PEC 1

UOC

NOMBRE:Juan Luis Acebal Rico

Introducción

Un periódico digital se dedica a realizar estudios sobre debates profundos como "tortilla con cebolla o sin cebolla", "agua con gas o sin gas", etc. Este mes ha tocado el debate relacionado con el tipo de café preferido, así como los análisis relacionados con costes y etc.

El fichero para realizar la PEC 1 es *"data_pac1"* y lo encontraréis en formato csv. Esta base de datos contiene información relacionada con los patrones de consumo de café en varios países.

Las variables que se encuentran en el dataset son las siguientes:

- *Country* : El nombre del país donde se recogieron los datos.
- *Year* : El año del registro.
- *Coffee.Consumption..kg.per.capita.per.year.* : Consumo per cápita de café.
- *Average.Coffee.Price..USD.per.kg.* : El precio medio del café por kilogramo en dólares.
- *Type.of.Coffee.Consumed* : Información sobre los tipos de café más populares en cada país
- *Population..millions.* : La población estimada de cada país.

Os puede ser útil consultar el siguiente material:

- Manuales de R
- Actividades Resueltas del Reto 1 (Estadística Descriptiva y Muestreo)

Hay que entregar la práctica en fichero pdf o html. Se recomienda generar el informe con Rmarkdown que genera automáticamente el html/pdf a entregar. Se puede utilizar el fichero .Rmd, que disponéis en la PEC, como plantilla para resolver los ejercicios.

Esta PEC debe realizarse de forma estrictamente individual, quedando del todo prohibido el uso de herramientas de IA. Cualquier indicio de copia será penalizado con un suspenso (D) por todas las partes implicadas y la posible evaluación negativa de la asignatura de forma íntegra.

Pregunta 1 (resolver con R). (4 puntos).

El periódico digital quiere preparar una campaña para diferentes redes sociales bajo el hashtag #megustaelcafémuchomucho. Así que se necesitan algunos gráficos así como información relacionada con el dataset data_pac1. Responde a las siguientes preguntas:

a) Describe brevemente la base de datos y los campos que contiene. (0.5 puntos).

Solución:

```
str(data_pac1)
```

```
## 'data.frame':    10000 obs. of  6 variables:
##  $ Country                                    : Factor w/ 50 levels "Country_1","Country_10",..: 33 1
##  $ Year                                       : int  2023 2011 2020 2005 2019 2004 2022 2008 2015 201
##  $ Coffee.Consumption..kg.per.capita.per.year.: num  9.25 9.98 3.31 2.44 4.64 ...
##  $ Average.Coffee.Price..USD.per.kg.          : num  6.47 4.35 8.77 11.75 9 ...
##  $ Type.of.Coffee.Consumed                    : Factor w/ 5 levels "Americano","Cappuccino",..: 1 5 4
##  $ Population..millions.                      : num  65.9 82.5 110.9 43.1 65.5 ...
```

```
summary(data_pac1)
```

```
##        Country           Year      Coffee.Consumption..kg.per.capita.per.year.
##  Country_26: 232   Min.   :2000   Min.   :2.000
##  Country_45: 225   1st Qu.:2006   1st Qu.:4.071
##  Country_49: 225   Median :2012   Median :6.094
##  Country_23: 223   Mean   :2012   Mean   :6.062
##  Country_28: 221   3rd Qu.:2018   3rd Qu.:8.061
##  Country_17: 220   Max.   :2023   Max.   :9.999
##  (Other)   :8654
##  Average.Coffee.Price..USD.per.kg. Type.of.Coffee.Consumed
##  Min.   : 4.001                    Americano :1975
##  1st Qu.: 6.728                    Cappuccino:2001
##  Median : 9.458                    Espresso  :1969
##  Mean   : 9.462                    Latte     :2071
##  3rd Qu.:12.136                    Mocha     :1984
##  Max.   :14.997
##
##  Population..millions.
##  Min.   :  1.002
##  1st Qu.: 37.466
##  Median : 75.022
##  Mean   : 75.167
##  3rd Qu.:112.596
##  Max.   :149.996
##
```

De lo observado, es un dataset del precio del cafe, su consumo, tipo, poblacion entre 2000 y 2023, tal y como esta escrito en el enunciado. Ademas, es un rango amplio de años, se puede ver la diferencia entre países, no se puede saber el país, pero se puede tener una idea de diferencias entre regiones (en principio, aunque realmente no las hay mucho)

b) Crea una nueva variable llamada "Habitantes.cantidad" de tipo factor sobre la cantidad de habitantes (variable Population..millions.) de cada fila del dataset. Esta nueva variable tiene 3 categorías: pequeño, medio o grande. Un tamaño pequeño es aquel que tiene menos de 37 millones de habitantes. Un tamaño medio es aquel que tiene entre 37 y 112 millones. Un tamaño grande es aquel que tiene más de 112 millones de habitantes. Obtenga la distribución de frecuencias absolutas y relativas de la variable "Habitantes.cantidad". (1.5 puntos).

Solución:

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

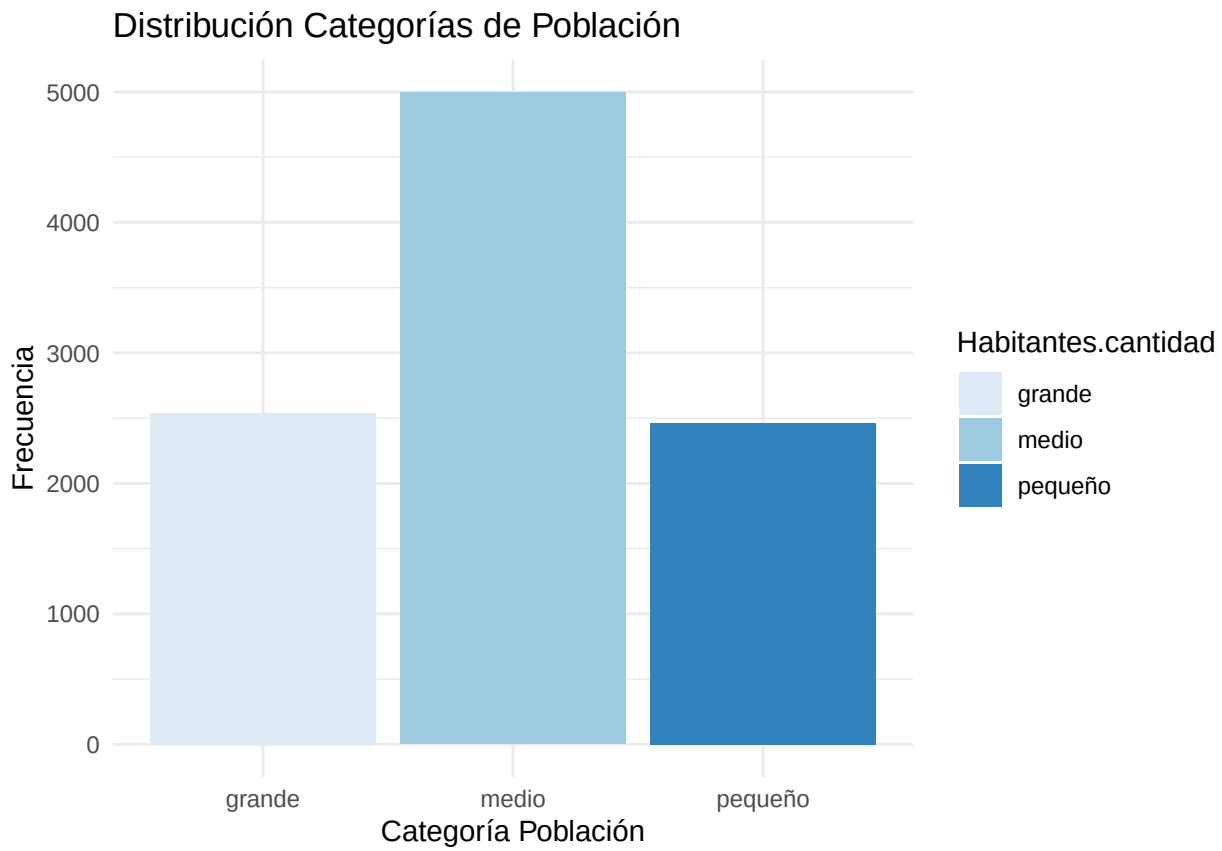
```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)

data_pac1 <- data_pac1 %>%
  mutate(Habitantes.cantidad = case_when(
    Population..millions. < 37 ~ "pequeño",
    Population..millions. >= 37 & Population..millions. <= 112 ~ "medio",
    Population..millions. > 112 ~ "grande"
  ))

# frecuencias absolutas y relativas
freq_table <- data_pac1 %>%
  count(Habitantes.cantidad) %>%
  mutate(Relative_Freq = n / sum(n))
freq_table
```

```
##   Habitantes.cantidad    n Relative_Freq
## 1              grande 2541        0.2541
## 2               medio 4997        0.4997
## 3            pequeño 2462        0.2462
```

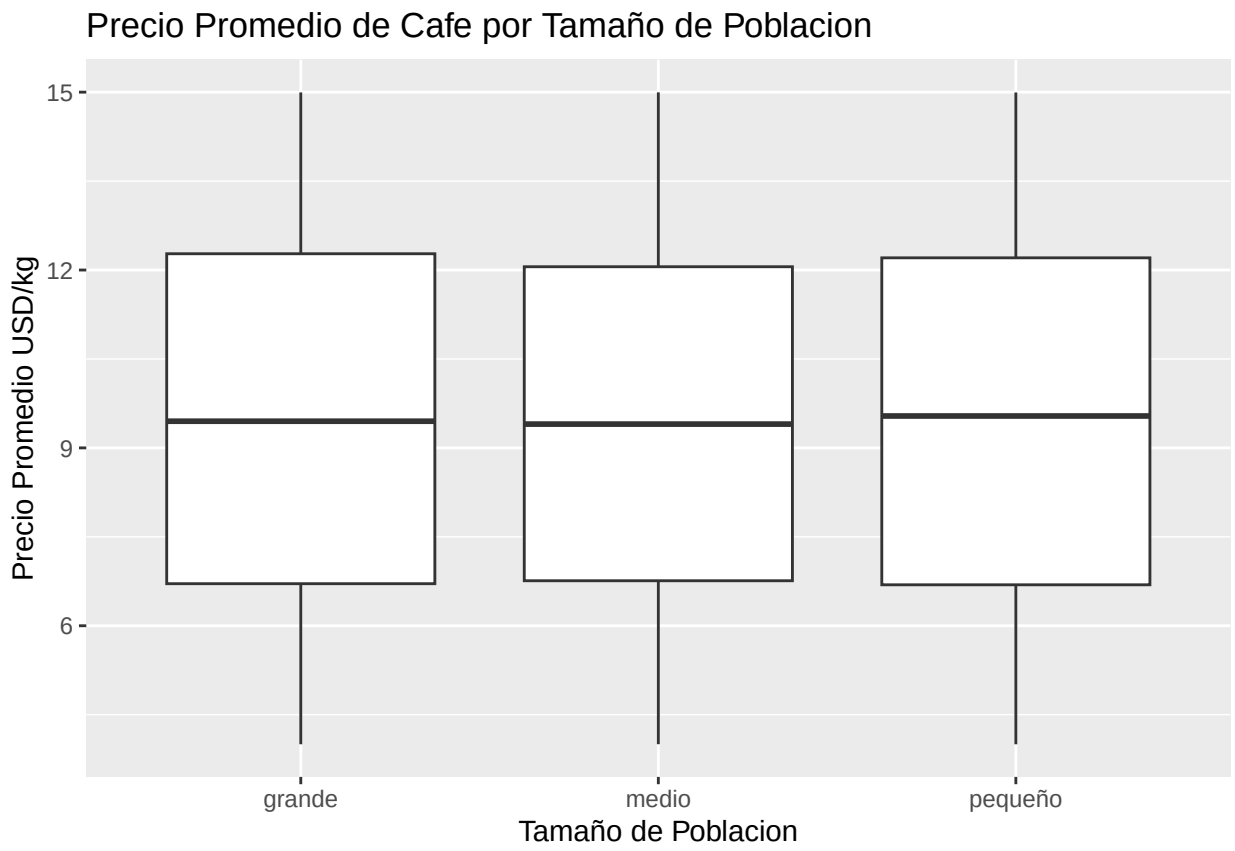
```
ggplot(data_pac1, aes(x = Habitantes.cantidad, fill = Habitantes.cantidad)) +
  geom_bar() +
  labs(title = "Distribución Categorías de Población",
       x = "Categoría Población",
       y = "Frecuencia") +
  scale_fill_brewer()+
  theme_minimal()
```



Bueno aquí vemos que lo que mas hay son paises medianos, con una frecuencia relativa de casi el 50%

c) Encontrad los resúmenes numéricos (media, mediana, cuartiles, desviación típica, mínimo y máximo) de la variable "Average.Coffee.Price..USD.per.kg." dependiendo de si "Habitantes.cantidad" es pequeño, medio o grande, y haced el diagrama de cajas correspondiente. Comentad los resultados. (2 puntos).

```r
# Calcular resumenes numericos
summary_table <- data_pac1 %>%
  group_by(Habitantes.cantidad) %>%
  summarize(
    media = mean(Average.Coffee.Price..USD.per.kg., na.rm = TRUE),
    mediana = median(Average.Coffee.Price..USD.per.kg., na.rm = TRUE),
    min = min(Average.Coffee.Price..USD.per.kg., na.rm = TRUE),
    max = max(Average.Coffee.Price..USD.per.kg., na.rm = TRUE),
    sd = sd(Average.Coffee.Price..USD.per.kg., na.rm = TRUE)
  )

# boxplot
library(ggplot2)
ggplot(data_pac1, aes(x = Habitantes.cantidad, y = Average.Coffee.Price..USD.per.kg.)) +
  geom_boxplot() +
  labs(title = "Precio Promedio de Cafe por Tamaño de Poblacion",
       x = "Tamaño de Poblacion",
       y = "Precio Promedio USD/kg")
```



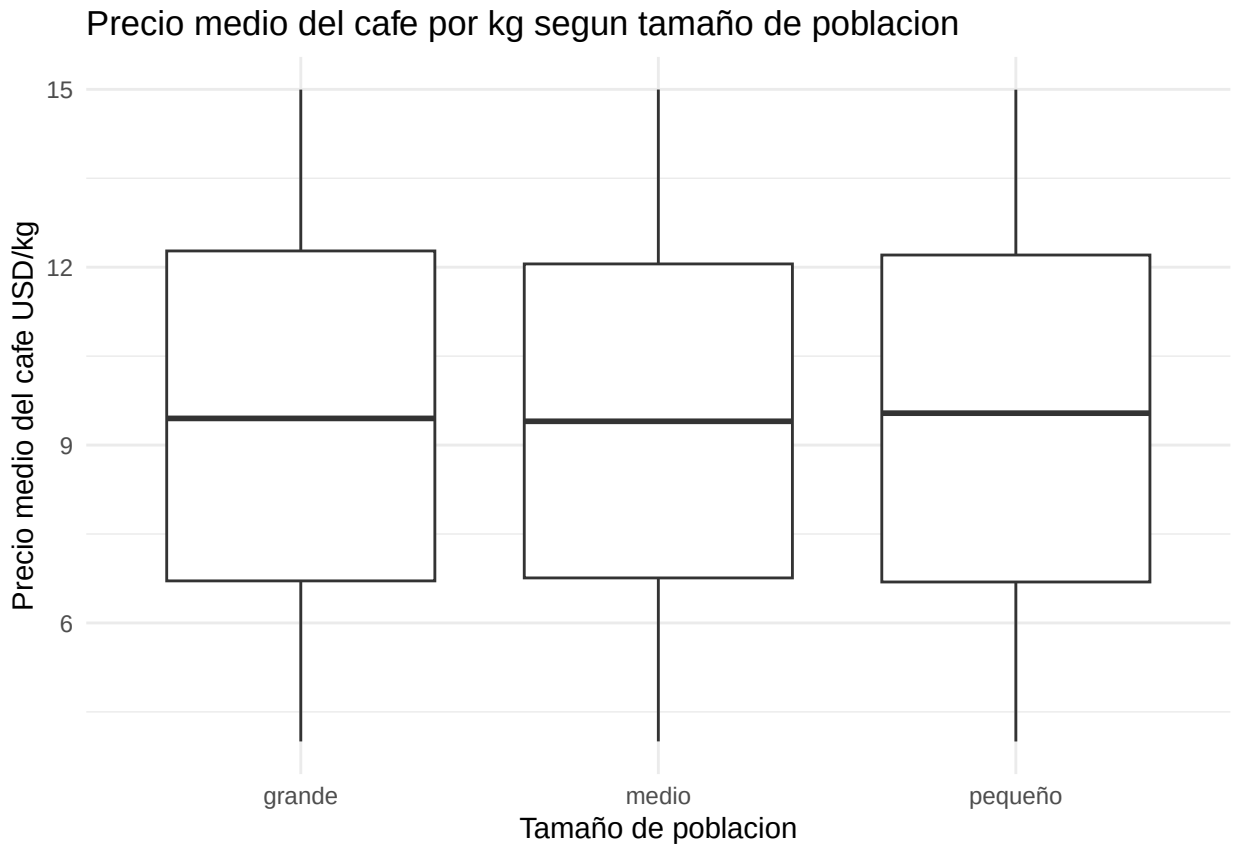
Solución:

```
library(ggplot2)
data_pac1 %>%
  group_by(Habitantes.cantidad) %>%
  summarize(
    media = mean(Average.Coffee.Price..USD.per.kg., na.rm = TRUE),
    mediana = median(Average.Coffee.Price..USD.per.kg., na.rm = TRUE),
    min = min(Average.Coffee.Price..USD.per.kg., na.rm = TRUE),
    max = max(Average.Coffee.Price..USD.per.kg., na.rm = TRUE),
    desviacion = sd(Average.Coffee.Price..USD.per.kg., na.rm = TRUE),
    q1 = quantile(Average.Coffee.Price..USD.per.kg., 0.25, na.rm = TRUE),
    q3 = quantile(Average.Coffee.Price..USD.per.kg., 0.75, na.rm = TRUE)
  )
```

```
## # A tibble: 3 x 8
##   Habitantes.cantidad media mediana   min   max desviacion    q1    q3
##   <chr>              <dbl>   <dbl> <dbl> <dbl>      <dbl> <dbl> <dbl>
## 1 grande              9.51    9.45  4.00  15.0       3.17  6.71  12.3
## 2 medio               9.42    9.40  4.00  15.0       3.13  6.76  12.1
## 3 pequeño             9.49    9.54  4.00  15.0       3.17  6.69  12.2
```

```
ggplot(data_pac1, aes(x = Habitantes.cantidad, y = Average.Coffee.Price..USD.per.kg.)) +
  geom_boxplot() +
  labs(
    title = "Precio medio del cafe por kg segun tamaño de poblacion",
```

```
    x = "Tamaño de poblacion",
    y = "Precio medio del cafe USD/kg"
  ) +
  theme_minimal()
```




```
head(data_pac1, n = 10)
```

```
##       Country Year Coffee.Consumption..kg.per.capita.per.year.
## 1  Country_39 2023                                    9.253939
## 2  Country_29 2011                                    9.981203
## 3  Country_15 2020                                    3.312916
## 4  Country_43 2005                                    2.436180
## 5   Country_8 2019                                    4.637849
## 6  Country_21 2004                                    5.693273
## 7  Country_39 2022                                    3.638570
## 8  Country_19 2008                                    4.411399
## 9  Country_23 2015                                    3.606966
## 10 Country_11 2021                                    8.194085
##    Average.Coffee.Price..USD.per.kg. Type.of.Coffee.Consumed
## 1                           6.467453               Americano
## 2                           4.346744                   Mocha
## 3                           8.767496                   Latte
## 4                          11.748750                Espresso
## 5                           8.999099                   Mocha
## 6                           9.059761                   Latte
```

```
## 7                          11.367855                  Latte
## 8                           6.798277                  Mocha
## 9                          12.270788                  Latte
## 10                          5.684249                  Mocha
##    Population..millions. Habitantes.cantidad
## 1               65.92948               medio
## 2               82.45668               medio
## 3              110.93886               medio
## 4               43.13721               medio
## 5               65.48426               medio
## 6              119.11866              grande
## 7              138.46090              grande
## 8              133.80712              grande
## 9               72.16701               medio
## 10              78.60900               medio
```

Pregunta 2 (resolver con R). (3 puntos).

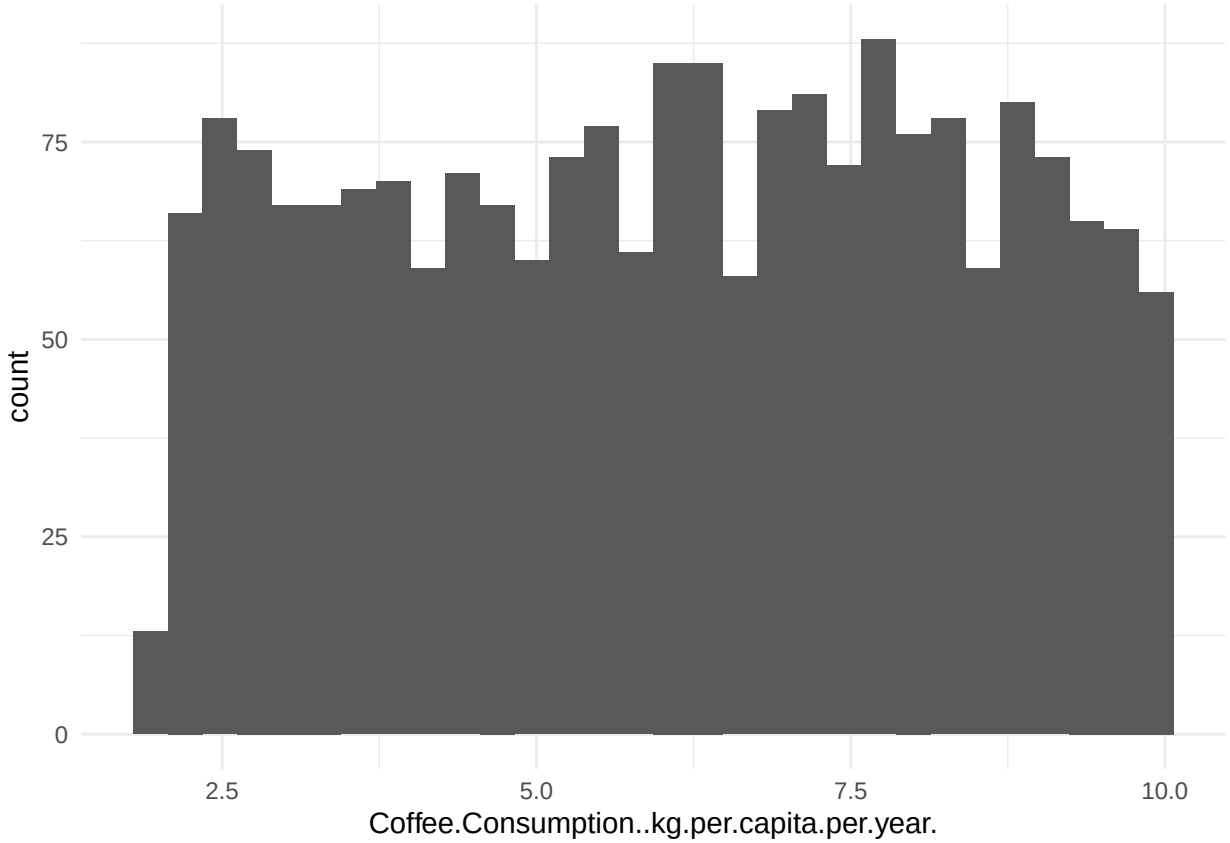
El webmaster del periódico digital quiere crear más contenido, ya que es necesario abrir el debate sobre el café. Responde a las siguientes preguntas:

a) Haga un histograma de la variable "Coffee.Consumption..kg.per.capita.per.year." solo para la categoría Latte. Interpreta el resultado.(1.5 puntos).

Solución:

```
latte_data <- filter(data_pac1, Type.of.Coffee.Consumed == "Latte")
ggplot(latte_data, aes(x = Coffee.Consumption..kg.per.capita.per.year.)) +
  geom_histogram()+
  theme_minimal()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
head(latte_data, n = 10)
```

```
##       Country Year Coffee.Consumption..kg.per.capita.per.year.
## 1  Country_15 2020                                    3.312916
## 2  Country_21 2004                                    5.693273
## 3  Country_39 2022                                    3.638570
## 4  Country_23 2015                                    3.606966
## 5  Country_11 2013                                    7.700624
## 6  Country_24 2021                                    7.079612
## 7  Country_36 2007                                    5.097743
## 8  Country_24 2018                                    9.756260
## 9  Country_38 2021                                    6.176589
## 10 Country_25 2006                                    2.531331
##    Average.Coffee.Price..USD.per.kg. Type.of.Coffee.Consumed
## 1                           8.767496                   Latte
## 2                           9.059761                   Latte
## 3                          11.367855                   Latte
## 4                          12.270788                   Latte
## 5                           8.353257                   Latte
## 6                          11.216089                   Latte
## 7                          13.112059                   Latte
## 8                          11.938542                   Latte
## 9                           4.835875                   Latte
## 10                          5.973760                   Latte
##    Population..millions. Habitantes.cantidad
```

```
## 1              110.93886              medio
## 2              119.11866             grande
## 3              138.46090             grande
## 4               72.16701              medio
## 5               79.83671              medio
## 6              130.43862             grande
## 7              126.15099             grande
## 8               25.26460            pequeño
## 9               13.88531            pequeño
## 10              86.56684              medio
```

```
# Cafa latte mas caro
latte_mas_caro <- latte_data %>%
  arrange(desc(Average.Coffee.Price..USD.per.kg.)) %>%
  select(Country, Year, Average.Coffee.Price..USD.per.kg.) %>%
  head(1)
latte_mas_caro
```

```
##      Country Year Average.Coffee.Price..USD.per.kg.
## 1 Country_21 2016                          14.99705
```

Pregunta 3 (resolver con R). (3 puntos)

El periódico digital quiere investigar en más detalle el refrán "Café cocido, café perdido", para ello quiere recoger muestras del dataset y hacer futuros análisis en función de la muestra.

a) Coja una muestra con muestreo aleatorio simple de tamaño 29 del dataset data_pac1 y calcula la media y desviación típica de la variable "Coffee.Consumption..kg.per.capita.per.year." (1.5 puntos).

Solución:

```
# Muestreo aleatorio simple de tamaño 29
set.seed(123456789)
sample_data <- sample_n(data_pac1, 29)
mean(sample_data$Coffee.Consumption..kg.per.capita.per.year., na.rm = TRUE)
```

```
## [1] 6.524224
```

```
sd(sample_data$Coffee.Consumption..kg.per.capita.per.year., na.rm = TRUE)
```

```
## [1] 1.96733
```

b) Se va a coger una nueva muestra del dataset data_pac1 de 40 cafés, estratificada por su tipología de café (Americano, capuccino, espresso, latte y mocha). Calcula el número de cafés capuccino que se necesitan para la muestra. (1.5 puntos).

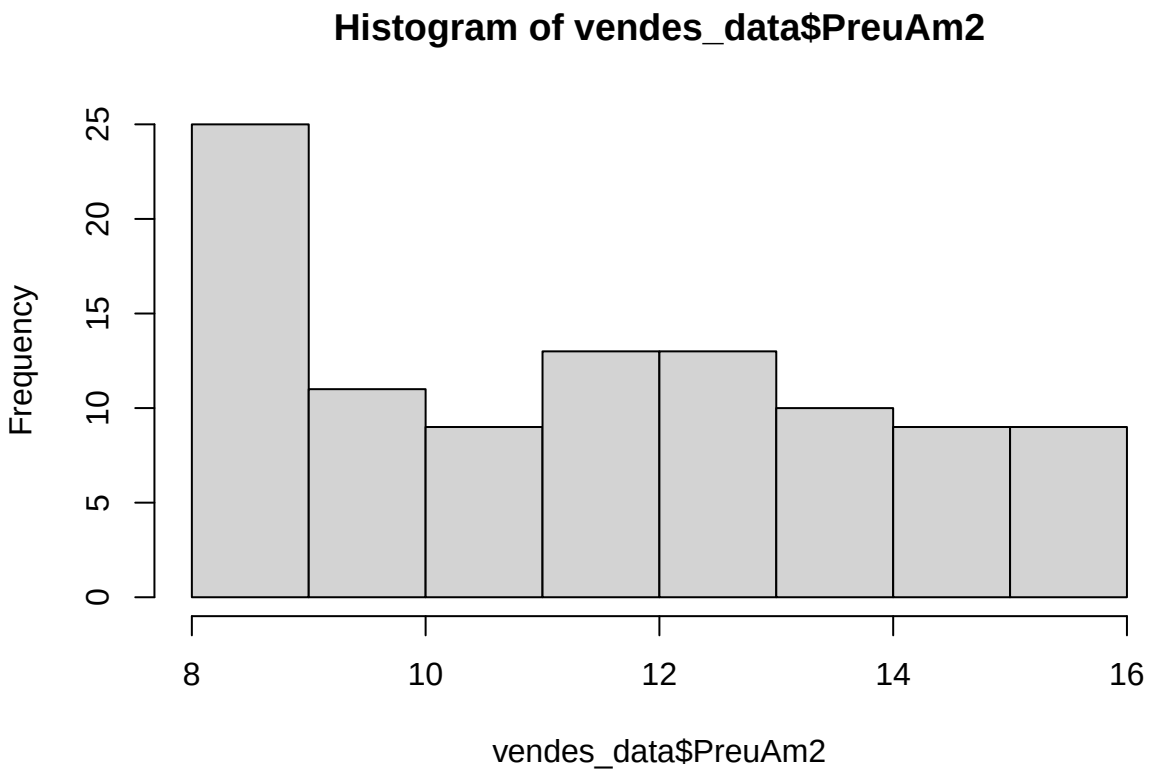
Solución:

```r
# Cargar librería
library(dplyr)

# calculo proporciones en base a la frecuencia.
proporciones <- data_pac1 %>%
  group_by(Type.of.Coffee.Consumed) %>%
  summarize(Frecuencia = n()) %>%
  mutate(Prop = Frecuencia / sum(Frecuencia))


tamano_muestra <- 40

# columma numero de columnas de muestras_Necesarias en base a la proporcion y
# tamaño de la muestra
proporciones <- proporciones %>%
  mutate(Muestras_Necesarias = round(Prop * tamano_muestra))

# Muestro las proporciones para cada tipo de cafe,
# aunque para nuestro caso son 8
proporciones
```

```
## # A tibble: 5 x 4
##   Type.of.Coffee.Consumed Frecuencia  Prop Muestras_Necesarias
##   <fct>                         <int> <dbl>               <dbl>
## 1 Americano                      1975 0.198                   8
## 2 Cappuccino                     2001 0.200                   8
## 3 Espresso                       1969 0.197                   8
## 4 Latte                          2071 0.207                   8
## 5 Mocha                          1984 0.198                   8
```

Datasets usados para los cuestionarios:

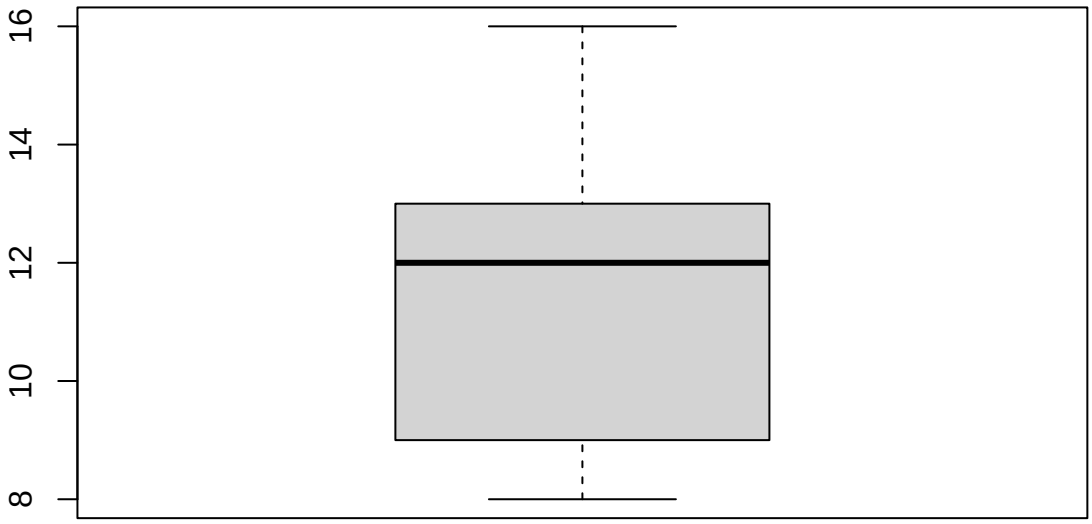
```r
vendes_data <- read.csv("vendes_pac1_P_15_1.csv", sep = ";", dec = ",")
hist(vendes_data$PreuAm2)
```




```
vendes_data2 <- read.csv("vendes_pac1_P_15_2.csv", sep = ";", dec = ",")
summary(vendes_data2$PreuDm2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.00   10.00   12.00   11.89   14.00   18.00
```

```
vendes_data4 <- read.csv("vendes_pac1_P_15_4.csv", sep = ";", dec = ",")
boxplot(vendes_data4$PreuAm2)
```


16
14
12
10
8