Movie Recommendation System

Blake Tolman

The Data

- GroupLens Research has collected and made available rating data sets from the MovieLens web site in both small and large datasets
- Small: 100,000 ratings and 3,600 tag applications applied to 9,000 movies by 600 users. Last updated 9/2018

	userId	movieId	rating	timestamp
0	1	1	4.0	964982703
1	1	3	4.0	964981247
2	1	6	4.0	964982224
3	1	47	5.0	964983815
4	1	50	5.0	964982931

	movieId	title	genres
0	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy
1	2	Jumanji (1995)	Adventure|Children|Fantasy
2	3	Grumpier Old Men (1995)	Comedy|Romance
3	4	Waiting to Exhale (1995)	Comedy|Drama|Romance
4	5	Father of the Bride Part II (1995)	Comedy

The Recommendation System

- The recommendation system uses collaborative filtering to match similar users together
- 11 different models were tested to find the best fit
  - BaselineOnly was the chosen algorithm due to its low error and run time.
- With any recommendation system the problem of the cold start is an issue: How to handle new users with no data.

Algorithm	test_rmse	fit_time	test_time
SVDpp	0.868845	882.926506	32.773629
BaselineOnly	0.876045	0.222493	0.456649
SVD	0.879474	9.959715	0.614917
KNNBaseline	0.881853	0.539366	7.746760
KNNWithMeans	0.905121	0.386333	6.592975
KNNWithZScore	0.905275	0.452219	6.905525
SlopeOne	0.912581	6.849113	21.746805
NMF	0.934993	11.795354	0.525171
CoClustering	0.955518	5.107392	0.581912
KNNBasic	0.960078	0.321203	5.173181
NormalPredictor	1.424901	0.260035	0.533820

Cold Start 1

- The first solution is to have new users answer a questionnaire about movies that they have seen
- The user ranks the presented movies to allow for the system to have a basis to build off initially
- Ex: *“How do you rate this movie on a scales of 1-5”*

```
      movieId                        title                 genres
4615     6874  Kill Bill: Vol. 1 (2003)  Action|Crime|Thriller
How do you rate this movie on a scale of 1-5, press n if you have not seen :
5
      movieId             title                                genres
4809     7164  Peter Pan (2003)  Action|Adventure|Children|Fantasy
How do you rate this movie on a scale of 1-5, press n if you have not seen :
4
      movieId                             title                          genres
5434    25946  Three Musketeers, The (1948)  Action|Adventure|Drama|Romance
How do you rate this movie on a scale of 1-5, press n if you have not seen :
3
```

```
Recommendation #  1 :  277    Shawshank Redemption, The (1994)
Name: title, dtype: object

Recommendation #  2 :  602    Dr. Strangelove or: How I Learned to Stop Worr...
Name: title, dtype: object

Recommendation #  3 :  906    Lawrence of Arabia (1962)
Name: title, dtype: object

Recommendation #  4 :  841    Streetcar Named Desire, A (1951)
Name: title, dtype: object

Recommendation #  5 :  2226    Fight Club (1999)
Name: title, dtype: object
```

Cold Start 2

- The second solution is to display a general list of movies that have a high number of ratings along with a high average movie rating
  - This will act as a general list of most liked movies
- A variation of this option is to adjust the general recommendation with any user data present.
  - Ex: Gender and or Age

Top Movies

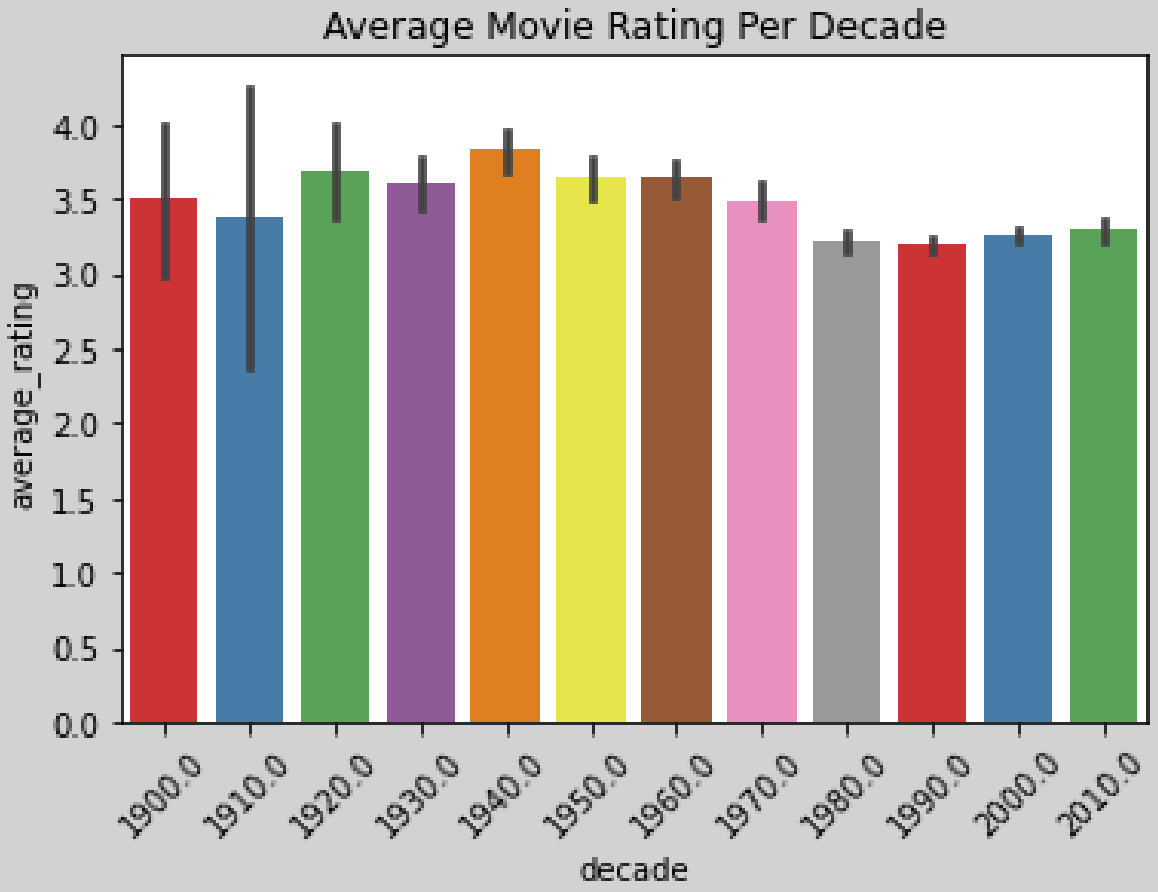
title	average_rating	number_of_ratings	movieId
Shawshank Redemption, The (1994)	4.429412	232.0	318.0
Forrest Gump (1994)	4.092784	232.0	356.0
Pulp Fiction (1994)	4.026882	214.0	296.0
Matrix, The (1999)	4.115942	209.0	2571.0
Silence of the Lambs, The (1991)	4.273333	204.0	593.0

Top Movies with User Data

	title	average_rating	number_of_ratings	movieId_x	release_year	decade	genres
26	13th Warrior, The (1999)	3.214286	19.0	2826.0	1999.0	1990.0	Action|Adventure|Fantasy
36	2 Fast 2 Furious (Fast and the Furious 2, The)...	3.000000	18.0	6383.0	2003.0	2000.0	Action|Crime|Thriller
37	2 Guns (2013)	4.000000	2.0	103883.0	2013.0	2010.0	Action|Comedy|Crime
21	13 Assassins (Jûsan-nin no shikaku) (2010)	4.333333	1.0	86142.0	2010.0	2010.0	Action
0	'Hellboy': The Seeds of Creation (2004)	4.000000	0.0	97757.0	2004.0	2000.0	Action|Adventure|Comedy|Documentary|Fantasy

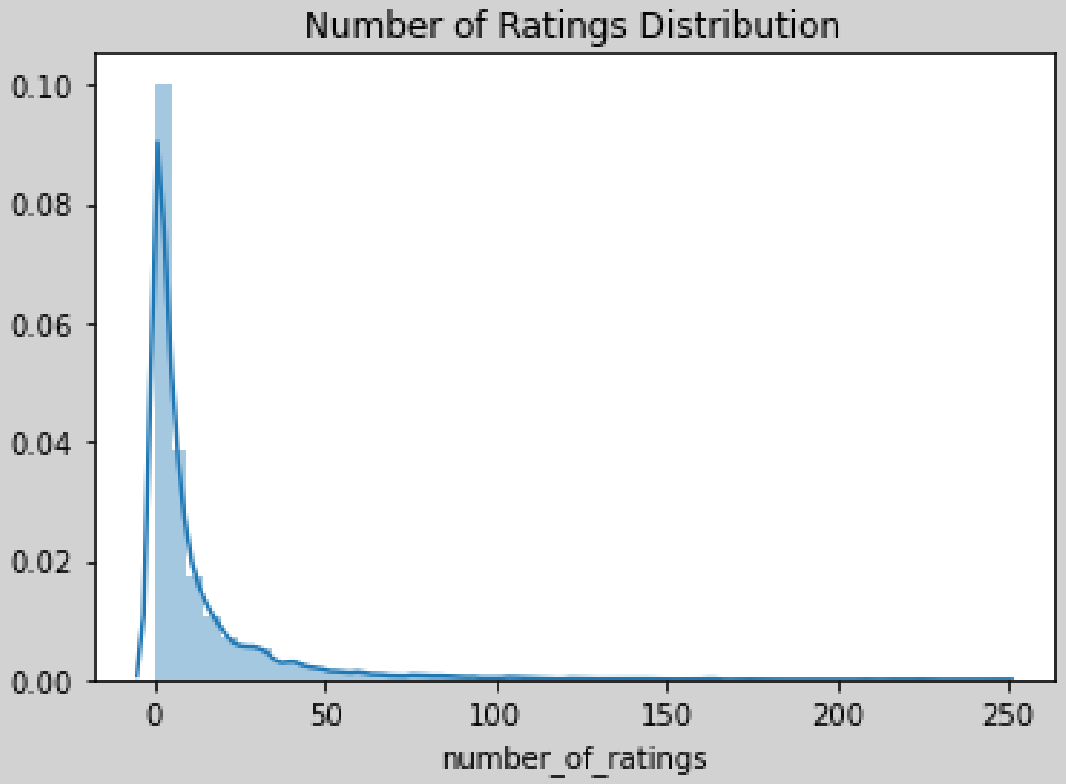
Best Decade for Movies

- The best decade for movies was from the 1920's to the 1960's
- Of the top 5 best movies, all 5 of them were made in the 90's
- Relative to the other decades the 90's was a particularly low time for movies



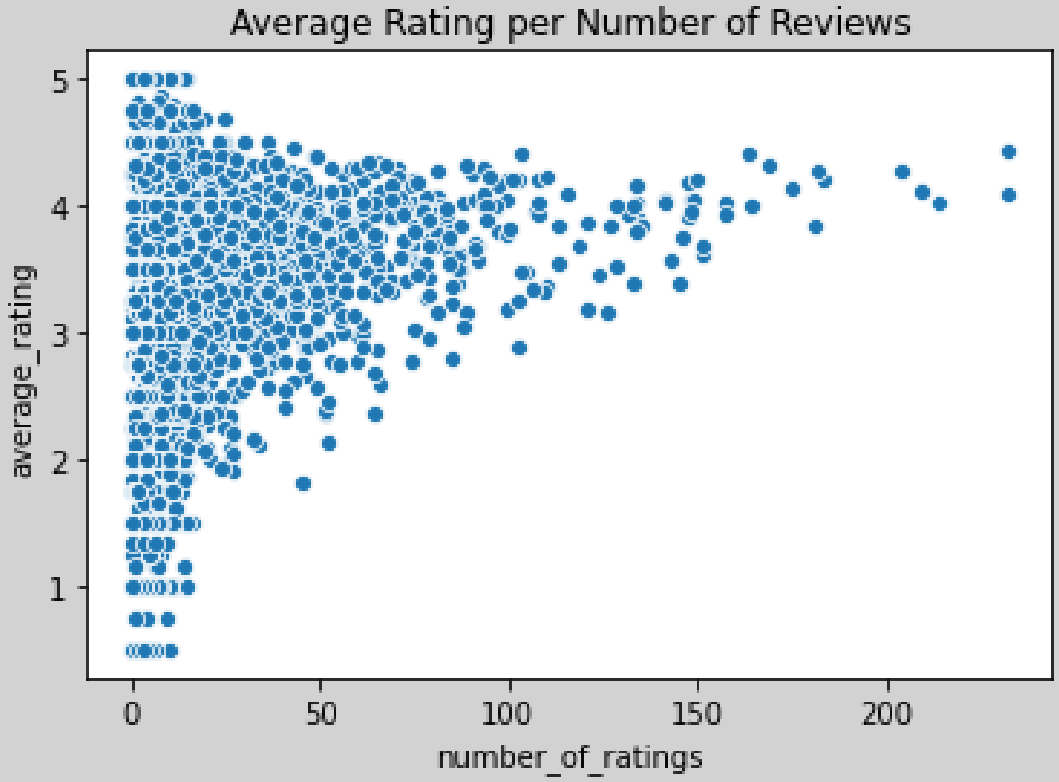
Distribution of Reviews

- 16.9% of movies have 1 reviews for them
- 70.3% of movies having less than 10 reviews
- The recommendation system runs into the issue of not enough users consistently rating movies



Ratings and Reviews

- Most movies are rated between 3 & 4 on average
- Movies that have achieved an average rating of 5 typically only have 1-5 reviews
  - Most movies cannot maintain a 5-star rating
- There is a cycle were movies with low ratings don't get any new viewers, so their ratings stay low, while highly rated moves get more reviews because of their rating



Recommendations and Future Work

- Recommendations

1. Try and increase the number of reviews users give
2. Don’t display the rating of the movie until it has at least 10 reviews to try and remove user bias
3. Collect as much data about the user as possible to personalize recommendations

- Future Work

1. Use the large data set for an improved model
2. Update the files with movies and ratings of the current year
3. Have an outside source like Rotten Tomatoes ratings available as a comparison against the individual opinion

Thank You