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: Chua Wen Soong **STUDENT ID: 18032573**

STUDENT NAME: Chan Wei Chee **STUDENT ID: 16052755**

STUDENT NAME: Chan Wei Wei **STUDENT ID: 16052748**

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 Chua Wen Soong(18032573) , Chan Wei Chee(16052755) , Chan Wei Wei (16052748) received the assignment and read the comments *SOONG CHEE WEI 19/11/2020*

Academic Honesty Acknowledgement

"We Chua Wen Soong(18032573) , Chan Wei Chee(16052755) , Chan Wei Wei (16052748) 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."
SOONG CHEE WEI 19/11/2020

Report: 30% contribute to final

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%

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%

Movie Analysis In Finding Preferences By Age-Group and Time

Chan Wei Chee 16052755
Chan Wei Wei 16052748
Chua Wen Soong 18032573

ABSTRACT

This paper aims to discover the patterns of the movie audience from the MoviesLens users who joined in the year 2000 [1] .There are 3 datasets which include movies, users and ratings. After conducting some data exploration, the average rating was computed to find the top genres by age-group and the top movies by quarterly basis. All age-groups were found to have a preference for Film-Noir, War and Documentary and a less favourable outlook on Horror. The quarterly results were mostly of classics.

INTRODUCTION

The film industry is booming and the cinema operators have invested a massive amount in creating demands for equally good profits. Thus, the success of a movie as a project and the factors surrounding it become an inevitable part of study before investing in the movie projects. The purpose of our analysis is to help the cinema operators to better understand the preference of their movie audiences and ultimately leads to increasing sales revenue.

After looking through the dataset, we have identified two areas of interest which could be beneficial to the cinema operators. They are the genre preference of each age-group and the quarterly analysis of ratings given. To be more specific, the first research question is “What are the top 3 movie genres preference of each age group?”. The second research question is “What are the top rated movies for each quarter?” focuses on the movie reviews according to the timeline.

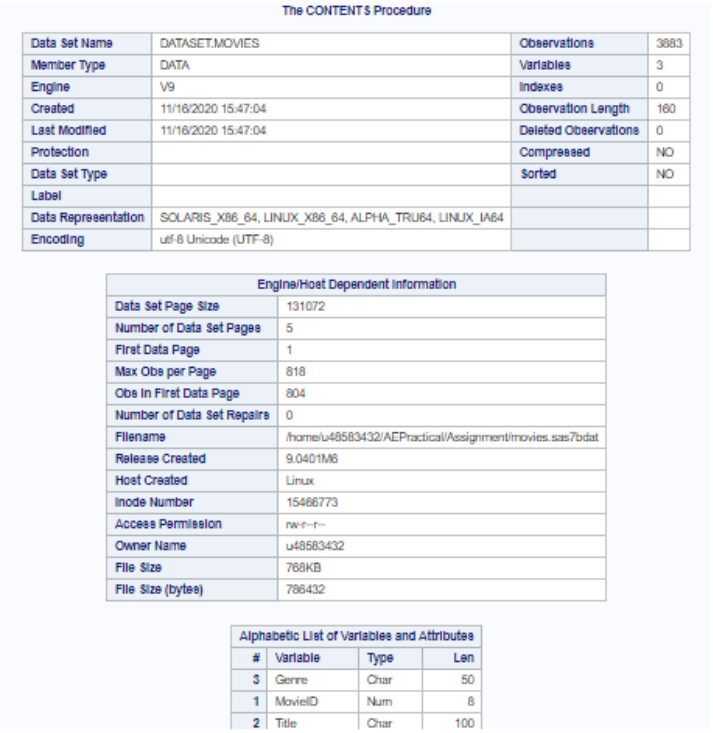
METHODOLOGY

All the SAS code statement mentioned will be included in the appendix.

A. Data Validation and Cleaning

In order to improve the accuracy of the results, validating tasks are performed. “PROC CONTENTS” statement is used to make sure the structure and content of the variables is correctly displayed.

Fig. 1 ,2 and 3 shows that all the datasets have the correct name,type and length.



The CONTENTS Procedure

Data Set Name	DATASET.MOVIES	Observations	3883
Member Type	DATA	Variables	3
Engine	V9	Indexes	0
Created	11/16/2020 15:47:04	Observation Length	160
Last Modified	11/16/2020 15:47:04	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
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	5
First Data Page	1
Max Obs per Page	818
Obs In First Data Page	804
Number of Data Set Repairs	0
Filename	/home/u48583432/AEPractical/Assignment/movies.sas7bdat
Release Created	9.0401M6
Host Created	Linux
Inode Number	15466773
Access Permission	rw-r--r--
Owner Name	u48583432
File Size	768KB
File Size (bytes)	786432

Alphabetic List of Variables and Attributes

#	Variable	Type	Len
3	Genre	Char	50
1	MovieID	Num	8
2	Title	Char	100

Fig. 1. PROC CONTENTS procedures of Movies datasets

The Movies dataset contains the movie information. There are a total of 3883 observations and 3 variables ,which are Genre of character type , MovieID of numeric type and Title of character type.

Data Set Name	DATASET.USERS	Observations	6040
Member Type	DATA	Variables	5
Engine	V9	Indexes	0
Created	11/18/2020 18:27:37	Observation Length	56
Last Modified	11/18/2020 18:27:37	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
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	3
First Data Page	1
Max Obs per Page	2334
Obs in First Data Page	2291
Number of Data Set Repairs	0
Filename	/home/u48583432/AEPractical/Assignment/users.sas7bdat
Release Created	9.0401M6
Host Created	Linux
Inode Number	15466773
Access Permission	rw-r--r--
Owner Name	u48583432
File Size	512KB
File Size (bytes)	524288

Alphabetic List of Variables and Attributes

#	Variable	Type	Len
3	Age	Num	8
2	Gender	Char	8
4	Occupation	Char	8
1	UserID	Num	8
5	Zipcode	Char	20

Fig. 2. PROC CONTENTS procedures of Users datasets

The User dataset consists of the demographic information of movie audiences. There are a total of 6040 observations and 5 variables which are Age of numeric type , Gender and Occupation of character type.

Data Set Name	DATASET.RATINGS	Observations	1000209
Member Type	DATA	Variables	6
Engine	V9	Indexes	0
Created	11/18/2020 18:27:36	Observation Length	48
Last Modified	11/18/2020 18:27:36	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
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	368
First Data Page	1
Max Obs per Page	2722
Obs in First Data Page	2670
Number of Data Set Repairs	0
Filename	/home/u48583432/AEPractical/Assignment/ratings.sas7bdat
Release Created	9.0401M6
Host Created	Linux
Inode Number	15466760
Access Permission	rw-r--r--
Owner Name	u48583432
File Size	46MB
File Size (bytes)	48365568

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format
5	Date	Num	8	DATE9.
2	MovieID	Num	8	
6	Quarter	Num	8	YYQ.
3	Rating	Num	8	
4	Timestamp	Num	8	
1	UserID	Num	8	

Fig. 3. PROC CONTENTS procedures of Ratings datasets

The Ratings dataset consists of the ratings which are made on a 5-star scale given by each user to the movies they watched. There are a total of 100209 observations and consists of 6 numeric type variables which are UserID, MovieID, Rating, Timestamp, Date with DATE9. format and Quarter with YYQ. format.

Before doing further analysis, validating tasks such as inspecting missing values is carried out to reduce biases in our analysis. “PROC FREQ” statement (line 44-52) is used to look up the number of missing values in each variable. At this stage of exploration, there are no missing values found in each dataset. However, we found that there are missing records in the merging dataset named “movie_ratings” where there are some movies that do not have ratings.

```
1          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
72
73         proc freq data=dataset.ratings;
74         where year(Date)<2000 or year(Date)>2003;
75         run;

NOTE: No observations were selected from data set DATASET.RATINGS.
NOTE: There were 0 observations read from the data set DATASET.RATINGS.
      WHERE  not (YEAR(Date)>=2000 and YEAR(Date)<=2003);
NOTE: PROCEDURE FREQ used (Total process time):
```

Fig. 4. Log file of “dataset.ratings”

The log file shows that validation on timestamp is carried out by subsetting the “WHERE” statement (line 54-57) to ensure that the dates are within the year 2000-2003 only.

```
1          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
72
73         proc sort data=dataset.movies nodupkey dupout=dup_movies;
74         by MovieID;
75         run;

NOTE: There were 3883 observations read from the data set DATASET.MOVIES.
NOTE: 0 observations with duplicate key values were deleted.
NOTE: The data set WORK.DUP_MOVIES has 0 observations and 3 variables.
NOTE: The data set DATASET.MOVIES has 3883 observations and 3 variables.
NOTE: PROCEDURE SORT used (Total process time):
```

Fig. 5. Log file of “dup_movies”

To ensure the quality of the results, “PROC SORT” statement (line 59-62) is used to check for any duplication. The log file shows that there are 0 observations in the “dup_movies” dataset , which means that there is no duplicate value for the Movies dataset.

B. Data Exploration

In this section, we are going to understand an overview distribution of each variable.

The FREQ Procedure

Genre	Frequency	Percent
Drama	1603	25.02
Comedy	1200	18.73
Action	503	7.85
Thriller	492	7.68
Romance	471	7.35
Horror	343	5.35
Adventure	283	4.42
Sci-Fi	276	4.31
Children's	251	3.92
Crime	211	3.29
War	143	2.23
Documentary	127	1.98
Musical	114	1.78
Mystery	106	1.65
Animation	105	1.64
Fantasy	68	1.06
Western	68	1.06
Film-Noir	44	0.69

Fig. 6(a). Frequency table of Movie Genres

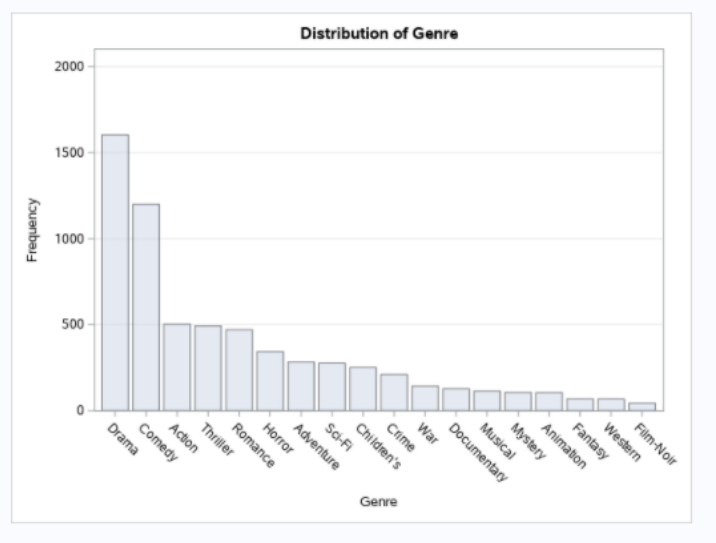


Fig. 6(b). Distribution of Movie Genres

The graph distribution of the movie genre is plotted as shown in Fig. 6(b).Notice that the pipe-separated genres are split into individual distinct genres for analysis by using "COUNTW" and "SCAN" function (line 64-74).
It shows that the movie genre of "Drama" contributes the highest percentage, followed by the movie genre of "Comedy" and "Action" to the overall movies in the datasets.

Occupation	Frequency	Percent
other or not specified	711	11.77
academic/educator	528	8.74
K-12 student	195	3.23
lawyer	129	2.14
programmer	388	6.42
retired	142	2.35
sales/marketing	302	5.00
scientist	144	2.38
self-employed	241	3.99
technician/engineer	502	8.31
tradesman/craftsman	70	1.16
unemployed	72	1.19
artist	267	4.42
writer	281	4.65
clerical/admin	173	2.86
college/grad student	759	12.57
customer service	112	1.85
doctor/health care	236	3.91
executive/managerial	679	11.24
farmer	17	0.28
homemaker	92	1.52

Fig. 7(a). Frequency table of Occupation

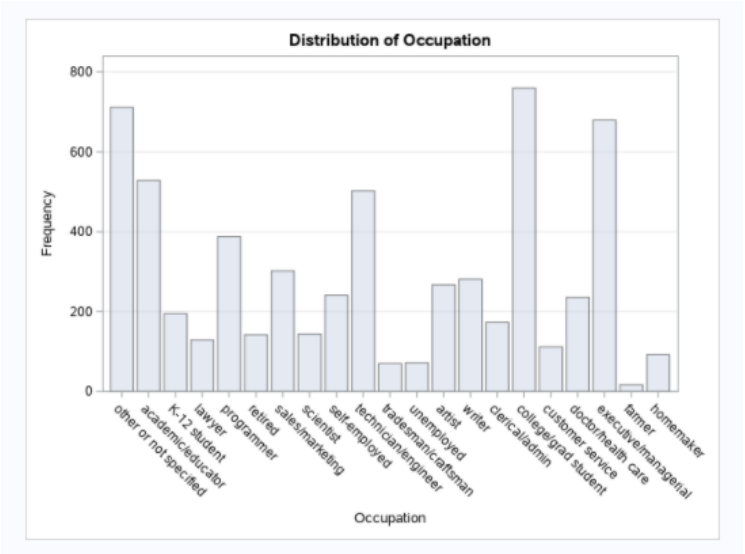


Fig. 7(b). Distribution of Occupation

The graph distribution of occupation is plotted as shown in Fig. 7(b). Notice that the occupation variable is formatted into character instead of numeric representation by using "PROC FORMAT" statement (line 81-106) for better visualization. It shows that college/grad student has the highest frequency, followed by other or not specified and executive/managerial.

The FREQ Procedure

Frequency

Table of Gender by Age

Gender	Age							
	Under 18	18-24	25-34	35-44	45-49	50-55	56+	Total
F	78	298	558	338	189	146	102	1709
M	144	805	1538	855	361	350	278	4331
Total	222	1103	2096	1193	550	496	380	6040

Fig. 8(a). Frequency table of Gender by Age

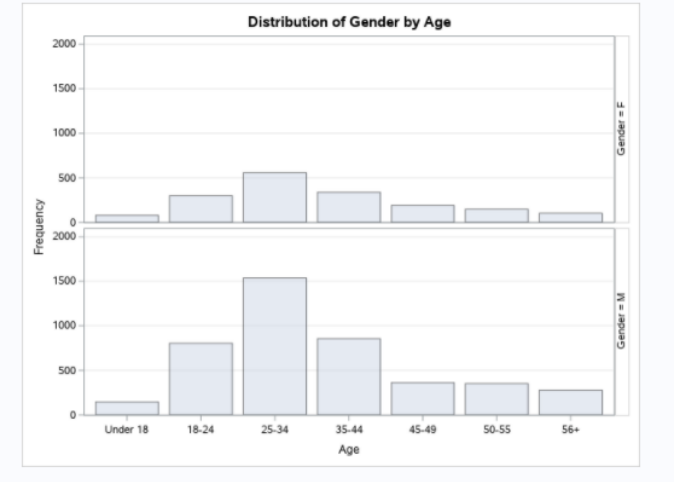


Fig. 8(b). Distribution of Gender by Age

Fig. 8(b) show that the graph distribution of Gender by Age. The Age variable is translated to the defined age group format using the "PROC FORMAT" statement. The graphs tell that the male audiences are almost 2.53 times more than the female audiences. The are more audiences who aged between 25-34 in this dataset.

The FREQ Procedure

Frequency

Table of Age by Rating

Age	Rating					
	1	2	3	4	5	Total
Under 18	2238	2983	6380	8606	6802	27211
18-24	13063	22073	47601	60241	40558	183536
25-34	23898	44817	104287	136824	85730	395556
35-44	9067	20253	52990	71983	44710	199003
45-49	3409	8437	22311	30334	19142	83633
50-55	2948	5993	18465	26484	18600	72490
56+	1551	3001	9163	14297	10768	38780
Total	56174	107557	261197	348971	226310	1000209

Fig. 9. Frequency table of Age by Rating

The frequency table of Age by Rating is plotted to visualize the distribution of rating (1-5) given by the movie audiences from each age group. Fig. 9 shows that most of the audiences have given a rating of 4 to the movie they watched.

C. Data Manipulation

In order to attempt our research questions, the movies, users and ratings datasets are merged into datasets namely 'movie_ratings'. The consequent analysis steps are carried out using the dataset. Furthermore, "IN" option is used to control the observations in the merged datasets to contain only records that exist in both datasets (line 151-157).
Otherwise, it will be output to the dataset named "Missing" as shown Fig. 10 below.

Partial Listing of Movies with No Ratings

Obs	MovieID	Title	Genre	UserID	Gender	Age	Occupation	Zipcode	Rating	Timestamp	Date	Quarter
1	51	Guardian Angel (1994)	Action|Drama|Thriller	.		.			.	.	.	.
2	109	Headless Body in Topless Bar (1995)	Comedy	.		.			.	.	.	.
3	115	Happiness Is in the Field (1995)	Comedy	.		.			.	.	.	.
4	143	Gospa (1995)	Drama	.		.			.	.	.	.
5	284	New York Cop (1996)	Action|Crime	.		.			.	.	.	.
6	285	Beyond Bedlam (1993)	Drama|Horror	.		.			.	.	.	.
7	395	Desert Winds (1995)	Drama	.		.			.	.	.	.
8	399	Girl in the Cadillac (1995)	Drama	.		.			.	.	.	.
9	400	Homage (1995)	Drama	.		.			.	.	.	.
10	403	Two Crimes (1995)	Comedy|Crime|Drama	.		.			.	.	.	.

Fig. 10. Output of "Missing" dataset

Fig. 10 shows the partial listing of movies with no ratings. There are 177 movies that do not have ratings from the movie audiences.

Top 10 Movies with Highest Number of Votes

MovieID	Title	_TYPE_	_FREQ_	AverageRating
2858	American Beauty (1999)	3	3428	4.31739
260	Star Wars: Episode IV - A New Hope (1977)	3	2991	4.45369
1196	Star Wars: Episode V - The Empire Strikes Back (1980)	3	2990	4.29296
1210	Star Wars: Episode VI - Return of the Jedi (1983)	3	2883	4.02289
480	Jurassic Park (1993)	3	2672	3.76385
2028	Saving Private Ryan (1998)	3	2653	4.33735
589	Terminator 2: Judgment Day (1991)	3	2649	4.05851
2571	Matrix, The (1999)	3	2590	4.31583
1270	Back to the Future (1985)	3	2583	3.99032
593	Silence of the Lambs, The (1991)	3	2578	4.35182

Fig. 11. Output of "avgRating" dataset

The average ratings of movies are computed by using the "PROC MEANS" with the "CLASS" statement (line 172-184). The result is then output to the "avgRating" dataset. Fig. 11 shows the listings of top 10 movies with the highest number of votes by movie audiences. Observed that MovieID 2858 American Beauty (1999) has the highest frequency while the movieID 260 Star Wars:Episode IV-A New Hope (1977) has the highest average rating of 4.45.

D. Research Findings

Research Question 1:

What are the top 3 favourite movie genres preferences of each age group?

Average Rating of Movie by Age Group

Title	AgeCat	AverageRating	RatingFrequency
American Beauty (1999)	Young Adults	4.39727	2049
American Beauty (1999)	Middle-Aged Adults	4.20702	855
American Beauty (1999)	Old Adults	4.13194	432
American Beauty (1999)	Teenager	4.43478	92
Star Wars: Episode IV - A New Hope (1977)	Young Adults	4.52426	1690
Star Wars: Episode IV - A New Hope (1977)	Middle-Aged Adults	4.36364	869
Star Wars: Episode IV - A New Hope (1977)	Old Adults	4.38671	331
Star Wars: Episode IV - A New Hope (1977)	Teenager	4.26733	101
Star Wars: Episode V - The Empire Strikes Back (1980)	Young Adults	4.40741	1755
Star Wars: Episode V - The Empire Strikes Back (1980)	Middle-Aged Adults	4.14927	824
Star Wars: Episode V - The Empire Strikes Back (1980)	Old Adults	4.06583	319
Star Wars: Episode V - The Empire Strikes Back (1980)	Teenager	4.18478	92
Star Wars: Episode VI - Return of the Jedi (1983)	Young Adults	4.10058	1720
Star Wars: Episode VI - Return of the Jedi (1983)	Middle-Aged Adults	3.90433	763
Star Wars: Episode VI - Return of the Jedi (1983)	Old Adults	3.84333	300
Star Wars: Episode VI - Return of the Jedi (1983)	Teenager	4.13000	100
Jurassic Park (1993)	Young Adults	3.71447	1541
Jurassic Park (1993)	Middle-Aged Adults	3.83832	736
Jurassic Park (1993)	Old Adults	3.82026	306
Jurassic Park (1993)	Teenager	3.80899	89
Saving Private Ryan (1998)	Young Adults	4.28974	1560
Saving Private Ryan (1998)	Middle-Aged Adults	4.38322	715
Saving Private Ryan (1998)	Old Adults	4.46622	296
Saving Private Ryan (1998)	Teenager	4.37805	82
Terminator 2: Judgment Day (1991)	Young Adults	4.03589	1616
Terminator 2: Judgment Day (1991)	Middle-Aged Adults	4.11018	717
Terminator 2: Judgment Day (1991)	Old Adults	4.01200	250
Terminator 2: Judgment Day (1991)	Teenager	4.22727	66

Fig. 12. Output of "SortedMeans" dataset

In this section, the user ratings for each movie was analysed to find the top 3 favourite movie genres preference of each age group. The initial 7 groups of age group is narrowed down and binned into 4 age categories as follow :

Under 18 = Teenager
18-34 = Young Adults
35-49 = Middle-Aged Adults
Above 50 = Old Adults

The technique used is the "IF-ELSE" statement (line 193-201). Next, the average rating of each movie title is computed and its average rating is classified into the age group of users accordingly using the "PROC MEANS" with "CLASS" statement. Fig. 12 gives an overview of how the audiences from different age groups rate one particular movie.

Partial Listing of Genre Categorised by Age Group

Genre	AgeCat	RatingFrequency	AverageRating
Action	Old Adults	25155	3.61113
Action	Middle-Aged Adults	69860	3.53546
Action	Teenager	6578	3.50638
Action	Young Adults	155864	3.45134
Adventure	Old Adults	13578	3.63507
Adventure	Middle-Aged Adults	37420	3.51921
Adventure	Teenager	3998	3.44997
Adventure	Young Adults	78957	3.43161
Animation	Old Adults	3115	3.77175
Animation	Middle-Aged Adults	11006	3.73905
Animation	Young Adults	26723	3.67156
Animation	Teenager	2449	3.47611
Children's	Old Adults	5778	3.57788
Children's	Middle-Aged Adults	19404	3.52098
Children's	Young Adults	42667	3.37427
Children's	Teenager	4337	3.24164
Comedy	Old Adults	35094	3.64826
Comedy	Middle-Aged Adults	97134	3.57054
Comedy	Teenager	11162	3.49749
Comedy	Young Adults	213190	3.48055
Crime	Old Adults	8494	3.81834
Crime	Middle-Aged Adults	20943	3.73862
Crime	Teenager	1701	3.71017
Crime	Young Adults	48403	3.67643
Documentary	Middle-Aged Adults	2395	3.95741
Documentary	Young Adults	4570	3.92757
Documentary	Old Adults	815	3.92515
Documentary	Teenager	130	3.73077
Drama	Old Adults	46516	3.89885
Drama	Teenager	7483	3.79473

Fig. 13. Output of "SortedGenreMeans" dataset

The analysis continues with digging into the genre of the movies.

The average rating of each movie genre is determined and the average rating is classified into the age group of the user accordingly by using the "PROC MEANS" with "CLASS" statement . Also the "PROC SORT" statement is used to sort the average rating by descending. Fig. 13 gives an overview of preferences of movie genres by the audiences from different age groups. For example, It showed that old adults (aged above 50) contribute the highest average ratings to the movie genre of Action while Middle-Aged Adults (aged 35-49) contribute the highest average ratings to the movie genre of Documentary.

Top 3 Favourite Movie Genres Preferences of Each Age Group

AgeGroup	Genre	AverageRating	RatingFrequency
Middle-Aged Adults	Film-Noir	4.07738	6035
Middle-Aged Adults	Documentary	3.95741	2395
Middle-Aged Adults	War	3.91979	21156
Old Adults	Film-Noir	4.15600	3077
Old Adults	War	4.00902	10089
Old Adults	Documentary	3.92515	815
Teenager	Film-Noir	4.14545	330
Teenager	War	3.89544	1578
Teenager	Drama	3.79473	7483
Young Adults	Film-Noir	4.04286	8819
Young Adults	Documentary	3.92757	4570
Young Adults	War	3.84486	35704

Fig. 14. Output of "Age_favourite" dataset

Fig. 14 shows the result of our first analysis. The table displays the top three favourite movie genres of each age group obtained from the code line 276-283.It shows that movie genres of Film-Noir, Documentary and War have gained a very good response from the movie audiences of nearly all the age groups. Teenagers prefer movie genre of Drama than Documentary.

Three Least Favourite Movie of Each Age Group

AgeGroup	Genre	AverageRating	RatingFrequency
Middle-Aged Adults	Horror	3.27203	21314
Middle-Aged Adults	Fantasy	3.49624	9701
Middle-Aged Adults	Sci-Fi	3.49648	45373
Old Adults	Horror	3.18504	6442
Old Adults	Sci-Fi	3.54353	15554
Old Adults	Fantasy	3.56650	3075
Teenager	Children's	3.24164	4337
Teenager	Horror	3.25418	2211
Teenager	Fantasy	3.31765	1360
Young Adults	Horror	3.19113	46419
Young Adults	Children's	3.37427	42667
Young Adults	Fantasy	3.41741	22165

Fig. 15. Output of "Age_dislike" dataset

Fig. 15 displays the three least favourite movie genres of each age group. It shows that movie genres of Horror,Fantasy,Sci-Fi are not favourable by the audience from nearly all the age groups. The observation of Teenagers dislike the movie genre of Children's is out of our expectation as we assumed that normally the users aged under 12 would prefer the movie genre of Children's. This could be due to most of the users taking this survey are aged above 13-18.

E. Discussion of Findings 1

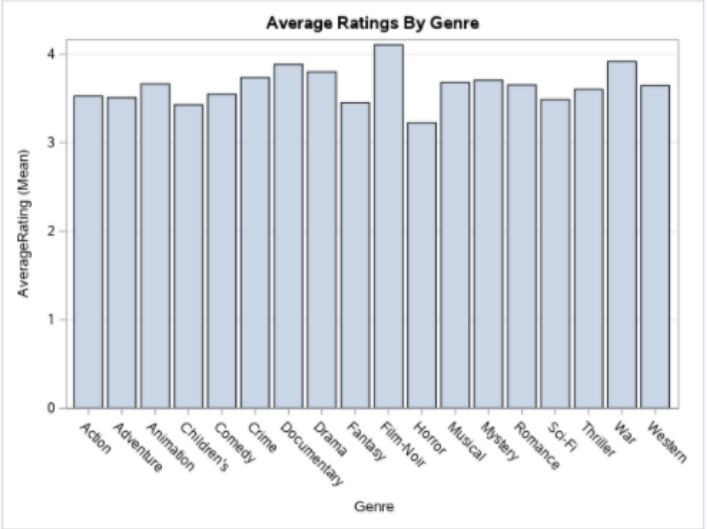


Fig. 16. Tables of Average Ratings by Genre

Out of all the movie genres, the movie genres of Film Noir, War and Documentary are well accepted and preferred by the audience from all levels of age group. It is interesting to find out that these three types of genre are considered as the most popular genres despite they contain only 5% rating frequency out of the movies genres. We have come to the conclusion that even though the rating frequency of the movie genre is high does not imply that it is the most popular movie. For example, genre Drama has the highest frequency in the dataset but its average rating is not as high as genre Film-Noir which only has a very low frequency.

Research Question 2:

What are the top rated movies for each quarter?

In this section, the movie reviews for each quarter in the time period of collection was analysed to find the top rated movies for each quarter. The time period of collection is between 25 April 2000 and 28 February 2003. The reviews for each movie were then binned into quarterly periods for quarterly analysis. 12 bins/quarters were created as a result.

The UNIVARIATE Procedure
Variable: AverageRating

Moments			
N	27697	Sum Weights	27697
Mean	3.27818885	Sum Observations	90795.9966
Std Deviation	0.85822431	Variance	0.73654896
Skewness	-0.4691451	Kurtosis	0.125561
Uncorrected SS	318045.884	Corrected SS	20399.4601
Coeff Variation	26.1798312	Std Error Mean	0.00515685

Fig. 17(a). Summary statistics of AverageRating

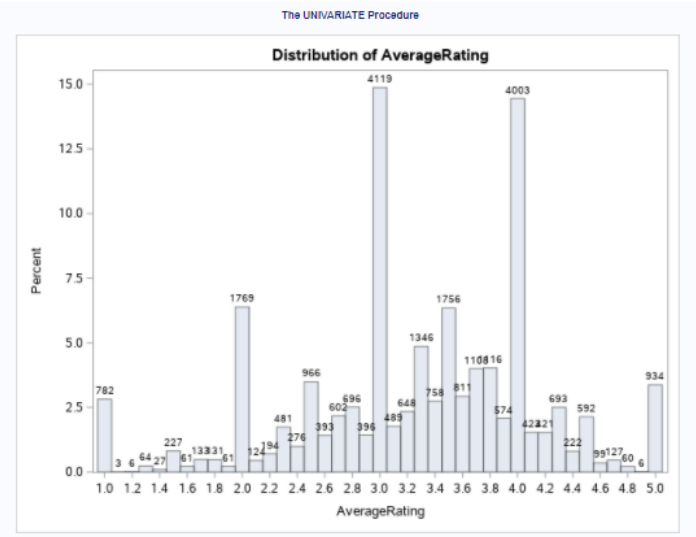


Fig. 17(b). Distribution of average rating after binning into 12 quarters.

Fig. 17(a) and (b) displays some simple statistics about the average rating for each movie that was binned into their quarters.After binning, there were 27,697 observations distributed into 12 bins.

The mean average rating for all the movies in each quarter is 3.278. However, the histogram shows some points of concern at the average rating of 3.0 and 4.0. These two ratings contain nearly 30% of the observations. The average rating of 5.0 also needs further investigation as there are 934 movies with a rating of 5.0.

The UNIVARIATE Procedure
Variable: RatingFrequency

Moments			
N	27697	Sum Weights	27697
Mean	11.4217063	Sum Observations	316347
Std Deviation	18.6907281	Variance	349.343316
Skewness	2.64874996	Kurtosis	6.83940856
Uncorrected SS	13288635	Corrected SS	9675412.47
Coeff Variation	163.64217	Std Error Mean	0.11230778

Fig. 18(a). Summary statistics of RatingFrequency

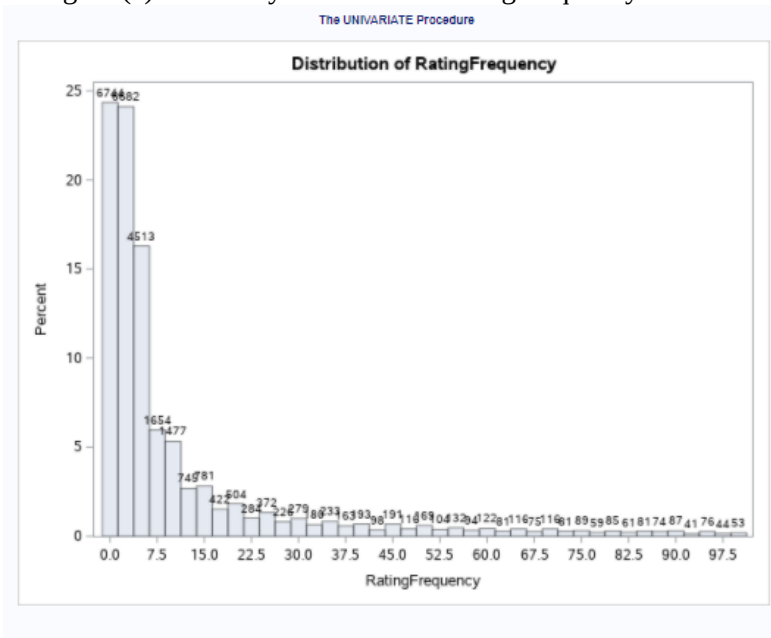


Fig. 18(b). Rating Frequency distribution for each quarter's movie. The histogram views from 0 to 100 rating frequency only.

The rating frequency histogram for each movie in every quarter in Fig. 18(b) shows an extremely right-skewed distribution. More than half of the movies have a rating frequency less than 10. The low frequency of ratings per movie will cause high bias in average rating. For example, a movie with 2 ratings of 5.0 will give an average rating of 5.0, whereas a movie with 90 ratings would in most cases have an average rating less than 5.0. Thus, a frequency threshold of 10 ratings was set to reduce this bias, while also accommodating movies for every quarter. The graphs is plotted using SAS statement in line 314-340.

Variable: AverageRating

Moments			
N	10171	Sum Weights	10171
Mean	3.39832902	Sum Observations	34564.4044
Std Deviation	0.61689254	Variance	0.38055641
Skewness	-0.5976964	Kurtosis	0.00743344
Uncorrected SS	121331.477	Corrected SS	3870.25869
Coeff Variation	18.1528198	Std Error Mean	0.00611685

Fig. 19(a). Summary statistics of AverageRating

Variable: RatingFrequency

Moments			
N	10171	Sum Weights	10171
Mean	92.2148265	Sum Observations	937917
Std Deviation	136.339371	Variance	18588.4241
Skewness	3.58533403	Kurtosis	18.6740465
Uncorrected SS	275534127	Corrected SS	189044274
Coeff Variation	147.849729	Std Error Mean	1.3518841

Fig. 20(a). Summary statistics of AverageRating

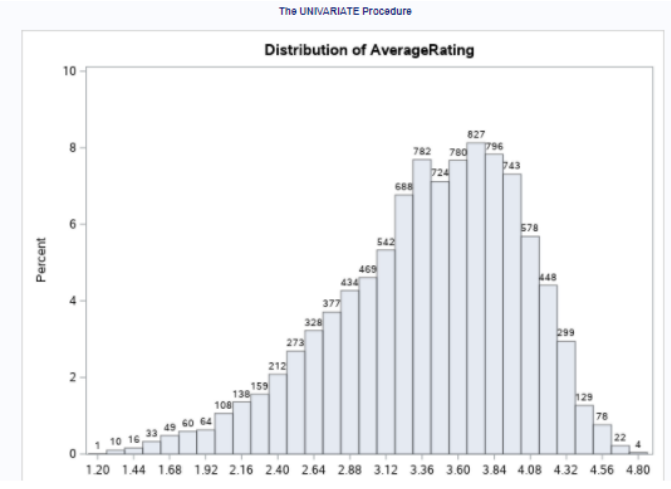


Fig. 19(b). Distribution of average rating for the average rating of movies with a frequency threshold of 10.

Fig. 19(b) shows the distribution of reduced dataset with movies having a rating frequency of 10 or more. The number of observations is reduced to nearly a third, at n = 10, 171. However, the distribution of average ratings is now an approximately normal distribution. There are also no movies with an average rating of 5.0. Thus, the movies can now be compared by their means and the top movies chosen. The graphs is plotted using SAS statement in line 342-357.

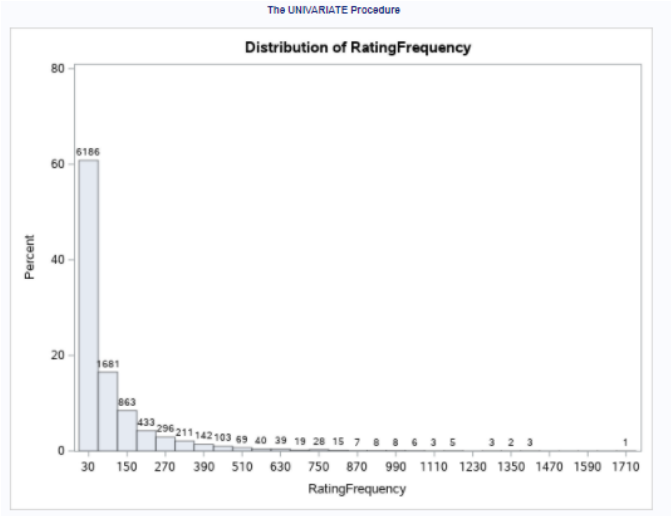


Fig. 20(b). Distribution of rating frequency for movies having 10 or more ratings.

Fig. 20(b) shows that the distribution of rating frequency is still extremely skewed to the right. The frequency threshold may be increased to improve fairness in comparing rating means but at the cost of further reducing the dataset. In this analysis, we proceed with movies having a rating frequency of 10 or more.

F. Discussion of Findings 2

Top Rated Movie for each Quarter by Year

Obs	Title	Quarter	RatingFrequency	AverageRating
1	Sanjuro (1962)	2000Q2	12	4.83333
2	World of Apu, The (Apur Sansar) (1959)	2000Q3	11	4.63636
3	Sanjuro (1962)	2000Q4	33	4.63636
4	Seven Samurai (The Magnificent Seven) (Shichinin no samurai) (1954)	2001Q1	13	4.76923
5	Iron Giant, The (1999)	2001Q2	10	4.70000
6	Sting, The (1973)	2001Q3	12	4.75000
7	Godfather, The (1972)	2001Q4	23	4.82609
8	Raiders of the Lost Ark (1981)	2002Q1	13	4.53846
9	Usual Suspects, The (1995)	2002Q2	16	4.68750
10	North by Northwest (1959)	2002Q3	11	4.54545
11	Star Wars: Episode V - The Empire Strikes Back (1980)	2002Q4	12	4.66667
12	Saving Private Ryan (1998)	2003Q1	10	4.50000

Fig. 21. The top rated movie for each quarter.

Fig. 21 shows the result of our analysis using SAS statement in line 359-370. The top movies for each quarter have an average rating of 4.50 or more. The rating frequencies for the movies is between 10 and 33.

The result of this analysis is quite surprising. From the data, most of the top movies are released 5 years before the time of review. This could imply that people rate old movies more favourably compared to new releases. Furthermore, the majority of these movies can be considered classics. It is not a stretch to say that these classics have a cult following and are rated highly by these groups of people. A point of contention is that the low rating frequency may be a cause of concern. A fairer comparison of top movies could be made with a higher rating frequency threshold, however it comes at the cost of not having movie entries for all quarters.

CONCLUSION

To conclude, we have performed an analysis of the dataset from MovieLens using SAS Studio and produced some insights on it. The first, is that the movie genres, Film-Noir, War and Documentary are highly rated among all age-groups. This highlights that generally, films with these genres do better than those without. Furthermore, the movie genre Horror, has a low average rating in all age-groups, possibly hinting that it only appeals to a niche market. In the quarterly analysis section, the top rated movies were released a few years before the time of rating. These movies have become classics over time and may have biased ratings by their fans. Ultimately, these findings are useful for the cinema operators to know which movie ranks top for each quarter of the year and facilitate decision making in determining which type of movie genres to be screened in the future for optimum ticket sales performances.

REFERENCES

[1] F.Maxwell Harper and Joseph A.Konstan. 2015.The MovieLens Datasets:History and Context. ACM Transactions on Interactive Intelligent Systems (TiiS) 5, 4,Article 19 (December 2015), 19 pages. DOI=http://dx.doi.org/10.1145/2827872

APPENDIX (Partial SAS code)

Line 44-52

```
*Check for missing value for user dataset;
proc freq data=dataset.users;
	tables UserID Gender Age Occupation ZipCode
/nocum nopercent;
run;

*Check for missing value for movie dataset;
proc freq data=dataset.movies ;
	tables MovieID Title Genre /nocum nopercent;
run;
```

```
Line 54-57
*Validate timestamp;
proc freq data=dataset.ratings;
	where year(Date)<2000 or year(Date)>2003;
run;

Line 59-62
*Check for duplicates movie;
proc sort data=dataset.movies nodupkey
dupout=dup_movies;
	by MovieID;
run;
```

Line 64-74

```
*Explore the genre of movie;
*seperate the pipe-seperated genre;
Data splitgenre;
	set dataset.movies;
	save=genre;
		do i=1 to countw(genre,'|');
			genre=scan(save,i,'|');
			output;
		end;
		drop i save;
run;
```

Line81-106

```
*Explore the occupation of the audience;
*Format occupation;
proc format;
	value $ OCP
                                "0"= 'other or not
specified'
                                "1"= 'academic/educator'
                                "2"= 'artist'
                                "3"= 'clerical/admin'
                                "4"= 'college/grad
student'
                                "5"= 'customer service'
                                "6"= 'doctor/health care'
                                "7"=
'executive/managerial'
                                "8"= 'farmer'
                                "9"= 'homemaker'
                                "10"= 'K-12 student'
                                "11"= 'lawyer'
                                "12"= 'programmer'
                                "13"= 'retired'
                                "14"= 'sales/marketing'
                                "15"= 'scientist'
                                "16"= 'self-employed'
                                "17"=
'technician/engineer'
                                "18"=
'tradesman/craftsman'
                                "19"= 'unemployed'
                                "20"= 'writer';
run;
```

Line 151-157

```
*Keep only records that exists in both datasets;
Data movie_ratings Missing;
	merge dataset.movies (IN=a) user_ratings
(IN=b);
		by MovieID;
	if not(a=1 and b=1) then output Missing;
		else output movie_ratings;
Run;
```

Line 172-184

```
*Count the average rating of each movies ordered
by descending frequency;
proc means data=movie_ratings noprint order=freq
n mean;
	class MovieID Title;
	var Rating;
	output out=avgRating mean=AverageRating;
run;

title 'Top 10 Movies with Highest Number of
Votes';
proc print data =avgRating (obs=10);
	ID MovieID;
	where title ^='' and MovieID ^=.;
run;
title;
```

Line 193-201

```
/** Binning age group**/
Data AgeCategory(keep=MovieID Title Genre UserID
Age AgeCat Rating);
	set movie_ratings;
	length AgeCat $20;
			if Age < 18 then AgeCat
="Teenager";
				else if 18<= Age
<=34 then AgeCat ="Young Adults";
					else if 35
<= Age <= 49  then AgeCat ="Middle-Aged Adults";

	else AgeCat = "Old Adults";
run;
```

```
Line 276-283
*Picking top 3 movie genre per age group;
data Age_favourite;
	do _n_ = 1 by 1 until (Last.AgeCat);
		set AgeGenreMeans;
		by AgeCat;
		if _n_ <= 3 then output;
	end;
Run;
```

Line 314-340

```
Data movie_quarter (keep= title genre rating
quarter);
	set movie_ratings;
run;

*sort the movie title by quarter;
proc sort data=movie_quarter
out=sortedMovieQuarter;
	by quarter ;
run;
proc means data=sortedMovieQuarter noprint
maxdec=2 order=freq;
	class quarter title;
	var Rating;
	output out=QuarterMeans n=RatingFrequency
mean=AverageRating;
run;
```

```
*removing unnecessary rows;
Data QuarterMeans_Clean;
   set QuarterMeans;
   if title ^='' and quarter ^=. then output;
run;

*Distribution of frequency and average rating;
proc univariate data= QuarterMeans_Clean;
        var AverageRating RatingFrequency;
        where RatingFrequency between 0 and 100;
        histogram / barlabel=count;
run;
```

Line 342-357

```
*Setting threshold=10 for rating frequency ;
Data QuarterMeans_over10;
        set QuarterMeans_Clean;
        if RatingFrequency >=10  then output;
run;

*Distribution of frequency and average rating with
threshold;
proc univariate data= QuarterMeans_over10;
        var AverageRating RatingFrequency;
        histogram/barlabel=count;
run;

*Sorting quarter by descending average rating;
proc sort data=QuarterMeans_over10
out=SortedQuarterMeans;
        by quarter descending AverageRating;
run;
```

Line 359-370

```
*Display only the movie title with the highest average
rating for each quarter of year;
Data Title_Genre;
        set SortedQuarterMeans;
        by quarter;
        if First.quarter;
run;

title 'Top Rated Movie for each Quarter by Year';
proc print data=Title_Genre;
        var title quarter RatingFrequency
AverageRating;
run;
title;
```