Final Project Report

INFS 580 FALL 2022

Movielens Dataset (ml-1m)
Data Analysis using Python, R & SQL

By Srividya Panchagnula

Due Date : Dec 5th, 2022

Abstract

This project and report contain methods and descriptions of data analysis using Python, R and SQL respectively. The dataset used is MovieLens ml-1m dataset containing 3 different datasets(movies, rating, users). Python is used for data converting extracting, merging, creation of new meaningful datasets, univariate datatype analysis, Feature engineering in which methods using python libraries such as NumPy, pandas...etc. are used. Python is also used for regression , linear analysis, manipulation, and evaluation of the dataset. The Python files are executed in google Colaboratory and the following code is such that to support the ide environment. R is used for visualizations using libraries such as ggplot2…etc. the R scries are run using R studio. SQL is used to show a schema for the dataset (, load your dataset, and to execute a few simple queries against the table to demonstrate relationships between columns of the datasets and for data summaries.

Introduction

Movielens ml-1m dataset has three file which are movies, rating and users.dat, related to the film industry and helps understand the ratings, pattern, public interests and popularity, gender, age groups as such. The dataset is generally used in research projects of Data analytics and recommender systems. MovieLens 1M movie ratings is a stable benchmark dataset with 1 million ratings from 6000 users on 4000 movies which was released in 2/2003, researched through 2000-2003. **The Movielens ml-datasets contain 1,000,209 anonymous ratings of approximately 3,900 movies made by 6,040 MovieLens users who joined MovieLens in 2000. [1]**

The below information is from the READ.ME file of the dataset zip. It explains each dataset of the ml-1m zip file.

RATINGS FILE DESCRIPTION [2]
All ratings are contained in the file "ratings.dat" and are in the
following format:

UserID::MovieID::Rating::Timestamp

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

USERS FILE DESCRIPTION [2]

User information is in the file "users.dat" and is in the following format:

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

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

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

- 1: "Under 18"
- 18: "18-24"
- 25: "25-34"
- 35: "35-44"
- 45: "45-49"
- 50: "50-55"
- 56: "56+"

- Occupation is chosen from the following choices:
- 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 FILE DESCRIPTION [2]

Movie information is in the file "movies.dat" and is in the following format:

MovieID::Title::Genres

- Titles are identical to titles provided by the IMDB (including year of release)
- Genres are pipe-separated and are selected from the following genres:

- Action
- Adventure
- Animation
- Children's
- Comedy
- Crime
- Documentary
- Drama
- Fantasy
- Film-Noir
- Horror
- Musical
- Mystery
- Romance
- Sci-Fi
- Thriller
- War
- 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 (this dataset was made in the early 2000s timeframe)

Literature Review [3]

First Paper
Exploratory Data Analysis using Python
Kabita Sahoo, Abhaya Kumar Samal, Jitendra Pramanik, Subhendu Kumar Pani

Data need to be analyzed so as to produce good result. Using the result decision can be taken. For example, recommendation system, ranking of the page, demand fore casting, prediction of purchase of the product. There are some leading companies where the review of the customer plays a great role to analyze the factor which influences the review rating. We have used exploratory data analysis (EDA) where data interpretations can be done in row and column format. We have used python for data analysis. it is object oriented ,interpreted and interactive programming language. it is open source with rich sets of libraries like pandas, MATplotlib, seaborn etc. We have used different types of charts and various types of parameters to analyze Amazon review data sets which contains the reviews of electronic data items. We have used python programming for the data analysis.
Keywords: Exploratory Data Analysis (EDA); MATPplotlib; Seaborn, Visualization; Pandas; Jupyter Notebook
Boxplots :A good graphical image of the concentration of data can be represented by the use of box plot. It shows the central tendency, symmetry, skew and outlier. It can be constructed from five values: the minimum, the first quartile, the median, the third quartile and the maximum value. These values are compared to show how close other data values are to them.

Second Paper [4]
On the Difficulty of Evaluating Baselines

The author has demonstrated in this [1] study that the ML benchmark findings for baselines that have been utilized in various papers over the past five years are not ideal. We were able to outperform not just the given results for the baselines but even the reported results of any recently proposed approach with careful setup of a simple vanilla matrix factorization baseline. Even more gain is offered by other well-known models like SVD++. These results are unexpected considering that the articles perform a respectable hyperparameter search, indicate statistical significance, and permit reproducibility all best practices in our community for ensuring reliable results. This suggests that it is challenging to execute baseline procedures correctly. The matrix factorization used in my final project differs from the results of the above research paper, the MF used in my project is TF – IDF. The term frequency-inverse document frequency, or TFIDF for short, is a numerical statistic used to measure how essential a word is to a document inside a corpus or collection. [1] In information retrieval, text mining, and user modeling searches, it is frequently employed as a weighting factor. To account for the fact that some words are used more frequently than others overall, the tf-idf value rises according to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the term. One of the most common term-weighting techniques used nowadays is tf-idf. In digital libraries, 83% of text-based recommender systems employ tf-idf, according to a 2015 survey.
This method will help us find movies public tend to like most for us to find movies with highest ratings (“Movies which are most liked or accepted”).

Research Questions

Movies have always been a major part of everyone's life, something that people enjoy watching on when sad, happy, with friends on weekends or just a solo pastime. This mundane habit of us humans do Movies a billion dollar business with significant amount contribution to US economy. The film industry in US that is the Hollywood contributed about $504 billon to the USGDP or 3.2% of the goods and services portion of GDP. In my opinion, with an Industry as such analysis to the data generated from industry is necessary to maintain an understanding and profitable graph in both overall and personal level.

What are the movies with the highest ratings?
This question helps to answer the most important and basic question which helps producers to understand the type of movie that hit popularity the most to understand what type of movies that being made in future will be an able to become popular as well.

Which genres have been unique over time?"
OR
What are emerging genre combinations from existing genres that are receiving rating?
This question helps to distinguish which genre to focus on while producing a movie to understand which has been always stayed on trend or have been liked by people for a duration of time and which are emerging new combinations of already popular or existing genres.

The ages actively giving the most rating?
Or
What is the avg user age?

This question helps focus on which age group to commonly focus on throughout the genres to have a common target in mind for differing combinations.

The average ratings received to the movies throughout.? This question gives us a rating to leverage public onions on the bases. If a movie receives less than avg rating it shows the movie isn't well liked

what is the gender ratio of among the users(m or f)?
gender specifications usually used to rate which gender is usually attracted to what and how a gender is responds ratings requests of which gender rates less.

Data Analysis using Python

Python is a popular programming language that can be used for a variety of purposes. Its adaptability and extensive collection of libraries, which are helpful for analytics and complex calculations, are two of the primary reasons it is so widely used. Because Python is so easily extended, it has thousands of libraries that are specifically geared toward analytics. One of the most popular of these is the Python Data Analysis Library (also known as Pandas) which is written for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series.
NumPy is a library written in Python that contains hundreds of mathematical calculations, operations, and functions. The vast majority of data analytics libraries written in Python are at least partially derived from NumPy.
This project's python code is executed using google Colaboratory.[5]

The project uses all the above explained libraries as follows,

Importing libraries
Numpy and pandas are imported for data processing and manipulation. The matplotlib and seaborn are imported for visualizations of the statistics.
Reading the data
The three datasets movies.dat, users.dat and ratings.dat are loaded into corresponding pandas dataframe called movies_df, users_df, and ratings_df. The datasets are stored in a directory called finalassgn which are converted into csv. These are helpful while processing data in R and SQL.
Univariate analysis:
First the statistics of the data are evaluated using describe() method.
Gender Ratio:
What is the gender ratio of among the users(m or f)?
In this analysis, first the gender ratio of users is derived by using value_count() method. The gender male is 71.7% and gender female is 28.2%.

	index	percentage
0	M	0.717053
1	F	0.282947

What are the movies with higher than avg ratings?
Movies with average rating more than 4 *(taking the base as 4 as the avg rating is among the lines of 3.5 – 3.6)*
The average rating of each movie is derived by using mean() method, then the number of movies with average rating more than 4 is displayed using count() method.
User age group and average user's age:
What is the avg user age?
Using the same methods, the user age group and average age of users is determined which is 31 years. Here, the gender is nominal data, rating is ordinal data, and age is descrete data.

```
25    2096
35    1193
18    1103
45     550
50     496
56     380
1      222
Name: Age, dtype: int64
```

Merging the datasets into one Master dataset

The movies_df and ratings_df dataframes are merged into single movie_rating dataframe using merge() method. Merge is done based on the key movieID. Then, the users_df is merged with the newly created dataframe.

In the master data, the zip-code and timestamp columns are dropped since they donot contribute to the multivariate analysis process.

The resultant master data has 8 columns MovieID, Title, Genres, UserID, Ratings, Gender, Age, and Occupation.

	MovieID	Title	Genres	UserID	Rating	Gender	Age	Occupation
0	1	Toy Story (1995)	Animation|Children's|Comedy	1	5.0	F	1	10
1	48	Pocahontas (1995)	Animation|Children's|Musical|Romance	1	5.0	F	1	10
2	150	Apollo 13 (1995)	Drama	1	5.0	F	1	10
3	260	Star Wars: Episode IV - A New Hope (1977)	Action|Adventure|Fantasy|Sci-Fi	1	4.0	F	1	10
4	527	Schindler's List (1993)	Drama|War	1	5.0	F	1	10

Feature engineering

Unique genres:

What are emerging genre combinations from existing genres that are receiving rating?

The value_count() method outputs the unique genre values. It is noted that most of the movies are combinations of different genres.

```
Comedy             10480
Drama               9544
Comedy|Romance      3915
Comedy|Drama        3707
Drama|Romance       2671
Name: Genres, dtype: int64
```

Genre and gender categories with one hot encoding

The genre column has multiple values, so onehotencoding is applied on each record with 0 if it does not belong to that sub-genre, 1 if it does. The gender is also one hot encoded with values 0 and 1 for the gender.

Features affecting rating:

The resultant dataframe with columns age, occupation, rating, 18 columns for genres and 2 columns for gender.

```
Age                    int64
Occupation             int64
Rating               float64
Genres_Action          int64
Genres_Adventure       int64
Genres_Animation       int64
Genres_Children's      int64
Genres_Comedy          int64
Genres_Crime           int64
Genres_Documentary     int64
Genres_Drama           int64
Genres_Fantasy         int64
Genres_Film-Noir       int64
Genres_Horror          int64
Genres_Musical         int64
Genres_Mystery         int64
Genres_Romance         int64
Genres_Sci-Fi          int64
Genres_Thriller        int64
Genres_War             int64
Genres_Western         int64
Gender_F               uint8
Gender_M               uint8
dtype: object
```

Linear regression

The updated dataframe is used in the linear regression model. The data is divided into train set and test set using train_test_split() method. The test size is taken as 20%. The rating column is used as the labels and remaining columns are used as parameters for the model. The linear regression method from sklearn library is used with train set as parameter. The fit() method is called to initiate the training. Then the trained model is used to make predictions on test set using the method predict().

Evaluation

We can derive y-intercept and coefficients from attributes of the model. The metric object is used to get MSE,RMSE, and MAE error values. The r2_score on the test set is 0.03. It is derived that age and occupation are the main features affecting the ratings of the movies.

```
y-intercept:  3.4446054978132232
Beta coefficients:  [ 0.00417635 -0.00290406 -0.05683106 -0.02663403  0.36776504 -0.27468965
 -0.02268764  0.0502738   0.33051552  0.24874651  0.06217411  0.41473244
 -0.26941255  0.14625133  0.01637325 -0.04751023 -0.00355147  0.02481103
  0.28429861  0.03682583  0.02463962 -0.02463962]
Mean Abs Error  MAE:  0.8993662992115887
Mean Sq Error  MSE:  1.1983237558247748
Root Mean Sq Error RMSE: 1.0946797503492858
r2 value:  0.037763005602315824
```

Data Analysis using R

From the extracted 'Master_data' dataset in the first part of python analysis which is a merged data of the MovieLens ml-1m datasets (movies.csv, rating.csv, users.csv files) which is taken and loaded into the r script of the r studio. [6]

R is a language that is designed for statistical computing, graphical data analysis, and scientific research. It is usually preferred for data visualization as it offers flexibility and minimum required coding through its packages ; library(ggplot2),library(grid),library(plyr)

The below visualizations are the results of using the above mentioned libraries, the

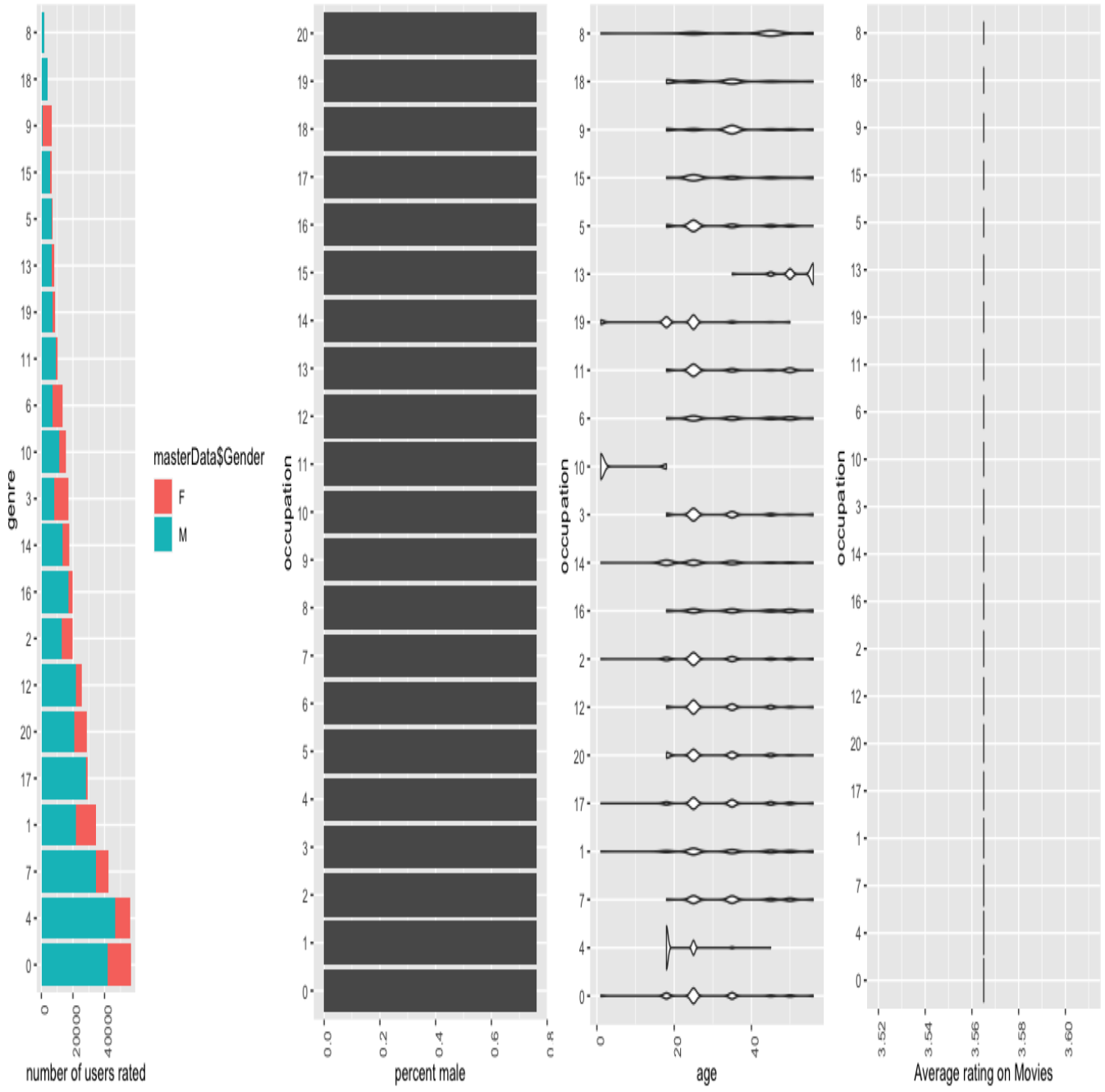
This result is python code result of created master data info used in this R Script
Note : Master_Data file is extracted using python and used for visulizations using R

```
Master_Data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 88829 entries, 0 to 88828
Data columns (total 8 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   MovieID     88829 non-null  int64
 1   Title       88829 non-null  object
 2   Genres      88829 non-null  object
 3   UserID      88829 non-null  int64
 4   Rating      88829 non-null  float64
 5   Gender      88829 non-null  object
 6   Age         88829 non-null  int64
 7   Occupation  88829 non-null  int64
dtypes: float64(1), int64(4), object(3)
memory usage: 6.1+ MB
```

The following visualizations are the result of r script using the mentioned libraries

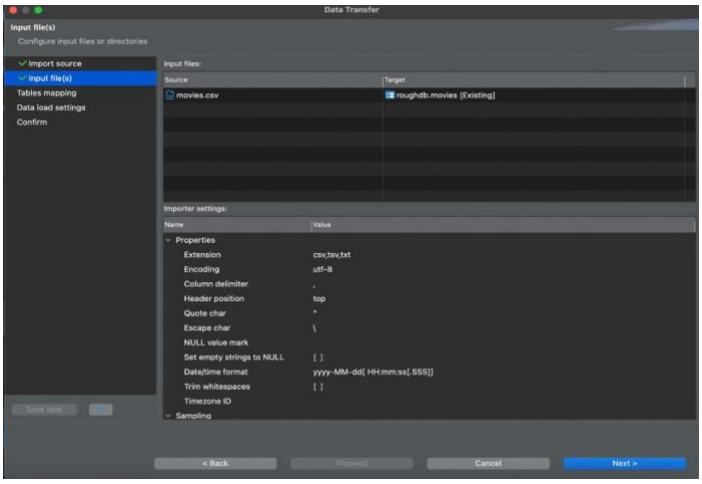
- The first graph explains the gender of the users rating vs the genres the specific genders rate
- The male percentage of rating to the occupation of the gender
- The age vs the occupation graph
- The avg rating visualization is 3.56 – 3.58



Data Analysis using SQL

SQL (Structured Query Language) is a programming language designed for managing data in a relational database. It's been around since the 1970s and is the most common method of accessing data in databases today. SQL has a variety of functions that allow its users to read, manipulate, and change data. [7]

Loading the dataset movies.csv into SQL IDE DB BEVER by uploading the dataset



SQL here is used to getting a particular userID meta data

```
mysql> select * from users where userid=100
    -> ;
+-----+--------+--------+------+------------+---------+---------+----------+
| id  | userid | gender | age  | occupation | zipcode | Column1 | zip-code |
+-----+--------+--------+------+------------+---------+---------+----------+
| NULL |    100 | M      |   35 |         17 | NULL    |      99 |    95401 |
+-----+--------+--------+------+------------+---------+---------+----------+
1 row in set (0.01 sec)
```

SQL query showing the count of gender specific ratings

```
mysql> select gender, count(*) AS counter from users
    -> GROUP BY gender
    -> ORDER BY counter DESC;
+--------+---------+
| gender | counter |
+--------+---------+
| M      |    4331 |
| F      |    1709 |
+--------+---------+
2 rows in set (0.01 sec)
```

Limitations

The dataset is minimal and does not take into account huge dataset. To data mine and use large dataset it is recommended to use spark as it supports big data processing.

Conclusions

The Data Analysis using python gives us an understanding of how to perform univariate and supervised learning models to predict ratings from the dataset using different parameters and methods

It is to be noted that the accuracy of the prediction can be improved using more complex models such as neural networks

Use of R was helpful in better and quick visualization of metadata and the relationship/correlation between varied data of movies, users, and their ratings.

Use of SQL was helpful in loading the dataset as an SQL schema and performing meaningful querying for data analysis

It is to be noted that the accuracy of the prediction can be improved using more complex models such as neural networks

References

[1]"MovieLens." 2019. GroupLens. April 26, 2019. https://grouplens.org/datasets/movielens/.

[2] 2015. Grouplens.org. 2015. https://files.grouplens.org/datasets/movielens/ml-1m-README.txt.

[3] Pramanik, Jitendra & Samal, Abhaya Kumar & Sahoo, Kabita & Pani, Dr. Subhendu. (2019). Exploratory Data Analysis using Python. International Journal of Innovative Technology and Exploring Engineering. 8. 4727-4735.

[4] Rendle, Steffen, Li Zhang, and Yehuda Koren. n.d. "On the Difficulty of Evaluating Baselines a Study on Recommender Systems." Accessed March 12, 2021. https://arxiv.org/pdf/1905.01395v1.pdf

[5] McKinney, Wes. 2017. *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. *Google Books*. "O'Reilly Media, Inc." https://books.google.com/books/about/Python_for_Data_Analysis.html?id=BCc3DwAAQBAJ&source=kp_book_description.

[6] "Data Visualization in R." 2020. GeeksforGeeks. April 12, 2020. https://www.geeksforgeeks.org/data-visualization-in-r/.

[7] Wikipedia Contributors. 2018. "SQL." Wikipedia. Wikimedia Foundation. December 11, 2018. https://en.wikipedia.org/wiki/SQL.