HDT1

AndresEmilioQ, Mirka M, Oscar de Leon

4/2/2022

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

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

## A continuacion los datos recopilados de la base de datos y sus caracteristicas

## El conjunto de datos esta compuesto por 10000 observaciones y 27 variables.

## 2. Diga el tipo de cada una de las variables (cualitativa ordinal o nominal, cuantitativa

continua, cuantitativa discreta)

variable <- c("id", "popularity", "budget", "revenue", "original\_title", "originalLanguage", "title", "homePage", "video", "director", "runtime", "genres", "genresAmount", "productionCompany", "productionCoAmount", "productionCompanyCountry", "productionCountry", "productionCountriesAmount", "releaseDate", "voteCount", "voteAvg", "actors", "actorsPopularity", "actorsCharacter", "actorsAmount", "castWomenAmount", "CastMenAmount")  
  
tipo <- c("cuantitativa continua", "cuantitativa continua", "cuantitativa continua ", "cuantitativa continua", "cualitativa nominal", "cualitativa nominal", "cualitativa nominal", "cualitativa nominal", "cualitativa nominal", "cualitativa nominal", "cuantitativa continua", "cualitativa nominal", "cuantitativa discreta", "cualitativa nominal", "cuantitativa discreta", "cualitativa nominal", "cualitativa nominal", " cuantitativa discreta", "cualitativa nominal", "cualitativa discreta", "cuantitativa continua", "cualitativa nominal", "cuantitativa continua", "cualitativa nominal", "cuantitativa discreta", "cuantitativa discreta", "cuantitativa discreta")  
  
numero <- 1:27  
  
DataFrame.Variables <- data.frame(numero, variable, tipo)  
  
print(DataFrame.Variables)

## numero variable tipo  
## 1 1 id cuantitativa continua  
## 2 2 popularity cuantitativa continua  
## 3 3 budget cuantitativa continua   
## 4 4 revenue cuantitativa continua  
## 5 5 original\_title cualitativa nominal  
## 6 6 originalLanguage cualitativa nominal  
## 7 7 title cualitativa nominal  
## 8 8 homePage cualitativa nominal  
## 9 9 video cualitativa nominal  
## 10 10 director cualitativa nominal  
## 11 11 runtime cuantitativa continua  
## 12 12 genres cualitativa nominal  
## 13 13 genresAmount cuantitativa discreta  
## 14 14 productionCompany cualitativa nominal  
## 15 15 productionCoAmount cuantitativa discreta  
## 16 16 productionCompanyCountry cualitativa nominal  
## 17 17 productionCountry cualitativa nominal  
## 18 18 productionCountriesAmount cuantitativa discreta  
## 19 19 releaseDate cualitativa nominal  
## 20 20 voteCount cualitativa discreta  
## 21 21 voteAvg cuantitativa continua  
## 22 22 actors cualitativa nominal  
## 23 23 actorsPopularity cuantitativa continua  
## 24 24 actorsCharacter cualitativa nominal  
## 25 25 actorsAmount cuantitativa discreta  
## 26 26 castWomenAmount cuantitativa discreta  
## 27 27 CastMenAmount cuantitativa discreta

## 3. Investigue si las variables cuantitativas siguen una distribución normal y haga una

tabla de frecuencias de las variables cualitativas. Explique todos los resultados.

Nos apoyamos de nuestros conocimientos previos en estadistica y utilizamos ademas la curtosis para determinar la concentracion de los datos entorno a la media. Encontramos que todas nuestras variables cuantitativas obtuvieron un coeficiente mayor a uno, es decir positivo (distribucion leptocúrtica)

Referencia y expliacion detallada: <https://economipedia.com/definiciones/curtosis.html>

#Libreria necesaria para utilizar la curtosis, por favor instalar  
library(e1071)  
  
## Popularidad  
  
popularity = datos$popularity  
"Kurtosis de popularidad"

## [1] "Kurtosis de popularidad"

#Presupuesto  
budget = datos$budget  
"Kurtosis de budget"

## [1] "Kurtosis de budget"

kurtosis(budget)

## [1] 13.20779

#Ingresos  
revenue = datos$revenue  
"Kurtosis de revenue"

## [1] "Kurtosis de revenue"

kurtosis(revenue)

## [1] 55.89168

#Duración  
runtime = datos$runtime  
"Kurtosis de runtime"

## [1] "Kurtosis de runtime"

kurtosis(runtime)

## [1] 35.52866

#Votos  
vote\_count = datos$voteCount  
"Kurtosis de VoteCount"

## [1] "Kurtosis de VoteCount"

kurtosis(vote\_count)

## [1] 22.76501

#Promedio de votos  
vote\_average = datos$voteAvg  
"Kurtosis de VoteAvg"

## [1] "Kurtosis de VoteAvg"

kurtosis(vote\_average)

## [1] 1.484826

#Cantidad de generos  
GenresA = datos$genresAmount  
"Kurtosis de GenresAmount"

## [1] "Kurtosis de GenresAmount"

kurtosis(GenresA)

## [1] 2.104824

#Cantidad de co productora   
ProductionCoA = datos$productionCoAmount  
"Kurtosis de productionCoAmount"

## [1] "Kurtosis de productionCoAmount"

kurtosis(ProductionCoA)

## [1] 140.2735

#Cantidad de productoras por pais  
ProducCountrieA = datos$productionCountriesAmount  
"Kurtosis de ProductCountriesAmount"

## [1] "Kurtosis de ProductCountriesAmount"

kurtosis(ProducCountrieA)

## [1] 751.2015

#Cantidad de actores  
ActorsA = datos$actorsAmount  
"Kurtosis de ActorsAmount"

## [1] "Kurtosis de ActorsAmount"

kurtosis(ActorsA)

## [1] 371.0959

#Cantiddad de actores Masculinos  
#Esta kurtosis la realizamos en excel debido a que la base de datos presenta datos no numericos en ciertas celdas  
ActoresM = datos$castMenAmount  
ActoresM <- na.omit(ActoresM)  
b <- 43.93458  
"Kurtosis de CastMenAmount"

## [1] "Kurtosis de CastMenAmount"

b

## [1] 43.93458

#kurtosis(ActoresM)  
  
#Cantiddad de actrices Femenina  
#Esta kurtosis la realizamos en excel debido a que la base de datos presenta datos no numericos en ciertas celdas  
ActricesF = datos$castWomenAmount  
a <- 108.7062  
"Kurtosis de CastWomenAmount"

## [1] "Kurtosis de CastWomenAmount"

a

## [1] 108.7062

#kurtosis(ActricesF)

##### Variables cualitativas tabla de frecuencias

Para las variables cualitativas realizamos tablas de frecuencia con la ayuda de data frames para poder observar y analizar los datos de mejor manera. Se ignoro la variable id debido a que se determino a que no aporta al analisis estadistico. Se noto que distintas variables contaban con valores vacios, se llenaron con NA y otros varios estaban acompañados por slash o pipe, por lo cual se decidio hacer un split para estos terminos.

library(tidyr)  
#Generos  
genres <- data.frame(table(do.call(c, lapply(datos$genres, function(x) unlist(strsplit(x, "\\|"))))))  
  
#Pagina  
homepage <- data.frame(table(datos$homePage))  
homepage[homepage == ""] <- NA  
homepage <- na.omit(homepage)  
  
#Companias de produccion  
productionCompany <- data.frame(table(do.call(c, lapply(datos$productionCompany, function(x) unlist(strsplit(x, "\\|"))))))  
  
#pais de compania de produccion  
ProductionCompanyCountry <- data.frame(table(do.call(c, lapply(datos$productionCompanyCountry, function(x) unlist(strsplit(x, "\\|"))))))  
  
#pais productora  
ProductionCountry <- data.frame(table(do.call(c, lapply(datos$productionCountry, function(x) unlist(strsplit(x, "\\|"))))))  
  
#video  
video <- data.frame(table(datos$video))  
video[video == ""] <- NA  
video <- na.omit(video)  
  
#Director  
director <- data.frame(table(datos$director))  
director[director == ""] <- NA  
director <- na.omit(director)  
  
#Actores  
actors <- data.frame(table(do.call(c, lapply(datos$actors, function(x) unlist(strsplit(x, "\\|"))))))  
actors <- data.frame(table(datos$actors))  
actors[actors == ""] <- NA  
actors <- na.omit(actors)  
  
#ActorsCharacter  
character <- data.frame(table(do.call(c, lapply(datos$actorsCharacter, function(x) unlist(strsplit(x, "\\|"))))))  
character[character == ""] <- NA  
character <- na.omit(character)  
  
#Titulo original  
original\_title <- data.frame(table(datos$originalTitle))  
  
#Titulo  
title <- data.frame(table(datos$title))  
  
#Lenguaje Original  
language <- data.frame(table(datos$originalLanguage))  
  
#fecha  
release\_date <- data.frame(table(datos$releaseDate))

## 4. Responda las siguientes preguntas:

### 4.4 ¿Cuál es la peor película de acuerdo a los votos de todos los usuarios?

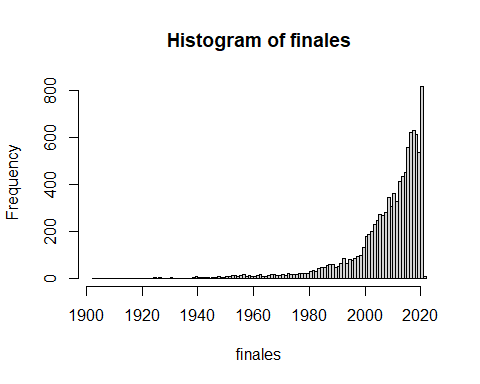
A continuacion las dos peor votadas

# peorvotadas <- datos[order(datos$voteAvg),]  
#peorvotadas <- peorvotadas[1,]  
#na.omit(peorvotadas)  
peorvotadas <- data.frame(datos$voteAvg, datos$title)  
lapeor <- data.frame(peorvotadas[order(peorvotadas$datos.voteAvg),])  
head(lapeor,2)

## datos.voteAvg  
## 9787 1.3  
## 9707 1.5  
## datos.title  
## 9787 DAKAICHI -I'm Being Harassed by the Sexiest Man of the Year- The Movie: In Spain  
## 9707 DRagON BALL P2 2wice dropda bbeet

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

Dates <- data.frame(datos$title, datos$releaseDate)  
PorAnioDesc <- data.frame(Dates[order(Dates$datos.releaseDate),])  
year1 <- data.frame(table(substring(PorAnioDesc$datos.releaseDate,1,4)))  
finales <- as.numeric(substring(PorAnioDesc$datos.releaseDate,1,4))  
hist(finales,breaks = 100)



### 4.6 ¿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

A continuacion el top20:

LicaPopular <- datos[order(datos$popularity,decreasing = TRUE),]  
top20 <- LicaPopular[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"

Para encontrar el genero principal predominante debemos separar la lista, debido a que las peliculas cuentan con mas de un genero.

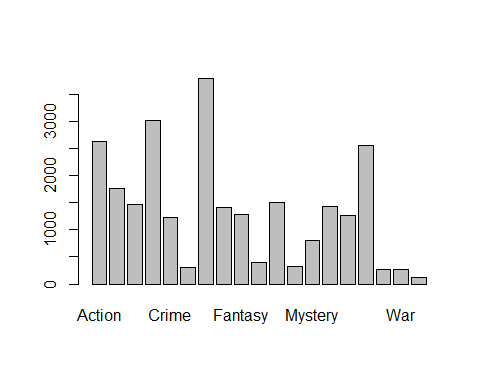
getmode <- function(v) {  
 uniqv <- unique(v)  
 uniqv[which.max(tabulate(match(v, uniqv)))]  
}  
topGenre20 <- getmode(genres20)  
"Se encontro que el genero principal de las peliculas es:"

## [1] "Se encontro que el genero principal de las peliculas es:"

topGenre20

## [1] "Action"

totalGenres <- unlist(strsplit(as.character(datos$genres), "\\|"))  
barplot(table(totalGenres))



### 4.7 ¿Las películas de qué genero principal obtuvieron mayores ganancias?

worth <- data.frame(datos$revenue, datos$genres, datos$title)  
ordenadas <- data.frame(worth[order(-datos$revenue), ])  
"Las peliculas del genero de accion fueron las que obtuvieron mayores ganancias"

## [1] "Las peliculas del genero de accion fueron las que obtuvieron mayores ganancias"

head(ordenadas, 5)

## datos.revenue datos.genres  
## 3211 2847246203 Action|Adventure|Fantasy|Science Fiction  
## 5953 2797800564 Adventure|Science Fiction|Action  
## 308 2187463944 Drama|Romance  
## 4948 2068223624 Action|Adventure|Science Fiction|Fantasy  
## 5954 2046239637 Adventure|Action|Science Fiction  
## datos.title  
## 3211 Avatar  
## 5953 Avengers: Endgame  
## 308 Titanic  
## 4948 Star Wars: The Force Awakens  
## 5954 Avengers: Infinity War