Lab 2 EDA

Austin Sanders + Martin Lim

11/15/2021

1. Load the Data

```
# Read the dataset
movies <- read.csv(file = "movies.csv")
```

Outcome variable:
`revenue` - Must be $10,000,000 or more

Explanatory Variables:
- `budget` - Might require a log-transform
- `rating` - Filter to indicator variables

2. Get Necessary Columns

```
# Retrieve only the needed columns

df_raw <- data.frame(movies)
```

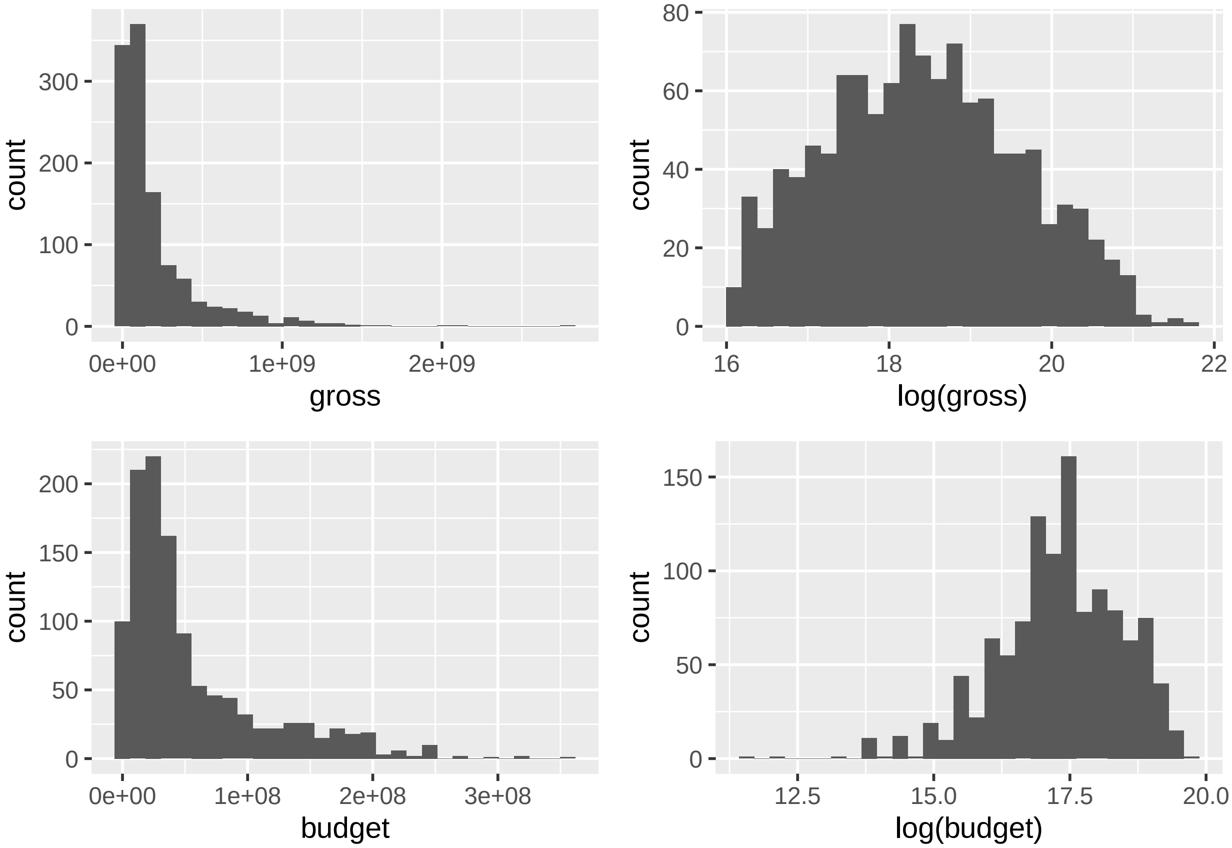
3. Filter out Low-Budget, Low-Revenue, and Duplicate Titles

```
# Remove world_revenue under MIN_REVENUE
df_raw <- subset(df_raw, df_raw$gross >= MIN_REVENUE & !is.na(df_raw$gross))

# Remove budget under MIN_BUDGET
df_raw <- subset(df_raw, df_raw$budget >= MIN_BUDGET & !is.na(df_raw$budget))

# Remove N/A ratings if desired
if (REMOVE_NA_RATING) {
  df_raw <- subset(df_raw, df_raw$rating != "Not Rated")
}

# Hash for title : budget
h <- hash()

# Clean dataframe
df = data.frame()

for(i in 1:nrow(df_raw)) {        # for-loop over rows
  title_key = df_raw[i,1]
  if (has.key( title_key, h )) {
    # Title Is already recorded

    # Search for existing row in clean dataframe with the same title
    for (k in 1:nrow(df)) {

      if (title_key == df[k,1]) {# tolower(title_key) == tolower(df[k,1])
        # Replace row if the budget of the new value is higher than that of the
        # budget of the recorded title
        if (df_raw[i, 12] > df[k, 12]) {

          # Delete found row in cleaned dataframe
          df = df[!k,]

          # Bind raw dataframe row to clean dataframe
          df <- rbind(df, df_raw[i,])

          # Revise title_key and budget to hash
          h[[title_key]] = df_raw[i,12]
        }

        break
      }

    }

  } else {
    # Add title_key and budget to hash
    h[[title_key]] = df_raw[i,12]

    # Bind raw dataframe row to clean dataframe
    df <- rbind(df, df_raw[i,])
  }
}

#df = df_raw
nrow(df)
```

```
## [1] 1155
```

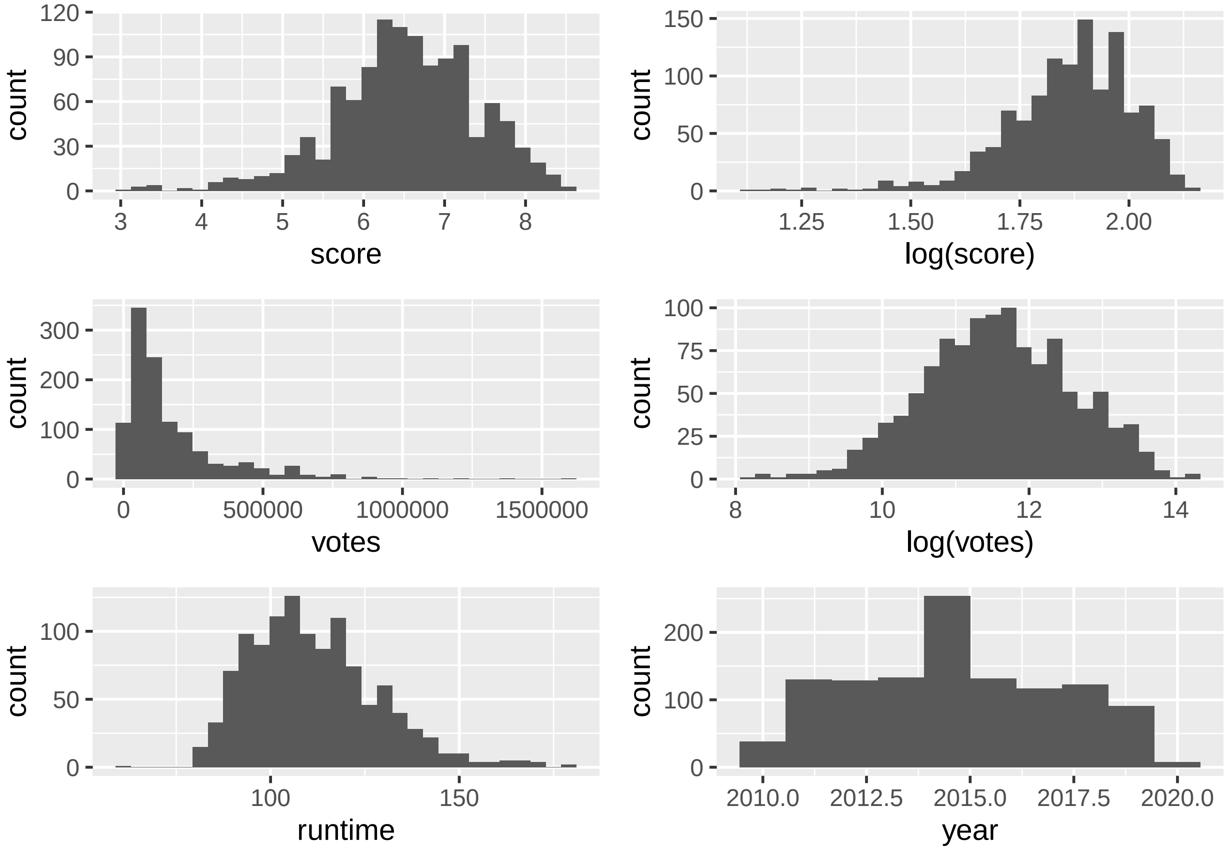
4. Examine Variables (apply transformations)

```
# CHECK MAIN NUMERIC VARIABLES
world_revenue_histogram <- df %>%
  ggplot(aes(gross)) +
  geom_histogram(bins=30)

log_world_revenue_histogram <- df %>%
  ggplot(aes(log(gross))) +
  geom_histogram(bins=30)

budget_histogram <- df %>%
  ggplot(aes(budget)) +
  geom_histogram(bins=30)

log_budget_histogram <- df %>%
  ggplot(aes(log(budget))) +
  geom_histogram(bins=30)

grid.arrange(world_revenue_histogram, log_world_revenue_histogram,
             budget_histogram, log_budget_histogram,
             nrow = 2, ncol = 2)
```



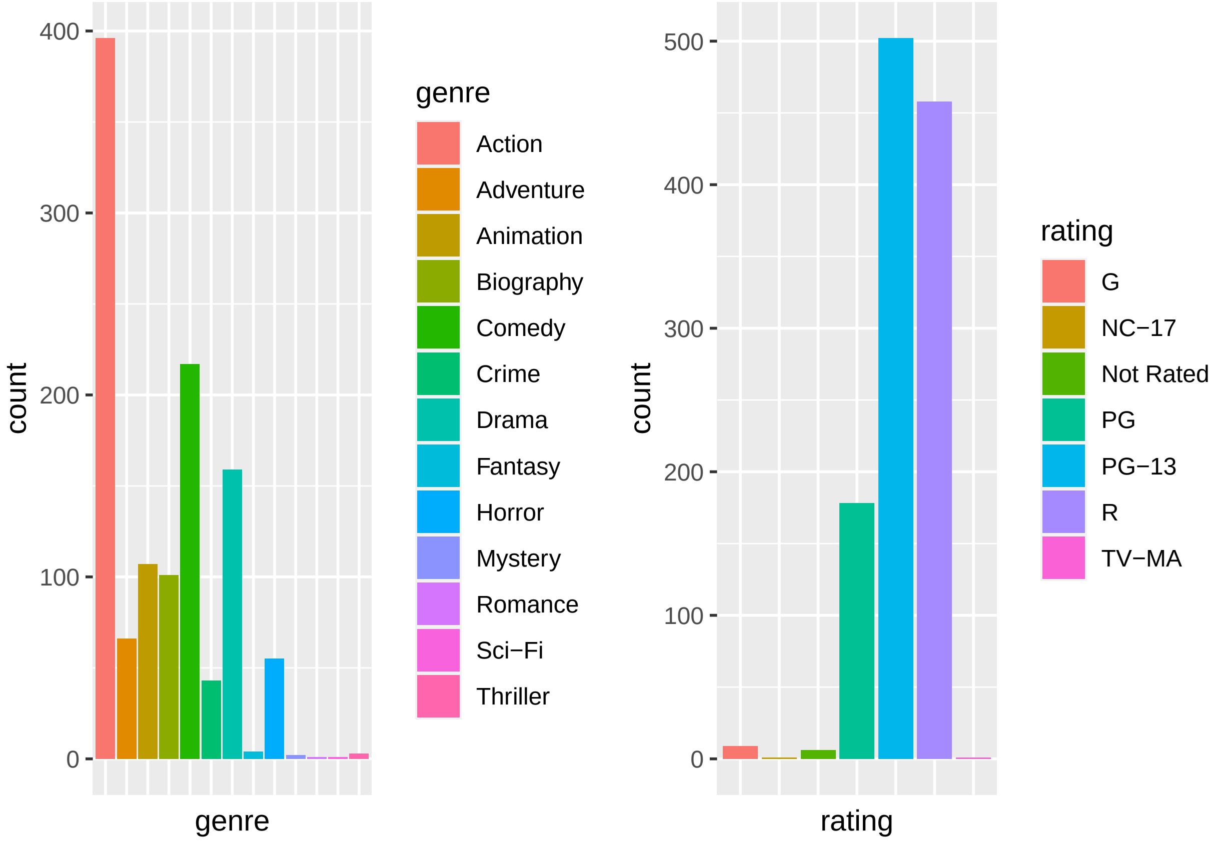
```
# CHECK NUMERIC EXPLANATORY VARIABLES
score_histogram <- df %>%
  ggplot(aes(score)) +
  geom_histogram(bins=30)

log_score_histogram <- df %>%
  ggplot(aes(log(score))) +
  geom_histogram(bins=30)

year_histogram <- df %>%
  ggplot(aes(year)) +
  geom_histogram(bins=10)

votes_histogram <- df %>%
  ggplot(aes(votes)) +
  geom_histogram(bins=30)
```

```r
log_votes_histogram <- df %>%
  ggplot(aes(log(votes))) +
  geom_histogram(bins=30)

runtime_histogram <- df %>%
  ggplot(aes(runtime)) +
  geom_histogram(bins=30)

country_plot <- df %>%
  ggplot(aes(x=rating, fill=country)) +
  geom_bar() +
  theme(axis.text.x=element_blank(),
        axis.ticks.x=element_blank())

grid.arrange(score_histogram, log_score_histogram,
             votes_histogram, log_votes_histogram,
             runtime_histogram, year_histogram,
             nrow = 3, ncol = 2)
```

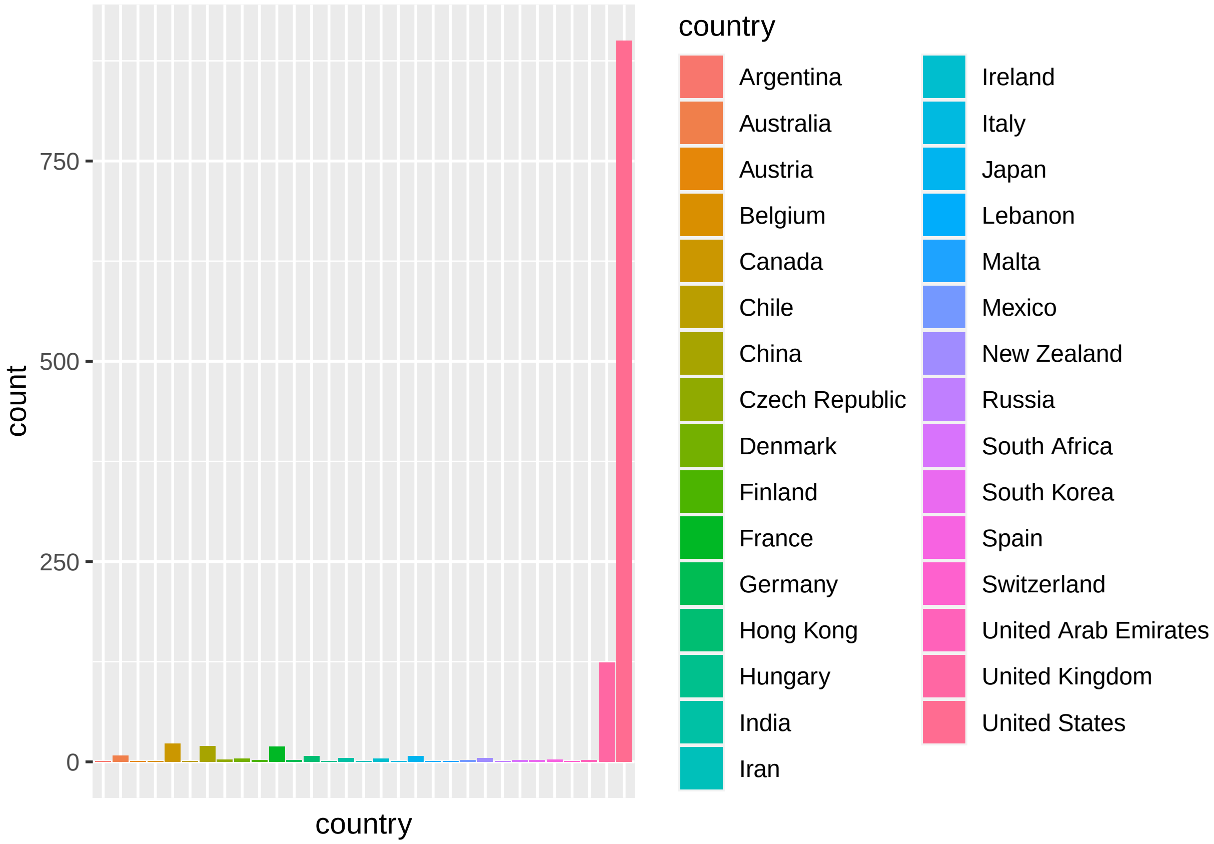


```r
# CHECK ORDINAL VARIABLES

rating_plot <- df %>%
  ggplot(aes(x=rating, fill=rating)) +
  geom_bar() +
  theme(axis.text.x=element_blank(),
        axis.ticks.x=element_blank())
```

```
genre_plot <- df %>%
  ggplot(aes(x=genre, fill=genre)) +
  geom_bar() +
  theme(axis.text.x=element_blank(),
        axis.ticks.x=element_blank())

grid.arrange(genre_plot, rating_plot,
             nrow = 1, ncol = 2)
```

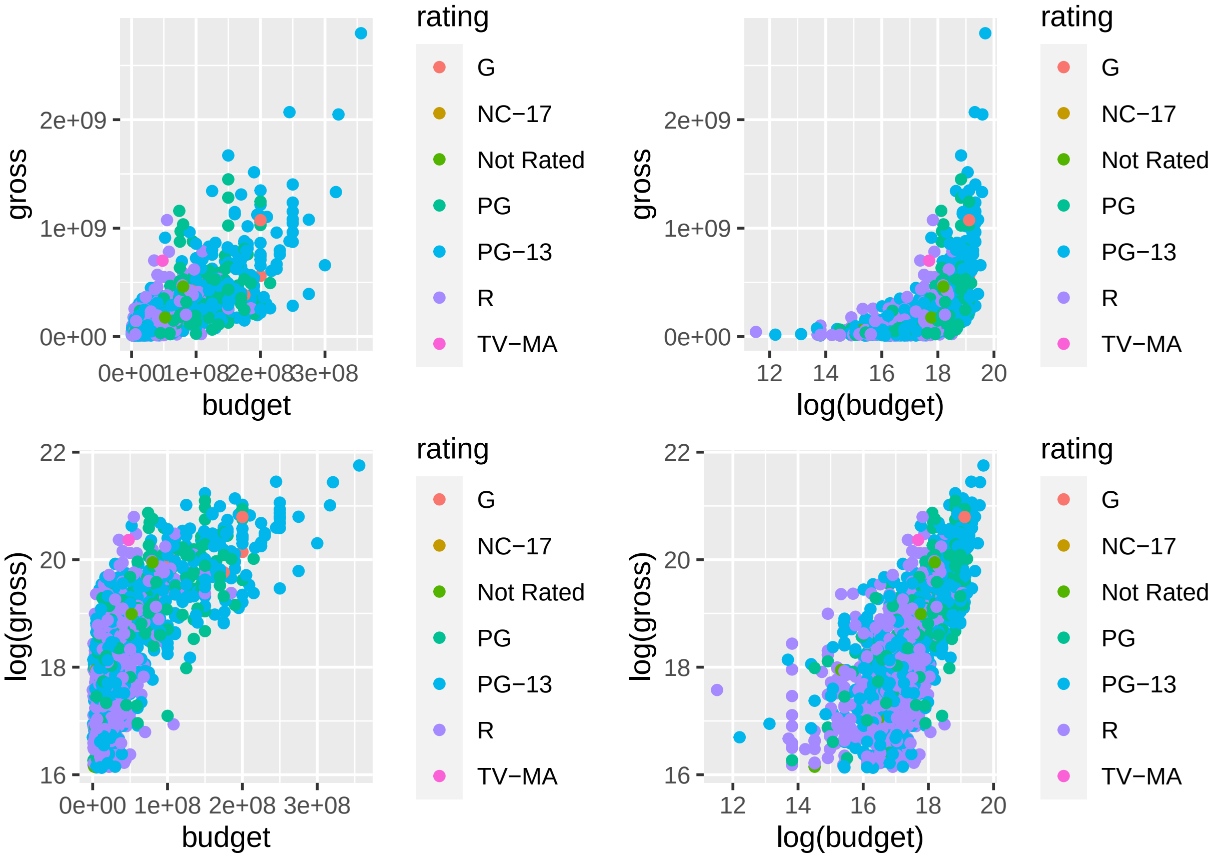


```
df %>%
  ggplot(aes(x=country, fill=country)) +
  geom_bar() +
  theme(axis.text.x=element_blank(),
        axis.ticks.x=element_blank())
```

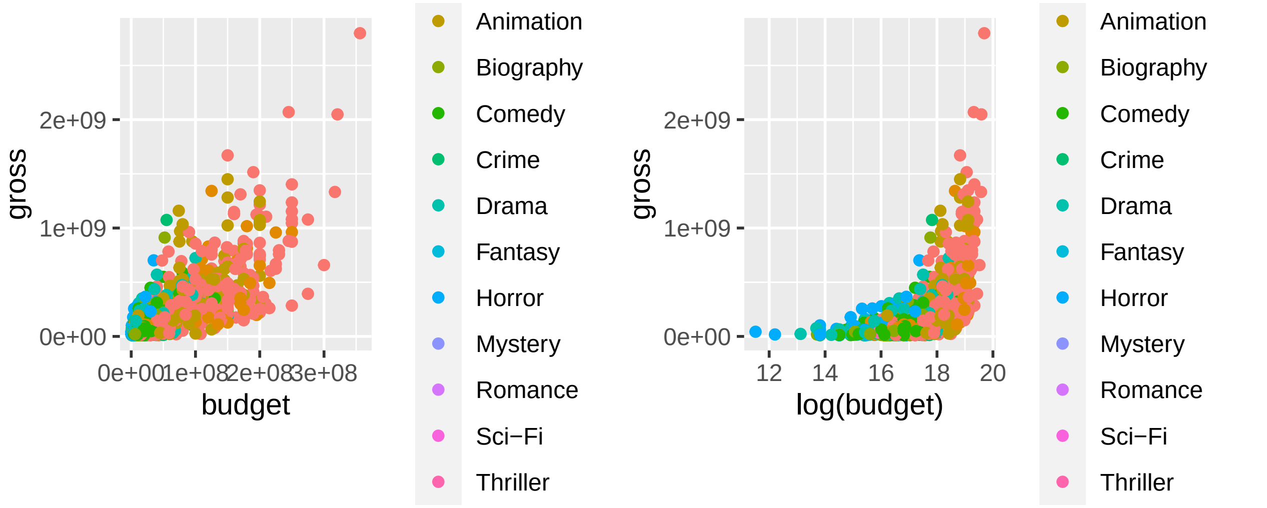


5. Compare plot of World Revenue to Budget (color by MPAA rating)

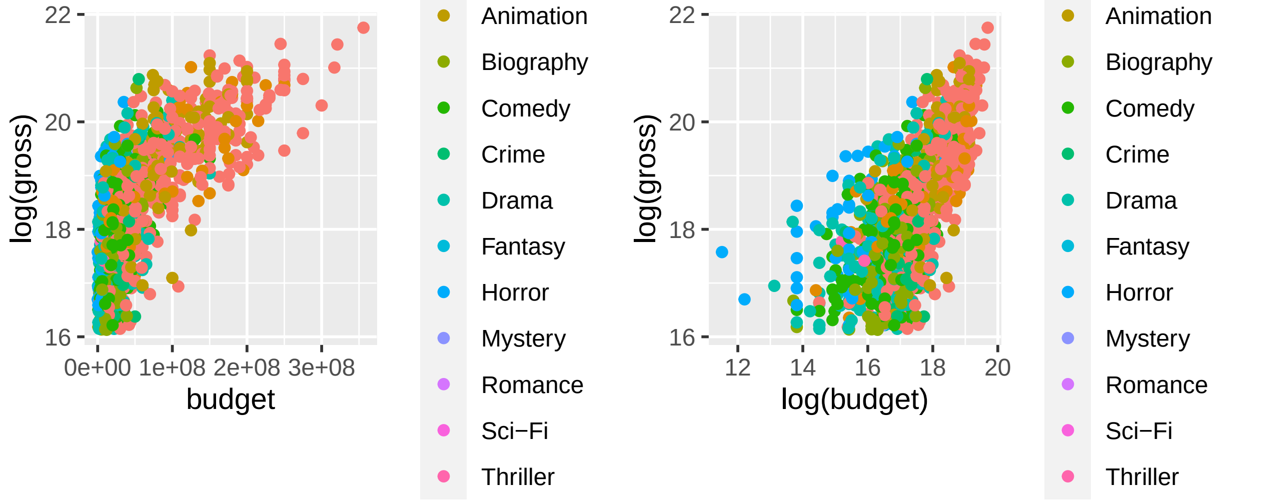
```
# COMPARE BUDGET TO GROSS COMPARE BY RATING
level_level_plot <- df %>%
  ggplot(aes(x=budget, y=gross, color=rating)) +
  geom_point()

level_log_plot <- df %>%
  ggplot(aes(x=log(budget), y=gross, color=rating)) +
  geom_point()

log_level_plot <- df %>%
  ggplot(aes(x=budget, y=log(gross), color=rating)) +
  geom_point()

log_log_plot <- df %>%
  ggplot(aes(x=log(budget), y=log(gross), color=rating)) +
  geom_point()

grid.arrange(level_level_plot, level_log_plot,
             log_level_plot, log_log_plot,
             nrow = 2, ncol = 2)
```



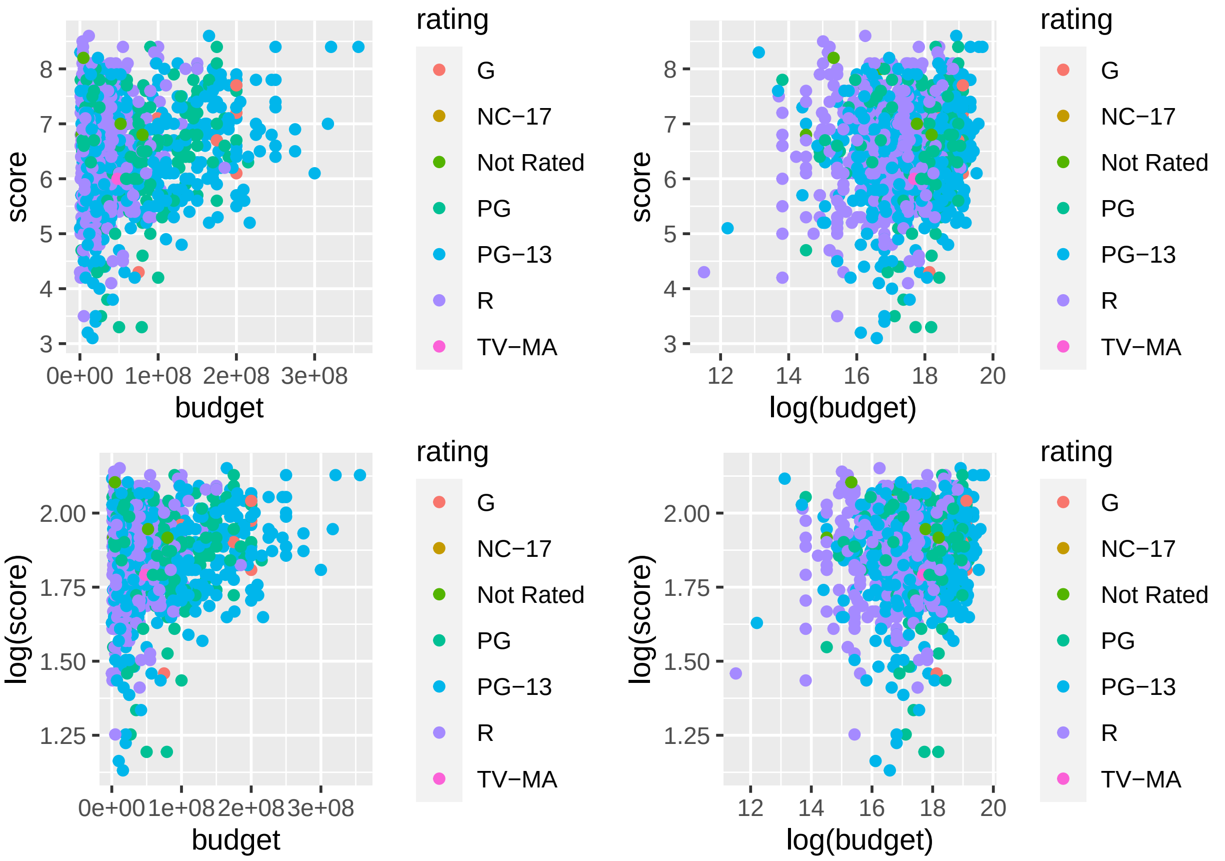
```
# COMPARE BUDGET TO GROSS COMPARE BY GENRE
level_level_plot <- df %>%
  ggplot(aes(x=budget, y=gross, color=genre)) +
  geom_point()

level_log_plot <- df %>%
  ggplot(aes(x=log(budget), y=gross, color=genre)) +
  geom_point()

log_level_plot <- df %>%
  ggplot(aes(x=budget, y=log(gross), color=genre)) +
  geom_point()

log_log_plot <- df %>%
  ggplot(aes(x=log(budget), y=log(gross), color=genre)) +
  geom_point()

grid.arrange(level_level_plot, level_log_plot,
             nrow = 2, ncol = 2)
```



```
grid.arrange(log_level_plot, log_log_plot,
             nrow = 2, ncol = 2)
```

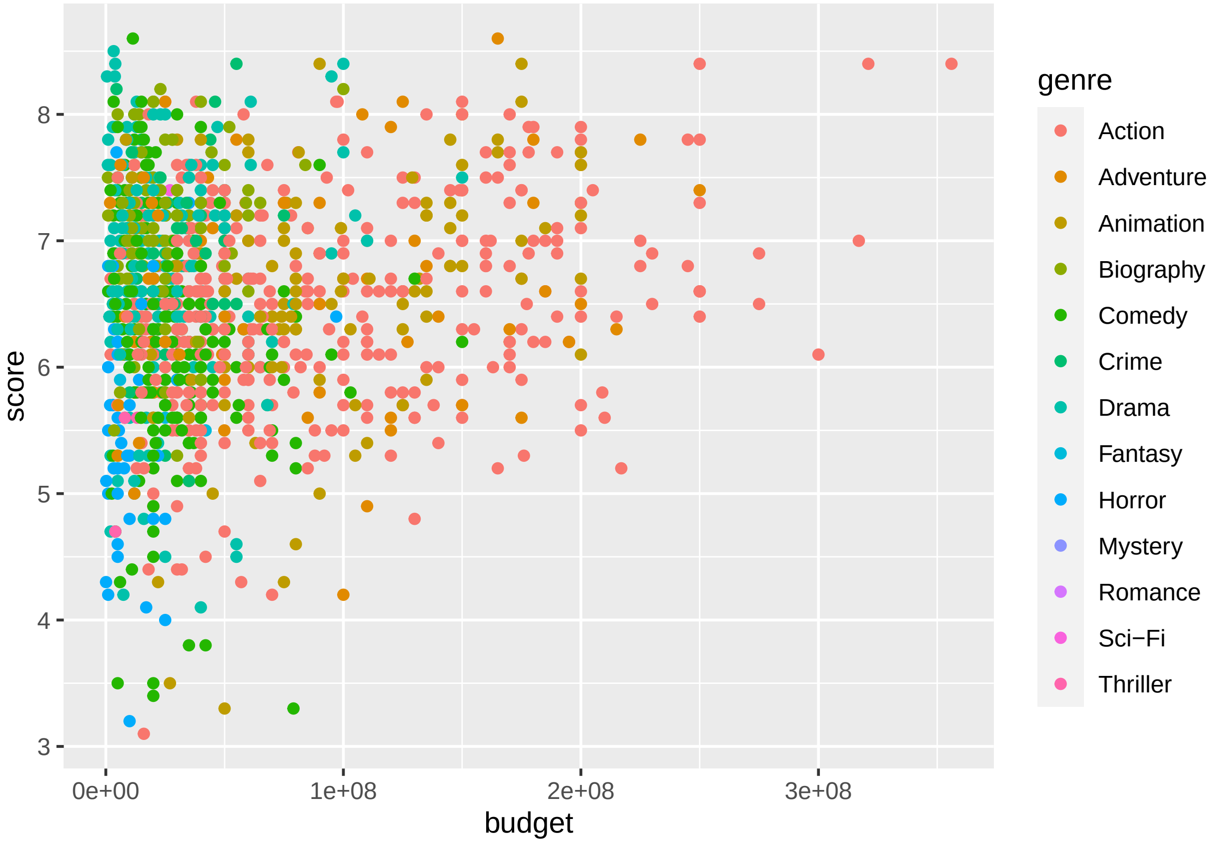


```
# COMPARE BUDGET TO SCORE COMPARE BY RATING
level_level_plot <- df %>%
  ggplot(aes(x=budget, y=score, color=rating)) +
  geom_point()

level_log_plot <- df %>%
  ggplot(aes(x=log(budget), y=score, color=rating)) +
  geom_point()

log_level_plot <- df %>%
  ggplot(aes(x=budget, y=log(score), color=rating)) +
  geom_point()

log_log_plot <- df %>%
  ggplot(aes(x=log(budget), y=log(score), color=rating)) +
  geom_point()

grid.arrange(level_level_plot, level_log_plot,
             log_level_plot, log_log_plot,
             nrow = 2, ncol = 2)
```

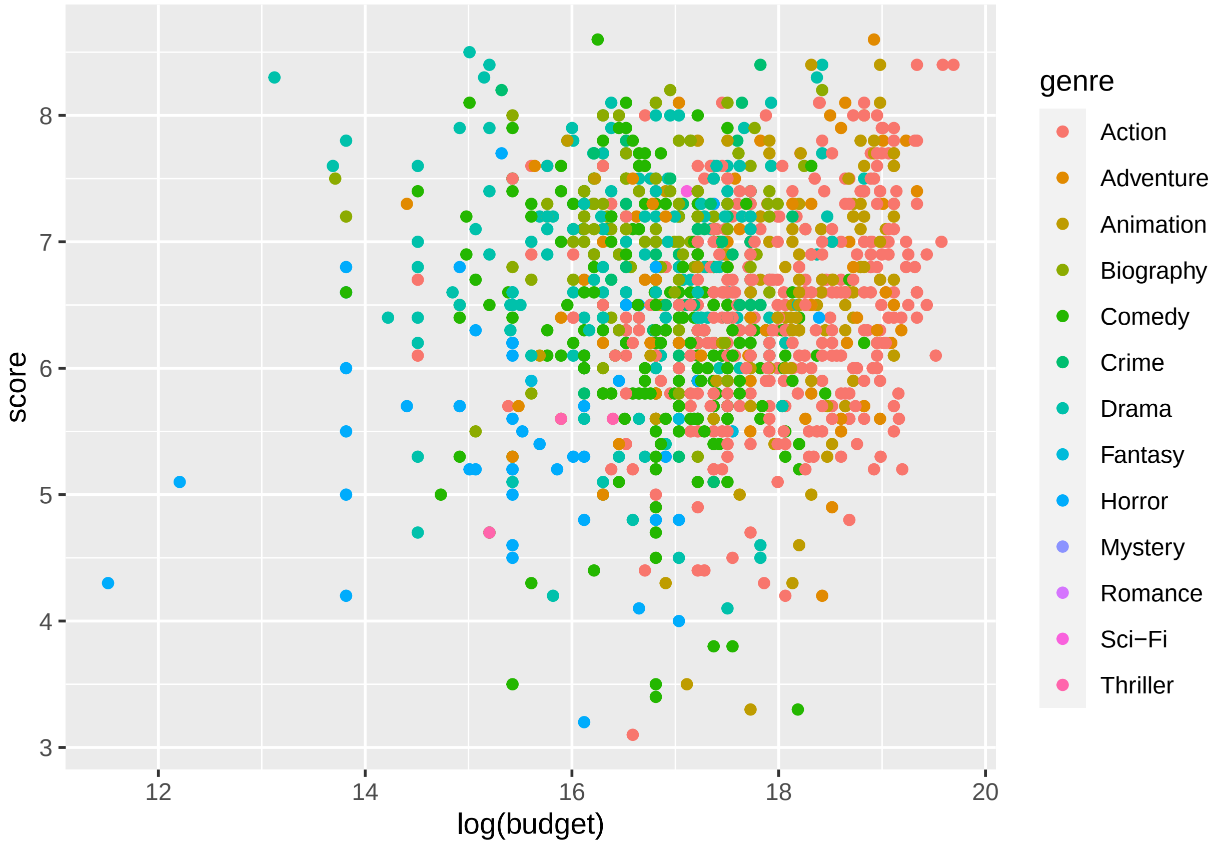


```
# COMPARE BUDGET TO SCORE COMPARE BY GENRE
level_level_plot <- df %>%
  ggplot(aes(x=budget, y=score, color=genre)) +
  geom_point()

level_log_plot <- df %>%
  ggplot(aes(x=log(budget), y=score, color=genre)) +
  geom_point()

log_level_plot <- df %>%
  ggplot(aes(x=budget, y=log(score), color=genre)) +
  geom_point()

log_log_plot <- df %>%
  ggplot(aes(x=log(budget), y=log(score), color=rating)) +
  geom_point()

#grid.arrange(level_level_plot, level_log_plot,
             #nrow = 1, ncol = 2)

#grid.arrange(log_level_plot, log_log_plot,
             #nrow = 1, ncol = 2)
```

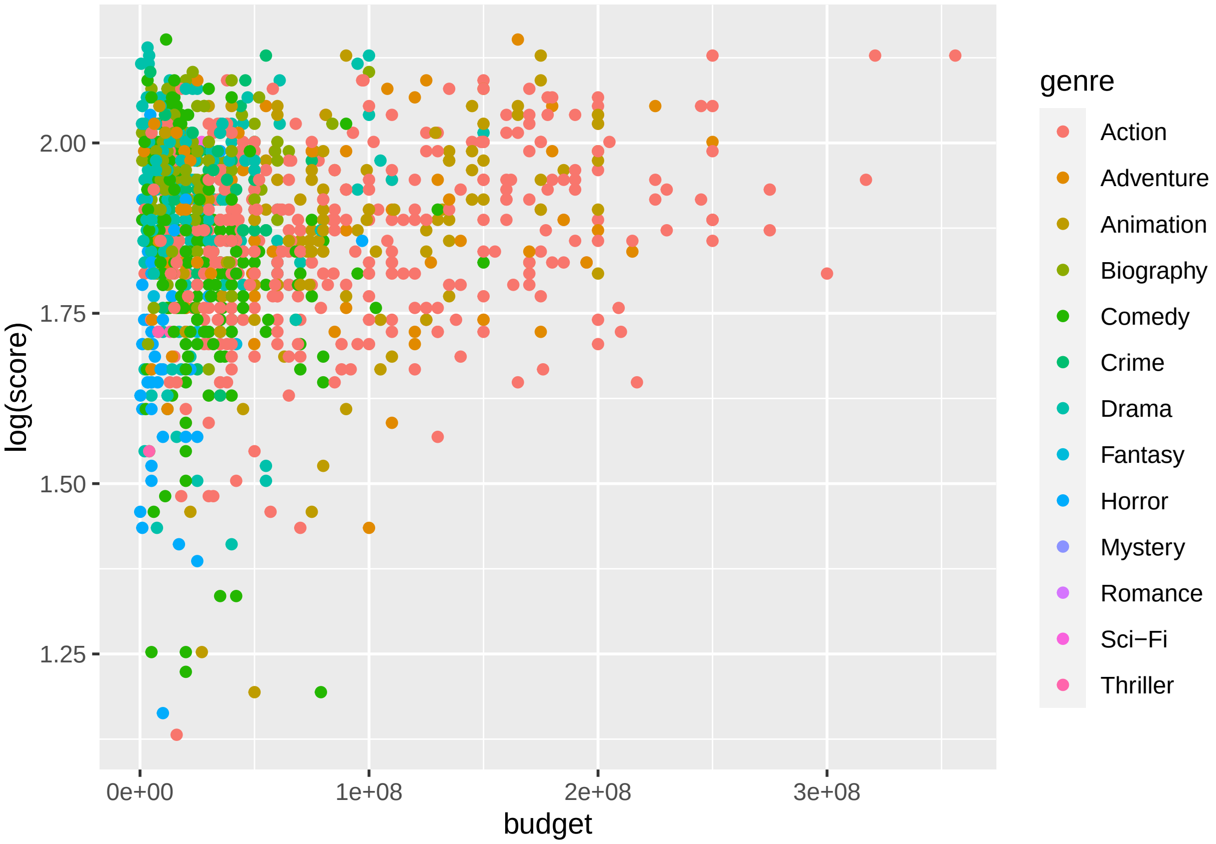
```
level_level_plot
```



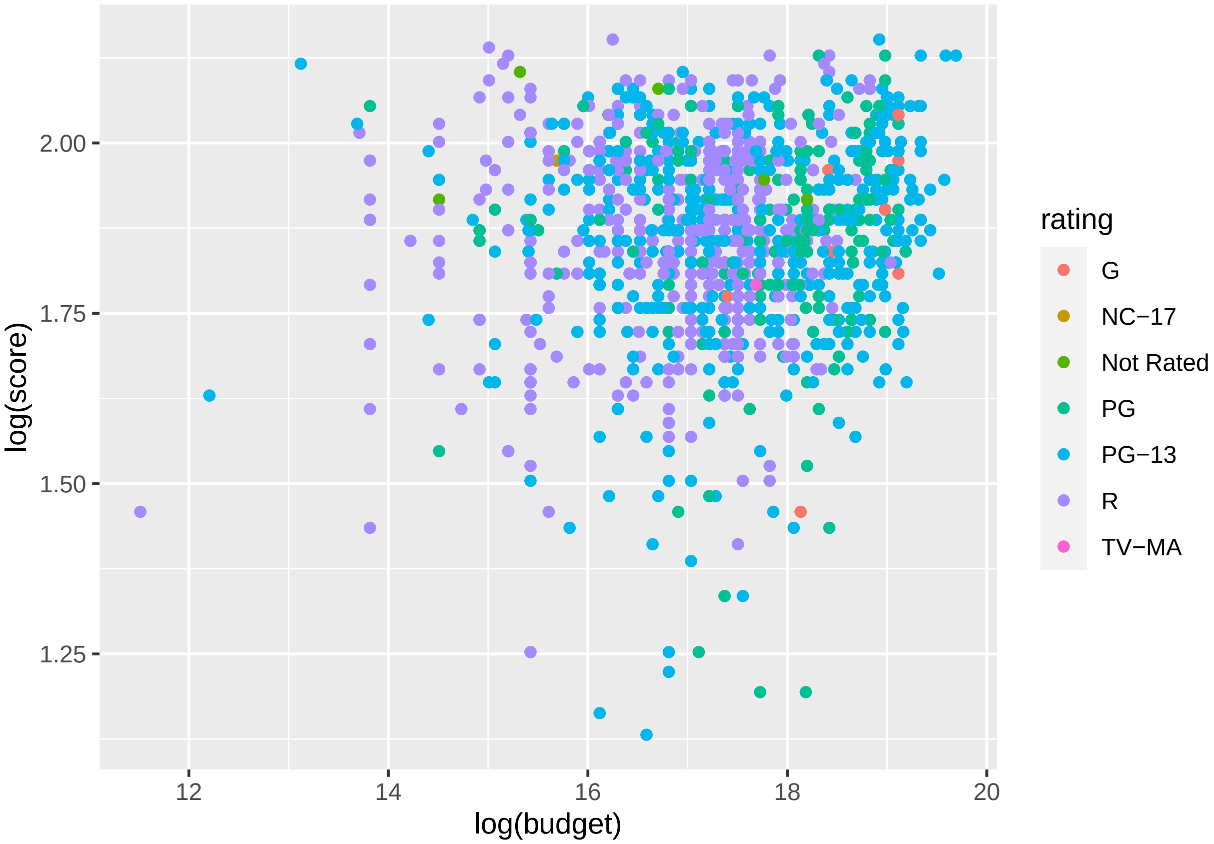
```
level_log_plot
```



```
log_level_plot
```


log(score)
2.00
1.75
1.50
1.25
0e+00
1e+08
2e+08
3e+08
budget
genre
Action
Adventure
Animation
Biography
Comedy
Crime
Drama
Fantasy
Horror
Mystery
Romance
Sci−Fi
Thriller

```
log_log_plot
```


log(score)
2.00
1.75
1.50
1.25
log(budget)
12
14
16
18
20
rating
G
NC−17
Not Rated
PG
PG−13
R
TV−MA