# Lab 2 EDA

Austin Sanders

11/14/2021

**Steps:**

1. Load The Data
2. Get Necessary Columns
3. Filter out Low-Budget, Low-Revenue, and Duplicate Titles - 285-295 rows
4. Examine Variables (apply transformations) - log-transformation makes them look normal
5. Compare plot of World Revenue to Budget (color by MPAA rating) - looks like there is a linear trend

## 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))
```

```r
# Remove N/A ratings if desired
if (REMOVE_NA_RATING) {
  df_raw <- subset(df_raw, df_raw$rating != "N/A")
}

df = df_raw
```

## 4. Examine Variables (apply transformations)

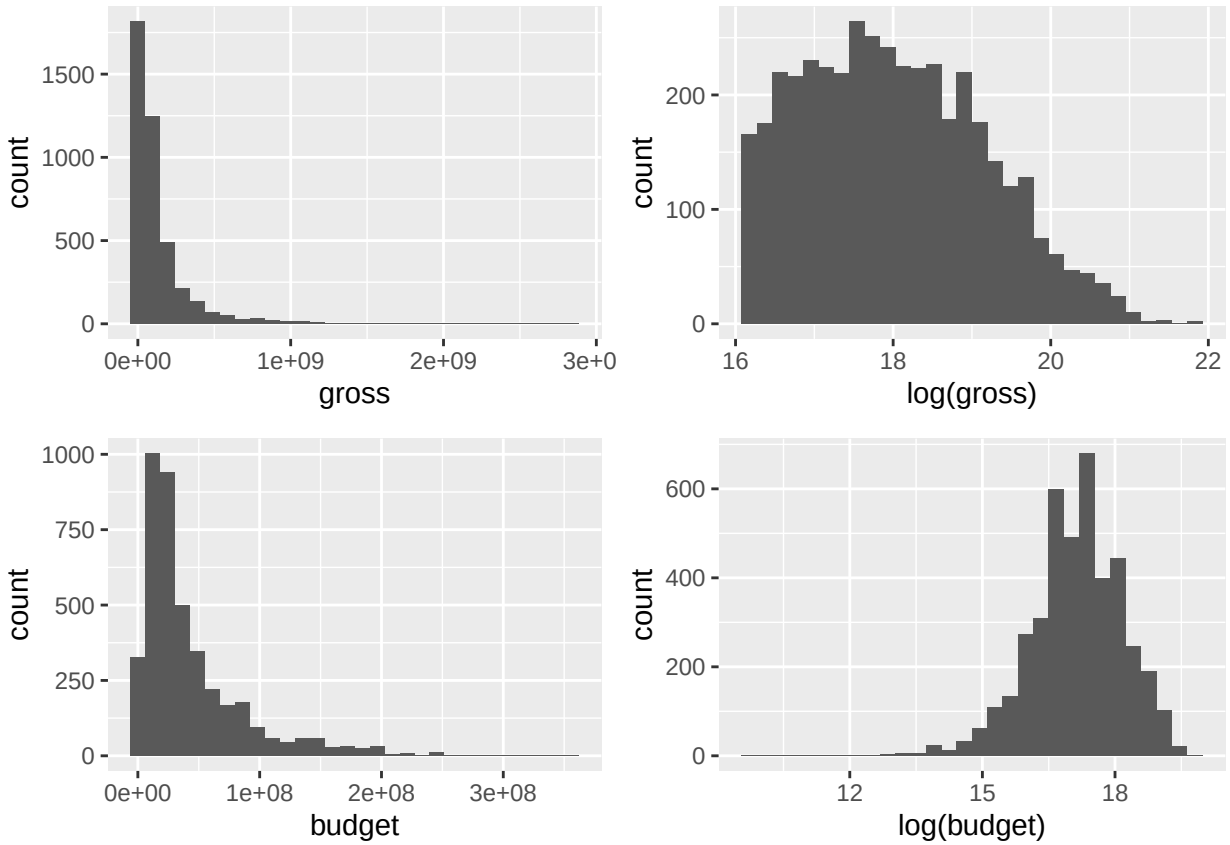
```r
# CHECK MAIN NUMERIC VARIABLES
world_revenue_histogram <- df %>%
  ggplot(aes(gross)) +
  geom_histogram(bins=30)

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)
```



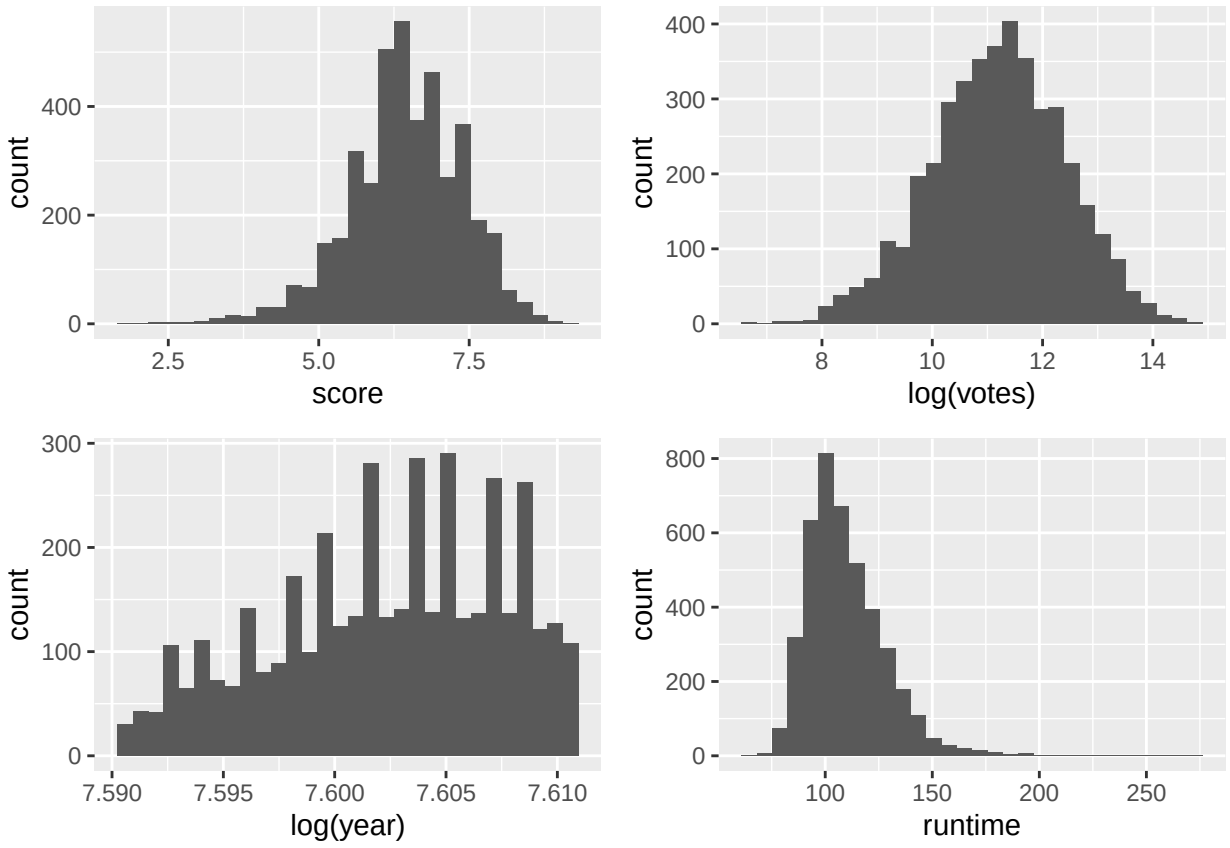

```
# CHECK NUMERIC EXPLANATORY VARIABLE
score_histogram <- df %>%
  ggplot(aes(score)) +
  geom_histogram(bins=30)

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

year_histogram <- df %>%
  ggplot(aes(log(year))) +
  geom_histogram(bins=30)

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

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




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

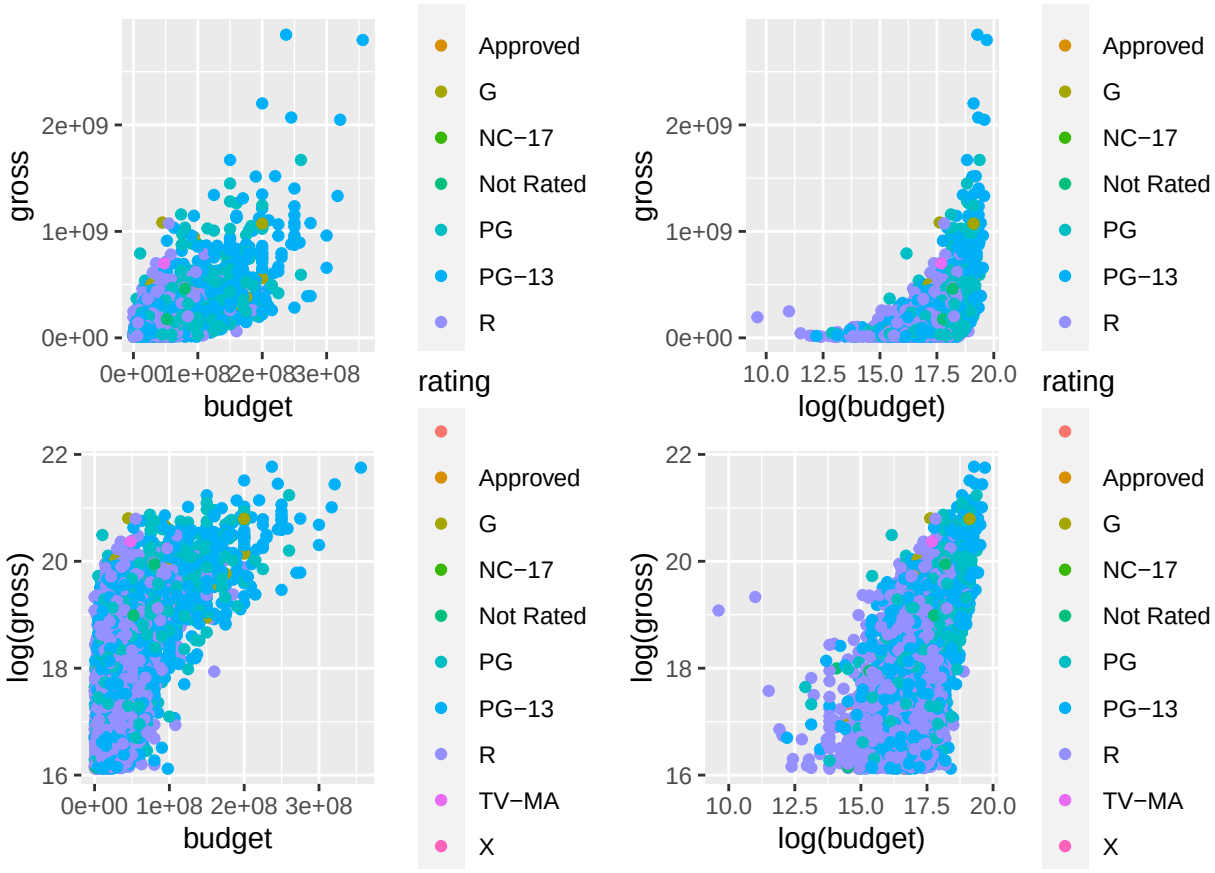
```
# COMPARE EXPLANATORY VARIABLES
level_level_plot <- df %>%
  ggplot(aes(x=budget, y=gross, color=rating)) +
  geom_point()

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)
```


gross
2e+09
1e+09
0e+00
0e+00 1e+08 2e+08 3e+08
budget
log(gross)
22
20
18
16
log(budget)
10.0 12.5 15.0 17.5 20.0
rating
Approved
G
NC−17
Not Rated
PG
PG−13
R
TV−MA
X