Final Project

John Black

5/10/2023

STA402 B

Dr. Stephan Wright

Original Statement:

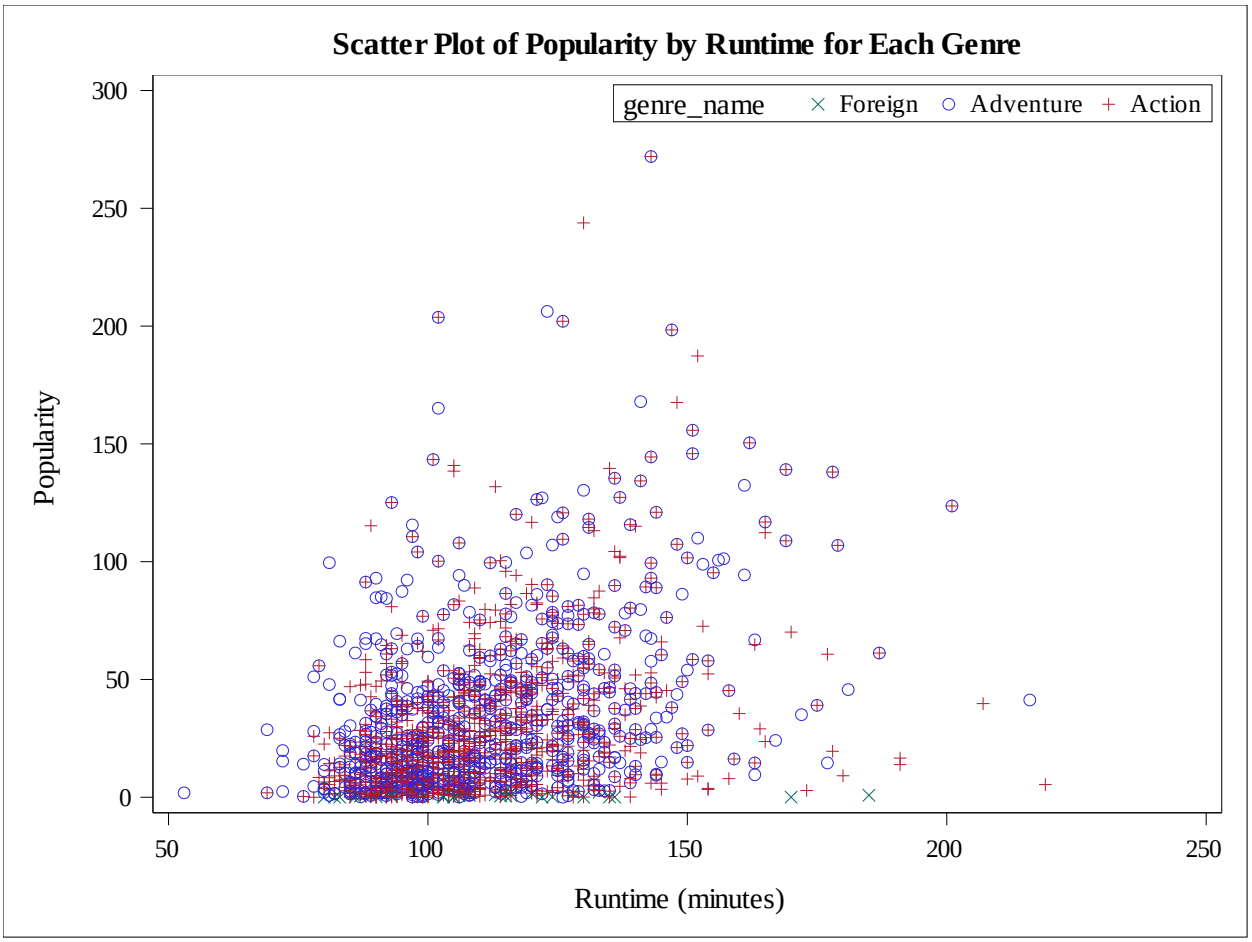
Write an SAS macro that calculates the popularity by each type of genre (using mean ratings or some other measure that you think makes sense) and reports the information for the top N number of genres, where N is a user-specified number. Your macro should also allow users to specify a list of genres that they'd like to compare. Use the ratings.csv dataset.

Tables and Graphs with Explanation:

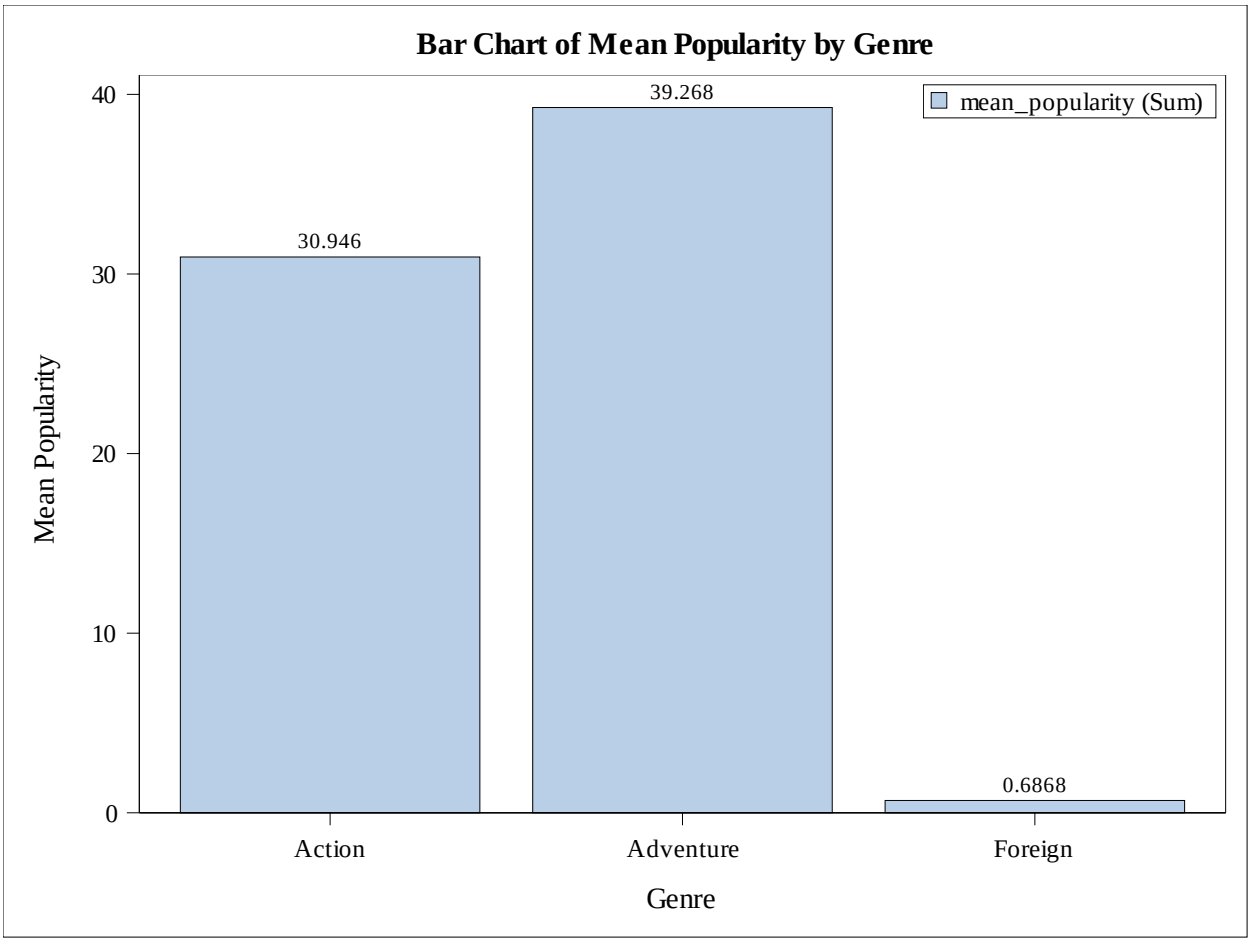
This code accomplishes the assigned task by writing an SAS macro that calculates the popularity by each type of genre (using mean ratings and mean popularity) and reports the information for the user-specified genres. The macro also allows users to specify a list of genres that they'd like to compare using the ratings.csv dataset. In the below example tables the Genres Adventure, Action, And Foreign are used in the macro.

genre_name	mean_popularity	mean_vote_average
Adventure	39.2680	6.15696
Action	30.9465	5.98994
Foreign	0.6868	6.35294

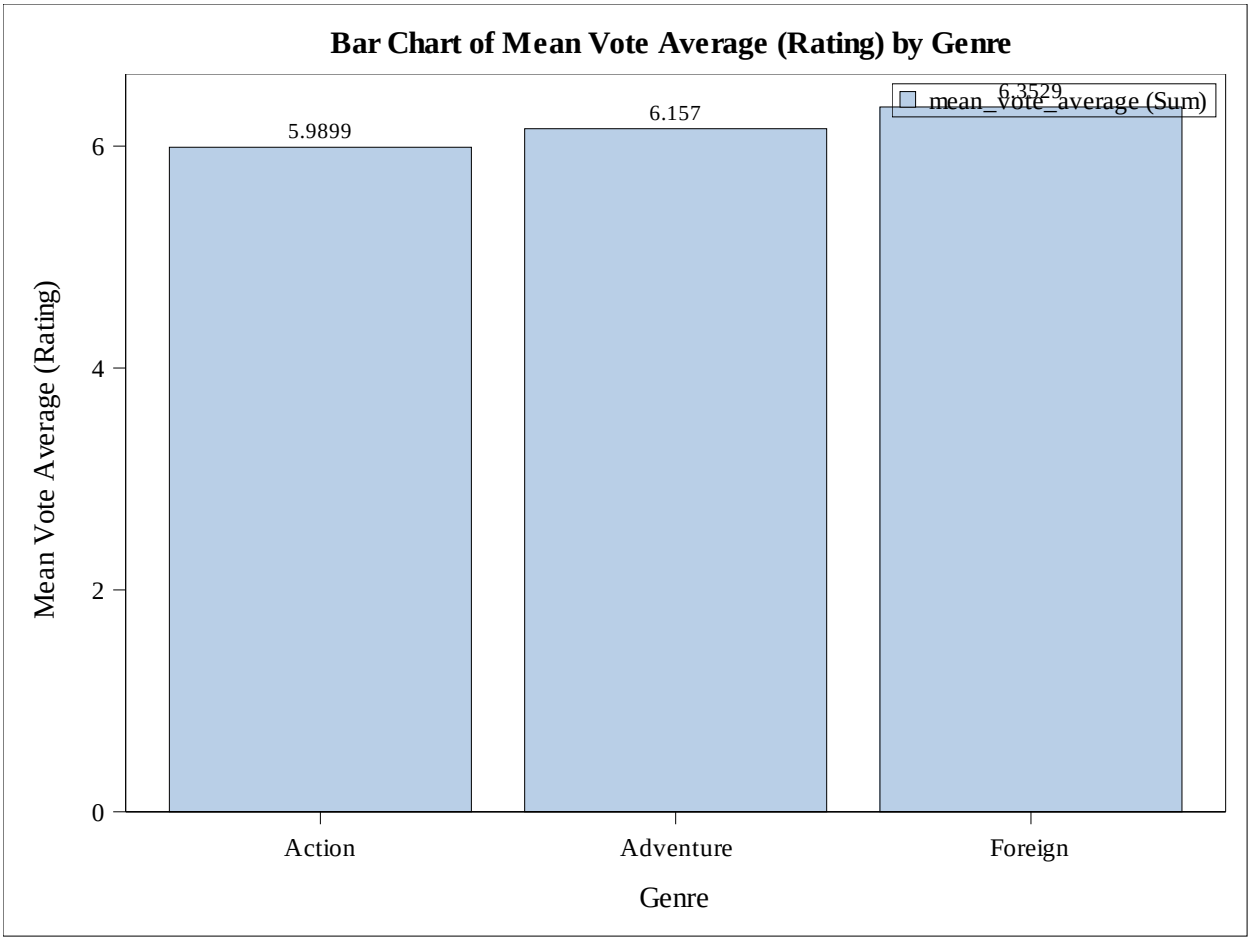
Genre Statistics Table: This table displays the mean popularity and mean vote average (rating) for the user-specified genres. It provides a clear understanding of the popularity and average rating for each genre in the list.



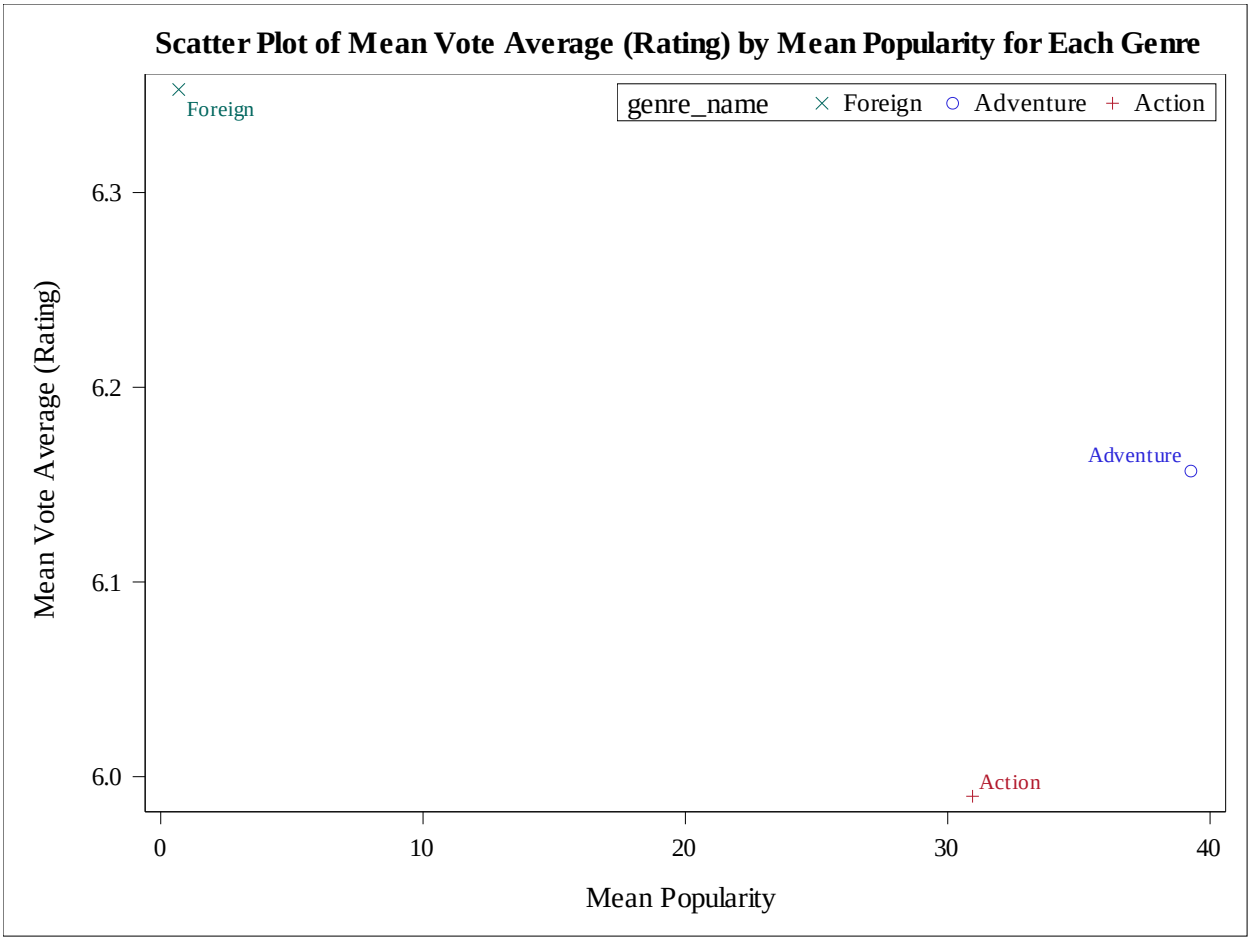
Scatter Plot of Popularity by Runtime for Each Genre: This scatter plot shows the relationship between movie runtime and popularity for the user-specified genres. The x-axis represents the movie runtime (in minutes) within the range of 50 to 250 minutes, and the y-axis represents the movie popularity within the range of 0 to 300. Each point on the graph represents a movie, and the points are color-coded based on their genre.



Bar Chart of Mean Popularity by Genre: This bar chart displays the mean popularity for each of the user-specified genres. The x-axis lists the genres, and the y-axis represents the mean popularity. This chart helps visualize how popular each genre is compared to the others.



Bar Chart of Mean Vote Average (Rating) by Genre: This bar chart shows the mean vote average (rating) for each of the user-specified genres. The x-axis lists the genres, and the y-axis represents the mean vote average (rating). This chart helps visualize how highly rated each genre is compared to the others.



Scatter Plot of Mean Vote Average (Rating) by Mean Popularity for Each Genre: This scatter plot displays the relationship between mean popularity and mean vote average (rating) for each of the user-specified genres. The x-axis represents the mean popularity, and the y-axis represents the mean vote average (rating). Each point on the graph represents a genre, and the genre names are used as data labels.

CODE

```
%LET folder = <your_folder_path>;

/*************************************************************************/
/* John Black: 5/10/2023
      */
/* Title: STA 402/502 Project SAS Macro                                  */
/* Purpose: Calculate and analyze popularity and                          */
/*          ratings for user-specified movie genres                       */
/*                                                                        */
/* Usage:                                                                 */
/* 1. Set the folder path containing the datasets:
*/
```

```
/*
*/
/*    %LET folder = <your_folder_path>;
*/
/*
*/
/* 2. Run the entire SAS file to set up the macros and
         */
/*    data sets.
         */
/*
          */
/* 3. Invoke the macro in a new SAS Editor window:
               */
/*
                */
/*  %theMacro(folder=<your_folder_path>,selectedGenres=%str("Genre1", "Genre2", ...));  */
/*
                 */
/* Example Macro Invocation I used
               */
/*
               */
/*    %theMacro(folder=M:\STA402\Project\,
             */
/*    selectedGenres=%str("Action", "Adventure",
          */
/*    "Foreign"));
         */
/*
        */
/*****************************************************************************/



%macro theMacro(folder=, selectedGenres=);
/* Import the data */
proc import datafile="&folder.tmdb_5000_movies.csv"
    out=tmdb_movies dbms=csv replace;
    getnames=yes;
run;

/* Extract the genres and put them in separate rows */
data tmdb_genres_raw(keep=id genre_id genre_name);
    set tmdb_movies(rename=(genres=movie_genres));
    length genre_string $200;
    genre_count = countc(movie_genres, '{"id":') + 1;

    do i = 1 to genre_count;
        genre_info = scan(movie_genres, i, '{}');

        if not missing(genre_info) then do;
            genre_start = find(genre_info, 'id":') + 4;
            genre_end = find(genre_info, ',', genre_start);
            genre_id = input(substr(genre_info, genre_start, genre_end - genre_start), 8.);

            name_start = find(genre_info, 'name": "') + 8;
            name_end = find(genre_info, '"', name_start);
            genre_name = substr(genre_info, name_start, name_end - name_start);

            output;
        end;
    end;
run;

/* Sort the raw dataset by genre_id */
proc sort data=tmdb_genres_raw out=tmdb_genres_sorted;
    by genre_id;
run;

/* Remove rows with missing genre_id and keep only unique genres */
```

```
data tmdb_genres_unique(keep=genre_id genre_name);
    set tmdb_genres_sorted;
    by genre_id;
    if not missing(genre_id) and first.genre_id;
run;

/*Create Genre_stats macro to calculate*/
%macro genre_stats;
    proc sql;
        create table genre_stats as
            select a.genre_name,
                avg(b.popularity) as mean_popularity,
                avg(b.vote_average) as mean_vote_average
            from tmdb_genres_unique as a
            left join tmdb_genres_raw as raw
            on a.genre_id = raw.genre_id
            left join tmdb_movies as b
            on raw.id = b.id
            where upcase(a.genre_name) in (%upcase(&selectedGenres)) /* Updated where clause */
            group by a.genre_name
            order by mean_popularity desc;
    quit;
%mend genre_stats;

%genre_stats;

ods rtf file="ProjectGraphs.rtf";
/* Print the genre statistics */
proc print data=genre_stats noobs;
    var genre_name mean_popularity mean_vote_average;
run;

/* Merge tmdb_genres_raw with tmdb_movies for the scatter plot */
proc sql;
    create table merged_movies as
        select a.id, a.runtime, a.popularity, b.genre_name
        from tmdb_movies as a
        left join tmdb_genres_raw as b
        on a.id = b.id
        where upcase(b.genre_name) in (%upcase(&selectedGenres)); /* Updated where clause */
quit;

/* Create a scatter plot of popularity by runtime for each genre */
proc sgplot data=merged_movies;
    scatter x=runtime y=popularity / group=genre_name;
    xaxis label="Runtime (minutes)" min=50 max=250; /* Set the x-axis range */
    yaxis label="Popularity" min=0 max=300;         /* Set the y-axis range */
    title "Scatter Plot of Popularity by Runtime for Each Genre";
    keylegend / location=inside position=topright sortorder=descending;
run;

/* Create a bar chart for mean popularity by genre */
proc sgplot data=genre_stats;
    vbar genre_name / response=mean_popularity datalabel;
    xaxis label="Genre";
    yaxis label="Mean Popularity";
    title "Bar Chart of Mean Popularity by Genre";
    keylegend / location=inside position=topright sortorder=descending;
run;

/* Create a bar chart for mean vote average (rating) by genre */
proc sgplot data=genre_stats;
    vbar genre_name / response=mean_vote_average datalabel;
    xaxis label="Genre";
    yaxis label="Mean Vote Average (Rating)";
    title "Bar Chart of Mean Vote Average (Rating) by Genre";
    keylegend / location=inside position=topright sortorder=descending;
run;

/* Create a scatter plot of mean vote average (rating) by mean popularity for each genre */
proc sgplot data=genre_stats;
```

```
    scatter x=mean_popularity y=mean_vote_average / group=genre_name datalabel=genre_name;
    xaxis label="Mean Popularity";
    yaxis label="Mean Vote Average (Rating)";
    title "Scatter Plot of Mean Vote Average (Rating) by Mean Popularity for Each Genre";
    keylegend / location=inside position=topright sortorder=descending;
run;

ods rtf close;
run;
%mend theMacro;
```