Anh Vo

STA 402 Section B

Dr. Steven Wright

I. Original statement for the assigned task:

The file genome-scores.csv contains tag relevance scores for different movies. For each movie, select the tags that correspond to the highest 5 relevance scores and then summarize them in a few keywords. Link these summarized keywords with the ratings the movie gets. Write a SAS macro so that when the user specifies a genre, your program generates graphs/tables that show the most popular keywords (you may use the average rating or other measures that make sense) and a few suggested movies for that genre).

Description of Dataset:

Six datasets in csv format were given with the following names: genome-scores.csv, genome-tags.csv, tags.csv, movies.csv, links.csv, and ratings.csv. After analyzing the datasets, I have made this summary table to easily identify what the datasets have in common for merging and selecting.

VARS CSV FILES	**movieId**	**tagId**	**tag**	**title**	**genre**	**ImdbId**	**tmdId**	**userId**	**timestamp**	**relevance**	**ratings**
genome_scores	X	X									
genome_tags		X	X							X	
tags	X		X					X	X		
links	X					X	X				
movies	X			X	X						
ratings	X							X	X		X

movieId : the ID given to the movie

tagId : the ID given to the tags

tags : the names of the tags, basically keywords to describe the movies

title : movie titles

genre : genre of the movies. One movie can fall to many different genres

ImdbId/tmbID: Id of IMDB scores and TMB

userId : ID of users that gave ratings to the movies

timestamp : timestamp of recorded ratings

relevance : relevance scores of the movies

ratings : ratings given by users

Google drive of the datasets and the README file:

https://drive.google.com/drive/folders/1oH4YluvtpvcubHdupYAA5_AhbkRuL6Nx.

II. Progress report

1. Project description:

I was given 6 data files, all in csv format, regarding the information of over thousands of movies and their additional information, such as genre, IMBD score, etc. My task was to create a program that accept a user specified genre, given them a user-specified number of tags/keywords that they are relevant to the genres, and a user specified number of movies relevant to the genre and keywords. I also have to create tables or graphs that reflect my findings.

2. Progress Updated:

A. Import datasets:

- Use proc import to import all six datasets in.
- Specify a datafile in proc import and declare the path for getting the csv files
- Specify an output file to use within the SAS program
- Use dbms to tell SAS the format of the data file
- Replace/ Overwrite any existing SAS data set with the name of the output file

IMPROVEMENT SUGGESTION:

- Create a macro to import datasets
- Save spaces in the SAS program
- Figure out the number of observations to reduce collisions and help the program run better

B. Merging genome_scores and genome_tags

- Sort genome_scores by tagId
- Merge the sorted genome_scores dataset and genome_tags dataset by tagId
- Sort the new dataset by movieId and descending relevance scores
- Dataset is named `tags_merged_sortted`

IMPROVEMENT SUGGESTION:

- Explore all parts of the new merged dataset
- Get information about rows and NAs

C. Transpose dataset to get top 5 tags

- Use do loops to transpose the tags_merged_sortted dataset
- Basically created 1128 arrays, as there are 1128 tags for each of the movie

- Get 5 tags related to each movies with the highest relevance scores
- Transfer this information to a new dataset called getTags

IMPROVEMENT SUGGESTION:

- Create a macro with user-specified number of tags that they want to generate
- Build a plot/table that display this information

D. Remove NAs

- Remove movies that doesn't have any tags
- These movies do have relevance scores, but there are no keywords

IMPROVEMENT SUGGESTION:

- Same with Section C, create a macro with user-specified number of tags that they want to generate, and these movies may have low relevant tags
- Generate how many NAs that were removed

E. Get average ratings for each movie

- Sort ratings of the movies by movieID, since these are rated at different timestamp
- Use proc means to compute the average ratings for each movies
- Remove/drop irrelevant variables
- The filtered dataset that the mean was sent to is meanRatings

IMPROVEMENT SUGGESTION:

- Use proc tabulate instead of proc means, as there are irrelevant data in proc means
- Figure out if there is any NAs

F. Build macro to plot average ratings

- Build a macro that takes a user-specified number of movies and the range of movies that the user wants to see the average ratings of
- Create a bar plot using proc sgplot

IMPROVEMENT SUGGESTION:

- Get the user-specified movie name instead of movieID
- Create a more lively plot

G. Merge tags and average ratings

- Merge the getTags dataset with the average ratings dataset, so that the ratings can link with the tags by movieId
- Merge this new dataset with the movies dataset, which contains the titles and genres of movies
- Final dataset should have movieId, title, genre, 5 relevant tags, average rating

IMPROVEMENT SUGGESTION:

- As in earlier suggesting, this final dataset should have user specified number of tags
- Remove NAs
- Figure out the number of observations

III. Plans for proceeding

- Create macros to reduce space in SAS program
- Build a user-specified macro with user-specified genre, and through that genre I will generate a user-specified number of movies and relevant keywords.
- Build a macro/data step that look into the genre of the movies and decide what genres this movie belongs to, since a movie can belong to many different genres, and when the user specify a genre it has to look for the genre inside a long string
- Build more plots related to the movies/keywords in order to be more trustworthy to the user
- Building a more user-friendly SAS program, which requires a lot of macros and user-specified section
- Using more concise and efficient steps rather than focusing on procs only

IV. Questions for the instructor

- I was having a hard time trying to think of what visuals rather than just tables that I can bring/visualize to the users. What type of plots and what kind of variables does it require to create more tables/visualization?
- Does proc tabulate, as in the last week's lessons, really more efficient than proc means? Proc means prints out a lot of unsuable variables, but it did compute the average ratings scores right.
- Is there a way to print only the first 20 observations of the table created in proc sql? I have known that proc sql is interactive, and when I used (obs = 20) I thought that I printed out only the first 20 observations of the table, but it printed all 1M of them, which almost crashed my program. I can try and select file (firstobs = obs =) of the data, but that won't perform the average on all observations in the dataset
- Is there a way, other than creating a macro, to import all 6 datasets using proc import?

V. Codes and visuals (Each Alphabet letter related to the Alphabet letter in Section II.

A. Import datasets:

- **Coding:**

```
/*The following 6 import statements import all 6
datasets: genome_scores, genome_tags, tags,
movies, links, ratings.
It takes in the path for datafile and specify
the type as CSV files*/
proc import
      datafile= "M:\sta402\TermProject\genome-scores.csv" /* this is the CSV
                                          file we want to read */
      out = genome_scores /* this is a new SAS data set */
      dbms=csv      /* tell SAS this really is a CSV file */
```

```
      replace;      /* tell SAS to overwrite any existing
                   SAS data set called genome_scores*/
run;

proc import
      datafile= "M:\sta402\TermProject\genome-tags.csv"
      out = genome_tags
      dbms=csv
      replace;
run;

proc import
      datafile= "M:\sta402\TermProject\movies.csv"
      out = movies
      dbms=csv
      replace;
run;

proc import
      datafile= "M:\sta402\TermProject\links.csv"
      out = links
      dbms=csv
      replace;
run;

proc import
      datafile= "M:\sta402\TermProject\tags.csv"
      out = tags
      dbms=csv
      replace;
run;

proc import
      datafile= "M:\sta402\TermProject\ratings.csv"
      out = ratings
      dbms=csv
      replace;
run;
```

B. Merging genome_scores and genome_tags

- **Coding**

```
/*Sort the genome scores dataset by tagId*/
proc sort data = genome_scores out = genome_scores_sortted;
      by tagId;

/*Merge the new sorted genome scores with genome tags
New dataset is called tags_merged*/
data tags_merged;
      merge genome_scores_sortted genome_tags;
      by tagId;
run;

/*Sort the tags_merged dataset by movieID and descending
```

```
relevance score*/
proc sort data = tags_merged out = tags_merged_sortted;
      by movieID descending relevance;

ods rtf bodytitle file =  "M:\sta402\TermProject\Tester.rtf";
/*Print out the first 10 observations*/
title "Tags sorted by relevance scores";
proc print data =  tags_merged_sortted(obs = 10);
run;
ods rtf close;
```

- **Table showing merged results:**

Tags sorted by relevance scores

Obs	movieId	tagId	relevance	tag
1	1	1036	0.9995	toys
2	1	244	0.999	computer animat
3	1	786	0.9955	pixar animation
4	1	64	0.98875	animation
5	1	589	0.9885	kids and family
6	1	588	0.9785	kids
7	1	785	0.96975	pixar
8	1	186	0.9585	cartoon
9	1	63	0.955	animated
10	1	204	0.9545	children

C. Transpose dataset to get top 5 tags

```
/*This data step transpose the tags_merged_sortted
dataset and get the 5 tags with highest relevance
score*/
data getTags;
      do k=1 to 187595;
            array C{1128} $;          /*array of characters*/
            do i=1 to 1128;
                  set tags_merged_sortted;
                  C{i} = tag;

            end;
```

```
            output;
      end;
      drop C6-C1128;                /*Top 5, we don't need
                                    6 to 1128*/
      drop i k tag tagId relevance;
run;
```

D. Remove NAs

```
/*This datastep remove the NAs variables within the
getTags dataset*/
data no_nas;
      set getTags;
      if cmiss(of C1-C5) then delete;
run;

/*This data step sort the ratings dataset
by movieId*/
proc sort data = ratings out = ratings_sortted;
      by movieID;
```

E. Get average ratings for each movie

```
/*This data step compute the average ratings group by
movieIDs*/
proc means data=ratings_sortted noprint;
      class movieId;
      var rating;
      output out = mean_ratings mean = Avg;
run;

/*This data step filter the mean_ratings dataset,
remove unrelated variables*/
data meanRatings;
      set mean_ratings;
      where movieId > 0;
      drop _TYPE_ _FREQ_;
run;
```

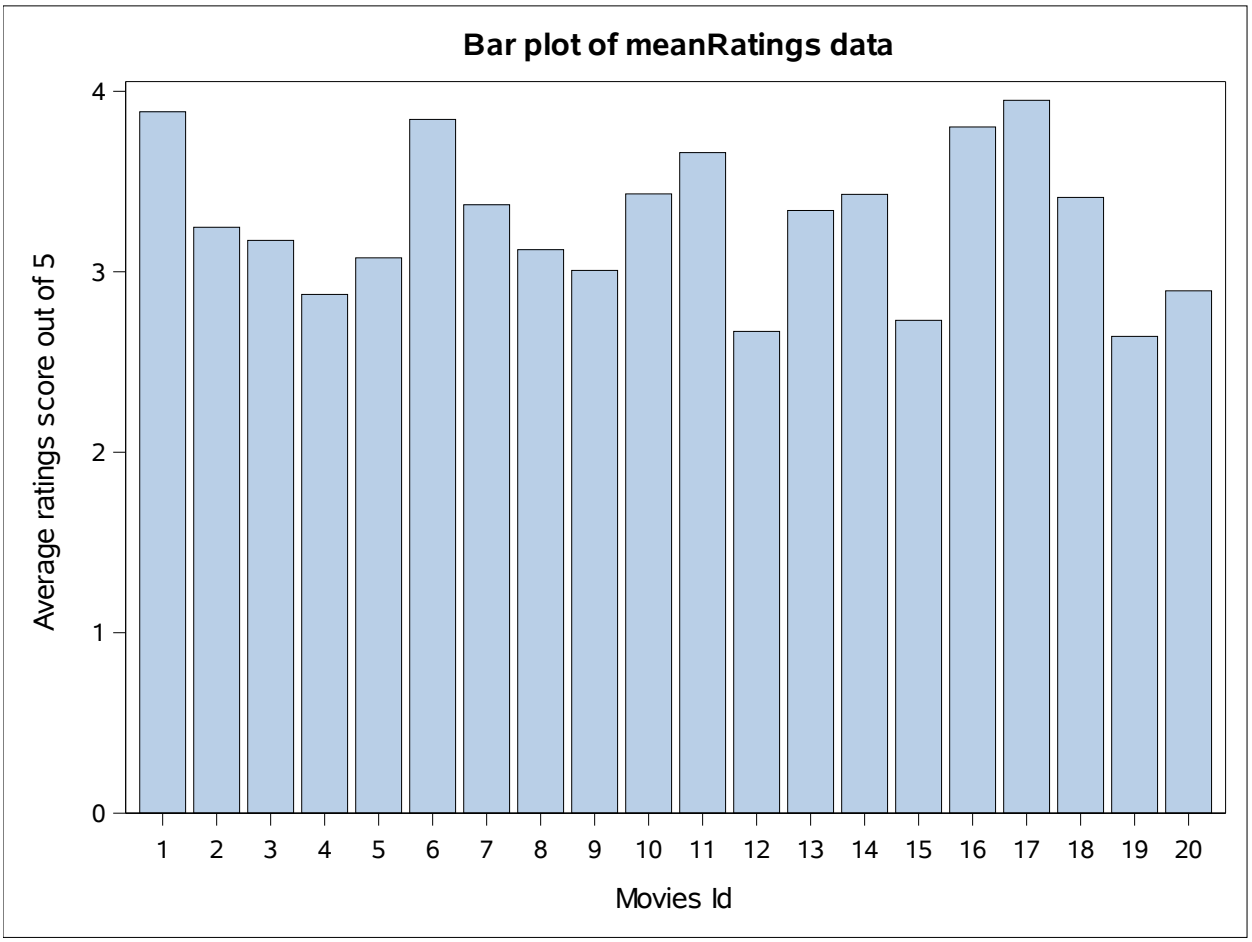
F. Build macro to plot average ratings

```
/*This macro create a user specified range
of mean ratings of movies for user to see*/
```

```
%macro rating_graphs(numMovies=, start=1);
      proc sgplot data=meanRatings(firstobs = &start obs = &numMovies);
            vbar movieId / response = Avg;
            yaxis label = "Average ratings score out of 5";
            xaxis label = "Movies Id";
      run;
%mend rating_graphs;

ods rtf bodytitle file =  "M:\sta402\TermProject\Tester.rtf";
title "Bar plot of meanRatings dataset";
%rating_graphs(numMovies= 20, start=1);
ods rtf close;
```

- **Plot:**



G. Merge tags and average ratings

```
/*This dataset merge the getTags dataset and meanRatings
to link tags to ratings by movieId*/
```

```
data ratings_merged;
      merge getTags meanRatings;
      by movieId;
run;

/*This dataset merge the movies dataset to the
ratings_merged dataset by movieId*/
data genres_merged;
      merge movies ratings_merged ;
      by movieId;
run;

ods rtf bodytitle file =  "M:\sta402\TermProject\Tester2.rtf";
/*Print out the first 20 observations of the final file*/
title "Final files for building macros";

proc print data = genres_merged(obs = 20);
run;
ods rtf close;
```

- **Table:**

Final files for building macros

Obs	movieId	title	genres
1	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy
2	2	Jumanji (1995)	Adventure|Children|Fantasy
3	3	Grumpier Old Men (1995)	Comedy|Romance
4	4	Waiting to Exhale (1995)	Comedy|Drama|Romance
5	5	Father of the Bride Part II (1995)	Comedy
6	6	Heat (1995)	Action|Crime|Thriller
7	7	Sabrina (1995)	Comedy|Romance
8	8	Tom and Huck (1995)	Adventure|Children
9	9	Sudden Death (1995)	Action
10	10	GoldenEye (1995)	Action|Adventure|Thriller
11	11	American President, The (1995)	Comedy|Drama|Romance
12	12	Dracula: Dead and Loving It (1995)	Comedy|Horror
13	13	Balto (1995)	Adventure|Animation|Children
14	14	Nixon (1995)	Drama
15	15	Cutthroat Island (1995)	Action|Adventure|Romance

Obs	movieId	title	genres
16	16	Casino (1995)	Crime|Drama
17	17	Sense and Sensibility (1995)	Drama|Romance
18	18	Four Rooms (1995)	Comedy
19	19	Ace Ventura: When Nature Calls (1995)	Comedy
20	20	Money Train (1995)	Action|Comedy|Crime|Drama|Thriller

-

Obs	C1	C2	C3	C4	C5	Avg
1	toys	computer	pixar an	animatio	kids and	3.8866494326
2	adventur	children	fantasy	kids	jungle	3.2465829127
3	sequel	good seq	sequels	comedy	original	3.1739813924
4	women	chick fl	girlie m	romantic	adultery	2.8745399799
5	good seq	sequel	sequels	pregnanc	father d	3.0772909396
6	crime	heist	great ac	bank rob	action	3.8442108566
7	remake	romantic	romantic	romance	paris	3.3713482779
8	adapted	based on	adaptati	literary	based on	3.1224821313
9	action	good act	lone her	fight sc	action p	3.007529782
10	007 (ser	007	bond	franchis	action	3.4316327147
11	presiden	politics	world po	politica	romantic	3.6602775942
12	spoof	parody	hilariou	comedy	vampire	2.6696507515
13	dog	talking	animatio	kids and	animated	3.3396516393
14	presiden	biograph	world po	biopic	politics	3.4289997075
15	treasure	pirates	swashbuc	action	adventur	2.7309765377
16	organize	gangster	mafia	mob	casino	3.8023623907
17	18th cen	adapted	costume	period p	romantic	3.9501262626
18	off-beat	hotel	storytel	tarantin	stylish	3.4120703437
19	goofy	comedy	silly fu	dumb	detectiv	2.6420142094
20	action	train	good act	chase	action p	2.8944826106

-

-