CMPT 353 D1 Project - On Wikidata, Movies, and Success

Jason Wang, Matthew Chan
Simon Fraser University

Project can be found on Github: https://github.com/period23dijason/wikidata

Abstract:

In this study, our group will be examining data on past movies and identifying some of the factors that may contribute a movie's success in terms of its profitability as well as its reception among the audience and critics. We will focus on attributes including a movie's genre, country of origin and cast members. We would also like to explore the trend of ratings over time to discover any possible occurrence of an inflation phenomenon for ratings. The insights in this paper is beneficial to our company, as we will be able to make decisions that will maximize our profitability.

Data:

This data was primarily gathered from Wikidata project which is dedicated to presenting information from Wikipedia using tags and metadata. Data was also grabbed from the Rotten Tomatoes database as well as the OMDB API based on the corresponding IDs from the Wikidata file where it contains attributes for a collection of movies. The refining process were achieved on Python scripts created by Professor Greg Baker. Additional cleaning was done before certain tests.

Techniques:

We have utilized various techniques taught in our Data Science course, starting with the import of the Python library Pandas to read the given JSON files and convert them into dataframe tables. Our first step was to filter out movie entries that do not have sufficient data such as those without information on profit. From there on, we aimed to translate the genres recorded in IDs from Wikidata into actual string values using the conversion chart. For a given movie provided by the Wikidata table, we have also combined it with the corresponding entry on the Rotten Tomatoes and

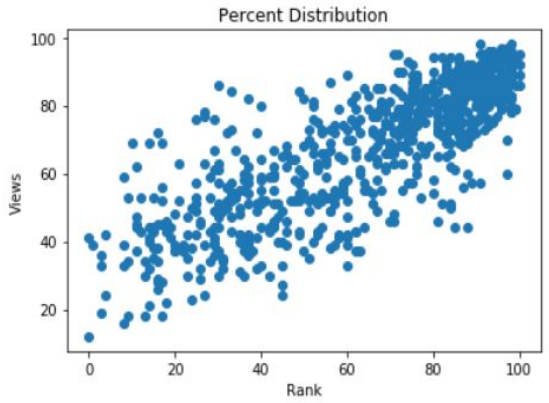
IDMB table to form one big table, leaving behind only relevant movies with information provided by all three sources.

With this combined table in hand, we were able to perform a number of tests. To compare the positivity ratings of audience and critics, we performed a Mann-Whitney U test. We used linear regression to see if movie ratings have inflated over time, and if a movie's rating correlates with its popularity. We did this not only to figure out if it correlated, but also the degree of the correlation. To find whether certain criteria had an effect on profitability, we used Chi-square tests. We made a machine learning model to predict profitability based on ratings. Finally, we attempted to use natural language processing to see if we could predict profitability or ratings based on a movie's summary.

Result and Visualizations:

Is there a difference between the positivity of critics and the audience?

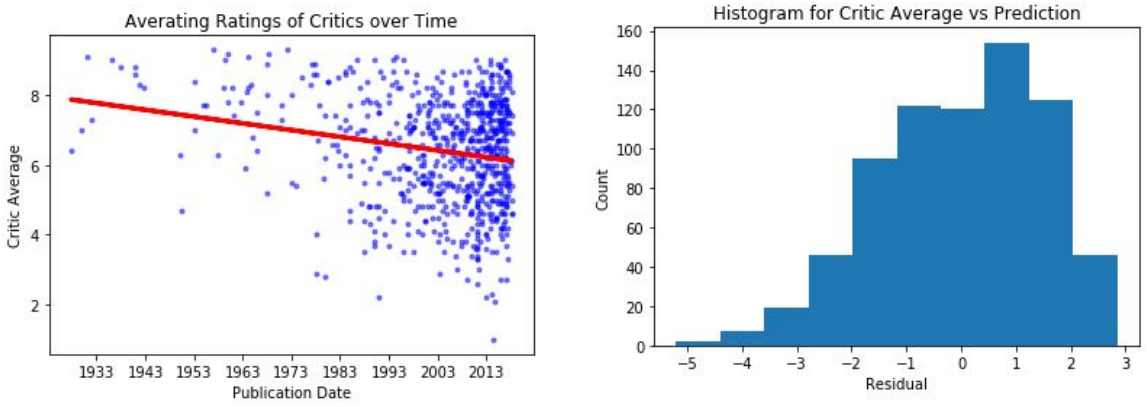
Firstly, we created a scatter-plot distribution of the critic and audience positivity percent (see below). At a glance we can see that plots with higher critic percent have a higher audience percent, as if sharing a positively linear relationship. Further experimenting with the Mann-Whitney U test validates this claim, returning us a p-value of approximately 0.4397286644629225, which is significantly less than our threshold of 0.05. This means that the positivity of the audience and critics are probably the same. For us, this means that we should equally favour the criticisms of both the audience and the critics. Note that we did not do this test for audience and critic rating averages. This is because they are done on different scales. A 10/10 rating is not necessarily equal to a 5/5 rating.



Have average ratings changed over time?

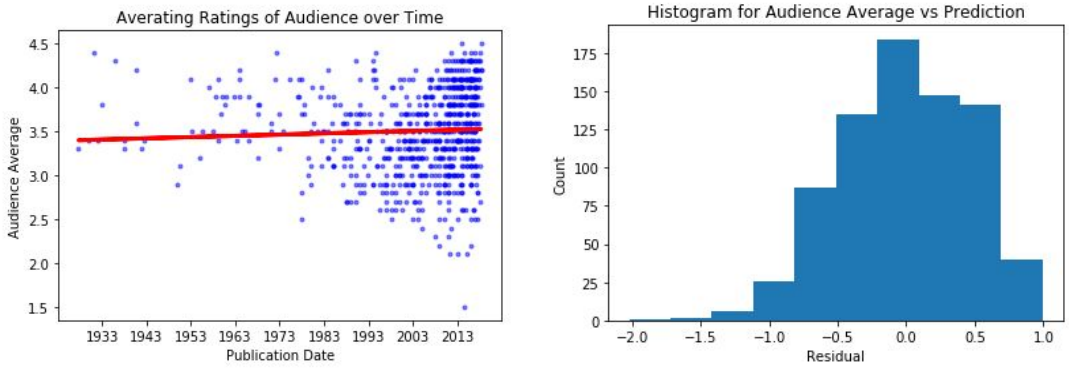
Next, we proceeded to calculate the linear least-squares regressions for the critic average (See diagram a) across an x-axis of publication date. When we fitted the values, the p-value and r-value returned are respectively 6.156831173958292e-08 and -0.19792833987738834. The p-value is significantly lower 0.05, so we conclude that the critic ratings are indeed decreasing over time, and this is further supported by the low negative r-value, which is a correlation coefficient that tells us here that there is a slight downhill linear relationship. Note that the correlation coefficient is still close to 0, indicating that the line is not very predictive, so we should not worry about our ratings too much as we release new movies in the future.

In red you can see a best-fit line created via ax+b = y where a is the slope and b is the y-intercept.



a)

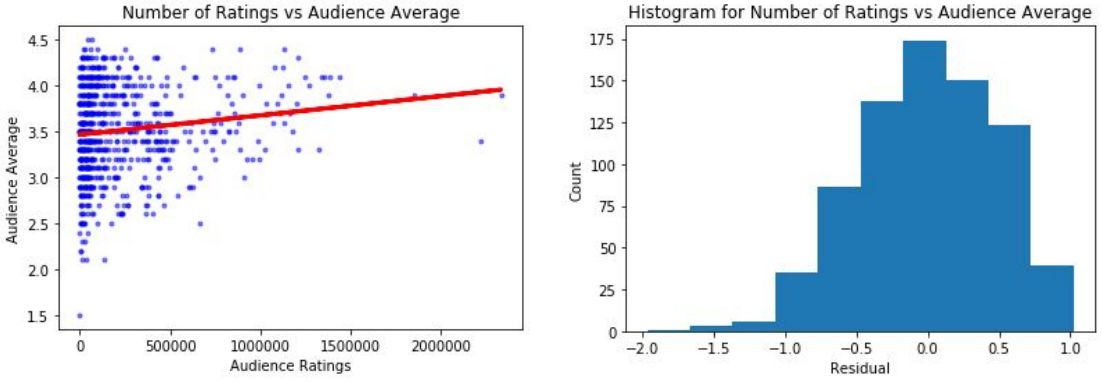
On the linear regression for the audience average across the publication date (See diagram b), we get a p-value of 0.20019655801512026 and an r-value of 0.046244255512890665. Therefore it is probable that audience ratings are not changing over time.



b)

Do average audience ratings change based on its popularity?

This time we apply a linear regression on the audience average with respect to the number of audience ratings. We removed a few outliers that had over 10,000,000 ratings, as this seemed unrealistic for both the data and for our company. This returned a p-value of 0.0003085653249741354 and an r-value 0.1309596810010819, meaning that more popular movies tend to be higher rated. Again, because of the low r-value, popularity is not a good predictor of average rating. Therefore, getting higher ratings should not be the goal of increasing the popularity of our movies.



c)

On the right side of a), b), and c), we have also included histograms indicating the difference between the actual average and the prediction from the best-fit line. As seen from all three histograms, a shape resembling that of a normal distribution is a common theme. With more than 40 data points, the Central Limit Theorem is safely established.

Do certain criteria effect movie profitability?

The three criteria we looked at were genre, country of origin, and cast members. Since this is a categorical test, we used Chi-squared tests to see if these criteria have an effect on profitability. For criteria that did have an effect, we took the probability of profitability to give us the best criteria. For each of these tests, we cleaned up the data by removing criteria of which we did not have enough data for, and by removing movies that did not have the necessary criteria data.

Testing genre gave us a p-value of 0.01956332775267009, so we can conclude that genre does have an effect on profitability. We then found that science fiction films are the most profitable, with a

91.5% chance of profitability. The next 9 genres are only slightly worse, with the 10th best genre, comedy films having a profitability rate of 85.7%.

	genre_label	wikidata_id	profit	total	loss	
0	science fiction film	Q471839	130	142	12	0.
1	romantic comedy	Q860626	53	59	6	0.
2	romance film	Q1054574	44	49	5	0.
3	horror film	Q200092	70	78	8	0.
4	war film	Q369747	40	45	5	0.
5	thriller film	Q2484376	108	122	14	0.
6	film based on literature	Q52162262	119	135	16	0.
7	adventure film	Q319221	104	119	15	0.
8	fantasy film	Q157394	109	125	16	0.

Does country of origin have an effect on profitability?

For country of origin, only for a few countries were there enough movies made in order to perform a proper Chi-Squared test. Regardless, testing country of origin gave us a p-value of 0.008871929501706175. Therefore, country of origin has an effect on profitability. Country Q16 has the highest chance of profitability at 87.5%, but the low number of movies made in that country makes it risky for us. Country Q30, which is presumably the United States, has only a slightly lower chance of 86.0%, but has a much greater sample size.

country_id	percent	total	profit	loss
Q16	0.875000	8	7.0	1.0
Q30	0.860465	602	518.0	84.0
Q142	0.838710	31	26.0	5.0
Q145	0.838710	62	52.0	10.0
Q159	0.677419	31	21.0	10.0
Q408	0.636364	11	7.0	4.0
Q183	0.625000	16	10.0	6.0

Does cast member have an effect on profitability?

For cast members, we got a p-value of 0.21051014915423152. Therefore cast members probably do not affect profitability. We hypothesize that even though cast members could increase revenue, having cast members also increases expenses.

For our company, in order to maximize the likelihood of profitability, we recommend that we base movies in country Q30, the United States, and to make science fiction films. It should not matter which cast member we use.

How well can we predict profitability based on ratings?

Using the movies critic and audience average ratings and their positivity percentages, we were able to train a model that predicts whether or not a movie is profitable. Using SVC with the RBF kernel and a standard scaler, we were able to make a decently accurate model. The accuracy score of our model is generally above 0.8, which is pretty accurate. For our company, to maximize the chance of profitability, our business model should be focused on quality over quantity, on making movies that our viewers will like, and not merely adhere to certain criteria.

Challenges:

During our investigation, we were met with several limitations beginning with the JSON files provided to us. Although there was some explanation given about the way the data was refined, we were missing several implementation details on the Python scraper scripts such as for the Rotten Tomato website. In addition, we were given to the conversion table with the genre ID and its corresponding genre word, albeit there was no table for converting the actor IDs from Wikidata to actual names which made it difficult to find useful relationships regarding which actors would affect profitability and ratings. The same applies to country as mentioned above earlier.

Since each movie tuple in our combined table may contain a genre section with multiple genres, such as The Godfather being both Crime and Drama, this made aggregate the genre column much more difficult to manipulate as it is no longer a 1xN matrix and forced us to write loop functions

to individually traverse through each subarray. Dealing with such a multifaceted dataset with multiple sources, we had to try our best to merge the tables to form one super table to be able to better locate each movie with its relevant attributes and gain a more holistic perspective. As a result, some incomplete movie tuples had to be filtered out in the process for the sake of consistency across all tests.

We have tried to create a machine learning model that could use genres to predict profitability, but we mistakenly used the same methodology that we used for using ratings to predict profitability. When we found out that X had to be a floating point number, and not a string for the genre, the problem revealed itself to be out of our area of expertise, and we did not have enough time to learn the proper methods.

Lastly, we were introduced to the notation of using natural language processing and attempted to apply the Term Frequency times Inverse Document Frequency (tf–idf) algorithm onto our column of omdb plot summaries as to find occurrences of keywords unique enough to be the defining characteristics of a movie plot summary. As an experiment, the audience average was used to be fitted into model as a y-Matrix via a naïve Bayes classifier. Unfortunately this process is still incomplete as of right now due to time constraints and lack of understanding in this field. Currently, we are only able to extract a list of keywords via tf–idf and display its corresponding score value.

Project Experience Summary

Final Project Jul-Aug 2018

Matthew Chan

- Discovered trends of movie ratings using linear regression.
- Predicted profitability based on movie ratings using a machine learning model.
- Determined best movie criteria for profitability using Chi-squared tests.
- Gave business recommendations to maximize the chance of profitability based on statistical analysis and machine learning on movie data

Project Experience Summary

Final Project Jul-Aug 2018

Jason Wang

Computational Data Science, SFU

- Collaborated on a Python program and extracted movie data from JSON files via Pandas and performed statistical analysis and predictive modelling with visualizations
- Created a report to document procedure and findings on factors attributed success in movies which assisted stakeholders with a new understanding of a the movie industry