MA678 Final Project

Zhiwei Liang

I. Overview

In this project, I will try to gain some insights into the movie industry.I am always a big fan of movies, so I choose it as the topic of my mid term project. I divide my work into three parts:

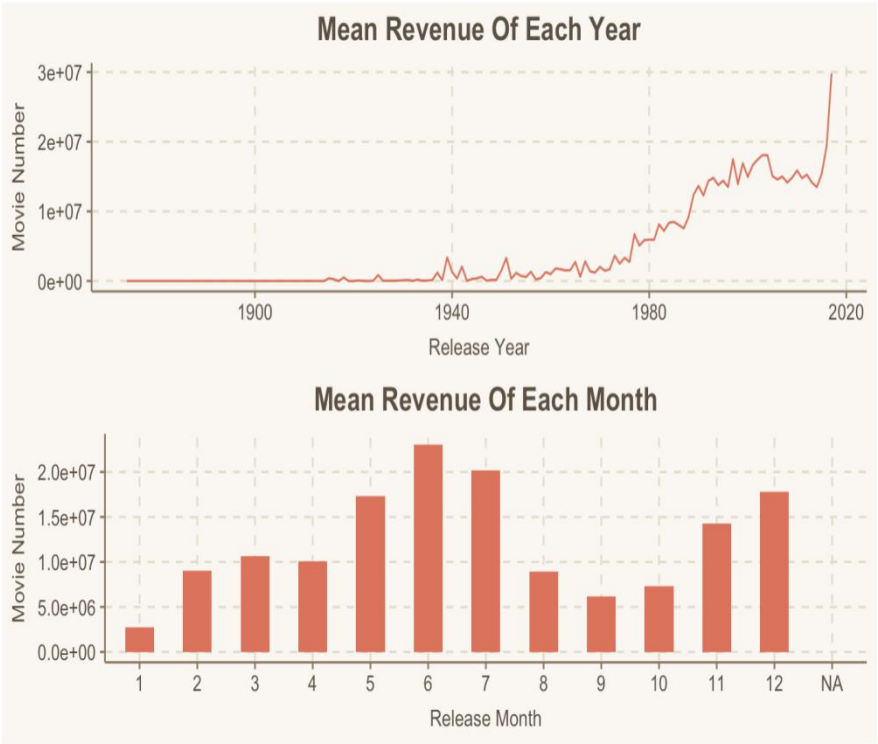
- Data Cleaning And Organization
- Exploratory Data Analysis
- Modeling And Validation

II. Data Description

My dataset contain metadata for all 45,000 movies listed in the Full MovieLens Dataset. The dataset consists of movies released on or before July 2017. Data points include cast, crew, plot keywords, budget, revenue, posters, release dates, languages, production companies, countries, TMDB vote counts and vote averages.I gain it from Kaggle.

III. Exploratory Data Analysis

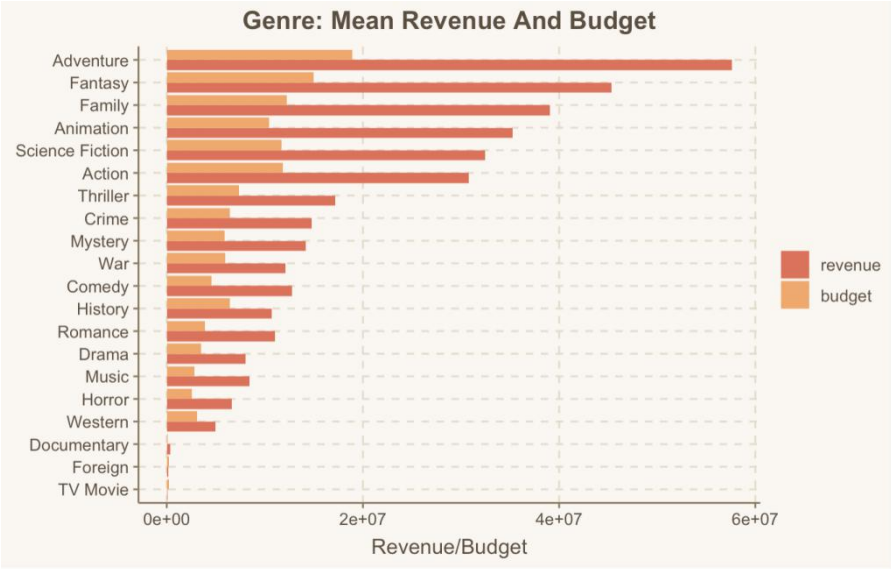
1. Which years and months have the highest revenue?



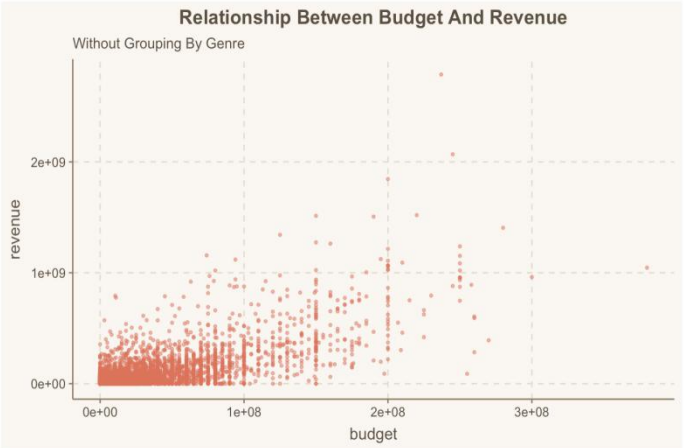
- The average revenue has an increasing overall trend among years. Especially in 2016, the yearly average revenue increased sharply.
- **May** to **July** have the highest average revenue. This can be attributed to the fact that blockbuster movies are usually released in the summer when the kids are out of school and the parents are on vacation and therefore, the audience is more likely to spend their disposable income on entertainment.

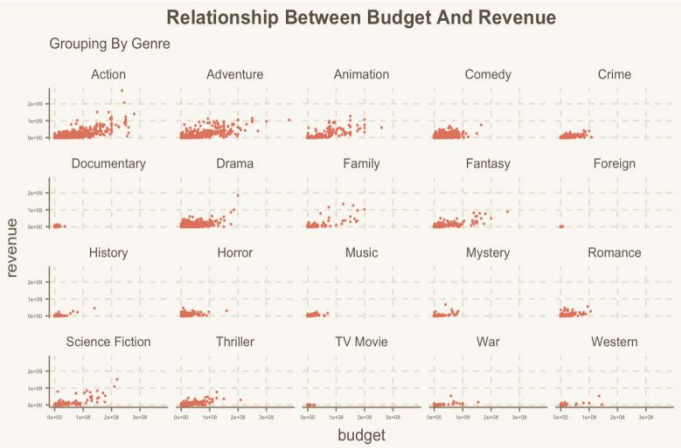
2. Which genres have the highest revenue?

- **Adventure** and **Fantasy** movies have the highest revenue and budget.
- **Documentary** and **Foreign** movies have the lowest revenue and budget



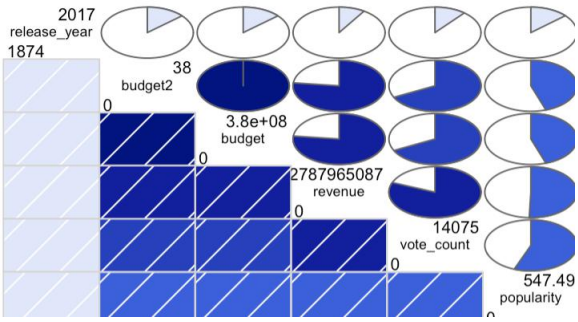
3. What is the relationship between revenue and other factors?





According to the correlation plot, we can find that **revenue** is strongly related with **budget** and **vote_count**.

Obviously, there are difference of revenue between different **genre** groups through the relationship plots.



IV. Modeling And Validation

Overview

The question I'd like to explore is **the relationship between movie revenue and budget**. If possible, I also want to gain a significant model to estimate movie revenue.

As the *Genres and Revenue plot* in EDA part, the group (genre) that a movie belongs to could determine their revenue in different ways. As the *Months and Revenue plot* in EDA part, it seems like group(month) could also determine their revenue in different ways.

In this part, I use the Multilevel Model (which is also called Hierarchical Model). The two grouping factors are genre and months, and covariates offset as budget and vote_count. (According to the correlation plot, I find vote_count is strongly related to revenue.The reason why I choose vote_count as a covariate).

Anova

Firstly, I calculate the total standard deviation of the revenue, which equals to 66378054.37. It seems like revenue varies widely.

Next, I want to know the group-class variation among genres and months. I use aov() to calculate the p value of the two-side variance test, and use TukeyHSD() to compute the exact differences between levels. According to the Anova output, there are significant inter-class differences of revenue in the two groups (genre and month). In this case, multilevel model works! In details, there are many genres have significant difference, many have not. For examples, group **Drama** and **Animation** have significant difference while **Family** and **Action** have not. group **Jan** and **Jul** have significant difference while **March** to **May** have not.

Modeling

- **Complete-pooling model**

One estimate would be the average that completely pools data across all genres. This ignores variation among genres.

According to the output, budget2,vote_count, Adventure, Animation, Comedy, Documentary, Drama, Family, Romance are all very significant. The estimate of coefficients represent the contribution of certain genre to 18the revenue. The model fits not bad since the R square equals to 0.75

- **Model with varying intercept**

Considering the case that each genre has a different revenue baseline, while the revenue increase rate at which budget increase is consistent across the genre groups. So, I would allow the intercept to vary by group(genre).
The formula is:

$$\text{revenue} - \text{budget2} + \text{vote\_count} + (1|\text{genre1})$$

The budget2 and vote are fixed effects (constant slopes), while the set of random intercepts is captured by genre1. This strategy allows us to capture variation in the revenue baseline of each genre. However, there may be additional information we want to incorporate into our model.

- **Model with varying intercept and slope**

Revenues of each genre have different baseline and change at different rates with budget depending on their genres. To incorporate both of these realities into our model, I allow both the budget slope and the intercept to vary depending on the movie genre.

The formula is:

$$revenue - budget2 + vote\_count + (1 + budget2|genre1)$$

- **Model with varying intercept among month and genre**

Here comes the more complex model, involving the grouping factor month. So there are two grouping factors in our formula. group1 genre has 20 levels and group2 month has 12 levels.
The formula is:

$$revenue - budget2 + vote\_count + (1|genre1) + (1|release\_month)$$

- **Model with varying intercept and slope among month and genre**

In my last model, I allow both the budget slope and the intercept to vary depending on the movie genre and month.
The formula is:

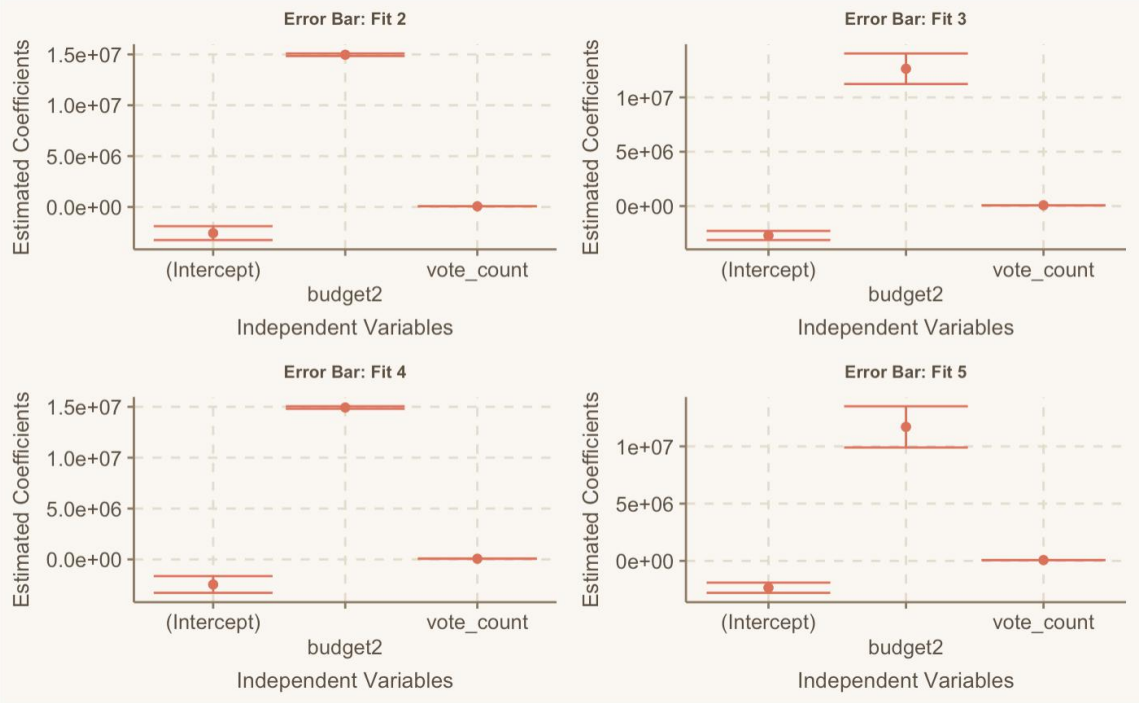
$$revenue - budget2 + vote\_count + (1 + budget2|genre1) + (1 + budget2|release\_month)$$

comparison

I plot error bar of all the model coefficients. It seems like Fit3 and Fit5 are better than other two through their estimate coefficient of budget2.

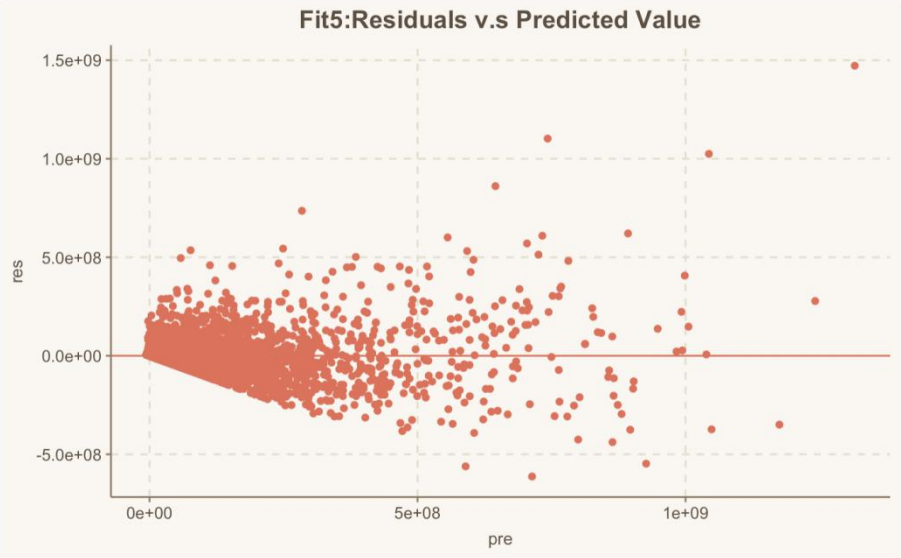
Then, I use anova() to compare the four models through the p value and AIC. According to the output, it seems like models with varying intercept and slope(Fit3 and Fit5) are better than other two(Fit2 andFit4). Compared the AIC of Fit3 and Fit5 again, AIC's Fit5 is the minimum.

So it's reasonable to regard **Fit5** as the goal model.



Validation

Finally, I plot the residuals and fitted value plot. The residuals are nearby zero when the fitted values stay in comparatively small range. *This is the coefficients table for each genre*



	(Intercept)	budget2	vote_count
Action	-4626109.4	13683051	69917.71
Adventur	-2461605.4	14560561	69917.71
Animatio	-2841508.7	20047983	69917.71
Comedy	-1358098.7	11926578	69917.71
Crime	-2360660.5	6427581	69917.71
Documer	-965274.8	11147132	69917.71
Drama	-1787922.1	8333675	69917.71
Family	-3143929.1	25965208	69917.71
Fantasy	-3456880.6	14535699	69917.71
Foreign	-2058059.4	11446505	69917.71
History	-2721402.5	14082258	69917.71
Horror	-2181515.4	8089920	69917.71
Music	-1368186.2	8872084	69917.71
Mystery	-2654600.7	13119453	69917.71
Romance	-1427625.3	13569307	69917.71
Science F	-3933762.1	13605286	69917.71
Thriller	-2944304.3	8535821	69917.71
TV Movie	-1795401	8672671	69917.71
War	-1530909.7	5728175	69917.71
Western	-1264434	1634828	69917.71

VI. Conclusion

fit5=lmer(revenue~budget2+vote_count+(1+budget2|release_month)+(1+budget2|genre1),data=main) is our final model.

In reality, movie revenue is not a simple thing to predict, instead, it relates to many complex factors. In fact, the AIC of Fit5 is also large. I cannot say this is a good model to estimate the movie revenue. The reasons I guess are the dependency of the covariates and the lack of predictors.

Next step, maybe I can involve more covariates (such as release_year, ratings) in the formula to improve the model.

VII. Reference

My dataset from Kaggle
Multilevel model example from Kaggle
Hierarchical model visualization
Multilevel introduction and RC method
ANOVA
The lme4 package (Bates, Maechler, Bolker, and Walker 2014a) for R (R Core Team 2015)
Multilevel parameter explanation