PEC1 Otoño 2025

UOC

En esta actividad no está permitido el uso de herramientas de inteligencia artificial. En el plan docente y en la web sobre integridad académica y plagio de la UOC encontraréis información sobre qué se considera conducta irregular en la evaluación y las consecuencias que puede tener.

Esta PEC se basará en los datos de las películas de Netflix estrenadas entre 1942 y 2019 (los datos se han obtenido de la web *ExcelDemy*). Hay incluida información de las siguientes variables:

1. *Name* = variable cualitativa que indica el título de la película
2. *Year* = año del estreno
3. *Age_Rating* = variable cualitativa que indica la clasificación de la película por edad
4. *Duration* = duración de la película en minutos
5. *Category* = variable cualitativa que indica la categoría de la película
6. *IMDb_Rating* = puntuación de la película (sobre 10)

Observación: las categorías de la variable *Age_Rating* son

1. *PG (Parental Guidance).* Se sugiere la supervisión de los padres; algunos materiales podrían no ser aptos para niños pequeños.
2. *PG-13 (Parents Strongly Cautioned).* Se advierte a los padres que algunos materiales podrían ser inapropiados para menores de 13 años.
3. *R (Restricted).* Los menores de 17 años necesitan la compañía de un padre o tutor adulto para ver la película.

Para importar los datos podéis usar las siguientes instrucciones:

```
library(readxl)
datos <- read_excel("Netflix-Movies-Sample-Data.xlsx", skip = 5)
```

Os puede ser útil consultar el siguiente material del reto 1:

1. El entorno estadístico R. Estructura, lenguaje y sintaxis
2. Análisis de datos y estadística descriptiva con R
3. Actividades Resueltas del Reto 1 (Estadística Descriptiva)

Hay que entregar la práctica en formato ".pdf" en esta misma tarea.

NOMBRE: José Carlos López Henestrosa

PEC1

Una vez importados los datos...

Pregunta 1 (40%)

1.1 Ordenad la base de datos según el orden decreciente de la variable *IMDb_Rating* y mostrad solo las 3 primeras filas de esta base de datos ordenada. (10%)

```
# 1) Cargar datos
library(readxl)
df <- read_excel("Netflix-Movies-Sample-Data.xlsx", skip = 5)

# 2) Ordenar en orden decreciente y mostrar solo las 3 primeras filas
ord <- order(df$IMDb_Rating, decreasing = TRUE)
top3 <- df[ord, ][1:3, ]

print(top3)
```

```
## # A tibble: 3 x 6
##   Name                      Year Age_Rating Duration Category     IMDb_Rating
##   <chr>                    <dbl> <chr>         <dbl> <chr>              <dbl>
## 1 The Shawshank Redemption  1994 R               142 Drama                9.3
## 2 The Godfather             1972 R               175 Crime/Drama          9.2
## 3 The Dark Knight           2008 PG-13           152 Action/Crime         9
```

Dad el resumen numérico (mínimo, Q1, mediana, media, Q3 y máximo), la varianza y la desviación estándar de la variable *IMDb_Rating* (10%).

```
# 1) Cargar datos
library(readxl)
df <- read_excel("Netflix-Movies-Sample-Data.xlsx", skip = 5)

# 2) Resumen numérico
s <- summary(df$IMDb_Rating)  # mínimo, Q1, mediana, media, Q3 y máximo

# 3) Tabla con todas las medidas
table <- as.data.frame(t(c(
  Min     = s["Min."],
  Q1      = s["1st Qu."],
  Mediana = s["Median"],
  Media   = s["Mean"],
  Q3      = s["3rd Qu."],
```

```
  Max      = s["Max."],
  Var      = var(df$IMDb_Rating),
  sd       = sd(df$IMDb_Rating)
)), check.names = FALSE)

colnames(table) <- c(
  "Min",
  "Q1",
  "Mediana",
  "Media",
  "Q3",
  "Max",
  "Varianza",
  "Des. estándar"
)

# 4) Mostrar tabla
print(table)
```

```
##   Min    Q1 Mediana Media  Q3 Max Varianza Des. estándar
## 1 7.3 7.925    8.35 8.262 8.6 9.3 0.247302     0.4972947
```

V

1.2 Dad el resumen numérico de la variable *IMDb_Rating*, pero solo cuando la variable *Age_Rating* vale *R*. Comentad los resultados obtenidos (20%).

```
# 1) Cargar datos
library(readxl)
df <- read_excel("Netflix-Movies-Sample-Data.xlsx", skip = 5)

# 2) Filtrar por `Age_Rating == "R"`
x <- df$IMDb_Rating[df$Age_Rating == "R"]

# 3) Resumen numérico
summary(x)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   7.300   8.000   8.400   8.297   8.600   9.300
```

V

En conjunto, las películas con *Age_Rating* = R ($n = 31$) tienen valoraciones altas y bastante homogéneas: la mitad central de títulos se concentra entre 8.0 y 8.6; el rango total es de 2.0 puntos (entre 7.3 y 9.3), y la media es ligeramente inferior a la mediana. Esto sugiere una ligera cola hacia valores más bajos, sin extremos problemáticos, lo que se traduce en una variabilidad moderada y ausencia de puntuaciones anómalas.

V

Pregunta 2 (10%)

Dad el valor mínimo de la variable *IMDb_Rating* junto con las variables *Name* y *Category* donde se da este valor mínimo.

```r
# 1) Cargar datos
library(readxl)
df <- read_excel("Netflix-Movies-Sample-Data.xlsx", skip = 5)

# 2) Valor mínimo de la variable `IMDb_Rating`
min_val <- min(df$IMDb_Rating)

# 3) Añade las variables `Name` y `Category` donde se da el valor mínimo
result <- df[which(df$IMDb_Rating == min_val),
             c("Name", "Category", "IMDb_Rating")]

# 4) Mostrar resultado
print(result)
```

```
## # A tibble: 2 x 3
##   Name               Category         IMDb_Rating
##   <chr>              <chr>                  <dbl>
## 1 The Shape of Water Adventure/Drama          7.3
## 2 Black Panther      Action/Adventure         7.3
```

V

Pregunta 3 (30%)

Dad la tabla de frecuencias absolutas de la variable *Age_Rating*, y otra tabla con los porcentajes de los diferentes niveles de esta misma variable *Age_Rating* (podéis usar la instrucción *prop.table*). Haced el gráfico adecuado de las frecuencias o de los porcentajes. Comentad los resultados obtenidos.

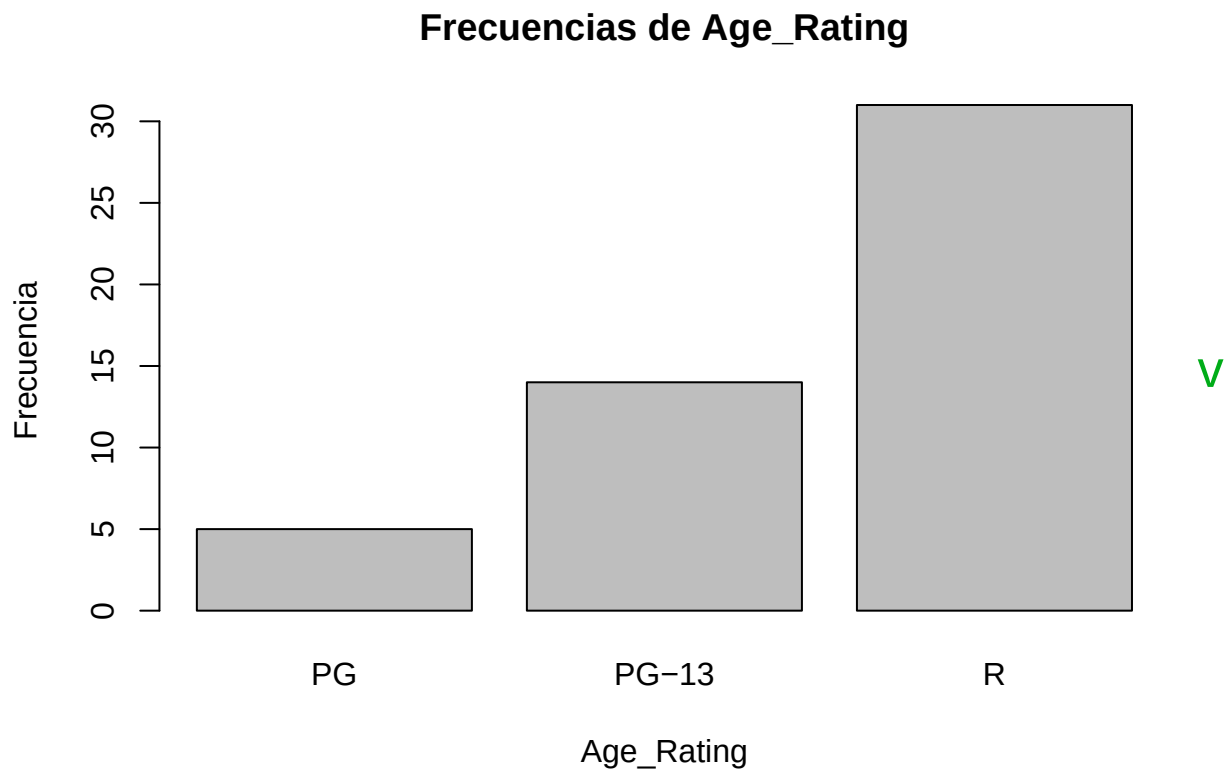
```
# 1) Cargar datos
library(readxl)
df <- read_excel("Netflix-Movies-Sample-Data.xlsx", skip = 5)

# 2) Frecuencias absolutas
tab_abs <- table(df$Age_Rating)
tab_abs
```

```
##
##    PG PG-13     R
##     5    14    31
```

V

```
barplot(tab_abs,
        main = "Frecuencias de Age_Rating",
        xlab = "Age_Rating",
        ylab = "Frecuencia")
```



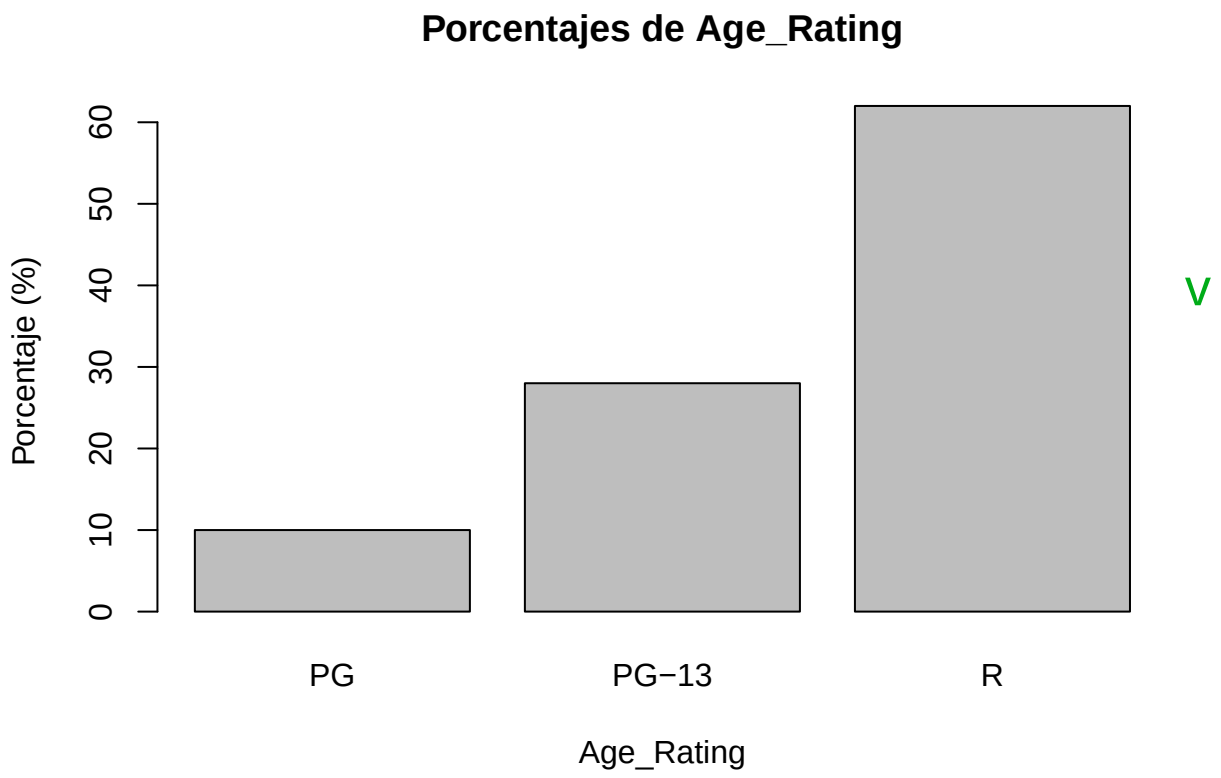
V

```r
# 3) Porcentajes
tab_pct <- prop.table(tab_abs) * 100
tab_pct_with_pct <- sprintf("%.0f%%", tab_pct) # with % for printing
names(tab_pct_with_pct) <- c("PG","PG-13","R")
tab_pct_with_pct
```

```
##    PG PG-13     R
## "10%" "28%" "62%"
```

V

```r
barplot(tab_pct,
        main = "Porcentajes de Age_Rating",
        xlab = "Age_Rating",
        ylab = "Porcentaje (%)")
```



V

En la muestra ($n = 50$) la distribución es claramente desigual. Podemos apreciar que R concentra la mayor parte de los casos con 31 títulos (62%), seguida de PG-13 con 14 (28%) y, a mucha distancia, PG con 5 (10%). El diagrama de barras de frecuencias absolutas muestra la dominancia de R en términos de recuentos, mientras que el de porcentajes replica el mismo patrón expresado sobre el total, lo que facilita la comparación relativa entre categorías. Esta asimetría sugiere que el catálogo analizado está sesgado hacia contenidos para público adulto, por lo que conviene tenerlo en cuenta al comparar otras variables por clasificación de edad, ya que los resultados para PG pueden ser más inestables debido a su reducido tamaño y las

V

conclusiones se verán principalmente influenciadas por el grupo R.

V

Pregunta 4 (20%)

Haced los boxplots de la variable *Duration* estratificando por la variable *Age_Rating*. Comentad el resultado obtenido.

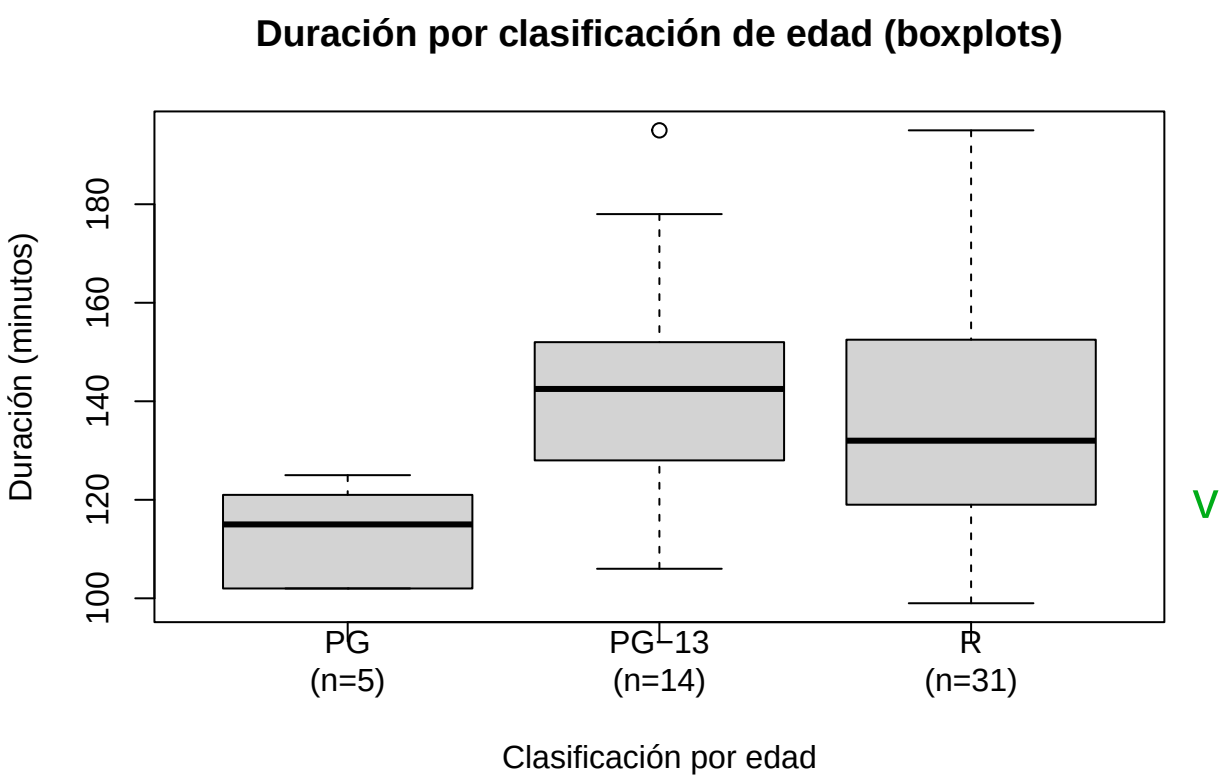
```r
# 1) Cargar datos
library(readxl)
df <- read_excel("Netflix-Movies-Sample-Data.xlsx", skip = 5)

# 2) Seleccionar columnas de interés
df <- df[, c("Duration", "Age_Rating")]

# 3) Asegurar tipo de dato correcto
df$Age_Rating <- as.factor(df$Age_Rating)
df$Duration   <- as.numeric(df$Duration)

# 4) Estratificación con el número de individuos por cada muestra
n_por_grupo <- table(df$Age_Rating)[levels(df$Age_Rating)]
nombres_ejes <- paste0(levels(df$Age_Rating), "\n(n=", n_por_grupo, ")")

# 5) Mostrar boxplots
boxplot(Duration ~ Age_Rating,
        data = df,
        main  = "Duración por clasificación de edad (boxplots)",
        xlab  = "Clasificación por edad",
        ylab  = "Duración (minutos)",
        names = nombres_ejes)
```



El diagrama compara la duración de las películas según su clasificación por edad. El grupo PG ($n = 5$) tiene las películas más cortas. Su mediana ronda los 110 minutos y la dispersión es pequeña (la caja es estrecha, aprox. de 105 a 120), lo que indica poca variabilidad. Además, el tamaño muestral es muy pequeño, así que cualquier conclusión que podamos sacar es frágil.

En PG-13 ($n = 14$) la mediana sube claramente, alrededor de 140 minutos, y el rango intercuartílico es mayor que el de PG. Aparece un valor atípico alto cercano a 190 minutos y el bigote superior más largo sugiere ligera asimetría a la derecha.

En R ($n = 31$) la mediana cae respecto a PG-13 y queda en torno a 125–130 minutos, pero es el grupo con mayor variabilidad, ya que la caja es ancha (aprox. de 120 a 150) y el rango total se extiende desde algo menos de 100 hasta casi 195 minutos.

Como conclusión, PG-13 reúne las duraciones típicamente más largas, PG las más cortas y R muestra la mayor heterogeneidad. Las diferencias aparentes deben interpretarse teniendo en cuenta que el tamaño de las muestras no son iguales y que PG tiene muy pocos casos.