CSCE 110: Programming I - Final Project

Texas A&M University, Spring 2019

Project title:

Netflix CSV Reader

Date: 05/01/2019

Team member name	**Team member UIN**
Ashlynn Droddy	225006165
Alexander Arias	127009894
Micaleb Johnson	127002572

Table of Contents

Introduction

This project takes in a csv file for movies released in 2016 including genre, release date, distributor, and ticket sales. The program takes that information and outputs statistics relating to one or more of the categories. While also giving text statistics, the program also gives visuals by using the module matplotlib.plyplot to graph some of those statistics.

Procedures

This project tasked us with using the python programming skills we have learned to answer five questions about a dataset. These questions required us to have knowledge of the packages matplotlib and csv for plotting and reading in the dataset as a .csv file. We also utilized a good understanding of data structures, conditional statements, and loops to complete this project.

Question 1.

This question asks us to obtain various details from the imported movie dataset. This required creating multiple lists and sets for each of the variables of interest from the dataset and then appending each value in the dataset to its respective list. Once the data structures were filled with the correct values, they were edited, for example, the header for each list was removed and for some lists, certain characters were removed so that the correct calculations could be made. The output for this question printed the wanted statistics from the dataset including the total number of movies, genres, MPAA ratings, distributors, and the number of tickets sold.

Question 2.

The second question tasks us to plot movies released each month in 2016, and find which month had the most movies released (Figure A-1). To start this question we created a dictionary initializing each month with a value of zero. Looping through the dates, each occurrence was added to its corresponding month in the dictionary. After that, we initialized zero as our max and created a loop to find the real max. With this information and matplotlib.pyplot a bar graph was created showing how many movies were released each month.

Question 3.

The third question requires us to produce a line plot that shows the number of movie ticket sales for each month of 2016. To begin this problem, we had to create empty lists for each month and then use a for loop to iterate through a list with all of the data in it. Inside this for loop, we included various if and elif statements conditioned on the month that a movie was released. These conditional statements would determine which empty list to assign the current iterations “number of tickets sold”. We then summed the values of the tickets sold inside each list and added them to a dictionary corresponding to their respective months. Another for loop was run looking through the dictionaries values determining which month had the most tickets sold. We then printed this month along with the actual number of tickets sold during this month and created a line plot showing the number of tickets sold in each month (Figure A-2).

Question 4.

For the fourth question, we had to find the percentage of ticket sales by a distributor. If the percentage was less than 1, then it would be put into a category ‘others’. To start this question we loop through the distinct set of distributors we previously created for the first question and we made a zero value dictionary for each distributor. We looped to add to the dictionary the ticket sales to the corresponding distributor. Then we found the percentage of each one by dividing by total ticket sales and multiplying by 100. After that, we printed in the order of greatest percent and joining those less than 1 percent to ‘others’. With all this information we plotted it to a pie chart. (Figure A-3)

Question 5.

The fifth question asks us to plot how many drama, horror, comedy, and action occurred in different months of 2016 (Figure A-4). To start this question we decided to start off with a function that would take a dictionary with each genre valued at zero and an integer to correspond to each month and with that it would return the dictionary with the number of occurrences for that month. Then we put each genre with its own list of occurrences per month as our y values. Also, we made a list with months as the x values. Finally plotting each genre line in the graph.

Results and Discussion

This program was successful in giving us the certain statistics that we wanted to see as well as giving great visual aid with the graphs. There were problems when first trying to output the complete data, but we eventually ended up with working code. The results gave us an insight into how movie companies might want to look at this to figure out when is the best time to release a movie or when to release a certain genre.

Functionality claims

In general, the statistics are accurate of the information given in the csv file. There is a possibility the csv file is set up in a different format then the output would likely be inaccurate.

Conclusion

Overall, the code was successful in answering the questions using the file, ‘2016_movie_data.csv’. The code broke down this dataset according to the different attributes of the data. This allowed access to each attribute according to the output needed. Collective data structures aided in storing data in order for the data to later be organized. By creating plots of the organized data, attributes of the movies could be better visualized. In order to answer the questions of the project, it was necessary to break down the data first. Majority of the code consists of breaking down and organizing information from the large dataset given. Once this information was organized it was simple to plot our information given it was already organized in a collective data structure. It was difficult to have three people write a single code as the code

worked differently on each computer. The problem was identified to be a problem in the way the data file was opened initially. In the end, the code simulated a movie tracker successfully.

Appendix

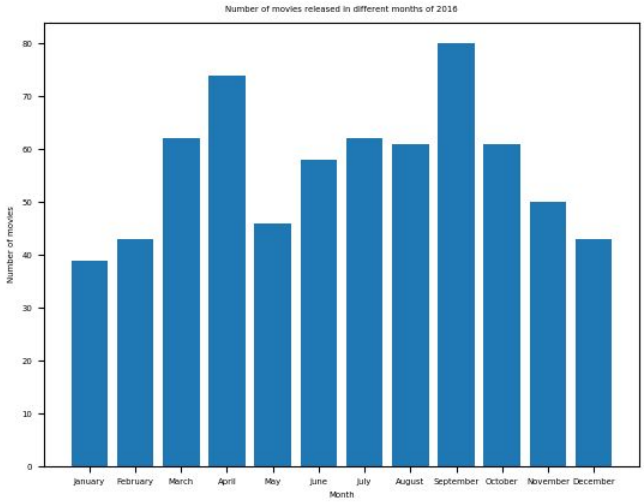

Figure A-1. Number of movies released in different months of 2016

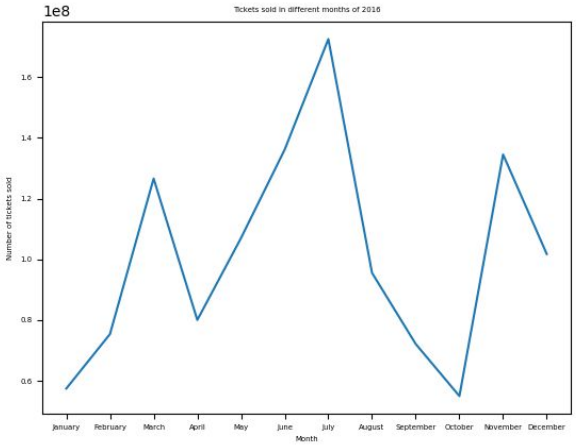

Figure A-2. Tickets sold in different months of 2016.

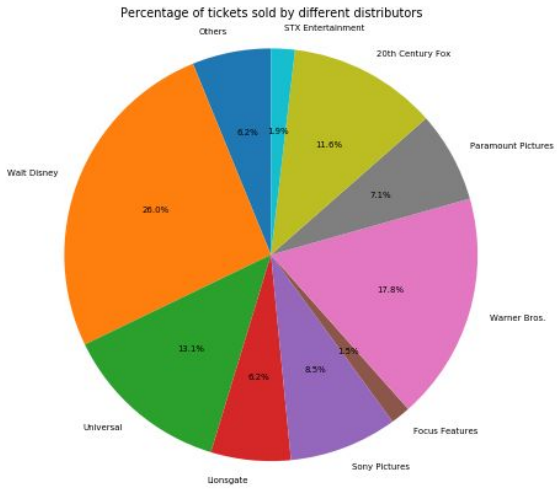

Figure A-3. Percentage of tickets sold by different distributors.

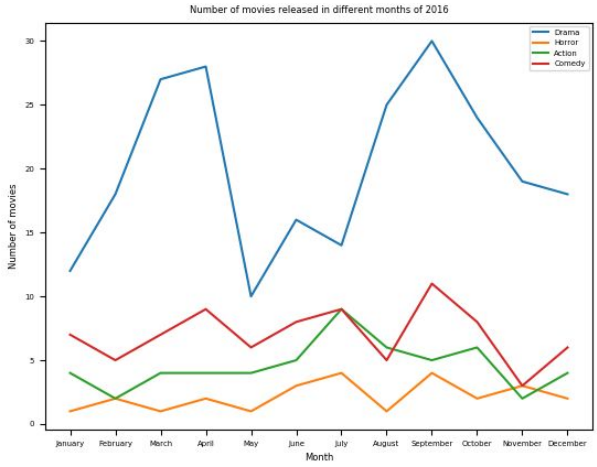

Figure A-3. Number of movies released in different months of 2016.