EDA

Katrina Truebebach

March 12, 2019

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

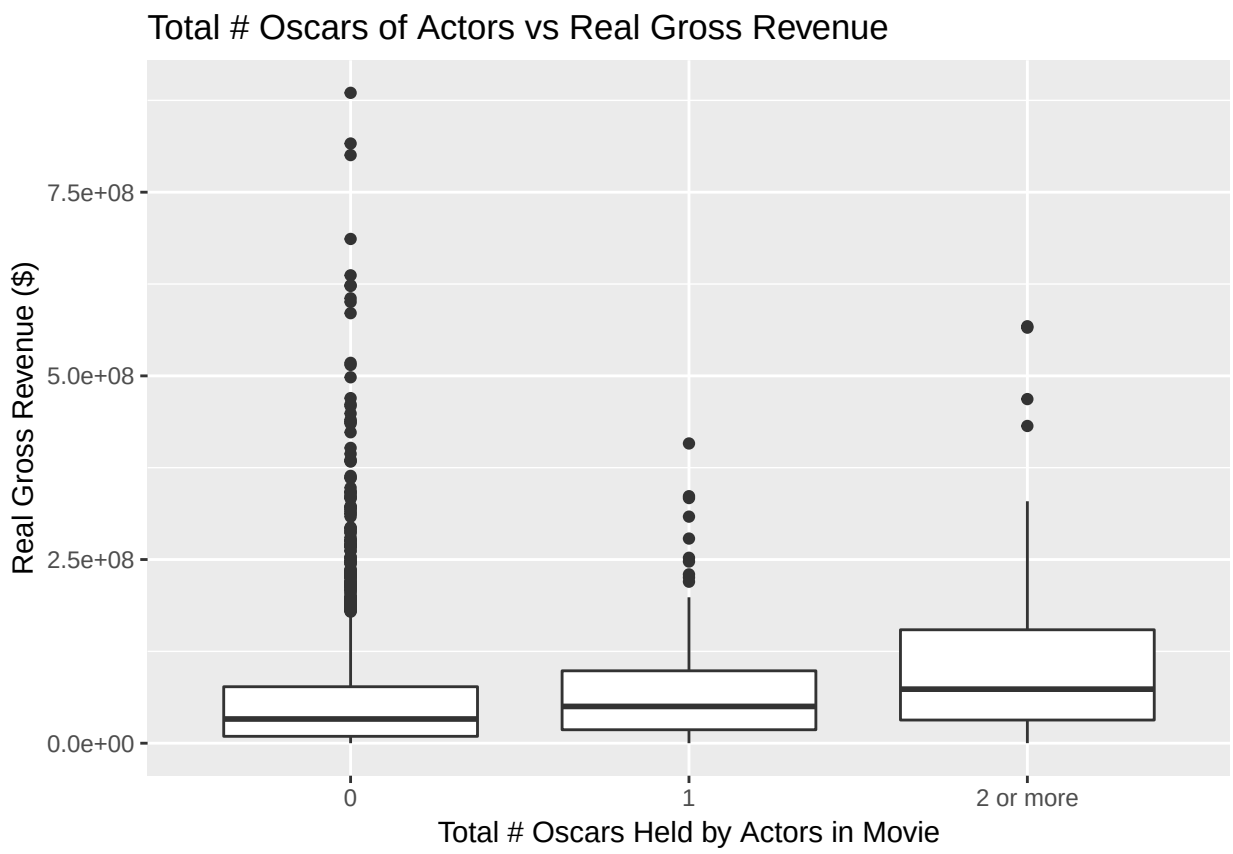
Load cleaned data

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

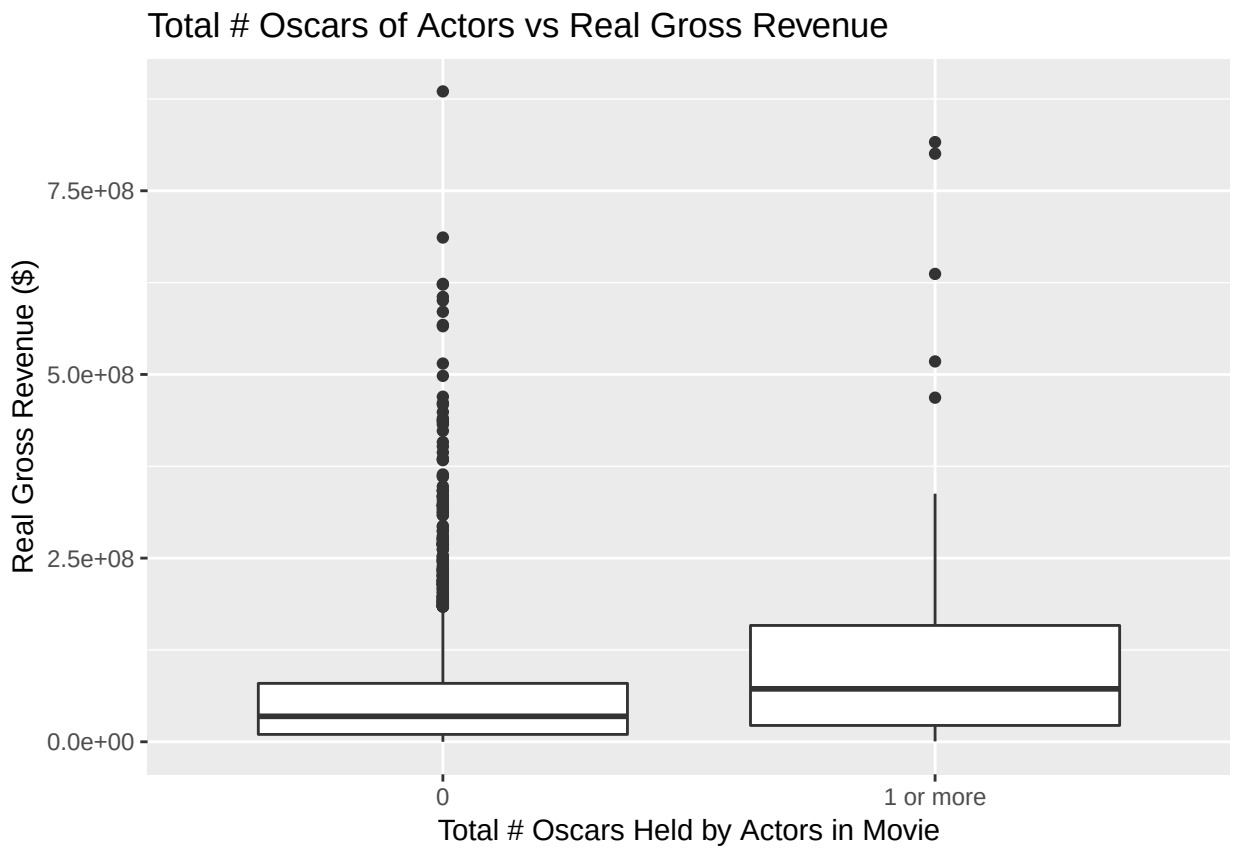
Graph number of Oscars for actors and directors against
Both are linear but very weak. Unclear if should include in model

```
# Function to graph number of Oscars held by actors in movie vs. real revenue
oscar_num <- 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 ($)')
}

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



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



Average real revenue vs year
Adeed APPROXIMATE recession shading. Annual data, so hard to do.
Real revenue 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 time



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 ($)')
```