**Factors That Affect Movie Revenue**

Qiang Fang, Minji Lee, Katrina Truebenbach, Mingfei Zhou

**Summary**

           We plan to use the IMDB 5000 Movie Dataset from Kaggle to answer the question: what factors affect movie revenue the most? The dataset includes 5,000 movies with 28 variables including: directors’ and actors’ names, gross revenue, genres, and duration. We hypothesize that genre impacts movie revenue the most out of a host of other variables such as movie duration, budget, IMDB score, content rating, and country.

           We will also use two external datasets: ‘Inflation, GDP deflator’ from The World Bank to convert nominal dollar values to real so that we can compare revenues and budgets consistently over time and across countries, and ‘The Academy Awards, 1927-2015’ from Kaggle to test if having ‘better’ directors and actors/actresses, measured by number of Oscars, affects movie revenue.

We will examine the relationship between revenue and these factors using both visualizations to plot them against each other and a regression model to separate out the relative effects on revenue.

**Proposed plan of research**

First, we need to process the data using tidyr. We need to separate the string variable genre that is currently multiple values separated by a delimiter into separate variables. Then we need to merge the dataset of number of Oscars held by actors and directors using a left join so we only include actors and directors that are in the IMDB dataset. We also want to merge in price deflators by country and year from the World Bank to make the revenue and budget numbers real so that they are comparable over time. We may need to merge in exchange rates as well if any values are not in US dollars. The last step is to separate the dataset into training, validation, and test datasets.

For the visualization of the numerical variables, such as budget and IMDB score, we will use scatter plots to visualize the variables versus real revenue. We will visualize the relationship between revenue and the categorical variables, such as genre and content rating, with boxplots or barplots. We also need a line graph of year vs revenue.

We also plan to run a regression model to parse out the relative effects. First, we will examine if the relationships are linear using the visualizations, and if they are not, consider log transforms and non-linear models. We may also consider interactions between variables: are comedy-dramas more profitable than straight comedies? Assuming linear, one possible model is as follows. We will likely fit many models on the training set and compare on the validation:

Real gross revenue = ꞵ0 + ꞵ1\*duration +ꞵ2\*real budget + ꞵ3\*IMDB score + ꞵ4\*genre + ꞵ5\*content rating + ꞵ6\*country + ꞵ7\*number of actor Oscars + ꞵ8\*number of director Oscars + ꞵ9\*year

The genre, content rating, and country variables will actually be a series of dummy variables (excluding one category per variable to prevent collinearity). It is important to separate out the effect of year because events such as the 2008 recession will have an effect on revenue.

**Preliminary results**

We have successfully loaded the data into R. Thus far we have removed duplicated rows and analyzed and removed missing values as seen in Figure 1. We also processed "genre" such that there is one variable per genre to create tidy data. We also created a couple of preliminary graphics, seen below. Figure 2 shows the average gross revenue by genre. Figure 3 shows that gross nominal revenue is positively related to nominal budget, albeit with some outliers. Figure 4 shows that nominal gross revenue has a slightly positive relationship with IMDB score.

**References**

1. Yueming. (2017, December 16). IMDB 5000 Movie Dataset. Retrieved February 21, 2019, from <https://www.kaggle.com/carolzhangdc/imdb-5000-movie-dataset>
2. Academy of Motion Picture Arts and Sciences. (2017, February 13). The Academy Awards, 1927-2015. Retrieved February 21, 2019, from https://www.kaggle.com/theacademy/academy-awards
3. Inflation, GDP deflator (annual %). (n.d.). Retrieved February 25, 2019, from https://data.worldbank.org/indicator/NY.GDP.DEFL.KD.ZG
4. Wang, Y., & Blei, D. M. (2018, June 19). The Blessings of Multiple Causes. Retrieved February 26, 2019, from <https://arxiv.org/abs/1805.06826>

**Figures**

Figure 1

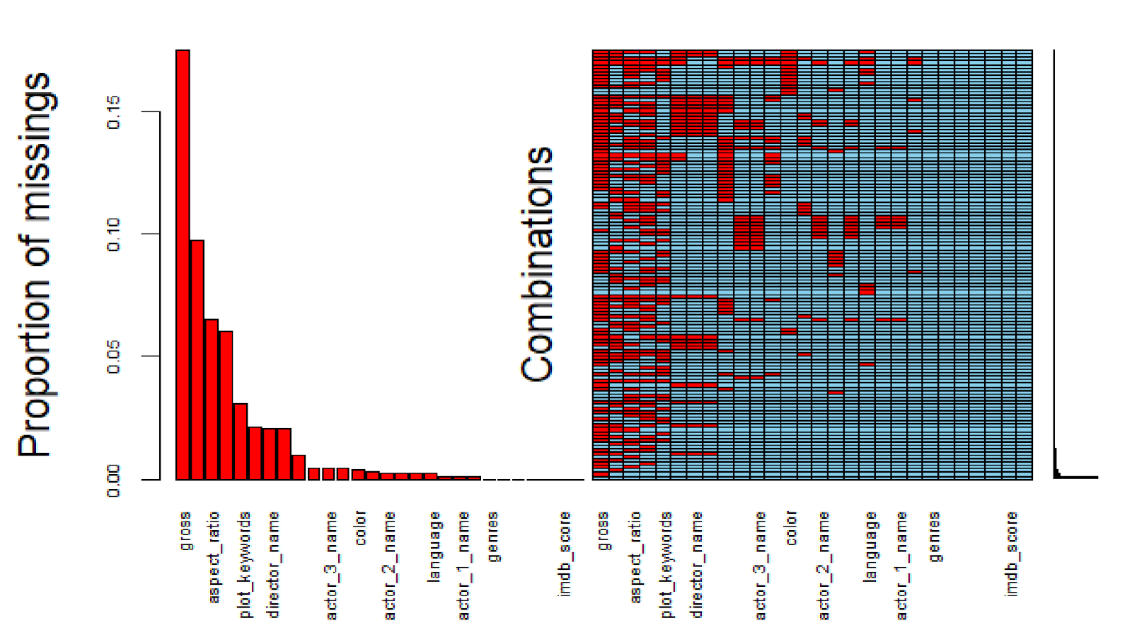


Figure 2

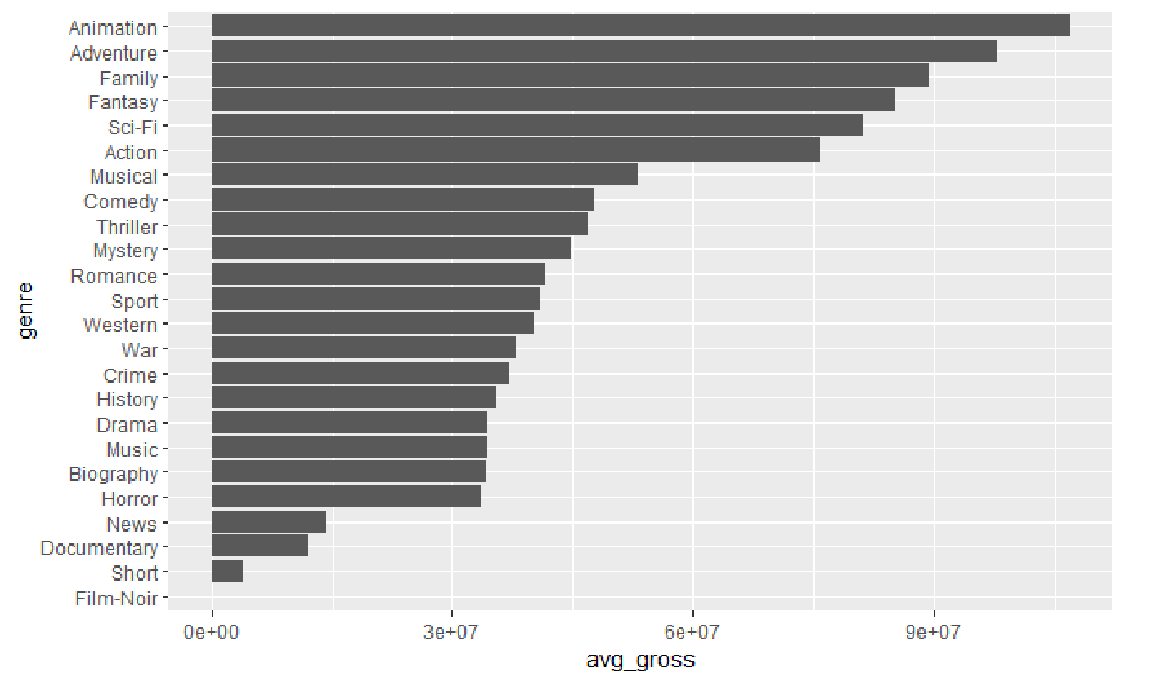


Figure 3

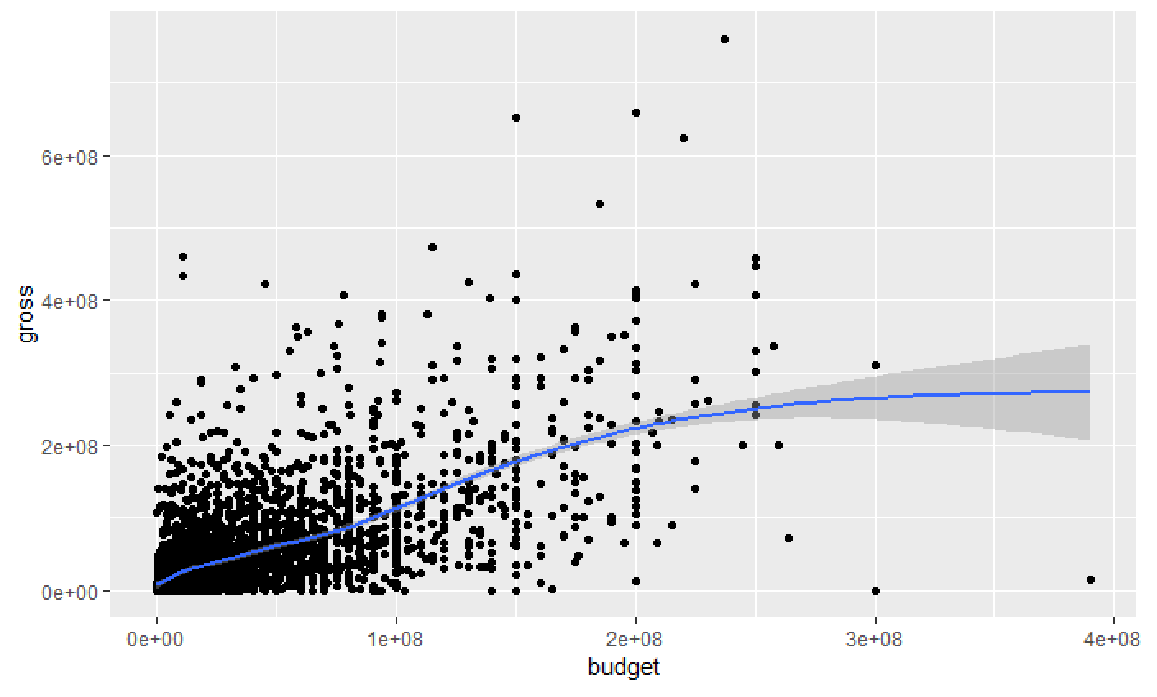


Figure 4

