# Final Project

Christina Yu, Damian Kim

## Read in the data

## Introduction and data

With the rise in the age of information, the processing speeds of data are accelerated to impossible levels. One of the densest forms of media content out there is the transcription of words into pixels. A book that previously took 2 weeks to finish now fits in the span of 2 hours in the insanely digestible form of: movies. The movie industry is worth 95 billion dollars in the US alone, and indeed there is a great demand for quality movies. Though it is not easily apparent to tell what makes a movie great, there are some basic data that might help determine some of the factors of a great movie. We decided to try and tackle this heavy yet essential question. What factors might help make a movie successful? Throughout this project we can consider this question through the lens of a potential movie investor.

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

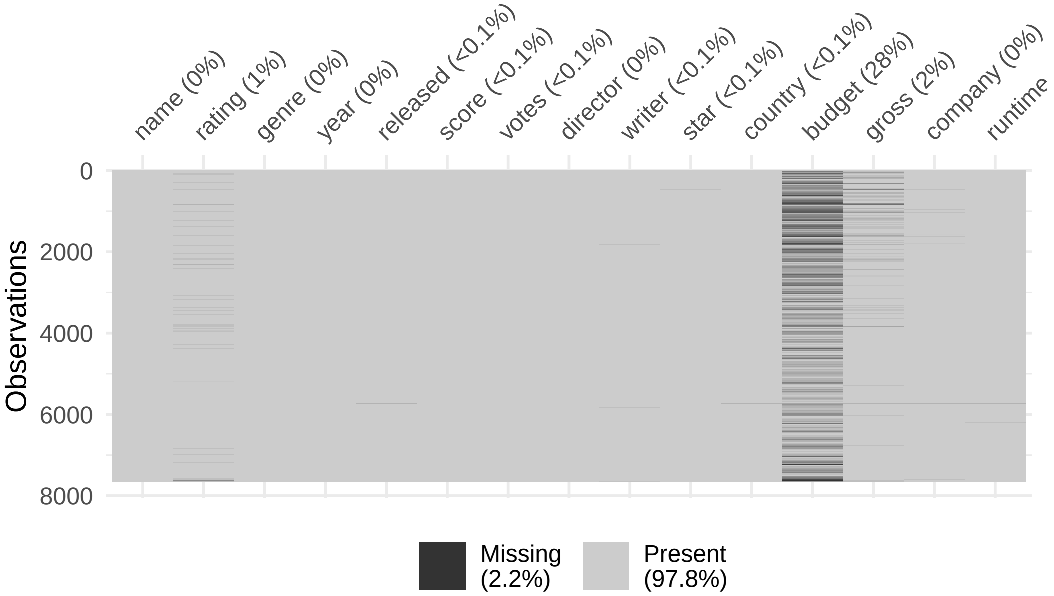
– `score`: IMDb user rating

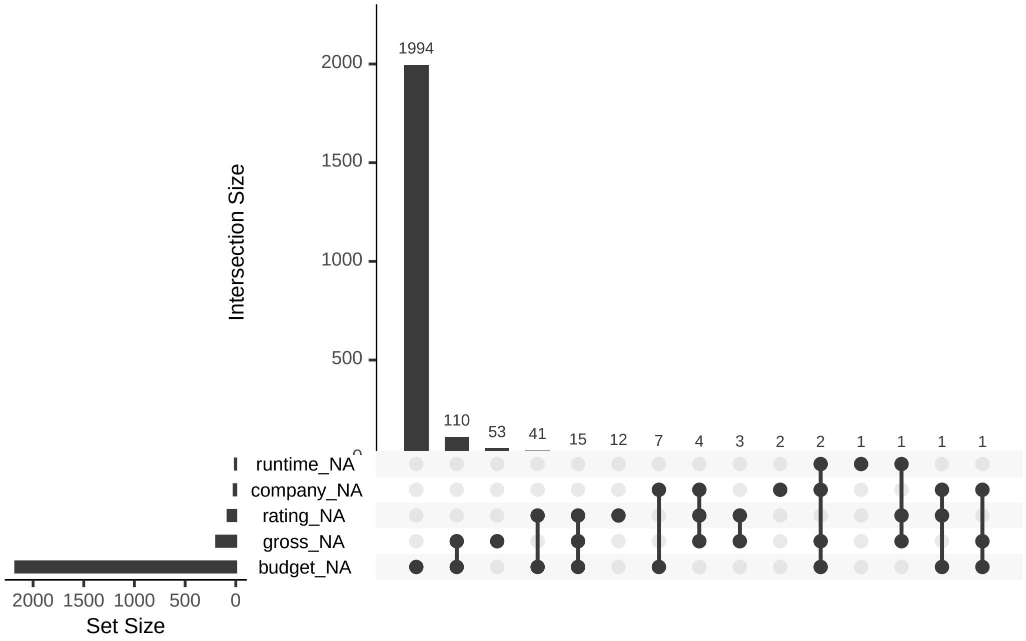
– `votes`: number of user votes

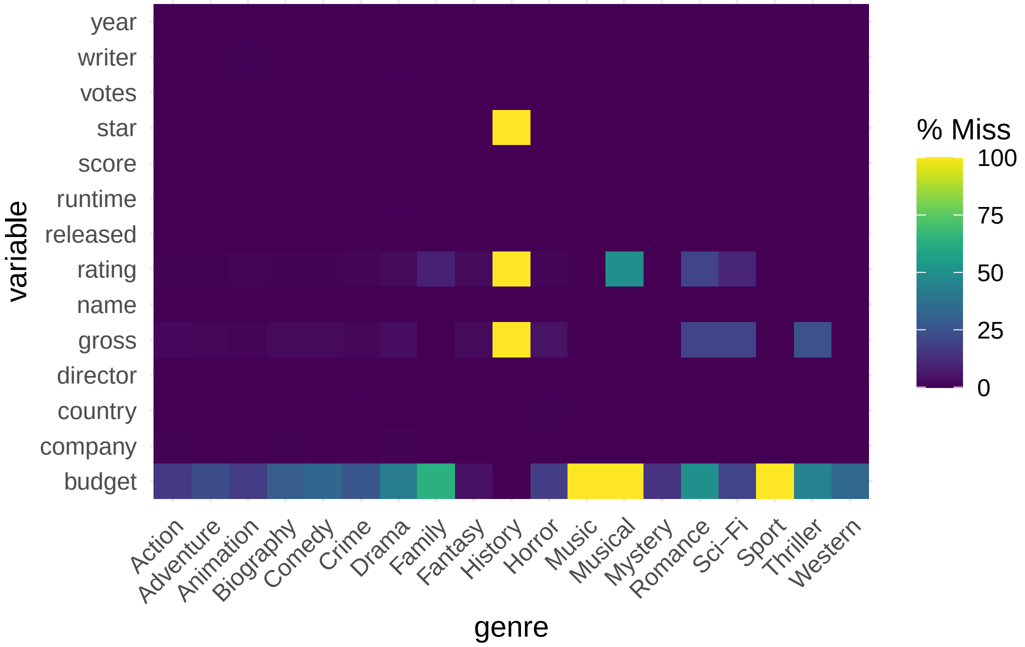
- `star`: main actor/actress
- `writer`: writer of the movie
- `year`: year of release

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

First we will explore missingness in our dataset.







The dataset includes movies that haven't been published yet (which causes a missing gross value). In our project we are considering gross revenue of a movie as the defining indicator of success. Considering that, we removed all observations that haven't been published yet or are missing gross values, because we are interested only in movies with a quantitative measure of success.

In fact, we also decided to rid our dataset of missing values for the other predictor variables as well. About 28% of the observations were missing a value for budget, whereas the next most missing variable was gross, with 2%. From elementary missingness analysis there did not seem to be very strong, numerically significant relationships between the missingness of each variable with other variables of interest. Though it is unlikely that the missingness of our budget data in particular was MCAR (generally unlikely in the real world, perhaps lower budgets were less public and thus less likely to have solid budget details), we decided to do a complete case analysis. From our data it was hard to tell if the dataset could be assumed to be MAR. Though there might be some bias introduced, since we were working with such a large number of movies, we decided it was okay to lose some validity of our model. Obviously this is a big limitation of our project, and if such a project were done again we would be more meticulous. Now we have 5423 observations in the dataset.

### The Predictor Variables

We will use the `budget`, `genre`, `rating`,`score votes`, `runtime`, and `country` variables as predictors. Among them, `budget`, `score votes`, and `runtime` are numerical variables, while `genre`, `rating`, and `country` are categorical variables.

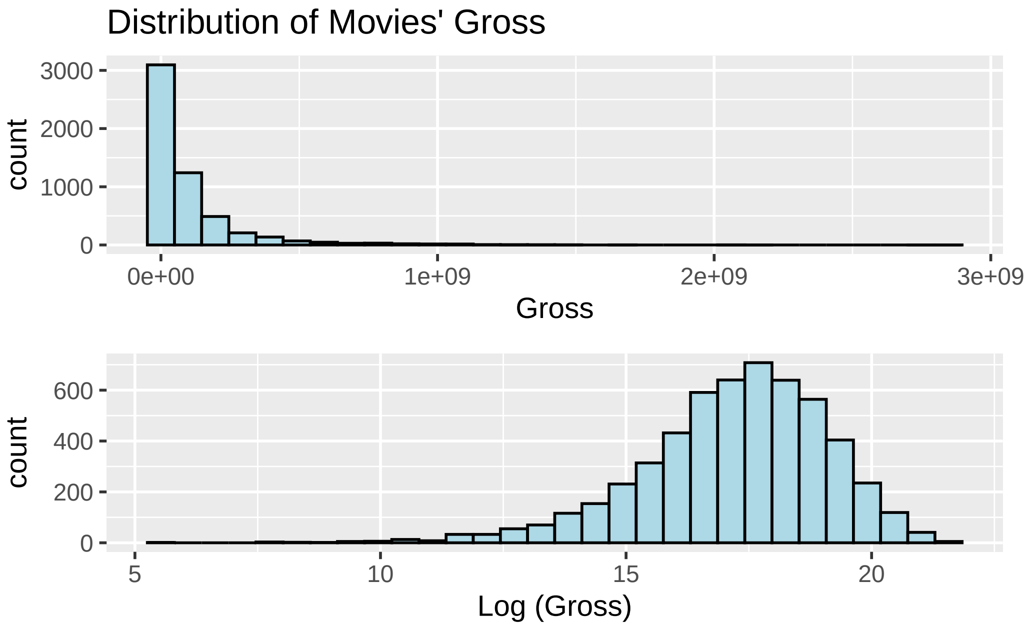
### The Response Variable

As described earlier, the `gross` numerical variable is our response variable. 1. Summary of the `gross` variable:

| mean_gross | median_gross | sd_gross | min_gross | max_gross |
|---:|---:|---:|---:|---:|
| 103192280 | 36850101 | 187278279 | 309 | 2847246203 |

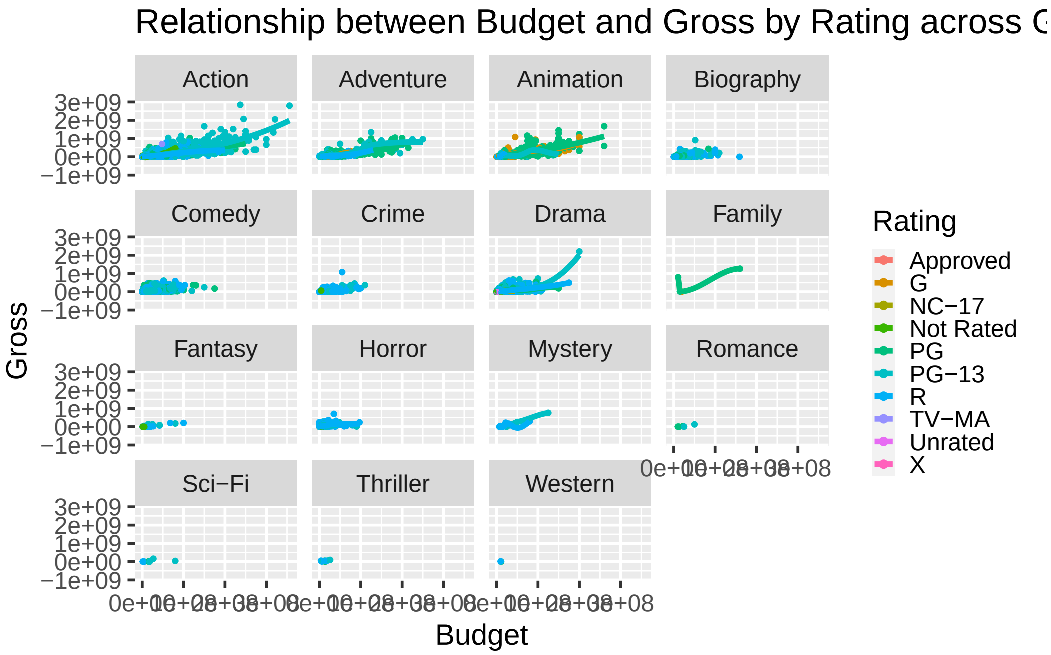
2. Log-transformation and Distribution of the `gross` variable:

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



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, following a roughly normal distribution, with a mean at 17.2102, and there exist some outliers on the left end.

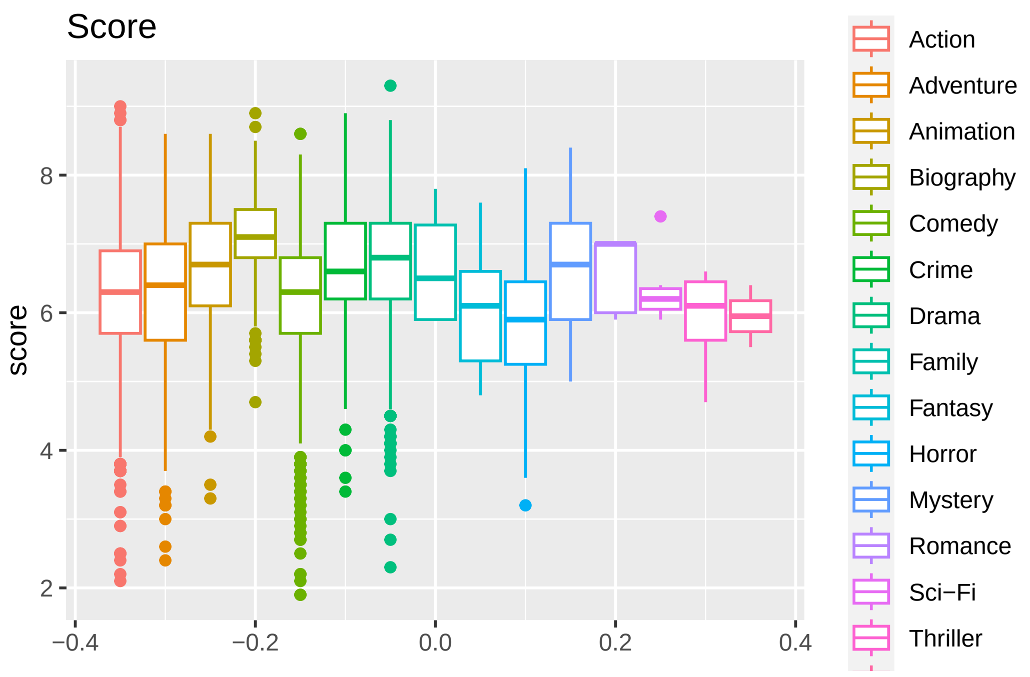
3. Relationship between Gross and Budget based on different Genres & Ratings:

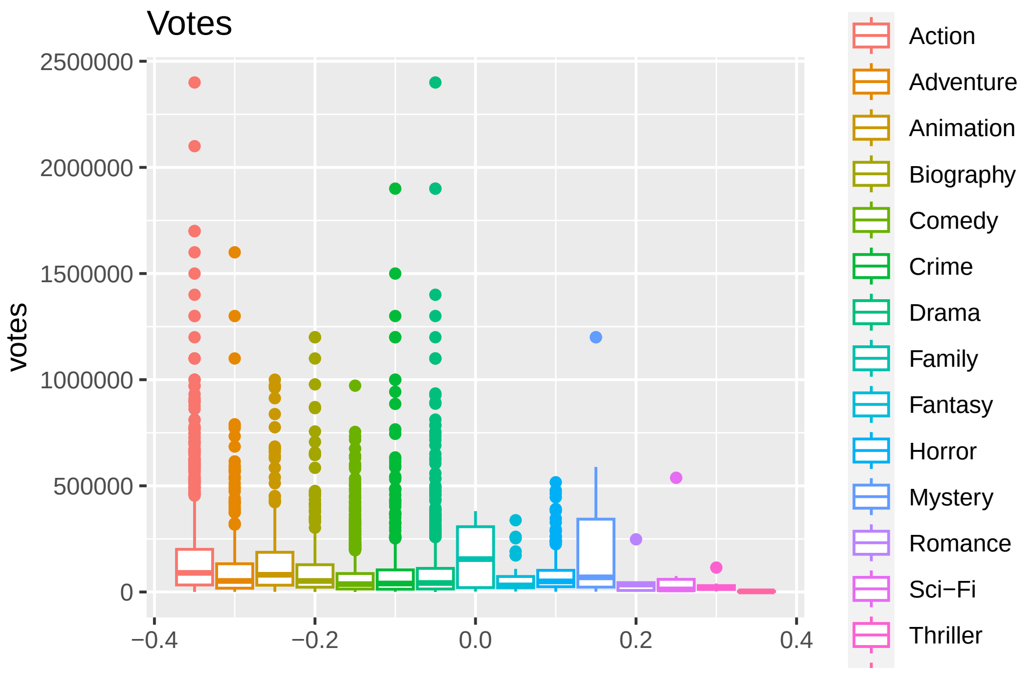


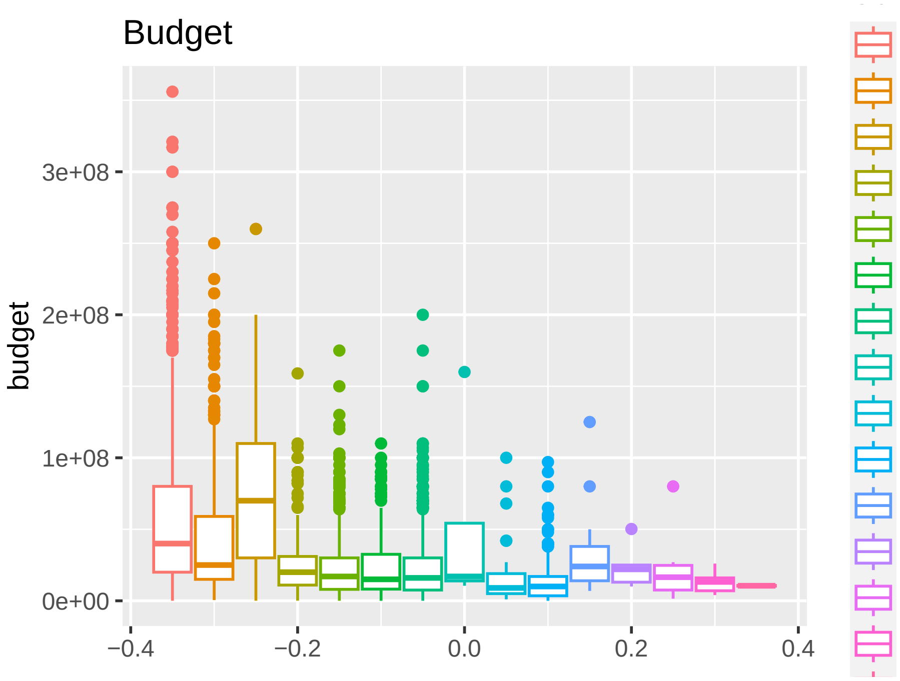
We observe that the relationship between budget and gross is vastly different across genres: action, adventure, animation movies have a steep slope and generally high budget ranges, with outliers which have exceedingly high budget and relatively high gross values. Noticeably, there are also a lot more movies in these three categories, with the most in action. On the other hand, genres such as horror, mystery and romance have a visibly flatter slope, which corresponds to the industry knowledge that certain genres are more conducive to low-budget film making than others, even when provided with additional funding. Thus, we're interested in further exploring the relationship between budget, genre, and success.

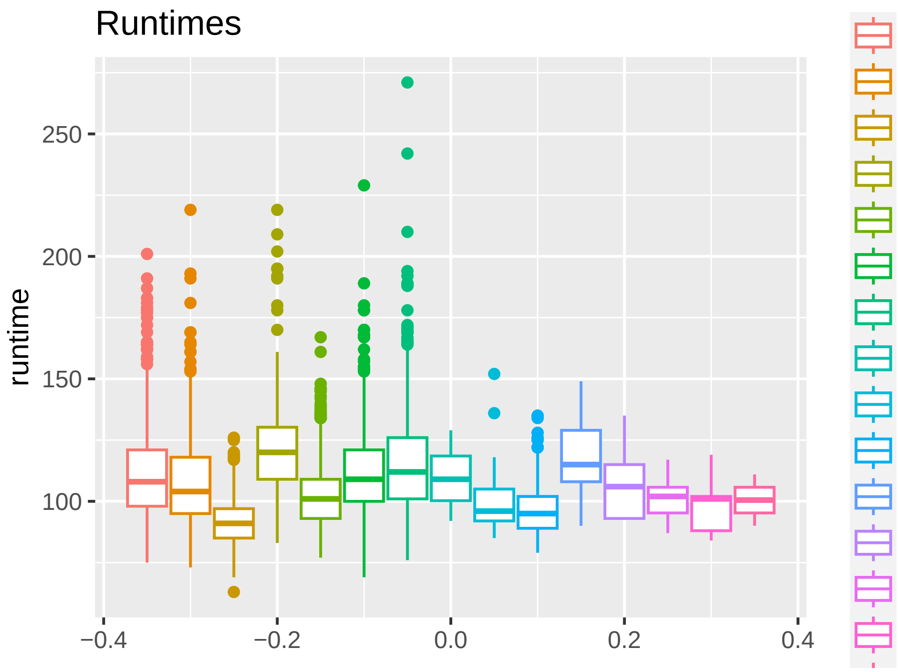
## EDA: Visualizatioins and Summary Statistics

The dataset contains many interesting variables that we want to explore fully. We will first do some elementary exploratory data analysis on our datasets to show potential insights that we can explore further.


Score
score
8
6
4
2
−0.4
−0.2
0.0
0.2
0.4
Action
Adventure
Animation
Biography
Comedy
Crime
Drama
Family
Fantasy
Horror
Mystery
Romance
Sci−Fi
Thriller

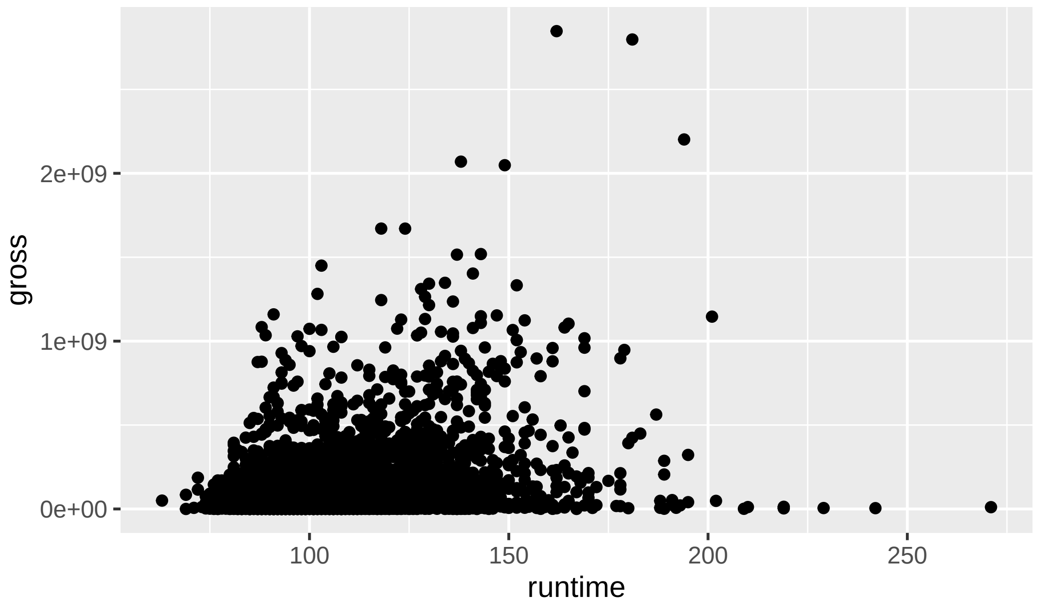

Votes
votes
2500000
2000000
1500000
1000000
500000
0
−0.4
−0.2
0.0
0.2
0.4
Action
Adventure
Animation
Biography
Comedy
Crime
Drama
Family
Fantasy
Horror
Mystery
Romance
Sci−Fi
Thriller



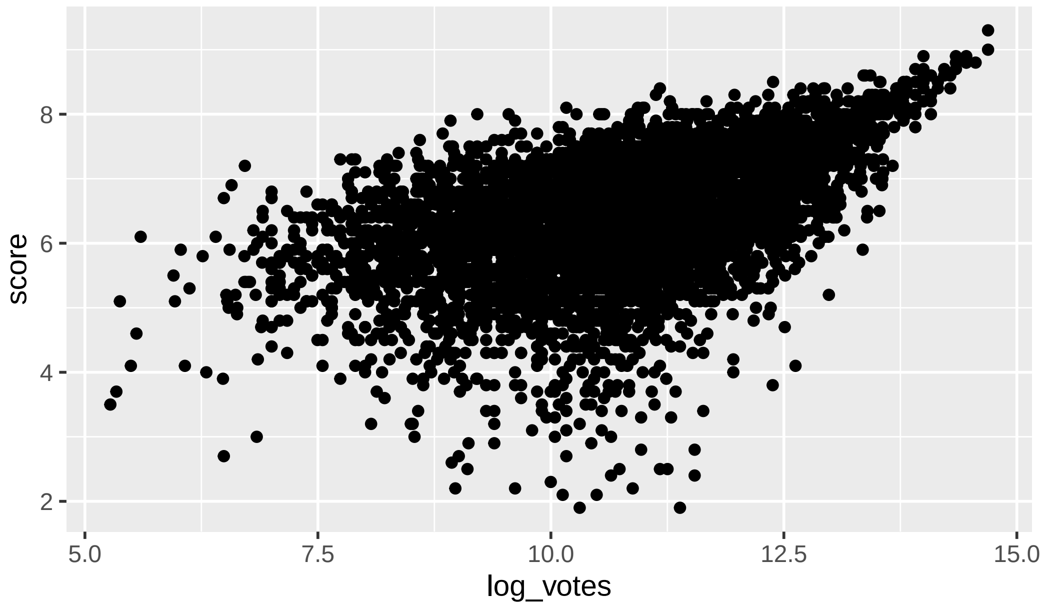


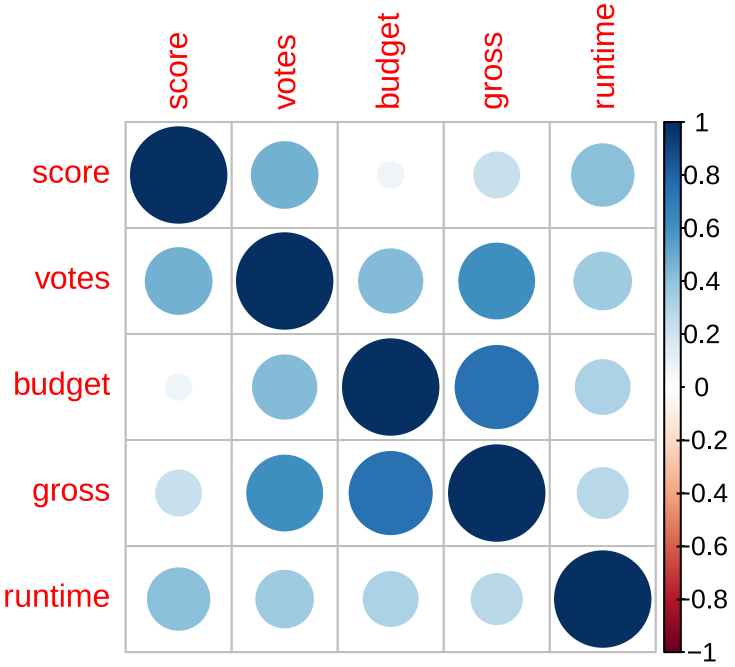
As can be seen, the score variable is relatively unskewed, whereas the votes, budget, and runtime variables are heavily positively-skewed.

Runtime has no relationship with Gross



Log of votes has a converging, positive relationship with Score





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. Though there seem to be no particularly strong correlations between predictor variables, the strongest correlation is between votes and score. We may consider adding this into our model because it also fits with our domain knowledge that a movie that is rated by more people might generally have a higher score because, in our experience, people are more motivated to rate a movie they enjoy than one they dislike.

## Methodology

As a potential movie investor, we're obviously curious about the prediction of movies' success, and the factors that lead to success. Therefore, we want to explore the following topics in detail in order for us to be better investors: 1) Prediction of movies' gross value 2) the factors that affect the success of the movie. Some of the methods by which we are going to tackle this is through linear regression, residual plots, linear mixed models, LASSO, and repeated k-fold validation.

## LASSO (Variable Selection):

As mentioned above we are interested in the relationships between the variables and success, but we still wish to further filter variables that are more likely to have a significant effect on our model. Therefore, we will apply a LASSO variable selection process in order to determine significant variables to explore.

Some coefficients of the categories of the variables of genre, rating, and country have coefficients were shrunk to zero. Our data is processed in a way where an observation can take only one “level” of a category, for example, a movie can only have one genre. Since even one of the factors of the genre, rating, or country variables has been excluded, we will remove those factor variables as a whole. This is because LASSO performs feature selection, and for a factor variable such as genre, which is set with a baseline of “action”, a singular comparison of, say, adventure vs action is not a feature, rather, the comparison of the entire variable versus action is a feature. This is somewhat of an all or nothing approach, which has its limitation discussed later. Similar logic applies to rating, with a baseline of “approved”, and country, with a baseline of “Argentina”. Our model therefore contains, after LASSO selection, the variables of budget, score, votes, and runtime.

Importantly, we decided to create our model from all of our original predictor variables of interest, even though LASSO suggested the opposite. Because the number of parameters (7) is much less than the number of observations in our dataset (5000+) we didn’t find it necessary to remove any predictor variables, there was not that much risk for overfitting because of our low ratio of parameters to observations.

### Linear Regression of all variables

Gross revenue is a continuous response variable, so we will first examine a linear regression model for our data. We’ve gained the equation of XX.

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

Residuals:
    Min      1Q  Median      3Q     Max
-9.9916 -0.4562  0.0940  0.6026  4.3977

Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                            1.65159    1.33444   1.238  0.21590
log(budget)                            0.48416    0.01680  28.827  < 2e-16 ***
score                                 -0.14146    0.02143  -6.600 4.51e-11 ***
log(votes)                             0.74079    0.01397  53.042  < 2e-16 ***
log(runtime)                           0.62197    0.12708   4.894 1.02e-06 ***
genreAdventure                        -0.09990    0.06886  -1.451  0.14688
```

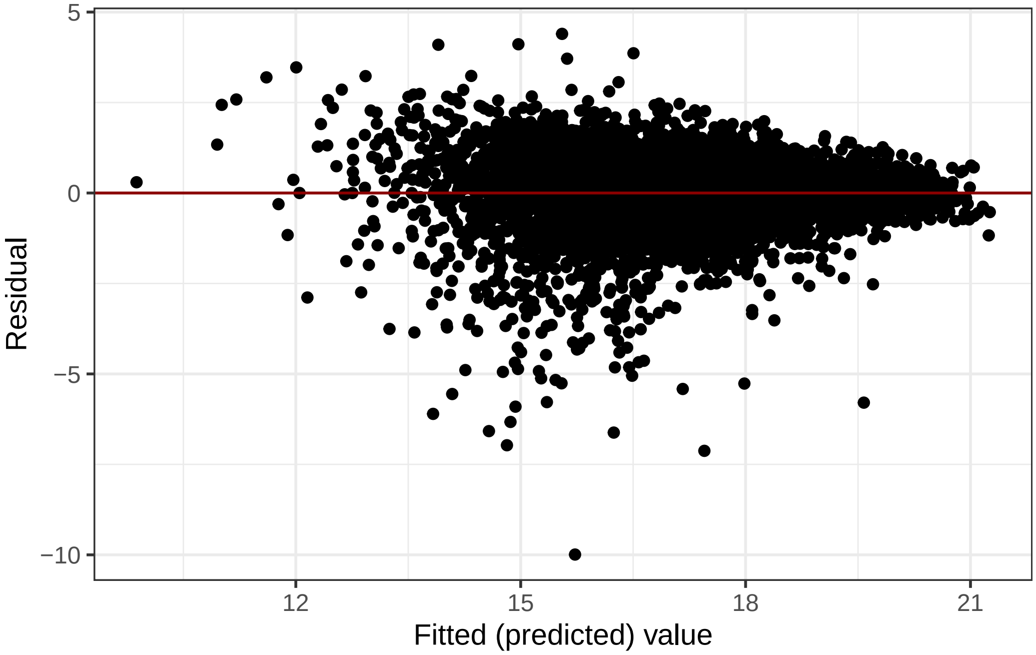
```
genreAnimation                         0.36501    0.08981   4.064 4.89e-05 ***
genreBiography                        -0.16805    0.07044  -2.386  0.01708 *
genreComedy                            0.04168    0.04211   0.990  0.32235
genreCrime                            -0.19741    0.06344  -3.112  0.00187 **
genreDrama                            -0.13719    0.04956  -2.768  0.00566 **
genreFamily                            0.24283    0.53302   0.456  0.64872
genreFantasy                          -0.02666    0.16931  -0.157  0.87489
genreHorror                            0.36552    0.07703   4.745 2.14e-06 ***
genreMystery                          -0.20585    0.26110  -0.788  0.43050
genreRomance                          -1.13613    0.47610  -2.386  0.01705 *
genreSci-Fi                            0.06207    0.43429   0.143  0.88635
genreThriller                          0.66646    0.40219   1.657  0.09757 .
genreWestern                           0.76878    0.75151   1.023  0.30636
ratingG                               -1.31831    1.06988  -1.232  0.21793
ratingNC-17                           -2.69747    1.10877  -2.433  0.01501 *
ratingNot Rated                       -3.23065    1.07863  -2.995  0.00276 **
ratingPG                              -1.48612    1.06390  -1.397  0.16251
ratingPG-13                           -1.78957    1.06419  -1.682  0.09270 .
ratingR                               -2.15700    1.06410  -2.027  0.04271 *
ratingTV-MA                           -3.53390    1.30565  -2.707  0.00682 **
ratingUnrated                         -2.70343    1.09542  -2.468  0.01362 *
ratingX                               -2.86686    1.50557  -1.904  0.05694 .
countryAruba                          -0.86525    1.22633  -0.706  0.48049
countryAustralia                      -0.68961    0.63525  -1.086  0.27771
countryAustria                        -0.69233    0.96881  -0.715  0.47488
countryBelgium                        -2.40101    0.81377  -2.950  0.00319 **
countryBrazil                         -0.78806    0.86622  -0.910  0.36298
countryCanada                         -0.78904    0.62207  -1.268  0.20471
countryChile                          -0.64255    1.22681  -0.524  0.60047
countryChina                           0.08394    0.64937   0.129  0.89715
countryColombia                        0.27768    1.22499   0.227  0.82068
countryCzech Republic                 -1.10920    0.73391  -1.511  0.13076
countryDenmark                        -1.33480    0.69940  -1.908  0.05638 .
countryFederal Republic of Yugoslavia -4.51586    1.23884  -3.645  0.00027 ***
countryFinland                         0.09465    0.97179   0.097  0.92242
countryFrance                         -1.13927    0.62229  -1.831  0.06719 .
countryGermany                        -0.88037    0.62518  -1.408  0.15913
countryHong Kong                      -0.31735    0.65134  -0.487  0.62612
countryHungary                        -0.68933    1.22649  -0.562  0.57411
countryIceland                        -2.76923    0.96937  -2.857  0.00430 **
countryIndia                           0.07689    0.73120   0.105  0.91625
countryIndonesia                      -1.43155    0.96846  -1.478  0.13942
countryIran                           -0.38380    0.87136  -0.440  0.65962
```

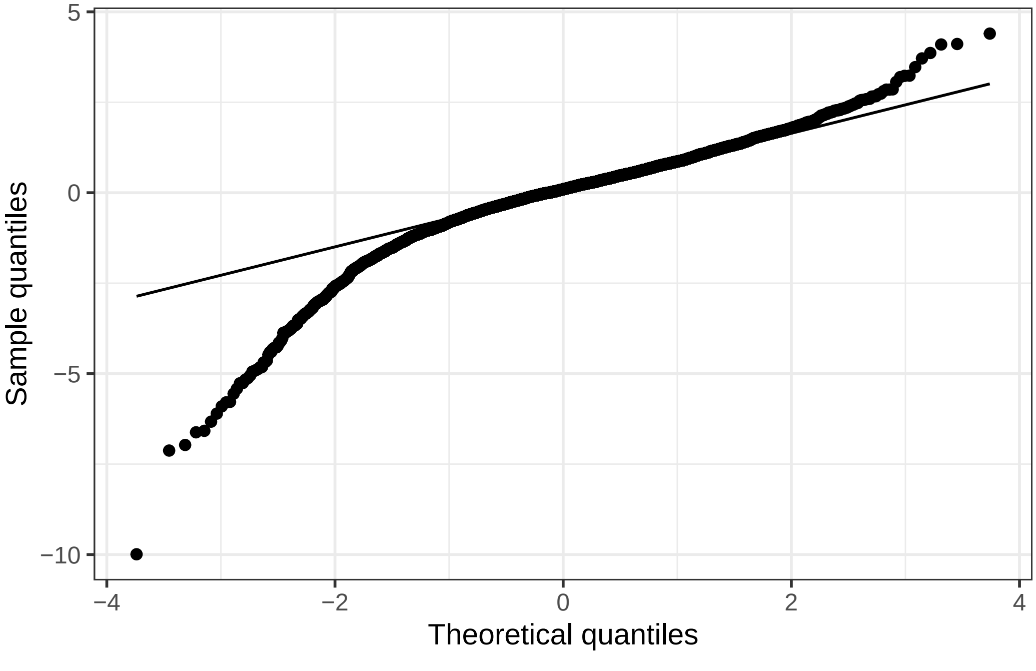
```
countryIreland                        -0.82094    0.65601  -1.251  0.21083  
countryIsrael                         -0.28140    1.22716  -0.229  0.81864  
countryItaly                          -1.64352    0.65951  -2.492  0.01273 *
countryJamaica                         0.77865    1.22534   0.635  0.52516  
countryJapan                          -0.70437    0.64585  -1.091  0.27549  
countryKenya                          -2.25745    1.22709  -1.840  0.06587 .
countryLebanon                         1.16873    1.22457   0.954  0.33992  
countryMalta                           0.24535    1.22609   0.200  0.84141  
countryMexico                          0.02592    0.69164   0.037  0.97011  
countryNetherlands                    -0.83310    0.86845  -0.959  0.33745  
countryNew Zealand                    -0.86672    0.66223  -1.309  0.19066  
countryNorway                         -1.04710    0.86832  -1.206  0.22792  
countryPanama                         -1.21186    1.22708  -0.988  0.32339  
countryPortugal                       -1.68623    1.22567  -1.376  0.16895  
countryRepublic of Macedonia          -0.61805    1.23786  -0.499  0.61760  
countryRussia                         -0.20727    0.81040  -0.256  0.79815  
countrySouth Africa                   -0.43243    0.81099  -0.533  0.59391  
countrySouth Korea                    -0.92666    0.68005  -1.363  0.17306  
countrySpain                          -0.64022    0.66271  -0.966  0.33406  
countrySweden                         -0.92469    0.77510  -1.193  0.23292  
countrySwitzerland                    -1.19138    0.81095  -1.469  0.14186  
countryTaiwan                          0.69653    0.87250   0.798  0.42472  
countryThailand                       -3.41402    1.22560  -2.786  0.00536 **
countryUnited Arab Emirates           -0.91297    0.96994  -0.941  0.34661  
countryUnited Kingdom                 -0.84947    0.61543  -1.380  0.16756  
countryUnited States                  -0.55506    0.61390  -0.904  0.36595  
countryWest Germany                   -0.68831    0.86694  -0.794  0.42726  
countryYugoslavia                     -0.29866    1.22571  -0.244  0.80750  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.06 on 5346 degrees of freedom
Multiple R-squared:  0.689, Adjusted R-squared:  0.6846 
F-statistic: 155.9 on 76 and 5346 DF,  p-value: < 2.2e-16
```

```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 rmse    standard        1.05
```

$$\widehat{Gross} = \beta_0 + \beta_1 * log(budget) + \beta_2 * score + \beta_3 * log(votes) + \beta_4 * log(runtime)$$

**Residual plots/Assumptions**





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 heavily skewed numeric variables, 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 variables 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.

```
                                  Estimate   Std. Error     t value
(Intercept)                     1.657265e+01 1.585551e+00 10.45229584
budget                          1.643599e-08 6.202083e-10 26.50075839
genreAdventure                 -2.077640e-01 8.852866e-02 -2.34685554
genreAnimation                  3.260631e-01 1.153802e-01  2.82598901
genreBiography                 -3.047205e-01 9.203402e-02 -3.31095462
genreComedy                    -1.100033e-01 5.490986e-02 -2.00334314
genreCrime                     -4.261092e-01 8.193284e-02 -5.20071334
genreDrama                     -4.716083e-01 6.419365e-02 -7.34665121
genreFamily                    -8.219951e-02 6.848484e-01 -0.12002585
genreFantasy                   -1.328181e-01 2.176435e-01 -0.61025535
genreHorror                     3.895676e-01 9.769604e-02  3.98754722
genreMystery                   -4.239937e-01 3.360492e-01 -1.26170110
genreRomance                   -1.411868e+00 6.118635e-01 -2.30748814
genreSci-Fi                    -7.913829e-01 5.577663e-01 -1.41884329
genreThriller                   2.413363e-01 5.168052e-01  0.46697735
genreWestern                   -8.924251e-01 9.647968e-01 -0.92498755
ratingG                        -1.000661e+00 1.373972e+00 -0.72829819
ratingNC-17                    -2.135006e+00 1.424776e+00 -1.49848543
ratingNot Rated                -3.672034e+00 1.386202e+00 -2.64899005
ratingPG                       -1.006539e+00 1.366645e+00 -0.73650337
ratingPG-13                    -1.060316e+00 1.367084e+00 -0.77560389
ratingR                        -1.647935e+00 1.367084e+00 -1.20543819
ratingTV-MA                    -2.821609e+00 1.677754e+00 -1.68177792
ratingUnrated                  -3.588825e+00 1.407751e+00 -2.54933303
```

```
ratingX                               -1.923063e+00 1.934240e+00 -0.99422159
score                                  1.378338e-01 2.546114e-02  5.41349620
votes                                  2.678037e-06 1.321454e-07 20.26583479
runtime                                7.046613e-03 1.358193e-03  5.18822562
countryAruba                          -6.223261e-01 1.574842e+00 -0.39516722
countryAustralia                      -4.285721e-01 8.154400e-01 -0.52557158
countryAustria                        -1.188163e-01 1.244522e+00 -0.09547151
countryBelgium                        -1.603143e+00 1.044727e+00 -1.53450854
countryBrazil                         -5.173868e-01 1.112711e+00 -0.46497864
countryCanada                         -4.918818e-01 7.984541e-01 -0.61604269
countryChile                          -4.332922e-01 1.575841e+00 -0.27495934
countryChina                           2.019743e-01 8.332550e-01  0.24239196
countryColombia                       -2.186150e-02 1.573876e+00 -0.01389023
countryCzech Republic                 -7.649493e-01 9.420817e-01 -0.81197767
countryDenmark                        -6.615747e-01 8.982454e-01 -0.73651889
countryFederal Republic of Yugoslavia -3.414024e+00 1.591548e+00 -2.14509639
countryFinland                         5.160677e-01 1.247993e+00  0.41351800
countryFrance                         -6.814080e-01 7.985923e-01 -0.85326137
countryGermany                        -3.226312e-01 8.021260e-01 -0.40222004
countryHong Kong                      -1.415802e-01 8.358754e-01 -0.16937952
countryHungary                         1.318828e-01 1.575764e+00  0.08369453
countryIceland                        -2.876264e+00 1.245524e+00 -2.30928035
countryIndia                           9.031688e-01 9.398206e-01  0.96100131
countryIndonesia                      -1.484632e+00 1.244476e+00 -1.19297747
countryIran                           -1.493119e+00 1.119420e+00 -1.33383221
countryIreland                        -7.172999e-01 8.420895e-01 -0.85180951
countryIsrael                         -9.326844e-01 1.576771e+00 -0.59151548
countryItaly                          -2.032464e+00 8.470625e-01 -2.39942681
countryJamaica                        -2.436770e-02 1.574285e+00 -0.01547858
countryJapan                          -5.948741e-01 8.295394e-01 -0.71711376
countryKenya                          -2.368203e+00 1.575844e+00 -1.50281565
countryLebanon                         1.226222e+00 1.573359e+00  0.77936595
countryMalta                           1.400113e+00 1.574963e+00  0.88898178
countryMexico                         -5.387707e-01 8.885889e-01 -0.60632164
countryNetherlands                     1.139056e-01 1.115184e+00  0.10214067
countryNew Zealand                    -1.531017e+00 8.509977e-01 -1.79908449
countryNorway                         -1.302300e+00 1.115601e+00 -1.16735328
countryPanama                         -1.262828e+00 1.575772e+00 -0.80140279
countryPortugal                       -1.456501e+00 1.573908e+00 -0.92540437
countryRepublic of Macedonia          -8.505413e-01 1.589854e+00 -0.53498076
countryRussia                         -3.252051e-01 1.041200e+00 -0.31233690
countrySouth Africa                    2.081995e-01 1.041582e+00  0.19988786
countrySouth Korea                    -1.015206e+00 8.735893e-01 -1.16210906
```

```
countrySpain                          -3.448539e-01 8.508364e-01 -0.40531160
countrySweden                         -7.928126e-01 9.960921e-01 -0.79592300
countrySwitzerland                    -2.103968e+00 1.041630e+00 -2.01988107
countryTaiwan                          4.772702e-01 1.120761e+00  0.42584494
countryThailand                       -3.810292e+00 1.574262e+00 -2.42036755
countryUnited Arab Emirates            3.517349e-01 1.245804e+00  0.28233576
countryUnited Kingdom                 -6.836693e-01 7.899414e-01 -0.86546836
countryUnited States                  -3.502972e-01 7.879711e-01 -0.44455593
countryWest Germany                   -2.723906e-01 1.113253e+00 -0.24467982
countryYugoslavia                     -1.079239e+00 1.575068e+00 -0.68520190
                                           Pr(>|t|)
(Intercept)                            2.505675e-25
budget                                1.668201e-145
genreAdventure                         1.896889e-02
genreAnimation                         4.730979e-03
genreBiography                         9.359492e-04
genreComedy                            4.519085e-02
genreCrime                             2.059222e-07
genreDrama                             2.338454e-13
genreFamily                            9.044672e-01
genreFantasy                           5.417186e-01
genreHorror                            6.765064e-05
genreMystery                           2.071114e-01
genreRomance                           2.106564e-02
genreSci-Fi                            1.560030e-01
genreThriller                          6.405351e-01
genreWestern                           3.550141e-01
ratingG                                4.664630e-01
ratingNC-17                            1.340662e-01
ratingNot Rated                        8.097011e-03
ratingPG                               4.614567e-01
ratingPG-13                            4.380172e-01
ratingR                                2.280875e-01
ratingTV-MA                            9.267032e-02
ratingUnrated                          1.082060e-02
ratingX                                3.201600e-01
score                                  6.450303e-08
votes                                  4.861018e-88
runtime                                2.201436e-07
countryAruba                           6.927352e-01
countryAustralia                       5.992077e-01
countryAustria                         9.239439e-01
countryBelgium                         1.249638e-01
```

```
countryBrazil                         6.419657e-01
countryCanada                         5.378926e-01
countryChile                          7.833581e-01
countryChina                          8.084858e-01
countryColombia                       9.889181e-01
countryCzech Republic                 4.168406e-01
countryDenmark                        4.614473e-01
countryFederal Republic of Yugoslavia 3.199007e-02
countryFinland                        6.792437e-01
countryFrance                         3.935526e-01
countryGermany                        6.875382e-01
countryHong Kong                      8.655045e-01
countryHungary                        9.333024e-01
countryIceland                        2.096596e-02
countryIndia                          3.365950e-01
countryIndonesia                      2.329311e-01
countryIran                           1.823157e-01
countryIreland                        3.943580e-01
countryIsrael                         5.542001e-01
countryItaly                          1.645478e-02
countryJamaica                        9.876509e-01
countryJapan                          4.733352e-01
countryKenya                          1.329457e-01
countryLebanon                        4.357987e-01
countryMalta                          3.740529e-01
countryMexico                         5.443269e-01
countryNetherlands                    9.186489e-01
countryNew Zealand                    7.206169e-02
countryNorway                         2.431198e-01
countryPanama                         4.229341e-01
countryPortugal                       3.547974e-01
countryRepublic of Macedonia          5.926854e-01
countryRussia                         7.547967e-01
countrySouth Africa                   8.415759e-01
countrySouth Korea                    2.452430e-01
countrySpain                          6.852647e-01
countrySweden                         4.261121e-01
countrySwitzerland                    4.344551e-02
countryTaiwan                         6.702380e-01
countryThailand                       1.553794e-02
countryUnited Arab Emirates           7.776970e-01
countryUnited Kingdom                 3.868206e-01
countryUnited States                  6.566587e-01
```

```
countryWest Germany                    8.067137e-01
countryYugoslavia                      4.932462e-01

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

REML criterion at convergence: 18920.6

Scaled residuals:
    Min      1Q  Median      3Q     Max
-8.0460 -0.4642  0.1673  0.6605  3.0305

Random effects:
 Groups   Name        Variance Std.Dev.
 country  (Intercept) 0.20145  0.4488
 genre    (Intercept) 0.08648  0.2941
 rating   (Intercept) 1.16189  1.0779
 Residual             1.85616  1.3624
Number of obs: 5423, groups:  country, 50; genre, 15; rating, 10

Fixed effects:
             Estimate Std. Error t value
(Intercept) 1.390e+01  4.425e-01  31.413
budget      1.663e-08  6.132e-10  27.114
score       1.365e-01  2.523e-02   5.411
votes       2.674e-06  1.316e-07  20.319
runtime     6.732e-03  1.335e-03   5.041

Correlation of Fixed Effects:
        (Intr) budget score  votes
budget   0.031
score   -0.275  0.184
votes    0.178 -0.357 -0.462
runtime -0.210 -0.347 -0.299 -0.053
fit warnings:
Some predictor variables are on very different scales: consider rescaling

# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
```

```
1 rmse    standard        1.35
```

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.

```
[1] 1.352356
```

```
[1] 1.357526
```

Compare model: since we can see that for linear regression, it has smaller RMSE as 1.3523 < 1.3575. We would conclude that model 1 is better than better 2 (linear regression better than the linear mixed model). However, the difference between RMSE is subtle.

## interaction terms

Since we conclude that the linear regression is better, we continue to apply linear regression and we're curious about whether the relationship between gross and budget depends on score/votes, while adjusting for all other variables.

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

Coefficients:
                    (Intercept)                           budget  
                      1.533e+01                        6.233e-08  
                 genreAdventure                   genreAnimation  
                     -1.785e-01                        4.381e-01  
                 genreBiography                      genreComedy  
```

```
     -2.953e-01                            -7.781e-02
     genreCrime                            genreDrama
     -4.155e-01                            -4.619e-01
    genreFamily                          genreFantasy
      3.958e-03                            -4.005e-02
    genreHorror                          genreMystery
      4.958e-01                            -4.617e-01
   genreRomance                           genreSci-Fi
     -1.361e+00                            -7.746e-01
  genreThriller                          genreWestern
      3.898e-01                            -7.493e-01
        ratingG                           ratingNC-17
     -1.472e+00                            -2.663e+00
ratingNot Rated                              ratingPG
     -4.141e+00                            -1.508e+00
    ratingPG-13                               ratingR
     -1.575e+00                            -2.143e+00
    ratingTV-MA                         ratingUnrated
     -3.386e+00                            -4.099e+00
        ratingX                                 score
     -2.569e+00                             3.320e-01
          votes                               runtime
      3.184e-06                             7.897e-03
   countryAruba                      countryAustralia
     -2.698e-01                            -1.177e-01
 countryAustria                        countryBelgium
      1.264e-01                            -1.249e+00
  countryBrazil                         countryCanada
     -3.579e-01                            -1.499e-01
   countryChile                          countryChina
     -9.122e-02                             4.203e-01
countryColombia                 countryCzech Republic
      1.814e-01                            -5.651e-01
 countryDenmark countryFederal Republic of Yugoslavia
     -4.766e-01                            -3.328e+00
 countryFinland                         countryFrance
      7.026e-01                            -3.935e-01
 countryGermany                      countryHong Kong
     -4.767e-02                             1.829e-01
 countryHungary                        countryIceland
      4.989e-01                            -2.498e+00
   countryIndia                      countryIndonesia
      1.013e+00                            -1.422e+00
```

```
                          countryIran                         countryIreland
                           -1.458e+00                             -4.456e-01
                        countryIsrael                           countryItaly
                           -9.517e-01                             -1.732e+00
                       countryJamaica                           countryJapan
                            4.031e-01                             -4.099e-01
                         countryKenya                         countryLebanon
                           -2.049e+00                              1.230e+00
                         countryMalta                          countryMexico
                            1.653e+00                             -2.062e-01
                   countryNetherlands                     countryNew Zealand
                            3.591e-01                             -1.041e+00
                        countryNorway                          countryPanama
                           -9.597e-01                             -9.140e-01
                      countryPortugal           countryRepublic of Macedonia
                           -1.104e+00                             -7.877e-01
                        countryRussia                    countrySouth Africa
                           -6.005e-02                              4.226e-01
                   countrySouth Korea                           countrySpain
                           -9.305e-01                             -1.573e-01
                        countrySweden                     countrySwitzerland
                           -6.318e-01                             -1.691e+00
                        countryTaiwan                        countryThailand
                            6.586e-01                             -3.315e+00
          countryUnited Arab Emirates                  countryUnited Kingdom
                            6.963e-01                             -3.695e-01
                 countryUnited States                    countryWest Germany
                           -4.272e-02                              5.280e-02
                    countryYugoslavia                           budget:score
                           -4.554e-01                             -7.001e-09


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

Coefficients:
                          (Intercept)                                 budget
                            1.671e+01                              2.628e-08
                       genreAdventure                         genreAnimation
                           -2.225e-01                              1.229e-01
                       genreBiography                            genreComedy
```

```
     -2.780e-01                             -5.439e-02
     genreCrime                             genreDrama
     -3.971e-01                             -4.340e-01
    genreFamily                           genreFantasy
     -1.750e-01                             -1.892e-02
    genreHorror                           genreMystery
      4.662e-01                             -5.331e-01
   genreRomance                            genreSci-Fi
     -1.319e+00                             -7.855e-01
  genreThriller                           genreWestern
      4.279e-01                             -6.417e-01
        ratingG                            ratingNC-17
     -1.170e+00                             -2.309e+00
ratingNot Rated                               ratingPG
     -3.765e+00                             -1.289e+00
    ratingPG-13                                ratingR
     -1.333e+00                             -1.911e+00
    ratingTV-MA                          ratingUnrated
     -2.954e+00                             -3.714e+00
        ratingX                                  score
     -2.208e+00                              1.162e-01
          votes                                runtime
      5.073e-06                              6.292e-03
   countryAruba                       countryAustralia
     -6.319e-01                             -3.861e-01
 countryAustria                         countryBelgium
     -7.550e-02                             -1.596e+00
  countryBrazil                          countryCanada
     -4.065e-01                             -4.953e-01
   countryChile                           countryChina
     -3.856e-01                             -2.665e-02
countryColombia                  countryCzech Republic
      2.077e-01                             -8.568e-01
 countryDenmark  countryFederal Republic of Yugoslavia
     -7.371e-01                             -3.453e+00
 countryFinland                          countryFrance
      4.079e-01                             -7.508e-01
 countryGermany                       countryHong Kong
     -4.645e-01                             -1.029e-01
 countryHungary                         countryIceland
      2.363e-01                             -2.667e+00
   countryIndia                       countryIndonesia
      9.495e-01                             -1.480e+00
```

```
                countryIran                         countryIreland
                 -1.429e+00                             -6.191e-01
              countryIsrael                           countryItaly
                 -5.049e-01                             -1.955e+00
             countryJamaica                           countryJapan
                  2.304e-01                             -5.158e-01
               countryKenya                         countryLebanon
                 -2.256e+00                              1.396e+00
               countryMalta                          countryMexico
                  1.100e+00                             -3.951e-01
         countryNetherlands                     countryNew Zealand
                  1.186e-01                             -1.075e+00
              countryNorway                          countryPanama
                 -1.177e+00                             -1.157e+00
            countryPortugal            countryRepublic of Macedonia
                 -1.525e+00                             -7.215e-01
              countryRussia                    countrySouth Africa
                 -1.405e-01                             -7.729e-02
         countrySouth Korea                           countrySpain
                 -1.078e+00                             -5.062e-01
              countrySweden                     countrySwitzerland
                 -7.400e-01                             -1.889e+00
              countryTaiwan                        countryThailand
                  5.142e-01                             -3.606e+00
countryUnited Arab Emirates                  countryUnited Kingdom
                  1.678e-01                             -6.815e-01
       countryUnited States                    countryWest Germany
                 -3.840e-01                             -1.770e-01
          countryYugoslavia                           budget:votes
                 -8.524e-01                             -3.566e-14
```

For interaction terms: The coefficient for score is 3.320e-01 and the coefficient for votes is 5.073e-06. Since the coefficient for votes is too low, we would only consider score here.

```
[1] 1.352356
```

```
[1] 1.357526
```

## 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.

```
[1] 5423   17
```

```
[1] 4338   17
```

```
Linear Regression

5423 samples
   7 predictor

No pre-processing
```

```
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 4883, 4879, 4882, 4881, 4881, 4880, ...
Resampling results:

  RMSE    Rsquared   MAE
  1.3736  0.4729455  1.0154

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

Linear Regression

5423 samples
   7 predictor

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

  RMSE      Rsquared   MAE
  1.083085  0.6717084  0.7602062

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.

## 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

Variable selection, specifically the lasso for categorical variables with 2+ levels, Rescaling predictor variables, linearity assumptions,

## Sources

https://towardsdatascience.com/feature-selection-in-machine-learning-using-lasso-regression-7809c7c2771a https://stackoverflow.com/questions/13646654/root-mean-square-error-in-r-mixed-effect-model https://psyarxiv.com/wc45u/