Estudio estadístico en R

Tema: Ranking de peliculas

Fuente:
https://raw.githubusercontent.com/cienciadedatos/datos-de-miercoles/master/datos/2020/2020-02-19/ranking_imdb.csv

Descripción del problema y estructura de datos utilizados

★ Se analizará qué sucede en la industria cinematográfica.

★ Conjunto de datos: Valoraciones de películas en el sitio web ***IMDB*** (Internet Movie Data Base).

★ La base contiene 10000 observaciones de películas.

Variable	Tipo	Clase	Descripción
ranking	Cuantitativa	Numérico (num)	Posición de la película en el ranking de IMBD
titulo	Cualitativa	Caracter (chr)	Título de la pelicula en ingles
anio	Cuantitativa	Numérico (num)	Año de estreno
puntaje	Cuantitativa	Numérico (num)	Puntaje de la pelicula del 1 al 10
genero	Cualitativa	Numérico(num)	Género o géneros a los que pertenece la película
votos	Cuantitativa	Caracter (chr)	Cantidad de votos recibidos
direccion	Cualitativa	Caracter (chr)	Nombre del director/a
duracion	Cuantitativa	Numérico (num)	Duración de película (min)
ganancias	Cuantitativa	Numérico (num)	Ganancias de la película en millones de dólares

Preguntas problemas

- ¿Los directores se inclinan por determinados tipos de géneros al realizar películas?
- ¿Las ganancias de las películas dependen del género?
- ¿El año de filmación de la película determina su duración?

Variable dependiente elegida: **duración**

Univariada

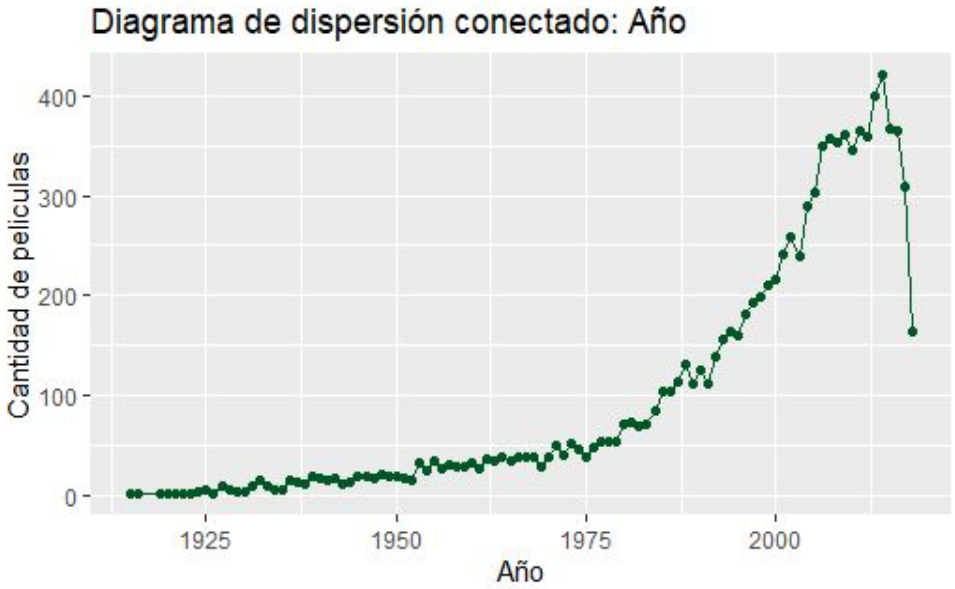
Variable	Media	Varianza	Desviación Típica	Coeficiente de Variación	Coeficiente Asimetría	Coeficiente Curtosis
anio	1998	333.05	18.25	0.0091	-1.56	5.27
duración	108.7	469.15	21.66	0.20	2.11	16.59
ganancias	15.09	3706.62	60.88	1.68	4.20	31.68

Multivariada (Bivariada)

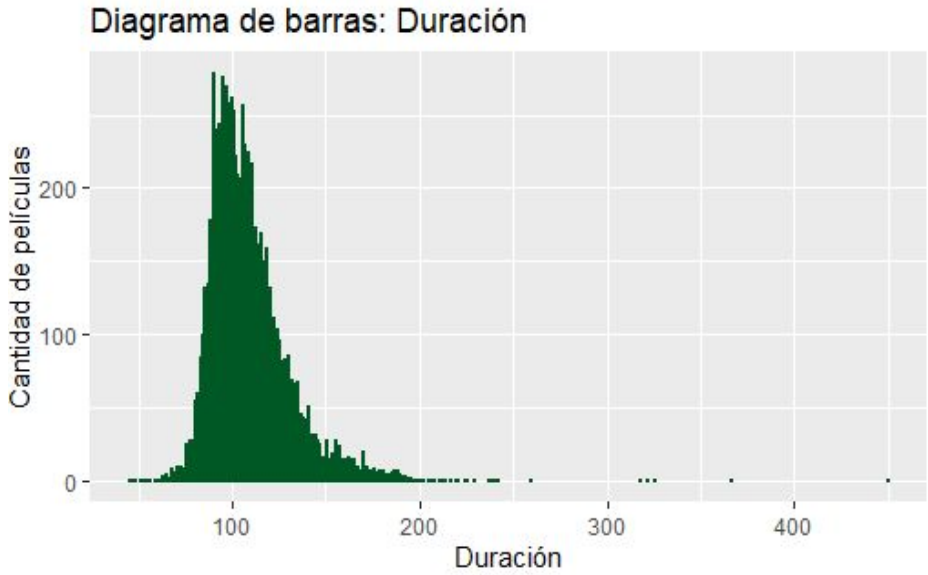
Coef. Correlación	duracion	ganancias
anio	0.001	0.084

Covarianza	duracion
anio	0.69

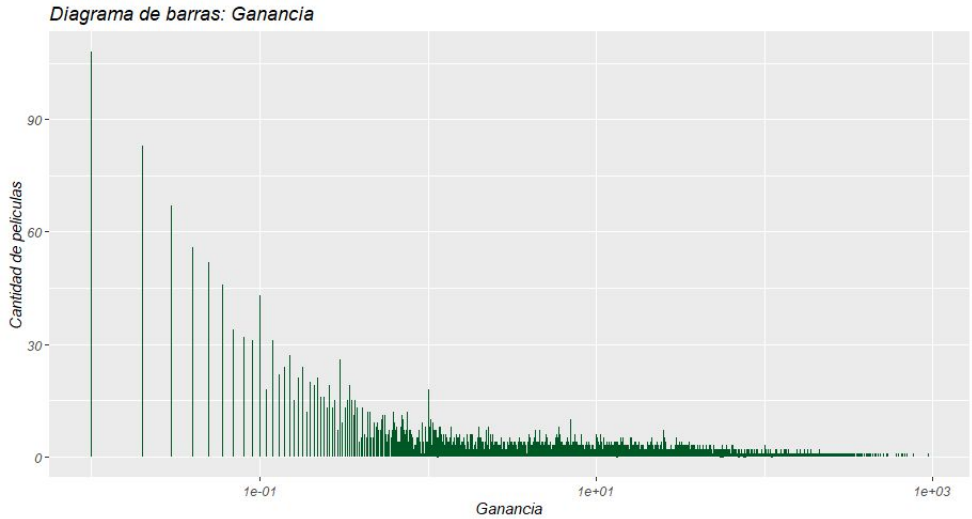
Algunos gráficos univariados:

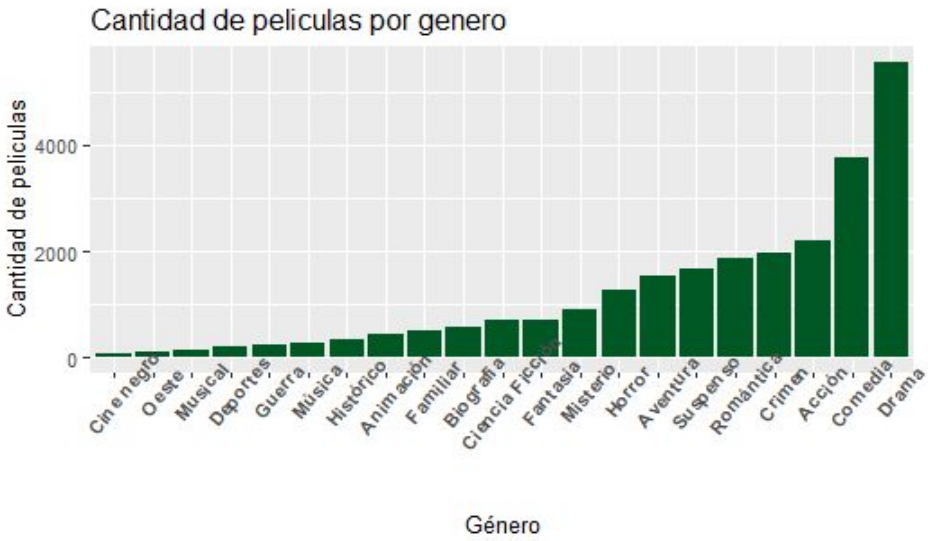


- Con el paso de los años cada vez se producen más películas
- A partir de 1980 hay un “boom” en la cantidad de películas producidas por año
- La curva comienza a crecer rápidamente hasta alcanzar su pico en 2015, después comienza a descender



- Alta concentración de datos cerca de la media
- Un valor se considera atípico cuando sea mayor a 154 min
- Los que más se destacan son los que superan los 300 min, superando las 5 horas de duración, hasta llegar a durar 450 min, lo cual son 7.5 hs

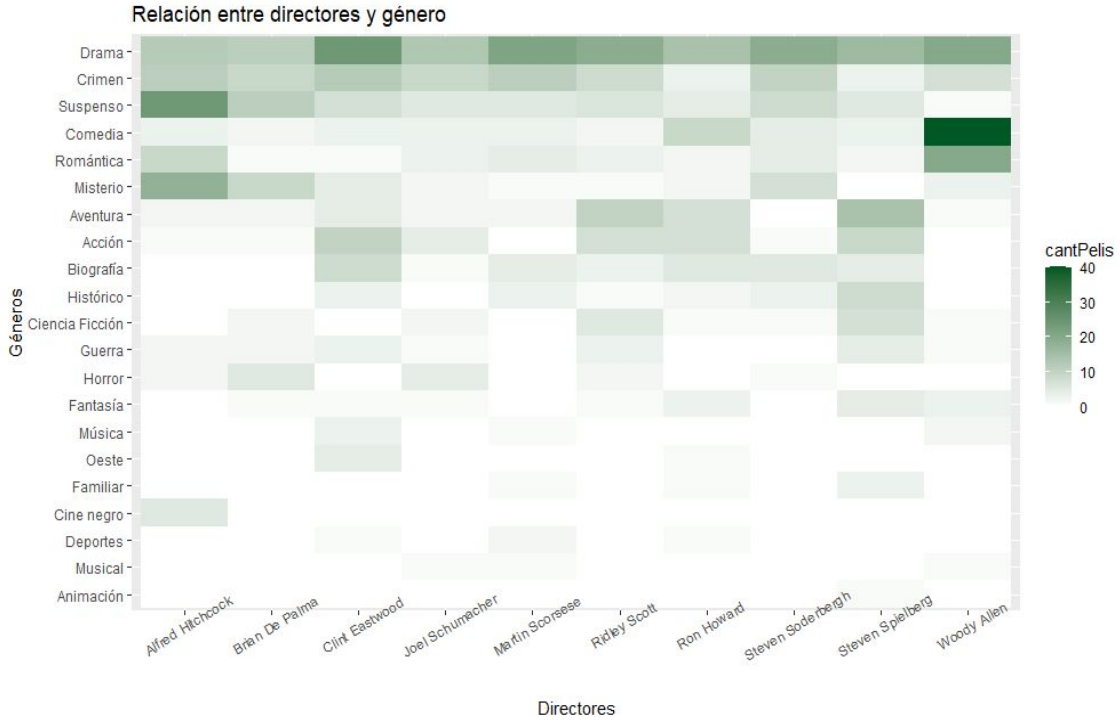




- Se utilizó una **Escala logarítmica** para una mejor visualización de la información.

- Los 2 géneros en los cuales hay más películas son Drama y Comedia.
- Los mismos se llevan más de la mitad de las películas registradas.

¿Los directores se inclinan por determinados tipos de géneros al realizar películas?



★ Gráfico utilizado: **geom_tile().**

★ Variable ***genero***: contiene más de un género por película

titulo	***genero***
WALL-E	Animación, Aventura, Familiar

→

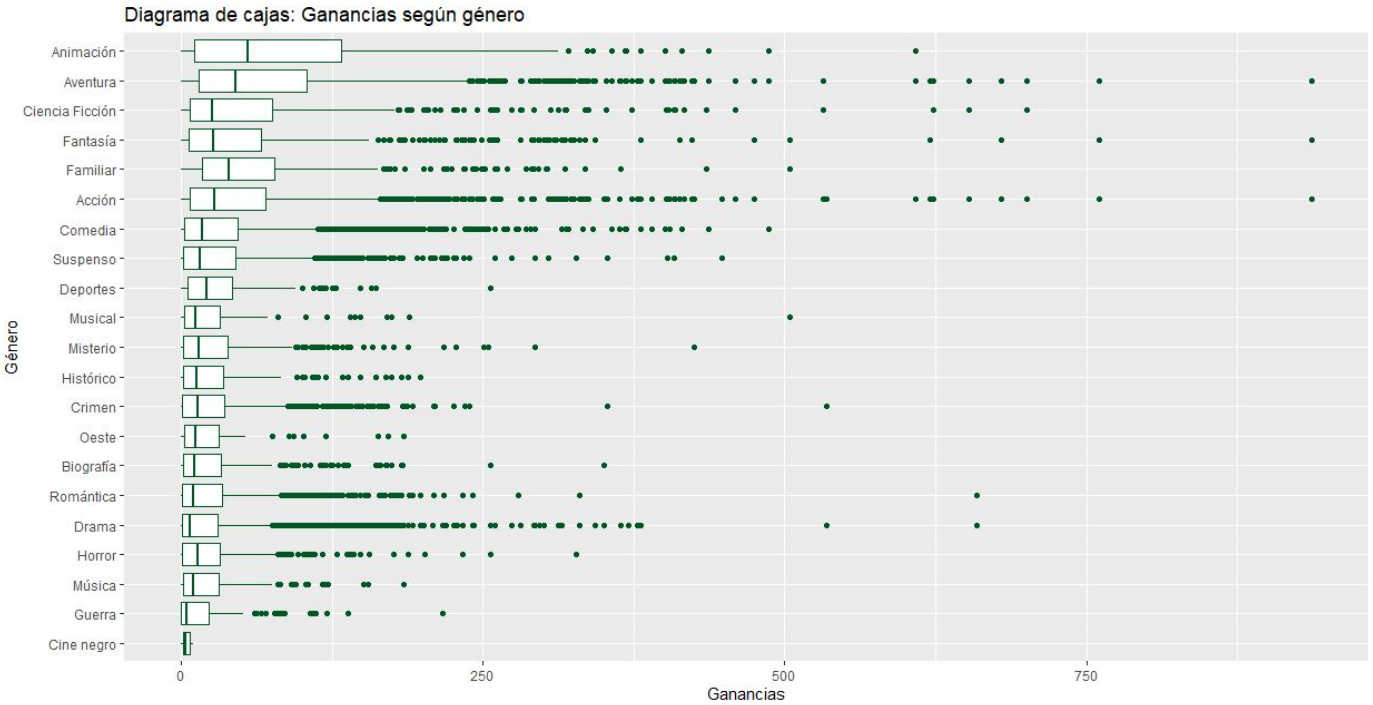
titulo	***genero***
WALL-E	Animación
WALL-E	Aventura
WALL-E	Familiar

★ Se realiza tratamiento de datos.

★ Código utilizado:

```
# ALGORTIMO GENEROS:
# Se detallan, de manera separada, cada uno de los generos que aparecen en el data.frame
# debido a que encontramos peliculas que pertenecen a mas de un genero
res <- NULL                                # Se incializa el arreglo de caracteres resultado
ngen <- strsplit(peliculas$genero, ", ")  # Se guarda en ngen todos los generos que contiene cada
pelicula
for (i in 1:length(ngen)){                 # Se recorre ngen fila por fila
  for (j in 1:length(ngen[[i]])) {         # Luego de entrar en la fila, se recorre cada genero de la
fila
    if (!ngen[[i]][j]%in%res){             # Si ya no se agrego al conjunto resultado, se agrega
      gen <- as.character(ngen[[i]][j])    # Primero se convierte a caracter
      res <- c(res,gen)                    # Se agrega el genero al resultado
    }
  }
}
res
```

¿Las ganancias de las películas dependen del género?



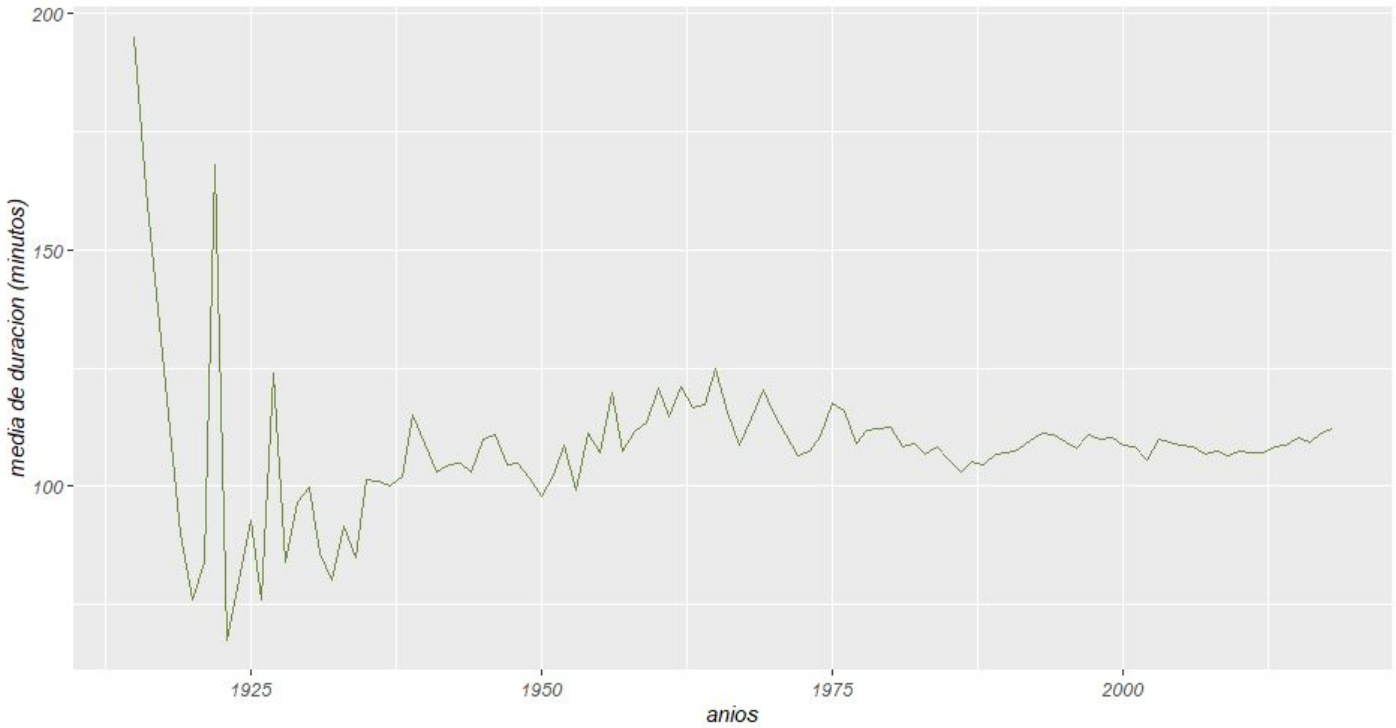
★ Gráfico utilizado: **geom_boxplot().**

★ Variable ***ganancia*** analizada a partir de las películas que no poseen NAs en dicho campo.

★ A partir del **DIAGRAMA DE CAJAS** se observa que:

- todas las películas reciben una ganancia similar (*mediana* alrededor de los 41 millones de dólares en general)
- algunos géneros poseen una ganancia mayor (***Aventura***, ***Familiar*** y ***Animación***)
- los géneros ***Cine Negro*** y ***Animación*** son los que reciben la menor y mayor retribución respectivamente.

¿El año de filmación de la película determina su duración?



★ Gráfico utilizado: **geom_line().**

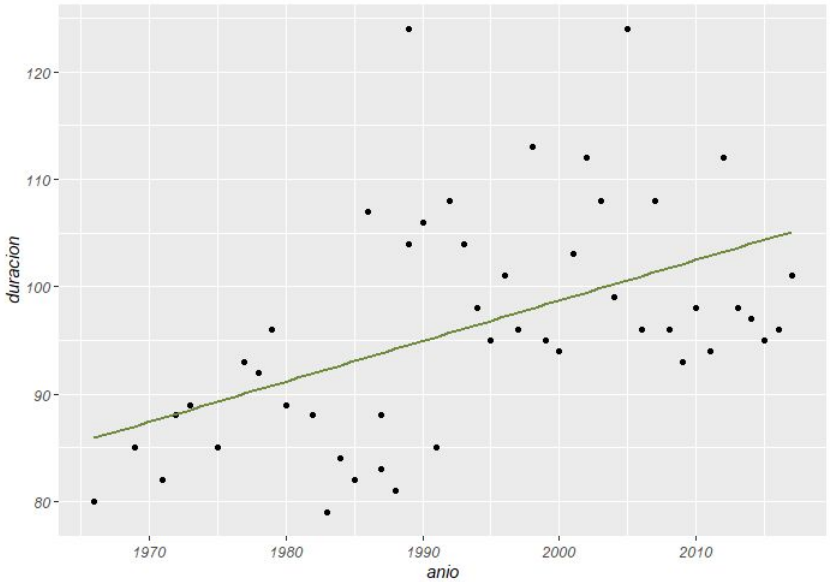
★ A partir del **GRÁFICO TEMPORAL** se observa:

- una tendencia a preservar la duración de las películas con respecto al transcurso del tiempo.
- leve fluctuación de la misma en torno a los 120 minutos (dos horas).
- dos etapas:
  - hasta 1930 aprox.: los años parecen influir, grandes oscilaciones en la duración y una tendencia al aumento.
  - a partir de 1930 aprox.: los años NO parecen influir, tendencia a estabilizarse.

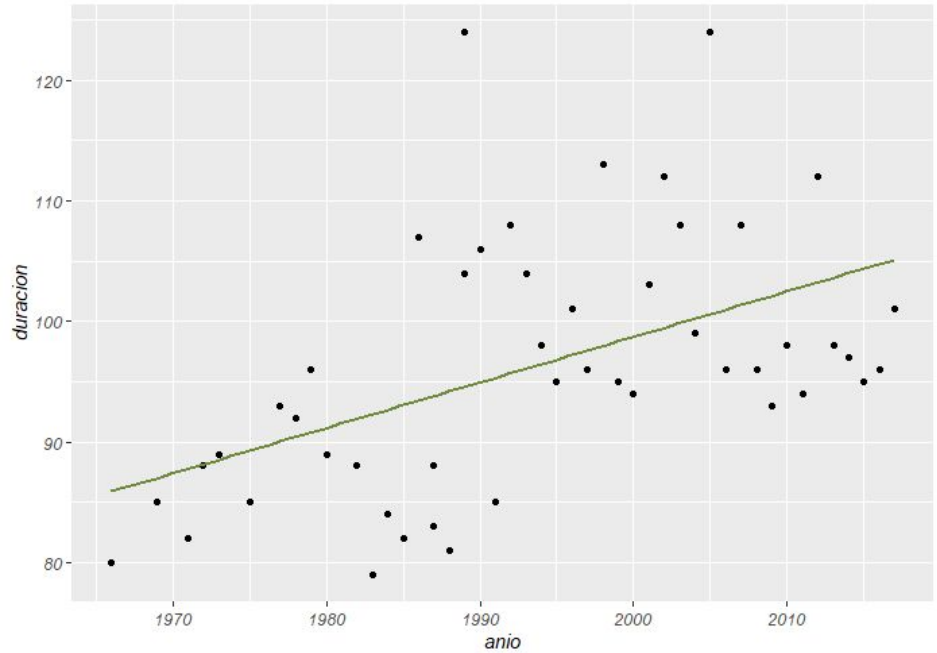
Recta de regresión

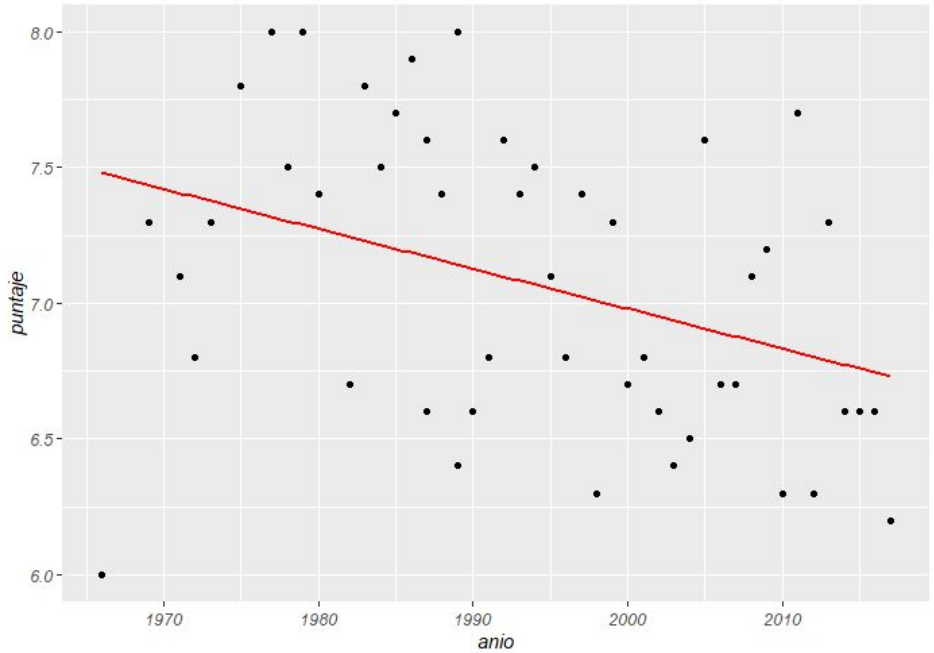
- ★ Variable dependiente: ***duracion*** (Y).
- ★ Variable independiente: ***anio*** (X).
- ★ Director elegido: Woddy Allen.
  - es quien más películas filmadas tiene.
  - 40 de las 48 películas contienen al género **Comedia**.

Y=0.38X-656.26



Duración vs Puntaje - Películas de Woddy Allen





Referencias:

★ **Bibliografía**

- Peña D. (2014) “Fundamentos de Estadística”
- “Introducción a R”, Notas sobre R: Un entorno de programación para Análisis de Datos y Gráficos Versión 1.0.1 (2000-05-16)
- “R4DS - R for Data Science”, Libro del curso. *Garrett Grolemund - Hadley Wickham - Diciembre 2016*

★ **Webgrafía**

- rpubs.com
- stackoverflow.com