CSCE 110: Programming I – Final Project

Texas A&M University, Fall 2019

Date: 12/10/19

Student Name:	Student UIN:	Student email:
Jacob Kastenschmidt	328000135	jacob_kastenschmidt@tamu.edu
Ryan Holloway	528007777	ryanholloway@tamu.edu
Stephen Shell	228004951	stephen6410@tamu.edu

Table of Contents

Introduction

The purpose of this project was to analyze a large data set in Microsoft Excel using Python. By reading the Excel file in a CSV format, our Python program turned information about 2016 Netflix movies into useful data and graphs. In this report, we will discuss our program's procedures: First, we organized the CSV file into lists of data, second, we printed a series of statistics regarding the quantities of said lists, and finally, we created four charts which are presented in this report's appendix. After explaining the procedures, we will reflect on our coding experience, the challenges we faced, and potential improvements to our program.

Procedures

Parsing the Excel file

Before we answered any questions or created any charts, we had to first open the CSV file, extract its data, and parse it into useful tools. We decided the best way to navigate our table of data in Excel was to organize each column into separate lists.

Question 1: Print data details

The first question asks for the total number of movies released in 2016. Since movie titles are unique, we put all of them in a list. Then, we simply took the length of the list and printed it as a string, which was 679. The second question asks for the number of different genres. Instead of placing every genre in a list as we did for the movies, we added each genre to a set. Sets cannot contain repeats, so the genre "comedy" was only placed in the set once, despite there being multiple comedy movies in 2016. So, the number of genres is the length of the set of genres, which was 13. The third question asks for the number of different MPAA ratings. As with genres, there are repeats, as multiple movies can have the same rating. Adding them all to a set and taking the length of the set gives us the number of different ratings, which was 6. The fourth question asks for the number of different distributors, which, like genres and ratings, are not unique, so we printed the length of the set of distributors, which was 145. The final question asks for the total number of tickets sold. Since we already made a list of the total number of tickets sold for each movie, we used the "sum" function to add every element of the list together, which we found to be 1,214,119,134.

*Question 2: Movies by each month (see Appendix: **Figure 1**)*

The next part of our program creates a bar chart for the number of movies per month with months on the x-axis and number of movies on the y-axis. First, we created a list of month names to label the x-axis of our bar graph. Then we assigned each month with the number of movies released within that month. This required us to iterate through our list of release dates, date by date. Using a gauntlet of 12 "if" statements, one for every month's name, our program added 1 to a month's number of movies every time the release date in question matched the month's name.

Once we had the correct number of movies assigned to each month, we added each month name to a list. By using the "max" function, we took extracted the month of which the most movies were released. September was the most common month for released movies, with April close behind.

Question 3: Tickets sold by each month (see Appendix: ***Figure 2****)*

Our next graph plots a line representing tickets sold each month with months on the x-axis and the number of tickets sold on the y-axis. For the last graph, we used only one list, namely the list of release dates. For this graph, however, we used three of our lists. We iterated through the length of the list of movie names, indexed the list of release dates, and added the corresponding number of tickets sold to the month of the release date. As with the for loop in the previous graph, we used 12 “if” statements to check which month the movie was released and added one to that month’s value for our chart.

The main struggle during this process was getting the computer to examine *rows* of data, even though our data was organized into *columns*. Our solution was something we referred to as symmetric indexing of parallel lists. Since our lists of movie names, release dates, and number of tickets sold retained their original order in the Excel file, the sixth movie in the list of movie titles should correspond with the sixth release date and sixth number of tickets sold in the other two lists respectively. This process works for the sixth movie or any number, so by using the variable, i, as our indexer, we essentially iterated through every row in the Excel file.

After plotting the line, we noticed that July had the most tickets sold by far. This number was calculated, like the max number of movies released in a month, using the “max” function. Strangely, September and April, the months with the most movies released, had two of the fewest numbers of tickets sold. We expected this line graph to match up at least somewhat with our previous bar graph, but the two graphs look completely different.

Question 4: Percentage of tickets by each distributor (see Appendix: ***Figure 3****)*

Our third graph asks for the percentage of tickets sold by each distributor in a pie chart. Since distributors such as Walt Disney produce multiple movies in a year, we iterated through our *set* of distributors. For each distributor, we divided the total number of tickets they sold by the total numbers of tickets sold by all distributors. Then, we made a dictionary with the distributors as keys and their percentage of 2016 tickets as values.

However, we ran into the issue of multiple distributors compiling too small of a percentage of the tickets to add to the pie chart. So, we compiled all distributors occupying less than 0.01% of the pie chart into one category called “Other.” Then, we removed all distributors which fell into the “Other” category from our dictionary of distributors and ticket percentage, and replaced them with one key, namely “Other”, with its value being the combined ticket percentages of all included distributors.

Only nine of the distributors sold enough tickets not to be placed in the “Other” category. One of those nine, Lionsgate, sold 6.2% of the tickets which is the same percentage as the entire “Other” category. But none were as impressive as Walt Disney, which soaked up just over a quarter of the pie at 26%. Warner Bros was the only one close, sitting at 17.8%.

*Question 5: Movies by genre (see Appendix: **Figure 4**)*

Our final graph asks to plot four lines, one for each major movie genre, that would track the number of movies of each genre per month. Before iterating through the movies, we created four lists: Drama, horror, action, and comedy. Each list was initially comprised of 12 zeros, or place holders, representing the number of movies released during each month. By iterating through the list of movies, our program checked both the release date and the movie genre for each movie, and it counted the number of movies per month for each genre list. For example, if a movie's genre was "drama," and its release date was in January, our code would add 1 to the place holder in the drama movies list's first position (drama[0]).

After the iteration was complete, we noticed that drama movies far exceeded the other movie genres released in 2016. As we expected, the plotline for drama movies mirrored our first bar chart for number of movies released per month. The two months with the most drama movies released were April and September. We also noticed that comedy was consistently the second most popular movie genre, but in August, action movies edged out comedy movies. Finally, we noticed that horror was consistently the least common movie genre, but in November, horror barely surpassed action and tied with comedy. We believe this was due to an increased number of horror movies released on Halloween.

Conclusion

Our experience writing this code was somewhat familiar. On one hand, we used our normal methods of loops and lists, but we had never written a code of this magnitude. The only "new" things we had to figure out were graphing with *matplotlib.pyplot* and parsing the Excel file. Graphing turned out to be easier than we thought, because it was very intuitive. Parsing the Excel file was not hard, but we needed a few tricks. For example, we replaced the commas in the number of tickets sold to "", so that we could turn those data numbers into "*ints*" and sum them. Another challenge we had was formatting our pie chart correctly. At first, the labels overlapped, but by adjusting the "start angle," we made space for all the words.

There are three main ways we wish we could improve our program. First, we all agree that it could probably be shorter, especially the section in which we counted the genres in each month. However, we could not figure out a shorter way to code it. Second, we wanted to make a code that could take the input of another similar Excel file, such as 2017's Netflix movie data. Our program should still work for such a file, but if there were any changes such as the columns being in a different order or there being a new column, our program would have issues. Finally, we are curious about what other data and graphs we could pull from the Excel file. One question that intrigued us was, "How would our current graphs change if we *only* included the more popular movies, or more specifically, the movies with 1,000,000 or more tickets sold?" We think that adding such charts to our program would interest most people and therefore improve our results.

Appendix

Figure 1. *Number of movies released in different months of 2016*

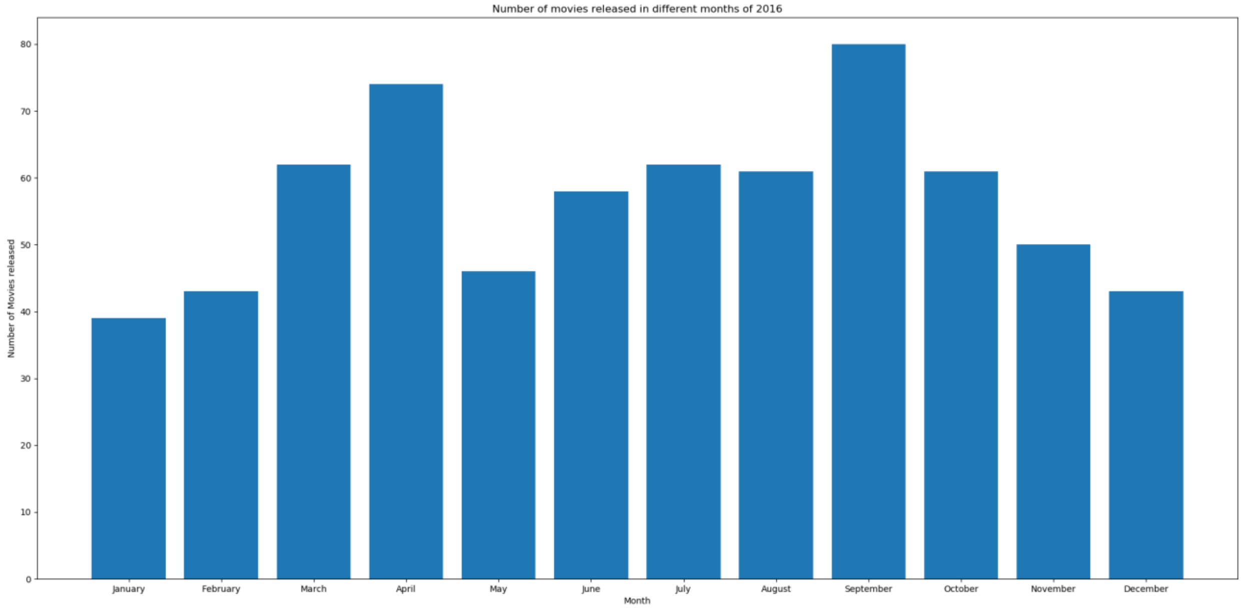


Figure 2. *Number of tickets sold in different months of 2016*

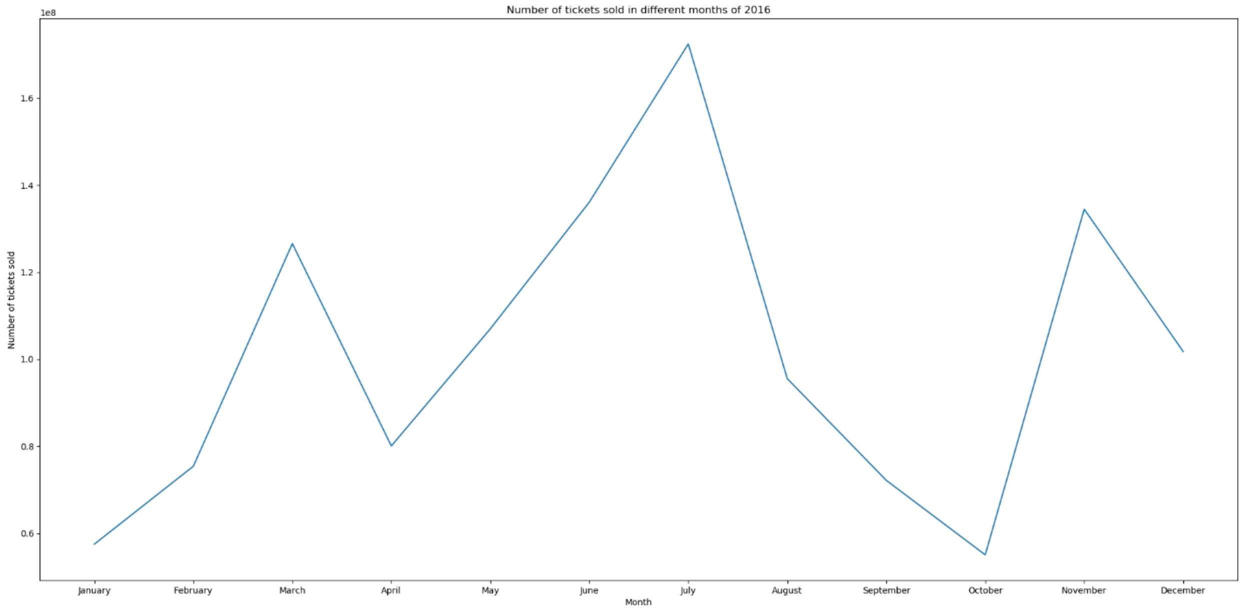


Figure 3. *Percentage of tickets sold by different distributors*

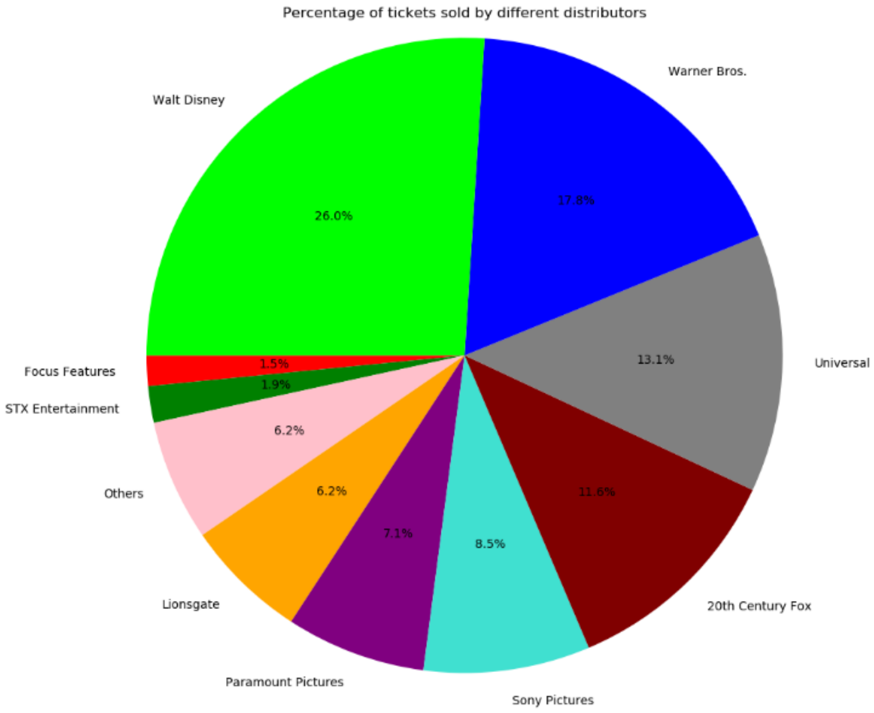
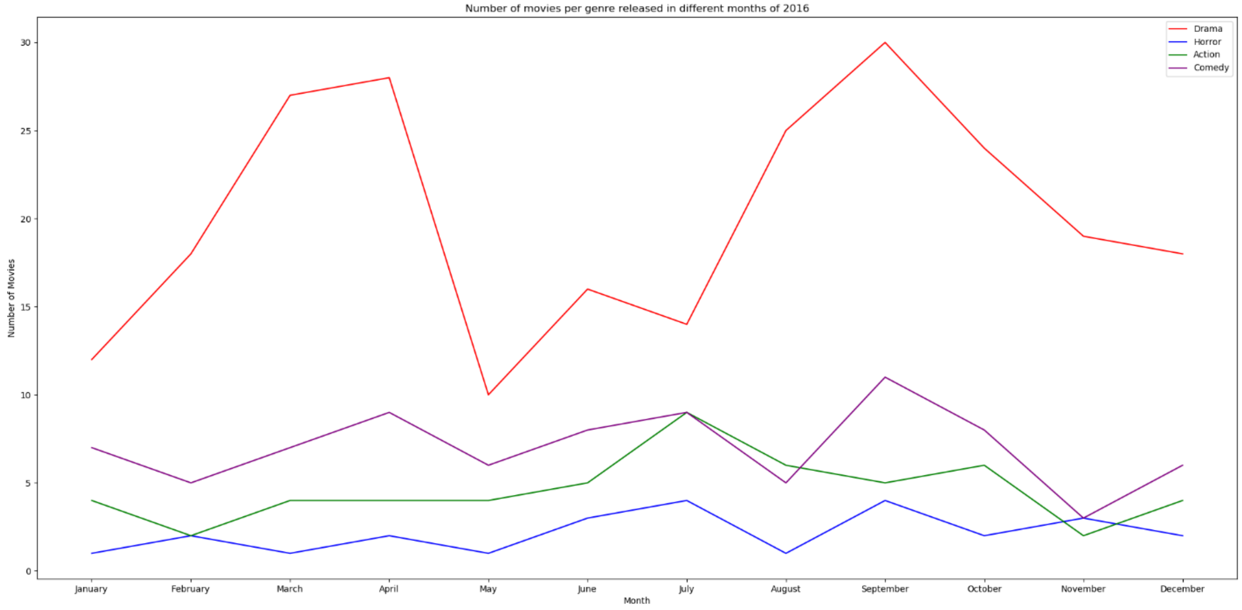


Figure 4. *Number of movies per genre released in different months of 2016*


Number of movies per genre released in different months of 2016
Drama
Horror
Action
Comedy
Number of Movies
January
February
March
April
May
June
July
August
September
October
November
December
Month