Tarea PIA02

Juan Manuel García Moyano
IABD
Informática y comunicaciones

Índice

Caso Práctico

Carla, como nueva integrante del departamento de informática de una gran empresa de marketing, ha recibido el encargo de explorar datos sobre el tratamiento de aguas residuales en España. Este análisis servirá de base para una campaña de concienciación sobre el uso responsable del agua.

El trabajo que tiene por delante es desafiante, ya que los datos que deberá analizar son muy amplios y provienen de diversas fuentes, lo que implica manejar grandes volúmenes de información y hacer frente a problemas relacionados con la calidad y la integración de los datos. Entre los obstáculos que Carla probablemente encontrará, está la variedad de formatos, la necesidad de limpiar registros erróneos o incompletos, y la integración de datos de diferentes orígenes.

El **lenguaje R** le ofrece una gran cantidad de herramientas que le facilitarán el proceso. Carla podrá utilizar R no solo para procesar los datos de manera eficiente, sino también para limpiar y transformar la información, realizar análisis estadísticos y crear visualizaciones claras y atractivas que destaquen patrones y tendencias. R le permitirá automatizar tareas repetitivas, facilitando que el análisis sea reproducible y escalable, lo que resultará fundamental en un proyecto de esta magnitud.

Gracias a la versatilidad de R, Carla estará en una posición ideal para superar los restos de manejar una gran cantidad de datos, y contribuir con información clave que impulse una campaña de concienciación efectiva sobre el consumo de agua en España.

¿Qué te pedimos que hagas?

Apartado 1: Instale el Intérprete y el R-Studio

Instale el intérprete de R y el IDE R-Studio, siga los pasos indicados en los contenidos de la unidad y familiarícese con el entorno. Trabaje con los ejercicios propuestos en la unidad y **aporte una captura de pantalla con alguno de estos ejercicios cargados.**

Este ejercicio no hay que hacerlo pero se puede descargar aquí.

Apartado 2: Explore los datos

Explore la página web del INE y busque la serie histórica 2000-2022 sobre *Suministros y Saneamiento de Agua en España por Comunidades Autónomas* y descargue los datos en forma de hoja de cálculo. Si no la encuentra, aquí tiene el enlace. Si el enlace está roto, puedes encontrar el fichero necesario en Información de Interés. Pero es importante que aprenda a buscar datos en las fuentes originales. Una vez tenga el fichero, ábralo y estudie su formato y composición. Normalmente, el formato de estos ficheros proporcionales por el INE suele ser bastante estable. **Explique los datos que contiene esta hoja de cálculo, su composición y formato, apoye su explicación con una captura.**

La composición de los datos que contiene la hoja de calculo son:

- El volumen de aguas residuales tratadas (m^3/día)
- El volumen total de agua reutilizada (m^3/día)
- El importe facturado por alcantarillado y depuración (miles €)
- Longitud de la red de alcantarillado (km)
- El volumen de lodos generados en el tratamiento de agua residuales (toneladas de materia seca/año).

Estos a su vez se dividen por el total nacional y por provincias. Las variables que se pueden observar son los años que va desde 2022 hasta 2000.

El formato que presenta el fichero es, en la filas podemos encontrarnos con alguna de las opciones anteriores y agrupadas por el total o la comunidad/ciudad autónoma, y en las columnas

tendría los años (2022 – 2000). Estos datos presentan datos numéricos y nulos. A la hora de realizar un estudio se debe tener especial cuidado en los datos nulos y en las unidades de medida que se utiliza en cada uno.

Apartado 3: Cargue los datos en un Dataframe

Es el momento de cargar los datos en un Dataframe. La librería R a utilizar dependerá del formato de los datos, si lo ha descargado en excel puede utilizar readxl y si lo ha descargado para Calc debe utilizar readODS, aunque hay otras librerías disponibles, sea como sea no olvide:

- Es necesario instalar las librerías (una vez) y cargarlas antes de hacer uso de las funciones que proporciona.
- Indique apropiadamente la ruta donde se encuentra el fichero y el rango de datos a cargar desde la hoja de cálculo (no nos interesa la cabecera ni el pie, si lo hubiera). Aunque puede variar según la librería utilizada, debe obtener algo parecido a lo siguiente, tenga en cuenta que en las capturas solo aparecen las primeras columnas.

```
> datos
# A tibble: 114 × 20
   ...1       `2022` `2020` `2018` `2016` `2014` `2013` `2012` `2011`
   <chr>       <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl> <chr>  <chr>
 1 Total Na…      NA     NA     NA     NA     NA     NA   NA     NA
 2 Volumen …  1.54e7 1.34e7 1.37e7 1.29e7 1.35e7 1.37e7 13592… 13497…
 3 Volumen …  1.24e6 1.46e6 1.53e6 1.35e6 1.45e6 1.46e6 15026… 16664…
 4 Importe …  2.41e6 2.48e6 2.50e6 2.49e6 2.54e6 2.38e6 23461… 20055…
 5 Longitud…  2.13e5 1.49e5 1.47e5 1.46e5 1.47e5 1.49e5 ..     ..
 6 Volumen …  1.16e6 1.15e6 1.21e6 1.17e6 1.13e6 1.12e6 12333… 13315…
 7 01 Andal…      NA     NA     NA     NA     NA     NA   NA     NA
 8 Volumen …  1.64e6 1.92e6 2.11e6 1.91e6 2.01e6 2.07e6 23308… 22567…
 9 Volumen …  8.91e4 1.00e5 1.02e5 1.13e5 1.57e5 1.72e5 240384 313820
10 Importe …  5.07e5 4.55e5 4.27e5 3.91e5 4.19e5 3.81e5 374806 331882
# i 104 more rows
# i 11 more variables: `2010` <chr>, `2009` <chr>, `2008` <chr>,
#   `2007` <chr>, `2006` <chr>, `2005` <chr>, `2004` <chr>, `2003` <chr>,
#   `2002` <chr>, `2001` <chr>, `2000` <chr>
# i Use `print(n = ...)` to see more rows
>
```

Miguel Ángel López Montero. *Captura sobre carga de los datos*

Enlace del fichero:
https://drive.google.com/drive/folders/1fxldjL8dISDFFkSFo3DWC33RpzUeQoa1?usp=sharing

Código:

```
install.packages("readxl") # Instalo la librería readxl para leer ficheros xls
library(readxl) # Cargo la librería readxl
datos <-
read_excel("E:/IABD/PIA/Tema2/R/Ejercicios/Deberes/Tarea1/Excel/Excel_Tarea_PIA02_Suministros.xls") # Leo el fichero y lo guardo en un dataframe
colnames(datos) [1] <- "Datos" # A la primera columna le cambio el valor a Datos
print(datos) # Imprimo por consola el dataframe
```

Captura:

```
> print(datos) # Imprimo por consola el dataframe
# A tibble: 114 x 20
   Datos      `2022` `2020` `2018` `2016` `2014` `2013` `2012` `2011` `2010` `2009` `2008` `2007` `2006` `2005`
   <chr>       <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl> <chr>  <chr>  <chr>  <chr>  <chr>  <chr>  <chr>  <chr>
 1 Total Na…      NA     NA     NA     NA     NA     NA   NA     NA     NA     NA     NA     NA     NA     NA
 2 Volumen …  1.54e7  1.34e7  1.37e7  1.29e7  1.35e7  1.37e7 13592… 13497… 13326… 12800… 12371… 12519… 13371… 13804…
 3 Volumen …  1.24e6  1.46e6  1.53e6  1.35e6  1.45e6  1.46e6 15026… 16664… 13460… 14642… 14398… 13721… 13351… 10835…
 4 Importe …  2.41e6  2.48e6  2.50e6  2.49e6  2.54e6  2.38e6 23461… 20055… 19910… 18850… 18514… 19408… 14266… 13295…
 5 Longitud…  2.13e5  1.49e5  1.47e5  1.46e5  1.47e5  1.49e5 ..     ..     ..     ..     ..     ..     ..     ..
 6 Volumen …  1.16e6  1.15e6  1.21e6  1.17e6  1.13e6  1.12e6 12333… 13315… ..     ..     ..     ..     ..     ..
 7 01 Andal…      NA     NA     NA     NA     NA     NA   NA     NA     NA     NA     NA     NA     NA     NA
 8 Volumen …  1.64e6  1.92e6  2.11e6  1.91e6  2.01e6  2.07e6 23308… 22567… 18384… 15667… 14885… 13872… 17673… 17456…
 9 Volumen …  8.91e4  1.00e5  1.02e5  1.13e5  1.57e5  1.72e5 240384 313820 338035 326892 264917 333936 124125 130757
10 Importe …  5.07e5  4.55e5  4.27e5  3.91e5  4.19e5  3.81e5 374806 331882 296192 284071 267801 283755 244127 275825
# i 104 more rows
# i Use `print(n = ...)` to see more rows
>
```

Juan Manuel García Moyano. Captura mostrando los datos

Apartado 4: Busque las autonomías que aparecen en la hoja de cálculo.

Deben estar todas pero, vamos a ver. Fíjese en que las autonomías vienen en una fila con ningún otro dato y que se encuentran en la primera columna, también podemos ver que viene precedida de un número. Debe extraer el nombre de todas las autonomías en un vector de cadenas, seleccionando correctamente los datos del dataframe, obteniendo algo similar a esto:

```
> autonomias
 [1] "01 Andalucía"                    "02 Aragón"
 [3] "03 Asturias, Principado de"      "04 Balears, Illes"
 [5] "05 Canarias"                     "06 Cantabria"
 [7] "07 Castilla y León"              "08 Castilla - La Mancha"
 [9] "09 Cataluña"                     "10 Comunitat Valenciana"
[11] "11 Extremadura"                  "12 Galicia"
[13] "13 Madrid, Comunidad de"         "14 Murcia, Región de"
[15] "15 Navarra, Comunidad Foral de" "16 País Vasco"
[17] "17 Rioja, La"
```

Miguel Ángel López Montero. Captura de la lista de autonomías

Código:

```
encontrarAutonomias <- function(datos) {
  # Hago una secuencia desde 1 hasta el número de filas. La segunda parte se encarga
  # de devolver el número de las filas y columnas que tienen un NAs. Elimino de la
  # secuencia las filas que coinciden.
  filasAutonomias <- seq(1, nrow(datos)) %in% which(is.na(datos), arr.ind = TRUE)["row"]
  # Me quedo con todas las filas y la primera columna
  autonomiasDf <- datos[filasAutonomias, 1]
  # Del Dataframe anterior, obtengo la cantidad de filas y le resto uno para que
  # luego tail me de las n filas - 1, es decir, elimino "Total Nacional"
  autonomias <- tail(autonomiasDf, n = (nrow(autonomiasDf) - 1))
  return (c(autonomias))
}

autonomias <- encontrarAutonomias(datos)
print(autonomias)
```

Captura:

```
> print(autonomias)
$Datos
 [1] "01 Andalucía"                    "02 Aragón"
 [3] "03 Asturias, Principado de"      "04 Balears, Illes"
 [5] "05 Canarias"                     "06 Cantabria"
 [7] "07 Castilla y León"              "08 Castilla - La Mancha"
 [9] "09 Cataluña"                     "10 Comunitat Valenciana"
[11] "11 Extremadura"                  "12 Galicia"
[13] "13 Madrid, Comunidad de"         "14 Murcia, Región de"
[15] "15 Navarra, Comunidad Foral de" "16 País Vasco"
[17] "17 Rioja, La"                    "Ceuta y Melilla"

>
```

Apartado 5: Implemente una función para extraer datos

Diseñe una función a la que pasaremos el dataframe y el nombre de una autonomía, dicha función devolverá todos los datos del dataframe referentes a dicha autonomía. Observe que los datos de una autonomía se encuentran en las 5 filas siguientes a la fila donde aparece el nombre de la autonomía.

```
> buscaAutonomia(datos,"Madrid")
# A tibble: 5 × 20
  ...1        `2022` `2020` `2018` `2016` `2014` `2013` `2012` `2011` `2010` `2009` `2008` `2007` `2006` `2005` `2004`
  <chr>        <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl> <chr>  <chr>  <chr>  <chr>  <chr>  <chr>  <chr>  <chr>  <chr>
1 Volumen de… 1.63e6 1.40e6 1.49e6 1.50e6 1.68e6 1.69e6 15709… 16568… 15499… 14286… 15153… 14323… 18438… 20438… 21025…
2 Volumen to… 6.17e4 3.63e4 3.48e4 3.42e4 3.99e4 2.92e4 31160  24874  18677  17051  16945  17352  15392  13526  12980
3 Importe fa… 3.57e5 2.95e5 2.92e5 3.13e5 3.12e5 3.06e5 293606 283401 266204 270007 257560 215085 192405 165388 158221
4 Longitud d… 2.20e4 1.52e4 1.51e4 1.48e4 1.42e4 1.22e4 ..     ..     ..     ..     ..     ..     ..     ..     ..
5 Volumen de… 1.10e5 1.00e5 1.09e5 1.03e5 1.15e5 1.17e5 98297  122881 ..     ..     ..     ..     ..     ..     ..
# ℹ 4 more variables: `2003` <chr>, `2002` <chr>, `2001` <chr>, `2000` <chr>
```

Miguel Ángel López Montero. Datos de la Comunidad de Madrid

Código:

```
# Función para extraer datos del dataframe, para ello se le pasa un dataframe y
# un string con el nombre de una comunidad/ciudad autónoma y devuelve un dataframe
buscarAutonomia <- function(datos, comunidad) {
  # datos <- datos[(grep(toupper(comunidad), toupper(datos$...1)) + 1):(grep(toupper(comunidad), toupper(datos$...1)) + 5), ]
  # Busco la fila donde se encuentre la comunidad en la primera columna,
  # a esa fila le sumo 1. Después hago lo mismo pero sumandole 5 y muestro una rango
  # entre la fila encontrada + 1 : fila encontrada + 5
  return (datos[(grep(toupper(comunidad), toupper(datos$Datos)) + 1):(grep(toupper(comunidad), toupper(datos$Datos)) + 5), ])
}
print(buscarAutonomia(datos, "Madrid"))
```

Captura:

```
> print(buscarAutonomia(datos, "Madrid"))
# A tibble: 5 x 20
  Datos `2022` `2020` `2018` `2016` `2014` `2013` `2012` `2011` `2010` `2009` `2008` `2007` `2006` `2005` `2004`
  <chr>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl> <chr>  <chr>  <chr>  <chr>  <chr>  <chr>  <chr>  <chr>  <chr>
1 Volu… 1.63e6 1.40e6 1.49e6 1.50e6 1.68e6 1.69e6 15709… 16568… 15499… 14286… 15153… 14323… 18438… 20438… 21025…
2 Volu… 6.17e4 3.63e4 3.48e4 3.42e4 3.99e4 2.92e4 31160  24874  18677  17051  16945  17352  15392  13526  12980
3 Impo… 3.57e5 2.95e5 2.92e5 3.13e5 3.12e5 3.06e5 293606 283401 266204 270007 257560 215085 192405 165388 158221
4 Long… 2.20e4 1.52e4 1.51e4 1.48e4 1.42e4 1.22e4 ..     ..     ..     ..     ..     ..     ..     ..     ..
5 Volu… 1.10e5 1.00e5 1.09e5 1.03e5 1.15e5 1.17e5 98297  122881 ..     ..     ..     ..     ..     ..     ..
# i 4 more variables: `2003` <chr>, `2002` <chr>, `2001` <chr>, `2000` <chr>
>
```

Juan Manuel García Moyano. Captura de los datos filtrados

Apartado 6: Prepare los datos

Ya podemos extraer los datos que nos interesan, pero antes de operar con ellos necesitamos transformarlos, primero marcaremos correctamente los datos perdidos, en lugar de con dos puntos seguidos (como hace el INE) nosotros debemos utilizar el valor **NA** de R, además pasaremos a número todas las columnas menos la primera. Diseña una función que realice estas tareas, debes obtener algo parecido a esto:

```
> madrid <- preparaDatos(buscaAutonomia(datos,"Madrid"))
> madrid
                                                                 Dato   X2022   X2020   X2018   X2016   X2014   X2013   X2012   X2011   X2010   X2009   X2008   X2007
1                                          Volumen de aguas residuales tratadas 1634765 1401292 1487112 1500326 1681915 1687026 1570983 1656841 1549957 1428676 1515336 1432331
2                                            Volumen total de agua reutilizada   61713   36320   34832   34178   39865   29175   31160   24874   18677   17051   16945   17352
3                                  Importe facturado por alcantarillado y depuración  356955  294720  291575  312980  311654  306426  293606  283401  266204  270007  257560  215085
4                                        Longitud de la red de alcantarillado (km)   22048   15183   15083   14841   14188   12233      NA      NA      NA      NA      NA      NA
5 Volumen de lodos generados en el tratamiento de aguas residuales (toneladas de materia seca/año)  109560  100111  109253  103141  115375  116992   98297  122881      NA      NA      NA      NA
```

Miguel Ángel López Montero. Datos preparados

Mira como no aparecen los dos puntos, sino NA. La primera columna ahora se llama Dato (le cambiamos el nombre) el resto de columnas se ha obtenido a pasar a numéricas las columnas cargadas en el dataframe, yo he utilizado la función **lapply,** pero puede hacerse de cualquier otra forma.

Código:

```
# Función para reemplazar los datos perdidos por NA y cambiar a número todas las
# columnas menos la primera
preparaDatos <- function(datos) {
  datos[datos == ".."] <- NA # Esto lo pongo para que no me salga los warning porque as.numeric
  # los datos que no puede pasar a número los pone como NA así podemos cambiar un string por
NA
  # Paso todos los datos a númericos menos la primera columna. Luego, el resultado
  # obtenido se lo asigno a datos excepto a la primera columna que se queda igual.
  # No hace falta pasar previamente los ".." a NA, porque los datos que no pueda pasar
  # as.numeric los convierte en NA.
  datos[, -1] <- apply(datos[, -1], 2, as.numeric)
  return (datos)
}
madrid <- preparaDatos(buscarAutonomia(datos, "Madrid"))
print(madrid)
```

Captura:

```
> print(madrid)
# A tibble: 5 × 20
  Datos          `2022` `2020` `2018` `2016` `2014` `2013`  `2012`  `2011`  `2010`  `2009`
  <chr>           <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
1 Volumen de … 1.63e6 1.40e6 1.49e6 1.50e6 1.68e6 1.69e6 1570983 1656841 1549957 1428676
2 Volumen tot… 6.17e4 3.63e4 3.48e4 3.42e4 3.99e4 2.92e4   31160   24874   18677   17051
3 Importe fac… 3.57e5 2.95e5 2.92e5 3.13e5 3.12e5 3.06e5  293606  283401  266204  270007
4 Longitud de… 2.20e4 1.52e4 1.51e4 1.48e4 1.42e4 1.22e4      NA      NA      NA      NA
5 Volumen de … 1.10e5 1.00e5 1.09e5 1.03e5 1.15e5 1.17e5   98297  122881      NA      NA
# i 1 more variable: `2000` <dbl>
> 
```

Figura 1: Juan Manuel García Moyano. Captura de los datos pasados a numéricos y NA

Apartado 7: Calcule las medias

Ahora estamos en disposición de operar con los datos. Diseña una función que, haciendo uso de las funciones anteriores, calcule la media de todos los datos de una comunidad dada. Aquí algunos ejemplos:

```
> mediasAutonomia(datos,"Madrid")
                                                                                  Dato      Media
1                                                    Volumen de aguas residuales tratadas 1529073.74
2                                                       Volumen total de agua reutilizada   22513.47
3                                             Importe facturado por alcantarillado y depuración  227921.21
4                                                     Longitud de la red de alcantarillado (km)   15596.00
5 Volumen de lodos generados en el tratamiento de aguas residuales (toneladas de materia seca/año)  109451.25
> mediasAutonomia(datos,"Andalucía")
                                                                                  Dato      Media
1                                                    Volumen de aguas residuales tratadas 1806841.74
2                                                       Volumen total de agua reutilizada  175958.95
3                                             Importe facturado por alcantarillado y depuración  303998.74
4                                                     Longitud de la red de alcantarillado (km)   31765.67
5 Volumen de lodos generados en el tratamiento de aguas residuales (toneladas de materia seca/año)  233083.62
```

Miguel Ángel López Montero. Medias por comunidad

Código:

```
# Función que calcula las medias de una comunidad/ciudad autónoma. Para ello le pasamos un
# dataframe y la comunidad que nos interesa. Devuelve un dataframe con los Datos y las Medias
mediaAutonomia <- function(datos, comunidad) {
  # datos <- colMeans(preparaDatos(buscarAutonomia(datos, comunidad)), na.rm = TRUE)
  autonomiaDatosRes <- preparaDatos(buscarAutonomia(datos, comunidad)) # Obtengo los datos
  # de la comunidad/ciudad autónoma con los datos bien formateados
  autonomiaDatosRes$Media <- c(rowMeans(autonomiaDatosRes[, -1], na.rm = TRUE)) # Calculo
la
  # media por filas y se lo asigno a una nueva columna
  return (autonomiaDatosRes[, c("Datos", "Media")])
}

print(mediaAutonomia(datos, "Madrid"))

print(mediaAutonomia(datos, "Andalucía"))
```

Captura:

```
> print(mediaAutonomia(datos, "Madrid"))
# A tibble: 5 × 2
  Datos                                                                                 Media
  <chr>                                                                                 <dbl>
1 Volumen de aguas residuales tratadas                                                1529074.
2 Volumen total de agua reutilizada                                                     22513.
3 Importe facturado por alcantarillado y depuración                                    227921.
4 Longitud de la red de alcantarillado (km)                                             15596
5 Volumen de lodos generados en el tratamiento de aguas residuales (toneladas de materia seca/año)  109451.
> print(mediaAutonomia(datos, "Andalucía"))
# A tibble: 5 × 2
  Datos                                                                                 Media
  <chr>                                                                                 <dbl>
1 Volumen de aguas residuales tratadas                                                1806842.
2 Volumen total de agua reutilizada                                                    175959.
3 Importe facturado por alcantarillado y depuración                                    303999.
4 Longitud de la red de alcantarillado (km)                                             31766.
5 Volumen de lodos generados en el tratamiento de aguas residuales (toneladas de materia seca/año)  233084.
>
```

Juan Manuel García Moyano. Captura de las medias de una Comunidad.

Bibliografía

- https://ine.es/
- https://ine.es/up/2Hs7okgKi2
- https://r-coder.com/colSums-rowSums-colMeans-rowMeans-en-r/