**Movielens Case Study**

Project 1

DESCRIPTION

**Background of Problem Statement :**

The GroupLens Research Project is a research group in the Department of Computer Science and Engineering at the University of Minnesota. Members of the GroupLens Research Project are involved in many research projects related to the fields of information filtering, collaborative filtering, and recommender systems. The project is led by professors John Riedl and Joseph Konstan. The project began to explore automated collaborative filtering in 1992 but is most well known for its worldwide trial of an automated collaborative filtering system for Usenet news in 1996. Since then the project has expanded its scope to research overall information by filtering solutions, integrating into content-based methods, as well as, improving current collaborative filtering technology.

**Problem Objective :**

Here, we ask you to perform the analysis using the Exploratory Data Analysis technique. You need to find features affecting the ratings of any particular movie and build a model to predict the movie ratings.

**Domain**: Entertainment

**Analysis Tasks to be performed:**

* Import the three datasets
* Create a new dataset [Master\_Data] with the following columns MovieID Title UserID Age Gender Occupation Rating. (Hint: (i) Merge two tables at a time. (ii) Merge the tables using two primary keys MovieID & UserId)
* Explore the datasets using visual representations (graphs or tables), also include your comments on the following:

1. User Age Distribution
2. User rating of the movie “Toy Story”
3. Top 25 movies by viewership rating
4. Find the ratings for all the movies reviewed by for a particular user of user id = 2696

* Feature Engineering:

            Use column genres:

1. Find out all the unique genres (Hint: split the data in column genre making a list and then process the data to find out only the unique categories of genres)
2. Create a separate column for each genre category with a one-hot encoding ( 1 and 0) whether or not the movie belongs to that genre.
3. Determine the features affecting the ratings of any particular movie.
4. Develop an appropriate model to predict the movie ratings

**Dataset Description :**

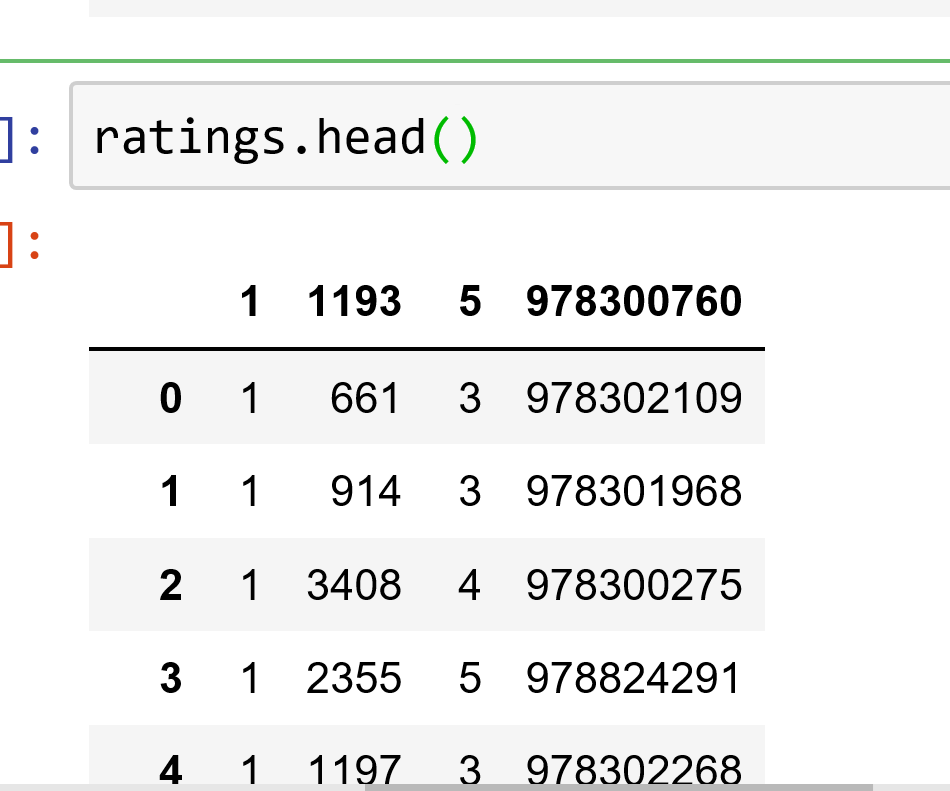
These files contain 1,000,209 anonymous ratings of approximately 3,900 movies made by 6,040 MovieLens users who joined MovieLens in 2000.

**Ratings.dat**

    Format - UserID::MovieID::Rating::Timestamp

|  |  |
| --- | --- |
| **Field** | **Description** |
| UserID | Unique identification for each user |
| MovieID | Unique identification for each movie |
| Rating | User rating for each movie |
| Timestamp | Timestamp generated while adding user review |

* UserIDs range between 1 and 6040
* The MovieIDs range between 1 and 3952
* Ratings are made on a 5-star scale (whole-star ratings only)
* A timestamp is represented in seconds since the epoch is returned by time(2)
* Each user has at least 20 ratings



Timestamp

**RATING**

**MOVIE ID**

**USER ID**

**Users.dat**

Format -  UserID::Gender::Age::Occupation::Zip-code

|  |  |
| --- | --- |
| Field | Description |
| UserID | Unique identification for each user |
| Genere | Category of each movie |
| Age | User’s age |
| Occupation | User’s Occupation |
| Zip-code | Zip Code for the user’s location |

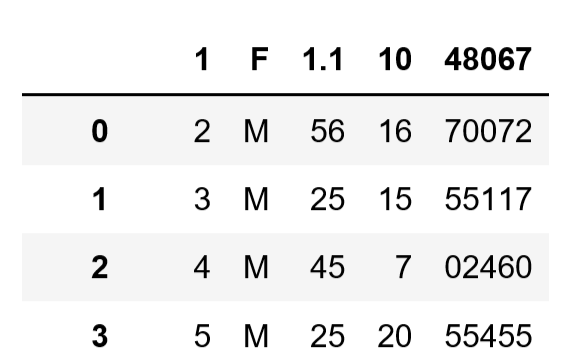
All demographic information is provided voluntarily by the users and is not checked for accuracy. Only users who have provided demographic information are included in this data set.

* Gender is denoted by an "M" for male and "F" for female
* Age is chosen from the following ranges:

|  |  |
| --- | --- |
| **Value** | **Description** |
| 1 | "Under 18" |
| 18 | "18-24" |
| 25 | "25-34" |
| 35 | "35-44"  **Occupation** |
| 45  **AGE** | "45-49" |
| 50  **GENDER** | "50-55" |
| 56 | "56+" |

**USER ID**

**ZIP CODE**



* Occupation is chosen from the following choices:

|  |  |
| --- | --- |
| **Value** | **Description** |
| 0 | "other" or not specified |
| 1 | "academic/educator" |
| 2 | "artist” |
| 3 | "clerical/admin" |
| 4 | "college/grad student" |
| 5 | "customer service" |
| 6 | "doctor/health care" |
| 7 | "executive/managerial" |
| 8 | "farmer" |
| 9 | "homemaker" |
| 10 | "K-12 student" |
| 11 | "lawyer" |
| 12 | "programmer" |
| 13 | "retired" |
| 14 | "sales/marketing" |
| 15 | "scientist" |
| 16 | "self-employed" |
| 17 | "technician/engineer" |
| 18 | "tradesman/craftsman" |
| 19 | "unemployed" |
| 20 | "writer” |

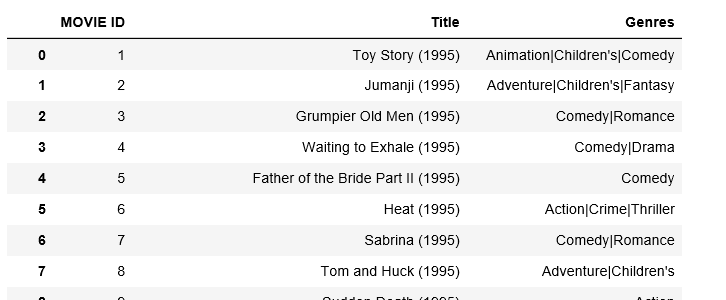
**Movies.dat**

Format - MovieID::Title::Genres

|  |  |
| --- | --- |
| Field | Description |
| MovieID | Unique identification for each movie |
| Title | A title for each movie |
| Genres | Category of each movie |

* Titles are identical to titles provided by the IMDB (including year of release)

* Genres are pipe-separated and are selected from the following genres:

1. Action
2. Adventure
3. Animation
4. Children's
5. Comedy
6. Crime
7. Documentary
8. Drama
9. Fantasy
10. Film-Noir
11. Horror
12. Musical
13. Mystery
14. Romance
15. Sci-Fi
16. Thriller
17. War
18. Western

* Some MovieIDs do not correspond to a movie due to accidental duplicate entries and/or test entries
* Movies are mostly entered by hand, so errors and inconsistencies may exist

Please download the dataset from [here](https://github.com/Simplilearn-Edu/Data-Science-with-Python-Project-One)

Good Luck!!

Program

1. **import pandas as pd**
   1. **1 ) Import the three datasets**
2. **Importing dat files**

users=pd.read\_csv("C:/Users/nilesh/Downloads/users.dat",header=None,sep='::')

movies=pd.read\_csv("C:/Users/nilesh/Downloads/movies.dat",header=None,sep='::')

ratings=pd.read\_csv("C:/Users/nilesh/Downloads/ratings.dat",header=None,sep='::')

1. **Assign Columns names**

**users.columns=[ ' USER ID ' , ' GENDER ' , ' AGE ' , ' Occupation ' , ' ZIP CODE ']**

**ratings.columns=['USER ID','MOVIE ID', 'RATING','Timestamp']**

**movies.columns=['MOVIE ID','Title','Genres']**

1. **“ RENAME ”**

**ratings.rename({"RATING" : "rating"},axis='columns')**

1. **“ REINDEX “**

**Master\_data = Master\_data.reindex(columns=['MOVIE ID','Title','USER ID',' AGE','GENDER','Occupation','RATING'])**

Q. 2) Create a new dataset [Master\_Data] with the following columns MovieID Title UserID Age Gende Occupation Rating.

1. **Merge**

**1 . data1= pd.merge(left=users,right=ratings,on='USER ID',how='outer')**

1000386 rows × 8 columns

**Master\_data= pd.merge(left=movies,right=data1,on='MOVIE ID',how='outer')**

**2. data1= pd.merge(left=users,right=ratings,on='USER ID',how='outer')**

1000209 rows × 8 columns

**Master\_data= pd.merge(left=movies,right=data1,on='MOVIE ID',how='outer')**

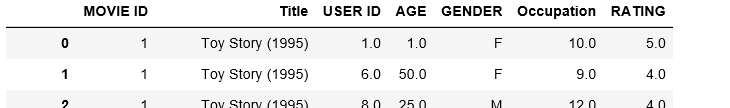
1. **Dropna()**

**Master\_data= Master.dropna(axis = 0, how ='any')**

Old data frame length: 1000386

New data frame length: 1000209

1. **Drop**

** Master\_data.drop(['Genres','Timestamp','ZIP CODE'],axis=1,inplace=True)**

Q. 3) Explore the datasets using visual representations

1. User Age Distribution

