Keeping it PG-13

Steve Hewitt, Martin Lim, Fidelia Nawar, & Austin Sanders

11/30/2021

1. Introduction

1a. Context

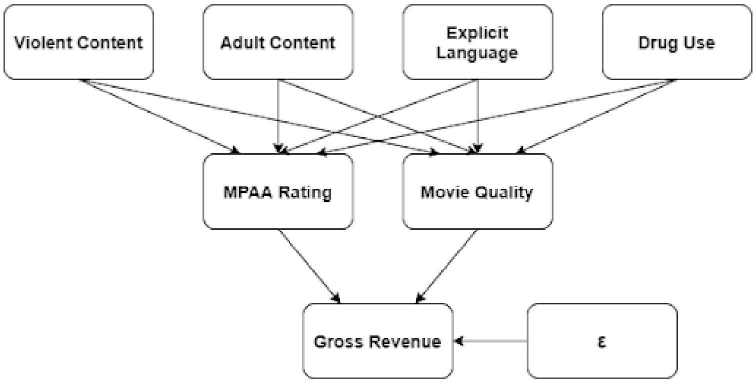
Acme Studios has spent a $50,000,000.00 budget on a superhero movie, and the director insists that the movie should include a scene where the main villain goes on an expletive-laden tirade. We know that including this scene will mean that the movie will be rated R, and cutting the scene will result in the movie being rated PG-13. The director is extremely upset that we want to cut the scene and says we're ruining the film's artistic integrity by trying to make editorial changes after the director's cut. So upset that he went directly to the studio head to complain. Now Acme Studios' executive team has to decide: do they modify the movie for a more family-friendly rating, or do they respect the director's wishes and release it as-is? As data scientists, we would have difficulty quantifying artistic integrity or the value of the relationship between the studio and the director. Still, we feel strongly that we can show the relationship between worldwide revenue for a PG-13 vs. R ratings (holding all other variables constant). The studio head wants to know: How much more money do they expect to make by defying the director's wishes and cutting the movie to make it PG-13?

1b. Research Question

Holding other factors constant, how much more money should a movie studio expect to make on a film that gets a PG-13 rating instead of an R rating from the MPAA?

Our research question intends to measure the impact *MPAA Rating* and *Movie Quality* on the *Gross Revenue* that it will generate. Given other available, quantifiable factors like the *Budget*, *Genre*, and *Runtime*, this study also intends to investigate if they affect the *MPAA Rating* and/or *Movie Quality* which in-turn affect the *Gross Revenue.*

1c. Causal Theory



Our research question seeks to measure the impact of the *MPAA Rating* (more specifically, PG-13 versus R) on the *Gross box office revenue.* A movie typically receives an R rating by the MPAA for some combination of violent content, adult content, explicit language, and drug use. In addition to these factors contributing to the MPAA rating, they also contribute to the quality of the movie. We expect to show that both MPAA rating and movie quality impact the gross revenue of the film. By adjusting the movie's content to secure the desired rating, we may also affect the quality of the movie. Therefore, this study will explore models that include a proxy for movie quality to attempt to minimize omitted variable bias.

2. Research Design and Data

After you have presented the introduction and the concepts that are under investigation, what data are you going to use to answer the questions? What type of research design are you using? What type of models are you going to estimate, and what goals do you have for these models?

2a. Data Source

The data used for this study is a movie dataset from Kaggle (https://www.kaggle.com/danielgrijalvas/movies). It contains 7512 unique movie titles ranging from the year 1986 to 2020. According to the description of the creator of the dataset, the data was scraped from IMDb.com so the extent of the movie title coverage can go as far as the available information posted on the IMBd website. Below are the important columns that were considered for this study:

- Outcome Variable
  - `gross` : Revenue of the movie in USD

- Explanatory Variables
  - `score` : IMDb user rating
  - `rating` : MPAA rating of the movie (R, PG, etc.)
  - `runtime` : Length of movie in minutes
  - `budget` : Budget of the movie in USD
  - `votes` : Number of user votes on the IMBd website

2b. Research Design

Using the data and variables above, this study measured the impact *MPAA Rating* and *Movie Quality* on the *Gross Revenue* that a movie will generate. Causal models were generated using the logic from the causal theory in 1c. `gross` is the main quantified success variable. According to the causal theory, two main explanatory quantities were included in the model. `rating` was used to quantify the MPAA rating of the film while `score` is the main quantified measurement of quality. `buzdget` and `runtime` both affect the the movie quality, but not the `rating` so they were considered as proxies for quality.

The outcome of interest is `gross`, the revenue of the movie in USD. This is the main outcome variable of all the causal models and is quantified in numeric form as is.

In order to properly quantify `rating` in the model, indicator variables were used. The data was divided into three (3) categories: G and PG, PG-13, and R and above. Since movies that are rated G and PG cater to similar audiences, they were classified as the base case and were given a value of 0 for all rating indicator variables. All movies rated PG-13 were given a separate indicator variable (`PG13`) while all movies rated R and more extreme were classified for a different indicator variable (`R`).

Measuring movie quality is primarily measured by the `score` variable. Variables like `budget` and `runtime` also have an effect on quality so they were considered as possible proxies for quality in the causal models.

In order to remove duplicate movie titles, the data point with the larger budget was retained. If both the budgets and titles were equal, the data point with the larger revenue was retained.

Based on our EDA, it seems likely that a log transformation of both gross revenue and budget may be helpful to produce a better-fitting linear model. We will divide our dataset into two parts: 30% for model building and 70% for model validation. We will use the first set portion of the dataset to produce multiple linear models and evaluate them using coefficient tests in R. We will then conduct a comparison using stargazer to determine which model best aligns with our data and our causal theory. Armed with this chosen model, we will evaluate how well its predictions stand up when applied to the second set of data that we saved for validation. Finally, we will apply our model to the specific case outlined in the overview section above to predict the revenue for both a PG-13 and R MPAA rating case.

2c. Data Cleaning

```
# Retrieve only the needed columns
df_raw = select(data, c('title', 'gross', 'budget', 'score', 'runtime', 'rating', 'year', 'votes'))

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

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

# Keep G, PG, PG-13, R, X
if (INCLUDE_X_FILMS) {
  df_raw <- subset(df_raw,
               df_raw$rating == "G"
             | df_raw$rating == "PG"
             | df_raw$rating == "PG-13"
             | df_raw$rating == "R"
             | df_raw$rating == "X")
} else {
  df_raw <- subset(df_raw,
               df_raw$rating == "G"
             | df_raw$rating == "PG"
             | df_raw$rating == "PG-13"
             | df_raw$rating == "R")
}

# Remove Duplicate Movie Titles
# Hash object 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,'title']

  if (TRUE == all(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,'title']) {
        # Replace row if the budget of the new value is higher than that of the
        # budget of the recorded title
        if (df_raw[i, 'budget'] > df[k, 'budget']) {

          # Delete found row in cleaned dataframe
          df = df[-c(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,'budget']
        } else if (df_raw[i, 'budget'] > df[k, 'budget']
                   & df_raw[i, 'gross'] > df[k, 'gross']) {
          # Delete found row in cleaned dataframe
          df = df[-c(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,'budget']
        }
        break
      }

    }

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

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

# Apply Indicator variable for PG-13 (G and PG is the base case)
df$PG13 = factor(ifelse(df$rating == "PG-13" , 1, 0))

# Apply Indicator variable for R and X (G and PG is the base case)
df$R = factor(ifelse(df$rating == "R" | df$rating == "X", 1, 0))
```

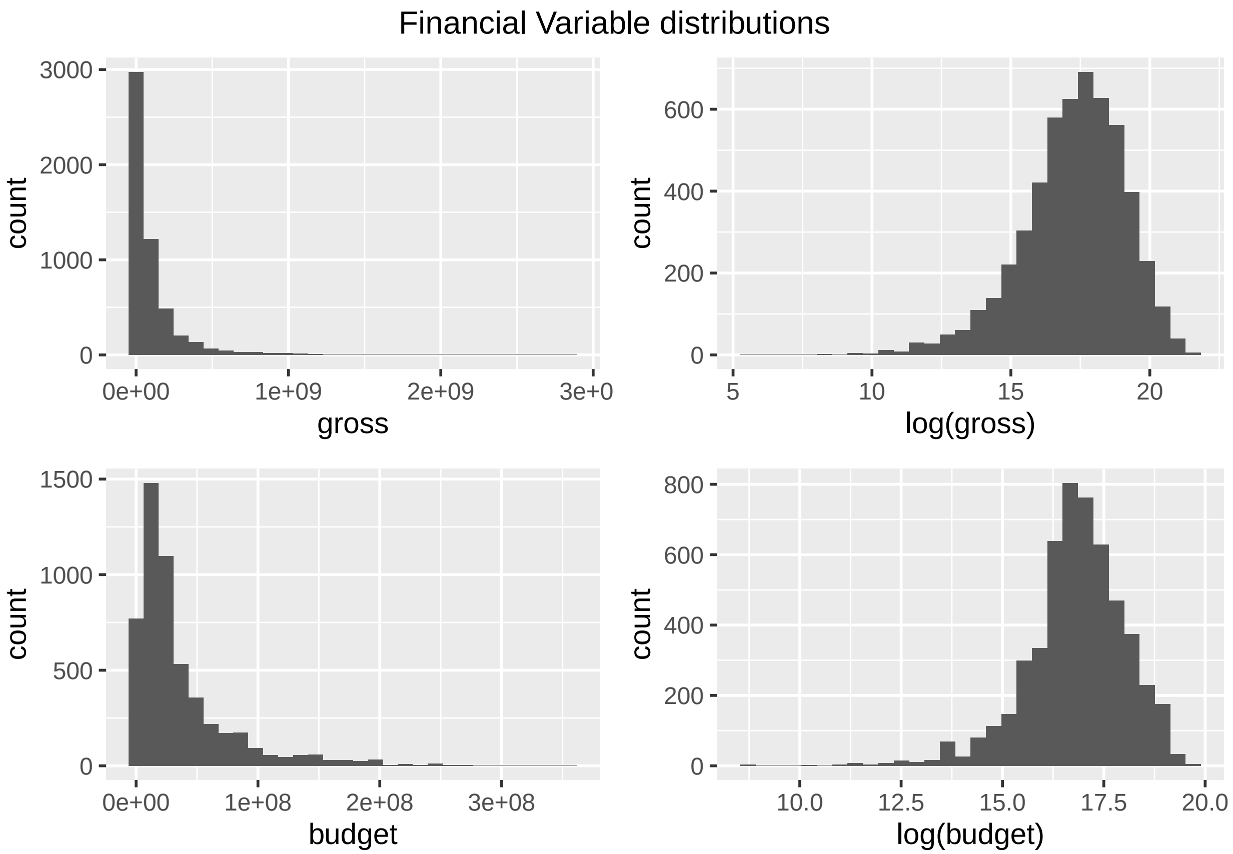
2d. Exploratory Data Analysis

```
nrow(df)
```

```
## [1] 5264
```

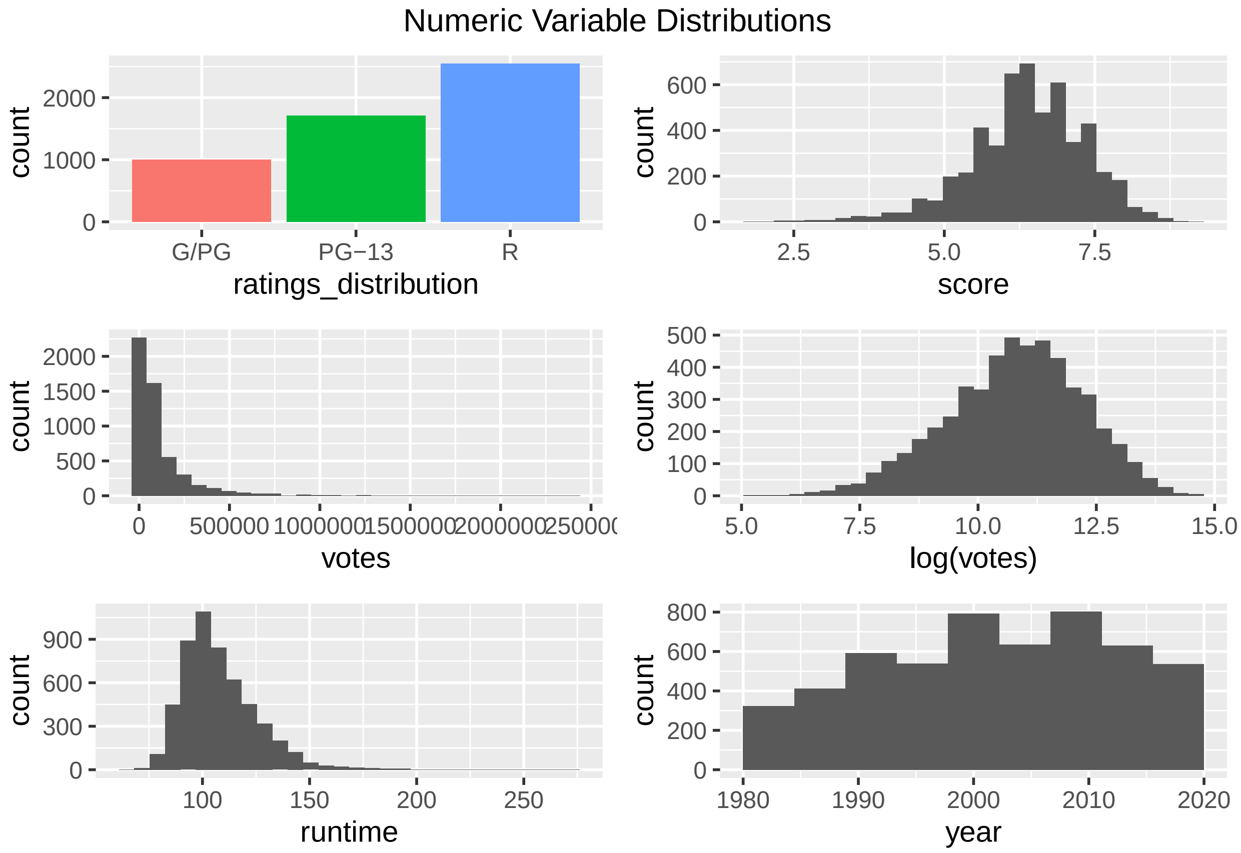
There are 5264 unique titles considered for this study.

```
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, top = "Financial Variable distributions")
```



```
# CHECK OTHER EXPLANATORY VARIABLES

df$ratings_distribution = factor(ifelse(df$rating == "R", "R", ifelse(df$rating == "PG-13", "PG-13", "G

ratings_distribution_histogram <- df %>%
  ggplot(aes(x=ratings_distribution, fill=ratings_distribution)) +
  geom_bar() +
  theme(legend.position = "none")

score_histogram <- df %>%
  ggplot(aes(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)

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

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

grid.arrange(ratings_distribution_histogram, score_histogram,
             votes_histogram, log_votes_histogram,
             runtime_histogram, year_histogram,
             nrow = 3, ncol = 2, top = "Numeric Variable Distributions")
```



```
# Correlation between log(gross) and year - weak correlation at best
cor(log(df$gross), df$year, method = "pearson")
```

```
## [1] 0.3280518
```

```
cor(log(df$gross), df$year, method = "spearman")
```

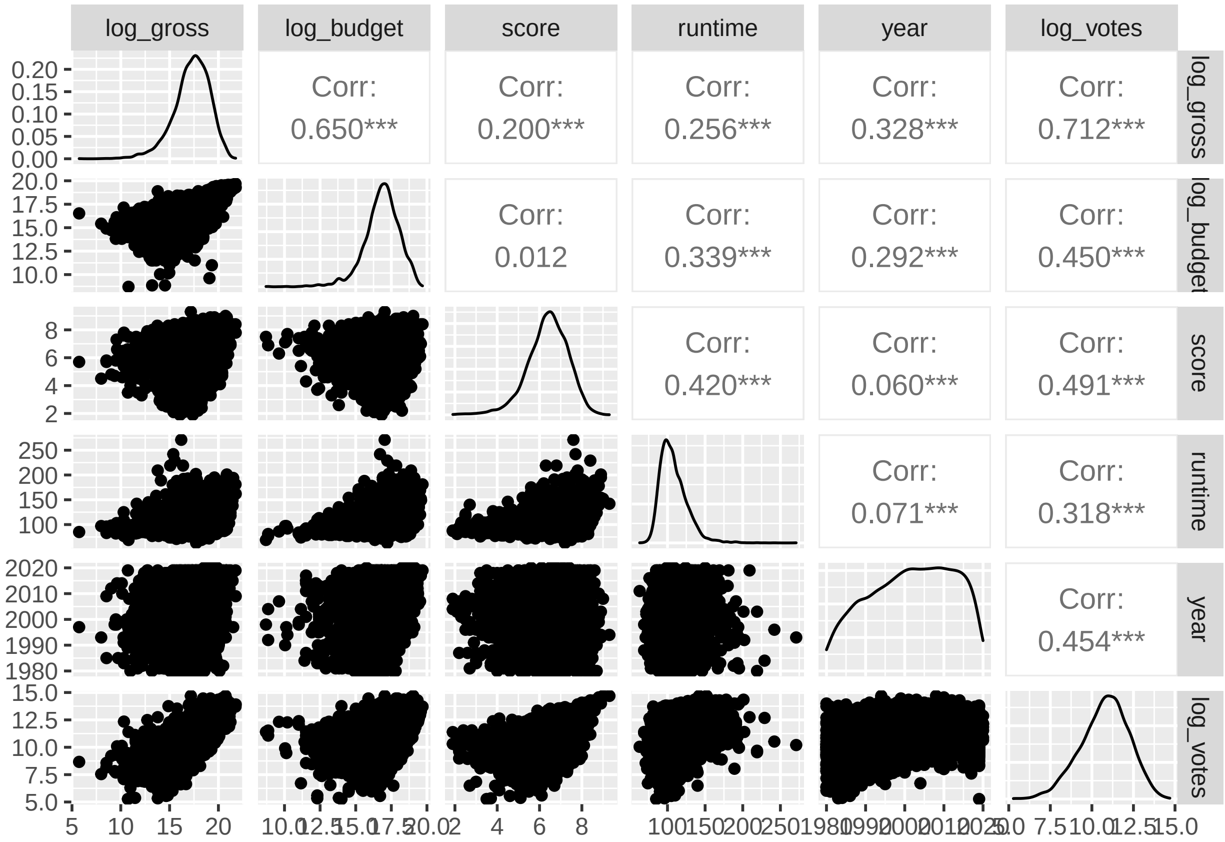
```
## [1] 0.3592937
```

```
cor(log(df$gross), df$year, method = "kendall")
```

```
## [1] 0.2457766
```

There doesn't seem to be a strong correlation between the year and revenue of the movie. So it might not be necessary to adjust the `gross` data for inflation.

```
df_collinearity = select(df, c('gross', 'budget', 'score', 'runtime', 'year', 'votes'))

# Log-Transform `gross`, `budget`, and `votes`
df_collinearity$log_gross = log(df_collinearity$gross)
df_collinearity$log_budget = log(df_collinearity$budget)
df_collinearity$log_votes = log(df_collinearity$votes)

# Remove `gross`, `budget`, and `votes`
df_collinearity = select(df_collinearity, c('log_gross', 'log_budget', 'score', 'runtime', 'year', 'log_

# check for collinearity
ggpairs(df_collinearity)
```



According to the correlation plots above, the explanatory variables do not seem to have any strong correlation to each other. The log-transformed data for `budget` and `votes` may have decent correlation with the log-transformed data for `gross` (the outcome variable), but this should not produce any problems for finding an BLP for this data.

3. Analysis

3a. A Model Building Process

You will next build a set of models to investigate your research question, documenting your decisions. Here are some things to keep in mind during your model building process:

1. *What do you want to measure*? Make sure you identify one, or a few, variables that will allow you to derive conclusions relevant to your research question, and include those variables in all model specifications. How are the variables that you will be modeling distributed? Provide enough context and information about your data for your audience to understand whatever model results you will eventually present.
2. What covariates help you achieve your modeling goals? Are there problematic covariates? either due to *collinearity*, or because they will absorb some of a causal effect you want to measure?
3. What *transformations*, if any, should you apply to each variable? These transformations might reveal linearities in the data, make our results relevant, or help us meet model assumptions.
4. Are your choices supported by exploratory data analysis (*EDA*)? You will likely start with some general EDA to *detect anomalies* (missing values, top-coded variables, etc.). From then on, your EDA should be interspersed with your model building. Use visual tools to *guide* your decisions. You can also leverage statistical *tests* to help assess whether variables, or groups of variables, are improving model fit.

At the same time, it is important to remember that you are not trying to create one perfect model. You will create several specifications, giving the reader a sense of how robust (or sensitive) your results are to modeling choices, and to show that you're not just cherry-picking the specification that leads to the largest effects.

At a minimum, you need to estimate at least three model specifications:

The first model you include should include *only the key variables* you want to measure. These variables might be transformed, as determined by your EDA, but the model should include the absolute minimum number of covariates (usually zero or one covariate that is so crucial it would be unreasonable to omit it).

Additional models should each be defensible, and should continue to tell the story of how product features contribute to product success. This might mean including additional right-hand side features to remove omitted variable bias identified by your casual theory; or, instead, it might mean estimating a model that examines a related concept of success, or a model that investigates a heterogeneous effect. These models, and your modeling process should be defensible, incremental, and clearly explained at all points.

Your goal is to choose models that encircle the space of reasonable modeling choices, and to give an overall understanding of how these choices impact results.

Proposed Models:

(1)

$$ln(Gross) = \beta_0 + \beta_1 * PG13 + \beta_2 * R$$

(2)

$$ln(Gross) = \beta_0 + \beta_1 * PG13 + \beta_2 * R + \beta 3 * Score$$

(3)

$$ln(Gross) = \beta_0 + \beta_1 * PG13 + \beta_2 * R + \beta 3 * Score + \beta 4 * PG13 * Score + \beta 5 * R * Score$$

(4)

$$ln(Gross) = \beta_0 + \beta_1 * PG13 + \beta_2 * R + \beta 3 * Score + \beta 4 * ln(Budget)$$

(5)

$$\begin{gathered} ln(Gross) = \beta_0 + \beta_1 * PG13 + \beta_2 * R + \beta 3 * Score + \beta 4 * PG13 * Score + \\ \beta 5 * R * Score + \beta 6 * ln(Budget) + \beta_7 * PG13 * ln(Budget) + \beta_8 * R * ln(Budget) \end{gathered}$$

```r
require(caTools)
set.seed(203)
sample = sample.split(df, SplitRatio = .30)
train = subset(df, sample == TRUE)
test  = subset(df, sample == FALSE)
```

```r
#Model 1: Rating Only
model_1 = lm(log(gross) ~ PG13 + R, data = train)

#Model 2: Rating and Quality
model_2 = lm(log(gross) ~ PG13 + R + score, data = train)

#Model 3: Rating and Quality with interaction terms

#Model 4: Rating and Quality with budget

#Model 5: Rating and Quality with budget and interaction terms
#model_5 = lm(log(gross) ~ PG13 + R + score + PG13 * score + R * score + log(budget) + log(budget) * PG


stargazer(
  model_1,
  model_2,
  #model_3,
  #model_4,
  #model_5,
  type = 'text', header = FALSE,
  star.cutoffs = c(0.05, 0.01, 0.001) # the default isn't in line with w203
)
```

```
##
## =====================================================================
##                                     Dependent variable:
##                     -------------------------------------------------
##                                         log(gross)
##                               (1)                      (2)
## ---------------------------------------------------------------------
## PG131                        0.085                    0.067
##                             (0.134)                  (0.130)
##
## R1                         -0.962***                -1.031***
##                             (0.127)                  (0.123)
##
## score                                                0.487***
##                                                      (0.048)
##
## Constant                   17.723***                14.647***
##                             (0.107)                  (0.321)
##
## ---------------------------------------------------------------------
## Observations                 1,435                    1,435
## R2                           0.075                    0.137
## Adjusted R2                  0.074                    0.135
## Residual Std. Error    1.785 (df = 1432)        1.725 (df = 1431)
## F Statistic         58.151*** (df = 2; 1432) 75.720*** (df = 3; 1431)
## =====================================================================
## Note:                                   *p<0.05; **p<0.01; ***p<0.001
```

3. A Results Section

You should display all of your model specifications in a regression table, using a package like `stargazer` to format your output. It should be easy for the reader to find the coefficients that represent key effects near the top of the regression table, and scan horizontally to see how they change from specification to specification. Make sure that you display the most appropriate standard errors in your table.

In your text, comment on both *statistical significance and practical significance.* You may want to include statistical tests besides the standard t-tests for regression coefficients. Here, it is important that you make clear to your audience the practical significance of any model results. How should the product change as a result of what you have discovered? Are there limits to how much change you are proposing? What are the most important results that you have discovered, and what are the least important?

4. Limitations of your Model

4a. Statistical limitations of your model

As a team, evaluate all of the large sample model assumptions. However, you do not necessarily want to discuss every assumption in your report. Instead, highlight any assumption that might pose significant problems for your analysis. For any violations that you identify, describe the statistical consequences. If you are able to identify any strategies to mitigate the consequences, explain these strategies.

Note that you may need to change your model specifications in response to violations of the large sample model.

4b. Structural limitations of your model

What are the most important *omitted variables* that you were not able to measure and include in your analysis? For each variable you name, you should *reason about the direction of bias* caused by omitting this variable and whether the omission of this variable calls into question the core results you are reporting. What data could you collect that would resolve any omitted variables bias?

5. Conclusion

Make sure that you end your report with a discussion that distills key insights from your estimates and addresses your research question.