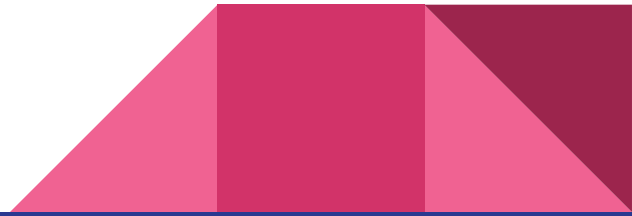
MOVIE RECOMMENDATIONS

CAPSTONE PROJECT 2

Background

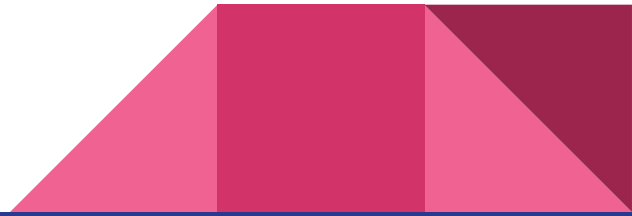
- Intrigued by Netflix Prize competition in 2009
- Avid movie goers
- First project involving recommendation system



Problems

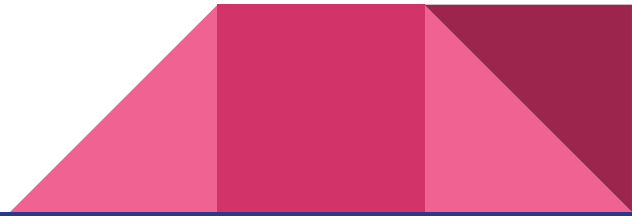
Many of the movie recommendation systems employed by various streaming platforms have been in **mature stage**

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



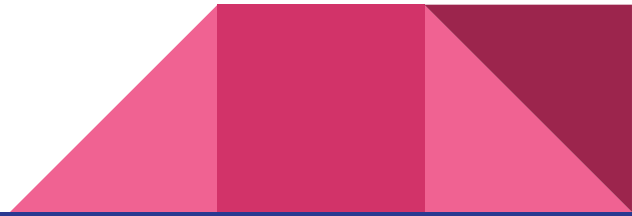
Outline

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



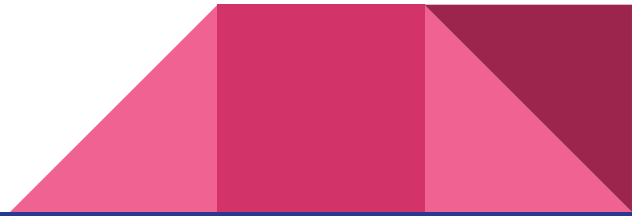
Importing and Data Cleaning

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

Exploratory Data Analysis (EDA)

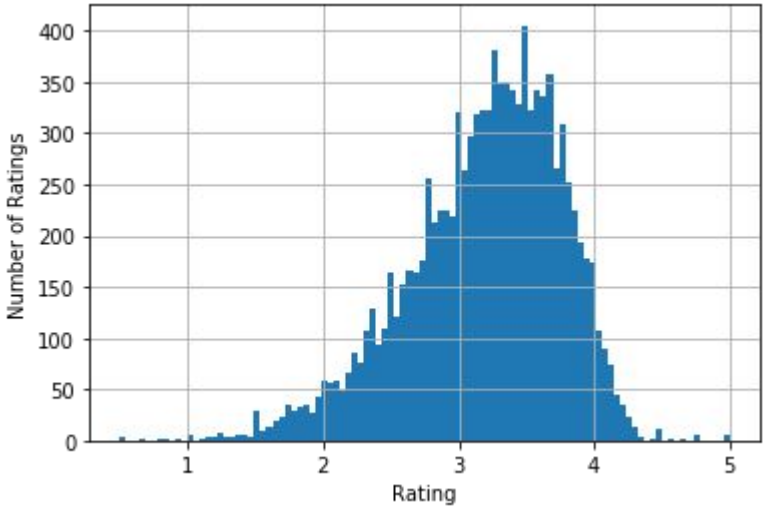
These sections are divided into 2 parts

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

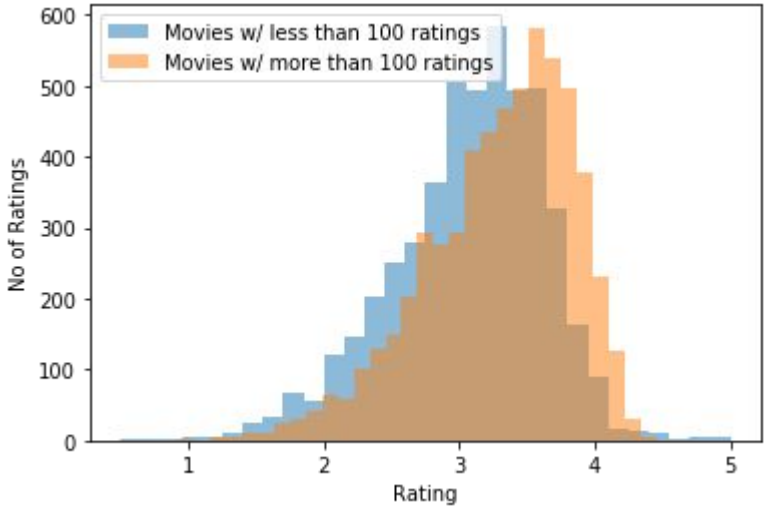
EDA (By Year)

- 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

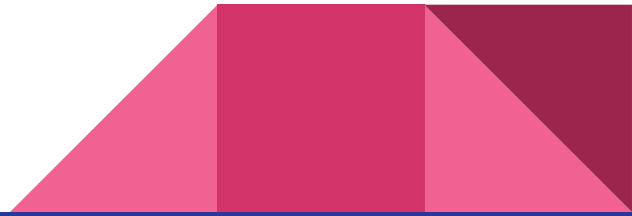
- Break the data into 2
  - 100 or less reviews per movie
  - More than 100 reviews per movie



Recommendation System

In this project, there will be 3 methods for the recommendation system:

1. Simple Correlation
2. Memory-based Method
3. Model-based Method



Simple Correlation (more than 100 reviews)

Example: Star Wars: Episode IV - A New Hope (a.k.a. Star Wars) (1977)

	title	Correlation	no_of_ratings	movieId	genres
0	Star Wars: Episode IV - A New Hope (a.k.a. Sta...	100.000000	28566	260	Action Adventure Sci-Fi
1	Star Wars: Episode V - The Empire Strikes Back...	72.164990	23091	1196	Action Adventure Sci-Fi
2	Star Wars: Episode VI - Return of the Jedi (1983)	66.312501	25098	1210	Action Adventure Sci-Fi
3	Beyond Silence (Jenseits der Stille) (1996)	48.432453	106	1893	Drama
4	Raiders of the Lost Ark (Indiana Jones and the...	46.240733	21803	1198	Action Adventure
5	Futurama: Bender's Game (2008)	46.023348	141	62956	Animation Comedy Sci-Fi
6	My Name Is Nobody (Il Mio nome Ã¨ Nessuno) (1973)	45.841934	119	26294	Comedy Western
7	Apartment, The (L'Appartement) (1996)	43.308614	132	6789	Drama Mystery Romance
8	Hairdresser's Husband, The (Mari de la coiffeu...	43.285454	138	8270	Comedy Drama Romance
9	Blackboard Jungle (1955)	42.578024	163	8451	Drama
10	Strange Love of Martha Ivers, The (1946)	42.047967	125	3965	Drama Film-Noir

Simple Correlation (more than 500 reviews)

	title	Correlation	no_of_ratings	movieId	genres
0	Star Wars: Episode IV - A New Hope (a.k.a. Sta...	100.000000	28566	260	Action Adventure Sci-Fi
1	Star Wars: Episode V - The Empire Strikes Back...	72.164990	23091	1196	Action Adventure Sci-Fi
2	Star Wars: Episode VI - Return of the Jedi (1983)	66.312501	25098	1210	Action Adventure Sci-Fi
3	Raiders of the Lost Ark (Indiana Jones and the...	46.240733	21803	1198	Action Adventure
4	Star Wars: Episode III - Revenge of the Sith (...	41.444946	5193	33493	Action Adventure Fantasy Sci-Fi
5	Star Wars: Episode I - The Phantom Menace (1999)	40.280006	15744	2628	Action Adventure Sci-Fi
6	Star Wars: Episode II - Attack of the Clones (...	39.495531	7934	5378	Action Adventure Sci-Fi
7	Lord of the Rings: The Two Towers, The (2002)	35.655860	14389	5952	Action Adventure Fantasy
8	Lord of the Rings: The Fellowship of the Ring,...	34.425024	15938	4993	Action Adventure Fantasy
9	Lord of the Rings: The Return of the King, The...	34.237021	12366	7153	Action Adventure Fantasy
10	Indiana Jones and the Last Crusade (1989)	33.121802	16145	1291	Action Adventure

Simple Correlation (more than 1,000 reviews)

	title	Correlation	no_of_ratings	movieId	genres
0	Star Wars: Episode IV - A New Hope (a.k.a. Sta...	100.000000	28566	260	Action Adventure Sci-Fi
1	Star Wars: Episode V - The Empire Strikes Back...	72.164990	23091	1196	Action Adventure Sci-Fi
2	Star Wars: Episode VI - Return of the Jedi (1983)	66.312501	25098	1210	Action Adventure Sci-Fi
3	Raiders of the Lost Ark (Indiana Jones and the...	46.240733	21803	1198	Action Adventure
4	Star Wars: Episode III - Revenge of the Sith (...	41.444946	5193	33493	Action Adventure Fantasy Sci-Fi
5	Star Wars: Episode I - The Phantom Menace (1999)	40.280006	15744	2628	Action Adventure Sci-Fi
6	Star Wars: Episode II - Attack of the Clones (...	39.495531	7934	5378	Action Adventure Sci-Fi
7	Lord of the Rings: The Two Towers, The (2002)	35.655860	14389	5952	Action Adventure Fantasy
8	Lord of the Rings: The Fellowship of the Ring,...	34.425024	15938	4993	Action Adventure Fantasy
9	Lord of the Rings: The Return of the King, The...	34.237021	12366	7153	Action Adventure Fantasy
10	Indiana Jones and the Last Crusade (1989)	33.121802	16145	1291	Action Adventure

Model-based Model

```
# Create a function to return 10 most recommended movies for a selected user

def predict_user(user_id):
    if user_id in df_f.userId.unique():
        ui_list = df_f[df_f.userId == user_id].movieId.tolist()
        d = {k: v for k,v in Mapping_file.items() if not v in ui_list}
        predicted_list = []
        for i, j in d.items():
            predicted = svd.predict(user_id, j)
            predicted_list.append((i, predicted[3]))
        pdf = pd.DataFrame(predicted_list, columns = ['movies', 'ratings'])
        pdf.sort_values('ratings', ascending=False, inplace=True)
        pdf.set_index('movies', inplace=True)
        return print(pdf.head(10))
    else:
        print("Cannot find User Id in the list")
        return None
```

Memory-based Method

	hybrid
Star Wars: Episode IV - A New Hope (a.k.a. Star Wars) (1977)	1.000000
Star Wars: Episode V - The Empire Strikes Back (1980)	0.874397
Star Wars: Episode VI - Return of the Jedi (1983)	0.866502
Star Wars: Episode I - The Phantom Menace (1999)	0.752074
Alien (1979)	0.735173
Aliens (1986)	0.727314
Raiders of the Lost Ark (Indiana Jones and the Raiders of the Lost Ark) (1981)	0.670736
Star Wars: Episode II - Attack of the Clones (2002)	0.660800
Star Trek II: The Wrath of Khan (1982)	0.640659
2001: A Space Odyssey (1968)	0.629786
Star Wars: Episode III - Revenge of the Sith (2005)	0.623952

Model-based Method

Example user: 1991

```
# Using the function above to display the top 10 recommended movies for a selected user

user_id = 1991
predicted_ratings = predict_user(user_id)
```

movies	ratings
Life Is Beautiful (La Vita Ã¨ bella) (1997)	4.232855
Braveheart (1995)	4.228899
Mr. Holland's Opus (1995)	4.180408
Shawshank Redemption, The (1994)	4.170698
Green Mile, The (1999)	4.165669
Lord of the Rings: The Return of the King, The ...	4.143111
Crash (2004)	4.095977
Lord of the Rings: The Two Towers, The (2002)	4.093613
Sixth Sense, The (1999)	4.079201
Schindler's List (1993)	4.075007

THANK YOU