National Technology of Mexico
Technological Institute of Tijuana

ACADEMIC SUBDIRECTION
Systems and Computing Department

SEMESTER
February – June 2021

CAREER
Information and Communication Technologies Engineer

SUBJECT AND KEY:
Data Mining BDD-1703TI9A

STUDENT'S NAME AND REGISTRATION:
Velázquez Farrera César Alejandro 17212937

NAME OF THE JOB:
Evaluative Practice – Unit 2

UNIT TO BE EVALUATED
Unit II

TEACHER'S NAME:
Mc. José Christian Romero Hernández

Due date: 5/14/2021

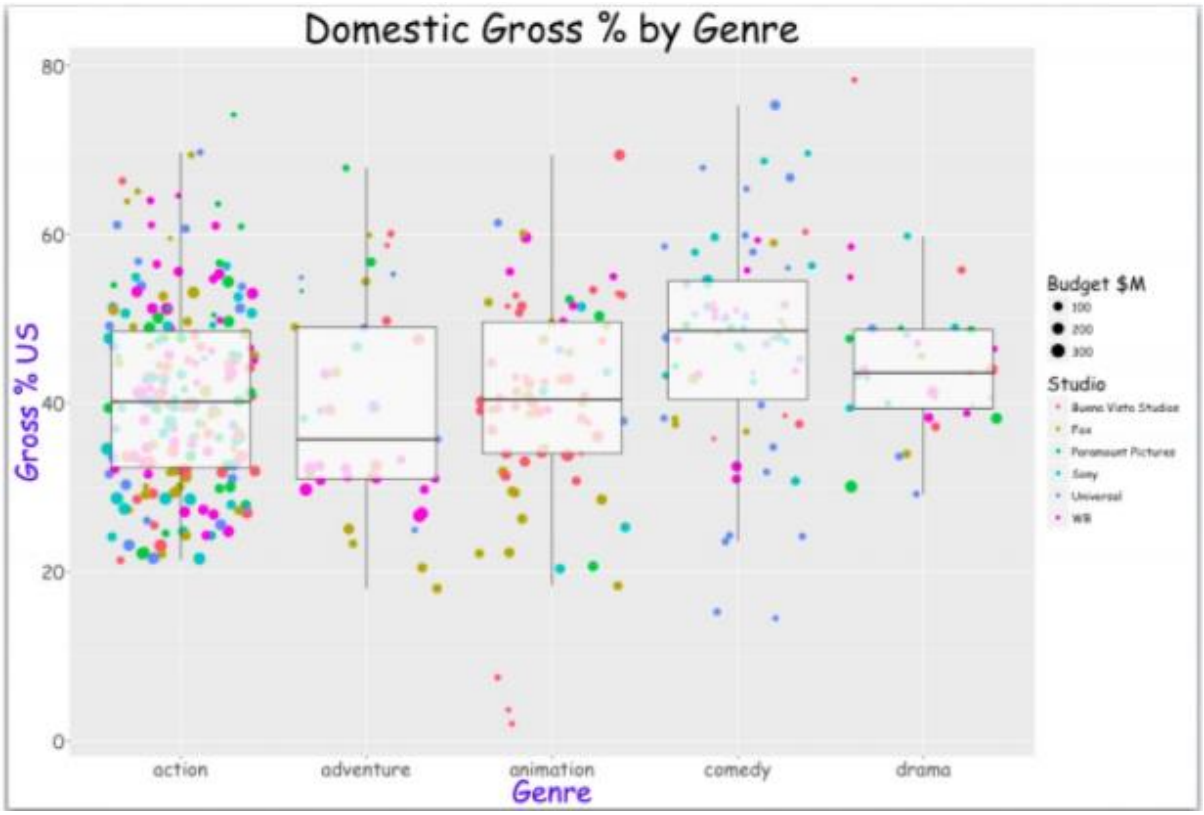
Introduction

In this evaluative practice, we will try to recreate the graph (which is shown below) using new code. The data provided for the exam is located in the following Github repository: [Project-Data.csv]

Evaluative practice

Scenario: Develop the following problem with R and RStudio for the knowledge extraction that the problem requires.

The managers of the movie review website are very happy with their previous installment and now they have a new requirement for you. The previous consultant had created a chart for them which is illustrated in the following image.



However, the R code used to create the graph has been lost and cannot be recovered.

Our task is to create the code that will recreate the same graph making it look as close to the original as possible.

Practice Development

1. After copying the '**Project-Data**' file with '**.csv**' extension, in the R file in RStudio you need to set the working directory.

```
setwd('\Desktop\\Tareas\\8vo Semestre\\Minería de
datos\\source\\Unit_2\\Evaluations')
```

2. A variable called "datos" is created to save the values of the previously mentioned csv file.

```
datos = read.csv('Project-Data.csv')
```

3. The data is explored with the following functions.

```
#Exploring Data
datos
head(datos)
tail(datos)
colnames(datos)
nrow(datos)
ncol(datos)
summary(datos)
```

4. Out of curiosity, we want to see what the dimensions of the data frame are. The following code is used to see the dimensions of the data frame.

```
#Shape of the Data Frame
cat(nrow(datos), " x ", ncol(datos))
```

5. The libraries 'dplyr' and 'ggplot2' are imported. With dplyr, the "filter" function is used for data filtering. With the ggplot2 library, it is intended to make the graphs.

```
library(dplyr)
library(ggplot2)
```

6. Data requires cleaning. Only movies of the following genres are required: **{"action", "adventure", "animation", "comedy", "drama"}**

```
Movies <- filter(datos, Genre == "action" | Genre == "adventure" |
Genre == "animation" |
```

```
        Genre == "comedy" | Genre == "drama")
```

7. Only records are needed where the column "Studio" equals one of the following records:
   {“Buena Vista Studios”, “Fox”, “Paramount Pictures”, “Sony”, “Universal”, “WB”}

```
DF_Movies_filtered <- filter(Movies, Studio == "Buena Vista
Studios" | Studio == "Fox" | Studio == "Paramount Pictures"
|Studio == "Sony" | Studio == "Universal" | Studio == "WB")
```

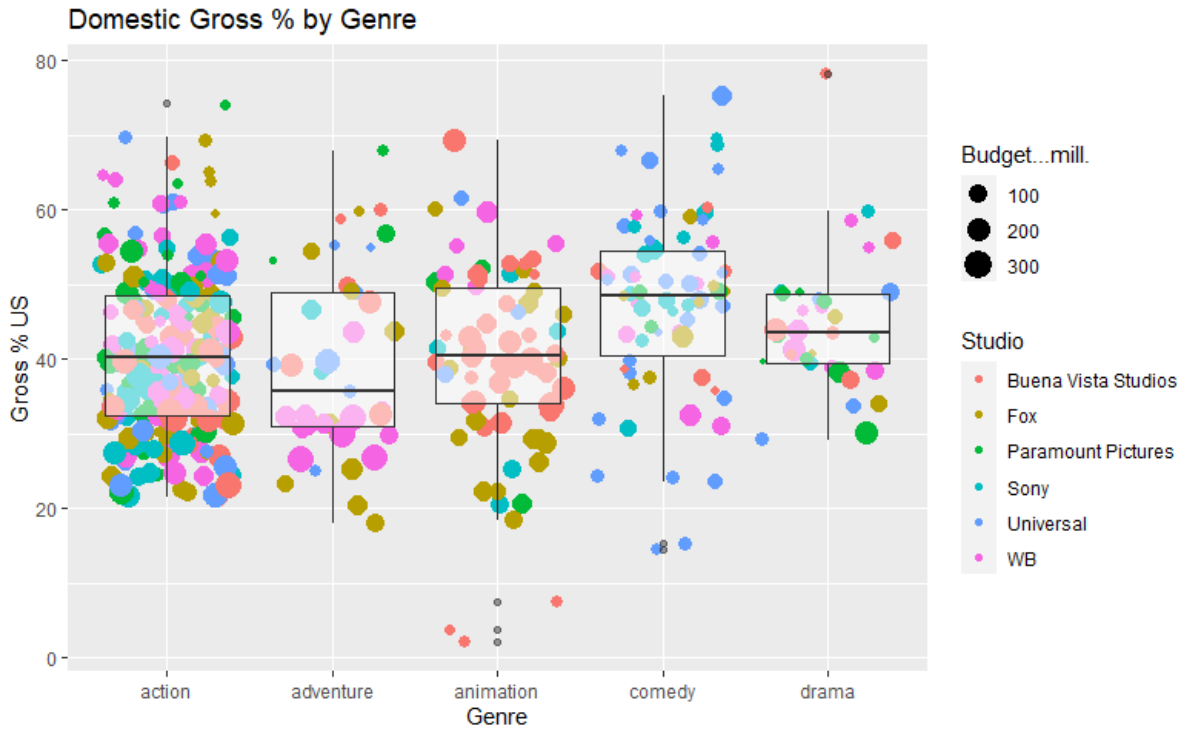
8. In the next three lines of code, the graph will be replicated. The first line indicates that the variable "u" will contain the graph, the gender data and the percentage of earnings in the United States.

```
u <- ggplot(DF_Movies_filtered, aes(x=Genre, y = Gross...US))
```

   The next lines adds the color of the data, the size and the boxplot.

```
u + geom_jitter(aes(color=Studio, size=Budget...mill.)) +
geom_boxplot(alpha=0.5) +
  xlab("Genre") + ylab("Gross % US") + ggtitle("Domestic Gross %
by Genre")
```

Generated Scatterplot



link to the repo: https://github.com/vcesar454/Data_Mining/tree/Unit_2/Evaluations
link to a YouTube video: https://youtu.be/UgRLqpHkPVE