1 Dataset Description:

The provided dataset files [1] i.e. ratings.dat, users.dat and movies.dat contain 1 million movie ratings. This set is a subset of a continous collection that has gone on since 1998 [1].

The ratings file contains 4 columns for each of the 1,000,209 instances where each defines a single rating by a user on a single movie. A snippet of the data can be seen in Table 1. In terms of number of ratings, each user have given at least 20 ratings from 1 to 5 on a set of movies which are associate via their unique movieid key/value relations i.e. the UserID, MovieID in the ratings file.

The users file as shown in Table 2 contains 6040 user instances, as 2 shows it includes their gender, age, occupation and zip-code. The ages are divided up into 7 different range bands, e.g. 18 means the user is between 18 and 24. Occupation on the other hand differentiates the users into 21 different categories made up of string values.

Lastly the movies.dat file (see Table 3) contains 3900 different movie instances, specifically each row contains an integer unique identifier and two string columns containing first a string with the title of the movie and secondly the category or categories that applies to this movie out of the 18 available.

Table 1: Ratings.dat file

ratings.dat	UserID	MovieID	Rating	Timestamp
	1::	1193::	5::	978300760
	1::	661::	3::	978302109
	1::	914::	3::	978301968
	1::	3408::	4::	978300275

Table 2: Users.dat file

users.dat	UserID	Gender	Age	Occupation	Zip-code
	1::	F::	1::	10::	48067
	2::	M::	56::	16::	70072
	3::	M::	25::	15::	55117
	4::	M::	45::	7::	02460

Table 3: Movies.dat file

movies.dat	MovieID	Title	Genres
	1::	Toy Story (1995)::	Animation—Children's—Comedy
	2::	Jumanji (1995)::	Adventure—Children's—Fantasy
	3::	Grumpier Old Men (1995)::	Comedy—Romance
	4::	Waiting to Exhale (1995)::	Comedy—Drama

Similar datasets include the book-crossing set [3] which contains a collection of around 1.1 million ratings from 278.000 users on 271.000 books collected in 2004. Similarities includes the 3 file layout where the users the books and the actual ratings are divided. Here relational identifiers are also used to tie movie books and ratings together via unique string values and a rating value between 0 and 10.

2 Technique for loading data:

Pandas is the best for loading these data files because it can combine these separate files using the relational identifiers into a single data structure as shown in [2]. Furthermore it is able to take separator values such as the "::" in this instance into account during this process. After loading the data a combined structure (DataFrame) is available that supports queries that can extract information such as "How many movies have an average rating over 4?

Comparing this to loading the data into 3 different file objects and then combining the columns manually with the unique identifiers is just plain overhead compared to the pandas approach.

3 Questions:

1. Movies with an average rating over 4? Answer: **370**
2. Movies rated over 4 on average by men? Answer: **364**
3. Movies rated below 4 on average by women? Answer: **2912**
4. Top-10 movies?

   Ratings statistical overview.

Figure 1: Statistics

```
Rating mean: 269.889098759
Rating std: 384.047837572
Rating median: 123.5
Rating skew: 2.81493596238
```

Figure 2: Rating Frequency

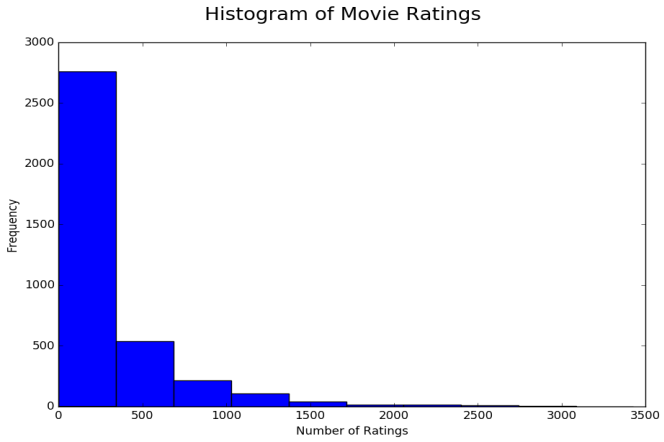

Figure 1 and Figure 2 shows that there is a positive skew in the distribution, i.e. the frequency of submitted ratings to each movie is bottom heavy. Because of this a top movie was defined as being a movie that has at least 300 ratings and the highest average rating. See the top 10 movies in 3.

Figure 3: Top 10 Movies

```
---- The top movies are ----
Number of applicable movies: 1058
title
Seven Samurai (The Magnificent Seven) (Shichinin no samurai) (1954)    4.560510
Shawshank Redemption, The (1994)                                       4.554558
Godfather, The (1972)                                                  4.524966
Close Shave, A (1995)                                                  4.520548
Usual Suspects, The (1995)                                             4.517106
Schindler's List (1993)                                                4.510417
Wrong Trousers, The (1993)                                             4.507937
Sunset Blvd. (a.k.a. Sunset Boulevard) (1950)                          4.491489
Raiders of the Lost Ark (1981)                                         4.477725
Rear Window (1954)                                                     4.476190
```

References

[1] F Maxwell Harper and Joseph A Konstan. The MovieLens Datasets : History and Context. 5(4):1–19, 2015.

[2] Wes McKinney. *Python for Data Analisys*. 2013.

[3] Cai-Nicolas C.N. Ziegler, Sean M. S.M. McNee, Joseph a. J.a. Konstan, and Georg Lausen. Improving recommendation lists through topic diversification. *Proceedings of the 14th international conference on World Wide Web WWW 05*, (January):22, 2005.