Final Project

Christina Yu, Damian Kim

Read in the data

```
library(tidyverse)
```

```
-- Attaching packages --------------------------------------- tidyverse 1.3.2 --
v ggplot2 3.4.0      v purrr   1.0.0
v tibble  3.1.8      v dplyr   1.0.10
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()
```

```
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 tidymodels_prefer() to resolve common conflicts.
```

```
library(corrplot)
```

```
corrplot 0.92 loaded
```

```
library(nnet)
library(car)
```

```
Loading required package: carData

Attaching package: 'car'

The following object is masked from 'package:dplyr':

    recode

The following object is masked from 'package:purrr':

    some
```

```
library(mice)
```

```

Attaching package: 'mice'

The following object is masked from 'package:stats':

    filter

The following objects are masked from 'package:base':

    cbind, rbind
```

```r
library(naniar)
library(UpSetR)
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

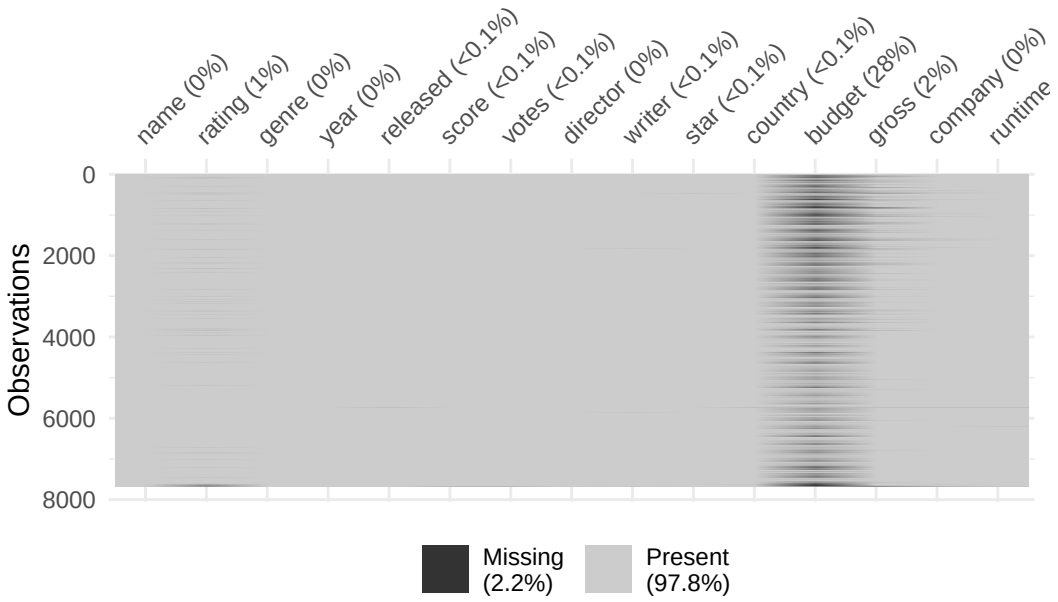
– `writer`: writer of the movie

– `year`: year of release

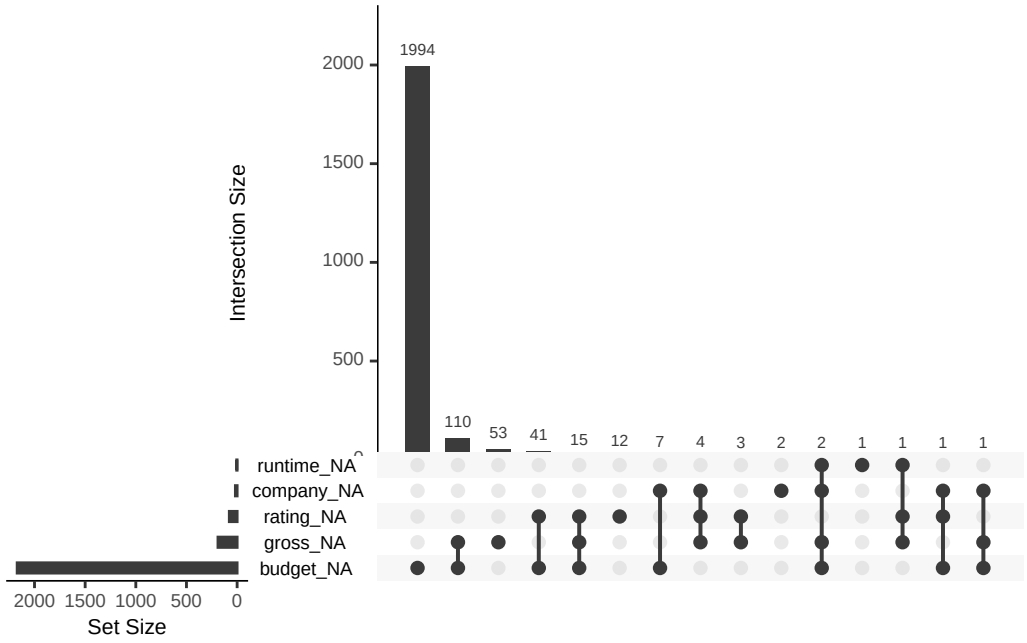
We will explore the factors that make a movie successful through examining the effects of `gross`, `budget`, `genre`, `rating`,`score` `votes`, `year`, and `runtime` for individual movie.

First we will explore missingness in our dataset.

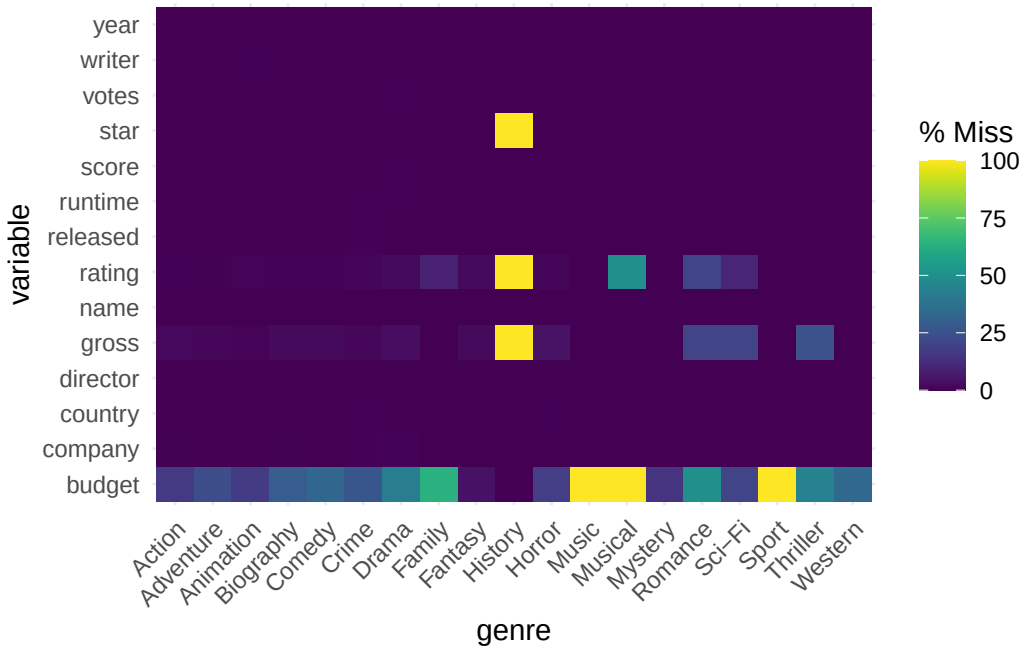
```
vis_miss(data)
```




```
gg_miss_upset(data)
```




```
gg_miss_fct(x = data, fct = genre)
```



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 filtered all the null values for the rest of predictor variables as well. About 28% of the observations were missing budget, whereas the next most missing variable was gross, with 2%. From elementary missingness analysis there did not seem to be numerically significant relationships between the missingness of variables and budget. We decided to do a complete case analysis, because even though MCAR missingness is unlikely in real world data, 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 (and because we would not have too much biasing, since in general missingness was less than 2% asides from budget). 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:

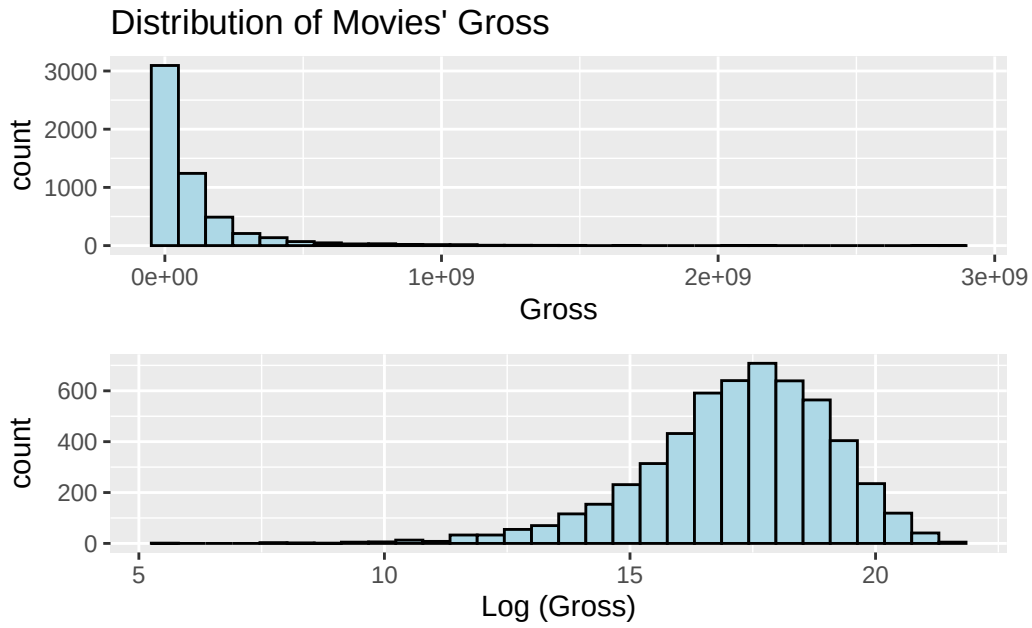
```
data %>%
  summarise(mean_gross = mean(gross),
            median_gross = median(gross),
            sd_gross = sd(gross),
            min_gross = min(gross),
            max_gross = max(gross)) %>%
  kable()
```

mean_gross	median_gross	sd_gross	min_gross	max_gross
103192280	36850101	187278279	309	2847246203

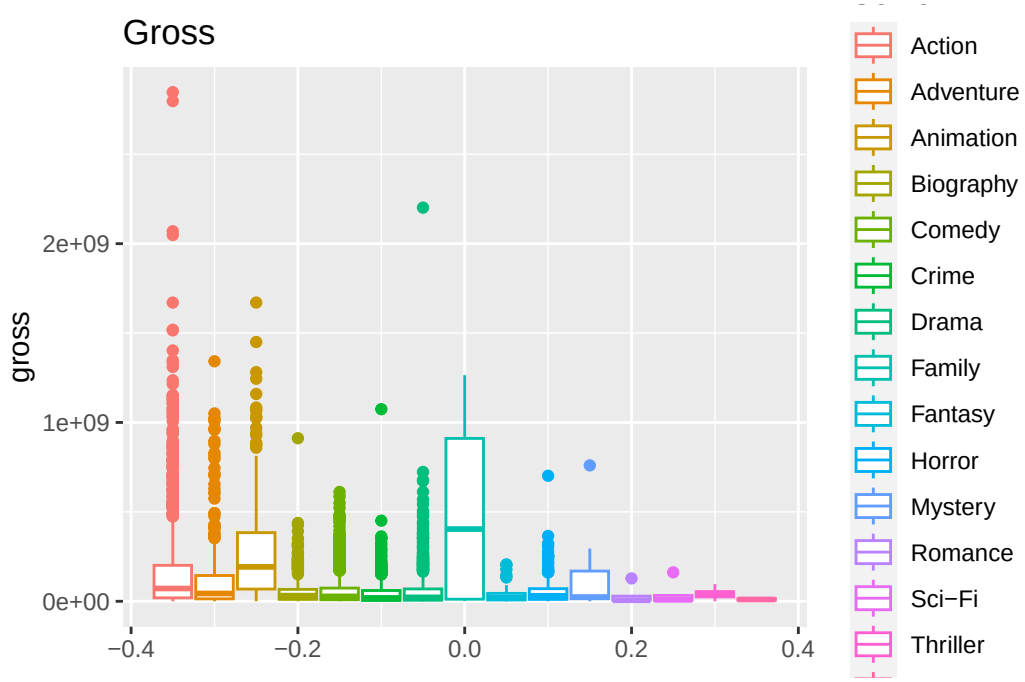
2. Log-transformation and Distribution of the `gross` variable:

```r
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
```

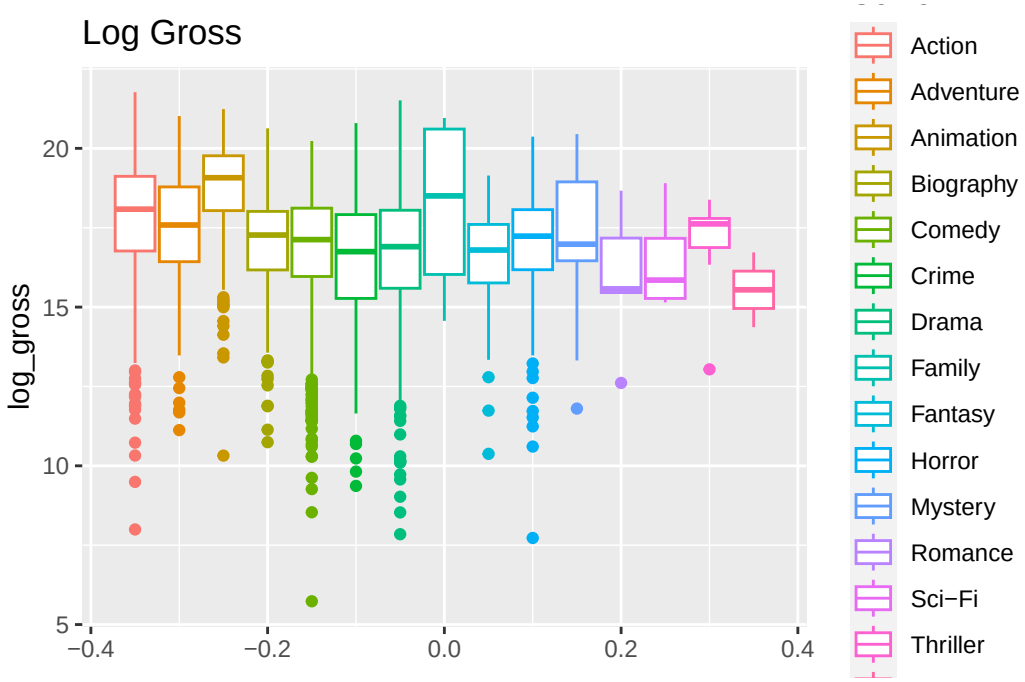
```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




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




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

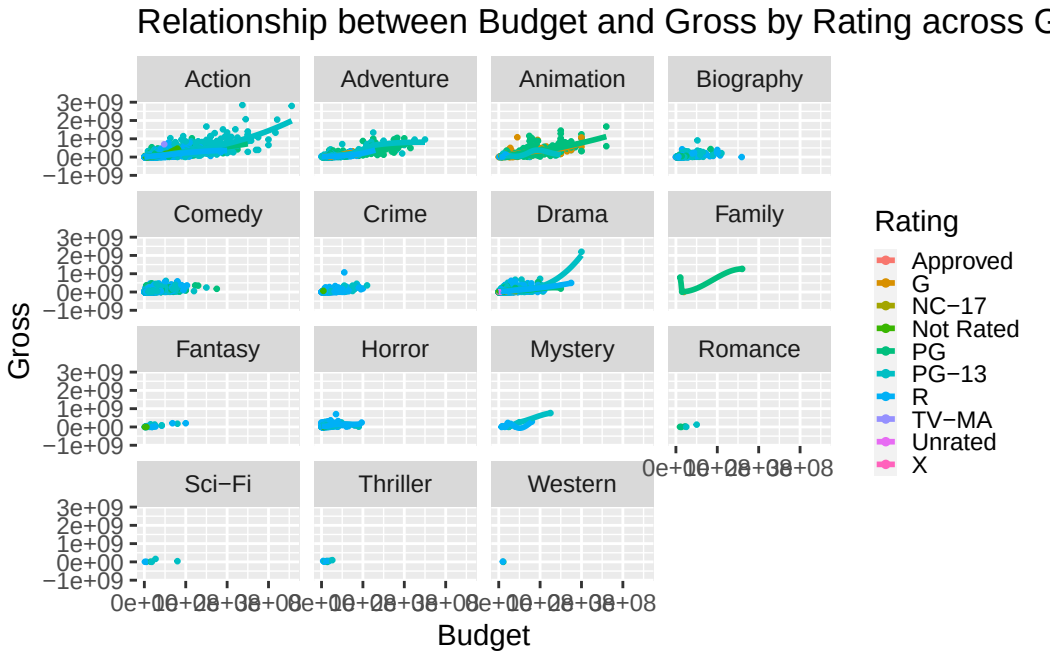


Since the response variable is significantly right skewed, we apply a log-transformation to it and will use log(gross) as our new response variable in the future analysis. Now, our response variable is unimodal, roughly a normal distribution, with the mean at 17.2102, and several outliers at left tail

keep the original graph for gross. XX genre has the most movies / some outliers a lot of movie

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

```
ggplot(data = data, aes(x=budget, y = 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("Gross")+
  scale_color_discrete(name = "Rating", guide = guide_legend(override.aes = list(size = 1)
  theme(panel.spacing.x = unit(2, "mm"))
```

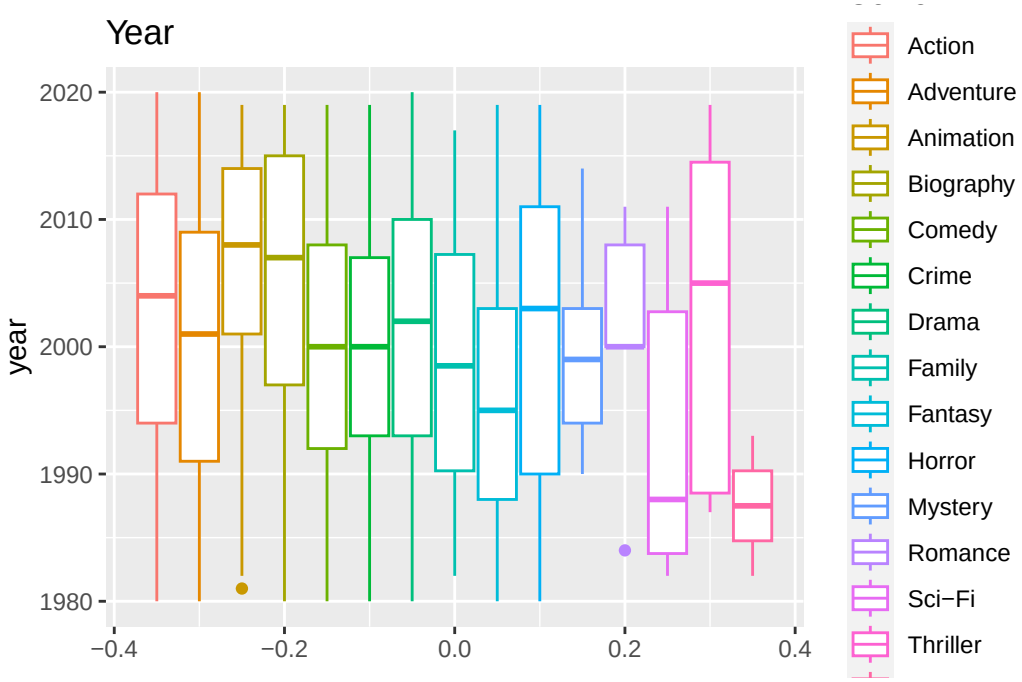


We observe that the relationship between budget and gross is vastly different across genres: Action, adventure, animation movies have a steep slope and generally high budget spans, with outliers which have exceedingly high budget and relatively high gross values. On the other hand, genres such as horror, 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.

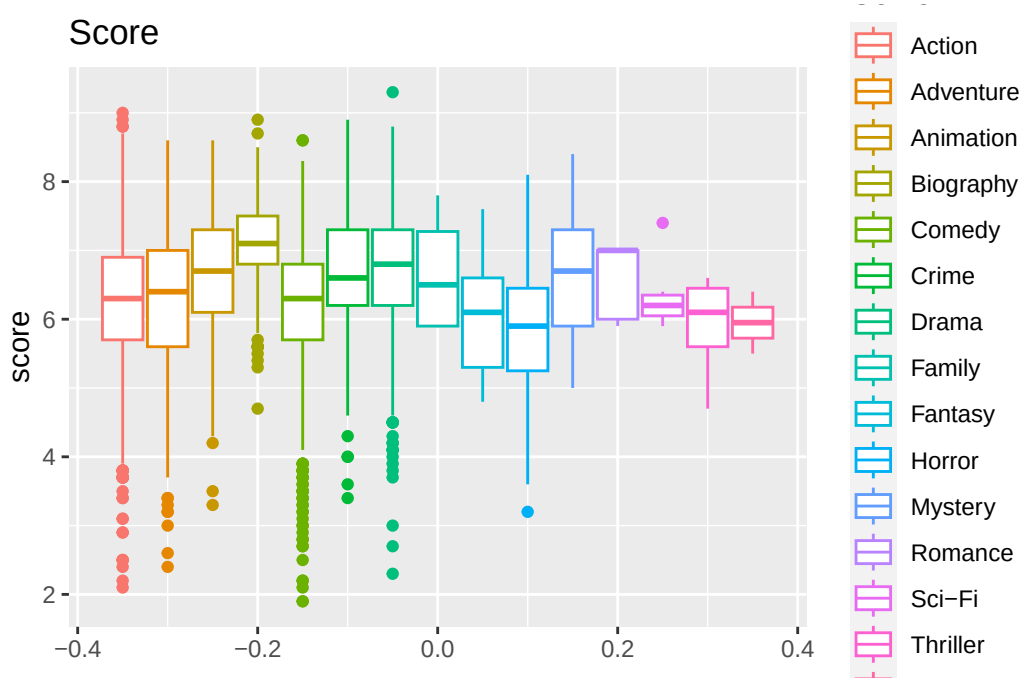
EDA: Visualizatioins and Summary Statistics

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.

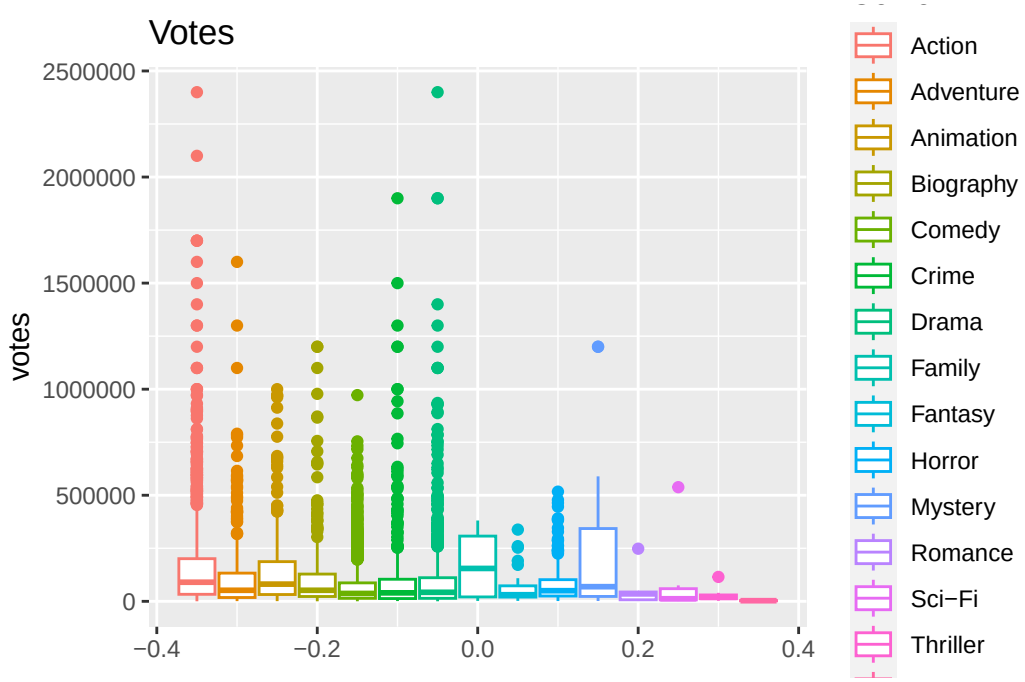
```
ggplot(data = data) +
  geom_boxplot(aes(y = year, color = genre)) +
  labs(title = "Year",
       color = "Genre")
```



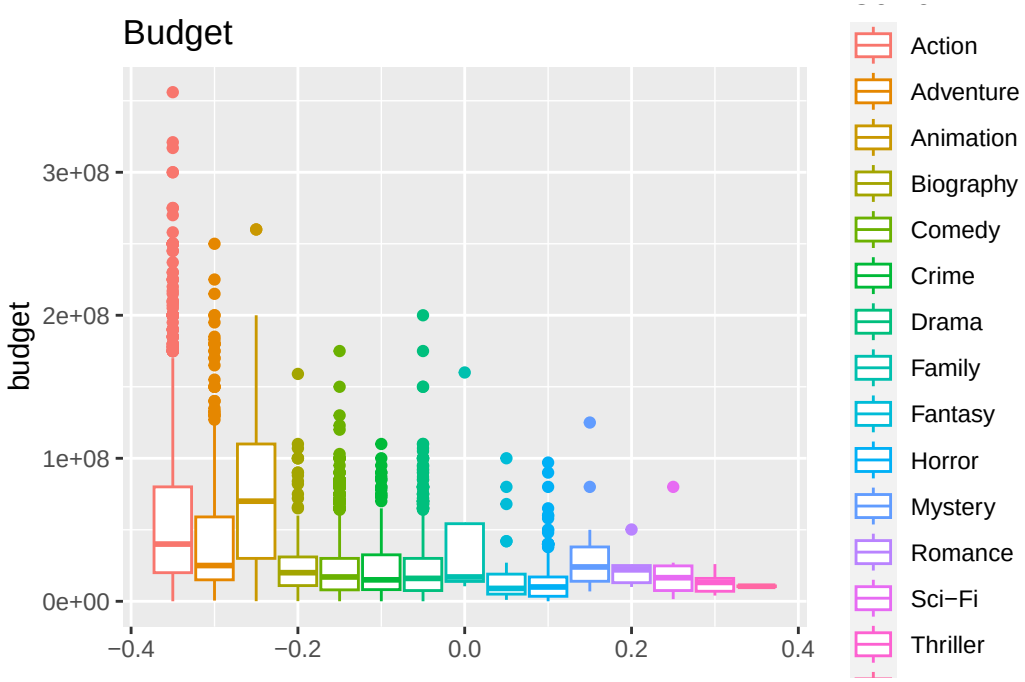

```
ggplot(data = data) +
  geom_boxplot(aes(y = score, color = genre)) +
  labs(title = "Score",
       color = "Genre")
```



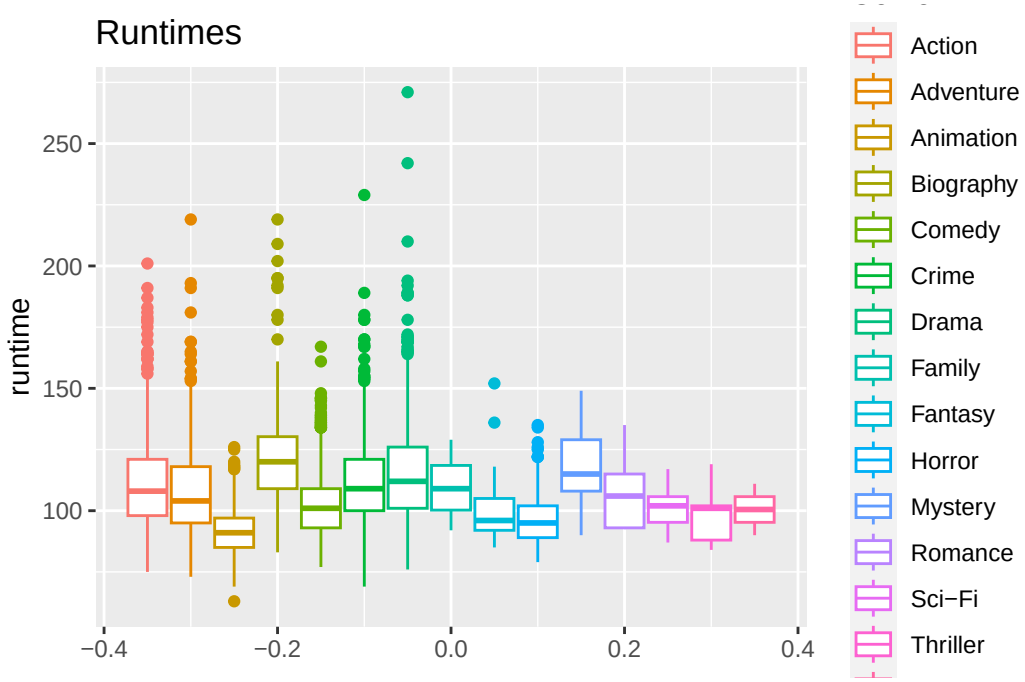

```
ggplot(data = data) +
  geom_boxplot(aes(y = votes, color = genre)) +
  labs(title = "Votes",
       color = "Genre")
```




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

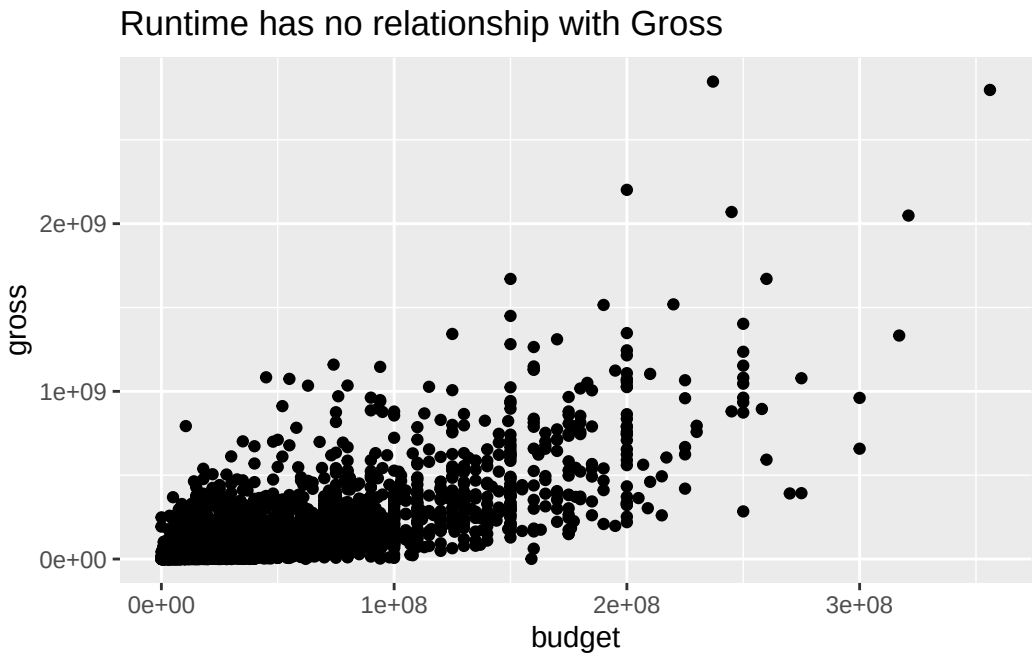



```
ggplot(data = data) +
  geom_boxplot(aes(y = runtime, color = genre)) +
  labs(title = "Runtimes",
       color = "Genre")
```

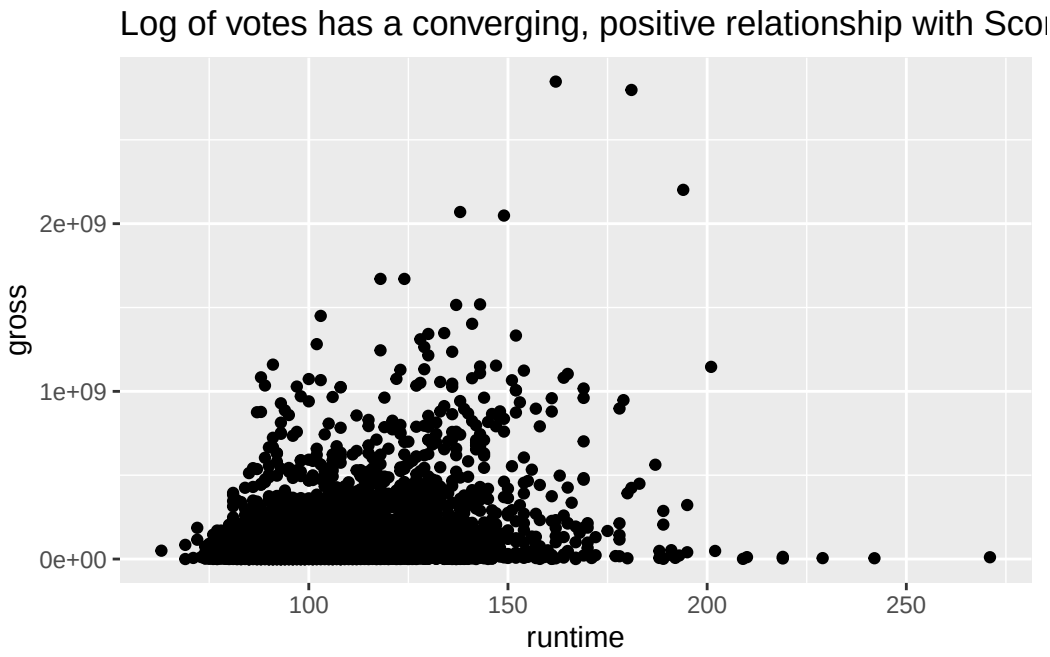


XXX can keep the interesting observations here. we can say that runtime across all the genress are similar around 100 – no need to include the plot – need to focus on relationship between response and predictor

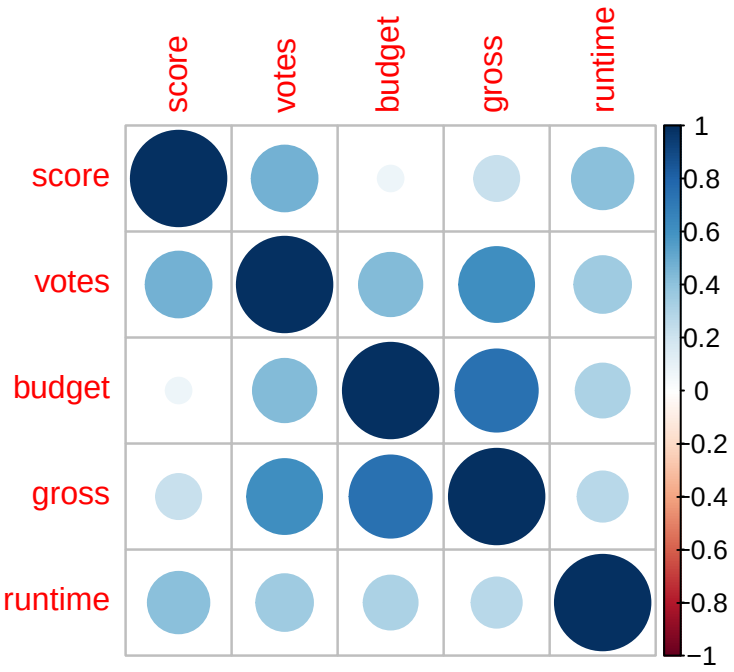
```
data2 <- data %>%
  mutate(log_votes = log(votes))

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




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

ggplot(data = data2, aes(x = runtime, y = gross)) +
  geom_point() +
  labs(title = "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, Residual Plots, linear mixed models, LASSO, Repeated k-fold validation, and Missing Data analysis for further data analysis.

LASSO (Variable Selection):

Since we're quite interested in the above variables as we just mentioned, but it's lack of technical proof to show that our selected variables indeed can lead to further data exploration. Therefore, we apply LASSO here as the variable selection process to see whether we con continue with our selected variables.

```r
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
```

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

```
78 x 1 sparse Matrix of class "dgCMatrix"
                                              s0
(Intercept)                                    .
budget                                         1.467877e-08
genreAdventure                                -1.297376e-01
genreAnimation                                 1.792094e-01
genreBiography                                -3.581104e-01
genreComedy                                   -6.988655e-02
genreCrime                                    -3.494242e-01
genreDrama                                    -4.453273e-01
genreFamily                                    .
genreFantasy                                   .
genreHorror                                    3.251064e-01
genreMystery                                  -1.483801e-01
genreRomance                                  -1.070428e+00
genreSci-Fi                                   -2.472153e-01
genreThriller                                  .
genreWestern                                  -6.241539e-02
ratingG                                        2.424686e-01
ratingNC-17                                   -7.285638e-01
ratingNot Rated                               -2.516918e+00
ratingPG                                       1.883001e-01
ratingPG-13                                    .
ratingR                                       -5.304342e-01
ratingTV-MA                                   -1.359022e+00
ratingUnrated                                 -2.251698e+00
```

```
ratingX                                .
score                                  1.270134e-01
votes                                  2.538886e-06
year                                   2.541135e-02
runtime                                7.579645e-03
countryAruba                           .
countryAustralia                       .
countryAustria                         .
countryBelgium                        -1.011077e+00
countryBrazil                          .
countryCanada                          .
countryChile                           .
countryChina                           3.129037e-01
countryColombia                        .
countryCzech Republic                 -1.534489e-01
countryDenmark                        -6.039812e-02
countryFederal Republic of Yugoslavia -2.047457e+00
countryFinland                         2.232914e-01
countryFrance                         -1.325658e-01
countryGermany                         8.925869e-02
countryHong Kong                       2.036604e-01
countryHungary                         .
countryIceland                        -2.205302e+00
countryIndia                           8.098359e-01
countryIndonesia                      -7.035011e-01
countryIran                           -6.233484e-01
countryIreland                        -5.688087e-02
countryIsrael                          .
countryItaly                          -1.194338e+00
countryJamaica                         .
countryJapan                           .
countryKenya                          -1.266646e+00
countryLebanon                         4.882455e-01
countryMalta                           7.906576e-01
countryMexico                          .
countryNetherlands                     .
countryNew Zealand                    -7.344671e-01
countryNorway                         -6.141330e-01
countryPanama                         -2.801313e-01
countryPortugal                        .
countryRepublic of Macedonia           .
countryRussia                          .
countrySouth Africa                    2.992477e-01
```

```
countrySouth Korea                    -3.625738e-01
countrySpain                           .
countrySweden                          .
countrySwitzerland                    -1.208691e+00
countryTaiwan                          4.654508e-01
countryThailand                       -2.978446e+00
countryUnited Arab Emirates            6.569088e-02
countryUnited Kingdom                 -1.532377e-01
countryUnited States                   1.879491e-01
countryWest Germany                    1.843197e-01
countryYugoslavia                      .
```

Explanation for LASSO: Based on this, dot means that those doesn't have a significant effect XXX. Put LASSO above – this is variable selection – we basically could say that because XXX these variables have coefficient – we keep these variables as our predictors.

Linear Regression of all variables

Since it's obvious that our data has a continuous response, we'll not use logistic or ordinal regression here or multinomial regression. Therefore, we gonna use linear regression for our model. 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 ~ log(budget) + genre + rating + log(score) + log(votes) + year + log(r
m1_aug <- augment(m1)
summary(m1)
```

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

Residuals:
     Min       1Q   Median       3Q      Max
-10.0255  -0.4551   0.0999   0.6049   4.4163

Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                         15.3283808  3.5386062   4.332 1.51e-05
log(budget)                          0.5067318  0.0165908  30.543  < 2e-16
genreAdventure                      -0.0959728  0.0689665  -1.392 0.164106
```

```
genreAnimation                          0.3751732  0.0907024   4.136 3.58e-05
genreBiography                         -0.1538538  0.0711766  -2.162 0.030695
genreComedy                             0.0420184  0.0421840   0.996 0.319259
genreCrime                             -0.2070222  0.0634371  -3.263 0.001108
genreDrama                             -0.1354841  0.0497604  -2.723 0.006495
genreFamily                             0.2546620  0.5334854   0.477 0.633129
genreFantasy                           -0.0404537  0.1694913  -0.239 0.811365
genreHorror                             0.3907570  0.0769649   5.077 3.96e-07
genreMystery                           -0.2307367  0.2613106  -0.883 0.377277
genreRomance                           -1.1319316  0.4765216  -2.375 0.017565
genreSci-Fi                             0.0159406  0.4346419   0.037 0.970745
genreThriller                           0.6951198  0.4025923   1.727 0.084296
genreWestern                            0.7256805  0.7520586   0.965 0.334626
ratingG                                -1.3416343  1.0712650  -1.252 0.210485
ratingNC-17                            -2.6492551  1.1104240  -2.386 0.017076
ratingNot Rated                        -3.1302871  1.0808245  -2.896 0.003792
ratingPG                               -1.4673304  1.0654285  -1.377 0.168502
ratingPG-13                            -1.7256348  1.0660203  -1.619 0.105557
ratingR                                -2.1073089  1.0658926  -1.977 0.048089
ratingTV-MA                            -3.4000215  1.3080414  -2.599 0.009366
ratingUnrated                          -2.6456124  1.0972869  -2.411 0.015940
ratingX                                -2.9689477  1.5070325  -1.970 0.048883
log(score)                             -0.6087812  0.1175503  -5.179 2.31e-07
log(votes)                              0.7428187  0.0145940  50.899  < 2e-16
year                                   -0.0065614  0.0016530  -3.969 7.30e-05
log(runtime)                            0.4607315  0.1255077   3.671 0.000244
countryAruba                           -0.8883772  1.2273438  -0.724 0.469207
countryAustralia                       -0.7251753  0.6358833  -1.140 0.254162
countryAustria                         -0.7212264  0.9695940  -0.744 0.457004
countryBelgium                         -2.3579208  0.8143623  -2.895 0.003802
countryBrazil                          -0.8147862  0.8668979  -0.940 0.347318
countryCanada                          -0.7900712  0.6225698  -1.269 0.204479
countryChile                           -0.6021238  1.2277256  -0.490 0.623844
countryChina                            0.1266840  0.6498922   0.195 0.845454
countryColombia                         0.2678266  1.2259260   0.218 0.827072
countryCzech Republic                  -1.0556990  0.7343929  -1.438 0.150631
countryDenmark                         -1.3356937  0.6999469  -1.908 0.056409
countryFederal Republic of Yugoslavia -4.6440893  1.2400396  -3.745 0.000182
countryFinland                          0.1620335  0.9725086   0.167 0.867680
countryFrance                          -1.1579110  0.6228494  -1.859 0.063074
countryGermany                         -0.9031245  0.6257676  -1.443 0.149015
countryHong Kong                       -0.3490576  0.6519056  -0.535 0.592367
countryHungary                         -0.5888385  1.2273649  -0.480 0.631419
```

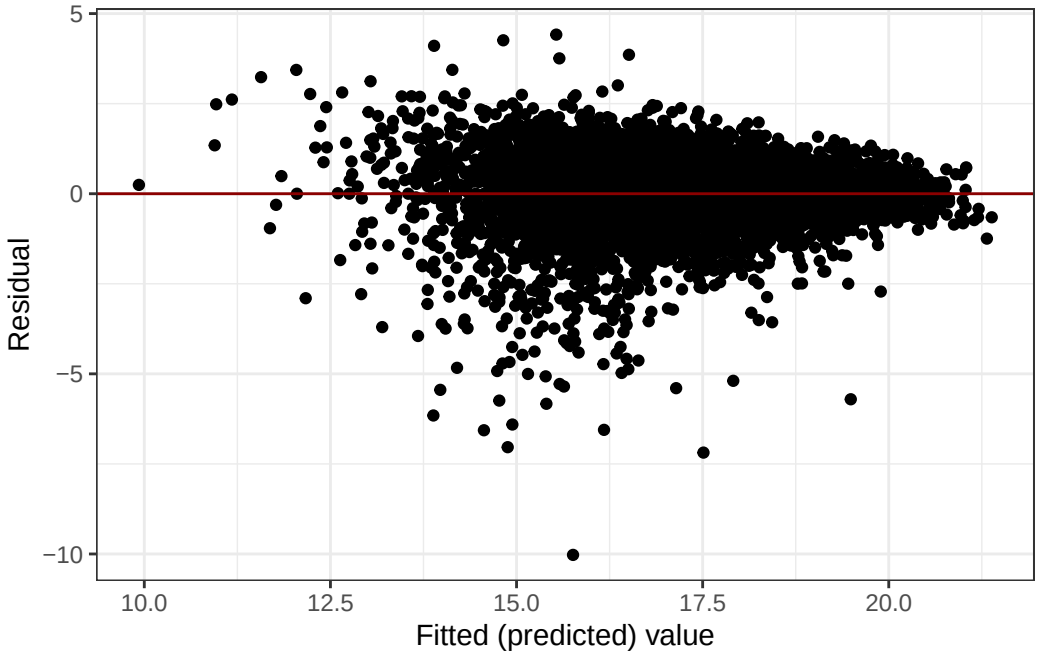
```
countryIceland                -2.6826215  0.9701351  -2.765 0.005708
countryIndia                   0.0929953  0.7317485   0.127 0.898877
countryIndonesia              -1.3843799  0.9692415  -1.428 0.153260
countryIran                   -0.4151541  0.8720346  -0.476 0.634040
countryIreland                -0.8462180  0.6565845  -1.289 0.197517
countryIsrael                 -0.3261548  1.2281112  -0.266 0.790577
countryItaly                  -1.6963700  0.6602466  -2.569 0.010217
countryJamaica                 0.7857689  1.2262718   0.641 0.521694
countryJapan                  -0.7401843  0.6464599  -1.145 0.252268
countryKenya                  -2.2684087  1.2280334  -1.847 0.064775
countryLebanon                 1.2027941  1.2256081   0.981 0.326447
countryMalta                   0.3049842  1.2269680   0.249 0.803705
countryMexico                 -0.0000268  0.6922985   0.000 0.999969
countryNetherlands            -0.8383497  0.8691333  -0.965 0.334798
countryNew Zealand            -0.9289969  0.6628940  -1.401 0.161145
countryNorway                 -0.9949058  0.8689564  -1.145 0.252284
countryPanama                 -1.1493398  1.2280397  -0.936 0.349360
countryPortugal               -1.7563660  1.2268082  -1.432 0.152301
countryRepublic of Macedonia  -0.7627100  1.2389749  -0.616 0.538186
countryRussia                 -0.1888589  0.8110093  -0.233 0.815872
countrySouth Africa           -0.4681655  0.8117030  -0.577 0.564120
countrySouth Korea            -0.9526599  0.6806058  -1.400 0.161654
countrySpain                  -0.6728193  0.6632945  -1.014 0.310457
countrySweden                 -0.9583404  0.7758246  -1.235 0.216790
countrySwitzerland            -1.1752302  0.8115445  -1.448 0.147637
countryTaiwan                  0.6495262  0.8732021   0.744 0.457004
countryThailand               -3.2829815  1.2267581  -2.676 0.007470
countryUnited Arab Emirates   -0.8629511  0.9705621  -0.889 0.373976
countryUnited Kingdom         -0.8709494  0.6159889  -1.414 0.157448
countryUnited States          -0.5800331  0.6144924  -0.944 0.345252
countryWest Germany           -0.8340880  0.8685676  -0.960 0.336946
countryYugoslavia             -0.3214551  1.2267952  -0.262 0.793310

(Intercept)                   ***
log(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                              *
log(score)                           ***
log(votes)                           ***
year                                 ***
log(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                         *
```

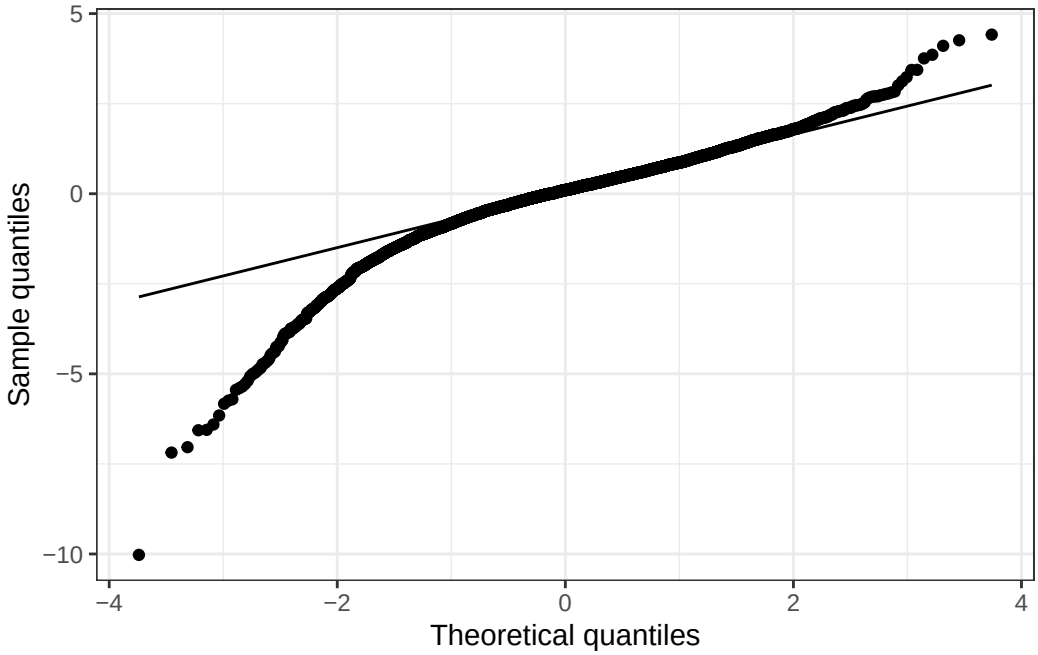
```
countryJamaica
countryJapan
countryKenya                            .
countryLebanon
countryMalta
countryMexico
countryNetherlands
countryNew Zealand
countryNorway
countryPanama
countryPortugal
countryRepublic of Macedonia
countryRussia
countrySouth Africa
countrySouth Korea
countrySpain
countrySweden
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: 1.061 on 5345 degrees of freedom
Multiple R-squared:  0.6886,    Adjusted R-squared:  0.6841
F-statistic: 153.5 on 77 and 5345 DF,  p-value: < 2.2e-16
```

Residual plots/Assumptions

```
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()
```




```
ggplot(m1_aug, aes(sample = .resid)) +
  stat_qq() +
  stat_qq_line() +
  theme_bw() +
  labs(x = "Theoretical quantiles",
       y = "Sample quantiles")
```



We may assume that each movie is independent of each other. We can see that the linear assumption is violated since the data is not symmetrically distributed observations around the horizontal axis. The linearity, constant variance, and normality is not satisfied from the above graph. In order to address the normality violation, we did log transformation for our response variable to be `log_gross`. In order to address this problem, we use log transformation to our numeric variabls, and in this way, the linearity residual plot is improved, while the constant variance and normality is still violated.

We acknowledge it – limitations XXX

Linear Mixed Model & Random Effect

Since we have the country, genre, and rating these three categorical variables have too many levels, for example, for United States, we have XXX data records, and for XX country, we only have XX variables, so it would make more sense for use to use these three varaibels as random effects, 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 random effect due to country, genre, and rating.

```
library(lme4)
m4 <- lm(log_gross ~ budget + genre + rating + score + votes + year + runtime + country, d
m5 <- lmer(log_gross ~ 1 + budget + (1 | genre) + (1 | rating) + score + votes + year + ru
```

```
summary(m4)$coef
```

```
                                Estimate   Std. Error     t value
(Intercept)                -3.744139e+01 4.180200e+00 -8.95684179
budget                      1.412419e-08 6.315283e-10 22.36509448
genreAdventure             -2.589934e-01 8.705133e-02 -2.97518069
genreAnimation              1.240931e-01 1.142776e-01  1.08589107
genreBiography             -4.964031e-01 9.145927e-02 -5.42758577
genreComedy                -1.475114e-01 5.401252e-02 -2.73105951
genreCrime                 -4.515406e-01 8.051431e-02 -5.60820272
genreDrama                 -5.446711e-01 6.328392e-02 -8.60678562
genreFamily                -1.943796e-01 6.728664e-01 -0.28888286
genreFantasy               -6.680695e-02 2.138729e-01 -0.31236756
genreHorror                 3.403503e-01 9.604496e-02  3.54365639
genreMystery               -3.751595e-01 3.301647e-01 -1.13627972
genreRomance               -1.522911e+00 6.011683e-01 -2.53325292
genreSci-Fi                -6.209023e-01 5.481052e-01 -1.13281585
genreThriller               1.328250e-01 5.077867e-01  0.26157635
genreWestern               -6.653857e-01 9.479893e-01 -0.70189163
ratingG                    -1.228217e+00 1.349935e+00 -0.90983415
ratingNC-17                -2.578972e+00 1.400111e+00 -1.84197653
ratingNot Rated            -4.263835e+00 1.362514e+00 -3.12938759
ratingPG                   -1.357827e+00 1.342875e+00 -1.01113416
ratingPG-13                -1.594916e+00 1.343619e+00 -1.18703053
ratingR                    -2.147455e+00 1.343549e+00 -1.59834579
ratingTV-MA                -3.676498e+00 1.649425e+00 -2.22895741
ratingUnrated              -4.000440e+00 1.383338e+00 -2.89187494
ratingX                    -2.046172e+00 1.900283e+00 -1.07677203
score                       1.486078e-01 2.502585e-02  5.93817174
votes                       2.511044e-06 1.303769e-07 19.25988301
year                        2.703754e-02 1.941757e-03 13.92426700
runtime                     8.846264e-03 1.340580e-03  6.59883391
countryAruba               -2.746412e-01 1.547380e+00 -0.17748789
countryAustralia           -1.585158e-01 8.013506e-01 -0.19781081
countryAustria              3.469289e-03 1.222692e+00  0.00283742
countryBelgium             -1.629259e+00 1.026377e+00 -1.58738809
countryBrazil              -3.999257e-01 1.093197e+00 -0.36583124
countryCanada              -2.929147e-01 7.845584e-01 -0.37334976
countryChile               -4.102243e-01 1.548160e+00 -0.26497533
countryChina                1.903264e-01 8.186183e-01  0.23249718
countryColombia             8.501925e-02 1.546248e+00  0.05498421
countryCzech Republic      -7.211523e-01 9.255383e-01 -0.77917074
```

```
countryDenmark                          -5.720809e-01 8.824901e-01 -0.64825763
countryFederal Republic of Yugoslavia -3.056663e+00 1.563801e+00 -1.95463618
countryFinland                           4.606788e-01 1.226077e+00  0.37573392
countryFrance                           -4.763101e-01 7.847024e-01 -0.60699460
countryGermany                          -7.834461e-02 7.882310e-01 -0.09939296
countryHong Kong                         4.031553e-02 8.212962e-01  0.04908769
countryHungary                          -2.945483e-02 1.548127e+00 -0.01902610
countryIceland                          -3.014754e+00 1.223685e+00 -2.46366809
countryIndia                             8.264847e-01 9.233280e-01  0.89511498
countryIndonesia                        -1.661492e+00 1.222681e+00 -1.35889194
countryIran                             -1.323624e+00 1.099823e+00 -1.20348733
countryIreland                          -5.043747e-01 8.274385e-01 -0.60956154
countryIsrael                           -7.438772e-01 1.549132e+00 -0.48018959
countryItaly                            -1.652584e+00 8.326299e-01 -1.98477625
countryJamaica                           2.626575e-01 1.546768e+00  0.16981054
countryJapan                            -3.554537e-01 8.151489e-01 -0.43605984
countryKenya                            -2.308265e+00 1.548169e+00 -1.49096461
countryLebanon                           9.549341e-01 1.545844e+00  0.61774304
countryMalta                             1.328329e+00 1.547305e+00  0.85847888
countryMexico                           -1.954613e-01 8.733279e-01 -0.22381203
countryNetherlands                       2.179255e-01 1.095620e+00  0.19890614
countryNew Zealand                      -1.199982e+00 8.363869e-01 -1.43472175
countryNorway                           -1.276532e+00 1.096006e+00 -1.16471312
countryPanama                           -1.287229e+00 1.548092e+00 -0.83149361
countryPortugal                         -9.858424e-01 1.546629e+00 -0.63741350
countryRepublic of Macedonia            -3.673155e-01 1.562312e+00 -0.23511023
countryRussia                           -2.674344e-01 1.022918e+00 -0.26144259
countrySouth Africa                      3.774146e-01 1.023357e+00  0.36880053
countrySouth Korea                      -8.707142e-01 8.583064e-01 -1.01445625
countrySpain                            -1.746791e-01 8.359798e-01 -0.20895136
countrySweden                           -5.264990e-01 9.787814e-01 -0.53791279
countrySwitzerland                      -1.853380e+00 1.023491e+00 -1.81084226
countryTaiwan                            6.764702e-01 1.101166e+00  0.61432171
countryThailand                         -4.093839e+00 1.546742e+00 -2.64675010
countryUnited Arab Emirates              3.824634e-01 1.223921e+00  0.31249014
countryUnited Kingdom                   -4.497461e-01 7.762470e-01 -0.57938538
countryUnited States                    -7.407536e-02 7.743836e-01 -0.09565720
countryWest Germany                      4.255948e-01 1.094846e+00  0.38872586
countryYugoslavia                       -5.590707e-01 1.547850e+00 -0.36119169
                                             Pr(>|t|)
(Intercept)                              4.534709e-19
budget                                  5.499708e-106
genreAdventure                           2.941262e-03
```

genreAnimation	2.775762e-01
genreBiography	5.964033e-08
genreComedy	6.333821e-03
genreCrime	2.147021e-08
genreDrama	9.756073e-18
genreFamily	7.726822e-01
genreFantasy	7.547734e-01
genreHorror	3.979915e-04
genreMystery	2.558905e-01
genreRomance	1.132929e-02
genreSci-Fi	2.573424e-01
genreThriller	7.936582e-01
genreWestern	4.827773e-01
ratingG	3.629510e-01
ratingNC-17	6.553395e-02
ratingNot Rated	1.761144e-03
ratingPG	3.119980e-01
ratingPG-13	2.352683e-01
ratingR	1.100252e-01
ratingTV-MA	2.585815e-02
ratingUnrated	3.844955e-03
ratingX	2.816307e-01
score	3.063933e-09
votes	5.669232e-80
year	2.556568e-43
runtime	4.544409e-11
countryAruba	8.591319e-01
countryAustralia	8.432006e-01
countryAustria	9.977362e-01
countryBelgium	1.124839e-01
countryBrazil	7.145055e-01
countryCanada	7.089029e-01
countryChile	7.910387e-01
countryChina	8.161608e-01
countryColombia	9.561531e-01
countryCzech Republic	4.359136e-01
countryDenmark	5.168462e-01
countryFederal Republic of Yugoslavia	5.067809e-02
countryFinland	7.071296e-01
countryFrance	5.438803e-01
countryGermany	9.208300e-01
countryHong Kong	9.608512e-01
countryHungary	9.848210e-01

```
countryIceland                          1.378361e-02
countryIndia                            3.707659e-01
countryIndonesia                        1.742382e-01
countryIran                             2.288411e-01
countryIreland                          5.421782e-01
countryIsrael                           6.311122e-01
countryItaly                            4.722041e-02
countryJamaica                          8.651656e-01
countryJapan                            6.628109e-01
countryKenya                            1.360298e-01
countryLebanon                          5.367711e-01
countryMalta                            3.906666e-01
countryMexico                           8.229121e-01
countryNetherlands                      8.423437e-01
countryNew Zealand                      1.514249e-01
countryNorway                           2.441871e-01
countryPanama                           4.057320e-01
countryPortugal                         5.238828e-01
countryRepublic of Macedonia            8.141322e-01
countryRussia                           7.937613e-01
countrySouth Africa                     7.122910e-01
countrySouth Korea                      3.104110e-01
countrySpain                            8.344942e-01
countrySweden                           5.906597e-01
countrySwitzerland                      7.022139e-02
countryTaiwan                           5.390289e-01
countryThailand                         8.150768e-03
countryUnited Arab Emirates             7.546802e-01
countryUnited Kingdom                   5.623536e-01
countryUnited States                    9.237964e-01
countryWest Germany                     6.974944e-01
countryYugoslavia                       7.179704e-01
```

```
summary(m5)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: log_gross ~ 1 + budget + (1 | genre) + (1 | rating) + score +
    votes + year + runtime + (1 | country)
   Data: data

REML criterion at convergence: 18739.2
```

```
Scaled residuals:
    Min      1Q  Median      3Q     Max
-8.0865 -0.4382  0.1522  0.6374  3.1248

Random effects:
 Groups   Name        Variance Std.Dev.
 country  (Intercept) 0.16078  0.4010
 genre    (Intercept) 0.08254  0.2873
 rating   (Intercept) 1.38342  1.1762
 Residual             1.79246  1.3388
Number of obs: 5423, groups:  country, 50; genre, 15; rating, 10

Fixed effects:
              Estimate Std. Error t value
(Intercept) -4.021e+01  3.896e+00 -10.322
budget       1.430e-08  6.249e-10  22.887
score        1.453e-01  2.479e-02   5.862
votes        2.515e-06  1.298e-07  19.377
year         2.695e-02  1.926e-03  13.992
runtime      8.514e-03  1.318e-03   6.462

Correlation of Fixed Effects:
        (Intr) budget score  votes  year
budget   0.268
score   -0.055  0.171
votes    0.106 -0.320 -0.462
year    -0.993 -0.266  0.025 -0.086
runtime -0.120 -0.359 -0.295 -0.061  0.097
fit warnings:
Some predictor variables are on very different scales: consider rescaling
```

While adjusting for the random intercept based on country variable and holding all other variables constant, we noticed that the estimated variance is 0.1623 which means that all the country-specific intercept is distributed around the model's overall intercept within estimated variance of 0.1623, given that our unit is log dollars, and our estimated variance is quite small, we may expect that the country is similar. In terms of our interest in knowing whether a movie is successful or not, production of country may not be a big effect on the movie's success.

Since we noticed that genre has the smallest estimated variance (0.08254), we may say that compared to the difference between ratings and countries, the difference between genres is smaller, while rating has the biggest difference. Given that our unit is log dollars, and all of

our estimated variance is quite small, the differences between these variables should be similar to each other.

Interested Hypothesis Test 1:

Question 1: Is there evidence to suggest that log(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: $p_1 \neq 0$ 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 log(budget) does have an effect on the success of the movies (log(gross)).

In terms of qualitatively addressing this issue: in our model, we know that every 1 million increase in our budget, the gross value is expected to increase by e^{cc} dollars. ## Interested Hypothesis Test 2: Aim 2: Is there evidence to suggest that log(Score) has an effect on the success of the movies? Null hypothesis: $p_2 = 0$ There isn't sufficient evidence to suggest that Score is associated with movies' gross, while controlling for all of the variables. Alternative hypothesis: $p_2 \neq 0$ There is sufficient evidence to suggest that Score 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 3.06e-09 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 genre does have an effect on the success of the movies.

Summary based on Hypothesis:

Based on our previous two hypothesis, we noticed that XXXXX.

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.

```
set.seed(123)
dim(data)
```

```
[1] 5423   17
```

```
indices <- sample(1:5423, size = 5423 * 0.8, replace = F)
train.data  <- data %>%
  slice(indices)
test.data <- data %>%
  slice(-indices)
dim(train.data)
```

```
[1] 4338   17
```

```
library(caret)
cv_method <- trainControl(method = "cv", number = 10,
                          repeats = 5)
m1 <- train(log_gross ~ budget + genre + rating + score + votes + year + runtime + country
m2 <- train(log_gross ~ budget + genre + rating + score + votes + year + runtime + country

print(m2)
```

```
Linear Regression

5423 samples
   8 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 4883, 4879, 4882, 4881, 4881, 4880, ...
Resampling results:

  RMSE      Rsquared   MAE
  1.349604  0.4908407  0.9867613

Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
library(caret)
cv_method <- trainControl(method = "cv", number = 10,
                          repeats = 5)
m3 <- train(log_gross ~ log(budget) + genre + rating + log(score) + log(votes) + year + lo
m4 <- train(log_gross ~ log(budget) + genre + rating + log(score) + log(votes) + year + lo
```

```r
print(m4)
```

```
Linear Regression

5423 samples
   8 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 4881, 4879, 4879, 4883, 4881, 4881, ...
Resampling results:

  RMSE      Rsquared   MAE
  1.081826  0.6724693  0.7606395

Tuning parameter 'intercept' was held constant at a value of TRUE
```

The RMSE from first model is XXX; while the RMSE from the second model is XXXX. Since the second one is smaller, the second one is better performing model in predicting the gross values of the movie. XXXX Since better, we might use log/ or fewer variables in predicting our gross values.

Missing Data Analysis:

XXXX

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 & Limitations

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