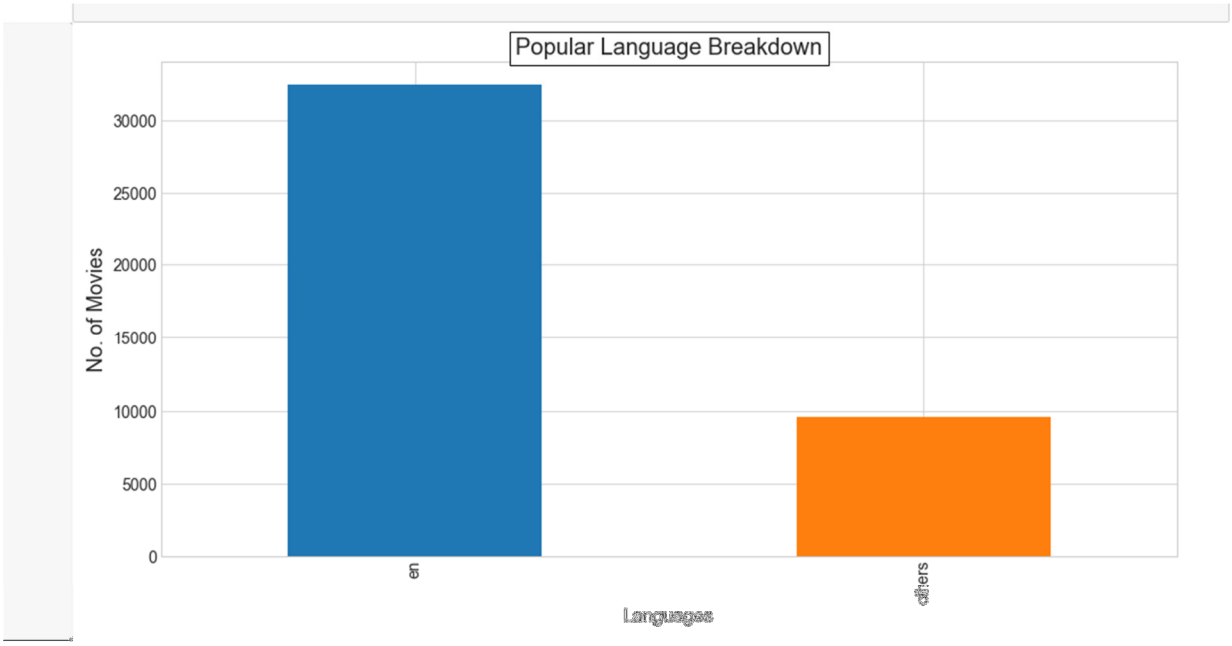
My class assignment was done on the Movielens data set from kaggle. This data set is a mix of data from TMDB and Grouplens. The dataset contained metadata of the movies, and also the TMDB popularity, voting average, and count of votes. Additionally, from Grouplens, the rating data of 26, 000, 000 records was available. However, due to computing powers on my end, I used the smaller sample of 100, 000 records. For exploratory purposes, many data fields represented the movie, some of the columns were: Genres, movieId (that linked back to the movieId of a keyword dataset for the movies and the rating dataset from Grouplens), original language, original title, Production Company, release date, spoken languages, production languages and the runtime. The numerical attributes of revenue, budget, average TMDB vote, TMDB vote count, TMDB popularity, revenue, and budget was also given.

To further explain some of the parameters, the TMDB voting system is on a scale of 10, the higher the vote, the better the movie. So the average vote is sum of all the votes divided by the number of votes (provided in another column). The popularity represents how many people saved this movie to watch later, put it in their favorites, or page views etc. It can be from 0 to infinity. The Grouplens rating on the other hand, is on a scale of 0 to 5. The higher the rate, the better liked the movie.

This rich dataset was quite interesting to explore. I do wonder, however, how the data would differ if it represented all movies from across the Globe equally. While I do believe it was a fair representation of US movies, I do not feel it was a fair representation of movies around the world. For example, a quick google search shows that an average of 250-350 Bollywood films are released per year, and an average of 600 movies are released in the US per year. Now, English movies are not only produced in the US, so it would make sense that there a lot more English movies than any other language, however, I do feel that TMDB would not have the data as complete for other languages.

Here is the dataset's breakdown by language:



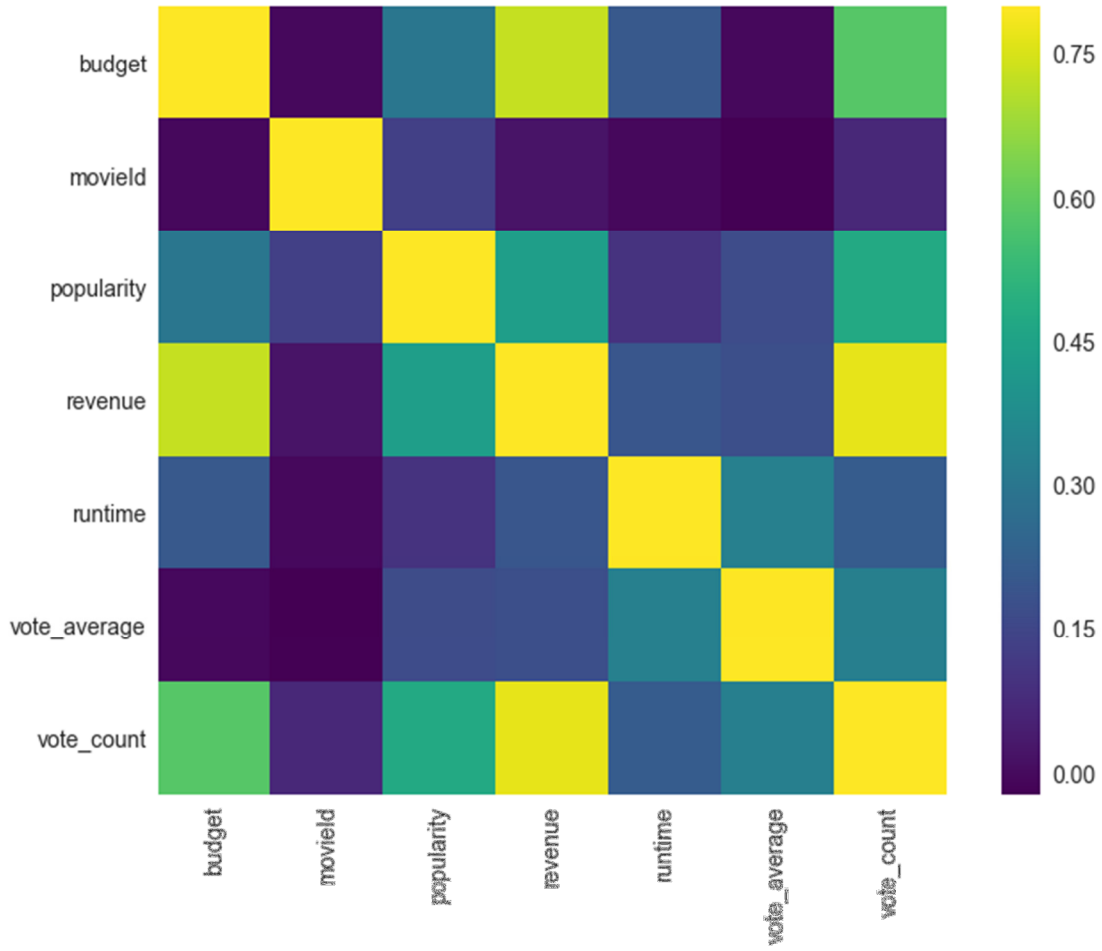
The blue represents English, and the orange represents, French, Italian, Japanese, German, Spanish, Russian, Hindi, Korean, and Zhōngwén (A Chinese language). Since the dataset had over 99 languages, I took the analysis of only the top 10. The numbers below show the spread of the dataset of the top 10 languages.

en	32464
fr	2467
It	1482
ja	1400
de	1048
es	988
ru	811
hi	526
ko	472
zh	416

Another interesting analysis was the most common Genres for the movies. And it the top 100 movies had the following Genre count:

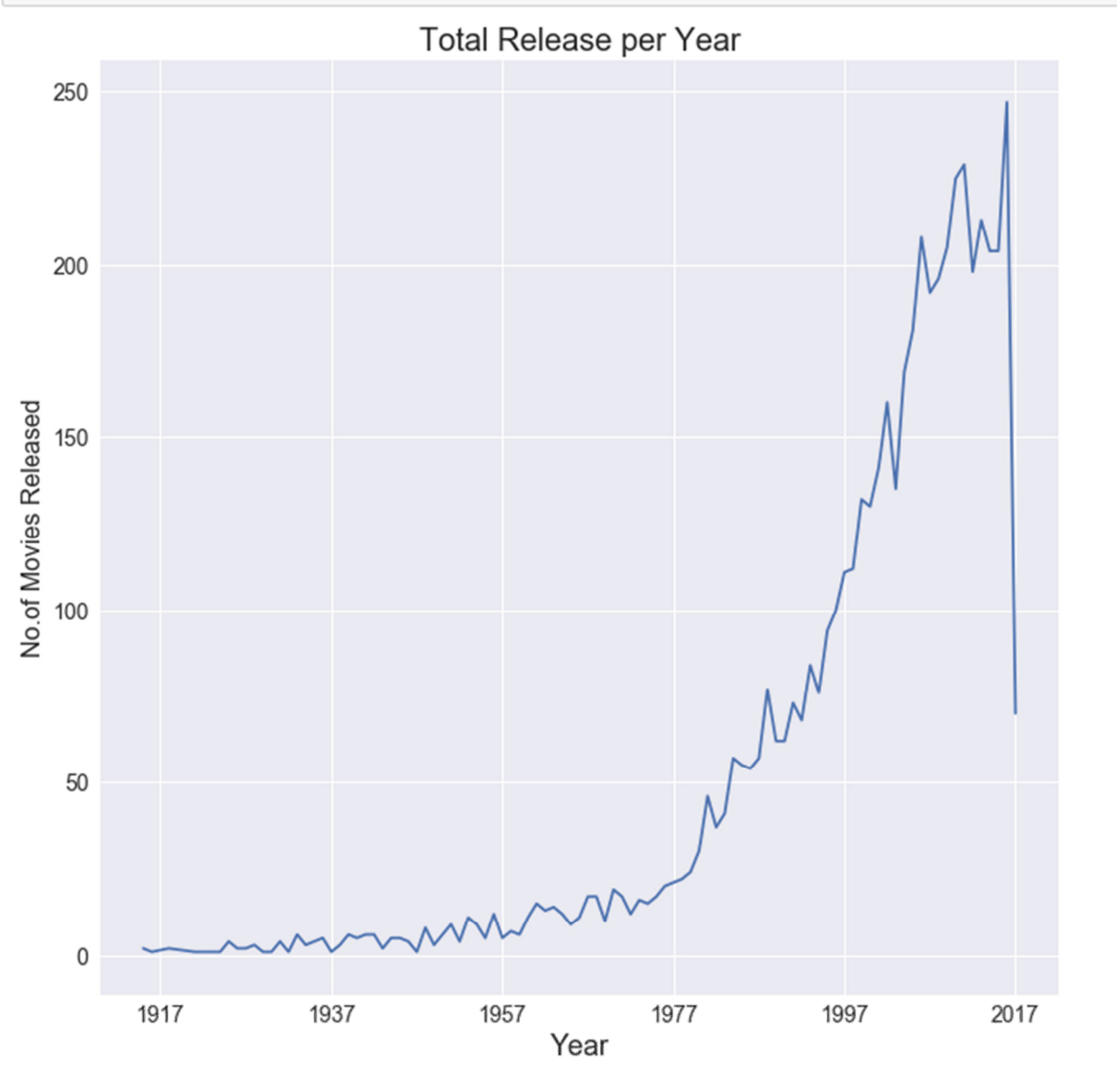
adventure	80
action	55
fantasy	44
family	35
science	31
fiction	31
animation	20
comedy	18
thriller	15
drama	12
romance	8
crime	6
mystery	3

A heatmap of the movies, also showed that there is a positive correlation between budget, revenue, popularity and vote counts. It makes sense that more people that go see the movie, the greater the popularity and number of votes. Additionally, the more money that goes into making a movie, the better it may be and the better it's revenue. Below is a heatmap of my analysis:



	budget	movieId	popularity	revenue	runtime	vote_average	vote_count
budget	1.00	-0.00	0.30	0.73	0.20	-0.00	0.58
movieId	-0.00	1.00	0.13	0.02	-0.01	-0.02	0.07
popularity	0.30	0.13	1.00	0.44	0.10	0.17	0.48
revenue	0.73	0.02	0.44	1.00	0.20	0.18	0.77
runtime	0.20	-0.01	0.10	0.20	1.00	0.33	0.21
vote_average	-0.00	-0.02	0.17	0.18	0.33	1.00	0.33
vote_count	0.58	0.07	0.48	0.77	0.21	0.33	1.00

Lastly, too understand how the number of movies released changed throughout the years, we also did an analysis on the movie releases by year. As expected, the number of movies increases with each passing year, however, the data of 2017 is not complete, and therefore there is a drop seen on the plot.



I would love to continue to analyse this data better, by creating different segments for budget/revenue as explained in the ipynb file. Additionally, I would like to see the trend in Grouplens raters. Do certain raters prefer certain Genres?

References:

https://www.themoviedb.org/documentation/api

https://grouplens.org/datasets/movielens/latest/

https://www.kaggle.com/rounakbanik/the-movies-dataset/home

https://www.themoviedb.org/talk/5141d424760ee34da71431b0