Data Viz de un modelo/dataset sobre Netflix

El enlace debajo es la fuente de los datos utilizados para las siguientes actividades de las que haré mención.

https://www.kaggle.com/datasets/shivamb/netflix-shows

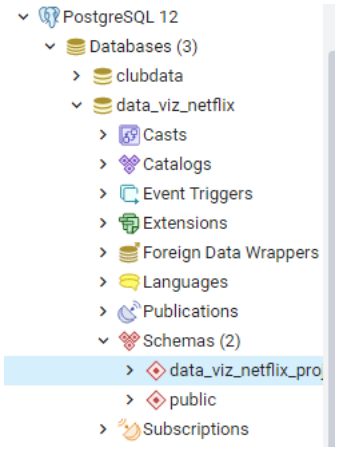
▼ **1- Una vez extraído la hoja de cálculo (csv) del que se utilizarán los datos, los migraremos a una base de datos PostgreSQL.**

1.1- Sin embargo, en caso de encontrar los datos de la fecha desalineados entre sus formatos, limpiaré en ambas formas (desde csv y la base de datos) para medir la rapidez y eficiencia entre cada opción.

1.2- Desde la hoja de cálculo; en éste caso LibreOffice, se colocaron todos los datos de la fecha en una sola columna con un simple copia y pega en una hoja distinta a la que están ubicados los datos (date_added - 1). Luego se copiaron los datos a un lado sin formato alguno, separando y distinguiendo sus días, meses y años (date_added - 2). Para terminar se juntan los valores con una fórmula y a un lado de éste último copiar los valores y formatos de la columna anterior (date_added - 3).

	A	B	C	D	E	F	G	H
1	date_added – 1		date_added – 2				date_added – 3	
2	9/25/2021		9	25	2021		25/9/2021	25/09/2021
3	9/24/2021		9	24	2021		24/9/2021	24/09/2021
4	9/24/2021		9	24	2021		24/9/2021	24/09/2021
5	9/22/2021		9	22	2021		22/9/2021	22/09/2021
6	9/24/2021		9	24	2021		24/9/2021	24/09/2021
7	9/24/2021		9	24	2021		24/9/2021	24/09/2021
8	9/24/2021		9	24	2021		24/9/2021	24/09/2021
9	05/01/2021		5	1	2021		1/5/2021	01/05/2021
10	9/23/2021		9	23	2021		23/9/2021	23/09/2021
11	05/01/2021		5	1	2021		1/5/2021	01/05/2021
12	05/01/2021		5	1	2021		1/5/2021	01/05/2021
13	05/01/2021		5	1	2021		1/5/2021	01/05/2021
14	5/21/2021		5	21	2021		21/5/2021	21/05/2021
15	7/13/2021		7	13	2021		13/7/2021	13/07/2021
16	06/12/2021		6	12	2021		12/6/2021	12/06/2021
17	06/12/2021		6	12	2021		12/6/2021	12/06/2021
18	05/07/2021		5	7	2021		7/5/2021	07/05/2021
19	9/24/2021		9	24	2021		24/9/2021	24/09/2021
20	9/22/2021		9	22	2021		22/9/2021	22/09/2021

▼ **2- Crear una base de datos dedicado a los datos que usaré en éste proyecto.**



2.1- Éste esquema no tendrá tablas así que crearemos la tabla con los formatos adecuados que tiene la hoja de cálculo.

2.2- Definiremos con el tipo de dato TEXT las variables que puedan contener carácteres especiales. Con los datos proporcionados se utilizará el formato VARCHAR con la variable show_id. Y se cambió el nombre de la variable cast a casting.

show_id	varchar
type	varchar
title	text
director	text
cast → casting	text
country	char
date_added	date
release_year	int
rating	varchar
duration	varchar
listed_in	text
description	text

2.3- Una vez definido los tipos de datos procedemos a crear la tabla.

2.4- Con los datos proporcionados basta con migrar todo como está señalado para no tener problemas y mayores retrasos que con el modelo anterior para tener las fechas limpias, sólo de ser necesario limpiar las fechas.

▼ **3- Organizar la base de datos de forma estándar con valores numéricos.**

3.1- Primero establecemos los valores para una actualización masiva. Agarrando los valores del dataset original, separando y copiando los valores numéricos de la variable show_id para colocarlos en lista de una hoja aparte seguido de copiar los textos a utilizar durante todos los valores.

```
UPDATE data_viz_netflix_project.netflix_viz
                                        SET id =      2 WHERE show_id = 's       2 ',
                                        SET id =      3 WHERE show_id = 's       3 ',
                                        SET id =      4 WHERE show_id = 's       4 ',
```

3.2- Bajando el comando UPDATE y copiando a lo largo de todos los valores necesarios para los scripts. Aún falta arreglar algunas cosas.

```
UPDATE data_viz_netflix_project.netflix_viz    SET id =      2 WHERE show_id = 's       2 ';
UPDATE data_viz_netflix_project.netflix_viz    SET id =      3 WHERE show_id = 's       3 ';
UPDATE data_viz_netflix_project.netflix_viz    SET id =      4 WHERE show_id = 's       4 ';
```

3.3- Los valores están ubicados en celdas diferentes, así postgre no lo ejecutará. Con la fórmula CONCAT de LibreOffice y luego pegarlos como texto sin formato tenemos todos los scripts correctamente. Luego de eliminar los datos extraídos a la derecha.

```
UPDATE data_viz_netflix_project.netflix_viz    SET id =2     WHERE show_id = 's2';
UPDATE data_viz_netflix_project.netflix_viz    SET id =3     WHERE show_id = 's3';
UPDATE data_viz_netflix_project.netflix_viz    SET id =4     WHERE show_id = 's4';
```

3.4- Y así está la base de datos capaz de organizarse por datos numéricos con la variable id aunque éste sea la última columna de la tabla coinciden sus valores (id = show_id).

```sql
SELECT id, show_id, title, type, duration
FROM data_viz_netflix_project.netflix_viz
ORDER BY id
LIMIT 20
;
```

Data Output | Explain | Messages | Notifications

	id integer	show_id character varying	title text	type character varying	duration character varying
1	1	s1	Dick Johnson Is Dead	Movie	90 min
2	2	s2	Blood & Water	TV Show	2 Seasons
3	3	s3	Ganglands	TV Show	1 Season
4	4	s4	Jailbirds New Orleans	TV Show	1 Season
5	5	s5	Kota Factory	TV Show	2 Seasons
6	6	s6	Midnight Mass	TV Show	1 Season
7	7	s7	My Little Pony: A New Generation	Movie	91 min

▼ **4- Establecer qué se necesita con ésta información para definir los datos a extraer.**

A- Visualizar series populares sin intervalo de tiempo definido (con al menos más de una temporada). Se distinguen las series de las películas con los datos de duración e identificado el número máximo de temporadas que tuvo unas series, tomo el valor cercano a la mitad hacia arriba para medirlo como popularidad. También con la columna type se puede filtrar como tipo 'película'; aunque la primera opción fue más complicada muestra que no está limitado a una única forma de conseguir los datos.

```sql
SELECT duration, title, director, rating, country
FROM data_viz_netflix_project.netflix_viz
WHERE duration IN ('8 Seasons', '9 Seasons', '10 Seasons', '11 Seasons', '12 Seasons', '13 Seasons',
                   '15 Seasons', '17 Seasons')
ORDER BY duration
;
```

Data Output | Explain | Messages | Notifications

	duration character varying	title text	director text	rating character varying	country text
1	10 Seasons	Friends	[null]	TV-14	United States
2	10 Seasons	Stargate SG-1	[null]	TV-MA	United States, Canada
3	10 Seasons	Danger Mouse: Classic Collection	[null]	TV-Y	United Kingdom
4	10 Seasons	LEGO Ninjago: Masters of Spinjitzu	[null]	TV-Y7	Denmark, Singapore, Canada, United States
5	10 Seasons	The Walking Dead	[null]	TV-MA	United States
6	10 Seasons	Dad's Army	[null]	TV-PG	United Kingdom

B.1- Inicialmente observamos que hay películas que no tienen registrados (con valores nulos) país alguno.

Data Output | Explain | Messages | Notifications

	country text	duration character varying
1	United States	90 min
2	South Africa	2 Seasons
3	[null]	1 Season
4	[null]	1 Season
5	India	2 Seasons
6	[null]	1 Season
7	[null]	91 min
8	United States, Ghana, Burkina Faso, United Kingdom, Germany, Ethiopia	125 min

B.2- Entonces seguimos descartando de todos los datos (las 8807 filas) los países con valores nulos y la duración en base a temporadas (pues eso sería la distinción de las series con las películas), dejando 5688 datos de películas con países registrados.

Data Output | Explain | Messages | Notifications

	country text	duration character varying
1	United States	90 min
2	United States, Ghana, Burkina Faso, United Kingdom, Germany, Ethiopia	125 min
3	United States	104 min
4	Germany, Czech Republic	127 min
5	India	166 min
6	United States	103 min
7	United States	97 min
8	United States, India, France	106 min

B.3- Utilizando tanto las funciones de agregación junto a condicionales internas, tenemos el número de películas en los países registrados, funciona exclusivamente por cada fila; habrán valores distintos para las películas únicamente en Estados Unidos al igual que valores distintos para películas que estén ubicadas en Estados Unidos junto a otros países.

Data Output | Explain | Messages | Notifications

	paises text	peliculas bigint
1	United States	2058
2	India	893
3	United Kingdom	206
4	Canada	122
5	Spain	97
6	Egypt	92
7	Nigeria	86
8	Indonesia	77

C.1- Mostrar la duración de las películas de los últimos 9 años por país, de los países que tienen mayor número de transmisión de películas de género 'Comedia' (de preferencia estadística: los 10 primeros en ranking de la búsqueda específica).

C.2- Con ésto tenemos las películas de comedia junto a los países que las sacaron en los últimos nueve años.

Data Output | Explain | Messages | Notifications

	country text	type character varying	listed_in text	release_year integer
1	Nigeria	Movie	Comedies, International Movies, Romantic Movies	2016
2	United States	Movie	Comedies, Dramas	2021
3	Nigeria	Movie	Action & Adventure, Comedies, Dramas	2020
4	United Kingdom, United States	Movie	Children & Family Movies, Comedies	2018
5	India	Movie	Action & Adventure, Comedies, Dramas	2017
6	India	Movie	Comedies, Dramas, Independent Movies	2015
7	United States	Movie	Comedies	2016
8	United States	Movie	Action & Adventure, Comedies	2014
9	India	Movie	Comedies, International Movies	2016
10	India, Nepal	Movie	Comedies, Dramas, International Movies	2018

C.3- El conteo de películas de los 10 países que han transmitido los mayores números de películas de comedia en los últimos nueve años, y hacer de ésto una vista (VIEW).

Data Output | Explain | Messages | Notifications

	country text	type character varying	listed_in text	release_year integer	peliculas bigint
1	India	Movie	Comedies, Dramas, International Movies	2019	12
2	United States	Movie	Comedies, Dramas, Independent Movies	2016	12
3	India	Movie	Comedies, Dramas, International Movies	2014	12
4	India	Movie	Comedies, Dramas, International Movies	2018	11
5	United States	Movie	Comedies	2019	11
6	United States	Movie	Comedies	2017	11
7	United States	Movie	Children & Family Movies, Comedies	2020	11
8	India	Movie	Comedies, Dramas, International Movies	2017	10
9	United States	Movie	Comedies	2020	10
10	United States	Movie	Comedies, Dramas, Independent Movies	2017	9

C.4- Al final se consigue el número de películas de comedia que lanzaron los 10 países que más estrenaron de éste género y tipo durante los últimos nueve años. Con ésta información extraemos el dataset para la visualización de datos en R.

Data Output | Explain | Messages | Notifications

	country text	conteo numeric
1	United States	301
2	India	177
3	Turkey	41
4	Nigeria	37
5	Spain	34
6	Canada	32
7	Egypt	26
8	Philippines	25
9	Indonesia	12
10	Germany	12

C.5-

```
SELECT
FROM netflix_project.netflix_viz

SELECT DISTINCT(listed_in)
FROM netflix_project.netflix_viz
```

```
WHERE type LIKE 'Movie' AND listed_in ILIKE '%comedies%'
-- Para distinguir las películas de género comedia
-- Con 1674 películas de comedia en TOTAL

SELECT country, type, listed_in, release_year
FROM netflix_project.netflix_viz
WHERE release_year >= 2013 AND country IS NOT NULL AND type LIKE 'Movie' AND listed_in LIKE '%Comedies%'
;
-- Con ésto tenemos las películas de comedia junto a los países que las sacaron en los últimos nueve años

SELECT country, type, listed_in, release_year, count(show_id) AS peliculas
FROM netflix_project.netflix_viz
GROUP BY country, type, listed_in, release_year
HAVING release_year >= 2013 AND country IS NOT NULL AND type LIKE 'Movie' AND listed_in LIKE '%Comedies%'
  AND country IN ('India', 'United States', 'Egypt', 'Canada', 'Philippines', 'Turkey', 'Spain',
             'Nigeria', 'Germany', 'Indonesia')
ORDER BY peliculas DESC
;
-- El conteo de películas de los 10 países que han transmitido los mayores números de películas de comedia en los últimos nueve años, y

CREATE OR REPLACE VIEW movies_countries AS
SELECT country, type, listed_in, release_year, count(show_id) AS peliculas
FROM netflix_project.netflix_viz
GROUP BY country, type, listed_in, release_year
HAVING release_year >= 2013 AND country IS NOT NULL AND type LIKE 'Movie' AND listed_in LIKE '%Comedies%'
  AND country IN ('India', 'United States', 'Egypt', 'Canada', 'Philippines', 'Turkey', 'Spain',
             'Nigeria', 'Germany', 'Indonesia')
ORDER BY peliculas DESC
;

SELECT country, sum(peliculas) AS conteo
FROM movies_countries
GROUP BY country
ORDER BY conteo DESC
;
-- Al final se consigue el número de películas de comedia que lanzaron los 10 países que más estrenaron de éste género y tipo durante l

-- Siguiendo con la duración de éstas
SELECT country, title, release_year, duration, CAST(LEFT(duration, 3) AS INTEGER) AS minutes
FROM netflix_project.netflix_viz
WHERE release_year >= 2013 AND country IS NOT NULL AND type LIKE 'Movie' AND listed_in LIKE '%Comedies%'
  AND country IN ('India', 'United States', 'Egypt', 'Canada', 'Philippines', 'Turkey', 'Spain',
             'Nigeria', 'Germany', 'Indonesia')

ORDER BY minutes asc
;
-- Con ésta información realizamos el dataset
```

D.1- Mostrar una progresión sobre cuantas películas se han subido a la plataforma respecto al año de estreno que les corresponde (cuantas películas están en la plataforma que se hayan estrenado X año).

D.2- Con estos únicos datos serán suficientes para formar un dataset que permita crear una progresión histórica sobre la publicación de películas en la plataforma para cada país.

Data Output | Explain | Messages | Notifications

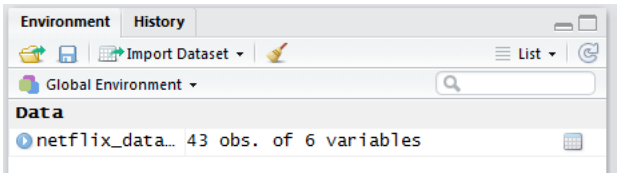
	countries text	years integer	counts bigint
1	Argentina	2021	1
2	Australia	2021	5
3	Belgium	2021	2
4	Brazil	2021	1
5	Canada	2021	3
6	Colombia	2021	1
7	Egypt	2021	1
8	France	2021	2
9	Germany	2021	4
10	Iceland	2021	1
11	India	2021	9
12	Israel	2021	2
13	Italy	2021	2
14	Japan	2021	10
15	Jordan	2021	1
16	Mexico	2021	6

D.3-

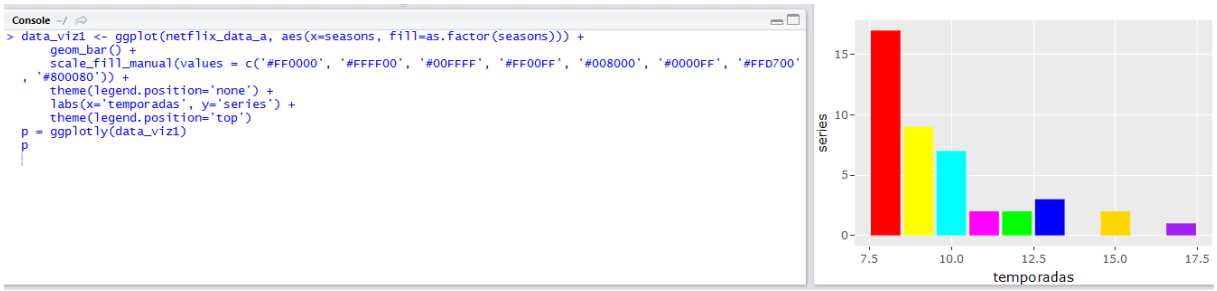
```
SELECT country AS countries, release_year AS years, COUNT(title) AS counts
FROM netflix_project.netflix_viz
  GROUP BY country, release_year, type
  HAVING type LIKE 'TV Show' AND country NOT ILIKE '%,%'
ORDER BY release_year DESC, countries
;
-- Con estos únicos datos serán suficientes para formar un dataset que permita crear una progresión histórica sobre la publicación de p
```

▼ **5- Visualizar los datos extraídos.**

A.1- Inicialmente, utilizando el software RStudio con el lenguaje de R para hacer las visualizaciones, importamos el dataset empleado (que puede ser modificado con anterioridad o tal cual está) desde la parte arriba a la derecha del programa en 'import dataset'.



A.2- Con los datos extraídos, realizamos un gráfico de barras para mostrar el número de series (eje Y) que cuentan con el número de temporadas correspondientes (eje X). Para eso guardo un gráfico de barras estándar como una variable.



A.3- Aunque en un principio el objetivo era aclarar las series populares, desde gráficos visuales que muestren como en éste caso que hay más de 40 series en general que cuentan con 8 temporadas o más y que cada vez menos series cuentan con un número más alto de temporadas.

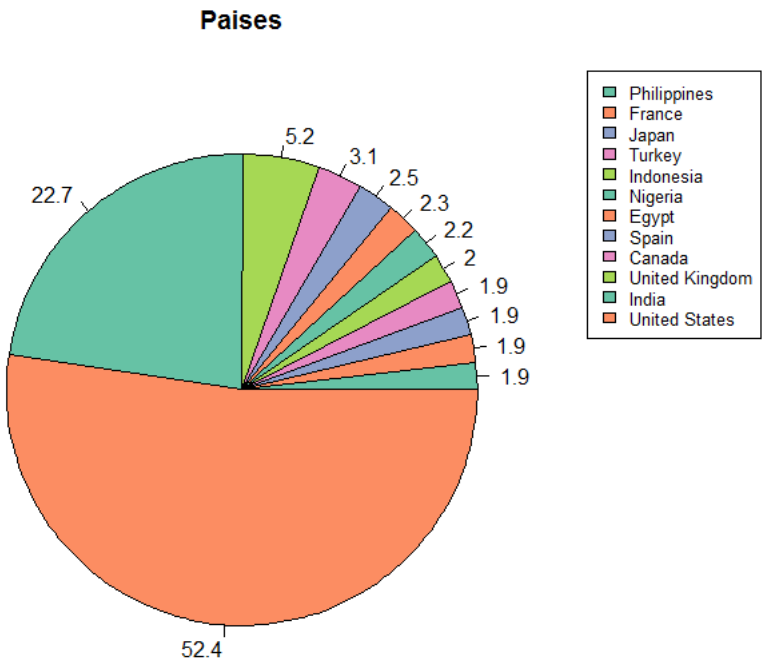
A.4- Se necesitaría de un gráfico de tablas para visualizar las series con sus respectivas temporadas. Particularmente creé un dataframe exceptuando los demás valores (director, rating…) para mostrar los datos concreta-mente necesarios.

	seasons	title
30	10	LEGO Ninjago: Masters of Spinjitzu
31	10	The Walking Dead
32	10	Dad's Army
33	10	Shameless (U.S.)
34	11	Frasier
35	11	Cheers
36	12	Trailer Park Boys
37	12	Criminal Minds
38	13	COMEDIANS of the world
39	13	Red vs. Blue

Showing 29 to 39 of 43 entries

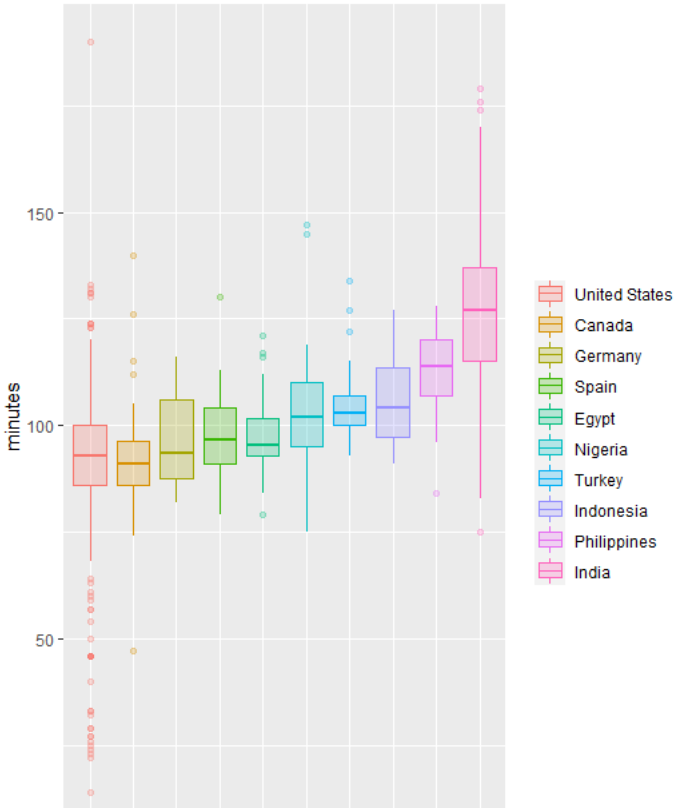
```
Console ~/
> view(newdataviz)
>
```

B- Los valores en el gráfico circular presentan los valores porcentuales, con los países a los que representa ascendente-mente, de modo que Philippines, France, Japan y Turkey cuentan con el 1.9% de contribución a las películas en Netflix cada uno mientras United States contribuye con el 52.4% de las películas. Aún así éstas no son todas las películas sino el total de los 12 países que aportan más películas a la plataforma en películas.



C.1- Realizamos un boxplot para mostrar la duración de las películas en datos utilizados que se pueden apreciar, donde los 5 países a la izquierda/arriba de la leyenda son aquellos que cuentan con las películas con una duración menor a los 100 minutos en su mayoría, mientras la otra mitad cuentan con su mayoría películas superiores a los 100 minutos, también se

puede observar como la India cuenta con una media mayor a los 125 minutos. Igualmente se pueden observar puntos/datos no habituales fuera de los diagramas que están por debajo o por encima de los datos aglomerados en los boxplot que reflejan un porcentaje bastante bajo (muy cercano a una única película).

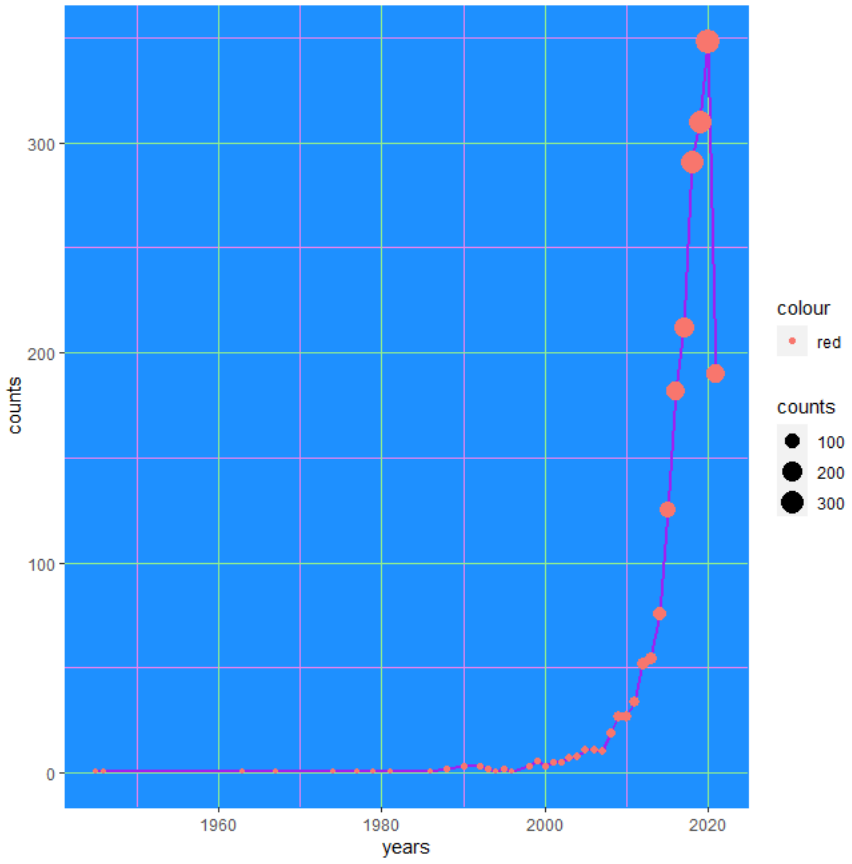


C.2-

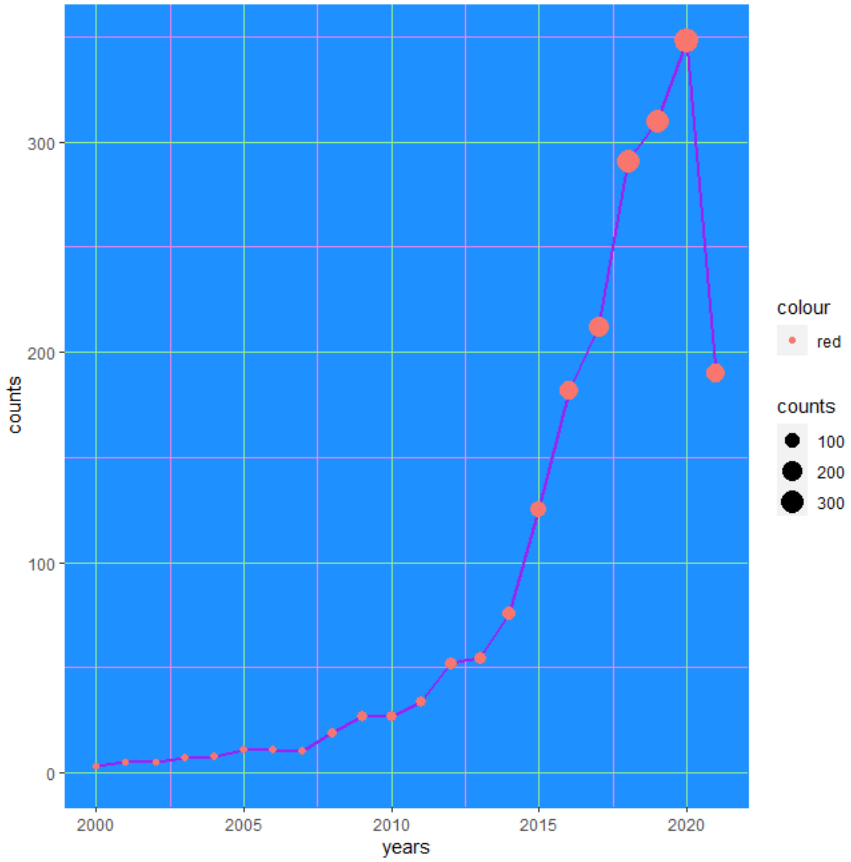
```
boxplot(datas$minutes,
  ylab='duracion (minutos)',
  main='duracion de peliculas')

ggplot(netflix_data_c, aes(x=as.factor(country), y=minutes)) +
  geom_boxplot(fill='slateblue', alpha=0.2) +
  xlab('country')

box_data <- ggplot(netflix_data_c,
          aes(x=reorder(as.factor(country),minutes),
            y=minutes,
            fill=as.factor(country),
            color=as.factor(country))) +
        geom_boxplot(alpha=0.25) +
        xlab('country')

box_data + scale_fill_manual(values=c(
              'red',
              'navy',
              'blue',
              'yellow',
              'cyan',
              'magenta',
              'orange',
              'maroon',
              'orange red',
              'purple'
              ))
# Un boxplot que muestra la aglomeración más común de los datos utilizados donde se puede apreciar que los 5 países a la izquierda/arri

ggplot(netflix_data_c, aes(x=reorder(as.factor(country),minutes),
```

```
              y=minutes,
              fill=reorder(as.factor(country),minutes),
              color=reorder(as.factor(country), minutes))) +
  geom_boxplot(alpha=0.25) +
  xlab('country')

indiaset <- netflix_data_c %>%
  filter(country %in% c('India'))
#Comprobando matemáticamente con el software los datos del boxplot sobre India
```

D.1- Se puede apreciar que son muy pocas las películas estrenadas por debajo de los años 2000 que se encuentran en la plataforma, aunque va aumentando el número de películas según el año que suben desde los 2000 no es sino hasta después del 2010 que aumenta significativamente (éstos datos pueden estar incompletos por lo que no es certeza de si la plataforma cuenta con menos películas del 2021 que la de varios años atrás en el momento en que se hizo la extracción de ésta información).



D.2- Descartando las películas que se estrenaron por debajo de los años 2000 puede apreciarse mejor y ligeramente con más detalle las películas publicadas según sus años de salida.


counts
300
200
100
0
2000
2005
2010
2015
2020
years
colour
red
counts
100
200
300