Final Project

Christina Yu, Damian Kim

Read in the data

```r
library(tidyverse)
```

```
-- Attaching packages --------------------------------------- tidyverse 1.3.2 --
v ggplot2 3.4.0      v purrr   1.0.0
v tibble  3.2.1      v dplyr   1.1.1
v tidyr   1.2.1      v stringr 1.5.0
v readr   2.1.3      v forcats 0.5.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
```

```r
library(ggfortify)
library(knitr)
library(broom)
library(patchwork)
library(tidymodels)
```

```
-- Attaching packages --------------------------------------- tidymodels 1.0.0 --
v dials        1.1.0      v rsample      1.1.1
v infer        1.0.4      v tune         1.0.1
v modeldata    1.0.1      v workflows    1.1.2
v parsnip      1.0.3      v workflowsets 1.0.0
v recipes      1.0.3      v yardstick    1.1.0
-- Conflicts ------------------------------------------ tidymodels_conflicts() --
x scales::discard() masks purrr::discard()
x dplyr::filter()   masks stats::filter()
x recipes::fixed()  masks stringr::fixed()
```

```
x dplyr::lag()      masks stats::lag()
x yardstick::spec() masks readr::spec()
x recipes::step()   masks stats::step()
* Use suppressPackageStartupMessages() to eliminate package startup messages
```

```
library(corrplot)
```

```
corrplot 0.92 loaded
```

```
data <- read_csv("data/movies.csv")
```

```
Rows: 7668 Columns: 15
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (9): name, rating, genre, released, director, writer, star, country, com...
dbl (6): year, score, votes, budget, gross, runtime

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Introduction and data

Nowadays, movie industries are definitely one of the most popular things for people, especially movie investors to look at. There are more factors that intervene in this kind of thing, like actors, genres, user ratings and more

This dataset was scraped from IMDb (Internet Movie Database). There are 6820 movies in the dataset (220 movies per year, 1986-2016). Each movie has the following attributes:

– **budget**: the budget of a movie. Some movies don't have this, so it appears as 0

– **company**: the production company

– **director**: the director

– **genre**: main genre of the movie.

– **gross**: revenue of the movie

– **name**: name of the movie

– **rating**: rating of the movie (R, PG, etc.)

- `released`: release date (YYYY-MM-DD)

- `runtime`: duration of the movie

- `score`: IMDb user rating

- `votes`: number of user votes

- `star`: main actor/actress

- `writer`: writer of the movie

- `year`: year of release

We will explore the factors that make a movie successful through examining the effects of `gross`, `budget`, `genre`, `rating`,`score` `votes`, `year`, and `runtime` for individual movie.

Since the dataset includes movies that are not missing or not published yet (therefore their gross values are missing), we filtered the dataset to include only observations with gross values not being NULL, as it does not make sense to include the movies that are not actually performed or lost real data in our analysis. We filter all the null values for the rest of predictor variables as well, since we decided that since we were working with such a large number of movies, that it was reasonable to get rid of movies that were missing some variable, for convenience of model-fitting. Now we have 5423 observations in the dataset.

```
data <- data %>%
  filter(!is.na(gross) & !is.na(budget) & !is.na(genre) & !is.na(rating) & !is.na(score) &
```

The Predictor Variables

We will use `budget`, `genre`, `rating`,`score` `votes`, `year`, and `runtime` as predictors. Among them, `budget`, `score` `votes`, `year`, and `runtime` are numerical variables, while `genre`and `rating` is a categorical variable.

The Response Variable

1. Summary of the `gross` variable:

```
data %>%
  summarise(mean_gross = mean(gross),
            median_gross = median(gross),
            sd_gross = sd(gross),
            min_gross = min(gross),
            max_gross = max(gross)) %>%
```

3

```
kable()
```

| mean_gross	median_gross	sd_gross	min_gross	max_gross
103192280	36850101	187278279	309	2847246203

2. Log-transformation and Distribution of the **gross** variable:

```
data <- data %>%
  mutate(log_gross = log(gross)) %>%
  mutate(mean = mean(log_gross))

p1 <- ggplot(data = data, aes(x = gross))+
  geom_histogram(fill = "light blue", color = "black")+
  labs(title = "Distribution of Movies' Gross",
       x = "Gross")
p2 <- ggplot(data = data, aes(x = log_gross))+
  geom_histogram(fill = "light blue", color = "black")+
  labs(x = "Log (Gross)")
p1/p2
```
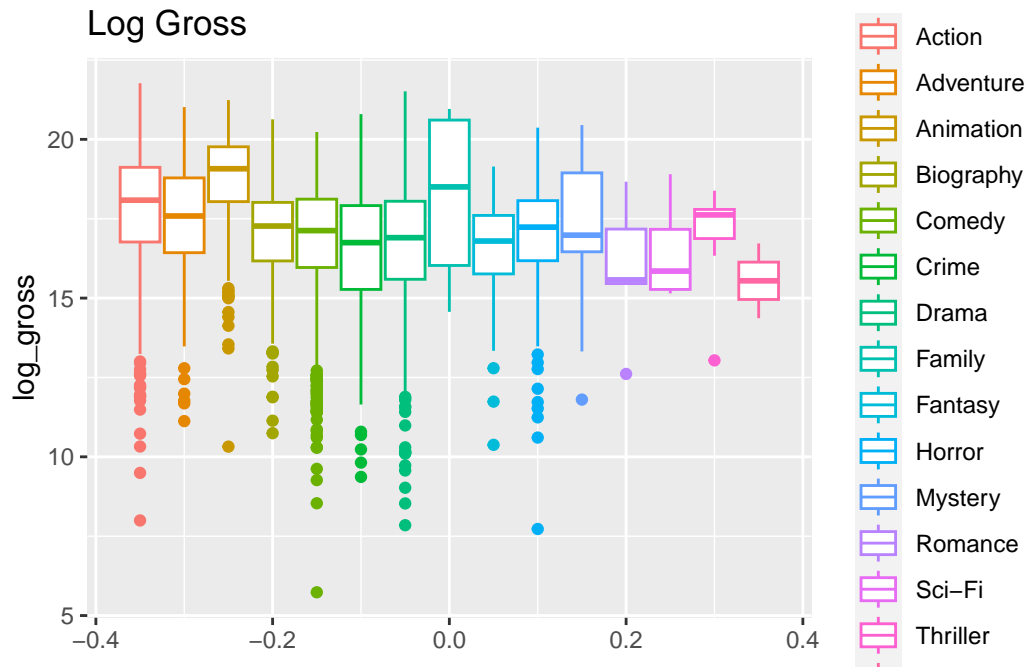
```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

Distribution of Movies' Gross
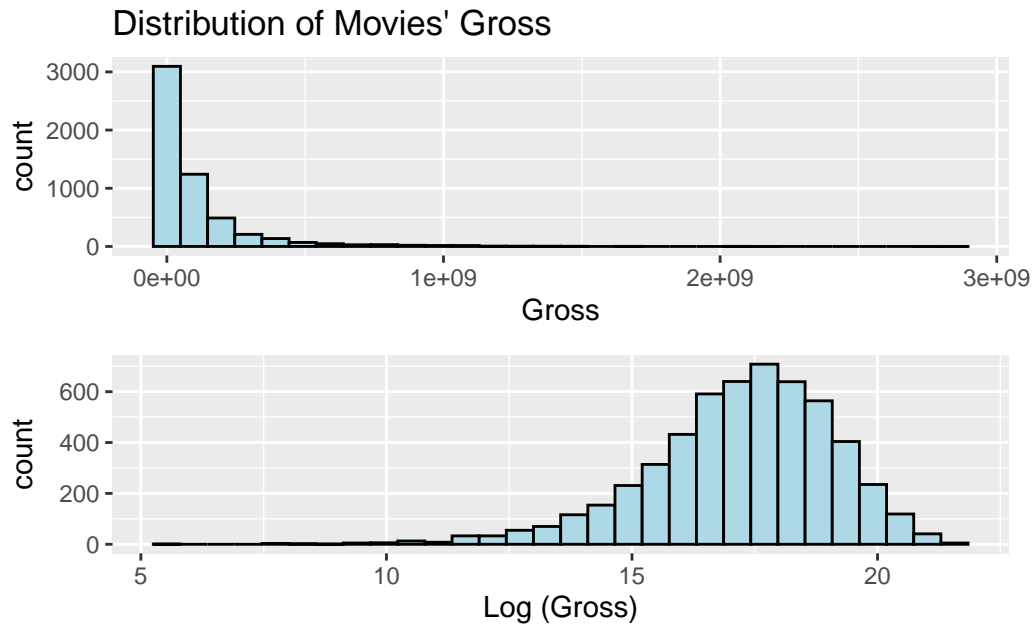

```
ggplot(data = data) +
  geom_boxplot(aes(y = gross, color = genre)) +
  labs(title = "Gross",
       color = "Genre")
```
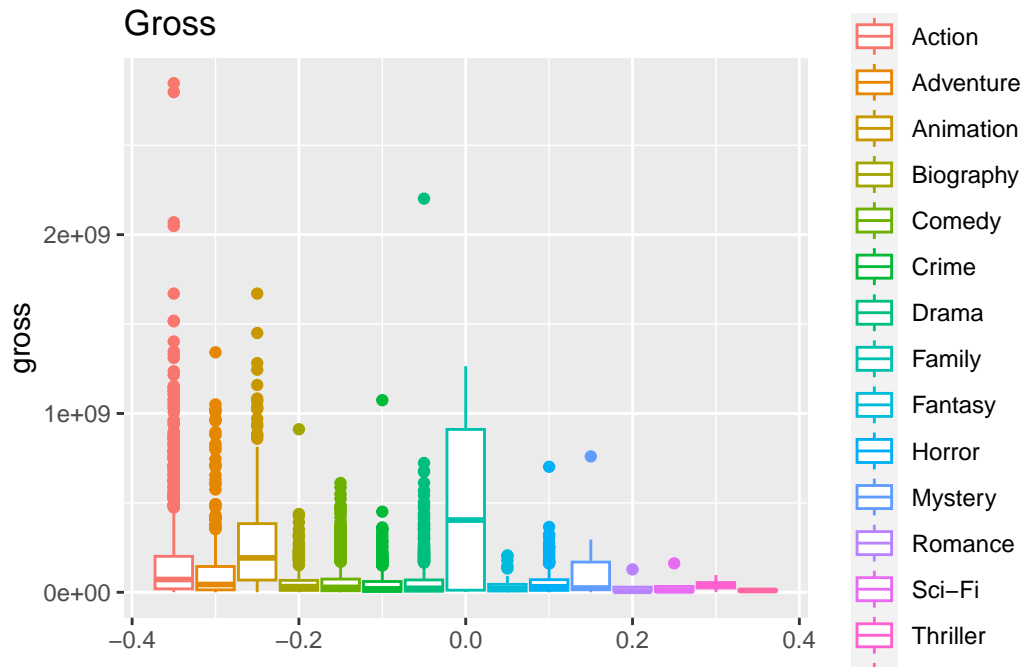
Gross

```
ggplot(data = data) +
  geom_boxplot(aes(y = log_gross, color = genre)) +
  labs(title = "Log Gross",
       color = "Genre")
```

Log Gross

Since the response variable is significantly right skewed, we apply a log-transformation to it and will use log(gross) as our new response variable in the future analysis. Now, our response variable is unimodal, roughly a normal distribution, with the mean at 17.2102, and several outliers at left tail.

3. Relationship between Gross and Budget based on different Genres & Ratings:

```
ggplot(data = data, aes(x=budget, y = log_gross, color = rating))+
  geom_point(size=0.5, fill=NA) +
  geom_smooth(fill=NA) +
  theme(legend.key.size = unit(0.3, "cm")) +
  facet_wrap(~ genre)+
  ggtitle("Relationship between Budget and Gross by Rating across Genres") +
  xlab("Budget") +
  ylab("Log(Gross)")+
  scale_color_discrete(name = "Rating", guide = guide_legend(override.aes = list(size = 1)
  theme(panel.spacing.x = unit(2, "mm"))
```

Relationship between Budget and Gross by Rating across Genre
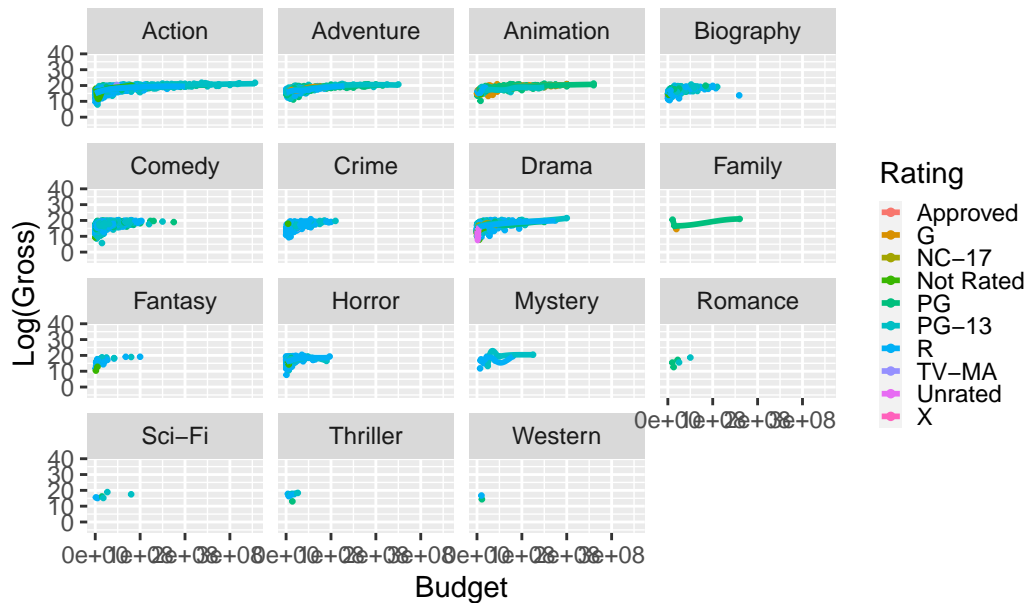
We observe that the relationship between budget and gross is vastly different across genres: Action, adventure, animation movies have a steep slope and generally high budget spans, with outliers which have exceedingly high budget and relatively high gross values. On the other hand, genres such as horror, mystery and romance have a much flatter slope, which corresponds to the industry knowledge that certain genres are more conducive to low-budget film making than others. Therefore, we're interested in further exploring the relationship between budget, genre, and our response variable.

EDA: Visualizatioins and Summary Statistics

Since the dataset contains many interesting variables that we want to explore. We first do some EDA on our datasets to show potential problems that we can explore further into.

```
ggplot(data = data) +
  geom_boxplot(aes(y = year, color = genre)) +
  labs(title = "Year",
       color = "Genre")
```

Year
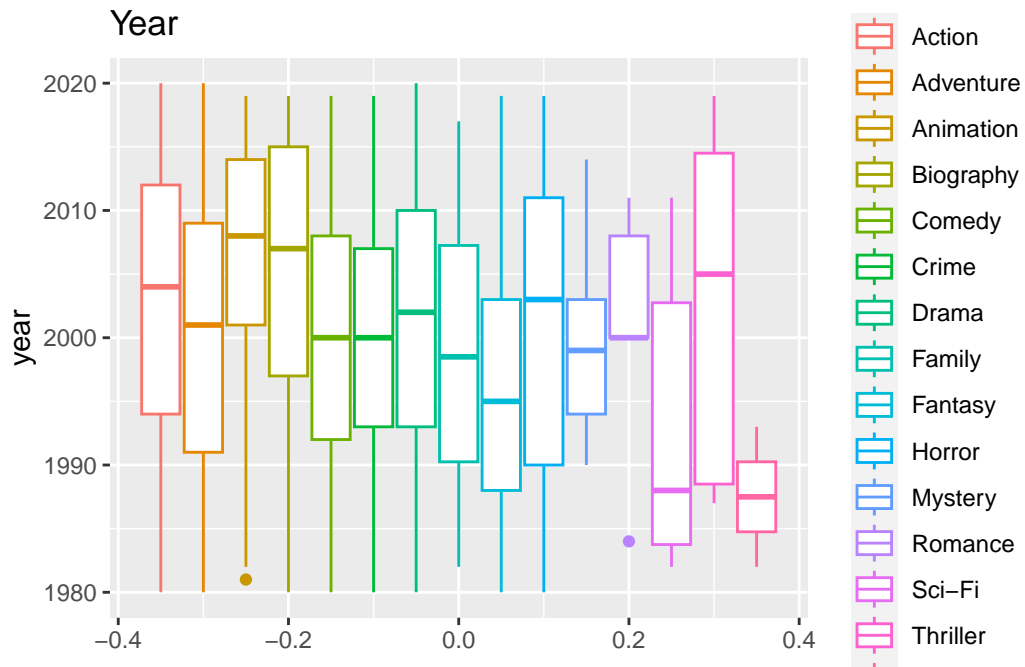

```r
ggplot(data = data) +
  geom_boxplot(aes(y = score, color = genre)) +
  labs(title = "Score",
       color = "Genre")
```

9
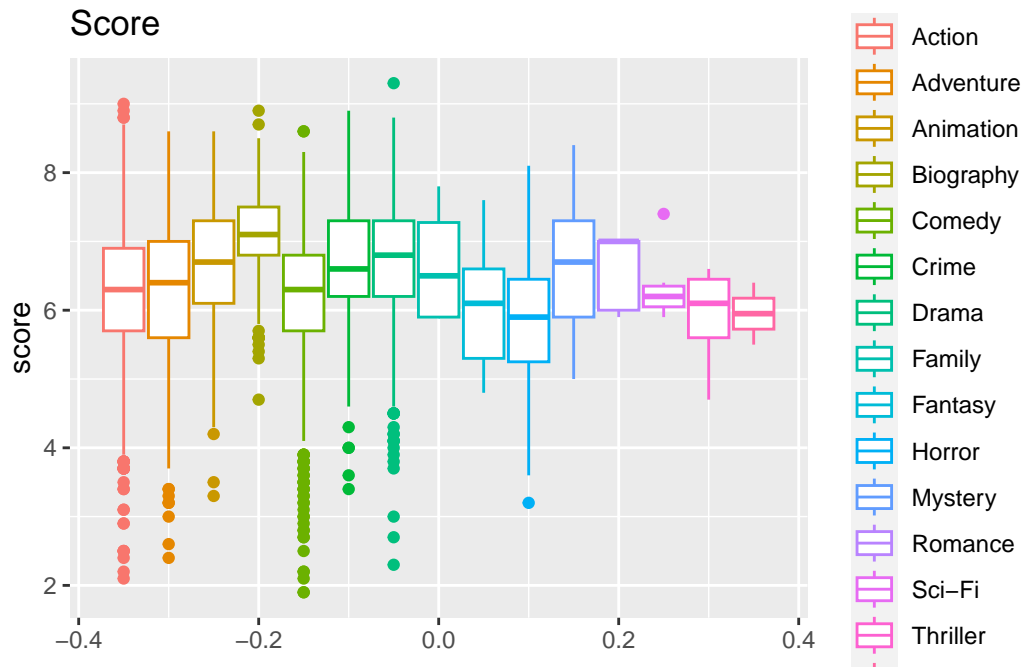
```
ggplot(data = data) +
  geom_boxplot(aes(y = votes, color = genre)) +
  labs(title = "Votes",
       color = "Genre")
```

Votes
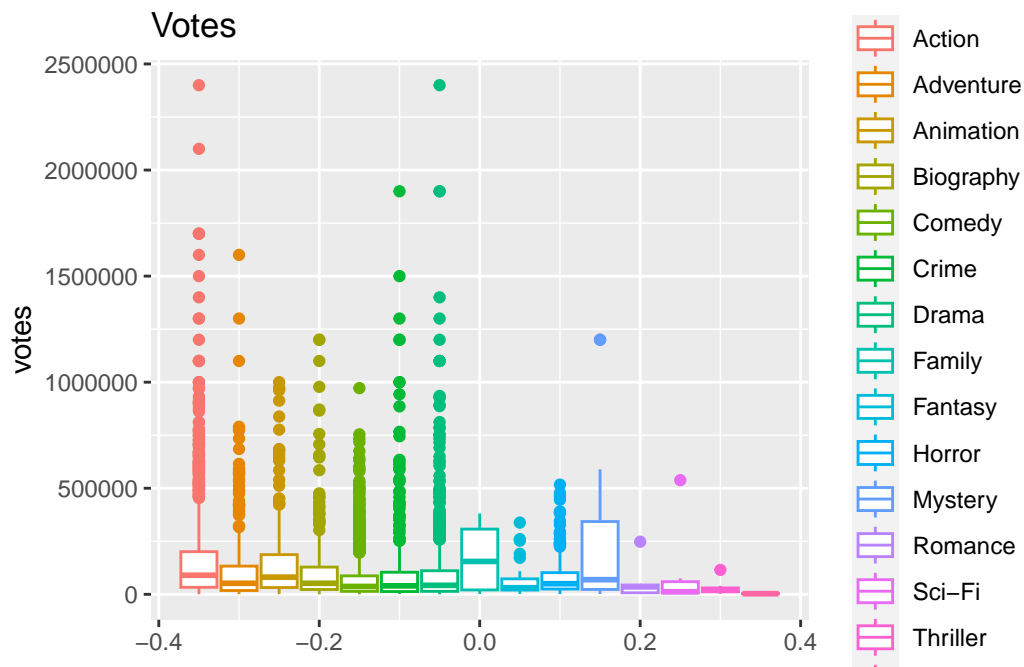
```
ggplot(data = data) +
  geom_boxplot(aes(y = budget, color = genre)) +
  labs(title = "Budget",
       color = "Genre")
```

Budget
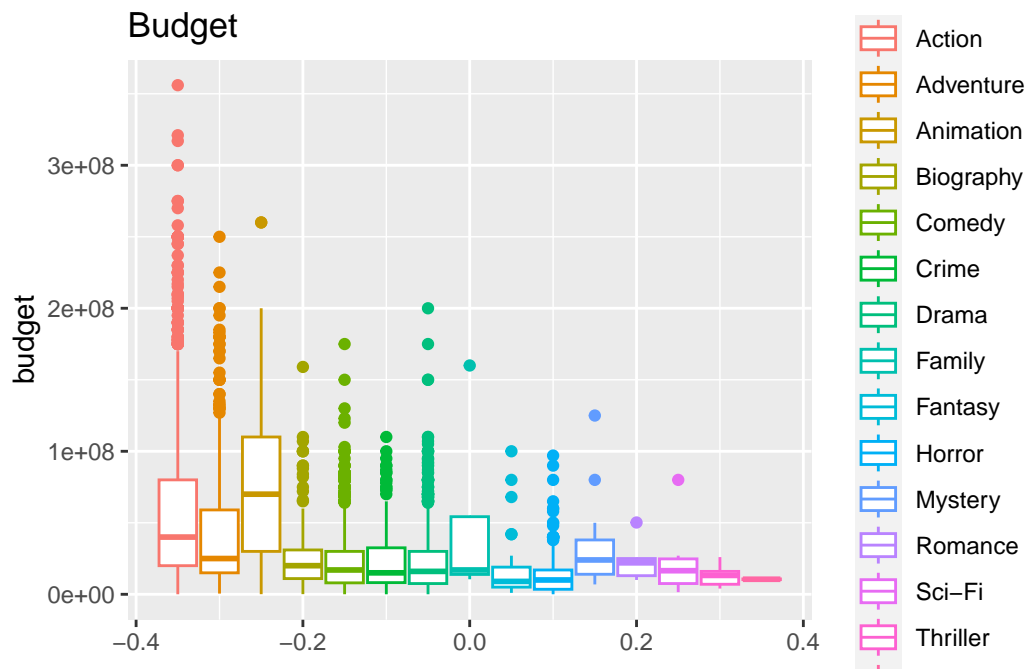

```
ggplot(data = data) +
  geom_boxplot(aes(y = runtime, color = genre)) +
  labs(title = "Runtimes",
       color = "Genre")
```
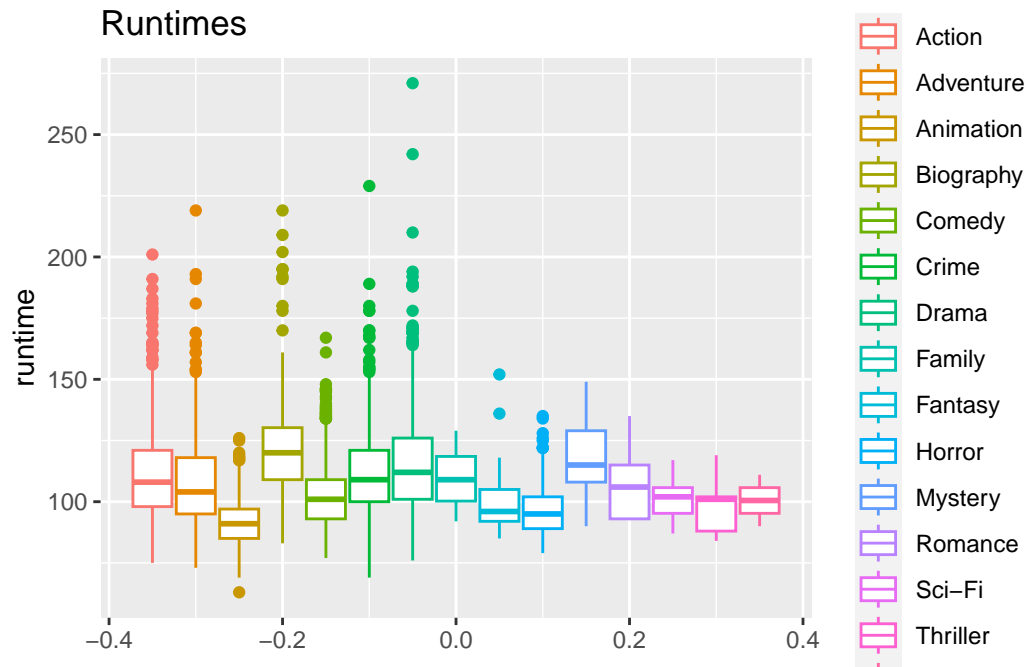
Runtimes


```
ggplot(data = data, aes(x = runtime, y = gross)) +
  geom_point() +
  labs(title = "Runtime has no relationship with Gross")
```

Runtime has no relationship with Gross
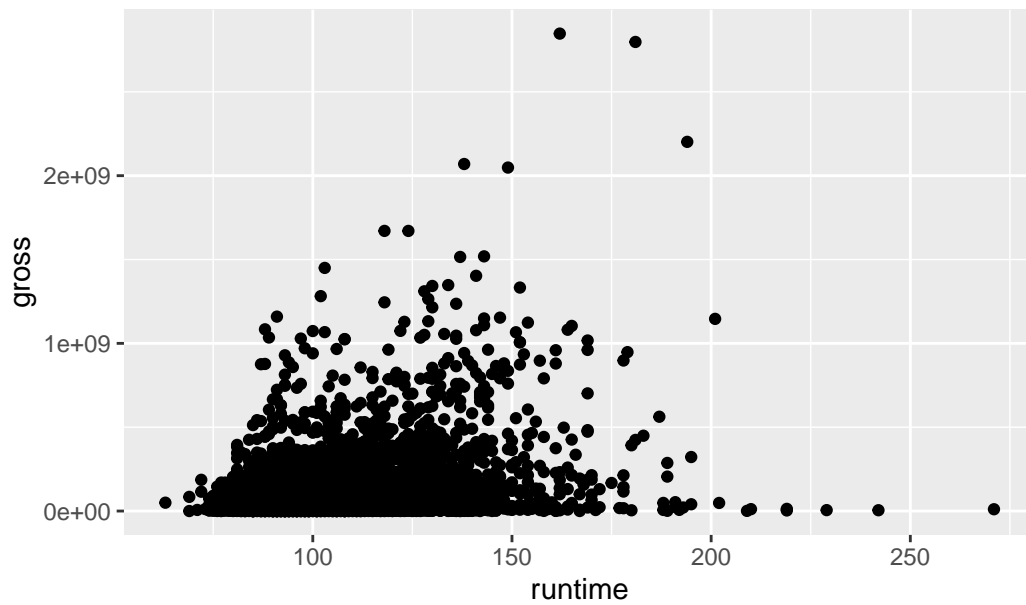

```
data2 <- data %>%
  mutate(log_votes = log(votes))

ggplot(data = data2, aes(x = log_votes, y = score)) +
  geom_point() +
  labs(title = "Log of votes has a converging, positive relationship with Score")
```

Log of votes has a converging, positive relationship with Score
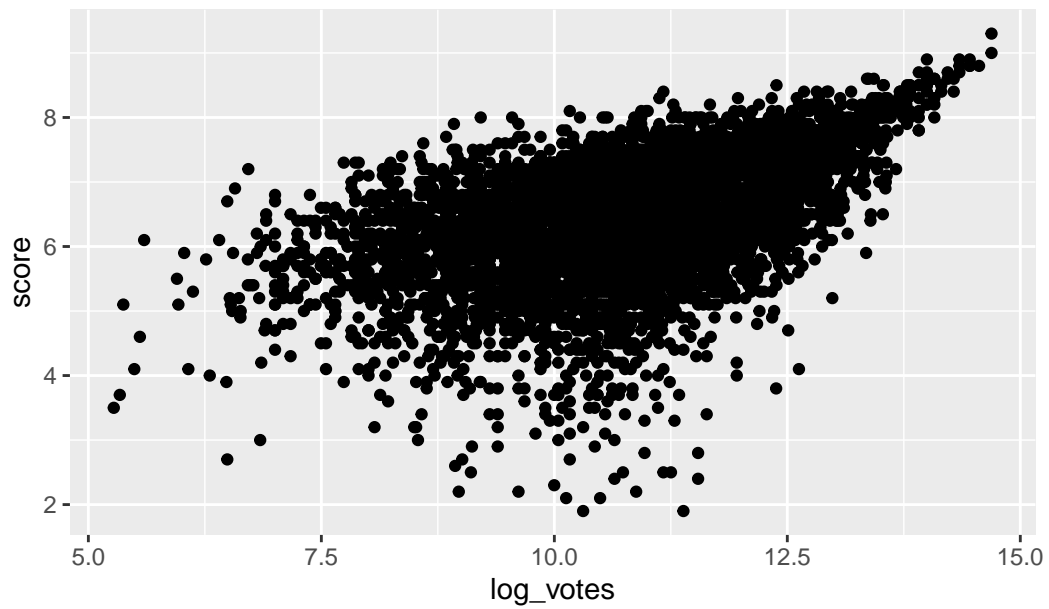
```
numeric_data <- data |>
  select("score", "votes", "budget", "gross", "runtime")
corrplot(cor(numeric_data))
```
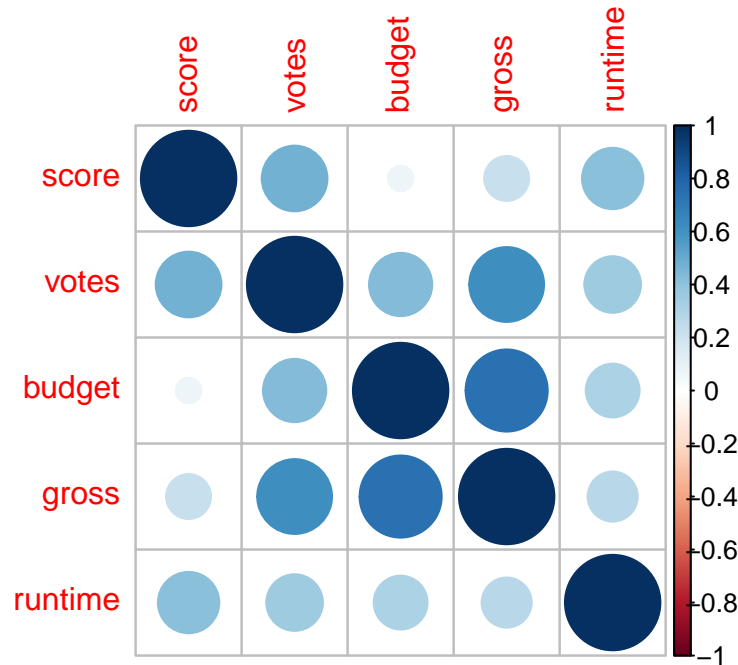
As can be seen, we have some correlated variables between our predictor variables, but none are particularly strong (0.8+) so we don't have to remove any of these predictor variables on the basis of correlation for our later models.

Methodology

As a potential movie investor, we're curious about the prediction of movies' gross based on all the variables we're interested in. We're very curious about the factors that affect success(gross) of the movie. Therefore, we want to explore the following subquestions in order for us to gain a better understanding of 1) Prediction of movies' gross value; and 2) the factors that affect the success of the movie. We choose to apply linear regression, linear mixed models, LASSO, and Ridge for further data analysis.

Linear Regression of all variables

Based on the above violations, we want to first apply linear regression model to our datasets to further explore our questions. We've gained the equation of XX.

```
m1 <- lm(log_gross ~ budget + genre + rating + score + votes + year + runtime + country, d
m1_aug <- augment(m1)
```

```
summary(m1)
```

```
Call:
lm(formula = log_gross ~ budget + genre + rating + score + votes +
    year + runtime + country, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-10.8282  -0.5829   0.1964   0.8493   4.2965

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)          -3.744e+01  4.180e+00  -8.957  < 2e-16
budget                1.412e-08  6.315e-10  22.365  < 2e-16
genreAdventure       -2.590e-01  8.705e-02  -2.975 0.002941
genreAnimation        1.241e-01  1.143e-01   1.086 0.277576
genreBiography       -4.964e-01  9.146e-02  -5.428 5.96e-08
genreComedy          -1.475e-01  5.401e-02  -2.731 0.006334
genreCrime           -4.515e-01  8.051e-02  -5.608 2.15e-08
genreDrama           -5.447e-01  6.328e-02  -8.607  < 2e-16
genreFamily          -1.944e-01  6.729e-01  -0.289 0.772682
genreFantasy         -6.681e-02  2.139e-01  -0.312 0.754773
genreHorror           3.404e-01  9.604e-02   3.544 0.000398
genreMystery         -3.752e-01  3.302e-01  -1.136 0.255890
genreRomance         -1.523e+00  6.012e-01  -2.533 0.011329
genreSci-Fi          -6.209e-01  5.481e-01  -1.133 0.257342
genreThriller         1.328e-01  5.078e-01   0.262 0.793658
genreWestern         -6.654e-01  9.480e-01  -0.702 0.482777
ratingG              -1.228e+00  1.350e+00  -0.910 0.362951
ratingNC-17          -2.579e+00  1.400e+00  -1.842 0.065534
ratingNot Rated      -4.264e+00  1.363e+00  -3.129 0.001761
ratingPG             -1.358e+00  1.343e+00  -1.011 0.311998
ratingPG-13          -1.595e+00  1.344e+00  -1.187 0.235268
ratingR              -2.147e+00  1.344e+00  -1.598 0.110025
ratingTV-MA          -3.676e+00  1.649e+00  -2.229 0.025858
ratingUnrated        -4.000e+00  1.383e+00  -2.892 0.003845
ratingX              -2.046e+00  1.900e+00  -1.077 0.281631
score                 1.486e-01  2.503e-02   5.938 3.06e-09
votes                 2.511e-06  1.304e-07  19.260  < 2e-16
year                  2.704e-02  1.942e-03  13.924  < 2e-16
runtime               8.846e-03  1.341e-03   6.599 4.54e-11
```

```
countryAruba                              -2.746e-01  1.547e+00  -0.177 0.859132
countryAustralia                          -1.585e-01  8.014e-01  -0.198 0.843201
countryAustria                             3.469e-03  1.223e+00   0.003 0.997736
countryBelgium                            -1.629e+00  1.026e+00  -1.587 0.112484
countryBrazil                             -3.999e-01  1.093e+00  -0.366 0.714505
countryCanada                             -2.929e-01  7.846e-01  -0.373 0.708903
countryChile                              -4.102e-01  1.548e+00  -0.265 0.791039
countryChina                               1.903e-01  8.186e-01   0.232 0.816161
countryColombia                            8.502e-02  1.546e+00   0.055 0.956153
countryCzech Republic                     -7.212e-01  9.255e-01  -0.779 0.435914
countryDenmark                            -5.721e-01  8.825e-01  -0.648 0.516846
countryFederal Republic of Yugoslavia     -3.057e+00  1.564e+00  -1.955 0.050678
countryFinland                             4.607e-01  1.226e+00   0.376 0.707130
countryFrance                             -4.763e-01  7.847e-01  -0.607 0.543880
countryGermany                            -7.834e-02  7.882e-01  -0.099 0.920830
countryHong Kong                           4.032e-02  8.213e-01   0.049 0.960851
countryHungary                            -2.945e-02  1.548e+00  -0.019 0.984821
countryIceland                            -3.015e+00  1.224e+00  -2.464 0.013784
countryIndia                               8.265e-01  9.233e-01   0.895 0.370766
countryIndonesia                          -1.661e+00  1.223e+00  -1.359 0.174238
countryIran                               -1.324e+00  1.100e+00  -1.203 0.228841
countryIreland                            -5.044e-01  8.274e-01  -0.610 0.542178
countryIsrael                             -7.439e-01  1.549e+00  -0.480 0.631112
countryItaly                              -1.653e+00  8.326e-01  -1.985 0.047220
countryJamaica                             2.627e-01  1.547e+00   0.170 0.865166
countryJapan                              -3.555e-01  8.151e-01  -0.436 0.662811
countryKenya                              -2.308e+00  1.548e+00  -1.491 0.136030
countryLebanon                             9.549e-01  1.546e+00   0.618 0.536771
countryMalta                               1.328e+00  1.547e+00   0.858 0.390667
countryMexico                             -1.955e-01  8.733e-01  -0.224 0.822912
countryNetherlands                         2.179e-01  1.096e+00   0.199 0.842344
countryNew Zealand                        -1.200e+00  8.364e-01  -1.435 0.151425
countryNorway                             -1.277e+00  1.096e+00  -1.165 0.244187
countryPanama                             -1.287e+00  1.548e+00  -0.831 0.405732
countryPortugal                           -9.858e-01  1.547e+00  -0.637 0.523883
countryRepublic of Macedonia              -3.673e-01  1.562e+00  -0.235 0.814132
countryRussia                             -2.674e-01  1.023e+00  -0.261 0.793761
countrySouth Africa                        3.774e-01  1.023e+00   0.369 0.712291
countrySouth Korea                        -8.707e-01  8.583e-01  -1.014 0.310411
countrySpain                              -1.747e-01  8.360e-01  -0.209 0.834494
countrySweden                             -5.265e-01  9.788e-01  -0.538 0.590660
countrySwitzerland                        -1.853e+00  1.023e+00  -1.811 0.070221
countryTaiwan                              6.765e-01  1.101e+00   0.614 0.539029
```

```
countryThailand                     -4.094e+00  1.547e+00  -2.647 0.008151
countryUnited Arab Emirates          3.825e-01  1.224e+00   0.312 0.754680
countryUnited Kingdom               -4.497e-01  7.762e-01  -0.579 0.562354
countryUnited States                -7.408e-02  7.744e-01  -0.096 0.923796
countryWest Germany                  4.256e-01  1.095e+00   0.389 0.697494
countryYugoslavia                   -5.591e-01  1.548e+00  -0.361 0.717970

(Intercept)                     ***
budget                          ***
genreAdventure                  **
genreAnimation
genreBiography                  ***
genreComedy                     **
genreCrime                      ***
genreDrama                      ***
genreFamily
genreFantasy
genreHorror                     ***
genreMystery
genreRomance                    *
genreSci-Fi
genreThriller
genreWestern
ratingG
ratingNC-17                     .
ratingNot Rated                 **
ratingPG
ratingPG-13
ratingR
ratingTV-MA                     *
ratingUnrated                   **
ratingX
score                           ***
votes                           ***
year                            ***
runtime                         ***
countryAruba
countryAustralia
countryAustria
countryBelgium
countryBrazil
countryCanada
countryChile
```

```
countryChina
countryColombia
countryCzech Republic
countryDenmark
countryFederal Republic of Yugoslavia .
countryFinland
countryFrance
countryGermany
countryHong Kong
countryHungary
countryIceland                              *
countryIndia
countryIndonesia
countryIran
countryIreland
countryIsrael
countryItaly                               *
countryJamaica
countryJapan
countryKenya
countryLebanon
countryMalta
countryMexico
countryNetherlands
countryNew Zealand
countryNorway
countryPanama
countryPortugal
countryRepublic of Macedonia
countryRussia
countrySouth Africa
countrySouth Korea
countrySpain
countrySweden
countrySwitzerland                         .
countryTaiwan
countryThailand                          **
countryUnited Arab Emirates
countryUnited Kingdom
countryUnited States
countryWest Germany
countryYugoslavia
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.338 on 5345 degrees of freedom
Multiple R-squared:  0.5046,     Adjusted R-squared:  0.4974
F-statistic:  70.7 on 77 and 5345 DF,  p-value: < 2.2e-16
```

Residual plots/Assumptions

```r
ggplot(m1_aug, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, color = "darkred") +
  labs(x = "Fitted (predicted) value", y = "Residual") +
  theme_bw()
```
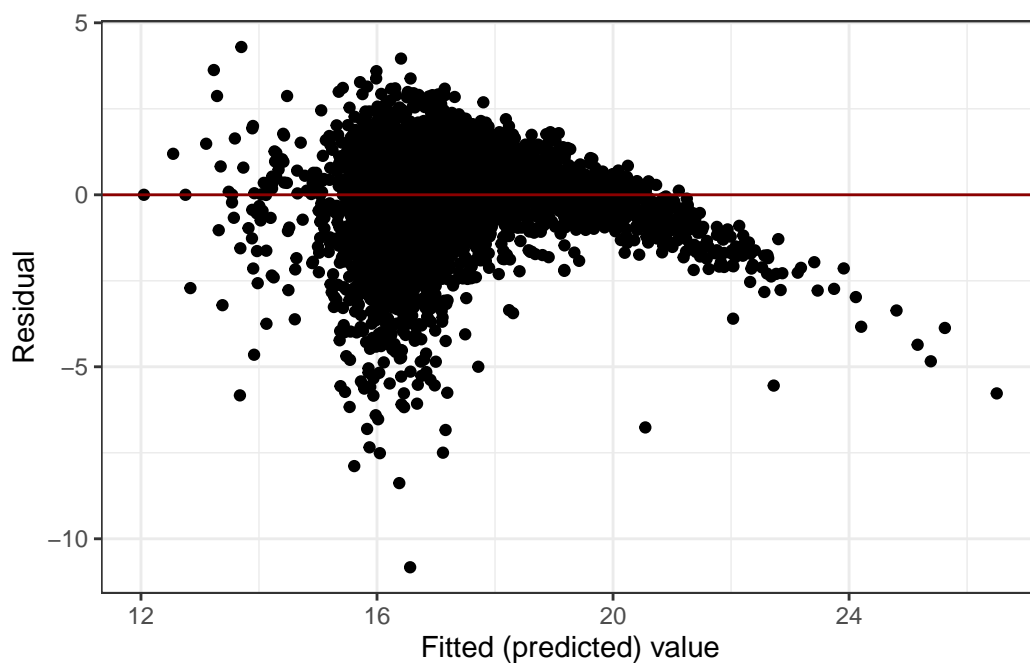

We can see that the linear assumption is violated since the data is not symmetrically dis-
tributed observations around the horizontal axis. We may assume that the dataset has some
independence. For example, the movie with the same director or writer, or the star in the
movie may lead to higher budget or gross values, such that this movie-specific relatioinship
violates independence.

Linear Mixed Model & Random Effect

Based on the above violations, we want to apply linear mixed model to our datasets to further explore our questions. We want to look at the associations between our interested variables, and randomeffect due to country.

Interested Hypothesis Test 1:

Question 1: Is there evidence to suggest that Budget has an effect on the success of the movies? Null hypothesis: $p\_1 = 0$ There isn't sufficient evidence to suggest that budget is associated with movies' gross, while controlling for all of the variables. Alternative hypothesis: p_1 0 There is sufficient evidence to suggest that budget is associated with movies' gross, while controlling for all of the variables.

We use significance level of 0.05. Since the t-statistics is 23.227 and the p-value is $< $ 2e-16 which is much smaller than our significance level, so we reject the null hypothesis since there's sufficient evidence, and thus there's sufficient evidence to suggest to budget does have an effect on the success of the movies.

Interested Hypothesis Test 2:

Aim 2: Is there evidence to suggest that Genre has an effect on the success of the movies? Null hypothesis: $p\_2 = 0$ There isn't sufficient evidence to suggest that Genre is associated with movies' gross, while controlling for all of the variables. Alternative hypothesis: p_2 0 There is sufficient evidence to suggest that Genre is associated with movies' gross, while controlling for all of the variables.

We use significance level of 0.05. Since the t-statistics is 23.227 and the p-value is $< $ 2e-16 which is much smaller than our significance level, so we reject the null hypothesis since there's sufficient evidence, and thus there's sufficient evidence to suggest to budget does have an effect on the success of the movies.

LASSO:

```
set.seed(919)
library(glmnet)
y <- data$log_gross
x <- model.matrix(log_gross ~ budget + genre + rating + score + votes + year + runtime + c

m_lasso_cv <- cv.glmnet(x, y, alpha = 1)
best_lambda <- m_lasso_cv$lambda.min
```

```
best_lambda
```

```
[1] 0.01073681
```

```
m_best <- glmnet(x, y, alpha = 1, lambda = best_lambda)
```

Explanation for LASSO: ## Ridge:

Explanation for Ridge:

Predicting moves' gross

Based on all the models we've done above, we've analyzed the relationship between different variables that we're interested in. In order to make our findings applicable for future use, we'd like to make a prediction on movies' gross based on our current variables.

Missing Data Analysis:

Results

Explain the reasoning for the type of model you're fitting, predictor variables considered for the model including any interactions. Additionally, show how you arrived at the final model by describing the model selection process, interactions considered, variable transformations (if needed), assessment of conditions and diagnostics, and any other relevant considerations that were part of the model fitting process.

Discussion

Summary + statistical arguments to support my conclusions + future limitations/future ideads