Abstract

In this report, our group of three have explored and discovered the patterns of movie audiences from MovieLens users who joined in the year 2000. Three datasets have been downloaded; movies.dat providing movies information; ratings.dat providing ratings information; and users.dat providing users information. We began by exploring and understanding these datasets. Then, we developed two research questions. The first one is “Does movie genre affect movie rating?”. Movie genres are regarded as reputational effects that assists users to rate the standard of a movie before watching. It guides their favourite movies and view in order with market offerings. Thus, making it a very possible relationship between movie genre and it’s movie rating by users. The second research question is “Do different user age groups have different preferred movie genres?”. There is a crucial distinction between users as the significance of age groups differ with their movie genre preferences. Therefore, we would like to discover the relationship between user age groups and their respective movie genre preference. For the research questions, new datasets had to be formed by combining variables from different datasets. Data validation, manipulating and cleaning was done by using statements such as “drop”, “and”, “where” and many more. The results that we found are staggering as Comedy and Drama seems to be receiving more number of higher ratings by users whilst Fantasy receives little to nothing.

Introduction

Frequently in time, watching movies has been known as an ‘escapism’ for people. While it is in fact that movies were made for entertainment purposes, to discard their ‘world-changing’ capability would be a mistake. Movies are able to have an immense influence than we ever dreamed as they can affect laws, culture, politics and primarily alter the course of history [1]. This report will be containing two research questions. The first will be “Does movie genre affect movie rating?”. Movie genres are important in movies because they build standards and expectations for viewers. When the viewer heads to a cinema to buy a ticket or sits down to enjoy a television movie, they choose for a different experience. In this report we are able to see which type of movie genre will affect the movie ratings. The following research question will be “Do different user age groups have different preferred movie genres?”. Users with younger age normally prefer family and animation movie genres while users at older age prefer action and comedy movie genres. In this report we are able to see which user age groups prefer watching which movie genre.

Background of Research Questions

1. Does movie genre affect movie rating?

   Movie genres are regarded as reputational effects that assists users to rate the standard of a movie before watching. It guides their favourite movies and view in order with market offerings [2]. Each movie is distinct in a way of having their own personalities. When a movie is categorized in a genre, it helps users to discover specific movies that they could enjoy watching. Numerous users prefer particular genres or two and would only watch those .Users might prefer to watch drama or horror films. They enjoy characters in a certain type of movie due to the proportion of horror or period of time setting. Users could prefer western movies for the reason that they desire they were living in the 19th century due to it being regarded as a simpler time [3].

2. Do different user age groups have different preferred movie genres?

   The crucial distinction between users is how the significance of age groups differ with their movie genre preferences [4]. In recent articles, it is proposed that users evaluate different movie genres at different ages due to the growth changes in emotional priorities. Young adults, respectively to middle-age and older adults were discovered to have more preference for watching movies to experience sadness and scaredness. They also have a great appeal to movies that provide scary, sad, dark and violent content. On the contrary, for younger adults respectively to late teens to early twenties, are more attracted to watching slapstick comedies. They are also more probable to mention that they watch movies to experience excitement, reduce boredom and to laugh [5].

Data Exploration

To be able to explore the data given, we first had to create a permanent library, “AES” under the pathing of “/home/u49646121/ASbar”. This can be seen in the *Code 1*. Then, we created three permanent datasets; “movies”, “ratings” and “users” under the permanent library, “AES”. This can be shown in *Code 2*, *Code 3*, and C*ode 4*. Next, we used code, “length” to specify the length of each category for each column. This is also shown in *Code 2, Code 3*, and *Code 4*. In order to have the datasets in the SAS Studio to read it, we uploaded the raw datasets that were in notepad form into the SAS Studio. From there, we used the “infile” statement with the pathing; “/home/u49646121/ASbar/movies.dat” for movies.dat file; “/home/u49646121/ASbar/ratings.dat” for ratings.dat file; and “/home/u49646121/ASbar/users.dat” for users.dat file to read the datasets. This is shown in *Code 2, Code 3*, and *Code 4*. In the raw datasets, the attributes were separated with the delimiter, “::”. Thus, we used the option, “dlm’::’ to separate the attributes to their respective columns which is shown in *Code 2, Code 3*, and *Code 4*. For only the movies dataset, we used the option, “dlmstr:’::’” as some movies had “:” to separate the movie title and sub movie title. What the “dlmstr” option does is that it enables us to specify a multi-character string as a delimiter so that it would not read “:” as a delimiter. Next, we used the “input” statement to input column names for our attributes respectively for each three datasets. For character columns, we put the “$” sign to set it as a character. For numerical columns, we do not put any sign to set it as a numerical.

Table I: Movies Dataset

	movieid	title	genres
1	1	Toy Story (1995)	Animation|Children's|Comedy
2	2	Jumanji (1995)	Adventure|Children's|Fantasy
3	3	Grumpier Old Men (1995)	Comedy|Romance
4	4	Waiting to Exhale (1995)	Comedy|Drama
5	5	Father of the Bride Part II (1995)	Comedy
6	6	Heat (1995)	Action|Crime|Thriller
7	7	Sabrina (1995)	Comedy|Romance
8	8	Tom and Huck (1995)	Adventure|Children's
9	9	Sudden Death (1995)	Action
10	10	GoldenEye (1995)	Action|Adventure|Thriller
11	11	American President, The (1995)	Comedy|Drama|Romance
12	12	Dracula: Dead and Loving It (1995)	Comedy|Horror
13	13	Balto (1995)	Animation|Children's
14	14	Nixon (1995)	Drama
15	15	Cutthroat Island (1995)	Action|Adventure|Romance
16	16	Casino (1995)	Drama|Thriller
17	17	Sense and Sensibility (1995)	Drama|Romance
18	18	Four Rooms (1995)	Thriller
19	19	Ace Ventura: When Nature Calls (1995)	Comedy
20	20	Money Train (1995)	Action
21	21	Get Shorty (1995)	Action|Comedy|Drama
22	22	Copycat (1995)	Crime|Drama|Thriller

Table I. shows the movies dataset which is made from the coding in *Code 2*. It shows three different column names; movieid, title and genres which provide 3,883 observations with single movie genres and combination of movie genres.

Table II: Ratings Dataset

	userid	movieid	ratings	timestamp
1	1	1193	5	978300760
2	1	661	3	978302109
3	1	914	3	978301968
4	1	3408	4	978300275
5	1	2355	5	978824291
6	1	1197	3	978302268
7	1	1287	5	978302039
8	1	2804	5	978300719
9	1	594	4	978302268
10	1	919	4	978301368
11	1	595	5	978824268
12	1	938	4	978301752
13	1	2398	4	978302281

Table II. shows the ratings dataset which is made from the coding in *Code 3*. It shows four different column names; userid; movieid; ratings; and timestamp which provide 1,000,209 observations.

Table III: Users Dataset

	userid	gender	age	occupation	zipcode
1	1	F	1	10	48067
2	2	M	56	16	70072
3	3	M	25	15	55117
4	4	M	45	7	2460
5	5	M	25	20	55455
6	6	F	50	9	55117
7	7	M	35	1	6810
8	8	M	25	12	11413
9	9	M	25	17	61614
10	10	F	35	1	95370
11	11	F	25	1	4093
12	12	M	25	12	32793
13	13	M	45	1	93304
14	14	M	35	0	60126
15	15	M	25	7	22903
16	16	F	35	0	20670
17	17	M	50	1	95350
18	18	F	18	3	95825

Table III. shows the users dataset which is made from *Code 4*. It shows five different column names: userid; gender; age; occupation; and zipcode which provide 6,040 observations.

Figure 1

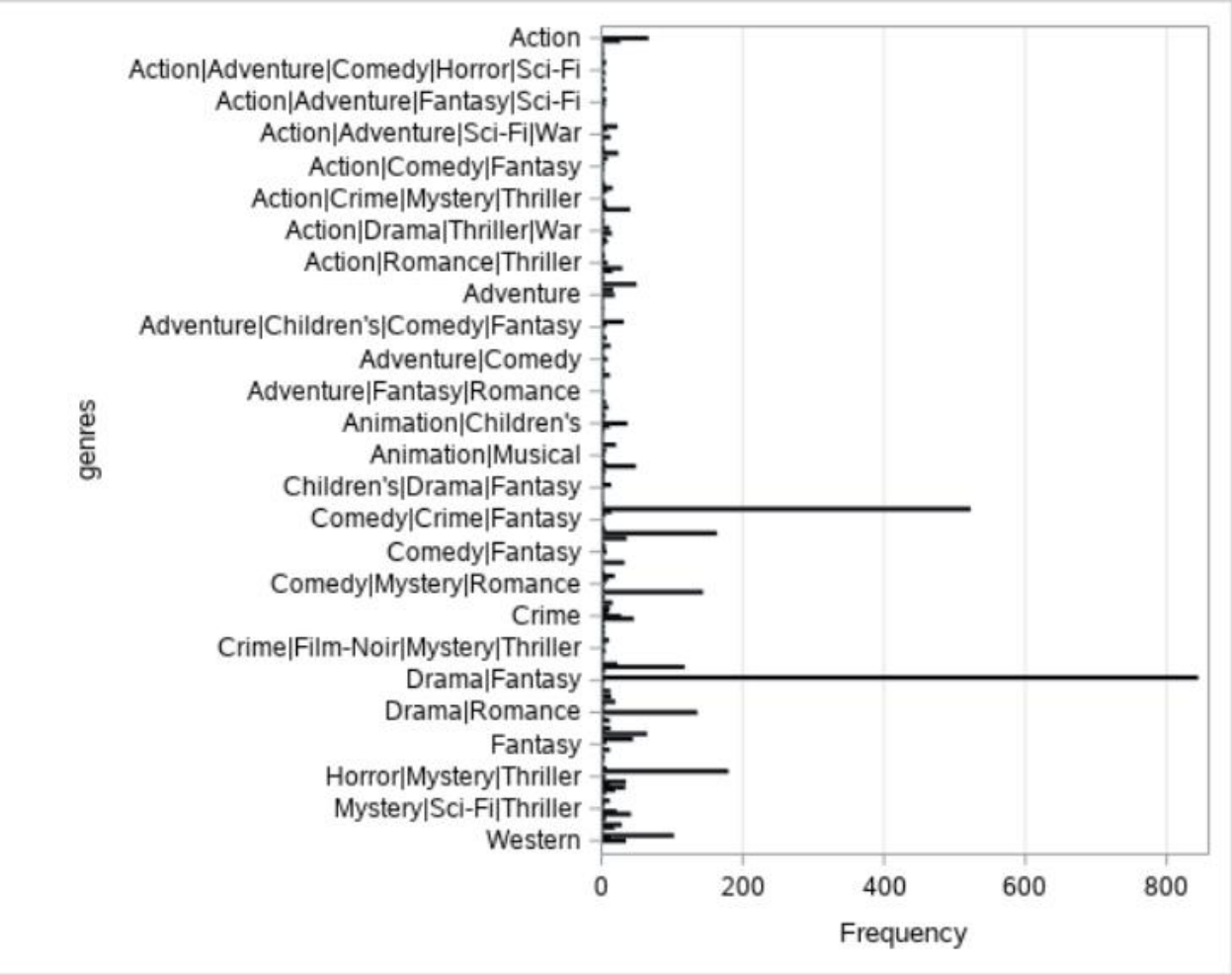

Fig. 1. is a basic plot derived from movie genres and its frequency. There are 26 different combinations of movie genres. The movie genre with the most frequency is Drama|Fantasy by far followed by Comedy|Drama|fantasy. Other than those two, the rest of the movie genres seem to far less in frequency compared to them as can be seen in the Fig. 1.

Figure 2

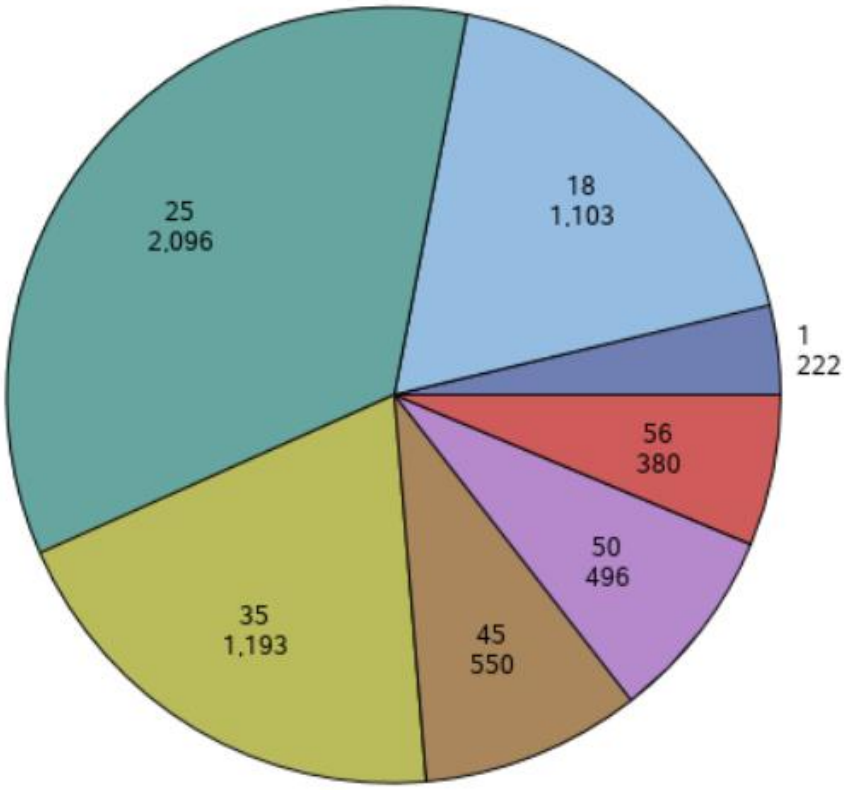

Fig. 2. is also a basic pie chart plot derived from users' age and its frequency. There are seven different age groups with the number on top representing the age group; 1 for age (under 18); 18 for age (18-24); 25 for age (25-34); 35 for age (35-44); 45 for age (45-49); 50 for age (50-55); and 56 for age (56+). The number below is the number of users that are within that age group. As can be seen in Fig. 2., age group 25 which is age (25-34) has the most frequency of 2,095 followed by age group 35 which is age (35-44) with frequency of 1,193.

Data Manipulation, Validation, Cleaning

For the first research question, a new data set had to be formed that consists of “movieID” and “genres” from movies.dat and “ratings” from ratings.dat. In order to combine these variables from different datasets into one without completely changing the structure of the original datasets, a new temporary dataset had to be created with just the necessary variables from each original datasets. In *Code 5* we can see two different temporary datasets being created which are indicated by the libref “work.” for each of the original dataset. The command “set” is being used to subset the original dataset into the temporary dataset. The command “drop” is used to omit unnecessary variables for the research question’s dataset. In this case the omitted variables for work.ratings is Timestamp and userid and for work.movies is title. The where statement used under the ratings dataset creation is used to manipulate the data output of work.ratings dataset and also as a tool of data cleaning. First the where statement filters the observations to the ratings that are only either “4” or “5”. This is done so that we can observe the amount of above average ratings each genre receives when the datasets are merged. The “and” statement refers to the further continuation of the “where” statement. The second expression is used to filter those observation with missing values in the ratings column.

Once these two temporary datasets are created, their quality has to be validated before being used for merging. Firstly, we can check for duplicate values present in the movies column in work.movies dataset. This is done because movieid is a unique value for each movie, having the same movieid for different movies defies the purpose of an unique identifier. In *Code 6*, the “proc freq” is used to get a frequency report on each variable present in work.movies.

The FREQ Procedure

movieid	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	1	0.03	1	0.03
2	1	0.03	2	0.05
3	1	0.03	3	0.08
4	1	0.03	4	0.10
5	1	0.03	5	0.13
6	1	0.03	6	0.15
7	1	0.03	7	0.18
8	1	0.03	8	0.21
9	1	0.03	9	0.23
10	1	0.03	10	0.26
11	1	0.03	11	0.28
12	1	0.03	12	0.31
13	1	0.03	13	0.33
14	1	0.03	14	0.36
15	1	0.03	15	0.39
16	1	0.03	16	0.41
17	1	0.03	17	0.44
18	1	0.03	18	0.46

Table iv

Table iv. Shows 18 of the 3,952 observations of frequency report on the movieid variable in work.movies.

When *Table iv*. was examined, no movieid was found to have a duplicate entry which validates the work.movies dataset. Secondly, we have to ensure that no missing values are present in the ratings column as it can make data noisy when merged. To do that, in *Code 6,* the “proc sort” statement is used to sort the data present in the work.ratings dataset in ascending order. The “out=” option prevents the dataset from being permanently sorted in that particular order by outputting the sorted data into a new temporary dataset created. In this case, the temporary dataset is called “work.sortedratings”. The “by” statement takes in the grouping variable which in this case is the ratings. Since the data has been sorted in ascending order and no missing values are present in the beginning of the dataset, we can ensure and validate that work.ratings has no missing data.

Table v

	movieid	ratings
1	3408	4
2	594	4
3	919	4
4	938	4
5	2398	4
6	2918	4
7	2791	4
8	2018	4
9	2797	4
10	1097	4

Table v shows 10 of 575,281 observations in the output of the ”proc sort” statement.

After both temporary datasets are validated, we can proceed on merging the two datasets. For that, both datasets must first be sorted by the same variable as shown in *code 7* . Then in a data step, a new dataset under the “AES” library will be created as shown in *code 8.* The “merge” statement is used to merge different datasets into one by grouping it with the same variable as we sorted both the datasets before merging. Finally, we can validate the overall structure of the new AES.ratings_genres dataset as shown in *Code 9* . This will produce a report that contains information about the dataset’s metadata,engine,etc as shown in Table vi

Table vi

The CONTENTS Procedure

Data Set Name	AES.RATINGS_GENRES	Observations	575631
Member Type	DATA	Variables	3
Engine	V9	Indexes	0
Created	11/19/2020 22:40:00	Observation Length	68
Last Modified	11/19/2020 22:40:00	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	YES
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Engine/Host Dependent Information

Data Set Page Size	131072
Number of Data Set Pages	300
First Data Page	1
Max Obs per Page	1923
Obs in First Data Page	1891
Number of Data Set Repairs	0
Filename	/home/u49646121/ASbar/ratings_genres.sas7bdat
Release Created	9.0401M6
Host Created	Linux
Inode Number	179308779
Access Permission	rw-r--r--
Owner Name	u49646121
File Size	38MB
File Size (bytes)	39452672

Alphabetic List of Variables and Attributes

#	Variable	Type	Len
2	genres	Char	60
1	movieid	Num	4
3	ratings	Char	1

Sort Information

Sortedby	ratings
Validated	YES
Character Set	ASCII

Table vi shows the contents procedure

For the second research question, a new data set had to be formed that consists of “movieID” and “genres” from movies.dat, “ratings” and “userid” from ratings.dat and “Userid” and “age” from users.dat. All these variables will be combined using similar techniques as the first research question’s dataset but with few changes. Firstly, new temporary datasets will be created for ratings.dat as “work.ratings2” and for users.dat as “work.users”. This is shown in Code 10 The “work.movies will be reused as nothing needed has changed. Both the new datasets will be created with added filters in the data step by using the where clause. This is to prevent missing data to occur in the userid column. Secondly, to create another temporary dataset similar to the Research Question 1 dataset but with the updated “work.ratings2” dataset as shown in code 11. Next, that new temporary dataset “work.ratings2_genres” and “work.users” are both sorted by userid as shown in Code 12. Finally creating the New Research question 2 datasets by merging “work.users” and “work.ratings2_genres” as shown in Code 13. Code 13 will also contain a where and if statement inside the merging program. This is to ensure further that no missing data are found.

Once the datasets are merged, it has to be validated. Firstly to validate the structure of the dataset, “proc contents” was used as shown in Code 14. The next two “proc sort”s are sorting the new dataset by userid (to check missing data in userid) and by ratings (to check missing data in ratings, movieid and genres). The reports were examined and no noisy data was found.

Result

For the first research question, “Does movie genre affect movie rating?”, we used *Code 15* to get the frequency report. The results show that the Comedy movie genre has received the most number of above average ratings of 62,293 followed by Drama with 72,695. These are the only two movie genres that surpassed every other movie genre by at least 50,000 above average ratings. The results also show that the Fantasy movie genre has the least number of above average ratings of 1 followed by Film-noir with 2. These are the only two movie genres that have less than 10 above average ratings.

For the second research question, “Do different user age groups have different preferred movie genres?”, we used *Code 16* to get the frequency results. The results show that for age group 1, the Comedy genre has received the most number of above average ratings of 2,089 while the least was Fantasy with 0. For age group 18, the movie genre with the most number of above average ratings is Comedy as well with 12,596 frequency while the least was Fantasy with 0. For age group 25, the Drama movie genre has the most above average rating with 27,233 while the least was Fantasy with 0. For age group 35, the Drama movie genre has the most number of above average ratings of 14,716 while the least was Fantasy with 0. For age group 45, the most number of above average ratings is Drama with 6,800 while the least is Fantasy with 0. For age group 50, the Drama movie genre as well has the most number of above average ratings with 6708 while the least is Fantasy with 0. Lastly, for age group 56, the most number of above average ratings is Drama with 4454 while the least is Fantasy with 0.

Conclusion

The two Research questions that have been asked in this research paper is “Does movie genre affects Rating?” and “Do different user age groups have different preferred movie genres?”. The results obtained for each research question were actually very logically predictable. For the first research question, it can be concluded that Comedy and Drama movie genres are the most widely and highly rated by most users with Comedy being the most while Fantasy and Film-noir is the least with fantasy being the most least. The first question’s results have said that certain genres have had more above average ratings than others. This clearly states that there will always be a genre that majority of people do not support and some that are liked by many. With this understanding, many cinema theater owners can predict the response of the audience based on the genre of movie which can help them in making decisions about the marketing of certain films.For the second research question, it can be concluded that Drama and Comedy as well is the most widely and highly rated by most users with Drama being the most while the least was Fantasy and Film-noir with fantasy being the most least. The second question’s results indicates that taste in movies do indeed evolve as a person’s age increases. The movies that a Child likes are not always liked by a middle aged person. The movies that an old man likes may not be liked by a teenager. Though these are known logic, the specificity of what type of movie that each age group prefers can actually be a surprise. In conclusion, movie genres Comedy and Drama will always seem to get high ratings by most users of all age groups while movie genre Fantasy and Film-noir will always be the least.

References

[1] K. O'Toole and K. O'Toole, "The Power of Cinema: 10 Films that Changed the World - Raindance", *Raindance*, 2020. [Online]. Available: https://www.raindance.org/the-power-of-cinema-10-films-that-changed-the-world/. [Accessed: 19- Nov- 2020]

[2] K. O'Toole and K. O'Toole, "The Power of Cinema: 10 Films that Changed the World - Raindance", *Raindance*, 2020. [Online]. Available: https://www.raindance.org/the-power-of-cinema-10-films-that-changed-the-world/. [Accessed: 19- Nov- 2020]

[3] J. Reich, "2. What Is Genre and How Is It Determined?", *Milnepublishing.geneseo.edu*, 2020. [Online]. Available: https://milnepublishing.geneseo.edu/exploring-movie-construction-and-production/chapter/2-what-is-genre-and-how-is-it-determined/#:~:text=Movies%20have%20their%20own%20personalities,watch%20movies%20in%20those%20genres. [Accessed: 19- Nov- 2020]

[4] N. Redfern, *Correspondence analysis of genre preferences in UK film audiences*. Participations, 2012 [Online]. Available: https://www.participations.org/Volume%209/Issue%202/4%20Redfern.pdf. [Accessed: 19- Nov- 2020]

[5] M. Mares and Y. Sun, *The Multiple Meanings of Age for Television Content Preferences*. Human Communication Research, 2020 [Online]. Available: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.387.2831&rep=rep1&type=pdf. [Accessed: 19- Nov- 2020]

Appendix

Code 1

```
Libname AES '/home/u49646121/ASbar';
```

Code 2

```
data AES.movies;
    length movieid 4 title $ 100 genres $ 60;
    infile '/home/u49646121/ASbar/movies.dat' DLMSTR='::' truncover;
    input movieid title $ genres $;
run;
```

Code 3

```
data AES.ratings;
    LENGTH userid movieid 4 ratings $ 1 timestamp $ 10;
    infile '/home/u49646121/ASbar/ratings.dat' DLM='::';
    input userid
          movieid
          ratings $
          Timestamp $ ;
Run;
```

Code 4

```
data AES.Users;
    infile '/home/u49646121/ASbar/users.dat' DLM= '::';
    input userid
          gender $
          age
          occupation
          zipcode;
run;
```

Code 5

```
*creating temporary ratings datasets with just the necesary variables;
data work.ratings;
    set AES.ratings(drop=Timestamp userid);
    where ratings in ('4','5') and ratings not eq .;
run;
*creating temporary movies datasets with just the necesary variables;
data work.movies;
    set AES.movies (drop=title);
run;
```

Code 6

```
*checking for duplicate in movieid;
proc freq data=movies;
run;
*Checking for missing value in ratings;
proc sort data= work.ratings out=work.sortedratings;
by ratings;
run;
```

Code 7

```
*Sorting both temporary datasets for merging;
proc sort data=ratings;
by movieid;
run;
proc sort data=movies;
by movieid;
run;
```

Code 8

```
*creating the Main dataset for Q1;
data AES.ratings_genres;
    merge movies ratings;
    by movieid;
run;
```

Code 9

```
*Validating structure;
proc contents data=AES.ratings_genres;
run;
```

Code 10

```
*Creating temporary users and ratings dataset with only the variables needed;
data work.users;
    set aes.users(drop= gender occupation zipcode);
    where userid not eq .;
run;

data work.ratings2;
    set aes.ratings(drop= timestamp);
    where userid not eq . and ratings in ('4','5') and movieid not eq .;
run;
```

Code 11

```
*Sorting new ratings dataset by movieid for merging it with movie datasets;
proc sort data=work.ratings2;
by movieid;
run;
*Merging new ratings dataset with movies dataset used in Q1 dataset;
data work.ratings2_genres;
    merge ratings2 movies;
    where movieid not eq .;
    by movieid;
run;
```

Code 12

```
*Sorting both user and new ratings_genre dataset by userid;
proc sort data=work.users;
by userid;
run;
proc sort data=work.ratings2_genres;
by userid;
run;
```

Code 13

```
*Creating Q2 dataset;
data work.age_genre_relationship;
    merge users ratings2_genres;
    where userid not eq .;
    if movieid eq . then delete;
    by userid;
run;
```

Code 14

```
proc contents data=work.age_genre_relationship;
run;
proc sort data=work.age_genre_relationship;
by userid;
run;
proc sort data=work.age_genre_relationship out=work.sortage_genreby_ratings;
by ratings;
run;
proc freq data=work.age_genre_relationship;
tables Age ratings genres;
run;
```

Code 15

```
proc freq data=AES.ratings_genres;
tables genres*ratings;
run;
```

Code 16

```
proc freq data=work.age_genre_relationship;
tables Genres*Age;
run;
```