Data Collection.

Scraping the websites

We scraped our data set from IMDB. We divided our data scraping work into two steps. First we scrape through lists of movies and fetch all the links to the movies we want. IMDB keeps its own list of films as "Top 250 Rated Films" and "Top Films by Genre". But to get on the lists a movie must have at least 25000 votes. All these lists add up to roughly 10,000 movies links. However, we observed that most of movies have three or more genres, which means we can only get around 3000 movies after removing all the duplicates. So we looked for some other websites that may keep links of movies at IMDB. Douban Movie is the Chinese version of IMDB, which keeps data about all movies in Chinese. It has lists of movies by genre and year, and the lists are almost exhaustive. For each movie record, Douban not only keeps information about films, but also a link to the film's page on IMDB. We opened a page like the one below to catch all the movie links. We specified the year to be each year during 2000 to 2015, and we added a tag “USA”, so that we get mostly movies here. To go through all pages of the list, we changed the number after "start=" to be 20, 40, 60, etc. Then we run through all movie links on Douban, to locate and fetch the IMDB link of each movie, and we have all links to movies on IMDB. To filter out some really unimportant ones we only selected the top 1000 films each year. After we have all the links, we run through all the links to fetch the data from IMDB. In data cleaning, we did some format parsing, and currency exchange for budget data that are not in USD.

To start out with, we want to have a brief direction of where we should be heading and what’s popular in the field of movies. We can observe the popular keywords in the movies. From the most popular words of 2000 to 2015, we are seeing some really typical keywords of movies, such as love, murder, and friend. Not really surprising. But if we look at the keywords year by year, we start to see some interesting trends. In year 2000 to 2010, the most popular keywords are all quite classical keywords. In recent years, however, we are seeing some interesting keywords such as “based on true story” and “box office flop”. I think this indicates that with the advance of Internet and online streaming services, people are not really interested by traditional elements for a good film. Rather than well-plotted love and hates, they would like to see more surprises or more real life related plots.

The first thing of producing a movie is to find a director. But who should we choose in order to make a profitable movie with only limited amount of budget. To answer that question, we need to plot a graph of average budget that each director spent when making a movie versus the average gross of all the movies directed by each director. For the cleaning process, we first selected the budget, gross, and director columns from the "movies\_imdb.csv" data frame. Then we grouped by the name of the director and used summarize function to calculate the mean of the budget and the gross for each director. The following graph is what we got when we only choose the top200 directors according to the gross they make with their movies.

Let's take a step back. Intuitively, without looking at the graph, one would probably choose some big names to be the director of his movie, such as Christopher Nolan, Peter Jackson, George Lucas, and James Cameron. All of these directors have made incredible movies and create history in movie box history. But since we are aiming to produce a fairly profitable movie, we need to choose a director who uses least amount of budget with the guarantee of the gross of the movie, which in this case should be represented as points located at the upper left portion of the graph. From figure-1, some of our potential candidates could be Pierre Coffin who directed all three "Despicable Me", George Lucas who is famous for producing and directing the "Star Wars" series, and James Cameron who directed "Titanic", and "Avatar". Figure-2 shows that when we only have about 120 million budget, we should choose James Cameron over Christopher Nolan. Similarly in figure-3, if we want to make a movie that has gross around 350 million, we should choose Pierre Coffin over David Yates who directed "Toy Stories", "Finding Nemo", and "Monster, Inc."

To conclude, the factor of director does have an influence on the gross of the move. Although, in general, directors who spent more money tend to make high box movies, yet there are still many talented and famous directors who are able to use only limited amount of money to create box miracle.

The next variable we are going to explore is the factor of actor/actress. Similar to what we did to the director data, we are going to do almost the same thing to the actor/actress data. However, the actor/actress data we collected from imdb are composite, which means all the actors/actresses from a movie are combined to a string. Therefore, what we did is that we first use str\_split function to separate the each entry of actor/actress information then unique the vector we got in order to get all the actor/actresses that cast in the movies we collected from imdb. Then we did the same thing as we did to the director data to get the following mean budget versus mean gross graph, in which each point represents a movie star. The graph only shows the top200 movies stars according to their mean gross.

First of all, let's compare the cleaned movie star arranged in descending order of mean gross with some online data--*top 100 stars in leading roles at the worldwide box office.* From this out side data we can see that Tom Hanks, Robert Downey, Jr., Johnny Depp, Daniel Radcliffe (Harry Potter) ranked very high in the table. We pulled the data from the website, and merged with the data we generated and cleaned. Although the rank are somehow different (not a huge difference), movies stars who rank high in the online source data also rank high in our actors/actresses data. The reason that there is some differences in the ranking is that first our data calculates the mean gross of each movie star, while the online data calculates the sum. Another reason is that out data only contains top1000 rated movies from year 2000 to 2015. But in general, our data is consistent.

Figure-1 shows some top actors/actresses whose movies have high gross but also require high budget, for example, Johnny Depp who is famous for starring Captain Jack Sparrow in the "Pirates of Caribbean" series, Robert Downey Jr. who is famous for starring Tony Stark in the "Iron Man" series, and Hugh Jackman who is famous for starring Logan in the "Wolverine" series. However, in order to produce a profitable movie we are looking for movie stars who located in the upper left portion of the graph just like what we expected for directors. Therefore, figure-2 shows some ideal actors/actresses that we could choose, for example Steve Carell who is famous for starring in the recent movie "The Big Short", Bradley Cooper, who is famous for starring in "American Sniper", "Guardians of the Galaxy" and the "Hangover", and Cameron Diaz who starring in "Sex Tape" and the "Shrek" series.

In general, movie stars does have a huge effect on the box office. Famous movie stars tend to help the movies they starred to gain earn more.

Having decided on directors, stars, and what genres of movies are popular, we want to find out the relationship between audiences’ thoughts about a film and the film’s box office. In our dataset, we have the ratings on IMDB and how many people gave that rating. The number of ratings a film get on IMDB is an indicator of how many people know about and care about this film. Since box office are mainly concerned with how many people went to see the film, we expect to draw some connections between number of people that rates a film and the actual box office of that film. And of course, rating score is also an indicator of audiences’ opinions towards a movie. Despite that there exists a number of commercially successful movies that are not highly rated, such as Warcraft, people would generally expect higher rated movies to receive higher box office. People usually consults websites like IMDB for good films to watch, so intuitively receiving a higher review would help the box office. We also included the Meta Score, which is also a kind of rating from metacritic. For the following discussion, we separate our movie data sets into low budget movies, which are movies with budget of lower than twenty-five thousand dollars, and normal/higher budget movies. There is no wide-acknowledged cutoff for low budget movies. This cutoff is given by Stephen Follows in his study carried out in UK on film professionals.

For low budget films, there is almost no association between number of ratings and box offices, while for higher budget films, there is clear indication that higher number of ratings leads to higher gross. We start with plotting the number of ratings against the gross box offices of films. Each dot represent a movie. This graph lacks proper association between these two variables. We can see the rough trend of more number of ratings leads to higher box offices. The main problem with this graph is that lots of dots are cramped on the lower-left corner. So for this graph, we have two things to do. We need to separate the data so that they do not appear as a whole black area on the graph, and we need to find out any possible relationships between the number of ratings and gross. For separating the data, we consider separating low-budget films out. We can draw two separate graphs for low budget films and comparatively higher budget one. Plus, for the higher budget group, we can observe that there is a general trend of number of ratings being proportional to gross. To show this, we group the films according to their number of ratings into seven groups and draw two box plots. The width of each box is proportional to the square root of number of films in this group. From these two plots, we can see quite clearly that for the low budget films, there is still quite little correlations between gross box office and the number of reviews. Most of low-budget falls into the “no one care” group where they receive quite limited reviews and gross. The seven mean lines of the seven box plots take on a snake shape. This lack of correlation can also be seen from the linear regression line of the low-budget films.

Gross = 0.1561 \* NumberOfReviews + 7878369, R-squared = 0.2276

The R-squared value is quite low. On the other hand, for the higher budget group, the boxplot clearly demonstrates that higher number of ratings is associated with higher gross.

We can get a linear regression line of

Gross = 0.3402 \* NumberOfReviews + 31424041, R-squared = 0.4096

For the movies that received zero to six hundred and fifty reviews, the growing trend is quite obvious. We have a r-squared value of 0.41. Although the correlation is still quite limited if we only look at the number, this factor is quite worth noticing given that box office is generally not easy to predict. Consider budget, which is probably the most important factor to consider for people who invest in films, only has a r-sqaured value of around 0.5. It is worth noticing that for the last two groups, we are seeing obvious drops in gross with increasing number of ratings. This might just be coincidental, because we are having much less sets of data about these two groups since they are usually big titles. Or this could be people tend to comment on bad movies which can thus explain the phenomenon. But in general this part of the graph does not give us too much information, if we want to generate more precise conclusion we need to do a hypothesis test or need to find more evidence to prove our guess.

We can see similar traits in ratings on IMDB. The correlation between ratings and gross are more substantial for higher budget films than low budget films. From the following boxplot, we can clearly see that for low budget films, they do not earn much regardless of their ratings, except for a couple of extreme outlier. On the contrary, for the higher budget films, the gross is very proportional to the rating on IMDB. We can draw similar graphs for gross vs Meta Score and observe very similar trend. The reason for this phenomenon might be that people look at the rating of the movies they are interested before they actually watch them in the theater. Also we can see that, in the middle range of the rating scale, the variance is larger and there exits more outliers. This could be the reason that some movies are the second or third movie in a series, in which the first one achieves great success. Audiences may as well follow up for the second or third episodes even if the movies are not as good themselves.

In fact, we cannot find out any clear relations between lower budget films and ratings nor number of ratings. In other words, it may be risky to make a low budget movie even though the low budget seems to be having a lower threshold to enter. For normal budget movies, audiences’ preferences affect the gross box office almost as efficiently as budget, and is definitely worth taken care of. If one is investing in a movie costing more than twenty-five millions of dollars, proper publication and advertisements are definitely needed. Actually, in 2012, Google published their model in estimating the box office of a film before the movies came in theater. They also wanted to look into what factors that reflex audiences’ opinions, but they took a different approach. They looked at the relationship between number of film related searches and box offices. In fact, doing a simple linear regression on these two variables would generate a R-squared value of 0.7, which is far better than what we could do. Ultimately, using the the volumes of trailer related search, together with some minor tweaks, they could reach a 94% accuracy in predicting a movie’s box office four weeks prior to on theater. This is clearly an demonstration on the importance of advertising in the field of films

The learning from this is that for low budget

Reference:

<https://stephenfollows.com/average-budget-low-micro-budget-film/>

https://ssl.gstatic.com/think/docs/quantifying-movie-magic\_research-studies.pdf