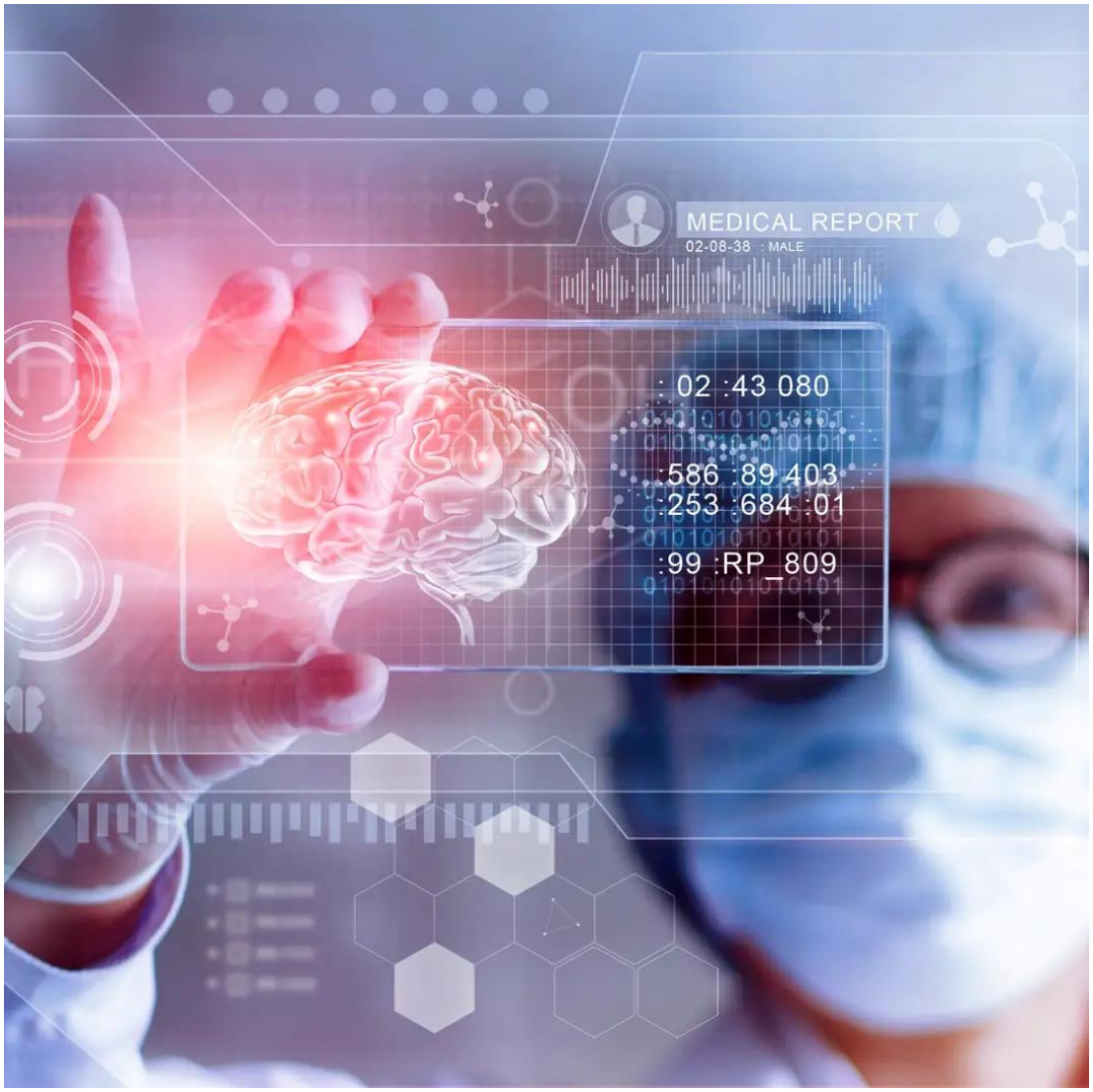
Berta Pelegrina Martínez y Laura Sánchez Garzón

GRUPO 6362

INFORME PRÁCTICA 1

Práctica de análisis de datos con R



Escuela Politécnica Superior - Universidad Autónoma de Madrid

GRADO EN INGENIERÍA BIOMÉDICA

CIENCIA DE DATOS BIOMÉDICOS

Versión del documento número 1

Práctica realizada por: ***Berta Pelegrina Martínez y Laura Sánchez Garzón***

Nº grupo: ***6362***

Fecha: ***06/03/2024***

Profesorado de la práctica: ***Jordi Porta Zamorano***

ÍNDICE

ÍNDICE 1
INTRODUCCIÓN 2
DESARROLLO DEL TRABAJO 3
APARTADO 1 3
APARTADO 2 6
APARTADO 3 7
APARTADO 4 8
APARTADO 5 9
APARTADO 6 10
APARTADO 7 12
APARTADO 8 13
APARTADO 9 16
APARTADO 10 24
APARTADO 11 25
CONCLUSIONES 29

INTRODUCCIÓN

Esta primera práctica va enfocada a analizar una tabla de datos, y trabajarlos utilizando el programa RStudio (o R en la nube), con el fin de realizar consultas de manera programada y obtener gráficas analíticas de las mismas.

Más concretamente, en esta práctica se pretende analizar el patrón de consumo de determinados alimentos según sexo y clase social basada en la ocupación de la persona de referencia, para la población mayor de 14 años.

Los datos originales tienen como fuente el Ministerio de Sanidad, Consumo y Bienestar Social (MSCBS) y el Instituto Nacional de Estadística (INE); obtenidos de: https://drive.google.com/file/d/1SvYB-Vv5sehNgB0_gVOjrdNEVcrlrp3U/view?usp=sharing 1 .

Estos datos se dividen según el sexo, el nivel económico, alimento y cantidad de veces que se consume dicho alimento a la semana (A diario, 3 o más veces, 1 o 2 veces, menos de 1 vez o nunca). También existe la columna no consta, que se ha supuesto que pertenece a aquellos participantes de los que no se llegó a preguntar por cierto alimento, no quiso responder, o no se anotó la respuesta adecuadamente.

Se dan dos posibilidades de sexo (HOMBRES o MUJERES), de alimento: fruta fresca (excluye zumos), carne, huevos, pescado, pasta, arroz y patatas, pan y cereales, verduras, ensaladas y hortalizas, legumbres, embutidos y fiambres, productos lácteos, dulces, refrescos con azúcar, comida rápida, aperitivos o comidas saladas de picar y zumo natural de frutas o verduras. Los niveles economicos presentados son los siguientes:

- I. Directores/as y gerentes de establecimientos de 10 o más asalariados/as y profesionales tradicionalmente asociados/as a licenciaturas universitarias
- II. Directores/as y gerentes de establecimientos de menos de 10 asalariados/as, profesionales tradicionalmente asociados/ as a diplomaturas universitarias y otros/as profesionales de apoyo técnico. Deportistas y artistas
- III. Ocupaciones intermedias y trabajadores/as por cuenta propia
- IV. Supervisores/as y trabajadores/as en ocupaciones técnicas cualificadas
- V. Trabajadores/as cualificados/as del sector primario y otros/as trabajadores/as semicualificados/as
- VI. Trabajadores/as no cualificados/as

Los objetivos de esta prática son, además de aprender a usar R, poder extraer conclusiones a partir de los datos, sobre determinados diagnósticos, distribuciones demográficas y su relación con niveles socioeconómicos.

DESARROLLO DEL TRABAJO

APARTADO 1

Carga y describe los datos con los que vas a trabajar: su formato, filas, columnas, valores, su número, etc. ¿Hay valores desconocidos? Independientemente de que los haya o no, ¿qué representan esos valores? ¿por qué motivo pueden producirse?, ¿qué relación puede haber con la columna "No consta"?

Antes de cargar los datos, se cargan las librerías *tidyverse* y *dplyr*.

La librería *tidyverse* es un conjunto de paquetes de R diseñados para trabajar de manera integrada y coherente para el análisis de datos. Incluye una serie de paquetes populares como *ggplot2* (para visualización), *dplyr* (para manipulación de datos), *tidyr* (para limpieza de datos), *readr* (para importar datos), y muchos otros.

Una vez instalados dichos paquetes, se cargan los datos del archivo *csv* donde se encuentra la información de interés. Dado que se ha comprobado que los datos quedan delimitados por comas, se utiliza para leer dicho *csv*. Guardamos los datos en un dataframe al que se le va a denominar *data_fr*. (NOTA: en este caso basta con poner el nombre del archivo .csv porque está dentro del path o porque hemos subido un archivo a la nube, sino habría que comprobar el directorio de trabajo con `getwd()`y `setwd()`).

```
library(tidyverse) # Cargamos la librería tidyverse
library(dplyr) # Cargamos la librería dplyr

data_fr <- read_csv("06007.csv") # Cargamos los datos del
                                 # archivo csv, delimitado por comas
```

Figura 1: Se cargan las librerías y los datos del csv

Se comprueba que los datos se han cargado correctamente.
Tanto `summarise_all(~sum(is.na(data_fr)))` como `summarise(sum(is.na(data_fr)))` nos permite contar si tenemos *missing values*. El primero los cuenta para cada variable y el segundo en todo el dataframe.

```
# APARTADO 1
data_fr
colnames(data_fr)
ncol(data_fr)
nrow(data_fr)

data_fr %>%
  summarise_all(~sum(is.na(data_fr)))

data_fr %>%
  summarise(sum(is.na(data_fr)))
```

Figura 2: *Código para resolver el apartado 1*

```
> data_fr
# A tibble: 210 × 9
   Sexo    Nivel Alimento    `A diario` 3 o más veces a la s…¹ 1 o 2 veces a la sem…²
   <chr>   <chr> <chr>            <dbl>                  <dbl>                  <dbl>
 1 HOMBRES I     Fruta fres…     1763.                   535.                   190.
 2 HOMBRES I     Carne            229.                  1740.                   572.
 3 HOMBRES I     Huevos            44.6                  829.                  1534.
 4 HOMBRES I     Pescado           24.8                 1090.                  1298.
 5 HOMBRES I     Pasta, arr…      236                   1663.                   661.
 6 HOMBRES I     Pan, cerea…     2127.                   256.                   132.
 7 HOMBRES I     Verduras, …      989.                  1311.                   250.
 8 HOMBRES I     Legumbres         16.9                  620.                  1675.
 9 HOMBRES I     Embutidos …      396.                  1026.                   714.
10 HOMBRES I     Productos …     2197.                   239.                    65.4
# i 200 more rows
# i abbreviated names: ¹`3 o más veces a la semana, pero no a diario`,
#   ²`1 o 2 veces a la semana`
# i 3 more variables: `Menos de 1 vez a la semana` <dbl>, Nunca <dbl>,
#   `No consta` <dbl>
# i Use `print(n = ...)` to see more rows
```

Figura 3: Visualización de data_fr

Se ha creado un tibble de 9 columnas (variables) y 210 filas u observaciones (comprobado también utilizando `ncol(data_fr)` y `nrow(data_fr)`); y dichas columnas tienen por nombre, el nombre por defecto del dataset original (comprobado mediante `colnames(data_fr)`).

```
> colnames(data_fr)
[1] "Sexo"
[2] "Nivel"
[3] "Alimento"
[4] "A diario"
[5] "3 o más veces a la semana, pero no a diario"
[6] "1 o 2 veces a la semana"
[7] "Menos de 1 vez a la semana"
[8] "Nunca"
[9] "No consta"
> ncol(data_fr)
[1] 9
> nrow(data_fr)
[1] 210
```

Figura 4: Visualización de los nombres de columnas, números de variables e instancias

Es interesante observar que las únicas columnas con valores no numéricos (*chr*) son las tres primeras (Sexo, Nivel, Alimento), lo que supondrá poder realizar operaciones matemáticas con el resto de columnas.

También se ha querido comprobar que no existen *missing values* en nuestro dataframe. Estos valores o bien indican que no se ha realizado una medida para esa variable, o bien que dicha variable es *Not Aplicable* (p. ej. que -0.3 miles de HOMBRES nivel II coman Pescado diariamente).

Utilizando la línea de código `summarise_all(~sum(is.na(data_fr)))`, se comprueba fácilmente si existen o no valores desconocidos, y el número por columna si los hubiera.

Por ejemplo, si en la variable “Sexo” hubiese un valor desconocido indicaría que para esa observación no se ha recogido el sexo del individuo; ocurre lo mismo en cualquiera de las otras variables. En este dataframe, parece ser que no hay *missing values* por lo que no es necesario limpiarlo o realizar cálculos predictivos.

```
> data_fr %>%
+   summarise_all(~sum(is.na(data_fr)))
# A tibble: 1 × 9
   Sexo Nivel Alimento `A diario` 3 o más veces a la se…¹ 1 o 2 veces a la sem…² Menos de 1 vez a la …³ Nunca `No consta`
  <int> <int>    <int>      <int>                  <int>                 <int>                 <int> <int>       <int>
1     0     0        0          0                      0                     0                     0     0           0
# ℹ abbreviated names: ¹`3 o más veces a la semana, pero no a diario`, ²`1 o 2 veces a la semana`,
#   ³`Menos de 1 vez a la semana`
```

Figura 5: Búsqueda de missing values

Que no haya *missing values* es indicativo de que se ha hecho correctamente el trabajo de encuestar a los individuos, y no ha faltado por preguntar a ningún hombre o mujer ningún nivel específico por algún alimento en específico.

La columna “No consta”, en cambio, sí parece que tenga varias filas con valor en dicha variable.

```
> print(select(data_fr, "No consta"))
# A tibble: 210 × 1
   `No consta`
         <dbl>
 1         0
 2         0
 3         0
 4         0
 5         5.1
 6         2
 7         0
 8         0
 9         1.7
10         0
# ℹ 200 more rows
# ℹ Use `print(n = ...)` to see more rows
```

Figura 6: Exploración de la columna “No consta”

Si bien se podría interpretar “No consta” como un valor desconocido, esto podría no ser así. Es decir, hay varios motivos por los que podrían producirse valores que “no constan”. Para empezar, la persona a la que se le ha interrogado para extraer los datos podría no saber cada cuánto ingiere un alimento concreto, por lo que, preferible a dar datos erróneos, opta por “no responder”. En cambio, podría ser un motivo completamente diferente, como que el interrogador no apuntó la respuesta del individuo, o bien porque no le entendió bien (si era, por ejemplo, una encuesta telefónica), o por complicaciones ajenas al estudio. Podría también ocurrir que se haya perdido información por errores de conexión a Internet o errores al pasar datos de una tabla a otra.

Por tanto, la columna “No consta” es necesaria para no tener en cuenta datos falsos que podrían perjudicar al análisis de datos y estadístico que se va a realizar; puede servir para extraer conclusiones como buscar métodos más eficientes en próximas ocasiones para reducir errores de transcripción o pérdida de datos o incluso porque si en una observación concreta, se “pierde” la respuesta de uno de los encuestados, esta pasaría a ser NA, convirtiendo a toda la observación en NA.

APARTADO 2

***Utilizando nombres más convenientes, renombra y acorta el nombre de las columnas que contienen espacios o que no son válidos en R. Esto te ayudará a nombrarlas más adelante sin problemas*.**

```
# APARTADO 2
(df <- transmute(data_fr, # se crea un nuevo dataframe, df,
    sexo = Sexo,          # copia del original, data_df, al
    nivel = Nivel,        # que se le renombran las columnas
    alimento = Alimento,  # usando transmute
    diario = `A diario`,
    tres = `3 o más veces a la semana, pero no a diario`,
    uno_dos = `1 o 2 veces a la semana`,
    uno = `Menos de 1 vez a la semana`,
    nunca = Nunca,
    no_consta = `No consta`
  ))
```

Figura 7: Código para resolver el apartado 2

Se hace uso de la función `transmute` para cambiar el nombre de las columnas. Se ha querido simplificar el nombre de las columnas, poniendo nombres fáciles de entender, y una o dos palabras por nombre de variable (si son dos, separadas por guion bajo); siempre en minúscula. Es importante el uso de coma invertida “` ” para especificar los nombres con espacios o caracteres numéricos.

```
> (df <- transmute(data_fr,
+     sexo = Sexo,
+     nivel = Nivel,
+     alimento = Alimento,
+     diario = `A diario`,
+     tres = `3 o más veces a la semana, pero no a diario`,
+     uno_dos = `1 o 2 veces a la semana`,
+     uno = `Menos de 1 vez a la semana`,
+     nunca = Nunca,
+     no_consta = `No consta`
+   ))
# A tibble: 210 × 9
   sexo    nivel alimento                       diario  tres uno_dos   uno nunca no_consta
   <chr>   <chr> <chr>                           <dbl> <dbl>   <dbl> <dbl> <dbl>     <dbl>
 1 HOMBRES I     Fruta fresca (excluye zumos)     1763.   535.   190.   72.4  43.9       0
 2 HOMBRES I     Carne                             229.  1740.   572.   29.9  33.3       0
 3 HOMBRES I     Huevos                             44.6  829.  1534.  170.   27.6       0
 4 HOMBRES I     Pescado                            24.8 1090.  1298.  143.   48.8       0
 5 HOMBRES I     Pasta, arroz, patatas             236   1663.   661.   33.8   5.3       5.1
 6 HOMBRES I     Pan, cereales                    2127.   256.   132.   56.1  31.4       2
 7 HOMBRES I     Verduras, ensaladas y hortalizas  989.  1311.   250.   41.3  13.2       0
 8 HOMBRES I     Legumbres                          16.9  620.  1675.  249    44         0
 9 HOMBRES I     Embutidos y fiambres              396.  1026.   714.  309.  158.        1.7
10 HOMBRES I     Productos lácteos                2197.   239.    65.4  34.1  69.3       0
# ℹ 200 more rows
# ℹ Use `print(n = ...)` to see more rows
```

Figura 8: Comprobación de que se cambia el nombre de las columnas correctamente

APARTADO 3

Añade una columna nueva a la tabla con el nombre "Total" con el total de casos sin considerar la columna etiquetada como "No consta".

```
# APARTADO 3
df <- df %>% # en el datafra,e creado, df, se genera una nueva columna, Total
  mutate(Total = rowSums(select(df, 4:8), na.rm = TRUE)) # se utiliza mutate

df

df <- df %>%
  mutate(Total = diario+tres+uno_dos+uno+nunca)
df
```

Figura 9: Código para resolver el apartado 3

Se utiliza la función `mutate` para crear una nueva columna, llamada "Total", que constará de la suma de las columnas de la 4 a la 8 (`rowSums(select(df, 4:8))`. `na.rm = TRUE` permite que sólo se tenga en cuenta valores que no son *missing values*.

```
> # APARTADO 3
> df <- df %>%
+   mutate(Total = rowSums(select(df, 4:8), na.rm = TRUE))
> df
# A tibble: 210 × 10
   sexo    nivel alimento            diario  tres uno_dos   uno nunca no_consta Total
   <chr>   <chr> <chr>                <dbl> <dbl>   <dbl> <dbl> <dbl>      <dbl> <dbl>
 1 HOMBRES I     Fruta fresca (excl… 1763.   535.   190.   72.4  43.9       0   2605.
 2 HOMBRES I     Carne                229.  1740.   572.   29.9  33.3       0   2605.
 3 HOMBRES I     Huevos                44.6  829.  1534.  170.   27.6       0   2605.
 4 HOMBRES I     Pescado               24.8 1090.  1298.  143.   48.8       0   2605.
 5 HOMBRES I     Pasta, arroz, pata…  236   1663.   661.   33.8   5.3       5.1 2600.
 6 HOMBRES I     Pan, cereales       2127.   256.   132.   56.1  31.4       2   2603.
 7 HOMBRES I     Verduras, ensalada…  989.  1311.   250.   41.3  13.2       0   2605.
 8 HOMBRES I     Legumbres             16.9  620.  1675.  249    44         0   2605.
 9 HOMBRES I     Embutidos y fiambr…  396.  1026.   714.  309.  158.        1.7 2603
10 HOMBRES I     Productos lácteos   2197.   239.    65.4  34.1  69.3       0   2605.
# ℹ 200 more rows
# ℹ Use `print(n = ...)` to see more rows
```

Figura 10: Creación de la columna "Total"

Se adjunta también en la Figura 11 la comprobación de código, utilizando una línea de código algo más manual; corroborando así que ambas opciones son equivalentes.

```
> df <- df %>%
+   mutate(Total = diario+tres+uno_dos+uno+nunca)
> df
# A tibble: 210 × 11
   sexo    nivel alimento          diario  tres uno_dos   uno nunca no_consta Total
   <chr>   <chr> <chr>              <dbl> <dbl>   <dbl> <dbl> <dbl>      <dbl> <dbl>
 1 HOMBRES I     Fruta fresca (ex… 1763.   535.   190.   72.4  43.9       0   2605.
 2 HOMBRES I     Carne              229.  1740.   572.   29.9  33.3       0   2605.
 3 HOMBRES I     Huevos              44.6  829.  1534.  170.   27.6       0   2605.
 4 HOMBRES I     Pescado             24.8 1090.  1298.  143.   48.8       0   2605.
 5 HOMBRES I     Pasta, arroz, pa…  236   1663.   661.   33.8   5.3       5.1 2600.
 6 HOMBRES I     Pan, cereales     2127.   256.   132.   56.1  31.4       2   2603.
 7 HOMBRES I     Verduras, ensala…  989.  1311.   250.   41.3  13.2       0   2605.
 8 HOMBRES I     Legumbres           16.9  620.  1675.  249    44         0   2605.
 9 HOMBRES I     Embutidos y fiam…  396.  1026.   714.  309.  158.        1.7 2603
10 HOMBRES I     Productos lácteos 2197.   239.    65.4  34.1  69.3       0   2605.
# ℹ 200 more rows
# ℹ 1 more variable: Media_semanal <dbl>
# ℹ Use `print(n = ...)` to see more rows
```

Figura 11: Comprobación de código

APARTADO 4

Utilizando la columna "Total" calculada en el apartado anterior, estima el número medio de días a la semana que se consume cada uno de los alimentos para cada fila de tabla utilizando la tabla.

Justifica el porqué de estos valores y explica qué fórmula vas a utilizar para realizar este cálculo. Llama a esta nueva columna "Media_semanal". ¿Tienen sentido los datos de esa columna? ¿Cómo deben interpretarse?

Se parte de una tabla con ciertas columnas en las que se han clasificado individuos, pertenecientes a cierto sexo y nivel socioeconómico, que comen un alimento específico, según el número de veces a la semana que lo comen. El objetivo en este ejercicio es hacer una media de veces que se come un cierto alimento, por tanto, y utilizando la columna "Total" del ejercicio anterior, se ha optado por ponderar la frecuencia y a partir de ahí realizar la media. La ponderación es sencilla, se contabiliza la frecuencia por número de días, por tanto, "a diario" pondera por 7, "3 o más veces" por 4.5... y así sucesivamente, hasta que "nunca" pondera por 0.

En nuestro código, se ha optado por considerar que la columna "no consta" pondera también por 0, es decir, no contabilizarla.

```
# APARTADO 4
df <- mutate(df, Media_semanal = (diario * 7 +                    # en df, creamos una nueva columna llamada
                                    tres * 4.5 + uno_dos * 1.5    # Media_semanal, que es la suma de la
                                  + uno * 0.5 + nunca * 0+        # ponderación de los individuos según cuántas
                                    + no_consta * 0)/Total)       # veces comen cierto alimento a la semana
df                                                                # (a diario tiene más ponderación que uno;
                                                                  # y nunca, cuenta por 0 por ejemplo)

(df %>% summarise(min_media = min(Media_semanal))) # comprobamos que el valor mínimo no sea menor que 0
(df %>% summarise(max_media = max(Media_semanal))) # comprobamos que el valor máximo no sea mayor que 7
```

Figura 12: Código para resolver el apartado 4

```
> df <- mutate(df, Media_semanal = (diario * 7 + tres * 4.5 + uno_dos * 1.5 + uno * 0.5 + nunca * 0+ + no_consta * 0)/Total)
> df
# A tibble: 210 × 11
   sexo    nivel alimento                      diario  tres uno_dos   uno nunca no_consta Total Media_semanal
   <chr>   <chr> <chr>                          <dbl> <dbl>   <dbl> <dbl> <dbl>      <dbl> <dbl>         <dbl>
 1 HOMBRES I     Fruta fresca (excluye zumos)    1763.   535.   190.   72.4  43.9        0   2605.          5.79
 2 HOMBRES I     Carne                            229.  1740.   572.   29.9  33.3        0   2605.          3.96
 3 HOMBRES I     Huevos                            44.6  829.  1534.  170.   27.6        0   2605.          2.47
 4 HOMBRES I     Pescado                           24.8 1090.  1298.  143.   48.8        0   2605.          2.73
 5 HOMBRES I     Pasta, arroz, patatas            236   1663.   661.   33.8   5.3        5.1 2600.          3.90
 6 HOMBRES I     Pan, cereales                   2127.   256.   132.   56.1  31.4        2   2603.          6.25
 7 HOMBRES I     Verduras, ensaladas y hortalizas 989.  1311.   250.   41.3  13.2        0   2605.          5.08
 8 HOMBRES I     Legumbres                         16.9  620.  1675.  249    44          0   2605.          2.13
 9 HOMBRES I     Embutidos y fiambres             396.  1026.   714.  309.  158.        1.7 2603           3.31
10 HOMBRES I     Productos lácteos               2197.   239.    65.4  34.1  69.3        0   2605.          6.36
# ℹ 200 more rows
# ℹ Use `print(n = ...)` to see more rows
```

Figura 13: Creación de una nueva columna, Media_semanal

Por último, se ha querido comprobar que el cálculo se ha realizado de forma correcta, buscando, a lo largo de la columna creada (Media_semanal), el máximo y el mínimo valor, para asegurar que el mínimo valor no es menor que 0 y el máximo, no mayor que 7.

```
> (df %>% summarise(min_media = min(Media_semanal))) # comprobamos que el valor mínimo no sea menor que 0
# A tibble: 1 × 1
  min_media
      <dbl>
1     0.700
> (df %>% summarise(max_media = max(Media_semanal))) # comprobamos que el valor máximo no sea mayor que 7
# A tibble: 1 × 1
  max_media
      <dbl>
1      6.54
```

Figura 14: Código para verificar que tienen sentido los resultados

APARTADO 5

Analiza los datos de las mujeres de nivel I, ordénalos de mayor consumo medio semanal a menor. ¿En qué consiste su dieta media diaria?

En el código, se buscan aquellas mujeres que, en la columna nivel, pertenecen al nivel I, y que en la columna sexo, pertenecen a "MUJERES" utilizando la función `filter`. Una vez se tiene dichas filas, se quieren ordenar de mayor a menor consumo medio semanal (`arrange(desc(Media_semanal))`). Por último, se muestra por pantalla sólo las dos columnas de interés, alimento y Media_semanal (haciendo uso de `select` y `print`).

```
# APARTADO 5
filter(df, sexo == "MUJERES" & nivel == "I") %>% # se filtran en el dataset df, aquellas mujeres, de nivel I,
  arrange(desc(Media_semanal)) %>%               # ordenadas de mayor a menor (según la media semanal),
  select(alimento, Media_semanal) %>%            # y se seleccionan las columnas alimento y media semanal
  print(n = 20)                                  # para mostrarlas por pantalla
```

Figura 15: Código para resolver el apartado 5

```
> filter(df, sexo == "MUJERES" & nivel == "I") %>%
+   arrange(desc(Media_semanal)) %>%
+   select(alimento, Media_semanal) %>%
+   print(n = 20)
# A tibble: 15 × 2
   alimento                              Media_semanal
   <chr>                                         <dbl>
 1 Productos lácteos                              6.26
 2 Fruta fresca (excluye zumos)                   6.21
 3 Pan, cereales                                  6.15
 4 Verduras, ensaladas y hortalizas               5.63
 5 Carne                                          3.75
 6 Pasta, arroz, patatas                          3.64
 7 Dulces                                         3.39
 8 Embutidos y fiambres                           2.98
 9 Pescado                                        2.82
10 Zumo natural de frutas o verduras              2.48
11 Huevos                                         2.41
12 Legumbres                                      2.04
13 Aperitivos o comidas saladas de picar          0.977
14 Comida rápida                                  0.829
15 Refrescos con azúcar                           0.803
```

Figura 16: Código para analizar los datos de las mujeres de nivel I, ordénalos de mayor consumo medio semanal a menor

A la vista de los datos obtenidos, se puede estimar que las mujeres directoras y gerentes de establecimientos de 10 o más asalariados/as y profesionales tradicionalmente asociados/as a licenciaturas universitarias tienden a consumir muchos productos lácteos, frutas y cereales, basando su dieta en estos alimentos y el consumo semanal de verduras, carne, pasta, arroz y patatas. Si bien el consumo de pescado debería ser mayor, en general mantienen una dieta saludable y balanceada, pudiéndose reducir el consumo de dulces y quizás de fiambres.

APARTADO 6

Utiliza la nueva columna "Media_semanal" calculada en el apartado anterior para estimar el consumo medio semanal general de cada uno de los alimentos. Para ello, genera una gráfica adecuada que permita visualizar los resultados ordenados de mayor a menor. ¿Qué conclusiones se obtienen?

```r
# APARTADO 6
comida <- df %>%                        # creamos una nueva variable llamda comida, a partir de df
  group_by(alimento) %>%                # queremos agrupar por los tipos de alimentos, y a partir
  summarise(                            # de ahí, calcular la media de las medias de cada tipo
    media_alimento = mean(Media_semanal) # de alimento
  ) %>%
  arrange(desc(media_alimento))               # ordenamos de mayor a menor

comida

comida %>%
  ggplot(aes(x = reorder(alimento, -media_alimento), y = media_alimento, fill = alimento)) + # orden descendente y cada alimento un color
  geom_bar(stat = "identity") +                          # los valores en el eje y representan directamente las alturas de las barras
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +                  # para no solapar los nombres de los alimentos en el eje x
  labs(title = "Consumo medio semanal general de cada uno de los alimentos")   # título
```

Figura 17: Código para resolver el apartado 6

Dado que se pretende encontrar cuántas veces se consume en promedio cada alimento, se debe agrupar primero por alimento (group_by(alimento); es como generar una nueva tabla para cada alimento). La función group_by en R, agrupa datos según una o más variables (en este caso, una). Una vez agrupados los datos, se puede operar con ellos, utilizando funciones como summarize(), mutate(), filter(); en el caso que nos ocupa, después de agrupar por alimento, se ha aplicado la función de resumen (summarize()) para poder así calcular la media de días que se consume cada alimento en específico (`summarise(mean(Media_semanal))`). Esta función nos permite obtener un resumen, como una tabla pequeñita a partir del `group_by` en el que se mostrará la media.

Además, tal y como pide el enunciado, se ordenan los resultados de mayor a menor frecuencia. La función `arrange()` como tal, ordena de menor a mayor, por tanto, se utiliza con ella `desc()`, que ordena por orden descendente, es decir, de mayor a menor.

```
> comida
# A tibble: 15 × 2
   alimento                                media_alimento
   <chr>                                            <dbl>
 1 Pan, cereales                                    6.39
 2 Productos lácteos                                6.34
 3 Fruta fresca (excluye zumos)                     5.63
 4 Verduras, ensaladas y hortalizas                 5.04
 5 Pasta, arroz, patatas                            3.99
 6 Carne                                            3.85
 7 Dulces                                           3.46
 8 Embutidos y fiambres                             3.08
 9 Pescado                                          2.57
10 Huevos                                           2.41
11 Legumbres                                        2.17
12 Zumo natural de frutas o verduras                1.94
13 Refrescos con azúcar                             1.36
14 Aperitivos o comidas saladas de picar            1.06
15 Comida rápida                                    0.979
```

Figura 18: Comprobación del código del apartado 6; salida

Se quiere representar un gráfico de barras con los datos obtenidos. Para ello, se hace uso de `ggplot2`, que en R es una librería de visualización de datos que permite, de una manera sencilla, representar gráficamente información. En este caso, viene acompañado de `geom_bar(stat = "identity")`, que representa los valores en el eje y como alturas numéricas. En el eje x se representan los distintos tipos de alimentos, ordenados de mayor a menor (se ha debido usar `x = reorder(alimento, -media_alimento)` para representarlo de esa manera; orden descendente de la media de *Media_semanal* para cada alimento), mientras que en el eje y, se representan las alturas según las medias de frecuencia con la que se consume cada alimento. Sin la línea de código `theme(axis.text.x = element_text(angle = 45, hjust = 1))`, se nombran los alimentos (eje x) en diagonal (sin ella, se hubieran solapado los nombres). Para evitar el problema del solape, se podría haber intercambiado los ejes, pero creemos que esta visualización es más adecuada y clara.

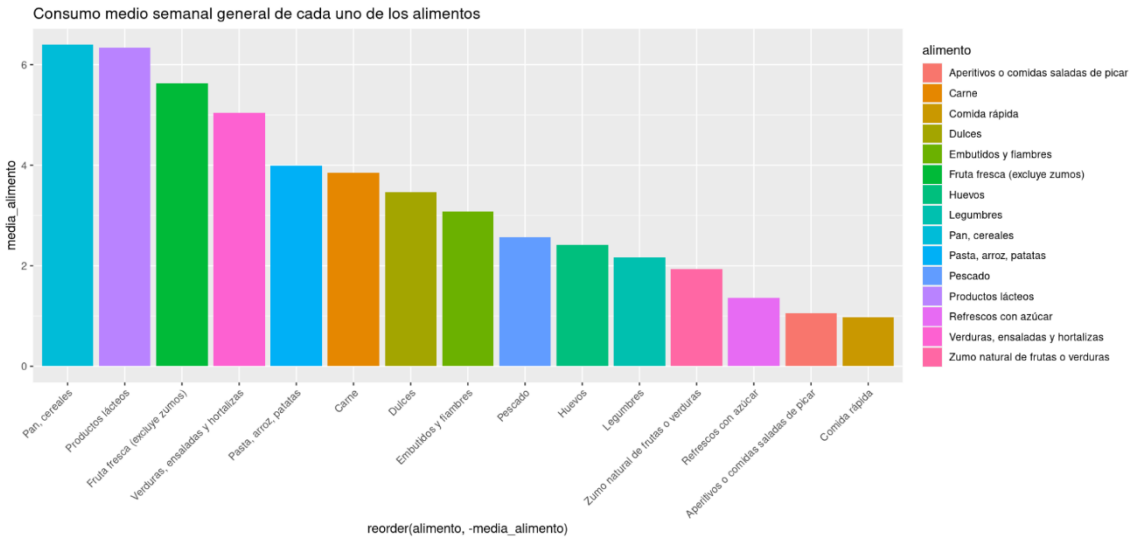


Figura 19: Gráfico de barras; visualización de resultados apartado 6

Conclusiones extraídas del gráfico serían que, por lo general, los alimentos que más se consumen a la semana son pan, cereales y productos lácteos. Seguidos por frutas, verduras y carbohidratos. Lo que menos se consume, y en muchísima desproporción son refrescos con azúcar, aperitivos y comida rápida. Si se compara con una pirámide alimenticia típica, podremos ver que, por lo general, coindicen ambos esquemas, lo que nos da una idea de que a la población le gusta cuidarse y es consciente de qué alimentos son los más saludables y en qué proporción deben consumirse.

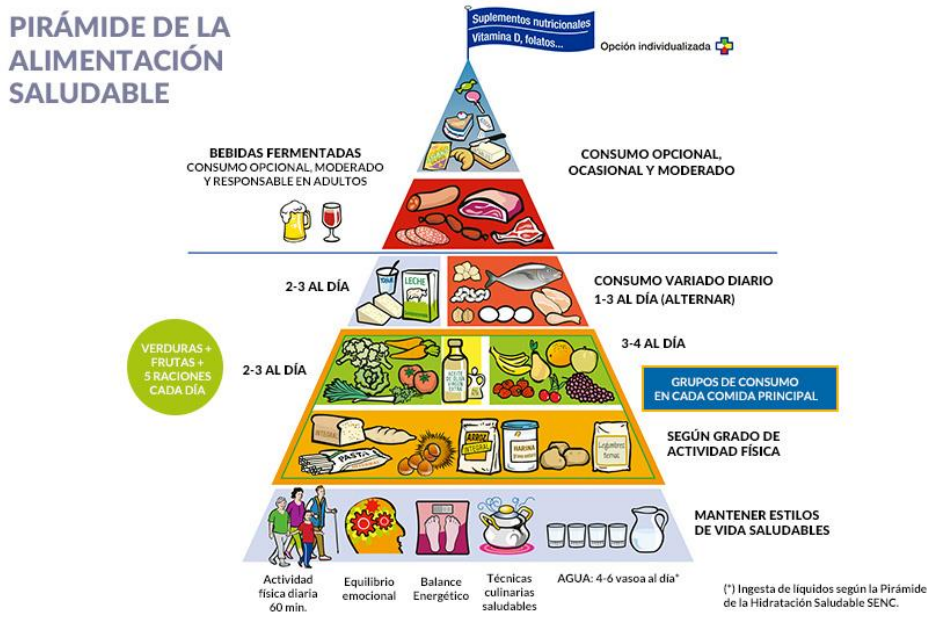


Figura 20: Ejemplo pirámide alimenticia; ayuda visual

Las diferencias más destacables son que los huevos, legumbres y pescado deberían consumirse más que carnes y embutidos, y en cambio, la carne lidera la lista de los alimentos más frecuentemente consumidos a la semana.

APARTADO 7

Obtén la media de días a la semana de consumo de comida rápida por grupo socioeconómico ordenado de mayor a menor. ¿Qué grupos consumen más y menos veces a la semana este tipo de alimento? ¿Encuentras alguna explicación?

```
# APARTADO 7
c_basura<-df %>%  # guardamos en una nueva variable, c_basura, y trabajamos con df
  group_by(nivel) %>% # agrupamos por los distintos niveles socioeconómicos
  filter(alimento=="Comida rápida") %>% # y aplicamos dos filtros, unos para buscar sólo aquellos alimentos que corresponden a comida rápida
  filter(nivel != "No consta") %>%      # y otro para no tener en cuenta el nivel "No consta"
  summarise(media_c_basura = mean(Media_semanal)) %>% # resumimos y aplicamos la operación mean (que realizará la media de la frecuencia
  arrange(desc(media_c_basura))                       # en la que hombres y mujeres de cada nivel, que ingieren Comida rápida
c_basura
```

Figura 21: Código para resolver el apartado 7

Para encontrar qué grupos consumen mayor cantidad de comida basura a la semana, se utilizan funciones ya vistas en anteriores apartados: se agrupa por niveles (`group_by(nivel)`) y se filtra para buscar sólo aquellas filas cuya columna alimento tiene el valor de “Comida rápida” (`filter(alimento=="Comida rápida")`), y un segundo filtro que rechaza la columna “No consta” (`filter(nivel != "No consta")`). Ahora, se resume la operación en que se busca la media de cada fila que consume “Comida rápida”, agrupando por nivel (`summarise(media_c_basura = mean(Media_semanal))`); es decir, se va a promediar la frecuencia con la que se consume “Comida basura” de hombres y mujeres del nivel I, nivel II y así sucesivamente.

Finalmente, como el enunciado solicita que se ordene de mayor a menor, se aplica `arrange(desc(media_c_basura))`.

```
> c_basura
# A tibble: 6 × 2
  nivel media_c_basura
  <chr>          <dbl>
1 VI             1.08
2 V              1.05
3 III            0.966
4 IV             0.964
5 II             0.908
6 I              0.886
```

Figura 22: Salida apartado 7

Es interesante observar que existe cierta tendencia directamente proporcional entre pertenecer a un nivel socioeconómico bajo, y consumir más comida basura (salvo el III y el IV que se intercambian). A la vista está que los pertenecientes al grupo I no llegan apenas a consumir una vez por semana comida basura, mientras que los grupos V y VI superan la vez por semana.

Algunas de los argumentos que pueden sustentar la idea de que, a menor poder adquisitivo, mayor tendencia a consumir comida basura son los siguientes: comer comida basura es relativamente barato y, por tanto, un fácil y económico plan para pasar en familia el fin de semana. Además, podría sugerir que cuanto menor poder adquisitivo, menos tiempo o ganas se tienen de cocinar a diario, por lo que es una opción fácil apoyarse de vez en cuando en la comida basura o comida rápida. Aun así, no se tienen datos concluyentes de que un mayor nivel económico suponga que se es más saludable, sólo que no se tiende tanto a consumir comida basura, pero no implica, ni se recoge en este estudio, que no se opte por ir los fines de semana a restaurantes de mayor categoría, algo más caros, pero más saludables.

APARTADO 8

Obtén las distribuciones en el consumo de alimentos entre hombres y mujeres, y represéntalos gráficamente mediante diferencias entre las medias. ¿Existe algún tipo de diferencia significativa? ¿Cuáles crees que son las causas y las consecuencias de estas diferencias? ¿Quién cuida más de su salud? Aporta algún dato que lo respalde.

```
# APARTADO 8
comida_diferencia <- df %>%  # guardamos en una nueva variable, llamada comida_diferencia
  group_by(sexo, alimento, Media_semanal) %>%  # agrupamos por sexo, alimento y Media_semanal
  pivot_wider(names_from = sexo, values_from = Media_semanal # pivotamos "a lo ancho", generando dos nuevas columnas para
  ) %>%              # la variable sexo (HOMBRES y MUJERES), que se rellenan con los valores de Media_semanal correspondientes
  summarise(
    Diferencia = mean(HOMBRES, na.rm = TRUE) - mean(MUJERES, na.rm = TRUE) # se resume la operación en crear una nueva variable,
  ) # Diferencia, que es la resta entre la media de los valores de la columna HOMBRES y la media de MUJERES (agrupado por alimento)

ggplot(comida_diferencia, aes(x = Diferencia, y =reorder(alimento,Diferencia))) + # ordneamos por diferencia
  geom_bar(stat = "identity", position = "dodge", fill="pink") + #
  labs(title = "Diferencia en el Consumo de Alimentos entre Hombres y Mujeres por Alimento", # "dodge" coloca las barras lado a lado
       x = "Diferencia Media", y = "Alimento") +
  theme_minimal() # eliminamos elementos innecesarios como líneas de fondo y cuadrículas
```

Figura 23*: Código para resolver el apartado 8*

En este apartado se busca la diferencia en el consumo de alimentos entre hombres y mujeres por alimento. Es decir, ya no se trata sólo de agrupar por sexo y alimento, sino que se quiere construir un gráfico que simplifique la comprensión de qué sexo se alimenta más de qué alimento, representando únicamente la diferencia de consumo.

Por tanto, tras agrupar por sexo, alimento y Media_semanal, se opta por utilizar `pivot_wider(names_from = sexo, values_from = Media_semanal)`, que pivota "a lo ancho", es decir, crea columnas según el tipo de sexo (en este caso, dos columnas, HOMBRES y MUJERES), y se rellenan dichas columnas con los valores correspondientes. Llegados a este punto del código, intrínsecamente se está viendo algo así:

```
> comida_diferencia <- df %>%  # guardamos en una nueva variable, llamada comida_diferencia
+     group_by(sexo, alimento, Media_semanal) %>%  # agrupamos por sexo, alimento y Media_semanal
+     pivot_wider(names_from = sexo, values_from = Media_semanal
+     )
> comida_diferencia
# A tibble: 210 × 11
# Groups:   alimento [15]
   nivel alimento                          diario  tres uno_dos   uno nunca no_consta Total HOMBRES MUJERES
   <chr> <chr>                              <dbl> <dbl>   <dbl> <dbl> <dbl>      <dbl> <dbl>   <dbl>   <dbl>
 1 I     Fruta fresca (excluye zumos)      1763.   535.   190.   72.4  43.9       0   2605.    5.79      NA
 2 I     Carne                              229.  1740.   572.   29.9  33.3       0   2605.    3.96      NA
 3 I     Huevos                              44.6  829.  1534.  170.   27.6       0   2605.    2.47      NA
 4 I     Pescado                             24.8 1090.  1298.  143.   48.8       0   2605.    2.73      NA
 5 I     Pasta, arroz, patatas              236   1663.   661.   33.8   5.3       5.1 2600.    3.90      NA
 6 I     Pan, cereales                     2127.   256.   132.   56.1  31.4       2   2603.    6.25      NA
 7 I     Verduras, ensaladas y hortalizas   989.  1311.   250.   41.3  13.2       0   2605.    5.08      NA
 8 I     Legumbres                           16.9  620.  1675.  249    44        0   2605.    2.13      NA
 9 I     Embutidos y fiambres               396.  1026.   714.  309.  158.       1.7 2603     3.31      NA
10 I     Productos lácteos                 2197.   239.    65.4  34.1  69.3       0   2605.    6.36      NA
# ℹ 200 more rows
# ℹ Use `print(n = ...)` to see more rows

104 No consta Aperitivos o comidas saladas d…    7.9   57.5   163.   135.   116.        1.1  479.   1.31   NA
105 No consta Zumo natural de frutas o verdu…   61.1   95.4    76.4   69.2  177         1.1  479.   2.10   NA
106 I         Fruta fresca (excluye zumos)    2034.   402.    109.    25     44        1.4 2614.  NA       6.21
107 I         Carne                            192.  1654     662.    39.9   62.6      4.8 2611.  NA       3.75
108 I         Huevos                            21.6  836    1523.   184.    40.8      9.5 2606   NA       2.41
109 I         Pescado                           33.8 1156.   1238.   137.    45.9      4.8 2611.  NA       2.82
110 I         Pasta, arroz, patatas            161.  1613.    723.    82.1   32.4      4.8 2611.  NA       3.64
111 I         Pan, cereales                   2075.   292.    146.    35.1   66.9      0   2616.  NA       6.15
112 I         Verduras, ensaladas y hortaliz… 1413.  1026.    138.    26.5   12.2      0   2615.  NA       5.63
113 I         Legumbres                         19.9  539.   1756.   251.    47.7      1.2 2614.  NA       2.04
114 I         Embutidos y fiambres             328.   909.    800.   369.   201.       8.1 2607.  NA       2.98
115 I         Productos lácteos               2194.   176.    122.    46.5   75.9      2   2614.  NA       6.26
116 I         Dulces                           677.   652.    638    423    221.       4.8 2611.  NA       3.39
```

Figura 24: *Salida pivot_wider para resolución apartado 8*

Donde se están distribuyendo los valores de Media_semanal según si se es hombre, o mujer (quedando la columna libre como NA). Ahora, se termina la operación realizando un resumen en el que se obtenga una nueva variable, Diferencia, correspondiente a restar la media de hombres a la media de mujeres (sin contabilizar los NA), clasificados por alimentos (`Diferencia = mean(HOMBRES, na.rm = TRUE) - mean(MUJERES, na.rm = TRUE)`).

```
comida_diferencia <- df %>%
  group_by(sexo, alimento, Media_semanal) %>%
  pivot_wider(names_from = sexo, values_from = Media_semanal
  ) %>%
  summarise(
    Diferencia = mean(HOMBRES, na.rm = TRUE) - mean(MUJERES, na.rm = TRUE)
  )

comida_diferencia
```

```
> comida_diferencia
# A tibble: 15 × 2
   alimento                              Diferencia
   <chr>                                      <dbl>
 1 Aperitivos o comidas saladas de picar      0.171
 2 Carne                                      0.262
 3 Comida rápida                              0.269
 4 Dulces                                     0.0797
 5 Embutidos y fiambres                       0.539
 6 Fruta fresca (excluye zumos)              -0.442
 7 Huevos                                     0.0514
 8 Legumbres                                  0.0581
 9 Pan, cereales                              0.0558
10 Pasta, arroz, patatas                      0.134
11 Pescado                                   -0.157
12 Productos lácteos                         0.0519
13 Refrescos con azúcar                      0.412
14 Verduras, ensaladas y hortalizas         -0.569
15 Zumo natural de frutas o verduras        -0.128
```

Figura 25: Cálculo y salida diferencias apartado 8

Para realizar la gráfica, se vuelve a hacer uso de `ggplot`, y toma el eje de las x para representar la variable numérica (la diferencia media por alimento), y en el eje y, el alimento. Para embellecer y facilitar la comprensión del gráfico, se utiliza `geom_bar(stat = "identity", position = "dodge", fill="pink")`, que especifica las alturas de los valores de cada alimento, pone las barras de lado a lado y pinta las barras de rosa.

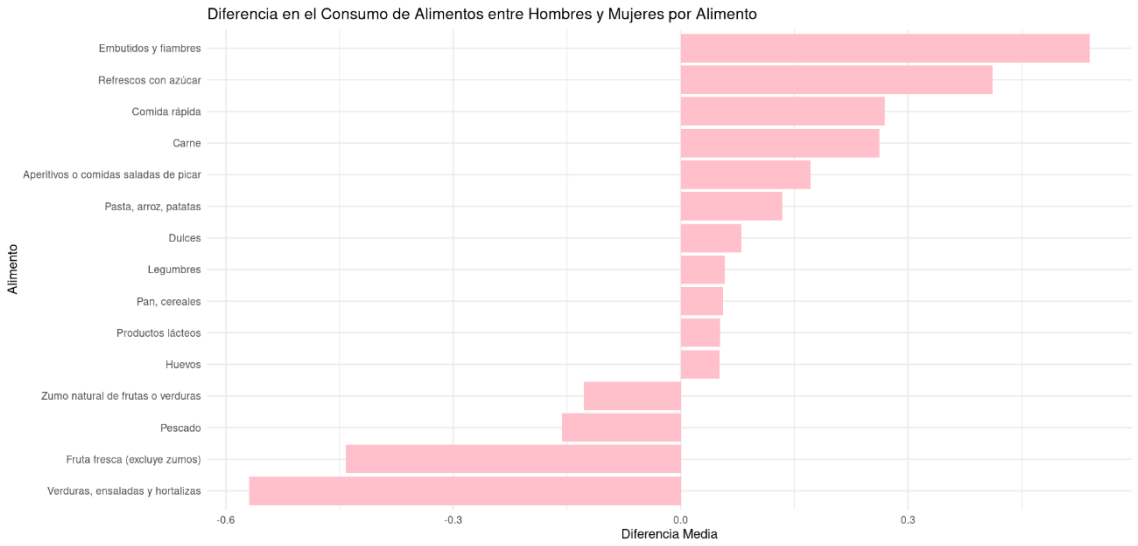


Figura 26: *Gráfico de barras, diferencias de consumo entre hombres y mujeres para cada alimento*

Destaca que, por lo general, los hombres comen más que las mujeres (barras positivas), probablemente porque los hombres por anatomía necesitan más energía, y, por tanto, comer más. Sin embargo, hay cuatro alimentos que las mujeres consumen más que los hombres: zumo, pescado, fruta y verdura. Son por excelencia los alimentos más saludables de la pirámide alimenticia, lo que nos sugiere que las mujeres comen menos, y además cuidan su dieta, buscando priorizar la salud al placer de comer alimentos más grasos. Las diferencias apenas son significativas en realidad, pero es cierto que los hombres no se privan de embutidos y bebidas azucaradas, consumiendo 0.5 más que las mujeres, que esa misma proporción sacan en diferencia a los hombres en el caso de las verduras.

Aunque no lo solicita el enunciado, se ha querido representar otro gráfico, que demuestra que las diferencias no son exageradas, salvo para los alimentos destacados previamente.

```
comida_sexo <- df %>%
  group_by(alimento, sexo) %>%
  summarise(
    frecuencia = mean(Media_semanal)
  ) %>%
  arrange(frecuencia)

ggplot(comida_sexo, aes(x = reorder(alimento, frecuencia), y = frecuencia, fill = sexo)) +
  geom_bar(stat = "identity", position = "dodge")+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

Figura 27: *Código gráfico adicional*

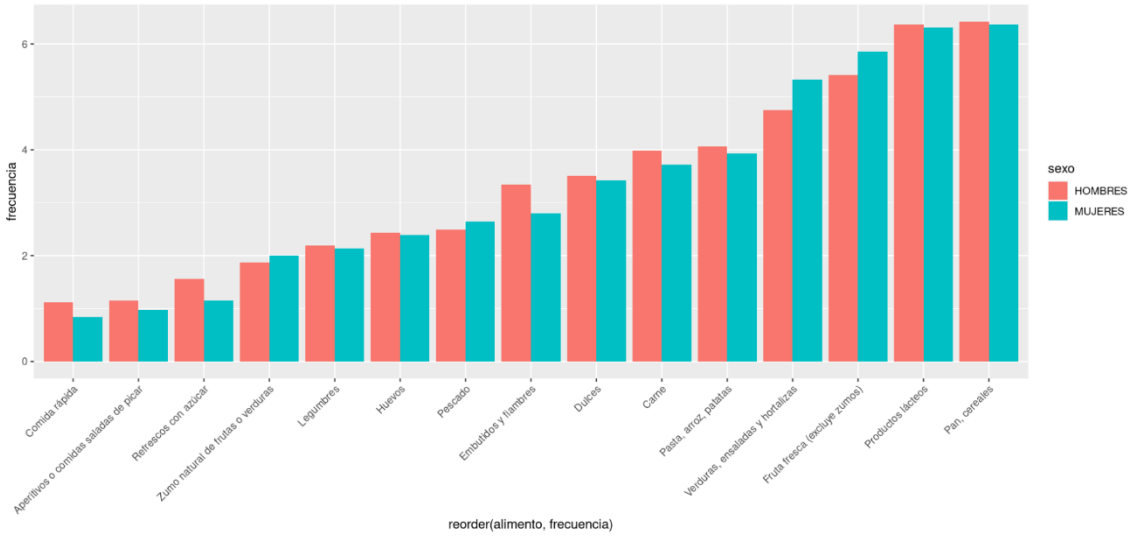


Figura 28: *Gráfico de barras adicional, diferencias de consumo de hombres y mujeres según cada alimento*

Cierto que existe cierta tendencia a que los hombres comen más que las mujeres, se puede observar cierto equilibrio, y por supuesto, tendencia a mantener la regla de la pirámide alimenticia.

APARTADO 9

Calcula el presupuesto semanal por nivel socioeconómico utilizando la siguiente tabla "coste_diario" en la que tendrás que rellenar los precios acordados en clase.

Para calcular el presupuesto semanal, se ha hecho un exhaustivo estudio en un supermercado español muy extendido, tanto en las grandes ciudades, como en regiones de menor extensión, el "Carrefour".

Accediendo a la página online del "Carrefour" (https://www.carrefour.es/?gad_source=1&gclid=CjwKCAiAuYuvBhApEiwAzq_YiU8E5atfHxlyx8fS1hV33KDyX9lWgdYqQP_Y6WKnq9ojS5qQG3VJiBoCVigQAvD_BwE&gclsrc=aw.ds

), se han comparado los precios de una variedad de alimentos, la mayoría de manera aleatoria, que se categorizarían en los tipos de alimentos de estudio que nos ocupa. Es decir, se ha realizado un muestreo, y se ha calculado la media de precios de varios alimentos para buscar el precio estimado de cada categoría.

Se ha generado la siguiente tabla, en la que se muestra cuánto se estima que cuesta comer al día, cierto alimento (el día que se consume dicho alimento). Observar que hay alimentos muy caros (pescado, carne y comida rápida), en comparación con otros muy baratos (huevos, pasta y cereales). A partir de esta tabla, y conociendo la media frecuencial con la que se consume cada alimento por nivel socioeconómico, se va a intentar estudiar qué nivel invierte más dinero en comida.

Alimento	**Coste diario**
Fruta fresca (excluye zumos)	*1.12€*
Carne	*2.67€*
Huevos	*0.37€*
Pescado	*3.89€*
Pasta, arroz, patatas	*0.17€*
Pan, cereales	*0.24€*
Verduras, ensaladas y hortalizas	*1.28€*
Legumbres	*0.43€*
Embutidos y fiambres	*0.36€*
Productos lácteos	*1.02€*
Dulces	*1.26€*
Refrescos con azúcar	*1.62€*
Comida rápida	*9.45€*
Aperitivos o comidas saladas de picar	*1.29€*
Zumo natural de frutas o verduras	*0.61€*

Previo a dicho análisis, se muestra cómo se ha realizado el muestreo:

FRUTA FRESCA (EXCLUYE ZUMOS)

- Plátano: 1.99€/ kg. Estimamos un peso de 150g de media. **0.285€**/ud.
- Naranja: 1.89€/kg. Estimamos un peso de 250g.**0.4725€**/ud.
- Manzana: 205€/kg. Estimamos un peso de 180g. **0.369**€/ud.
- Kiwi: 4.99€/kg. Estimamos un peso de 130g. **0.6487€**/ud.
- Mango: 3.39€/kg. Estimamos un peso de 300g. **1.017€**/ud.

FRUTA		
€/kg	€/ud	MEDIA
1,99 €	0,30 €	0,56 €
1,89	0,47 €	
2,05	0,37 €	
4,99	0,65 €	
3,39	1,02 €	

Media: 0.56€ por pieza de fruta consumida

Asumimos que el día que se come fruta se comen 2 piezas de fruta: 1.12€

GRUPO 6362

CARNE

Estimamos la ración de carne es de 200g.

CARNE		
€/kg	€/200g	MEDIA
9,47 €	1,89 €	2,67 €
18,95	3,79 €	
8,32	1,66 €	
15,19	3,04 €	
14,89	2,98 €	

- Carne de ternera picada: 9,47 €/kg. **1.89€**/200g
- Lomo de añojo: 18.95€/kg. **3.79€**/200g.
- Solomillos de pollo campero: 15,19 €/kg. **3.04€**/200g.
- Pechuga de Pollo campero: 14,89 €/kg/kg. **2.98€**/200g
- Escalopín de lomo de cerdo fresco marinado: 8.32€/kg.**1.66€**/200g.
-

Media: 2.67€ por ración individual.

HUEVOS

Asumimos un consumo de 2 huevos al día. **0.365€**.

PESCADO

Estimamos un consumo de 200g por ración individual.

PESCADO		
€/kg	€/200g	MEDIA
23,95 €	4,79 €	3,89 €
14,95	2,99 €	
21,90	4,38 €	
17,50	3,50 €	
18,99	3,80 €	

- Salmón: 23,95 €/kg. **4.79€**/200g.
- Rodaballo: 14.95€/kg. **3.99€**/200g.
- Mejillones: 21.90€/kg. **4.38 €**/200g.
- Merluza en filete: 17,50 €/kg. **3.50€**/200g.
- Cola de rape: 18.99€/kg. **3.798€**/200g.

Media: 3.89€ por ración de pescado.

PASTA, ARROZ, PATATAS

- Espaguetis/ macarrones: 1.20€/kg. Asumiendo que la ración de pasta que se come al día es de 100g: **0.2€**/100g.
- Arroz: 1.80€/Kg. Asumiendo que la ración de arroz que se come al día es de 100g: **0.18€**/100g.
- Patatas simples: **9,50€** cada 10 Kg. Asumiendo que la ración de patata que se come al día es de 150g: **0.1425€**/150g.

Media: 0.174€ por ración de pasta, arroz y patatas.

PAN, CEREALES

- Pan: 0.75€/ud. Asumiendo que al día se come media barra: **0.375€**/ración.
- Corn Flakes: 1.50€/500g. Asumiendo que por ración se comen 40g de cereales: **0.12€**/ración.
- Cereales rellenos de cacao y avellanas: 2.25€/400g. Asumiendo que por ración se comen 40g de cereales: **0.225€**/ración.

Media: 0.24€ por ración individual de pan y cereales al día.

VERDURAS, ENSALADAS Y HORTALIZAS

- Ensalada de huerto: **1.99**€/unidad
- Tomate pera rama: 4.50€/kg. Estimamos un peso de 115g.**0.5175€**/ud.
- Calabaza: 4.97€/kg. Estimamos un peso de 400g. **1.988€**/ud
- Puré de verduras: 2.20€/l. Estimamos un consumo de 300g. **0.66€**/plato.

Media: 1.28€ por ración individual de verduras, ensaladas y hortalizas al día.

LEGUMBRES

- Garbanzo cocido: 1.30€/400g. Asumiendo que la ración de legumbres cocidas que se come al día es de 200g: **0.65€**/ración.
- Garbanzo categoría extra: 2.15€/Kg. Asumiendo que la ración de legumbres frescas que se come al día es de 80g: **0.172€**/ración.
- Alubia cocida: 1.30€/400g. Asumiendo que la ración de legumbres cocidas que se come al día es de 200g: **0.65€**/ración.
- Alubia blanca: 1.85€/ Kg. Asumiendo que la ración de legumbres cocidas que se come al día es de 80g: **0.148€**/ración.
- Lenteja cocida: 1.30€/400g. Asumiendo que la ración de legumbres cocidas que se come al día es de 200g: **0.65€**/ración.
- Lenteja pardina categoría extra: 3.79€/Kg. Asumiendo que la ración de legumbres frescas que se come al día es de 80g: **0.3032€**/ración.

Media: 0.428€ por ración individual de legumbres.

EMBUTIDOS Y FIAMBRES

- Jamón York/Cocido: 10,12€/kg
- Jamón Serrano: 20€/kg
- Lomo embuchado: 18,5€/kg
- Chorizo: 11€/kg
- Salchichón: 12€/kg

Media: 14,3€/kg. Al día unos 25g (menos de 200g/semana): 0,36€/día.

PRODUCTOS LÁCTEOS

Suponiendo tres raciones de lácteos al día,

- Queso: 8€/kg (teniendo en cuenta queso fresco, curados, etc). Ración: 65g=0,52€
- Leche: 1€/l. Ración: 250mL=0,25€
- Yogur: 2€/kg. Ración: 125g=0,25€

Sumando todo al suponer tres raciones y que sea una de cada, al día: 1,02€/día.

DULCES

DULCES	
€	MEDIA
0,10 €	1,26 €
0,53 €	
0,45 €	
4,50 €	
0,72 €	

- Galletas María dorada: 1,90 €/kg. Una galleta pesa alrededor de 6.5 g. Asumimos un consumo de 8 galletas. **0.099€**
- Galletas Dinosaurus. 8,88 €/kg. Tenemos en cuenta que en la caja vienen 6 paquetes de 3 galletas de 10g. Asumimos un consumo de 2 paquetes. **0.53€**
- Chocolate: 15,83 €/kg. Tenemos en cuenta una onza de 28.35g. **0.449€**
- Bollería (croissants, magdalenas, napolitanas): **4,5€/kg**
- Donut: 2.89€ el pack de 4. Asumimos un consumo de 1 donut. **0.723€**/ud.

Media: 1,26€/kg. Consumo medio al día de 60g (una ración de cereales son 30g y una de galletas otras 30), entonces precio al día: 0,33€/dia.

REFRESCOS CON AZÚCAR

- Cocacola Zero: 2,6€/l
- Fanta Naranja: 1,94€/l
- Nestea: 3,3€/l
- Refrescos marca blanca: 1€/l

Media: 2,21€/L. Suponiendo refresco de 33cl y que no es de bar (compra de latas en el Carrefour), sale 0,73€/día. En caso de que sea de bar, costaría de media 2,5€/día.

Haciendo la media entre las dos, un refresco sale por 1,62€/día.

APERITIVOS

APERITIVOS	
€	MEDIA
1,30 €	1,29 €
1,93 €	
1,07 €	
1,20 €	
0,96 €	

- Jumpers: 1.30€
- Patatas churrería: 1,93€
- Nachos marca blanca: 1.07 €
- Pipas girasol: 1.20€
- Aceitunas verdes rellenas de anchoa: 0.96€

Se estima que el día que se comen aperitivos, se comen bolsas individualizadas (que realmente se abren varias y se comparte entre todos los comensales). Media: 1.29€.

ZUMOS

ZUMO		
€/L	€/250mL	MEDIA
0,86 €	0,22 €	0,61 €
3,49	0,87 €	
3,15	0,79 €	
1,68	0,42 €	
2,95	0,74 €	

- Néctar de naranja: 0,86 €/l
- Zumo de naranja Carrefour botella 1 l: 3.49€/l
- Granini naranja : 3.09 €/l
- Marca blanca: 1.68 €/l
- Don simón naranja: 2.95 €/l

Media por ración (250 mL): 0.61€.

```r
# APARTADO 9
nuevo_df <- data.frame( # creamos una nueva tabla, nuevi_df con las columnas alimento y Coste_diario
  alimento = c("Fruta fresca (excluye zumos)", "Carne", "Huevos", "Pescado", "Pasta, arroz, patatas",
               "Pan, cereales", "Verduras, ensaladas y hortalizas", "Legumbres", "Embutidos y fiambres",
               "Productos lácteos", "Dulces", "Refrescos con azúcar", "Comida rápida",
               "Aperitivos o comidas saladas de picar", "Zumo natural de frutas o verduras"),
  Coste_diario = c(1.12, 2.67, 0.37, 3.89, 0.17, 0.24, 1.28, 0.43, 0.36, 1.02, 1.26, 1.62, 9.45, 1.29, 0.61)
)

# Calculamos la media de Media_semanal para cada alimento y nivel socioeconómico, tomando en cuenta el sexo
presupuesto_por_nivel <- df %>%
  group_by(nivel, alimento) %>%
  summarise(media_semanal_promedio = mean(Media_semanal, na.rm = TRUE)) %>%
  inner_join(nuevo_df, by = "alimento") %>%
  mutate(semanal_alimento = media_semanal_promedio * Coste_diario) %>%
  group_by(nivel) %>%
  summarise(presupuesto_semanal = sum(semanal_alimento))%>%
  arrange(desc(presupuesto_semanal))

# Imprimimos el presupuesto semanal por nivel socioeconómico
print(presupuesto_por_nivel)

ggplot(presupuesto_por_nivel, aes(x = presupuesto_semanal, y = reorder(nivel,presupuesto_semanal), fill = nivel)) +
  geom_bar(stat = "identity") +
  labs(title = "Presupuesto Semanal por Nivel Socioeconómico",
       x = "Gasto Semanal",
       y = "Nivel Socioeconómico") +
  theme_minimal() +
  theme(legend.position = "none")
```

Figura 29: Código para resolver el apartado 9

Primero, se crea una nueva tabla, nuevo_df, con las columnas de la tabla de muestreo de precios anterior.

```
> nuevo_df
                                alimento Coste_diario
1           Fruta fresca (excluye zumos)         1.12
2                                 Carne         2.67
3                                Huevos         0.37
4                               Pescado         3.89
5                  Pasta, arroz, patatas         0.17
6                          Pan, cereales         0.24
7       Verduras, ensaladas y hortalizas         1.28
8                              Legumbres         0.43
9                   Embutidos y fiambres         0.36
10                     Productos lácteos         1.02
11                                Dulces         1.26
12                  Refrescos con azúcar         1.62
13                          Comida rápida         9.45
14 Aperitivos o comidas saladas de picar         1.29
15     Zumo natural de frutas o verduras         0.61
```

Figura 30: Salida del nuevo dataframe creado con los valores de los precios de coste de consumo de un tipo de alimento concreto para un día.

Ahora, partiendo de la tabla df, se crea una nueva variable, presupuesto_por_nivel, que va a calcular la cantidad de dinero que se deja cada nivel socioeconómico a lo largo de una semana, por alimento. Para ello, se comienza agrupando por nivel y alimento, y se obtiene una media_semanal_promedio, que simboliza la frecuencia media con la que se consume cada alimento a la semana (`summarise(media_semanal_promedio = mean(Media_semanal, na.rm = TRUE))`).

Ahora, se va a aplicar la función `inner_join(nuevo_df, by = "alimento")` para unificar las dos tablas (es como el NATURAL JOIN en SQL) con las que se está operando (df y nuevo_df), tomando como referente para ordenar la tabla generada, “alimento”.

```
> presupuesto_por_nivel <- df %>%
+     group_by(nivel, alimento) %>%
+     summarise(media_semanal_promedio = mean(Media_semanal, na.rm = TRUE)) %>%
+     inner_join(nuevo_df, by = "alimento")
`summarise()` has grouped output by 'nivel'. You can override using the `.groups` argument.
> presupuesto_por_nivel
# A tibble: 105 × 4
# Groups:   nivel [7]
   nivel alimento                              media_semanal_promedio Coste_diario
   <chr> <chr>                                                 <dbl>        <dbl>
 1 I     Aperitivos o comidas saladas de picar                  1.06         1.29
 2 I     Carne                                                  3.86         2.67
 3 I     Comida rápida                                          0.886        9.45
 4 I     Dulces                                                 3.39         1.26
 5 I     Embutidos y fiambres                                   3.14         0.36
 6 I     Fruta fresca (excluye zumos)                           6.00         1.12
 7 I     Huevos                                                 2.44         0.37
 8 I     Legumbres                                              2.08         0.43
 9 I     Pan, cereales                                          6.20         0.24
10 I     Pasta, arroz, patatas                                  3.77         0.17
# ℹ 95 more rows
# ℹ Use `print(n = ...)` to see more rows
```

Figura 31*: Salida coste diario para cada alimento y el promedio semanal de consumo de cada alimento para cada nivel socioeconómico*

En este punto, se genera una nueva columna, semanal_alimento, haciendo uso de `mutate(semanal_alimento = media_semanal_promedio * Coste_diario)`, que, multiplica el precio de cada alimento por la frecuencia con la que es consumido cada semana.

```
> presupuesto_por_nivel <- df %>%
+     group_by(nivel, alimento) %>%
+     summarise(media_semanal_promedio = mean(Media_semanal, na.rm = TRUE)) %>%
+     inner_join(nuevo_df, by = "alimento") %>%
+     mutate(semanal_alimento = media_semanal_promedio * Coste_diario)
`summarise()` has grouped output by 'nivel'. You can override using the `.groups` argument.
> presupuesto_por_nivel
# A tibble: 105 × 5
# Groups:   nivel [7]
   nivel alimento                              media_semanal_promedio Coste_diario semanal_alimento
   <chr> <chr>                                                 <dbl>        <dbl>            <dbl>
 1 I     Aperitivos o comidas saladas de picar                  1.06         1.29             1.37
 2 I     Carne                                                  3.86         2.67            10.3
 3 I     Comida rápida                                          0.886        9.45             8.38
 4 I     Dulces                                                 3.39         1.26             4.27
 5 I     Embutidos y fiambres                                   3.14         0.36             1.13
 6 I     Fruta fresca (excluye zumos)                           6.00         1.12             6.72
 7 I     Huevos                                                 2.44         0.37             0.903
 8 I     Legumbres                                              2.08         0.43             0.896
 9 I     Pan, cereales                                          6.20         0.24             1.49
10 I     Pasta, arroz, patatas                                  3.77         0.17             0.641
# ℹ 95 more rows
# ℹ Use `print(n = ...)` to see more rows
```

Figura 32*: Salida coste semanal por cierto alimento para cada nivel socioeconómico*

Ahora, sólo queda agrupar por niveles (`group_by(nivel)`), realizar la suma de lo que se paga por alimento durante una semana (`summarise(presupuesto_semanal = sum(semanal_alimento))`) y ordenar de mayor a menor (`arrange(desc(presupuesto_semanal))`).

```
> presupuesto_por_nivel <- df %>%
+   group_by(nivel, alimento) %>%
+   summarise(media_semanal_promedio = mean(Media_semanal, na.rm = TRUE)) %>%
+   inner_join(nuevo_df, by = "alimento") %>%
+   mutate(semanal_alimento = media_semanal_promedio * Coste_diario) %>%
+   group_by(nivel) %>%
+   summarise(presupuesto_semanal = sum(semanal_alimento))%>%
+   arrange(desc(presupuesto_semanal))
`summarise()` has grouped output by 'nivel'. You can override using the `.groups` argument.
> presupuesto_por_nivel
# A tibble: 7 × 2
  nivel     presupuesto_semanal
  <chr>                   <dbl>
1 V                        63.3
2 IV                       63.2
3 III                      63.2
4 I                        63.1
5 VI                       63.1
6 II                       62.7
7 No consta                62.5
```

Figura 33*: Salida ordenada de mayor a menor gasto semanal por persona*

Parece que el nivel que más gasta en comida es el nivel V, con realmente poca diferencia con respecto a los niveles que le siguen (IV, III, I, VI). El nivel II es el único que gasta menos, pero se ha de comprobar que son diferencias significativas. Para ello, se podría recurrir a estudios más precisos, recurriendo, por ejemplo, al p-valor.

No consta no se tendrá en cuenta para hacer suposiciones, porque no aporta información constatada.

Finalmente, se grafican los datos.

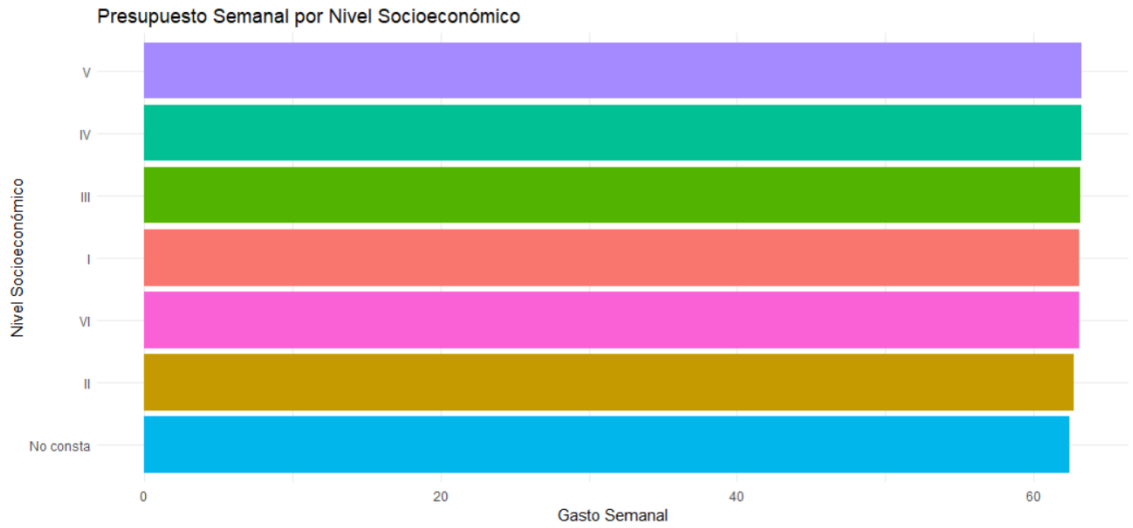


Figura 34: Gráfico de barras consumo por nivel socioeconómico

APARTADO 10

Haz un boxplot de la media semanal de consumo de los distintos alimentos tomando todos los datos calculados en el Apartado 4. Explica qué hace un boxplot e interpreta los resultados.

```
# APARTADO 10
ggplot(df, aes(x = Media_semanal, y = alimento, color = alimento)) +
  geom_boxplot() + # para graficar de manera automática los boxplots en función de los ejes x e y
  labs(title = "Consumo semanal de alimentos",
       x = "Alimento",
       y = "Media Semanal de Consumo")
```

Figura 35: Código para resolver el apartado 10

En el apartado 4 se creó la columna Media_semanal, que estima la frecuencia media con la que se consume cada alimento por nivel socioeconómico. Haciendo uso de la librería `ggplot`, esta vez, se recurre a `geom_boxplot()`, que graficará automáticamente los boxplots de las medias semanales según cada tipo de alimento, como puede observarse en la Figura 36.

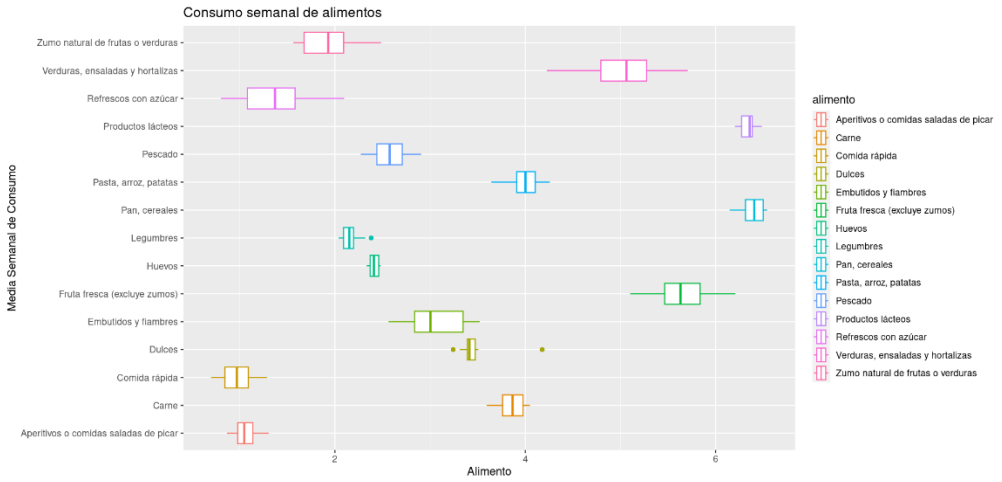


Figura 36: Diagrama de cajas

Un boxplot es un diagrama de cajas y bigotes representa la distribución de un conjunto de datos numéricos a través de sus cuartiles. Se interpreta de la siguiente manera:

- La línea central de la caja corresponde a la mediana, es decir, el valor que divide al conjunto de datos en dos partes iguales, con el 50% de los datos por debajo y el 50% por encima de este valor.
- La longitud de la caja, viene dada según su rango intercuartílico. Los bordes inferior y superior del cuadro corresponden al primer y tercer cuartil (Q1 es el valor que deja al 25% de los datos por debajo y Q3 es el valor que deja al 75% de los datos por debajo).
- Los bigotes a su vez, son los valores más extremos que se encuentran dentro de 1.5 veces el rango intercuartílico por encima del tercer cuartil y por debajo del primer cuartil. Los valores que se encuentran más allá de los bigotes se consideran valores atípicos y se representan como puntos o asteriscos individuales.

Así, con la ayuda del gráfico creado, se pueden comparar distribuciones entre los diferentes alimentos. Por ejemplo, si estudiamos los alimentos más consumidos con frecuencia, nos encontramos con que "Productos lácteos" y "Pan, cereales" son los únicos alimentos que sobrepasan las 6 veces de consumo a la semana. En el caso de productos lácteos, hay poca variabilidad, pues la caja graficada es pequeña y los bigotes cortos, es decir, hay mucha tendencia a ser un producto muy consumido. Para el caso de la caja más grande, "Verduras, ensaladas y hortalizas", la mediana supera la frecuencia de 5 veces, pero sobresalen bigotes que llegan casi a las 4 por la izquierda y a las 6 por la derecha, es decir, es un alimento que hay mucha diferencia de consumo (visto previamente que las mujeres lo consumen con mayor frecuencia y posible diferencia en caso de dietas veganas y vegetarias). Para el caso de "Legumbres" y "Dulces", se grafican unos puntos, indicativos de que se tratan de *outliers*, es decir, hay grupos que se salen de la distribución y los consumen con mucha mayor frecuencia que el resto de los participantes.

APARTADO 11

¿Qué otros estudios se te ocurren que podrían hacerse con los datos de esta práctica?

Este conjunto de datos parece interesante para estudios que relacionen el efecto del consumo de ciertos alimentos para la salud. Por ejemplo, si se pudiese obtener otro conjunto con datos que relacionen a estos grupos con enfermedades, como diabetes, obesidad, falta de vitaminas, etc, podríamos intentar buscar un modelo que relacionase estas enfermedades con el hábito de consumo de ciertos grupos.

Si bien a día de hoy se sabe que el consumo prologando y recurrente de alimentos ricos en azúcares puede provocar la aparición de enfermedades tan comunes como es la diabetes tipo II, que a su vez es la causa más común del siglo XXI de problemas cardiovasculares, podría estudiarse la probabilidad de padecerlas según el nivel socioeconómico.

Relacionado con ello, para próximos estudios semejantes podría preguntarse por el peso de los participantes del estudio, así como su edad o región. Podrían ser datos interesantes para estudiar si varía o no el consumo de alimentos según dichas características o para evaluar si el gasto es mayor en grandes ciudades que en el área rural. Del mismo modo, se podría pedir un precio estimado en rangos (25-50;50-75;75-100,...) para estudiar el gasto por zonas.

Otro posible enfoque es el estudio de dietas. ¿Qué nivel socioeconómico y qué sexo lleva una dieta *keto*? Este tipo de dietas se han vuelto muy famosas debido a su eficacia a la hora de bajar de peso. En una dieta keto se comen menos carbohidratos, se mantiene un consumo moderado de proteína y puede que aumente la ingesta de grasa. La reducción de carbohidratos, fuente a corto plazo de energía del organismo, pone al cuerpo en un estado metabólico llamado cetosis, en donde, debido a la ausencia de glucosa, la grasa se transforma en cuerpos cetónicos para obtener energía.

```r
# APARTADO 11
umbral_bajo_glucidos <- 7
umbral_alto_proteinas <- 14
keto <- df %>%
  group_by(sexo, nivel) %>%
  summarise(
    consumo_bajo_glucidos = sum(Media_semanal[alimento %in% c("Pasta, arroz, patatas", "Pan, cereales", "Dulces", "Refrescos con azúcar")]),
    consumo_alto_proteinas = sum(Media_semanal[alimento %in% c("Carne", "Pescado", "Huevos", "Embutidos y fiambres")]),
    .groups = 'drop'  # Establecer el argumento .groups para desagrupar
  ) %>%
  print(n=100) %>%
  filter(consumo_bajo_glucidos <= umbral_bajo_glucidos & consumo_alto_proteinas >= umbral_alto_proteinas) %>%
  arrange(desc(consumo_bajo_glucidos),consumo_alto_proteinas)
keto
```

Figura 37: Código propuesto para resolver un posible ejercicio de estudio de la dieta keto

```
# A tibble: 14 × 4
   sexo    nivel     consumo_bajo_glucidos consumo_alto_proteinas
   <chr>   <chr>                     <dbl>                  <dbl>
 1 HOMBRES I                          14.6                   12.5
 2 HOMBRES II                         14.9                   12.1
 3 HOMBRES III                        15.3                   12.5
 4 HOMBRES IV                         15.5                   12.3
 5 HOMBRES No consta                  16.2                   12.2
 6 HOMBRES V                          16.0                   12.3
 7 HOMBRES VI                         16.3                   12.0
 8 MUJERES I                          14.0                   12.0
 9 MUJERES II                         14.2                   11.6
10 MUJERES III                        14.7                   11.7
11 MUJERES IV                         15.1                   11.9
12 MUJERES No consta                  15.1                   11.1
13 MUJERES V                          15.3                   11.5
14 MUJERES VI                         15.7                   11.1
> keto
# A tibble: 0 × 4
# ℹ 4 variables: sexo <chr>, nivel <chr>, consumo_bajo_glucidos <dbl>, consumo_alto_proteinas <dbl>
```

Figura 38: Vista del consumo de glúcidos y proteínas de cada grupo

Dado que ningún grupo cumple estas condiciones, se podría concluir que no es posible asociar un perfil de los estudiados con esta dieta. Cosa que es lógica puesto que, como se ha visto antes, el consumo de pan y cereales encabezaba la lista tanto para hombres como para mujeres.

Otra idea de estudio podría ser encuestar a los participantes según cuántos días a la semana invierten al ocio o disfrute personal. Supongamos que la gente que toma muchos aperitivos pasa un mayor tiempo de calidad con familiares y amigos. Investiguemos qué grupo es el que invierte más en aperitivos.

```r
# APARTADO 11
ocio<-df %>%
  group_by(nivel,sexo) %>%
  filter(alimento=="Aperitivos o comidas saladas de picar") %>%
  filter(nivel != "No consta") %>%
  summarise(mediasuma = mean(Media_semanal)) %>%
  arrange(desc(mediasuma))
ocio
```

Figura 39: Código propuesto para encontrar aquellos participantes con mayor nivel de ocio y vida social

```
> ocio
# A tibble: 12 × 3
# Groups:   nivel [6]
   nivel sexo    mediasuma
   <chr> <chr>       <dbl>
 1 VI    HOMBRES     1.19
 2 III   HOMBRES     1.16
 3 I     HOMBRES     1.14
 4 II    HOMBRES     1.13
 5 V     HOMBRES     1.11
 6 VI    MUJERES     1.08
 7 IV    MUJERES     1.02
 8 IV    HOMBRES     0.989
 9 V     MUJERES     0.983
10 I     MUJERES     0.977
11 III   MUJERES     0.961
12 II    MUJERES     0.943
```

Figura 40: Vista de los grupos que podrían estar llevando mayor vida de ocio en relación con la cantidad de aperitivos que ingieren por semana

Podemos ver que los hombres lideran el ranking, si bien la diferencia entre los primeros y los últimos es baja, ambos rondan el consumo de un día a la semana de este tipo de aperitivos. En el caso de los hombres, siempre se cumple este día destinado (a excepción de los hombres de nivel IV), según nuestra interpretación, al ocio. En cambio, en el caso de las mujeres, sobre todo para niveles socioeconómicos altos, esto no siempre se cumple.

Finalmente, a raíz del apartado 10, a la vista de los *outliers* encontrados en "Legumbres" y "Dulces", se ha querido encontrar a qué nivel y sexo pertenecían dichos valores. Nos hemos centrado en el *outlier* por la derecha de "Dulces", dado que queda excesivamente fuera de lugar.

```
# APARTADO 11
dulces_data <- df %>% # Queremos que aparezcan sólo las filas relacionadas con el consumo de Dulces
  filter(alimento == "Dulces")

dulces_data <- dulces_data %>%
  group_by(nivel, sexo) %>%  # Agrupamos por nivel y sexo para promediar las diferentes
  summarise(Media_semanal_promedio = mean(Media_semanal)) # Media_semanal de cada grupo y sexo

dulces_outliers <- dulces_data %>% # Visto en el boxplot del ejercicio anterior que un outlier se encuentra
  filter(Media_semanal_promedio > 4.0)  # claramente a partir de una Media_semanal de 4, buscamos ese valor

dulces_outliers
```

Figura 41: Código propuesto para encontrar los outliers de Dulces

```
> dulces_outliers
# A tibble: 1 × 3
# Groups:   nivel [1]
  nivel     sexo    Media_semanal_promedio
  <chr>     <chr>                    <dbl>
1 No consta HOMBRES                   4.18
```

Figura 42: Vista de a qué grupo pertenece el outlier en cuestión

Resulta que pertenecía al nivel "No consta". Por tanto, se ha decidido realizar una mejora respecto al apartado 10, consistente en suprimir las instancias con nivel == "No consta", con el fin de analizar los datos de los que se tiene el 100% de la fiabilidad.

```
# MEJORA DEL APARTADO 10
df_sin_no_consta <- df%>% # Creamos nuevo dataframe, sin las instancias que pertenecen a "No_consta"
  filter(nivel != "No consta")

ggplot(df_sin_no_consta, aes(x = Media_semanal, y = alimento, color = alimento)) +
  geom_boxplot(data = df_sin_no_consta) +
  labs(title = "Consumo semanal de alimentos",
       x = "Media Semanal de Consumo",
       y = "Alimento")
```

Figura 43: Código propuesto para mejorar el apartado 10, eliminando las instancias "No_consta"

Obteniéndose el gráfico de la Figura 44. Desaparecen los *outliers* de "Legumbres" y "Dulces" de la Figura 36. Se generan unos pequeños *outliers* en "Dulces", pero no se considerarán como tal, puesto que se encuentran prácticamente dentro de los límites del alimento.

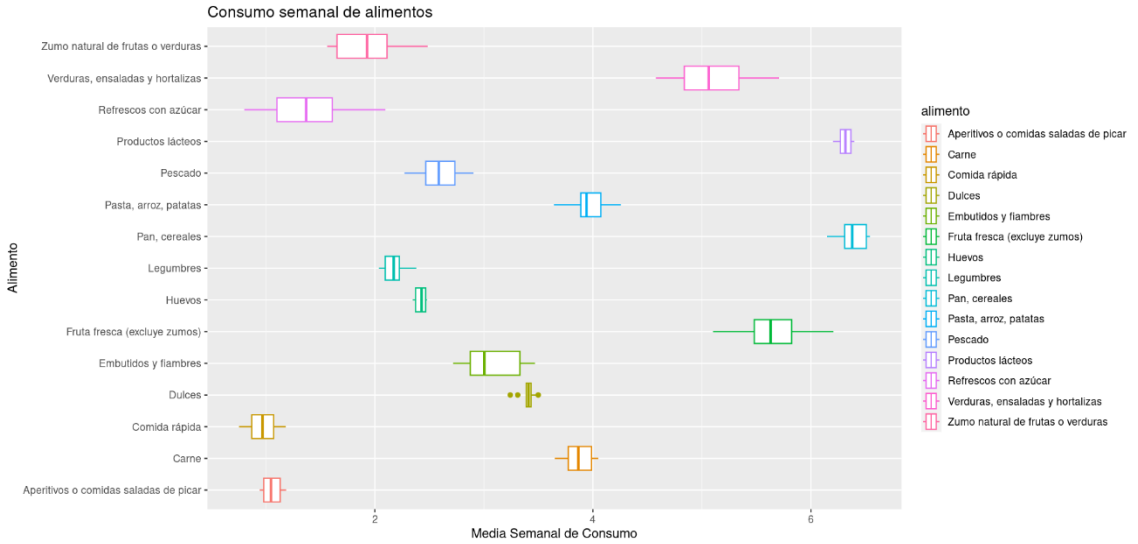


Figura 43: Diagrama de cajas sin tener en cuenta "No_consta"

CONCLUSIONES

Esta práctica demuestra que, con unas pocas variables, tan sencillas como nivel socioeconómico, sexo y cantidad de veces que se consume un alimento por semana, se puede extraer gran cantidad de conclusiones e información acerca de qué nivel gasta más dinero en comida, qué porción de la población cuida más lo que se lleva a la boca, o qué grupo podría incluso tener más vida social.