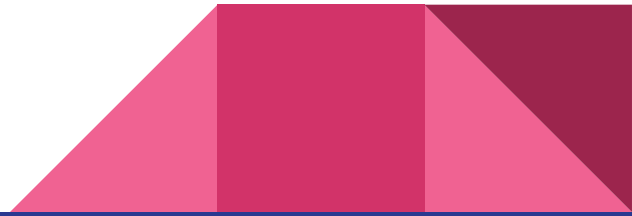
MOVIE RECOMMENDATIONS

CAPSTONE PROJECT 2

Background

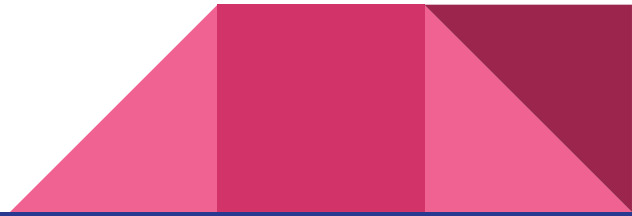
- Intrigued by Netflix Prize competition in 2009
- Avid movie goers
- First project involving recommendation system



Problems

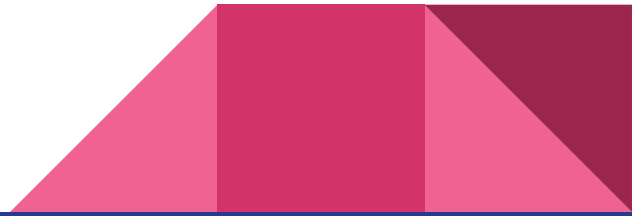
Many of the movie recommendation systems employed by various streaming platforms have been in **mature stage**

I believe that there is still room for improvement on recommending a person a movie that is out of his usual preference but still relatable



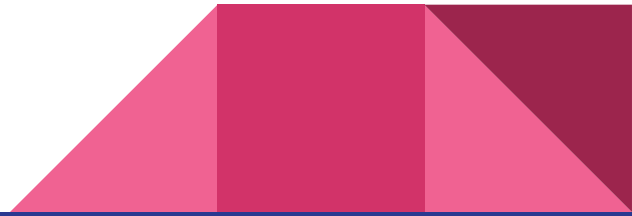
Outline

1. Data Source
2. Importing and Cleaning Data
3. Exploratory Data Analysis (EDA)
4. Inferential Statistics



Data Source

- MovieLens 10M Dataset (https://grouplens.org/datasets/movielens/)
  - movies.dat (510 KB)
  - ratings.dat (258,893 KB)
  - tags.dat (3,501 KB)
- 10,676 movies
- 10 million movie ratings



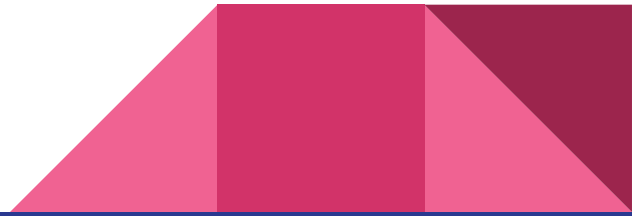
Importing and Data Cleaning

- Check for missing values (N/A)
  - replace with blank
- Check types of data
  - float64 for ratings

Exploratory Data Analysis (EDA)

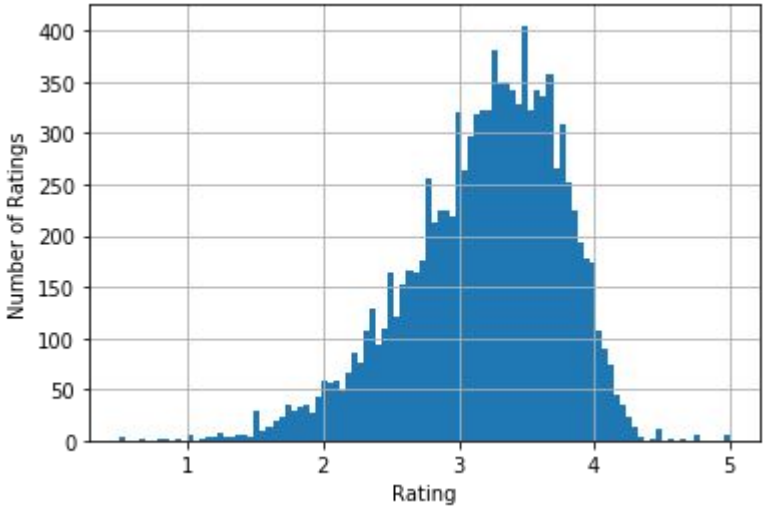
These sections are divided into 2 parts

1. By Title of the Movies
2. By Year of the Movies are Released



EDA (By Title)

- Distribution of Ratings by Movie Title



EDA (By Title)

- The ratings range from 0.5 to 5 with mean of 3.19 and standard deviation of 0.567
- On average, there are 936 reviews per movie with standard deviation of 2,487.43
- The movie with most rating is Pulp Fiction (2004) with 34,864 reviews and rating of 4.15

title	rating	no_of_ratings
Pulp Fiction (1994)	4.157426	34864
Forrest Gump (1994)	4.013582	34457
Silence of the Lambs, The (1991)	4.204200	33668
Jurassic Park (1993)	3.661564	32631
Shawshank Redemption, The (1994)	4.457238	31126

EDA (By Year)

- The movies released year span from 1915 to 2007 (92 years)
- The ratings range from 0.5 to 5 with mean of 3.72 and standard deviation of 0.211
- On average, there are 106,383.55 reviews per year with standard deviation of 172,558.88
- The year with most rating is 1995 with 874,436 reviews and rating of 3.44

year	rating	no_of_ratings
1995	3.442817	874436
1994	3.472097	746042
1996	3.360754	659425
1999	3.453832	543990
1993	3.496386	534899

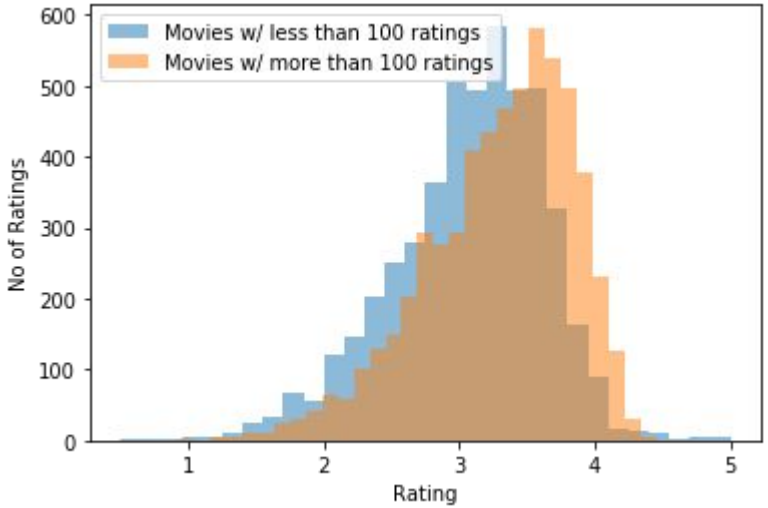
EDA (By Year)

- The year with the highest rating is 1946 with 18,719 reviews and rating of 4.05

year	rating	no_of_ratings
1946	4.054036	18719
1934	4.051894	6600
1942	4.043820	22353
1931	4.025816	8483
1941	4.013690	26589

Inferential Statistics

- Break the data into 2
  - 100 or less reviews per movie
  - More than 100 reviews per movie



THANK YOU