SCHOOL OF SCIENCE AND TECHNOLOGY

ASSIGNMENT FOR THE
BSC (HONS) IS; YEAR 2
BSC (HONS) IS (BUSINESS ANALYTICS); YEAR 2
BSC (HONS) IS (DATA ANALYTICS); YEAR 2

ACADEMIC SESSION AUGUST 2020
IST2034: ANALYTICS ENGINEERING

DEADLINE: Group Report - Week 13 (20 Nov, Fri, 5pm)

STUDENT NAME:	**SHAMALAN RAJESVARAN**	**STUDENT ID: 18042945**
STUDENT NAME:	**AHMED MOHAMMED GHANEM**	**STUDENT ID: 11035201**
STUDENT NAME:	**NG WEI XIANG**	**STUDENT ID: 18033167**

INSTRUCTIONS TO CANDIDATES

- This assignment will contribute 30% to your final grade.

IMPORTANT

The University requires students to adhere to submission deadlines for any form of assessment. Penalties are applied in relation to unauthorized late submission of work.

- Coursework submitted after the deadline but within 1 week will be accepted for a maximum mark of 40%.
- Work handed in following the extension of 1 week after the original deadline will be regarded as a non-submission and marked zero.

Lecturer's Remark (Use additional sheet if required)

We …... (Name) ………..…………………........std. ID received the assignment and read the comments ....……………………..……… (Signature/date)

Academic Honesty Acknowledgement

SHAMALAN RAJESVARAN	**18042945**
AHMED MOHAMMED GHANEM	**11035201**
NG WEI XIANG	**18033167**

"We ……….. (student name) verify that this paper contains entirely our own work. We have not consulted with any outside person or materials other than what was specified (an interviewee, for example) in the assignment or the syllabus requirements. Further, We have not copied or inadvertently copied ideas, sentences, or paragraphs from another student. We realize the penalties *(refer to page 16, 5.5, Appendix 2, page 44 of the student handbook diploma and undergraduate programme)* for any kind of copying or collaboration on any assignment."

Xiang Ahmed **[20 NOV 2020]**

…………………………………..... (Student's signature / Date) ……………………………….....

<u>Report: 30% contribute to final</u>

Assessment Criteria:

- Research questions derived from preliminary data exploration : 5%
- Data validation, cleaning and manipulation to address the issues : 10%
- Output and discussion of the finding : 10%
- Programs with internal documentation : 5%

<u>Evaluation Rubric:</u>

Assessment criteria	Weight	10-9	8-7	6-4	3-1	0	Mark Awarded
Research questions derived from preliminary data exploration	5%	Creating excellent research questions and support with preliminary data exploration techniques.	Creating good research questions and support with preliminary data exploration techniques.	Creating weak research questions and acceptable support preliminary data exploration techniques.	Creating weak research questions and minimum support with preliminary data exploration techniques.	No assignment submitted.	
Data validation, cleaning and manipulation to address the issues	10%	Excellent executed with no errors.	Well executed with few errors.	Poorly executed with many errors.	Very poorly executed and very difficult to read.	No assignment submitted.	
Output and discussion of the finding	10%	Addressed all of the research questions with additional components. Clearly illustrates the idea and well thought out the response.	Addressed all of the research questions. It is relatively detailed to illustrate the idea and relatively well thought out the response.	Addressed many of the research questions. Not detailed and poorly thought out the response.	Not address all research questions. No evidence of having given the assignment real thought or submitted late.	No assignment submitted.	
Programs with internal documentation	5%	Excellent executed with no errors.	Well executed with few errors.	Poorly executed with many errors.	Very poorly executed and very difficult to read.	No assignment submitted.	
Remark:						**Total**	/ 30%

MovieLens-1M: Analysis of Movie Audience's Patterns

Shamalan Rajesvaran
18042945

Ahmed Mohammed Ghanem
11035201

Ng Wei Xiang
18033167

Abstract: The vast amount of data collected by GroupLens Research has made it possible for deep analysis to be conducted with regards to the trends amongst the movie audience. Many studies have been conducted to determine the growth of movie ratings and the factors influencing it. The objective of this analysis is to empirically investigate the impact of age groups, gender, and genre on movie ratings. Throughout the analysis, SAS Programming and Python programming was utilised to generate the findings of MovieLens dataset. Different functions and instructions were implemented to generate the required results of the research. Detailed data analysis and visualization are employed to best enhance the comprehension of the datasets. The analysis began at the preprocessing stage followed by focused analysis to fulfil 2 Research Questions. Results indicate that a pattern of high rated movies are discovered in which we will discuss in detail in the report. All in all, we used different methods to examine our research questions which express how movie ratings are impacted by gender, users' age, and movie genres. We were able to ascertain the correlation between the different variables in the MovieLens datasets.

INTRODUCTION

Movie rating is known as a classification system designed to classify films based on their suitability for the audience. It began as a necessary and strict procedure, and over time has morphed into something more complicated and secretive, while allowing movie-makers to create what they want [3]. Despite the numerous studies investigating the dimensions of influencing movie ratings, there are still some notable gaps studying age groups and their relationship with movie ratings. The three datasets, Users, Ratings and Movies are interconnected. Hence, this report aims to bring to light any noticeable patterns and relationships to best understand the movie audiences.

The significance of the relationship between various age and movie ratings and users' preferences are based on their age and emotions [2]. Although the study was based on age groups, it had insufficient demonstration about which precise age group contributed heavily to ratings.

Therefore, the first question of this analysis can be summarized as the Correlation between movie rating and users' age range. *"Are movie ratings influenced by age groups?"*. The analysis was further extended to identify the popular genre based on the specific age groups. Considering users' age groups is a dominant factor in determining audience preferences, it is important to comprehend which genre has a positive/higher relationship with users' preferences. Movie genre preference cannot always result in better performance; their preference might shift to more comprehensible movies. Thus, the second research question of this analysis is to list the movies that fall above the upper quartile range of the variable ratings. From there we will further analyse the relationship between the said movies and the gender of the users who rated them as well as the year of release.

Using Movielens-1M dataset, this report will extensively discuss the relationship between movie ratings and age groups, and which movies genre is the most popular among the chosen age groups. The MovieLens-1M dataset can be obtained from Appendix G.

DATA EXPLORATION, CLEANING AND MANIPULATION

A. Data Exploration

During the data exploration process, we focused on gaining familiarity with the datasets provided through thorough analysis and data visualizations. We took the effort to detail out the characteristics of the data and understand the general structure. A major part of our analysis included the use of Python Programming with certain elements utilising SAS Programming. The code for this section is entailed in **Appendix A**. We used the Pandas which is a popular python library for data manipulation. Throughout this process, we utilized Python functions to understand and analyze our datasets further. The Pandas data types that we have identified are objects and int64, which Python type equivalents are strings and integers respectively. In this assignment, we were given a set of datasets from the MovieLens user who joined in the year 2000 and the set consists of 3 datasets which were "**movies**", "**ratings**", and "**users**" dataset.

The "**movies**" dataset consists of 3 variables which are MovieID, Title, and Genres. We used the "COUNT()" function in python to count the number of rows for the dataset. Based on our analysis, we were able to find out that there are 3883 rows. We then used the "INFO()" function to get a general idea on what the dataset is all about. Through this process, we were able to find out the variables, data type as well as memory usage. The Pandas Data Types in this dataset are 1 int64 data type (MovieID), and 2 object data types (Title, Genre). We also ran the "HEAD()" and "TAIL()" function to get a brief idea of the content within the dataset. The genres of the movies are combined as a single variable and pipe-separated. To conduct our research questions, we performed some manipulation and managed to successfully separate it individually. We then identified that there are 18 different genres. We performed two data visualization to best understand the "movies" dataset. Firstly, we generated a word cloud in Python to identify the genre with the highest frequency in the dataset.

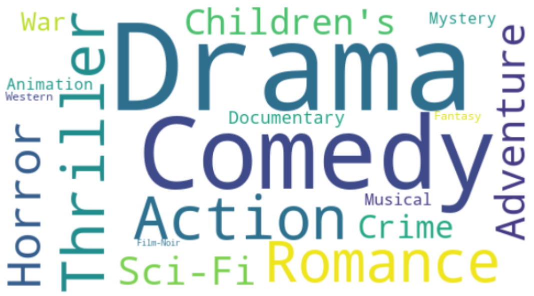


Figure 1. Wordcloud representation of the genres.

Based on the word cloud, it is clear that the movie genre drama and comedy has the highest frequency in this dataset. We defined a function to calculate the number of times each genre appears. Based on the output, we found out that the Drama genre has appeared a total of 1603 times while the Comedy genre has appeared a total of 1200 times. Hence, the word cloud is shown to be an accurate representation of the movie genre representation in the "movies" dataset.

Next, we ran an analysis to obtain a general idea about the distribution of movies per year. We generated a graph on SAS Studio utilising the "PROC CHART" function. This allowed us to plot the graph of the frequency of the movies released against the year of the movies' release. Based on the graph generated, we noticed that a vast majority of the movies in the "movies" dataset is from the year 1994 to 1997 with a total of 1259 movies. The second highest group of movies in the "movies" dataset spans from the year 1998 to 2000 with a total of 776 movies.

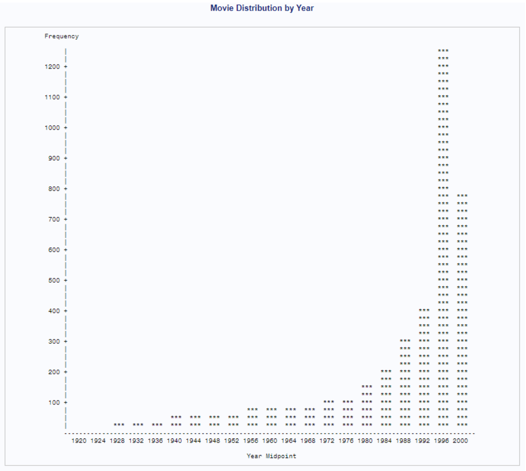


Figure 2. Movie Distribution by Year

The "**ratings**" dataset consists of 4 variables which are User ID, Movie ID, Rating, and Timestamp. Similarly, we used the "COUNT()" function and was able to find out that there are a total of 1000208 rows. The ratings are scaled from 1 to 5 whole star ratings and each had a timestamp on it. We further analysed the variable ratings in the "ratings" dataset using the "MIN()" and "MAX()" function to identify the minimum and maximum value values. We identified that the minimum rating given by a user is 1 while the maximum rating given by a user is 5. The "INFO()" function also indicated that all 4 variables are of the int64 Panda data type. We also ran the "HEAD()" and "TAIL()" function to get a brief idea about the general structure of the dataset.

Lastly, the "**users**" dataset consists of 5 variables which are User ID, Gender, Age, Occupation, and Zip-Code. Similarly, we used the "COUNT()" function in python to identify the tidiness of the dataset and the result once again showed that it is all good to go. There are a total of 6040 rows. The gender in the dataset is represented with "M" for male and "F" for female. Age is categorized into 7 categories which are "1", "18", "25", "35", "45", "50", and "56". The occupation of the users was also categorized and rephrased to digits from 0 to 20 for easier analysis. The detailed categorization criteria can be viewed from appendix A. We ran the "INFO()" function which indicated that there was

3 int64 Panda data type (UserID, Age and Occupation) and 2 object Panda data type (Gender and Zip-Code).

We conducted further analysis into the “user” dataset. We performed data visualization to obtain the user age distribution. We utilised the imported library matplotlib to best visualize our required data into a bar graph. We then defined “xlabel” and “ylabel”. The graph is shown in Figure 3.

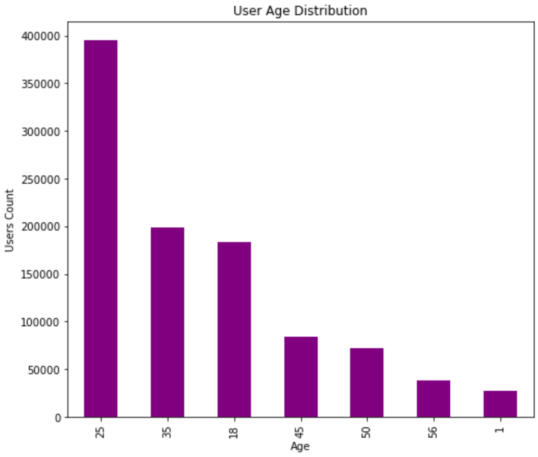


Figure 3. User Age Distribution.

Based on Figure 3, the plot shows that user in the age group “25” has the highest user count at 395556 of them. This is followed by user age group “35” at 199003, the age group “18” at 83633, the age group “45” at 72490, the age group “50” at 72490, the age group “56” at 38780 and lastly age group “1” at the lowest with 27211 of them.

We then focused on obtaining a sample movie and its overall ratings from the users. For this purpose, we randomly selected a movie and chose Jumanji (1995). We had to use the “GROUPBY()” function to obtain the necessary information for our graph. Additionally, we also had to specify “SIZE()” to obtain the total number of ratings for each respective levels. Figure 4 shows the resulting output.

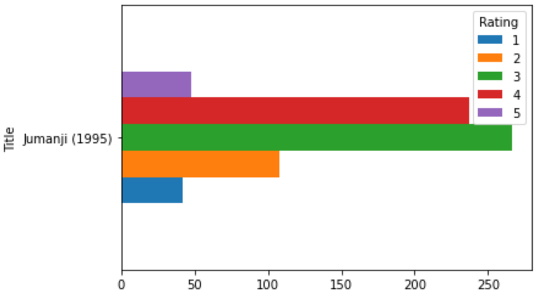


Figure 4. Total rating by Users for Jumanji (1995)

For the graph, it is seen that the most number of ratings given out by users was a rating of 3.

Finally, we were also interested in identifying the movies which have been rated the most out of the users. This would subsequently indicate the most popular movie out of the most famous movies among others in this dataset by counting the numbers of ratings. Similarly, we performed the “GROUPBY()” and “SIZE()” function in python to plot out a bar chart. We have chosen to output the 10 most popular movies by indicating beside the ‘GROUPBY()” function “[:10]”. Based on the figure, it is seen that “American Beauty(1999)” is the most popular movie with 3428 number of users rating the movie.

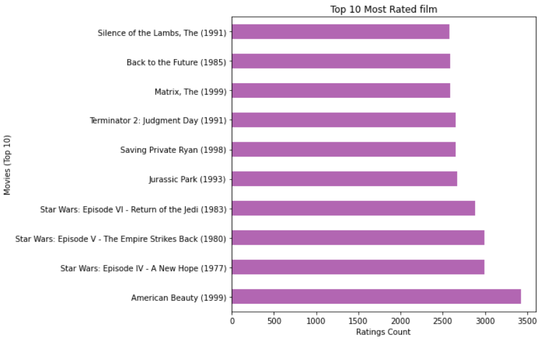


Figure 5. Top 10 most rated film

Lastly, we wanted to find out as much as we can from a UserID. Hence, the random UserID that we picked was UserID 4827. We wanted to find the ratings of him/her on the movies viewed. We matched the UserID to the database and obtained information that the User had reviewed 72 movies in total which is much lesser than the average of 165.

B. Data Cleaning

In data cleaning, our team had performed various techniques in cleaning and tidying the dataset. The code for this section is entailed in Appendix B.

. Firstly, the dataset is read into SAS using “dlmstr=’::’” because the data are separated from each other by a string of delimiter “::”. Using “dlmstr=’::’” would allow us to read the data file without any errors. Encoding of “wlatin1” was also used to be able to read – (dashes) and other special characters for the movie name. Next, while investigating the movies file we found out that the MovieID are not consecutive values which there is missing of MovieID no 91. However, the purpose of MovieID is to assign a special value to it which it does not need to be consecutive as long as there are no duplicates in the movie ID. Thus, we use the “COUNT()” and “ISNULL()” function in python to find the number of rows and check whether the dataset has null values for all 3 datasets. The result in python showed 0 which means that our dataset is clean and does not carry null values. Additionally, we also ran the “DUPLICATED()” function to identify if there are any duplicated rows. Once again, we did not encounter any duplicated row. Till this point, we did not manage to figure out any flaws within the dataset that might affect our analysis and it is now cleaned and ready for manipulation.

C. Data Manipulation

Upon completing the data cleaning process, we moved on to the data manipulation segment of our analysis. The purpose of the data manipulation process is to prepare the datasets before further analysis is conducted to fulfil the research questions. The code for this section is entailed in **Appendix C.**

After visualizing and understanding the content of Movielens datasets, we used SAS Studio to manipulate the data to make it organized and easier to analyze. We had 2 methods to the data manipulation process. Firstly, we merged the Ratings, Users and Movies dataset to create a Master dataset. We then followed it with further analysis of the dataset to improve readability and analysis.

During the merging phase, we combined both Ratings and Users datasets into one dataset called “UserRatings”. The method used in merging each dataset was MERGE SAS function. The common key between the Ratings and Users dataset is UserID. To ensure a successful merge sequence, we first had to sort the Ratings dataset with the common key, UserID. During the sorting process, we used ‘PROC SORT’ followed by the ‘BY’ function. We then applied the function MERGE to combine both datasets by the common variable “UserID”.

Next, we started the process to combine both the “UserRatings” and “Movies” dataset into one dataset called “MoviesCombined”. Once again, we had to sort the “UserRatings’ dataset with its common key with the ‘Movies” dataset. The common key, in this case, was MovieID. We used “PROC SORT” followed by the ‘BY’ function. Finally, we applied the function MERGE to combine both datasets by the common variable “MovieID”. This concluded the first phase of our data manipulation.

The next step was to conduct further analysis of the master dataset. Firstly, our focus was to split the genre into separate columns to improve readability. Furthermore, we also aim to extract the year of the films from the Movie name and give it a designated column to allow further analysis which is shown in Appendix C.

In order to split the genre, we used the ‘ARRAY’ as well as the ‘SCAN’ function. The ‘SCAN’ and ‘ARRAY’ function allowed us to parse out each genre from the character string into separate variables. Since the genre values are delimited between each other by ‘|’, the ‘SCAN’ function will split the respective genres once the system encountered a ‘|’. Additionally, the ‘ARRAY’ function will work hand in hand in assigning the genres that are being split into its separate column (titled G1 – G5).

Next, to extract the year value from the movie name, we used the ‘SCAN’ function as well as the ‘INPUT’ function. The year can be identified as the last word of the Movie name (ie. Toy Story (1995)). Hence, to return the last character we specified for the ‘SCAN’ function to read the ‘-1’ character. Hence, SAS would scan the last character and place it into a new variable named ‘Year’. However, the year value that has been extracted was in the character data type. We required for the year value to be a numeric variable so that we can make a meaningful analysis. Hence, we used the ‘INPUT’ function with an informat of ‘.8’ to output to a new variable ‘YearNum’. This would mean that SAS would convert it to a numeric form. We then use the ‘DROP’ function to remove the Year in character format and ‘RENAME’ function to rename the ‘YearNum’ to ‘Year’.

From this data manipulation, we have prepared our datasets sufficiently so that we can fulfil our Research Questions.

Our first research question “The correlation between movie rating and age groups”, our goal was to find how movie ratings are affected by age groups and which group is more interested in rating movies. Therefore, we were required to combine both Ratings and Users datasets into one dataset called “UserRatings”. Furthermore, we would also be analyzing the popular genre based on age category. Hence, we would require the genres to be split up and assigned to its variable.

Next, our second research question is “List of movies that fall under the upper quartile of the ratings (4.0)”. From there we are analyzing further between the movies with a rating higher than 4.0 with gender and year. Hence, we would require the master dataset as well as the year value to be extracted from the movie name.

At the end of the data manipulation process, we have successfully merged all the 3 datasets and performed further analysis for greater understanding. The master dataset is prepared and ready for the research questions’ analysis.

Results and Discussion

I. Research Question 1

As a result of the initial measures of the combination of exploration and manipulation of the datasets. We have organized datasets that are accessible for analysis. Our first research question "**Correlation between movie ratings and age groups**", discusses the relationship between high rating scores of movies for different users' age groups. The code for this section is entailed in **Appendix D**. Based on the findings of the correlation coefficient, we constructed further analysis "ANOVA" to compare how the classified variable "Age" is influencing the fluctuation of movie ratings, and which age group participates more in rating movies.

We plotted a boxplot with the idea of identifying the distribution of rating in each age group. As shown in Figure 6, it is seen that all the age groups have similar ratings provided at a mean of around 3.5. However, an interesting pattern was identified based on Figure 6. It is seen that people in the age group of 1, 18, 25, 35, and 45 provide ratings of 5 at most and 2 at least. Age group 50 and 56 provides ratings at the minimum rating value of 1. This shows that users at younger ages tend to not provide more extreme negative feedback. The mean rating for each age group is somewhat similar amongst all the 7 different age groups.

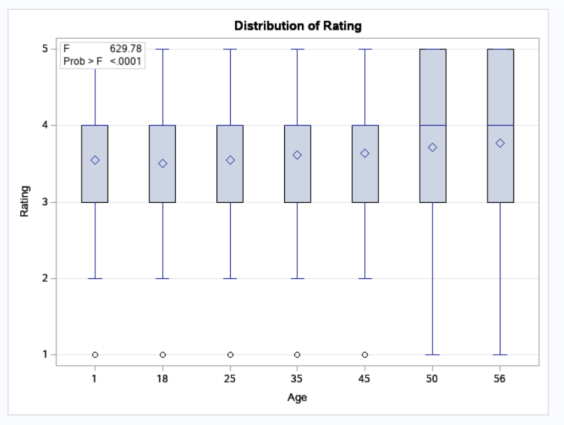


Figure 6. Distribution of ratings among each age group

Then, we constructed a Bar Chart as shown in Figure 7 to identify the frequency of age groups participated in rating movies. As shown, the most prominent information we can obtain is that within the age group 25, they have given a majority of the movies that they have rated a rating of 4. This is followed by ratings 3, 5, 2 and 1. After careful inspection, we have identified that the remaining age groups (age group 1, 18, 35, 45, 50 and 56) exhibit similar patterns to that of the age group18. That is that the ratings given in descending order are ratings 4, 3, 5, 2 and 1.

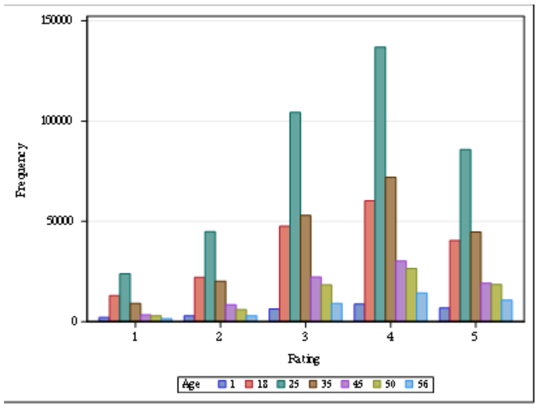
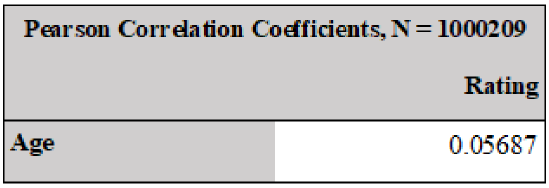


Figure 7. Rating Breakdown based on Age.

Hence, from Figure 6 and Figure 7 we can identify the range of ratings that each of the users of the specific age groups provides as well the frequency distribution. Based on Figure 7, it is obvious that Age Group 25 has the highest frequency of ratings. This is because they have the largest number of users in this MovieLens data with 395556 users.

Based on that, we used correlation analysis "PROC CORR" to identify the strength of the relationship between the two variables. Figure 8 shows that the correlation coefficient between movie rating and age group is 0.05687. Since the value of Pearson Correlation Coefficients is close to 0, this indicates that there is no linear relationship between the variables [4].



Pearson Correlation Coefficients, N = 1000209

	Rating
Age	0.05687

Figure 8. Pearson Correlation Coefficients

Appertaining to our first results suggesting that most users who rate movies are in the age group of 25, we wanted to conduct more in-depth analysis on which genre is more popular among this age group. Accordingly, we have generated a few other graphs to determine the precise movie genre users are interested to watch. Figure 10 illustrates the favourite genres of each age group counted by the genre that has the highest frequency of votes in their age group. The results show that as the age grows to 34, people tend to prefer the genre "Comedy" but as age grows passing 34, the users tend to prefer drama instead. This transition of the preferred genre is seen more detailed in Figure 10.1 and 10.2 as we can see the percentage of users preferring drama movies increased from 16.69% to 17.15% which overtook comedy.

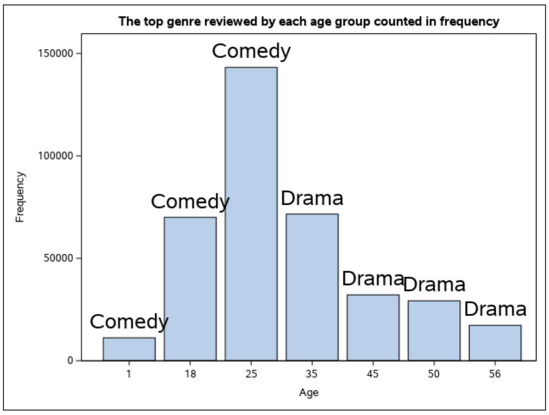


Figure 9. Top Genre based on age groups.

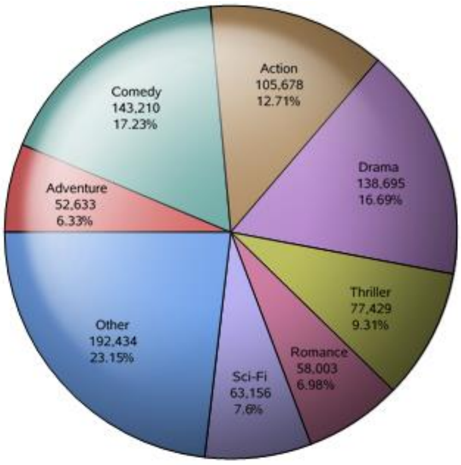


Figure 10 Genre preference in Age Group 25.

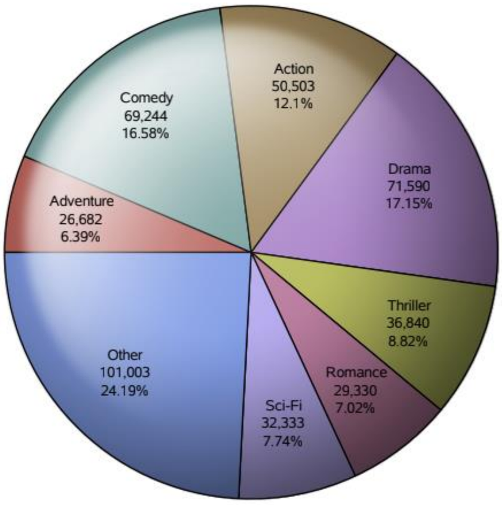


Figure 10.1 Genre preference in Age Group 35.

In Figure 10.1 and 10.2, it is noted that the category "Other" has the highest percentage. This is because to keep the pie chart clean and more readable , the genres that are lower than 5% which were Animation, Children's, Crime, Documentary, Fantasy, Film-Noir, Horror, Musical, Mystery, War and Western were categorized together as "Others" in the pie chart. Hence, we will observe the 2nd largest category as the most popular category in the specific age group. In this case, it would be Comedy and Drama for the age group 25 and 35 respectively,

II. Research Question 2

Research Question 2 is "**List of movies that fall above the upper quartile of the variable ratings**". The code for this section is entailed in Appendix E.

The purpose of this research question is to identify which movie is rated highly amongst the users and consequently indicate which movies are of good quality. We have decided to set the benchmark to that of the upper limit of the variable ratings. We used the "PROC MEANS" function equipped with "Q3" as its statistical keyword. Through this analysis, we were able to identify that the upper quartile of the variable ratings is 4.0.

We then proceeded to our analysis to fulfil our research question 2. We used "PROC SQL" to filter out our desired output. We created a table and then started our SQL sequence. Since each movie has been rated by multiple users, we used the average of all the ratings that have been collected from the users. To calculate the average rating for the respective movie and to avoid any duplication of the Movie Name, we utilized the function "DISTINCT()" on the variable name and "AVG()" on the variable rating. We also filtered the table to only display average ratings of greater than or equal to ("GE") 4. We were also interested to find out how many users have rated a specific movie. We used the "COUNT()" and 'DISTINCT()" on the variable UserID. The results of this analysis showed that there are a total of **430 Movies** that fell above the upper quartile of the variable ratings (4.0). It is noted that 10 Movies are rated at 5.0. However, upon close inspection, we noticed that those 10 Movies had only been rated once or twice. This is compared to other movies that have been rated by thousands of users. Hence, that is something we can take a look at in the future.

Based on the movies that we have identified to have fallen to the upper limit of the rating benchmark, we conducted further analysis to identify the "**Gender breakdown based on User Rating**". The purpose of this analysis was to identify the gender breakdown of the users who had rated the movies that had an average rating of greater than or equal to 4. We visualized our findings into a pie chart in Figure 1. Based on our findings, a total of 75.51% (179716) of the users are male and 24.49% (58301) of them are females. There were a total of 238017 users who had rated for movies that had averaged a rating greater than or equal to 4.0.

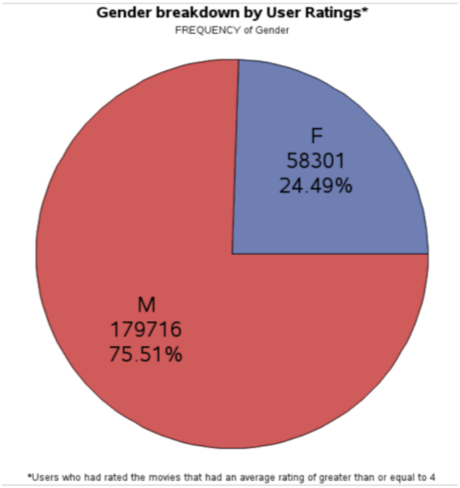


Figure 11. Gender breakdown based on User Rating

This goes back to our initial finding which is that the majority of the users are male in comparison to the female users.

Finally, to end our analysis for Research Question 2, we conducted a sub-analysis to identify the "**Average rating based on the Year of Release**". The objective of this analysis was to identify the relationship between the movies that had an average rating of greater than or equal to 4 and the year of its release. In order to obtain a meaningful output, we ran a regression model using "PROC REG" and generated a fit plot. The fit plot in Figure 12 shows us that there is no obvious relationship between the Average rating and the Year of Release. The fit plot also indicates that there is an adjusted R-square of 0.0143. Hence, we can conclude that there is an insignificant relationship between the average rating and the year of release.

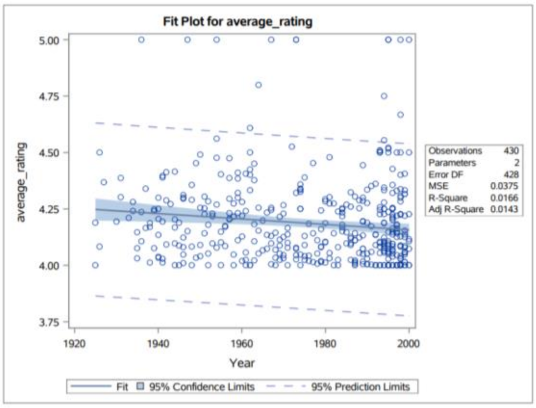


Figure 12. Average rating based on the Year of Release

CONCLUSION

This- work -has- shown -how -evolutionary users' age, gender, and movie genre influence movie ratings. We were able to properly understand the datasets and identify key trends amongst the movie audience.

In this analysis, we ran several tests to prove how users' age group impacts movie ratings, which genre is more popular among users, and the list of movies that fall above the upper quartile of movie ratings. The results indicate that a majority user under the age group 25 are rated between 3.5 and 4, and their most reviewed genres are comedy and drama. On the other hand, we found that at least 75.51% of the users are male and 24.49% are female, which concludes our research that men are significantly more involved in reviewing movies they watch.

Based on the result of research question 1, we can identify that age group does not significantly affect the ratings of a movie. However, we can identify the general range of ratings that each age group gives a movie. In general, we can say that the overall rating is at the middle to a positive range which may indicate moderate satisfaction rate. Apart from that, it is also seen that the age group below age group 25 prefers comedy while the age group above 25 prefers drama. The age group above 25 also tend to have a more diverse range of feedback. They have given low as well as high feedback.

Additionally, from the research question 2, we can conclude that the majority of users that rated for the movies that fall above the upper quartile of ratings are males. However, it is noted that there is no significant relationship between the movies and the year of release. The gender breakdown and year analysis allowed us to learn more about the set of movies that fell above the upper limit. There is also a small proportion of movies that have averaged a score of 5.0.

Overall, if our analysis was to be presented to any movie producers, the general advice based on the result of our analysis would be as follows. Content and storyline of the movie should still be the priority as there is insufficient evidence to say that there are factors that could affect the ratings of a given movie. If the quality of the content and storyline is maintained, a general pattern of a well-rated movie will need to satisfy male viewers, in the comedy genre and target audience of age lesser than 45.

REFERENCES

[1] Barza, S. (2014, May). *Movie Genre Preference and Culture*. Retrieved from https://www.sciencedirect.com/science/article/pii/S187704281402518X

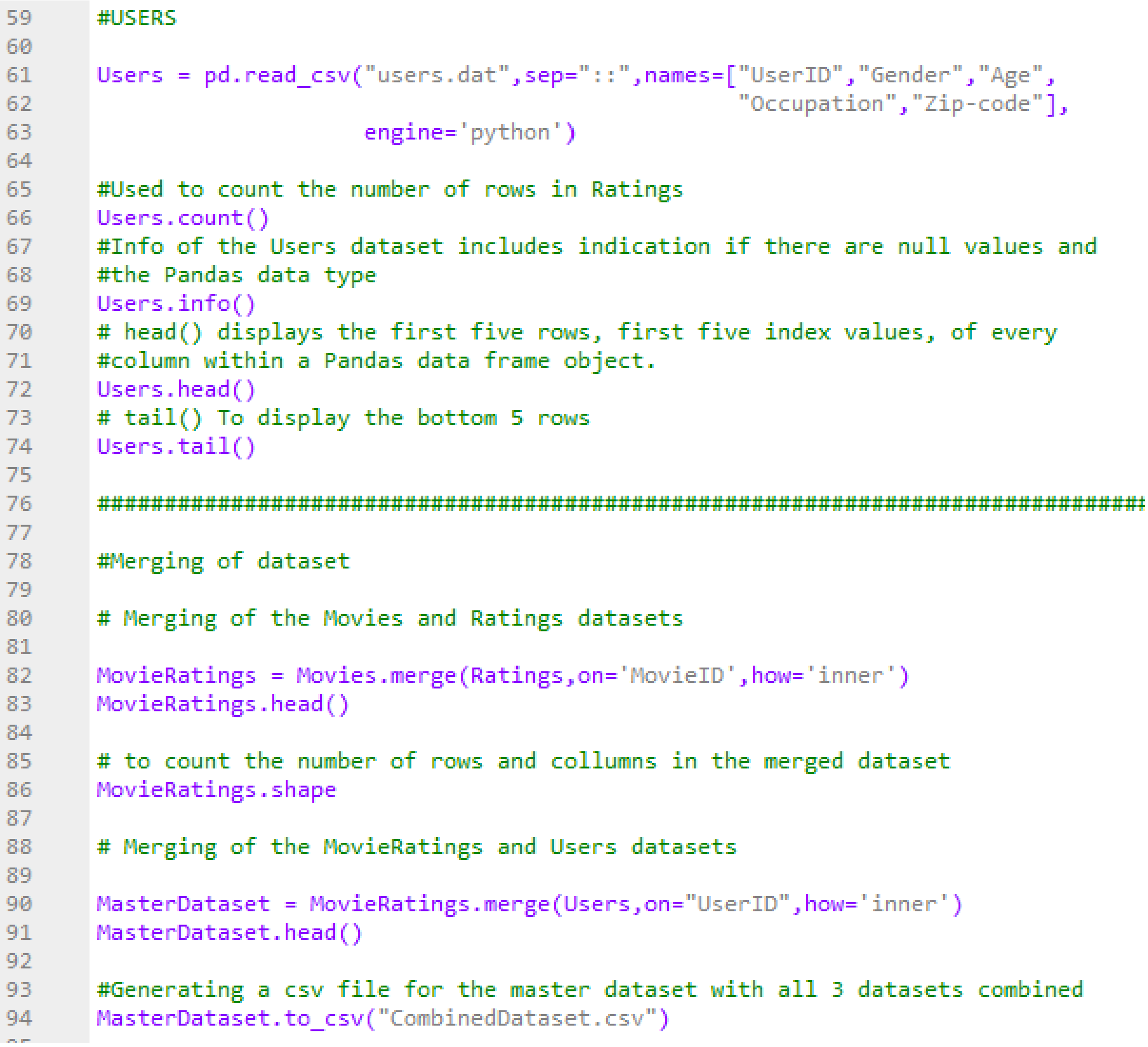
[2] Hazer D., M. X. (2015). *Emotion Elicitation Using Film Clips: Effect of Age Groups on Movie Choice and Emotion Rating*. Retrieved from springer:

https://link.springer.com/chapter/10.1007/978-3-319-21380-4_20#citeas

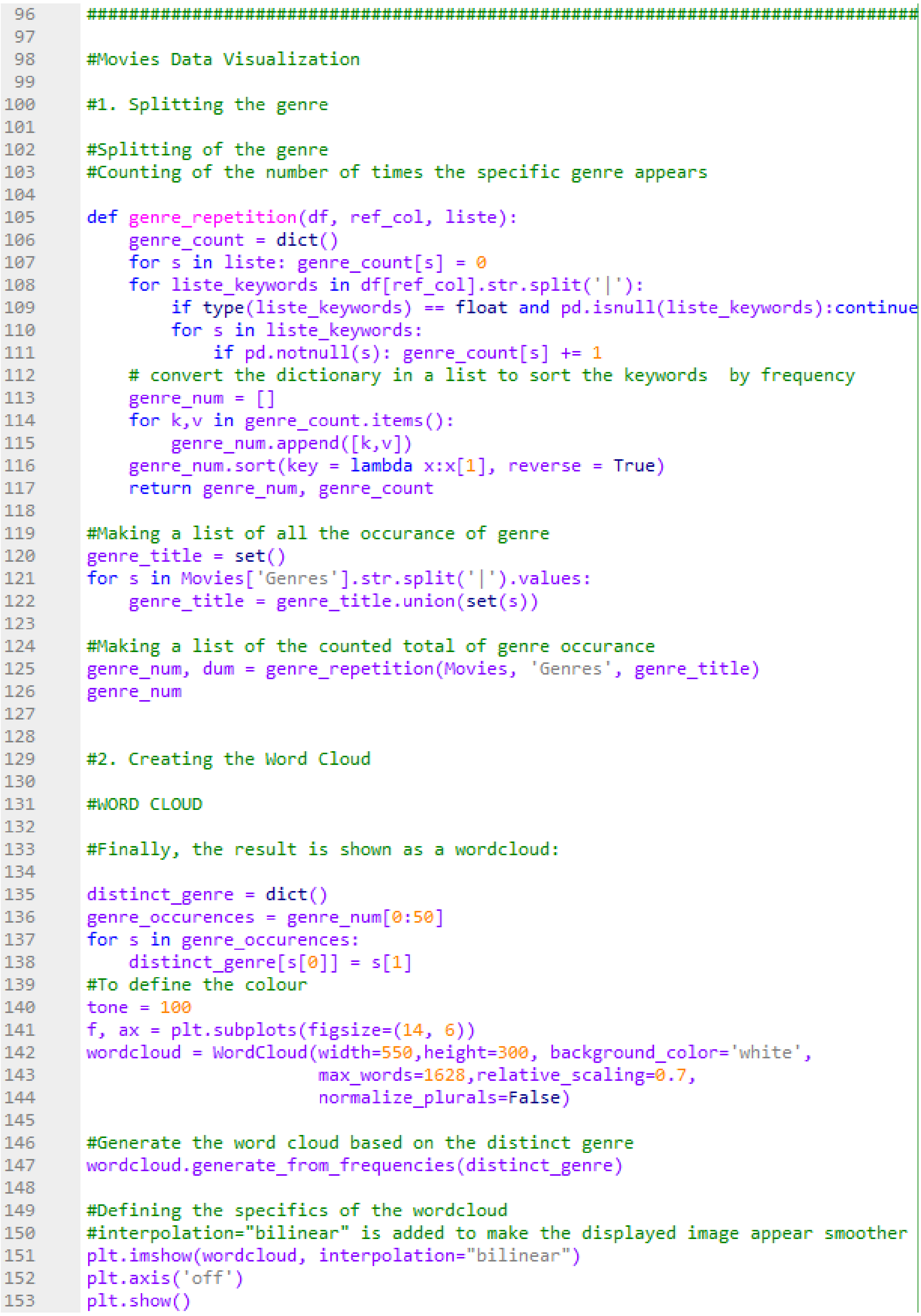
[3] Rieckmann, R. (2017, May 22). *The Importance of Film Ratings*. Retrieved from neenah satellite: https://neenahsatellite.com/15528/student-life/creative-writing/magazines/the-importance-of-film-ratings/

[4] Chu, J. (2019). *Interpret the key results for Correlation*. Minitab. Retrieved November 17, 2020, from https://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/regression/how-to/correlation/interpret-the-results/#:~:text=For%20the%20Pearson%20correlation%2C%20an,linear%20relationship%20between%20the%20variables.&text=If%20both

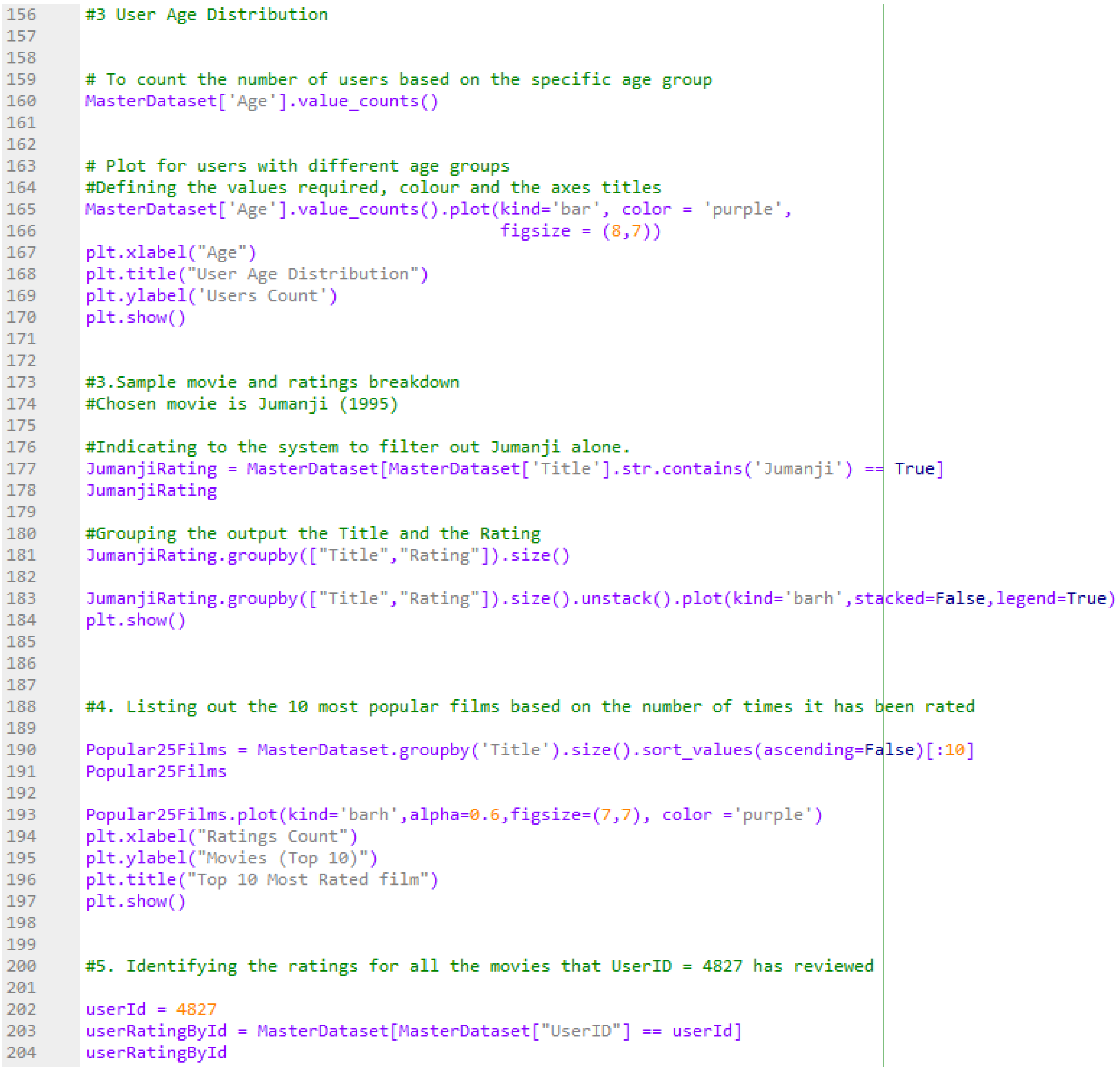
Appendix A - Data Exploration



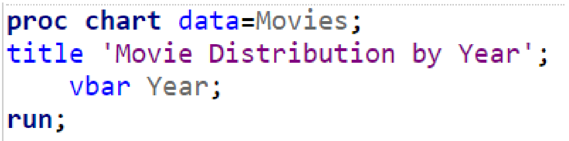
```
#USERS

Users = pd.read_csv("users.dat",sep="::",names=["UserID","Gender","Age",
                                                "Occupation","Zip-code"],
                    engine='python')

#Used to count the number of rows in Ratings
Users.count()
#Info of the Users dataset includes indication if there are null values and
#the Pandas data type
Users.info()
# head() displays the first five rows, first five index values, of every
#column within a Pandas data frame object.
Users.head()
# tail() To display the bottom 5 rows
Users.tail()

###############################################################################

#Merging of dataset

# Merging of the Movies and Ratings datasets

MovieRatings = Movies.merge(Ratings,on='MovieID',how='inner')
MovieRatings.head()

# to count the number of rows and collumns in the merged dataset
MovieRatings.shape

# Merging of the MovieRatings and Users datasets

MasterDataset = MovieRatings.merge(Users,on="UserID",how='inner')
MasterDataset.head()

#Generating a csv file for the master dataset with all 3 datasets combined
MasterDataset.to_csv("CombinedDataset.csv")
```



```python
###############################################################################

#Movies Data Visualization

#1. Splitting the genre

#Splitting of the genre
#Counting of the number of times the specific genre appears

def genre_repetition(df, ref_col, liste):
    genre_count = dict()
    for s in liste: genre_count[s] = 0
    for liste_keywords in df[ref_col].str.split('|'):
        if type(liste_keywords) == float and pd.isnull(liste_keywords):continue
        for s in liste_keywords:
            if pd.notnull(s): genre_count[s] += 1
    # convert the dictionary in a list to sort the keywords  by frequency
    genre_num = []
    for k,v in genre_count.items():
        genre_num.append([k,v])
    genre_num.sort(key = lambda x:x[1], reverse = True)
    return genre_num, genre_count

#Making a list of all the occurance of genre
genre_title = set()
for s in Movies['Genres'].str.split('|').values:
    genre_title = genre_title.union(set(s))

#Making a list of the counted total of genre occurance
genre_num, dum = genre_repetition(Movies, 'Genres', genre_title)
genre_num


#2. Creating the Word Cloud

#WORD CLOUD

#Finally, the result is shown as a wordcloud:

distinct_genre = dict()
genre_occurences = genre_num[0:50]
for s in genre_occurences:
    distinct_genre[s[0]] = s[1]
#To define the colour
tone = 100
f, ax = plt.subplots(figsize=(14, 6))
wordcloud = WordCloud(width=550,height=300, background_color='white',
                      max_words=1628,relative_scaling=0.7,
                      normalize_plurals=False)

#Generate the word cloud based on the distinct genre
wordcloud.generate_from_frequencies(distinct_genre)

#Defining the specifics of the wordcloud
#interpolation="bilinear" is added to make the displayed image appear smoother
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis('off')
plt.show()
```

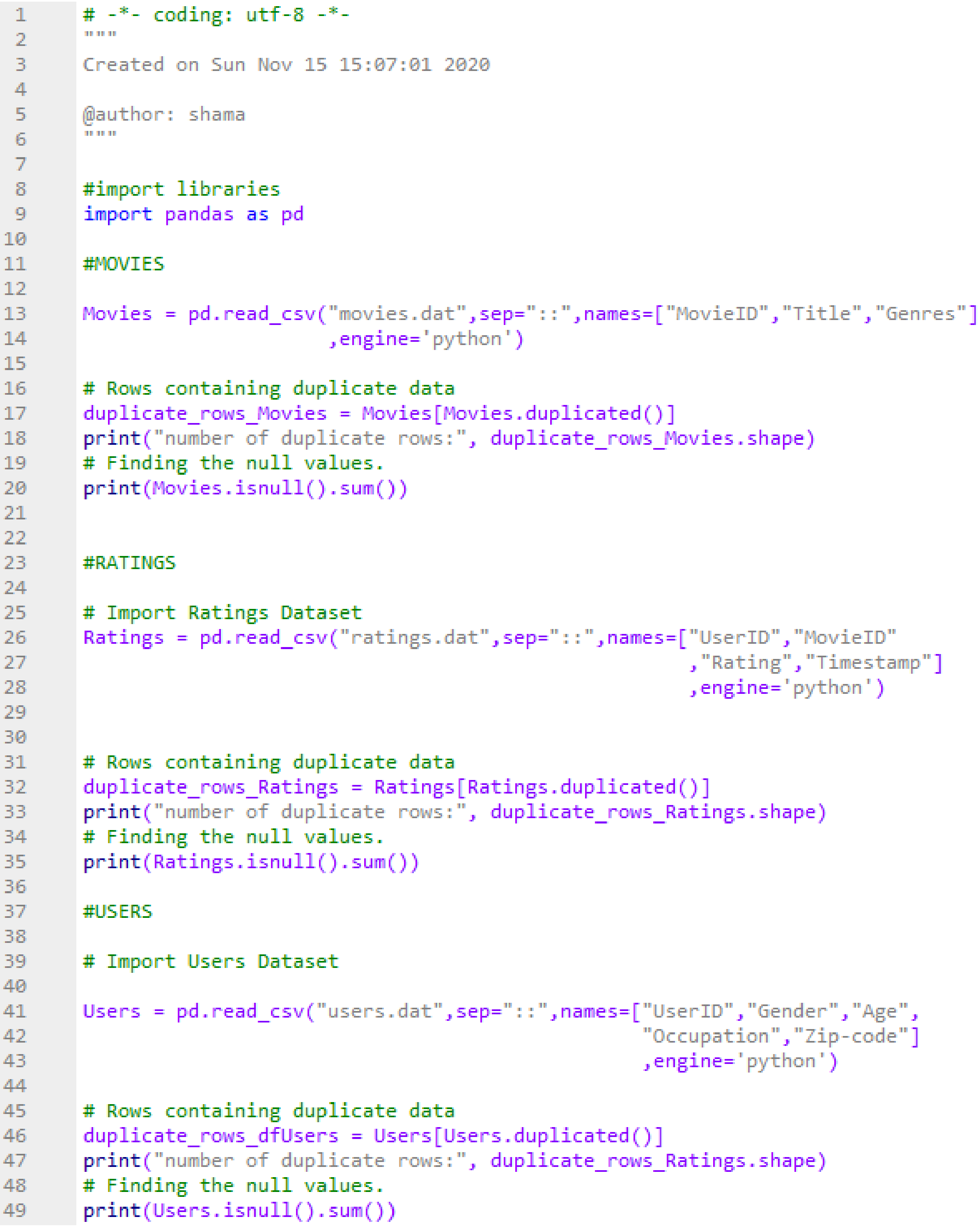


```
#3 User Age Distribution


# To count the number of users based on the specific age group
MasterDataset['Age'].value_counts()


# Plot for users with different age groups
#Defining the values required, colour and the axes titles
MasterDataset['Age'].value_counts().plot(kind='bar', color = 'purple',
                                          figsize = (8,7))
plt.xlabel("Age")
plt.title("User Age Distribution")
plt.ylabel('Users Count')
plt.show()


#3.Sample movie and ratings breakdown
#Chosen movie is Jumanji (1995)

#Indicating to the system to filter out Jumanji alone.
JumanjiRating = MasterDataset[MasterDataset['Title'].str.contains('Jumanji') == True]
JumanjiRating

#Grouping the output the Title and the Rating
JumanjiRating.groupby(["Title","Rating"]).size()

JumanjiRating.groupby(["Title","Rating"]).size().unstack().plot(kind='barh',stacked=False,legend=True)
plt.show()



#4. Listing out the 10 most popular films based on the number of times it has been rated

Popular25Films = MasterDataset.groupby('Title').size().sort_values(ascending=False)[:10]
Popular25Films

Popular25Films.plot(kind='barh',alpha=0.6,figsize=(7,7), color ='purple')
plt.xlabel("Ratings Count")
plt.ylabel("Movies (Top 10)")
plt.title("Top 10 Most Rated film")
plt.show()


#5. Identifying the ratings for all the movies that UserID = 4827 has reviewed

userId = 4827
userRatingById = MasterDataset[MasterDataset["UserID"] == userId]
userRatingById
```



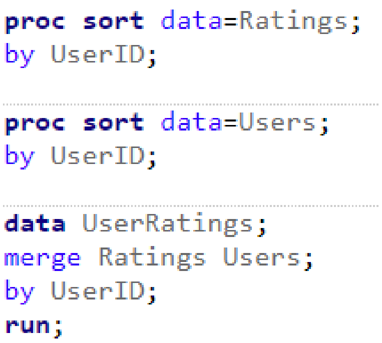
```
proc chart data=Movies;
title 'Movie Distribution by Year';
    vbar Year;
run;
```

Appendix B - Data Cleaning



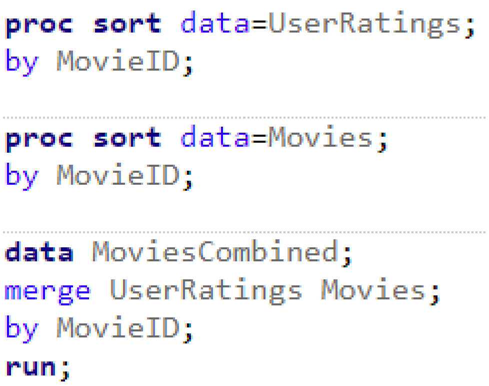
```python
# -*- coding: utf-8 -*-
"""
Created on Sun Nov 15 15:07:01 2020

@author: shama
"""

#import libraries
import pandas as pd

#MOVIES

Movies = pd.read_csv("movies.dat",sep="::",names=["MovieID","Title","Genres"]
                     ,engine='python')

# Rows containing duplicate data
duplicate_rows_Movies = Movies[Movies.duplicated()]
print("number of duplicate rows:", duplicate_rows_Movies.shape)
# Finding the null values.
print(Movies.isnull().sum())


#RATINGS

# Import Ratings Dataset
Ratings = pd.read_csv("ratings.dat",sep="::",names=["UserID","MovieID"
                                                    ,"Rating","Timestamp"]
                                                    ,engine='python')


# Rows containing duplicate data
duplicate_rows_Ratings = Ratings[Ratings.duplicated()]
print("number of duplicate rows:", duplicate_rows_Ratings.shape)
# Finding the null values.
print(Ratings.isnull().sum())

#USERS

# Import Users Dataset

Users = pd.read_csv("users.dat",sep="::",names=["UserID","Gender","Age",
                                                "Occupation","Zip-code"]
                                                ,engine='python')

# Rows containing duplicate data
duplicate_rows_dfUsers = Users[Users.duplicated()]
print("number of duplicate rows:", duplicate_rows_Ratings.shape)
# Finding the null values.
print(Users.isnull().sum())
```

Appendix C - Data Manipulation



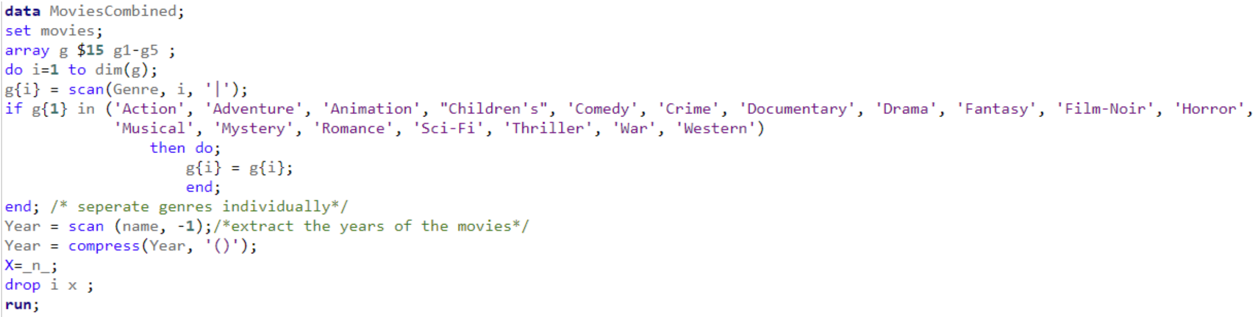
```
proc sort data=Ratings;
by UserID;

proc sort data=Users;
by UserID;

data UserRatings;
merge Ratings Users;
by UserID;
run;
```

Source code to perform sorting and merging of Users and Ratings dataset.



```
proc sort data=UserRatings;
by MovieID;

proc sort data=Movies;
by MovieID;

data MoviesCombined;
merge UserRatings Movies;
by MovieID;
run;
```

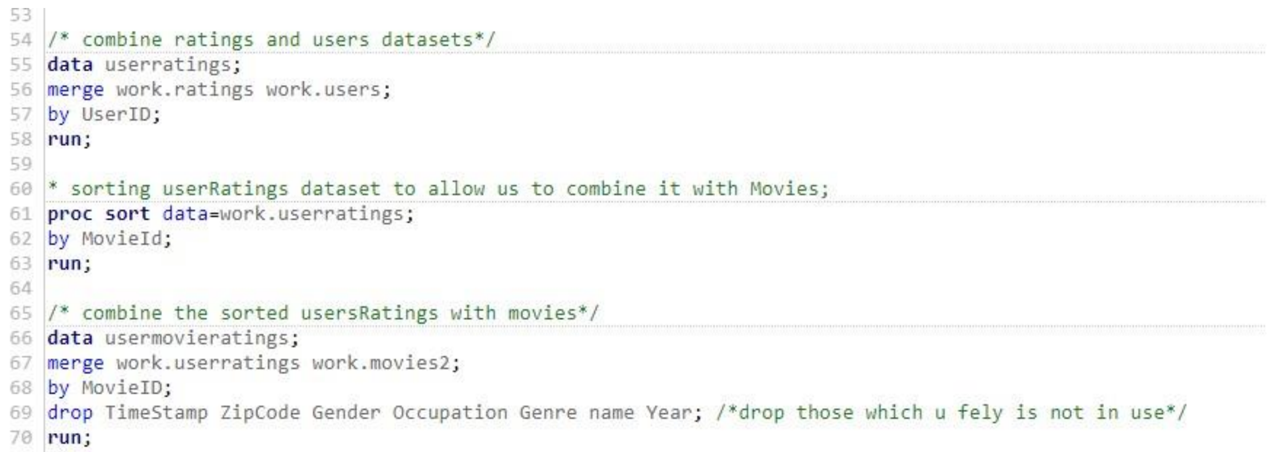
Source code to perform courting and merging of Movies and UserRatings dataset.



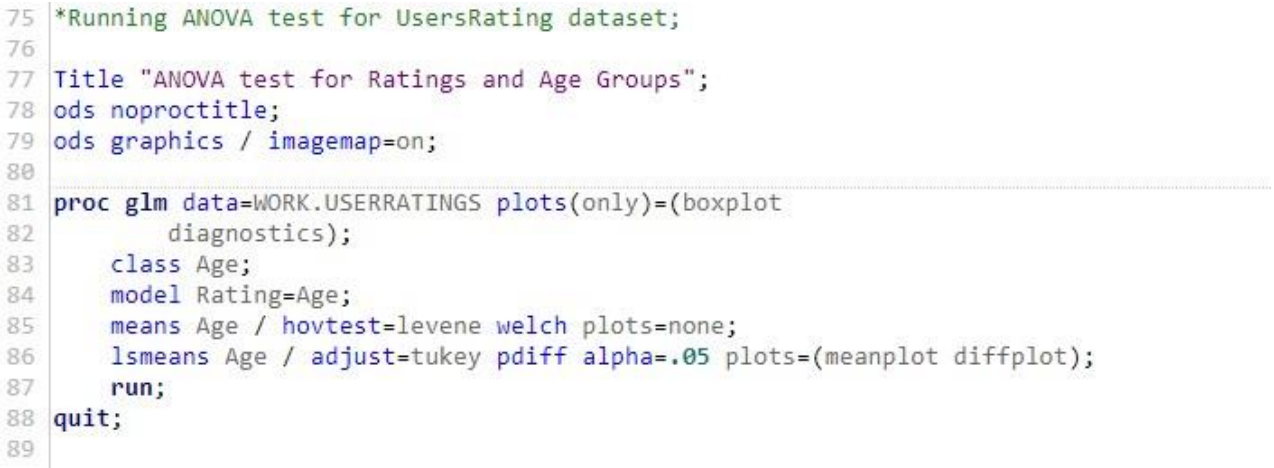
```
data MoviesCombined;
set movies;
array g $15 g1-g5 ;
do i=1 to dim(g);
g{i} = scan(Genre, i, '|');
if g{1} in ('Action', 'Adventure', 'Animation', "Children's", 'Comedy', 'Crime', 'Documentary', 'Drama', 'Fantasy', 'Film-Noir', 'Horror',
            'Musical', 'Mystery', 'Romance', 'Sci-Fi', 'Thriller', 'War', 'Western')
                then do;
                    g{i} = g{i};
                    end;
end; /* seperate genres individually*/
Year = scan (name, -1);/*extract the years of the movies*/
Year = compress(Year, '()');
X=_n_;
drop i x ;
run;
```

Source code to perform genre and years extraction.

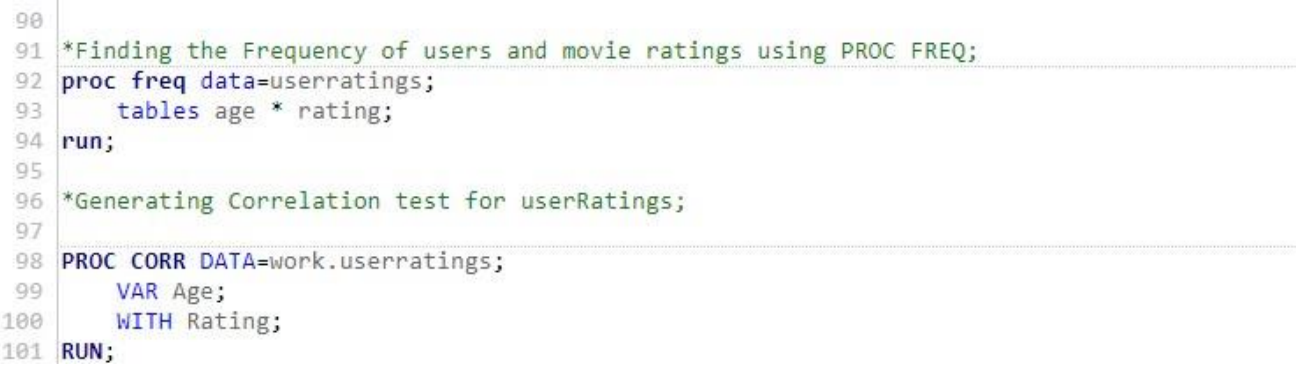
Appendix D - Research Question 1 Analysis



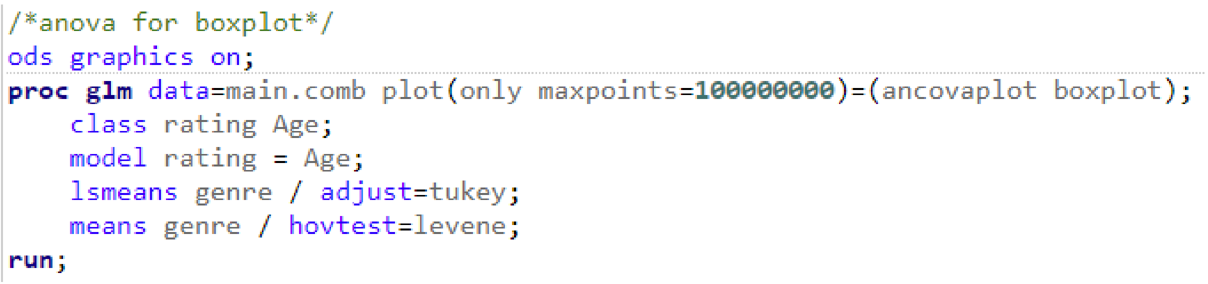
```
/* combine ratings and users datasets*/
data userratings;
merge work.ratings work.users;
by UserID;
run;

* sorting userRatings dataset to allow us to combine it with Movies;
proc sort data=work.userratings;
by MovieId;
run;

/* combine the sorted usersRatings with movies*/
data usermovieratings;
merge work.userratings work.movies2;
by MovieID;
drop TimeStamp ZipCode Gender Occupation Genre name Year; /*drop those which u fely is not in use*/
run;
```



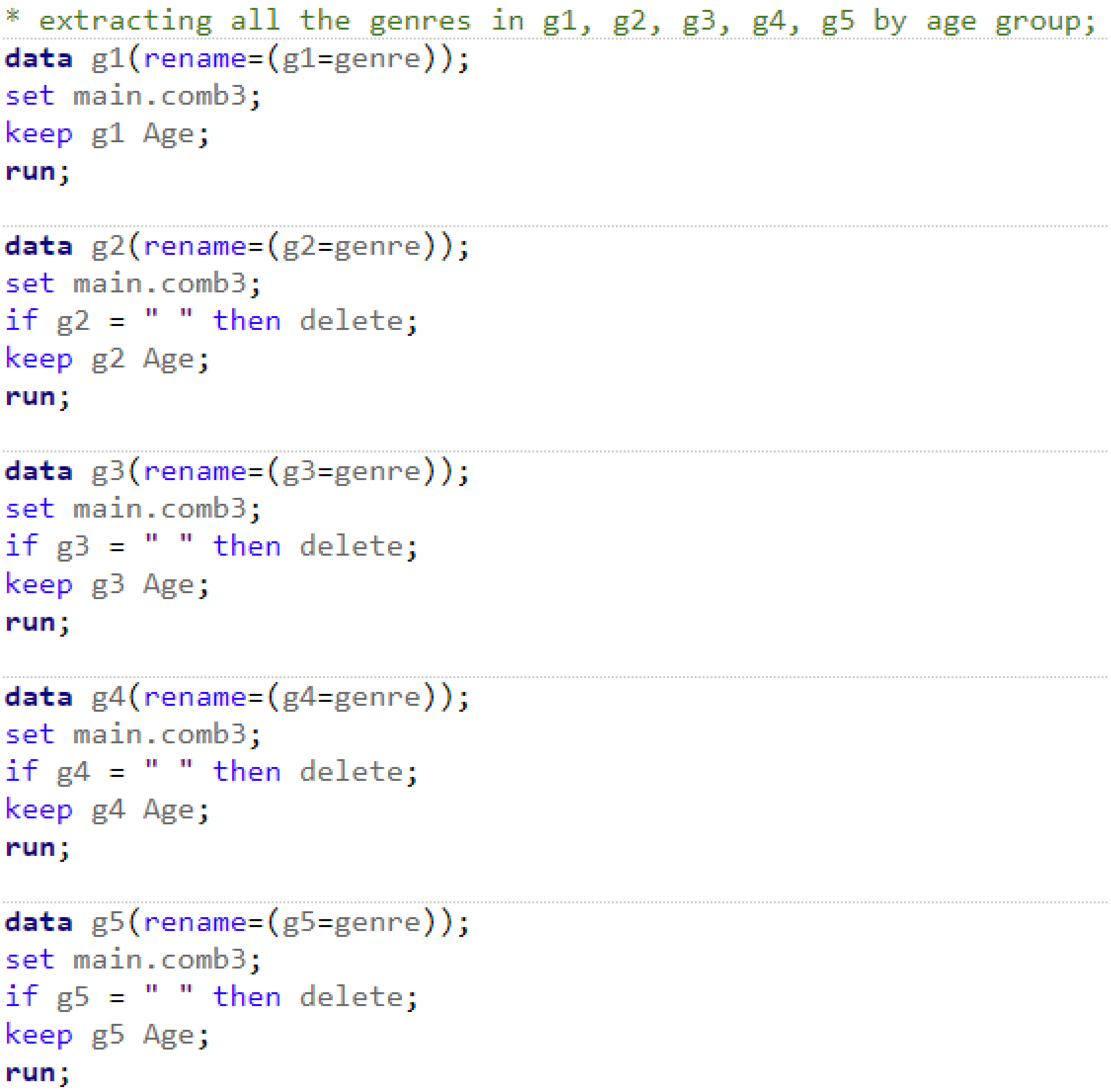
```
*Running ANOVA test for UsersRating dataset;

Title "ANOVA test for Ratings and Age Groups";
ods noproctitle;
ods graphics / imagemap=on;

proc glm data=WORK.USERRATINGS plots(only)=(boxplot
        diagnostics);
    class Age;
    model Rating=Age;
    means Age / hovtest=levene welch plots=none;
    lsmeans Age / adjust=tukey pdiff alpha=.05 plots=(meanplot diffplot);
    run;
quit;
```



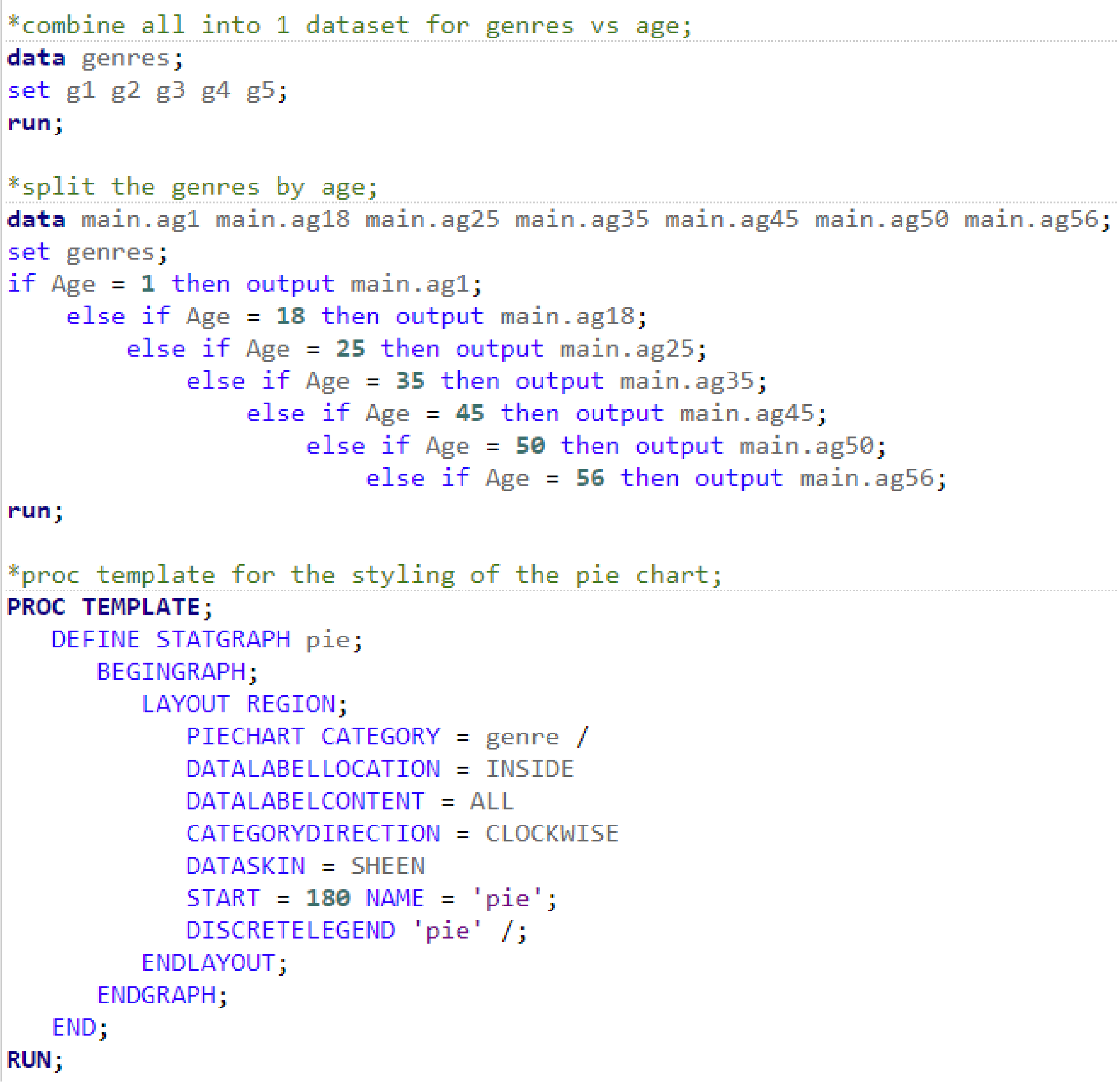
```

*Finding the Frequency of users and movie ratings using PROC FREQ;
proc freq data=userratings;
    tables age * rating;
run;

*Generating Correlation test for userRatings;

PROC CORR DATA=work.userratings;
    VAR Age;
    WITH Rating;
RUN;
```



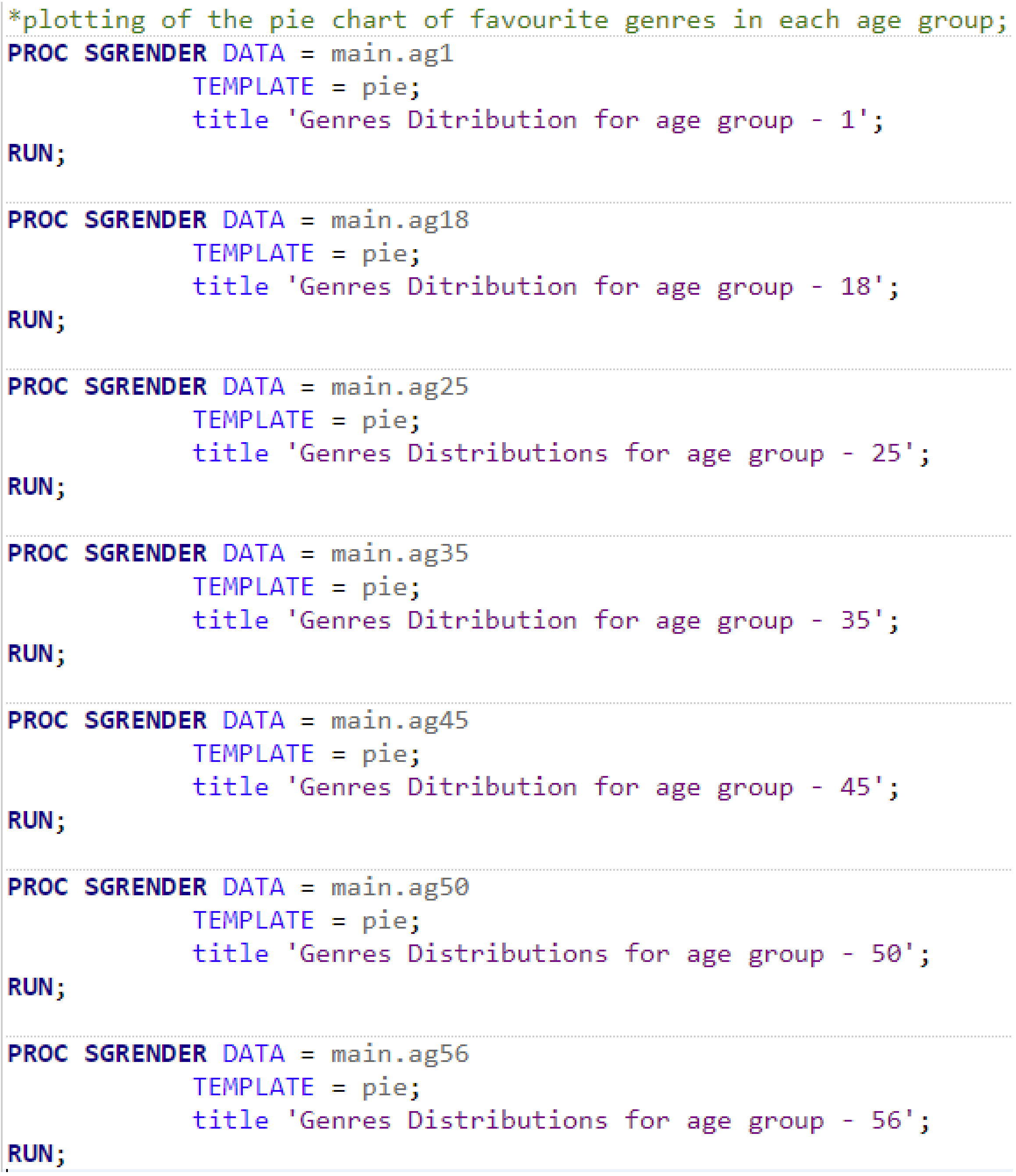
```
/*anova for boxplot*/
ods graphics on;
proc glm data=main.comb plot(only maxpoints=100000000)=(ancovaplot boxplot);
    class rating Age;
    model rating = Age;
    lsmeans genre / adjust=tukey;
    means genre / hovtest=levene;
run;
```



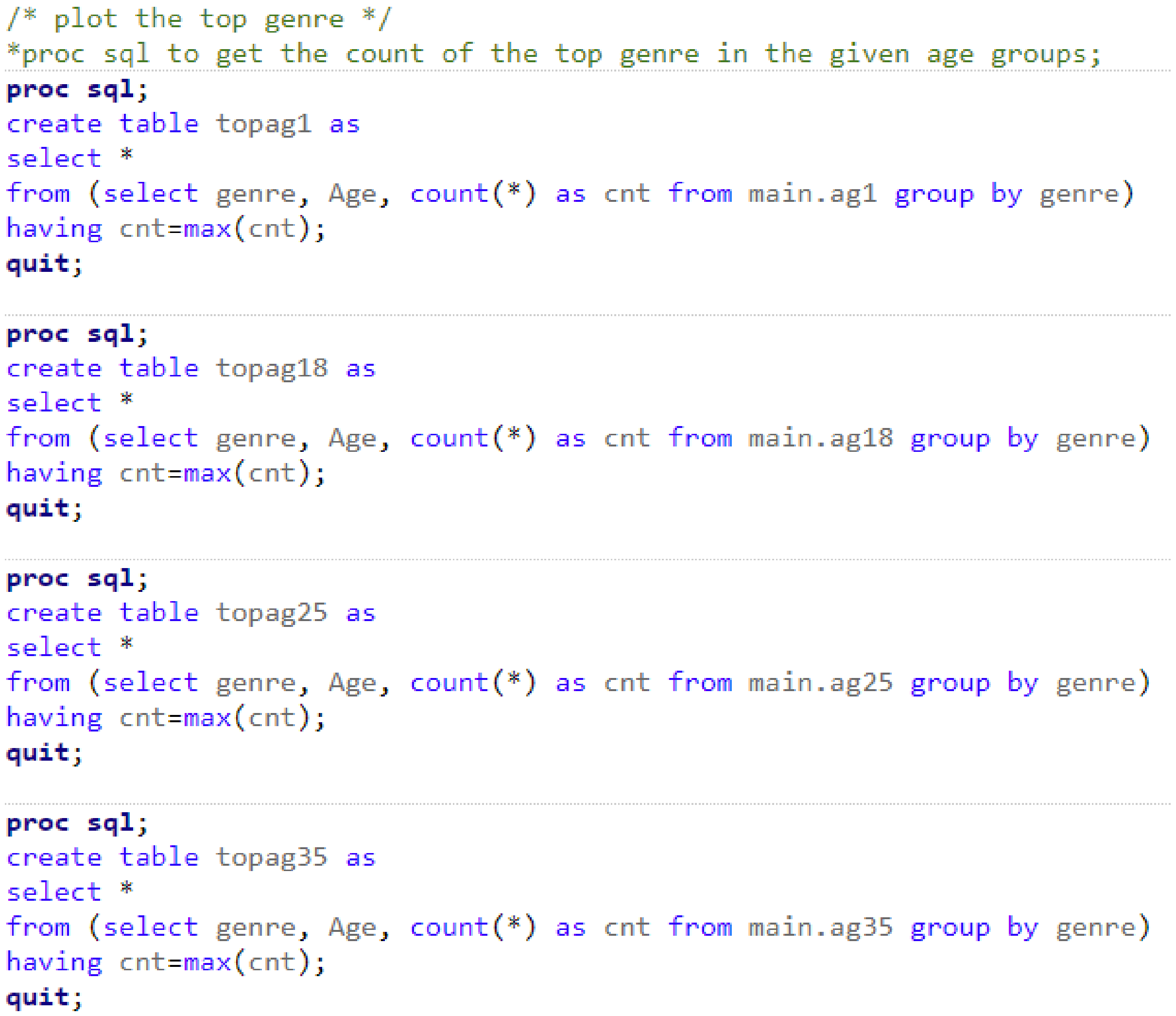
```
* extracting all the genres in g1, g2, g3, g4, g5 by age group;
data g1(rename=(g1=genre));
set main.comb3;
keep g1 Age;
run;

data g2(rename=(g2=genre));
set main.comb3;
if g2 = " " then delete;
keep g2 Age;
run;

data g3(rename=(g3=genre));
set main.comb3;
if g3 = " " then delete;
keep g3 Age;
run;

data g4(rename=(g4=genre));
set main.comb3;
if g4 = " " then delete;
keep g4 Age;
run;

data g5(rename=(g5=genre));
set main.comb3;
if g5 = " " then delete;
keep g5 Age;
run;
```



```
*combine all into 1 dataset for genres vs age;
data genres;
set g1 g2 g3 g4 g5;
run;

*split the genres by age;
data main.ag1 main.ag18 main.ag25 main.ag35 main.ag45 main.ag50 main.ag56;
set genres;
if Age = 1 then output main.ag1;
    else if Age = 18 then output main.ag18;
        else if Age = 25 then output main.ag25;
            else if Age = 35 then output main.ag35;
                else if Age = 45 then output main.ag45;
                    else if Age = 50 then output main.ag50;
                        else if Age = 56 then output main.ag56;
run;

*proc template for the styling of the pie chart;
PROC TEMPLATE;
   DEFINE STATGRAPH pie;
      BEGINGRAPH;
         LAYOUT REGION;
            PIECHART CATEGORY = genre /
            DATALABELLOCATION = INSIDE
            DATALABELCONTENT = ALL
            CATEGORYDIRECTION = CLOCKWISE
            DATASKIN = SHEEN
            START = 180 NAME = 'pie';
            DISCRETELEGEND 'pie' /;
         ENDLAYOUT;
      ENDGRAPH;
   END;
RUN;
```



```
*plotting of the pie chart of favourite genres in each age group;
PROC SGRENDER DATA = main.ag1
            TEMPLATE = pie;
            title 'Genres Ditribution for age group - 1';
RUN;

PROC SGRENDER DATA = main.ag18
            TEMPLATE = pie;
            title 'Genres Ditribution for age group - 18';
RUN;

PROC SGRENDER DATA = main.ag25
            TEMPLATE = pie;
            title 'Genres Distributions for age group - 25';
RUN;

PROC SGRENDER DATA = main.ag35
            TEMPLATE = pie;
            title 'Genres Ditribution for age group - 35';
RUN;

PROC SGRENDER DATA = main.ag45
            TEMPLATE = pie;
            title 'Genres Ditribution for age group - 45';
RUN;

PROC SGRENDER DATA = main.ag50
            TEMPLATE = pie;
            title 'Genres Distributions for age group - 50';
RUN;

PROC SGRENDER DATA = main.ag56
            TEMPLATE = pie;
            title 'Genres Distributions for age group - 56';
RUN;
```



```
/* plot the top genre */
*proc sql to get the count of the top genre in the given age groups;
proc sql;
create table topag1 as
select *
from (select genre, Age, count(*) as cnt from main.ag1 group by genre)
having cnt=max(cnt);
quit;

proc sql;
create table topag18 as
select *
from (select genre, Age, count(*) as cnt from main.ag18 group by genre)
having cnt=max(cnt);
quit;

proc sql;
create table topag25 as
select *
from (select genre, Age, count(*) as cnt from main.ag25 group by genre)
having cnt=max(cnt);
quit;

proc sql;
create table topag35 as
select *
from (select genre, Age, count(*) as cnt from main.ag35 group by genre)
having cnt=max(cnt);
quit;
```

```
proc sql;
create table topag45 as
select *
from (select genre, Age, count(*) as cnt from main.ag45 group by genre)
having cnt=max(cnt);
quit;
```

```
proc sql;
create table topag50 as
select *
from (select genre, Age, count(*) as cnt from main.ag50 group by genre)
having cnt=max(cnt);
quit;
```

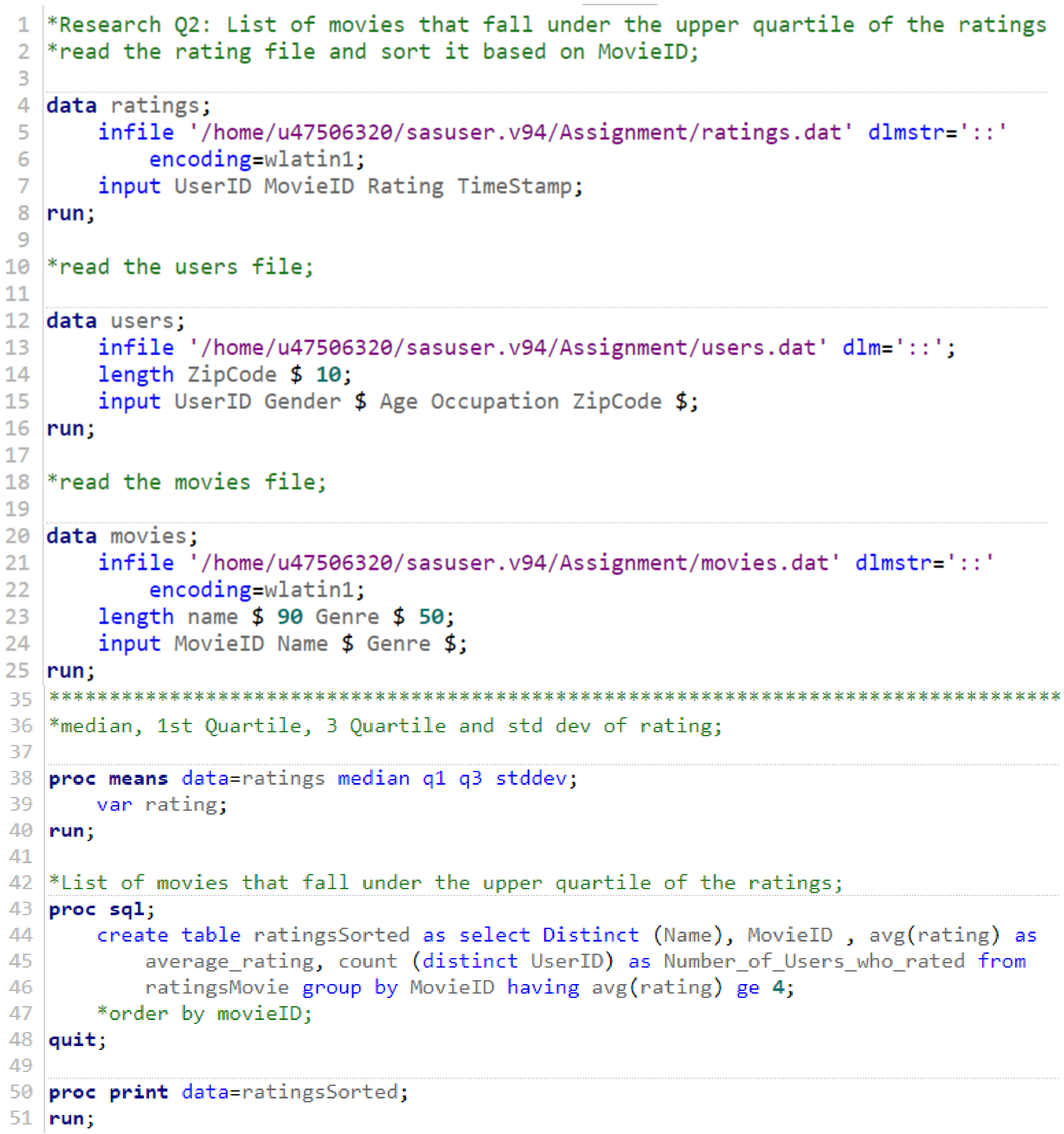
```
proc sql;
create table topag56 as
select *
from (select genre, Age, count(*) as cnt from main.ag56 group by genre)
having cnt=max(cnt);
quit;
```

```
data topgva;
set topag1 topag18 topag25 topag35 topag45 topag50 topag56;
run;
```

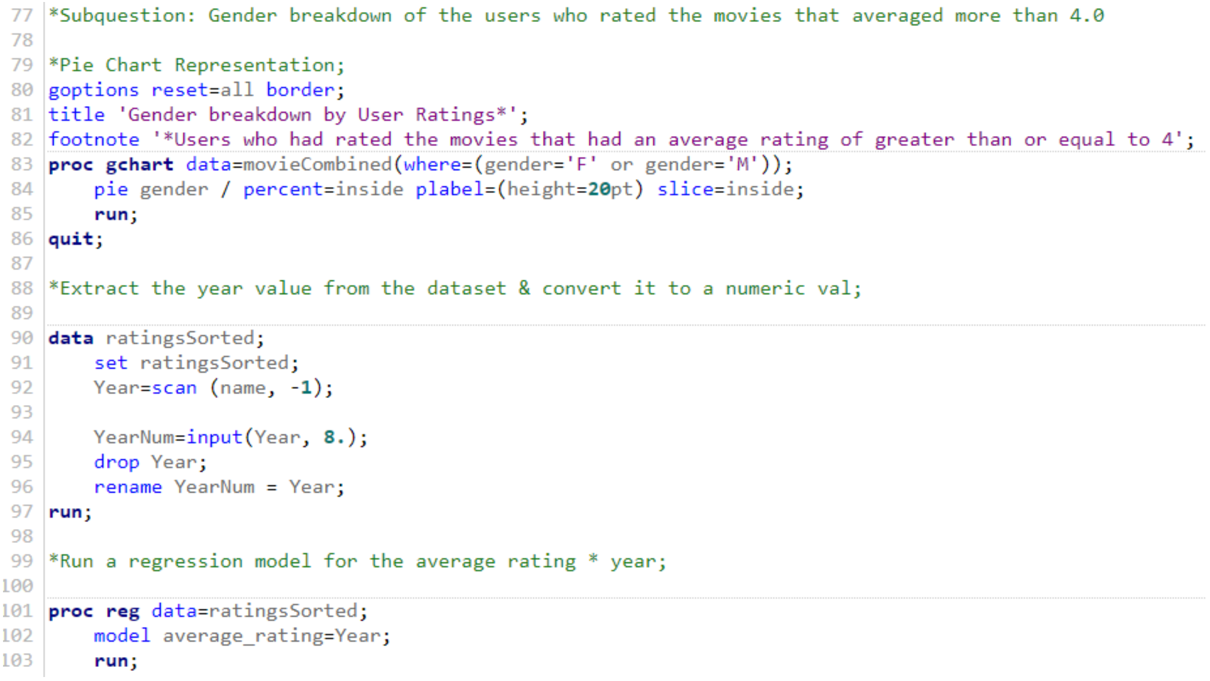
```
proc template;
define style styles.mystyle;
parent=styles.listing;
style GraphDataText from GraphDataText /
fontsize=18pt;
end;
run;

ods html style=tyle;
```

```
PROC SGPLOT DATA = topgva;
title 'The top genre reviewed by each age group counted in frequency';
VBAR age /
datalabel = genre;
RUN;
```

Appendix E - Research Question 2 Analysis



```
*Research Q2: List of movies that fall under the upper quartile of the ratings
*read the rating file and sort it based on MovieID;

data ratings;
    infile '/home/u47506320/sasuser.v94/Assignment/ratings.dat' dlmstr='::'
        encoding=wlatin1;
    input UserID MovieID Rating TimeStamp;
run;

*read the users file;

data users;
    infile '/home/u47506320/sasuser.v94/Assignment/users.dat' dlm='::';
    length ZipCode $ 10;
    input UserID Gender $ Age Occupation ZipCode $;
run;

*read the movies file;

data movies;
    infile '/home/u47506320/sasuser.v94/Assignment/movies.dat' dlmstr='::'
        encoding=wlatin1;
    length name $ 90 Genre $ 50;
    input MovieID Name $ Genre $;
run;
************************************************************************************
*median, 1st Quartile, 3 Quartile and std dev of rating;

proc means data=ratings median q1 q3 stddev;
    var rating;
run;

*List of movies that fall under the upper quartile of the ratings;
proc sql;
    create table ratingsSorted as select Distinct (Name), MovieID , avg(rating) as
        average_rating, count (distinct UserID) as Number_of_Users_who_rated from
        ratingsMovie group by MovieID having avg(rating) ge 4;
    *order by movieID;
quit;

proc print data=ratingsSorted;
run;
```



```
*Subquestion: Gender breakdown of the users who rated the movies that averaged more than 4.0

*Pie Chart Representation;
goptions reset=all border;
title 'Gender breakdown by User Ratings*';
footnote '*Users who had rated the movies that had an average rating of greater than or equal to 4';
proc gchart data=movieCombined(where=(gender='F' or gender='M'));
    pie gender / percent=inside plabel=(height=20pt) slice=inside;
    run;
quit;

*Extract the year value from the dataset & convert it to a numeric val;

data ratingsSorted;
    set ratingsSorted;
    Year=scan (name, -1);

    YearNum=input(Year, 8.);
    drop Year;
    rename YearNum = Year;
run;

*Run a regression model for the average rating * year;

proc reg data=ratingsSorted;
    model average_rating=Year;
    run;
```

Appendix F - Figures

Figure 1. Wordcloud representation of the genres.

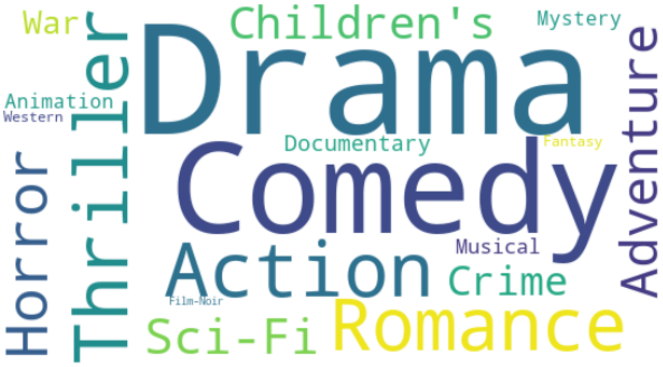


Figure 2. Movie Distribution by Year

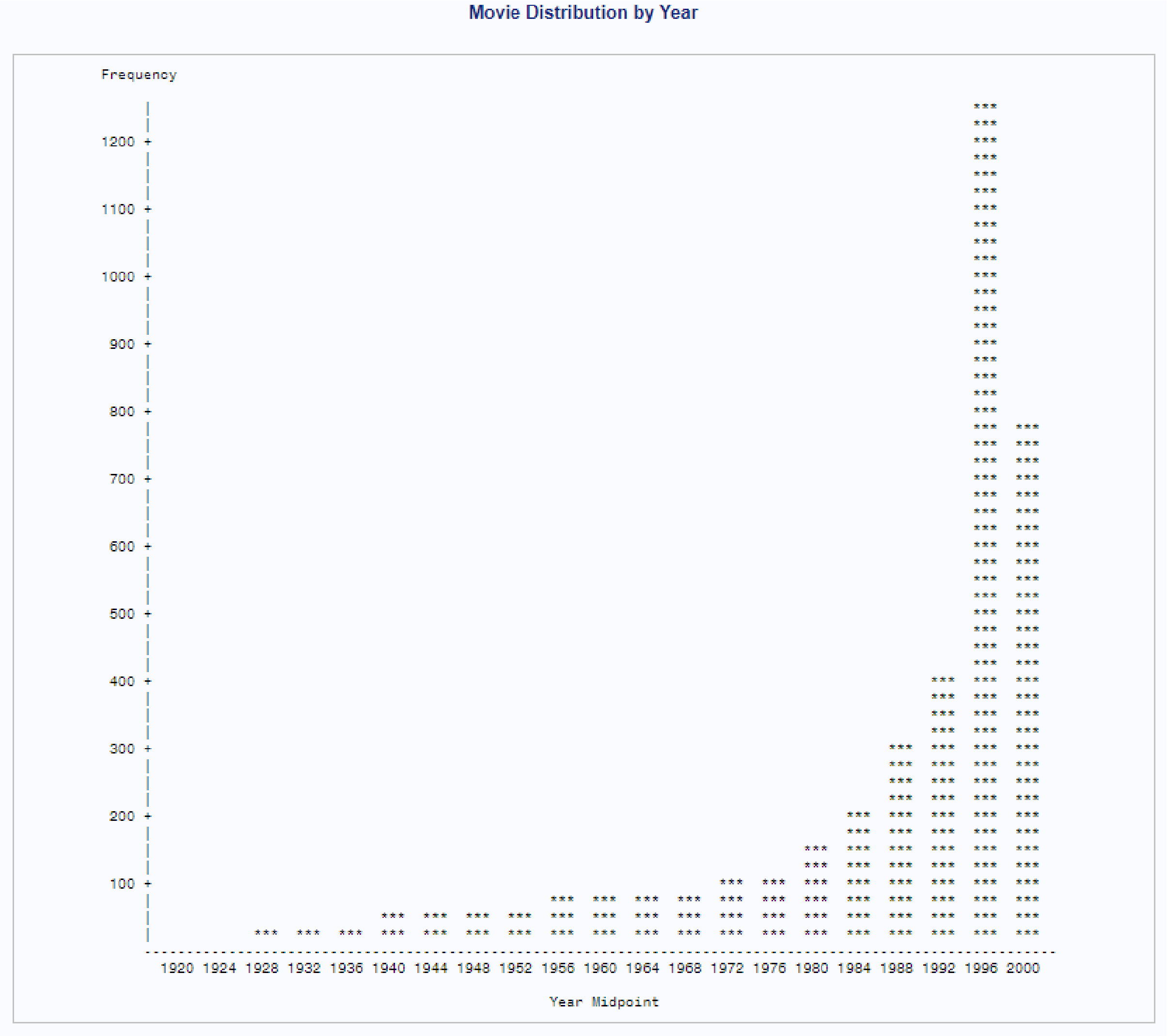


Figure 3. User Age Distribution.

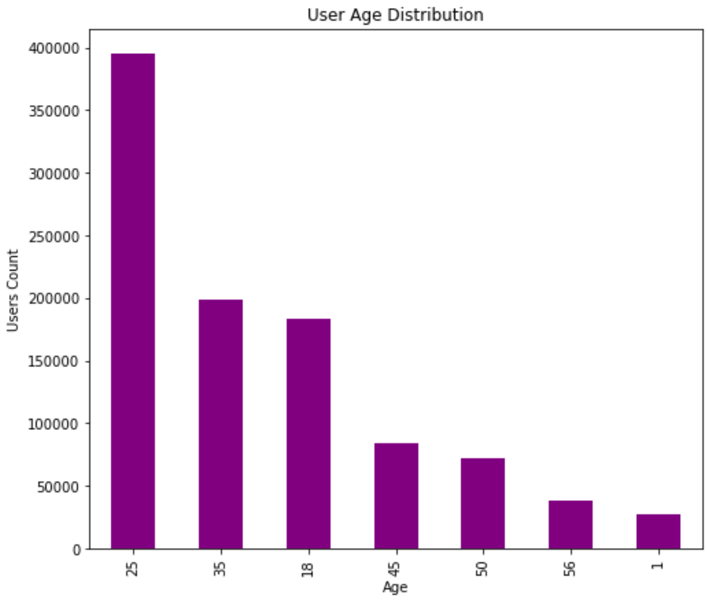


Figure 4. Total rating by Users for Jumanji (1995)

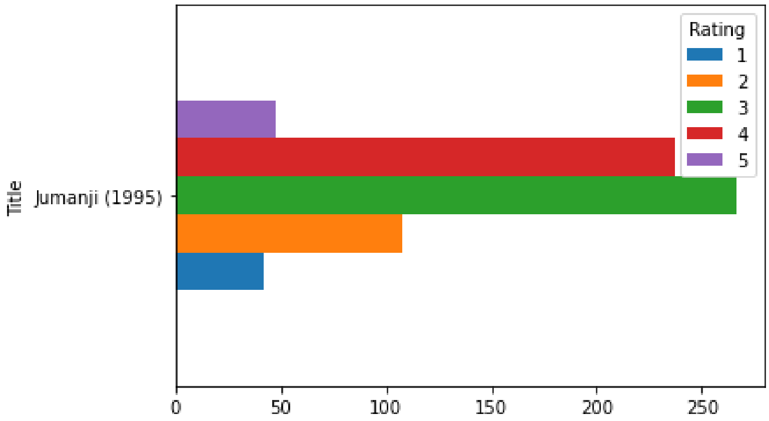


Figure 5. Top 10 most rated film

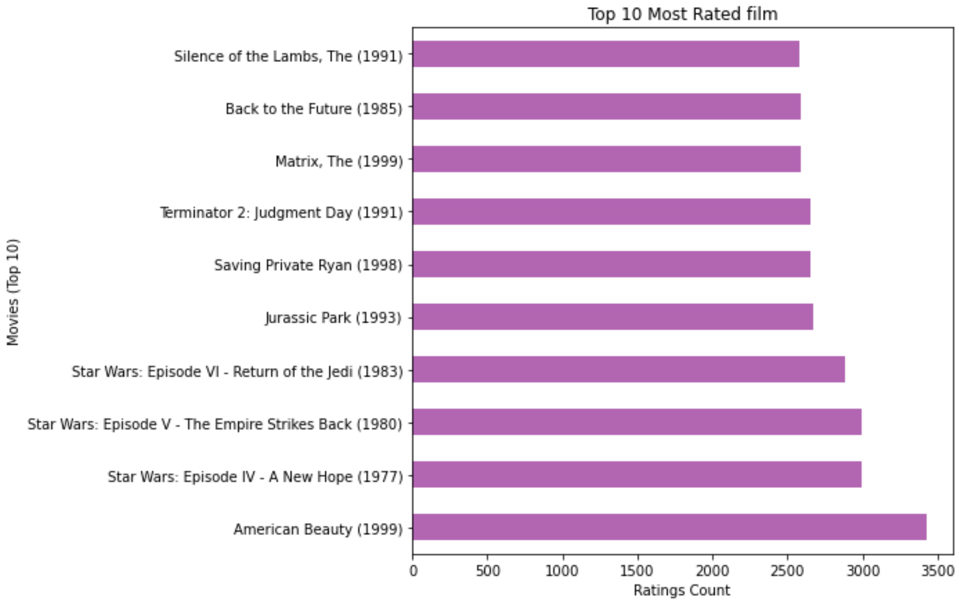


Figure 6. Rating Breakdown based on Age.

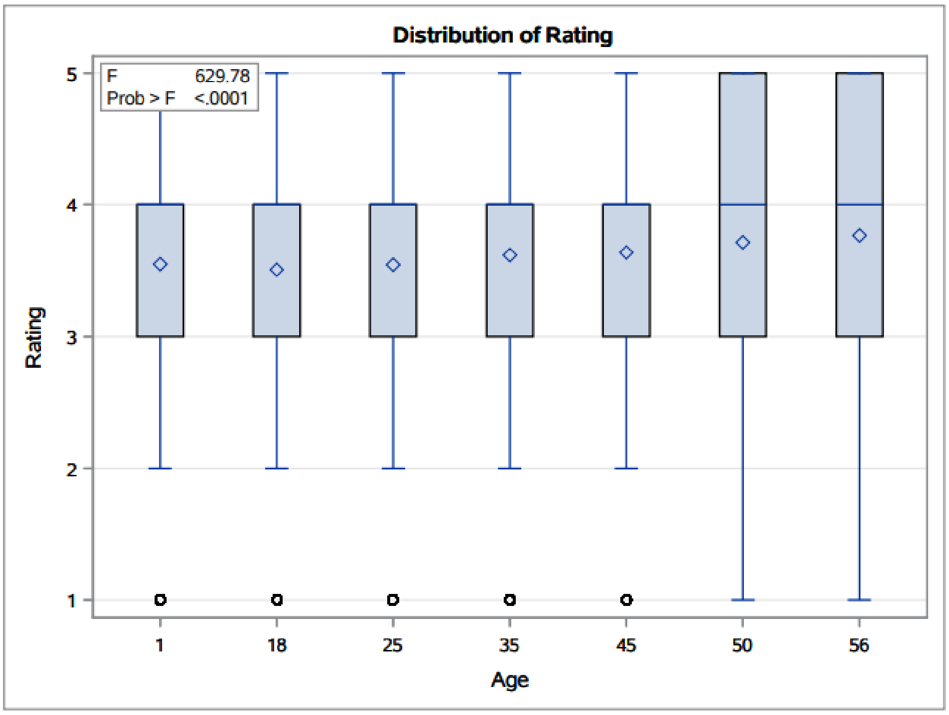


Figure 7. Rating Comparison for Age.

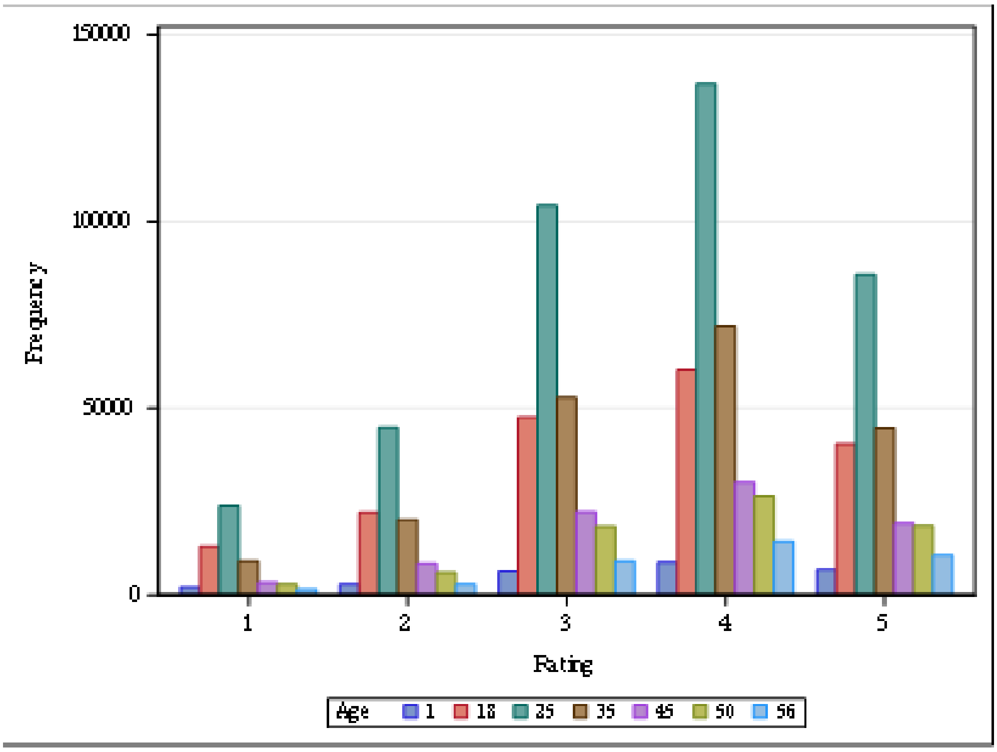
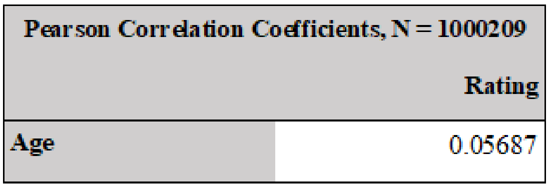


Figure 8. Pearson Correlation Coefficients



Pearson Correlation Coefficients, N = 1000209	
	Rating
Age	0.05687

Figure 9. Top Genre based on age groups.

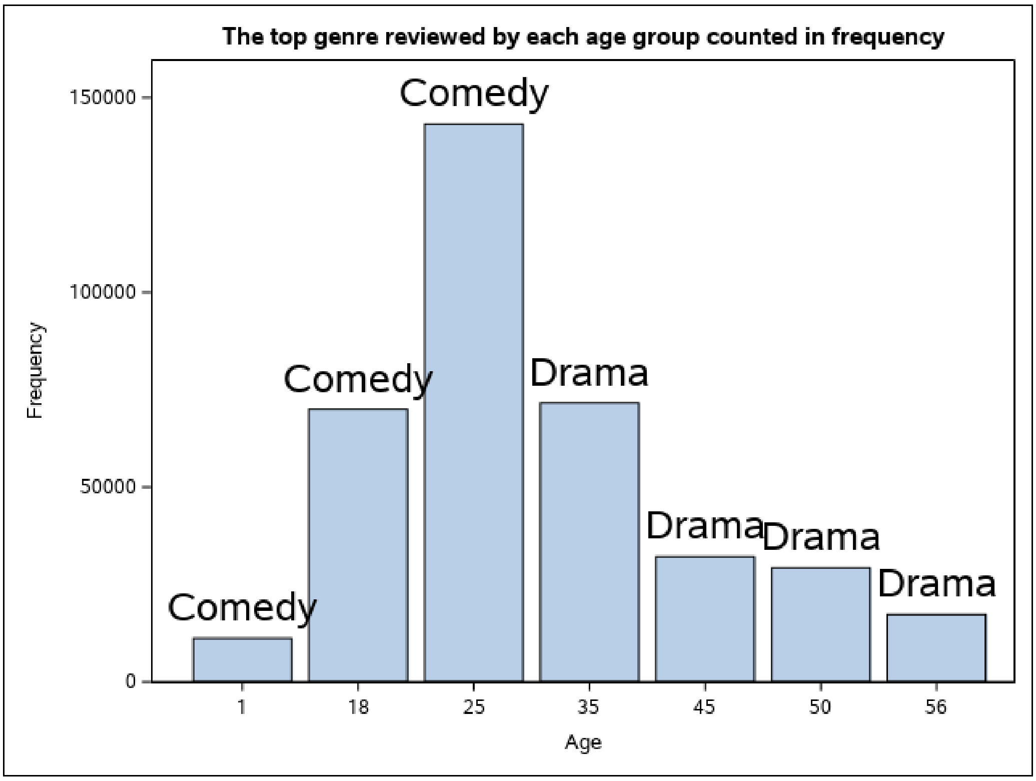


Figure 10. Genre preference in Age Group 25.

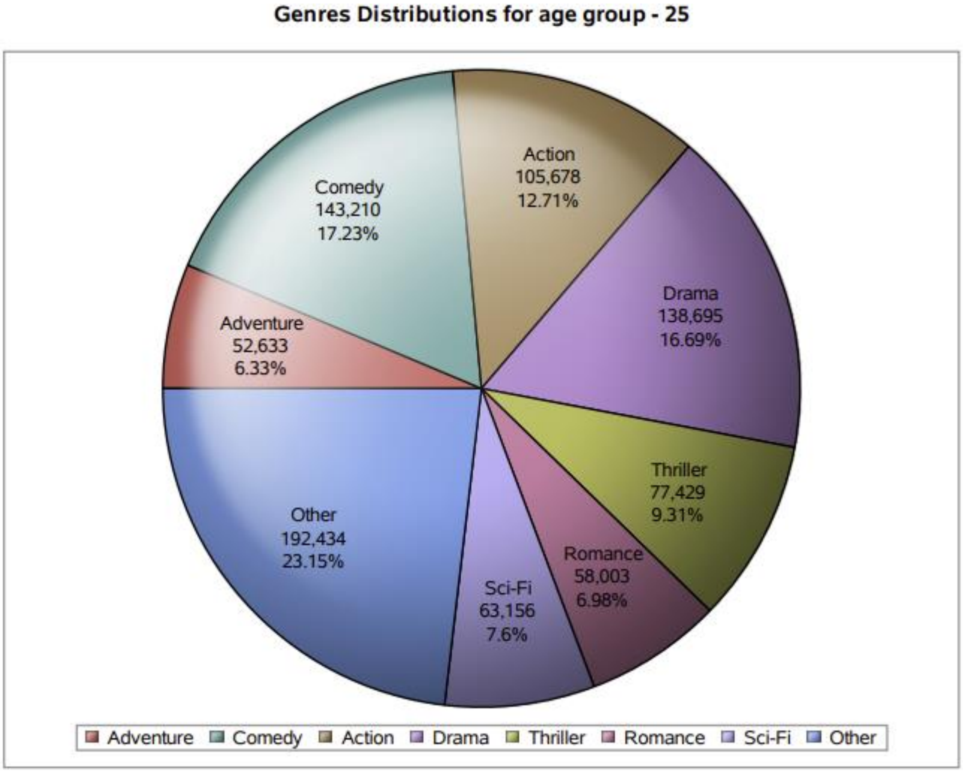


Figure 10.1 Genre preference in Age Group 35.

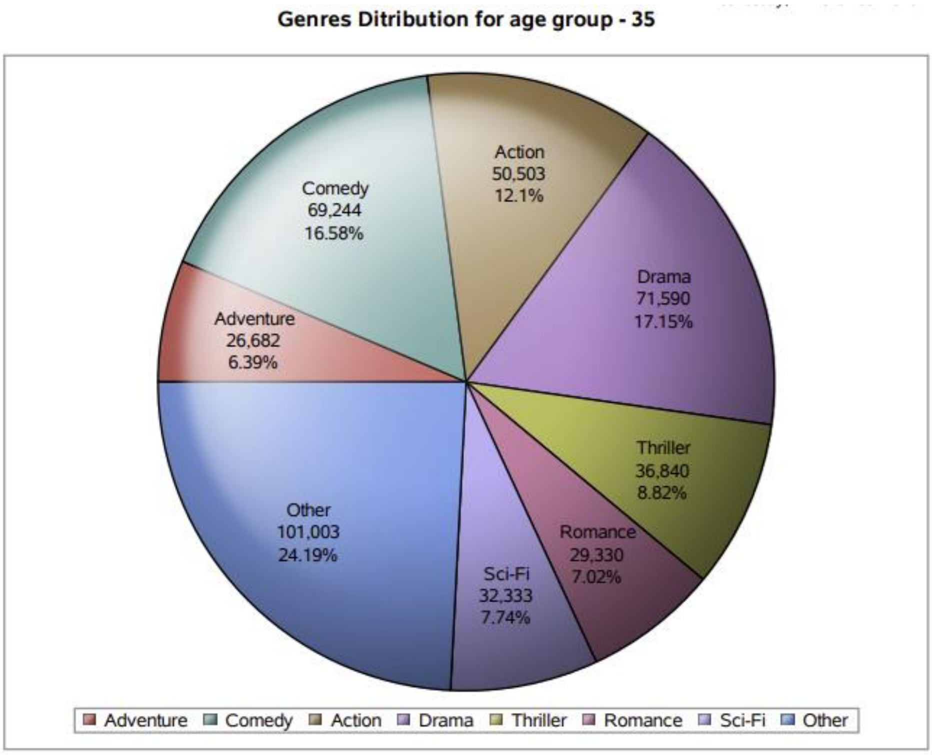


Figure 11. Gender breakdown based on User Rating

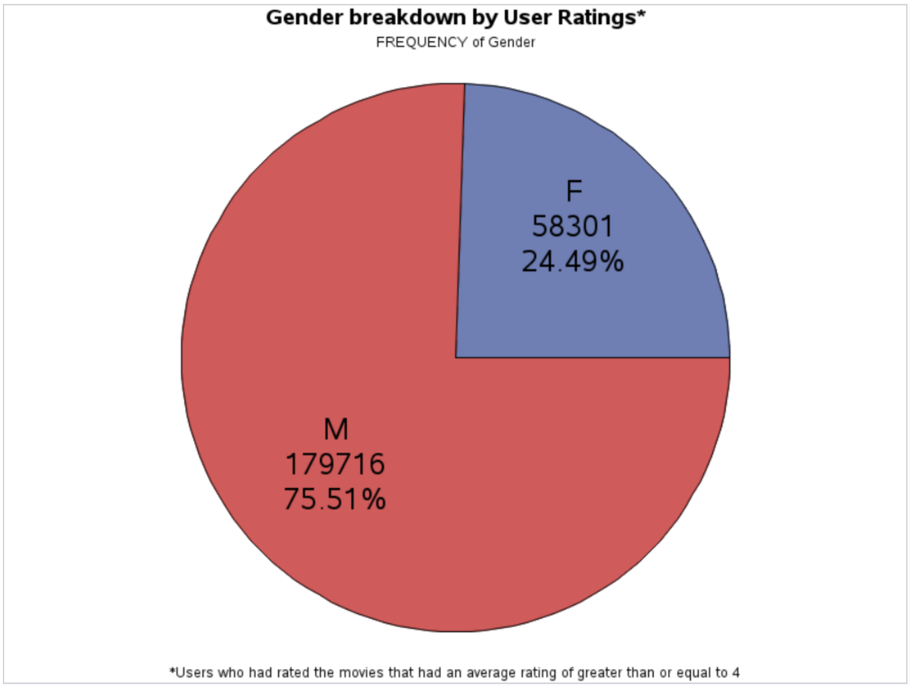
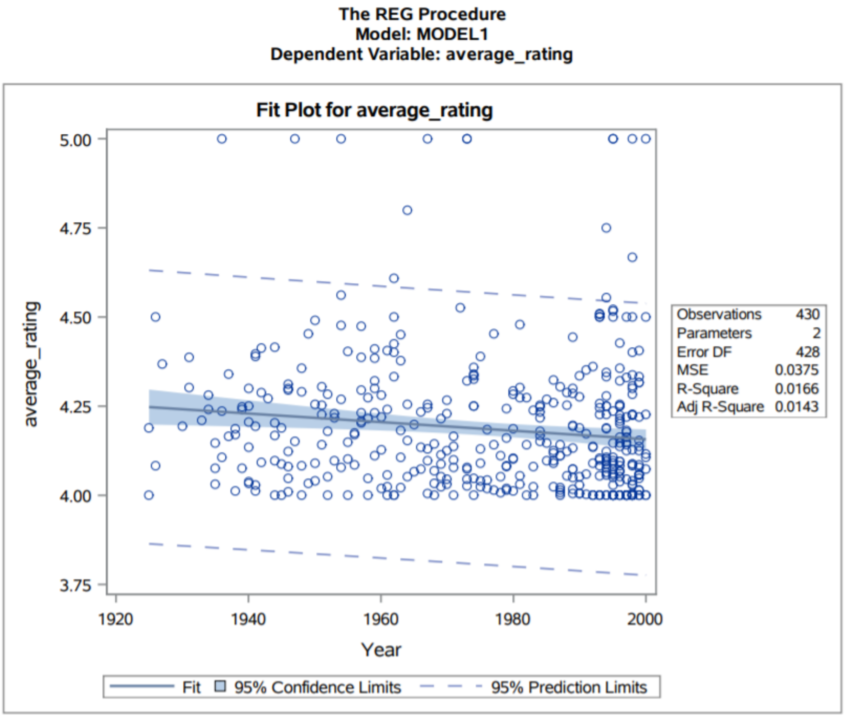


Figure 12. Average rating based on the Year of Release



The REG Procedure
Model: MODEL1
Dependent Variable: average_rating

Observations	430
Parameters	2
Error DF	428
MSE	0.0375
R-Square	0.0166
Adj R-Square	0.0143

Appendix G – MovieLens data

Categorization and representation of data in each variable are stated in the following file.

Proceed to this link