Parcial1

Camilo Alvarez

11/3/2021

Caso: Netflix

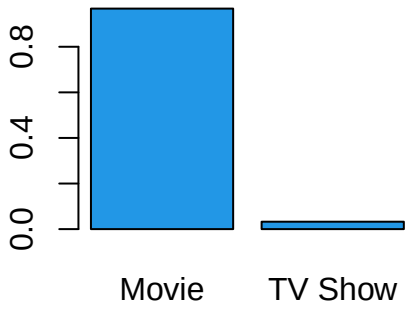
Se tiene una base de datos con registros de de 7787 películas/series/programas de Netflix con las siguientes características: “show_id” “type” “title” “director” “cast” “country” “date_added” “release_year” “rating” “duration” “listed_in” “description”

Por simplicidad, excluiremos las variables cast y description, además por decisión del analista, se **eliminaron** las observaciones que presentaran algún dato en blanco en alguna de las variables de interés anteriormente mencionadas.

1.

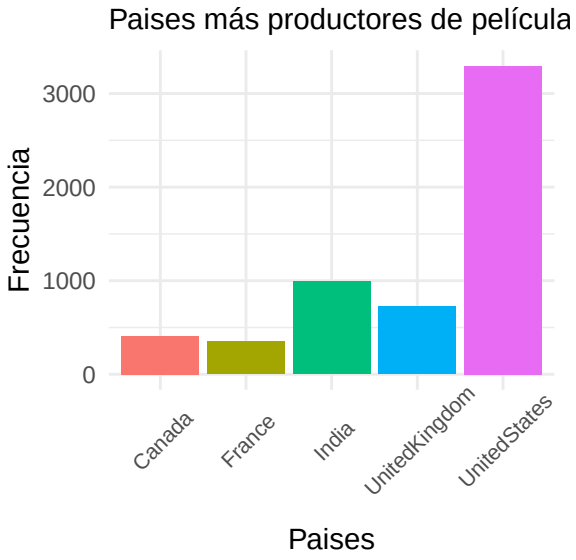
Comenzamos el análisis con tres variables clave como son type, country y release_year

Type:



Se puede concluir que de los datos seleccionados para analizar, la mayoría de registros son de información relacionada a Peliculas y una muy pequeña parte relacionada con TV Shows.

Country:



Luego, se puede inferir que de los países productores de estas peliculas, Estados Unidos es con gran diferencia el que más produce, seguido de la India y Reino Unido, entre otros paises que se pueden observar en la gráfica.

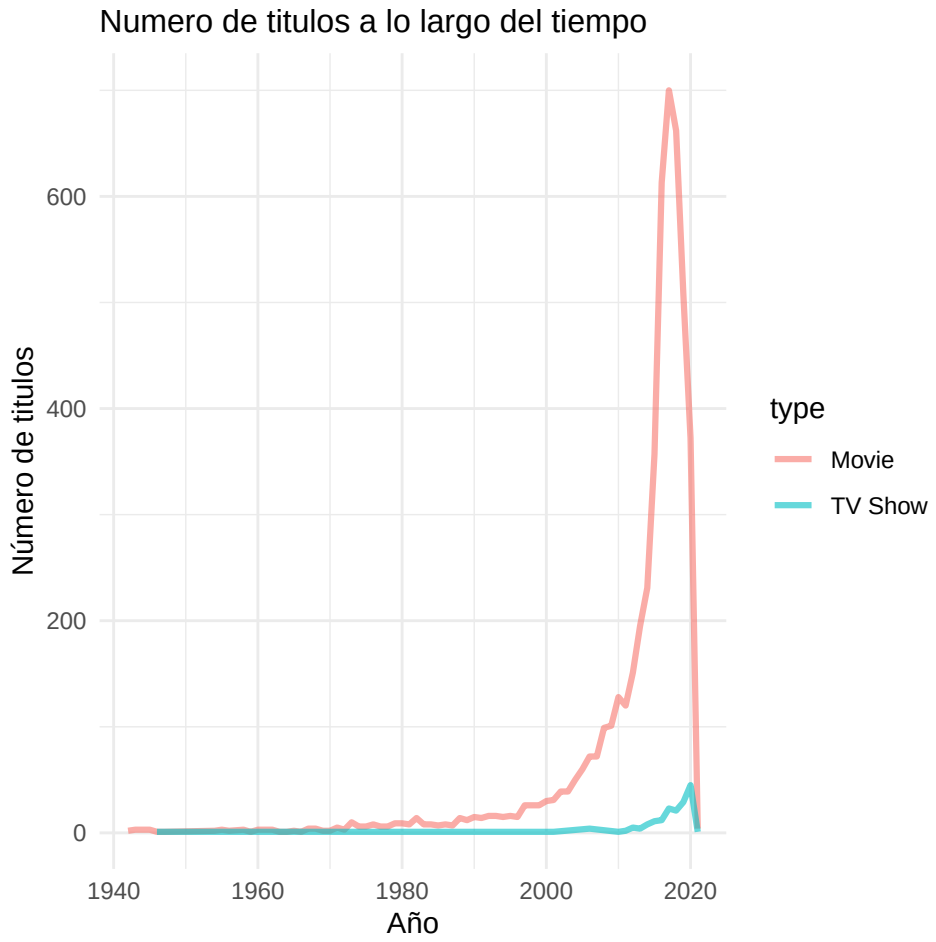
Release_year:

Con este gráfico de frecuencia de lanzamientos se nota un claro incremento en los ultimos años especialmente en la última decada. Además con la siguiente tabla se puede concluir que de los TvShow y Peliculas de los datos, por lo menos el 50% tuvo una fecha de lanzamiento mayor o igual al año 2017

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1925    2013    2017    2014    2018    2021
```

2.

Para esta segunda parte se revisó el número de titulos por año y el año en el que se presentaron (release_year)

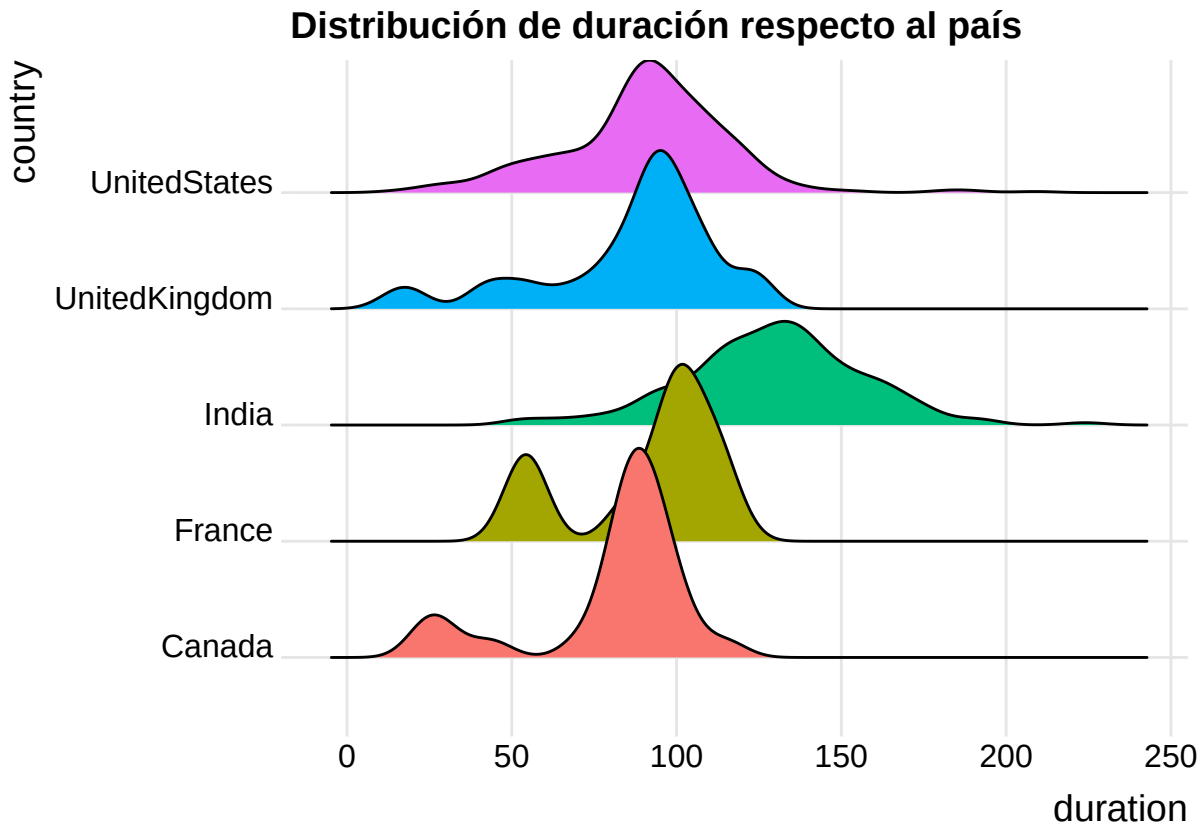



```
##         Antes de 2010 Desde de 2010 Incremento porcentual
## TV Show             7           162              2214.2857
## Movie             979          4041               312.7681
```

Para el caso de las Movies se observa un claro patrón con tendencia a la alza en la ultima decada, mientras que para los TV Show aunque no se ve tan pronunciado, también existe una leve tendencia a la alza en los ultimos años. Cabe mencionar que para ambas lineas de tiempo se ve observa un bajón abrupto para el año 2020, lo cual tiene totalmente sentido debido a la situación sanitaria. Se puede afirmar que tanto los Tv Show como para las Movies se ha triplicado desde 2010 con respecto a antes de este año, lo podemos ver en la tabla donde el incremento porcentual de ambos es mayor al 300%. Sin embargo, es importante aclarar que debido a que se eliminaron observaciones por tener datos en blanco, es muy probable que esa falta de información se haya tratado por ser registros de fechas antiguas lo cual inlfuye en tener estos incrementos porcentuales tan elevados, como es el caso de los TvShow.

3.

Para esta tercera parte nos enfocaremos solamente en los registros de Movies, y estudiaremos las variables de duration y country para estos registros.



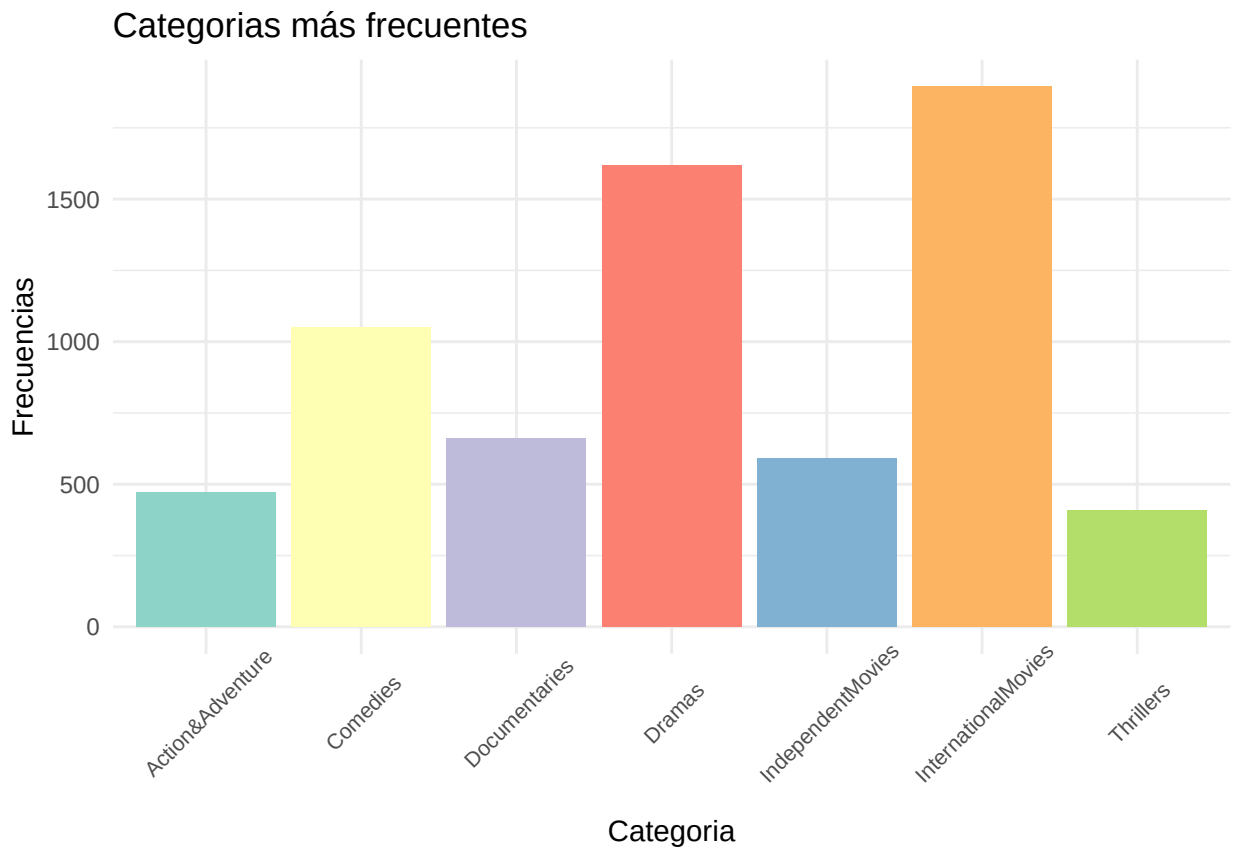
Con el gráfico anterior podemos ver como es la distribución de la duración de las peliculas para los paises con mayor número de registros. Podemos observar que las peliculas de Estados Unidos y United Kingdom tienen distribuciones bastantes similares. Las gráficas para la mayoría de paises lucen relativamente simetrícas a excepción de la de Francia que pareciera tener tendencia a ser bimodal, lo cual sería interesante averiguar el por qué. Adicionalmente cabe destacar que la media de India es el país que produce peliculas con la mayor duración en comparación con el resto de paises.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     3.0    88.0    99.0   100.8   115.0   253.0
```

En el resumen anterior se ven medidas de posición para la duración en general (para todos los paises) y podemos sacar conclusiones interesantes, como por ejemplo que la pelicula de mayor duración de todos los tiempos duró 253 minutos y fue The School of Mischief mientras que la pelicula de menor duración tuvo 3 minutos y fue Silent.

4.

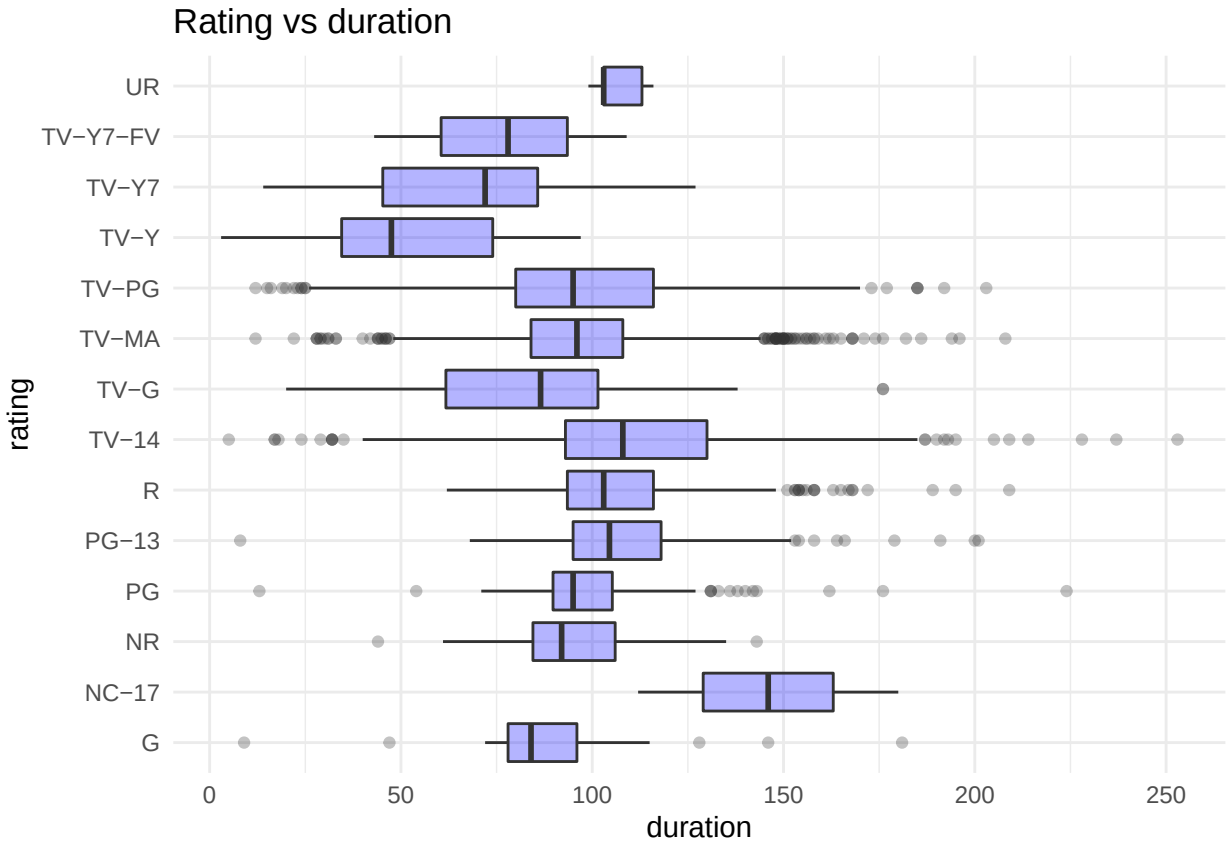
Para esta cuarta parte, estudiaremos la variable listed_in, para analizar las diferentes listas de categorías en las que más se ubican los registros de peliculas.



Este resultado presentado en el gráfico anterior es muy importante porque nos indica cuales son las categorias o los generos de peliculas que más se están produciendo en la última decada. Si se desa hacer una inversión para nuevo contenido, se debería enfocar en las categorías de International Movies, Drama y Comedia, que tienen buenas opciones en ser exitosas. Por otro lado si se desea hacer campaña de publicidad para impulsar algún tipo de genero, lo ideal estaría entre las categorías de Action, Documentaries, Independient y Thrillers.

5

En esta quinta parte, se estudio las variables rating y duración, además para está ultima se hicieron unas hipotesis respecto a los años 2019 y 2020 que serán debidamente validadas o rechazadas.




```
##
##  Welch Two Sample t-test
##
## data:  x1 and x2
## t = 2.1387, df = 738.3, p-value = 0.03278
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.3334517 7.7913473
## sample estimates:
## mean of x mean of y
##  96.47638  92.41398
```

```
##
##  F test to compare two variances
##
## data:  x1 and x2
## F = 0.77677, num df = 507, denom df = 371, p-value = 0.008498
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.6416882 0.9376370
## sample estimates:
## ratio of variances
##          0.7767696
```

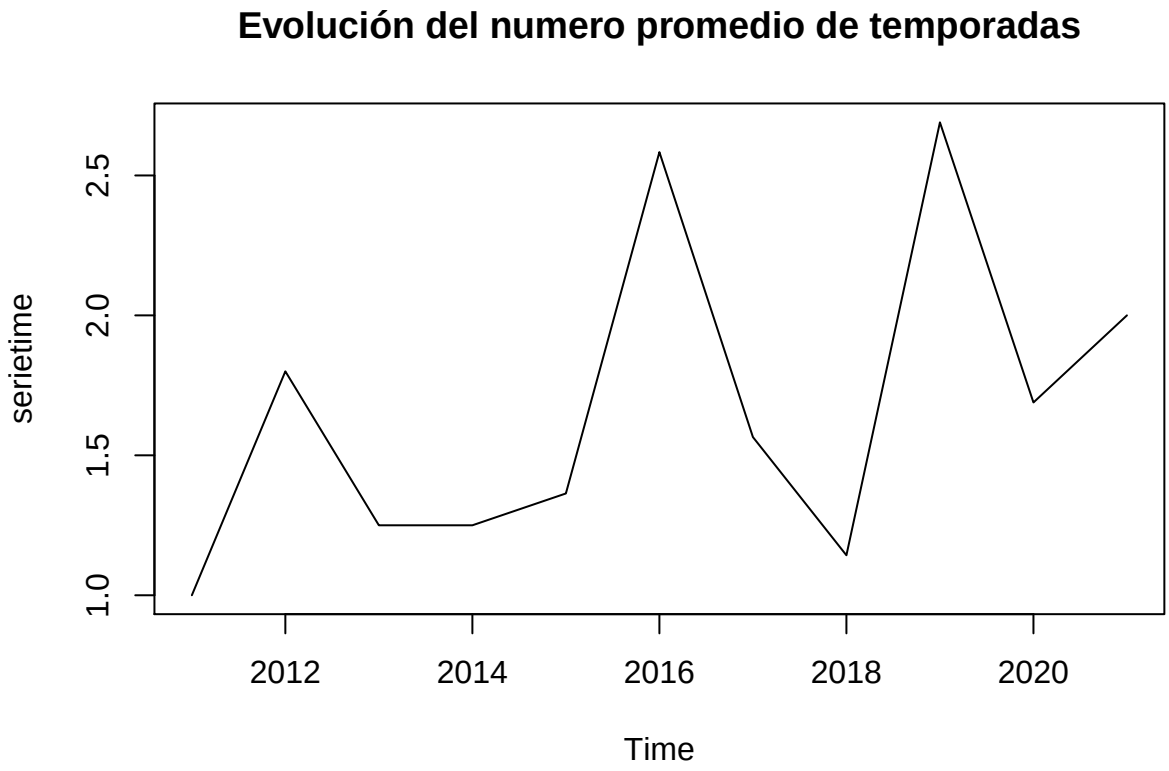
Con base en el gráfico anterior se puede decir que si existe relación entre las categorías de rating y la duración, debido a que a diferentes niveles de rating también varía la mediana de la duración.

Se utilizó pruebas de hipotesis para probar si la duración promedio y la variabilidad de la duración para los

años 2019 y 2020 se puede considerar igual, tomando un nivel de significancia del 0.05, para ambos casos se rechaza la hipoteisis nula, por tanto se puede concluir que tanto la duración promedio como la variabilidad de la misma son diferentes para los años 2019 y 2020.

6

Por ultimo, se quiso estudiar para los TV Show, el número promedio de temporadas y la evolución de este periodo en el tiempo.



Si bien son pocos datos para dar una conclusión más fundamentada, se puede decir que en el numero de temporadas promedio de cada año, se observa un patron con una ligera tendencia al alza y además una aparente estacionalidad cada 4 años.