# MAST 6251: HW2 - Movies

Chris Graham(31540773), Yan Li(48081999),Pinhsuan Wu(48079490),Lifen Zeng(48082142), Wenjie Ni(48072067)

11/20/2019

This assignment is split into three parts:

- **Part I:** Upload the datasets and merge.
- **Part II:** Perform EDA analysis.
- **Part III:** See if we can find any correlations and observe our findings!
- **Part IV:** Summary and conclusion
- **Part V:** Appendix

## Part I: Merging Datasets

Our dataset consisted of four files. The files of interest we wish to use for this excercise are `ratings.csv`, `movies_metadata.csv` and `credits.csv`. All three of these files are linked via movie_id. We will be pulling ratings information from `ratings.csv`, cast information from `credits.csv` and movie information from `movies_metadata.csv`. Before merging the datasets we will be adding two dummy variables to `ratings.csv`: * mean_rating - this will provide us with a single rating value per movie * label - this will provide us with a binary indicator whether a movie is “good” or “bad” We will also be parsing the `credits.csv` as the cast information is currently unreadable in its original format. Let’s go!

Let’s take a peek on how everything turned out:

`rating.csv`

```
##   movieId mean_rating label
## 1       1        3.89     1
## 2       2        3.24     0
## 3       3        3.18     0
```

`credits.csv`

```
##      id        name movie_id
## 1    31   Tom Hanks      862
## 2 12898   Tim Allen      862
## 3  7167 Don Rickles      862
```

Looking good! Now all that is left is to join the data and choose our variables of interest.

Now let's look at our finalized dataset after our join:

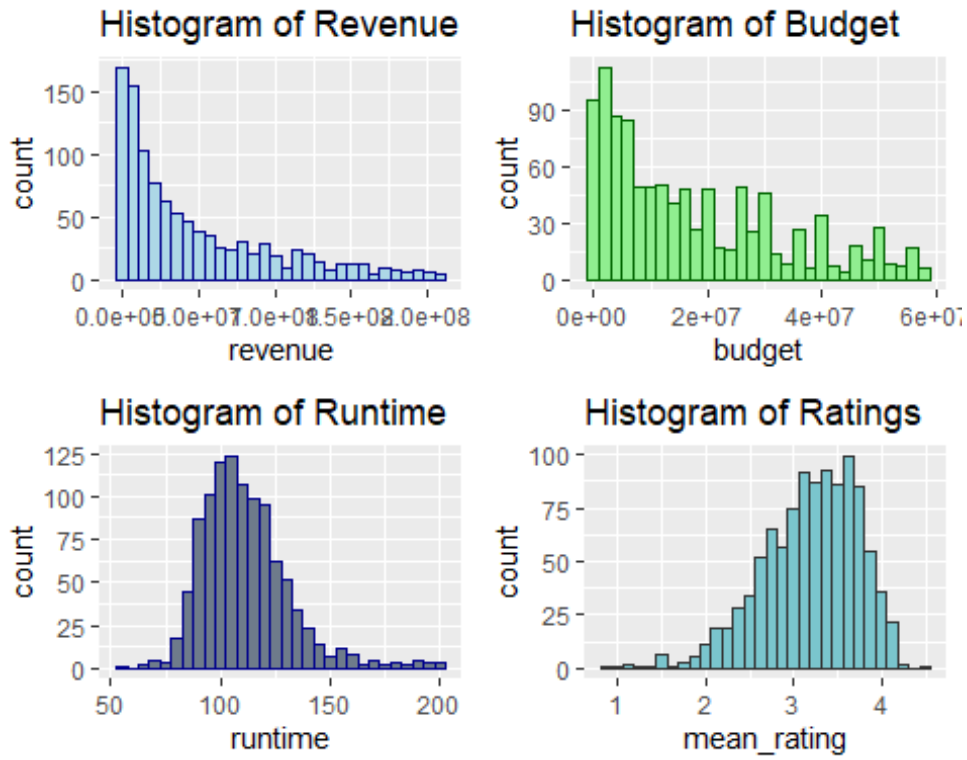
```
Com_Data
```

```
##    movie_id            title           name  budget   revenue runtime
## 1       862        Toy Story      Tom Hanks 3.0e+07 373554033      81
## 14     8844          Jumanji Robin Williams 6.5e+07 262797249     104
## 40    15602 Grumpier Old Men Walter Matthau 0.0e+00         0     101
##    vote_count mean_rating label
## 1        5415        3.60     1
## 14       2413        3.76     1
## 40         92          NA    NA
```

## Part II: EDA

Our EDA will be fairly standard for this dataset. We will be detecting outliers in our numeric variables, converting those outlier values to missing values and then removing all rows with missing values. There are also a fair amount of 0 values in our dataset and since this will heavily skew our distributions for the numeric variables in question we will be removing those as well. Here we go!

How is our post - EDA dataset looking? Let's take a look at the distributions of our numeric variables:




A little skewed but definitely manageable. We are now ready for Part III!

## Part III: Observations

What we decided to observe was whether the `Budget`, `Revenue` or `Runtime` significantly influenced the eventual rating of movies in our dataset. Prior to running our analysis we believed that `Revenue` of a movie would be highly influential in its rating as successful movies are usually well liked by their movie-goers. We understood that some endogeniety could occur from this assumption so we had to be careful! Let's run our regression:

Let's observe our results and come to a conclusion about our relationships:

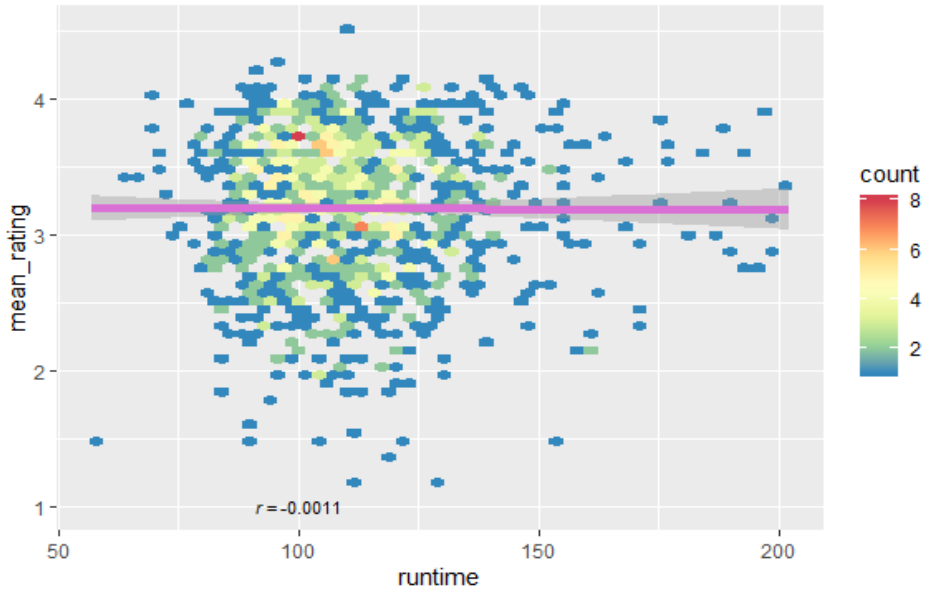
`Coefficients`

```
## (Intercept)      budget     revenue     runtime
##        1.15        0.00        0.00        0.00
```

`P-Values`

```
## (Intercept)      budget     revenue     runtime
##        0.00        0.14        0.84        0.84
```

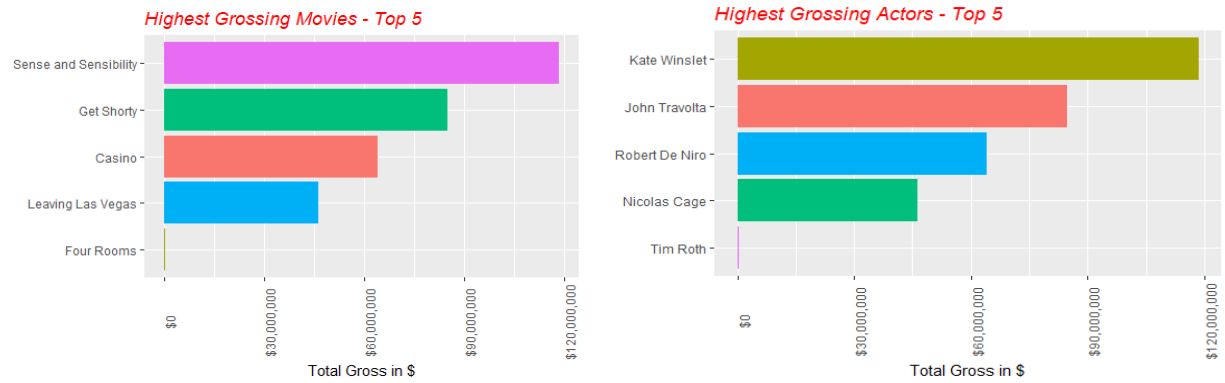
Overall our model is **not** statistically significant as we are getting an R-squared of only 0.2%. Therefore we cannot make any concrete determinations on whether `Budget`, `Revenue` or `Runtime` truly have any significant impact on a movie's future rating in our dataset. Let's confirm by looking at our most statistically significant variable of the three (`Budget`) in linear form:




Unfortunately we just do not have much there. So we can not make any accurate predictions moving forward. But let's look at some additional components of the dataset so we can have a better understanding of what it is telling us!

### What are our highest grossing movies and actors? Do they have high ratings? Let's see!

First we will need to add a dummy variable `Gross` to our dataset which is simple calculation of (Revenue - Budget). Let's look at our most profitable players:




Sense and Sensibility and Kate Winslet reign supreme. But do they have good ratings? Let's take a look.

`Movies`

```
##                   title         gross2 mean_rating
## 1 Sense and Sensibility 118,500,000.00        2.72
## 2            Get Shorty  84,851,622.00        3.84
## 3                Casino  64,112,375.00        3.52
## 4     Leaving Las Vegas  46,200,000.00        3.05
## 5            Four Rooms     300,000.00        3.08
```

`Actors`

```
##             name         gross2 mean_rating
## 1   Kate Winslet 118,500,000.00        2.72
## 2  John Travolta  84,851,622.00        3.84
## 3 Robert De Niro  64,112,375.00        3.52
## 4   Nicolas Cage  46,200,000.00        3.05
## 5       Tim Roth     300,000.00        3.08
```

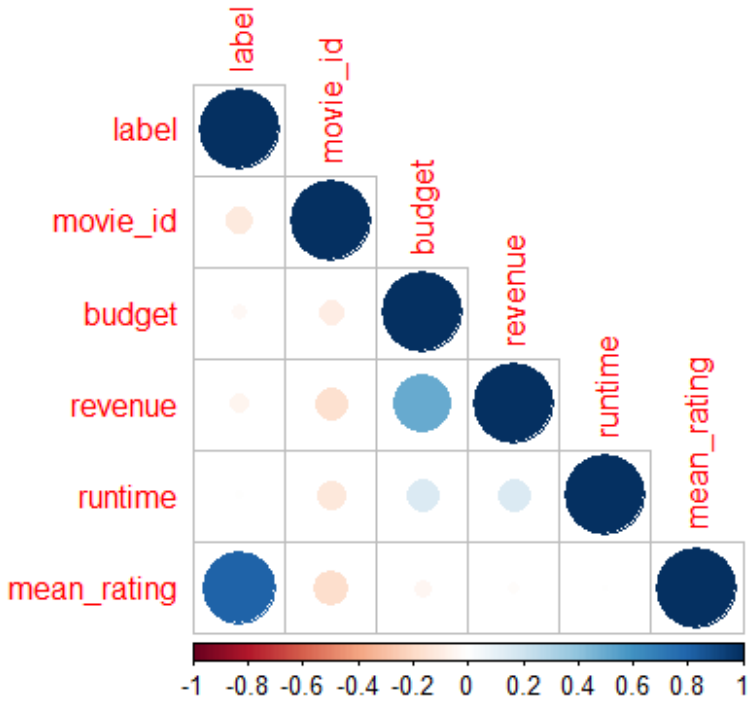
## Part IV: Conclusion

That's it! When we started the project we thought that we were going to uncover some statistically significant indicators of movie ratings but unfortunately our analysis told us otherwise. However we did find some interesting facts! According to our dataset, a movie or actor does not need a particularly good rating in order to be commercially successful. We will leave you with some words to live by: "Not everyone likes me but not everyone matters".

## Part V: Appendix

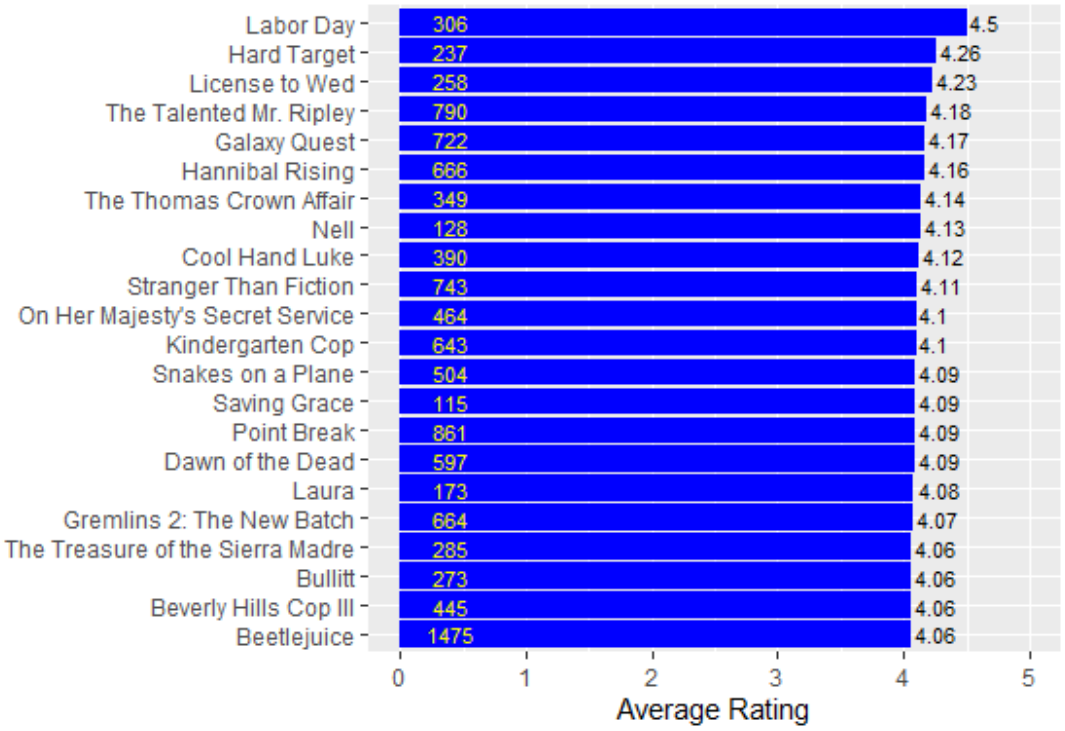
### How does our regression look in a corrplot format?

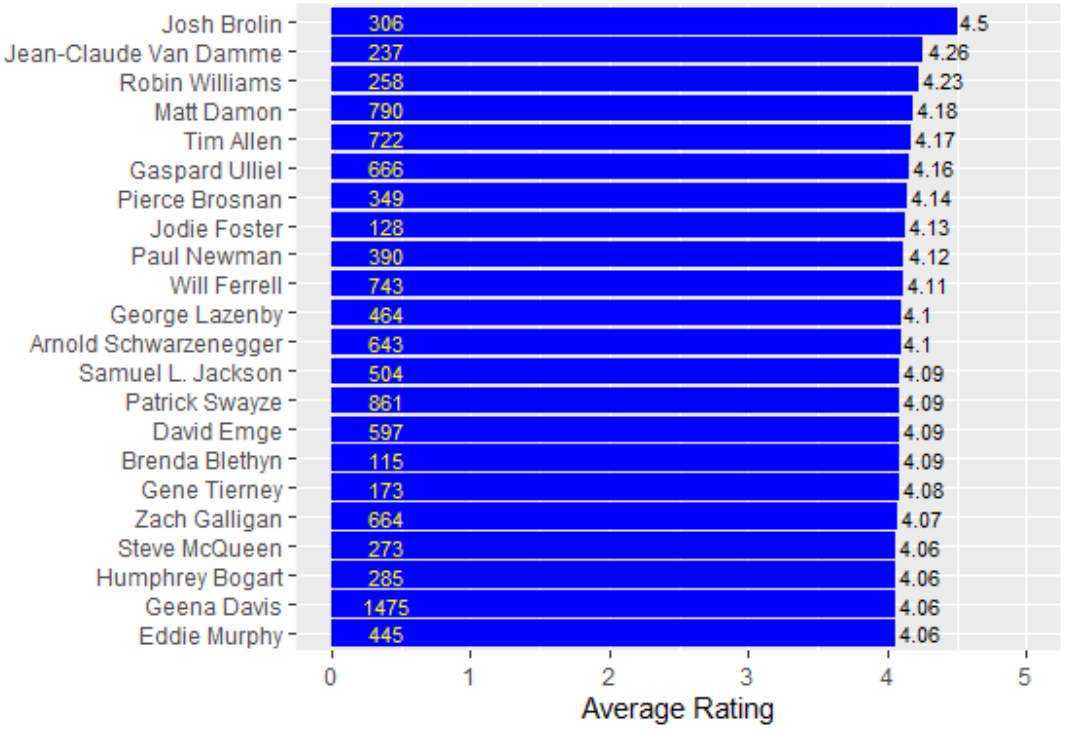
We wanted to include this into our original report but ran out of room. We decided that the linear graph of budget was more aesthetically pleasing.




### What are our highest ranking movies and actors?

We are going to only look at variables with vote counts over 100 so we can keep our analysis focused on statistically significant players.




Pretty cool! If you are going to the movies make sure it has `Josh Brolin` in it and if you are perusing Netflix try `Labor Day`!