1st Capstone project milestone report:

"Exploring Netflix's movie recommendation system"

1. The Problem:

Recommender systems are a class of information filtering that is employed in multiple online services/platforms (e.g. movies, music, books, etc.), where the extremely large number of available options makes it impossible for a single consumer to have explored and evaluated all of the possible products. For this capstone project, I plan to explore different algorithms used in a movie recommender system.

2. Who is the client?

Netflix is one of the most popular online entertainment platforms. The main product of Netflix is its on-demand video streaming service, which allows users to stream movies or television series through multiple electronic devices. As such, Netflix relies on a movie recommender system to provide users with a selection of digital content that users are most likely to enjoy. Therefore, having an accurate movie recommender system is of special interest to Netflix.

3. Where can the data be obtained?

The initial dataset comprising the Netflix prize dataset can be downloaded from kaggle (link: https://www.kaggle.com/netflix-inc/netflix-prize-data)

The original dataset was preprocessed and regrouped into 4 .txt files, comprising a total of 100480507 ratings by 480189 users for a collection of 17770 movies. A separate .csv file contains the list of name of movies and year of release.

4. Data Wrangling Steps:

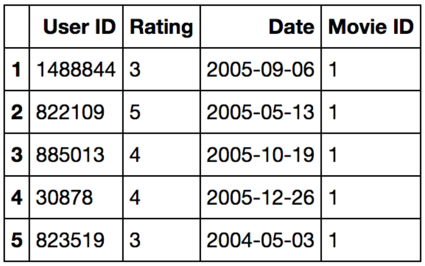
(*: all of the ipython notebooks mentioned here are included in the same folder as this report)**

The data comes in the form of four .txt files, each with the following format:

Movie X:
Customer ID_A, rating, date of rating
Customer ID_B, rating, date of rating
....
Movie Y:
Customer ID_C, rating, date of rating
Customer ID_D, rating, date of rating

My first step was to use the code in "**data_processing.ipynb**" to compile the 4 txt file into a single dataframe, where each row corresponds to 1 movie rating by one specific user. The columns of the resulting data frame are: "User ID", 'Rating", "Date", "Movie ID".

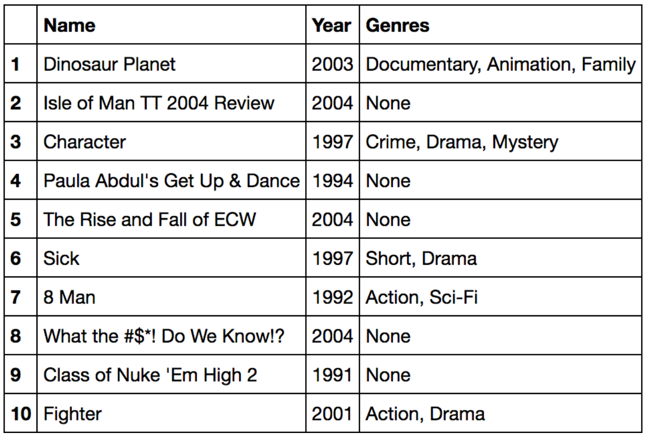
The following image shows how this dataframe looks like:



	User ID	Rating	Date	Movie ID
1	1488844	3	2005-09-06	1
2	822109	5	2005-05-13	1
3	885013	4	2005-10-19	1
4	30878	4	2005-12-26	1
5	823519	3	2004-05-03	1

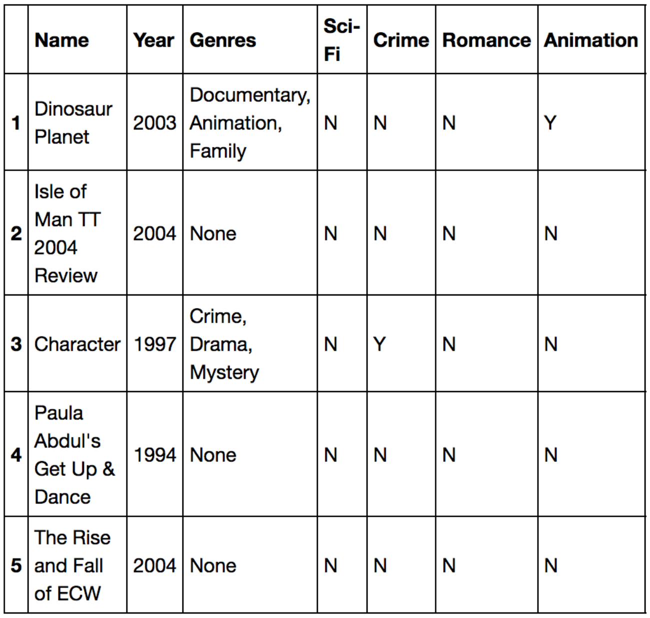
I also used the code in "**movie_omdb_request.ipynb**" to retrieve the genre to which each movie belongs. For this, I used the "omdb" python package, which allows performing queries to the "Open Movie Database (OMDb)" api (http://www.omdbapi.com/). Unfortunately, this database does not contain information for all of the movies in the Netflix Prize data set. Regardless, I was able to retrieve the genre information for ~72 % of the all the movies (17770), which corresponds to ~90% of the total ratings.

The final steps for data wrangling are contained in "**data_processing_2.ipynb**". The genre information retrieved from the omdb api came in the form of a single string containing the respective genres. That information was added to the original csv file with the list of movies. The resulting dataframe is shown below:



	Name	Year	Genres
1	Dinosaur Planet	2003	Documentary, Animation, Family
2	Isle of Man TT 2004 Review	2004	None
3	Character	1997	Crime, Drama, Mystery
4	Paula Abdul's Get Up & Dance	1994	None
5	The Rise and Fall of ECW	2004	None
6	Sick	1997	Short, Drama
7	8 Man	1992	Action, Sci-Fi
8	What the #$*! Do We Know!?	2004	None
9	Class of Nuke 'Em High 2	1991	None
10	Fighter	2001	Action, Drama

For subsequent data analysis, I opted to add to the list of movies dataframe one column for each available genre. For this columns, ‘Y’ means that the respective movie belongs to that genre, and ‘N’ that it doesn’t. The following image shows what the resulting data frame looks like:



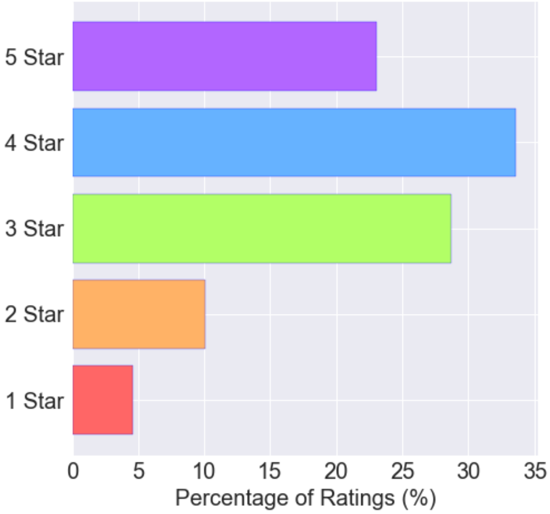
	Name	Year	Genres	Sci-Fi	Crime	Romance	Animation
1	Dinosaur Planet	2003	Documentary, Animation, Family	N	N	N	Y
2	Isle of Man TT 2004 Review	2004	None	N	N	N	N
3	Character	1997	Crime, Drama, Mystery	N	Y	N	N
4	Paula Abdul's Get Up & Dance	1994	None	N	N	N	N
5	The Rise and Fall of ECW	2004	None	N	N	N	N

The index of this dataframe corresponds to the movie id from the original dataset.

5. Exploratory data analysis:

The purpose of this exploratory data analysis is to identify possible variables in the Netflix Prize Data set that could potentially act as important factors in predicting movie ratings. The specific details of this analysis are contained in “**EDA.ipynb**”.

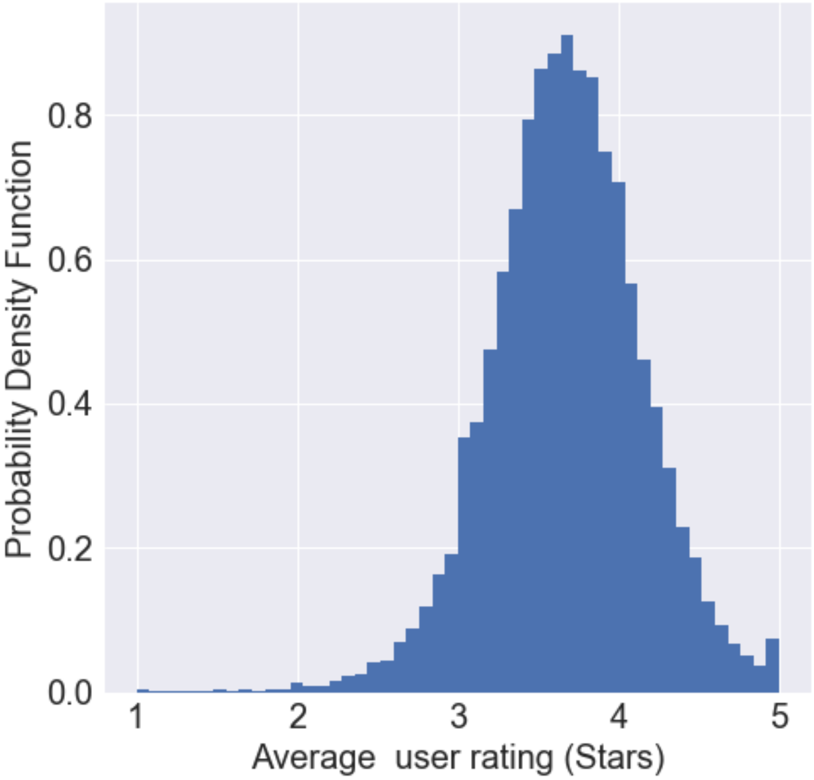
To start with, since we want to eventually predict movie ratings, I looked at the overall distribution of the movie ratings, as shown in the following bar graph:



The previous bar plot shows that the most frequent movie rating is 4 stars. I also calculated that the majority of the ratings (~ 56%) received a good rating (4 or 5 Stars). Overall, the average movie rating is ~ 3.6 Stars, with a standar deviation of ~1.09 Stars.

Do some users have bias when giving ratings?

To answer this question, I looked at the distribution average user rating (plot below):

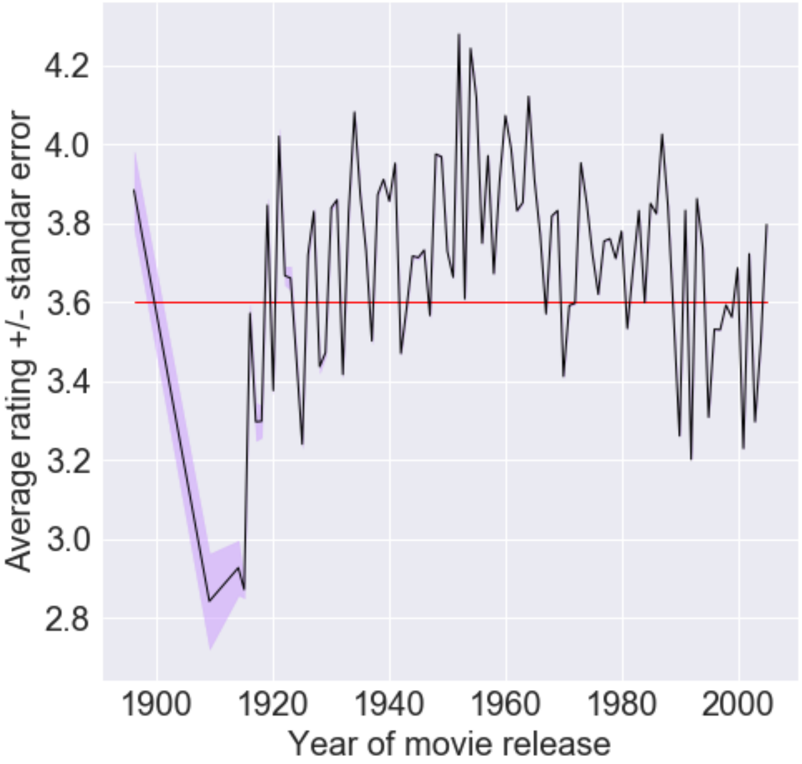


Looking at the above histogram, it seems that the average user rating is normally distributed. Indeed, running a normality test (pvalue < 0.05; normaltest() from scipy.stats), confirms this observation. If we assume that users with an average ratings above 4 Stars or below 3 Stars average have a positive or negative bias when evaluating movies, I saw that ~30% of users have a bias when rating movies. Thus, it could potentially be informative to include a user-specific bias in a movie rating predictive model.

Of note doing a quick calculation, turns out that on average each user gave ~209 ratings. Surprisingly, the biggest number of ratings given by a single user was 17653!!!

Does the movie release year have an effect on ratings?

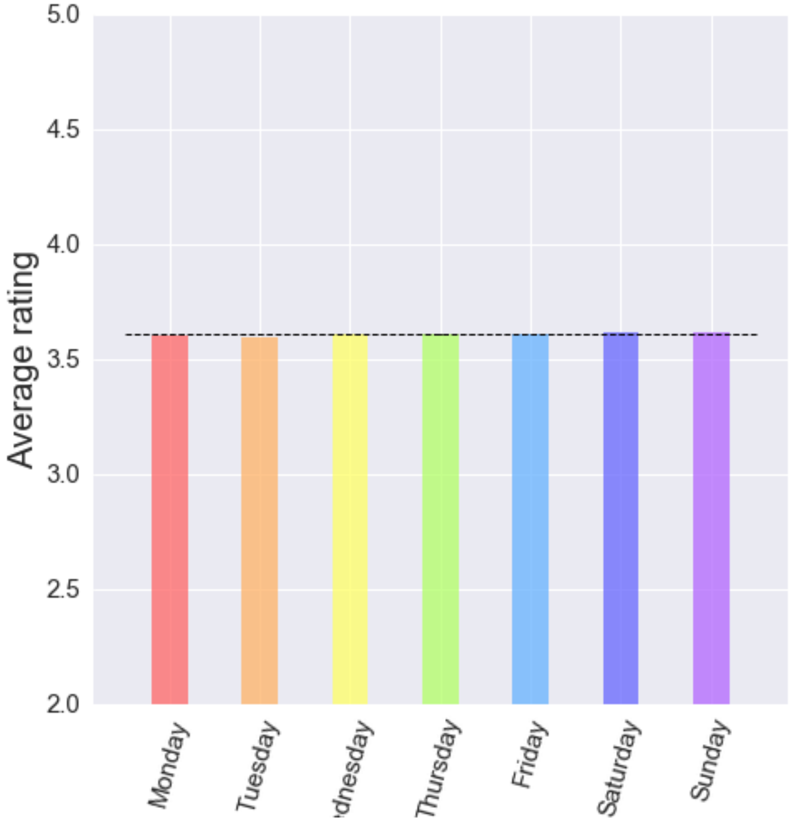
Since the Netflix prize data set also includes the year in which a movie is released, I also analyzed if movies released on different years had different ratings. For this I grouped the data by movie's release year, and plotted the average rating for each year.



The previous plot shows the average rating for each year (black line) +/- the standard error (shaded region around black line). The red line represents the overall average rating (~3.6 Stars). The small standard error for most years shows that the average rating on that year is different than the overall average rating. This suggests that the year in which the movie was released may be an important factor affecting the rating predictions. In fact, using a one-sample t-test to compare them to the overall average (considering pvalues < 0.001 as indicating significance), showed that 89 out of the 94 different movie release year are significantly different than the overall average. Thus, the year in which a particular movie was release may have significant effect on the predictions of movie ratings.

Does the day of the week have an effect on the ratings?

Another piece of information on the Netflix prize dataset is the date in which a rating was given. I used this information to test if the days of the week could have a significant effect in the movies rating. For this, I used the “datetime” module to convert the date into a day of the week, and grouped the data accordingly.



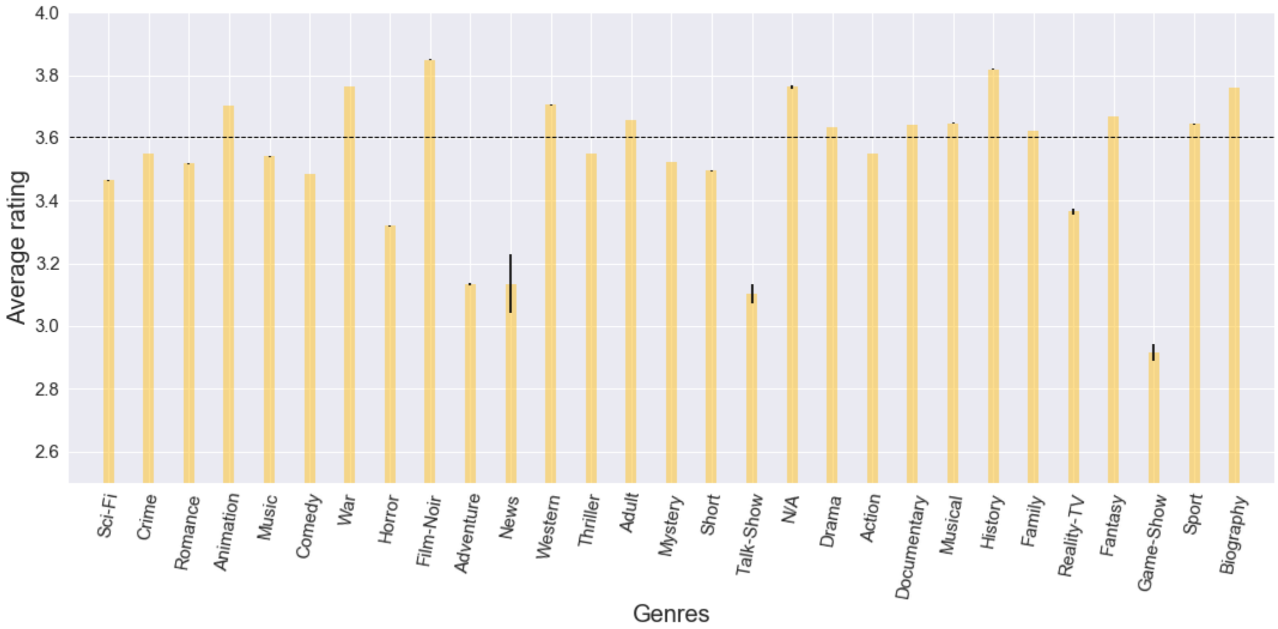
When extracting the data from the main dataframe, a one sample t-test was used to asses if the ratings for a respective day of the week were different from the overall mean (~3.6). Significant differences where considered when pvalue <0.001.

The one sample t-test revealed that Monday, Tuesday, Friday, Saturday and Sunday where different than the overall average. However when we look at the bar plot above we can see that, although some days had ratings different than the average (~3.6; dashed black line in the bar plot graph), the average rating for all days of the week is very similar to the overall average (~3.6; see the printed values above).

This suggests that the day of the week might not be an influential factor when predicting movie ratings.

Is the rating of a movie affected by its genre?

I originally retrieved the movie's genre from the Open Movie database API (OMDb) with the idea of investigating if different genres tend to have different ratings.



The bar plot above (average genre rating +/- standard error), strongly suggests that the genre of a movie affects the rating of that type of movie. In fact running a 1-sample t-test run on previous steps, it was possible to determine that all of the genres differ significantly (pvalue < 0.001) from the overall mean (black dashed line). While some genres, like Family, don't differ much from the overall average rating, other genres, like Game-Show, show much bigger differences.

6. Conclusion:

After performing this EDA, it can be concluded that factors such as 'User specific bias', 'Year of movie release' and 'Movie's Genre' significantly affect the rating of movies, thus, they can potentiality be incorporated into a predictive model for movie preferences and ratings. In further steps, I would like to also calculate more specific user biases. For example, one could calculate the bias that a user might have when rating a particular genre.