Homework assignment Steven Bowler 20562494 UTRGV CSCI6370 Dr. Lei

Submit: a report in word doc format answering 4 questions:

1. Describe the dataset
2. Describe how the data is loaded
3. What is the average rating for movie ID 1001?
4. What is the average rating that user ID 20001 gives to movies?

Homework Question 1

Netflix study data is provided in three principal parts

```
1.Training Data - in for files titled 'combined_data_*.txt total 100MM records
2.Test Data - in the file probe.txt
3.Qualifying Data - used for the competition to provide predictions against
```

The above 3 files are in the same general format, that each need to be parsed out into a table: Movie_Id: Cust_Id,Rating,Date

There is also a Movie Titles file that contains Movie_Id and Movie_Title.

Homework Question 2

Currently the data is loaded using two Jupyter notebooks, see Github repo (https://github.com/stevenbowler/netflixstudy) stevenbowler/netflixstudy:

1. Data-Wrangling: Load, clean, store as .csv see Github Data Wrangling (https://github.com/stevenbowler/netflixstudy/blob/master/reports/netflixstudyDataWranglingForCSCI6370.pdf)
2. Preliminary EDA: this same file, see Github Preliminary EDA (https://github.com/stevenbowler/netflixstudy/blob/master/reports/netflixstudyEDAforCSCI6370.pdf)

Homework Question 3

Pick a movie id and show average rating for that movie

In [39]:
```
Movie_id = 1001
movie_average = df[df['Movie_Id']==Movie_id].Rating.mean()
print('Movie ',Movie_id,'had an average rating of',np.round(movie_average,2))
```

```
Movie  1001 had an average rating of 3.29
```

Homework Question 4

pick a customer id and show average rating for that customer

In [38]:
```
# per homework question 4, pick a customer id and show average rating for that customer
Customer_id = 97
customer_average = df[df['Cust_Id']==Customer_id].Rating.mean()
print('Customer ',Customer_id,'had an average rating of',np.round(customer_average,2))
```

```
Customer  97 had an average rating of 3.23
```