PEC1: Estadística Descriptiva

Marco Russo

October, 2025

Contents

1 Información del Estudiante

Nombre	Marco Russo
Email	mrussorb@uoc.edu
GitHub	https://github.com/marcusRB/uoc-ub-scientific-programming
LinkedIn	https://www.linkedin.com/in/marcusrb/
Fecha	October 28, 2025

2 Sección 1. Importación, exportación y gestión de datos (2 puntos)

2.1 (1 punto) Ejercicio 1

Se pide de descargar el dataset desde Kaggle disponible a la ruta indicada:

```
# Descargamos el dataset
# https://www.kaggle.com/datasets/amitvkulkarni/hair-health
```

Para ello utilizaremos el paquete `RKaggle` para descargar el dataset directamente desde R Studio

Preparamos el entorno cargando el resto de librerías que serán útiles para realizar un análisis exploratorio de los datos.

2.1.1 1.1

Guardamos el dataset en un formato dataframe y comprobaremos

```
# Guardamos el dataset en formato dataframe
HairFallH <- get_dataset('amitvkulkarni/hair-health')
```

```
# Comprobamos si es un dataframe
is.data.frame(HairFallH)
```

```
## [1] TRUE
```

Finalmente mostramos los primeros datos y la naturaleza de las características.

```
# Mostramos los primeros datos con head()
head(HairFallH, 10)
```

```
## # A tibble: 10 x 13
##        Id Genetics `Hormonal Changes` `Medical Conditions`
##     <dbl> <chr>    <chr>              <chr>
##  1 133992 Yes      No                 No Data
##  2 148393 No       No                 Eczema
##  3 155074 No       No                 Dermatosis
##  4 118261 Yes      Yes                Ringworm
##  5 111915 No       No                 Psoriasis
##  6 139661 Yes      No                 Psoriasis
##  7 169255 Yes      Yes                No Data
##  8 112032 Yes      No                 Dermatosis
##  9 140785 Yes      No                 Eczema
## 10 187999 No       Yes                Ringworm
## # i 9 more variables: `Medications & Treatments` <chr>,
## #   `Nutritional Deficiencies` <chr>, Stress <chr>, Age <dbl>,
## #   `Poor Hair Care Habits` <chr>, `Environmental Factors` <chr>,
## #   Smoking <chr>, `Weight Loss` <chr>, `Hair Loss` <dbl>
```

```
# Mostramos los últimos datos también con tail()
tail(HairFallH, 10)
```

```
## # A tibble: 10 x 13
##        Id Genetics `Hormonal Changes` `Medical Conditions`
##     <dbl> <chr>    <chr>              <chr>
##  1 144786 No       No                 Ringworm
```

```
##  2 127532 Yes      No                Alopecia Areata
##  3 131739 No       Yes               Thyroid Problems
##  4 181854 Yes      Yes               Dermatosis
##  5 196218 No       Yes               Scalp Infection
##  6 184367 Yes      No                Seborrheic Dermatitis
##  7 164777 Yes      Yes               No Data
##  8 143273 No       Yes               Androgenetic Alopecia
##  9 169123 No       Yes               Dermatitis
## 10 127183 Yes      Yes               Psoriasis
## # i 9 more variables: `Medications & Treatments` <chr>,
## #   `Nutritional Deficiencies` <chr>, Stress <chr>, Age <dbl>,
## #   `Poor Hair Care Habits` <chr>, `Environmental Factors` <chr>,
## #   Smoking <chr>, `Weight Loss` <chr>, `Hair Loss` <dbl>
```

2.1.2 1.2

Realizaremos un exploratorio genérico del dataset. Mostrando información básica del dataset, para pasar luego a los estadísticos básico y comenzaremos a interactuar con las características luego.

Verificamos la estructura del juego de datos principal. Vemos el número de columnas que tenemos y ejemplos de los contenidos de las filas.

```
str(HairFallH)
```

```
## spc_tbl_ [999 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Id                      : num [1:999] 133992 148393 155074 118261 111915 ...
##  $ Genetics                : chr [1:999] "Yes" "No" "No" "Yes" ...
##  $ Hormonal Changes        : chr [1:999] "No" "No" "No" "Yes" ...
##  $ Medical Conditions      : chr [1:999] "No Data" "Eczema" "Dermatosis" "Ringworm" ...
##  $ Medications & Treatments: chr [1:999] "No Data" "Antibiotics" "Antifungal Cream" "Antibiotics" ...
##  $ Nutritional Deficiencies: chr [1:999] "Magnesium deficiency" "Magnesium deficiency" "Protein defic
##  $ Stress                  : chr [1:999] "Moderate" "High" "Moderate" "Moderate" ...
##  $ Age                     : num [1:999] 19 43 26 46 30 37 40 35 19 49 ...
##  $ Poor Hair Care Habits   : chr [1:999] "Yes" "Yes" "Yes" "Yes" ...
##  $ Environmental Factors   : chr [1:999] "Yes" "Yes" "Yes" "Yes" ...
##  $ Smoking                 : chr [1:999] "No" "No" "No" "No" ...
##  $ Weight Loss             : chr [1:999] "No" "No" "Yes" "No" ...
##  $ Hair Loss               : num [1:999] 0 0 0 0 1 1 1 0 1 0 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Id = col_double(),
##   ..   Genetics = col_character(),
##   ..   `Hormonal Changes` = col_character(),
##   ..   `Medical Conditions` = col_character(),
##   ..   `Medications & Treatments` = col_character(),
##   ..   `Nutritional Deficiencies` = col_character(),
##   ..   Stress = col_character(),
##   ..   Age = col_double(),
##   ..   `Poor Hair Care Habits` = col_character(),
##   ..   `Environmental Factors` = col_character(),
##   ..   Smoking = col_character(),
##   ..   `Weight Loss` = col_character(),
##   ..   `Hair Loss` = col_double()
##   .. )
```

```
##  - attr(*, "problems")=<externalptr>
```

```
# Observamos su composición
dim(HairFallH)
```

```
## [1] 999  13
```

Vemos que tenemos **13** características o variables y **999** registros. De las características observamos que solamente existen 3 variables numéricas del tipo double, y el resto, 10 variables del tipo objeto o character.

A simple vista las que podrían ser variables categoróricas, en realidad son características binarias (TRUE o FALSE), como podríamos interpolar por 1,0 si deseamos.

Podemos revisar la descripción de las variables contenidas en el fichero y si los tipos de variables se corresponden con las que hemos cargado. Las organizamos lógicamente para darles sentido y construimos un pequeño diccionario de datos utilizando la documentación auxiliar.

```
# Extraemos los nombres de las variables
names(HairFallH)
```

```
##  [1] "Id"                       "Genetics"
##  [3] "Hormonal Changes"         "Medical Conditions"
##  [5] "Medications & Treatments" "Nutritional Deficiencies"
##  [7] "Stress"                   "Age"
##  [9] "Poor Hair Care Habits"    "Environmental Factors"
## [11] "Smoking"                  "Weight Loss"
## [13] "Hair Loss"
```

- **Genetics** : Indicates whether the individual has a family history of baldness (Yes/No).
- **Hormonal Changes**: Indicates whether the individual has experienced hormonal changes (Yes/No).
- **Medical Conditions**: Lists specific medical conditions that may contribute to baldness, such as Alopecia Areata, Thyroid Problems, Scalp Infection, Psoriasis, Dermatitis, etc.
- **Medications & Treatments**: Lists medications and treatments that may lead to hair loss, such as Chemotherapy, Heart Medication, Antidepressants, Steroids, etc.
- **Nutritional Deficiencies**: Lists nutritional deficiencies that may contribute to hair loss, such as Iron deficiency, Vitamin D deficiency, Biotin deficiency, Omega-3 fatty acid deficiency, etc.
- **Stress**: Indicates the stress level of the individual (Low/Moderate/High).
- **Age**: Represents the age of the individual.
- **Poor Hair Care Habits**: Indicates whether the individual practices poor hair care habits (Yes/No).
- **Environmental Factors**: Indicates whether the individual is exposed to environmental factors that may contribute to hair loss (Yes/No).
- **Smoking**: Indicates whether the individual smokes (Yes/No).
- **Weight Loss**: Indicates whether the individual has experienced significant weight loss (Yes/No).
- **Baldness (Target)**: Binary variable indicating the presence (1) or absence (0) of baldness in the individual.

Esta última variable, la variable predictora siendo binaria [0,1], nos indica que está renombrada como `Hair Loss`.

Realizamos además un pequeño resumen estadístico de cada una de las variables

```
# Con esta función observamos la distribución de los datos a grande rasgos.
summary(HairFallH)
```

```
##        Id           Genetics         Hormonal Changes   Medical Conditions
##  Min.   :110003   Length:999         Length:999         Length:999
##  1st Qu.:131868   Class :character   Class :character   Class :character
##  Median :152951   Mode  :character   Mode  :character   Mode  :character
##  Mean   :153355
##  3rd Qu.:174969
```

```
##  Max.   :199949
##  Medications & Treatments Nutritional Deficiencies    Stress
##  Length:999               Length:999               Length:999
##  Class :character         Class :character         Class :character
##  Mode  :character         Mode  :character         Mode  :character
##
##
##
##       Age        Poor Hair Care Habits Environmental Factors   Smoking
##  Min.   :18.00   Length:999            Length:999             Length:999
##  1st Qu.:26.00   Class :character      Class :character       Class :character
##  Median :34.00   Mode  :character      Mode  :character       Mode  :character
##  Mean   :34.19
##  3rd Qu.:42.00
##  Max.   :50.00
##  Weight Loss          Hair Loss
##  Length:999         Min.   :0.0000
##  Class :character   1st Qu.:0.0000
##  Mode  :character   Median :0.0000
##                     Mean   :0.4975
##                     3rd Qu.:1.0000
##                     Max.   :1.0000
```

El siguiente paso será observar si necesitaremos tratar los datos con la tareas de limpieza o imputaciones de los valores, por ejemplo, nulos o vacíos

```
print('NA')
```

```
## [1] "NA"
```

```
colSums(is.na(HairFallH))
```

```
##                      Id                 Genetics         Hormonal Changes
##                       0                        0                        0
##       Medical Conditions Medications & Treatments Nutritional Deficiencies
##                       0                        0                        0
##                  Stress                      Age    Poor Hair Care Habits
##                       0                        0                        0
##    Environmental Factors                  Smoking              Weight Loss
##                       0                        0                        0
##               Hair Loss
##                       0
```

```
print('Blancos')
```

```
## [1] "Blancos"
```

```
print('Empty')
```

```
## [1] "Empty"
```

```
print('NAN')
```

```
## [1] "NAN"
```

```
colSums(HairFallH=="")
```

```
##                      Id                 Genetics         Hormonal Changes
##                       0                        0                        0
```

```
##       Medical Conditions Medications & Treatments Nutritional Deficiencies
##                        0                        0                        0
##                   Stress                      Age    Poor Hair Care Habits
##                        0                        0                        0
##    Environmental Factors                  Smoking              Weight Loss
##                        0                        0                        0
##                Hair Loss
##                        0
```

```
# Verificamos más en detalle los valores na
sum(is.na(HairFallH))
```

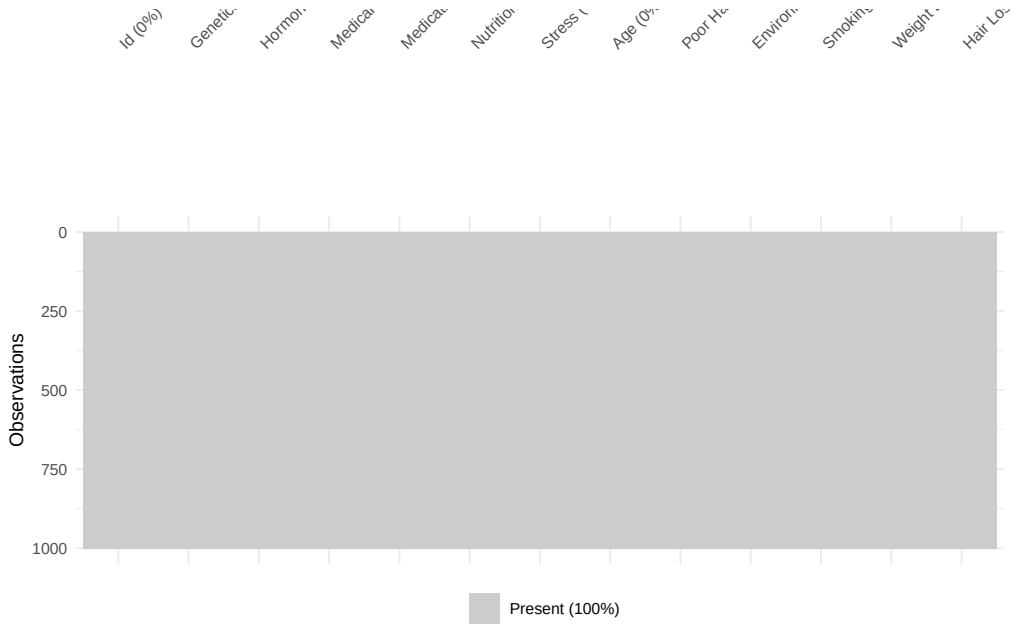
```
## [1] 0
```

```
colSums(is.na(HairFallH))
```

```
##                       Id                 Genetics         Hormonal Changes
##                        0                        0                        0
##       Medical Conditions Medications & Treatments Nutritional Deficiencies
##                        0                        0                        0
##                   Stress                      Age    Poor Hair Care Habits
##                        0                        0                        0
##    Environmental Factors                  Smoking              Weight Loss
##                        0                        0                        0
##                Hair Loss
##                        0
```

También de una manera más gráfica, podemos observar si existen valores nulos.

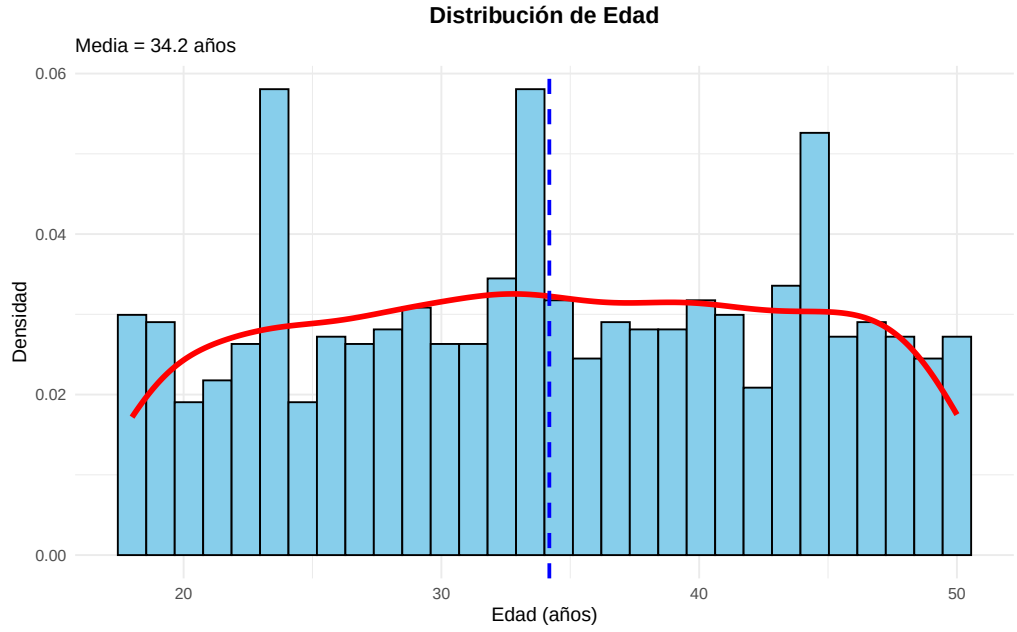
```
# Visualizamos con el comando vis_miss del paquete visdat
vis_miss(HairFallH, )
```



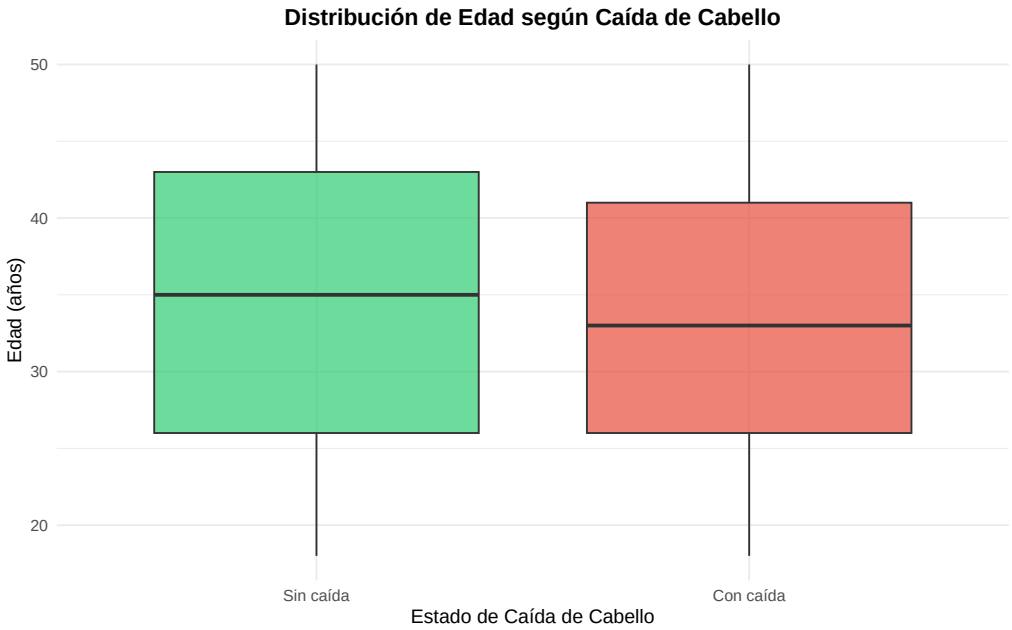
Seguimos sin apreciar los datos nulos o vacíos. Por lo general nos tocará realizar un exploratorio más exhaustivo para observar la distribución de los datos, y, en su caso ver si los datos de las variables categóricos, estén ocultos.

Vamos a crear histogramas y describir los valores para ver los datos en general de estos atributos para hacer una primera aproximación a los datos:

```r
# Histograma de edad con densidad
ggplot(HairFallH, aes(x = Age)) +
  geom_histogram(aes(y = after_stat(density)), bins = 30,
                 fill = "skyblue", color = "black") +
  geom_density(color = "red", size = 1.5) +
  geom_vline(xintercept = mean(HairFallH$Age),
             color = "blue", linetype = "dashed", size = 1) +
  labs(
    title = "Distribución de Edad",
    subtitle = paste("Media =", round(mean(HairFallH$Age), 1), "años"),
    x = "Edad (años)",
    y = "Densidad"
  ) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))
```




```r
# Box plot: Edad por Hair Loss
ggplot(HairFallH, aes(x = factor(`Hair Loss`, labels = c("Sin caída", "Con caída")),
                      y = Age, fill = factor(`Hair Loss`))) +
  geom_boxplot(alpha = 0.7) +
  scale_fill_manual(values = c("#2ecc71", "#e74c3c")) +
  labs(
    title = "Distribución de Edad según Caída de Cabello",
    x = "Estado de Caída de Cabello",
    y = "Edad (años)"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    legend.position = "none"
  )
```



Mostramos la distrubicón mostrando la mediana, y la distribución de frecuencia de edad. Del otro gráfico se muestra la distribución de edad por caída y si caída. Vemos que está bien equilibrado, por lo que entendemos que la muestra está homogénea y no parecer haya sesgo.

2.2 (1 punto) Ejercicio 2. Exploración del entorno RStudio y paquetes asociados.

Comprobamos que tenemos el paquete `datasets` descargado y obtenemos el conjunto BOD

```
if (!require('datasets')) install.packages('datasets')
library('datasets')
```

Obtenemos el conjunto BOD y listamos el dataset

```
bod <- datasets::BOD
```

```
summary(bod)
```

```
##       Time           demand
##  Min.   :1.000   Min.   : 8.30
##  1st Qu.:2.250   1st Qu.:11.62
##  Median :3.500   Median :15.80
##  Mean   :3.667   Mean   :14.83
##  3rd Qu.:4.750   3rd Qu.:18.25
##  Max.   :7.000   Max.   :19.80
```

```
str(bod)
```

```
## 'data.frame':    6 obs. of  2 variables:
##  $ Time  : num  1 2 3 4 5 7
##  $ demand: num  8.3 10.3 19 16 15.6 19.8
##  - attr(*, "reference")= chr "A1.4, p. 270"
```

Observamos que tenemos solo *2 variable* del tipo númerico s y *6 registros*.

```
# Veámos su composición
BOD
```

```
##   Time demand
## 1    1    8.3
## 2    2   10.3
## 3    3   19.0
## 4    4   16.0
## 5    5   15.6
## 6    7   19.8
```

```
# Buscamndo la información oficial
help(BOD)
```

Efectivamente desde el dataset `BOD` oficial, segúna la descripción de este dataset incluido con el paquete en R `datasets::BOD` nos indica:

> The BOD data frame has 6 rows and 2 columns giving the biochemical oxygen demand versus time in an evaluation of water quality.

Las dos variables nos dan información de:

> Time A numeric vector giving the time of the measurement (days). demand A numeric vector giving the biochemical oxygen demand (mg/l).

Seguramente es un dataset resultado de una prueba de concepto que está disponible y referenciado:

Source Bates, D.M. and Watts, D.G. (1988), *Nonlinear Regression Analysis and Its Applications*, Wiley, Appendix A1.4.

Originally from Marske (1967), *Biochemical Oxygen Demand Data Interpretation Using Sum of Squares Surface M.Sc. Thesis*, University of Wisconsin – Madison.

2.2.1 2.2

Ahora podemos guardar el dataframe resultante a un formato csv a nuestra carpeta `working directory` de nuestro ordenador.

```
# Podemos guardar este dataset directamente a nuestro entorno con el comando write_csv
write_csv(bod, "bod.csv")
```

Podemos comprobar como el fichero ha sido guardado correctamente

```
WD <- getwd()
list.files(WD)
```

```
## [1] "bod.csv"                              "PEC1_estadistica_descriptiva_files"
## [3] "PEC1_estadistica_descriptiva.pdf"     "PEC1_estadistica_descriptiva.Rmd"
```

---

3 Sección 2: Análisis básico del conjunto de datos (3,5 puntos)

3.1 Ejercicio 3

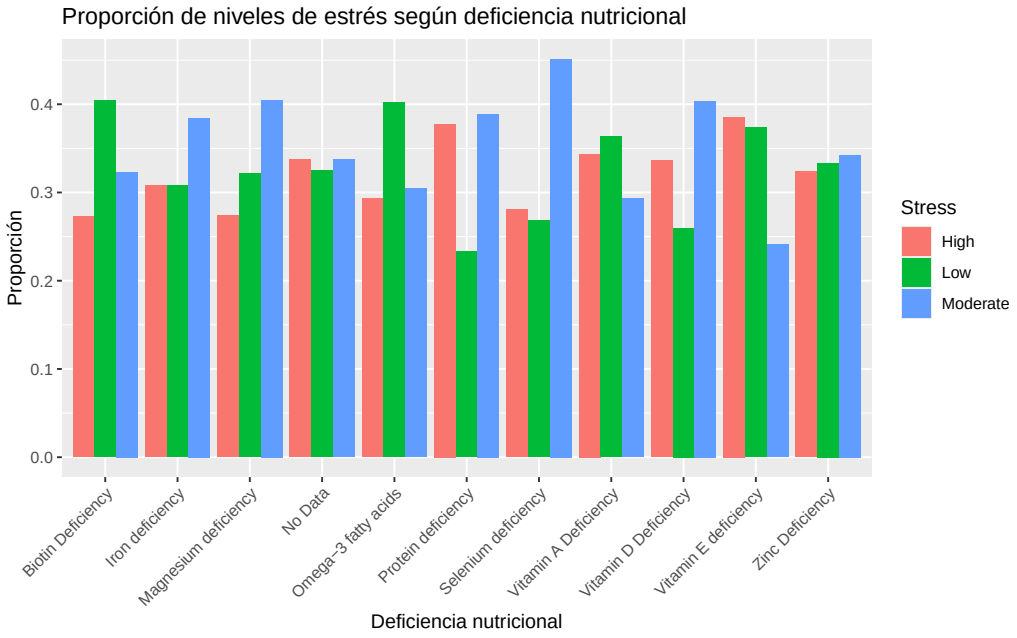
Realizaremos un análisis exploratorio más detallado y comentaremos cada uno de los insights de las preguntas del enunciado.

3.1.1 3.1

Calculad la proporción de personas según el nivel de estrés, agrupadas por el tipo de deficiencia nutricional. Comentad los resultados.

Utilizaremos ggplot para poder mostrar los datos en barras verticales por cada una de las 3 categorías de la variable `Stress` indicando por colores rojos HIGH, verde LOW, azule MODERATE

```
# Mostramos los resultados de esta proporción
stress_prop <- HairFallH %>%
  dplyr::group_by(`Nutritional Deficiencies`, Stress) %>%
  dplyr::summarise(count = n(), .groups = "drop") %>%
  dplyr::group_by(`Nutritional Deficiencies`) %>%
  dplyr::mutate(prop = count / sum(count))

ggplot(stress_prop, aes(x = `Nutritional Deficiencies`, y = prop, fill = Stress)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Proporción de niveles de estrés según deficiencia nutricional",
       x = "Deficiencia nutricional", y = "Proporción") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



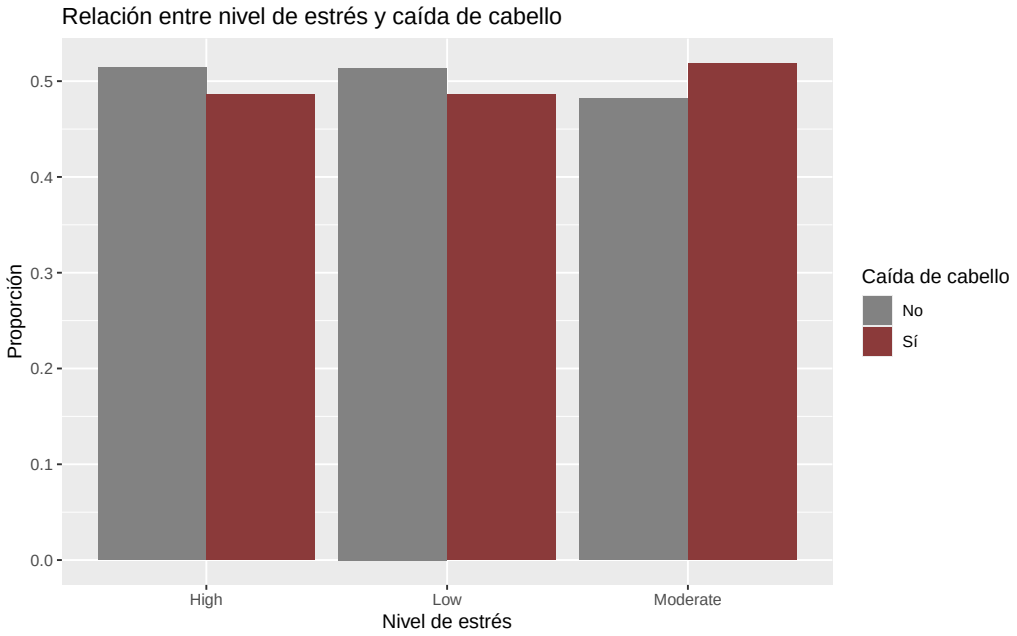
Podemos apreciar como los valores de `Selenium deficiency` se encuentra en una proporción del 40% más altarespecto a la `Vitamin D Deficiency`. Por lo que la correlación entre estas dos variables pueda tener su importancia. Los valores de stress HIGH mayor se encuentra con `Protein deficiency` y `Vitamin E deficiency` en un casi 35-38%. El valor bajo LOW de stress está por `Biotin` y `Omega 3`. Sería oportuno realizar unos cálculos más exhaustivos para poder correlacionar las dos variables y poder extraer unas conclusiones más detalladas.

3.1.2 3.2

Mostrad la relación entre el estrés y la caída de cabello . Mostrad un gráfico que compare dichas proporciones

Utilizamos siempre el paquete `ggplot2` para poder graficar y visualizar la relación que hay entre estrés y la caída de cabello.

```
stress_hairloss <- HairFallH %>%
  dplyr::group_by(Stress, `Hair Loss` = `Hair Loss`) %>%
  dplyr::summarise(count = n(), .groups = "drop") %>%
  dplyr::group_by(Stress) %>%
  dplyr::mutate(prop = count / sum(count))

ggplot(stress_hairloss, aes(x = Stress, y = prop, fill = factor(`Hair Loss`))) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Relación entre nivel de estrés y caída de cabello",
       x = "Nivel de estrés", y = "Proporción",
       fill = "Caída de cabello") +
  scale_fill_manual(values = c("#828282", "#8B3A3A"), labels = c("No", "Sí"))
```



Aunque los tres niveles de estrés puedan tener la misma proporción de pérdida de cabello, un nivel moderado tiene una mayor proporción respecto al resto. Asimismo la caída de cabello no influye por un nivel estrés alto ni bajo.

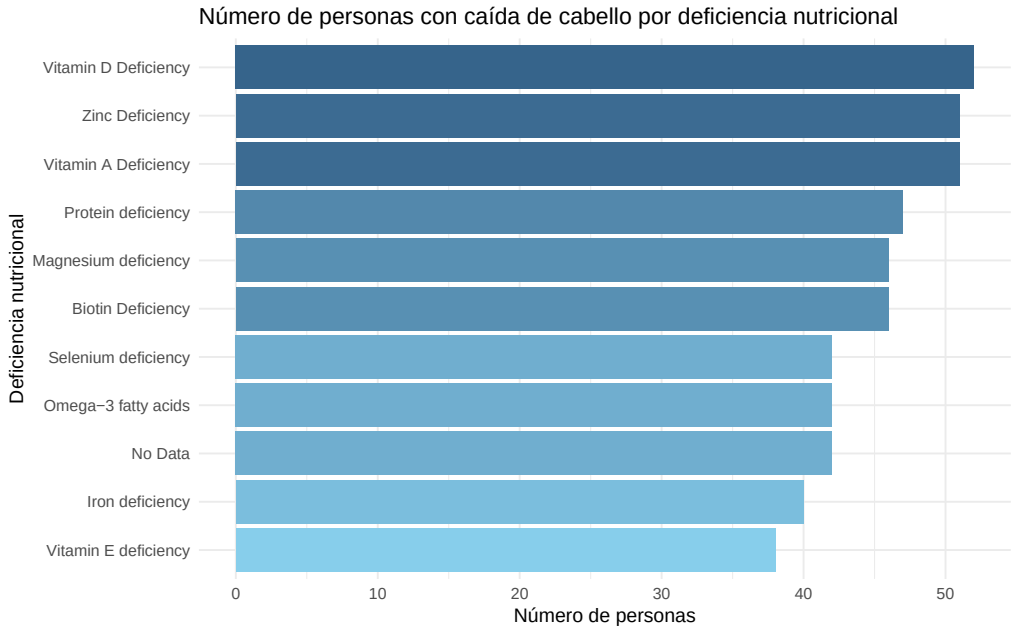
Al igual que ejercicio anterior, son conclusiones a simple vista de un gráfico. Con unos cálculos estadístico de correlación, quizás podríamos ver más detalles que con la gráfica podría ocultar.

3.1.3 3.3

Evaluad la caída de cabello con deficiencias de nutrición específicas, contando el número de personas por cada tipo de deficiencia nutricional.

```
nutri_hairloss <- HairFallH %>%
  dplyr::filter(`Hair Loss` == 1) %>%
  dplyr::count(`Nutritional Deficiencies`, sort = TRUE)
```

```
ggplot(nutri_hairloss, aes(x = reorder(`Nutritional Deficiencies`, n), y = n, fill = n)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  scale_fill_gradient(low = "#87CEEB", high = "#36648B") +
  labs(title = "Número de personas con caída de cabello por deficiencia nutricional",
       x = "Deficiencia nutricional", y = "Número de personas") +
  theme_minimal() +
  theme(legend.position = "none")
```



Este gráfico de barras horizontales nos muestra el número de personas que por deficiencia nutricional sí tienen caída de pelo, en valores absolutos. Podemos apreciar que en orden tenemos la carencia de `Vitamina D`, `Zinc`, `Vitamin A Protein`, `Magnesium`, `Biotin`, valores por encima de los 45. El resto se encontrarían por debajo de este valor.

3.1.4 3.4

Mostrad en una tabla el número de personas con caída de cabello, agrupadas por nivel de estrés, que sean fumadoras y cuya edad sea inferior a la media de la distribución

Podemos realizar la tarea por parte. Primero analizamos la media de la distribución

```
# Calcular media distribución
mean_age <- mean(HairFallH$Age, na.rm = TRUE)
mean_age
```

```
## [1] 34.18819
```

Entonces tenemos que realizar el filtro de aquellas personas solo con caída de cabello, agrupamos por estrés YES, fumadoras YES y la edad es menor que 34.19

```
tabla <- HairFallH %>%
  dplyr::filter(`Hair Loss` == 1,
         Smoking == "Yes",
         Age < mean_age) %>%
```

```
  dplyr::group_by(Stress) %>%
  dplyr::summarise(count = n(), .groups = "drop")


# La tabla mostraría este resultado.
tabla
```

```
## # A tibble: 3 x 2
##   Stress   count
##   <chr>    <int>
## 1 High        37
## 2 Low         38
## 3 Moderate    56
```

Teniendo en cuenta la media de 34, aquellas personas que se encuentran por debajo del promedio, con un nivel de estrés alto tenemos 37 personas; nivel bajo 38 y un nivel moderado 56 personas.

3.1.5 3.5

Mostrad las personas con edad mínima que presentan caída de cabello, fuman, tienen el nivel de estrés alto y tienen algún tratamiento médico asociado. ¿Qué deficiencia nutricional presentan?

Primero calcularemos la edad mínima, y aplicaríamos los filtros.

```
min_age <- min(HairFallH$Age[HairFallH$`Hair Loss` == 1], na.rm = TRUE)

tabla_filtrada <- HairFallH %>%
  filter(Age == min_age,
         `Hair Loss` == 1,
         Smoking == "Yes",
         Stress == "High",
         `Medications & Treatments` != "No Data") %>%
  select(Id, Age, `Nutritional Deficiencies`, `Medications & Treatments`)


tabla_filtrada
```

```
## # A tibble: 3 x 4
##       Id   Age `Nutritional Deficiencies` `Medications & Treatments`
##    <dbl> <dbl> <chr>                      <chr>
## 1 167156    18 Zinc Deficiency            Steroids
## 2 125449    18 Vitamin A Deficiency       Antidepressants
## 3 165112    18 Omega-3 fatty acids        Heart Medication
```

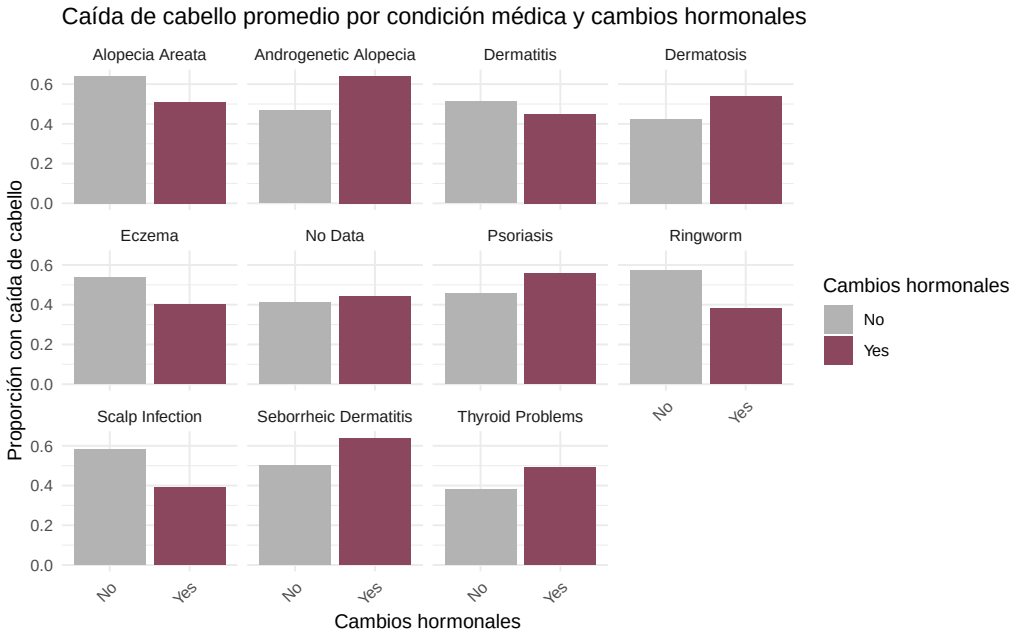
Mostramos la tabla con la edad mínima de 18 años, en particular 3 individuos con las respectivas deficiencias nutricional y que están siendo tratados.

3.1.6 3.6

En base a las cuestiones realizadas anteriormente, añadid una nueva cuestión que incluya una representación gráfica.

Aplicamos una nueva cuestión de caída de cabello por condición médica y cambios hormonales. Queremos ver si hay ciertas correlaciones según estos dos factores.

```r
combo_hairloss <- HairFallH %>%
  dplyr::group_by(`Medical Conditions`, `Hormonal Changes`) %>%
  dplyr::summarise(Mean_Hair_Loss = mean(`Hair Loss`), .groups = "drop")

ggplot(combo_hairloss, aes(x = `Hormonal Changes`, y = Mean_Hair_Loss, fill = `Hormonal Changes`)) +
  geom_bar(stat = "identity", position = "dodge") +
  facet_wrap(~`Medical Conditions`) +
  scale_fill_manual(values = c("Yes" = "#8B475D", "No" = "grey70")) +
  labs(title = "Caída de cabello promedio por condición médica y cambios hormonales",
       x = "Cambios hormonales", y = "Proporción con caída de cabello",
       fill = "Cambios hormonales") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



A simple vista, un cambio hormonal NO afecta en los casos de “Alopecia Areata”, “Dermatitis”, “Eczema”, “Ringworm”, “Scalp Infection”. Sin embargo, sí vemos que un cambio hormonal afecta a los síntomas de “Androgenetic Alopecia”, “Dermatosis” “Seborrheic Dermatisis” y “Psioriasis” más que al resto. Se podría estudiar más el caso y cada síntoma para poder ver su causa-efecto para poder dar una respuesta correcta a los gráficos actuales.

---

4 Sección 3: Estadística descriptiva y gráficos (4,5 puntos)

El objetivo de este apartado trabajar los conceptos de estadística descriptiva y representación de gráficos. Para ello, seguiremos trabajando con el conjunto de datos de la sección anterior.

4.1 (4,5 puntos) Ejercicio 4

4.1.1 4.1. (0,5 puntos)

A partir del conjunto de datos citado, mostrad los estadísticos descriptivos más relevantes y comentad los resultados, teniendo en cuenta el tipo de variables del conjunto de datos.

```
# Estadísticos descriptivos para variables numéricas
num_summary <- HairFallH %>%
  select(`Age`,`Genetics`, `Hormonal Changes`, `Medical Conditions`,
         `Medications & Treatments`, `Nutritional Deficiencies`,
         Stress, `Poor Hair Care Habits`, `Environmental Factors`,
         Smoking, `Weight Loss`) %>%
  psych::describe()

num_summary
```

```
##                            vars   n  mean   sd median trimmed   mad min max
## Age                           1 999 34.19 9.38     34   34.25 11.86  18  50
## Genetics*                     2 999  1.52 0.50      2    1.53  0.00   1   2
## Hormonal Changes*             3 999  1.51 0.50      2    1.51  0.00   1   2
## Medical Conditions*           4 999  5.87 3.24      6    5.84  4.45   1  11
## Medications & Treatments*     5 999  5.79 3.24      6    5.74  4.45   1  11
## Nutritional Deficiencies*     6 999  6.15 3.21      6    6.19  4.45   1  11
## Stress*                       7 999  2.03 0.82      2    2.04  1.48   1   3
## Poor Hair Care Habits*        8 999  1.49 0.50      1    1.49  0.00   1   2
## Environmental Factors*        9 999  1.51 0.50      2    1.51  0.00   1   2
## Smoking*                     10 999  1.52 0.50      2    1.52  0.00   1   2
## Weight Loss*                 11 999  1.47 0.50      1    1.47  0.00   1   2
##                            range  skew kurtosis   se
## Age                           32 -0.03    -1.15 0.30
## Genetics*                      1 -0.09    -1.99 0.02
## Hormonal Changes*              1 -0.04    -2.00 0.02
## Medical Conditions*           10  0.05    -1.24 0.10
## Medications & Treatments*     10  0.16    -1.20 0.10
## Nutritional Deficiencies*     10 -0.08    -1.23 0.10
## Stress*                        2 -0.06    -1.51 0.03
## Poor Hair Care Habits*         1  0.03    -2.00 0.02
## Environmental Factors*         1 -0.03    -2.00 0.02
## Smoking*                       1 -0.08    -2.00 0.02
## Weight Loss*                   1  0.11    -1.99 0.02
```

```
# Frecuencias para variables categóricas
cat_summary <- HairFallH %>%
  select(`Genetics`, `Hormonal Changes`, `Medical Conditions`,
         `Medications & Treatments`, `Nutritional Deficiencies`,
         Stress, `Poor Hair Care Habits`, `Environmental Factors`,
         Smoking, `Weight Loss`) %>%
  dplyr::summarise(across(everything(), ~length(unique(.x))))

cat_summary
```

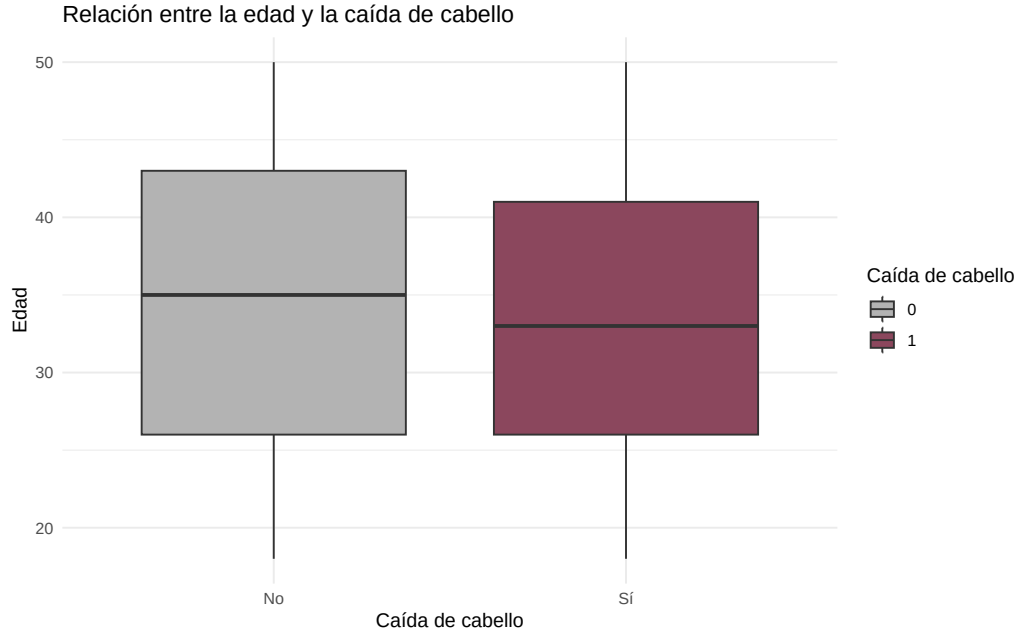
```
## # A tibble: 1 x 10
##   Genetics `Hormonal Changes` `Medical Conditions` `Medications & Treatments`
##      <int>              <int>                <int>                       <int>
## 1        2                  2                   11                          11
## # i 6 more variables: `Nutritional Deficiencies` <int>, Stress <int>,
## #   `Poor Hair Care Habits` <int>, `Environmental Factors` <int>,
## #   Smoking <int>, `Weight Loss` <int>
```

4.1.2 4.2. (1 punto)

Realizad un boxplot que relacione la edad con la caída de cabello, ¿qué podéis concluir?

Mostramos el gráfico de esta relación con un boxplot a través del paquete ggplot2.

```
ggplot(HairFallH, aes(x = factor(`Hair Loss`, labels = c("No", "Sí")),
                      y = Age,
                      fill = factor(`Hair Loss`))) +
  geom_boxplot() +
  scale_fill_manual(values = c("0" = "grey70", "1" = "#8B475D")) +
  labs(title = "Relación entre la edad y la caída de cabello",
       x = "Caída de cabello", y = "Edad",
       fill = "Caída de cabello") +
  theme_minimal()
```



A simple vista, no parece ser un factor ni tener correlación que la edad pueda influir del todo en la pérdida del cabello. Se puede apreciar que la mediana, así como el mínimo y el máximo son valores con diferencia poco significativa. De hecho, necesitaríamos tener en cuenta otras características, como stress, factores ambientales, o hormonal y genética. Sin embargo, podemos ver también que el dataset incluye solo una muestra de 999 individuos de 18 a 50, equilibrando los casos positivos y negativo de pérdida de pelo en una muestra homogénea, quizás para estudiar otros factores y no tener el sesgo por edad.

```
# Mostramos efectivamente la distribución de esta característica
summary(HairFallH$'Age')
```

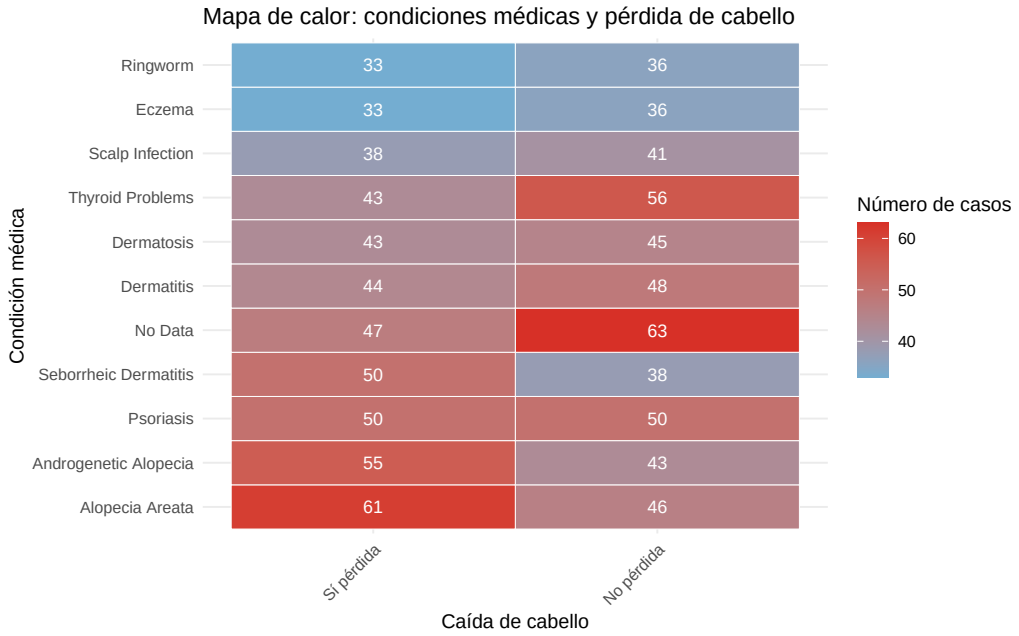
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##    18.00    26.00    34.00    34.19    42.00    50.00
```

4.1.3 4.3. (1 punto)

Realizad una tabla de frecuencias que relacione las condiciones médicas de las personas con el factor de caída del cabello y realizad gráfico tipo 'Mapa de calor' para visualizar cuales son los factores más relevantes.

```r
# Se crea la tabla de frecuencias
tabla_medical <- HairFallH %>%
  dplyr::group_by(`Medical Conditions`, `Hair Loss`) %>%
  dplyr::summarise(count = n(), .groups = "drop") %>%
  pivot_wider(names_from = `Hair Loss`, values_from = count, values_fill = 0)

# Ordenamos por mayor número de "Sí" (Hair Loss = 1) para tener una mejor visualización
tabla_medical <- tabla_medical %>%
  dplyr::arrange(desc(`1`))

# Convertimos a formato largo para el heatmap
heat_data <- tabla_medical %>%
  reshape2::melt(id.vars = "Medical Conditions", variable.name = "Hair_Loss", value.name = "count")

# Vamos a ordenar los factores para que aparezcan de mayor a menor
heat_data$`Medical Conditions` <- factor(heat_data$`Medical Conditions`,
                                         levels = tabla_medical$`Medical Conditions`)

# Creamos las etiquetas legibles para el eje x
heat_data$Hair_Loss <- factor(heat_data$Hair_Loss,
                              levels = c("1", "0"),
                              labels = c("Sí pérdida", "No pérdida"))

# Creamos el gráfico con ggplot
ggplot(heat_data, aes(x = Hair_Loss, y = `Medical Conditions`, fill = count)) +
  geom_tile(color = "white") +
  geom_text(aes(label = count), color = "white", size = 3.5) +
  scale_fill_gradient(low = "#74add1", high = "#d73027") +  # azul→rojo
  labs(title = "Mapa de calor: condiciones médicas y pérdida de cabello",
       x = "Caída de cabello", y = "Condición médica",
       fill = "Número de casos") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



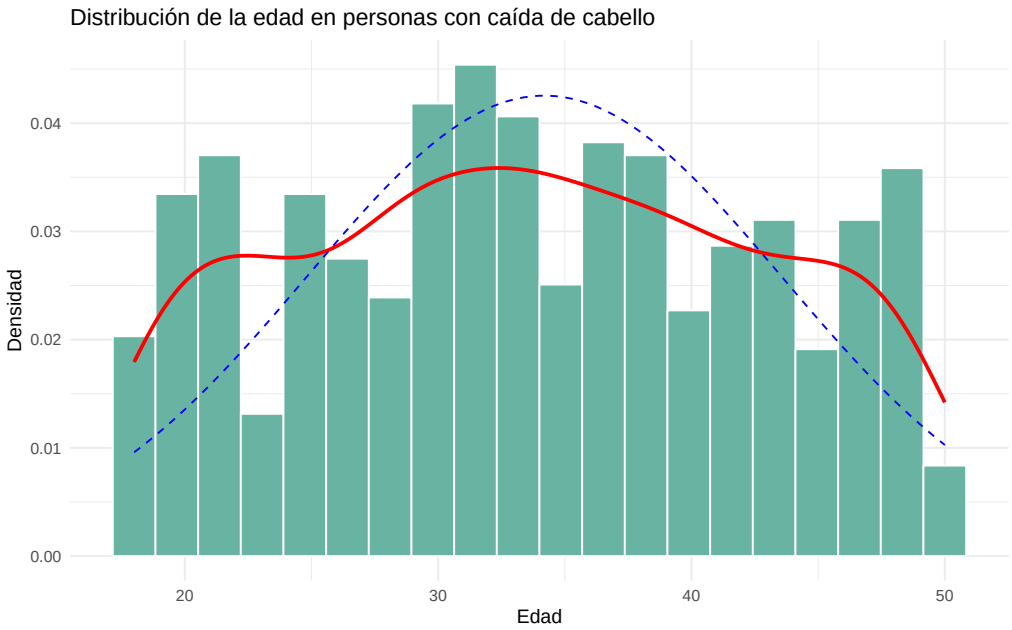
Quitando los valores de No Data, donde tenemos un conteo de 63 individuos, vemos que la "Alopecia Areata" y "Androgenetic Alopecia" inciden bastante , seguido por "Seborrheic Dermatisis" y "Psioriasis" . Seguramente afectarán la pérdida de cabello más que otros síntomas. Mientras que la no caída de cabello no está afectada por ejemplo con problemas de "Thyroid", más que el resto y un 50% con "Psioriasis". ***

4.1.4 4.4. (1 punto)

Representad un histograma de la edad de las personas con caída de cabello. A partir de aquí, y teniendo en cuenta los conceptos de normalidad trabajados al LAB1, representad la curva de densidad normal y evaluad la normalidad de la distribución y comentad los resultados.

Mostramos un gráfico donde mostraremos la distribución, tiene que mostrar una forma de campana más hacía el centro donde viene representada la mediana y la moda. Finalmente vemos la densidad normal.

```
ggplot(HairFallH %>% filter(`Hair Loss` == 1), aes(x = Age)) +
  geom_histogram(aes(y = ..density..), bins = 20, fill = "#69b3a2", color = "white") +
  geom_density(color = "red", size = 1) +
  stat_function(fun = dnorm,
                args = list(mean = mean(HairFallH$Age, na.rm = TRUE),
                            sd = sd(HairFallH$Age, na.rm = TRUE)),
                color = "blue", linetype = "dashed") +
  labs(title = "Distribución de la edad en personas con caída de cabello",
       x = "Edad", y = "Densidad") +
  theme_minimal()
```



A primera vista tenemos la densidad más hacía el promedio, estamos entorno a 30 años y la mediana toca justo a los 34. Tenemos valores de frecuencia más altos en un primer rango entre 20-25, un segundo entre 30-35, otro casi a los 40-45

Parece indicar que la muestra está bien repartida de forma homogénea para que se puedan tomar en cuenta los valores y no tener sesgos.

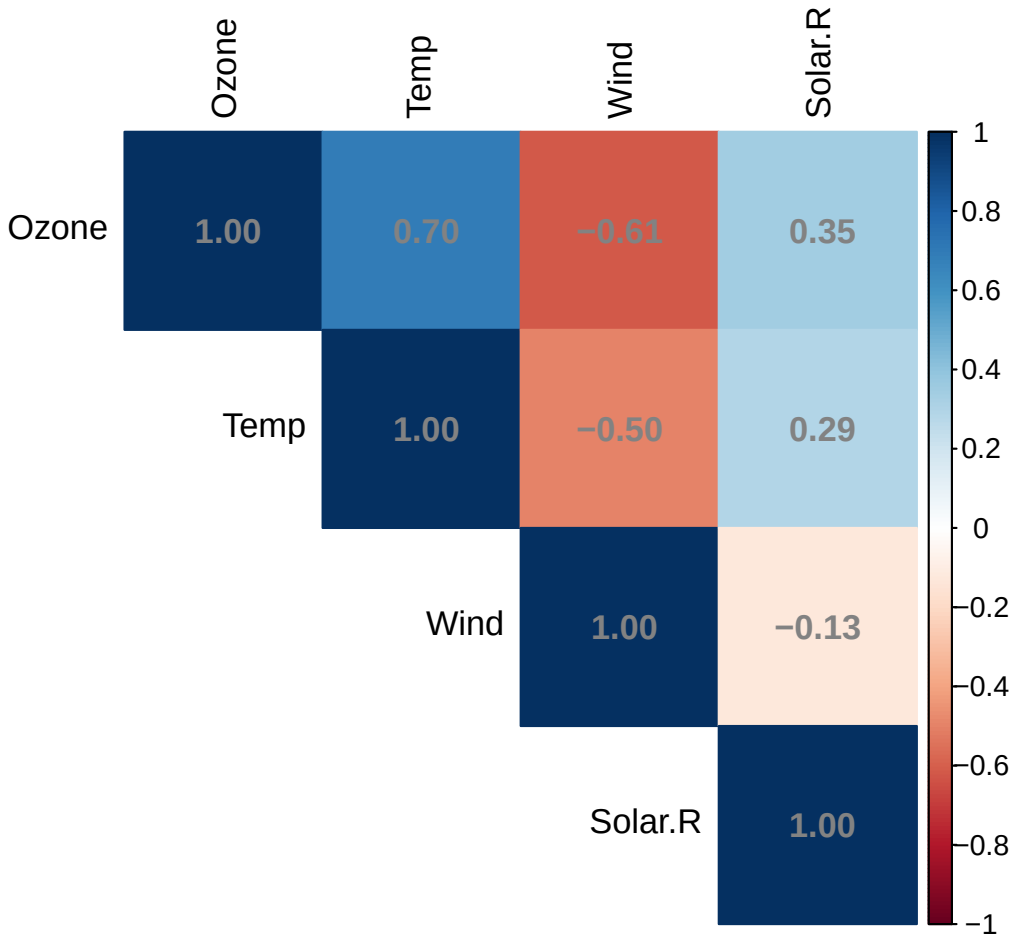
4.1.5 4.5. (1 punto)

A partir del ejercicio 5 del LAB1 que trabaja con los datos Airquality del paquete datasets, realiza un análisis de correlación lineal sobre dos de las variables del conjunto de datos y realiza la representación gráfica que más se ajuste.

```
data("airquality")

# Creamos y seleccionaremos las variables numéricas relevantes
cor_data <- airquality %>%
  select(Ozone, Temp, Wind, Solar.R) %>%
  drop_na()

# Creamos una matriz de correlación
cor_matrix <- cor(cor_data)
round(cor_matrix, 2)
```

```
##         Ozone  Temp  Wind Solar.R
## Ozone    1.00  0.70 -0.61    0.35
## Temp     0.70  1.00 -0.50    0.29
## Wind    -0.61 -0.50  1.00   -0.13
## Solar.R  0.35  0.29 -0.13    1.00
```

```
# Mostramos gráficamente
corrplot(cor_matrix, method = "color", type = "upper",
         addCoef.col = "#828282", tl.col = "black",
         title = "Correlación entre variables del dataset Airquality",
         mar = c(0,0,2,0))
```



Observamos que en la diagonal tenemos los valores 1, entre sus propias características. Pero las que nos importan son `Tempy` `Ozone` tienen una correlación fuerta positiva de 0,70 mientras por el opuesto, `Wind` y `Ozone` tiene una correlación casi fuerte negativa de -0,61. Esto significa que si crece el nivel de ozono la temperatura se eleva, es directamente proporcional. Mientras que si hay viento, no es porque hay más ozono o menos, es indirectamente proporcional. Por este motivo al crecer unos valores de una característica, baja el otro y viceversa.