Homework 4 Answers

Code by Jason Roman

Q1. How many movies were released for every year within the dataset?

```
+----+-----+
|year|count|
+----+-----+
|1891|    1|
|1893|    1|
|1894|    2|
|1895|    2|
|1896|    2|
|1898|    5|
|1899|    1|
|1900|    1|
|1901|    1|
|1902|    1|
|1903|    1|
|1905|    1|
|1909|    3|
|1910|    3|
|1912|    5|
|1913|    5|
|1914|   13|
|1915|   17|
|1916|   17|
|1917|   12|
|1918|    8|
|1919|   17|
|1920|   19|
|1921|   27|
|1922|   25|
|1923|   17|
|1924|   30|
|1925|   32|
|1926|   40|
|1927|   31|
|1928|   48|
|1929|   50|
|1930|   59|
|1931|   69|
|1932|   96|
|1933|   98|
|1934|  101|
|1935|  107|
|1936|  107|
|1937|  104|
|1938|   95|
|1939|   93|
|1940|  117|
|1941|  107|
|1942|  101|
|1943|  117|
|1944|  101|
|1945|  105|
|1946|   92|
|1947|   99|
|1948|  102|
|1949|  126|
|1950|  122|
|1951|  125|
|1952|  131|
|1953|  136|
|1954|  113|
|1955|  146|
|1956|  136|
|1957|  163|
|1958|  146|
|1959|  151|
|1960|  148|
|1961|  123|
|1962|  151|
|1963|  148|
|1964|  173|
|1965|  166|
|1966|  199|
|1967|  173|
|1968|  203|
|1969|  177|
|1970|  204|
|1971|  205|
|1972|  219|
|1973|  211|
|1974|  195|
|1975|  196|
|1976|  199|
|1977|  198|
|1978|  192|
|1979|  201|
|1980|  243|
|1981|  248|
|1982|  238|
|1983|  223|
|1984|  234|
|1985|  254|
|1986|  266|
|1987|  313|
|1988|  325|
|1989|  310|
|1990|  314|
|1991|  312|
|1992|  335|
|1993|  371|
|1994|  432|
|1995|  474|
|1996|  509|
|1997|  528|
|1998|  555|
|1999|  542|
|2000|  613|
|2001|  633|
|2002|  678|
|2003|  655|
|2004|  706|
|2005|  741|
|2006|  855|
|2007|  902|
|2008|  979|
|2009| 1113|
|2010|  962|
|2011| 1016|
|2012| 1022|
|2013| 1011|
|2014|  740|
|2015|  120|
+----+-----+
```

I first had to format the dataset using a UDF to find these values. This UDF would find the year variable inside each title variable and add it to an added year column I create. I then filter the dataset to remove entries without a year value, group all of the movies by year, and then return the number of movies made per year.

Q2. What is the average number of genres for movies within this dataset?

```
+------------------+
|   avg(genreCount)|
+------------------+
|1.9945010631277953|
+------------------+
```

To find the average I first used a UDF to count the number of genres associated with each movie. I did this by treating genres as a delimited string separated by pipelines ("|"). I then create another dataframe to house the average genre count and calculate the average.

Q3. Rank the genres in the order of their ratings? Again, a movie may span multiple genres; such a movie should be counted in all the genres.

```
+------------------+------------------+
|             genre|        avg_rating|
+------------------+------------------+
|         Film-Noir|  3.96538126070082|
|               War|3.8095307347384844|
|       Documentary|3.7397176834178865|
|             Crime|3.6745276025631113|
|             Drama|3.6742955093068264|
|           Mystery| 3.663508921312903|
|              IMAX| 3.655945983272606|
|         Animation|3.6174939235897994|
|           Western|3.5704980246109406|
|           Musical| 3.558090628821412|
|           Romance| 3.541802581902903|
|          Thriller|  3.50711121809216|
|           Fantasy|3.5059453358738244|
|         Adventure|3.5018926565473865|
|          Action|  3.44386376493354|
|          Sci-Fi|3.4367726714455005|
|          Comedy|3.4260113054324886|
|        Children|3.4081137685270444|
|          Horror|3.2772238097518307|
|(no genres listed)|3.0069252077562325|
+------------------+------------------+
```

To answer this question I first need to grab the individual genres. To do that, I used the explode function to split the genres into individual values for each movie. Next, I joined the exploded genres with the ratings DataFrame, grouped them by genre, and then calculated the average rating for each genre. The final result is sorted in descending order.

Q4. What are the top-3 combinations of genres that have the highest ratings?

```
+------------------------------------------+----------+
|genres                                    |avg_rating|
+------------------------------------------+----------+
|Adventure|Drama|Fantasy|Musical           |5.0       |
|Adventure|Children|Comedy|Documentary|Drama|5.0       |
|Adventure|Children|Mystery                |5.0       |
+------------------------------------------+----------+
```

To find the top 3 combinations, I directly grouped movies by their genre combinations and then calculated the average ratings for movies with these combinations. The results are in the top-3 list.

Q5. How many movies have been tagged as “comedy”? Ignore the “case” information (i.e. both “Comedy” and “comedy” should be considered).

Number of movies tagged as comedy: 8374

To find the number of comedies in the data, I exploded the genres DataFrame to count how many movies were tagged as “comedy” or “Comedy” and returned the final comedy count.

Q6. What are the different genres within this dataset? How many movies were released within different genres? A movie may span multiple genres; in such cases, that movie should be counted in all the Genres.

```
+------------------+-----+
|             genre|count|
+------------------+-----+
|             Drama|13344|
|            Comedy| 8374|
|          Thriller| 4178|
|           Romance| 4127|
|            Action| 3520|
|             Crime| 2939|
|            Horror| 2611|
|       Documentary| 2471|
|         Adventure| 2329|
|            Sci-Fi| 1743|
|           Mystery| 1514|
|           Fantasy| 1412|
|               War| 1194|
|          Children| 1139|
|           Musical| 1036|
|         Animation| 1027|
|           Western|  676|
|         Film-Noir|  330|
|(no genres listed)|  246|
|              IMAX|  196|
+------------------+-----+
```

To find the number of movies per game, I exploded the genres DataFrame to count how many movies fell into each genre. The output is in descending order of count.

Q7. According to the dataset, what tags are most relevant to rating?

```
+------------------------------------------------------------+------------------+
|tag                                                         |avg_rating        |
+------------------------------------------------------------+------------------+
|brilliant                                                   |4.193353416105184 |
|perfect                                                     |4.170106494856577 |
|photographer                                                |4.154815115806582 |
|kurosawa                                                    |4.14102774361077  |
|awesome                                                     |4.111752597868822 |
|afi 100                                                     |4.099186567876109 |
|francis ford copolla                                                  |4.0946153143761554|
|rio de janeiro                                              |4.094241649228861 |
|italy                                                       |4.092857535715759 |
|cathartic                                                   |4.091316218028547 |
|flashbacks                                                  |4.0870243135077615|
|miyazaki                                                    |4.0825326365866905|
|studio ghibli                                               |4.081722498658922 |
|genius                                                      |4.081144601998759 |
|italian                                                          |4.080379038153669 |
|tolkien                                                          |4.06793301300667  |
|moving                                                           |4.066388124562134 |
|new zealand                                                      |4.062143922820634 |
|mozart                                                           |4.061347461839697 |
|oscar (best foreign language film)                               |4.054581248282963 |
|photography                                                      |4.052403102894233 |
|movielens top pick                                               |4.050395682681373 |
|oscar (best writing - screenplay written directly for the screen)|4.0413885793648285|
|neo-nazis                                                        |4.035894115954539 |
|holocaust                                                        |4.033983126030806 |
|marx brothers                                                    |4.028364542016419 |
|black and white                                                  |4.024767193743608 |
|afi 100 (movie quotes)                                           |4.022662504949675 |
|poland                                                           |4.02244035690124  |
|morality                                                         |4.021750161063898 |
|noir                                                             |4.019163780588793 |
|brazil                                                           |4.018235677791064 |
|skinhead                                                         |4.01813528108125  |
|hannibal lecter                                                  |4.015966687104026 |
|unusual plot structure                                           |4.01546907189772  |
|short                                                            |4.009492817461031 |
|fighting the system                                              |4.008922345849952 |
|amazing photography                                              |4.008778530498346 |
|nonlinear                                                        |4.006277342250715 |
|vienna                                                           |4.005964905315812 |
|nazi                                                             |4.001197441989668 |
|ironic                                                           |3.9980522948971515|
|prohibition                                                      |3.9927312298590736|
|oscar (best cinematography)                                      |3.992593021248947 |
|cynical                                                          |3.9921826096192934|
|intelligent                                                      |3.991764774689694 |
|notable soundtrack                                               |3.990451586523495 |
|oscar (best picture)                                             |3.9893983389723973|
|idealism                                                         |3.9882732421131966|
|oscar (best actor)                                               |3.9873201893950427|
+-----------------------------------------------------------------+------------------+
```

To find the most relevant tags I had to first filter the genon_scores to only input tages with a relevance greater than 0.5. I then joined the dataset with genome-tags.csv, movies.csv, and ratings.csv to link each movie with its respective ratings. I then grouped the new DataFrame by tag, calculated the average rating per tag, and then ordered the results by the average ratings. The output is the top 50 most relevant tags according to rating.