MARK5828 Advertising Analytics

Individual Research Project - FINAL REPORT

Analysis of Movies Dataset

Danica Huei Ye YONG

Z5142161

Tutor Name: Sunny LEE

Tutorial Class: 11164

Student ID	**Group Member Name**
z5235375	Yuejia WANG
z5189830	Suyue ZHANG
z5199917	Wanting JIANG

DATE: 2 MAY 2019

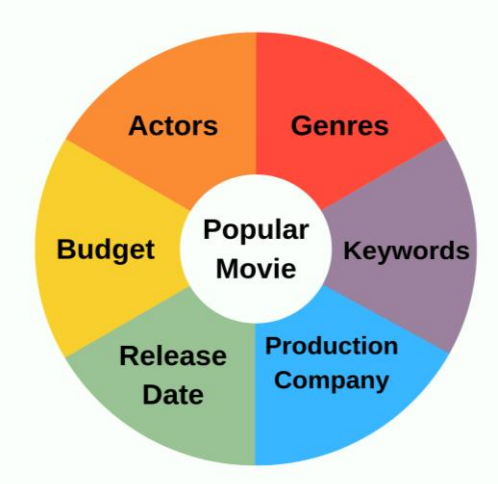
I. Project Goal / Research Question

With movie tickets costing $20 each, movie makers need to make sure that the movies they produce are both beneficial to the audience and those who invest in the movies. Carrilat, Legoux and Hadida (2018) states that due to movies being ‘experimental’, there are many variables that contributes to a movie’s popularity as the audience’s preferences may vary.

Asad, Ahmed and Rahman (2012) mentioned 4 factors that influence the popularity of a movie are viewers, start actors, market trends and budget, while the specific variables are director, budget, main actors, release date, genres and keywords. This is supported by Liu, Chen and Guo (2016) who also added that majority of the previous studies on movie prediction focuses on similar specific variables.

Without a doubt to determine what makes a movie popular and finding the right balance between having the right elements of a movie and influencing audiences to see the movie is challenging. Therefore, in order to understand which key elements are essential in making a movie popular as well as helping production companies determine which movies are worthwhile in investing. I would like to create a movie marketing strategy for production companies to take note of what key elements is essential in a movie, thus leading to a higher probability of the movie becoming popular.

My research question is to explore the six factors: actors, genres, budget, release date, production company and keywords, if they contribute to a movies’ popularity as seen in Picture 1,



Picture 1: Six Areas of Focus for Popularity of Movie

I've made several hypotheses to explore this research question:

- ➔ Budget is significant.
- ➔ Warner Bros is a significant production company.
- ➔ The 18th, 29th days of the month and the month of November are significant.
- ➔ The keyword during credit stinger is a significant keyword.
- ➔ The genres Action and Adventure are significant.
- ➔ The actors Jennifer Lawrence, Scarlett Johansson, Josh Hutcherson and Margot Robbie are significant.

After further analysis, I will either accept or reject the hypothesis which will be supported by the linear regression result.

II. Data Collection

This dataset has been collected from IMDb for 291 movies that were released from 1937 to 2017, the IMDb website is known for movie and television show ratings.

The table below shows the description for each dependent and independent variable. New columns such as actor_xxx, director and release_xxx has been created to see if it contributes to the popularity of movies.

Columns	**Descriptions**	**Notes**
popularity	Popularity rate of the movie	Dependent Variable
genres_xxx	The dummy variables of genres	Independent Variable
budget	The total money spent on making the movie	
production_companies_xxx	The dummy variables of the production companies	
release_xxx	The dummy variables of release day/month	
Keyword_xxx	The dummy variable of the keywords	
actor_xxx	The main two leads of the movie	

director	The director of the movie	
overview	The overview for movie	Other String columns
crew	The crew of the movie	
cast	The cast of the movie	
tagline	The tagline of the movie	
title	The title of the movie	
Keywords	The keyword of the movie	
release_date	The release year of the movie	
production_companies	The movie production companies	
original_language	The original language of the movie	
original_title	The original title of the movie	
genres	The genre of the movie	
prod_countries	The country where the production companies are	
runtime	The duration of the movie	

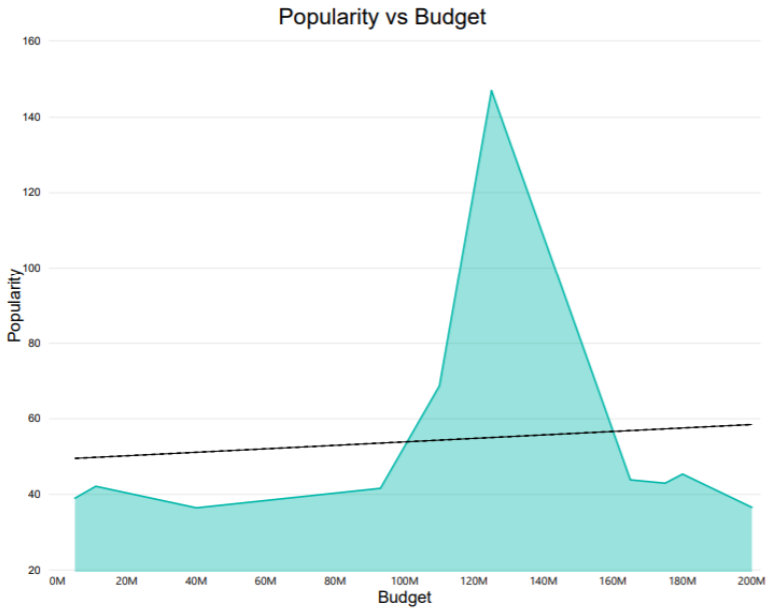
Table 1: Description of Each Column in Dataset

III. Data Exploration

For Plot1, the word cloud shows the top genres based on popularity and genres_Action and genres_Adventure are the most significant genres.

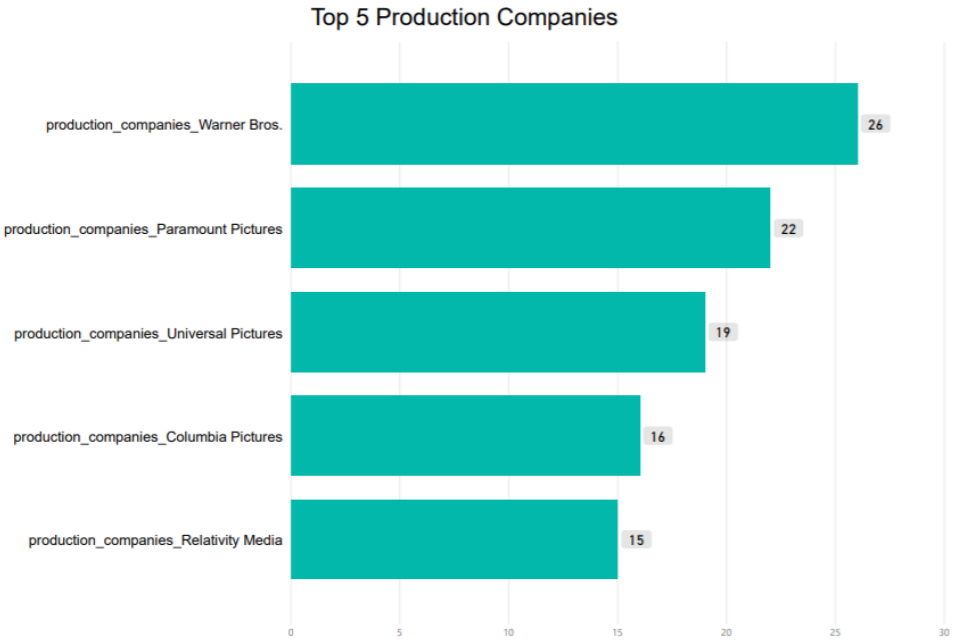
Plot 1: Genres by Popularity

Plot 2 shows the popularity against budget; which allows us to see that a bigger budget does not mean a more popular movie, but it shows that budget does have significance in making a movie popular.



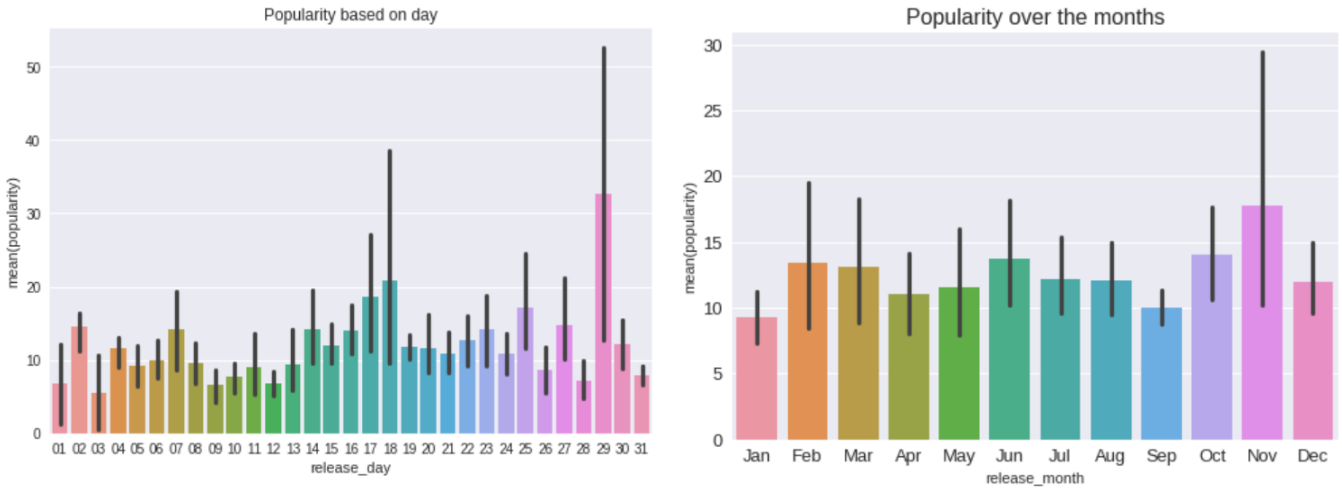
Plot 2: Budget by Popularity

Plot3 shows the top 5 production companies and production_ companies_Warner Bros. is the most significant production company.



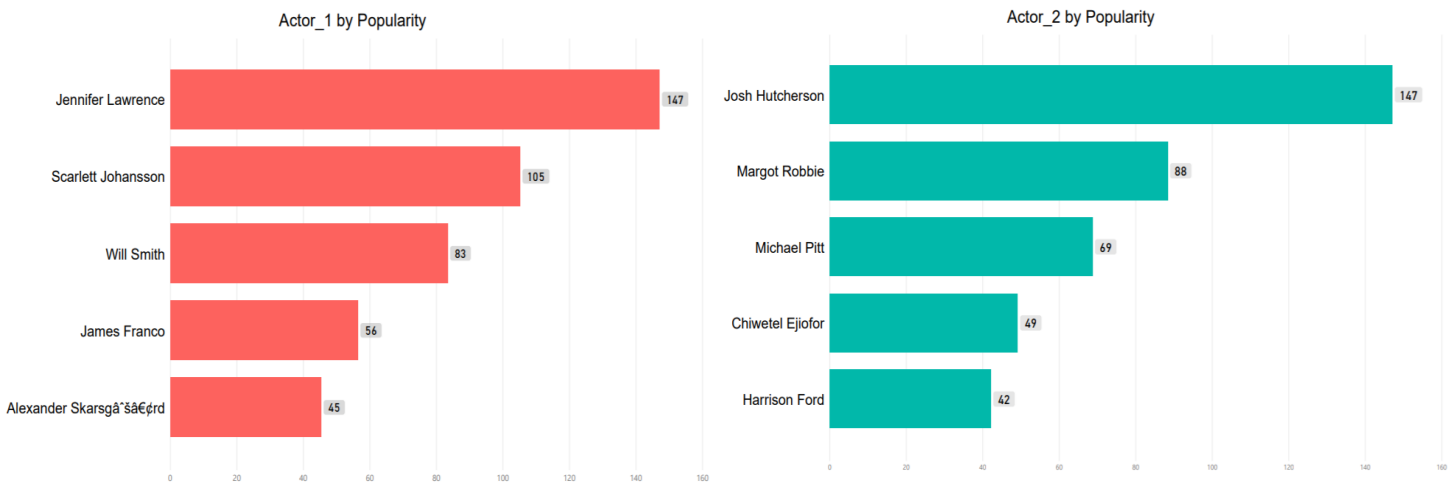
Plot 3: Top 5 Production Companies

Plot 4 shows us two bar charts for both the release_day and release_month respectively based on popularity, from the left-hand plot the days 18 and 29 are significant and on the right-hand plot the month November is significant.



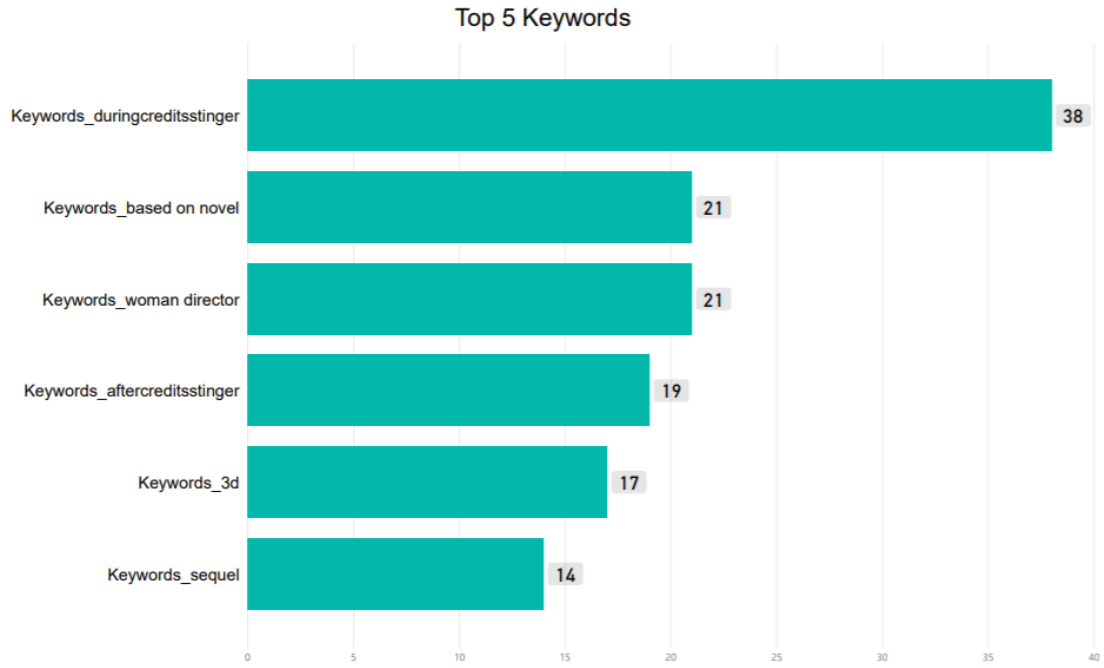
Plot 4: Release day and Release Month by Popularity

Plot 5 shows us the top 5 main two actors based on popularity, for actors Jennifer Lawrence, Scarlett Johansson, Josh Hutcherson and Margot Robbie



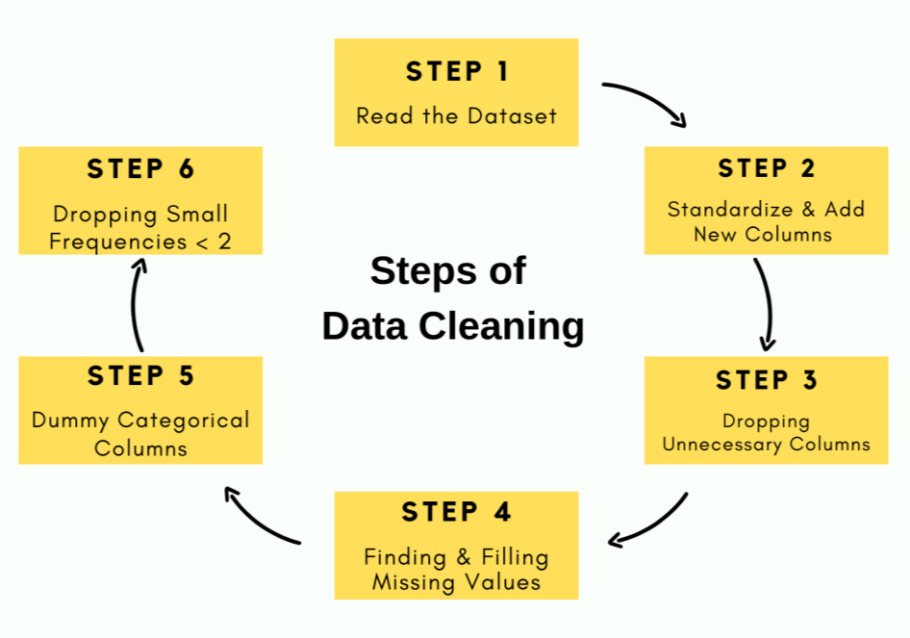
Plot 5: Top 5 Actors by Popularity

Plot 6 shows the top 5 Keywords based on popularity and Keywords_duringcreditsstinger is a significant keyword. During Credit Stinger means that after the movie has ended, a behind the scene video or a video to drive up interest for a sequel is running at the same time as the credits.



Plot 6: Top 5 Keywords Based on Popularity

IV. Data Cleaning



Picture 1: Stages of Data Cleaning

These are the steps of Data Cleaning:

1. Reading the dataset
   Using the pandas function to read the dataset and assign it to a variable.

2. Standardizing and Adding New Columns
   For the release_date column, the dates were not standardized thus I used a pandas function to fix it and separated day and month into release_day and release_month respectively.
   In addition, I extracted extra information from columns such as Crew and Cast to get new columns such as actor_1, actor_2, director.

3. Dropping Unnecessary Columns
   Irrelevant columns such as “Unnamed: 0”, “Unnamed: 0.1”, “Unnamed: 0.1.1”, “id”, “homepage”, “original_title”, “overview”, “spoken_languages”, “title”, “tagline”, “Keywords”, “genres”, “production_companies”, “prodiction_countries”, “cast”, “crew”, “release_date”, “The sum of production_companies”, “The sum of Keywords”, “The sum of genres”, “Director_name(if more than 2 for the movie)_Clint Eastwood” were dropped as it was not relevant to the analysis.

4. Finding and Filling Missing Values
   After dropping the unnecessary columns, there were only two columns that had missing values, actor_1 and actor_2, which was filled with ‘no_actor’.

5. Dummy Category Columns
   The pandas function “get_dummies” was used for the new categorical columns such as actor_xx, director, release_day and release_month.

6. Dropping Smalle Frequencies < 2
   For columns that only appear once in the dataset is removed as well as the columns ‘actor_1_no_actor’ and ‘actor_2_no_actor’ which is used as the baseline for actors.

V. Data Analysis

As new categorical columns were created such as actor_1, actor_2, director, release_day, release_year, we need to use ‘get_dummies’ to dummy the variables for linear regression. The baseline for the columns is ‘actor_1_no_actor’ and ‘actor_2_no_actor’ that is removed after dummying the categorical variables.
The following pictures show the process of using ‘get_dummies’ and removing columns that only have a sum of one.

```
m1 = pd.get_dummies(m_drop)
m1 = m1.drop(columns=['actor_2_no_actor', 'actor_1_no_actor'])
```

Picture 1: Get Dummies of Categorical Columns and Removing Baseline

```
#removing low frequency columns
sum_df = m1.sum()
list_col = list(m1)

counter = 0
for x in sum_df:
  if x < 2:
    m1.pop(list_col[counter])
  counter=counter+1

m1.columns
```

```
Index(['budget', 'popularity', 'runtime', 'genres_Action', 'genres_Adventure',
       'genres_Animation', 'genres_Comedy', 'genres_Crime',
       'genres_Documentary', 'genres_Drama',
       ...
       'actor_2_Kristen Bell', 'actor_2_Lena Headey', 'actor_2_Margot Robbie',
       'actor_2_Meryl Streep', 'actor_2_Morgan Freeman',
       'actor_2_Natalie Portman', 'actor_2_Philip Seymour Hoffman',
       'actor_2_Robert Pattinson', 'actor_2_Samuel L. Jackson',
       'actor_2_Theo James'],
      dtype='object', length=238)
```

Picture 2: Removing Low-Frequency Columns that have a sum of one

After doing the above stages of data cleaning (refer to Picture 1 and Picture 2), I ran linear regression. The Y-variable is ‘popularity’ and the X-variables are ‘budget’, ‘genres_xx’, ‘release_day_xx’, ‘release_month_xx’, ‘production_companies_xx’, ‘actor_1_xx’, ‘actor_2_xx’.

In picture 3, there are x-variables that have p-values of less than 0.05 but there is evidence of multicollinearity therefore VIF > 5 should be removed.

	Coefficients	p	vif
Keywords_based on young adult novel	27.0622	0.00127511	4.56079e-35
actor_2_Theo James	-21.8854	0.00330815	2.76181e-35
actor_1_Shailene Woodley	-21.8854	0.00330815	3.23534e+12
Keywords_war	31.9542	0.0038546	2.358e-35
actor_1_Dwayne Johnson	29.7869	0.0128485	6.91846e-35
actor_1_Scarlett Johansson	46.8686	0.0274952	10.0154
Keywords_dystopia	17.6737	0.0276036	6.93383
release_month_10	6.60768	0.0414556	9.44105e-36

Picture 3: Initial VIF

After removing VIF > 5, the following image show the final regression results. There are 16 variables in total that have a p-value of less than 0.05 and VIF < 5.

	Coefficients	p	vif
production_companies_Lionsgate	21.0751	2.17808e-05	1.42094
actor_1_Scarlett Johansson	44.4533	3.4056e-05	1.96877
actor_2_Margot Robbie	33.333	0.00058649	1.63961
Keywords_sequel	14.1845	0.000923637	2.15974
actor_2_Theo James	-27.7495	0.00313015	1.55745
release_day_18	9.81743	0.00343914	1.45829
genres_Adventure	6.85823	0.00367393	2.62382
genres_Science Fiction	7.18988	0.00469368	2.1875
genres_Action	-6.76424	0.0123343	3.36394
production_companies_Walt Disney Pictures	10.8996	0.0174624	2.13981
production_companies_Legendary Pictures	13.3167	0.0233185	2.40067
actor_1_James McAvoy	-20.3116	0.03046	1.56465
actor_1_James Franco	19.4238	0.0420117	1.62353

Picture 4: After removing VIF > 5

There are 10 variables that have positive coefficients and they are ‘production_companies_Lionsgate’, ‘production_companies_Walt Disney Pictures’, ‘production_companies_Legendary Pictures’, ‘Keywords_sequel’, ‘release_day_18’, ‘genres_adventure’, ‘genres_Science Fiction’, ‘ actor_1_Scarlett Johansson’, ‘actor_1_James Franco’, ‘actor_2_Margot Robbie’.

In conclusion, there are a total of 10 variables that have positive coefficients.

1. Movies that are produced by companies such as Lionsgate, Walt Disney Pictures and Legendary Pictures will have more popularity as these production companies are known to have produced top movies such as The Hunger Games, Star Wars and Inception respectively. Vr and Pb (2014) tells us that production companies control the film industry as they are the ones with the money therefore production companies are crucial to a movies' popularity.

2. Movies that have genres such as Adventure and Science Fiction, or even both genres, are deemed popular from the results. Without a doubt movie that contain adventure and science fiction have always fascinated audiences from young to old, for example movies such as The Matrix, Inception and Star Wars are all considered classics and critically acclaimed movies that are popular. Kim (2018) has stated that having a multi-genre movie is more successful therefore having both genres in a movie will be good.

3. Movies that have sequels as supported by Apala, Jose and Motnam (2013) is a recipe for a successful movie, therefore movie sequels are popular as they want to relive the magic of the first movie.

4. Movies should also be released on the 18th of the month as the results shows that it is a good day to release movies.

5. According to the results, actors such as Scarlett Johansson, James Franco and Margot Robbie will give a movie more popularity due to their star power and brand. These actors are A-listers and having them as main leads of the movie will guarantee an instant hit, they have appeared in movies such as The Avengers, Rise of the Planet of the Apes and Suicide Squad respectively all which have been huge success in the box offices and are widely popular. Carrilat, Legoux and Hadida (2018) states that popular actors especially Oscar winners will appeal more to audiences and perform better in theaters, also Apala, Jose and Motnam (2013) mentions that the leading actress popularity is pivotal to the popularity of a movie.

Variables	P > |t|	Hypothesis(< 0.05%)
genres_action, genres_adventure	0.012 0.004	✓ Significant and Accept Hypothesis
Budget	0.147	✗ Not significant and Reject the hypothesis
release_day_18, release_day_25, release_day_29	0.003 0.086 0.822	✓ Release_day_18 is ✗ significant, partially accept the hypothesis
Production_companies_Warner Bros	0.726	✗ Not significant and Reject the hypothesis
Keywords_during credit stinger	0.742	✗ Not significant and reject the hypothesis
actor_1_Jennifer Lawrence, actor_1_Scarlett Johansson, actor_2_Josh Hutcherson, actor_2_Margot Robbie	0.000 0.001	✓ Scarlett Johansson, Margot Robbie are ✗ significant and partially accept the hypothesis

Table 1: Revised Hypotheses Based on Conclusion

Therefore a movie that contains the genre ‘adventure’ and ‘science fiction’,produced by Lionsgate, Walt Disney Pictures and Legendary Pictures acted by Scarlett Johansson, Margot Robbie and James Franco. In addition, released on the 18th day of any month and is a sequel will be popular in movie theatres.

VI. Future Plan

As this dataset has only 291 movies ranging from 1937 to 2017, it is insufficient as there are too little movies to analyze it completely. It would be good to have more movies in the dataset as well as more variables like gross profit, number of theatres, marketing budget and the rating for each movie. Therefore, I would like to use the dataset from Kaggle as there is a larger dataset on IMDB movies that I can use for further analysis or even web-scrap from the IMDB website itself, which would be a good learning experience in the future.

Possibly I would like to write a blog post about the factors that make a movie popular and delve deeper into the gender inequality of lead actors and directors. As I found that for directors, the male directors are trusted with a bigger budget than female directors thus I think that it would be good to analyze what is the reason.

VII. Future Career Goals

Currently, I am an undergraduate student and my main career goal was to become a data scientist before taking this course. I took this course as I am very interested in combining digital marketing as well as data science as it is a passion of mine. In addition, I wanted to further strengthen my Python skills and have more hands-on learning through the various datasets and group and individual projects.

I learnt a lot from both the group and individual projects, in many ways my confidence in doing data cleaning has improved tremendously and been able to do simple linear regression is good to learn from this course. By doing two projects, I am also able to put this on my resume. I do hope that I will be able to do the other datasets given in Moodle as projects as well.

Another thing I really enjoyed is the group project, even though we were strangers beforehand all four of us really came together and tackled Python even though we had not touched it before and grew in confidence as we finished our group project. Moving on to the individual project, we were not as worried or doubtful which was refreshing. In the future I am eager to do more projects, I will start from Kaggle as there are already datasets there that has been web-scraped and cleaned, also I plan on learning web-scraping using Python as I find it a useful skill to have as a data scientist. Furthermore, I plan on taking MARK5827 Product Analytics and MARK5827 Customer Analytics as I find this course quite exciting and relevant to enhancing my skills as a data scientist or hopefully a Marketing Scientist.

References

1. Apala, K.R. et al., 2013. Prediction of movies box office performance using social media. 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013), pp.1209–1214.
2. Carrillat, F., Legoux, A. & Hadida, R., 2018. Debates and assumptions about motion picture performance: a meta-analysis. Journal of the Academy of Marketing Science, 46(2), pp.273–299.
3. Kim, Iksuk & Kim, Hannearl, 2018. THE MORE, THE BETTER? MOVIE GENRE AND PERFORMANCE ANALYSIS. Journal of Business and Educational Leadership, 7(1), pp.105–113.
4. Liu, Ting et al., 2016. Predicting movie Box-office revenues by exploiting large-scale social media content. Multimedia Tools and Applications, 75(3), pp.1509–1528.
5. Wanderer, J.J., 2011. When Film Critics Agree: Does Film Genre Matter? Empirical Studies of the Arts, 29(1), pp.39–50.
6. Vr, Nithin. & Babu PB, S. 2014. Predicting Movie Success Based on IMDB Data.