MARK5828 - Individual Research Project

Interim Report

Danica YONG

z5142161

Members
Wanting JIANG z5199917
Yuejia WANG z5235375
Suyue Zhang z5189830

Research Question

- Exploring which x variables affect the popularity of a movie(y-axis)
  - genres/budget/language/actors/director/production_company
  - production_country/release_day/month/year

Data Collection

- I collected the data from web scraping the IMBD Movies Website
- A total of 291 rows

Columns	Descriptions
budget	The total money spent on making the movie
genres	The genre of the movie
genres_xxx	The dummy variables of genres
original_language	The original language of the movie
original_title	The original title of the movie
popularity	Popularity rate of the movie
production_companies	The movie production companies
production_companies_xxx	The dummy variables of the production companies
prod_countries	The country where the production companies are
release_year	The release year of the movie
release_xxx	The dummy variables of release day/month/year
runtime	The duration of the movie
tagline	The tagline of the movie
title	The title of the movie
Keywords	The keyword of the movie
Keyword_xxx	The dummy variable of the keywords
cast	The cast of the movie
actor_xxx	Took the first 2 actors from the cast

crew	The crew of the movie
director	The director of the movie
overview	The overview for movie

Data Cleaning

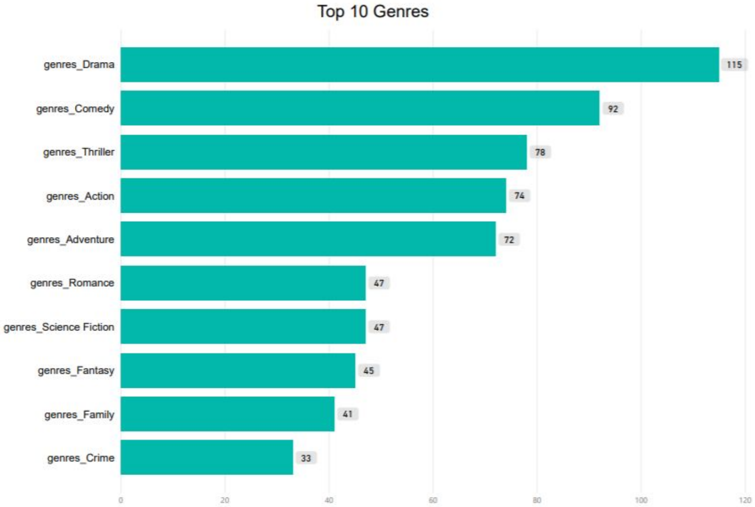
1. Filled in the missing data
2. Dropped columns that aren’t relevant
    - Homepage, Unnamed 0.1, Unnamed 0.1.1, original_title,id,tagline
3. Added new columns such as
    - actor_1,actor_2,director,prod_countries,

5. Fixing the date to get release year/month/day separately
6. Fixing the original language of the movie

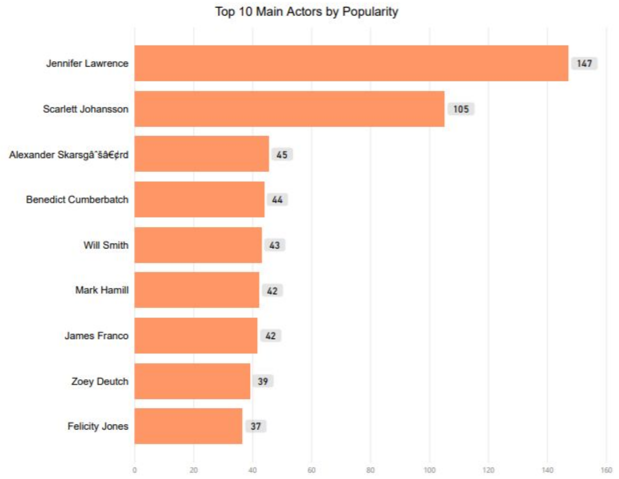
Data Exploration

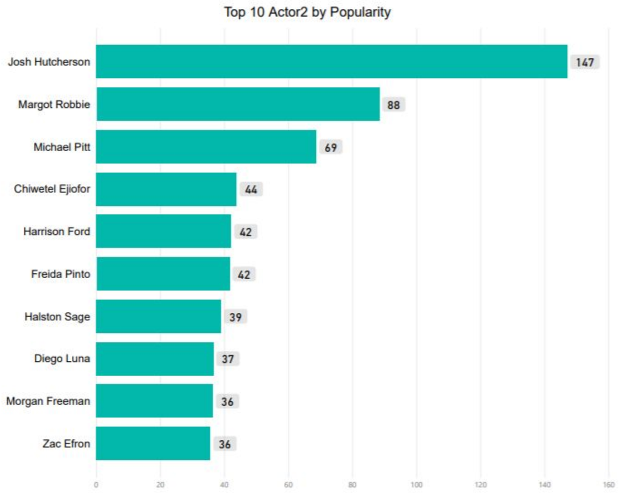
1. Top 10 Genres

This graph shows us the Top 10 Genres based on popularity, where we can see that genre_Drama is the top genre.



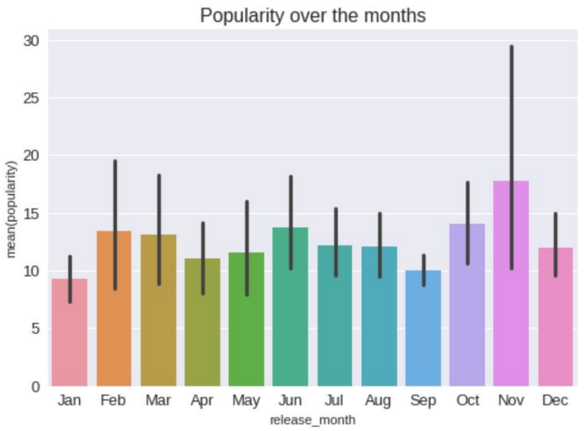
2. Top 10 Actors

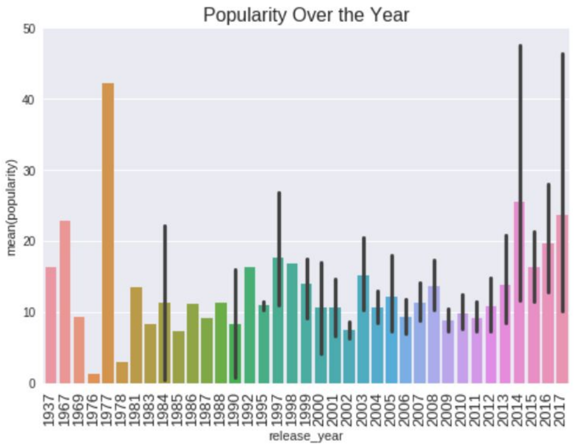




Both plots show the main two actors based on popularity, with both actors from The Hunger Games coming out on top. Jennifer Lawrence and Josh Hutcherson are significant x-variable.

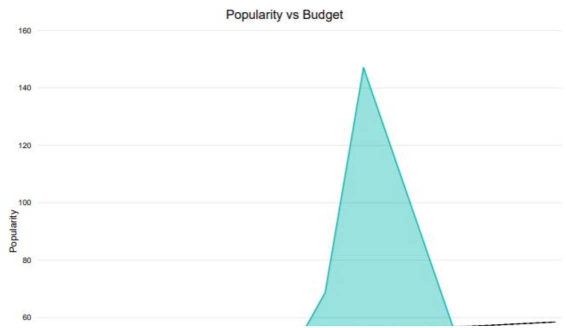
3. Popularity over the Months and Years





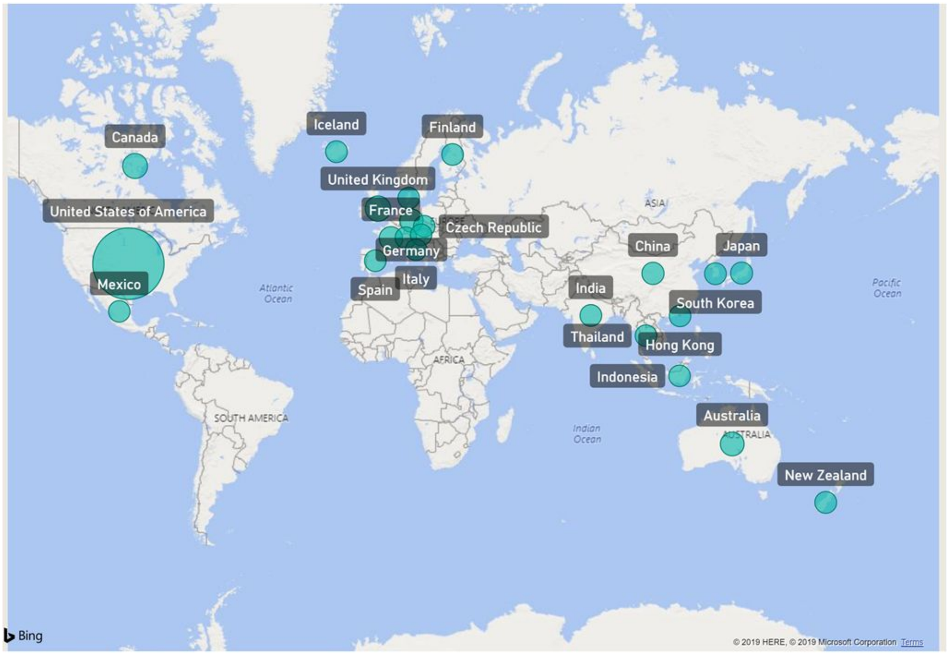
November is considered as a popular month to release movies and 1977 is the year with the highest popularity. We can see that for the release years plot, recently 2014 and 2017 are the years with the most releases.

4. Budget vs Popularity



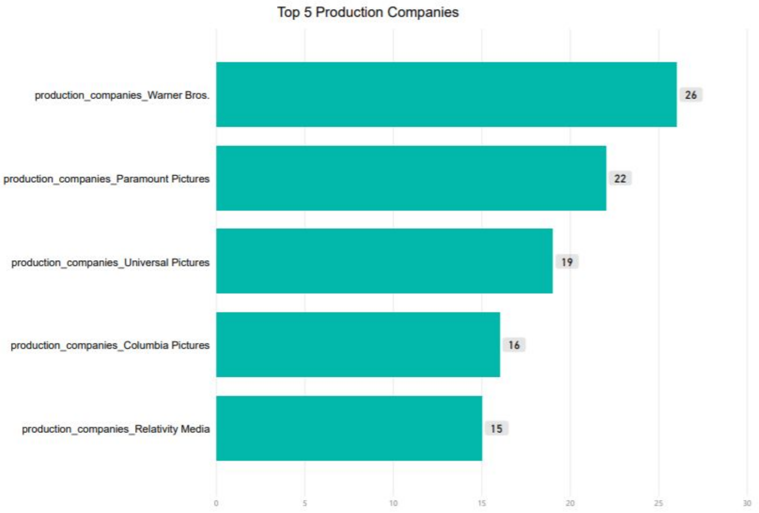
This plot shows to us that a higher budget doesn't equal to a popular movie, but a budget between 100 Million and 160 Million based on the dataset will be a good area for the budget of the movie.

5. Countries where movies are produced based on Popularity



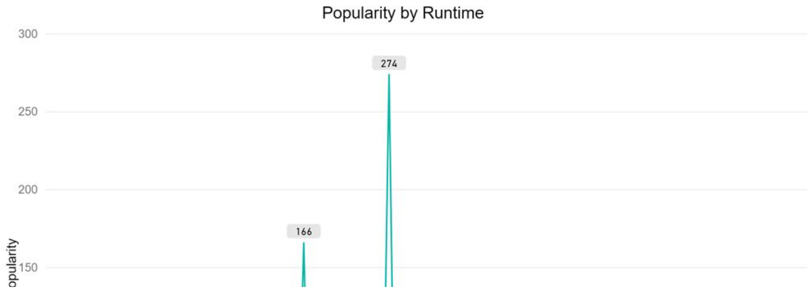
From the map, we can see that the United States of America has the most number of production companies that produce movies. Also, Europe has a good cluster of production companies.

6. Production Companies based on Popularity



We can see that Warner Bros is the most popular production company, therefore one factor that might affect the popularity of the movie is the production company.

7. Popularity vs Runtime



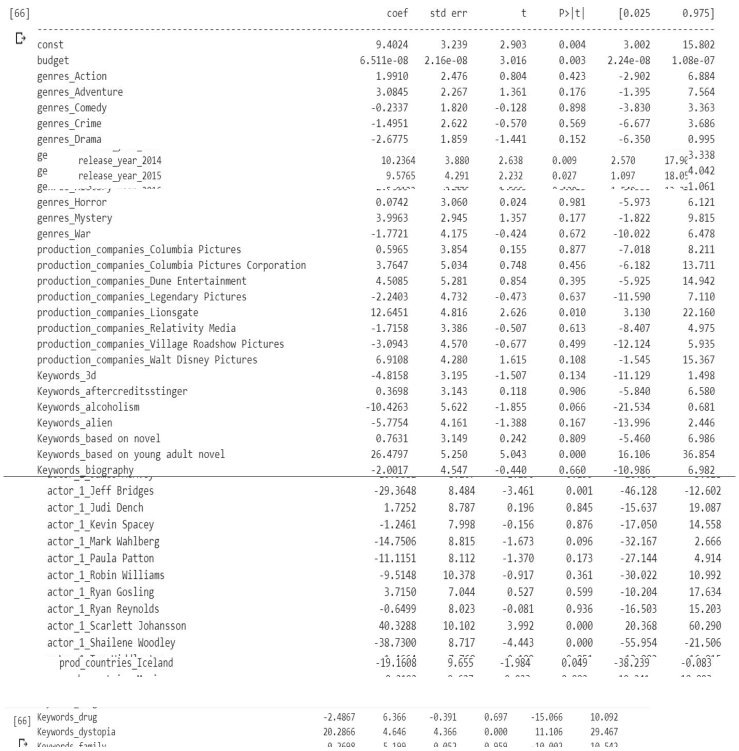
From this line plot, we can see that a longer movie does not mean that the movie will be more popular.

A good runtime is between 120minutes and 125minutes.

Hypothesis

1. genre_Drama, genre_Comedy, genre_Thriller are significant variables.
2. November is a significant variable.
3. 1977,2014,2017 are significant variables.
4. Jennifer Lawrence and Josh Hutcherson are significant variables.
5. For a successful movie, a budget between 100Mil and 160Mil.
6. A production company based in the United States is a significant variable.
7. Warner Bros production company is a significant variable
8. A good runtime is between 120miutes and 125minutes

Initial Result[vif<=5]

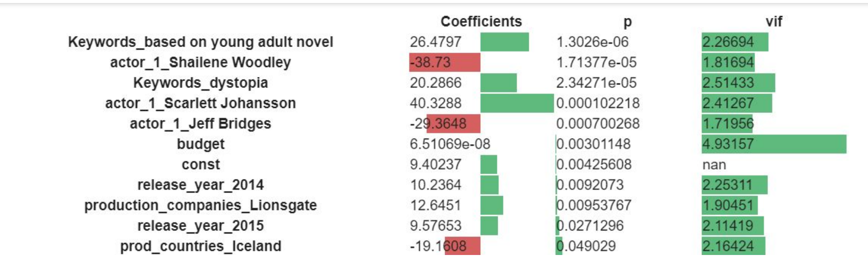


```
[66]                                                 coef    std err          t      P>|t|      [0.025      0.975]
      --------------------------------------------------------------------------------------------------------------
      const                                        9.4024      3.239      2.903      0.004       3.002      15.802
      budget                                    6.511e-08   2.16e-08      3.016      0.003    2.24e-08    1.08e-07
      genres_Action                                1.9910      2.476      0.804      0.423      -2.902       6.884
      genres_Adventure                             3.0845      2.267      1.361      0.176      -1.395       7.564
      genres_Comedy                               -0.2337      1.820     -0.128      0.898      -3.830       3.363
      genres_Crime                                -1.4951      2.622     -0.570      0.569      -6.677       3.686
      genres_Drama                                -2.6775      1.859     -1.441      0.152      -6.350       0.995
      ge    release_year_2014                     10.2364      3.880      2.638      0.009       2.570   17.9(3.338
      ge    release_year_2015                      9.5765      4.291      2.232      0.027       1.097   18.0!4.042
      ge...                                                                                                1.061
      genres_Horror                                0.0742      3.060      0.024      0.981      -5.973       6.121
      genres_Mystery                               3.9963      2.945      1.357      0.177      -1.822       9.815
      genres_War                                  -1.7721      4.175     -0.424      0.672     -10.022       6.478
      production_companies_Columbia Pictures       0.5965      3.854      0.155      0.877      -7.018       8.211
      production_companies_Columbia Pictures Corporation 3.7647 5.034    0.748      0.456      -6.182      13.711
      production_companies_Dune Entertainment      4.5085      5.281      0.854      0.395      -5.925      14.942
      production_companies_Legendary Pictures     -2.2403      4.732     -0.473      0.637     -11.590       7.110
      production_companies_Lionsgate              12.6451      4.816      2.626      0.010       3.130      22.160
      production_companies_Relativity Media       -1.7158      3.386     -0.507      0.613      -8.407       4.975
      production_companies_Village Roadshow Pictures -3.0943  4.570     -0.677      0.499     -12.124       5.935
      production_companies_Walt Disney Pictures    6.9108      4.280      1.615      0.108      -1.545      15.367
      Keywords_3d                                 -4.8158      3.195     -1.507      0.134     -11.129       1.498
      Keywords_aftercreditsstinger                 0.3698      3.143      0.118      0.906      -5.840       6.580
      Keywords_alcoholism                        -10.4263      5.622     -1.855      0.066     -21.534       0.681
      Keywords_alien                              -5.7754      4.161     -1.388      0.167     -13.996       2.446
      Keywords_based on novel                      0.7631      3.149      0.242      0.809      -5.460       6.986
      Keywords_based on young adult novel         26.4797      5.250      5.043      0.000      16.106      36.854
      Keywords_biography                          -2.0017      4.547     -0.440      0.660     -10.986       6.982
        actor_1_Jeff Bridges                     -29.3648      8.484     -3.461      0.001     -46.128     -12.602
        actor_1_Judi Dench                         1.7252      8.787      0.196      0.845     -15.637      19.087
        actor_1_Kevin Spacey                      -1.2461      7.998     -0.156      0.876     -17.050      14.558
        actor_1_Mark Wahlberg                    -14.7506      8.815     -1.673      0.096     -32.167       2.666
        actor_1_Paula Patton                     -11.1151      8.112     -1.370      0.173     -27.144       4.914
        actor_1_Robin Williams                    -9.5148     10.378     -0.917      0.361     -30.022      10.992
        actor_1_Ryan Gosling                       3.7150      7.044      0.527      0.599     -10.204      17.634
        actor_1_Ryan Reynolds                     -0.6499      8.023     -0.081      0.936     -16.503      15.203
        actor_1_Scarlett Johansson                40.3288     10.102      3.992      0.000      20.368      60.290
        actor_1_Shailene Woodley                 -38.7300      8.717     -4.443      0.000     -55.954     -21.506
          prod_countries_Iceland                 -19.1608      9.655     -1.984      0.049     -38.239      -0.083

[66]  Keywords_drug                 -2.4867      6.366     -0.391      0.697     -15.066      10.092
      Keywords_dystopia             20.2866      4.646      4.366      0.000      11.106      29.467
```

X-variables below 5%

- Budget
- production_companies_Lionsgate
- Keywords_based_on_young_adult_movie
- Keyword_dystopia
- actor_1_Jeff_Bridges
- actor_1_Scarlett_Johanson
- actor_1_Shailene_Woodley
- Release_year_2014
- Release_year_2015
- prod_countries_iceland



	Coefficients	p	vif
Keywords_based on young adult novel	26.4797	1.3026e-06	2.26694
actor_1_Shailene Woodley	-38.73	1.71377e-05	1.81694
Keywords_dystopia	20.2866	2.34271e-05	2.51433
actor_1_Scarlett Johansson	40.3288	0.000102218	2.41267
actor_1_Jeff Bridges	-29.3648	0.000700268	1.71956
budget	6.51069e-08	0.00301148	4.93157
const	9.40237	0.00425608	nan
release_year_2014	10.2364	0.0092073	2.25311
production_companies_Lionsgate	12.6451	0.00953767	1.90451
release_year_2015	9.57653	0.0271296	2.11419
prod_countries_Iceland	-19.1608	0.049029	2.16424

From the p-values and after removing the vif > 5, no genres are significant, the month&days aren't significant, the year 2014 is significant but not the years 1977&2017, the actor Scarlett Johansson is significant, budget is significant, Lionsgate is significant.

Plan for Final Presentation & Report

- Revise the hypotheses
- Check if there are any research papers with a similar topic of movies
- Check Kaggle for more data if possible
- Possibly focus on Franchises

	Task
Wk6	Data Exploration
Wk7	Data Cleaning
Wk8	Focus on Interim Report
Wk9	Focus on Final Presentation
Wk10	Finish up report