Epsilon – *Revista de Educación Matemática* *2024, nº 118, 91-100*
***Sección**: Aportes desde la investigación en Educación Matemática*

Visualización de datos con R

Carlos Santiago Sánchez Muñoz
Universidad de Granada, *carlossamu7@correo.ugr.es*

Manuel Alejandro Jiménez Morales
Universidad de Granada, *majm86@correo.ugr.es*

Resumen: *Este artículo explora el potencial del lenguaje R en la enseñanza de visualización y análisis de datos a nivel de Educación Secundaria y Bachillerato. Se introduce la herramienta y se proveen diversas fuentes susceptibles de guiar al profesorado en la creación de experiencias de clase ligadas a la enseñanza de habilidades de programación, análisis y visualización de datos, y pensamiento computacional. Por último, se proponen un conjunto de actividades que introducen, promoviendo un aprendizaje activo y experimental, el empleo de una base de datos real, ilustrando la gran versatilidad de R para la manipulación, análisis y representación gráfica de la información relevante.*

Palabras clave: *visualización de datos, ciencia de datos, estadística, pensamiento computacional, programación en R.*

Data Visualization with R

Abstract: *This article explores the potential of the R programming language in teaching data visualization and analysis at the secondary education level. The tool is introduced and various resources are provided to guide teachers in creating class experiences focused on programming, data analysis, visualization skills, and computational thinking. Lastly, it proposes a set of activities that promotes active, hands-on learning, by using a real database, illustrating R's versatility for data manipulation, analysis, and graphical representation of relevant information.*

Key words: *data visualization, data science, statistics, computational thinking, R programming.*

1. INTRODUCCIÓN

El uso de R en el entorno educativo para el tratamiento y la visualización de datos se ha convertido en una herramienta poderosa que fomenta el aprendizaje de la estadística, la ciencia de datos y la programación de manera práctica y accesible. R es un lenguaje de programación de código abierto, ampliamente utilizado tanto en el ámbito académico como profesional, que permite a los estudiantes y educadores analizar grandes volúmenes de datos y representarlos gráficamente de manera clara y comprensible.

Una de las principales ventajas de R es su versatilidad y la amplia gama de paquetes y librerías disponibles para el manejo y visualización de datos, como *dplyr*, *ggplot2* y *shiny*. Estos paquetes permiten a los usuarios crear gráficos interactivos y personalizados que ayudan a ilustrar patrones y tendencias en los datos, facilitando su interpretación y análisis.

En el contexto educativo, R proporciona una oportunidad única para que los estudiantes desarrollen habilidades de pensamiento crítico y resolución de problemas a través del manejo de

datos reales. Al igual que el resto de herramientas, R se puede utilizar para representar visualmente la información, pero con el valor añadido de ofrecer una mayor flexibilidad y control sobre el proceso de análisis y visualización.

Además, R puede adaptarse a diferentes niveles educativos, desde la enseñanza secundaria hasta la educación superior. Su implementación en el aula puede ser progresiva, comenzando con la introducción a la programación básica y la creación de gráficos sencillos, hasta llegar a la manipulación avanzada de datos y la construcción de aplicaciones interactivas. Este lenguaje cuenta con una comunidad activa que proporciona tutoriales, foros y recursos educativos accesibles para estudiantes y docentes. Sin embargo, a diferencia de otras herramientas basadas en la web, el uso de R requiere un mayor grado de familiaridad con la programación, lo cual puede ser un reto, pero también una ventaja educativa, ya que fomenta el desarrollo de competencias digitales avanzadas.

Esta combinación de análisis de datos y visualización permite a los estudiantes no solo entender conceptos matemáticos y estadísticos, sino también explorar y comunicar sus hallazgos de manera efectiva a través de gráficos dinámicos y reportes reproducibles, alineándose con las demandas actuales de trabajo.

Para el desarrollo de esta práctica resulta imprescindible contar con una instalación del ecosistema R & RStudio, la cual se puede obtener de forma gratuita en su web oficial https://posit.co/download/rstudio-desktop/.

2. ANTECEDENTES

La gran capacidad y versatilidad de R, en combinación con el hecho de tratarse de un lenguaje de código abierto, ha propiciado su explotación desde una miríada de aplicaciones, tanto fuera como dentro del aula, si bien estos mismos rasgos lo ha hecho tradicionalmente más adecuado para entornos de educación superior. Buena muestra de ello lo dan Cuevas et al. (2019) o Chan et al. (2020), donde se exploran sus ventajas en comparación con otras similares y se contrasta el grado de implantación de la herramienta a nivel universitario. No obstante, trabajos como el de Ruiz et al. (2009) o Martínez et al. (2022) ofrecen un primer acercamiento a las características y funcionalidades de R que servirían para introducir al profesorado eventualmente más novel en habilidades de programación o visualización de datos.

De entre una gran variedad de recursos en línea, el libro de Estrellado et al. (2020) constituye un compendio excelente que abarca no sólo al lenguaje, sino también la forma de abordar el estudio de Ciencia de Datos en educación. Un paso más allá lo proporciona el recurso web de Bean (2021) que, si bien extiende la complejidad matemática más allá del objetivo que aquí se persigue, expone gran variedad de ejemplos y casos de estudio que podrían inspirar la creación de actividades y otras formas más heterodoxas de tratar los contenidos curriculares en clase.

Las aplicaciones de R en Educación Secundaria han sido menos reportadas, si bien hay ejemplos que dan buena muestra de las posibilidades de este recurso a pie de aula. Al tratarse de un lenguaje no ideado especialmente para la enseñanza, su empleo requiere de cierta elaboración por parte del docente para hacerlo más amigable. Aun así, estudios como los de Calahorra et al. (2019) y de Briz et al. (2018), que exponen la experiencia de uso con adolescentes a nivel de E.S.O. y Bachillerato, confirman el potencial de esta herramienta en la enseñanza de las matemáticas a nivel preuniversitario, así como hacia la integración del pensamiento computacional en el currículum de matemáticas.

3. METODOLOGÍA

A modo de ejemplo ilustrativo del uso de R, seguidamente se exponen una serie de actividades de visualización y análisis de datos susceptibles de ser empleadas en clase. El nivel de dificultad es creciente, y con ello se persigue un acercamiento tanto a las particularidades del lenguaje R como a la exploración de datos en general.

La base de datos que se va a emplear es la misma que en el resto de capítulos. Se trata de un documento en formato CSV (*Comma Separated Values*), que se leerá y formateará correspondientemente utilizando este software libre. Este capítulo propone tres grandes retos:

1. El primero se corresponde a la lectura y preprocesamiento del conjunto de datos. Para ello, es necesario alterar la gran cantidad de variables categóricas que incluye este dataset. Se realizarán los siguientes cambios:
   1.1. Para las variables dicotómicas 1=*"Sí"* , 6=*"No"* y 9=*"NS/NR"*.
   1.2. Para la variable *INTEFOR*, 1=*"Fija"* , 2=*"Solo móvil"* y 9=*"NS/NR"*.
   1.3. Para la variable *SEXO*, 1=*"Hombre"* y 6=*"Mujer"*.
2. Con el conjunto de datos adecuadamente procesado y cargado en memoria, nos encontramos en condiciones de construir cualquier gráfico deseado. Como se ha indicado anteriormente, se utilizará la librería *ggplot2*. Las tareas correspondientes a este aportado consistirán en:
   2.1. Construir un gráfico de barras que refleje el número de participantes de cada comunidad autonómica. Guarda esta imagen en una carpeta "img" ubicada en la misma ruta de los datos, con el nombre "participantes_CA.png" y con un ancho de 8cm y un alto de 6cm.
   2.2. Dibujar otro gráfico de barras con las edades de los participantes. Guarda esta imagen en una carpeta "img" ubicada en la misma ruta de los datos, con el nombre "participantes_EDAD.png" y con un ancho de 8cm y un alto de 6cm.
   2.3. Pintar un gráfico de sectores con la participación de cada sexo en la encuesta. Guarda esta imagen en una carpeta "img" ubicada en la misma ruta de los datos, con el nombre "sectores_SEXO.png" y con un ancho de 8cm y alto de 6cm.
   2.4. Construir un gráfico de barras que indique para cada sexo si se ha usado internet en los últimos tres meses. De forma análoga se realizará para la edad. Pista: utilizar el argumento en la función *fill*. Guarda estas imágenes en una carpeta "img" ubicada en la misma ruta de los datos, con el nombre "INT_segun_SEXO.png" y "INT_segun_EDAD.png", y con un ancho de 8cm y un alto de 6cm.
   2.5. Prueba a personalizar los gráficos anteriores tanto como desees. Por ejemplo, se puede cambiar:
      - el color de las barras/sectores.
      - el título del gráfico y de los ejes.
      - el alineamiento del título (puedes centrarlo).
      - el tamaño de letra del título y de los ejes.
      - el color de fondo de la imagen. Pista: en ggsave utiliza el parámetro "bg".
3. Finalmente, se va aumentar levemente la complejidad solicitando cálculos estadísticos que implican algunos filtros o selecciones. Más concretamente se plantean dos actividades:
   3.1. Calcular una tabla de contingencia entre las variables TELEF1 y TELEF2. Interpreta los resultados.

3.2. Calcula el porcentaje de personas con 15 años que no utilizan internet. Para ello, identifica las variables relevantes, realiza el filtro correspondiente y, por último, calcula el porcentaje.

4. RESULTADOS

Esta sección expone detalladamente los pasos a seguir para completar los retos planteados. Se indicarán los pormenores del proceso a la vez que los resultados que se van obteniendo. Comenzamos abriendo un nuevo script dentro de RStudio: *File* > *New File* > *R Script* y lo guardamos con el nombre que deseemos. La presentación de esta herramienta consiste en 4 paneles correspondientes al código fuente, la consola, el entorno y los otros recursos (como archivos, paquetes o menú de ayuda entre otros).

Las subsecciones siguientes abordan cada una de las tres grandes tareas propuestas en la Metodología.

4.1. Lectura y preprocesamiento de datos

Todos los lenguajes de programación tienen una guía de buenas prácticas con sugerencias y consejos que evitan errores y permiten realizar mantenimientos más eficientes. El primer consejo en el uso de R es siempre incluir una línea que borre cualquier información cargada en el sistema previamente mediante la ejecución de rm(list=ls()). Más adelante, se suelen incluir las librerías necesarias para poder ejecutar el software (si no dispone de alguna puede instalarlas ejecutando install.package("nombre_libreria")). En estos ejemplos se van a emplear las librerías *dplyr* para el procesamiento de datos y *ggplot2* para la creación de gráficas. A continuación, se insertan la(s) ruta(s) de trabajo (cuidado con los ordenadores con sistema operativo Windows ya que copian las barras de forma invertida). Por último, se carga el conjunto de datos mediante read.csv, función contenida en el software base de R. Obsérvese que se indica que el elemento separador es el carácter ";" y la codificación "latin1".

```
rm(list=ls())

library(dplyr)
library(ggplot2)

# Rellenar ruta adecuadamente
ruta <- ""

# Lectura del conjunto de datos
data <- read.csv(paste0(ruta, "datospaper.csv"), sep=";", encoding = "latin1")
```

Tras esta lectura del fichero, hemos de formatear aquellas variables que vamos utilizar para el análisis:

```
T_SINO_lev <- c(1, 6, 9)
T_SINO_lab <- c("Sí", "No", "NS/NR")

# Cambiamos la variable categóricas por su factor correspondiente
data <- data %>%
  mutate(SEXO = factor(SEXO, levels = c(1, 6), labels = c("Hombre", "Mujer")),
         ORD = factor(ORD, levels = T_SINO_lev, labels = T_SINO_lab),
         TABLET = factor(TABLET, levels = T_SINO_lev, labels = T_SINO_lab),
         TELEF1 = factor(TELEF1, levels = T_SINO_lev, labels = T_SINO_lab),
         TELEF2 = factor(TELEF2, levels = T_SINO_lev, labels = T_SINO_lab),
         VIV_INTER = factor(VIV_INTER, levels = T_SINO_lev,
```

```
                              labels = T_SINO_lab),
          INTEFOR = factor(INTEFOR, levels = c(1, 2, 9),
                           labels = c("Fija", "Solo móvil", "NS/NR")),
          PC = factor(PC, levels = T_SINO_lev, labels = T_SINO_lab),
          INT = factor(INT, levels = T_SINO_lev, labels = T_SINO_lab),
          MOVIL = factor(MOVIL, levels = T_SINO_lev, labels = T_SINO_lab)
  )
```

4.2. Visualización de gráficas

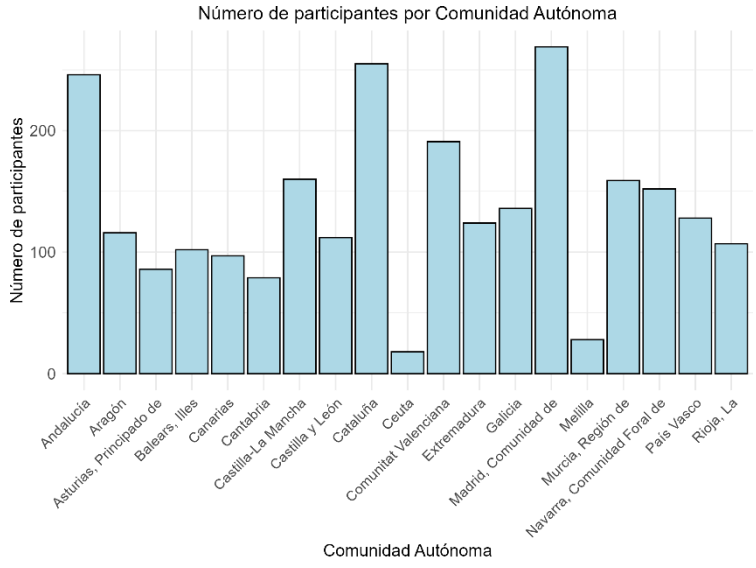
El lenguaje de programación R tiene una gran potencia para la visualización de datos gracias a la librería ggplot. Ésta, permite crear gráficos altamente personalizables de manera intuitiva. Para empezar, se debe usar la función ggplot() donde se especifica el dataframe (por ejemplo, data) y, dentro de aes(), se definen las variables estéticas, como los ejes.

Para la consecución de la gráfica pedida en 2.1, indicaremos que x es la Comunidad y utilizaremos la función geom_bar. La función theme_minimal() le da un aspecto más minimalista eliminando los ejes. La función labs permite establecer el valor del título y de los ejes. Con theme es posible personalizar aspectos estéticos del gráfico. Por último, se almacena con ggsave, indicando la ruta (se añade "/img" y previamente se crea la carpeta manualmente) y el ancho y alto (con "width" y "height" respectivamente).

```
# Graficar el número de participantes por Comunidad Autónoma
ggplot(data, aes(x = Comunidad)) +
  geom_bar(fill = "lightblue ", color = "black ") +
  theme_minimal() +
  labs(title = "Número de participantes por Comunidad Autónoma",
       x = "Comunidad Autónoma", y = "Número de participantes") +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1, size = 11),
    axis.text.y = element_text(size = 12),
    axis.title.x = element_text(size = 13),
    axis.title.y = element_text(size = 13),
    plot.title = element_text(hjust = 0.5, size = 14)
  )
ggsave(paste0(ruta, "participantes_CA.png"), width = 8, height = 6)
```

Figura 1

Gráfico de barras del número de participantes según la Comunidad Autónoma.



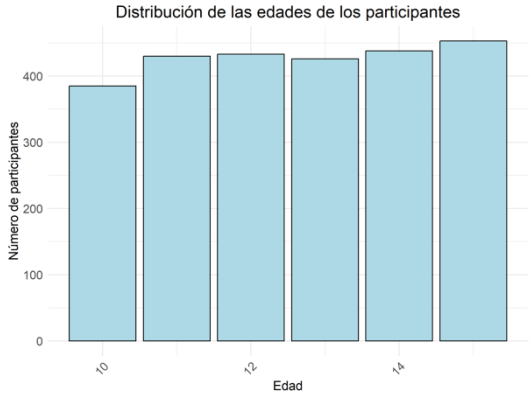
En la Figura 1 se puede observar que las comunidades con menor participación son Ceuta y Melilla y la de mayor participación es la Comunidad de Madrid.

El procedimiento es idéntico para la edad, con la salvedad de cambiar el atributo correspondiente.

```
ggplot(data, aes(x = EDAD)) +
  geom_bar(fill = "lightblue", color = "black") +
  theme_minimal() +
  labs(title = "Distribución de las edades de los participantes",
       x = "Edad", y = " Número de participantes")
ggsave(paste0(ruta, "participantes_EDAD.png"), width = 8, height = 6)
```

Figura 2

Gráfico de barras con la edad de los participantes de la encuesta.

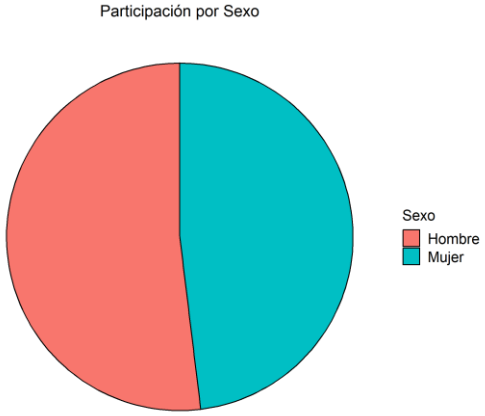


Se observa que la distribución por edades es uniforme. Para el gráfico de sectores es necesario pasarlo a coordenadas polares, utilizando la función coord_polar(theta = "y"). Además, eliminamos todos los ejes con theme_void().

```
# Graficar el gráfico de sectores
ggplot(data, aes(x = "", fill = factor(SEXO))) +
  geom_bar(width = 1, color = "black") +
  coord_polar(theta = "y") +
  labs(title = "Participación por Sexo", fill = "Sexo") +
  theme_void()  # Elimina los ejes
ggsave(paste0(ruta, "sectores_SEXO.png"), width = 8, height = 6)
```

Figura 3

Gráfico de sectores que recoge la participación de cada sexo en la encuesta.



El gráfico de sectores de la Figura 3 muestra que existe una ligera mayor participación de hombres que de mujeres en la encuesta.

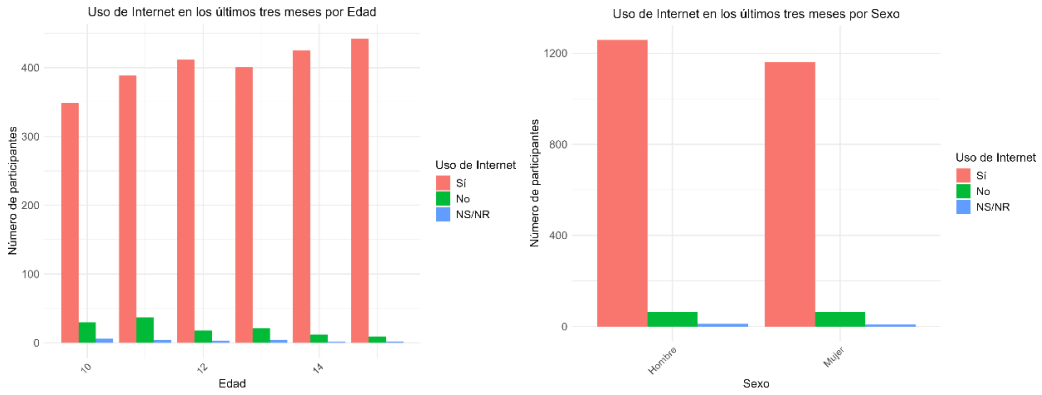
Para graficar una variable en función de otra se emplean parámetros de estética, como “fill”, “color” o “size” entre otros. Además, se ha empleado “dodge” para la posición de las barras de modo que una se ubique al lado de la otra.

```
# Graficar el uso de Internet por sexo
ggplot(data, aes(x = factor(SEXO), fill = factor(INT))) +
  geom_bar(position = "dodge") +
  labs(title = "Uso de Internet en los últimos tres meses por Sexo",
       x = "Sexo", y = "Número de participantes", fill = "Uso de Internet") +
  theme_minimal()
ggsave(paste0(ruta, "INT_segun_SEXO.png"), width = 8, height = 6)

# Graficar el uso de Internet por edad
ggplot(data, aes(x = EDAD, fill = factor(INT))) +
  geom_bar(position = "dodge") +
  labs(title = "Uso de Internet en los últimos tres meses por Edad",
       x = "Edad", y = "Número de participantes", fill = "Uso de Internet") +
  theme_minimal()
ggsave(paste0(ruta, "INT_segun_EDAD.png"), width = 8, height = 6)
```

Figura 4

Gráfico de barras que expresa el uso de internet en las familias españolas según edad (izquierda) y según sexo (derecha).



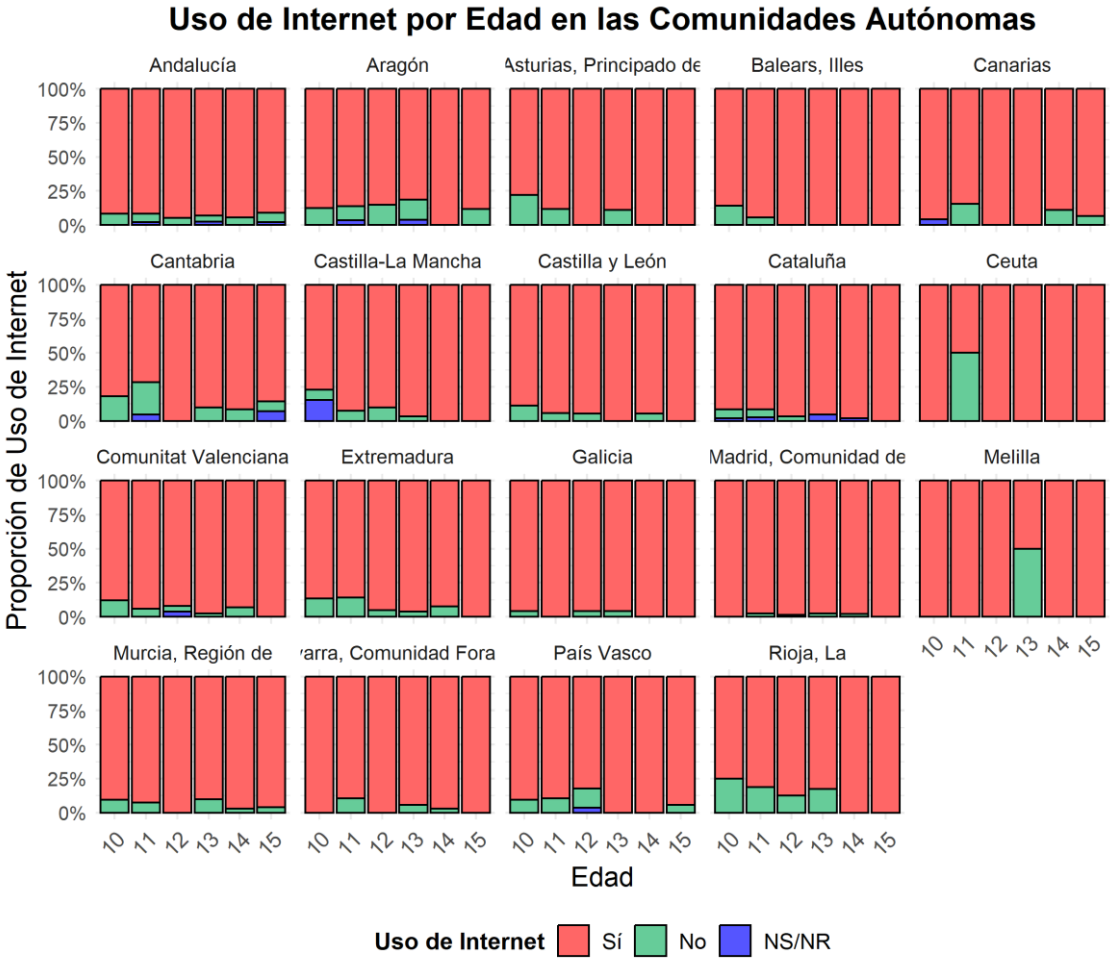
Por último, se puede personalizar al gusto la gráfica. Encuentre a continuación el código de un ejemplo:

```
ggplot(data, aes(x = EDAD)) +
  geom_bar(fill = "orange", color = "darkred") +
  theme_minimal() +
  labs(title = "Distribución de las Edades de los Participantes",
       subtitle = "Encuesta de Uso de Internet",
       x = "Edad de los Participantes", y = "Número de participantes") +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1, size = 11),
    axis.text.y = element_text(size = 12),
    axis.title.x = element_text(size = 13),
    axis.title.y = element_text(size = 13),
    plot.title = element_text(size = 14)
  )
```

A modo adicional, se va a elaborar un gráfico avanzado que combina múltiples variables para ilustrar el potencial y la versatilidad de R en el ámbito de las ciencias de datos. Este gráfico demuestra cómo R facilita el análisis y la visualización de datos complejos de forma efectiva.

Figura 5

Gráfico de barras que refleja el porcentaje de uso de internet en las familias españolas según la edad, facetando la Comunidad Autónoma.



A simple golpe de vista la gráfica expone que la inmensa mayoría de jóvenes de España sí usan internet.

4.3. Cálculos estadísticos

Este lenguaje nació en el contexto de la Estadística por lo que otro de sus puntos fuertes es contar con funcionalidades y herramientas que permitan hallar casi cualquier cálculo deseado.

El ejercicio 3.1 requiere del empleo de "table", que tabula una variable en función de la otra. El uso de "prop.table" realiza el mismo cálculo pero en porcentaje.

```
# Calcular la tabla de contingencia
tabla_contingencia <- table(data$TELEF1, data$TELEF2)
print(tabla_contingencia)

# Mostrar la tabla en porcentaje sobre el total
tabla_contingencia_prop <- prop.table(tabla_contingencia) * 100
print(tabla_contingencia_prop)
```

La tabla resultante de la ejecución anterior es la siguiente:

Tabla 1

Tabla de contingencia de las variables TELEF1 y TELEF2.

		TELEF2		
		Sí	No	NS/NR
TELEF1	Sí	1542	2	0
	No	1011	0	0
	NS/NR	10	0	0

La interpretación de esta tabla permite identificar varias ideas:

- Casi todos los hogares cuentan con móviles.
- El uso de teléfono fijo se está perdiendo (un 40% de hogares no tienen).
- Cuando el hogar no cuenta con teléfono fijo es porque hay al menos un móvil.

Para resolver la tarea 3.2, se construye un nuevo conjunto de datos filtrando (función "filter") a que la edad sean 15 años. Se cuenta el número de participantes con esa edad que son 453. Se aplica un nuevo filtro que indica que no se usa internet, resultando en 9 participantes. Finalmente, se halla el porcentaje con el cálculo 9/453*100, obteniendo un 1,99% de participantes con 15 años que no han utilizado internet en los últimos tres meses.

```
# Filtrar las personas que tienen 15 años
personas_15 <- data %>%
  filter(EDAD == 15)

total_personas_15 <- nrow(personas_15)

# Calcular el número de personas de 15 años que no utilizan internet
personas_15_sin_internet <- personas_15 %>%
  filter(INT == "No")

total_personas_15_sin_int <- nrow(personas_15_sin_internet)

# Calcular el porcentaje de personas de 15 años que no utilizan internet
porcentaje_sin_internet <- (total_personas_15_sin_int / total_personas_15) * 100

# Mostrar el porcentaje
print(paste("El porcentaje de personas de 15 años que no utilizan internet es:",
            round(porcentaje_sin_internet, 2), "%"))
```

5. CONCLUSIONES

La gran difusión que el lenguaje R ha alcanzado en la exploración y visualización de datos da buena muestra de sus posibilidades en cuanto a su uso a pie de aula. Especialmente en la enseñanza de la Estadística, esta herramienta gratuita y de código abierto permite a estudiantes y profesores acceder a potentes funcionalidades sin coste alguno, fomentando además un enfoque colaborativo y de código reproducible.

Las actividades aquí presentadas persiguen la adquisición de nociones fundamentales sobre manipulación y visualización de datos en el contexto educativo, haciendo uso de R. A través de estas tareas, los estudiantes no solo se familiarizan con herramientas útiles para su futura inserción laboral, sino que también desarrollan habilidades de pensamiento crítico y computacional. El análisis de un conjunto de datos sobre el uso de la tecnología e internet en diferentes regiones de España enriquece el aprendizaje, ofreciendo una aplicación práctica y

relevante para el alumnado, hacia una experiencia educativa más rica, práctica y adaptable a la diversidad de disciplinas existente.

REFERENCIAS BIBLIOGRÁFICAS

Bean, J. (2021). Using R for Educational Research.

Briz, A. y Serrano, Á. (2018). Aprendizaje de las matemáticas a través del lenguaje de programación R en Educación Secundaria. *Educación Matemática*, *30*(1), 133-162.

Calahorra, J., Aguilar, T., Diciembre, S. y Sanchiz, D. (2019). Aproximación didáctica a las matemáticas a través de la programación en R. *Números*, *102*, 161-184.

Chan, D. y Galli M.G. (2020). Aplicación de técnicas estadísticas multivariadas con el lenguaje de programación R en investigaciones educativas del nivel superior. *Revista Argentina de Educación Superior*, *20*, 123-136.

Cuevas, H., Solís, C. y Silva, I. (2019). Programación computacional y análisis de datos en educación estadística. *Areté: Revista Digital del Doctorado en Educación de la Universidad Central de Venezuela*, *5*(9), 11-27.

Estrellado, R.A., Freer, E.A., Motsipak, J., Rosenberg, J.M. y Velásquez, I.C. (2020). *Data science in education using R*. Londres, ed. Routledge.

Martínez J., Joel F., Herrera, J.B. y Paredes, F.A. El programa R: Una estrategia inicial para su entendimiento y aprendizaje. *Revista Digital Universitaria*, *23*(4) (http://doi.org/10.22201/cuaieed.16076079e.2022.23.4.4)

Puig, L. (1997). Análisis fenomenológico. En L. Rico (Coord.) *La educación matemática en la enseñanza secundaria*, 61-94. Horsori.

Ramírez, R., Brizuela, B. y Blanton, M. (2020). Kindergarten and first-grade students' understandings and representations of arithmetic properties. *Early Childhood Education Journal*, *50*, 1-12.

Ruiz, M. y López, E. (2009). El entorno estadístico R: ventajas de su uso en la docencia y la investigación. *Revista Española de Pedagogía*, *243*, 255-274.