Netflix

Ing. José Miguel Pereira

23/1/2022

# Show de Netflixs

Se posee la información de 1000 ejemplos de shows que posee la plataforma de Netflix.

Las variables que se poseen de la misma base son:

* Puntuación de calificación de usuario (User Rating Score)
* Calificación de descripción (Rating Description)
* Información Técnica:
  + Año de lanzamiento (Release Year)
  + Título (Title)

# Tabla de contenido:

1. Librerías
2. Data
3. Manipulación de datos y resumen
4. Correlación
5. Visualización

# 1. Librerías

Se procede a importar las librerías a emplear para el procesamiento y visualización de los datos.

# Librerias   
library(dplyr)  
# Gráficos  
library(ggplot2)  
# Gráficos de correlación  
library(corrplot)  
# Resumen personalizado  
library(gtsummary)  
# cluster result visualization with network diagram  
library(igraph)  
library(vioplot)

# 2. Data

Con las librerias cargadas, se procede a cargar la base de datos.

title rating  
1 White Chicks PG-13  
2 Lucky Number Slevin R  
3 Grey's Anatomy TV-14  
4 Prison Break TV-14  
5 How I Met Your Mother TV-PG  
6 Supernatural TV-14  
 ratingLevel  
1 crude and sexual humor, language and some drug content  
2 strong violence, sexual content and adult language  
3 Parents strongly cautioned. May be unsuitable for children ages 14 and under.  
4 Parents strongly cautioned. May be unsuitable for children ages 14 and under.  
5 Parental guidance suggested. May not be suitable for all children.  
6 Parents strongly cautioned. May be unsuitable for children ages 14 and under.  
 ratingDescription release.year user.rating.score user.rating.size  
1 80 2004 82 80  
2 100 2006 NA 82  
3 90 2016 98 80  
4 90 2008 98 80  
5 70 2014 94 80  
6 90 2016 95 80

Tal como se aprecia en la descripción anterior, se tienen 7 variables:

* Título
* Rating
* Rating Level
* Description del rating
* Año de aparición
* Puntaje de rating de usuario
* Tamaño del rating del usuario.

# 3. Manipulación de datos y resumen

Vamos a observar una breve descripción de las metricas por categoría

| **Characteristic** | **N = 1,000** |
| --- | --- |
| rating |  |
| G | 138 (14%) |
| NR | 14 (1.4%) |
| PG | 170 (17%) |
| PG-13 | 15 (1.5%) |
| R | 19 (1.9%) |
| TV-14 | 234 (23%) |
| TV-G | 52 (5.2%) |
| TV-MA | 148 (15%) |
| TV-PG | 59 (5.9%) |
| TV-Y | 68 (6.8%) |
| TV-Y7 | 38 (3.8%) |
| TV-Y7-FV | 44 (4.4%) |
| UR | 1 (0.1%) |
| ratingDescription | 60 (35, 90) |
| release.year | 2,015 (2,007, 2,016) |
| user.rating.score | 88 (75, 95) |
| Unknown | 395 |
| user.rating.size |  |
| 80 | 605 (60%) |
| 81 | 7 (0.7%) |
| 82 | 388 (39%) |

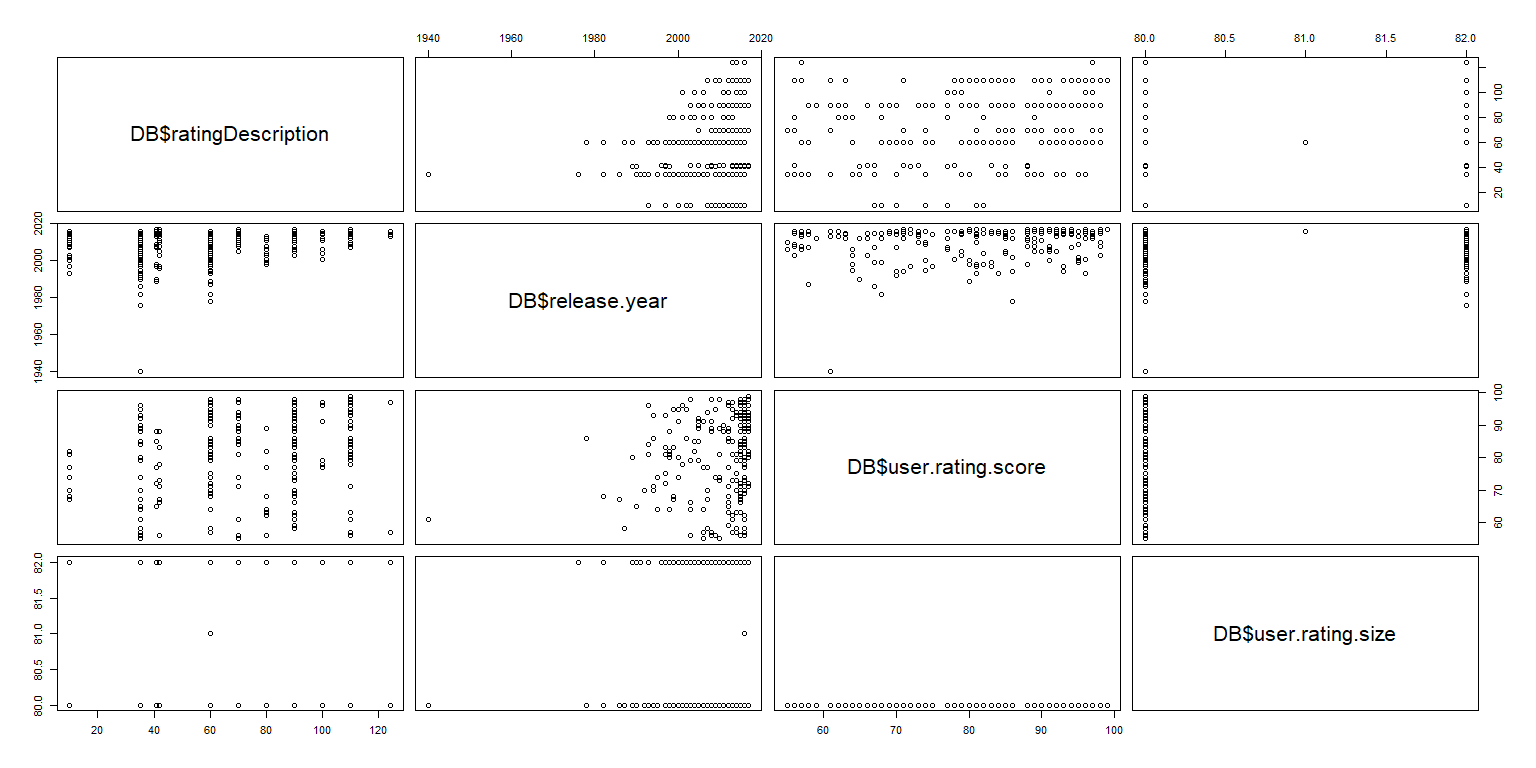
Con estos datos se puede observar más o menos como se encuentran los datos de la base de datos.

# 4. Correlación

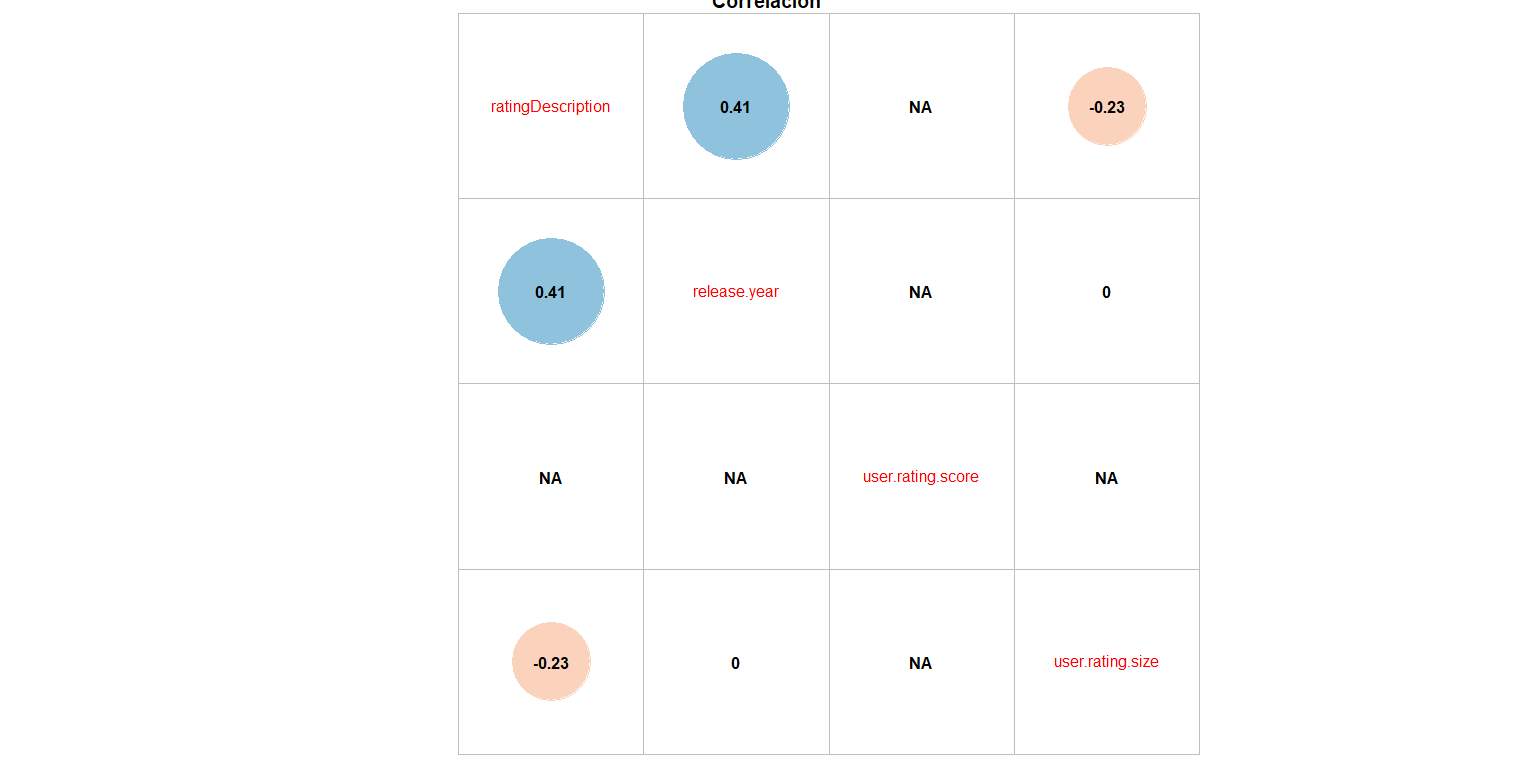
Es importante visualizar si la información que se tiene posee un grado de correlación entre las variables.

Primero se obtiene los datos de correlación de toda la tabla y el gráfico de pares.

ratingDescription release.year user.rating.score  
ratingDescription 1.0000000 0.408019919 NA  
release.year 0.4080199 1.000000000 NA  
user.rating.score NA NA 1  
user.rating.size -0.2262070 -0.003279772 NA  
 user.rating.size  
ratingDescription -0.226206952  
release.year -0.003279772  
user.rating.score NA  
user.rating.size 1.000000000



Observemos la correlación de la información dentro de un gráfico más amigable.

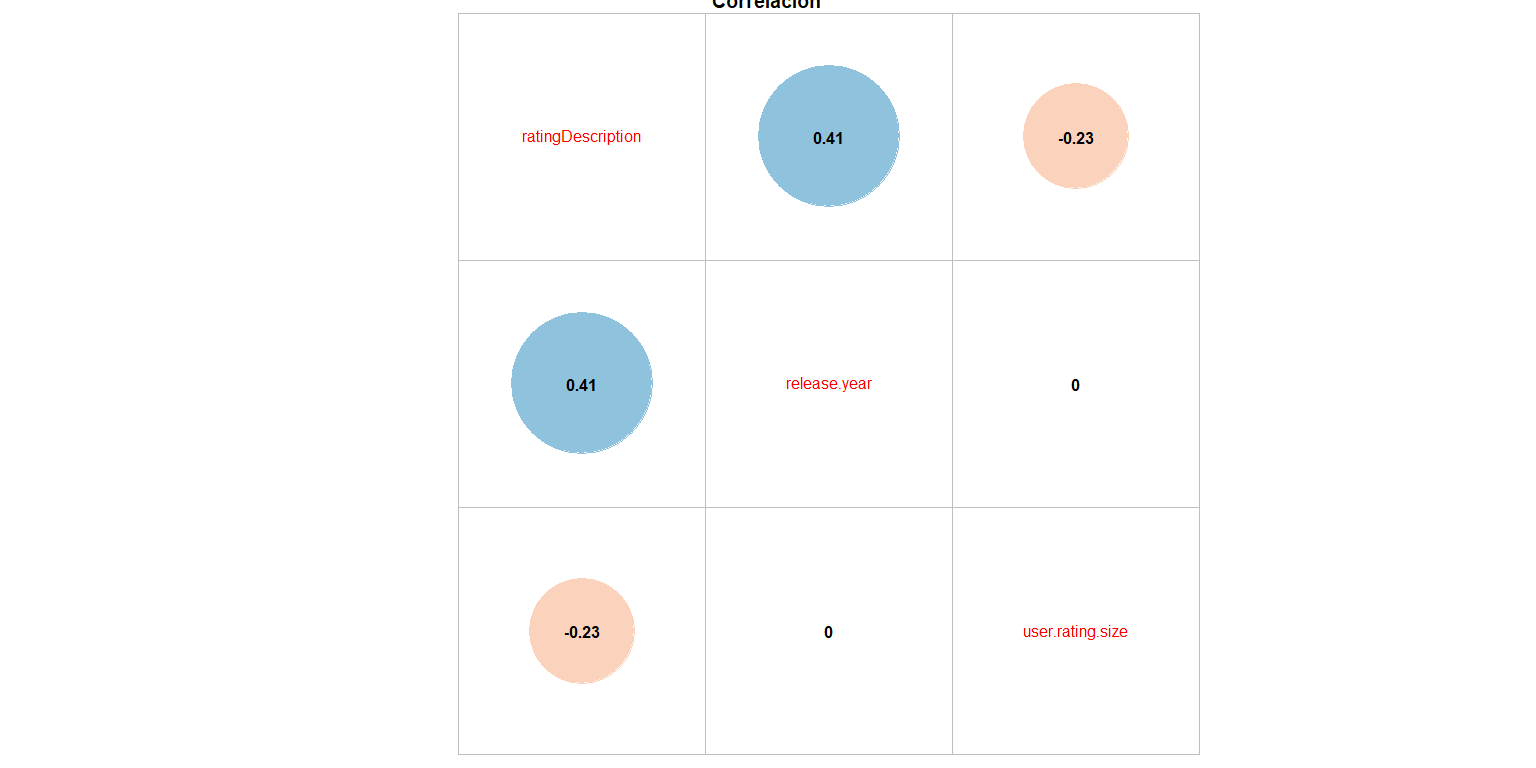


Con respecto a la variable de user.rating.score se tienen valores NA, lo que implica que se poseen variables faltantes en esa categoría.

Sin embargo, el observar los datos completos nos llevaría a tomar decisiones erróneas.

Por lo cual, se observará la correlación existente entre las variables discriminando la variable user.ratinf.score.

ratingDescription release.year user.rating.size  
ratingDescription 1.0000000 0.408019919 -0.226206952  
release.year 0.4080199 1.000000000 -0.003279772  
user.rating.size -0.2262070 -0.003279772 1.000000000

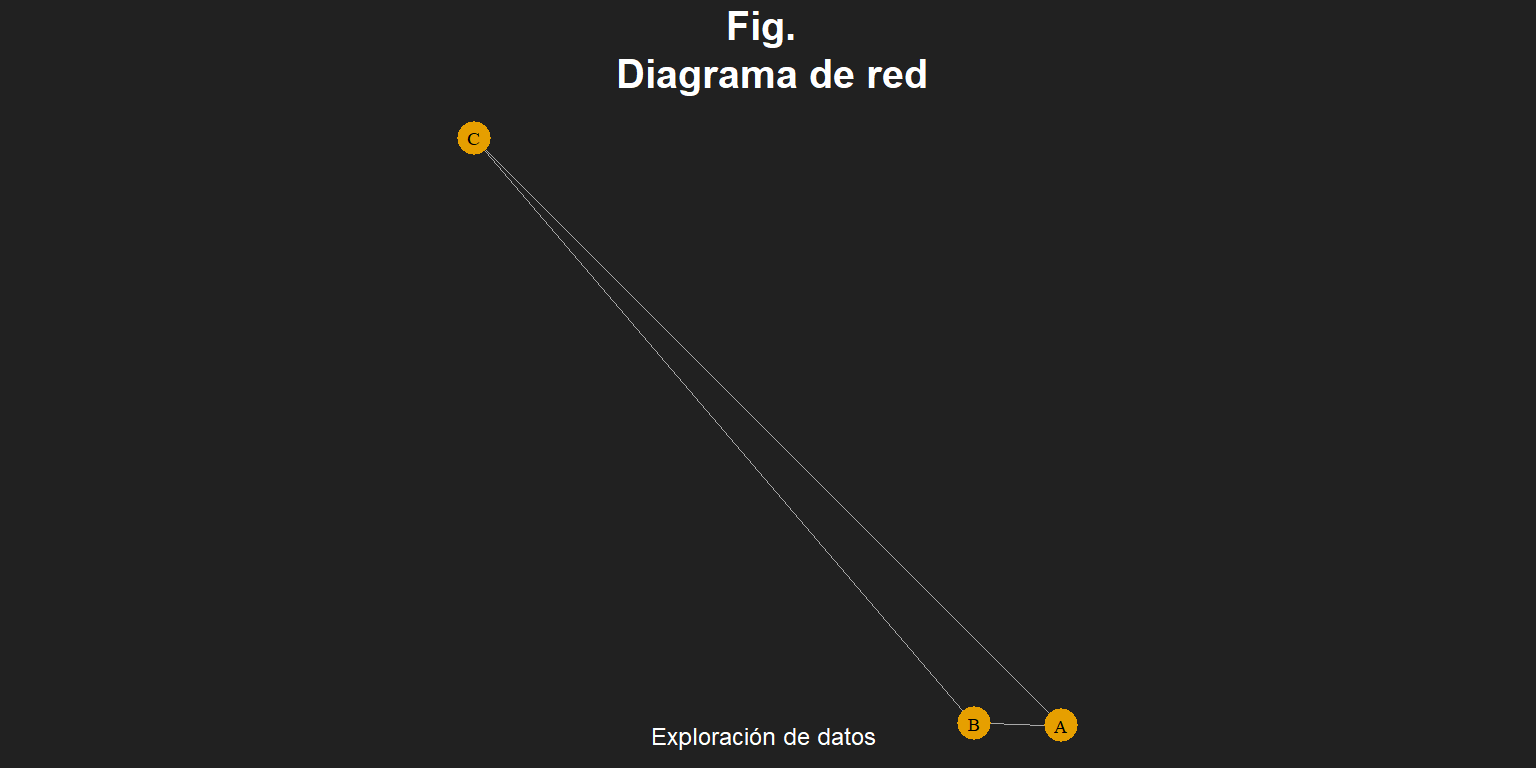


Se logra apreciar la siguiente información:

* El año de lanzamiento con la descripción del rating se puede indicar que poseen un porcentaje ligero de correlación positiva.
* Mientras que la descripción del rating frente al tamaño del rating del usuario poseen una mínima correlación negativa.

## Diagrama de Red

Cada matriz de correlación se la puede visualizar como un diagrama de red, en donde, cada entidad de la matriz es un nodo.



Donde las letras representan:

Nombres Letras  
[1,] "ratingDescription" "A"   
[2,] "release.year" "B"   
[3,] "user.rating.size" "C"

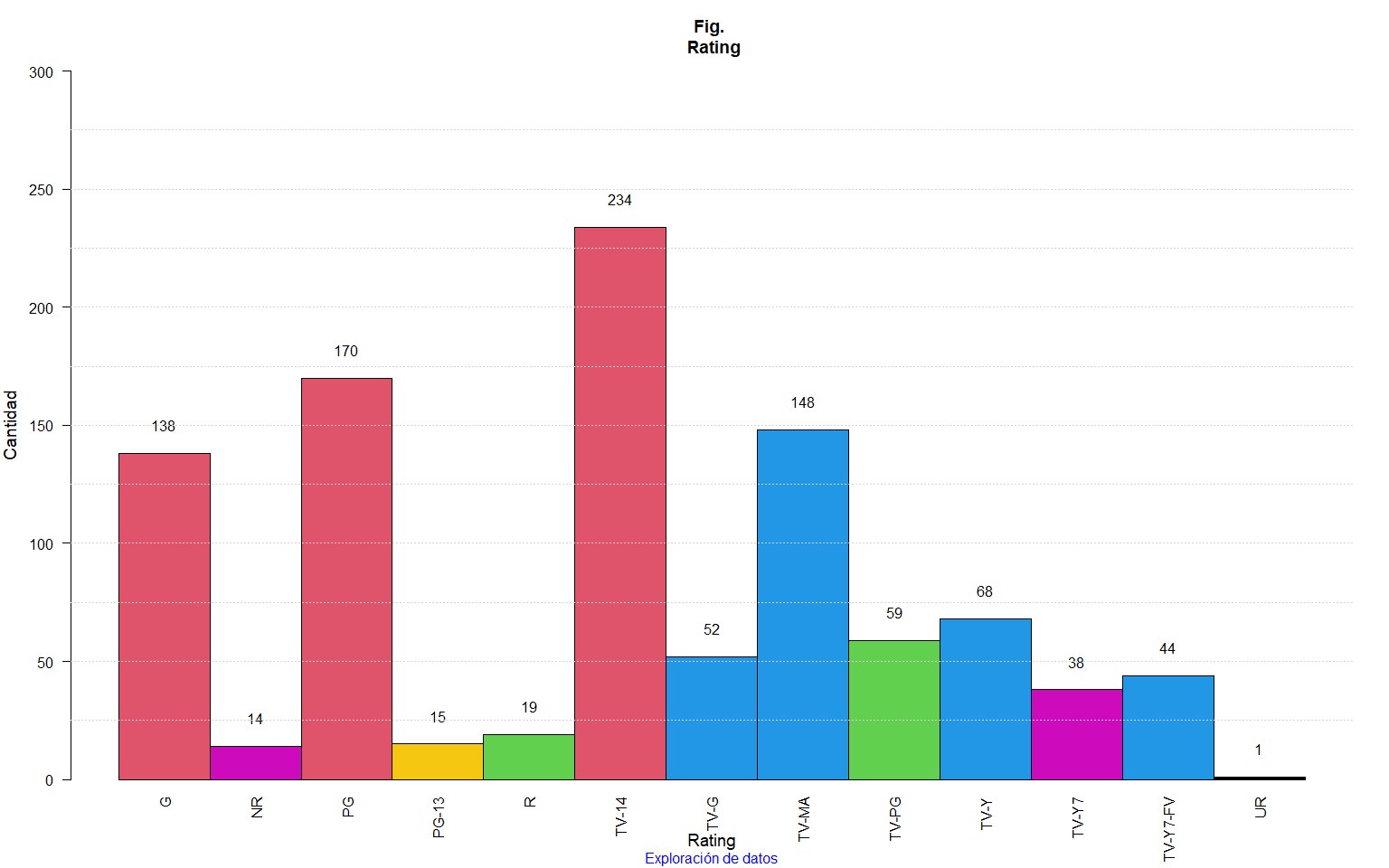
Se logra apreciar que las variables ratingDescription y release.year poseen mayor relación entre ellas que con la variable user.rating.size.

# 5. Visualizaciones

A continuación se realizarán las agrupaciones por categorías y ver su comportamiento.

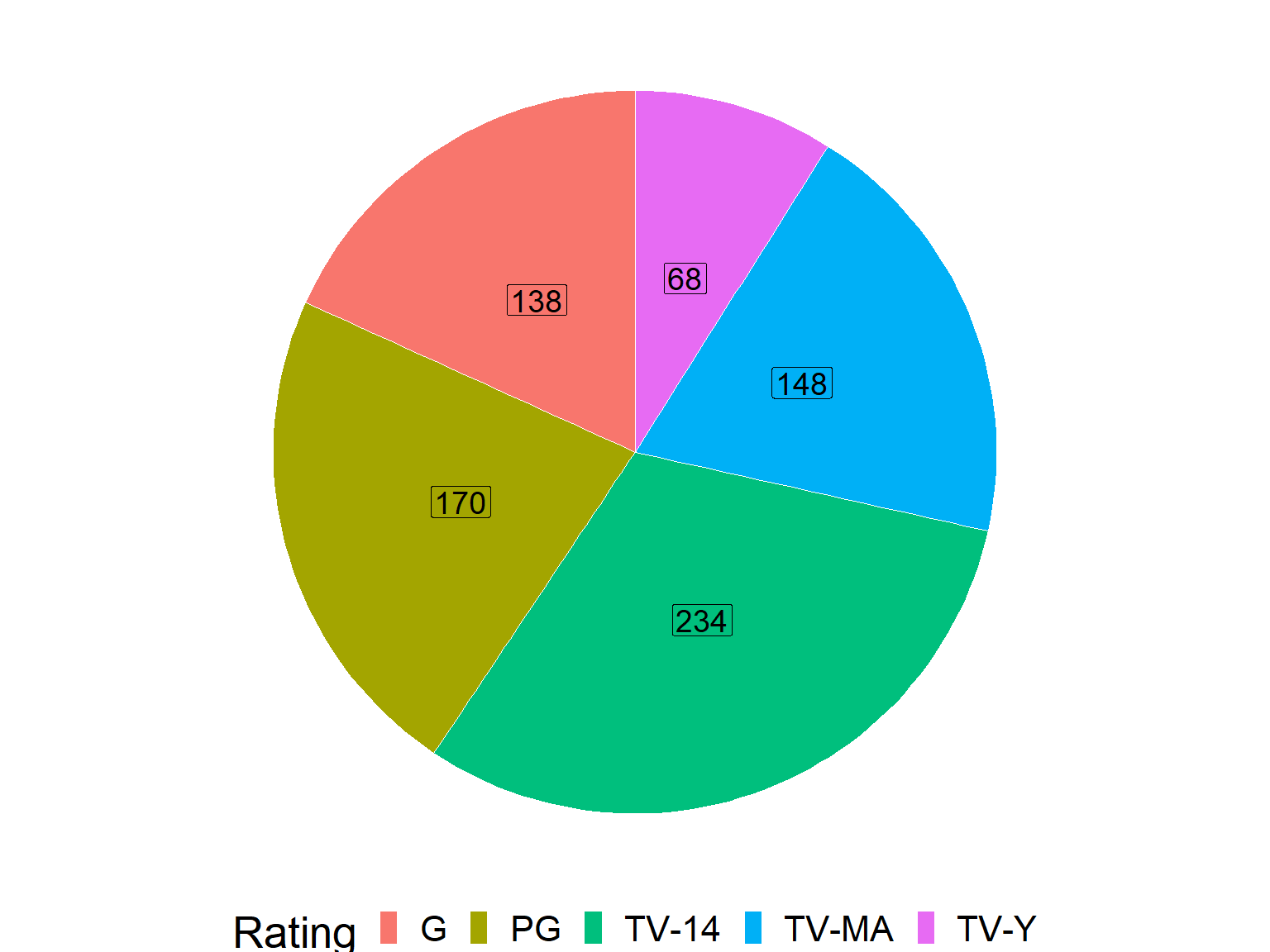
## Rating

El primer análisis gráfico realizado será a la categoría rating.



En la figura previa se observa la cantidad de elementos que se posee por cada rating.

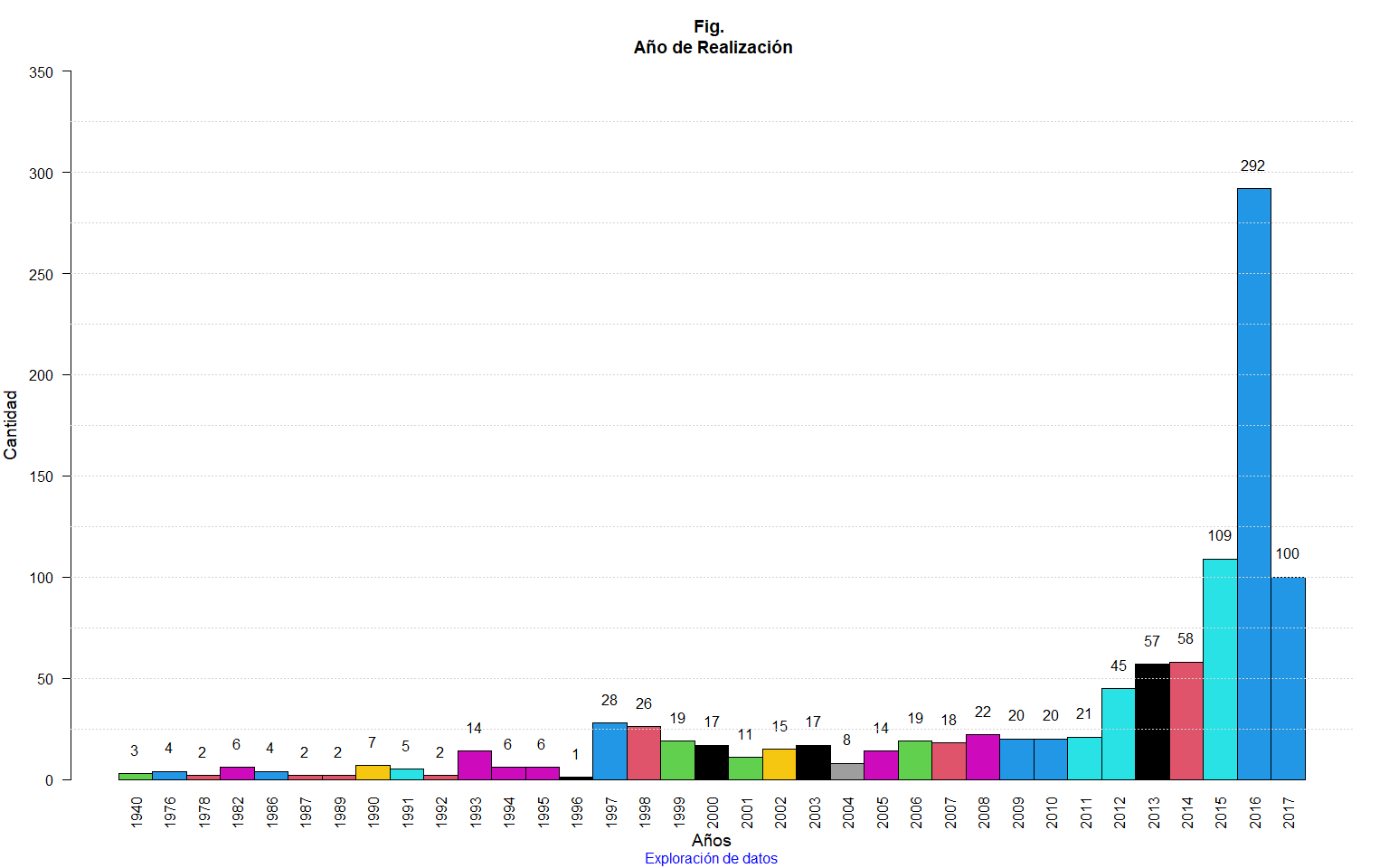
### Top 5 de Rating



Entre los ratings, predomina las clasificaciones TV-14.

## Release Year

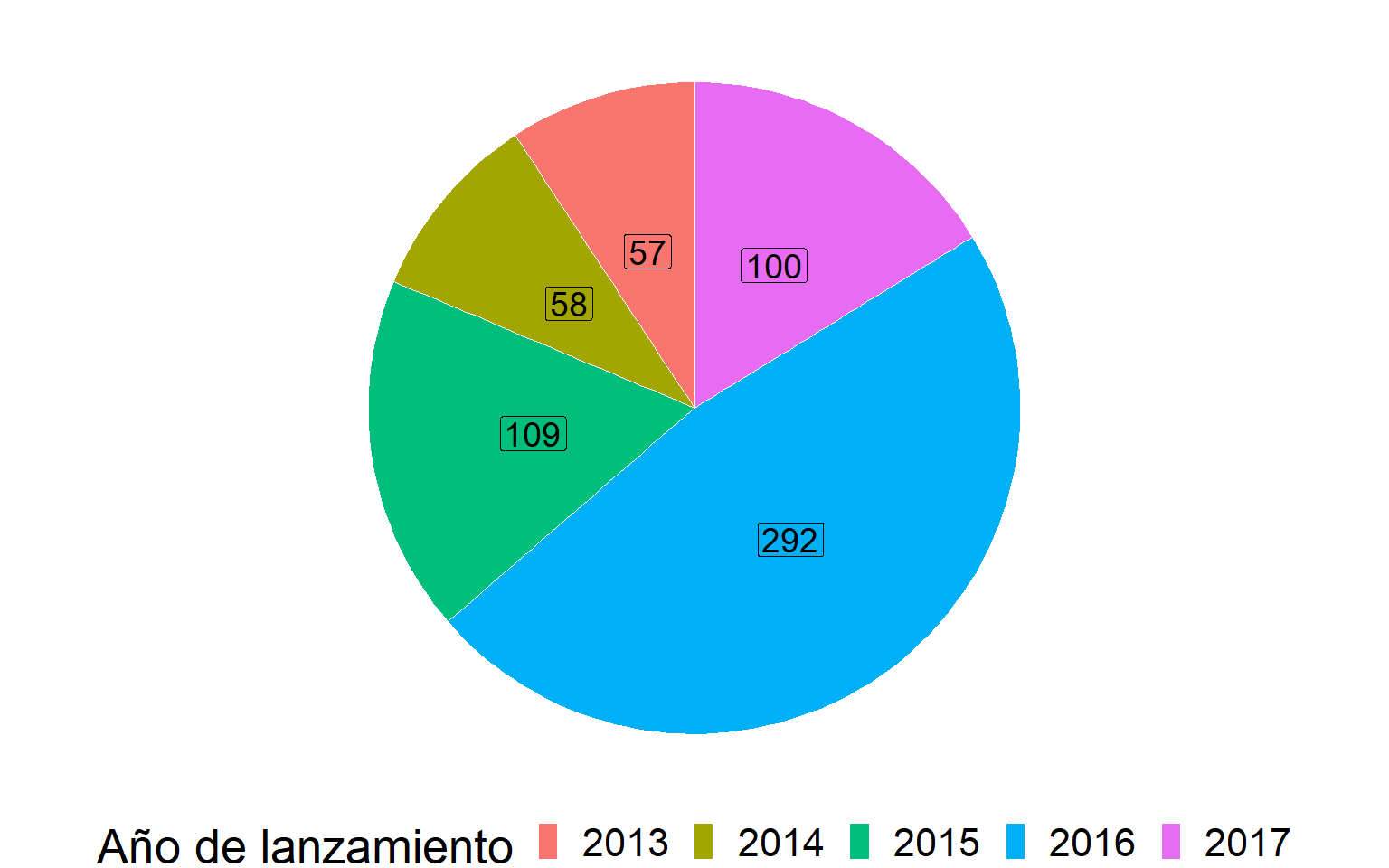
La segunda categoría a analizar es el año de lanzamiento.



Se toman la cantidad de cintas fueron realizadas por cada uno de los años, desde 1940 hasta el año 2017.

### Top 5 de años con mayor publicación

releaseYear\_Ordered Freq  
1 2016 292  
2 2015 109  
3 2017 100  
4 2014 58  
5 2013 57

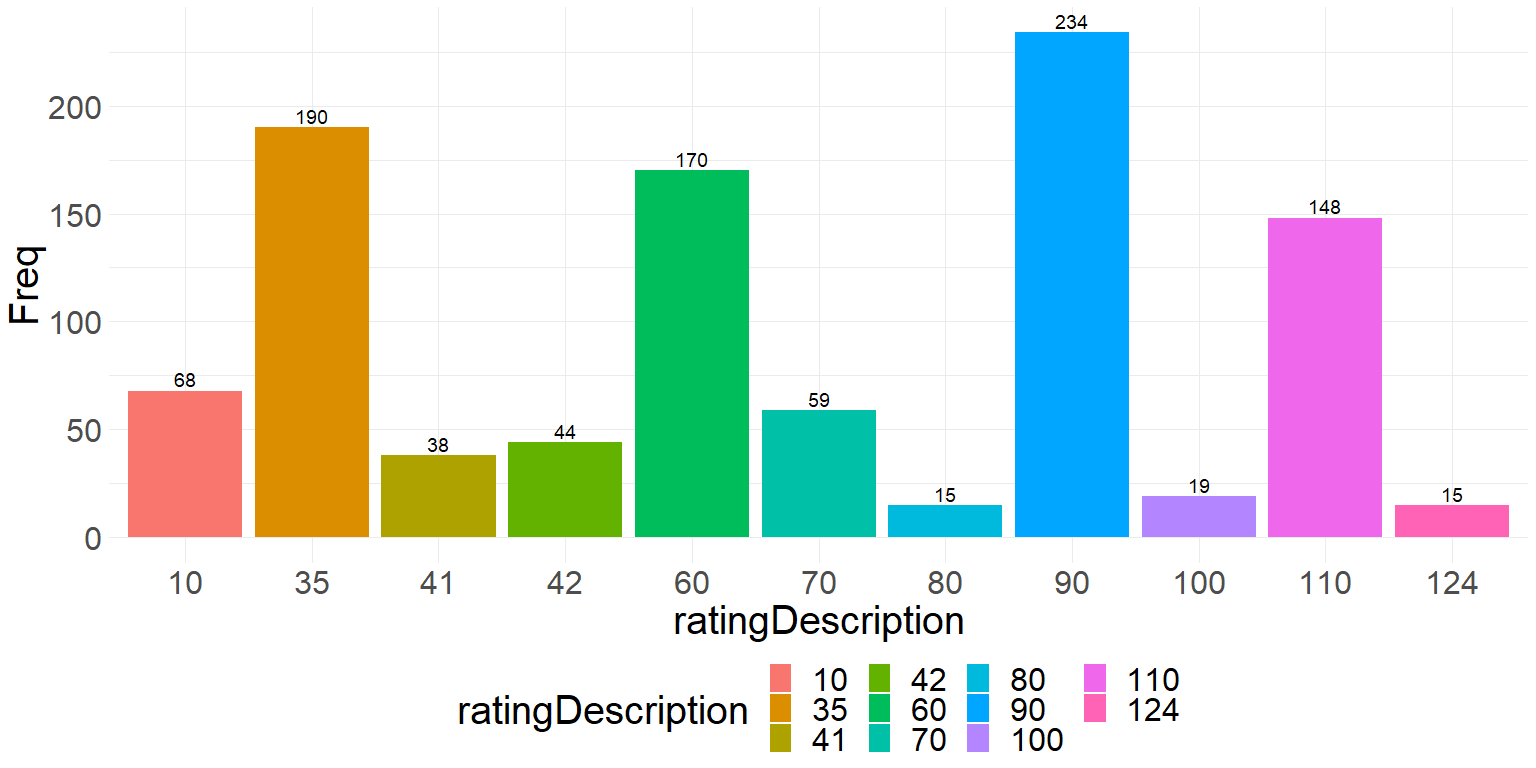


El año que predomina en esta lista de elementos con mayor cantidad de cintas, es el año 2016.

## Rating Description

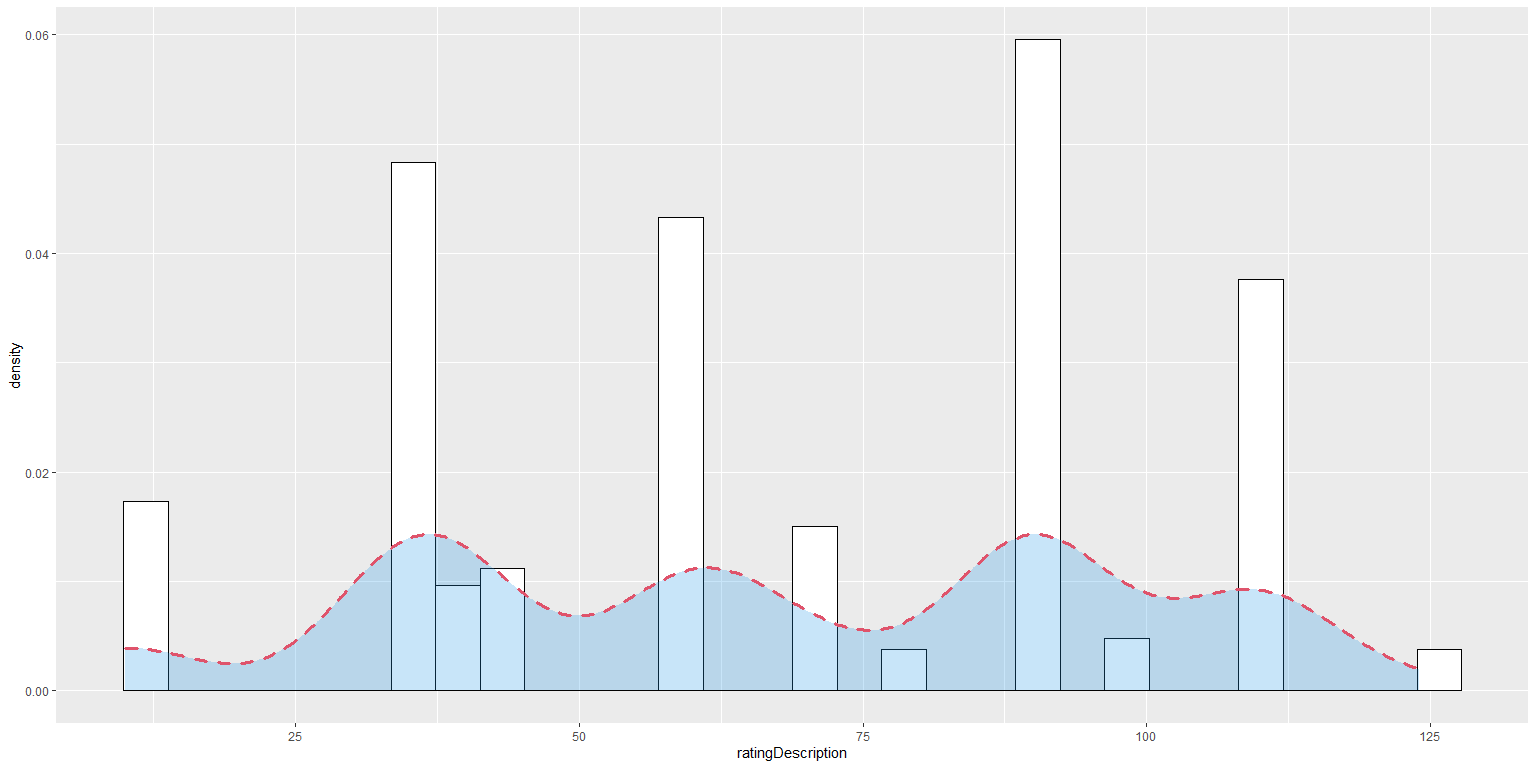
La tercera categoría observada es el rating description.

ratingDescription Freq  
1 10 68  
2 35 190  
3 41 38  
4 42 44  
5 60 170  
6 70 59  
7 80 15  
8 90 234  
9 100 19  
10 110 148  
11 124 15

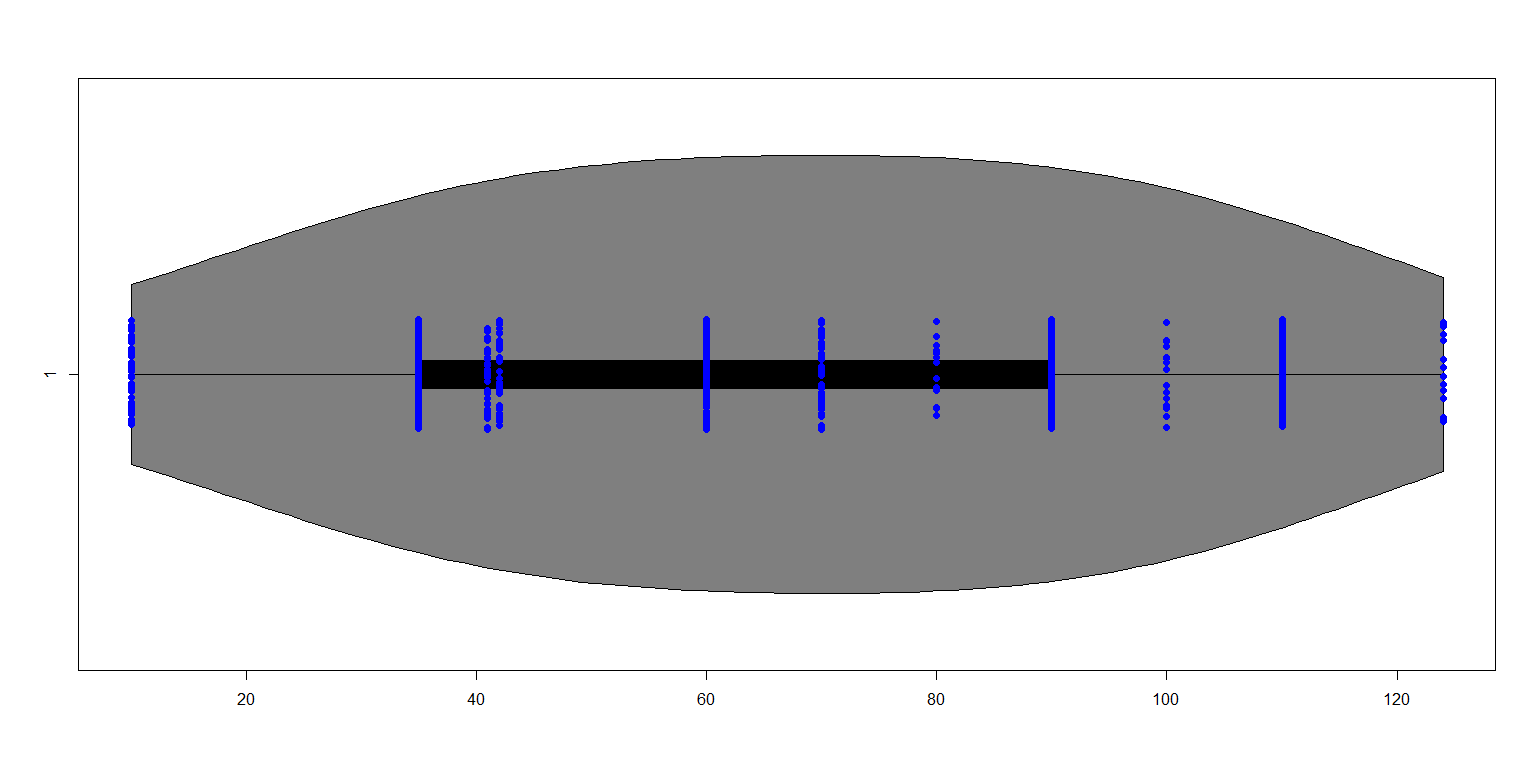


La moda en la categoría de rating description, se aprecia que es 90 entre las 1000 muestras.

### Histograma + densidad



### Diagrama de Violin



## Title

Vamos a analizar cuales fueron los títulos con mayor puntaje por el usuario.

# A tibble: 10 x 4  
# Groups: title [10]  
 title rating user.rating.score release.year  
 <chr> <chr> <dbl> <int>  
 1 13 Reasons Why TV-MA 99 2017  
 2 Grey's Anatomy TV-14 98 2016  
 3 Prison Break TV-14 98 2008  
 4 The Walking Dead TV-MA 98 2015  
 5 Once Upon a Time TV-PG 98 2016  
 6 Marvel's Iron Fist TV-MA 98 2017  
 7 The Flash TV-PG 98 2016  
 8 Family Guy TV-MA 98 2015  
 9 Criminal Minds TV-14 98 2016  
10 Friends TV-14 98 2003

Top 1:



Top 2:



Top 3:

