PEC1 Otoño 2025 - Solución

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 una **muestra** de datos de 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:

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. Dad el resumen numérico (mínimo, Q1, mediana, media, Q3 y máximo), la varianza y la desviación estándar de la variable *IMDb_Rating* (20%).

```
head(datos[order(datos$IMDb_Rating, decreasing = TRUE),], 3)
```

```
## # 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
```

```
summary(datos$IMDb_Rating)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   7.300   7.925   8.350   8.262   8.600   9.300
```

```
var(datos$IMDb_Rating)
```

```
## [1] 0.247302
```

```
sd(datos$IMDb_Rating)
```

```
## [1] 0.4972947
```

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

```r
summary(datos$IMDb_Rating[datos$Age_Rating == "R"])
```

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

Los valores de la media y de la mediana no son muy parecidos (8.297 y 8.400 respectivamente) lo que indicaría que nos encontramos ante una distribución ligeramente asimétrica, concretamente con cola hacia la izquierda. El 50% central de las películas tiene una puntuación entre 8.0 y 8.6, el 25% inferior tiene una puntuación menor de 8.0 y el 25% superior restante tiene una puntuación mayor de 8.6. El rango intercuartílico es pequeño, lo que significa que los datos centrales están apretados, hay poca dispersión.

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
datos[datos$IMDb_Rating == min(datos$IMDb_Rating),
      c("IMDb_Rating", "Name", "Category")]
```

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

Atención: ¡hay 2 filas!

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.

```r
table(datos$Age_Rating)
```

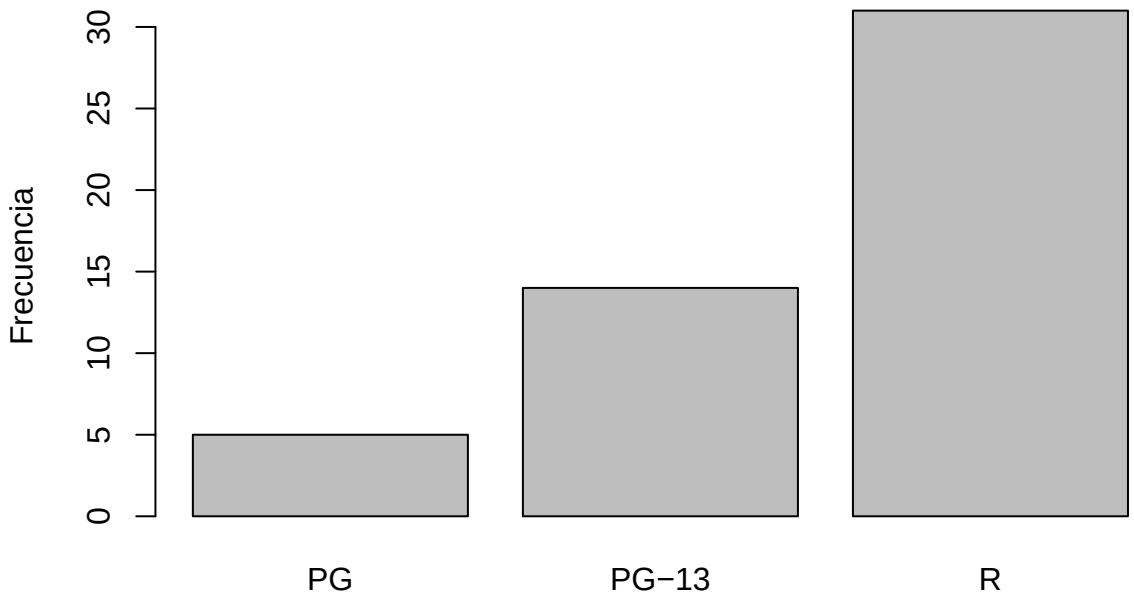
```
##
##    PG PG-13     R
##     5    14    31
```

```r
prop.table(table(datos$Age_Rating)) * 100
```

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

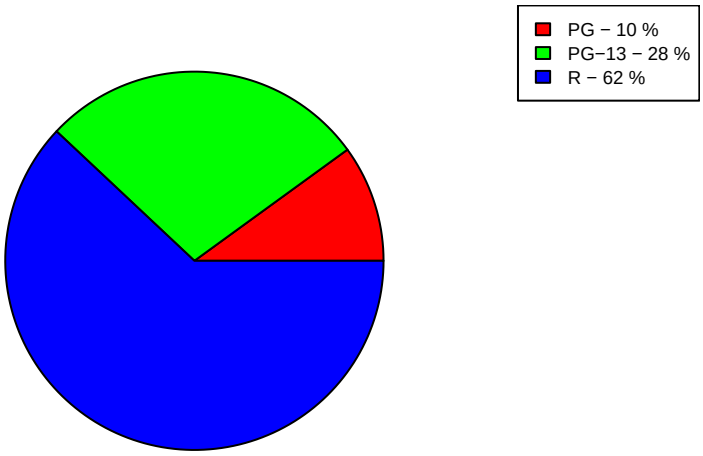
Podéis hacer este gráfico:

```r
barplot(table(datos$Age_Rating), ylab = "Frecuencia")
```



O podéis hacer este otro:

```r
valores <- table(datos$Age_Rating)
etiquetas <- names(table(datos$Age_Rating))
porcentajes <- round(prop.table(table(datos$Age_Rating))*100, 2)
etiquetas_con_porcentajes <- paste(etiquetas, "-", porcentajes, "%")
colores <- rainbow(length(valores))
pie(valores, labels = NA, col = colores)
legend("topright", legend = etiquetas_con_porcentajes, fill = colores,
       cex = 0.62)
```

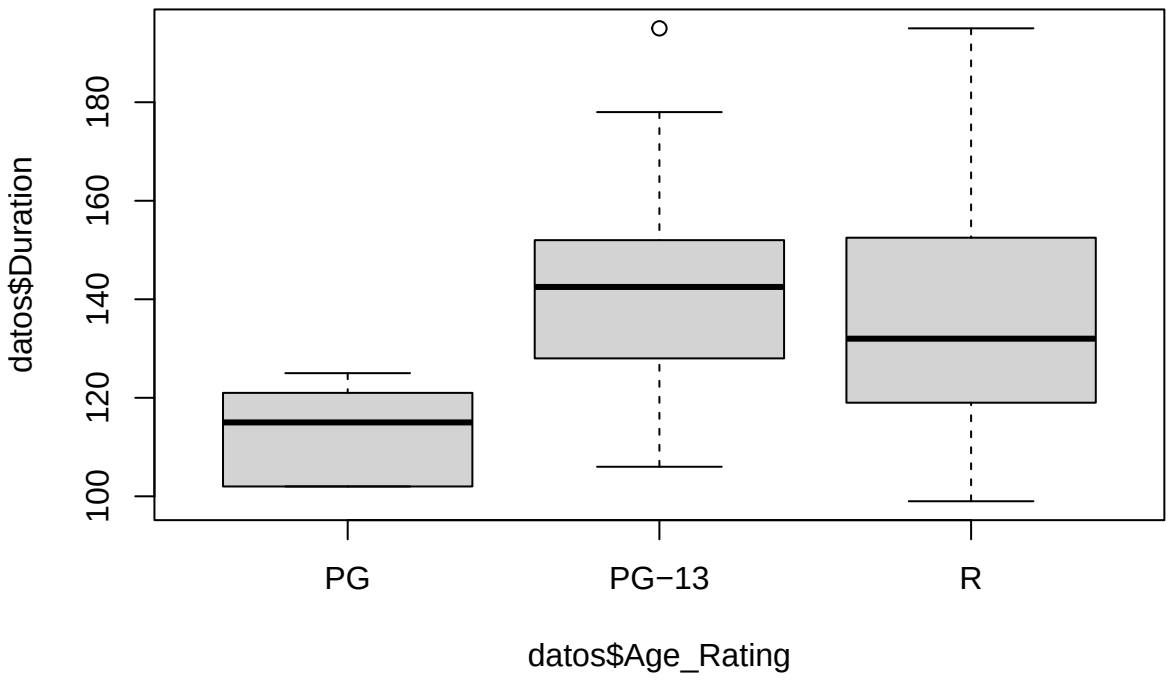


La clasificación mayoritaria es “R” (62%), seguido de “PG-13” (28%), y la clasificación minoritaria es “PG” (10%).

Pregunta-4 (20%)

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

```
boxplot(datos$Duration ~ datos$Age_Rating)
```



La distribución con menos dispersión se da en la clasificación “PG”. La distribución más simétrica se da en la clasificación “PG-13”. La duración de las películas con clasificación “PG” es claramente inferior a las otras dos categorías (“PG-13” y “R”): las películas “PG” como mucho duran aproximadamente 125 minutos; el 75% de las películas “PG-13” duran más de 130 minutos aproximadamente; y el 75% de las películas “R” duran más de 120 minutos aproximadamente. Si comparamos las medianas, tenemos de más grande a más pequeña “PG-13” (aproximadamente 140), “R” (aproximadamente 130) y “PG” (aproximadamente 115).