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  3.1.8      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()
```

```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()
* Search for functions across packages at https://www.tidymodels.org/find/
```

```r
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. 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()
```

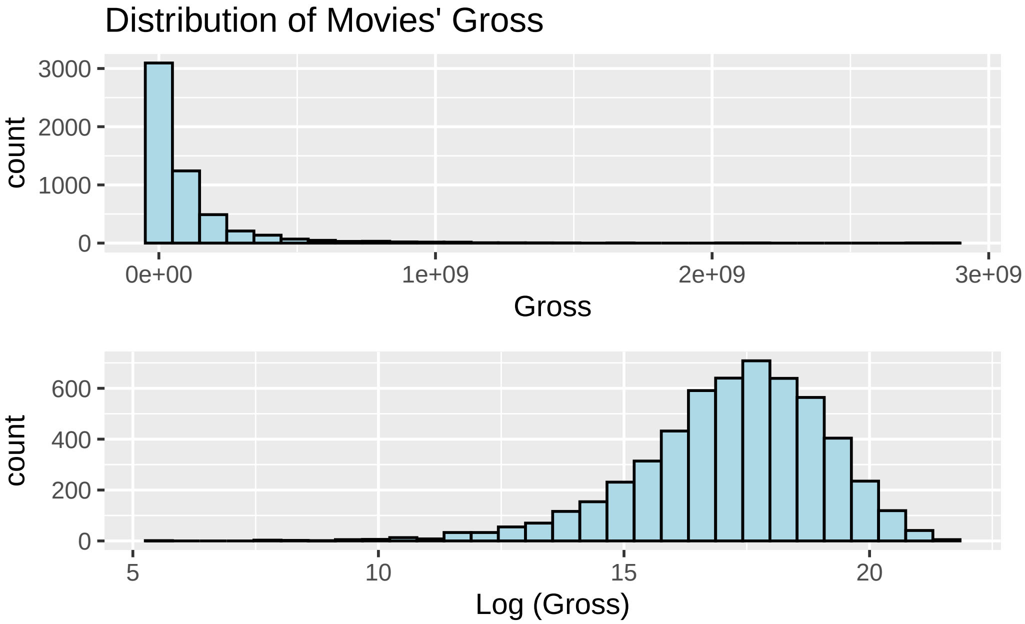
mean_gross	median_gross	sd_gross	min_gross	max_gross
103192280	36850101	187278279	309	2847246203

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

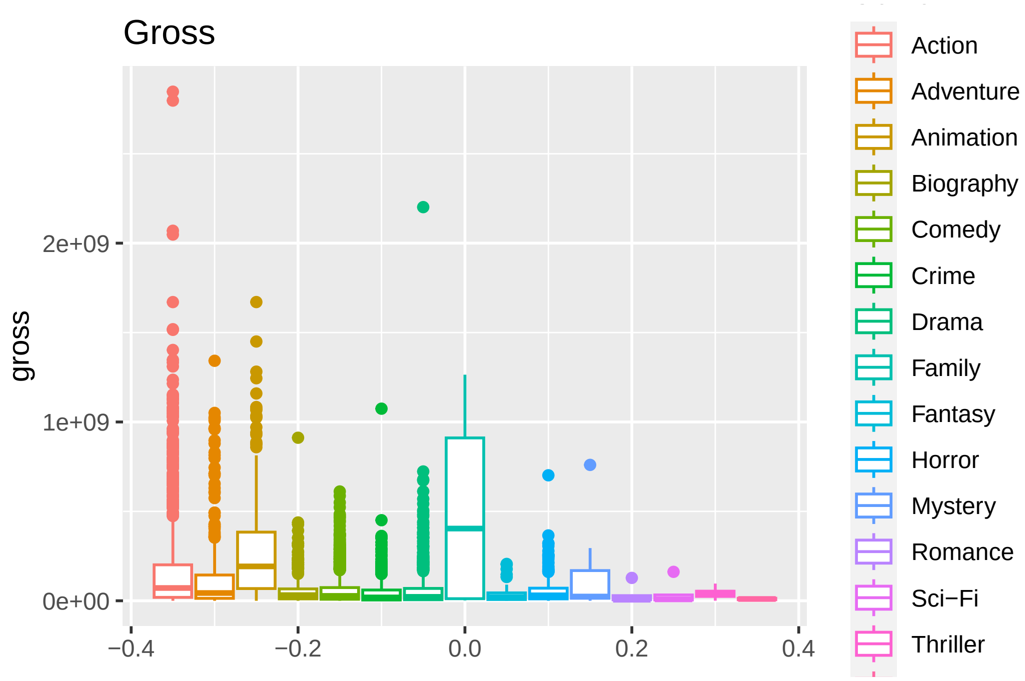
```
data <- data %>%
  mutate(log_gross = log(gross)) %>%
  mutate(mean = mean(log_gross))

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
```

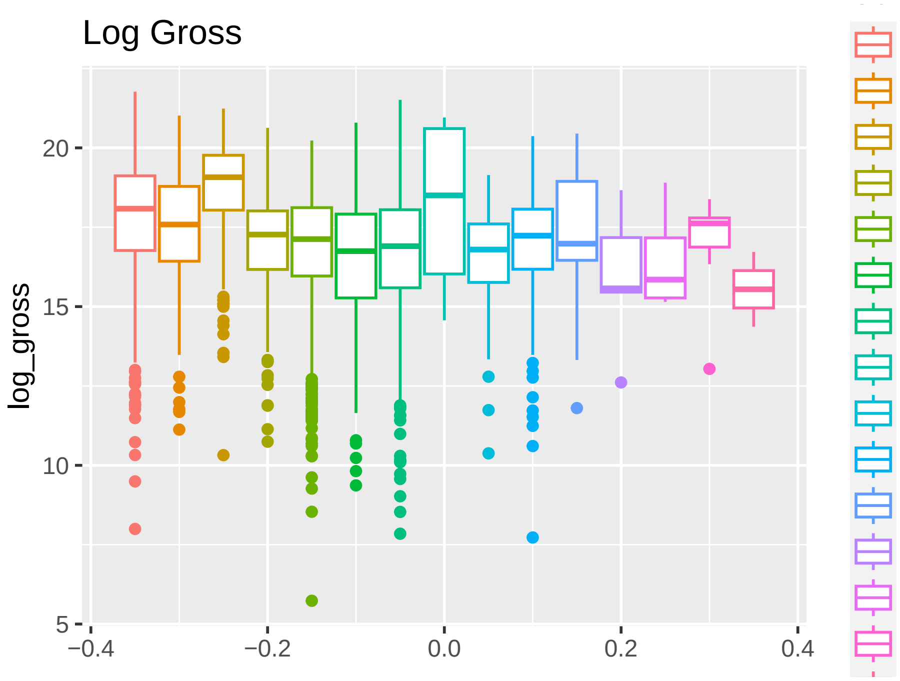
```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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



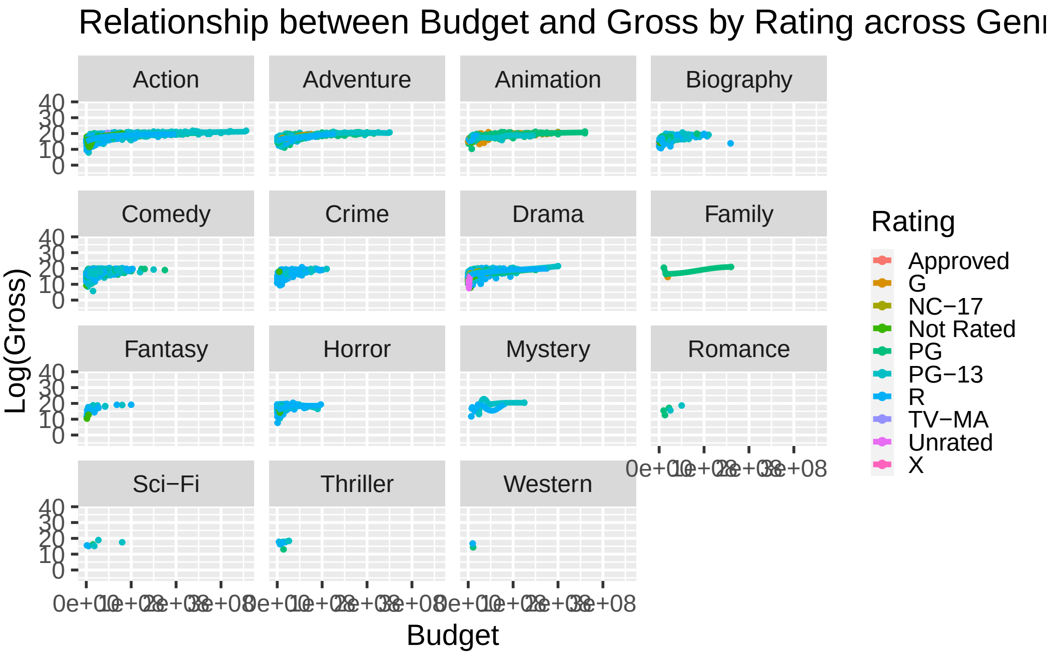
```
ggplot(data = data) +
  geom_boxplot(aes(y = log_gross, color = genre)) +
  labs(title = "Log Gross",
       color = "Genre")
```



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.

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

```
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"))
```

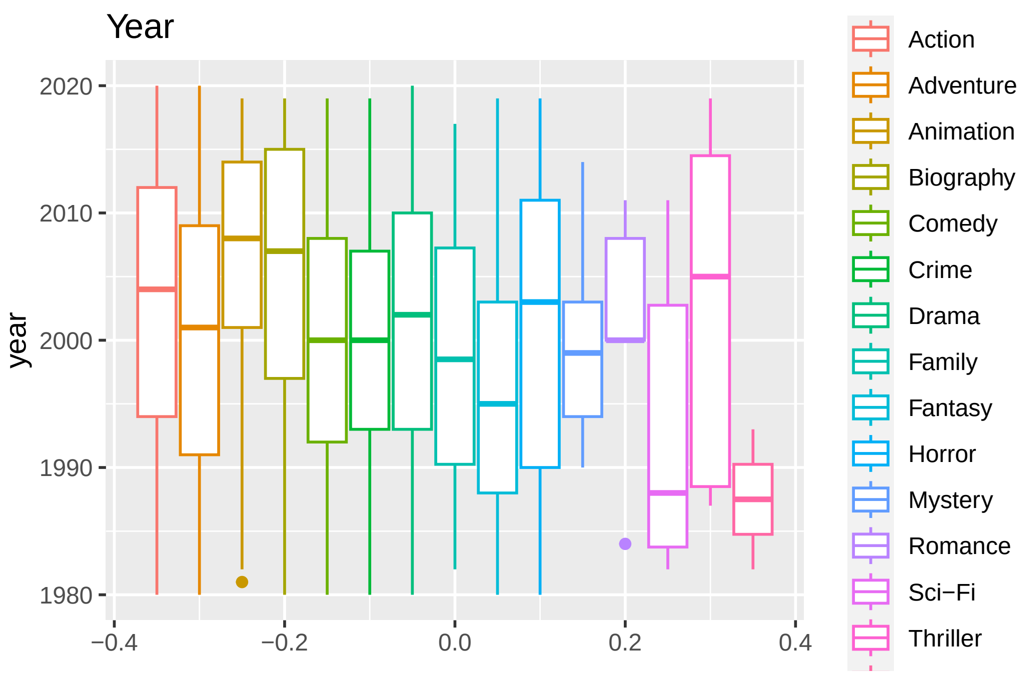


We observe that the relationship between budget and gross is vastly different across genres: Action, adventure, animation movies have a steep slope and generally high budget spans, with outliers which have exceedingly high budget and relatively high gross values. On the other hand, genres such as horror,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

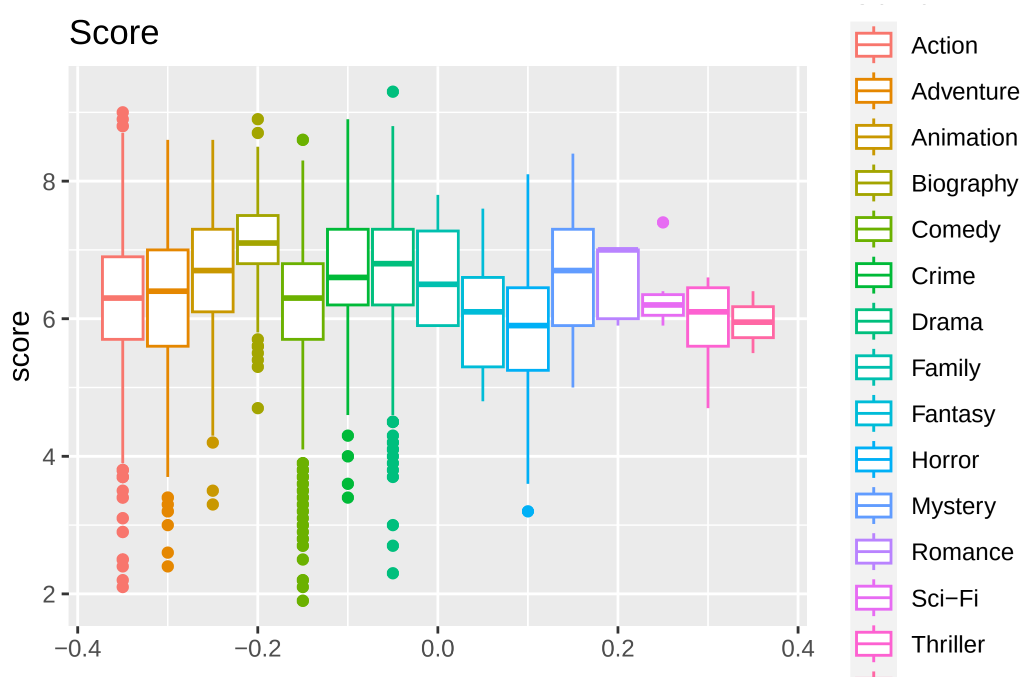
EDA: Visualizatioins and Summary Statistics

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.

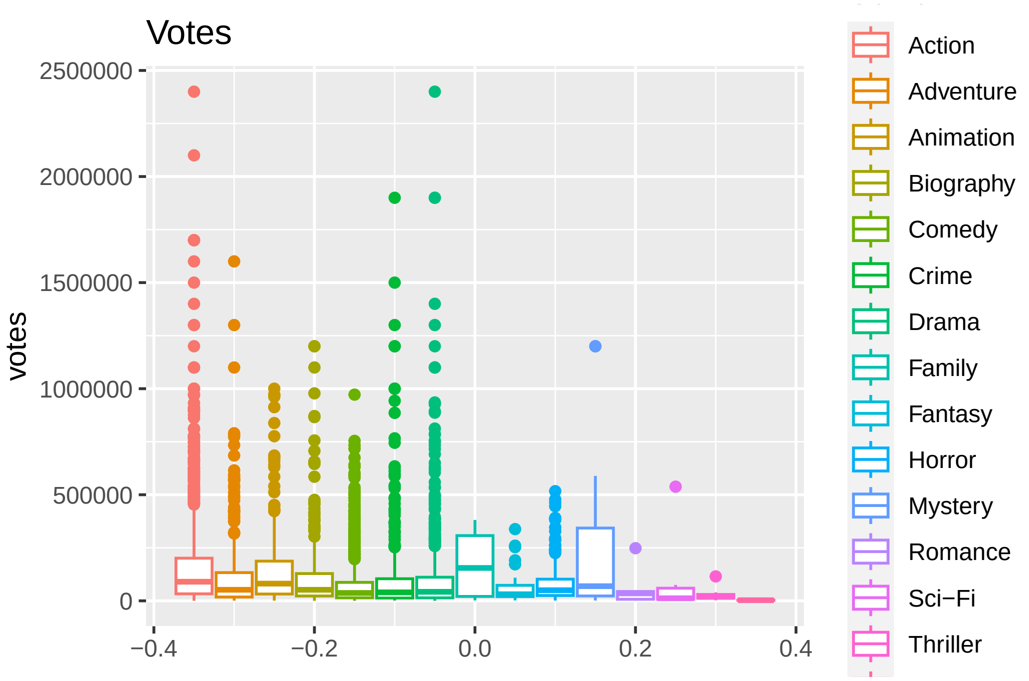
```
ggplot(data = data) +
  geom_boxplot(aes(y = year, color = genre)) +
  labs(title = "Year",
       color = "Genre")
```



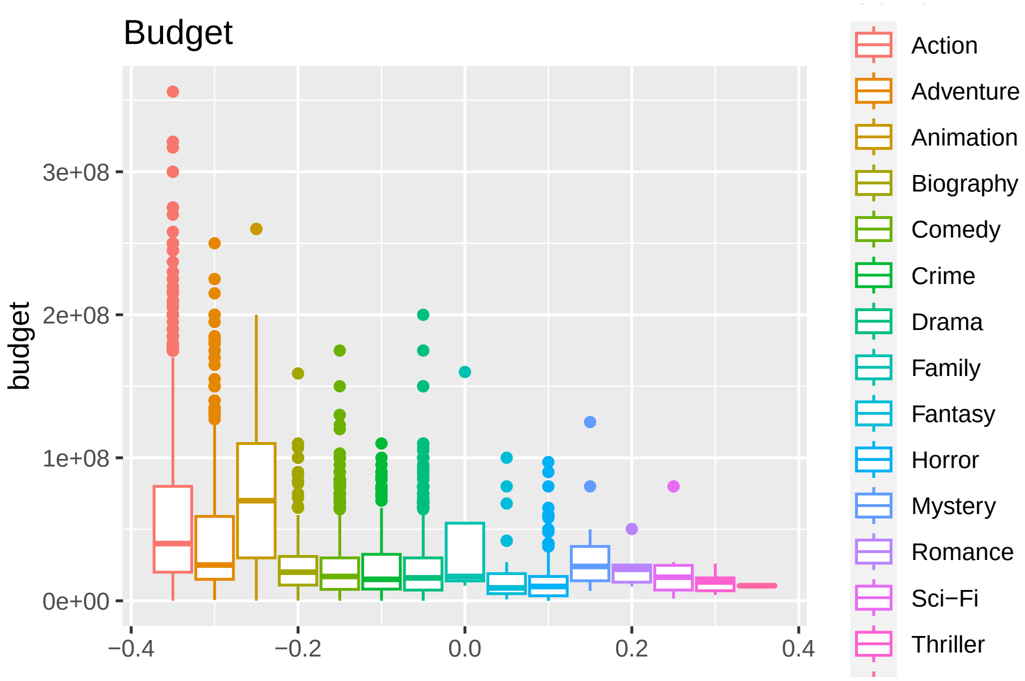
```
ggplot(data = data) +
  geom_boxplot(aes(y = score, color = genre)) +
  labs(title = "Score",
       color = "Genre")
```



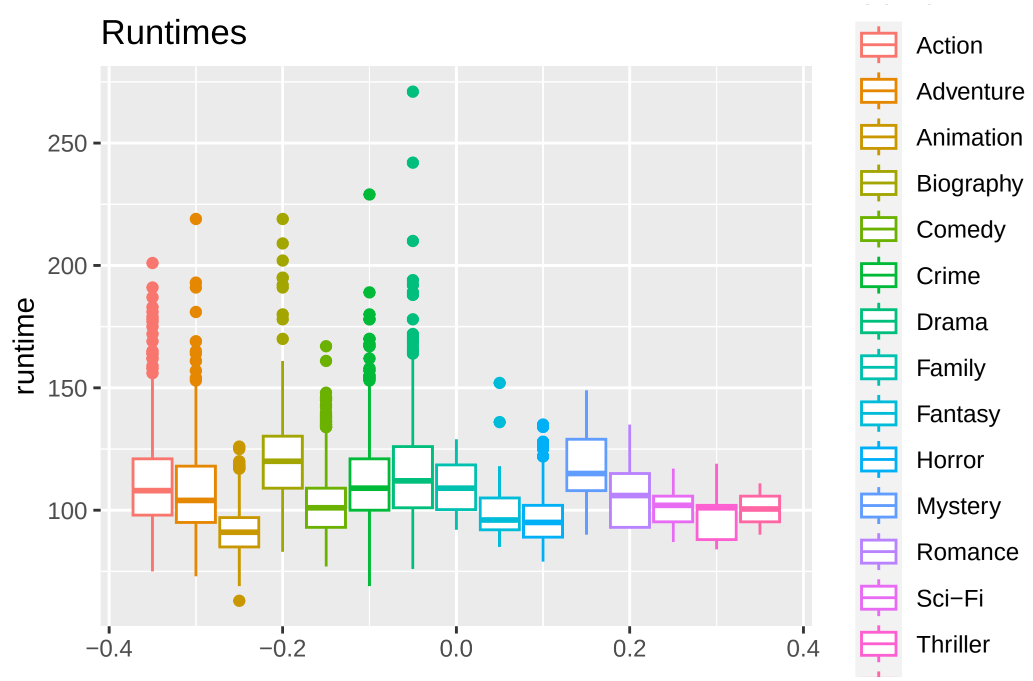
```
ggplot(data = data) +
  geom_boxplot(aes(y = votes, color = genre)) +
  labs(title = "Votes",
       color = "Genre")
```



```
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")
```



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.

Linear Regression of all variables

```
lm1 <- lm(log_gross ~ budget + genre + rating + score + votes + year + runtime, data = dat
summary(lm1)
```

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

Residuals:
     Min       1Q   Median       3Q      Max
```

```
-10.8061  -0.6074   0.2093   0.8735   4.0636 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -3.570e+01  4.070e+00  -8.772  < 2e-16 ***
budget           1.460e-08  6.286e-10  23.227  < 2e-16 ***
genreAdventure  -3.098e-01  8.703e-02  -3.560 0.000374 ***
genreAnimation   1.146e-01  1.138e-01   1.007 0.313913    
genreBiography  -5.381e-01  9.088e-02  -5.921 3.40e-09 ***
genreComedy     -1.338e-01  5.375e-02  -2.489 0.012848 *  
genreCrime      -4.354e-01  8.028e-02  -5.423 6.11e-08 ***
genreDrama      -5.520e-01  6.306e-02  -8.754  < 2e-16 ***
genreFamily     -1.275e-01  6.785e-01  -0.188 0.850906    
genreFantasy    -8.261e-02  2.151e-01  -0.384 0.700991    
genreHorror      3.453e-01  9.628e-02   3.586 0.000338 ***
genreMystery    -4.626e-01  3.302e-01  -1.401 0.161379    
genreRomance    -1.605e+00  6.060e-01  -2.648 0.008110 ** 
genreSci-Fi     -6.298e-01  5.527e-01  -1.139 0.254582    
genreThriller    1.402e-01  5.121e-01   0.274 0.784189    
genreWestern    -7.965e-01  9.557e-01  -0.833 0.404654    
ratingG         -1.345e+00  1.361e+00  -0.988 0.323171    
ratingNC-17     -2.675e+00  1.410e+00  -1.897 0.057867 .  
ratingNot Rated -4.386e+00  1.371e+00  -3.199 0.001385 ** 
ratingPG        -1.434e+00  1.354e+00  -1.059 0.289615    
ratingPG-13     -1.664e+00  1.355e+00  -1.228 0.219431    
ratingR         -2.231e+00  1.355e+00  -1.646 0.099760 .  
ratingTV-MA     -3.512e+00  1.659e+00  -2.117 0.034292 *  
ratingUnrated   -4.146e+00  1.395e+00  -2.973 0.002960 ** 
ratingX         -2.427e+00  1.915e+00  -1.267 0.205273    
score            1.346e-01  2.490e-02   5.404 6.80e-08 ***
votes            2.536e-06  1.304e-07  19.452  < 2e-16 ***
year             2.622e-02  1.931e-03  13.575  < 2e-16 ***
runtime          8.040e-03  1.324e-03   6.075 1.32e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.35 on 5394 degrees of freedom
Multiple R-squared:  0.4914,	Adjusted R-squared:  0.4887 
F-statistic: 186.1 on 28 and 5394 DF,  p-value: < 2.2e-16
```

Predicting moves' gross

Residual Plots

Hypothesis Test 1:

Aim 1: Budget does NOT have an effect on the success of the movies. Null hypothesis: Alternative hypothesis:

Hypothesis Test 2:

Aim 2: Genre has an effect on the success of the movies Null hypothesis: Alternative hypothesis:

Hypothesis Test 3:

Aim 3: Longer runtime has an effect the success of the movies.

Other Models:

Forward-backward model, LASSO, Linear mixed models

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