Tecnológico Nacional de México
Instituto Tecnológico de Tijuana

Subdirección Académica
Departamento de Sistemas y Computación

Semestre:
Febrero – Junio 2021

Carrera:
Ingeniería en Tecnologías de la Información y Comunicaciones

Materia y serie:
Minería de datos BDD-1703TI9A

Unidad a evaluar: Unidad II

Nombre de la Tarea:
Práctica Evaluatoria - Unidad 2

Nombre del Alumno:
Hernandez Negrete Juan Carlos

Nombre del docente:
José Christian Romero Hernández

Develop the following problem with R and RStudio using dataframes for the extraction of knowledge that the problem required.

The objective of this practice is to apply the different layers seen throughout unit two in a real environment, recreating a graph seen in class, which represented the percentage of gross profit with respect to genres, showing differences in colors and sizes according to the studies and budgets invested. To fulfill the proposed task, a CSV file was provided, but it contained additional data to that used to build the previous graph, so, as an additional step, a data filtering had to be done. All the steps used to solve the problem are explained below:

To load the CSV file on the data related to the movies into memory, the read.csv function is used in conjunction with file.choose, so as not to use the full path in the which the file is hosted, but to do it dynamically. The head, tail, str, and summary functions are to perform a quick analysis on the data and determine if additional actions must be performed in order to build a graph that behaves correctly with the desired data.

```
movies <- read.csv (file.choose ())

head (movies)
tail (movies)
str (movies)
summary (movies)
```

Afterwards, the libraries necessary to carry out the practice are loaded into memory. To filter the data, dplyr is used, extrafont is used to use more fonts than those provided by default by R studio and ggplot2 is what allows using the ggplot function to create the graphs used.

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

To use additional fonts, they must be downloaded separately, so a folder was created and the directory changed through the setwd function. To import and load into memory the font_import function is used, in this way it is possible to make use of these fonts. To see the available ones, use windowsFonts.

```
getwd ()
setwd ("../Desktop/Subjects 8th / Data mining / Practices / U2 /")
getwd ()
font_import ("fonts /", prompt = F)
windowsFonts ()
```

The columns belonging to the CSV are rewritten to avoid errors in their writing due to not having control of the previously assigned name. The data filtering is carried out through the filter function, in which the data source is specified as the first parameter, and then only the conditions through which the already filtered result will be delivered must be determined, in

this case, based on Under the given conditions, only the genres and studies for which information is desired are determined.

```
colnames (movies) <- c ("Day","Director",
"Genre","Movie","Date","Studio","Adjusted","Budget","Gross","IMDb",
"MovieLR ", " Overseas ", " OverseasPercent ", " Profit ", "
ProfitPercent ", " Runtime ", " US ", " GrossPercentUS ")

movies <- filter (movies, Genre%in% c (" action ", " adventure ",
"animation", "comedy", "drama"), Studio%in% c ("Buena Vista Studios",
"Fox", "Paramount Pictures", "Sony", "Universal", "WB"))
```

Arrived on When constructing the structure of the graph, with the help of the aesthetics layer within the ggplot function, the axes "x" and "y" are first determined, which is subsequently saved in the variable "u" to maintain a better order and compression the moment of constructing the graph. To observe data within the graph, the geom_jitter and geom_boxplot functions are used, the first one allowing to show the points belonging to the budget of the films, modifying the color through the aesthetics layer so that it differs according to the different existing studies and the size depending on how much is the budget invested amount, and the second function is for the representation of the candle graph of the mean with respect to all the data, using the alpha parameter for the transparency of the graph and to be able to observe the points of the above function.

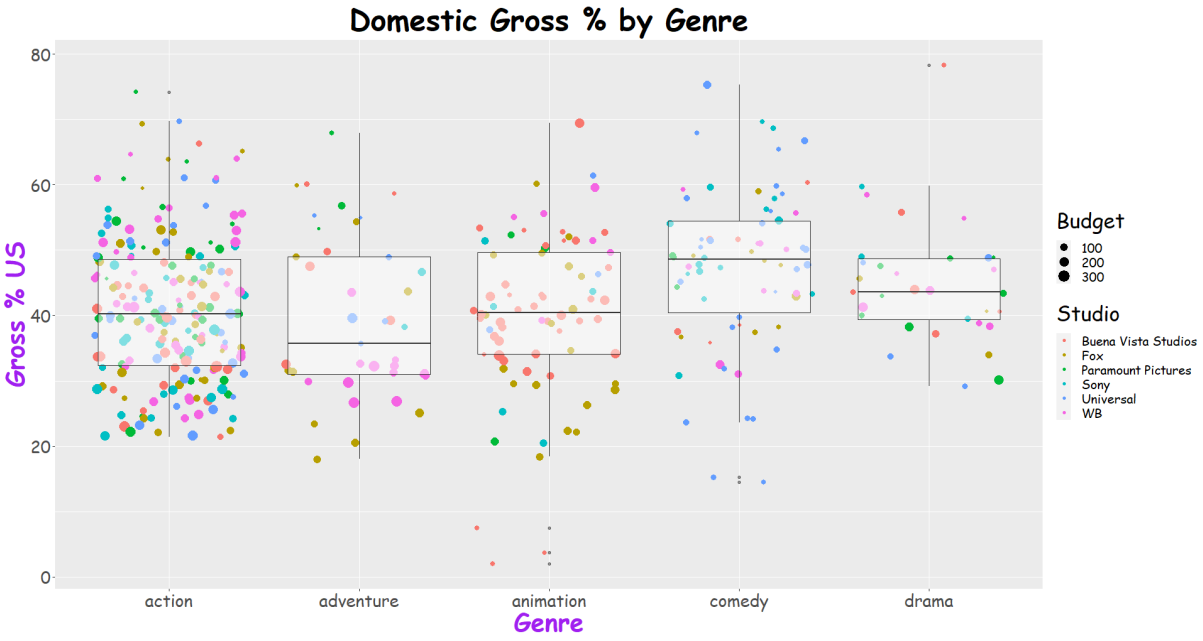
```
u <- ggplot (movies, aes (x = Genre, y = GrossPercentUS))
t <- u + geom_jitter (aes (color = Studio, size = Budget)) +
geom_boxplot (alpha =0.5) In
```

addition to the previously constructed graph, The theme layer is used so that the presentation is made in a more personalized way and according to the needs, being able to change the size, type, color related to the letters of the title, the legend and numbers of the graph, among many other parameters that enables the theme layer.

```
t +
  xlab ("Genre") +
  ylab ("Gross% US") +
  ggtitle ("Domestic Gross% by Genre") +
  theme (axis.title.x = element_text (color = "Purple", size =30, face =
"bold"),
         axis.title.y = element_text (color = "Purple", size =30, face =
"bold"),
         axis.text.x = element_text (size = 20),
         axis.text.y = element_text (size = 20),
         legend.title = element_text (size = 25),
         legend.text = element_text (size = 15),
```

```
        legend.justification = c (1,.5),
        text = element_text (family = "Comic Sans MS "),
        plot.title = element_text (color = " Black ",
                                   size = 35,
                                   hjust = 0.5,
                                   face = " bold "))
```

The resulting graph is as follows, asking to observe the relationship between genders and the gross percentage of profit according to the percentage invested per study.



Repository

https://github.com/JuanCarlos-Negrete/Data-Mining/tree/Unit_2/Unit_2/Evaluation

Video

https://youtu.be/MsnnE4dNkB0