Problemas 1-4\_3

Pablo Noack

2/2/2021

library(nortest)  
data <-read.csv("C:/Users/LENOVO/Desktop/Clases/Minería de datos/HT/HT1DataMining/data/tmdb-movies.csv")

# Haga una exploración rápida de sus datos,para eso haga un resumen de su conjunto de datos.

## Solución:

summary(data)

## id imdb\_id popularity budget   
## Min. : 5 Length:10866 Min. : 0.00006 Min. : 0   
## 1st Qu.: 10596 Class :character 1st Qu.: 0.20758 1st Qu.: 0   
## Median : 20669 Mode :character Median : 0.38386 Median : 0   
## Mean : 66064 Mean : 0.64644 Mean : 14625701   
## 3rd Qu.: 75610 3rd Qu.: 0.71382 3rd Qu.: 15000000   
## Max. :417859 Max. :32.98576 Max. :425000000   
## revenue original\_title cast homepage   
## Min. :0.000e+00 Length:10866 Length:10866 Length:10866   
## 1st Qu.:0.000e+00 Class :character Class :character Class :character   
## Median :0.000e+00 Mode :character Mode :character Mode :character   
## Mean :3.982e+07   
## 3rd Qu.:2.400e+07   
## Max. :2.782e+09   
## director tagline keywords overview   
## Length:10866 Length:10866 Length:10866 Length:10866   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## runtime genres production\_companies release\_date   
## Min. : 0.0 Length:10866 Length:10866 Length:10866   
## 1st Qu.: 90.0 Class :character Class :character Class :character   
## Median : 99.0 Mode :character Mode :character Mode :character   
## Mean :102.1   
## 3rd Qu.:111.0   
## Max. :900.0   
## vote\_count vote\_average release\_year budget\_adj   
## Min. : 10.0 Min. :1.500 Min. :1960 Min. : 0   
## 1st Qu.: 17.0 1st Qu.:5.400 1st Qu.:1995 1st Qu.: 0   
## Median : 38.0 Median :6.000 Median :2006 Median : 0   
## Mean : 217.4 Mean :5.975 Mean :2001 Mean : 17551040   
## 3rd Qu.: 145.8 3rd Qu.:6.600 3rd Qu.:2011 3rd Qu.: 20853251   
## Max. :9767.0 Max. :9.200 Max. :2015 Max. :425000000   
## revenue\_adj   
## Min. :0.000e+00   
## 1st Qu.:0.000e+00   
## Median :0.000e+00   
## Mean :5.136e+07   
## 3rd Qu.:3.370e+07   
## Max. :2.827e+09

# Diga el tipo de cada una de las variables(cualitativaordinal o nominal, cuantitativa continua, cuantitativa discreta)

## Solucion

str(data)

## 'data.frame': 10866 obs. of 21 variables:  
## $ id : int 135397 76341 262500 140607 168259 281957 87101 286217 211672 150540 ...  
## $ imdb\_id : chr "tt0369610" "tt1392190" "tt2908446" "tt2488496" ...  
## $ popularity : num 32.99 28.42 13.11 11.17 9.34 ...  
## $ budget : int 150000000 150000000 110000000 200000000 190000000 135000000 155000000 108000000 74000000 175000000 ...  
## $ revenue : num 1.51e+09 3.78e+08 2.95e+08 2.07e+09 1.51e+09 ...  
## $ original\_title : chr "Jurassic World" "Mad Max: Fury Road" "Insurgent" "Star Wars: The Force Awakens" ...  
## $ cast : chr "Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vincent D'Onofrio|Nick Robinson" "Tom Hardy|Charlize Theron|Hugh Keays-Byrne|Nicholas Hoult|Josh Helman" "Shailene Woodley|Theo James|Kate Winslet|Ansel Elgort|Miles Teller" "Harrison Ford|Mark Hamill|Carrie Fisher|Adam Driver|Daisy Ridley" ...  
## $ homepage : chr "http://www.jurassicworld.com/" "http://www.madmaxmovie.com/" "http://www.thedivergentseries.movie/#insurgent" "http://www.starwars.com/films/star-wars-episode-vii" ...  
## $ director : chr "Colin Trevorrow" "George Miller" "Robert Schwentke" "J.J. Abrams" ...  
## $ tagline : chr "The park is open." "What a Lovely Day." "One Choice Can Destroy You" "Every generation has a story." ...  
## $ keywords : chr "monster|dna|tyrannosaurus rex|velociraptor|island" "future|chase|post-apocalyptic|dystopia|australia" "based on novel|revolution|dystopia|sequel|dystopic future" "android|spaceship|jedi|space opera|3d" ...  
## $ overview : chr "Twenty-two years after the events of Jurassic Park, Isla Nublar now features a fully functioning dinosaur theme"| \_\_truncated\_\_ "An apocalyptic story set in the furthest reaches of our planet, in a stark desert landscape where humanity is b"| \_\_truncated\_\_ "Beatrice Prior must confront her inner demons and continue her fight against a powerful alliance which threaten"| \_\_truncated\_\_ "Thirty years after defeating the Galactic Empire, Han Solo and his allies face a new threat from the evil Kylo "| \_\_truncated\_\_ ...  
## $ runtime : int 124 120 119 136 137 156 125 141 91 94 ...  
## $ genres : chr "Action|Adventure|Science Fiction|Thriller" "Action|Adventure|Science Fiction|Thriller" "Adventure|Science Fiction|Thriller" "Action|Adventure|Science Fiction|Fantasy" ...  
## $ production\_companies: chr "Universal Studios|Amblin Entertainment|Legendary Pictures|Fuji Television Network|Dentsu" "Village Roadshow Pictures|Kennedy Miller Productions" "Summit Entertainment|Mandeville Films|Red Wagon Entertainment|NeoReel" "Lucasfilm|Truenorth Productions|Bad Robot" ...  
## $ release\_date : chr "6/9/15" "5/13/15" "3/18/15" "12/15/15" ...  
## $ vote\_count : int 5562 6185 2480 5292 2947 3929 2598 4572 2893 3935 ...  
## $ vote\_average : num 6.5 7.1 6.3 7.5 7.3 7.2 5.8 7.6 6.5 8 ...  
## $ release\_year : int 2015 2015 2015 2015 2015 2015 2015 2015 2015 2015 ...  
## $ budget\_adj : num 1.38e+08 1.38e+08 1.01e+08 1.84e+08 1.75e+08 ...  
## $ revenue\_adj : num 1.39e+09 3.48e+08 2.72e+08 1.90e+09 1.39e+09 ...

# Investigue si las variables cuantitativas siguen una distribución normal y haga una tabla de frecuencias de las variables cualitativas. Explique todos los resultados.

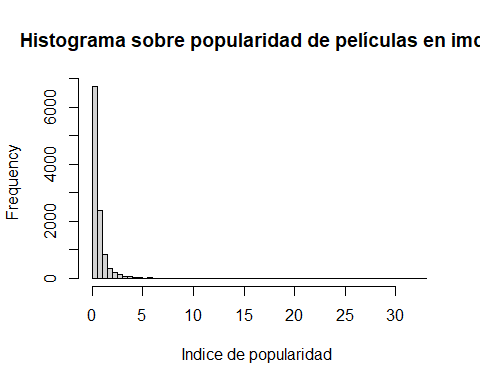
## Solución:

Antes que todo, es importante remarcar que las variables id, imbd\_id, original\_title, cast, tagline y overview, son meras variables cuantitativas que no exiben ningún comportamiento estadístico. Por ejemplo, si hicieramos una tabla de frecuencias o un test de normalidad estas variables no exibirían ningún comportamiento debido a que son etiquetas para cada película. ### Variables cuantitativas.

#### Popularidad:

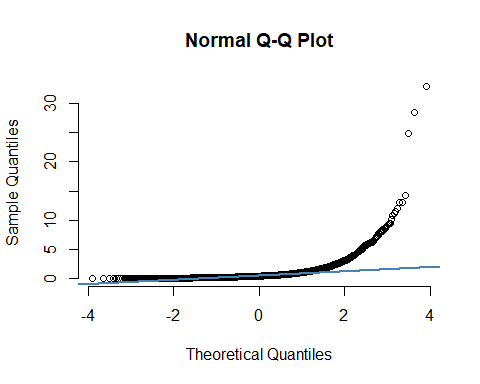
* Histograma: muestra un sesgo hacia el 0, esto implica que la mayoría de películas que salen al cine no son reconocidas y son muy pocas las que triunfan.Por otro lado, no muestra la simetría que cabría esperar de una distribución normal.

hist(data$popularity[data$popularity != 0],  
 main = "Histograma sobre popularidad de películas en imdb",  
 xlab = "Indice de popularidad ",  
 breaks = 105)



* QQnormal: dado ese comportamiento y el sesgo que tienen los datos, se aplicó una gráfica de qqnorm.En la que se oberva que los extremos no se comportan de manera normal.

qqnorm(data$popularity, pch = 1, frame = FALSE)  
qqline(data$popularity, col = "steelblue", lwd = 2)



* Test de Lilliefors: tenemos que el valor p es menor a 0.05 con lo que se confirma que la popularidad **No exibe un comportamiento normal**

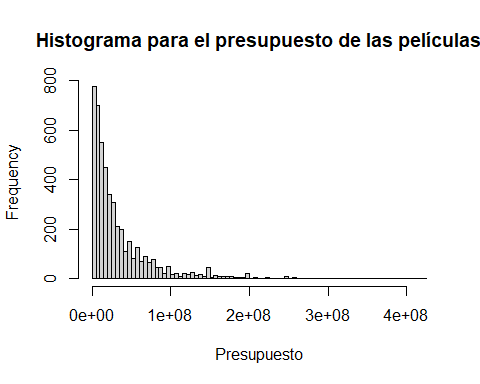
lillie.test((data$popularity))

##   
## Lilliefors (Kolmogorov-Smirnov) normality test  
##   
## data: (data$popularity)  
## D = 0.25907, p-value < 2.2e-16

#### Presupuesto

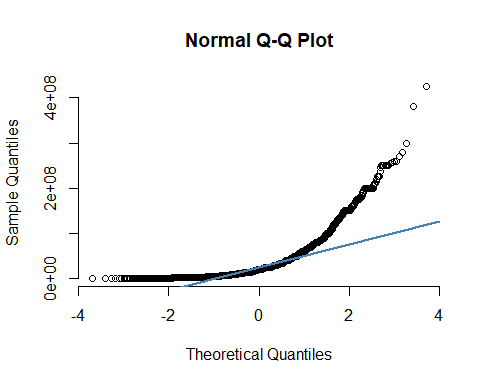
Para el estudio de estos datos, se analizaron películas que hayan costado más de $1,000,000 USD debido a que se detectaron 2 escalas distintas en el set de datos. Además de esto no se sabe si la escala fue cambiada alguna otra vez. \* Histograma: En el histograma nos damos cuenta que esta no sigue ninguna distribución de normalidad, nos damos cuenta que no posee simetría.

data2 = data[data$budget > 1000000,]  
hist(data2$budget,  
 main = "Histograma para el presupuesto de las películas",  
 xlab = "Presupuesto",  
 breaks = 105)



* QQnormal: dado ese comportamiento y el sesgo que tienen los datos, se aplicó una gráfica de qqnorm.En la que se oberva que los extremos no se comportan de manera normal.

qqnorm(data2$budget, pch = 1, frame = FALSE)  
qqline(data2$budget, col = "steelblue", lwd = 2)



* Test de Lilliefors: tenemos que el valor p es menor a 0.05 con lo que se confirma que la popularidad **No exibe un comportamiento normal**

lillie.test((data2$budget))

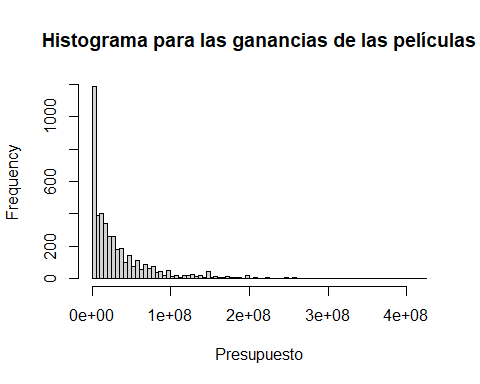
##   
## Lilliefors (Kolmogorov-Smirnov) normality test  
##   
## data: (data2$budget)  
## D = 0.20528, p-value < 2.2e-16

### Ganancias

Para el estudio de estos datos, se analizaron películas que hayan ganado más de $1,000,000 USD debido a que se detectaron 2 escalas distintas en el set de datos. Además de esto no se sabe si la escala fue cambiada alguna otra vez.

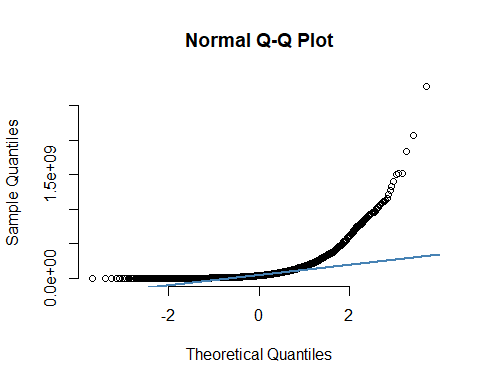
* Histograma: En el histograma nos damos cuenta que esta no sigue ninguna distribución de normalidad, nos damos cuenta que no posee simetría.

data21 = data[data$revenue > 1000000,]  
hist(data21$budget,  
 main = "Histograma para las ganancias de las películas",  
 xlab = "Presupuesto",  
 breaks = 100)



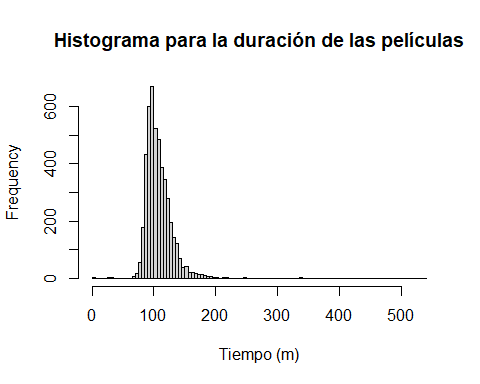
* QQnormal: dado ese comportamiento y el sesgo que tienen los datos, se aplicó una gráfica de qqnorm.En la que se oberva que la distribución no se comporta de manera normal.

qqnorm(data21$revenue, pch = 1, frame = FALSE)  
qqline(data2$revenue, col = "steelblue", lwd = 2)

 ### Duración Para estos datos se utilizaron solamente las películas que duran más de 30 minutos, esto debido a que se observó que la base de datos no se solo peliculas, sino que también cortos y en ciertos casos se toma como la duración del corto como 0.Es por esto que, con el fin de evitar utilizar datos de películas con longitud de 0 minutos, se utilizó la cota inferior de 30 minutos. De cualquier forma, esto hace que los datos pasen de 10866 a 10625. Por lo que se considera que no afecta al estudio de la población.

\*Histograma: Podemos darnos cuenta que en este caso el tiempo parece comportarse con normalidad, como uno cabría esperar ya que, según nuestra experiencia la mayoría de películas rondan la hora y media. Esto nos lo comprueba el analisis exploratorio del punto 1. Donde el tiempo promedio es de 102 minutos (1hr y 42 minutos).

data3 = data[data$runtime > 30,]  
hist(data2$runtime,  
 main = "Histograma para la duración de las películas",  
 xlab = "Tiempo (m)",  
 breaks = 105)



# Variables cualitativas: