Analítica de Datos

Caso Hollywood

Carlos Cardona Andrade

Hollywood Rules

1. Todo lo que hagamos hoy, escríbanlo en un script con anotaciones.
2. Guarden el script.
3. Enviénmelo al correo con el nombre: nombre_apellido.R

Tengan un estilo para su script

```
## ---------------------------
##
## Script name:
##
## Purpose of script:
##
## Author: Carlos Cardona
##
## Date Created: `r paste(Sys.Date())`
##
## Email: ccardonaa@gmail.com
##
## ---------------------------
##
## Notes:
##
##
## ---------------------------

```

Importemos los datos

```
# Recuerden instalar los paquetes primero!!
# Este lo hemos usado siempre así que no hay lío
library(tidyverse)

# Estos dos son nuevos:
# install.packages("readxl")
# install.paclages("janitor")
# install.paclages("Hmisc")

library(readxl)
library(janitor)
library(Hmisc)


# Establecemos el directorio de trabajo

setwd("C:/Users/ccard/Downloads")

```

¿Qué hay en nuestros datos?

```
  1  glimpse(data)
```

```
Rows: 75
Columns: 18
$ Movie                  <chr> "16 Blocks", "Accepted", "Apocalypto", "Arthur ~
$ `Opening Gross`        <dbl> 11855260, 10023835, 15005604, 4294936, 11554404~
$ `Total U.S. Gross`     <dbl> 36895141, 36323505, 50866635, 15132763, 2117056~
$ `Total Non-U.S. Gross` <dbl> 65664721, 2146261, 69309076, 97854413, 0, 10102~
$ Budget                 <dbl> 4.50e+07, 2.30e+07, 4.00e+07, 8.60e+07, 2.00e+0~
$ `Opening Theatres`     <dbl> 2706, 2914, 2465, 2247, 1602, 1251, 3311, 3261,~
$ `Known Story`          <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0,~
$ Sequel                 <dbl> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0,~
$ `Origin_United States` <dbl> 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1,~
$ Genre                  <chr> "Action", "Comedy", "Adventure", "Animation", "~
$ Summer                 <dbl> 1, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0,~
$ Holiday                <dbl> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0,~
$ Christmas              <dbl> 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0,~
$ MPAA                   <chr> "PG-13", "PG-13", "R", "PG", "PG-13", "R", "PG"
```

Limpiemos los nombres usando janitor

```
1 data <- data %>%
2   clean_names()
3 glimpse(data)
```

```
Rows: 75
Columns: 18
$ movie                <chr> "16 Blocks", "Accepted", "Apocalypto", "Arthur an~
$ opening_gross        <dbl> 11855260, 10023835, 15005604, 4294936, 11554404, ~
$ total_u_s_gross      <dbl> 36895141, 36323505, 50866635, 15132763, 21170563,~
$ total_non_u_s_gross  <dbl> 65664721, 2146261, 69309076, 97854413, 0, 1010273~
$ budget               <dbl> 4.50e+07, 2.30e+07, 4.00e+07, 8.60e+07, 2.00e+07,~
$ opening_theatres     <dbl> 2706, 2914, 2465, 2247, 1602, 1251, 3311, 3261, 1~
$ known_story          <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0, 1~
$ sequel               <dbl> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0~
$ origin_united_states <dbl> 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1~
$ genre                <chr> "Action", "Comedy", "Adventure", "Animation", "Dr~
$ summer               <dbl> 1, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0~
$ holiday              <dbl> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 1~
$ christmas            <dbl> 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0~
$ mpaa                 <chr> "PG-13", "PG-13", "R", "PG", "PG-13", "R", "PG"
```

Estadísticas Descriptivas

```r
# Cambiemos los nombres de estas variables!
data <- data %>% rename(us_gross = total_u_s_gross)
data <- data %>% rename(non_us_gross = total_non_u_s_gross)

# Usamos summary para ver la distribución de las variables de interés
summary(data$opening_gross)
```

```
   Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
4120497 10014865 14503650 17468466 21569368 68033544 
```

```r
summary(data$us_gross)
```

```
    Min.   1st Qu.    Median      Mean   3rd Qu.      Max. 
13090630  33880974  52330111  59620651  74345586 198000317 
```

```r
summary(data$non_us_gross)
```

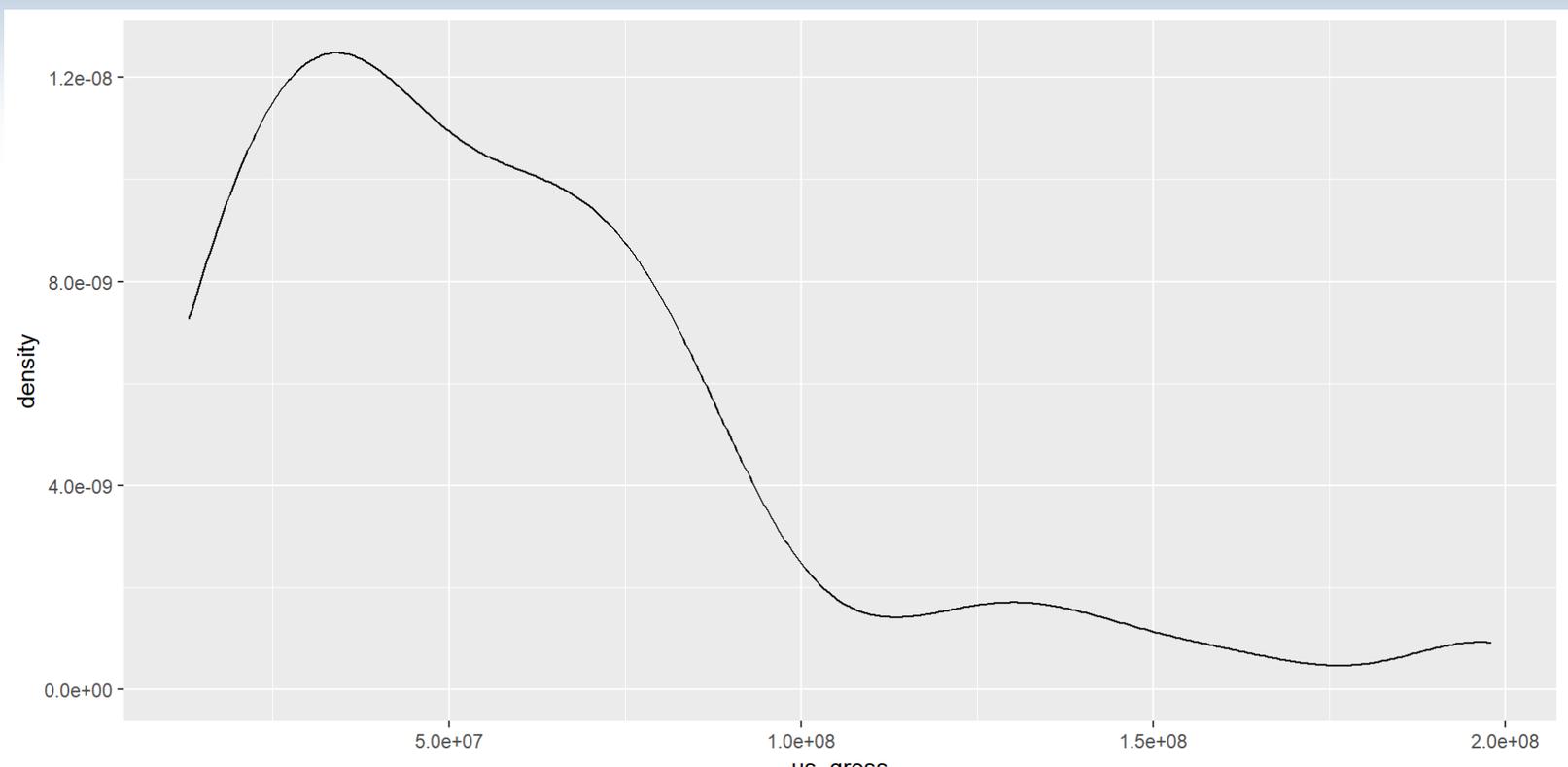
```
    Min.   1st Qu.    Median      Mean   3rd Qu.      Max. 
       0  15433097  42950069  59560983  75985298 456235122 
```

```r
summary(data$opening_theatres)
```

```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   852    2490    2880    2766    3209    3964 
```

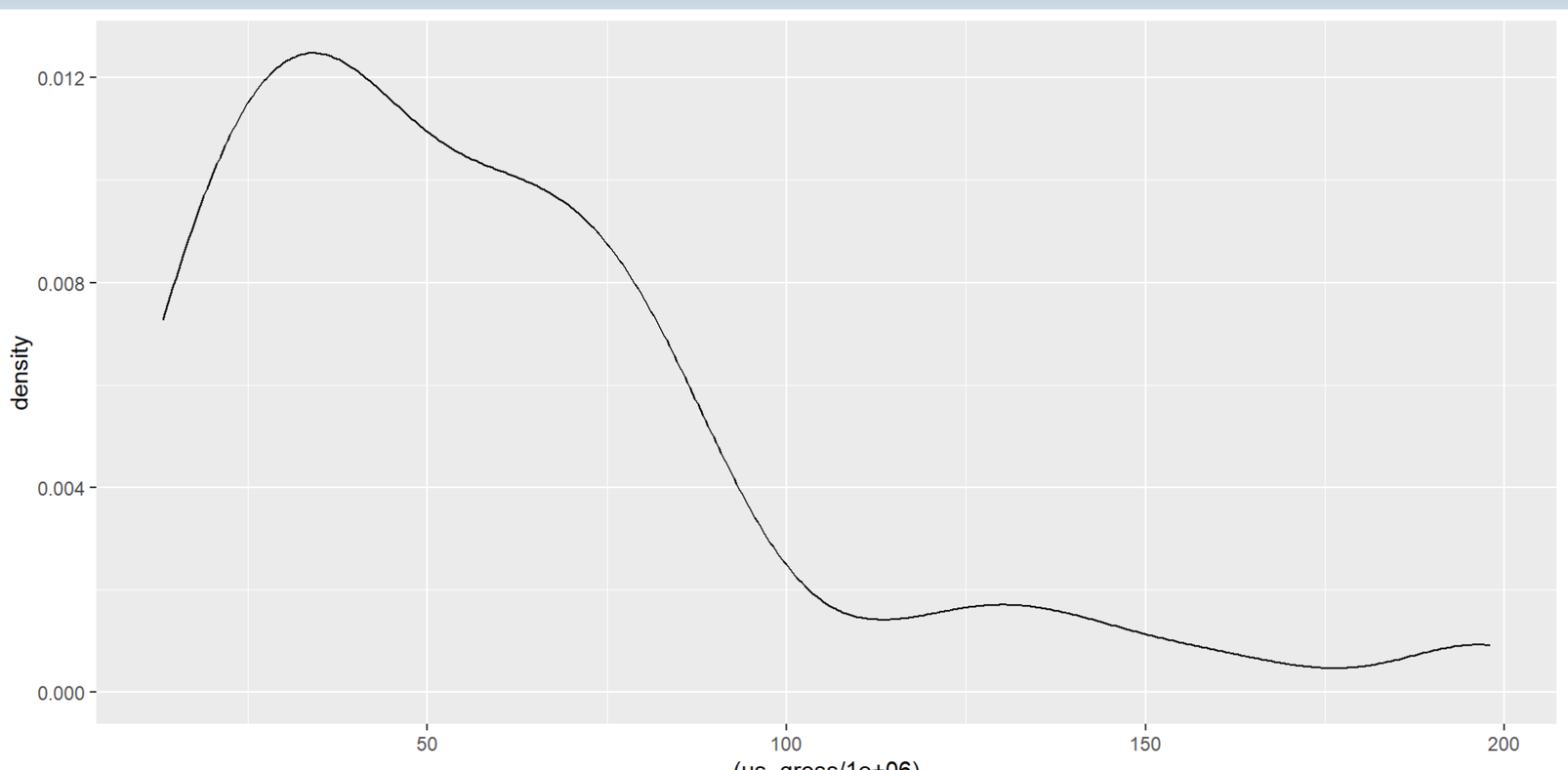
Exploremos el recaudo en US

```
1  # Qué tal es la distribución?
2
3  ggplot(data, aes(x=us_gross)) + geom_density()
```



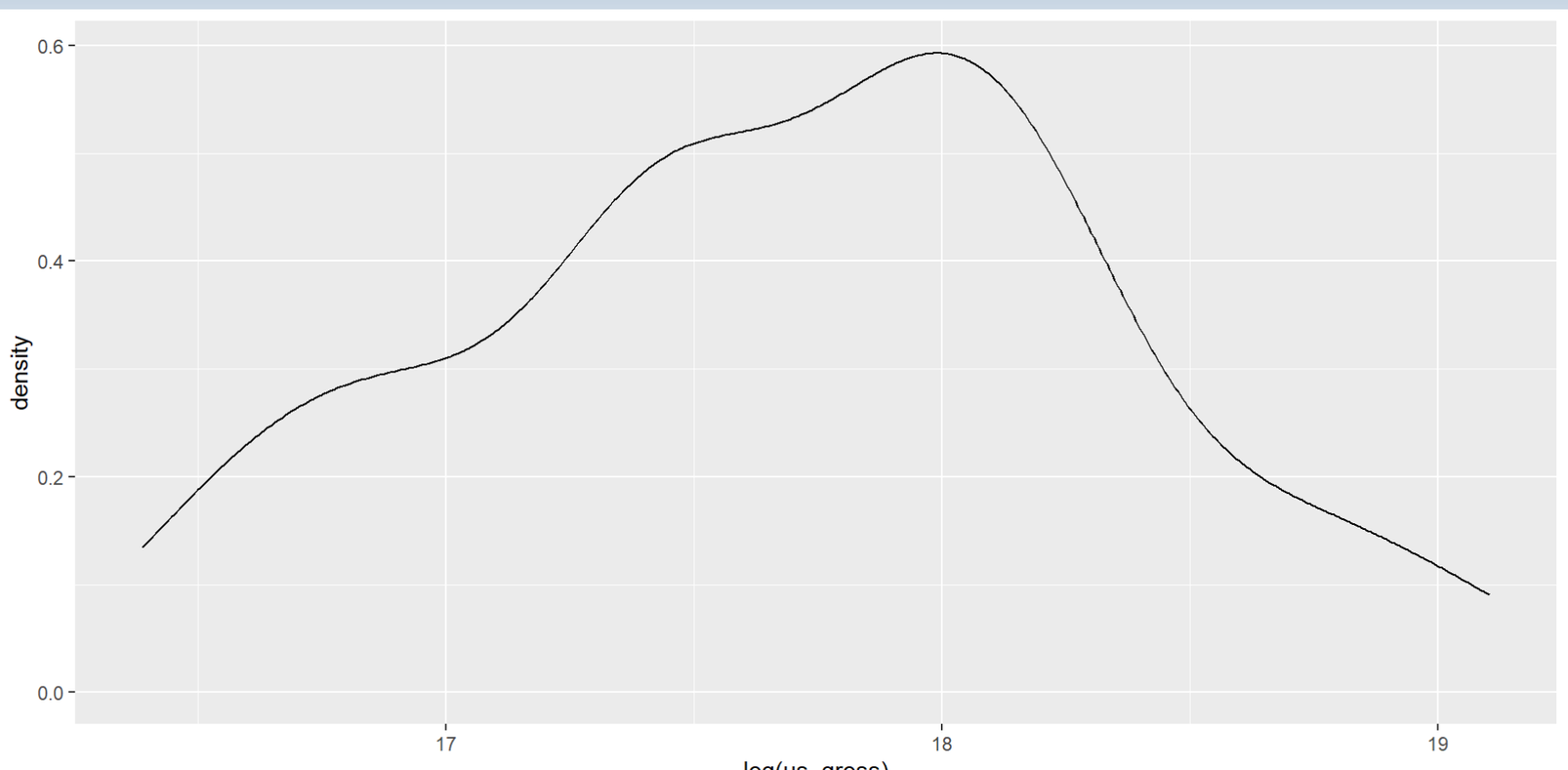
Exploremos el recaudo en US (en millones)

```
1  # Cambiemos los valores de los ejes
2
3  ggplot(data, aes(x=(us_gross/1000000))) + geom_density()
```



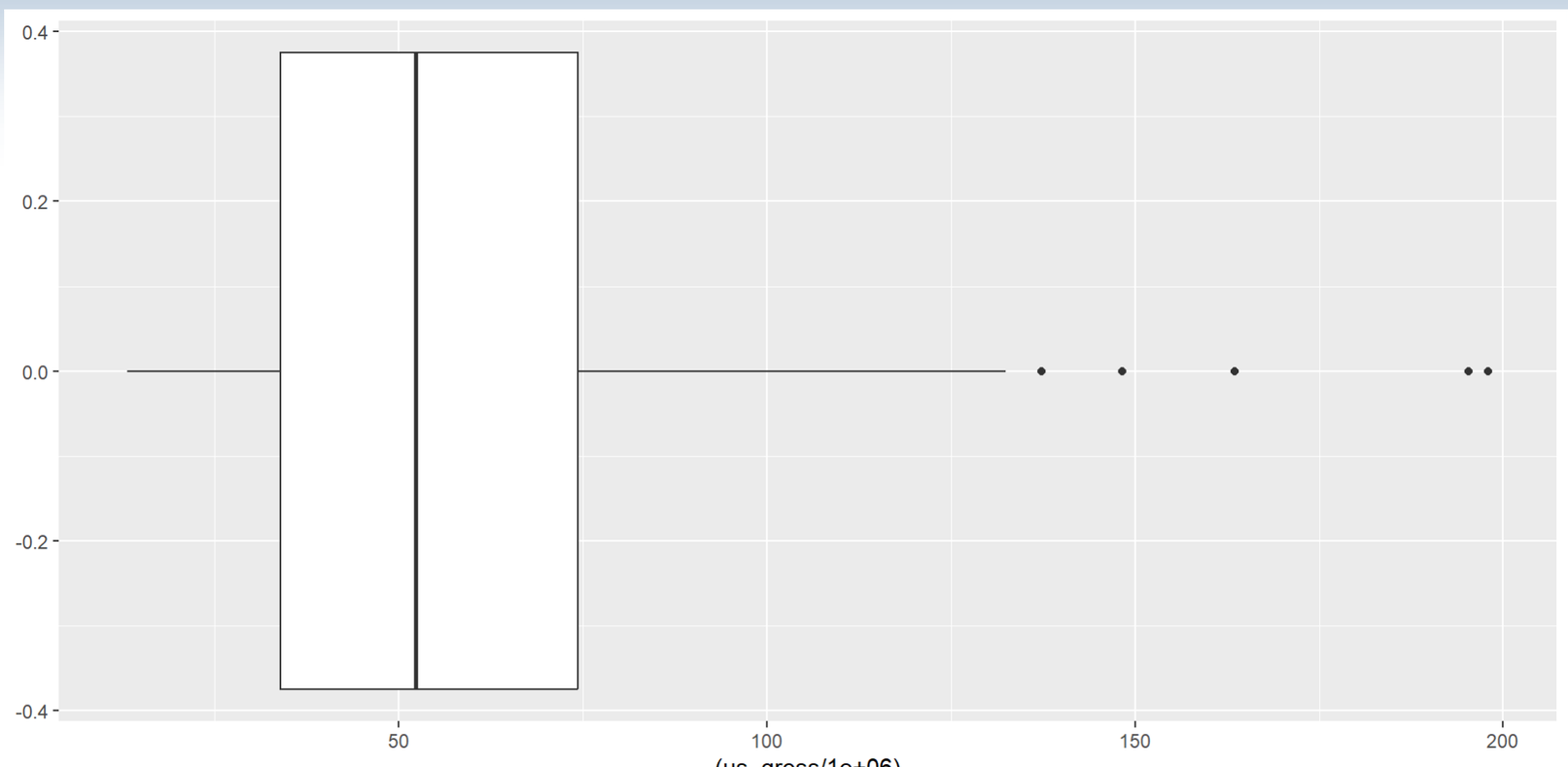
Exploremos el recaudo en US (en logs)

```
1  # Qué tal es la distribución?
2
3  ggplot(data, aes(x=log(us_gross))) + geom_density()
```



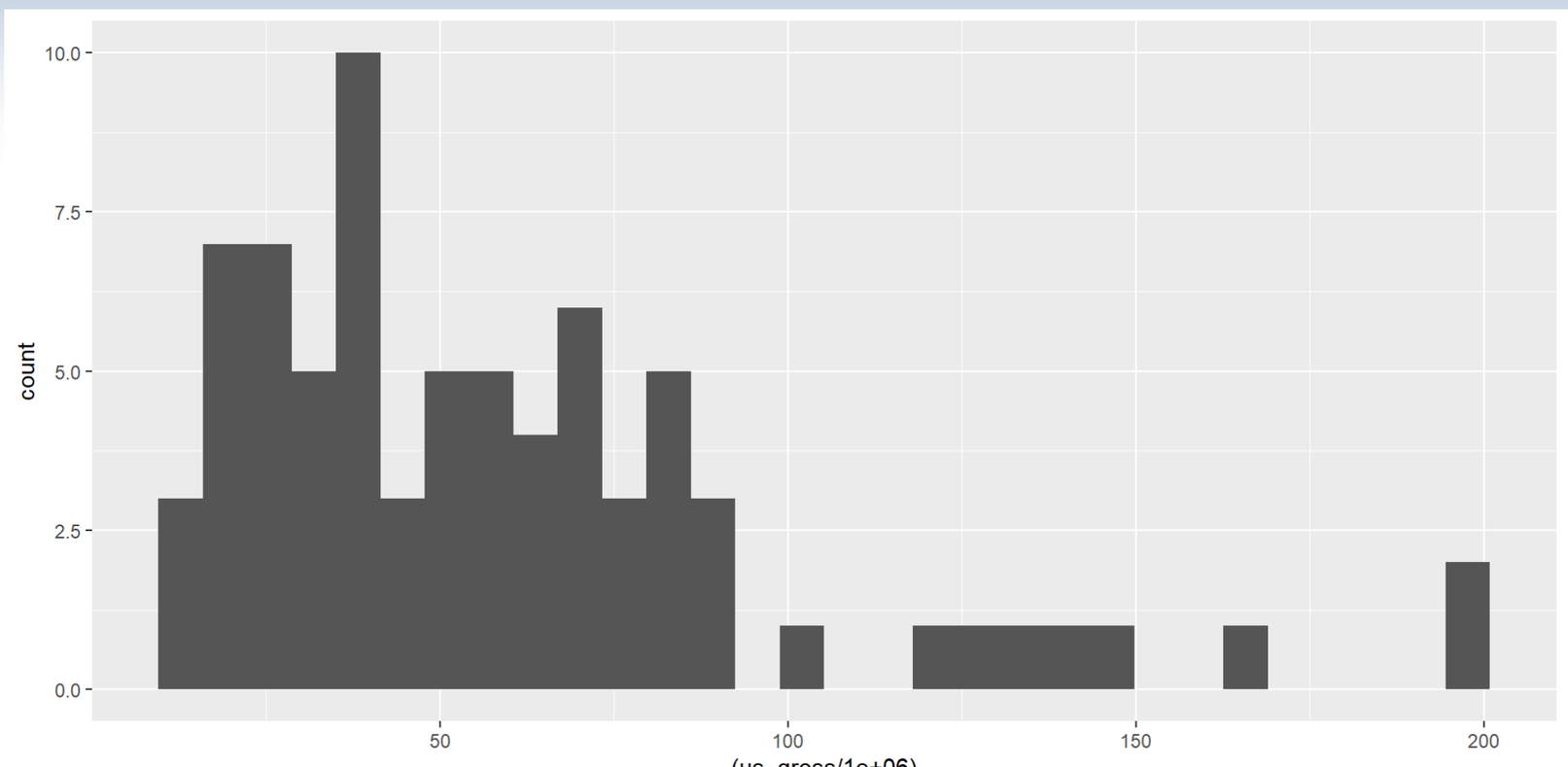
Exploremos el recaudo en US

```
1  # Qué tal es el diagrama de caja?
2
3  ggplot(data, aes(x=(us_gross/1000000))) + geom_boxplot()
```



Exploremos el recaudo en US

```
1  # Qué tal es el histograma?
2
3  ggplot(data, aes(x=(us_gross/1000000))) + geom_histogram()
```



Comedias y R-Rated

```
1  # Usamos la función xtabs
2  # ¿Cuántas comedias?
3  xtabs(~genre, data=data)
```

```
genre
   Action Adventure Animation    Comedy     Drama    Horror  Thriller
        8         2         8        23        19         9         6
```

```
1  # Y cuántas R-rated?
2  xtabs(~mpaa, data=data)
```

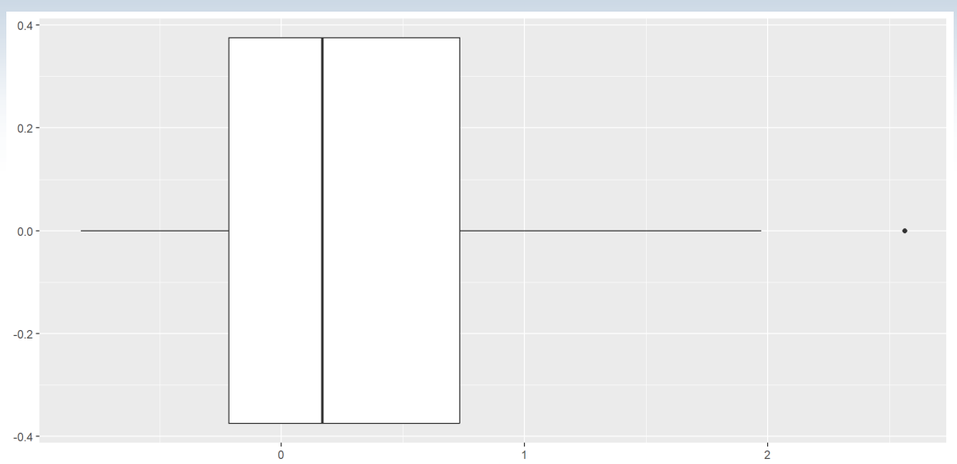
```
mpaa
    G    PG PG-13     R
    3    20    37    15
```

Calculemos el ROI

```r
# Calculamos ROI usando "mutate"
data <- data %>%
  mutate(roi=(us_gross-budget)/budget)

# Es 12%?
mean(data$roi)
```

```
[1] 0.2929317
```

```r
# Veamos su distribución
ggplot(data, aes(x=roi)) + geom_boxplot()
```



Calculemos un IC para ROI al 95%

```r
# Usamos el comando "t.test"
result <- t.test(data$roi, conf.level = 0.95)
result
```

```
    One Sample t-test

data:  data$roi
t = 3.6914, df = 74, p-value = 0.0004237
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.1348149 0.4510486
sample estimates:
mean of x
0.2929317
```

```r
# En IC guardamos el intervalo
ic<- result$conf.int
ic
```

```
[1] 0.1348149 0.4510486
attr(,"conf.level")
[1] 0.95
```

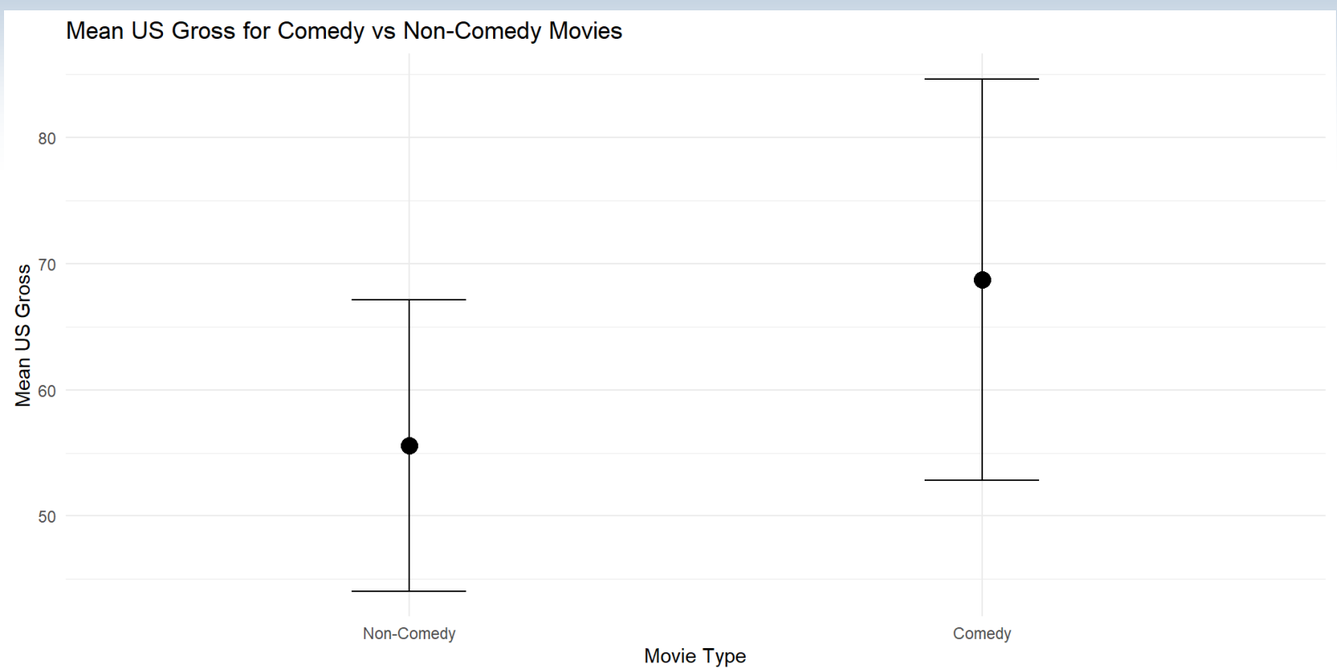
Comedy vs No Comedy

```r
# Construimos una variable indicando si es comedia o no
data <- data %>%
  mutate(comedy = ifelse(genre == "Comedy", TRUE, FALSE))

# Hacemos un test comparando 'us_gross' para comedias vs no comedias
t_test_result <- t.test(us_gross ~ comedy, data = data)

# ¿Cuál es el resultado?
t_test_result
```

```
    Welch Two Sample t-test

data:  us_gross by comedy
t = -1.3728, df = 47.176, p-value = 0.1763
alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
95 percent confidence interval:
 -32437354   6122596
sample estimates:
mean in group FALSE  mean in group TRUE
           55585721             68743100
```

Comedy vs No Comedy

```r
ggplot(data, aes(x = factor(comedy, labels = c("Non-Comedy", "Comedy")),
                 y = (us_gross/1000000))) +
  stat_summary(fun = mean, geom = "point", size = 4) +  # Graficamos la media como puntos
  stat_summary(fun.data = mean_cl_normal, geom = "errorbar", width = 0.2) +  # Añadimos la barra de e
  labs(x = "Movie Type", y = "Mean US Gross", title = "Mean US Gross for Comedy vs Non-Comedy Movies"
  theme_minimal() # Esto es para que el fondo sea blanco y no gris
```



Comedy vs No Comedy (ROI)

```r
# Hacemos un test comparando 'roi' para comedias vs no comedias
t_test_roi <- t.test(roi ~ comedy, data = data)

# ¿Cuál es el resultado?
t_test_roi
```

```
    Welch Two Sample t-test

data:  roi by comedy
t = -2.0471, df = 38.965, p-value = 0.04743
alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
95 percent confidence interval:
 -0.708944798 -0.004248791
sample estimates:
mean in group FALSE  mean in group TRUE
          0.1835754           0.5401722
```

R-Rated vs NonR-Rated

```r
# Creamos el indicador para R-rated movies
data <- data %>%
  mutate(r_rated = ifelse(mpaa == "R", TRUE, FALSE))

# Hacemos el test
t_test_rated <- t.test(us_gross ~ r_rated, data = data)

# Qué nos dice?
t_test_rated
```

```
    Welch Two Sample t-test

data:  us_gross by r_rated
t = 0.85686, df = 31.986, p-value = 0.3979
alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
95 percent confidence interval:
 -10829257  26555104
sample estimates:
mean in group FALSE  mean in group TRUE
           61193236             53330312
```

R-Rated vs NonR-Rated

```
ggplot(data, aes(x = factor(r_rated, labels = c("NonR-Rated", "R-Rated")),
                 y = (us_gross/1000000))) +
  stat_summary(fun = mean, geom = "point", size = 4) +  # Graficamos la media como puntos
  stat_summary(fun.data = mean_cl_normal, geom = "errorbar", width = 0.2) +  # Añadimos la barra de e
  labs(x = "Movie Type", y = "Mean US Gross", title = "Mean US Gross for R-Rated vs NonR-Rated Movies
  theme_minimal() # Esto es para que el fondo sea blanco y no gris
```