Name: Kelly Choy

DSCI 510: Final Project

Abstract:

My final project is to analyze the original films Netflix released that ended. Using my datasets, I want to explore which genres tend to have a higher or lower rating and to monitor Netflix's stock prices throughout the years since their first released film. The 3 sources that will help me with this project are the ended Netflix original programming on Wikipedia, OMDb API, and stock prices through Yahoo Finance API. The final dataset with some data cleaning is a CSV file called "final.csv." This dataset is created using 3 data sources listed below in the data sources section. In the provided zip file, "dataset1.csv", "dataset2.csv", and "dataset3.csv" are CSV files that led to the creation of the final dataset version. To sum up my findings, the dataset ranges from 2/6/2012 to 4/14/2023, and I've identified out of the 163 genres Netflix has released, the genres that have a high rating (based on IMDb) were supernatural/action comedy, a slice of life/anthology series, stop motion, dark fantasy action, workplace comedy/science fiction, and western. Out of the years that Netflix has been releasing original films, 2021 was the year that had the highest average close price.

Motivation:

Netflix has been a top-rated entertainment platform for people. I recently watched many Netflix shows and am curious about what makes Netflix stand out compared to other movie viewing platforms. I want to see what genres Netflix has released to the public and what is the average IMDB ratings these genres have and compare them to one another. I'm unsure how many genres they have, but I'm hypothesizing that action, comedy, and romance have high ratings. I will also look at the stock prices and see when the highest and lowest Close Prices of Netflix were. I'm curious to see the growth from how Netflix did before to today.

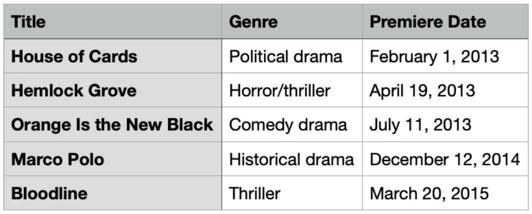
Data Sources/Technical Solutions:

Dataset1: Web Scraping - Netflix Films

URL: https://en.wikipedia.org/wiki/List_of_ended_Netflix_original_programming

From the ended Netflix original programming on Wikipedia, I web-scraped the content (title of the film, genre, premiere date). Like what we've learned in class, I used the request library and beautifulSoup to extract the necessary information on the web page. This website provides me with the names of the Netflix films that I will be doing my research on. From this data source, I will have 3 columns and 1240 rows. In this dataset, I did not capture the last remaining table (continuation) because those films are duplicate titles from above, and I wanted to avoid this in my analysis. With the films' names, I can connect them to my API data, which is collected in my second data source.

Below are the first few rows of my dataset from this data source.



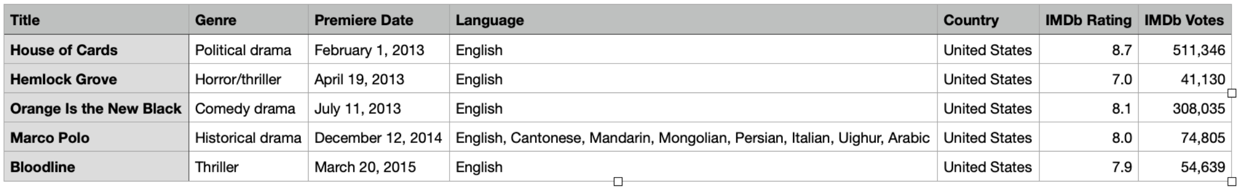
Title	Genre	Premiere Date
House of Cards	Political drama	February 1, 2013
Hemlock Grove	Horror/thriller	April 19, 2013
Orange Is the New Black	Comedy drama	July 11, 2013
Marco Polo	Historical drama	December 12, 2014
Bloodline	Thriller	March 20, 2015

Dataset2: OMDb API

URL: https://www.omdbapi.com/

From the OMDb API, I can get more information about the specific movie I'm looking into, such as getting the language, country, IMDB ratings and IMDB Votes. To get the information from the API, an API key is needed. You will need to request for a unique API key through their website. I have already added the new columns to the original dataset, which creates a new CSV that looks like this:

(3 columns from dataset1, added 4 new columns from this dataset)



Title	Genre	Premiere Date	Language	Country	IMDb Rating	IMDb Votes
House of Cards	Political drama	February 1, 2013	English	United States	8.7	511,346
Hemlock Grove	Horror/thriller	April 19, 2013	English	United States	7.0	41,130
Orange Is the New Black	Comedy drama	July 11, 2013	English	United States	8.1	308,035
Marco Polo	Historical drama	December 12, 2014	English, Cantonese, Mandarin, Mongolian, Persian, Italian, Uighur, Arabic	United States	8.0	74,805
Bloodline	Thriller	March 20, 2015	English	United States	7.9	54,639

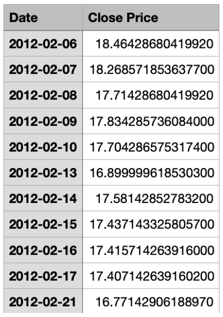
Dataset3: Yahoo Finance - API

URL: https://finance.yahoo.com/quote/nflx/history/

The third dataset is to get the stock close price through Yahoo Finance API. From this dataset, I can see what the close price is for Netflix within a specific timeframe and connect it to the premiere date of the Netflix films to see if there are any patterns. In my project, I identified what was the earliest and latest released date and used those are my maximum and minimum date.

Netflix Close prices:

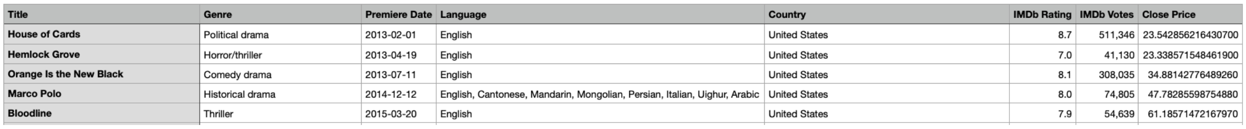
netflix_close_prices



Date	Close Price
2012-02-06	18.46428680419920
2012-02-07	18.268571853637700
2012-02-08	17.71428680419920
2012-02-09	17.834285736084000
2012-02-10	17.704286575317400
2012-02-13	16.899999618530300
2012-02-14	17.58142852783200
2012-02-15	17.437143325805700
2012-02-16	17.415714263916000
2012-02-17	17.407142639160200
2012-02-21	16.77142906188970

Final merged sample:

(3 columns from dataset1, 4 columns from dataset2, and 1 new column from this dataset



Title	Genre	Premiere Date	Language	Country	IMDb Rating	IMDb Votes	Close Price
House of Cards	Political drama	2013-02-01	English	United States	8.7	511,346	23.542856216430700
Hemlock Grove	Horror/thriller	2013-04-19	English	United States	7.0	41,130	23.338571548461900
Orange Is the New Black	Comedy drama	2013-07-11	English	United States	8.1	308,035	34.88142776489260
Marco Polo	Historical drama	2014-12-12	English, Cantonese, Mandarin, Mongolian, Persian, Italian, Uighur, Arabic	United States	8.0	74,805	47.78285598754880
Bloodline	Thriller	2015-03-20	English	United States	7.9	54,639	61.18571472167970

From the first dataset, there were 3 columns, then in the second dataset, we added 4 more columns to the original, and from the third dataset, we will add 1 more to complete it. In total, there are 8 columns in the final output file.

Technical:

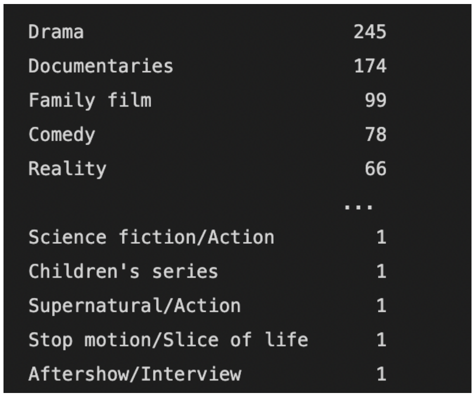
This paragraph refers to my Python code in HW#4_Kelly_Choy.ipynb. Getting the data from source #1 and source #3 was simple, but data from source #2 was challenging due to the limitations of the API Key. The API Key has a limit of 1,000, so if you are trying to gather more than 1,000 requests, you may need to request 2 different API KEYs (using different email addresses) or pay ($$) to get a larger limit daily. So for 2 API Keys, my max would be 2,000 records. If your dataset has more than this, you may need to request additional API Keys and add another argument to the function for a third key. In my code, I have a dataset from 1,000-2,000, so I need 2 API Keys; My two API keys are '24bea5c3','b17bad04'. In my code, I've included conditional statements where if the dataset is <1000 records, it will utilize 1 API key or the other condition with the 2 API keys. For homework #4, I have already merged all the content (using Pandas) I needed into 1 CSV file. The result was 8 columns and 1240 rows. Check out the ReadMe file in the "Homework #4 info" folder inside the finalProject_Kelly_Choy folder for more detailed information.

The main Python code I will refer to in this section is finalproject_Kelly_Choy.ipynb. The first step was to do additional necessary data cleaning on top of what I've done separately above for the 3 individual datasets. In my dataset, I've noticed that some genres were all the same but not categorized the same due to capitalization, spelling, punctuation, etc. For example, there were "Docu-series" and "Docuseries" in the data, so I wanted them grouped in the same category. This applied to the drama genre (ex: some were listed as "drama" and others were "Drama") as well as the horror genre. Cleaning and categorizing the data better helped me with my project analysis, such as identifying how many movies were released in a specific genre. After data cleaning, the new dataset is labeled as "final.csv." My analysis was based on this CSV file.

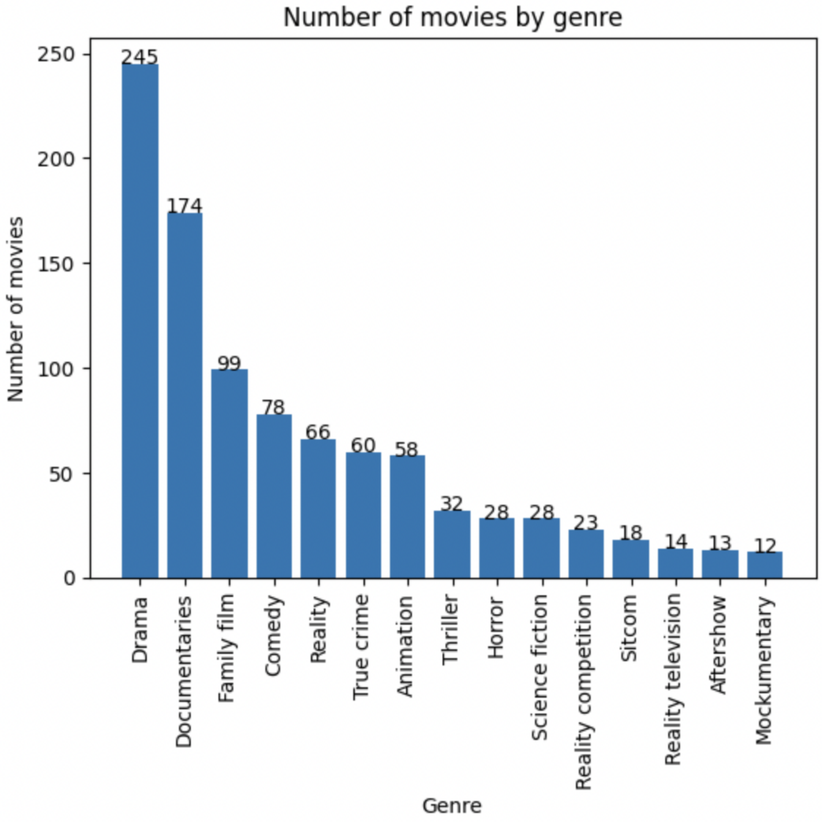
In my code for analysis, I mainly used the Pandas library to create data frames and perform some statistical analysis. I used Matplotlib to help generate graphs so visually see the analysis. Lastly, I incorporated the use of sqlite3 to help create a SQL table where I could run queries that answered my questions. Because the dataset consists of many rows and values, sometimes the visuals were not portrayed in a clear way, so I had to change up my analysis to only select a limited amount of records. For example, there were 163 genres that Netflix had, but if I were to display this in a bar chart visual, the graph would be too small and long to analyze. Therefore, I decided to focus on the top 15 genres instead of analyzing all 163. I saw more value in exploring this group. Overall, there weren't extreme technical difficulties that were obstacles to me. More detailed notes will be written for each section of my code in the Jupyter Notebook file(finalproject_Kelly_Choy.ipynb).

Analysis:

Using the first data source, I identified the number of genres Netflix had and how many movies belonged to each genre. There are 163 genres that Netflix released between 2/6/2012 to 4/14/2023 (this is the date range of the dataset). Using pandas, I could count the number of movies per genre. As I mentioned in the technical solution section, to have a clear vision of our analysis, we will be focusing on the top 15 genres. The bar chart below shows the top 15 genres that Netflix has. From this chart, we can see that Netflix has many drama genres put out in their library for the public, and documentaries and family films come in second and third.

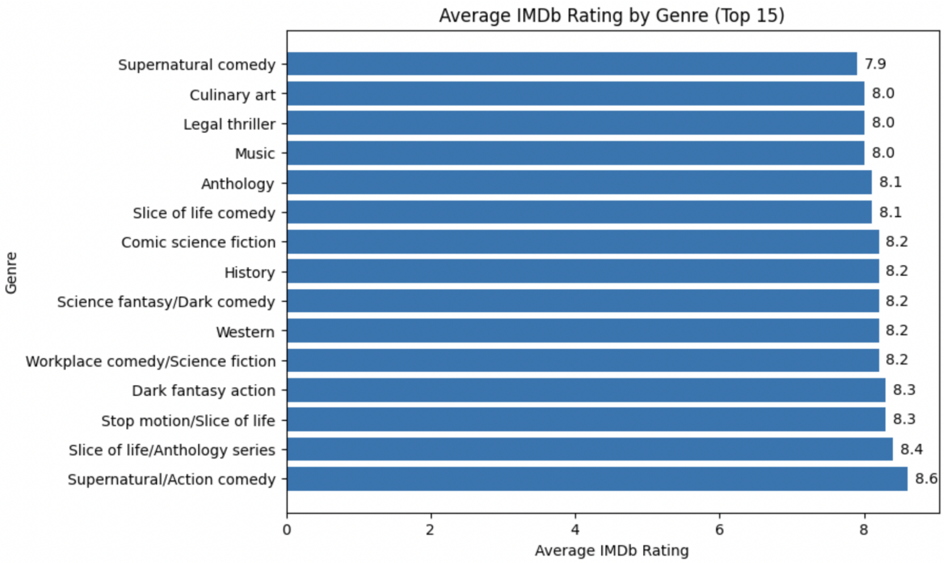


```
Drama                          245
Documentaries                  174
Family film                     99
Comedy                          78
Reality                         66
                              ...
Science fiction/Action           1
Children's series                1
Supernatural/Action              1
Stop motion/Slice of life        1
Aftershow/Interview              1
```

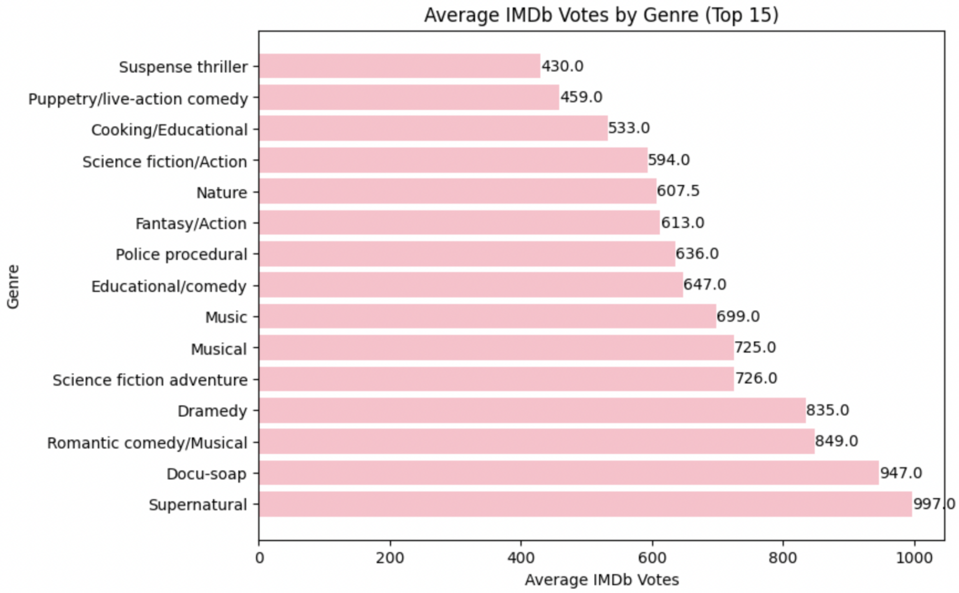


From the second data source, I calculated the average IMDb ratings and votes of the movies per genre. I created an SQL table and inputted the CSV data as values. By running a select statement, I can get the IMDB ratings averages for each genre group. After this, I created a bar chart to visualize which genres had the highest ratings. Then, I do the same steps but use the IMDB Votings values instead. The average ratings for genres are close, but the top 3 genres that perform best (according to the IMDB Ratings) are supernatural/comedy, a slice of life/Anthology series, and stop motion/slice of life.

Below is a graph that represents the top 15 genres and their average IMDb Ratings

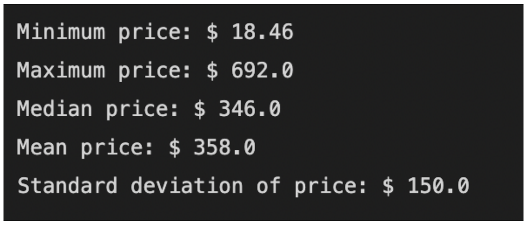


Below represent the top 15 genres and their average IMDb Votes



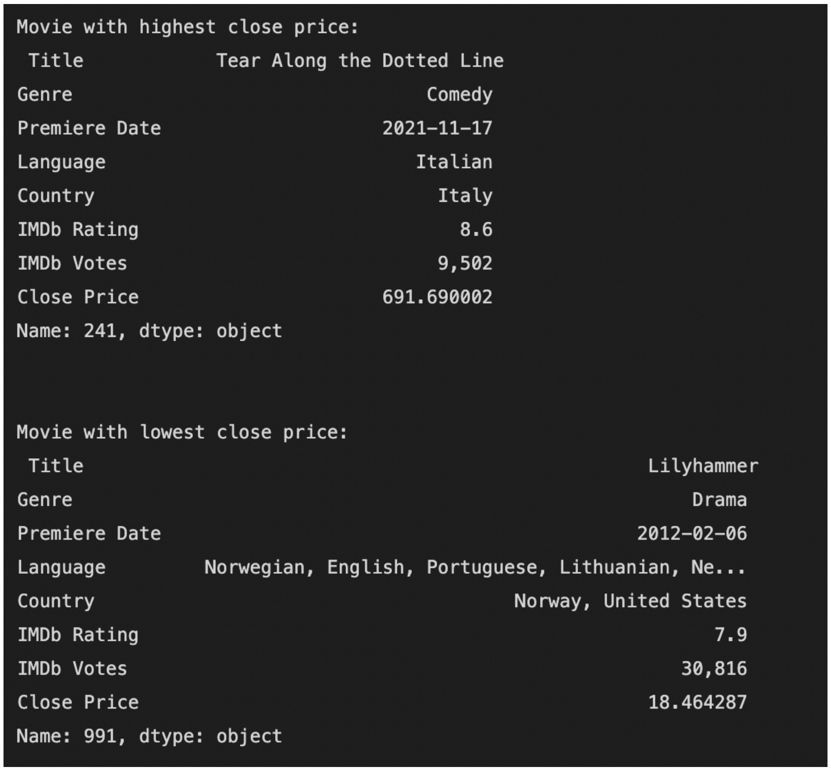
From the third data source, I want to analyze the close prices of Netflix. I have calculated several statistical measures (mean, median, standard deviation, etc.) for the close price column. I was also curious to see which film had the minimum close price value and which one had the maximum close price value. I have the answers listed below, and I want to compare them to see if the genres of these movies were part of the top 15 I gathered earlier. One last analysis I tried to collect from the third dataset was to see how the close prices change year to year (we are looking at 2012-2023 because that's the time range of our dataset). I group the records by year (get this from the premiere date column) and then find the average close price. These values are plotted on a line chart below.

Descriptive measures of the close price column:



```
Minimum price: $ 18.46
Maximum price: $ 692.0
Median price: $ 346.0
Mean price: $ 358.0
Standard deviation of price: $ 150.0
```

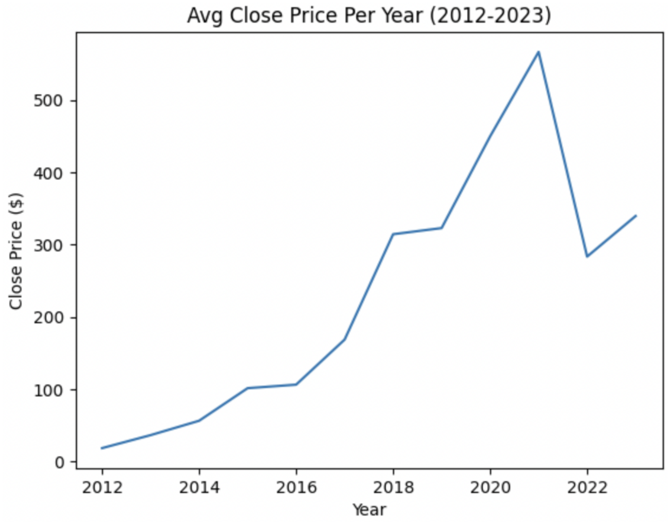
Movie with the highest and lowest close price:



```
Movie with highest close price:
 Title           Tear Along the Dotted Line
Genre                                Comedy
Premiere Date                    2021-11-17
Language                            Italian
Country                               Italy
IMDb Rating                             8.6
IMDb Votes                            9,502
Close Price                      691.690002
Name: 241, dtype: object


Movie with lowest close price:
 Title                                               Lilyhammer
Genre                                                     Drama
Premiere Date                                        2012-02-06
Language         Norwegian, English, Portuguese, Lithuanian, Ne...
Country                                   Norway, United States
IMDb Rating                                                 7.9
IMDb Votes                                               30,816
Close Price                                           18.464287
Name: 991, dtype: object
```

Average Close Price Chart:



Conclusion:

My project taught me much about Netflix and answered some lingering questions. Through my analysis, I learned that Netflix has 163 genres in the library, and the drama genre was the highest category, followed by documentaries and family films. It can be concluded that the top 3 genres that perform best (according to the IMDB Ratings) are supernatural/comedy, a slice of life/Anthology series, and stop motion/slice of life, which have a score of 8.6, 8.4, and 8.3, respectively. My initial thought would be action, comedy, and romance. I was surprised to find out that these were genres that Netflix has. The results from using IMDb votes showed different genres at the top; however, it depends on what the user wants to use to analyze the genres. Both ratings can be used when concluding, depending on the user's preference. With the statistical measures, I can see Netflix's highest close price was $692, and the lowest was $18.46. The film "Tear Along the Dotted Line" had the highest Close Price, which makes me believe maybe it was a good film due to its high IMDb rating (8.6). As for the lowest close price, the film is called "Lilyhammer" and premiered on February 6, 2012, when Netflix first released its first original film. It does sound reasonable to have a lower close price because of this. Looking at the Avg Close Price by Year graph, we see the highest average close price was in 2021. I believe this may be due to the increase in staying home and watching films from the covid-19 pandemic. Some future work ideas would be that Netflix can utilize this data and see what new films to produce to appeal more to the audience. We have identified which genres have a high rating, so focusing on creating films within those categories can help Netflix expand more.