EDA

Katrina Truebebach

March 16, 2019

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

Load cleaned data

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

Log

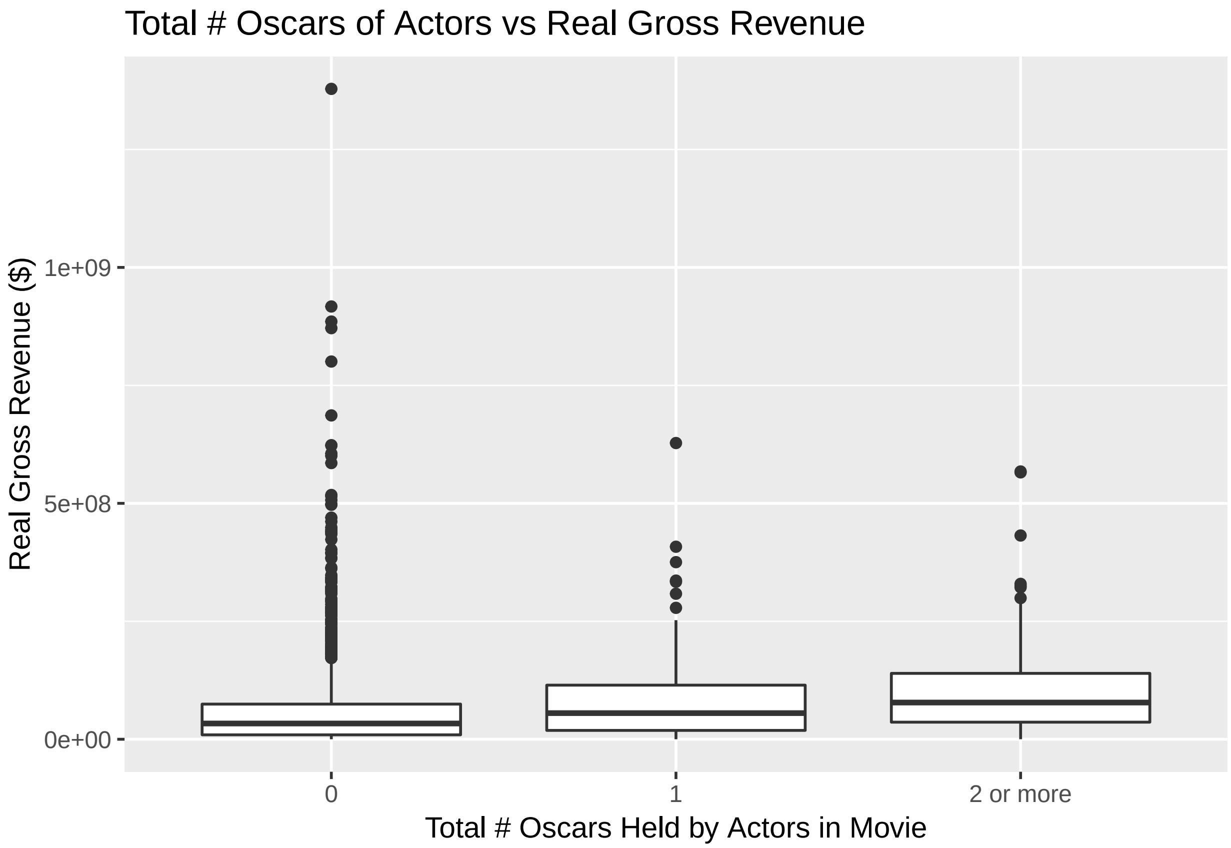
```r
train <- train %>%
  mutate_at(vars(real_gross, real_budget, director_facebook_likes, cast_total_facebook_likes, imdb_scor
```

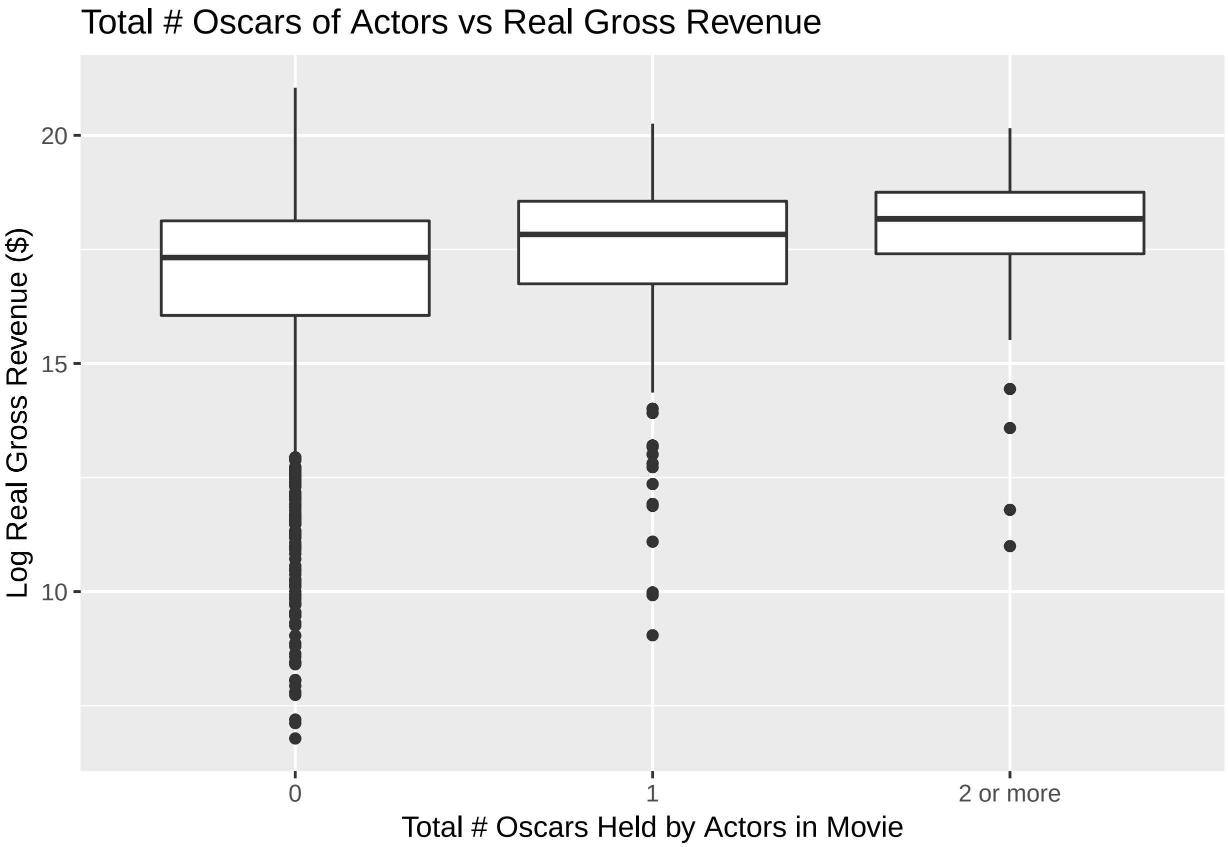
Oscars

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*

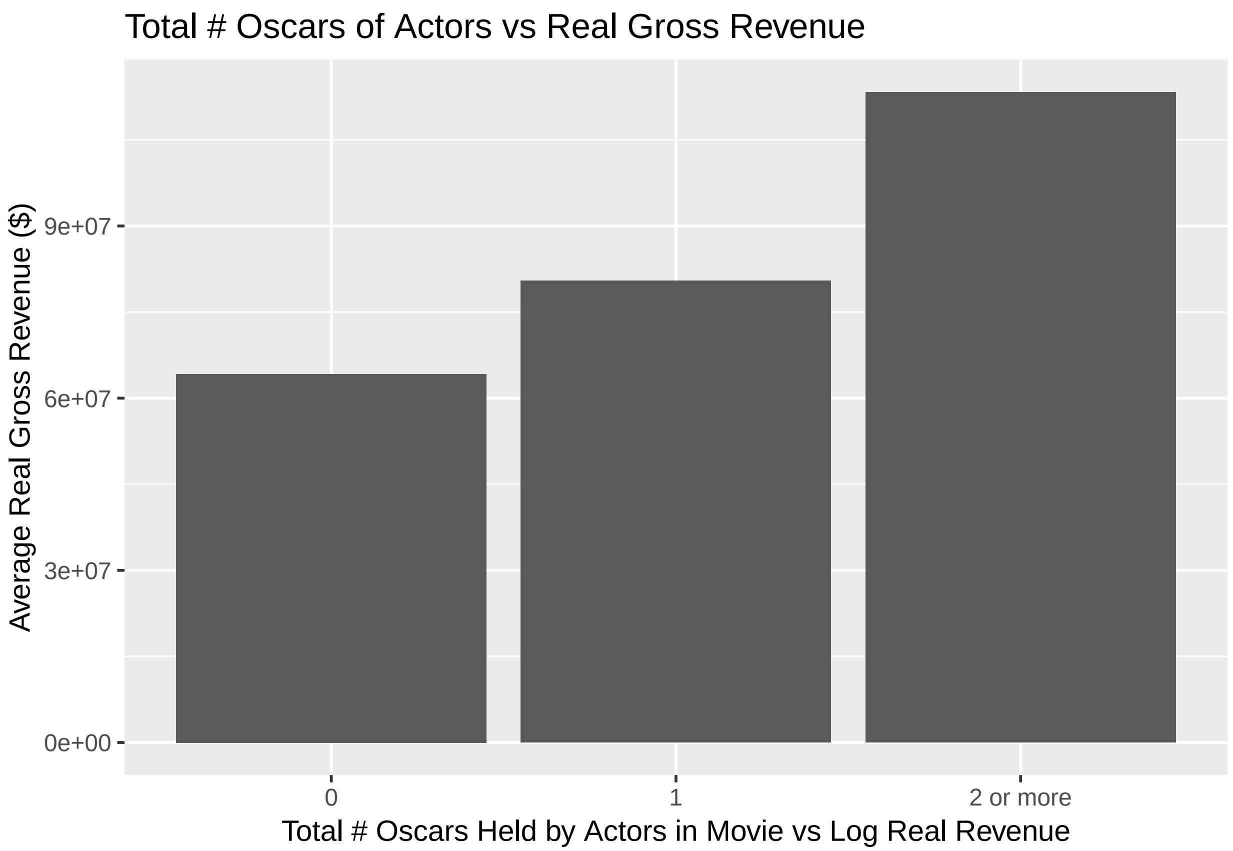
```r
# Versions of data with average revenue by number of oscars
train_oscar_actor <- train %>%
  group_by(total_oscars_actor) %>%
  summarize(avg_real_gross = mean(real_gross),
            avg_real_gross_log = mean(real_gross_log))
train_oscar_director <- train %>%
  group_by(total_oscars_director)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

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

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

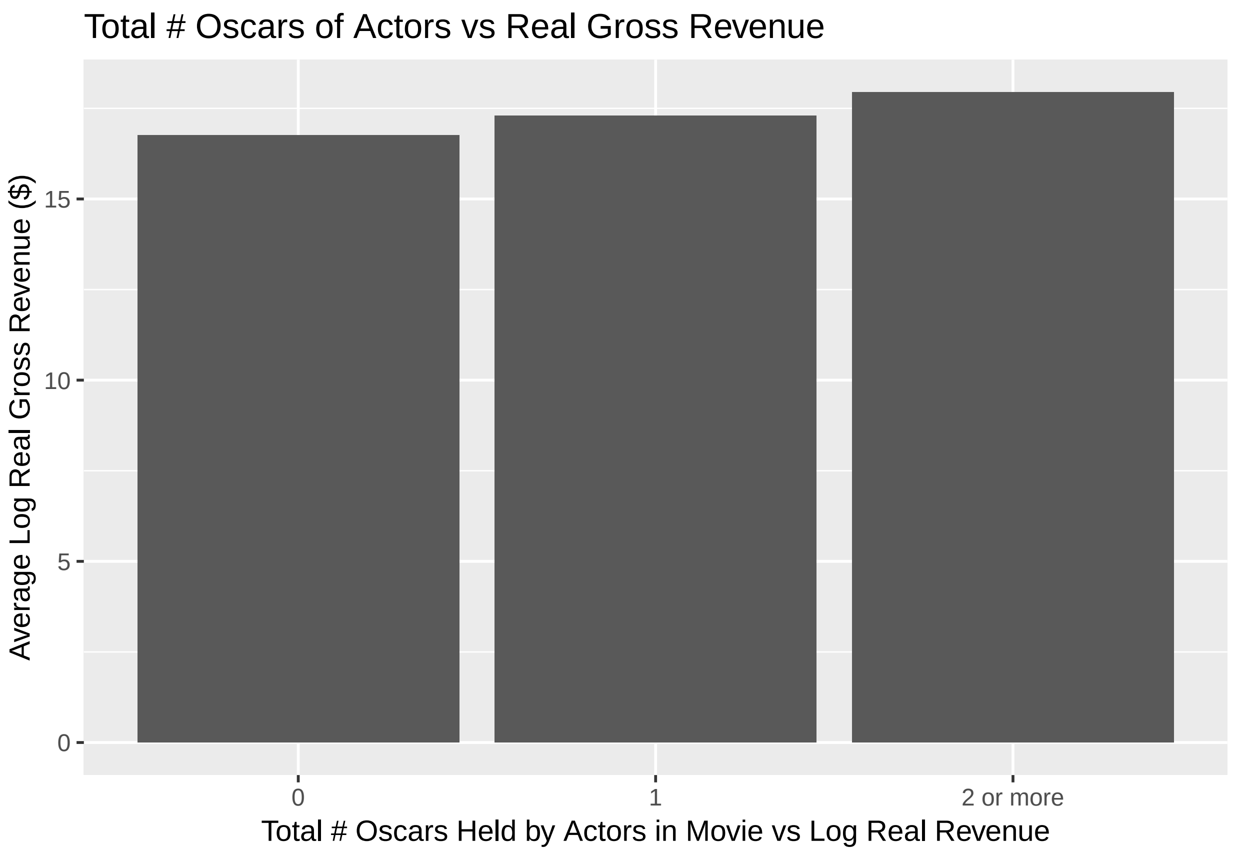
```
# actors
oscar_box(train, 'total_oscars_actor', 'Total # Oscars of Actors vs Real Gross Revenue',
          'Total # Oscars Held by Actors in Movie')
```



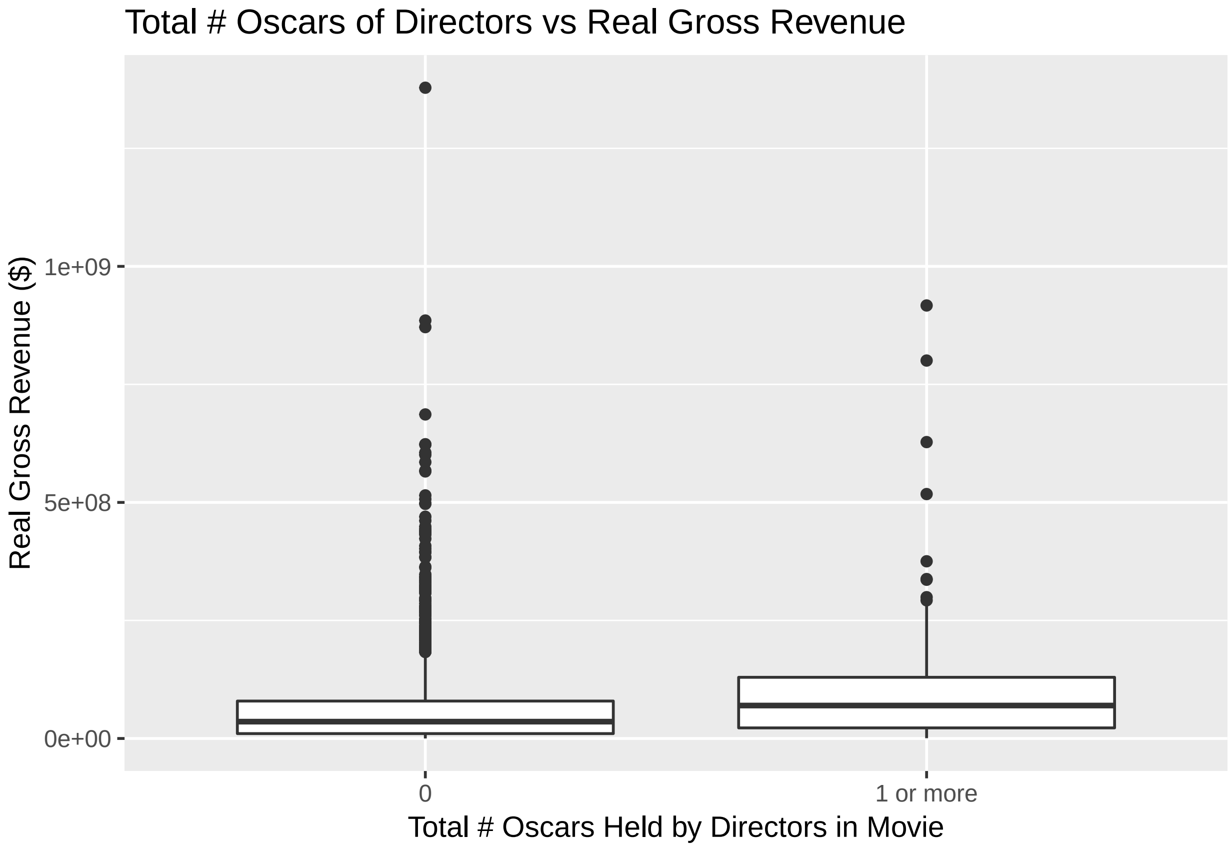


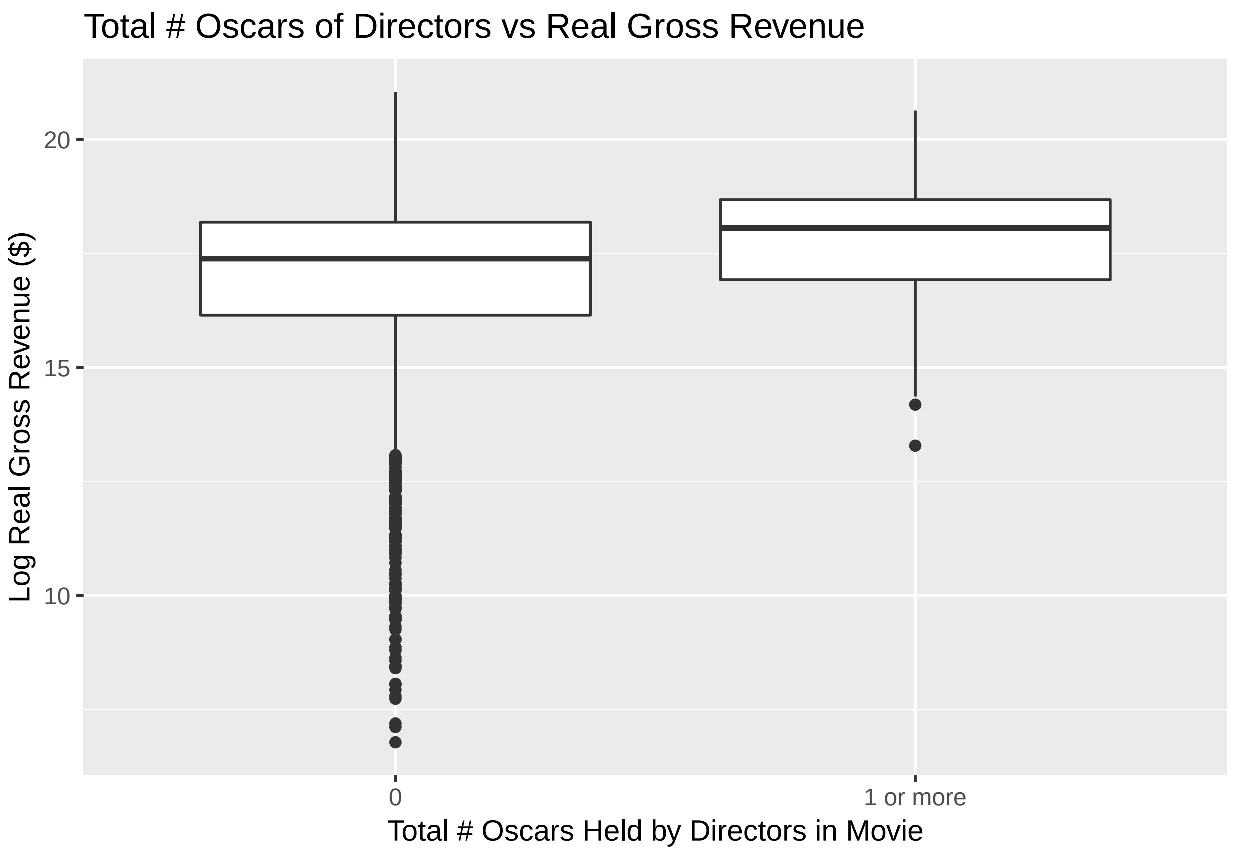
```
oscar_bar(train_oscar_actor, 'total_oscars_actor', 'Total # Oscars of Actors vs Real Gross Revenue',
          'Total # Oscars Held by Actors in Movie vs Log Real Revenue')
```


Total # Oscars of Actors vs Real Gross Revenue
Average Real Gross Revenue ($)
9e+07
6e+07
3e+07
0e+00
0
1
2 or more
Total # Oscars Held by Actors in Movie vs Log Real Revenue

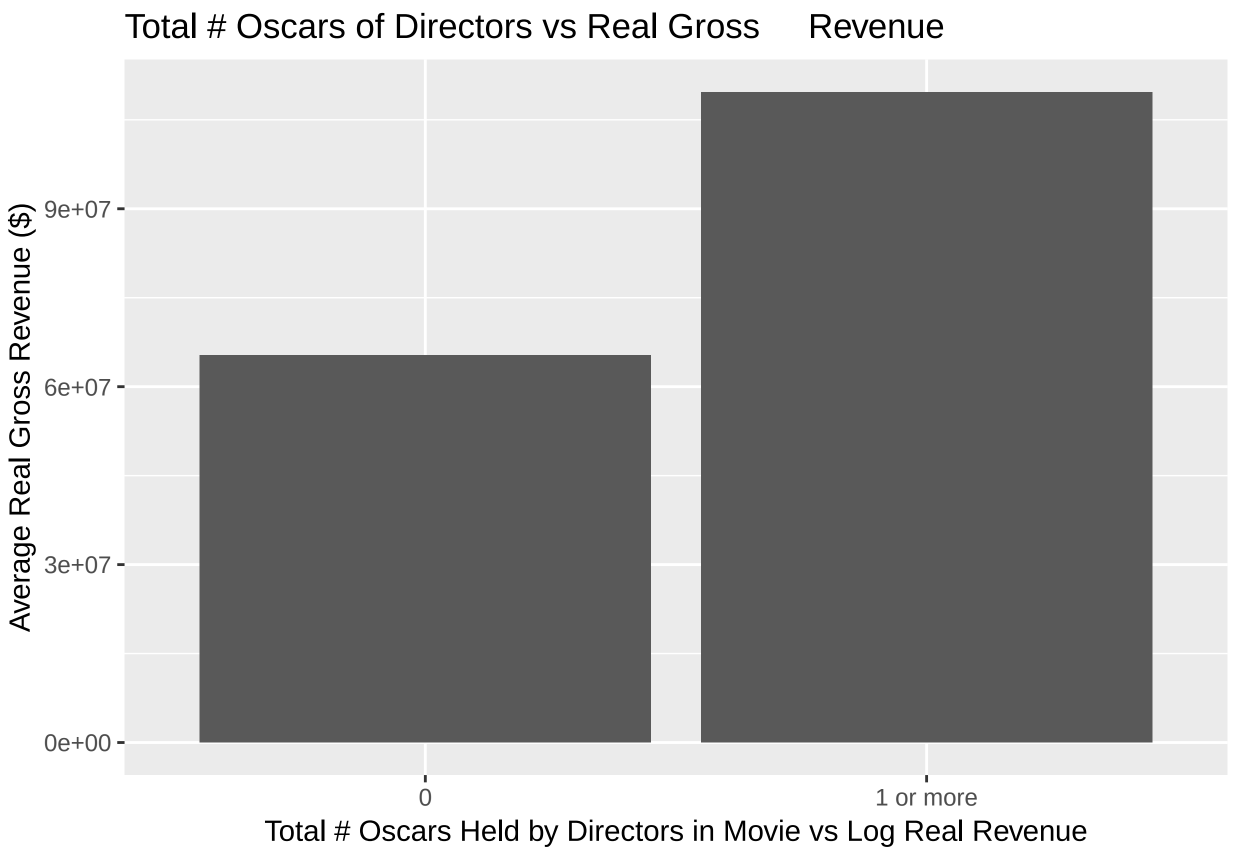


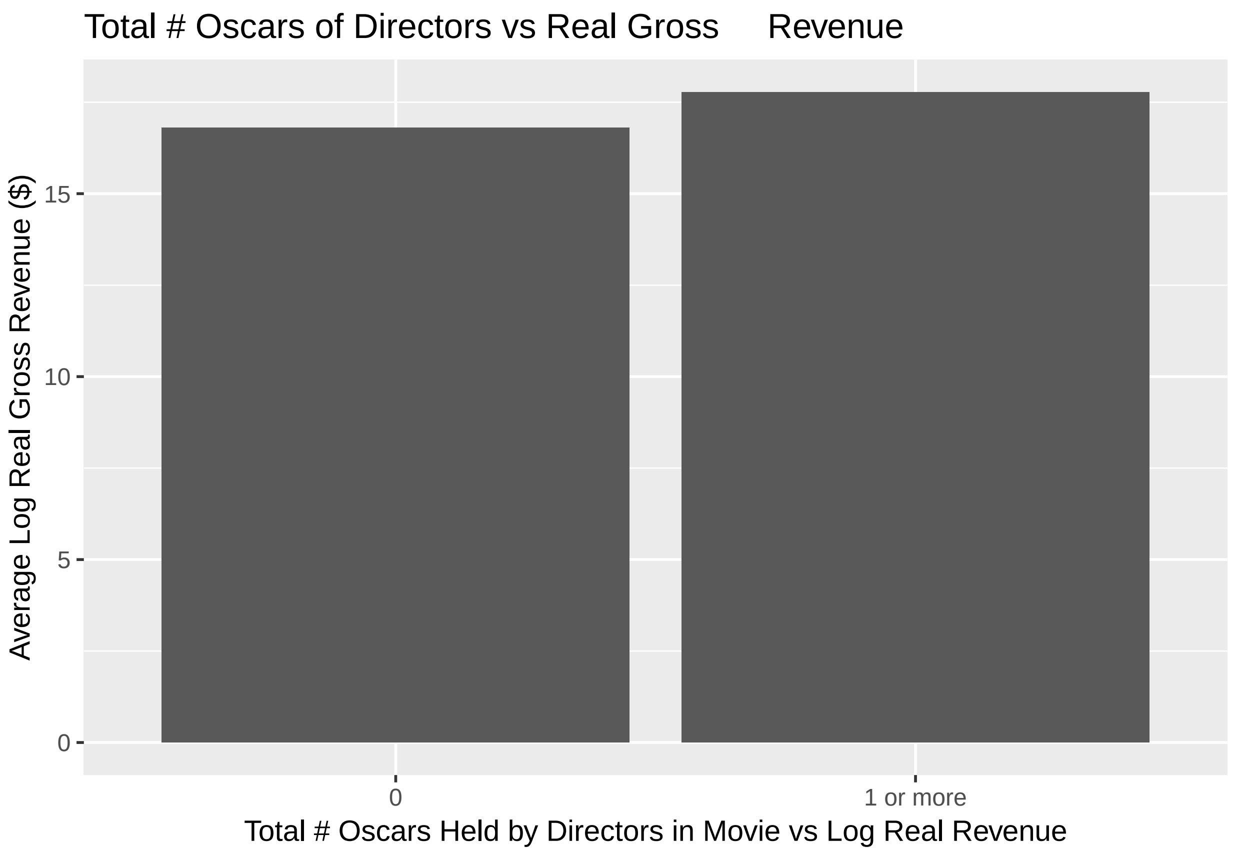
```
# directors
oscar_box(train, 'total_oscars_director', 'Total # Oscars of Directors vs Real Gross Revenue',
          'Total # Oscars Held by Directors in Movie')
```


Total # Oscars of Directors vs Real Gross Revenue
Real Gross Revenue ($)
1e+09
5e+08
0e+00
0
1 or more
Total # Oscars Held by Directors in Movie



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


Total # Oscars of Directors vs Real Gross   Revenue
Average Real Gross Revenue ($)
9e+07
6e+07
3e+07
0e+00
0
1 or more
Total # Oscars Held by Directors in Movie vs Log Real Revenue



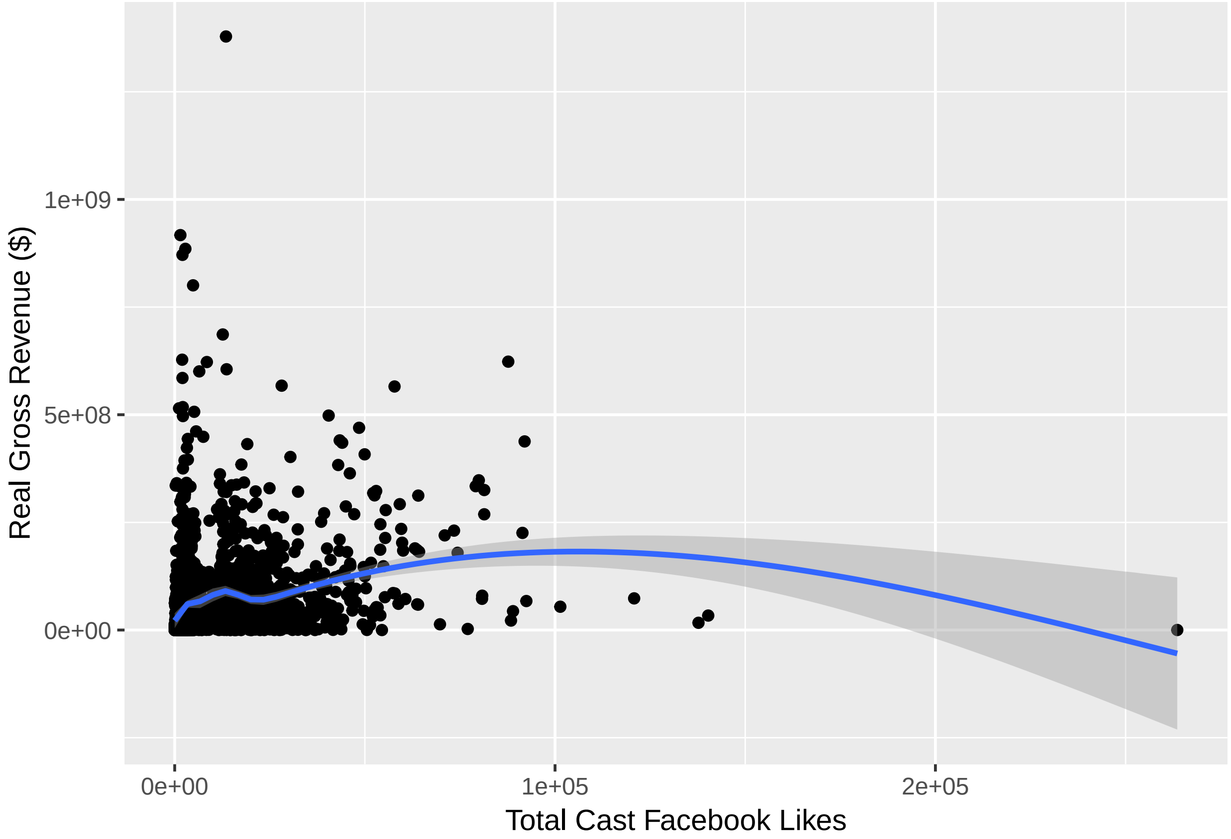
Facebook Likes

Graph total cast facebook likes and director facebook likes vs real gross revenue. Continuous, so scatter plot. Trying log versions too.

Not as strong a relationship as I originally expected. Not sure this is worth including either.

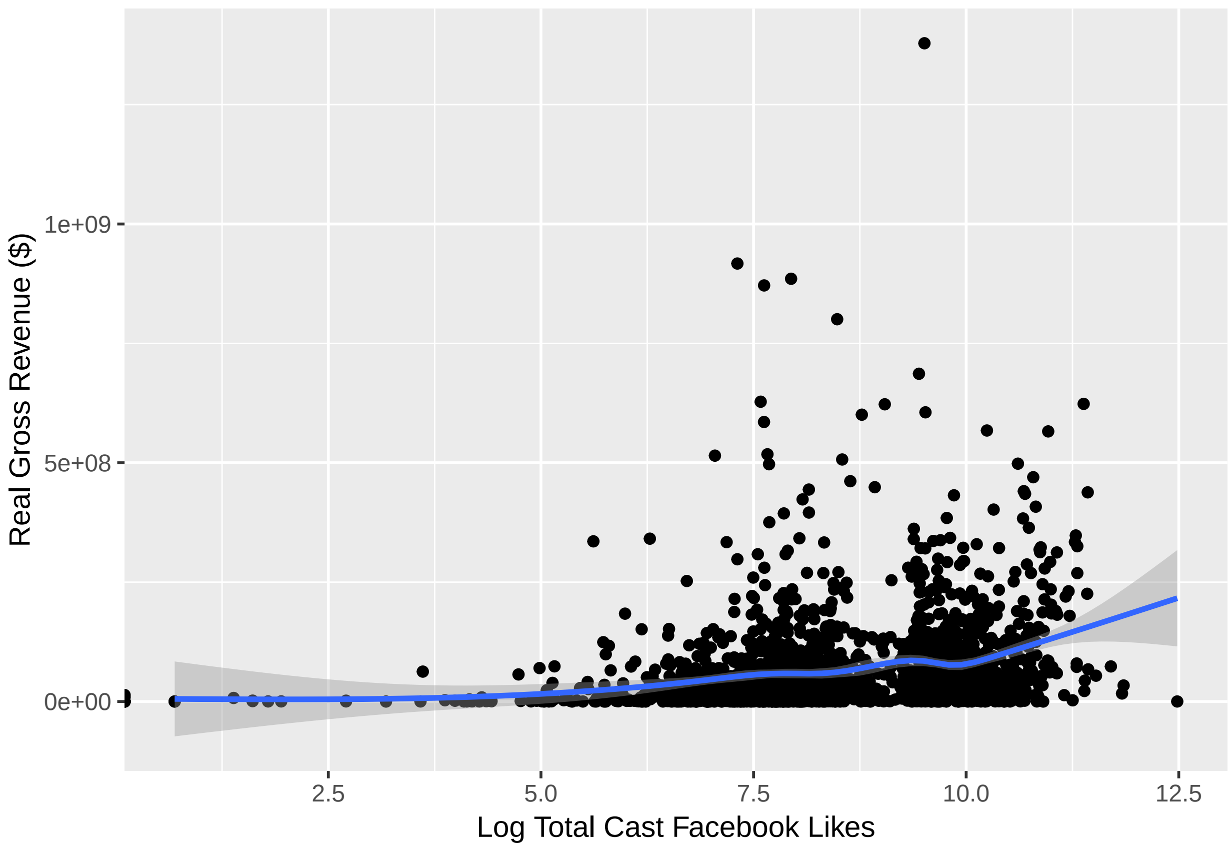
```r
facebook_plot <- function(xvar, xlab) {
 print(train %>%
   ggplot(aes_string(x = xvar, y = 'real_gross')) +
   geom_point() +
   geom_smooth() +
   labs(x = xlab, y = 'Real Gross Revenue ($)'))
 print(train %>%
   ggplot(aes_string(x = str_c(xvar, '_log'), y = 'real_gross')) +
   geom_point() +
   geom_smooth() +
   labs(x = str_c('Log ', xlab), y = 'Real Gross Revenue ($)'))
 print(train %>%
   ggplot(aes_string(x = xvar, y = 'real_gross_log')) +
   geom_point() +
   geom_smooth() +
   labs(x = xlab, y = 'Log Real Gross Revenue ($)'))
 print(train %>%
   ggplot(aes_string(x = str_c(xvar, '_log'), y = 'real_gross_log')) +
   geom_point() +
   geom_smooth() +
    labs(x = str_c('Log ', xlab), y = 'Log Real Gross Revenue ($)'))
}

facebook_plot('cast_total_facebook_likes', 'Total Cast Facebook Likes')
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

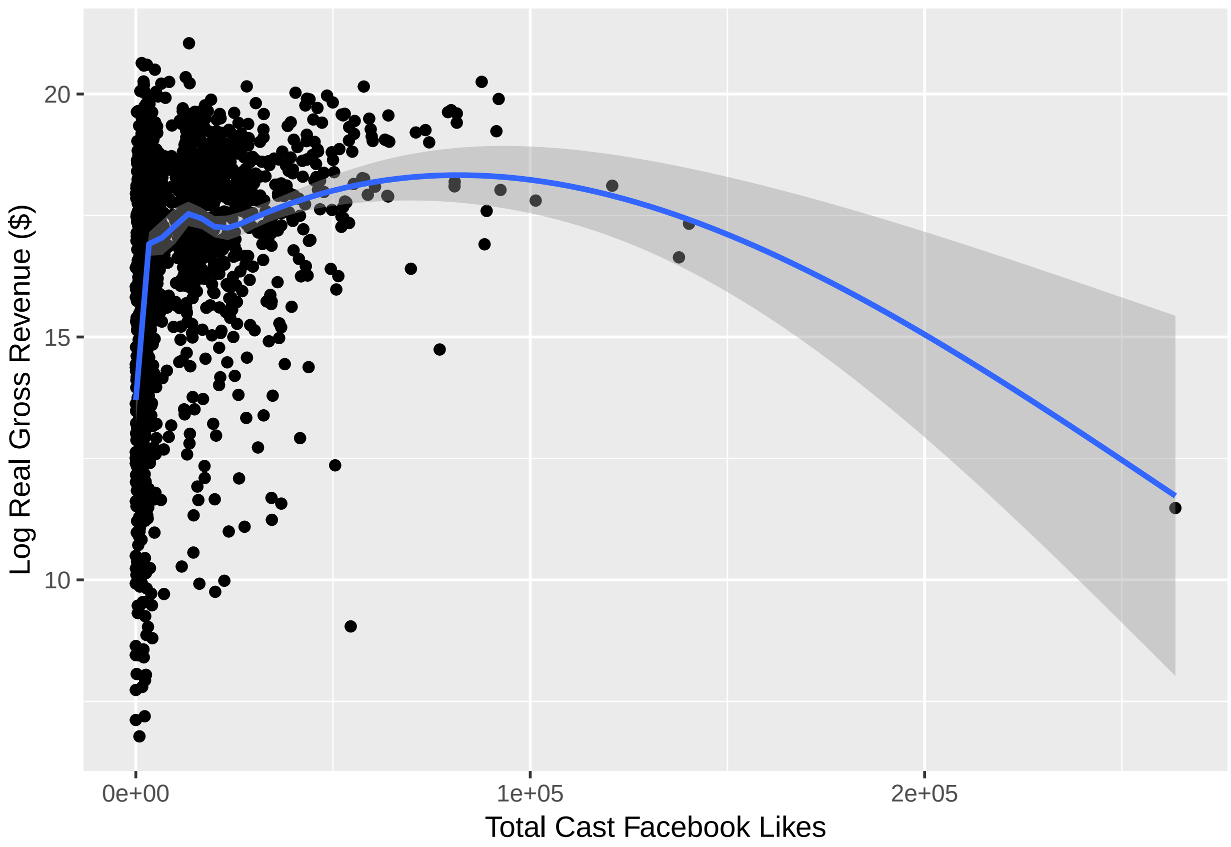


```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

```
## Warning: Removed 8 rows containing non-finite values (stat_smooth).
```

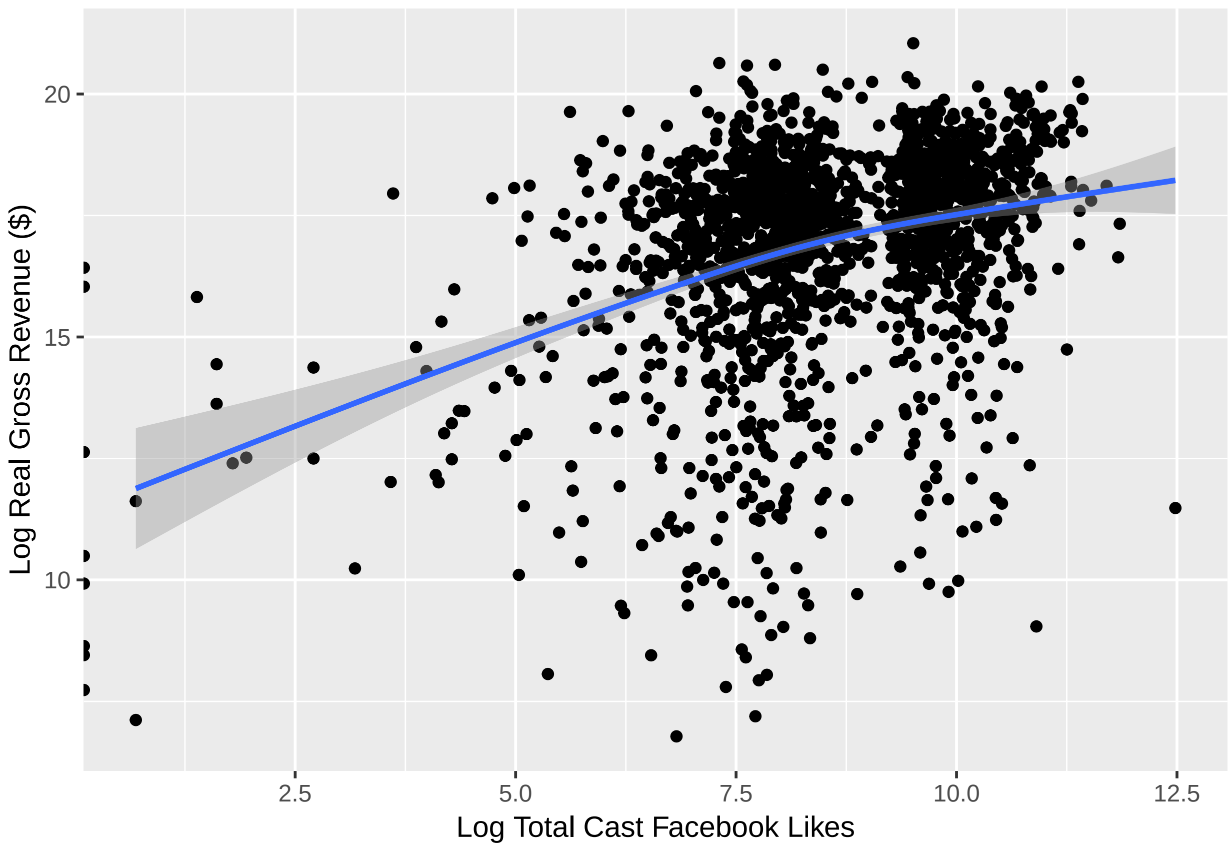


```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



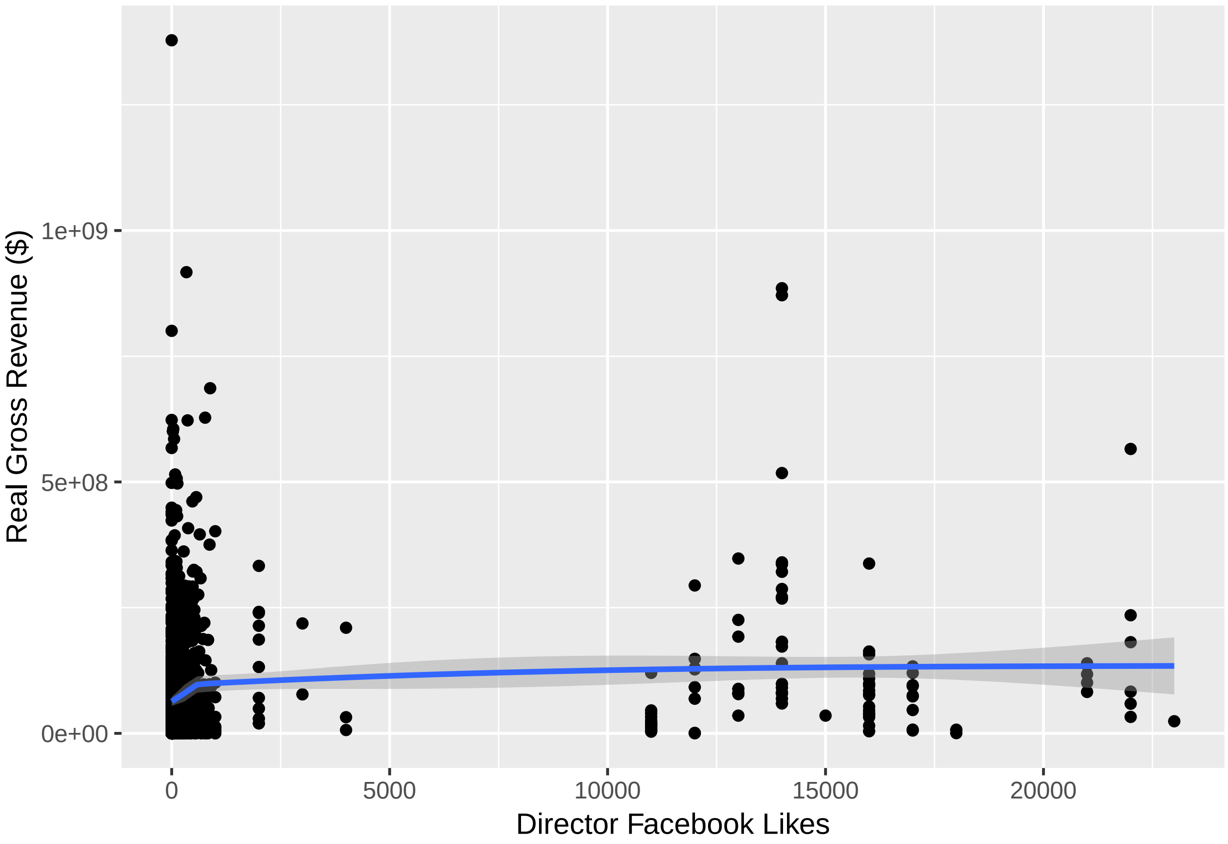
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

```
## Warning: Removed 8 rows containing non-finite values (stat_smooth).
```

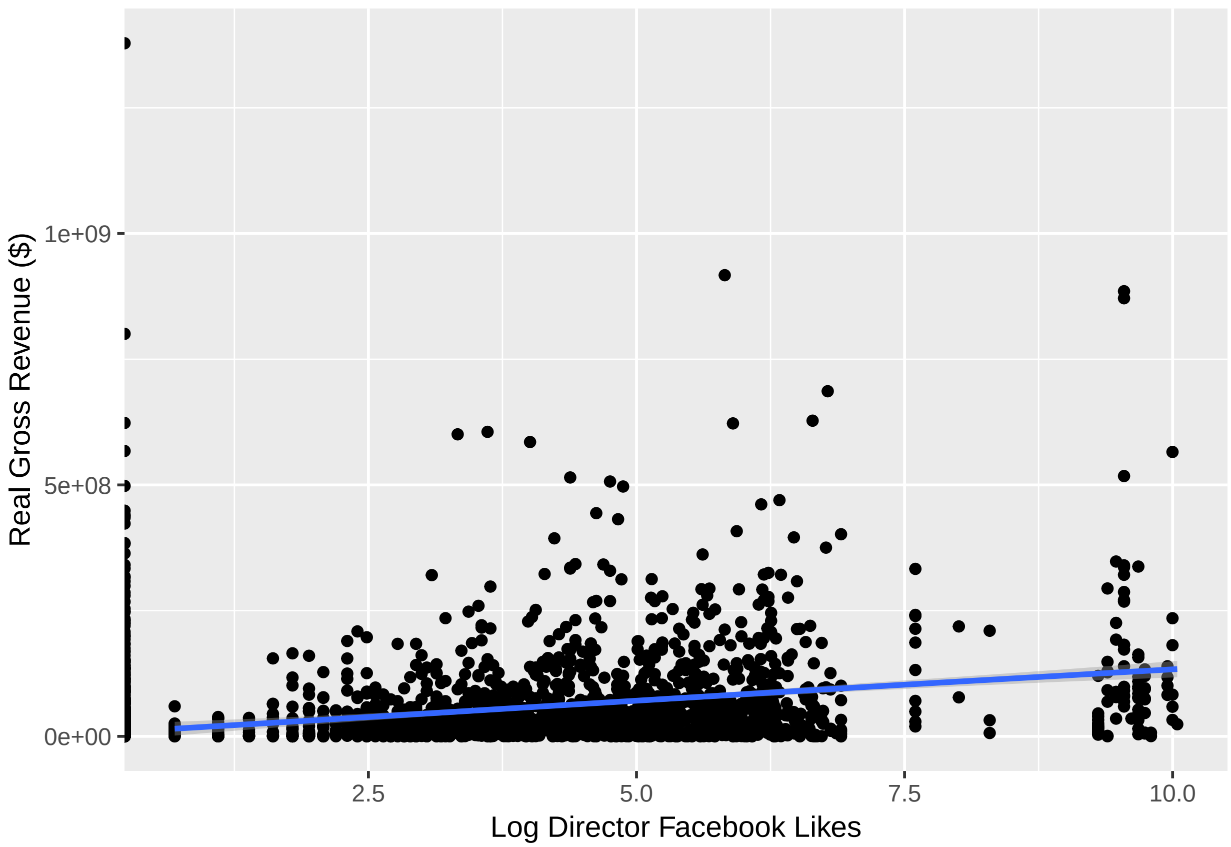


```
facebook_plot('director_facebook_likes', 'Director Facebook Likes')
```

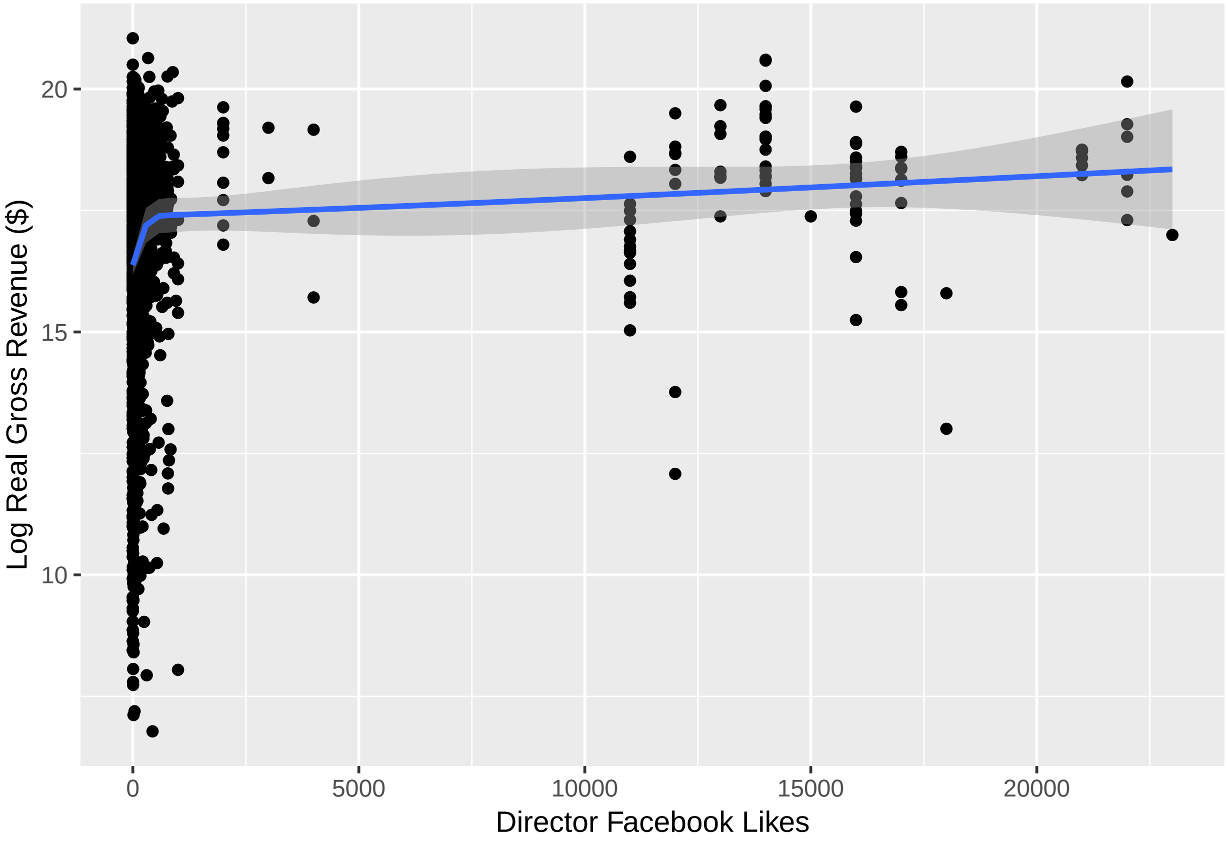
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



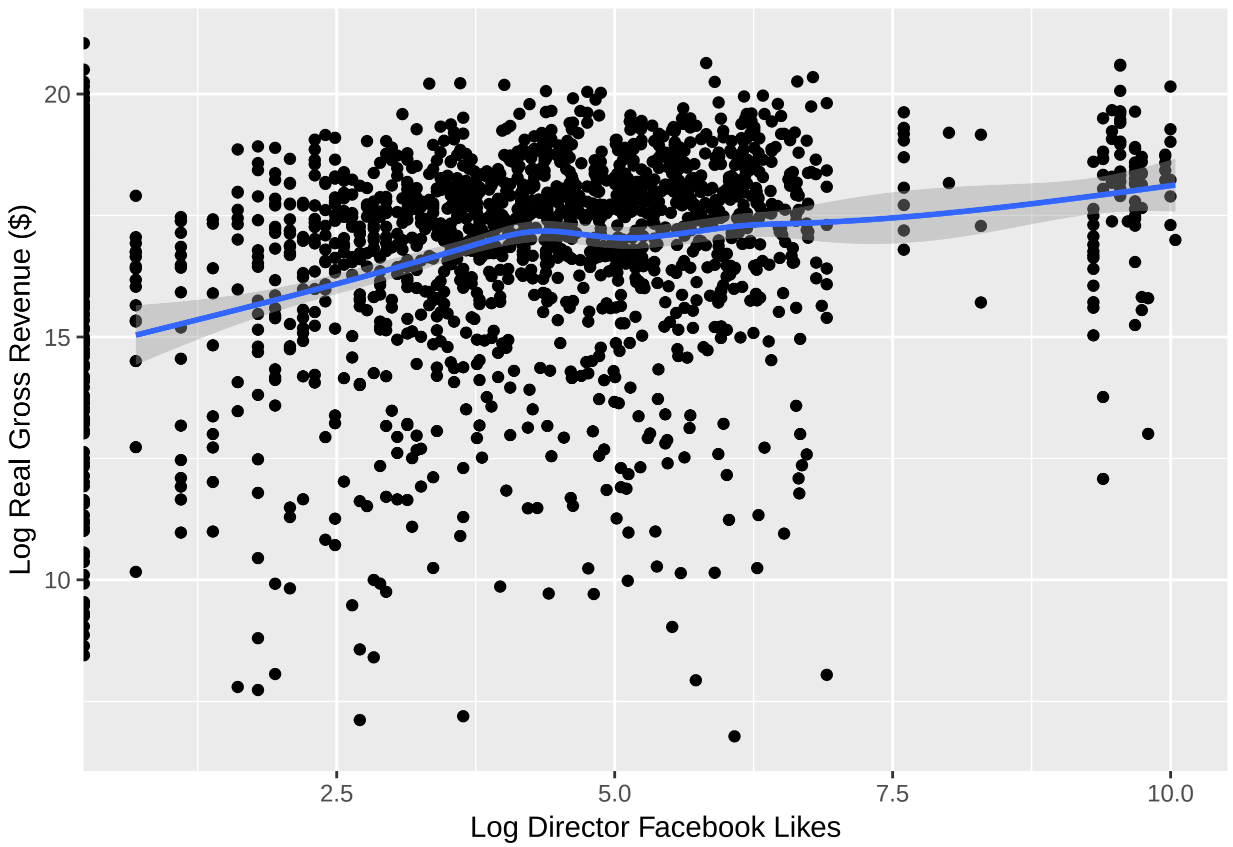
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'

## Warning: Removed 340 rows containing non-finite values (stat_smooth).
```



```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



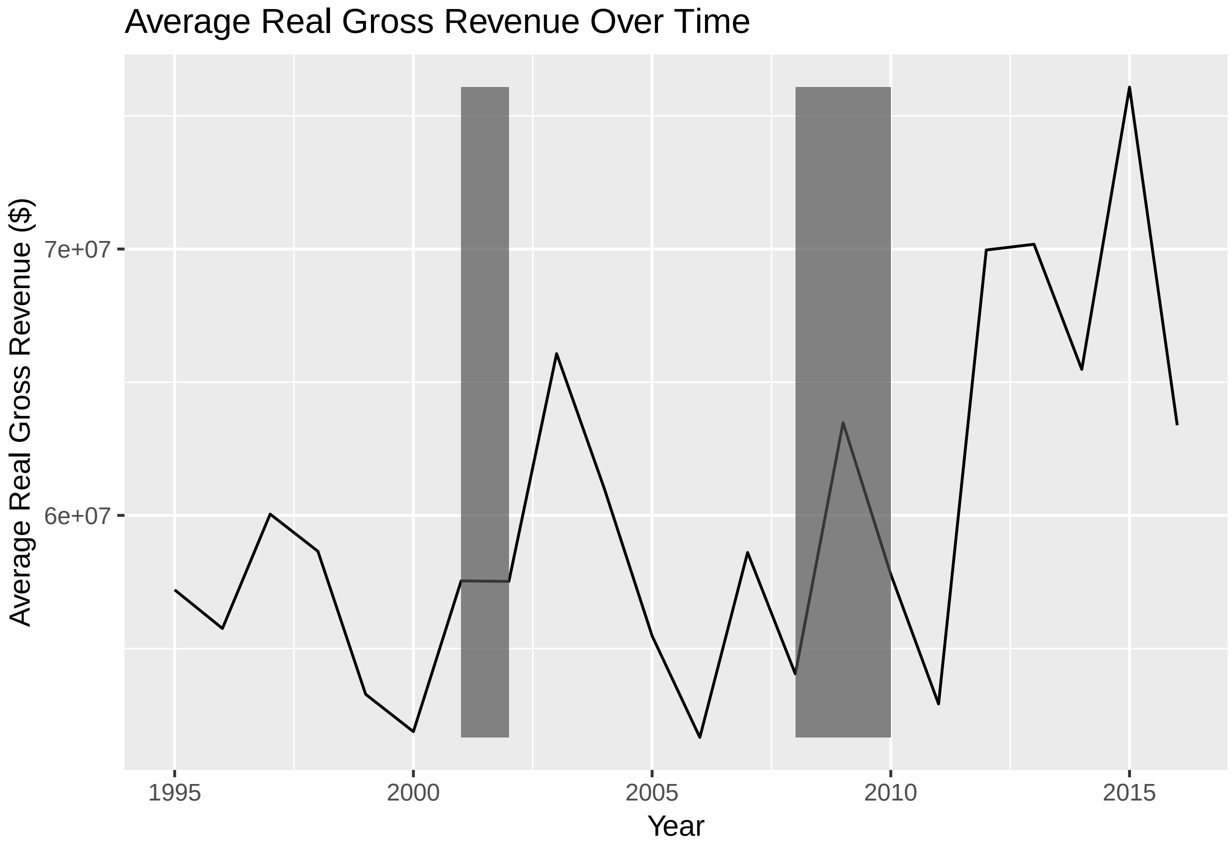
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'

## Warning: Removed 340 rows containing non-finite values (stat_smooth).
```



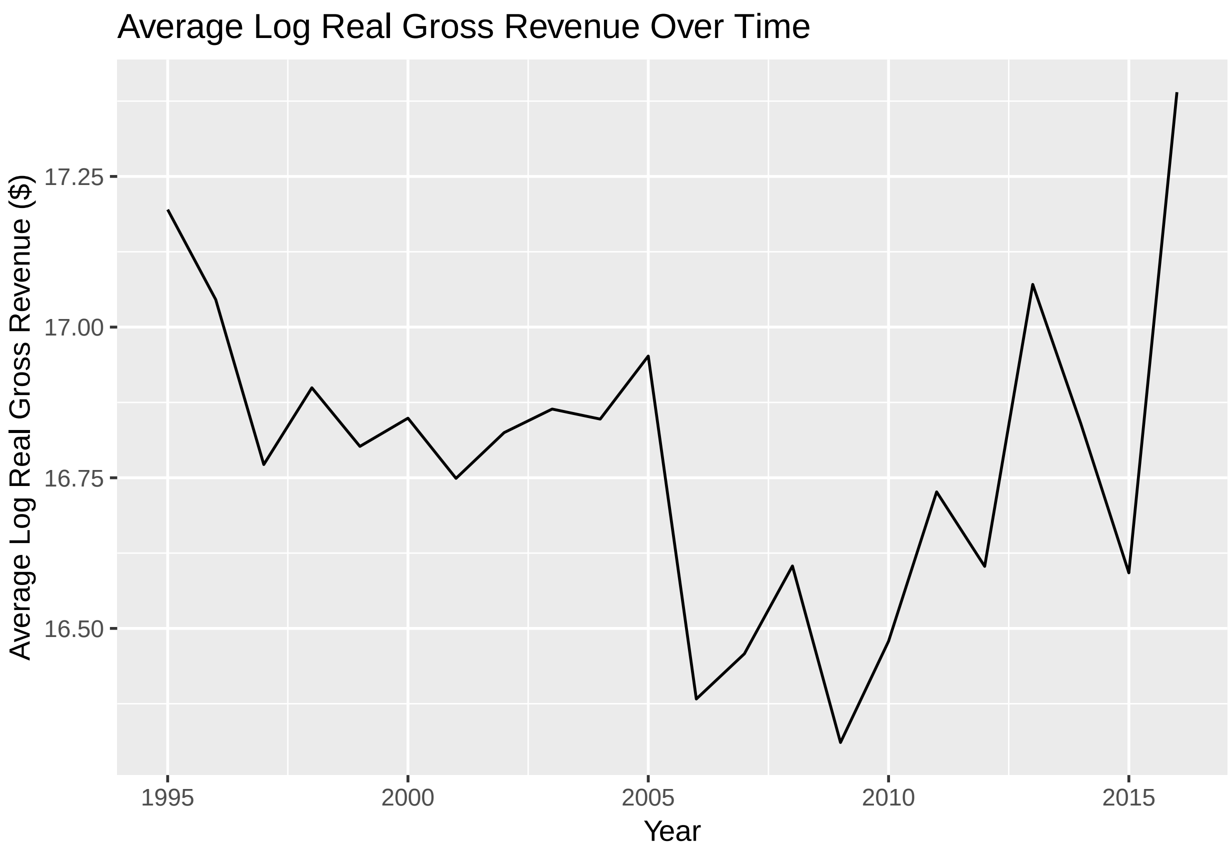
Year

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

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

# 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

plt <- ggplot(data = train_sum %>% filter(year >= 1995))
plt +
  geom_line(aes(x = year, y = real_gross_avg)) +
  labs(title = 'Average Real Gross Revenue Over Time', x = 'Year', y = 'Average Real Gross Revenue ($)'
    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)
```



```
plt +
  geom_line(aes(x = year, y = real_gross_avg_log)) +
  labs(title = 'Average Log Real Gross Revenue Over Time', x = 'Year', y = 'Average Log Real Gross Rever
```

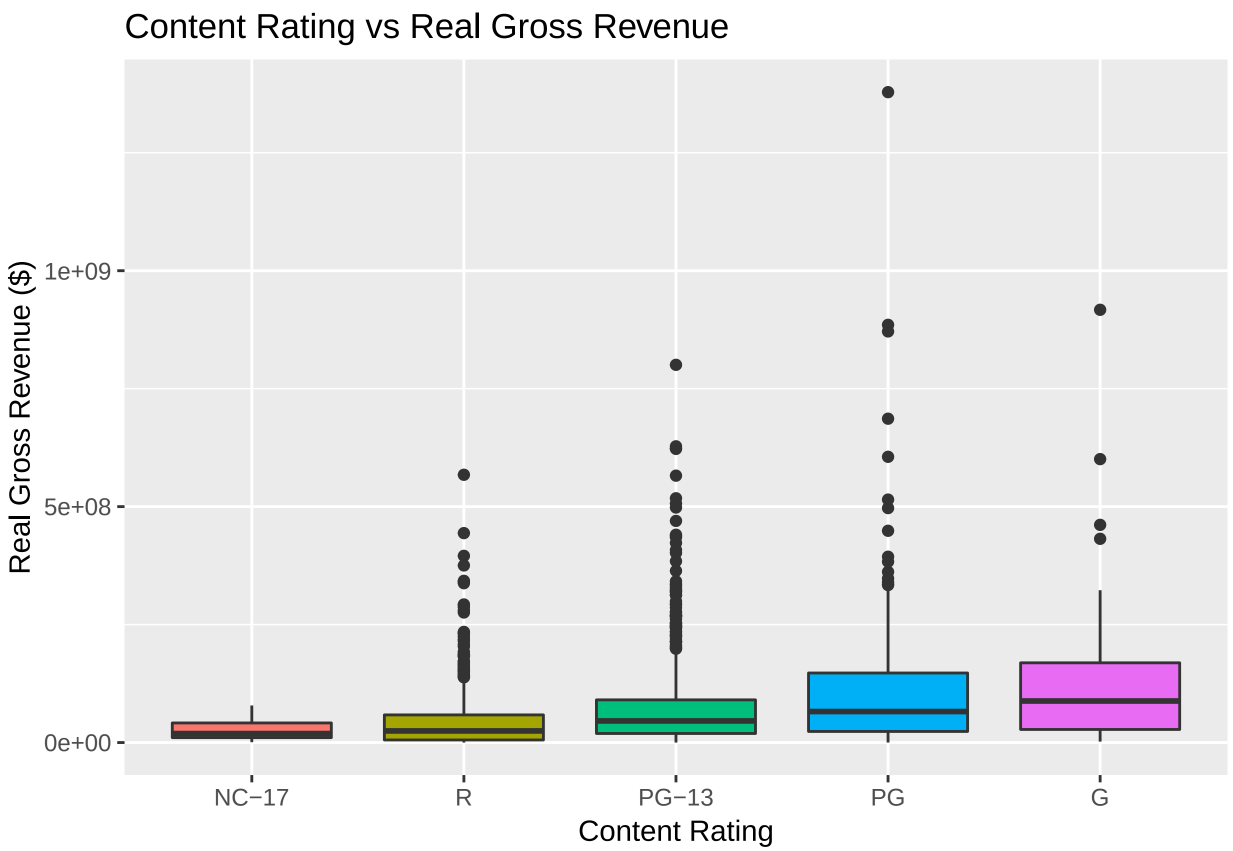


Content Rating

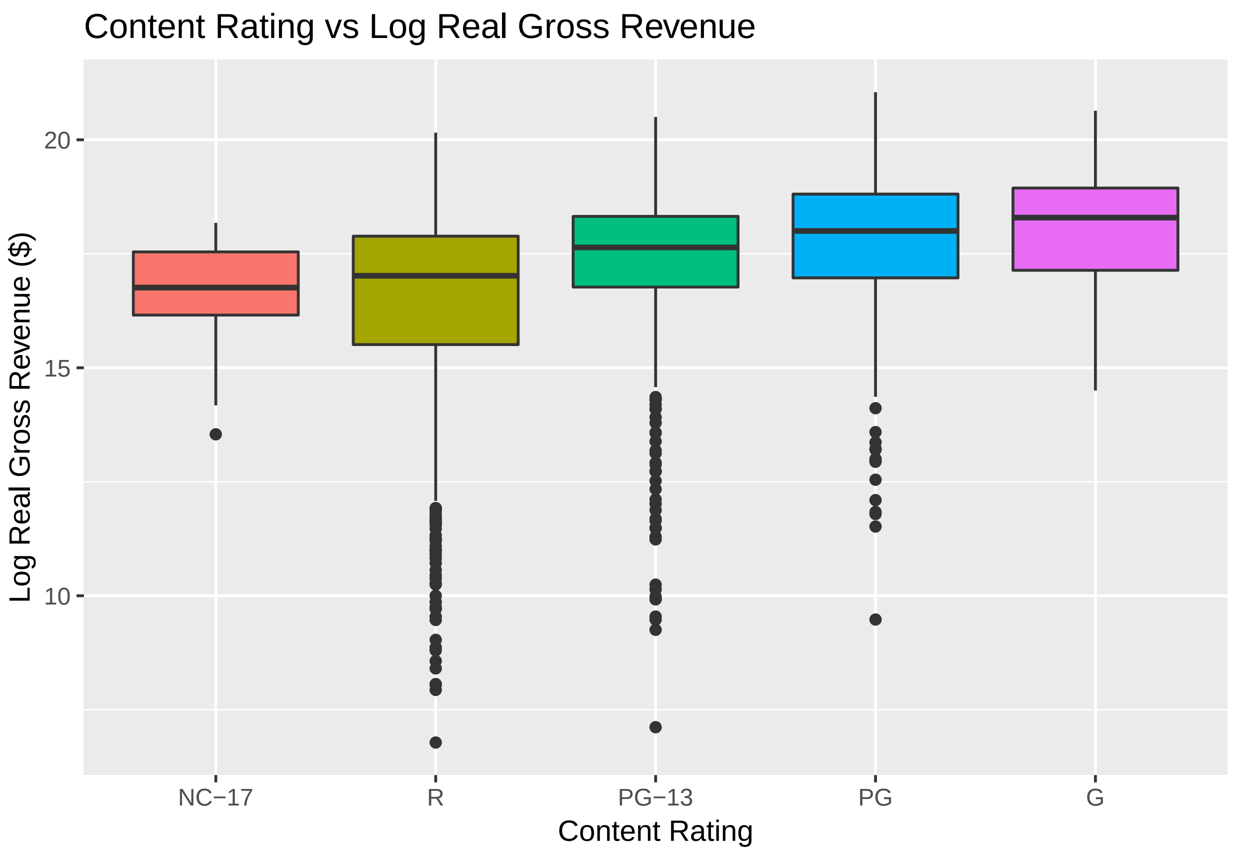
Bar graph of average real revenue and boxplot

Linear relationship. Good candidate to include in the model Each individual rating has an significantly different average mean real revenue

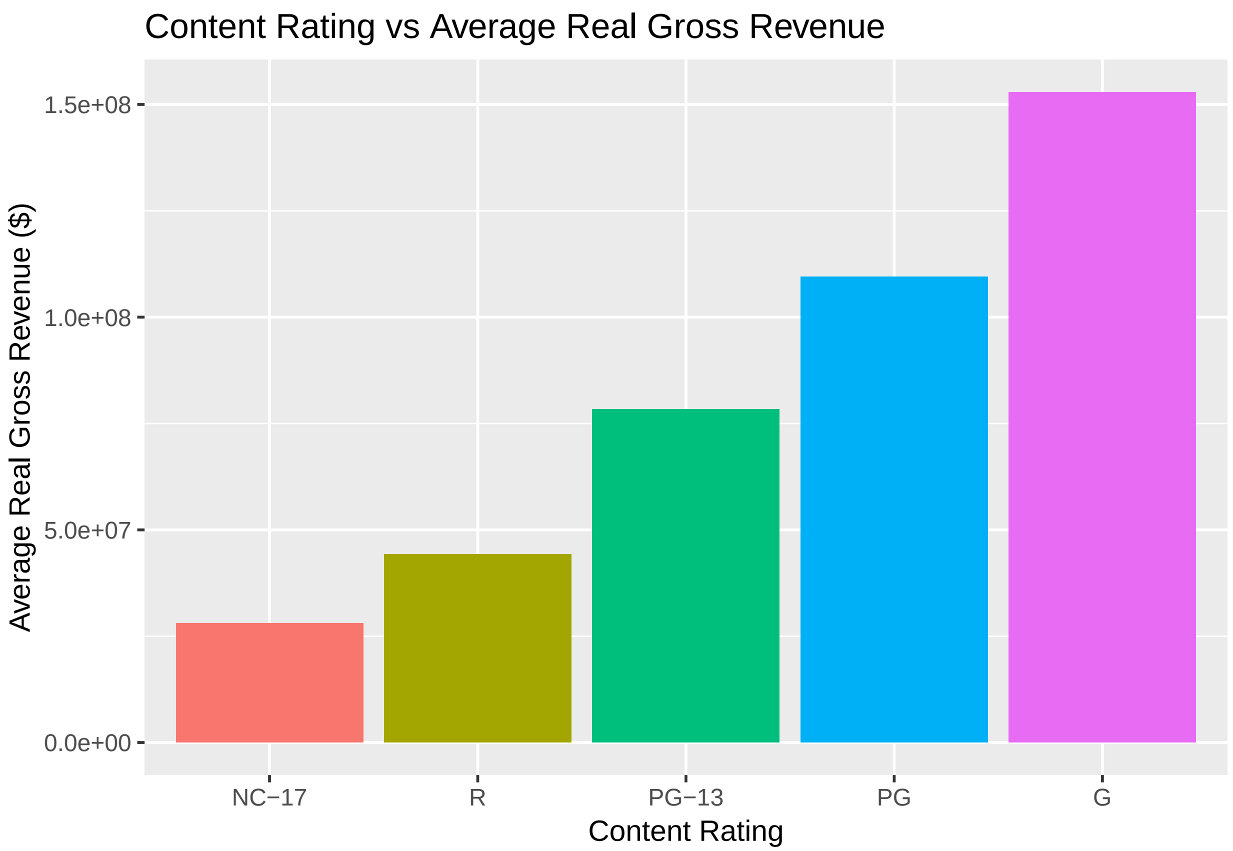
```
# data manipulation. Factor.
train_content <- train %>%
  # filter out missing
  filter(!is.na(content_rating)) %>%
  # reorder content rating based on gross
  mutate(content_rating = reorder(content_rating, real_gross))

# 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')
```



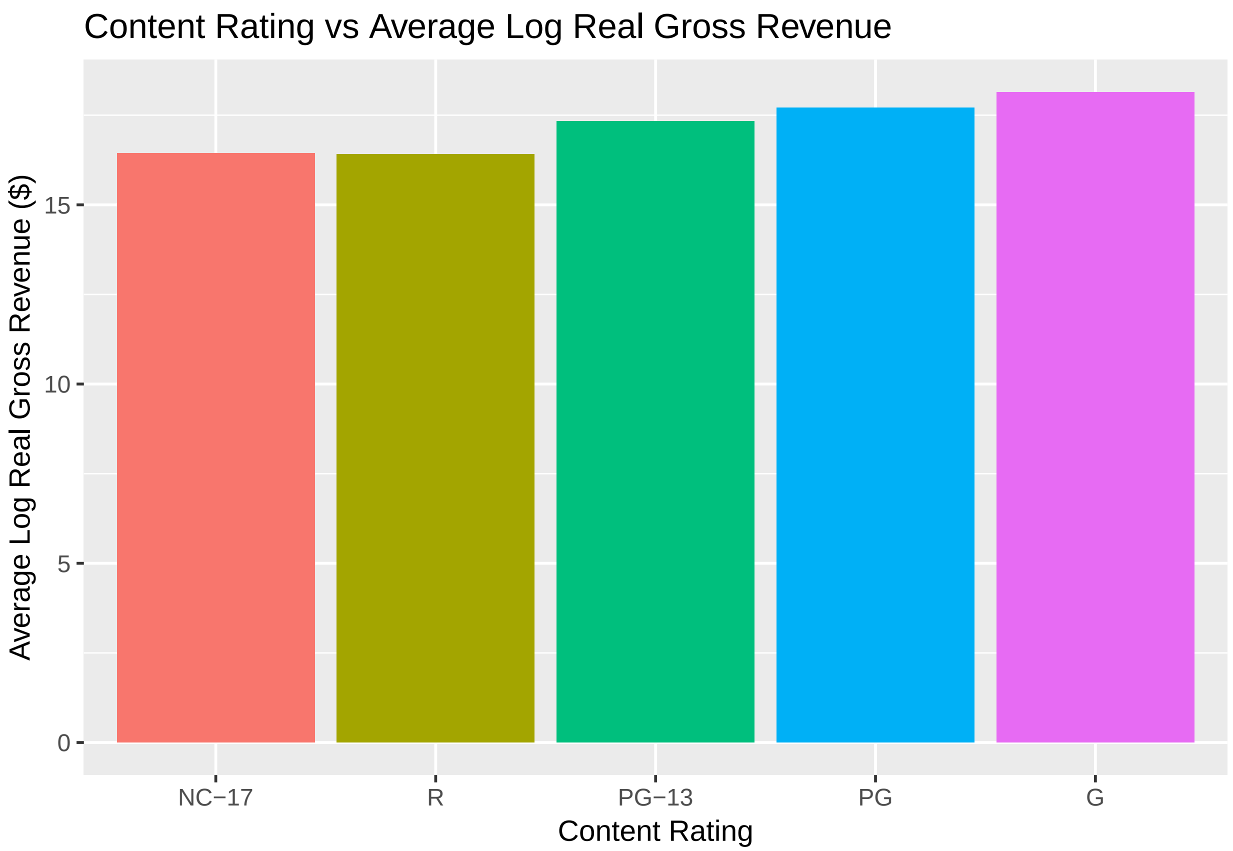
```
train_content  %>%
  ggplot() +
  geom_boxplot(aes(x = content_rating, y = real_gross_log, fill = content_rating)) +
  labs(x = 'Content Rating', y = 'Log Real Gross Revenue ($)',
       title = 'Content Rating vs Log Real Gross Revenue') +
  theme(legend.position = 'none')
```



```
# 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')
```

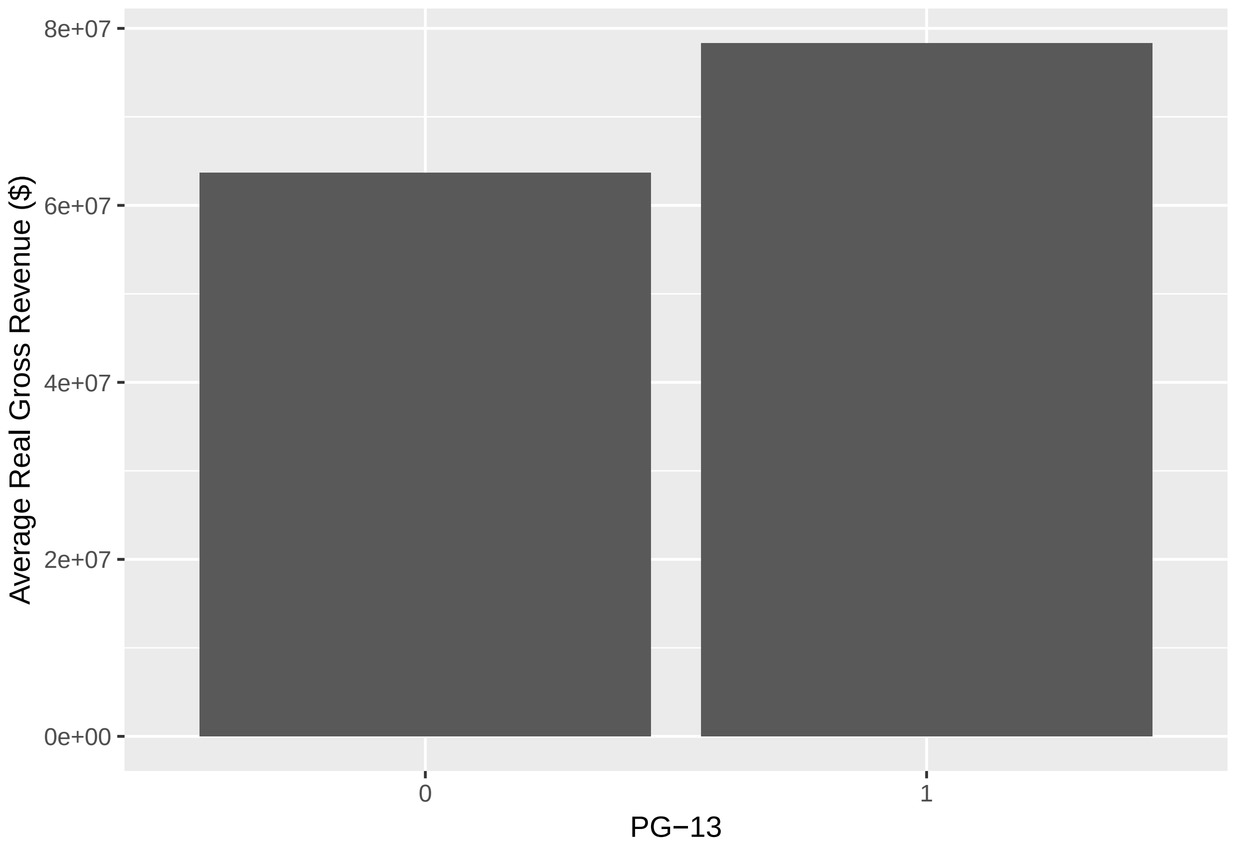


```r
train_content  %>%
  # average revenue by content rating
  group_by(content_rating) %>%
  summarize(avg_real_gross_log = mean(real_gross_log)) %>%
  ggplot() +
  geom_col(aes(x = content_rating, y = avg_real_gross_log, fill = content_rating)) +
  labs(x = 'Content Rating', y = 'Average Log Real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

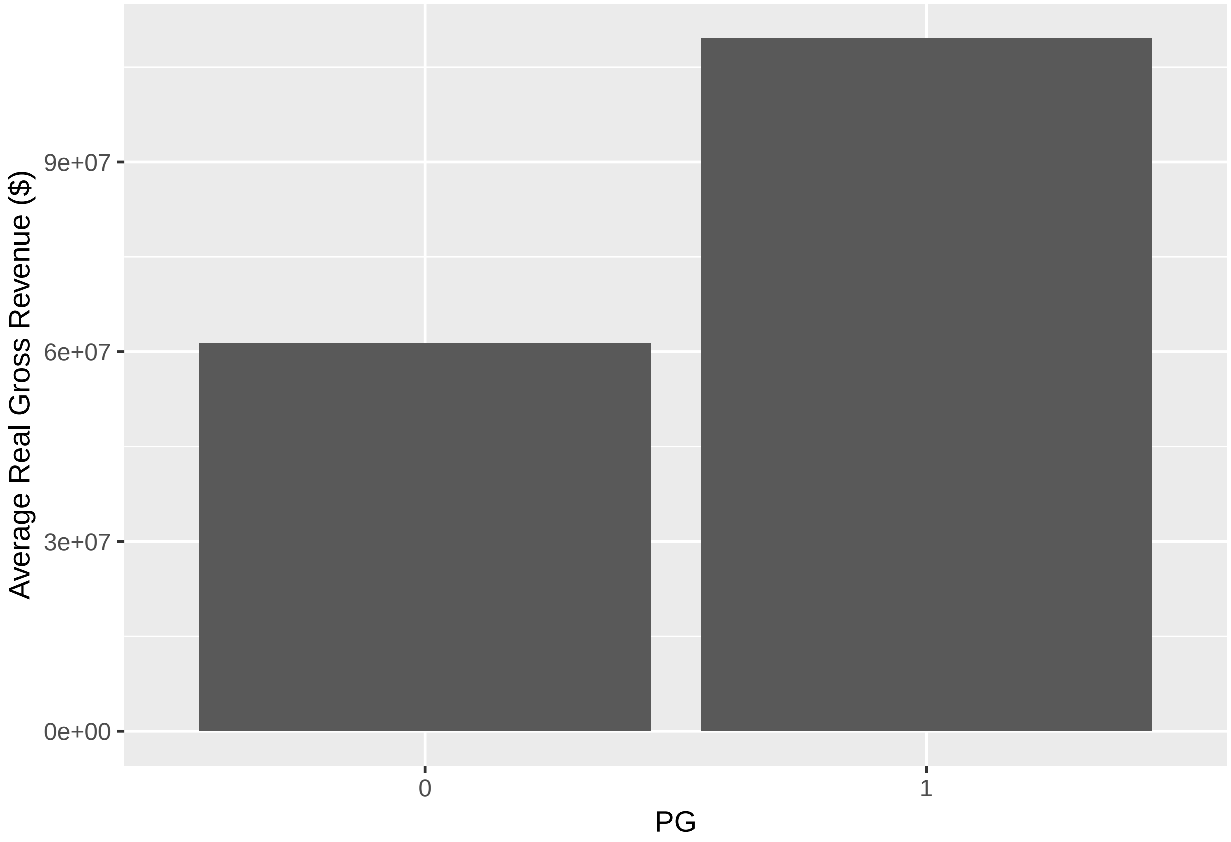


```r
# graph each content rating 0/1 against mean revenue: is there a difference?
lapply(unique(train_content$content_rating), function(r) {
  train_content %>%
    # get 1 if this rating, 0 else. Make factor
    mutate(rating_dum = as.factor(ifelse(content_rating == r, 1, 0))) %>%
    # mean revenue for 0 vs 1 for that content rating
    group_by(rating_dum) %>%
    summarize(avg_real_gross = mean(real_gross)) %>%
    ggplot() +
    geom_col(aes(x = rating_dum, y = avg_real_gross)) +
    labs(x = r, y = 'Average Real Gross Revenue ($)')
})
```

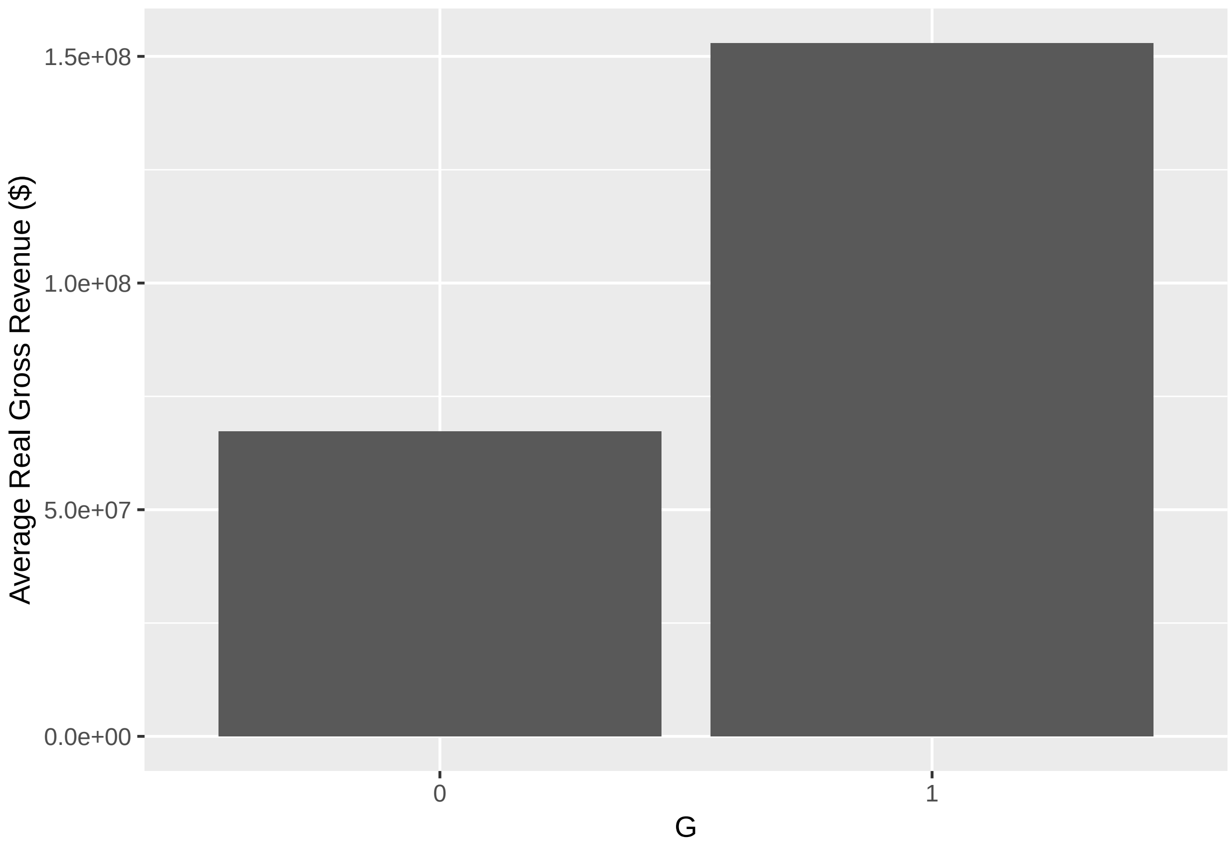
```
## [[1]]
```



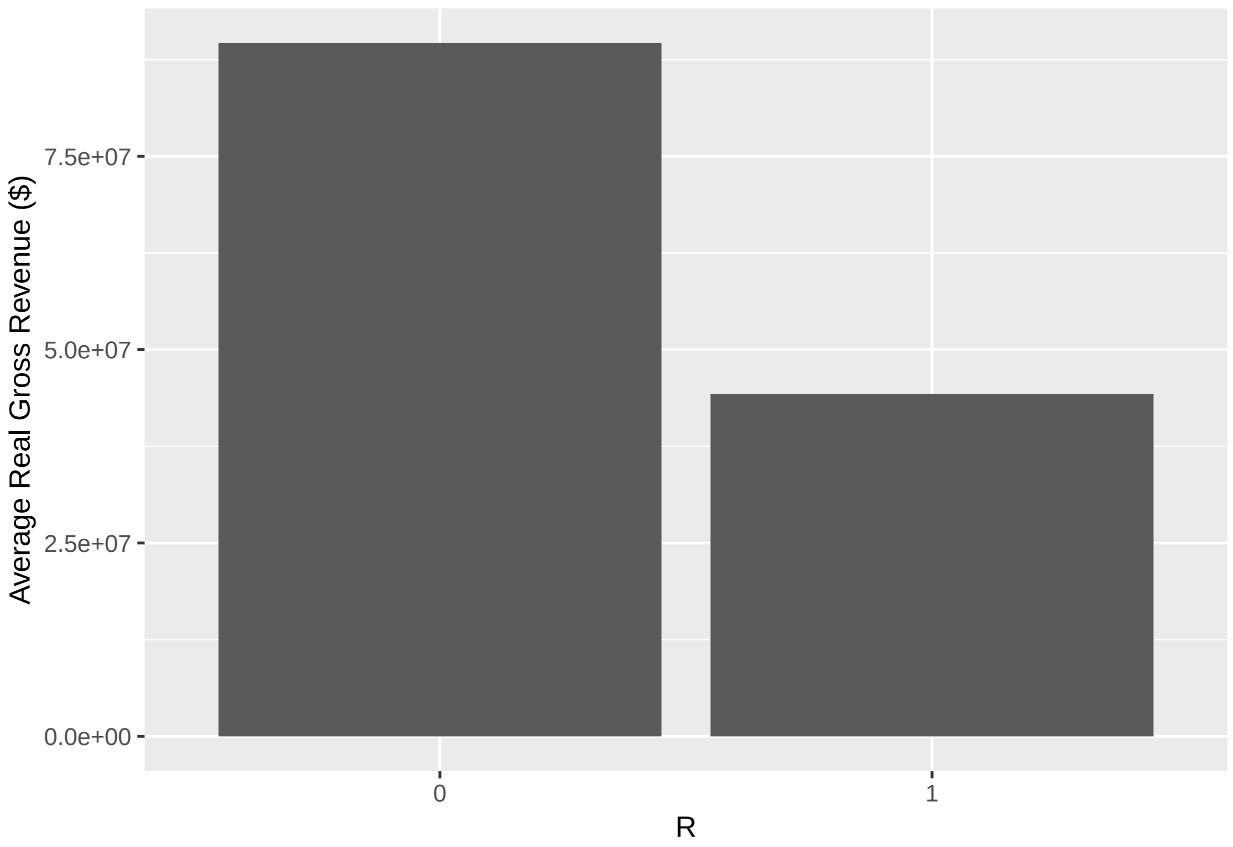
```
##
## [[2]]
```



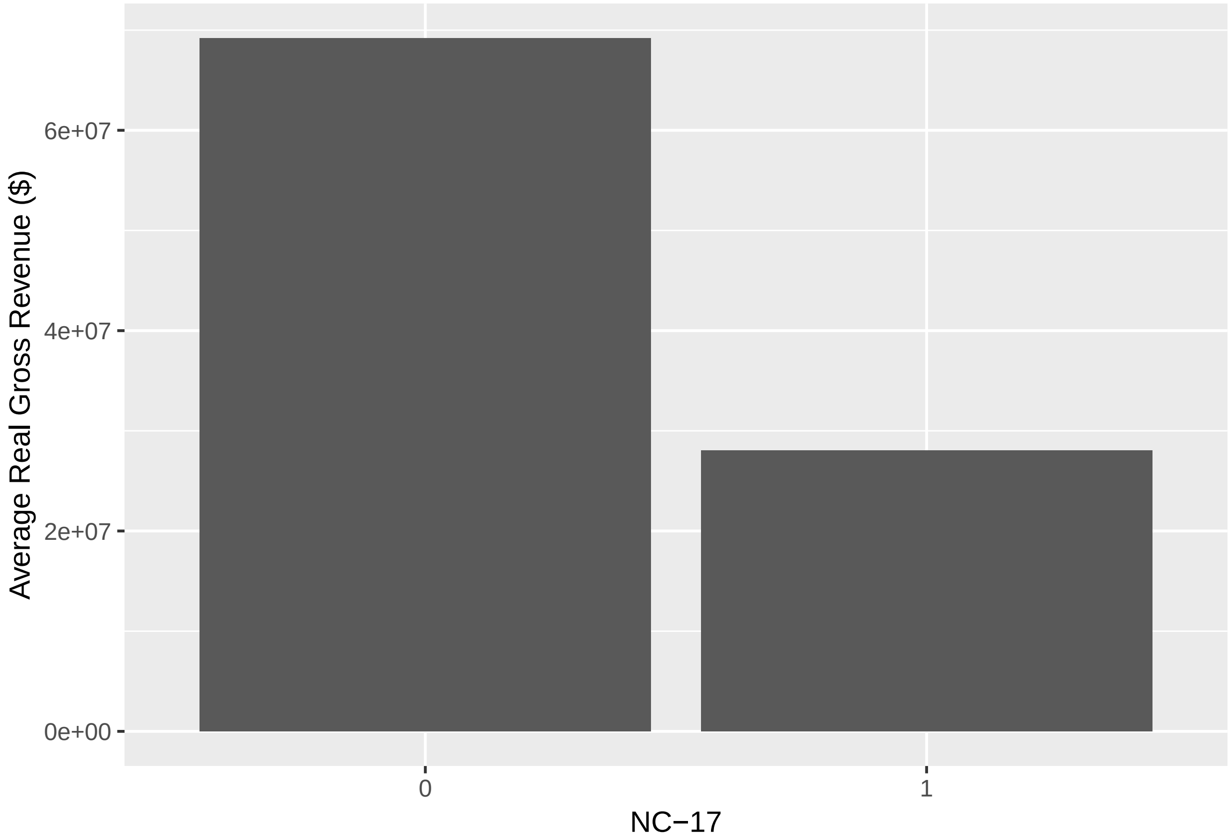
```
##
## [[3]]
```



```
##
## [[4]]
```

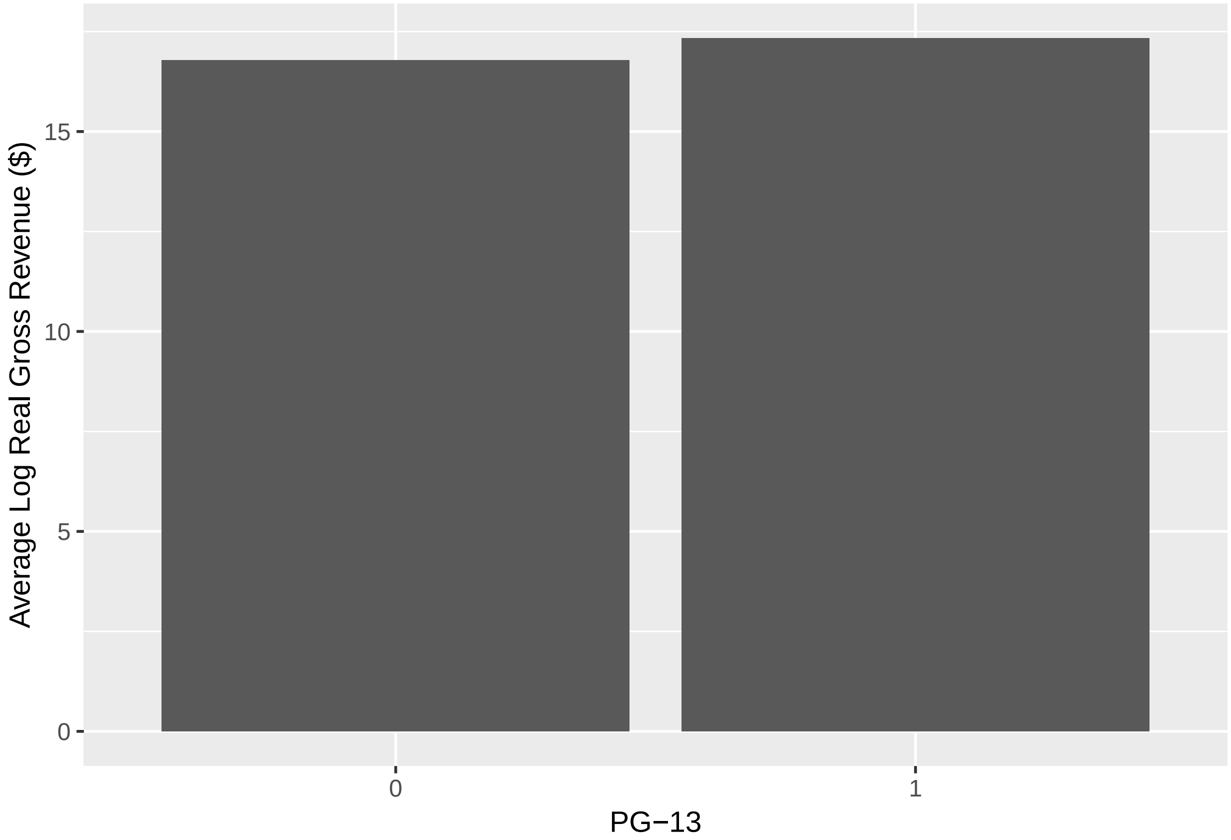


```
##
## [[5]]
```

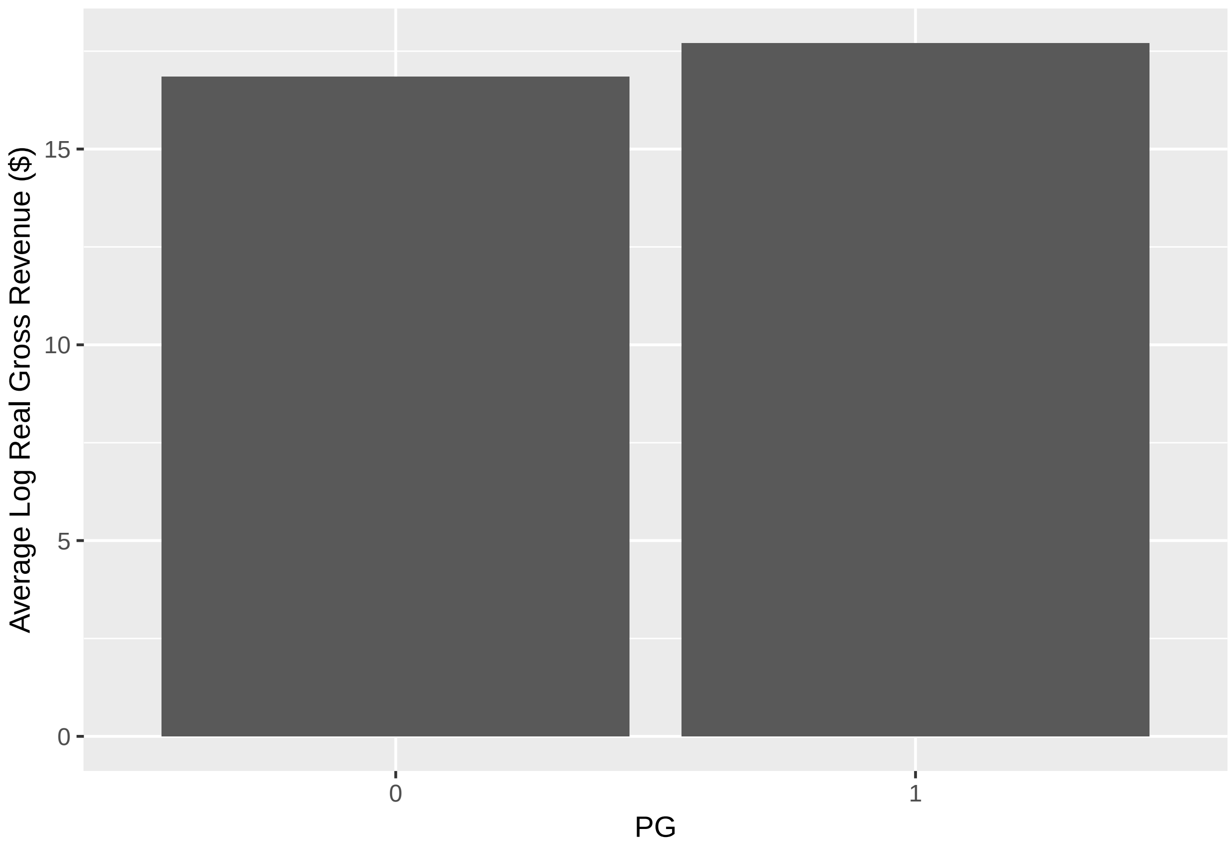


```
lapply(unique(train_content$content_rating), function(r) {
  train_content %>%
    # get 1 if this rating, 0 else. Make factor
    mutate(rating_dum = as.factor(ifelse(content_rating == r, 1, 0))) %>%
    # mean revenue for 0 vs 1 for that content rating
    group_by(rating_dum) %>%
    summarize(avg_real_gross_log = mean(real_gross_log)) %>%
    ggplot() +
    geom_col(aes(x = rating_dum, y = avg_real_gross_log)) +
    labs(x = r, y = 'Average Log Real Gross Revenue ($)')
})
```

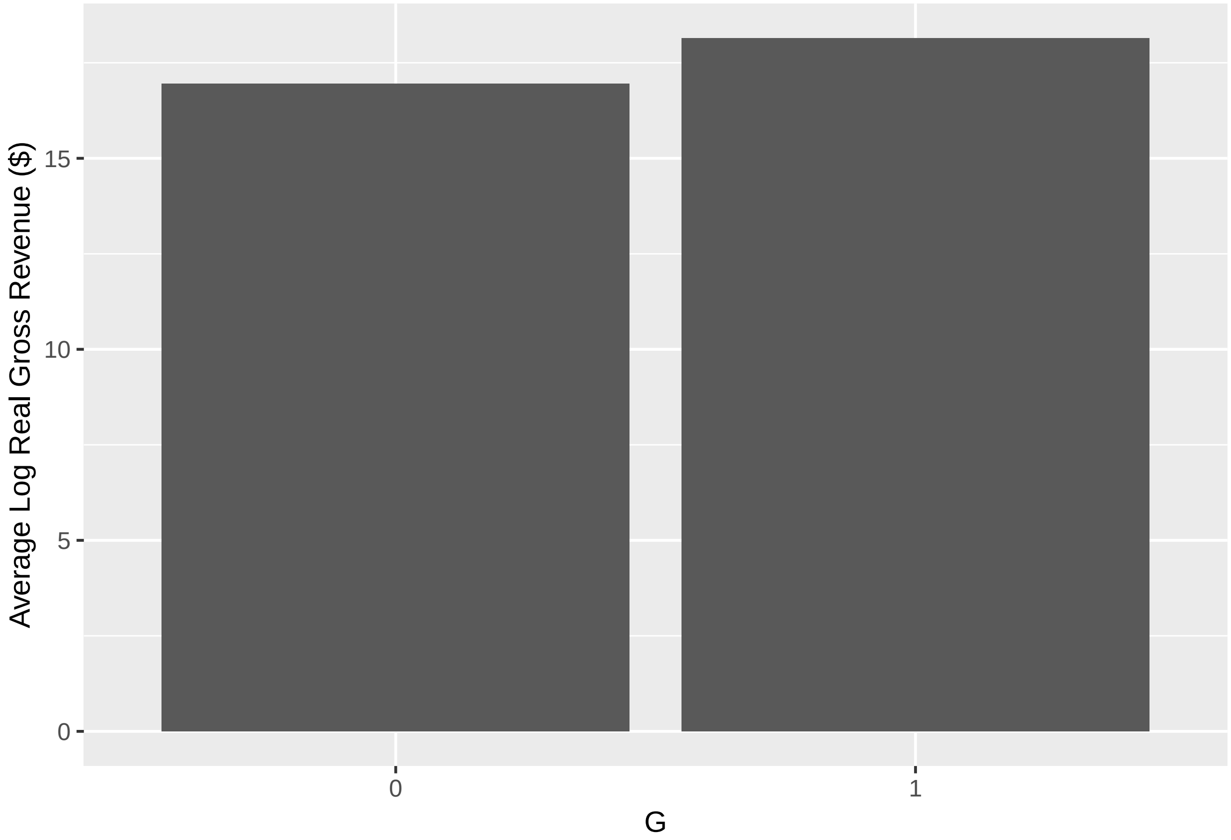
```
## [[1]]
```



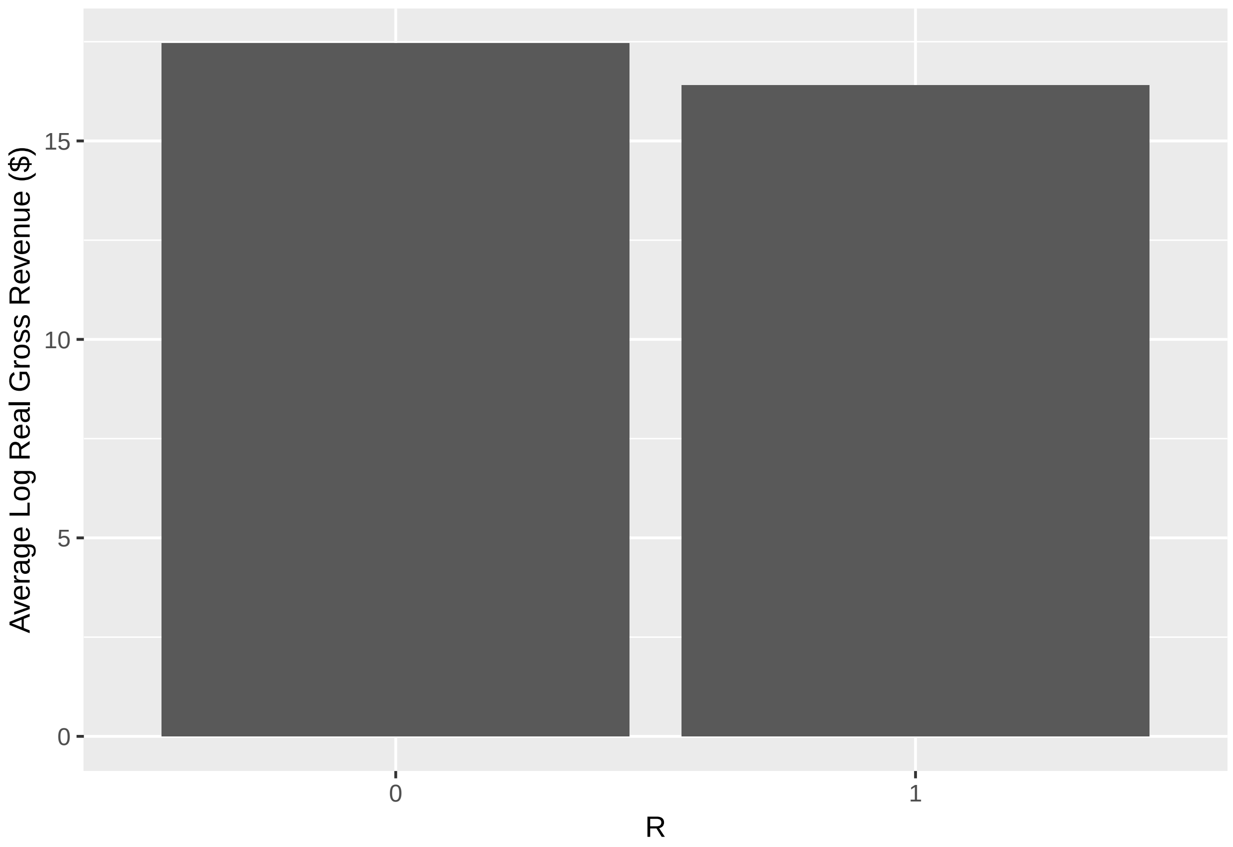
```
##
## [[2]]
```



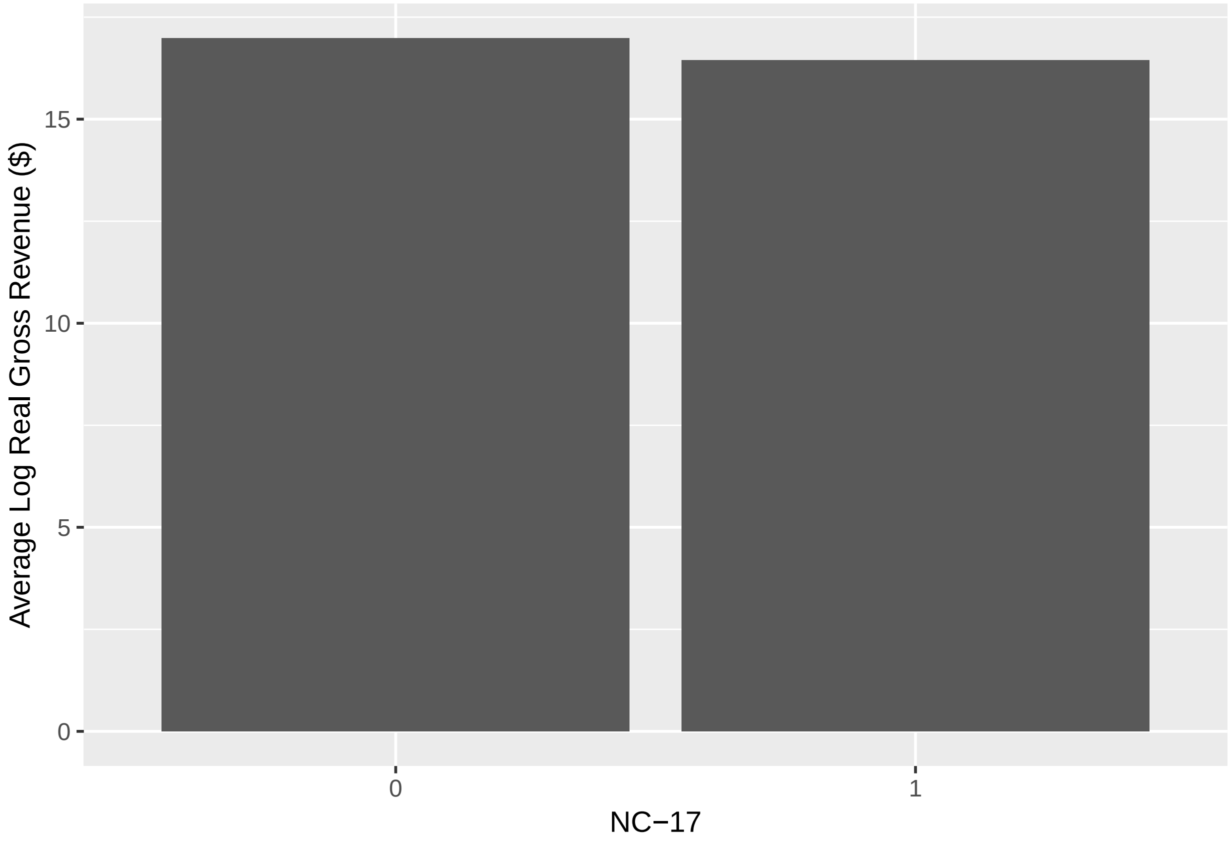
```
##
## [[3]]
```



```
##
## [[4]]
```



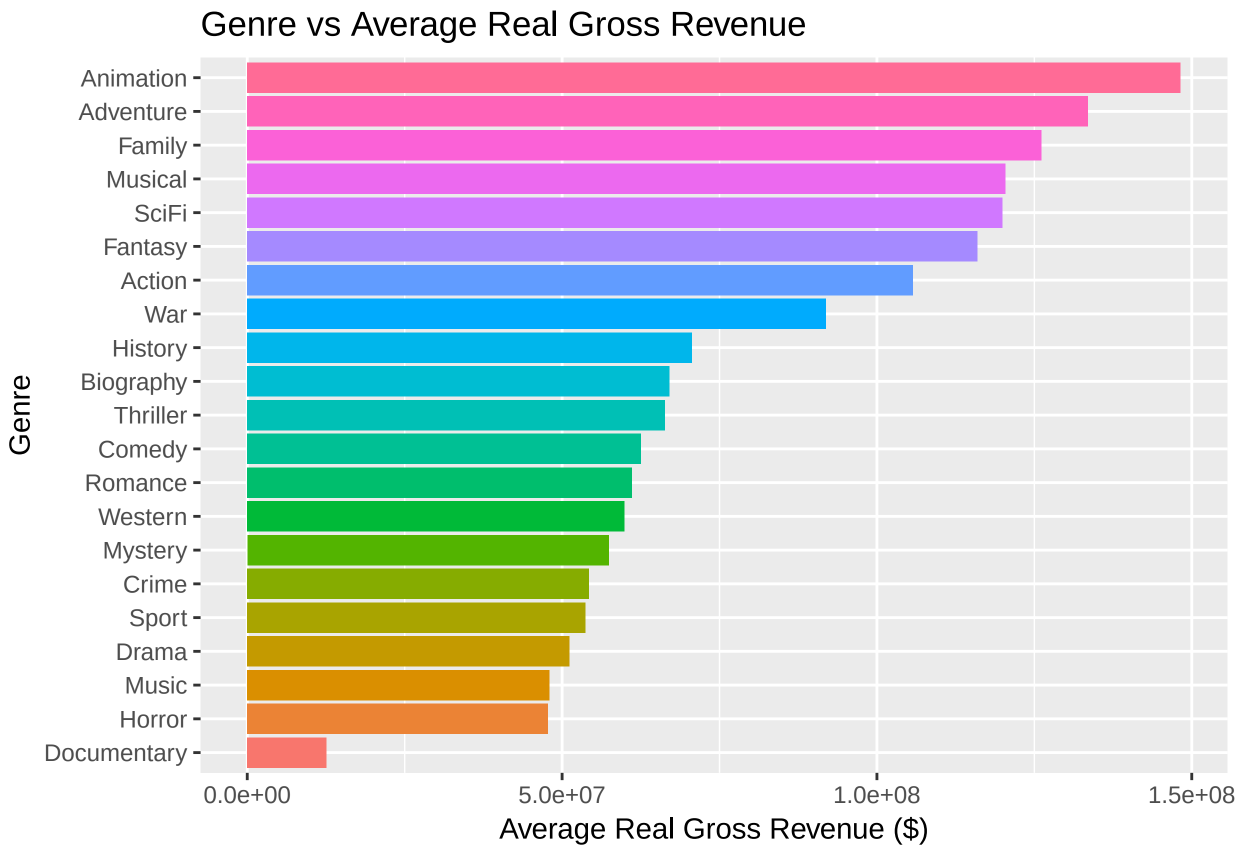
```
##
## [[5]]
```



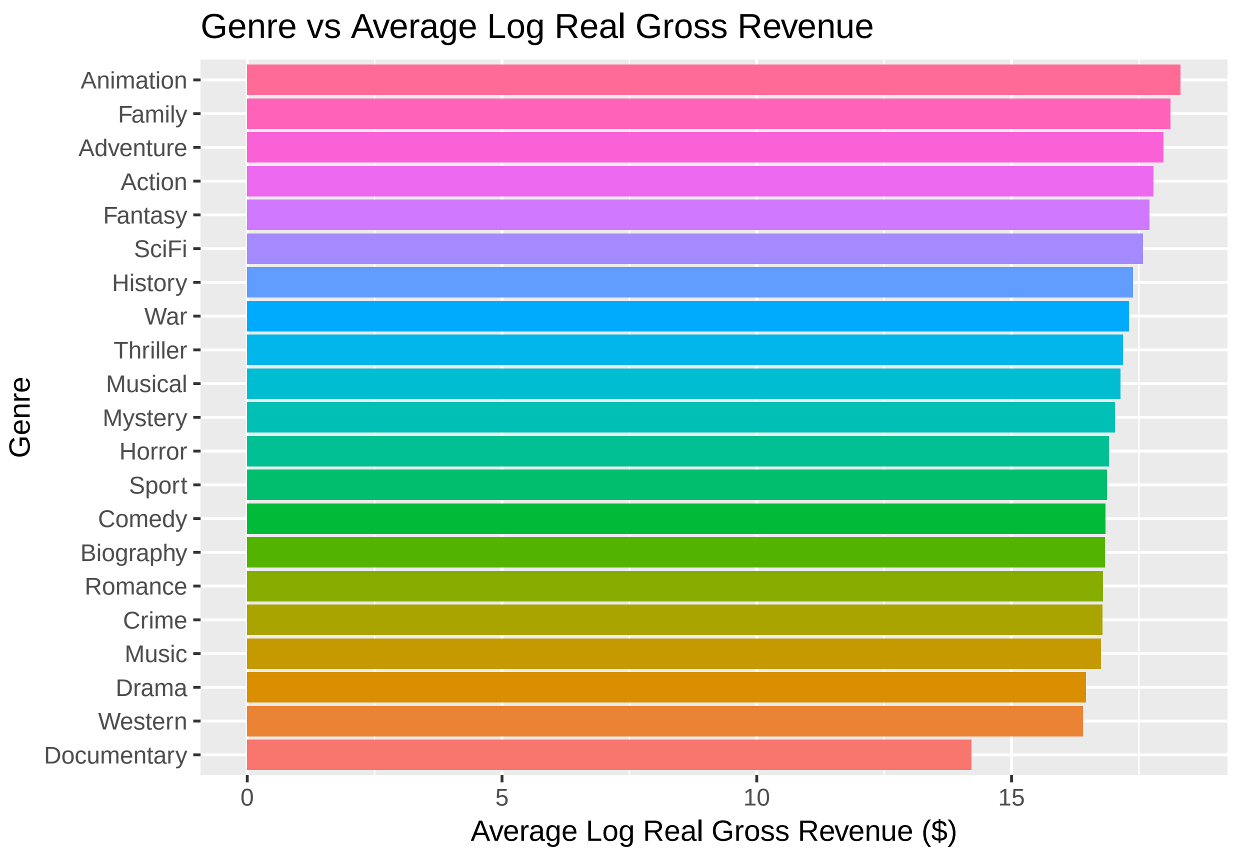
Genre

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.

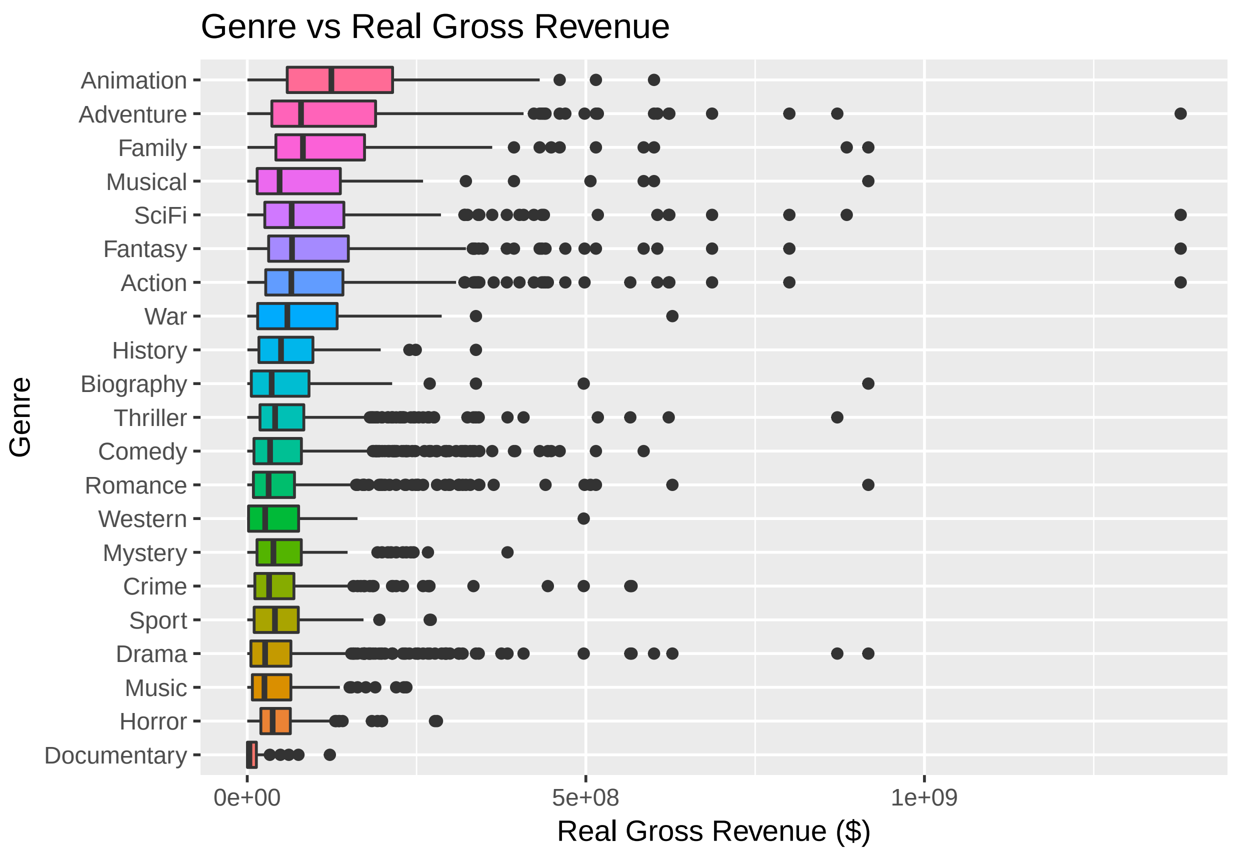
```
# untidy the genre data such that one observation is spread across many rows. Easier to graph
genre_cols <- c('Action', 'Adventure', 'Animation', 'Biography', 'Comedy', 'Crime', 'Documentary',
                'Drama', 'Family', 'Fantasy', 'History', 'Horror', 'Music', 'Musical', 'Mystery',
                '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) %>%
  # reorder genre by real_gross
  mutate(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')
```



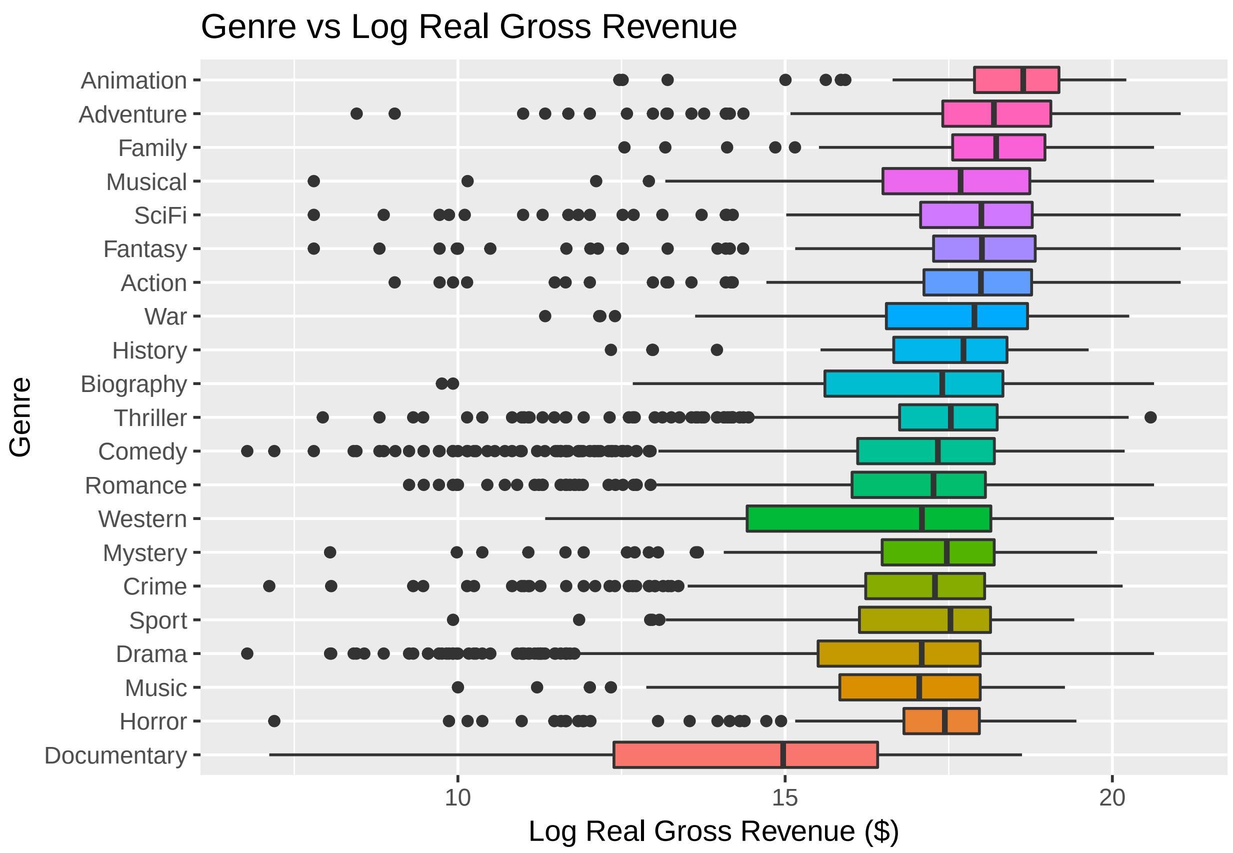
```
train_genre %>%
  # reorder factor by log
  mutate(genre = reorder(genre, real_gross_log)) %>%
  # average by genre
  group_by(genre) %>%
  summarize(avg_real_gross_log = mean(real_gross_log)) %>%
  # graph
  ggplot() +
  geom_col(aes(x = genre, y = avg_real_gross_log, fill = genre)) +
  coord_flip() +
  labs(x = 'Genre', y = 'Average Log Real Gross Revenue ($)',
       title = 'Genre vs Average Log 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 Real Gross Revenue') +
  theme(legend.position = 'none')
```



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


Genre vs Log Real Gross Revenue
Animation
Adventure
Family
Musical
SciFi
Fantasy
Action
War
History
Biography
Thriller
Comedy
Romance
Western
Mystery
Crime
Sport
Drama
Music
Horror
Documentary
Genre
10
15
20
Log Real Gross Revenue ($)