**Final Report (Group Project Phase 3)**

**Introduction**

An average of 708 films have been released over the past 20 years, but according to Forbes, 80 percent of them are losing money. In other words, it is impossible to hang 300 or 500 movies simultaneously that consumers do not require. This result drove a curiosity to ask the question “Then what is the secret of the 20% of movies that are profitable?”. Supply and demand are closely tied to the success of a movie. Compared to the size of the market, excessive volume is released at once, leading to a decrease in the quality of the movie, which leads to a decrease in audiences except for some. Also, since there are many movies released, it is a problem to quickly change the movie to another new movie in the theater for the next movie. Therefore, audiences have many options for movie choices, and only a few popular movies attract a lot of audiences. To reach a break-even point is also critical as operating at a deficit is risky to production companies and directors. Hence, identifying the desired scenarios and genres of the audience in the process of designing a movie scenario can pave the way for box office success. Moreover, recognizing external variables that negatively affect the success of a movie, movie producers and directors can minimize their risk.

Our team aimed to utilize the movies dataset, to find out several factors impact on film’s commercial success so that production companies and directors can maximize their competitive advantage by creating a movie with high demand.

**Data Assessment**

The dataset, ‘The Movies Dataset Metadata on over 45,000 movies’, was collected from Kaggle and it gained 6 million ratings from over 270,000 users. It contains metadata for movies that were released on or before July 2017.

We put revenue as a dependent variable and selected budget, holiday and genre as independent variables. We initially chose 5 independent variables, runtime, vote average, genre, budget and holiday, but the runtime and vote average proved to be insignificant impact on the revenue, while the budget, genre, and release date of the movie proved to be significant indicators of revenue.

‘Revenue’ is an important independent factor for determining commercial success. It is a primary component of the breakeven point, a figure that measures the success of a project, the commercial success of a movie. ‘Genre’ variable can be the criterion for selecting a movie for audience as a initial step to choose a movie to watch, ‘Budget’ can impact on a quality of movie and ‘Holiday’, which is Christmas season, has relatively more leisure time given to people so that its able to attract more audience. These 3 dependent variable, Genre, Budget and Holiday shows the most significant relationship from our dataset.

**Data: The Movies dataset (‘movies\_metadata.csv’, ‘movies\_credits.csv’)**

**Preparation**

1. **Data pre-processing**

**1) Cleaning Process**

Step 1. Examine dataset

: Removed null values that were shown over multiple rows and columns in ‘movies\_credits’.

Step 2. Remove unnecessary columns

: Removed unrelated attributes such as ‘adult’, ‘belongs\_to\_collection’,

‘homepage’, ‘imdb\_id’, ‘overview’, ‘poster\_path’, ‘spoken\_language’,

‘tagline’, ‘video’, ‘vote\_count’, ‘production\_countries’ and ‘original\_title’.

Step 3. Remove duplicate rows

: Removed duplicates rows using dataframe.drop\_duplicates()

Step 4. Remove missing values

: Deleted missing values in original\_language(11), popularity(3), production\_companies(3), production\_countries(3), release\_date(87), revenue(3), runtime(260), status(84), title(3), vote\_average(3) by using dataframe.dropna().

Step 5. Remove non-English films in ‘orginial\_laungage’ column

: Removed all non-English values by using command to remain only English values in the column. Created a new dataframe that included all rows where the value of a row was equal to “en”.

< Movies\_dropen= movies\_dropdups\_dropmissing

[movies\_dropdups\_dropmissing.orginal\_language== ‘en’]>

Step 6. Remove movie status that is not ‘Released’

: Deleted values of ‘Canceled’, ‘In Production’, ‘Planned’, ‘Post Production’, and ‘Rumored’. By deleting the other status to focus on released movies.

Step 7. Remove [] in genre, production\_companies, and crew

: Deleted missing values in genre, production\_companies, and crew that had only square brackets and no values within by using dataframe= dataframe[dataframe.column != ‘[]’].

Step 8. Create dummy variable for genre

: To make genre values into dummy variables, converted the column from a string to a list of dictionaries and using the following code:

<genre\_column = []

for row in movies1['genres']:

row = eval(row)

if len(row) != 0:

genre\_column.append(row[0]['name'])

else:

genre\_column.append(None)

movies1['single\_genre'] = genre\_column>

Then, for ‘sinlge\_genre’ generated above, created dummy variables by using pd.get\_dummies.

Step 9. Create holiday(Christmas) dummy variable from release\_date

: To make a dummy variable for movies that was released in December, extracted months from the released data and used if/else statements to find whether the month is December or not. Converted into ‘1’ if it is in holiday season, and ‘0’ if not by using the following code:

<date\_column = []

for row in movies1['release\_date']:

if str(row) == 'nan' or len(str(row)) not in {8, 9, 10}:

date\_column.append(None)

else:

try:

date\_column.append(datetime.strptime(str(row), '%Y.%m.%d'))

except:

date\_column.append(datetime.strptime(str(row), '%Y-%m-%d'))

movies1['release\_date'] = date\_column

movies1 = movies1.dropna()

movies1['holiday'] = pd.DatetimeIndex(movies1['release\_date']).month

movies1['holiday'] = np.where(movies1['holiday'] == 12, 1, 0)>

Step 10. Remove all movies that has zero amount of revenue

: To examine the correlation between revenue and independent variables, removed values that had revenue =0.

**2) Data type conversion**

1st conversion

: When merging two datasets, ‘credit.csv’ file data type was not aligned with ‘movies\_metadata.csv’.

Converted credits type: int64 to float(object) by using the following code:

<print (credits.columns)

credits = credits.astype(str)

print(credits.dtypes)>

|  |  |
| --- | --- |
| movies\_metadata.csv | object |
| movies\_credits.csv | cast\_dtype: object, crew\_dtype: object,  *id\_dtype: int64 (converted\*)* |

2nd conversion

: When printing out the regression model using <regression = OLS(dv,iv).fit()>, an error code “Pandas data cast to numpy dtype of object. Check input data with np.asarray(data).” has occurred.

Therefore, changed the data type to float and then ran a regression model. <movies1\_regression = OLS(dv.astype(float), iv.astype(float)).fit()>

**3) Data transformation**

: Import two separate datasets and merge into a new dataframe using the following code.

< metadata\_credits\_merge = pd.merge(metadata, credits, how='inner', on = 'id') >

**4) Variable selection**

**Model: Regression Model**

* dv(dependent variables): revenue, popularity
* iv(independent variables) budget, runtime, vote average, holiday genre

1. **Statistics**

**1) Statistical model for analysis in business context**

1. **Statistical model:** Regression Model
2. **How can it be used to answer business question?**

* Business question: “Which factors lead to a successful movie?”
* By using regression analysis to find out the relationship between variables. Finding out which independent variables affect on the dependent variables to verify the success factor of a movie.

**2) Summary of data reports**

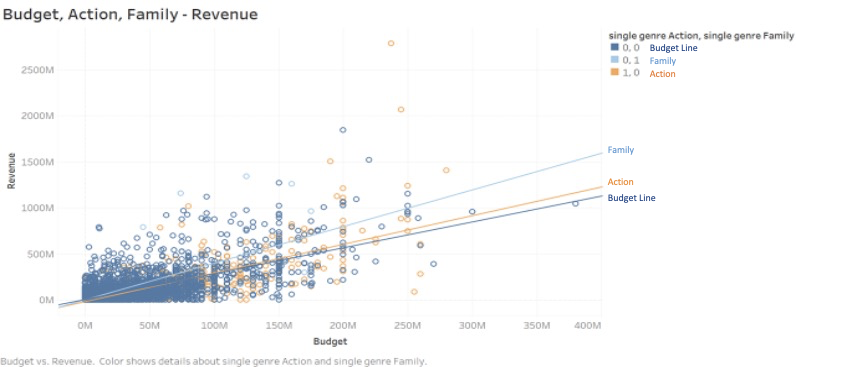
1. **Business Report 1**

- DV: Revenue

- IV: Budget, Action(Genre), Family(Genre)

Revenue = 3.05+(2.96\*Budget)+(-1.31\*Action)+(5.37\*Family)

|  |
| --- |
|  |
| For instance, every 1 million USD in budget, a movie is expected to return 3.05 million in revenue.  - If a movie is of the Action genre, the revenue is expected to decrease 1.31 million.  -If a movie is of the Family genre, the revenue is expected to increase by 5.37 million.  This set of independent variables explains 54.9% of variance in revenue. |

****

This graph displays the relationship of budget, genre(action, family) and revenue. The y-axis represents the revenue and the x-axis represents the budget. The business report 1 visualization explains if a production company has an equal amount of budget for each genre, family movie will be likely to generate more revenue than action movie.

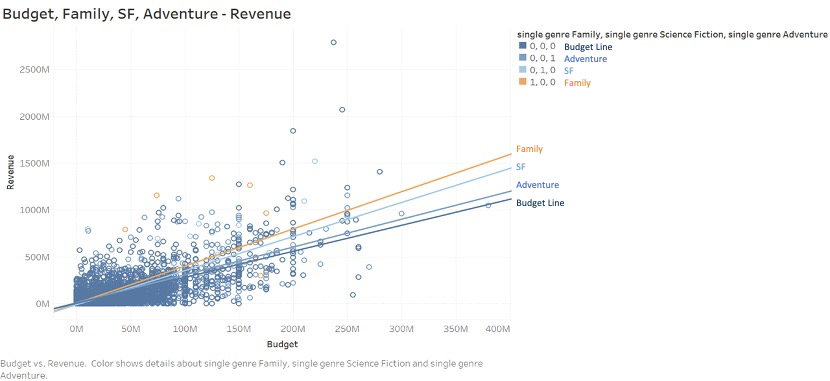
1. **Business Report 2**

- DV: Revenue

- IV: Budget, Family(Genre), Adventure(Genre), Science Fiction(Genre)

Revenue=6.37+(2.90\*Budget)+(1.87\*Adventure)+(5.88\*Family) +(2.36\*SciFi)

|  |
| --- |
|  |
| For instance every 1 million USD in budget, a movie is expected to return 6.37 million in revenue.  -If a movie is of the Adventure genre, the revenue is expected to increase 1.87 million.  -If a movie is of the Family genre, the revenue is expected to increase by 5.88 million.  -If a movie is of the SciFi genre, the revenue is expected to increase by 2.36 million.  This set of independent variables explains 54.9% of variance in revenue. |



Above the graph shows the relationship of budget, genre (Family, SF, Adventure) and revenue. The y-axis represents revenue and the x-axis represents budget. Through this visualization for business report 2, if a production company has an equal amount of budget for each genre, the family genre movie will be more likely to generate the highest revenue among the three genres.

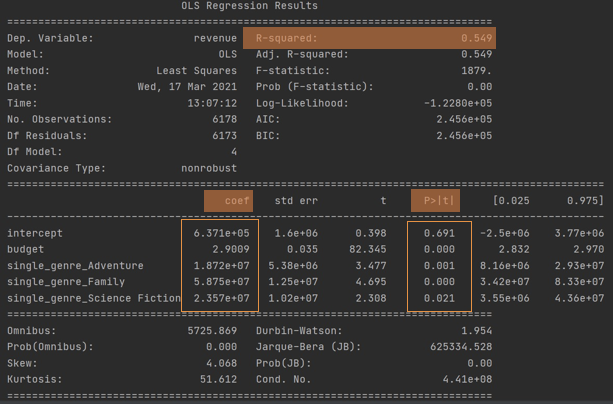
**Business Report 3**

- DV: Popularity

- IV: Budget, Action(Genre), Romance(Genre), Comedy(Genre)

|  |  |
| --- | --- |
| Before  y-intercept | Popularity=(1.58\*Budget)+(4.53\*Action)+(4.55\*Romance)  +(4.8\*Comedy) |
| After  y-intercept | Popularity=6.89+(1.01\*Budget)+(-.0.03\*Action)+(-1.33\*Romance)+(-1.16\*Comedy) |

When running the regression analysis with the y-intercept, the coefficient of independent variables has dramatically changed. Therefore, our group decided to exclude the result.



1. **Business Report 4**

- DV: Revenue

- IV: Holiday, Budget, Adventure(Genre), Animation(Genre), Family(Genre), Science Fiction(Genre)

Revenue = -9.71+(1.21\*Holiday)+(2.86\*Budget)+(2.13\*Adventure)+

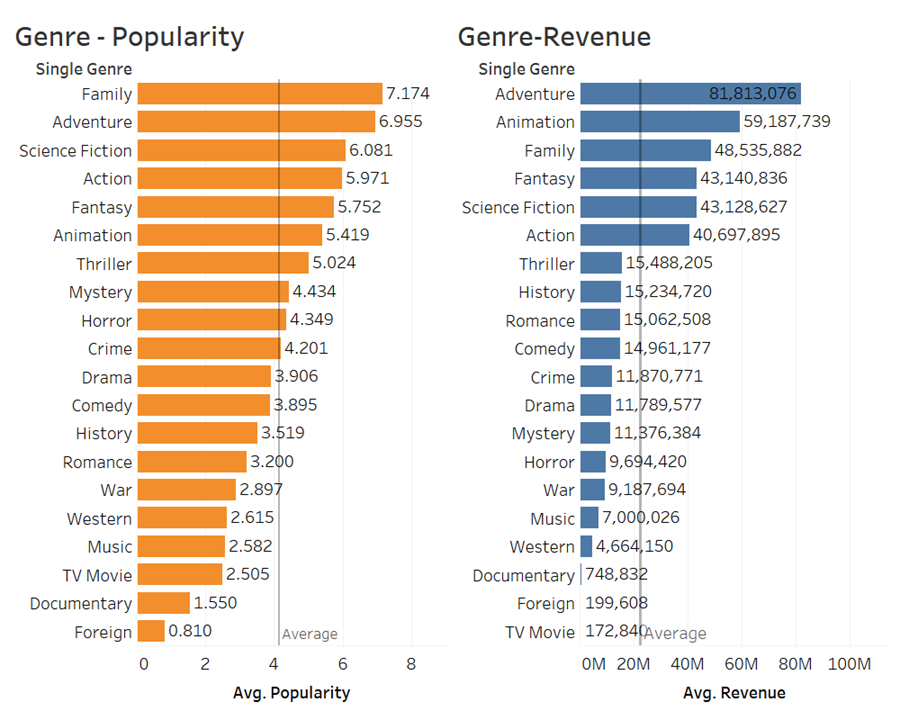
(5.44\*Animation)+(6.20\*Family)+(2.63\*SciFi)

|  |
| --- |
|  |
| For instance, if a movie is released in the holiday season, the revenue is expected to increase by 1.21 million. Additional 1 million USD allotted to the budget is expected to increase the revenue by 2.86 million.  -If a movie is of the genre Adventure, the revenue is expected to increase 2.13 million. -If a movie is of the genre Animation, the revenue is expected to increase 5.44 million.  -If a movie is of the genre Family, the revenue is expected to increase by 6.20 million.  -Lastly, if a movie is of the genre Science Fiction, the revenue is expected to increase 2.63 million.  This regression model explains 55.2% of the variance in revenue. |

**3) Evaluation of the model**

Most of the movies we can actually watch in theaters are movies that have the primary purpose of ‘a box office hit’ which have a tendency to be the opposite of an artistic film. Movies require astronomical capital compared to other media industries, so it is natural to pursue ‘commerciality.’ Generally, the more budget the production companies/directors have, they will have a better chance to bring more revenue as it improves the film quality. However, in reality, the movie production party has always constrained their budget and can not help considering its Return on Investment (ROI).

People who like movies have their own preference, especially for their favorite genre, however, popular genres are more likely to be welcomed by the audience. Independent film festival’s official, Elliot Grove emphasized that Scenario writers and filmmakers should prioritize understanding genres. It has become more and more critical in today’s crowded movie market. Average 700s of new movies are released each year, but only a few are available on online distribution and fewer are available to the theater.[[1]](#footnote-1)  
 Our group utilized tableau to see the relationships between popularity of gere and revenue of genre and the result is as follows:

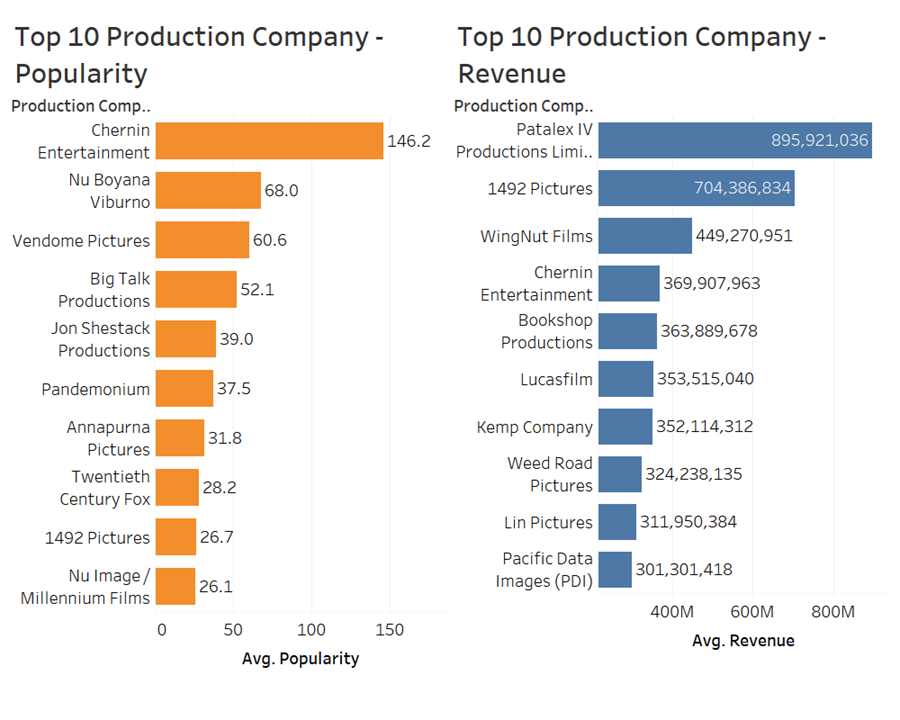


The top three genres considering popularity are Family, Adventure, and Science Fiction in order while the top three genres considering revenue are Adventure, Animation, and Family. Through the graphs, Adventure and Family are great to select. Moreover, our group tried to find some relationships between genre and revenue under the same budget and found that some models show that when targeting the right genre movie the revenue actually grows compared to selecting the other genre. The results can mean three important things for the business; First, the results give an answer on which genre is better to produce or invest on. Second, For directors and movie production companies, the result will give answers for which genre generates more revenue or popularity. Third, production companies can get insights about which types of genres should be released in December. Especially, according to the regression model, Family genre was shown the most welcoming genre in general, however, the revenue can be more maximized depending on adding which type of genre. This result is also supported by the research from Claremont McKenna College.[[2]](#footnote-2)

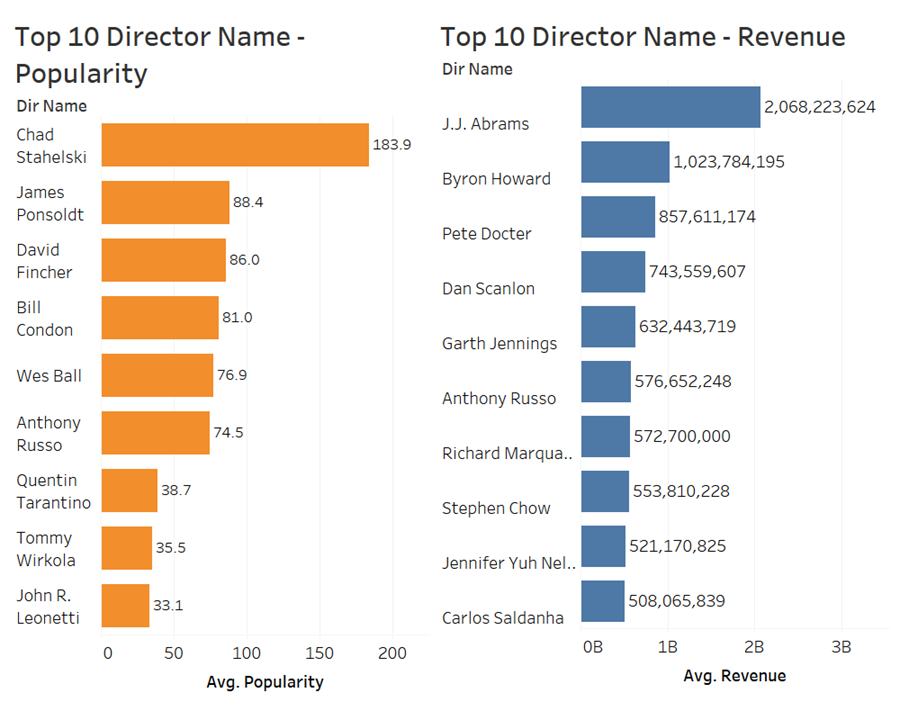
1. **Application & Knowledge**

**1) Application for business stakeholders** The proposed models can improve the business function for directors and production companies. For directors, through analyzing which type of film generates more revenue, they can utilize the result to present it as a proof for attaining external investments. By showing the result, they can bring more investors to invest in their films and this will result in an increase in budget. Additionally, directors can use the result to understand the audiences’ preferences and produce high-quality movies. For production companies, the result will work as a source of determining the best movie scenarios which facilitate effective investment. Moreover, the company can increase revenue by eliminating the loss by using the analytic strategy presented by the result.

The further analysis visualized by Tableau also can intrigue business stakeholders. Besides regression models, our group used extra primary variables and this visualization process allows us to identify the stakeholders’ characteristics and their relationship. The ranking of the movies based on the revenue or popularity was driven by genre, production companies, and directors for each. The value of revenue and popularity is set as the average to avoid a particular genre dominantly having a number of movies.

Production Company - Popularity, Revenue

Above the two graphs represents the relationship between a production company and popularity and the relationship between production company and revenue. The top three of the production companies in terms of popularity are Chernin Entertainment, Nu Boyana Viburno, and Vendome Pictures in order. In contrast, Patalex IV Productions Limited, 1492 Pictures, and WingNut Films are the top three production companies in terms of revenue.

Director Name - Popularity, Revenue  
  


These graphs show the relationship between director and popularity and the relationship between director and revenue. Chad Stahelski, Jame Ponsoldt, and David Fincher are the top three directors in terms of popularity while J.J Abrams, Byron Howard, and Pete Doctor are the top three directors in terms of revenue.

Three sections of the statistics visualization results represent that revenue and popularity are not proportional relationships. It is easy to think that they have a direct co-relationship but they are not perfectly proportional relationships in our data. Marketing and having business in other industries would be the factors it leads to.

**2) Review the project**

The report showed the result of the question “Which factor leads to a successful move?”. Genre played a key role in the revenue. Depending on the genre type, the revenue generated could be estimated. However, to examine thoroughly, the report can add more improvements

1. Sentiment analysis

The sentiment analysis can be done by extracting reviews and social media posts of movie audiences. By classifying the polarity of given text, the emotional tone embedded in the words can be collected. Using the sentiment analysis, our group might identify which genre acts as emotive triggers that change the customer mood. For example, examine the different emotions after watching romance comedy and thriller.

1. Keyword analysis

The keyword analysis can be done by using the movies keywords dataset by finding out the common keywords with high revenue or popularity. The keywords can be utilized in further actions such as marketing operations. Specifically, the keywords can be inserted in advertisements, descriptions, and any other sources related to the movie to take audiences’ intention and generate more revenue.

1. Examine other variables which are not given in the report

Other variables such as actors, marketing expenses, and motion picture content rating system. can be taken into consideration for further analysis. The independent variables might have a strong correlation with the revenue or popularity. By accounting more variables to examine the question will enhance the quality of the report.

1. Elliot Grove. “Genre’s Importance to Screenwriters and Filmmakers”, Raindance, 11 May 2017, https://www.raindance.org/genres-importance-screenwriters-filmmakers/ [↑](#footnote-ref-1)
2. https://scholarship.claremont.edu/cgi/viewcontent.cgi?article=1237&context=cmc\_theses [↑](#footnote-ref-2)