**Movielens Case Study**

**Project Report**

**Aparna Nair**

**aparna.nair.vijay@gmail.com**

**Business Scenario**

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.

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

**Objectives**

**Analysis Tasks**

1. Import the three datasets
2. Create a new dataset [Master\_Data] with the following columns MovieID Title UserID Age Gender Occupation Rating.
3. Explore the datasets using visual representations

* User Age Distribution
* User rating of the movie “Toy Story”
* Top 25 movies by viewership rating
* 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 |

Users.dat  
Format -  UserID::Gender::Age::Occupation::Zip-code

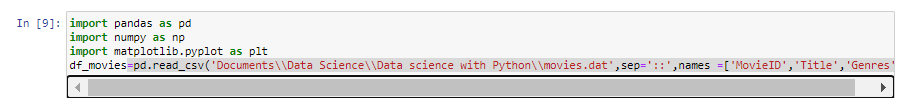
|  |  |
| --- | --- |
| Field | Description |
| UserID | Unique identification for each user |
| Genere | Category of each movie |
| Age | User’s age |
| Occupation | User’s Occupation |
| Zip-code | Zip Code for the user’s location |

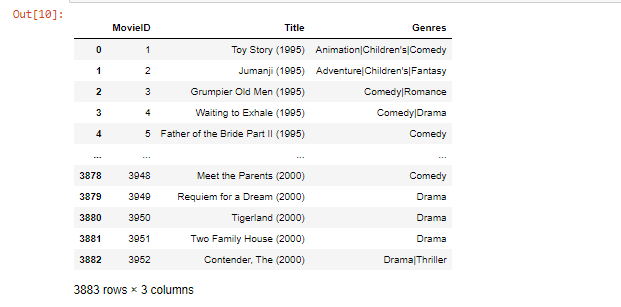
Movies.dat  
Format - MovieID::Title::Genres

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

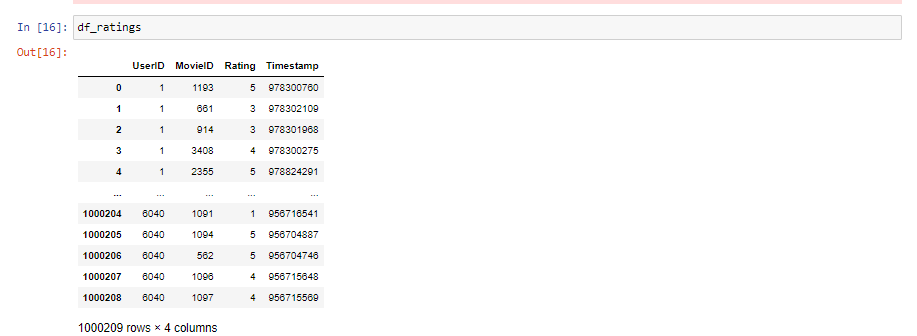
**Statistical algorithm execution – Python code and outputs**

Objective1)Import the three datasets

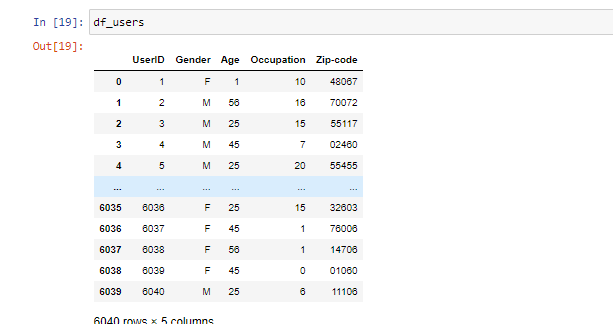






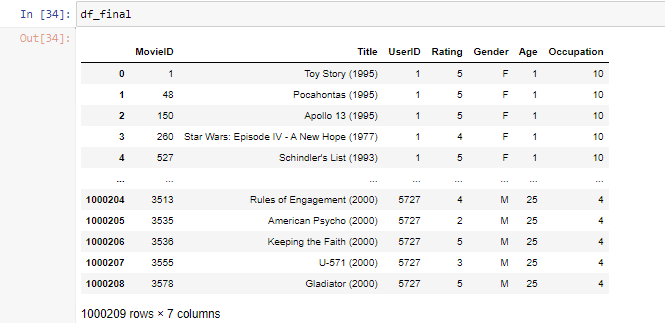






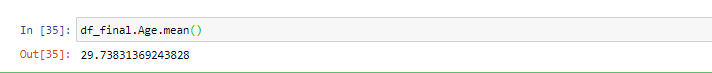
Objective 2)Create a new dataset [Master\_Data] with the following columns MovieID Title UserID Age Gender Occupation Rating

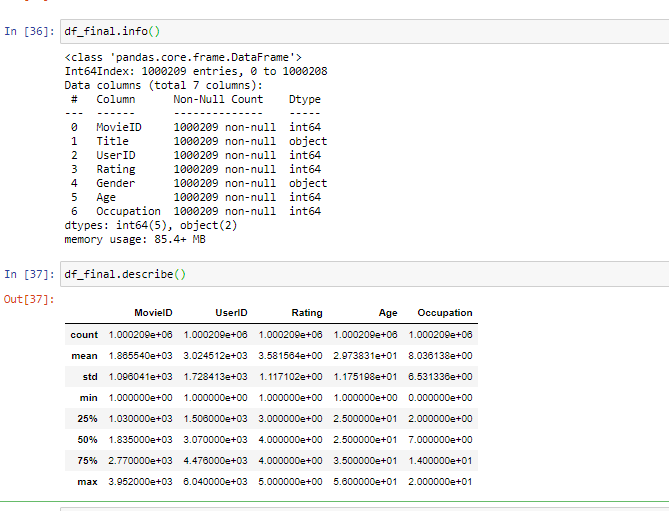


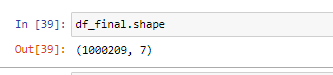


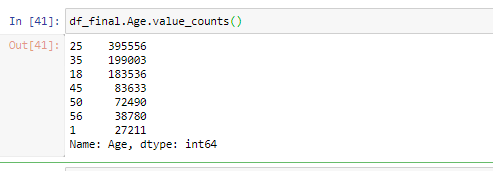
Objetive 4 )Explore the datasets using visual representations (graphs or tables), also include your comments on the following:

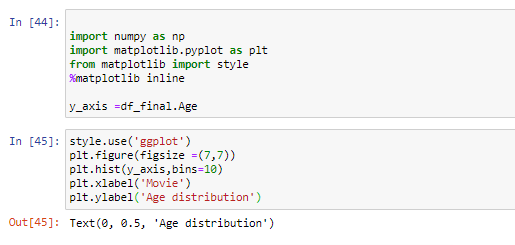
* User Age Distribution

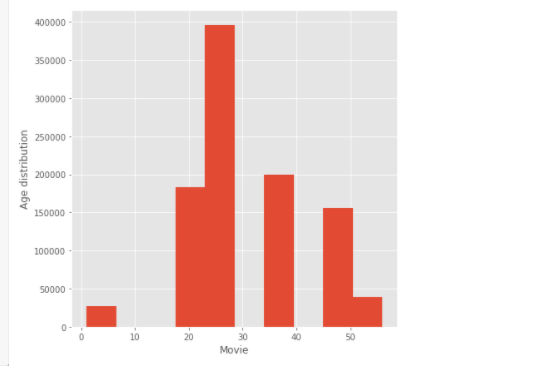


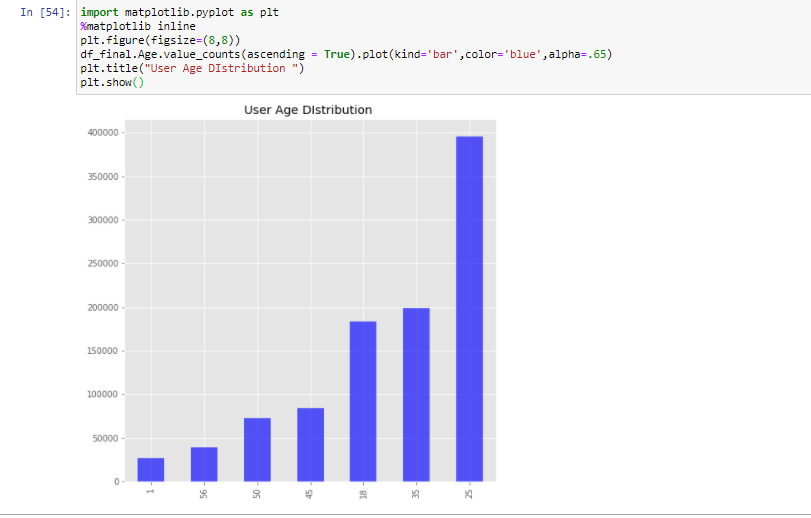






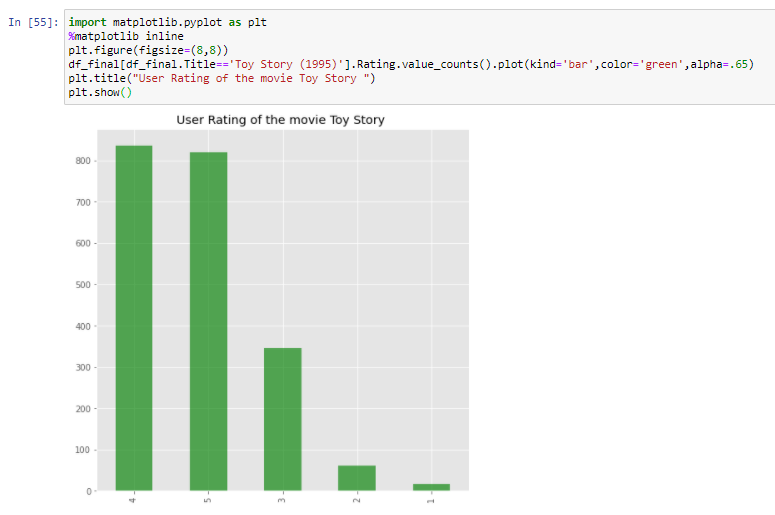






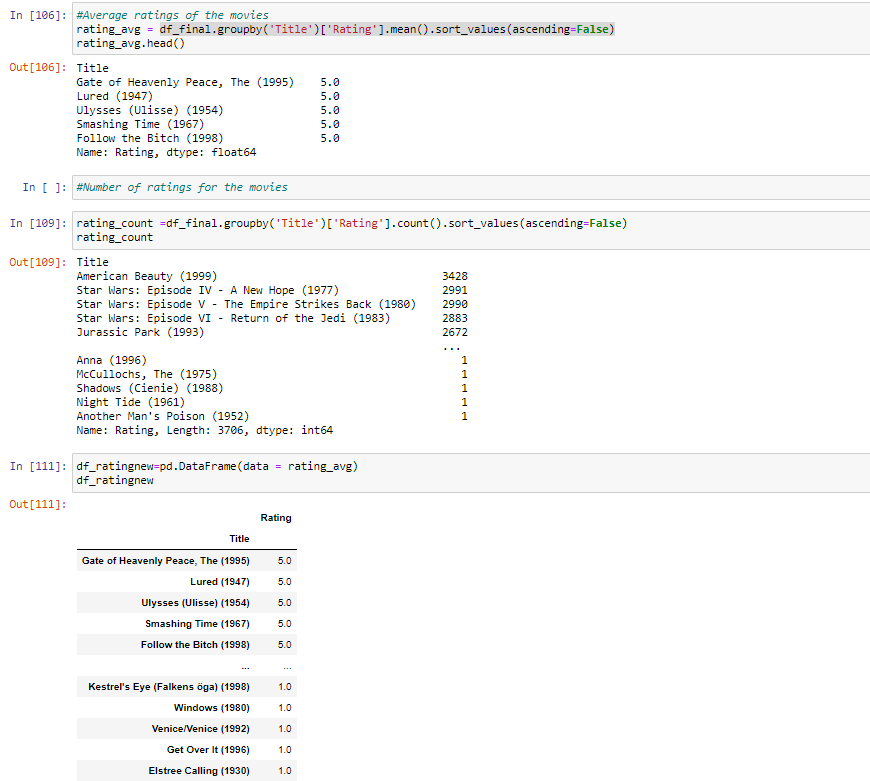
Age group 25-34 is having the highest number of movies and and age group under 18 having the lowest .

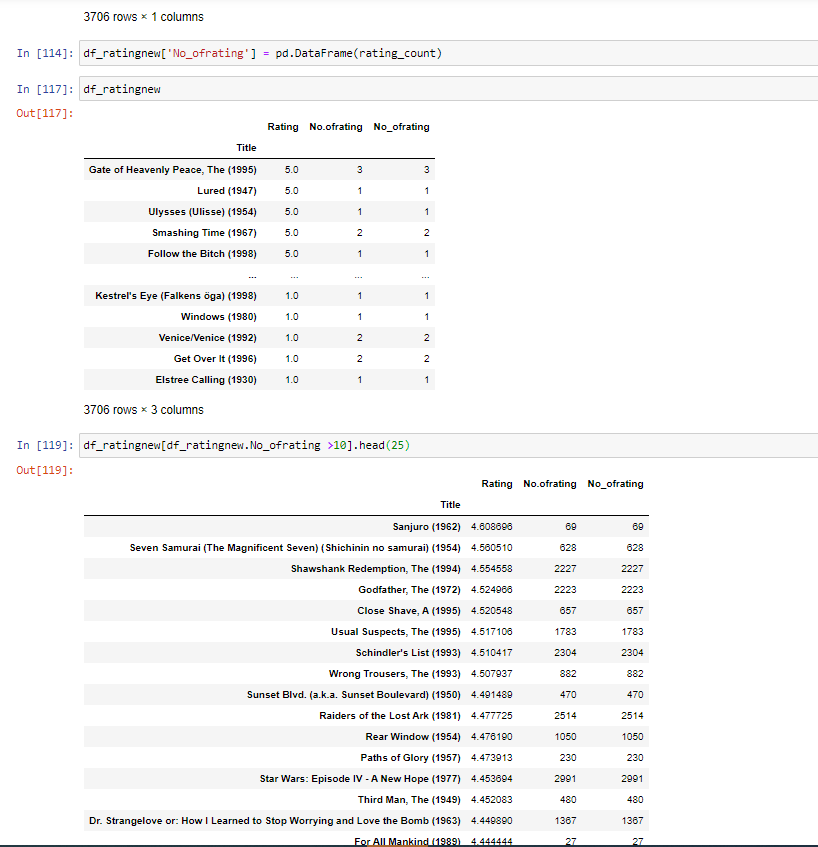
User rating of the movie “Toy Story”



Most of the viewers have rated the movie 4 closely followed by rating 5.

Top 25 movies by viewership rating



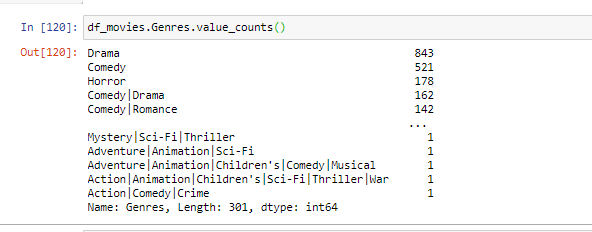


Find the ratings for all the movies reviewed by for a particular user of user id = 2696

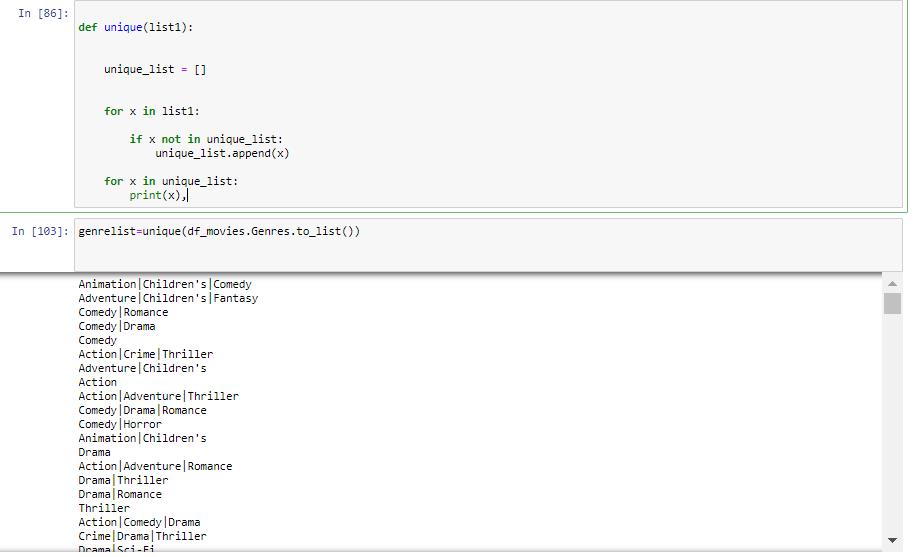


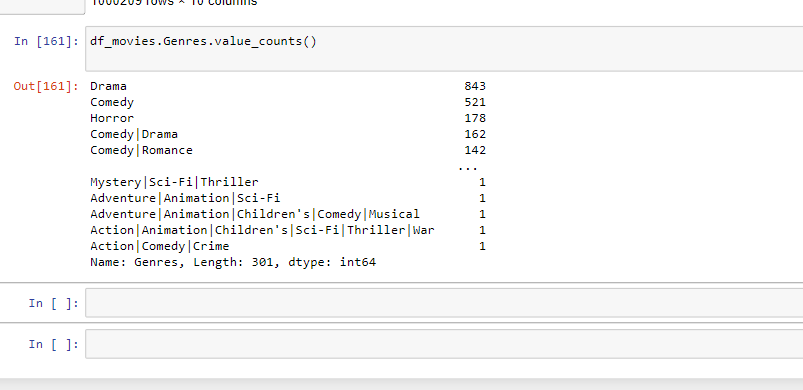
Feature Engineering:

Objective 4)Find out all the unique genres

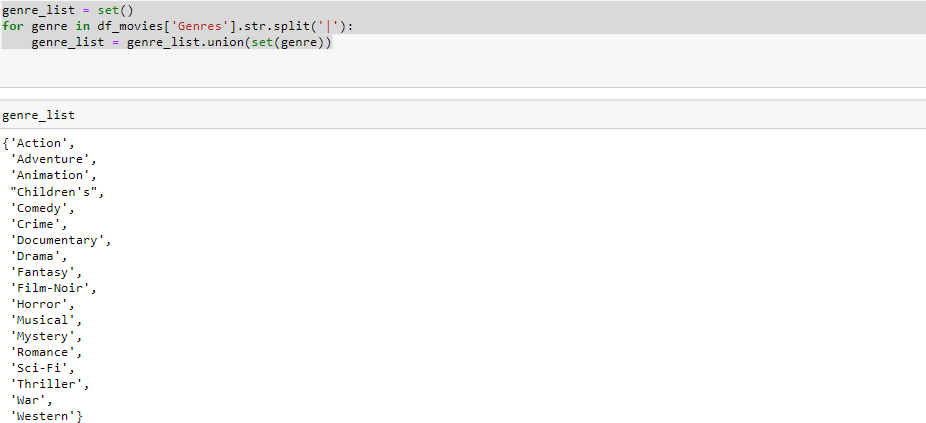


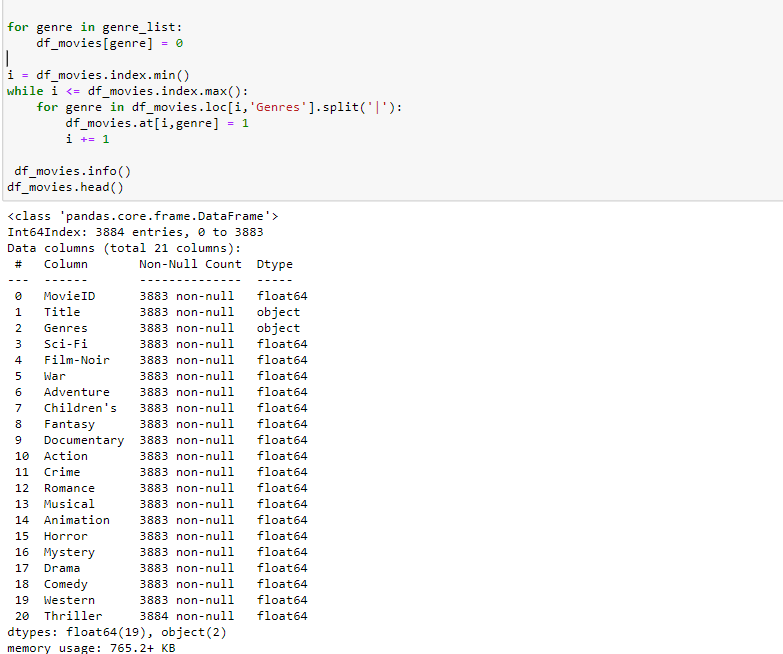


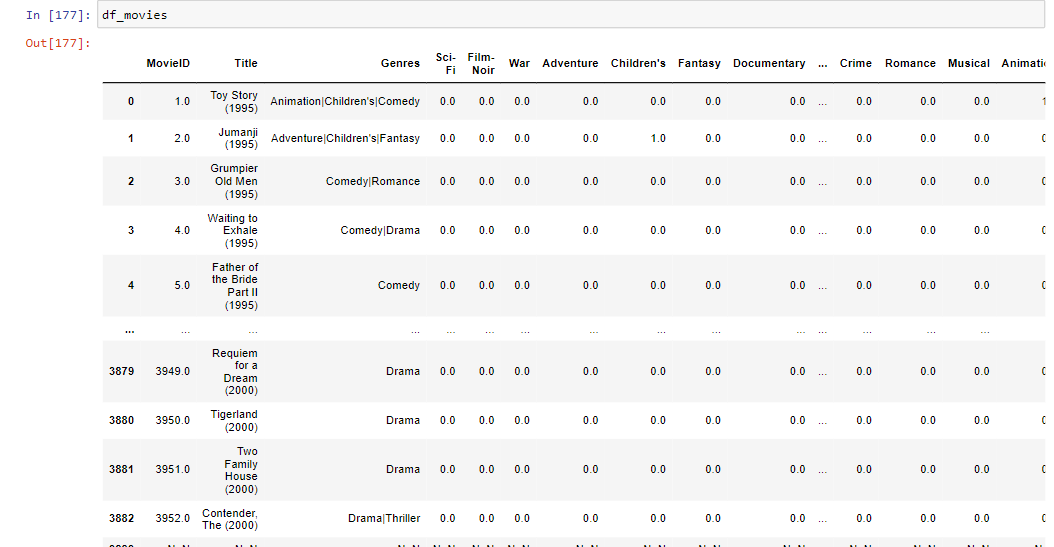




Objective 5)Create a separate column for each genre category with a one-hot encoding ( 1 and 0) whether or not the movie belongs to that genre.







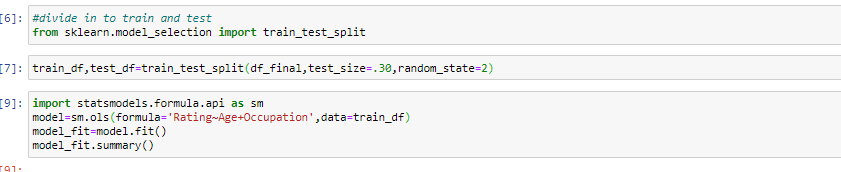
Objective 6)Determine the features affecting the ratings of any particular movie.

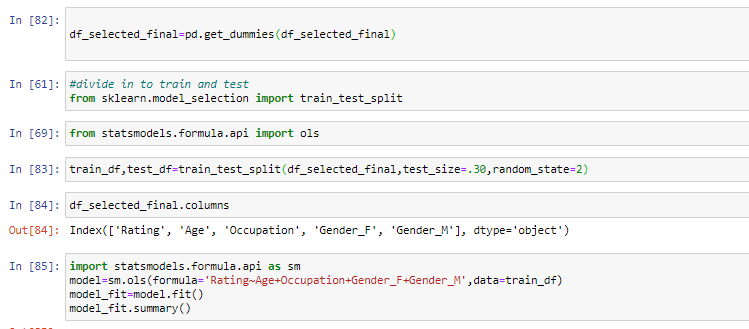
Develop an appropriate model to predict the movie ratings

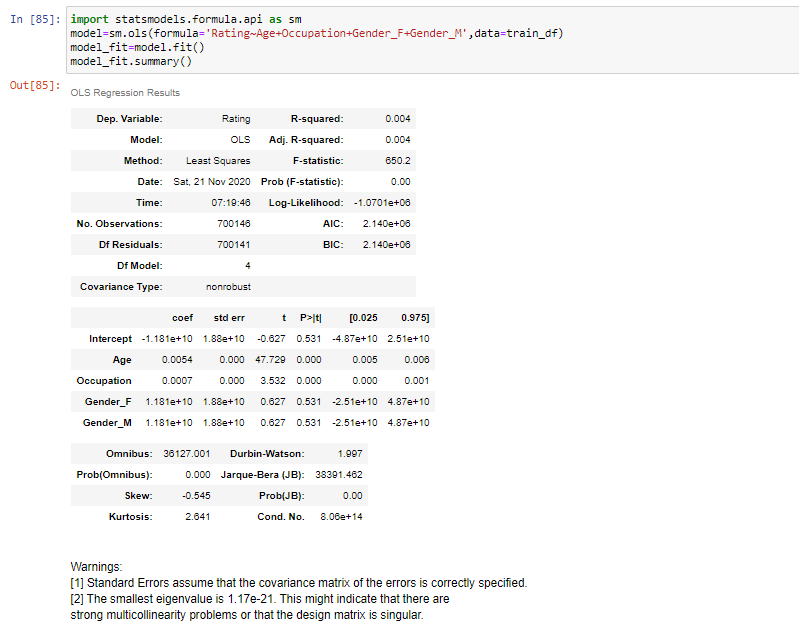
Dependent variable – Rating

Independent Variable – Age,Occupation,Gender



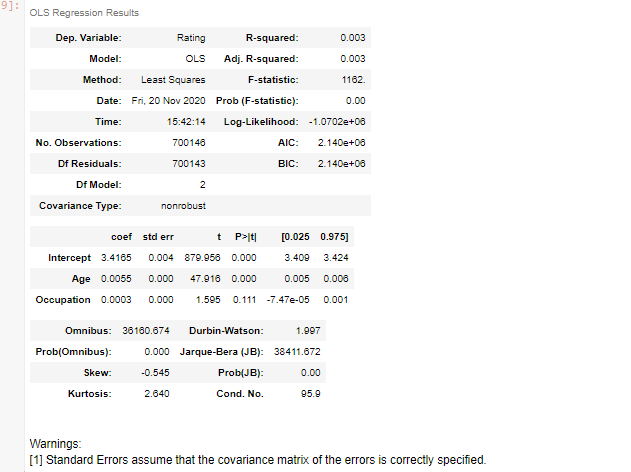




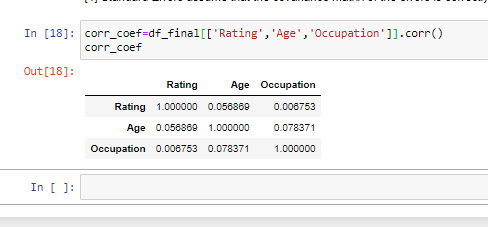


Since Gender is having high p value dropping variable Gender\_M and Gender\_F

Independent variable – Age and Occupation



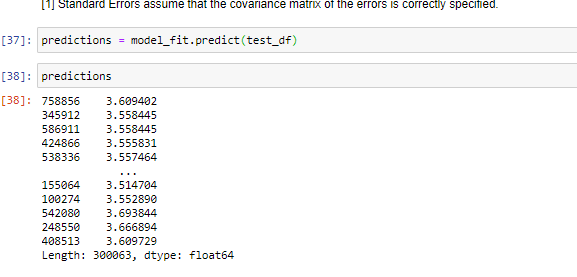
Rating is dependent on Age (Since the pvalue is low , Age is a significant variable) and not dependent on Occupation(Since p value is high , Occupation is an insignificant variable)



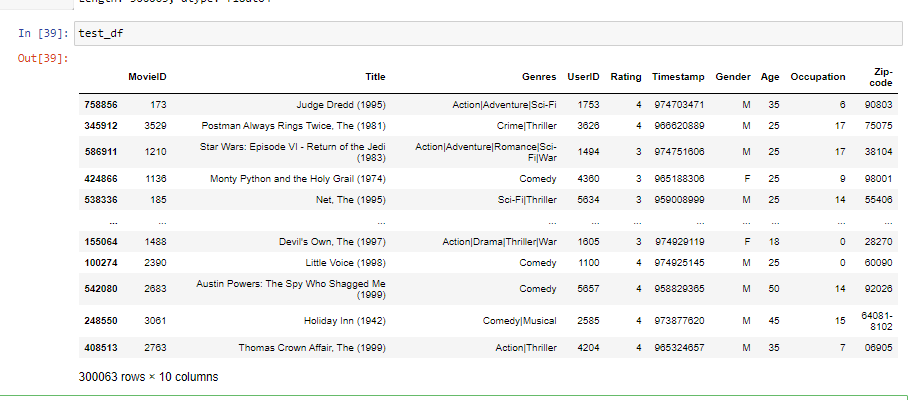
Rating and Age is having positive good correlation .

Rating and Occupation is having a positive very weak correlation.

# Predicted values

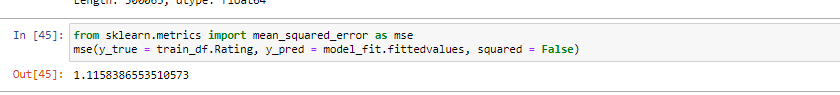


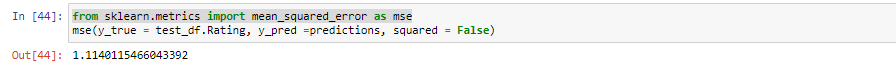
Actual values



|  |  |  |
| --- | --- | --- |
|  | Test | Predicted |
| 75856 | 4 | 3.60 |
| 345912 | 4 | 3.56 |
| 586911 | 3 | 3.56 |
| 424866 | 3 | 3.6 |

MSE Error





Result

Analysis Task

* Objective 1 - Import the three datasets

Three data sets were imported using pandas (Refer pages 5-6)

* Objective 2 - Create a new dataset [Master\_Data] with the following columns MovieID Title UserID Age Gender Occupation Rating

The three datasets were merged to created new dataset (Refer pages 6-7)

* Objective 3 - Explore the datasets using visual representations
* User Age Distribution
* User rating of the movie “Toy Story”
* Top 25 movies by viewership rating
* Find the ratings for all the movies reviewed by for a particular user of user id = 2696

Age group 25-34 is having the highest number of movies and and age group under 18 having the lowest .

Most of the viewers have rated the Toy Story movie 4 and 5.

Top 25 movies by viewership rating was analysed

Find the ratings for all the movies reviewed by for a particular user of user id = 2696(Refer pages 7-13)

* Objective 4-Find out all the unique genres

301 unique genres were found (Refer page 13-14)

* Objective 5-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.

Separate columns for each genre created(Refer page 15-16)

Objective 6 - Determine the features affecting the ratings of any particular movie.

Develop an appropriate model to predict the movie ratings

It was found that Rating is dependent on Age (Refer pages 17-21)