CS838 Project Stage 5 Report

Fangzhou Mu, Nafisah Islam, Meera George

I. Datasets

We combined samples of *tracks.csv* and *songs.csv* and obtained *merged_data.csv* in stage 4. In stage 5, we further merged *movies.csv* into it and obtained *final_merged_data.csv*. The data integration pipeline is similar to the one reported in stage 4. The final dataset contains 430 tuples.

The schema of this table includes the following attribute:
movie_title, year, length, budget, rating, votes, mpaa, Action, Animation, Comedy, Drama, Documentation, Romance, Short, episode, song_title

Here are four representative tuples in the table.

movie_title	year	length	budget	rating	votes	mpaa	Action	Animation	Comedy	Drama	Documentar	Romance	Short	episode	song_title	artists
the forger	2012	94		5.4	4010	pg-13	0	0	0	1	0	0	0		moanin'	art blakey and the jazz messengers
the forger	2012	94		5.4	4010	pg-13	0	0	0	1	0	0	0		prelude to a	horace silver
vacation	2012	2					0	0	1	0	0	0	1		chariots of fi	vangelis
killing bono	2011	114		6.4	5155	r	0	0	1	0	0	0	0		you spin me	dead or alive

All datasets and Jupyter notebooks are available at Github Link.

II. Data analysis: tasks, steps and conclusions

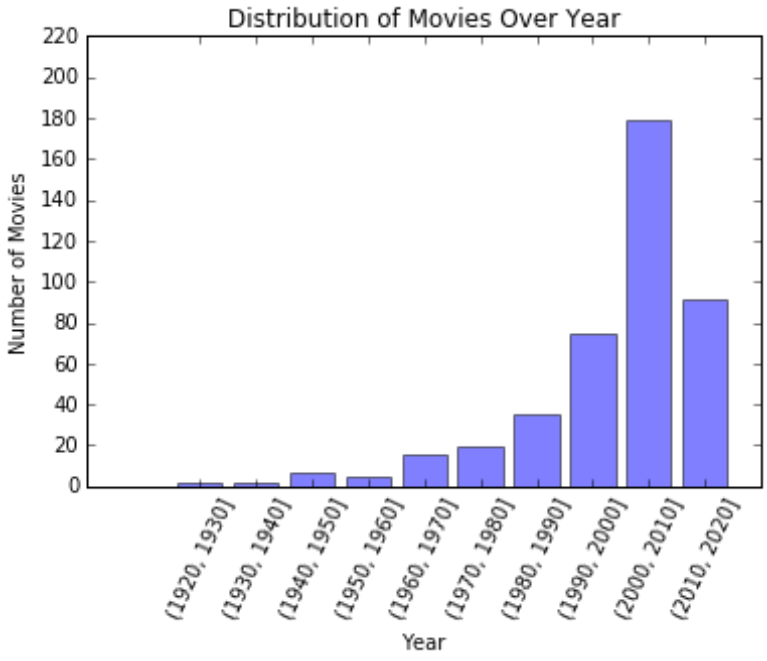
We do **OLAP-style** queries. We propose the following questions. Notice that we are only interested in the specific sample we obtained rather than the full datasets. Details of how we answer these questions and what we conclude from the answer are discussed below each question.

(1) How many tracks are mapped onto each movie on average?

We group movies by their titles. They are grouped into **286** groups, which means that there are 286 unique movies in our dataset. We then examine the size of each group. To get a more global idea, we compute the mean, max and min of the number of tracks mapped to each movie. The corresponding values are **1.5**, **10** and **1**. In particular, the movie *The Soul of A Man* is associated with 10 tracks.

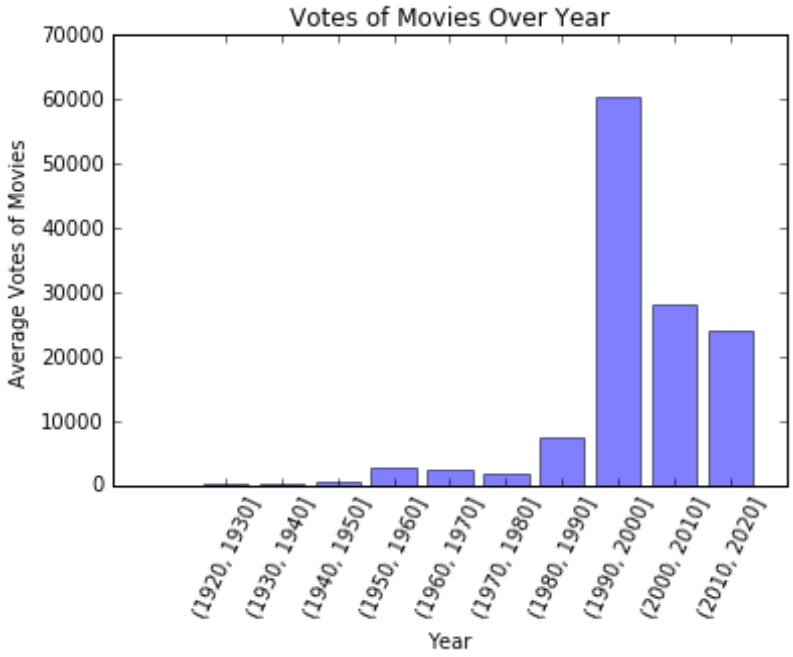
(2) How many movies are produced every ten years? How popular are they?

We count the number of movies within each 10-year time interval.



This trend is totally expected as the movie market grows over years. The last bar is not meaningful since we miss data from 2017 to 2020.

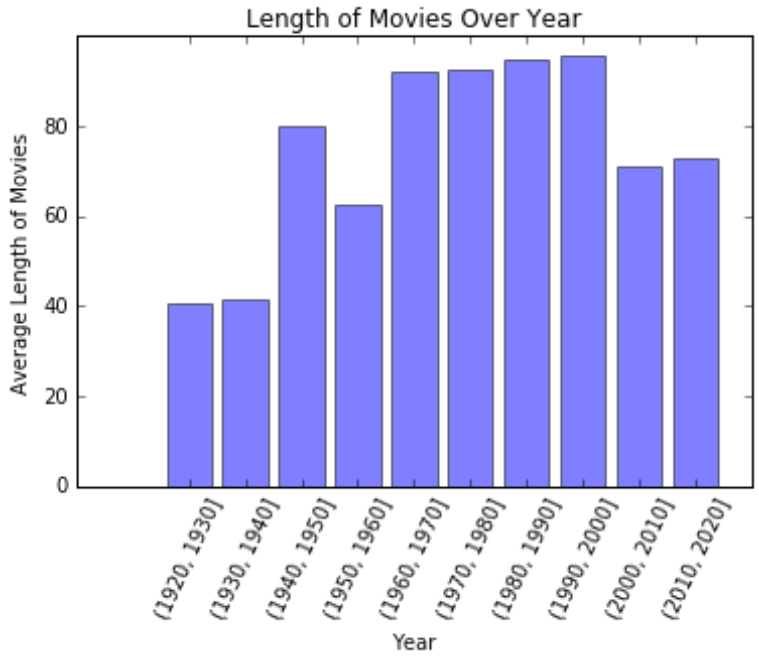
Then we compute the average votes of the movies in each category.



Clearly, the interval [1990, 2000] is interesting. The average number of votes for movies within this period is particularly high. We then ask which movies are responsible for most of the votes. We checked this particular interval and sort the tuples according to the number of votes in descending order. The top three movies are *American Beauty* (**831630** votes), *Die Hard with A Vengeance* (**301541** votes), and *Me, Myself & Irene* (**187297** votes). They are indeed popular movies. They together contribute to over 75% of the total votes.

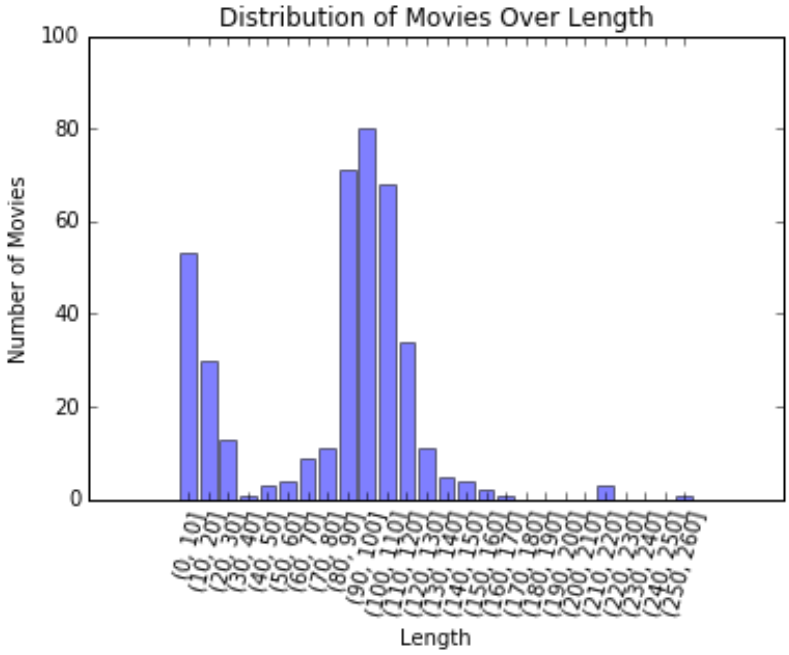
(3) Do movies tend to become longer or shorter over years? How are their lengths distributed?

We again group movies produced in each 10-year interval and compute their average length.



This tells us that movie length grew from about **40** minutes to over **90** minutes from the 1920s to 2000s, and then dropped slightly to about **70** minutes later.

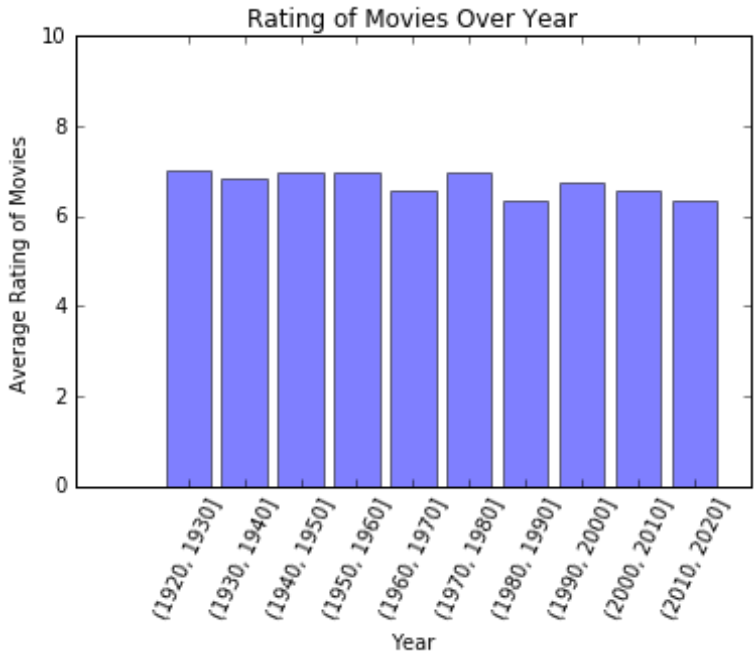
We then estimate the distribution of movie length over the entire timeline by histograms.



The first peak corresponds to short movies. All movies of other types form a nice Gaussian-like distribution centered at [90-100] minutes. This agrees with our intuition.

(4) Do movie ratings drop or increase over years?

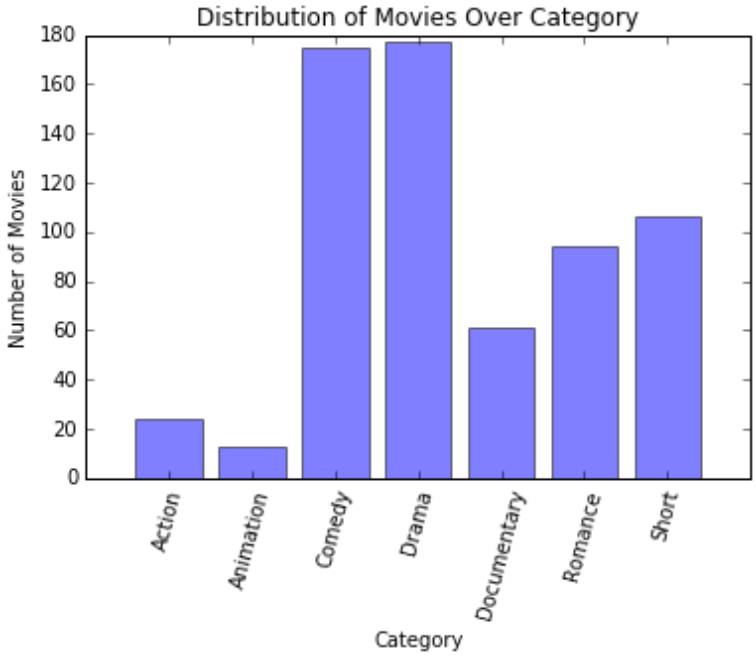
We perform similar operations as in (3).



This tells us that the average rating of movies every ten years is very stable. It is about **7** out of 10.

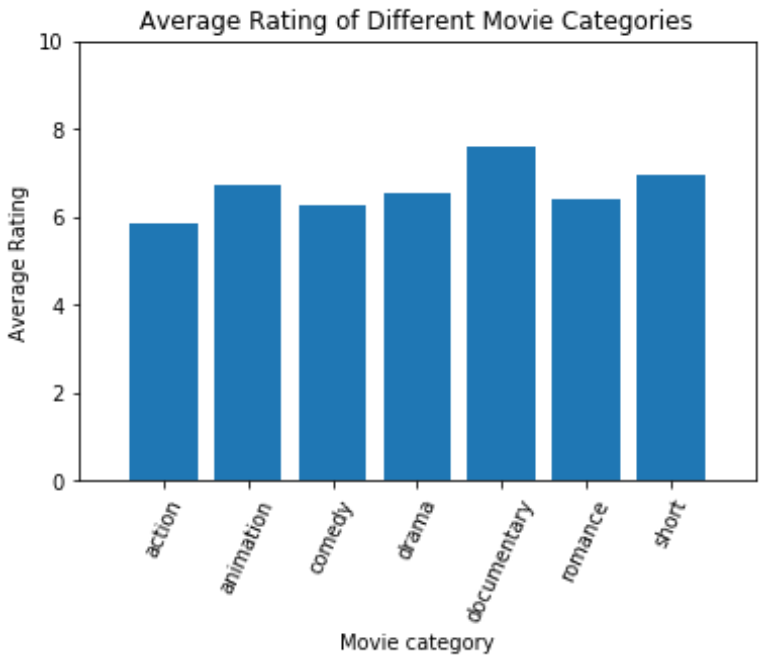
(5) How many movies are in each category? Which movie category receives the best rating?

We group movies of the same category and count the movies in each category.



Clearly, our dataset mostly consists of comedies and dramas, while it contains very few action movies and animations.

We then look into each movie category to see if there is any variation across different categories. We compute the average rating over the movies in each category.



There is not too much difference, although documentary movies receive a moderately higher score. We go into this category and look for the movie with the highest rating, which turns out to be *Led Zeppelin: Celebration Day*.

(6) If someone is interested in a drama produced in the 1990s whose rating is over 8.0, which movies can he/she choose?

We selected the movies that satisfy the three criteria. *American Beauty* is his/her only option.

(7) If someone is interested in a romance movie whose music is composed by Holly Johnson, which movies can he/she choose?

We selected the movies that satisfy the two criteria. *Made of Honor* is his/her only option.

(8) What are the top 10 popular movies in the past 20 years whose ratings are better than 8.0?

We select all movies produced in the past 20 years and sort them based on their ratings. We then select those whose ratings are greater than 8.0 and output the names of the top 10 movies. They are *American Beauty*, *The Help*, *The Cove*, *Led Zeppelin: Celebration Day*, *A Walk to Beautiful, Prison Break, Blur, Middle of Nowhere, The Great Everything & The Nothing,* and *Jam.*

These queries involve the four most common operations used in OLAP: roll-up, drill-down, slice and dice.

Some problems in our data analysis is listed as follows.

(1) Our table is full of missing values.

We avoid making use of columns which contain too many missing values to generate insights. However, it would be beneficial if we can figure out a reasonble way to use the information available in those columns or to fill out the missing entries.

(2) We only analyzed a sample of the full dataset.

Unfortunately, our final table only has over 400 tuples. This significantly limits the number and types of insights we can generate. For example, we may be able to create a classifier to predict movie catergory based on other information provided. The columns corresponding to movie

categories naturally fit the format of class labels. However, given the number of tuples we have right now, it is hard to obtain a sufficiently large training set for training a good classifier.

(3) Some data quality issues.

Beyond missing values, the quality of our final merged table may not be the best. This is due to two rounds of data merging, which could propagate errors.

III. Future directions

Depending on the application, we can further accomplish a lot more tasks on this dataset, assuming we have a sufficiently large table for analysis. In fact, we can do classification, clustering, association rule mining, anomaly detection, and many other OLAP queries different from what are described in this report. We describe two possible tasks in more detail.

(1) Supervised learning:

As discussed in the previous section, we can train a classifer to predict movie categories if we have enough training and test data. This classifier needs to predict class labels of multiple dimensions. Random Forests, Neural Networks, SVMs, among others, are all good candidate algorithms.

(2) Association rule mining:

Given a sufficiently large table (e.g. by merging the three full tables), we may be able to learn associations among different attributes. For example, we may wonder how rating and number of votes are correlated. We can generate a scatter plot for all (rating, votes) pairs in the dataset and estimate a correlation coefficient.