EDA

Katrina Truebebach

March 16, 2019

```
rm(list = ls())
```

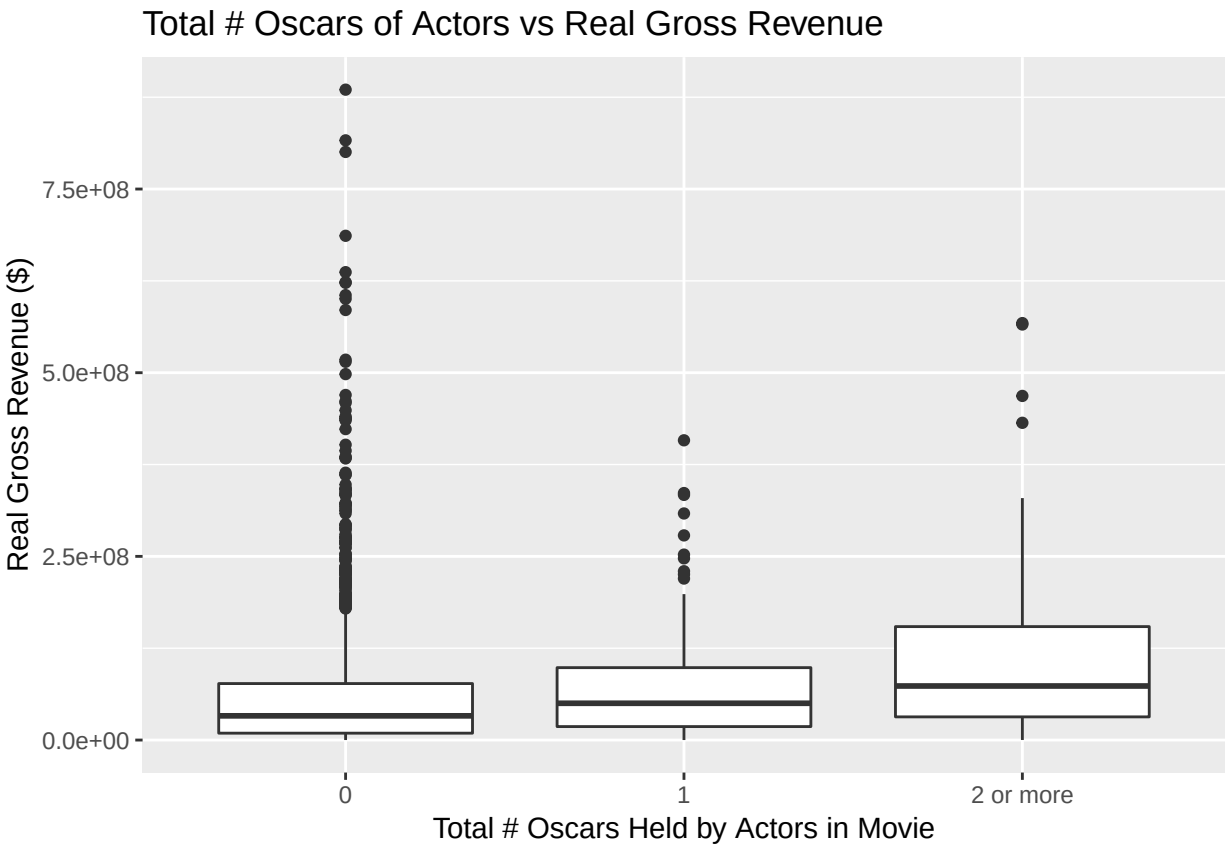
Load cleaned data

```
load(file = '~/DS5110/data/proj_cleaned_dta.RData')
```

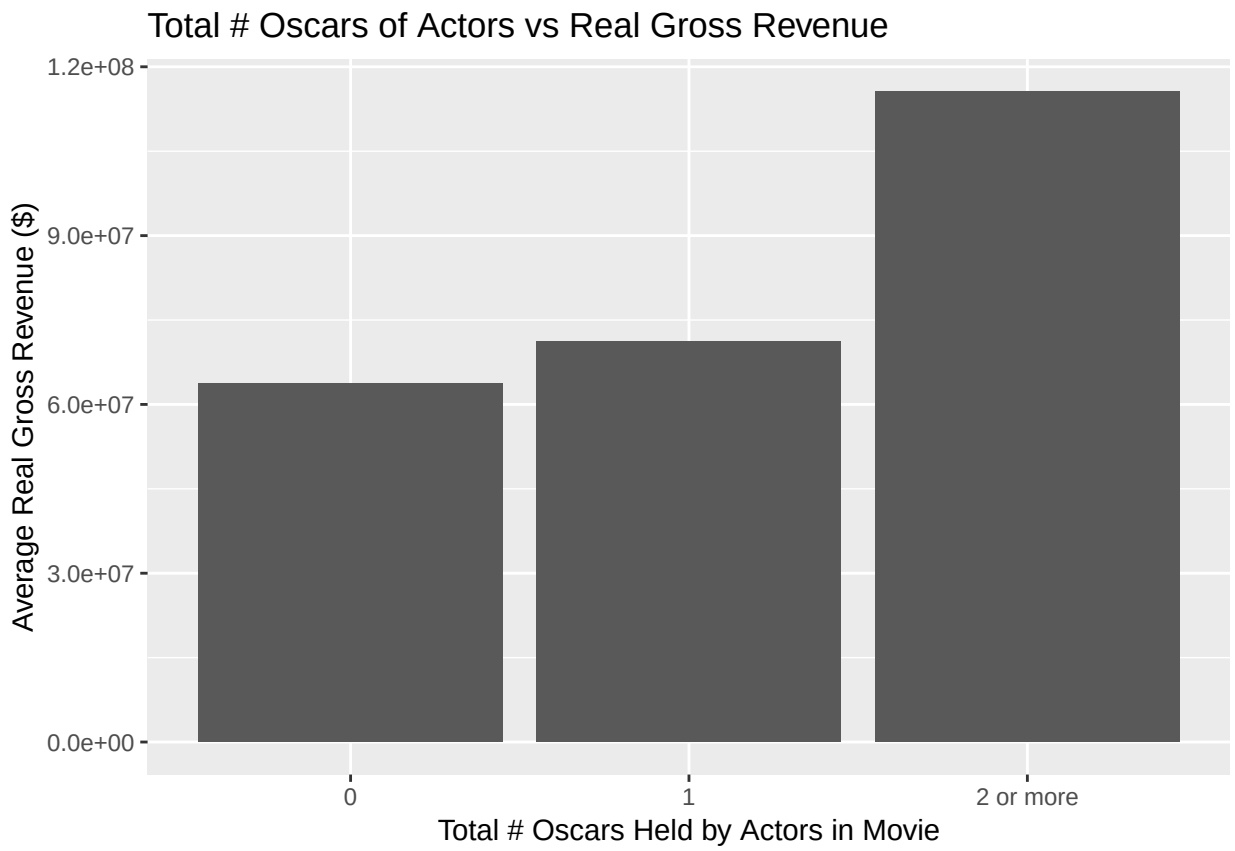
Oscars

Graph number of Oscars for actors and directors against real revenue. Boxplot and bar plot (average revenue)
Both are linear but very weak. Unclear if should include in model

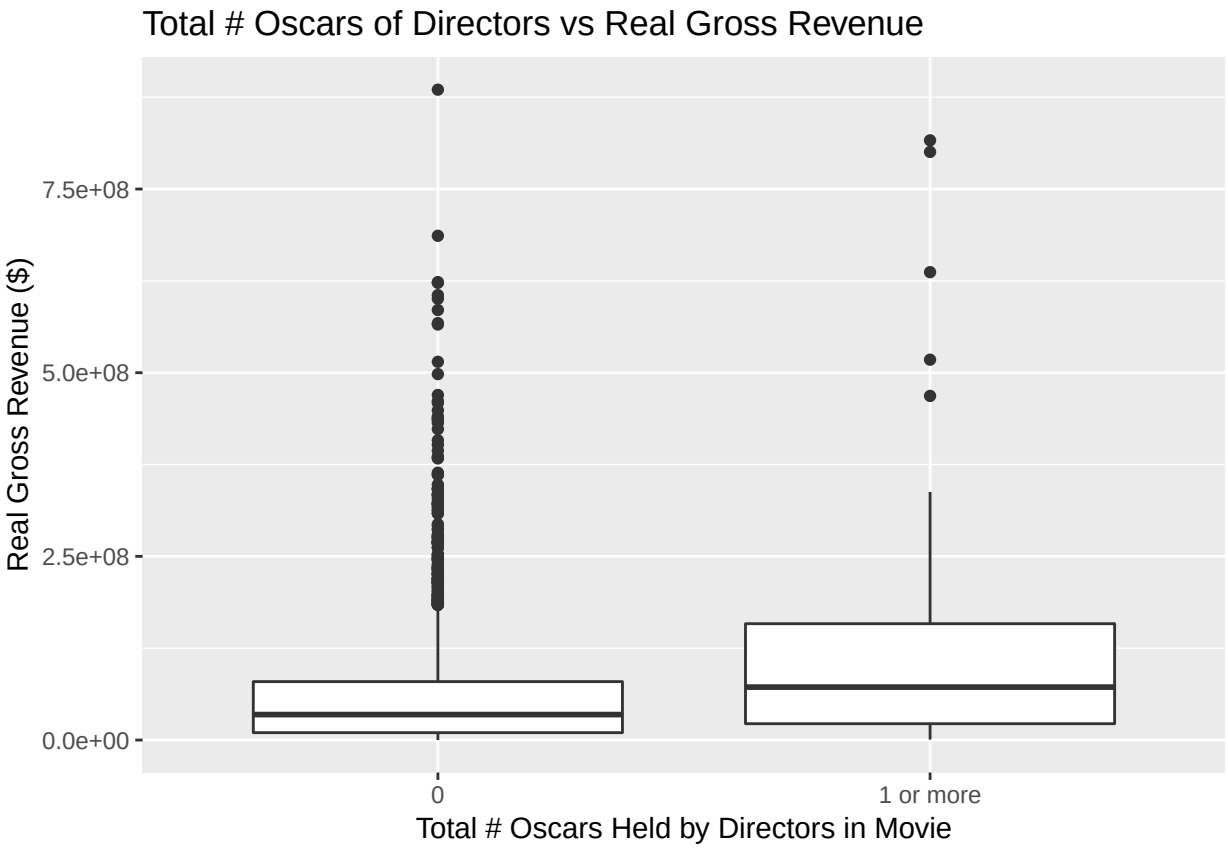
```
# Versions of data with average revenue by number of oscars
train_oscar_actor <- train %>%
  group_by(total_oscars_actor) %>%
  summarize(avg_real_gross = mean(real_gross))
train_oscar_director <- train %>%
  group_by(total_oscars_director) %>%
  summarize(avg_real_gross = mean(real_gross))

# Functions to graph number of Oscars held by actors in movie vs. real revenue
# boxplot
oscar_box <- function(df, var, title_str, x_str) {
  ggplot(df, aes_string(var, "real_gross")) +
    geom_boxplot() +
    labs(title = title_str, x = x_str, y = 'Real Gross Revenue ($)')
}

# bar graph
oscar_bar <- function(df, var, title_str, x_str) {
  ggplot(df, aes_string(var, "avg_real_gross")) +
    geom_col() +
    labs(title = title_str, x = x_str, y = 'Average Real Gross Revenue ($)')
}

# actors
oscar_box(train, 'total_oscars_actor', 'Total # Oscars of Actors vs Real Gross Revenue',
          'Total # Oscars Held by Actors in Movie')
```



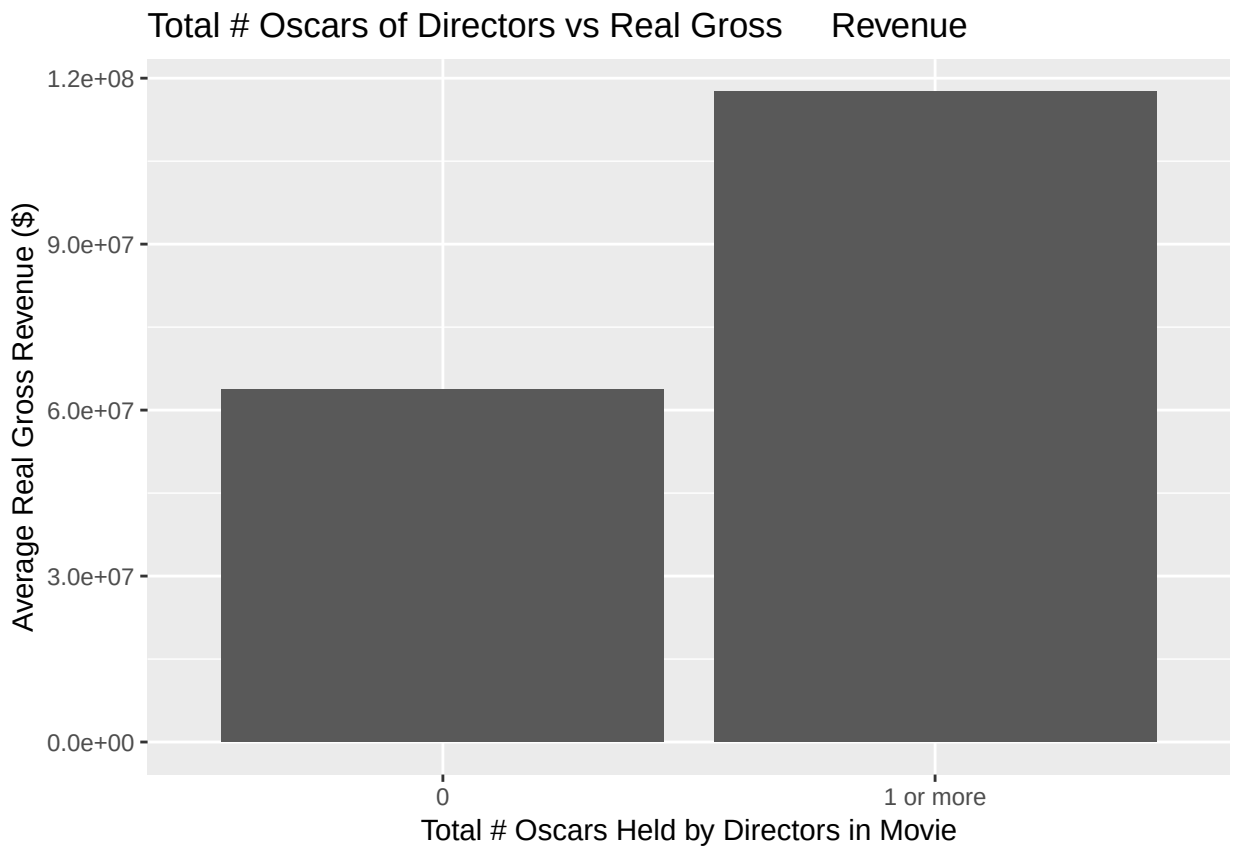

```
oscar_bar(train_oscar_actor, 'total_oscars_actor', 'Total # Oscars of Actors vs Real Gross Revenue',
          'Total # Oscars Held by Actors in Movie')
```




```
# directors
oscar_box(train, 'total_oscars_director', 'Total # Oscars of Directors vs Real Gross Revenue',
          'Total # Oscars Held by Directors in Movie')
```



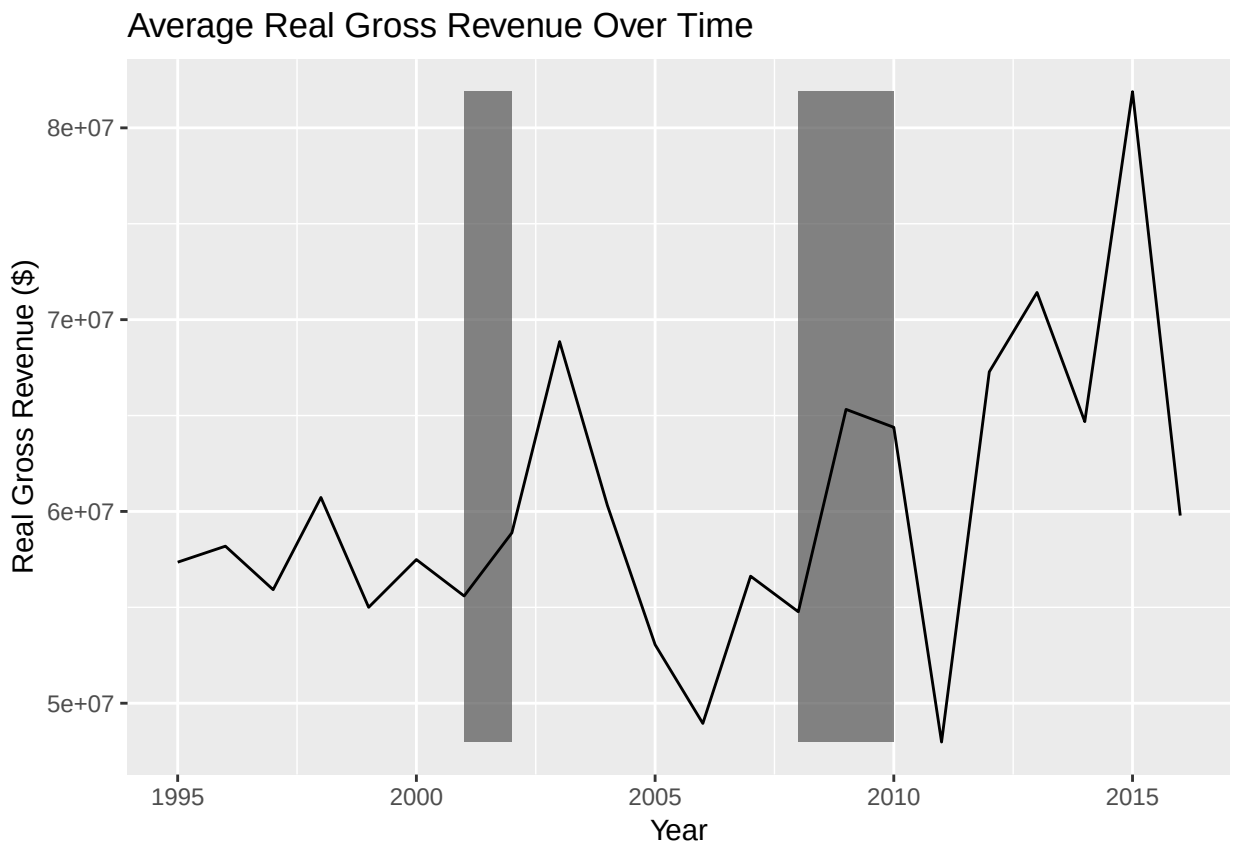

```
oscar_bar(train_oscar_director, 'total_oscars_director', 'Total # Oscars of Directors vs Real Gross
```



Year

Average real revenue vs year
Adeed APPROXIMATE recession shading. Annual data, so hard to do.
Real revenue increase during recessions (have seen this before with Great Depression - numerous articles we can reference)
Regardless, clear that year could have an effect

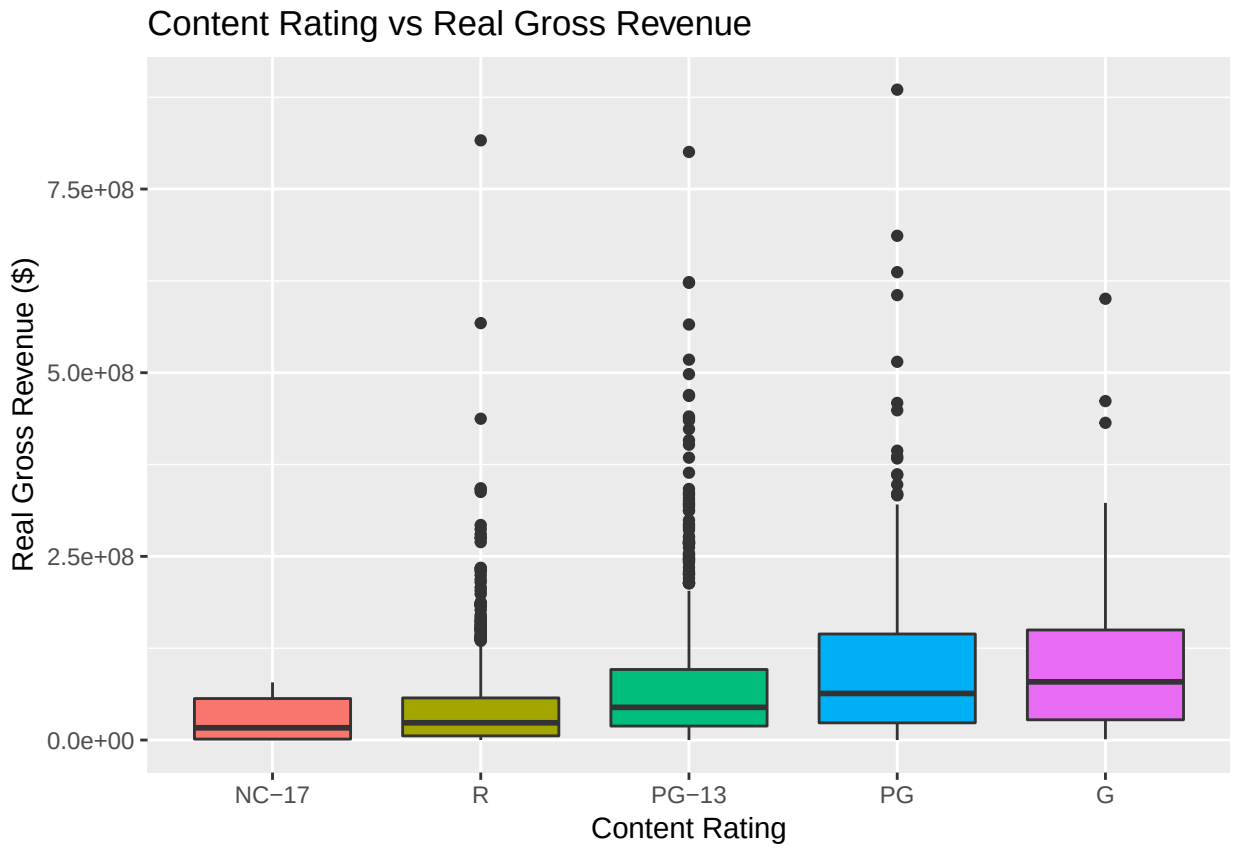
```
# take average of revenue per year
train_sum <- train %>%
  group_by(year) %>%
  summarize(real_gross_avg = mean(real_gross))

# need to limit because before 1995 there are very few observations per year (< 10 usually).
# this causes large spikes because one high earning or low earning movie influences the average heavily
# Starting at 1995, where have at least 30 (or very close) movies per year. Now can see movements over

ggplot(data = train_sum %>% filter(year >= 1995)) +
  geom_rect(aes(xmin = 2008, xmax = 2010,
                ymin = min(real_gross_avg, na.rm = T),
                ymax = max(real_gross_avg, na.rm = T)), alpha = .05) +
  geom_rect(aes(xmin = 2001, xmax = 2002,
                ymin = min(real_gross_avg, na.rm = T),
                ymax = max(real_gross_avg, na.rm = T)), alpha = .05) +
  geom_line(aes(x = year, y = real_gross_avg)) +
  labs(title = 'Average Real Gross Revenue Over Time', x = 'Year', y = 'Real Gross Revenue ($)')
```



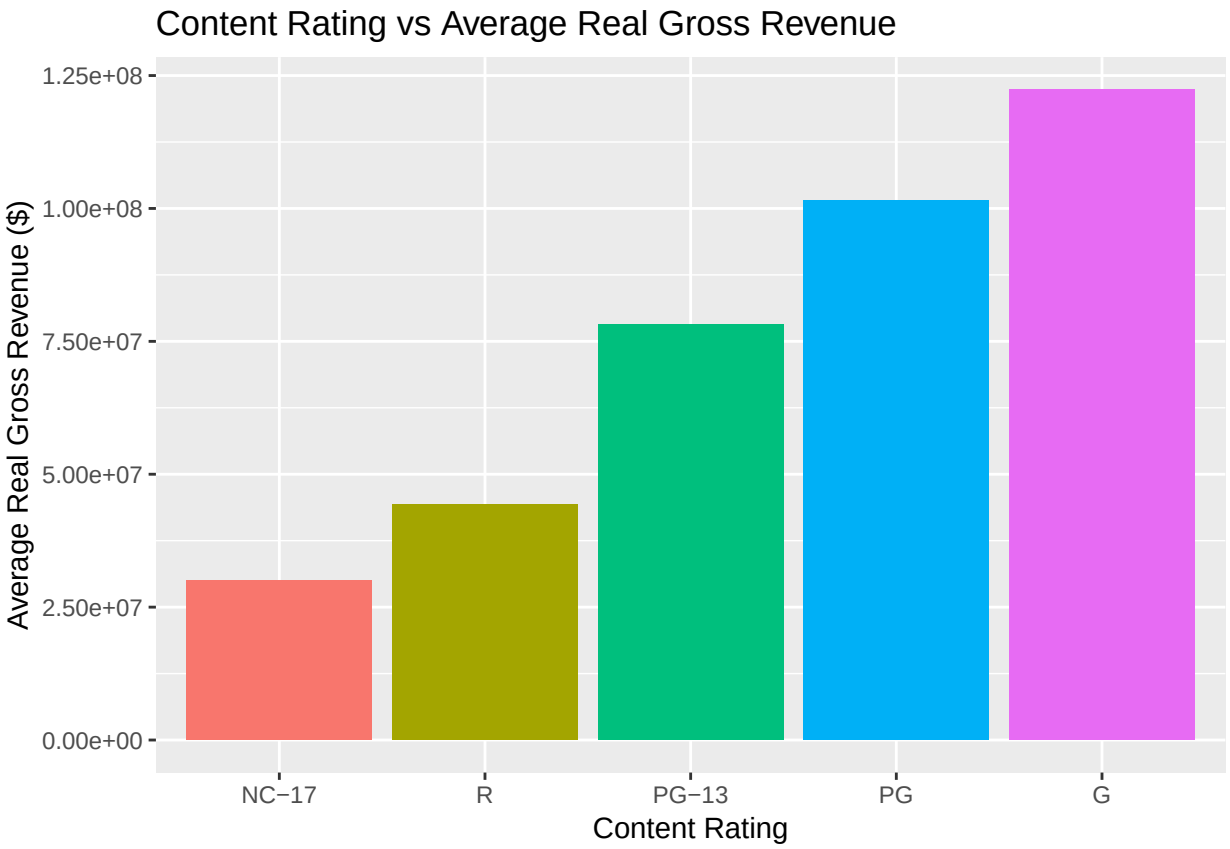
Content Rating

Bar graph of average real revenue and boxplot
Linear relationship. Good candidate to include in the model

```
# data manipulation. Factor.
train_content <- train %>%
  # filter out missing
  filter(!is.na(content_rating)) %>%
  # Make content rating a factor so can order in graphics
  mutate(content_rating = as.factor(content_rating),
         content_rating = reorder(content_rating, real_gross))

# boxplot
train_content  %>%
  ggplot() +
  geom_boxplot(aes(x = content_rating, y = real_gross, fill = content_rating)) +
  labs(x = 'Content Rating', y = 'Real Gross Revenue ($)',
       title = 'Content Rating vs Real Gross Revenue') +
  theme(legend.position = 'none')
```




```
# bar graph
train_content  %>%
  # average revenue by content rating
  group_by(content_rating) %>%
  summarize(avg_real_gross = mean(real_gross)) %>%
  ggplot() +
  geom_col(aes(x = content_rating, y = avg_real_gross, fill = content_rating)) +
  labs(x = 'Content Rating', y = 'Average Real Gross Revenue ($)',
       title = 'Content Rating vs Average Real Gross Revenue') +
  theme(legend.position = 'none')
```

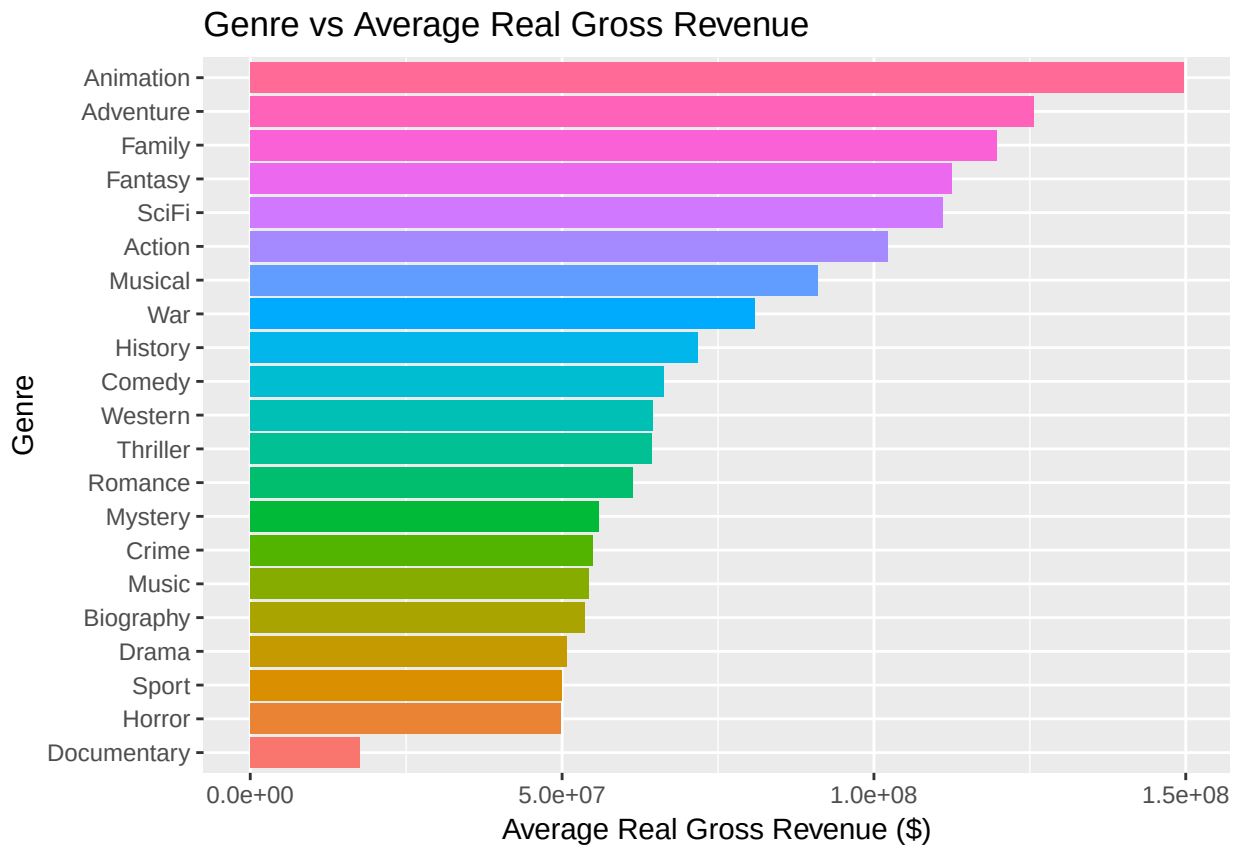


Genre

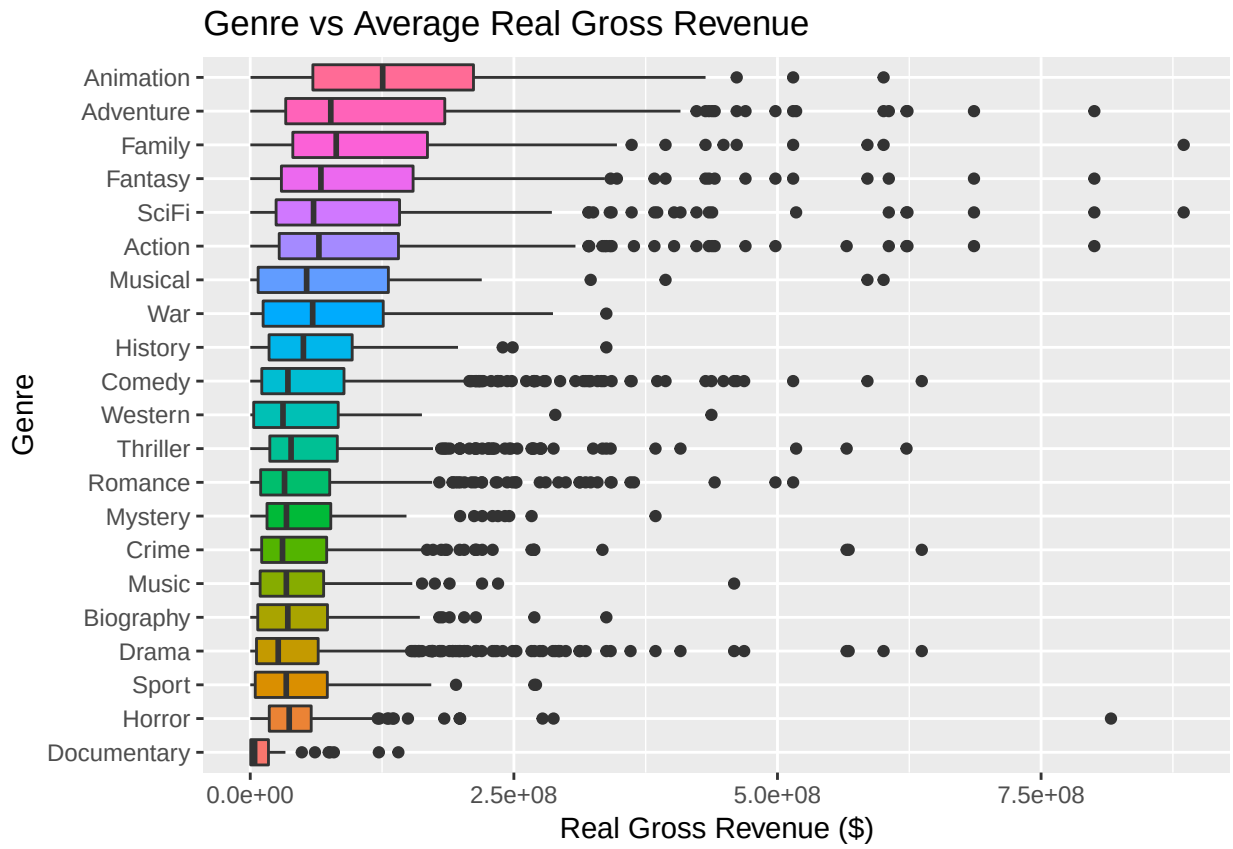
Bar graph of genre vs real revenue. Try boxplot and bar graph against average real revenue.
Fairly linear relationship. Good candidate to include in model.

```
# untidy the genre data such that one obseration is spread across many rows. Easier to graph
genre_cols <- c('Action', 'Adventure', 'Animation', 'Biography', 'Comedy', 'Crime', 'Documentary',
                'Drama', 'Family', 'Fantasy', 'History', 'Horror', 'Music', 'Musical', 'Mystery',
                'Romance', 'SciFi', 'Sport', 'Thriller', 'War', 'Western')
train_genre <- train %>%
  # gather: one row per genre-movie combo
  gather(genre_cols, key = genre, value = yes) %>%
  # only keep when 'yes' is 1 (yes it is of that genre) %>%
  filter(yes == 1) %>%
  # make genre a factor and order by real_gross
  mutate(genre = as.factor(genre), genre = reorder(genre, real_gross))

# bar graph
train_genre %>%
  # average by genre
  group_by(genre) %>%
  summarize(avg_real_gross = mean(real_gross)) %>%
  # graph
  ggplot() +
  geom_col(aes(x = genre, y = avg_real_gross, fill = genre)) +
  coord_flip() +
```

```
  labs(x = 'Genre', y = 'Average Real Gross Revenue ($)',
       title = 'Genre vs Average Real Gross Revenue') +
  theme(legend.position = 'none')
```




```
# boxplot
train_genre %>%
  ggplot() +
  geom_boxplot(aes(x = genre, y = real_gross, fill = genre)) +
  coord_flip() +
  labs(x = 'Genre', y = 'Real Gross Revenue ($)',
       title = 'Genre vs Average Real Gross Revenue') +
  theme(legend.position = 'none')
```


Genre vs Average Real Gross Revenue
Animation
Adventure
Family
Fantasy
SciFi
Action
Musical
War
History
Comedy
Western
Thriller
Romance
Mystery
Crime
Music
Biography
Drama
Sport
Horror
Documentary
Genre
0.0e+00
2.5e+08
5.0e+08
7.5e+08
Real Gross Revenue ($)