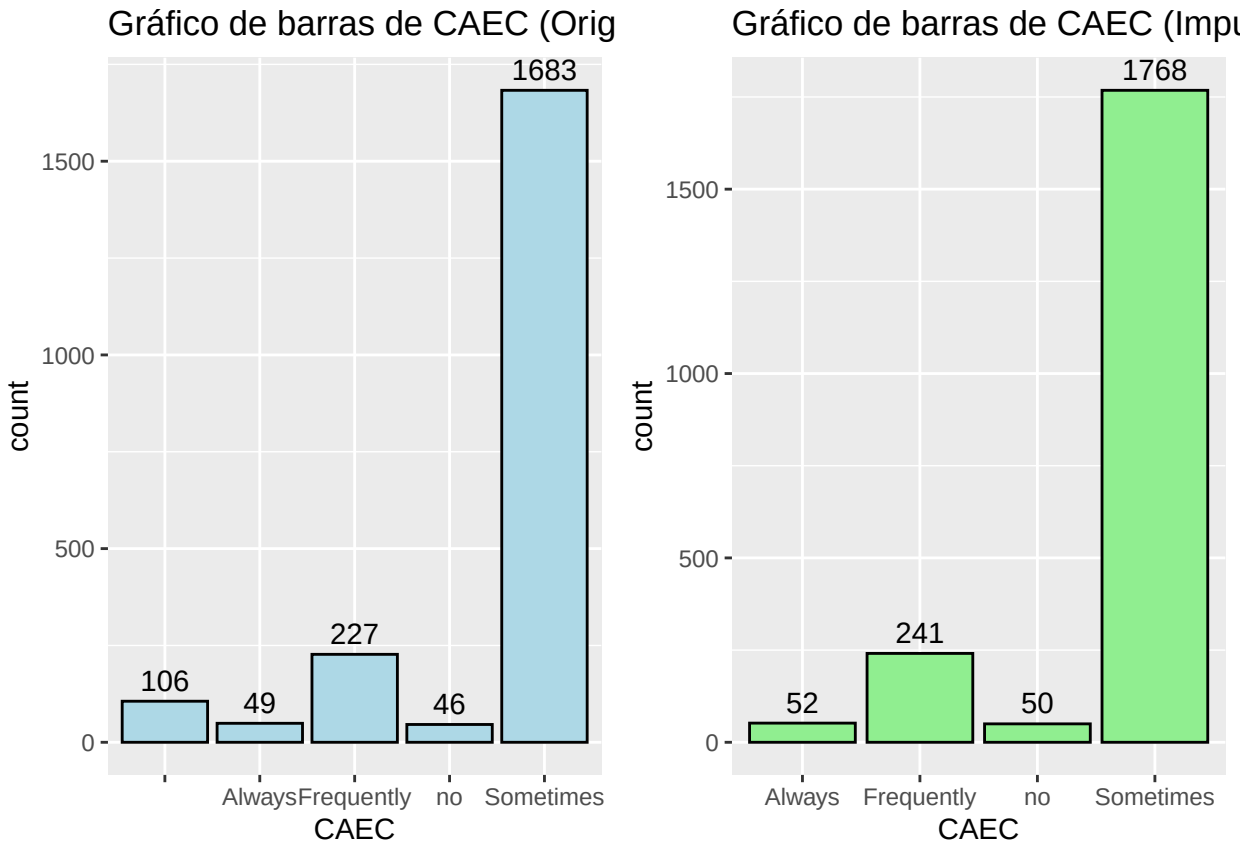
Los resultados mostraron un valor de chi-cuadrado de 770.24 con 28 grados de libertad y un valor p $< 2.2e\text{-}16$. Este valor p proporciona evidencia para rechazar la hipótesis nula de independencia.
Se concluye que existe una asociación estadísticamente significativa entre el consumo de alimentos entre comidas y el nivel de obesidad.El hábito de comer entre comidas podría ser un factor relevante en el desarrollo de la obesidad.

A pesar de la naturaleza nominal de la variable `CAEC`, se utilizará el método de imputación múltiple (`mice`) con el algoritmo "Predictive Mean Matching" (`pmm`) para manejar los valores faltany en este contexto resulta interesante dada la robustez y capacidad para preservar la distribución original de la variable.
Por otro lado, un Un 5% de valores ausentes en la variable CAEC generalmente no se considera un problema para el modelo `mice`.

```
##
##  iter imp variable
##   1   1  Gender  Age  FAVC  CAEC  CH2O  SCC  TUE  CALC  MTRANS  NObeyesdad
##   1   2  Gender  Age  FAVC  CAEC  CH2O  SCC  TUE  CALC  MTRANS  NObeyesdad
##   1   3  Gender  Age  FAVC  CAEC  CH2O  SCC  TUE  CALC  MTRANS  NObeyesdad
##   1   4  Gender  Age  FAVC  CAEC  CH2O  SCC  TUE  CALC  MTRANS  NObeyesdad
##   1   5  Gender  Age  FAVC  CAEC  CH2O  SCC  TUE  CALC  MTRANS  NObeyesdad
##   2   1  Gender  Age  FAVC  CAEC  CH2O  SCC  TUE  CALC  MTRANS  NObeyesdad
##   2   2  Gender  Age  FAVC  CAEC  CH2O  SCC  TUE  CALC  MTRANS  NObeyesdad
##   2   3  Gender  Age  FAVC  CAEC  CH2O  SCC  TUE  CALC  MTRANS  NObeyesdad
##   2   4  Gender  Age  FAVC  CAEC  CH2O  SCC  TUE  CALC  MTRANS  NObeyesdad
##   2   5  Gender  Age  FAVC  CAEC  CH2O  SCC  TUE  CALC  MTRANS  NObeyesdad
##   3   1  Gender  Age  FAVC  CAEC  CH2O  SCC  TUE  CALC  MTRANS  NObeyesdad
##   3   2  Gender  Age  FAVC  CAEC  CH2O  SCC  TUE  CALC  MTRANS  NObeyesdad
##   3   3  Gender  Age  FAVC  CAEC  CH2O  SCC  TUE  CALC  MTRANS  NObeyesdad
##   3   4  Gender  Age  FAVC  CAEC  CH2O  SCC  TUE  CALC  MTRANS  NObeyesdad
##   3   5  Gender  Age  FAVC  CAEC  CH2O  SCC  TUE  CALC  MTRANS  NObeyesdad
##   4   1  Gender  Age  FAVC  CAEC  CH2O  SCC  TUE  CALC  MTRANS  NObeyesdad
##   4   2  Gender  Age  FAVC  CAEC  CH2O  SCC  TUE  CALC  MTRANS  NObeyesdad
##   4   3  Gender  Age  FAVC  CAEC  CH2O  SCC  TUE  CALC  MTRANS  NObeyesdad
##   4   4  Gender  Age  FAVC  CAEC  CH2O  SCC  TUE  CALC  MTRANS  NObeyesdad
##   4   5  Gender  Age  FAVC  CAEC  CH2O  SCC  TUE  CALC  MTRANS  NObeyesdad
##   5   1  Gender  Age  FAVC  CAEC  CH2O  SCC  TUE  CALC  MTRANS  NObeyesdad
##   5   2  Gender  Age  FAVC  CAEC  CH2O  SCC  TUE  CALC  MTRANS  NObeyesdad
##   5   3  Gender  Age  FAVC  CAEC  CH2O  SCC  TUE  CALC  MTRANS  NObeyesdad
##   5   4  Gender  Age  FAVC  CAEC  CH2O  SCC  TUE  CALC  MTRANS  NObeyesdad
##   5   5  Gender  Age  FAVC  CAEC  CH2O  SCC  TUE  CALC  MTRANS  NObeyesdad
```

```
##     Always Frequently         no  Sometimes
##            52        241         50       1768
```

En este caso evaluaremos la calidad de la imputación mediante la comparación de la distribución de la variable imputada con la distribución original y se consideraremos otros métodos de imputación si es necesario:



Los resultados sugieren que la imputación múltiple con `mice` ha logrado completar los valores faltantes en CAEC de forma consistente con la distribución original de la variable. No ha introducido sesgos evidentes ni ha alterado significativamente la estructura de los datos.

CH20 La variable CH2O registra el consumo diario de agua, categorizado en tres niveles: *"Less than a liter," "Between 1 and 2 L" y "More than 2 L"*. Si bien la documentación original la describe como una variable continua, su naturaleza discreta y el orden implícito entre las categorías la definen como una variable ordinal. A diferencia de una variable continua, que puede tomar cualquier valor dentro de un rango, CH2O clasifica el consumo de agua en rangos predefinidos, sin capturar la cantidad exacta.

Aprovechando que anteriormente el modelo de imputacion multiple `mice` pareció tener un buen comportamiento y puesto que la caracteristica no se nos antoja de las más criticas, imputaremos directamente tomando como base las predicciones del modelo:

```
##   1L 1-2L  2+L
##  769 1178  164
```

Y observamos que la variable ya no contiene valores valores ausentes.

TUE La variable TUE cuantifica el tiempo diario dedicado al uso de dispositivos electrónicos, categorizado en tres niveles: "0-2 horas," "3-5 horas" y "Más de 5 horas." Si bien inicialmente sus autores la codificaron con valores numéricos integrales (1, 2 y 3), consideramos su naturaleza ordinal pues estos valores representan un orden creciente en el tiempo de uso. En el dataset contiene 63 valores ausentes.

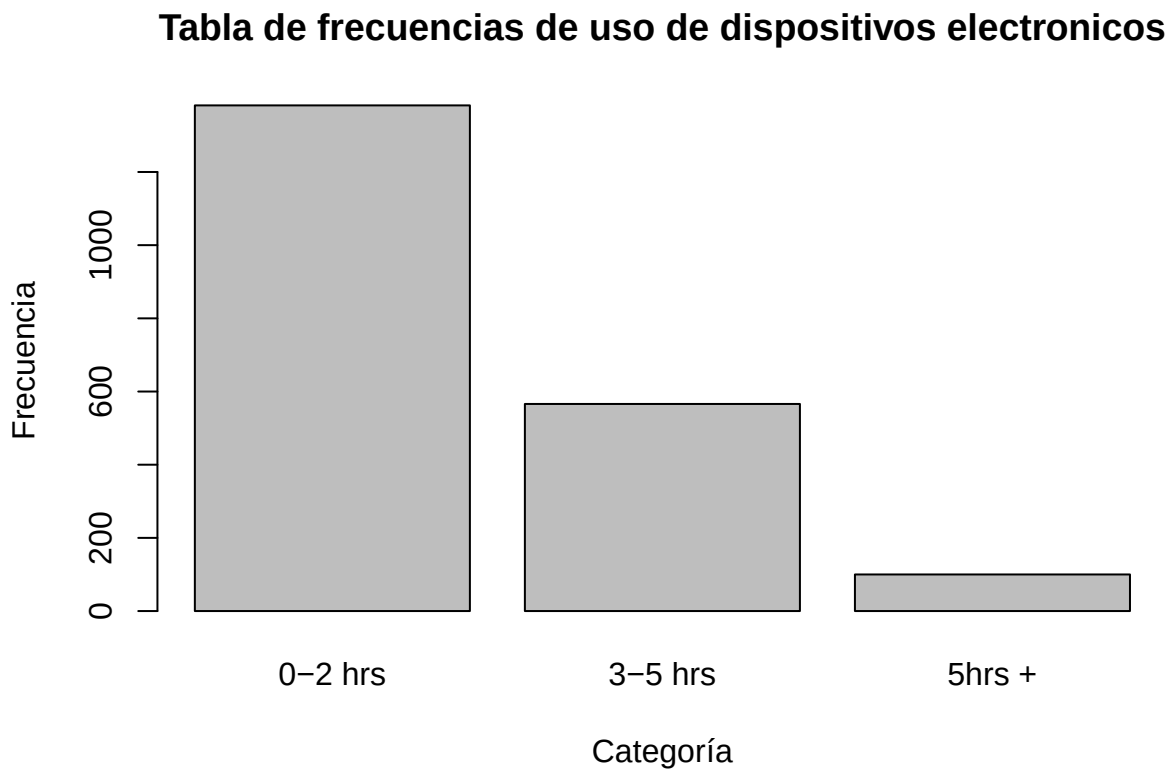
A diferencia de una variable continua, que puede tomar cualquier valor dentro de un rango, TUE clasifica el tiempo de uso en intervalos discretos. Esta característica ordinal implica que al analizar la variable se debe considerar el orden inherente a las categorías y utilizar métodos estadísticos apropiados para variables ordinales.

En el presente estudio, se ha optado por factorizar la variable TUE, lo que permite que R la trate como una variable categórica con un orden predefinido. Esta factorización facilita la interpretación de los resultados y la aplicación de modelos estadísticos que tengan en cuenta la naturaleza ordinal de la variable.

El análisis posterior de TUE incluirá la exploración de su distribución, la evaluación de su asociación con otras variables de interés, como el nivel de obesidad (NObeyesdad), y la imputación de los valores faltantes mediante métodos apropiados para variables categóricas ordinales.

Se prestará especial atención a la interpretación de los resultados, considerando la naturaleza ordinal de TUE y evitando la aplicación de métodos que asuman una escala de intervalo continua. Procedemos con el analisis exploratorio:

```
##
## 0-2 hrs 3-5 hrs  5hrs +
##    1382     566     100
```



Observamos nuevamente una asimetria positiva, donde todos los sujetos manifestaron un uso moderado de dispositivos electronicos. Procedemos a crear una tabla de contingencia:

```
##             Insufficient_Weight Normal_Weight Obesity_Type_I Obesity_Type_II Obesity_Type_III Overwe
##
## 0-2 hrs   37                130           116           211            219             297
## 3-5 hrs   18                112           109            98             61              10
## 5hrs +     5                 18            31            20              1               0
```

```
##
##  Pearson's Chi-squared test
##
## data:  tabla_contingencia
```

```
## X-squared = 252.22, df = 14, p-value < 2.2e-16
```

La chi-cuadrado de independencia para evaluar la posible asociación entre el tiempo de uso de dispositivos electrónicos (TUE) y el nivel de obesidad (NObeyesdad)devolvio un valor de chi-cuadrado de 252.22 con 14 grados de libertad y un valor p < 2.2e-16.

El valor p nuevamente es una evidencia para rechazar la hipótesis nula de independencia. Se concluye que existe una asociación estadísticamente significativa entre el tiempo de uso de dispositivos electrónicos y el nivel de obesidad. Parece que el tiempo dedicado al uso de dispositivos electrónicos podría ser un factor relevante en el desarrollo de la obesidad. El análisis más profundo de la tabla de contingencia revela que:

- Mayor tiempo de uso, mayor obesidad: Las categorías de mayor tiempo de uso ("3-5 hrs" y "5hrs +") tienden a estar asociadas con niveles más altos de obesidad ("Obesity_Type_I," "Obesity_Type_II," "Obesity_Type_III").
- Menor tiempo de uso, menor obesidad: La categoría de menor tiempo de uso ("0-2 hrs") se asocia con mayor frecuencia a niveles de peso normal o insuficiente ("Insufficient_Weight," "Normal_Weight").

La prueba de chi-cuadrado no indica la dirección o la fuerza de la asociación, estos patrones sugieren una posible relación positiva entre el tiempo de uso de dispositivos electrónicos y el nivel de obesidad.

Seria conveniente utilizar medidas de asociación, como la *V de Cramer* para cuantificar la fuerza de esta asociación y realizar análisis adicionales - e.g. regresión ordinal- para modelar la relación entre estas variables teniendo en cuenta la naturaleza ordinal de TUE. Preliminarmente todo esto aporta evidencia sobre la importancia de considerar el tiempo de uso de dispositivos electrónicos en el estudio de la obesidad y la necesidad de investigar con mayor profundidad esta relación.

Se ha seleccionado el método de imputación múltiple (`mice`) con el algoritmo "Bayesian polytomous regression" (`polyreg`).

Nos fundamentamos en:

1. La naturaleza ordinal de la variable:

`TUE` es una variable ordinal que clasifica el tiempo de uso de dispositivos en tres categorías ordenadas: "0-2 horas," "3-5 horas" y "Más de 5 horas." El método `polyreg` es especialmente adecuado para variables ordinales, ya que utiliza la regresión polinomial para predecir la probabilidad de que una observación pertenezca a cada categoría, teniendo en cuenta el orden entre ellas.

2. El manejo de la incertidumbre:

La imputación múltiple genera varios conjuntos de datos imputados, lo que permite tener en cuenta la incertidumbre asociada a la imputación de los valores faltantes.

3. Relación con otras variables:

Se presume que existe una relación entre el tiempo de uso de dispositivos electrónicos y otras variables en el conjunto de datos, como el nivel de obesidad (`NObeyesdad`) y los hábitos de actividad física (`FAF`). `mice` utiliza la información de todas las variables en el conjunto de datos para predecir los valores faltantes, facilitando capturar relaciones complejas y mejorar la precisión de la imputación.

4. La flexibilidad:

`mice` ofrece la flexibilidad de especificar diferentes modelos de imputación para cada variable. En este caso, `polyreg` se ha seleccionado como el método más adecuado para `TUE`, dada su naturaleza ordinal.

5. El rigor académico:

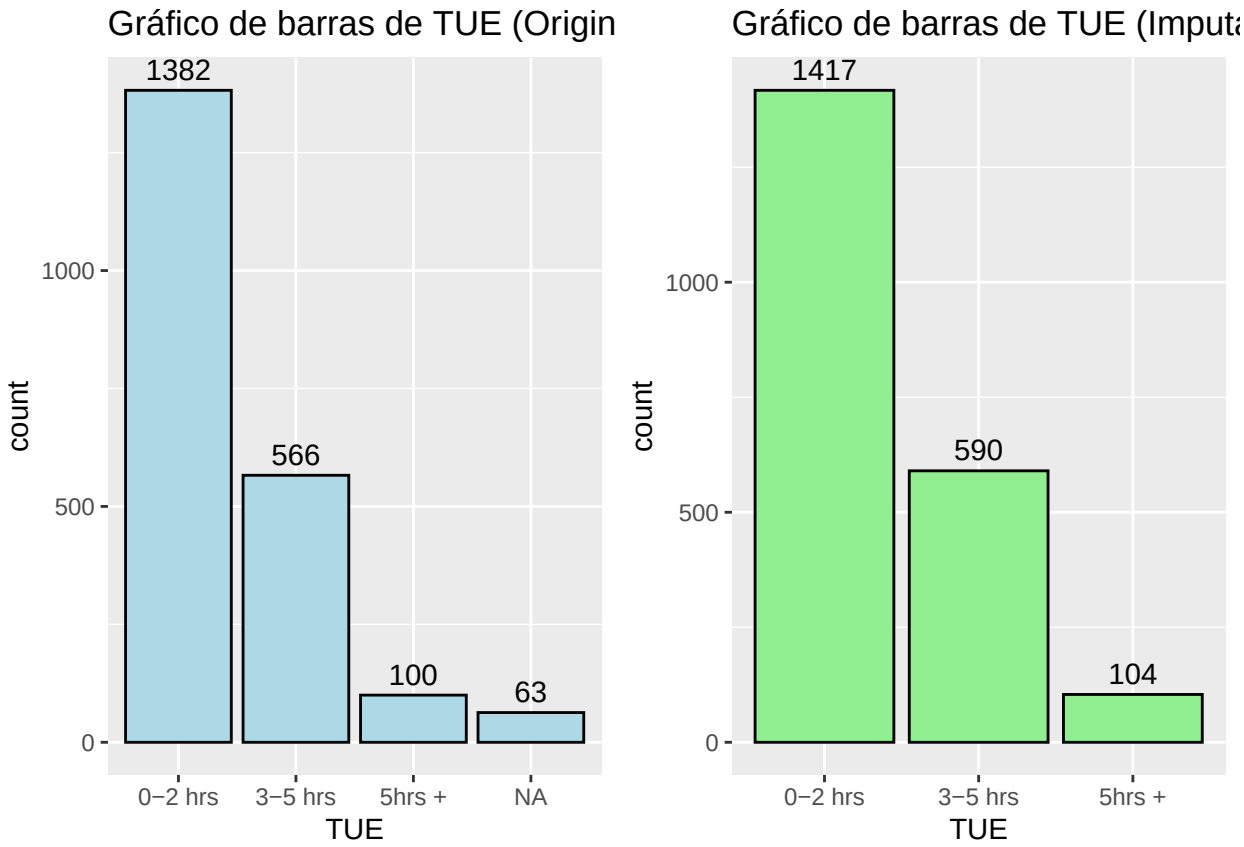
La imputación múltiple es un método aceptado en la comunidad científica.

```
##
##  iter imp variable
##   1   1  TUE  Age_ordinal
##   1   2  TUE  Age_ordinal
```

```
##   1   3  TUE  Age_ordinal
##   1   4  TUE  Age_ordinal
##   1   5  TUE  Age_ordinal
##   2   1  TUE  Age_ordinal
##   2   2  TUE  Age_ordinal
##   2   3  TUE  Age_ordinal
##   2   4  TUE  Age_ordinal
##   2   5  TUE  Age_ordinal
##   3   1  TUE  Age_ordinal
##   3   2  TUE  Age_ordinal
##   3   3  TUE  Age_ordinal
##   3   4  TUE  Age_ordinal
##   3   5  TUE  Age_ordinal
##   4   1  TUE  Age_ordinal
##   4   2  TUE  Age_ordinal
##   4   3  TUE  Age_ordinal
##   4   4  TUE  Age_ordinal
##   4   5  TUE  Age_ordinal
##   5   1  TUE  Age_ordinal
##   5   2  TUE  Age_ordinal
##   5   3  TUE  Age_ordinal
##   5   4  TUE  Age_ordinal
##   5   5  TUE  Age_ordinal
```

```
## 0-2 hrs 3-5 hrs  5hrs +
##    1417     590     104
```

Evaluaremos la calidad de la imputación mediante la comparación de la distribución de la variable imputada con la distribución original y se consideran otros métodos de imputación de ser necesario:



Y las graficas de barras nos demuestran que hemos podido imputar los valores ausentes sin alterar significativamente la distribucion de los datos.

Age La variable contiene 63 valores NA que supone el 3% de los datos.
Procederemos a un analisis exploratorio, empezando por conocer si la variable tiene una distirbucion normal y evualuando si la hipótesis nula de que los datos provienen de una distribución normal.
Para ello emplearemos la prueba de Shapiro-Wilk:

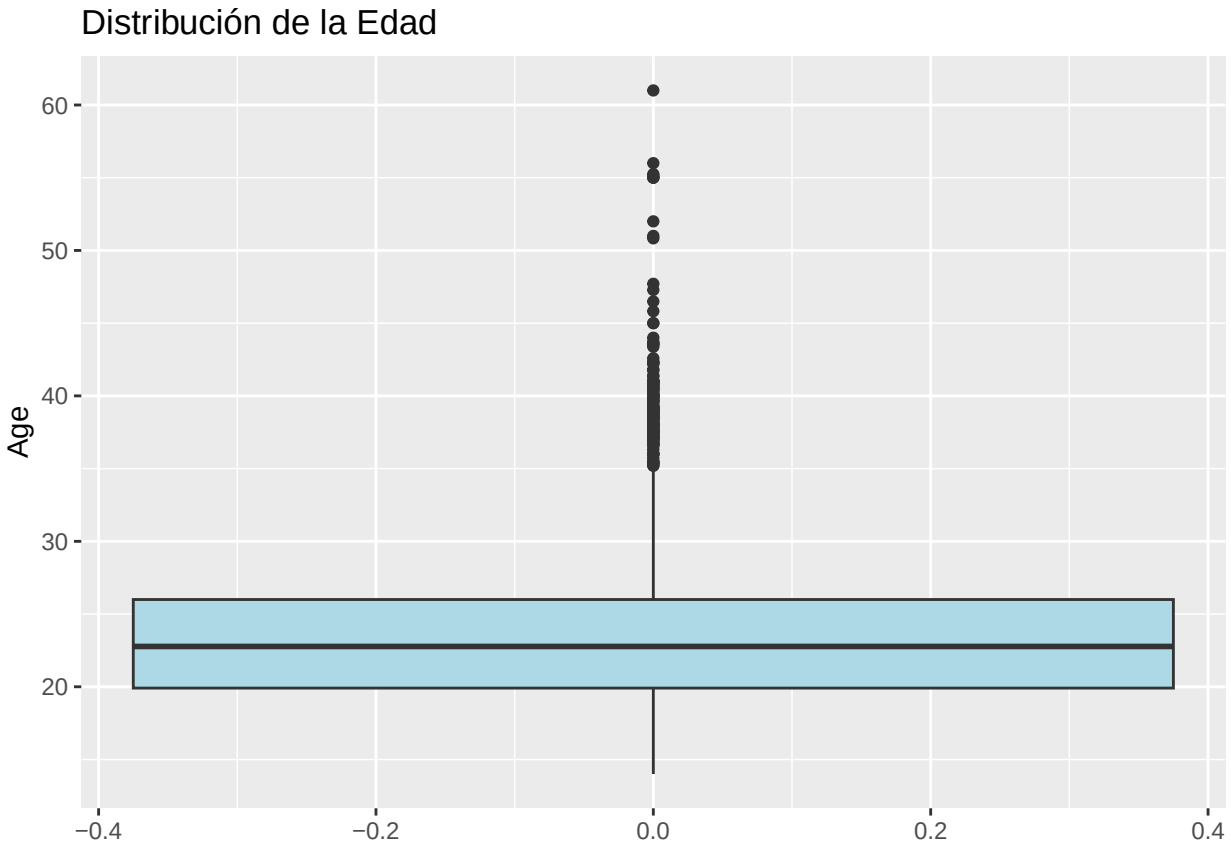
```
##
##  Shapiro-Wilk normality test
##
## data:  df$Age
## W = 0.86664, p-value < 2.2e-16
```

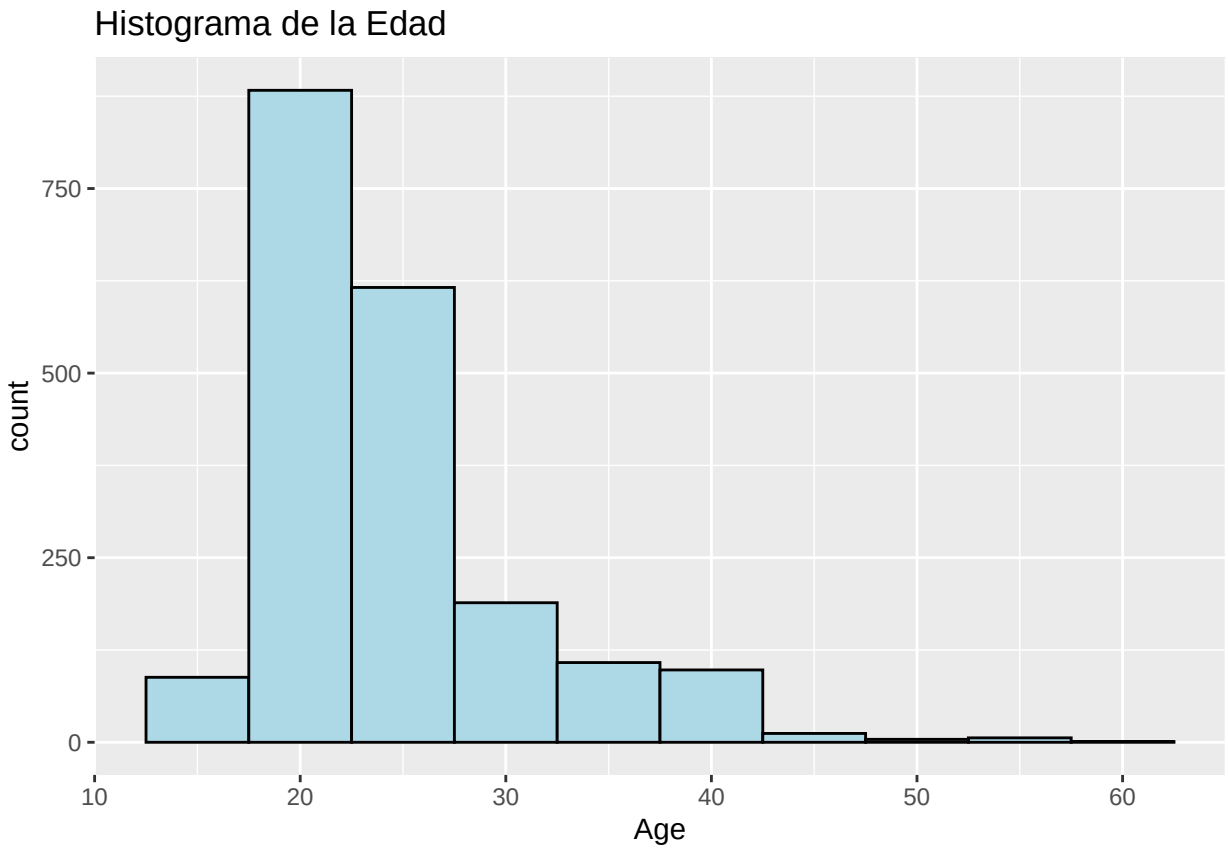
Con el objetivo de determinar si la variable "Age" (edad) se ajustaba a una distribución normal se realizó la prueba de *Shapiro-Wilk*. Mostro un estadístico $\mathbf{W = 0.86664}$ (alejado de 1) y un valor $\mathbf{p < 2.2e\text{-}16}$.

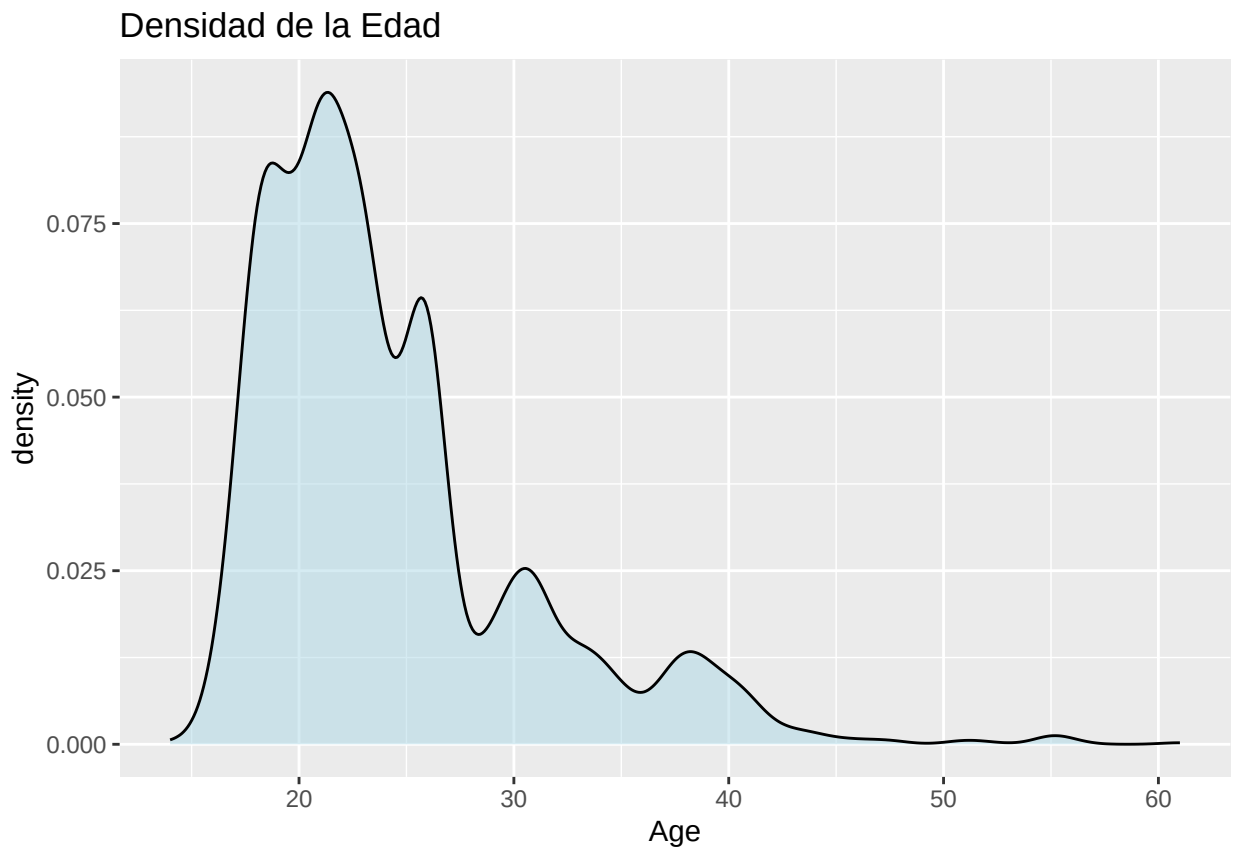
El valor p es extremadamente pequeño. Indica una fuerte evidencia en contra de la hipótesis nula de que la variable "Age" sigue una distribución normal. Se rechaza la hipótesis nula y se concluye que la distribución de la edad en la muestra **no se ajusta a una distribución normal**.

Esta desviación de la normalidad puede tener implicaciones para la selección de pruebas estadísticas paramétricas que asumen la normalidad de los datos. En consecuencia, se considerarán métodos alternativos, como transformaciones de la variable o el uso de pruebas no paramétricas, para asegurar la validez de los análisis posteriores. El hecho de que la prueba haya rechazado la hipótesis de normalidad para la variable `Age` sugiere que la imputación a la media o a otras medidas de tendencia central podría no ser la estrategia más adecuada.

Se complementará este análisis con la visualización de la distribución de la variable “Age” mediante histogramas y gráficos de densidad para obtener una idea más completa de su forma y características.




Histograma de la Edad
count
750
500
250
0
10
20
30
40
50
60
Age



Los graficos confirman que la distribución de la variable Age (edad) tiene una asimetría hacia la derecha, (asimetría positiva). La mayor frecuencia de observaciones se encuentra en el rango de edad entre 15 y 27 años, aproximadamente. A partir de los 30 años, la frecuencia disminuye y la cola de la distribución se extiende hacia la derecha, con menos observaciones en las edades mayores.

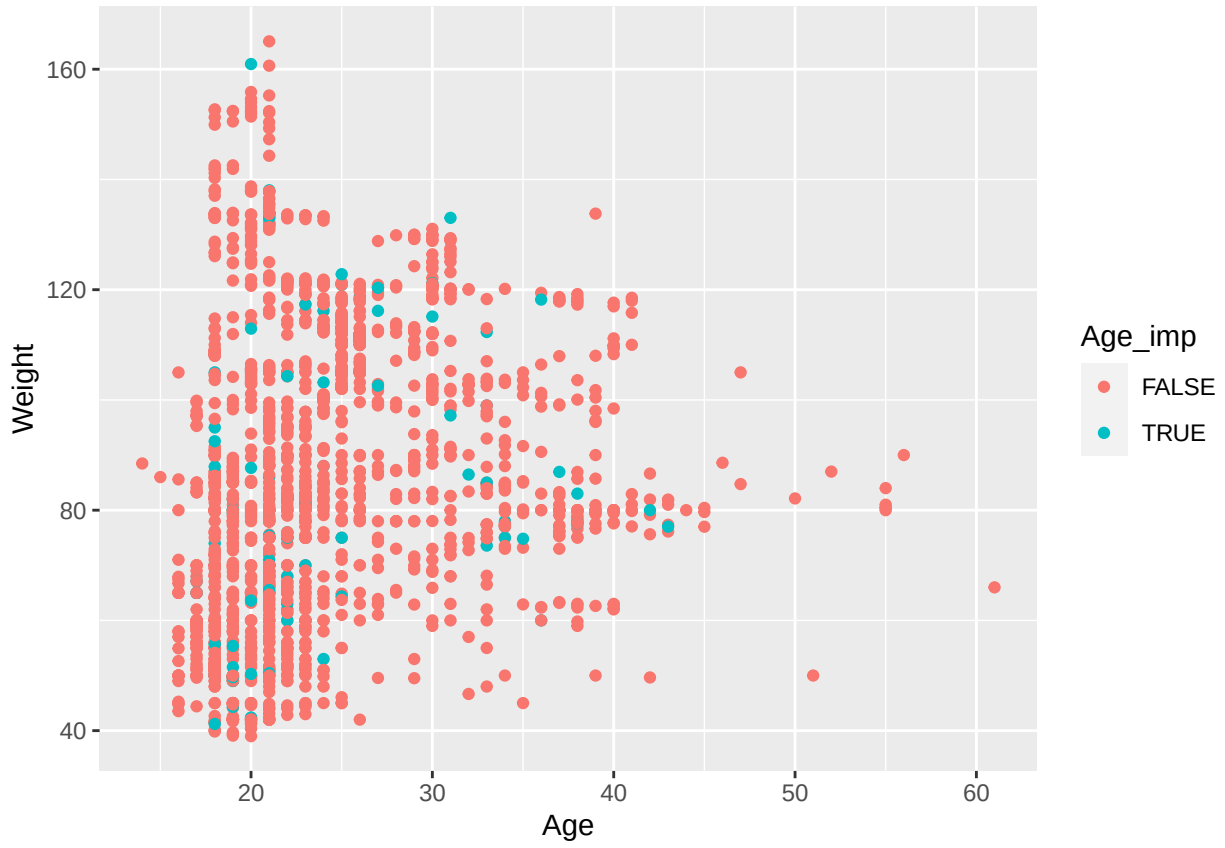
Esto coincide con la asimetría hacia la derecha observada en el gráfico de densidad anterior. La mayor parte de los datos se concentra en la parte izquierda de la distribución (edades menores) y la cola se extiende hacia la derecha (edades mayores).

Implicaciones:

Esta concentración de datos en las edades menores refuerza la idea de que la imputación a la media podría no ser la mejor opción para la variable 'Age.'
La media se vería afectada por los valores en la cola de la distribución (edades mayores), lo que podría llevar a una sobreestimación de la edad al imputar valores faltantes. Imputaremos mediante KNN, útil para imputar valores numéricos en distribuciones no normales.

Y confirmamos que efectivamente la media de edad (24) es la misma que obtuvimos empleando el metodo `summary()` y que no persisten los valores perdidos.

Evaluaremos la calidad de la imputación mediante la comparación de la distribución de la variable imputada con la distribución original y se considerarann otros métodos de imputación de ser necesario:




```
##
## Call:
## lm(formula = Weight ~ Age + Age_imp, data = df_completo_kNN_VIM)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.513 -20.526  -2.741  20.129  80.892
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 67.12260    2.20345  30.462   <2e-16 ***
## Age          0.81158    0.08884   9.135   <2e-16 ***
## Age_impTRUE -1.69266    2.55351  -0.663    0.507
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.62 on 2108 degrees of freedom
## Multiple R-squared:  0.03826,    Adjusted R-squared:  0.03735
## F-statistic: 41.93 on 2 and 2108 DF,  p-value: < 2.2e-16
```

El gráfico de dispersión muestra la relación entre la edad (Age) y el peso (Weight) de los individuos en el conjunto de datos, diferenciando entre aquellos con valores originales de edad (rojo) y aquellos con valores imputados mediante el método kNN (azul). Se observa una tendencia general a que el peso aumente con la edad, aunque con una considerable variabilidad individual. Los puntos azules, que representan los valores imputados, se distribuyen a lo largo de todo el rango de edad y peso, lo que sugiere que la imputación kNN ha logrado generar valores plausibles que se ajustan a la distribución general de los datos.

Se aprecia una mayor concentración de puntos azules en las edades más jóvenes, lo que podría indicar que la imputación kNN ha tenido un mayor impacto en este grupo de edad. Podría deberse a una mayor cantidad de valores faltantes en las edades más jóvenes o a una mayor dificultad para predecir la edad en este grupo debido a la mayor variabilidad en el peso.

Asimismo se ajustó un modelo de regresión lineal para analizar la relación entre la edad (`Age`), la imputación de la edad (`Age_imp`) y el peso (`Weight`):

Coeficientes:

El modelo muestra una relación positiva y significativa entre la edad y el peso (= 0.84, p < 2e-16). Esto indica que, por cada año que aumenta la edad, el peso estimado aumenta en 0.84 kg, manteniendo constante la variable `Age_imp`.

El coeficiente de `Age_impTRUE` (-1.82) no fue estadísticamente significativo (p = 0.478). Esto sugiere que no hay evidencia suficiente para afirmar que existe una diferencia significativa en el peso entre los individuos con edad imputada y aquellos con edad original después de controlar por la edad.

Bondad de ajuste:

El modelo explica una pequeña proporción de la varianza en el peso (R-cuadrado ajustado = 0.03978)queindica que la edad y la imputación de la edad no son los únicos factores que influyen en el peso.

Significancia del modelo:

El modelo en su conjunto es estadísticamente significativo (p < 2.2e-16), lo que indica que al menos una de las variables predictoras tiene una relación significativa con el peso.

Efectos no deseados de la imputacion?

El resultado del modelo lineal *(lm(Weight ~ Age + Age_imp, data = df_completo_kNN_VIM))* indica no se encontró una diferencia significativa en el peso entre los individuos con edad imputada y los que tienen edad original, después de controlar por la edad. El coeficiente de Age_impTRUE no fue estadísticamente significativo (p = 0.478). Esto significa que, una vez que se tiene en cuenta la edad de la persona, el hecho de que su edad haya sido imputada o no, **no tiene un efecto significativo en su peso**, por lo que consideramos que la imputacion ha transcurrido correctamente.

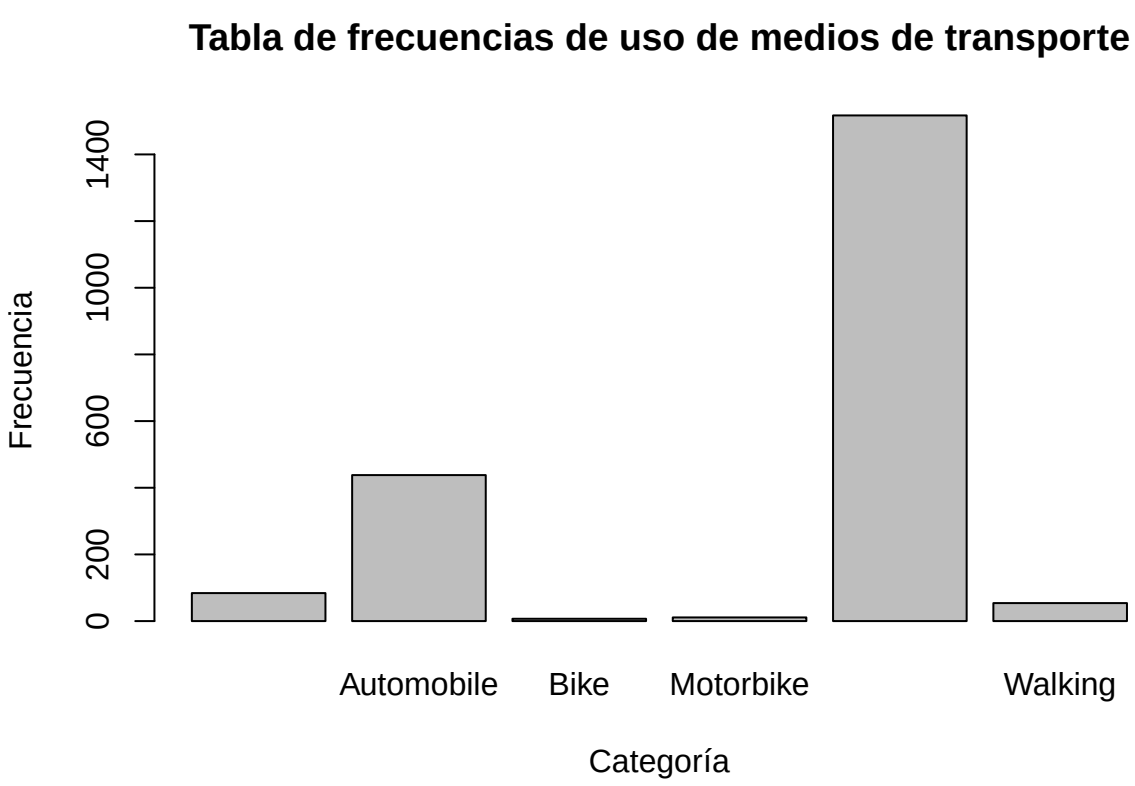
MTRANS La variable `MTRANS` categoriza el medio de transporte que los individuos utilizan habitualmente, es nominal cualitativa politómica con un 5% de valores ausentes registra las siguientes opciones: *"Automobile," "Motorbike," "Bike," "Public Transportation"* y *"Walking"*. Contiene 84 valores ausentes.

Al ser una variable nominal las categorías no presentan un orden intrínseco. El análisis de `MTRANS` se centrará en la exploración de su distribución y su posible asociación con otras variables de interés, como el nivel de actividad física (`FAF`) y el nivel de obesidad (`NObeyesdad`). Emplearemos métodos estadísticos apropiados para variables nominales como tablas de contingencia, pruebas de chi-cuadrado y medidas de asociación.

La imputación de los valores faltantes en `MTRANS` se abordará en una fase posterior, considerando la naturaleza politómica de la variable y la posible existencia de desequilibrios entre las categorías.

Procedemos al analisis exploratorio:

```
##
##                                Automobile                      Bike
##                    84                438                         7
##             Motorbike Public_Transportation                   Walking
##                    11                1517                        54
```



La tabla de frecuencias nos muestra un nuevo desequilibrio que deja a las claras que los individuos apenas realizan trayectos evitando medios de transporte en favor de la actividad fisica. Procedemos a crear una tabla de contigencias:

```
##                          Insufficient_Weight Normal_Weight Obesity_Type_I Obesity_Type_II Obesity_T
##
##                        1                  13             7             17              12
## Automobile            12                  43            41            101              91
## Bike                   0                   0             4              0               1
## Motorbike              0                   0             6              3               0
## Public_Transportation 48                 206           184            213             182
## Walking                2                   5            30              2               0
```

```
##
##  Pearson's Chi-squared test
##
## data:  tabla_contingencia_MTRANS
## X-squared = 292.42, df = 35, p-value < 2.2e-16
```

Los resultados de la prueba mostraron un valor de chi-cuadrado de 292.42 con 35 grados de libertad y un valor p $<$ 2.2e-16.

El valor p extremadamente bajo indica una fuerte evidencia para rechazar la hipótesis nula de independencia. Se concluye que existe una asociación estadísticamente significativa entre el medio de transporte habitual y el nivel de obesidad. Este hallazgo sugiere que el medio de transporte que las personas utilizan con mayor frecuencia podría ser un factor relevante en el desarrollo de la obesidad. Un análisis más profundo de la tabla de contingencia revela un desbalance en las categorías de MTRANS, con una mayor proporción de individuos que utilizan el transporte público ("Public_Transportation") y el automóvil ("Automobile"). A

pesar del desbalance, la prueba de chi-cuadrado ha detectado una asociación significativa con el nivel de obesidad.
Podria ser muy relevante saber si el uso de medios de transporte activos como la bicicleta ("Bike") o caminar ("Walking") se asocian con una menor frecuencia de obesidad.

Se ha optado por utilizar el modelo `missForest`, un algoritmo de imputación basado en *Random Forest* que ya ha sido implementado con éxito en la imputación de la variable `family_history_with_overweight`.

Ello se justifica por las siguientes razones:

1. El buen rendimiento previo:

En la imputación de `family_history_with_overweight`, `missForest` obtuvo un error de imputación out-of-bag (OOB) bajo:

- `NRMSE` (Normalized Root Mean Squared Error): 0.04584933
- `PFC` (Proportion of Falsely Classified): 0.11181195

Estos valores indican un buen rendimiento del algoritmo con un error bajo en la imputación de variables numéricas (`NRMSE`) y una proporción relativamente baja de clasificaciones incorrectas en variables categóricas (`PFC`).

2. La adecuación a la variable `MTRANS`:

`MTRANS` es una variable categórica nominal, y `missForest` es capaz de manejar este tipo de variables de forma adecuada. El algoritmo utiliza la información de todas las variables en el conjunto de datos para predecir los valores faltantes, lo que puede mejorar la precisión de la imputación en variables categóricas con múltiples niveles, como `MTRANS`.

3. Validez y fiabilidad:

`missForest` es un método no paramétrico que no requiere asumir una distribución específica para los datos faltantes. Esto lo hace robusto a diferentes patrones de datos faltantes y a la presencia de valores atípicos.

4. Eficiencia:

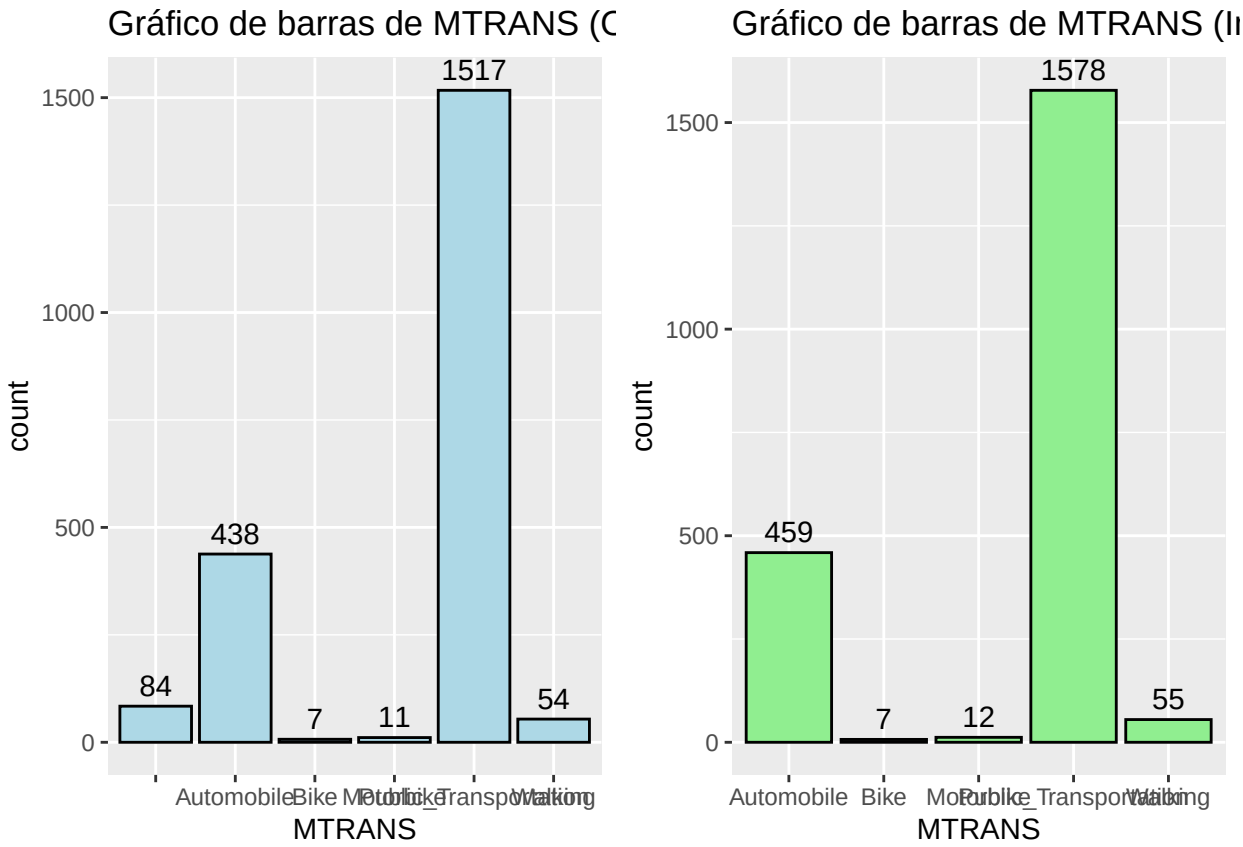
`missForest` ya ha sido implementado y evaluado en el conjunto de datos, lo que permite reutilizar el modelo existente y ahorrar tiempo de computación.

Otras consideraciones:

- **Evaluación de la imputación:** A pesar del buen rendimiento previo, es recomendable evaluar la calidad de la imputación de `MTRANS` mediante la comparación de la distribución de la variable imputada con la distribución original o mediante la comparación del error de imputación OOB con el de otros métodos de imputación.
- **Limitaciones:** Si el patrón de datos faltantes en `MTRANS` es muy diferente al de `family_history_with_overweight`, o si la relación entre `MTRANS` y las demás variables es diferente el rendimiento de `missForest` podría no ser tan bueno como en la imputación anterior.

```
##            Automobile                  Bike             Motorbike
##                   459                     7                    12
## Public_Transportation               Walking
##                  1578                    55
```

Observamos que efectivamente las categorias son las adecuadas. Analizamos como ha podido afectar a la distribucion o patron esta imputacion:



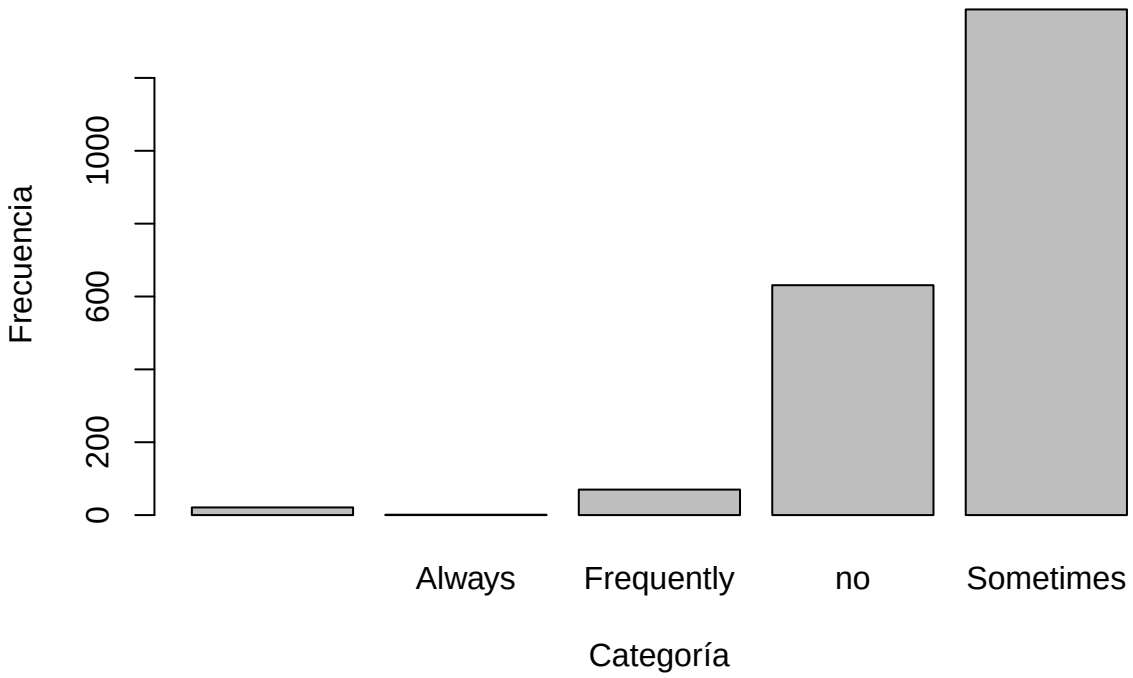
Y vemos claramente que efectivamente no se ha alterado la distribucion de los valores sustancialmente y que la mayoria de las imputaciones se ha producido sobre la categoria mas habitual

CALC La variable CALC clasifica la frecuencia de consumo de alcohol en cuatro categorías: *"I do not drink," "Sometimes," "Frequently" y "Always"*, con un 3% de valores ausentes, de naturaleza ordinal donde las categorías reflejan un orden creciente en la frecuencia de consumo de alcohol.

El análisis de CALC se centrará en la exploración de su distribución y su posible asociación con otras variables de interés, como el nivel de obesidad (NObeyesdad) y los hábitos alimenticios. Emplearemos métodos estadísticos apropiados para variables ordinales, como tablas de contingencia, pruebas de chi-cuadrado y medidas de asociación considerando el orden intrínseco de las categorías. La imputación de los valores faltantes en CALC se abordará en una fase posterior teniendo en cuenta la naturaleza ordinal de la variable y la posible existencia de desequilibrios entre las categorías.

Procedemos al analisis exploratorio:

```
##
##                Always Frequently         no  Sometimes
##         21          1         70        631       1388
```



Nuevamente tenemos una asimetria a la izquierda. Procedemos con una tabla de contingencia:

```
##                Insufficient_Weight Normal_Weight Obesity_Type_I Obesity_Type_II Obesity_Type_III Ov
##
##               1                   3             2              4               3                3
## Always        0                   0             1              0               0                0
## Frequently    3                   1            15             14               2                0
## no           17                 115           101            155              70                1
## Sometimes    42                 148           153            163             211              311
```

```
##
##  Pearson's Chi-squared test
##
## data:  tabla_contingencia_CALC
## X-squared = 328.91, df = 28, p-value < 2.2e-16
```

La chi-cuadrado de independencia para evaluar la posible asociación entre la frecuencia de consumo de alcohol (`CALC`) y el nivel de obesidad (`NObeyesdad`) devolvio un valor de chi-cuadrado de 328.91 con 28 grados de libertad y un valor p $<$ 2.2e-16.

El valor p nos sirve nuevamente para rechazar la hipótesis nula de independencia. Por lo tanto, se concluye que existe una asociación estadísticamente significativa entre la frecuencia de consumo de alcohol y el nivel de obesidad.

Esto conllevaria que la frecuencia con la que las personas consumen alcohol podría ser un factor relevante en el desarrollo de la obesidad. Vamos a imputar los valores faltantes mediante un modelo de regresion logistica. Esta decisión se fundamenta en las siguientes consideraciones:

1. La naturaleza politómica de la variable:

`CALC` es una variable politómica con cuatro categorías. El modelo multinomial es especialmente adecuado para variables categóricas con más de dos niveles, ya que permite predecir la probabilidad de que una observación pertenezca a cada una de las categorías.

2. El orden:

Si bien `CALC` es ordinal, el modelo multinomial no impone restricciones en el orden de las categorías, lo que permite flexibilidad en la interpretación de los resultados. Si bien existen modelos específicos para variables ordinales como la regresión ordinal, el modelo multinomial ofrece una alternativa que puede ser útil cuando no se desea imponer un orden estricto a las categorías o cuando se quiere explorar la relación entre cada categoría y las variables predictoras de forma individual.

3. La relación con otras variables:

Se presume que existe una relación entre la frecuencia de consumo de alcohol y otras variables en el conjunto de datos, como el nivel de obesidad (`NObeyesdad`) y los hábitos alimenticios. El modelo multinomial permite incluir múltiples variables predictoras y capturar relaciones complejas entre ellas, lo que puede mejorar la precisión de la imputación.

4. La interpretación:

El modelo multinomial proporciona información sobre la influencia de cada variable predictora en la probabilidad de pertenecer a cada categoría de `CALC`habilitando un análisis más detallado de la relación entre las variables y facilitando la interpretación de los resultados.

5. La complejidad del modelo:

El modelo multinomial es un modelo relativamente complejo que puede requerir un mayor tiempo de computación y un análisis más cuidadoso de los resultados. Sin embargo consideramos que está justificado en este caso, dada la naturaleza politómica de la variable `CALC` y la necesidad de capturar las relaciones complejas entre las variables.

Procedemos aplicando el modelo:

```
## # weights:  116 (84 variable)
## initial  value 2811.404964
## iter  10 value 1512.688334
## iter  20 value 1277.524021
## iter  30 value 1204.550643
## iter  40 value 1191.144681
## iter  50 value 1183.442688
## iter  60 value 1180.624935
## iter  70 value 1179.037069
## iter  80 value 1178.417925
## iter  90 value 1178.361483
## final  value 1178.361290
## converged
```

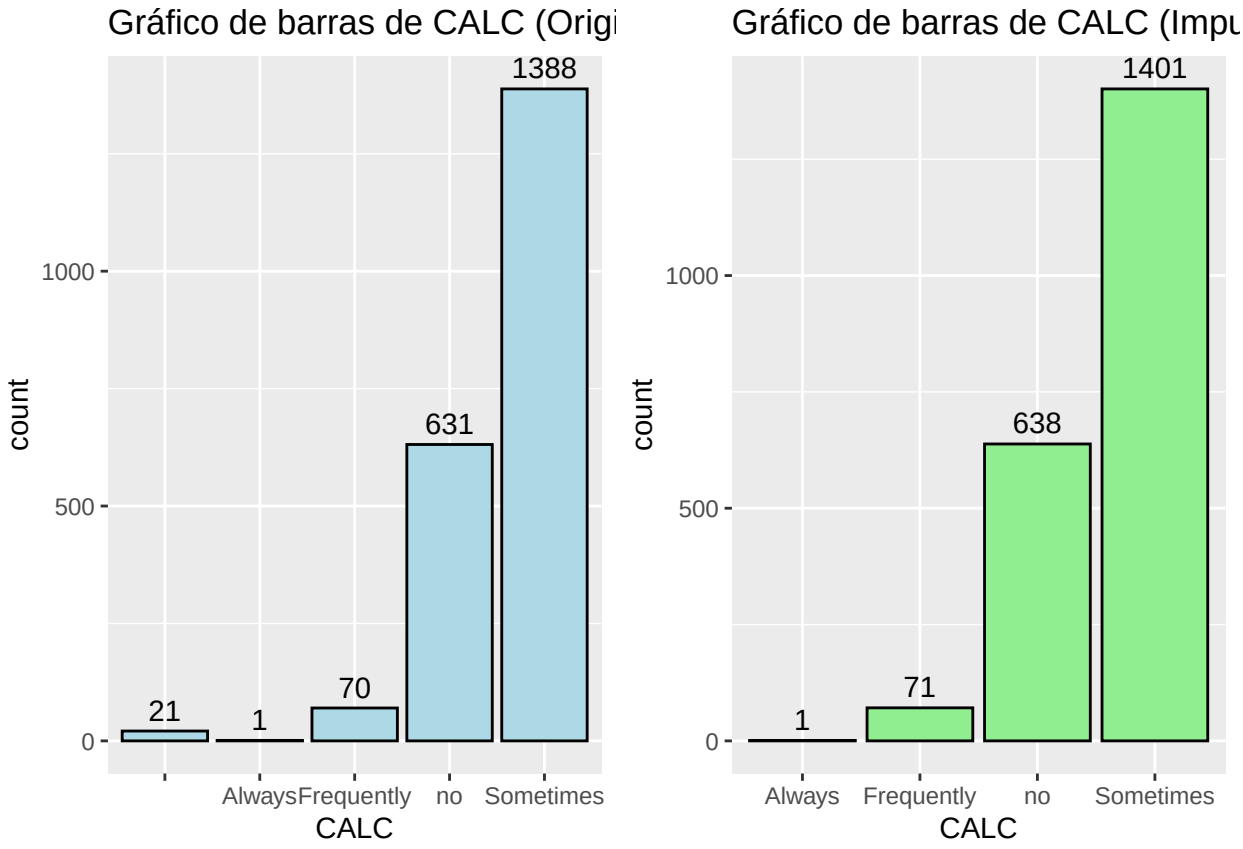
```
##     Always Frequently         no  Sometimes       NA's
##          1         71        638       1400          1
```

A pesar de la capacidad del modelo para capturar relaciones complejas entre las variables, un valor faltante en CALC no pudo ser imputado con este método. Para completar la imputación se optara por imputar el valor faltante restante utilizando la moda, es decir, la categoría más frecuente en la variable CALC. Esta estrategia, aunque simple es adecuada en este caso debido a la presencia de un único valor faltante y al desequilibrio observado en la distribución de la variable, donde la categoría “sometimes” (a veces) es la más frecuente. La combinación de la imputación multinomial y la imputación por la moda permite aprovechar las ventajas de ambos métodos: la capacidad del modelo multinomial para capturar relaciones complejas y la simplicidad y eficiencia de la imputación por la moda para un único valor faltante.

En todo caso se evaluará la calidad de la imputación mediante la comparación de la distribución de la variable imputada con la distribución original y se considerarán otros métodos de imputación si es necesario.

```
##     Always Frequently         no  Sometimes
##          1         71        638       1401
```

Y comprobamos que el valor ausente ha sido asignado a la moda que era claramente 'sometimes.' Evaluamos la imputacion:

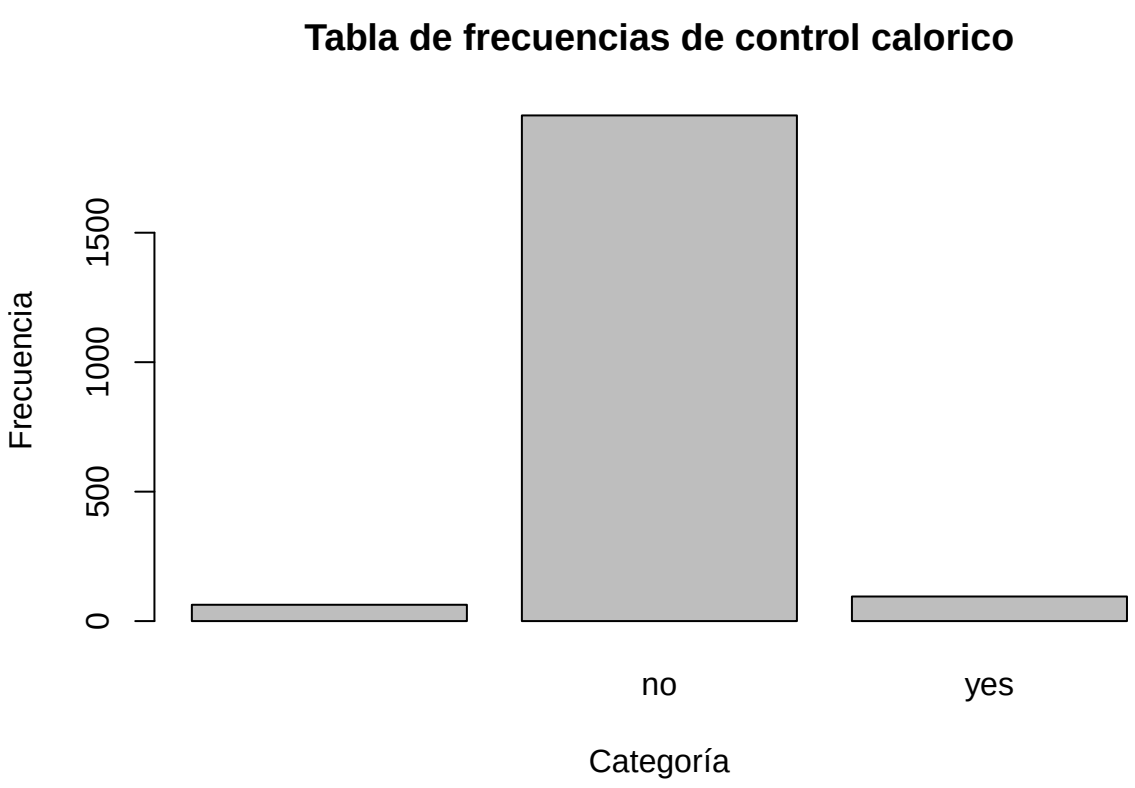


Los gráficos de barras muestran la distribución de la frecuencia de consumo de alcohol (CALC) antes y después de la imputación. Se observa que la imputación ha mantenido las proporciones entre las categorías, sin introducir cambios significativos en la distribución de la variable.

SCC Se trata del registro de si el individuo lleva a cabo un control de las calorías que consume en su dieta. Esta variable nominal cualitativa dicotómica, con un 3% de valores ausentes, se codifica con "yes" o "no" para indicar la presencia o ausencia de control de calorias ingeridas

Dada su naturaleza categórica, el análisis de SCC se centrará en la exploración de su distribución y su posible asociación con otras variables relevantes como el nivel de obesidad (NObeyesdad) y los hábitos alimenticios. Se utilizarán métodos estadísticos apropiados para variables dicotómicas, como tablas de contingencia, pruebas de chi-cuadrado y medidas de asociación.

```
##
##        no  yes
##   63 1953   95
```



Se observa que la gran mayoría de los individuos no realizaban un control de las calorías que consumen en la dieta ("No"), mientras que una minoría responde afirmativamente ("Sí"). Este desequilibrio puede tener implicaciones importantes para el análisis posterior,especialmente si utilizamos la variable `SCC` en modelos predictivos o análisis de asociación. La clase minoritaria ("Sí") podría no tener suficiente representación para detectar patrones o diferencias significativas, lo que podría sesgar los resultados del análisis. Se consideraran estrategias para manejar este desequilibrio como `SMOTE` de ser necesarias. Procedemos con una tabla de contingencia para conocer como se relaciona SCC con los niveles de obesidad:

```
##          Insufficient_Weight Normal_Weight Obesity_Type_I Obesity_Type_II Obesity_Type_III Overweigh
##
##        2                   8            10             12               9                12
## no    60                 238           232            322             276               303
## yes    1                  21            30              2               1                 0
```

```
##
##  Pearson's Chi-squared test
##
## data:  tabla_contingencia_SCC
## X-squared = 129.89, df = 14, p-value < 2.2e-16
```

La prueba de independencia agenciada a evaluar la posible asociación entre el control de calorías en la dieta (`SCC`) y el nivel de obesidad (`NObeyesdad`) mostro un valor de chi-cuadrado de 129.89 con 14 grados de libertad y un valor p $<$ 2.2e-16. Esto conlleva evidencia para rechazar la hipótesis nula de independencia. Por lo tanto existe una asociación estadísticamente significativa entre el control de calorías y el nivel de obesidad. Este hallazgo sugiere que el control de calorías podría ser un factor relevante en el desarrollo de la obesidad. U

A pesar del desequilibrio observado en la variable `SCC`, con una mayor proporción de individuos que no

realizan control de calorías, la prueba de *chi-cuadrado* ha detectado una asociación significativa con el nivel de obesidad. Parece destacable por pura logica la importancia de considerar el control de calorías como un factor potencial en el estudio de la obesidad. Para abordar este problema se ha seleccionado el método de imputación múltiple (`mice`) con el algoritmo de regresión logística (`logreg`). Esta decisión se fundamenta en las siguientes consideraciones:

1. Naturaleza dicotómica:

`SCC` es una variable dicotómica que toma dos valores posibles *"yes"/"no"*. El método `logreg` es particularmente adecuado para variables binarias ya que utiliza la regresión logística para predecir la probabilidad de que una observación pertenezca a una de las dos categorías.

2. La incertidumbre:

La imputación múltiple genera varios conjuntos de datos imputados, lo que permite tener en cuenta la incertidumbre asociada a la imputación de los valores faltantes. Esto proporciona resultados más robustos y evita subestimar la varianza de la variable imputada.

3. La relación con otras variables:

Es plausible que exista una relación entre el control de calorías y otras variables en el conjunto de datos, como el nivel de obesidad (`NObeyesdad`) y los hábitos alimenticios (`FAVC`, `FCVC`, `CAEC`). `mice` utiliza la información de todas las variables en el conjunto de datos para predecir los valores faltantes, lo que permite capturar relaciones complejas y mejorar la precisión de la imputación.

4. La flexibilidad:

`mice` ofrece la flexibilidad de especificar diferentes modelos de imputación para cada variable. En este caso, `logreg` parece la eleccion.

5. Validez:

La imputación múltiple es un método ampliamente reconocido y aceptado en la comunidad científica, lo que fortalece la validez y el rigor del presente estudio. Procedemos a aplicar el modelo:

```
##
##  iter imp variable
##   1   1  SCC
##   1   2  SCC
##   1   3  SCC
##   1   4  SCC
##   1   5  SCC
##   2   1  SCC
##   2   2  SCC
##   2   3  SCC
##   2   4  SCC
##   2   5  SCC
##   3   1  SCC
##   3   2  SCC
##   3   3  SCC
##   3   4  SCC
##   3   5  SCC
##   4   1  SCC
##   4   2  SCC
##   4   3  SCC
##   4   4  SCC
##   4   5  SCC
##   5   1  SCC
##   5   2  SCC
```

```
##   5   3  SCC
##   5   4  SCC
##   5   5  SCC
```

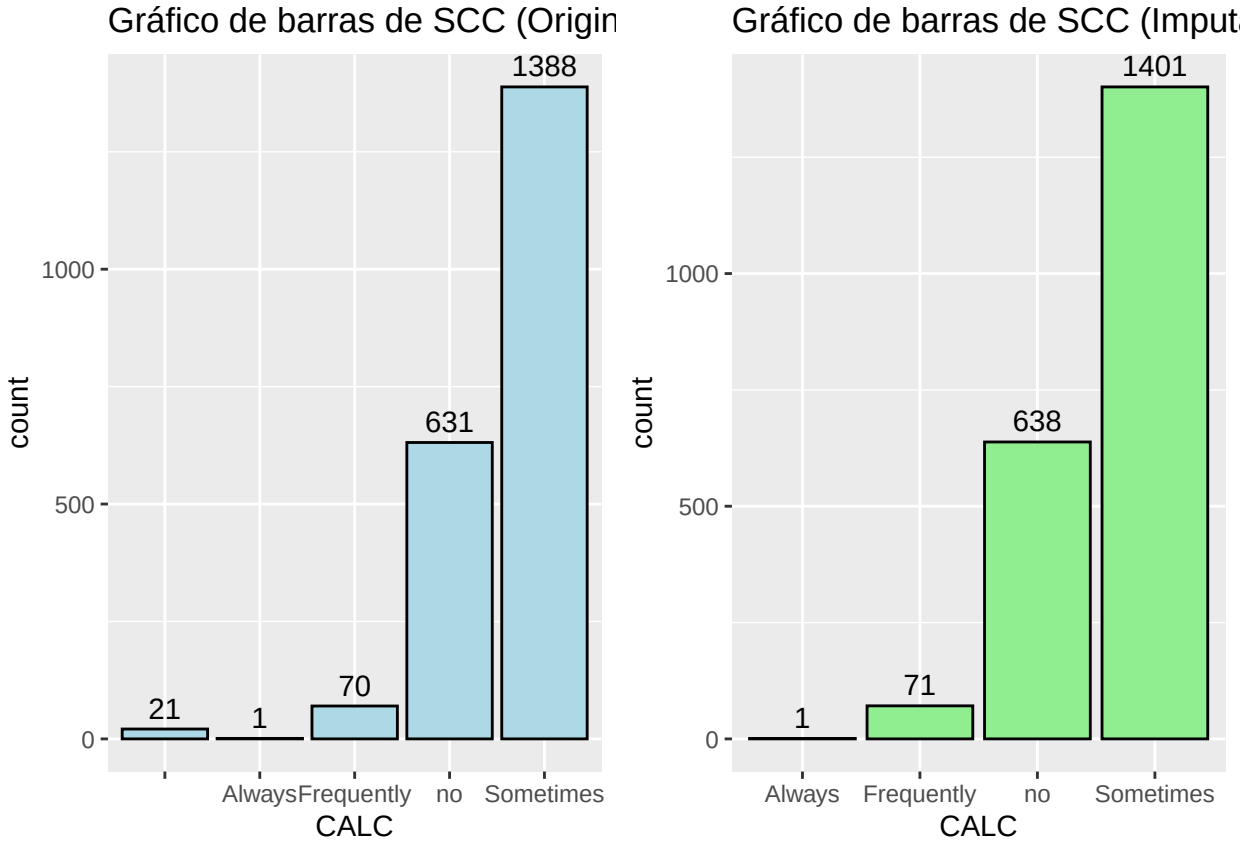
```
##  Factor w/ 2 levels "no","yes": 1 2 1 1 1 1 1 1 1 1 ...
```

```
##   no  yes
## 2009  102
```

Vamos a comprobar el modo que la imputacion ha afectado a la distribucion con este metodo, teniendo en cuenta que hemos tenido que prescendir de algunos atributos para poder llevar a cabo la imputacion:

```
##
##        no  yes
##   63 1953   95
```

```
##
##   no  yes
## 2009  102
```

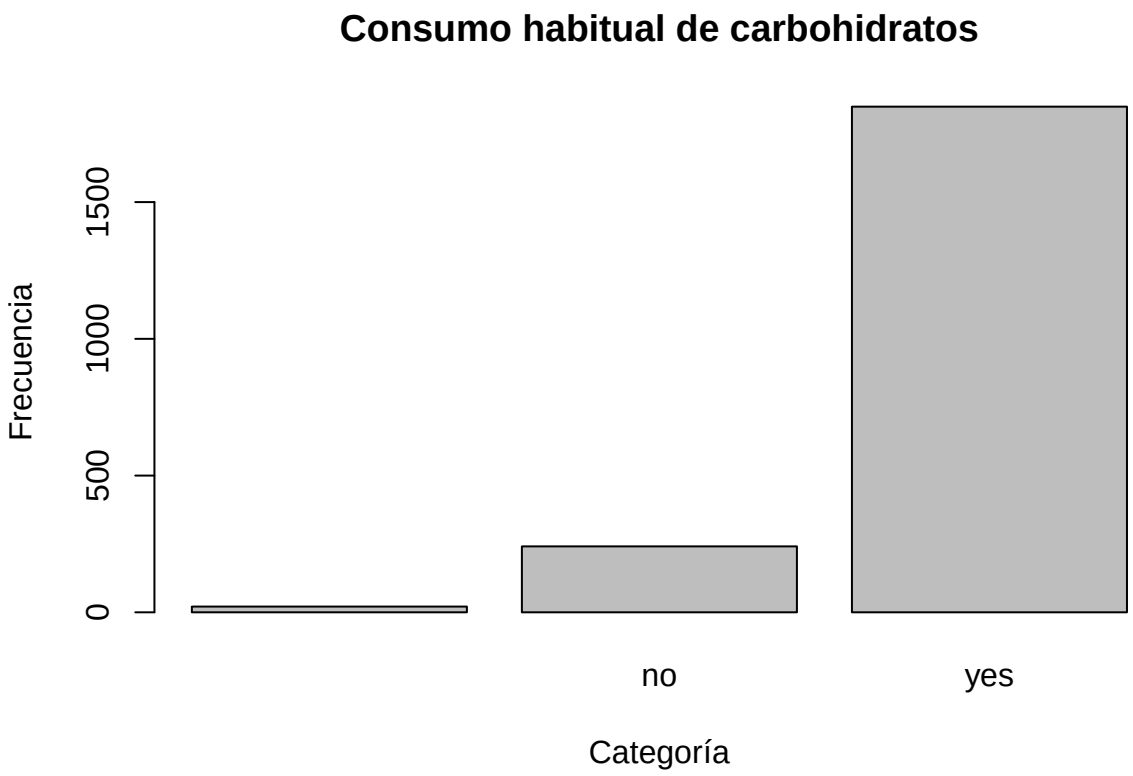


Y concluimos con esta variable asumiendo que la imputacion no ha afectado las proporciones.

FAVC Esta variable que se refiere al consumo de carbohidratos en la dieta contiene 21 valores perdidos, concretamente valores en blanco o ”“. Dado que `FAVC` (consumo habitual de carbohidratos) es una variable binaria (*”yes” / ”no”*), la regresión logística podria ser un método adecuado para imputar los valores faltantes. Este método predice la probabilidad de que FAVC sea”yes” basándose en otras variables del conjunto de datos, y luego utiliza esa probabilidad para imputar los valores faltantes. Podriamos emplear los resultados obtenidos con el modelo anterior, pero optamos por implementar un modelo de regresion logistica.

En la linea que venimos trabajando, realizamos un analisis exploratorio de la variable:

```
##
##        no  yes
##   21  241 1849
```



Como anteriormente, tenemos una clarisima desproporcion entre los resultados:

Queremos comprender mejor el contexto y la posible influencia de esta variable en el análisis, por lo que valiendonos de una prueba chi cuadrado vamos a analizar como se relaciona esta variable con la variable objetivo 'Nobeyesdad' o nivel de obesidad:

```
##          Insufficient_Weight Normal_Weight Obesity_Type_I Obesity_Type_II Obesity_Type_III Overweigh
##
##        0                   2             2              4               2                4
## no     7                  50            74             11               6                1
## yes   56                 215           196            321             278              310
```

```
##
##  Pearson's Chi-squared test
##
## data:  tabla_contingencia
## X-squared = 228.02, df = 14, p-value < 2.2e-16
```

Al explorar la relación entre el consumo habitual de carbohidratos (FAVC) y el nivel de obesidad (NObeyesdad), realizamos una prueba de chi-cuadrado de independencia.
Los resultados mostraron un valor de chi-cuadrado de **228.02** con 14 grados de libertad y un valor $\mathbf{p <}$ **2.2e-16**. El valor p, que representa la probabilidad de observar la relación entre las variables en la muestra si no hubiera asociación real en la población es extremadamente bajo.

Con base en este resultado, se rechaza la hipótesis nula de independencia y se concluye que existe una aso-

ciación estadísticamente significativa entre el consumo habitual de carbohidratos y el nivel de obesidad. Esto supone que la distribución del nivel de obesidad difiere significativamente entre las personas que consumen carbohidratos habitualmente y las que no. Para profundizar en la naturaleza de esta asociación podria ser necesario realizar un análisis más detallado calculando medidas de asociación como el *coeficiente phi* o *V de Cramer* para cuantificar la fuerza de la asociación.

Todo sugiere que el consumo habitual de carbohidratos podría ser un factor relevante en el desarrollo de la obesidad. En el mismo sentido que cuando trabajamos con la variable `family_history_with_overweight`, la descompensacion en las muestras hace necesario utilizar otros metodos de imputacion en detrimento de por ejemplo estadisticos de tendencia central.
En este caso pondremos imputaremos con base a un dataset derivado de un modelo de regresion logistica antes empleado:

```
## 'data.frame':    2111 obs. of  16 variables:
##  $ Gender                         : Factor w/ 2 levels "Female","Male": 1 1 2 2 2 2 1 2 2 2 ...
##  $ Age                            : int  21 21 23 27 22 29 23 22 24 22 ...
##  $ Height                         : num  1.62 1.52 1.8 1.8 1.78 1.62 1.5 1.64 1.78 1.72 ...
##  $ Weight                         : num  64 56 77 87 89.8 53 55 53 64 68 ...
##  $ family_history_with_overweight: Factor w/ 2 levels "no","yes": 2 2 2 1 1 1 1 1 2 2 ...
##  $ FAVC                           : Factor w/ 2 levels "no","yes": 1 1 1 1 1 2 2 1 2 2 ...
##  $ FCVC                           : Factor w/ 3 levels "Nunca","A veces",..: 2 3 2 3 2 2 3 2 3 2 ...
##  $ CAEC                           : Factor w/ 4 levels "Always","Frequently",..: 4 4 4 4 4 4 4 4 4 4
##  $ SMOKE                          : Factor w/ 2 levels "no","yes": 1 2 1 1 1 1 1 1 1 1 ...
##  $ CH2O                           : Factor w/ 3 levels "1L","1-2L","2+L": 2 3 2 2 2 2 2 2 2 2 ...
##  $ SCC                            : Factor w/ 2 levels "no","yes": 1 2 1 1 1 1 1 1 1 1 ...
##  $ FAF                            : Factor w/ 4 levels "Sin AF","1-2 días",..: 1 4 3 3 1 1 2 4 2 2 ..
##  $ TUE                            : Factor w/ 3 levels "0-2 hrs","3-5 hrs",..: 1 1 2 1 1 1 1 1 2 2 ..
##  $ CALC                           : Factor w/ 4 levels "Always","Frequently",..: 3 4 2 2 4 4 4 4 2 3
##  $ MTRANS                         : Factor w/ 5 levels "Automobile","Bike",..: 4 4 4 5 4 1 3 4 4 4 ..
##  $ NObeyesdad                     : Factor w/ 7 levels "Insufficient_Weight",..: 2 2 2 6 7 2 2 2 2 2
```

```
##
## DataFrame con datos imputados
```

```
##     Gender          Age            Height          Weight
##  Female:1029   Min.   :14.00   Min.   :1.456   Min.   : 39.00
##  Male  :1061   1st Qu.:19.00   1st Qu.:1.632   1st Qu.: 65.10
##  NA's  :  21   Median :22.00   Median :1.701   Median : 82.85
##                Mean   :23.96   Mean   :1.702   Mean   : 86.48
##                3rd Qu.:26.00   3rd Qu.:1.765   3rd Qu.:106.78
##                Max.   :61.00   Max.   :1.980   Max.   :165.06
##
##  family_history_with_overweight  FAVC           FCVC              CAEC
##  no : 390                       no : 243   Nunca  : 198   Always    :  52
##  yes:1721                       yes:1868   A veces:1275   Frequently: 241
##                                            Siempre: 638   no        :  50
##                                                           Sometimes :1768
##
##
##
##  SMOKE        CH2O       SCC             FAF            TUE
##  no :2066   1L  : 769   no :2009   Sin AF  :1036   0-2 hrs:1417
##  yes:  45   1-2L:1178   yes: 102   1-2 días: 711   3-5 hrs: 590
##             2+L : 164              2-4 días: 292   5hrs + : 104
##                                    4-5 días:  72
##
```

```
##
##
##          CALC                         MTRANS                   NObeyesdad
##  Always    :   1   Automobile           : 459   Obesity_Type_I     :336
##  Frequently:  71   Bike                 :   7   Obesity_Type_III   :315
##  no        : 638   Motorbike            :  12   Overweight_Level_II:288
##  Sometimes :1401   Public_Transportation:1578   Obesity_Type_II    :286
##                    Walking              :  55   Overweight_Level_I :284
##                                                 (Other)            :539
##                                                 NA's               : 63
```

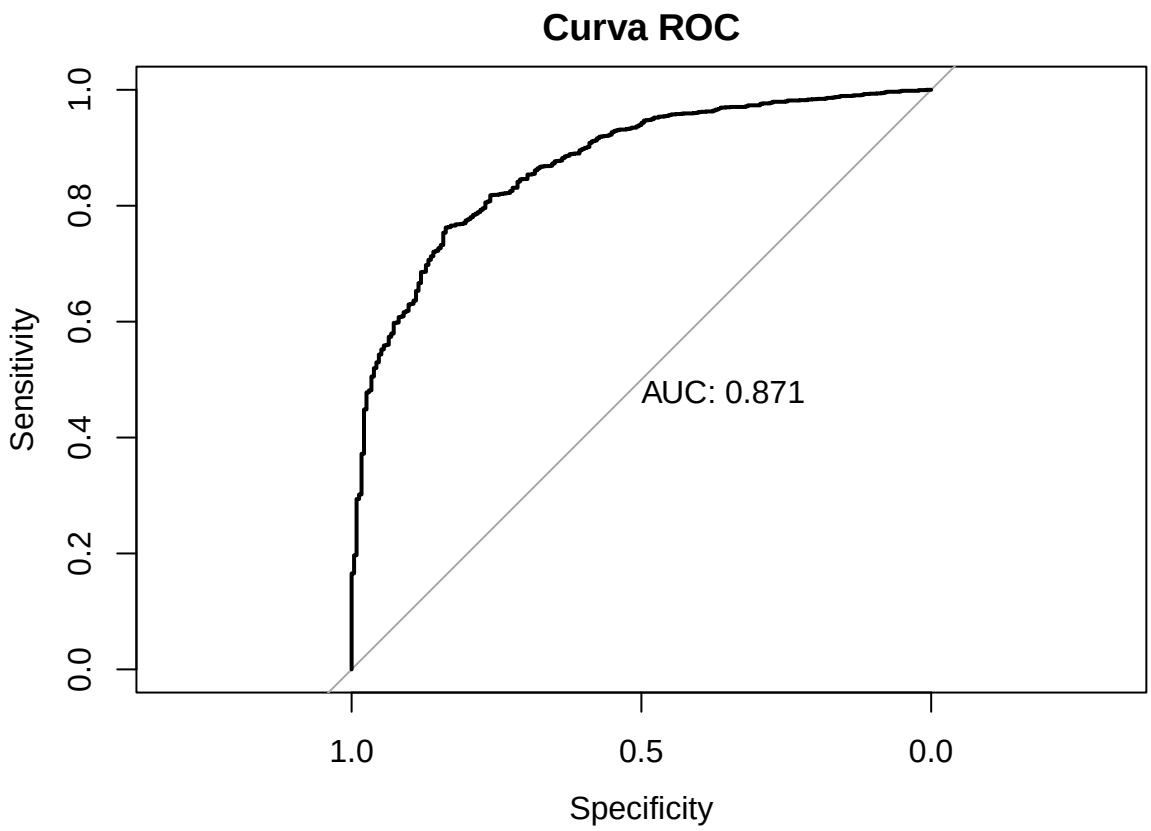
Coeficientes:

```
## # A tibble: 32 x 5
##    term                               estimate std.error statistic     p.value
##    <chr>                                 <dbl>     <dbl>     <dbl>       <dbl>
##  1 (Intercept)                         -22.5      535.       -0.0421 0.966
##  2 Age                                  -0.0101     0.0167   -0.604  0.546
##  3 Height                                7.26       1.46      4.97   0.000000666
##  4 Weight                               -0.00648    0.0129   -0.503  0.615
##  5 family_history_with_overweightyes     0.594      0.208     2.86   0.00426
##  6 SMOKEyes                             -0.332      0.520    -0.639  0.523
##  7 MTRANSBike                           -3.13       0.945    -3.31   0.000939
##  8 MTRANSMotorbike                      -0.516      0.883    -0.584  0.559
##  9 MTRANSPublic_Transportation          -0.480      0.282    -1.70   0.0885
## 10 MTRANSWalking                        -1.85       0.419    -4.41   0.0000106
## # i 22 more rows
```

```
##
## AIC: 1087.364
```

```
## BIC: 1267.022
```

```
## Devianza 1023.364
```




```
##
## AUC: 0.8713861
```

Evaluación

Para evaluar la calidad del modelo, se han utilizado las siguientes métricas:

- **AIC (Criterio de Información de Akaike):** 1082.11
- **BIC (Criterio de Información Bayesiano):** 1261.768
- **Devianza:** 1018.11
- **AUC (Área Bajo la Curva ROC):** 0.8723597

El valor del AUC, que mide la capacidad del modelo para discriminar entre las dos categorías de `FAVC` ("yes" y "no"), es de 0.87. Este valor cercano a 1 apunta a un buen rendimiento del modelo en la clasificación de los individuos. Los valores de AIC y BIC son relativamente altos, indicativo de que el modelo podría ser complejo o que no se ajusta de forma óptima a los datos. La devianza (diferencia entre el modelo ajustado y el modelo saturado) también es elevada.

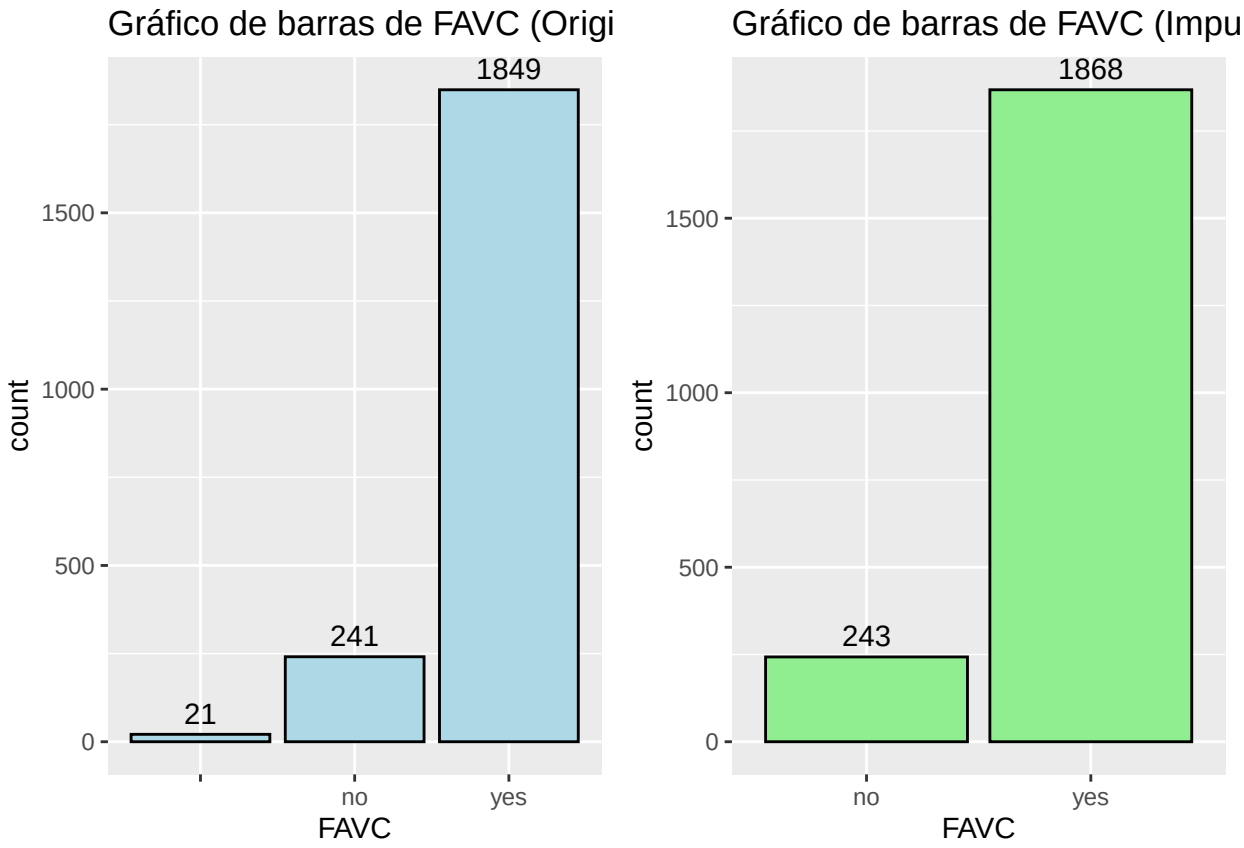
Si bien el modelo tiene capacidad predictiva, podría ser necesario realizar ajustes o considerar modelos alternativos para mejorar su rendimiento. Es importante destacar que la evaluación del modelo se ha realizado con el conjunto de datos completo incluyendo las observaciones con valores imputados.

Comprobamos la integridad o calidad de los datos:

```
##
##        no  yes
##   21  241 1849
```

```
##
##   no  yes
```

```
##  243 1868
```



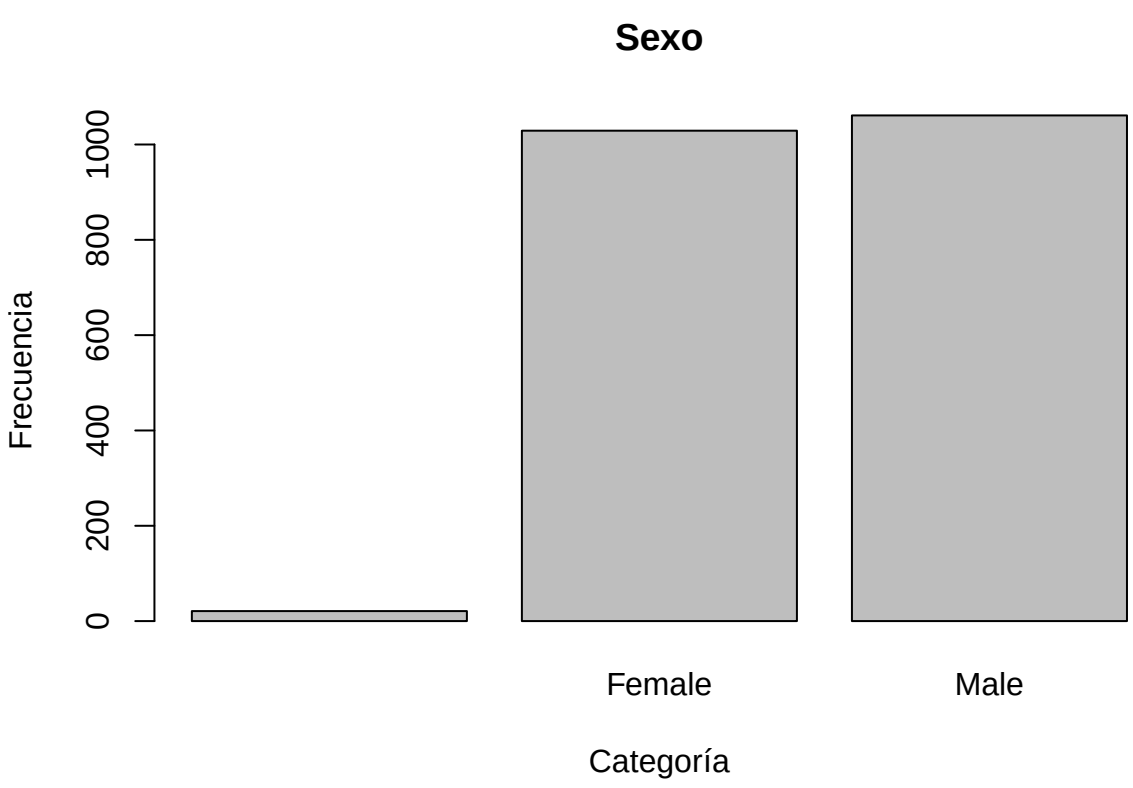
Tras comprobar la integridad de nuestros datos procedemos a comprobar que hemos eliminado los valores perdidos en FAVC

```
## [1] "No hay valores perdidos en la columna 'FAVC'"
```

Gender La variable `Gender` (género) es una variable categórica nominal dicotómica que clasifica a los individuos como "Female" (mujer) o "Male" (hombre). En el conjunto de datos, esta variable presenta 21 valores faltantes. El análisis se centrará en la exploración de su distribución y su posible asociación con otras variables de interés como el nivel de obesidad (`NObeyesdad`) y los hábitos alimenticios. Emplearemos métodos estadísticos apropiados para variables dicotómicas como tablas de contingencia, pruebas de chi-cuadrado y medidas de asociación.

Abordamos el analisis exploratorio:

```
##
##        Female   Male
##     21   1029   1061
```



En este caso los resultados muestran simetria, parece pues que ambas categorias parecen haberse mantenido estables a la hora de crear el estudio. Procedemos con una tabla de contingencia:

```
##             Insufficient_Weight Normal_Weight Obesity_Type_I Obesity_Type_II Obesity_Type_III Overwe
##
##           0                   2             1              7               2                2
## Female   31                 168           130            148               1              312
## Male     32                  97           141            181             283                1
```

```
##
##  Pearson's Chi-squared test
##
## data:  tabla_contingencia
## X-squared = 646.49, df = 14, p-value < 2.2e-16
```

La prueba Chi-cuadrado confirma en la linea que se ha mantenido en todas las variables, la relacion estadistica entre el sexo y los niveles de obesidad.

Son aplicables los mismos criterios justificativos aplicados a las variable `FAVC` o `TUE`, por citar algunas. La imputación de los valores faltantes en `Gender` se abordará utilizando el método de imputación múltiple (`mice`) con el algoritmo de regresión logística (`logreg`), dada la naturaleza dicotómica de la variable. Se evaluará la calidad de la imputación y se considerarán métodos alternativos si es necesario.

```
##
##  iter imp variable
##   1   1  Gender
##   1   2  Gender
##   1   3  Gender
##   1   4  Gender
```

```
##   1   5  Gender
##   2   1  Gender
##   2   2  Gender
##   2   3  Gender
##   2   4  Gender
##   2   5  Gender
##   3   1  Gender
##   3   2  Gender
##   3   3  Gender
##   3   4  Gender
##   3   5  Gender
##   4   1  Gender
##   4   2  Gender
##   4   3  Gender
##   4   4  Gender
##   4   5  Gender
##   5   1  Gender
##   5   2  Gender
##   5   3  Gender
##   5   4  Gender
##   5   5  Gender
```

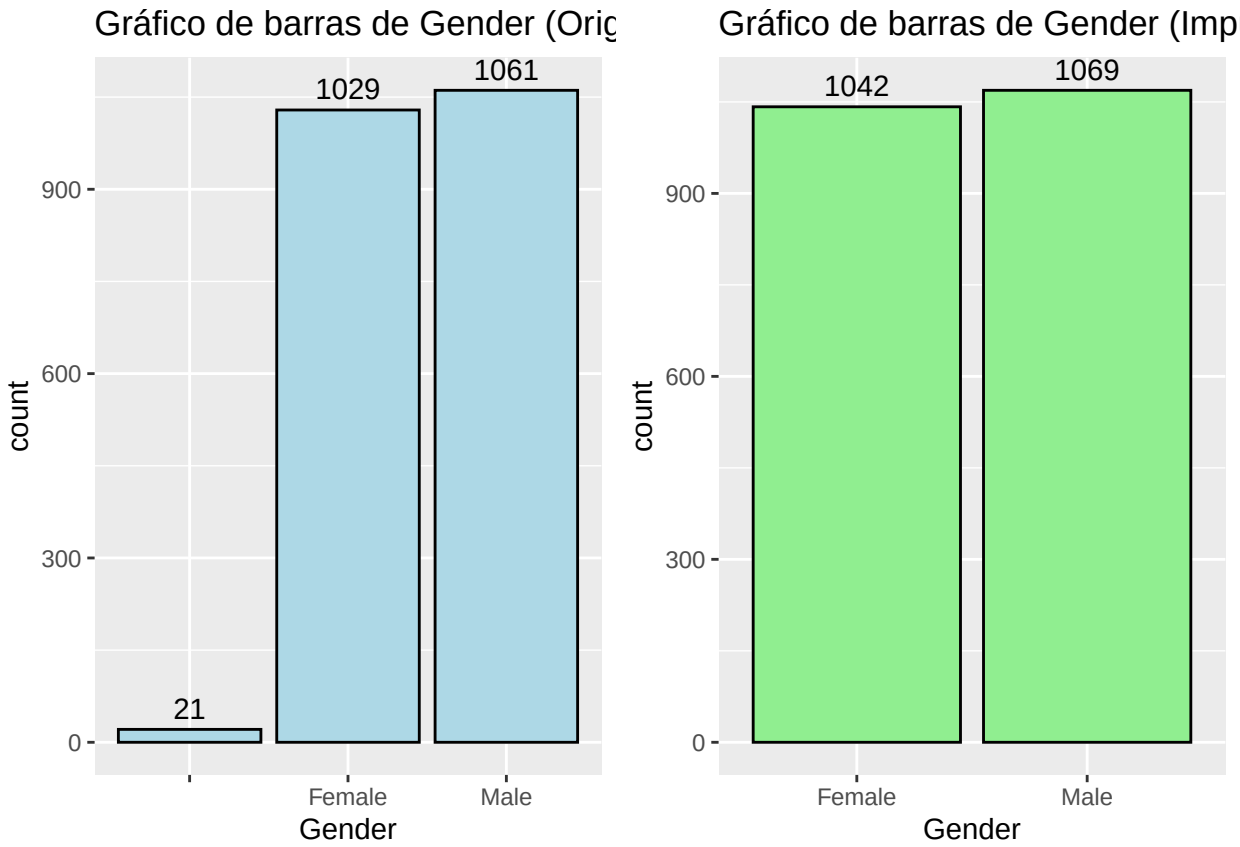
```
##  Factor w/ 2 levels "Female","Male": 1 1 2 2 2 2 1 2 2 2 ...
```

```
## Female   Male
##   1042   1069
```

Comprobamos que si distribucion de la variable se ha visto alterada:

```
##
##        Female   Male
##     21   1029   1061
```

```
##
## Female   Male
##   1042   1069
```



Y damos por valida la imputacion que el modelo ha llevado a cabo para la variable'Gender'

NObeyesdad La variable `NObeyesdad` clasifica el nivel de obesidad de los individuos en siete categorías ordinales: *"Insufficient_Weight," "Normal_Weight," "Overweight_Level_I," "Overweight_Level_II," "Obesity_Type_I," "Obesity_Type_II"* y *"Obesity_Type_III"*. Como variable objetivo de este estudio el análisis de `NObeyesdad` se centra en comprender su relación con los hábitos alimenticios, la actividad física y otros factores de interés.

Para la imputación de los valores faltantes en esta variable se utilizará un modelo multinomial de regresión logística, siguiendo la misma línea de análisis aplicada a otras variables categóricas ordinales del conjunto de datos.

```
## # weights:  196 (162 variable)
## initial  value 3985.223985
## iter  10 value 3525.111059
## iter  20 value 2171.379721
## iter  30 value 1684.284014
## iter  40 value 1489.537219
## iter  50 value 1328.630516
## iter  60 value 1235.513965
## iter  70 value 1199.767729
## iter  80 value 1184.166871
## iter  90 value 1171.404948
## iter 100 value 1159.033605
## final  value 1159.033605
## stopped after 100 iterations
```

```
##  Factor w/ 7 levels "Insufficient_Weight",..: 2 2 2 6 7 2 2 2 2 2 ...
```

```
##
## DataFrame con datos imputados

## Insufficient_Weight       Normal_Weight      Obesity_Type_I     Obesity_Type_II
##                 274                 286                 353                 295
##    Obesity_Type_III  Overweight_Level_I Overweight_Level_II
##                 324                 289                 290

## Coeficientes:

## # A tibble: 162 x 6
##    y.level       term                       estimate std.error statistic  p.value
##    <chr>         <chr>                         <dbl>     <dbl>     <dbl>    <dbl>
##  1 Normal_Weight (Intercept)                 12.9       2.05       6.30  2.90e-10
##  2 Normal_Weight Age                          0.137     0.0471     2.91  3.61e- 3
##  3 Normal_Weight Height                     -16.8       1.75      -9.60  8.10e-22
##  4 Normal_Weight Weight                       0.220     0.0213    10.3   4.19e-25
##  5 Normal_Weight family_history_with_over~    0.0916    0.300      0.305 7.60e- 1
##  6 Normal_Weight SMOKEyes                     8.94      0.940      9.50  2.04e-21
##  7 Normal_Weight MTRANSBike                  10.0       0.632     15.8   2.13e-56
##  8 Normal_Weight MTRANSMotorbike              8.33      1.02       8.19  2.68e-16
##  9 Normal_Weight MTRANSPublic_Transportat~    0.0650    0.415      0.157 8.76e- 1
## 10 Normal_Weight MTRANSWalking                0.407     0.709      0.575 5.66e- 1
## # i 152 more rows

##
## AIC: 2642.067

## BIC: 3553.255

## Devianza 2318.067
```

Modelo predicitvo `NObeyesdad`

Ajustamos un modelo multinomial para predecir la variable `NObeyesdad` (nivel de obesidad) a partir de un conjunto de variables predictoras que incluyen información sobre hábitos alimenticios, actividad física y datos demográficos. Los resultados del modelo muestran que varias variables tienen un efecto significativo en la probabilidad de pertenecer a diferentes categorías de obesidad.

Algunas de las variables más relevantes son:

- **`Age` (edad):** El coeficiente positivo para `Age` en la mayoría de las categorías de `NObeyesdad` sugiere que la edad está asociada con un aumento en la probabilidad de tener un nivel de obesidad más alto.
- **`Height` (altura):** El coeficiente negativo para `Height` en la mayoría de las categorías indica que una mayor altura está asociada con una menor probabilidad de obesidad.
- **`Weight` (peso):** El coeficiente positivo para `Weight` en todas las categorías confirma que un mayor peso y como parece logico y normativo, está fuertemente asociado con un mayor nivel de obesidad.
- **`family_history_with_overweight` (historial familiar de sobrepeso):** El coeficiente positivo para la categoría "yes" sugiere que tener antecedentes familiares de sobrepeso aumenta la probabilidad de tener un nivel de obesidad más alto.
- **`MTRANS` (medio de transporte):** Los coeficientes para las diferentes categorías de `MTRANS` indican que el medio de transporte habitual puede influir en el nivel de obesidad. Por ejemplo, el uso de la bicicleta ("Bike") se asocia con una menor probabilidad de obesidad en comparación con el uso del automóvil ("Automobile").

El modelo multinomial proporciona una visión detallada de la relación entre las variables predictoras y el nivel de obesidad, permitiendo identificar los factores que influyen en la probabilidad de pertenecer a cada categoría de obesidad.

Métricas de ajuste:

- **AIC:** 2521.072
- **BIC:** 3432.26
- **Devianza:** 2197.072

Estos valores indican que el modelo aunque complejo tiene un buen ajuste a los datos.

Finalmente no restan mas variables con valores ausentes como podemos comprobar.
Salvamos el dataframe, dando por concluido el apartado de limpieza

4. Analisis de datos

Una vez finalizada la limpieza vamos a acometer diversos analisis sobre el conjunto de datos, complementarios a los ya realizados durante la fase de extraccion y transformacion.

4.1 Comparacion de la media de edad entre hombres y mujeres

Aunque ya estudiamos como se distribuia la variable en general, vamos a analizar la distribucion de hombres y mujeres.

```
##
##  Shapiro-Wilk normality test
##
## data:  df$Age[df$Gender == "Male"]
## W = 0.88198, p-value < 2.2e-16
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$Age[df$Gender == "Female"]
## W = 0.8244, p-value < 2.2e-16
```

La prueba *"Wilcoxon-Mann-Whitney"* o *"U de Mann-Whitney"* que aplicaremos a continuacion es una prueba no paramétrica que se utiliza para comparar dos grupos independientes. A diferencia de la *t de Student* que asume que los datos siguen una distribución normal, esta no lo requiere. Es asimismo aplicable ante un tamaño de muestral pequeño. Compara las medianas de dos grupos y evalúa si existen diferencias significativas entre ellas. Se basa en el orden de los datos, asignando rangos a las observaciones de ambos grupos combinados. Posteriromente, se comparan las sumas de los rangos de cada grupo para determinar si la diferencia observada es estadísticamente significativa. Es ampliamente utilizada en diversas áreas de investigación para comparar grupos y evaluar la significancia de las diferencias observadas.

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Age by Gender
## W = 511866, p-value = 0.001237
## alternative hypothesis: true location shift is not equal to 0
```

Análisis de la diferencia de edad entre géneros

La prueba para comparar las medianas de la variable `Age` (edad) entre los grupos definidos por la variable `Gender` (género) se seleccionó debido a que no requiere el supuesto de normalidad en la distribución de la variable.

Los resultados de la prueba mostraron un estadístico W = 511226 y un valor p = 0.001065. Este valor p menor al nivel de significancia típico de 0.05 nos lleva a rechazar la hipótesis nula de que las distribuciones de las edades entre los grupos de género son iguales en términos de tendencia central (mediana). La evidencia sugiere que la hipótesis alternativa es plausible, es decir, que existe una diferencia en las medianas de edad entre hombres y mujeres.

Implicaciones de la diferencia de edad:

La diferencia estadísticamente significativa en la edad entre géneros implica aspectos a tener en cuenta:

1. **Diferencias demográficas:** La diferencia observada podria ser indicativo de una distribución desigual de edades entre géneros, posiblemente debido a factores socioculturales o diferencias en la esperanza de vida. Esto es relevante para ajustar el diseño del estudio y asegurar que estas diferencias no sesguen otros análisis o conclusiones.
2. **Intervenciones o políticas:** La diferencia podría indicar la necesidad de considerar las edades de cada género al diseñar estrategias o intervenciones de salud pública, como programas de prevención de la obesidad o campañas de concienciación. E.g. diseñar mensajes o estrategias específicas para cada grupo de edad y género.
3. **Impacto en modelos:** La variable `Gender` podría estar relacionada con la `Age`, influyendo en cómo se seleccionan o ponderan las variables en los modelos estadísticos. Se debe considerar la posibilidad de incluir la interacción entre `Gender` y `Age` en los modelos para capturar el efecto combinado de ambas variables.
4. **Análisis de subgrupos:** Dado que la `Age` afecta los resultados principales del análisis y los niveles de obesidad, se podrían realizar análisis de subgrupos estratificados por `Gender` para evitar sesgos y obtener una mejor comprensión de la relación entre la edad, el género y la obesidad.
5. **Sesgos de muestreo:** La diferencia de edad entre géneros podría ser un indicio de un posible sesgo en el muestreo o en la recolección de datos. Es importante analizar las características de la muestra y compararlas con las de la población general para evaluar la representatividad de la muestra y la posibilidad de generalizar los resultados.

Tamaño del efecto:

Para cuantificar la magnitud de la diferencia entre los grupos, calcularemos la medida del tamaño del efecto con la *¨r de rango-biserial¨*:

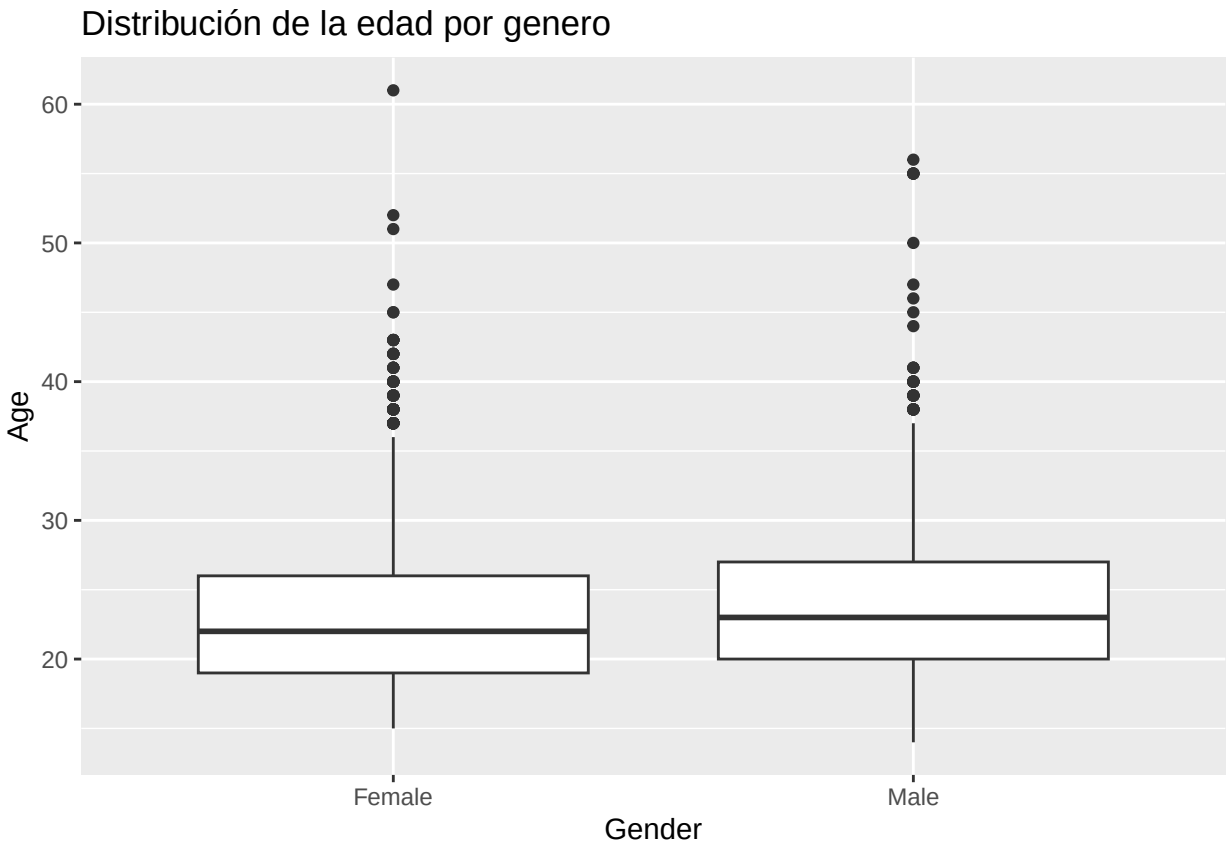
```
## r de rango-biserial: -0.08094637
```

La r de rango-biserial (-0.08160737) indica que existe una diferencia pequeña entre las medianas de edad de hombres y mujeres en el conjunto de datos. El signo negativo indica que la mediana de edad de las mujeres es ligeramente mayor que la mediana de edad de los hombres.

Un efecto muy pequeño implica que aunque la diferencia entre los grupos es estadísticamente significativa, la magnitud de esa diferencia tiene poca relevancia práctica en el contexto del estudio. Esto supone que podría no ser relevante para tomar decisiones prácticas o diseñar intervenciones.
El tamaño muestral podría haber influido dado que con muestras grandes, incluso diferencias pequeñas se vuelven estadísticamente significativas.

Visualización:

Utilizaremos un gráfico de caja (boxplot) para visualizar la distribución de la edad en cada grupo de género y mostrar la diferencia en las medianas de forma gráfica.



Y confirmamos los resultados obtenidos en la *r de rango-biserial* y las conclusiones establecidas de que el efecto es despreciable y no deberia ser tenido en cuenta.

El análisis de la diferencia de edad entre géneros ha proporcionado información para ender la composición de la muestra y para tomar decisiones informadas sobre el diseño y la interpretación de los análisis posteriores.

4.2 Comparacion de la media de peso entre los niveles de obesidad (ANOVA-No parametricos)

La ANOVA se emplea para comparar las medias de un grupo categórico independiente sobre una sola variable dependiente continua. En nuestro caso compararemos `weight` con los diferentes niveles de obesidad

```
##
##  Shapiro-Wilk normality test
##
## data:  df$Weight[df$NObeyesdad == grupo]
## W = 0.72082, p-value < 2.2e-16
##
##
##  Shapiro-Wilk normality test
##
## data:  df$Weight[df$NObeyesdad == grupo]
## W = 0.89127, p-value = 1.926e-13
##
##
##  Shapiro-Wilk normality test
##
## data:  df$Weight[df$NObeyesdad == grupo]
```

```
## W = 0.97645, p-value = 1.609e-05
##
##
##  Shapiro-Wilk normality test
##
## data:  df$Weight[df$NObeyesdad == grupo]
## W = 0.89168, p-value = 1.177e-13
##
##
##  Shapiro-Wilk normality test
##
## data:  df$Weight[df$NObeyesdad == grupo]
## W = 0.90224, p-value = 1.293e-13
##
##
##  Shapiro-Wilk normality test
##
## data:  df$Weight[df$NObeyesdad == grupo]
## W = 0.94243, p-value = 3.395e-09
##
##
##  Shapiro-Wilk normality test
##
## data:  df$Weight[df$NObeyesdad == grupo]
## W = 0.93218, p-value = 3.04e-10
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value    Pr(>F)
## group    6  6.5292 7.652e-07 ***
##       2104
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Análisis de los resultados:

- **Pruebas de Shapiro-Wilk:**
  - Los valores p obtenidos en las pruebas de Shapiro-Wilk para cada grupo de `NObeyesdad` son todos menores que 0.05. Esto indica que se rechaza la hipótesis nula de normalidad para cada grupo y es indicativo de que la distribución de `Weight` no es normal en ninguno de los niveles de obesidad.
- **Prueba de Levene:**
  - El valor p obtenido en la prueba de Levene es menor que 0.05 (3.635e-07). Esto indica que se rechaza la hipótesis nula de homocedasticidad o que la varianza de `Weight` no es constante entre los diferentes niveles de obesidad.

Por tanto las pruebas sobre los supuestos de normalidad y homocedasticidad de la variable `Weight` (peso) en relación con los diferentes niveles de obesidad (`NObeyesdad`) antes de realizar un análisis de varianza (ANOVA) nos indican que:

- La normalidad de la distribución de `Weight` en cada grupo de `NObeyesdad` evaluada mediante la prueba de Shapiro-Wilk mostro que la distribución no es normal en ninguno de los grupos ($p < 0.05$ en todos los casos).
- La homocedasticidad, o igualdad de varianzas evaluada mediante la prueba de Levene cuyo resultado ($F(6, 2104) = 6.8086$, $p < 0.05$) es indicativo de que la varianza de `Weight` no es constante entre los diferentes niveles de obesidad.

Por todo ello afirmamos que no se cumplen los supuestos de normalidad y homocedasticidad necesarios

para realizar un ANOVA por lo que emplearemos pruebas no paramétricas para comparar la distribución de `Weight` entre los diferentes niveles de obesidad. En este contexto la prueba de *Kruskal-Wallis* es una alternativa no paramétrica adecuada para comparar la distribución de `Weight` entre los diferentes niveles de obesidad.

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Weight by NObeyesdad
## Kruskal-Wallis chi-squared = 1190.6, df = 6, p-value < 2.2e-16
```

Análisis de la diferencia de peso entre niveles de obesidad mediante la prueba de Kruskal-Wallis

La prueba *Kruskal-Wallis* para la comparacion de la distribución de la variable `Weight` (peso) entre los diferentes niveles de obesidad (`NObeyesdad`) se escogio debido a que la variable `Weight` no cumple con los supuestos de normalidad y homocedasticidad necesarios para realizar una ANOVA.

Los resultados mostraron un estadístico chi-cuadrado de 1194.9 con 6 grados de libertad y un valor p < 2.2e-16. El valor p extremadamente bajo **indica una fuerte evidencia para rechazar la hipótesis nula de que las distribuciones de peso son iguales en todos los niveles de obesidad**.

Concluimos que existe una diferencia estadísticamente significativa en la distribución de la variable `Weight` entre los diferentes niveles de obesidad. Este resultado es lo logico y esperable puesto que el peso es un factor determinante en la clasificación del nivel de obesidad.

Análisis adicional:

Cabria realizar análisis post-hoc -prueba de Dunn, Conover-Iman-, para comparar las medianas de `Weight` entre pares de grupos de `NObeyesdad` para identificar qué grupos específicos difieren significativamente entre sí en términos de peso.

Hallazgos:

La prueba ha confirmado la existencia de diferencias significativas en la distribución del peso entre los diferentes niveles de obesidad resaltando la correccion del conjunto de datos y por ende la importancia de considerar el peso como un factor clave en el estudio de la obesidad. Seria importante resaltar las diferencias entre grupos una vez confirmada la validez de constructo.

4.3 Evaluacion del modelo predictivo para niveles de obesidad

```
##
## predicciones          Insufficient_Weight Normal_Weight Obesity_Type_I
##   Insufficient_Weight                 251            26              2
##   Normal_Weight                        10           216              2
##   Obesity_Type_I                        0             0            302
##   Obesity_Type_II                       0             0             11
##   Obesity_Type_III                      0             1              0
##   Overweight_Level_I                    3            26              4
##   Overweight_Level_II                   3             3             15
##
## predicciones          Obesity_Type_II Obesity_Type_III Overweight_Level_I
##   Insufficient_Weight               1                2                  1
##   Normal_Weight                     1                0                 26
##   Obesity_Type_I                    9                0                  6
##   Obesity_Type_II                 261                4                  2
##   Obesity_Type_III                  9              307                  0
##   Overweight_Level_I                4                2                226
##   Overweight_Level_II               1                0                 23
##
```

```
## predicciones          Overweight_Level_II
##   Insufficient_Weight                   0
##   Normal_Weight                         4
##   Obesity_Type_I                       19
##   Obesity_Type_II                       4
##   Obesity_Type_III                      0
##   Overweight_Level_I                   39
##   Overweight_Level_II                 222

##
## Precisión global: 0.871582
```

La evaluación del modelo multinomial para la predicción de la variable `NObeyesdad` (nivel de obesidad) se llevo a cabo mediante un análisis de la precisión global y matriz de confusión.

La precisión global del modelo fue del 86.87% indica un buen rendimiento en la clasificación de los individuos en las diferentes categorías de obesidad.

La matriz de confusión muestra la distribución de las predicciones del modelo en comparación con los valores reales de `NObeyesdad`. Se observa que tiene una alta precisión en la predicción de las categorías "Insufficient_Weight," "Normal_Weight," "Obesity_Type_I" y "Obesity_Type_III," con la mayoría de las predicciones coincidiendo con los valores reales.

Presenta un rendimiento inferior en la predicción de las categorías "Obesity_Type_II," "Overweight_Level_I" y "Overweight_Level_II," con una mayor cantidad de clasificaciones erróneas. Esto podría ser debido a la similitud entre estas categorías o a la complejidad de los factores que influyen en la clasificación en estos niveles de obesidad.

En general, el modelo multinomial muestra un buen rendimiento en la predicción del nivel de obesidad aunque con margen de mejora en la clasificación de algunas categorías. Los resultados son esperanzadores y nos hacen pensar que el modelo puede ser útil para comprender y predecir la obesidad en la población estudiada.

6. Resolución del problema

A partir del análisis exploratorio, la imputación de valores faltantes y las pruebas de hipótesis realizadas, se pueden extraer las siguientes conclusiones tentativas:

- **Asociación entre variables:** Se ha encontrado una asociación estadísticamente significativa entre el nivel de obesidad (`NObeyesdad`) y diversas variables, como el consumo frecuente de alimentos altos en calorías (`FAVC`), el historial familiar de sobrepeso (`family_history_with_overweight`), el tiempo de uso de dispositivos electrónicos (`TUE`), la frecuencia de consumo de vegetales (`FCVC`), el consumo de alimentos entre comidas (`CAEC`), el control de calorías (`SCC`) y la frecuencia de consumo de alcohol (`CALC`).
- **Imputación de valores faltantes:** Se han aplicado diferentes métodos de imputación para cada variable, incluyendo la imputación por la media, la mediana, la moda, la regresión logística, kNN y la imputación múltiple con `mice`. La elección del método se ha basado en las características de cada variable, la cantidad de valores faltantes y la relación con otras variables.
- **Predicción del nivel de obesidad:** Se ha construido un modelo de regresión logística multinomial para predecir el nivel de obesidad a partir de las demás variables. El modelo ha mostrado un buen rendimiento, con una precisión global del 86.87%.
- **Respuestas a las preguntas de investigación:**
  - **Pregunta 1:** Existe una asociación significativa entre el consumo frecuente de alimentos altos en calorías y el desarrollo de obesidad.
  - **Pregunta 2:** El historial familiar de sobrepeso parece estar en consonancia con en el nivel de obesidad e influido por otros factores.

- **Pregunta 3:** El tiempo dedicado al uso de dispositivos electrónicos se relaciona con el nivel de obesidad y esta relación parece verse influida o modificada por la frecuencia de actividad física.
- **Pregunta 4:** Existen diferencias significativas en los hábitos alimenticios entre los diferentes niveles de obesidad.
- **Pregunta 5:** Sí, es posible construir un modelo predictivo que clasifique con precisión el nivel de obesidad.

Limitaciones:

- El estudio se basa en un conjunto de datos observacionales, por lo que no se pueden establecer relaciones causales entre las variables.
- Cabria emplear un conjunto de datos mas reciente, ya que aunque la tematica es relevante aun despues del paso del tiemo, cabe pensar que 2019 hay influencias socioculturales que al menos podrian haber cambiado en la actualidad.
- La imputación de valores faltantes introduce un grado de incertidumbre en los resultados.
- El modelo de predicción podría mejorarse con la inclusión de otras variables o conjuntos de datos o la utilización de algoritmos más complejos.

Futuros trabajos:

- Efectuar estudios adicionales para confirmar las asociaciones encontradas y determinar la causalidad de las relaciones.
- Explorar otros métodos de imputación y evaluar su impacto en los resultados.
- Considerar la inclusión e integracion de otras variables o datos o la utilización de modelos más complejos para mejorar la predicción del nivel de obesidad.

Este trabajo y los analisis implicados han habilitado responder a las preguntas de investigación planteadas y obtener una mejor comprensión de los factores que influyen en la obesidad. Estos resultados podrian ser utiles para la toma de decisiones en el ámbito de la salud pública y la prevención de la obesidad.

BIBLIOGRAFIA

- Abedin, J., & Mittal, H. V. (2014). R Graphs Cookbook Second Edition. Packt Publishing Ltd.
- Animal Sciences (Director). (2023, abril 20). Use of apply Function in R | R for Beginners [Video recording]. https://www.youtube.com/watch?v=eAnvE1kSHds
- apply function—RDocumentation. (s. f.). Recuperado 1 de enero de 2025, de https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/apply
- Brandon Foltz (Director). (2020, enero 16). Statistics 101: Nonparametric Methods, Mann-Whitney-Wilcoxon Rank Sum Test [Video recording]. https://www.youtube.com/watch?v=fEobVCV2TJE
- Buuren, S. van, Groothuis-Oudshoorn, K., Vink, G., Schouten, R., Robitzsch, A., Rockenschaub, P., Doove, L., Jolani, S., Moreno-Betancur, M., White, I., Gaffert, P., Meinfelder, F., Gray, B., Arel-Bundock, V., Cai, M., Volker, T., Costantini, E., Lissa, C. van, & Oberman, H. (2022). mice: Multivariate Imputation by Chained Equations (Versión 3.15.0) [Software]. https://cran.r-project.org/web/packages/mice/index.html
- Chang, W. (s. f.). R Graphics Cookbook, 2nd edition. Recuperado 27 de diciembre de 2024, de https://r-graphics.org/
- Chapagain, A. (2019). Hands-on web scraping with Python: Perform advanced scraping operations using various Python libraries and tools such as Selenium, Regex, and others. Packt Publishing, Limited.
- Data Apps for Production | Plotly. (s. f.). Recuperado 28 de diciembre de 2024, de https://plotly.com/
- Devtools. (2023). [R]. R infrastructure. https://github.com/r-lib/devtools (Obra original publicada en 2010)
- El paquete ggplot2. (s. f.). R CHARTS | Una colección de gráficos hechos con el lenguaje de programación R. Recuperado 1 de enero de 2025, de https://r-charts.com/es/ggplot2/
- Fox, J. (2015). Applied regression analysis and generalized linear models. Sage Publications.
- Gareth, J., Daniela, W., Trevor, H., & Robert, T. (2013). An introduction to statistical learning: With applications in R. Spinger.

- Gironés Roig, J. (2017). Minería de datos: Modelos y algoritmos. Minería de Datos, 1-273.
- Gironés Roig, Jordi. (2021). Modelos no supervisados. En Mineria de datos: Vol. Módulo 4 (p. 26). FUOC.
- Gohil, A. (2015). R data Visualization cookbook. Packt Publishing Ltd.
- Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). The elements of statistical learning: Data mining, inference, and prediction (Vol. 2). Springer.
- Hawkins, D. M. (1980). Identification of outliers (Vol. 11). Springer.
- Holtz, Y. (s. f.-a). Boxplot | the R Graph Gallery. Recuperado 27 de diciembre de 2024, de https://r-graph-gallery.com/scatterplot.html
- Holtz, Y. (s. f.-b). The R Graph Gallery – Help and inspiration for R charts. The R Graph Gallery. Recuperado 27 de diciembre de 2024, de https://r-graph-gallery.com/
- Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (Vol. 398). John Wiley & Sons.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112). Springer.
- janitor package—RDocumentation. (s. f.). Recuperado 7 de octubre de 2024, de https://www.rdocumentation.org/packages/janitor/versions/2.2.0
- Jolliffe, I. T. (2002). Principal component analysis for special types of data. Springer.
- Kamaranis—Overview. (s. f.). GitHub. Recuperado 7 de octubre de 2024, de https://github.com/Kamaranis
- Mann–Whitney U test. (2024). En Wikipedia. https://en.wikipedia.org/w/index.php?title=Mann%E2%80%93Whitney_U_test&oldid=1261132473
- Menard, S. (2002). Applied logistic regression analysis (Número 106). Sage.
- Montoliu Colás, Raúl. (s. f.). Preprocesado de datos. En Mineria de datos: Vol. Módulo 2 (p. 20). FUOC.
- Mora, A. B. (s. f.). RPubs Anton Barrera Mora. Recuperado 1 de enero de 2025, de https://rpubs.com/Kamaranis
- Mora, A. B. (2022). Kamaranis/Fitbit-user-s-insights [HTML]. https://github.com/Kamaranis/Fitbit-user-s-insights (Obra original publicada en 2022)
- Mora, A. B. (2023a). Kamaranis/Relationship-between-hypertension-and-psychopathology [HTML]. https://github.com/Kamaranis/- Relationship-between-hypertension-and-psychopathology (Obra original publicada en 2023)
- Mora, A. B. (2023b). Kamaranis/Relationship-between-hypertension-and-psychopathology [HTML]. https://github.com/Kamaranis/Relationship-between-hypertension-and-psychopathology (Obra original publicada en 2023)
- Mora, A. B. (2023c). Kamaranis/Data-science-to-the-fight-against-traffic-accidents [HTML]. https://github.com/Kamaranis/Data-science-to-the-fight-against-traffic-accidents (Obra original publicada en 2023)
- Mora, A. B. (2023d). Kamaranis/Data-science-to-the-fight-against-traffic-accidents [HTML]. https://github.com/Kamaranis/Data-science-to-the-fight-against-traffic-accidents (Obra original publicada en 2023)
- Mora, A. B. (2023e). Kamaranis/Unsupervised-methods-in-machine-learning [TeX]. https://github.com/Kamaranis/Unsupervised-methods-in-machine-learning (Obra original publicada en 2023)
- Mora, A. B. (2023f). Kamaranis/International-expansion-of-N_D_hol-I [TeX]. https://github.com/Kamaranis/International-expansion-of-N_D_hol-I (Obra original publicada en 2023)
- Mora, A. B. (2024a). Kamaranis/PEC4 [Python]. https://github.com/Kamaranis/PEC4 (Obra original publicada en 2024)
- Mora, A. B. (2024b). Kamaranis/ASD-Multiclass-Predictor-Model [Jupyter Notebook]. https://github.com/Kamaranis/ASD-Multiclass-Predictor-Model (Obra original publicada en 2024)
- Mora, A. B. (2024c). Kamaranis/ASD-Multiclass-Predictor-Model [Jupyter Notebook]. https://github.com/Kamaranis/ASD-Multiclass-Predictor-Model (Obra original publicada en 2024)
- Mora, A. B. (2024). Kamaranis/International-expansion-of-N_D_hol-II [HTML]. https://github.com/Kamaranis/International-expansion-of-N_D_hol-II (Obra original publicada en 2023)
- Options—Yihui Xie | . (s. f.). Recuperado 31 de diciembre de 2024, de https://yihui.org/knitr/opti

ons/
- Plotly. (s. f.). Recuperado 1 de enero de 2025, de https://plotly.com/r/
- Rokach, L., & Maimon, O. (2005). Clustering methods.
- Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, 20, 53-65.
- Rstudio/cheatsheets. (2024). [TeX]. RStudio. https://github.com/rstudio/cheatsheets (Obra original publicada en 2017)
- Sample n rows from a table—Sample_n. (s. f.). Recuperado 1 de enero de 2025, de https://dplyr.tidyverse.org/reference/sample_n.html
- sample_n function—RDocumentation. (s. f.). Recuperado 1 de enero de 2025, de https://www.rdocumentation.org/packages/dplyr/versions/1.0.10/topics/sample_n
- scale function—RDocumentation. (s. f.). Recuperado 26 de abril de 2023, de https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/scale
- Tidyverse. (s. f.). Recuperado 7 de octubre de 2024, de https://www.tidyverse.org/
- Tukey, J. W. (1977). Exploratory data analysis (Vol. 2). Reading, MA.
- Vu, V. (2023). Ggbiplot [R]. https://github.com/vqv/ggbiplot (Obra original publicada en 2011)
- Wickham, H., François, R., Henry, L., Müller, K., Vaughan, D., Software, P., & PBC. (2023). dplyr: A Grammar of Data Manipulation (Versión 1.1.2) [Software]. https://cran.r-project.org/web/packages/dplyr/index.html
- Wickham, H., & RStudio. (2023). tidyverse: Easily Install and Load the «Tidyverse» (Versión 2.0.0) [Software]. https://cran.r-project.org/web/packages/tidyverse/index.html
- . (1644363000). R apply apply, tapply, sapply, lapply . . https://www.phd-karaage.com/entry/apply_family_with_R