EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

Tecnológico Nacional De México

Instituto Tecnológico De Tijuana

Subdirección Académica

Departamento de Sistemas y Computación

Semestre Enero - Junio 2022

Ingeniería Informática

Mineria De Datos

Práctica Evaluatoria

Unidad 2

Perez Ortega Victoria Valeria **No.18210718**

Díaz Ruiz Uriel **No.18210839**

JOSE CHRISTIAN ROMERO HERNANDEZ

Tijuana, B.C. a 04 de Mayo de 2022.

TECNOLÓGICO NACIONAL DE MÉXICO

INSTITUTO TECNOLÓGICO DE TIJUANA

SUBDIRECCIÓN ACADÉMICA

Departamento de Sistemas y Computación

EXAMEN

Carrera: Ingeniería En Sistemas Computacionales/ Tecnologías de la información/ Informática Período: Febrero-Junio 2022 Materia: Minería de datos Grupo: BDD-1703SC9C Salón: Unidad (es) a evaluar: Unidad 1 Tipo de examen:Practico Fecha: Catedrático: José Christian Romero Hernandez Firma del maestro: Calificación:

Alumno: Díaz Ruiz Uriel y Perez Ortega Victoria Valeria No. Control: 18210839, 18210718

Instrucciones

Desarrolle el siguiente problema con R y RStudio para la extracción de conocimiento

que el problema requiere.

Los directivos del sitio web de reseñas de películas están muy contentos con su entrega anterior y ahora tienen una nueva solicitud para usted.

El consultor anterior había creado una gráfica para ellos que se ilustra en el siguiente imagen.

Sin embargo, el código R utilizado para crear la gráfica se ha perdido y no puede ser recuperado .

Su tarea es crear el código que volverá a crear la misma gráfica haciendo que se vea lo más cerca posible del original.

Se le proporcionará con un nuevo conjunto de datos el cual puedes encontrar en este link:

https://github.com/jcromerohdz/DataMining/blob/master/Datasets/Project-Data.csv

O si ya tienen clonado mi repositorio entonces se encuentre en la carpeta:

DataMining/Datasets/Project-Data.csv

Pista

- Tenga en cuenta que no todo los Géneros (Genre) y estudios (Studio) son usados.
- Necesitarás filtrar tu dataframe después de importar el archivo csv.

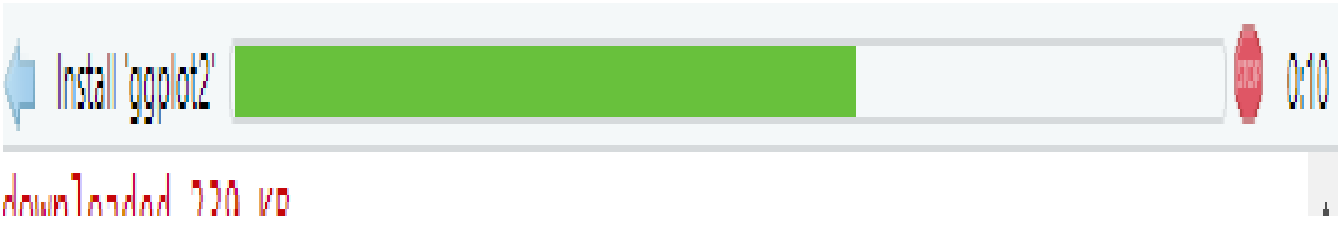
Instrucciones de evaluación

- Tiempo de entrega 4 días
- Al terminar poner el código y la documentación con su explicación en el branch correspondiente de su github, así mismo realizar su explicación de la solución en su google drive en documento de google (Portada, Introducción, Desarrollo, etc).
- Finalmente defender su desarrollo en un video de 6-8 min explicando su solución y observaciones, este servirá para dar su calificación de esta práctica evaluatoria, este video debe subirse a youtube para ser compartido por un link público (Utilicen algún software de captura de video con las cámaras encendidas y graben su defensa para elaborar el video).

Hay que importar primero nuestra biblioteca mediante el comando **ggplot2,** al correr esa línea de código se instalará automáticamente para así poder utilizar el comando de gráficos de RStudio.
que sería así:

```
install.packages("ggplot2")

library(ggplot2)
```



Configuración de nuestro directorio para la práctica evaluatoria para poder comprobar donde se encuentra localizada en nuestro directorio para cargar el archivo CSV. :

```
getwd()

setwd("C:/Users/udr97/Desktop/Unit_2")

getwd()

movies <- read.csv("Project-Data.csv")
```

Selección de columnas que utilizaremos para la graficación:

```
movies <- movies [c(3,6,8,18)]
```

Implementación de las columnas que utilizaremos:

```
movies<-movies[
movies$Genre=="action"|
movies$Genre=="adventure"|
movies$Genre=="animation"|movies$Genre=="comedy"|
movies$Genre=="drama",]

movies<-movies[
movies$Studio=="Buena Vista Studios"|
movies$Studio=="Fox"| movies$Studio=="Paramount Pictures"|
```

```
movies$Studio=="Sony"|movies$Studio=="Universal" |
movies$Studio=="WB",]
```

Debemos cambiar los nombres de nuestras variables de nuestro archivo CSV. a las que tenemos como en nuestra práctica de RStudio.

```
colnames(movies) <- c("Genre", "Studio", "BudgetInMillions",
"GrossInUS")
```

```
> colnames(movies) <- c("Genre", "Studio", "BudgetInMillions", "GrossUS")
```

Debemos obtener los datos mediante el conjunto de datos mediante el archivo de CSV. por medio de una estructura:

```
str(movies)
```

```
'data.frame':    423 obs. of  4 variables:
 $ Genre           : Factor w/ 15 levels "action","adventure",..:
1 1 5 1 1 2 1 1 3 8 ...
 $ Studio          : Factor w/ 36 levels "Art House Studios",..: 2
2 25 25 25 2 31 31 34 25 ...
 $ BudgetInMillions: num  170 66 42 150 80 50 85 70 80 60 ...
 $ GrossUS         : num  44.6 21.4 68.7 35.6 40.5 60.1 60.7 26.1
49.7 39.4 ...
```

Con el sigueinte comando obtenemos un resumen de los datos:

```
summary(movies)
```

Los siguientes datos nos demuestra que solo se muestra los datos ingresados para poder empezar a graficar.

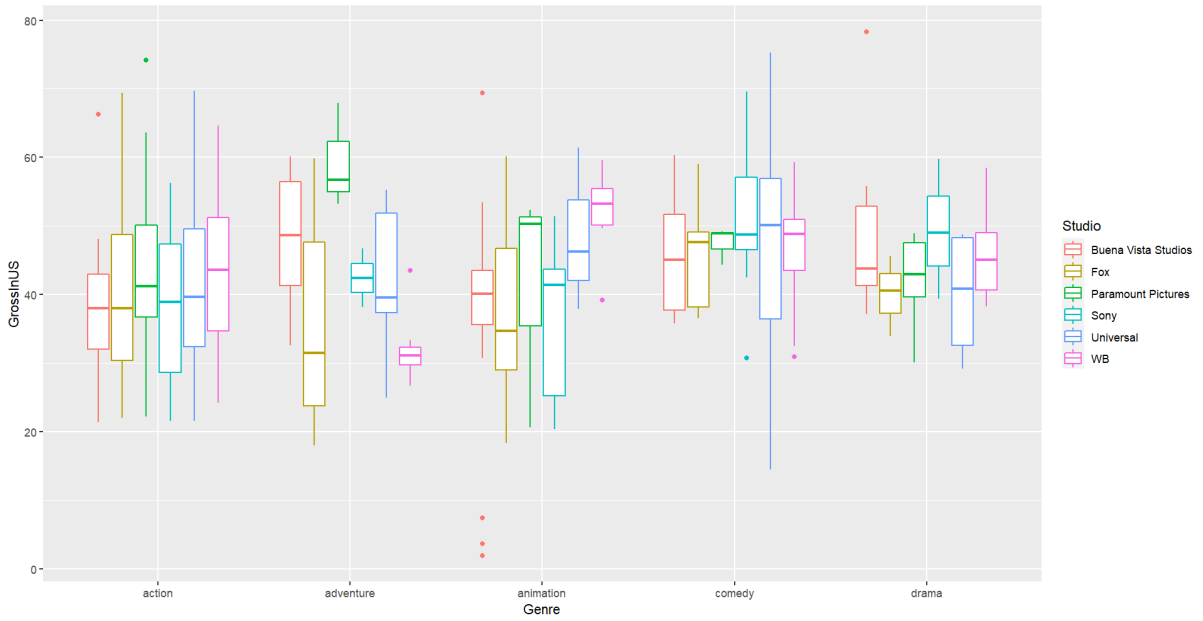
```
> summary(movies)
       Genre                     Studio   BudgetInMillions
GrossUS
 action    :203   Buena Vista Studios:88   Min.    :  3.5    Min.
: 2.00
 animation: 76   WB                  :78   1st Qu.: 55.0    1st
Qu.:33.60
 comedy    : 70   Fox                 :75   Median : 90.0    Median
:41.40
 adventure: 41   Universal           :73   Mean   :101.5    Mean
:41.98
 drama     : 33   Sony                :58   3rd Qu.:140.0    3rd
Qu.:49.75
 biography:  0   Paramount Pictures  :51   Max.   :300.0    Max.
:78.30
 (Other)   :  0   (Other)             : 0
>
```

Ya para poder graficar nuestro archivo CSV. Debemos de definir nuestra variable gráfica mediante nuestra biblioteca ggplot que nos permita establecer nuestros ejes “X” y “Y”, necesitaremos aplicar un color para los Studios y con eso poder utilizar nuestra información de nuestra variable de BudgetInMillions.

```
BoxPlot <- ggplot(movies, aes(x=Genre, y=GrossInUS, color=Studio,
size=BudgetInMillions))
```

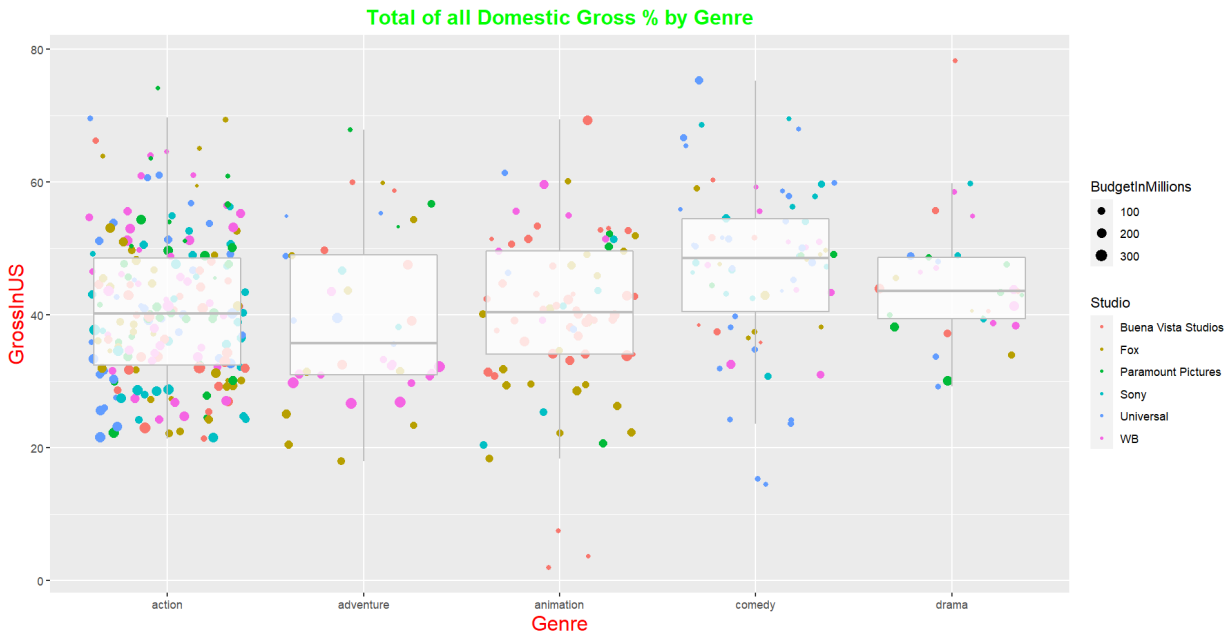
Aplicaremos el siguiente comando para poder indicar a nuestra gráfica su tipo el cual utilizamos el de diagrama de caja.

```
BoxPlot + geom_boxplot()
```



Ahora faltaría implementar los siguientes comandos para que nos muestre los resultados obtenidos del archivo CSV. debemos de establecer una estimación del límite de opacidad de nuestros diagramas de caja para que con eso pueda ser visible su dispersión de los puntos que son necesarios para nuestra práctica, además de poner etiquetas que nos identifique los ejes “X” y “Y”.

```
BoxPlot + geom_jitter(shape=20) +
geom_boxplot(size=0.5,alpha=0.8,color="Grey",
outlier.shape = NA)+ theme(plot.title =
element_text(color="Green", size=16, face="bold", hjust = 0.5),
axis.title.x = element_text(color="Red", size=16), axis.title.y =
element_text(color="Red", size=16))+  ggtitle("Total of all
Domestic Gross % by Genre") + ylab("GrossInUS") + xlab("Genre")
```



URL: https://youtu.be/V7tjtX1UjvE