Hoja de Trabajo I
Minería de datos
Lynette García

3 de febrero del 2023
Universidad del Valle

Hoja de Trabajo 1

Integrantes:
Alejandra Guzmán Dominguez 20262
Jorge Caballeros Perez 20009
Mariana David Sosa 201055

1. (3 puntos) Haga una exploración rápida de sus datos, para eso haga un resumen de su conjunto de datos.

```
> summary(data)
       id              budget              genres             homePage         productionCompany  productionCompanyCountry
 Min.   :     5   Min.   :        0   Length:10000       Length:10000       Length:10000       Length:10000
 1st Qu.: 12286   1st Qu.:        0   Class :character   Class :character   Class :character   Class :character
 Median :152558   Median :   500000   Mode  :character   Mode  :character   Mode  :character   Mode  :character
 Mean   :249877   Mean   : 18551632
 3rd Qu.:452022   3rd Qu.: 20000000
 Max.   :922260   Max.   :380000000
 productionCountry     revenue               runtime          video           director           actors
 Length:10000       Min.   :0.000e+00   Min.   :  0.0   Mode :logical   Length:10000       Length:10000
 Class :character   1st Qu.:0.000e+00   1st Qu.: 90.0   FALSE:9430      Class :character   Class :character
 Mode  :character   Median :1.631e+05   Median :100.0   TRUE :84        Mode  :character   Mode  :character
                    Mean   :5.674e+07   Mean   :100.3   NA's :486
                    3rd Qu.:4.480e+07   3rd Qu.:113.0
                    Max.   :2.847e+09   Max.   :750.0
 actorsPopularity   actorsCharacter    originalTitle         title          originalLanguage     popularity
 Length:10000       Length:10000       Length:10000       Length:10000       Length:10000       Min.   :    4.258
 Class :character   Class :character   Class :character   Class :character   Class :character   1st Qu.:   14.578
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :   21.906
                                                                                                Mean   :   51.394
                                                                                                3rd Qu.:   40.654
                                                                                                Max.   :11474.647
 releaseDate           voteAvg         voteCount      genresAmount    productionCoAmount productionCountriesAmount
 Length:10000       Min.   : 1.300   Min.   :    1   Min.   : 0.000   Min.   : 0.000     Min.   :  0.000
 Class :character   1st Qu.: 5.900   1st Qu.:  120   1st Qu.: 2.000   1st Qu.: 2.000     1st Qu.:  1.000
 Mode  :character   Median : 6.500   Median :  415   Median : 3.000   Median : 3.000     Median :  1.000
                    Mean   : 6.483   Mean   : 1342   Mean   : 2.596   Mean   : 3.171     Mean   :  1.751
                    3rd Qu.: 7.200   3rd Qu.: 1316   3rd Qu.: 3.000   3rd Qu.: 4.000     3rd Qu.:  2.000
                    Max.   :10.000   Max.   :30788   Max.   :16.000   Max.   :89.000     Max.   :155.000
  actorsAmount    castWomenAmount    castMenAmount
 Min.   :     0   Length:10000       Length:10000
 1st Qu.:    13   Class :character   Class :character
 Median :    21   Mode  :character   Mode  :character
 Mean   :  2148
 3rd Qu.:    36
 Max.   :919590
```

Hoja de Trabajo I
Minería de datos 3 de febrero del 2023
Lynette García Universidad del Valle

El resumen del conjunto de datos es este:

```
##        id             budget             genres            homePage
##  Min.   :     5   Min.   :        0   Length:10000       Length:10000
##  1st Qu.: 12286   1st Qu.:        0   Class :character   Class :character
##  Median :152558   Median :   500000   Mode  :character   Mode  :character
##  Mean   :249877   Mean   : 18551632
##  3rd Qu.:452022   3rd Qu.: 20000000
##  Max.   :922260   Max.   :380000000
##  productionCompany  productionCompanyCountry productionCountry
##  Length:10000       Length:10000             Length:10000
##  Class :character   Class :character         Class :character
##  Mode  :character   Mode  :character         Mode  :character
##
##
##
##     revenue             runtime        video           director
##  Min.   :0.000e+00   Min.   :  0.0   Mode :logical   Length:10000
##  1st Qu.:0.000e+00   1st Qu.: 90.0   FALSE:9430      Class :character
##  Median :1.631e+05   Median :100.0   TRUE :84        Mode  :character
##  Mean   :5.674e+07   Mean   :100.3   NA's :486
##  3rd Qu.:4.480e+07   3rd Qu.:113.0
##  Max.   :2.847e+09   Max.   :750.0
##     actors          actorsPopularity   actorsCharacter    originalTitle
##  Length:10000       Length:10000       Length:10000       Length:10000
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##     title           originalLanguage     popularity        releaseDate
##  Length:10000       Length:10000       Min.   :    4.258   Length:10000
##  Class :character   Class :character   1st Qu.:   14.578   Class :character
##  Mode  :character   Mode  :character   Median :   21.906   Mode  :character
##                                        Mean   :   51.394
##                                        3rd Qu.:   40.654
##                                        Max.   :11474.647
```

2. (5 puntos) Diga el tipo de cada una de las variables (cualitativa ordinal o nominal, cuantitativa continua, cuantitativa discreta)

Variables:

ID: variable cuantitativa nominal.
Popularidad: variable cuantitativa continua.
Budget : Cuantitativa continua
Genres: Cualitativa nominal
HomePage: Cualitativa nominal
Production company: Cualitativa nominal
Production Country: Cualitativa nominal
Revenue: Cuantitativa continua
Runtime: Cuantitativa discreta
Video: Cualitativa nominal
Director: Cualitativa nominal
Actors: Cualivativa nominal
Actors Popularity: Cuantitativa continua
Actors Character: Cualitativa nominal
Title: Cualitativa nominal
original Language: Cualitativa nominal

Hoja de Trabajo I
Minería de datos 3 de febrero del 2023
Lynette García Universidad del Valle

original title: Cualitativa nominal
original language: Cualitativa nominal
vote count: cuantitativa discreta
vote average: cuantitativa continua
Release date: Cualitativa ordinal
Vote Avg: Cuantitativa continua
Genres Amount: Cuantitativa discreta
Production Co Amount: Cuantitativa discreta
Production Countries Amount: Cuantitativa discreta
Actors Amount: Cuantitativa discreta
CastWomenAmount: Cuantitativa discreta
CastMenAmount: Cuantitativa discreta

3. (6 puntos) Investigue si las variables cuantitativas siguen una distribución normal y haga una tabla de frecuencias de las variables cualitativas. Explique todos los resultados.

Pasos para verificar si una variable sigue una distribución normal:

- Se presenta el histograma de Xk.
- Se realiza un QQplot para observar la distribución de los valores de Xk.
- Si es posible observar que la distribución de Xk se aleja significativamente de la distribución teórica. Es una señal de que la distribución no es normal.
- Para corroborar esto se realiza la prueba de normalidad de Kolmogorov-Smirnov, usando la media y desviación estándar del budget del dataset. También se realiza la prueba de normalidad de lLilliefors.
- Si para ambos resultados el valor de p es menor que el nivel de significancia (0.05) se puede afirmar que la distribución de Xk no es normal.
    - Buget: No es normal, el p en los niveles de significancia es menor que 0.05.
    - Runtime: No es normal, los valores de significancia son menores que 0.05.
    - Revenue: No es normal, los valores de significancia menores que 0.05
    - VoteCount:No es normal, los valores de significancia menores que 0.05.
    - Genres: No es normal, los valores de significancia no son mayores que 0.05.
    - Production Co Amount: No es normal, los valores de significancia no son mayores que 0.05.
    - Production countries amount: No es normal, los valores de significancia no son mayores que 0.05.
    - Actors amount: No es normal, los valores de significancia no son mayores de 0.05.

4. Responda las siguientes preguntas:
4.1. (3 puntos) ¿Cuáles son las 10 películas que contaron con más presupuesto?

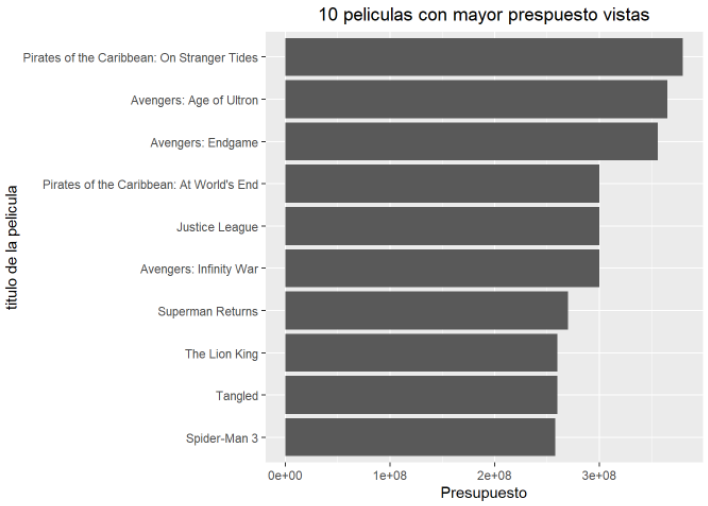
```
Ejercicio 4.1 : Peliculas con mayor presupuesto

```{r echo=FALSE}
library(ggplot2)
movies_sorted <- movies[order(-movies$budget), ]
highest_budget_movies <- head(movies_sorted, n = 10)
highest_budget_movie_title <- highest_budget_movies$title[1:10]
print(paste("Esta es una de las 10 peliculas con mayor presupuesto:", highest_budget_movie_title))


ggplot(data = highest_budget_movies, aes(x = reorder(title, +budget), y = budget)) + geom_bar(stat = "identity") +
xlab("titulo de la pelicula") + ylab("Presupuesto") + ggtitle("10 peliculas con mayor prespuesto vistas") +
theme(plot.title = element_text(hjust = 0.5)) + coord_flip()
```

Ejercicio 4.1 : Peliculas con mayor presupuesto

```
##  [1] "Esta es una de las 10 peliculas con mayor presupuesto: Pirates of the Caribbean: On Stranger Tides"
##  [2] "Esta es una de las 10 peliculas con mayor presupuesto: Avengers: Age of Ultron"
##  [3] "Esta es una de las 10 peliculas con mayor presupuesto: Avengers: Endgame"
##  [4] "Esta es una de las 10 peliculas con mayor presupuesto: Pirates of the Caribbean: At World's End"
##  [5] "Esta es una de las 10 peliculas con mayor presupuesto: Justice League"
##  [6] "Esta es una de las 10 peliculas con mayor presupuesto: Avengers: Infinity War"
##  [7] "Esta es una de las 10 peliculas con mayor presupuesto: Superman Returns"
##  [8] "Esta es una de las 10 peliculas con mayor presupuesto: Tangled"
##  [9] "Esta es una de las 10 peliculas con mayor presupuesto: The Lion King"
## [10] "Esta es una de las 10 peliculas con mayor presupuesto: Spider-Man 3"
```



4.2. (3 puntos) ¿Cuáles son las 10 películas que más ingresos tuvieron?

```
Ejercicio 4.2: Las 10 peliculas con mayor cantidad de ingresos.

```{r echo =FALSE}
movies_sorted <- movies[order(-movies$revenue), ]
highest_budget_movies <- head(movies_sorted, n = 10)
highest_budget_movie_title <- highest_budget_movies$title[1:10]
print(paste("Esta es una de las 10 peliculas con mayor cantidad de ingresos:", highest_budget_movie_title))

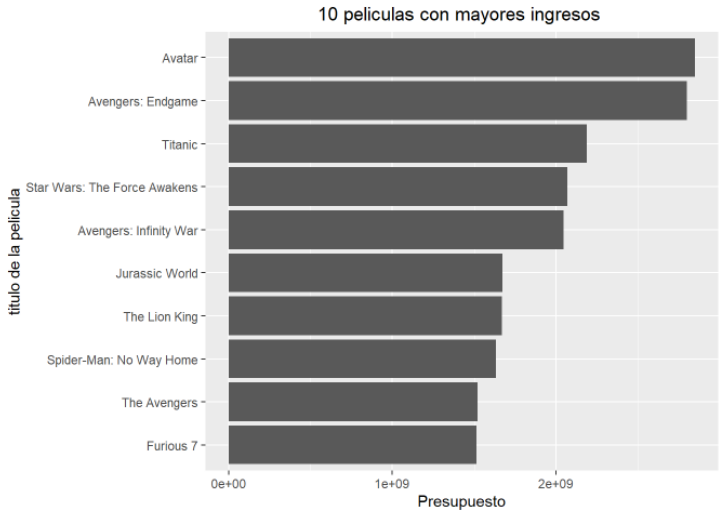
ggplot(data = highest_budget_movies, aes(x = reorder(title, +revenue), y = revenue)) + geom_bar(stat = "identity") +
xlab("titulo de la pelicula") + ylab("Presupuesto") + ggtitle("10 peliculas con mayores ingresos ") + theme(plot.title
= element_text(hjust = 0.5)) + coord_flip()


```
```
```

Presupuesto

Ejercicio 4.2: Las 10 peliculas con mayor cantidad de ingresos.

```
##  [1] "Esta es una de las 10 peliculas con mayor cantidad de ingresos: Avatar"
##  [2] "Esta es una de las 10 peliculas con mayor cantidad de ingresos: Avengers: Endgame"
##  [3] "Esta es una de las 10 peliculas con mayor cantidad de ingresos: Titanic"
##  [4] "Esta es una de las 10 peliculas con mayor cantidad de ingresos: Star Wars: The Force Awakens"
##  [5] "Esta es una de las 10 peliculas con mayor cantidad de ingresos: Avengers: Infinity War"
##  [6] "Esta es una de las 10 peliculas con mayor cantidad de ingresos: Jurassic World"
##  [7] "Esta es una de las 10 peliculas con mayor cantidad de ingresos: The Lion King"
##  [8] "Esta es una de las 10 peliculas con mayor cantidad de ingresos: Spider-Man: No Way Home"
##  [9] "Esta es una de las 10 peliculas con mayor cantidad de ingresos: The Avengers"
## [10] "Esta es una de las 10 peliculas con mayor cantidad de ingresos: Furious 7"
```



4.3. (3 puntos) ¿Cuál es la película que más votos tuvo?

Presupuesto

Ejercicio 4.3: La peliculas que mayor cantidad de votos ha obtenido

```
movies_sorted <- movies[order(-movies$voteCount), ]
bestmovie <- head(movies_sorted, n=1)
bestmovietile <- bestmovie$title[1]
print(paste("Esta es la pelicula con mayor cantidad de votos:", bestmovietile))
```

```
## [1] "Esta es la pelicula con mayor cantidad de votos: Inception"
```

4.4. (3 puntos) ¿Cuál es la peor película de acuerdo a los votos de todos los usuarios

Ejercicios 4.4: La peor pelicula segun los usuarios.

```
min_votes <- movies$voteCount
min_index <- which.min(min_votes)
min_movie <- movies[min_index, "title"]
print(paste("Esta es la peor pelicula segun los usuarios:", min_movie))
```

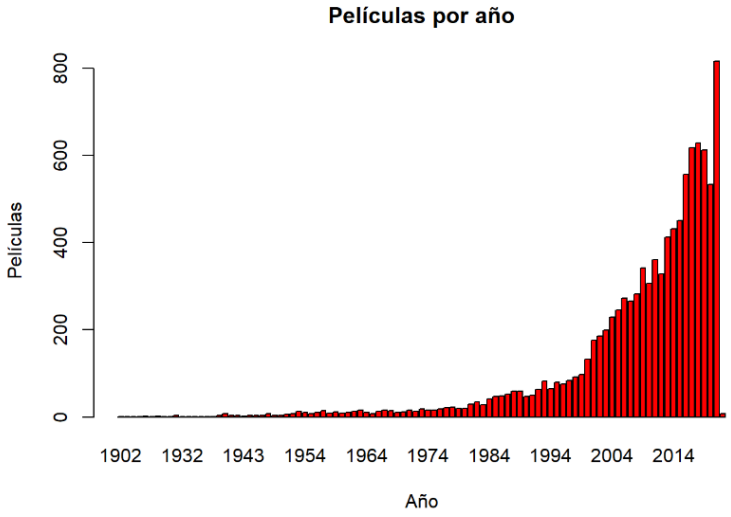
```
## [1] "Esta es la peor pelicula segun los usuarios: La Mera Reyna del Sur"
```

Hoja de Trabajo I
Minería de datos 3 de febrero del 2023
Lynette García Universidad del Valle

4.5. (8 puntos) ¿Cuántas películas se hicieron en cada año? ¿En qué año se hicieron más películas? Haga un gráfico de barras

¿Cuántas películas se hicieron en cada año? ¿En qué año se hicieron más peñículas? Haga un gráfico de barras ¿Cuántas películas se hicieron en cada año? ¿En qué año se hicieron más películas? Haga un gráfico de barras

```
movies$releaseDate <- as.Date(movies$releaseDate)
movies$year <- format(movies$releaseDate, "%Y")
movies$year <- as.numeric(movies$year)
movies$year <- as.factor(movies$year)
moviesByYear <- table(movies$year) # nolint
barplot(moviesByYear, main = "Películas por año", xlab = "Año", ylab = "Películas", col = "red") # nolint
```



4.6. (9 puntos) ¿Cuál es el género principal de las 20 películas más recientes? ¿Cuál es el género principal que predomina en el conjunto de datos? Represéntelo usando un gráfico

¿Cuál es el género principal de las 20 peliculas más recientes? ¿Cuál es el género principal que predomina en el conjunto de movies? Represéntelo usando un gráfico.

```
popularMovie <- movies[order(movies$popularity,decreasing = TRUE),]
top20 <- popularMovie[1:20,c("genres")]
genres20 <- unlist(strsplit(as.character(top20), "\\|"))
genres20
```

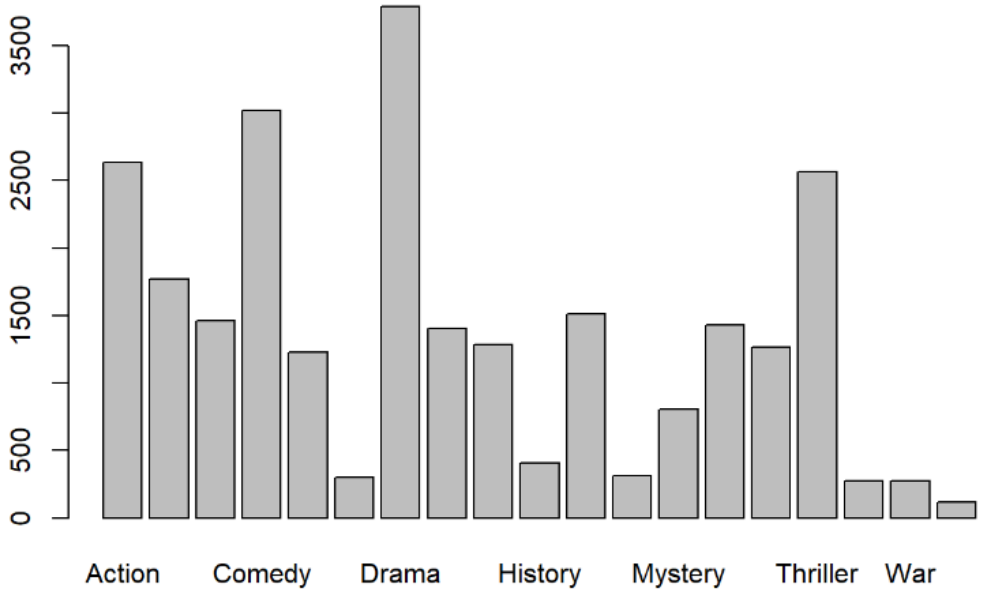
```
##  [1] "Action"          "Adventure"       "Fantasy"         "Science Fiction"
##  [5] "Action"          "Adventure"       "Science Fiction" "Animation"
##  [9] "Comedy"          "Family"          "Music"           "Horror"
## [13] "Action"          "Science Fiction" "Animation"       "Comedy"
## [17] "Family"          "Fantasy"         "Comedy"          "Fantasy"
## [21] "Adventure"       "Action"          "Thriller"        "Science Fiction"
## [25] "Action"          "Adventure"       "Science Fiction" "Action"
## [29] "Adventure"       "Science Fiction" "Thriller"        "Action"
## [33] "Thriller"        "Action"          "Comedy"          "Crime"
## [37] "Thriller"        "Action"          "Adventure"       "Fantasy"
## [41] "Action"          "Thriller"        "Crime"           "Drama"
## [45] "Horror"          "Mystery"         "Animation"       "Comedy"
## [49] "Family"          "Crime"           "Action"          "Thriller"
## [53] "Animation"       "Action"          "Adventure"       "Fantasy"
## [57] "Drama"           "History"         "Adventure"       "Horror"
## [61] "Thriller"
```

```
getmode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}

top20 <- getmode(genres20)
top20
```

```
## [1] "Action"
```

```
total <- unlist(strsplit(as.character(movies$genres), "\\|"))
barplot(table(total))
```



4.7. (8 puntos) ¿Las películas de qué genero principal obtuvieron mayores ganancias?

¿Las películas de qué genero principal obtuvieron mayores ganancias?

```
movies <- read.csv("movies.csv")
sum_by_genre <- aggregate(movies$revenue, by=list(movies$genres), sum)
genre_with_max_earnings <- sum_by_genre[which.max(sum_by_genre$x), ]
genre_with_max_earnings
```

```
##                          Group.1          x
## 94 Action|Adventure|Science Fiction 19780390887
```

4.8. (3 puntos) ¿La cantidad de actores influye en los ingresos de las películas?¿se han hecho películas con más actores en los últimos años?

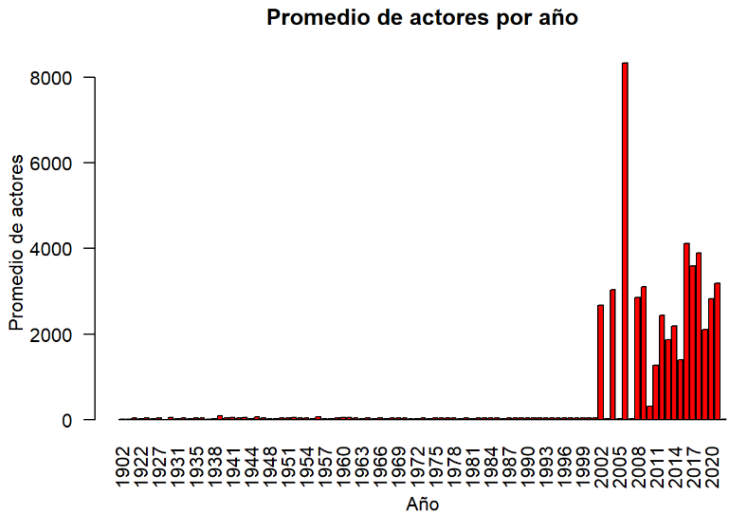
¿La cantidad de actores influye en los ingresos de las películas?

```
cor(movies$actorsAmount, movies$revenue)
```

```
## [1] -0.01955488
```

¿se han hecho películas con más actores en los últimos años?

```
movies$releaseDate <- as.Date(movies$releaseDate)
movies$year <- format(movies$releaseDate, "%Y")
movies$year <- as.numeric(movies$year)
movies$year <- as.factor(movies$year)
actorsByYear <- data.frame(movies$year, movies$actorsAmount)
actorsByYear <- aggregate(actorsByYear$movies.actorsAmount, by = list(actorsByYear$movies.year), FUN = mean)
barplot(actorsByYear$x, main = "Promedio de actores por año", xlab = "Año", ylab = "Promedio de actores", col = "red", las = 2, names.arg = actorsByYear$Group.1)
```



Ejercicio 4.9 ¿Es posible que la

4.9. (3 puntos) ¿Es posible que la cantidad de hombres y mujeres en el reparto influya en la popularidad y los ingresos de las películas?

Ejercicio 4.9

¿Es posible que la cantidad de hombres y mujeres en el reparto influya en la popularidad y los ingresos de las películas?

```
cor(movies$actorsAmount, movies$popularity)
```

```
## [1] -0.006230412
```

```
cor(movies$actorsAmount, movies$revenue)
```

```
## [1] -0.01955488
```

4.10. (8 puntos) ¿Quiénes son los directores que hicieron las 20 películas mejor calificadas?

¿Quiénes son los directores que hicieron las 20 películas mejor calificadas?

```
movies_without_na <- movies[!is.na(movies$voteCount) & is.numeric(movies$voteCount), ] # nolint: line_length_linter.
movies_sorted <- movies_without_na[order(movies_without_na$voteCount, decreasing = TRUE), ] # nolint: line_length_linter.
top_20_movies <- movies_sorted[1:20, ]
top_20_directors <- unique(top_20_movies$director)
top_20_directors
```

```
##  [1] "Christopher Nolan"             "Joss Whedon"
##  [3] "Tim Miller"                    "James Cameron"
##  [5] "James Gunn"                    "Anthony Russo|Joe Russo"
##  [7] "David Fincher"                 "Quentin Tarantino"
##  [9] "Jon Favreau"                   "Robert Zemeckis"
## [11] "Chris Columbus"                "Lilly Wachowski|Lana Wachowski"
## [13] "Frank Darabont"                "Peter Jackson"
## [15] "Todd Phillips"
```

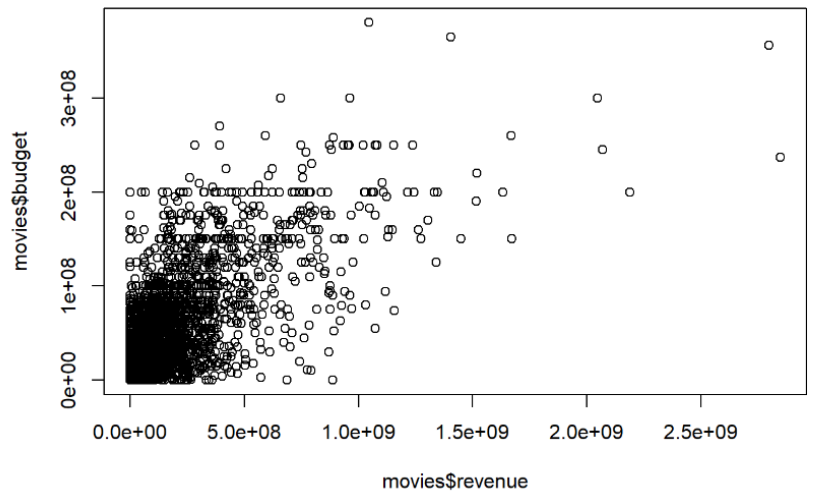
4.11. (8 puntos) ¿Cómo se correlacionan los presupuestos con los ingresos? ¿Los altos presupuestos significan altos ingresos? Haga los gráficos que necesite, histograma, diagrama de dispersión

¿Cómo se correlacionan los presupuestos con los ingresos? ¿Los altos presupuestos significan altos ingresos? Haga los gráficos que necesite, histograma, diagrama de dispersión?

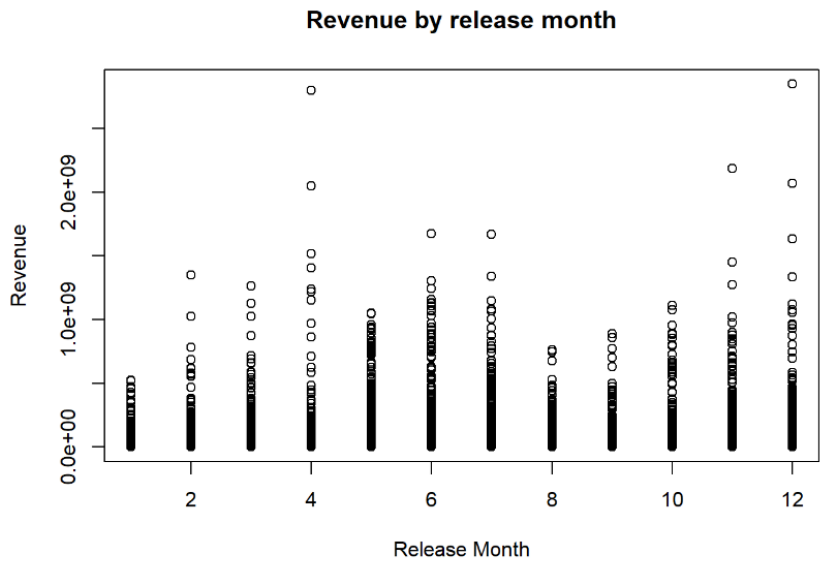
```
movies <- read.csv("movies.csv")
cor(movies$revenue, movies$budget)
```

```
## [1] 0.757454
```

```
# Graficando
plot(movies$revenue, movies$budget)
```



4.12. (7 puntos) ¿Se asocian ciertos meses de lanzamiento con mejores ingresos?



Como podemos observar , hay una relación entre los meses y los mejores ingresos, dado que los mejores ingresos se dan en los meses de abril, diciembre y noviembre, podríamos argumentar que estos son las mejores épocas para el lanzamiento de una pelicula.

4.13. (8 puntos) ¿En qué meses se han visto los lanzamientos con mejores ingresos?

- Abril, Noviembre y diciembre. (ver tabla anterior)

¿cuantas películas, en promedio, se han lanzado por mes?

¿Cuantas películas, en promedio, se han lanzado por mes?

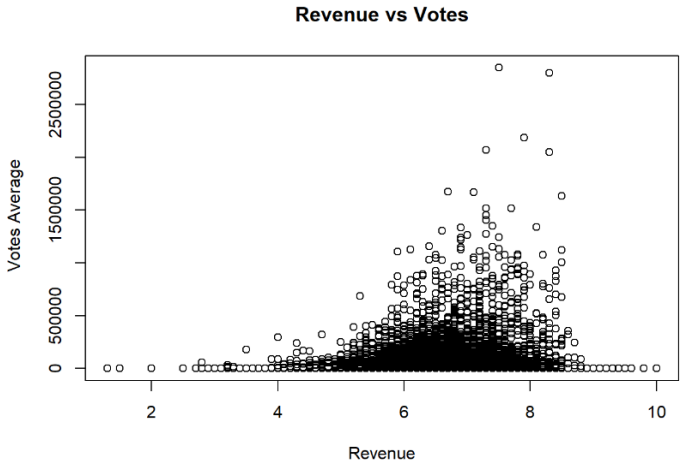
```
monthlyCount <- table(releaseMonths)

mean(monthlyCount)
```

```
## [1] 833.3333
```

El output determina que el promedio de peliculas lanzadas por mes es de `833.33`.

4.14. (7 puntos) ¿Cómo se correlacionan las calificaciones con el éxito comercial?



Los votos de la pelicula son directamente proporcionales a los ingresos, por lo que se puede decir que las calificaciones y el éxito comercial tienen la misma relación.

4.15. (5 puntos) ¿A qué género principal pertenecen las películas más largas?

Problema 4.15

¿A qué género principal pertenecen las películas más largas?

```
df <- data.frame(runtime = movies$runtime, genre = movies$genres)

df_sorted <- df[order(df$runtime, decreasing = TRUE),]

df_sorted[1:10,]
```

```
##      runtime                                    genre
## 9348     750                              Documentary
## 5359     400                              Documentary
## 3886     333                      Drama|History|War
## 963      317                          Drama|History
## 1264     248                  Drama|History|Romance
## 7066     247                  Action|Crime|Thriller
## 1949     242                                  Drama
## 9687     242 Action|Adventure|Fantasy|Science Fiction
## 3741     240                              Documentary
## 5593     240                             Action|Drama
```

Como podemos observar el genero con la duracion mas larga es el de documentales.

Extra:

4.16. ¿La película con mayor cantidad de productoras?

```
¿Cual es la pelicula donde tiene un cast con mayor cantidad de productoras?

```{r echo =FALSE}

movies_sorted <- movies[order(-movies$productionCoAmount), ]
highest_pro_movies <- head(movies_sorted, n = 1)
highest_pro_movie_title <- highest_pro_movies$title[1]
print(paste("La pelicula con mayor cantidad de prodcutoras", highest_pro_movie_title))



```
```

4.17. ¿Cual es el país más frecuentado para grabar películas?

```
¿Cual es el pais mas frecuentado para grabar peliculas?
```{r echo =FALSE}

most_repeated_country <- names(which.max(table(movies$productionCountry)))

print(paste("El pais mas frecuentado para grabar peliculas es: ", most_repeated_country))


```
```

Preguntas extras!!!

¿Cual es la pelicula donde tiene un cast con mayor cantidad de productoras?

```
## [1] "La pelicula con mayor cantidad de prodcutoras Goal! III : Taking On The World"
```

¿Cual es el pais mas frecuentado para grabar peliculas?

```
## [1] "El pais mas frecuentado para grabar peliculas es:  United States of America"
```

4.18. ¿Cuáles son las 3 películas que han tenido *menos* ganancias?

¿Cuales son las 3 peliculas que han tenido menos ganancias?

```
movies[order(movies$revenue), c("revenue", "title")][1:3, ]
```

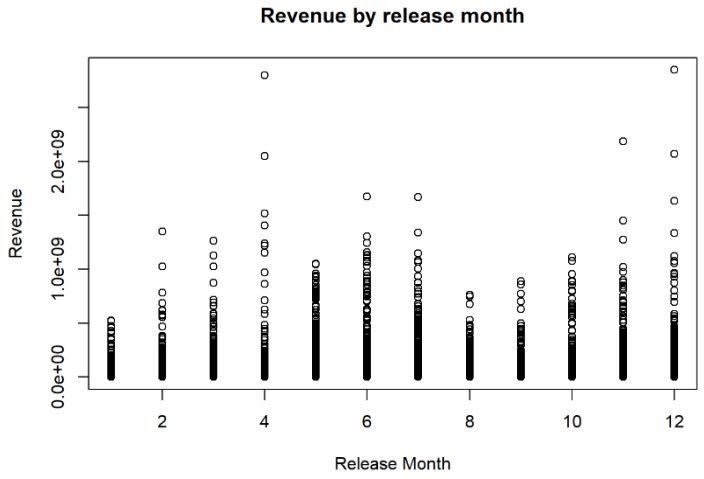
```
##    revenue              title
## 26       0             Brazil
## 57       0  Three Colors: Blue
## 58       0 Three Colors: White
```

4.19. ¿Cuál es la película que cuenta con más cantidad de actores hombres?

```
movies$castMenAmount <- suppressWarnings(as.numeric(movies$castMenAmount))
movies[order(movies$castMenAmount, decreasing = TRUE), c("castMenAmount", "title")][1, ] # nolint: line_length_linter.
```

```
##      castMenAmount                         title
## 9998        922017 Chief Daddy 2: Going for Broke
```

4.20 ¿En qué meses se han visto los lanzamientos con peores ingresos?



Enero, Agosto y Septiembre.

4.21 ¿Qué película tiene el mayor presupuesto?

Pelicula con mayor presupuesto

```
library(ggplot2)

movies <- read.csv("movies.csv")

df <- data.frame(budget = movies$budget, title = movies$title)

df_sorted <- df[order(df$budget, decreasing = TRUE),]

df_sorted[1:10,]
```

```
##         budget                                         title
## 717  380000000 Pirates of the Caribbean: On Stranger Tides
## 4711 365000000                     Avengers: Age of Ultron
## 5953 356000000                           Avengers: Endgame
## 164  300000000    Pirates of the Caribbean: At World's End
## 4954 300000000                              Justice League
## 5954 300000000                      Avengers: Infinity War
## 608  270000000                             Superman Returns
## 3792 260000000                                     Tangled
## 7135 260000000                                The Lion King
## 281  258000000                                  Spider-Man 3
```