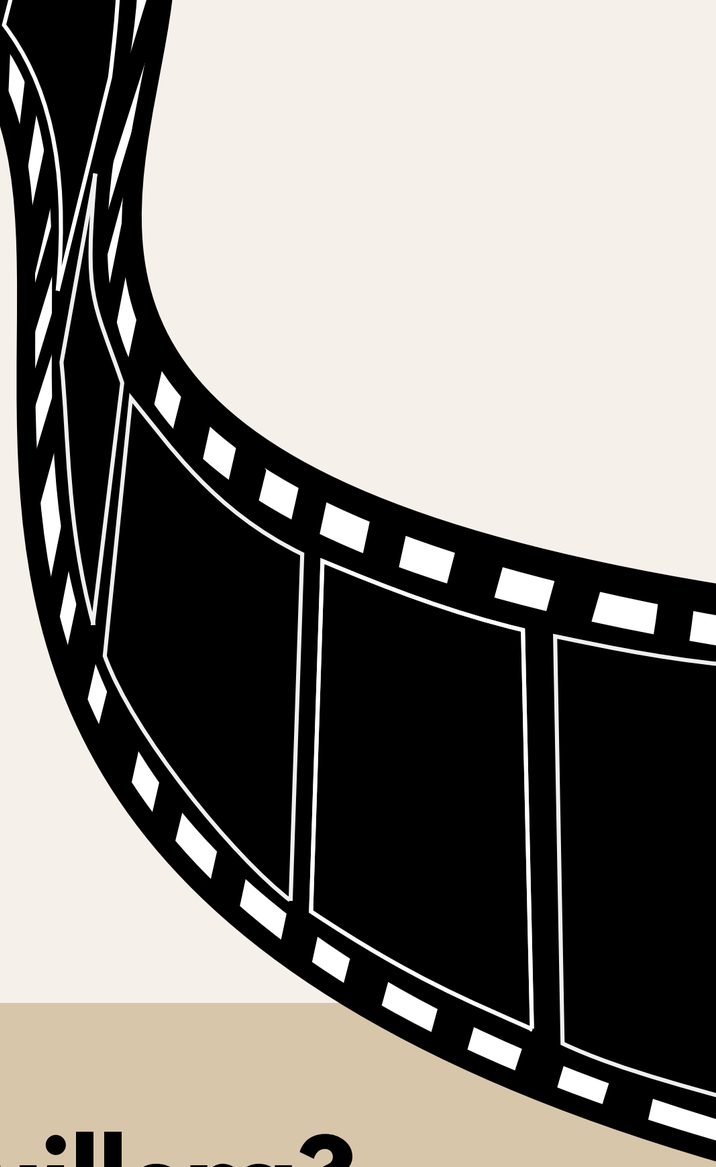
Introducción a Ciencia de Datos

Industria del Cine

Joan Espada

03/07/23



¿Como hacer una película taquillera?

¿Como hacer una película taquillera?

¿De que depende el éxito de una pelicula hoy en dia?

¿Tiempo de pelicula?

¿Director? ¿Presupuesto?

¿Puntuación ? ¿Genero?

Pelicula Exitosa

=

Beneficio Económico

↓

Ganancia Real

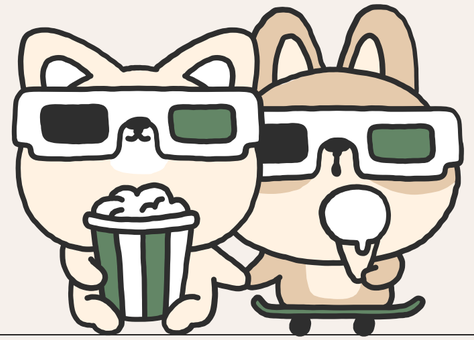
¿Que datos usaremos?

Dataset:

Movie industry

- 1980 -- 2020
- 7800 Peliculas
- **IMDb**
  220 mejores peliculas por año
- Fuente:

https://www.kaggle.com/datasets/danielgrijalvas/movies



15 Columnas:

Numéricas	Categóricas
budget	genre
gross	director
year	writer
runtime	star
votes	country
score	company
released	rating
	name

Limpieza / Filtrado de datos

Eliminamos N/A:

budget
gross
votes
score

Filtramos Cantidad:

Genero > 100 Películas

7800 Películas

↓

5400 Películas

Nuevas variables

Puntuación

↓

Mala (0 : 4)
Media (4 : 7)
Buena (7 : 10)

Ganancia Real

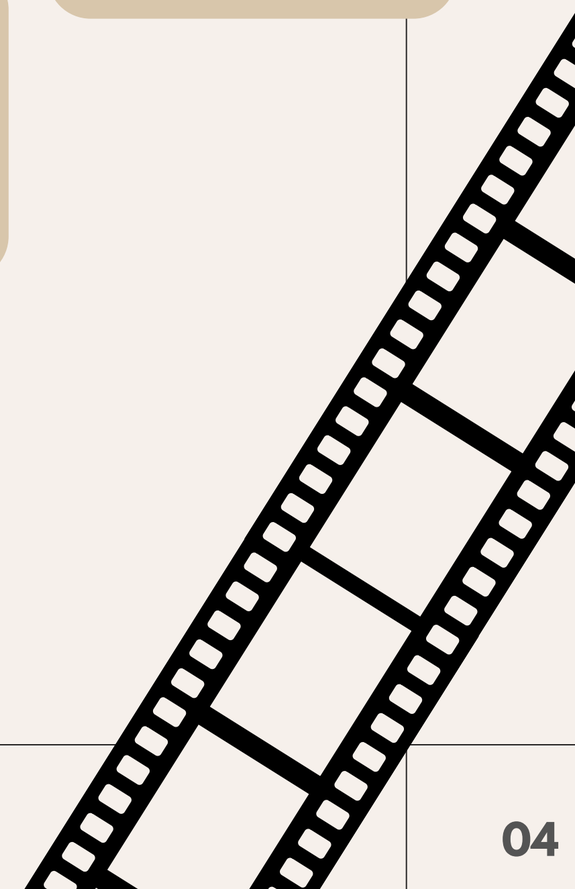
↓

Ganancia - Presupuesto

Popularidad

↓

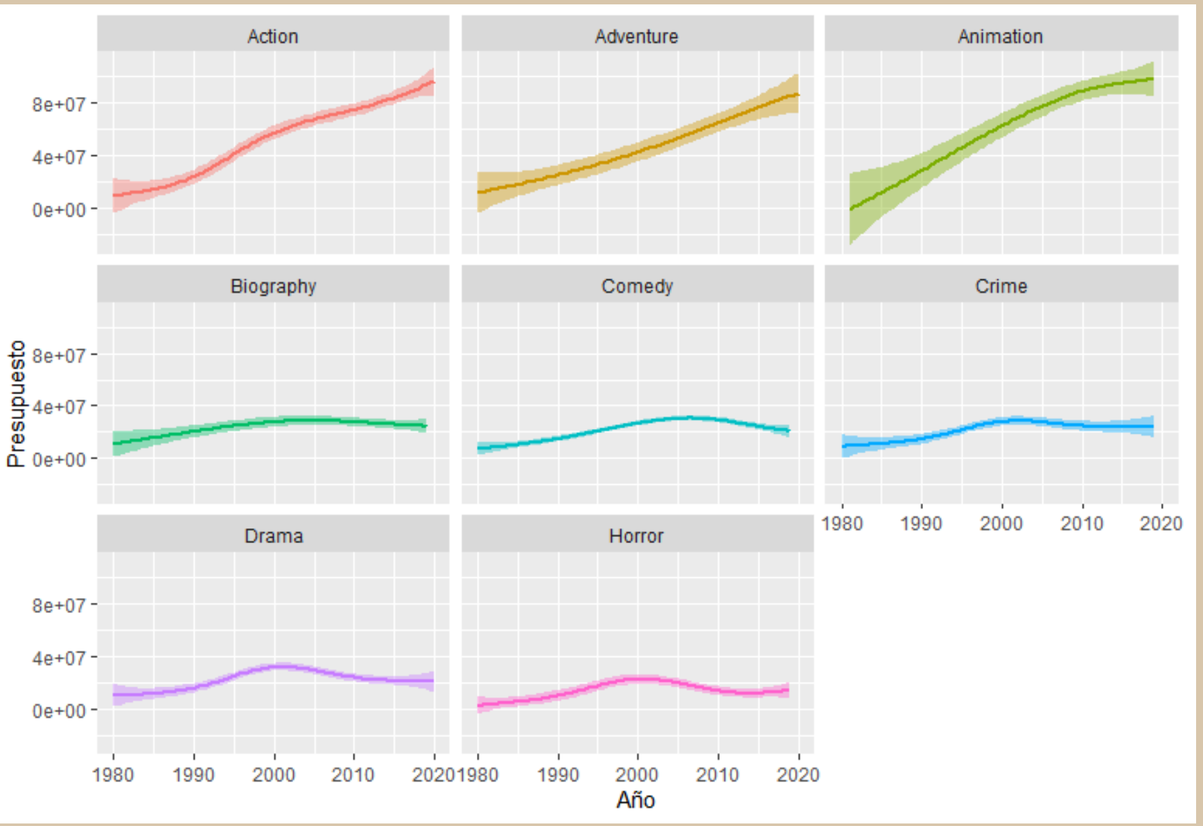
No populares (0: 27000)
Populares (27000: 92100)
Muy Populares (92100: 2500000)

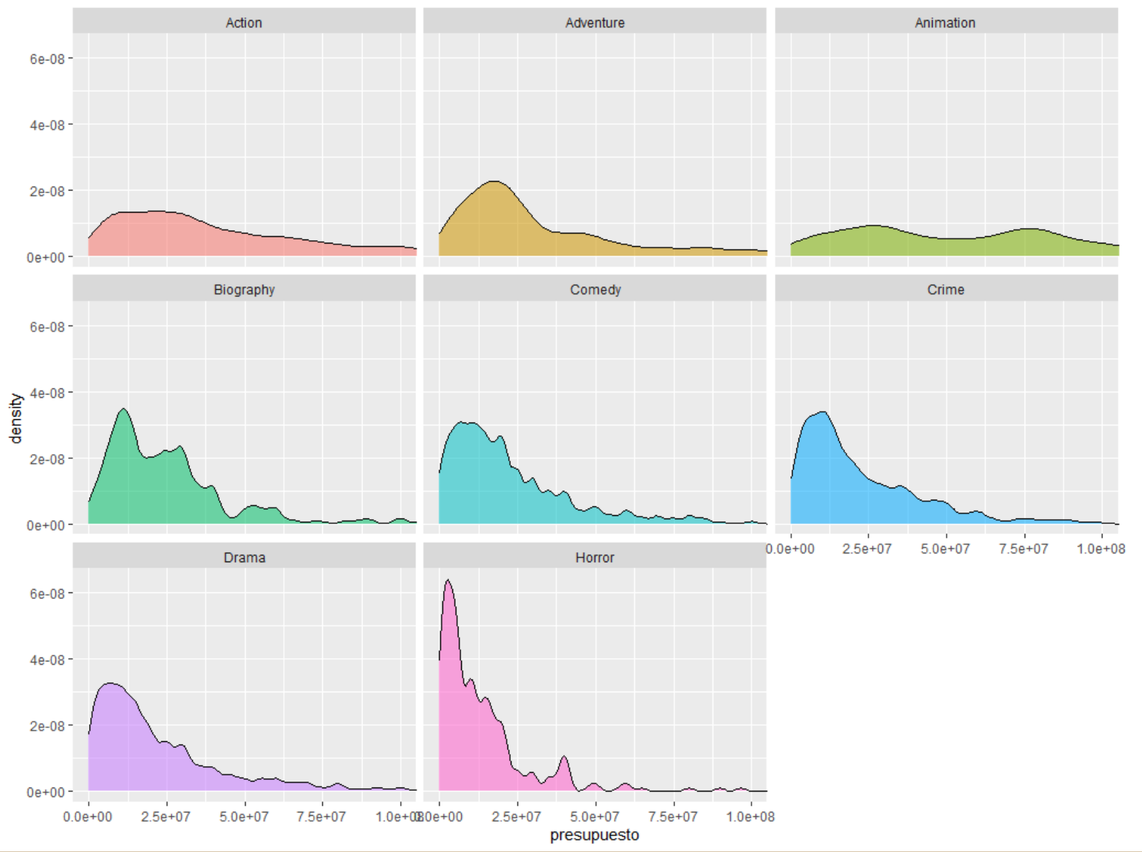


Análisis de Variables

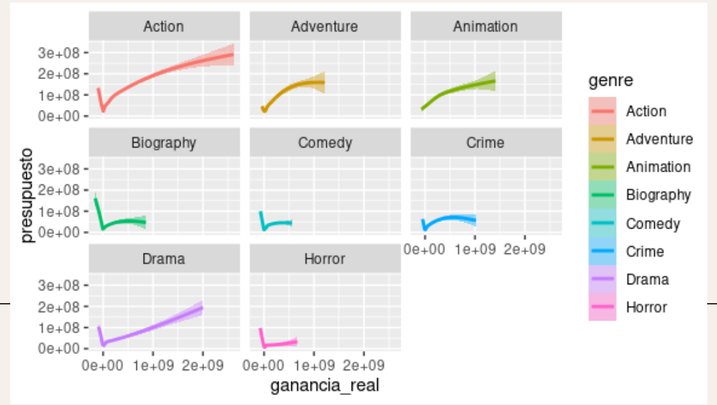
Presupuesto en los generos

¿En que géneros se invierte mas?





Como se distribuye esta inversion?



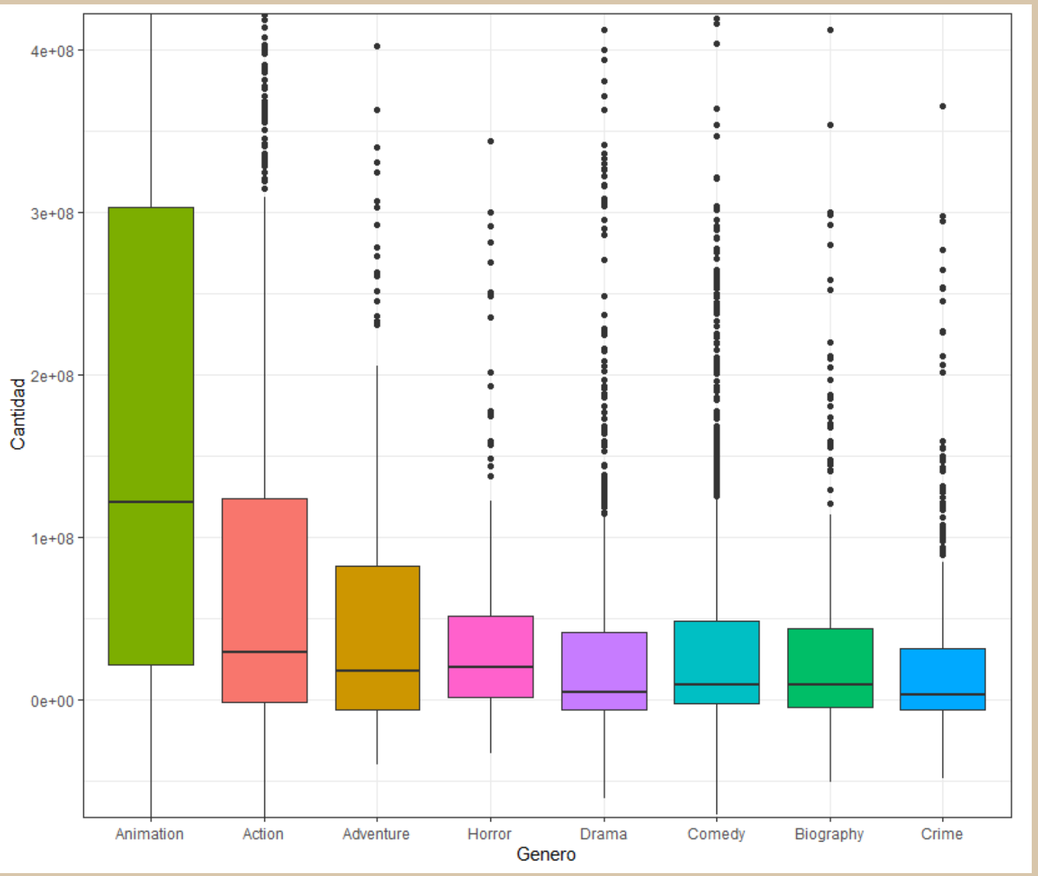
Relación ascendente: Acción, Aventura, Animación

Relación constante: Biografía, Comedia, Crimen, Drama, Año

Análisis de Variables

Ganancia real segun genero

¿Qué géneros son exitosos en taquilla?



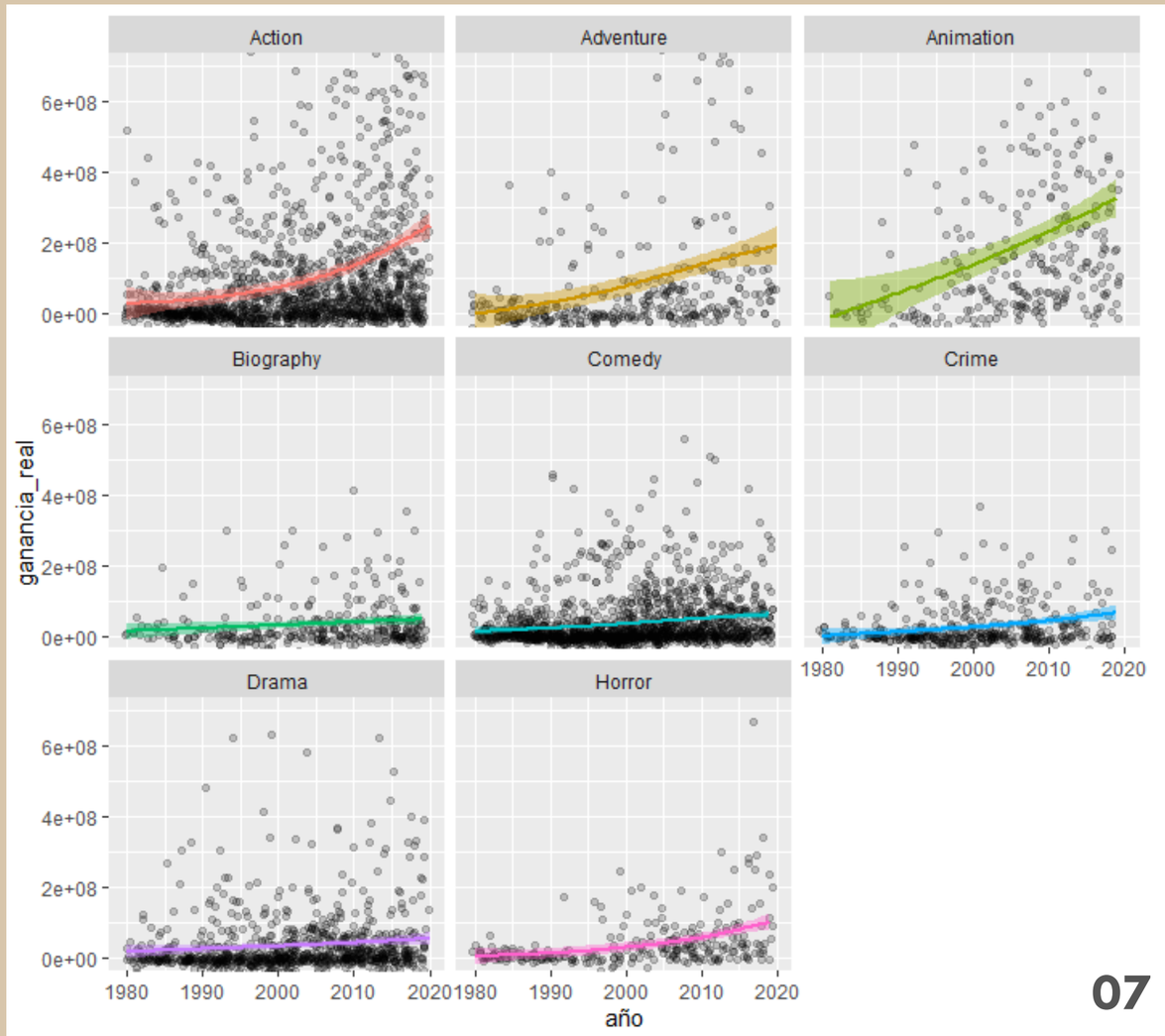
Se destacan: Animación, acción y aventura

Ya nos damos una idea de como va a ser el modelado...

Otra vez...

Relación ascendente: Acción, Aventura, Animación

Relación constante: Biografía, Comedia, Crimen, Drama, Horror



Análisis de Variables:

Variable Director

2100 directores distintos

10 directores con mas ganancia real generada

Director	Cantidad	Ganancia
Anthony Russo	5	903337496
Peter Jackson	11	712182865
David Yates	7	684454096
Michael Bay	13	309220945
James Cameron	7	263882411
J.J. Abrams	6	262922372
Steven Spielberg	27	238993951
Christopher Nolan	11	223661946
Jon Favreau	9	190853736
Chris Columbus	13	131497209

Variable Actor

1800 actores distintos

10 actores con mas ganancia real generada

Actor	Cantidad	Ganancia
Robert Downey Jr.	19	9557937746
Tom Hanks	38	7607944903
Tom Cruise	34	7000252559
Daniel Radcliffe	11	6575311520
Will Smith	23	5657179234
Leonardo DiCaprio	20	5391867648
Johnny Depp	31	4711625354
Vin Diesel	15	4625312568
Dwayne Johnson	22	4099270624
Ben Stiller	24	3613015011

Problemas con el modelado:

- Muy pocas entradas
- Poca significancia
- Demasiadas categorías
- No ajusta

Descartamos

Modelado I

Ganancia Real respecto al año de salida

Modelo Cuadratico

Filtrado por:

- Acciòn
- Comedia
- Drama

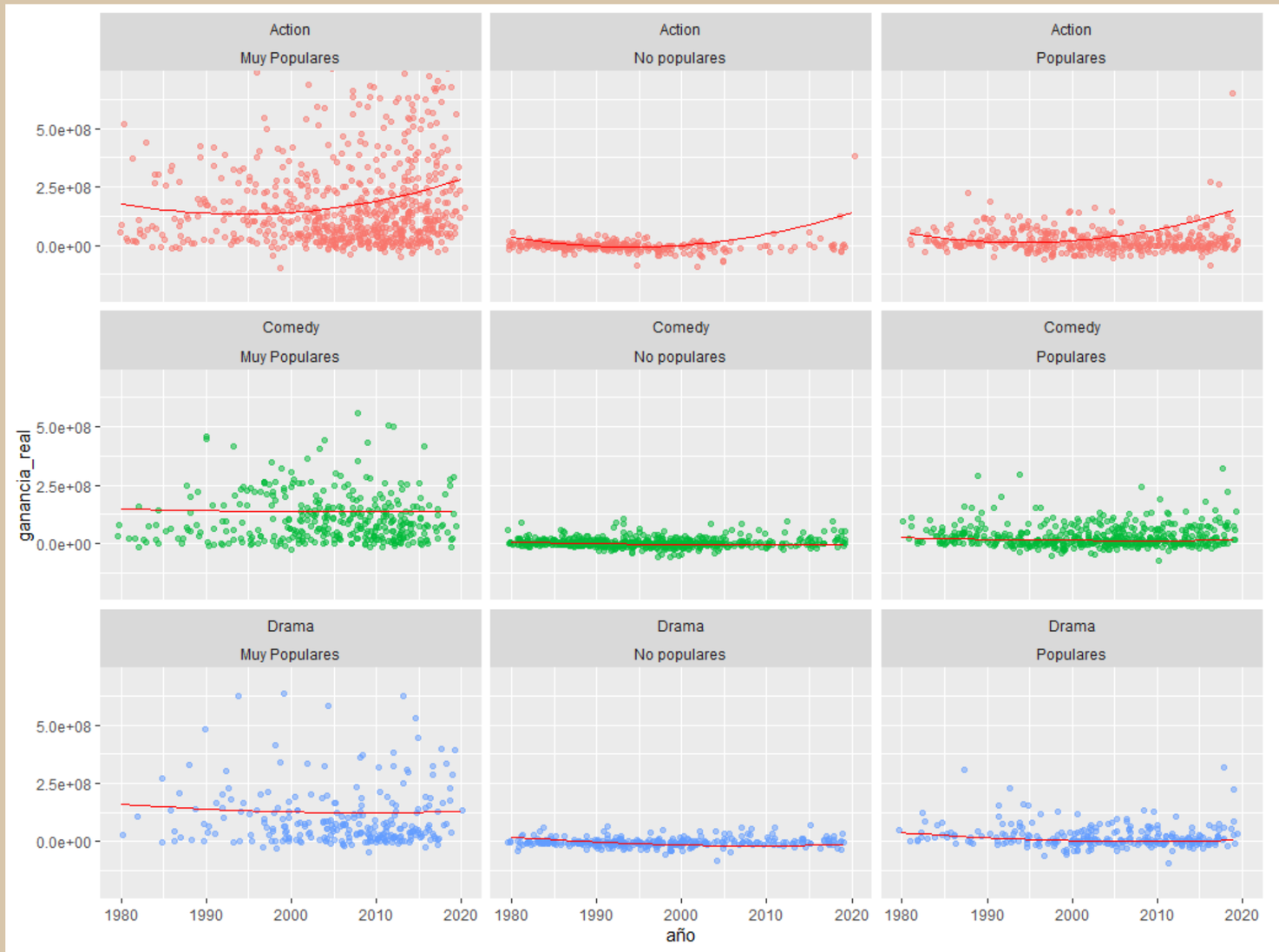
```
lm(formula = ganancia_real ~ poly(año, 2) * genre + popularidad -
    1, data = cuad_sum_genero)
```

Ajustan mejor cuadráticamente

```
Residuals:
       Min         1Q     Median         3Q        Max
-276376747  -43583863   -5808187   15127559 2427375941

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
poly(año, 2)1                1.855e+09  2.256e+08   8.223 2.70e-16 ***
poly(año, 2)2                1.450e+09  2.186e+08   6.633 3.75e-11 ***
genreAction                  1.695e+08  4.505e+06  37.631  < 2e-16 ***
genreComedy                  1.378e+08  5.217e+06  26.425  < 2e-16 ***
genreDrama                   1.314e+08  5.862e+06  22.416  < 2e-16 ***
popularidadNo populares     -1.415e+08  6.093e+06 -23.218  < 2e-16 ***
popularidadPopulares        -1.228e+08  5.561e+06 -22.081  < 2e-16 ***
poly(año, 2)1:genreComedy   -2.010e+09  3.131e+08  -6.420 1.53e-10 ***
poly(año, 2)2:genreComedy   -1.353e+09  3.116e+08  -4.343 1.44e-05 ***
poly(año, 2)1:genreDrama    -2.330e+09  3.602e+08  -6.468 1.12e-10 ***
poly(año, 2)2:genreDrama    -1.137e+09  3.605e+08  -3.153  0.00163 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 135600000 on 3771 degrees of freedom
Multiple R-squared:  0.359,     Adjusted R-squared:  0.3571
F-statistic:   192 on 11 and 3771 DF,  p-value: < 2.2e-16
```



Modelado I

Ganancia Real respecto al año de salida

Modelo Lineal

Filtrado por:

- Aventura
- Animación
- Biografia
- Horror
- Crimen

```
lm(formula = ganancia_real ~ año * genre + popularidad - 1,
    data = lin_sum_genero)
```

Ajustan mejor linealmente

```
Residuals:
       Min         1Q     Median         3Q        Max
-414868778  -40542817   -5373469   23622118 1108071024

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
año                      3.523e+06  6.273e+05   5.617 2.30e-08 ***
genreAdventure          -6.740e+09  1.256e+09  -5.368 9.17e-08 ***
genreAnimation          -1.786e+10  1.614e+09 -11.065  < 2e-16 ***
genreBiography           2.434e+08  1.350e+09   0.180 0.856921
genreCrime               7.801e+08  1.351e+09   0.577 0.563821
genreHorror             -2.512e+09  1.329e+09  -1.890 0.058996 .
popularidadNo Populares -2.798e+08  1.175e+07 -23.801  < 2e-16 ***
popularidadPopulares    -2.237e+08  1.096e+07 -20.409  < 2e-16 ***
año:genreAnimation       5.585e+06  1.015e+06   5.504 4.34e-08 ***
año:genreBiography      -3.517e+06  9.082e+05  -3.872 0.000112 ***
año:genreCrime          -3.787e+06  8.995e+05  -4.210 2.70e-05 ***
año:genreHorror         -2.130e+06  9.031e+05  -2.359 0.018445 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 123400000 on 1560 degrees of freedom
Multiple R-squared:  0.5379,    Adjusted R-squared:  0.5343
F-statistic: 151.3 on 12 and 1560 DF,  p-value: < 2.2e-16
```

Modelado II

Ganancia Real respecto al presupuesto

```
lm(formula = ganancia_real ~ poly(presupuesto, 2) * genre + popularidad +
    punt_categorica - 1, data = sum_genero)
```


```
Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
poly(presupuesto, 2)1                    5.305e+09  1.900e+08  27.922  < 2e-16 ***
poly(presupuesto, 2)2                    2.842e+09  1.592e+08  17.855  < 2e-16 ***
genreAction                              1.420e+08  4.682e+06  30.334  < 2e-16 ***
genreAdventure                           1.479e+08  7.144e+06  20.702  < 2e-16 ***
genreAnimation                           1.900e+08  9.629e+06  19.728  < 2e-16 ***
genreBiography                           9.399e+07  9.792e+06   9.599  < 2e-16 ***
genreComedy                              1.289e+08  6.037e+06  21.356  < 2e-16 ***
genreCrime                               1.388e+08  1.331e+07  10.422  < 2e-16 ***
genreDrama                               1.513e+08  5.549e+06  27.256  < 2e-16 ***
genreHorror                              1.142e+08  2.358e+07   4.844 1.31e-06 ***
popularidadNo populares                 -9.102e+07  4.843e+06 -18.794  < 2e-16 ***
popularidadPopulares                    -7.909e+07  4.273e+06 -18.510  < 2e-16 ***
punt_categoricaMalas                    -4.249e+07  1.169e+07  -3.636  0.00028 ***
punt_categoricaMedia                    -2.861e+07  4.187e+06  -6.834 9.21e-12 ***
poly(presupuesto, 2)1:genreAdventure     9.591e+08  4.244e+08   2.260  0.02386 *
poly(presupuesto, 2)2:genreAdventure    -9.125e+08  4.983e+08  -1.831  0.06714 .
poly(presupuesto, 2)1:genreAnimation     1.051e+09  4.305e+08   2.441  0.01470 *
poly(presupuesto, 2)2:genreAnimation    -2.022e+09  4.805e+08  -4.208 2.62e-05 ***
poly(presupuesto, 2)1:genreBiography    -8.282e+09  1.870e+09  -4.429 9.66e-06 ***
poly(presupuesto, 2)2:genreBiography    -5.903e+09  1.615e+09  -3.655  0.00026 ***
poly(presupuesto, 2)1:genreComedy       -4.990e+09  1.098e+09  -4.543 5.67e-06 ***
poly(presupuesto, 2)2:genreComedy       -4.295e+09  8.657e+08  -4.961 7.23e-07 ***
poly(presupuesto, 2)1:genreCrime         2.545e+09  3.126e+09   0.814  0.41559
poly(presupuesto, 2)2:genreCrime         1.341e+09  2.341e+09   0.573  0.56677
poly(presupuesto, 2)1:genreDrama         5.069e+09  8.823e+08   5.745 9.70e-09 ***
poly(presupuesto, 2)2:genreDrama         3.779e+09  7.512e+08   5.030 5.06e-07 ***
poly(presupuesto, 2)1:genreHorror       -9.495e+09  5.632e+09  -1.686  0.09188 .
poly(presupuesto, 2)2:genreHorror       -5.607e+09  3.786e+09  -1.481  0.13873
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 112600000 on 5326 degrees of freedom
Multiple R-squared:  0.5766,    Adjusted R-squared:  0.5744
F-statistic: 259.1 on 28 and 5326 DF,  p-value: < 2.2e-16
```

Conclusiones

El Exito es dificil de medir

- Modas cambiantes
- Industria Impredecible
- Factores dificiles de predecir...

Modelos

- Ajuste Cuadrado bajo
- Pocas variables significativas
- Gran margen de error

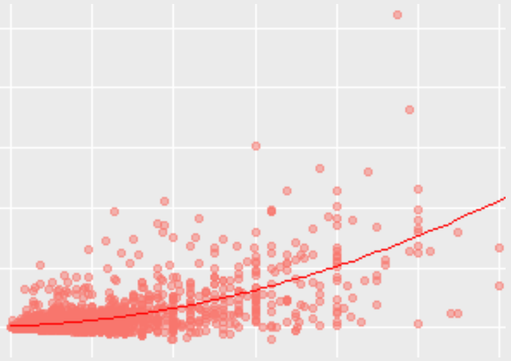
Aun así...¿Qué descubrimos?

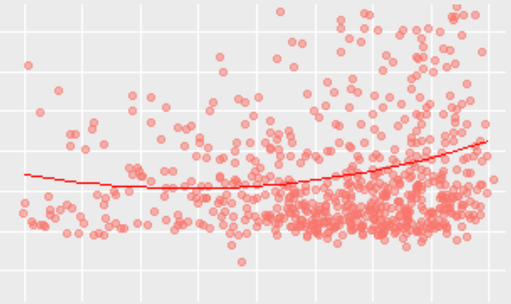
Conclusiones

¿Qué descubrimos?

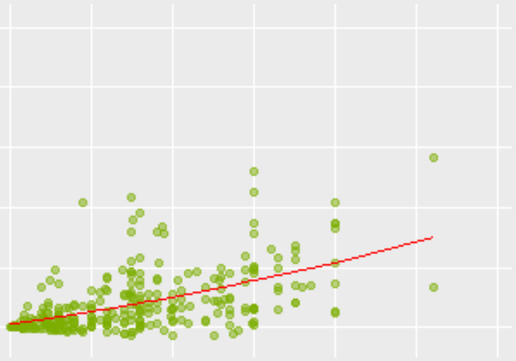
Coincidencias en los generós:

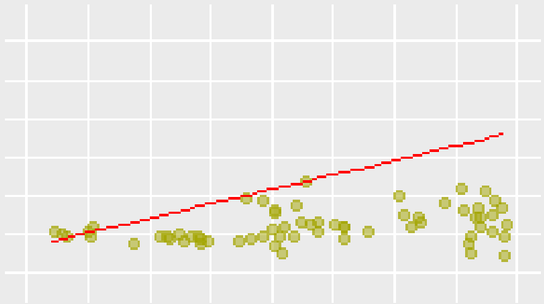
Acción



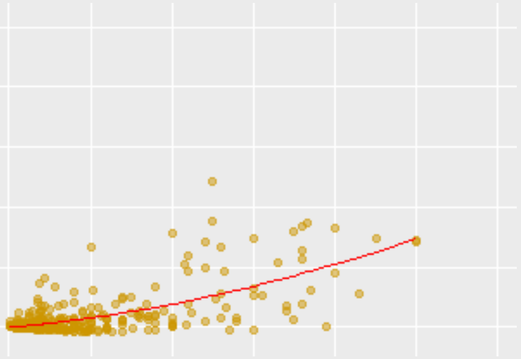


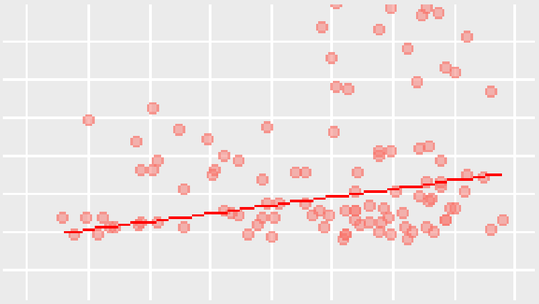
Animación





Aventura





Modelado Presupuesto

Mas presupuesto indica una mayor ganancia en estos generós.

Vale la pena invertir en películas de este tipo.

Modelado Tiempo

Se ve una tendencia en la actualidad en estos generos.

Podemos predecir que será así por los próximos años.

Trabajo a Futuro

Aspecto a Mejorar:

En el dataset:

- Mostrar división de gastos en el presupuesto.
- Incluir Datos de los gastos en los distintos aspectos de la películas.
- Mas entradas de peliculas
- Rango de tiempo mas amplio
- Mas metricas de evaluacion

En el analisis:

- Poder llegar a un modelo mas preciso


END
THE END
THE


¡Gracias por escuchar!
¿Preguntas?
joanespadaarg@gmail.com
ezecoggiola@gmail.com

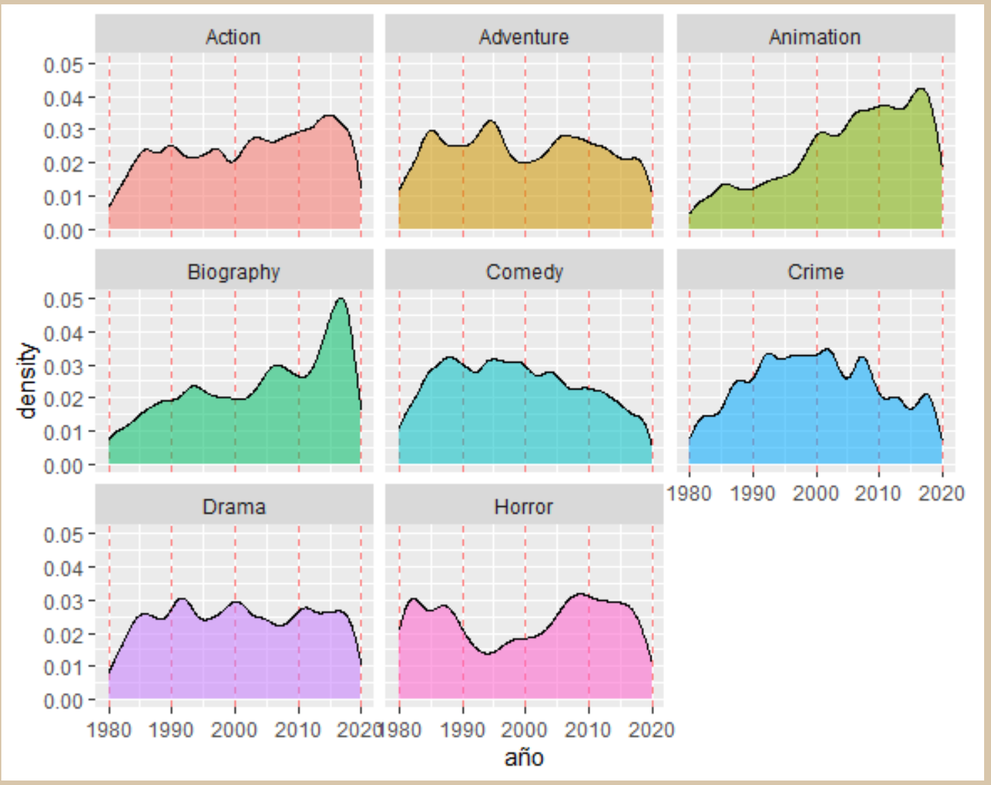
Variables que se quedaron atras....

Variable	Compania	Ranking	Pais	Duración de la película	
R^2	-0.0157	+0.0192	+0.0045	+0.0205	

Análisis de Variables

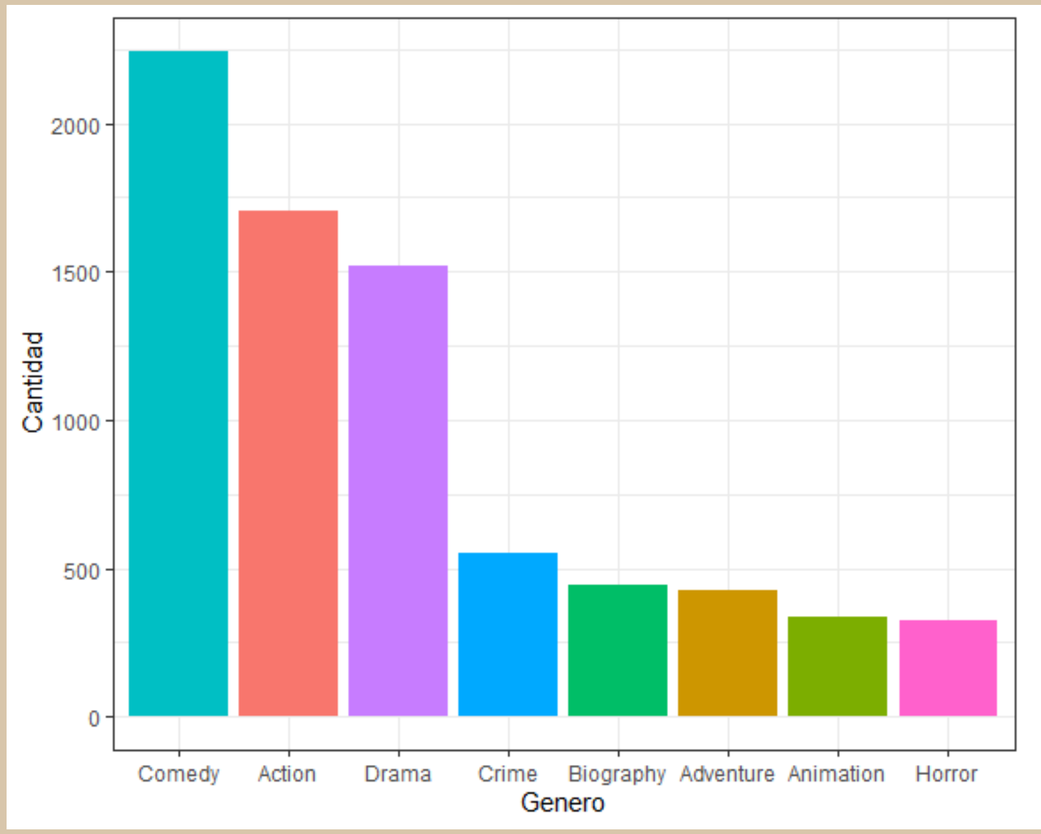
Los generos

Distribución películas x año en cada genero



Genero	Cantidad
Comedy	2245
Action	1705
Drama	1518
Crime	551
Biography	443
Adventure	427
Animation	338
Horror	322

Cantidad de peliculas en cada genero



Peliculas destacadas:

Aventura : The Hobbit

Accion : Mad Max, RoboCop, Predator

Biografía : Goodfellas, Schindler's List

Animación : Cars 3, Toy Story 3

Drama : Gran Torino, 1917

Crimen : Seven, Joker

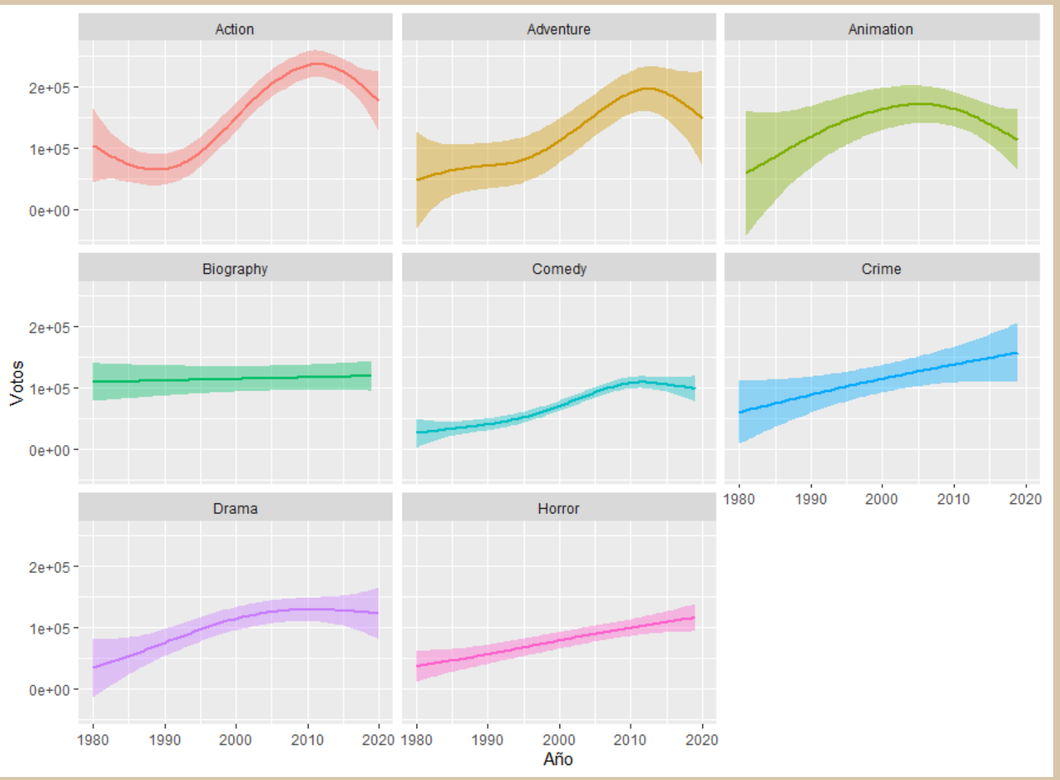
Comedia : Scary Movie, American Pi

Horror : Scream, Halloween

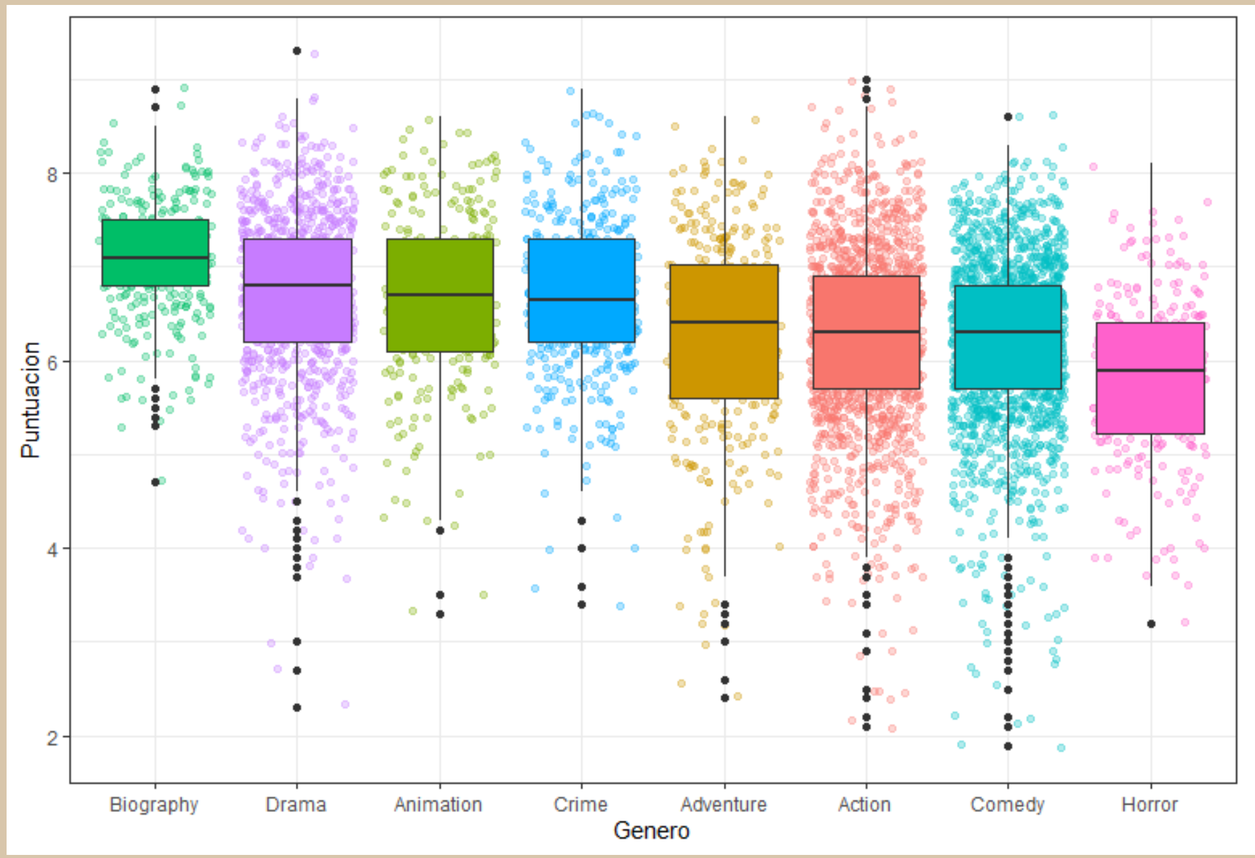
Análisis de Variables

Votos y puntuacion

Cantidad de votos al pasar los años



Puntuación x Genero

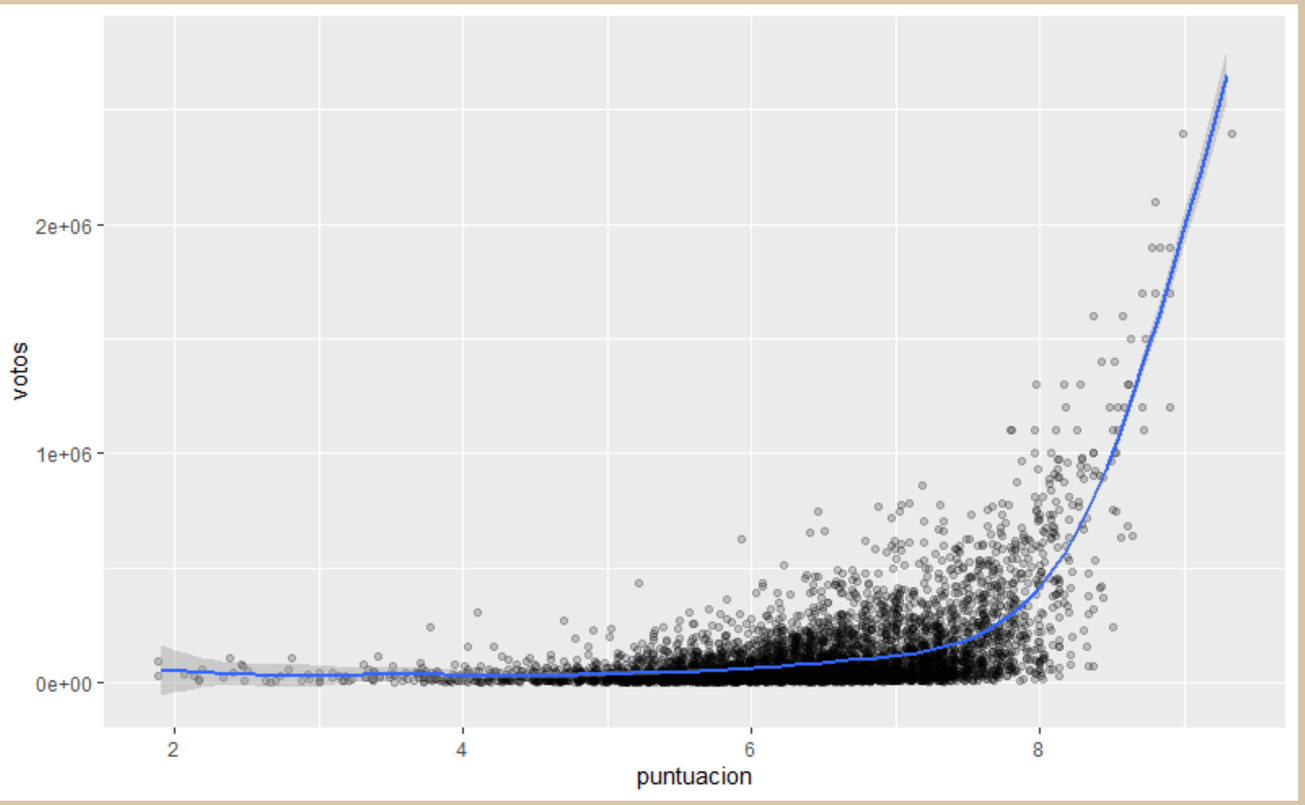


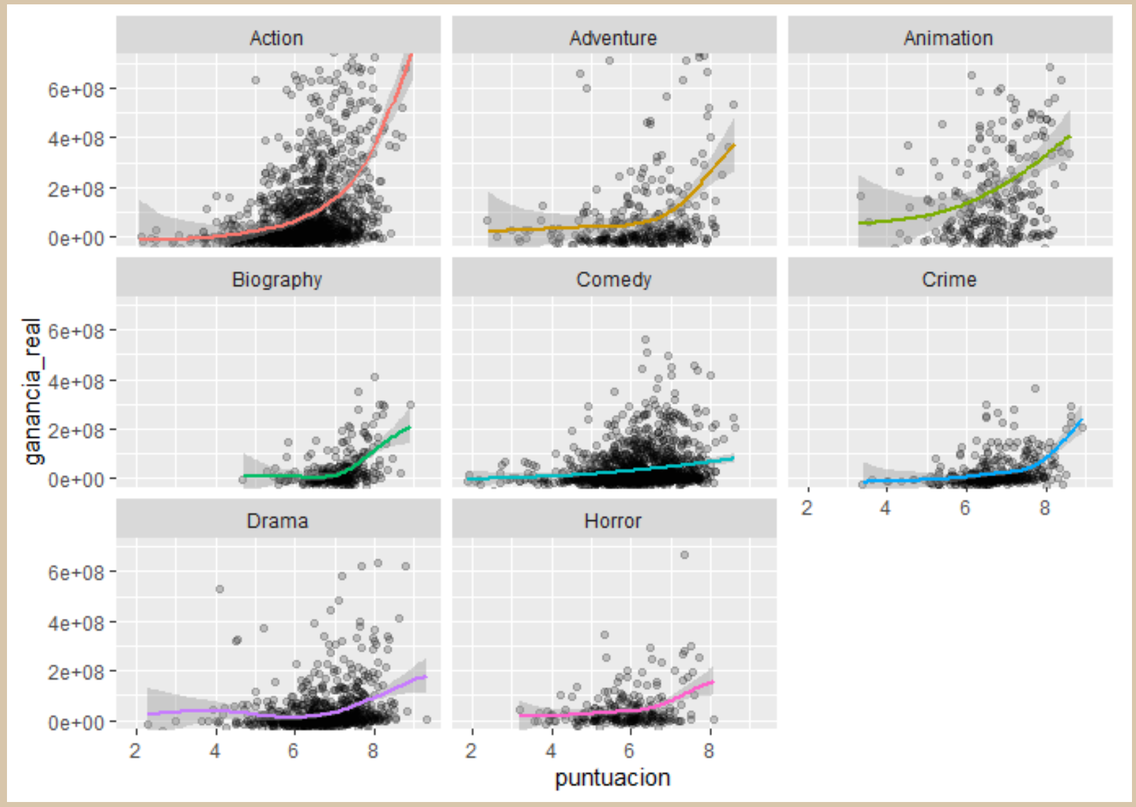
DIsclaimer: Imdb (Internet Movie Database), es una base de datos de películas en Internet que fue creada en 1990, Y se observa que los votos de películas mas viejas a esta fecha puede no ser muy informativa, ya que es mas raro que el usuario vote estas películas.

Análisis de Variables

Votos en los distintos generos

Relación ascendente entre Votos y Puntuación





Relación entre Puntuación y Ganancia Real