CS51 Final Project: Netflix Movie Analysis

Diane Nguyen

Project Overview

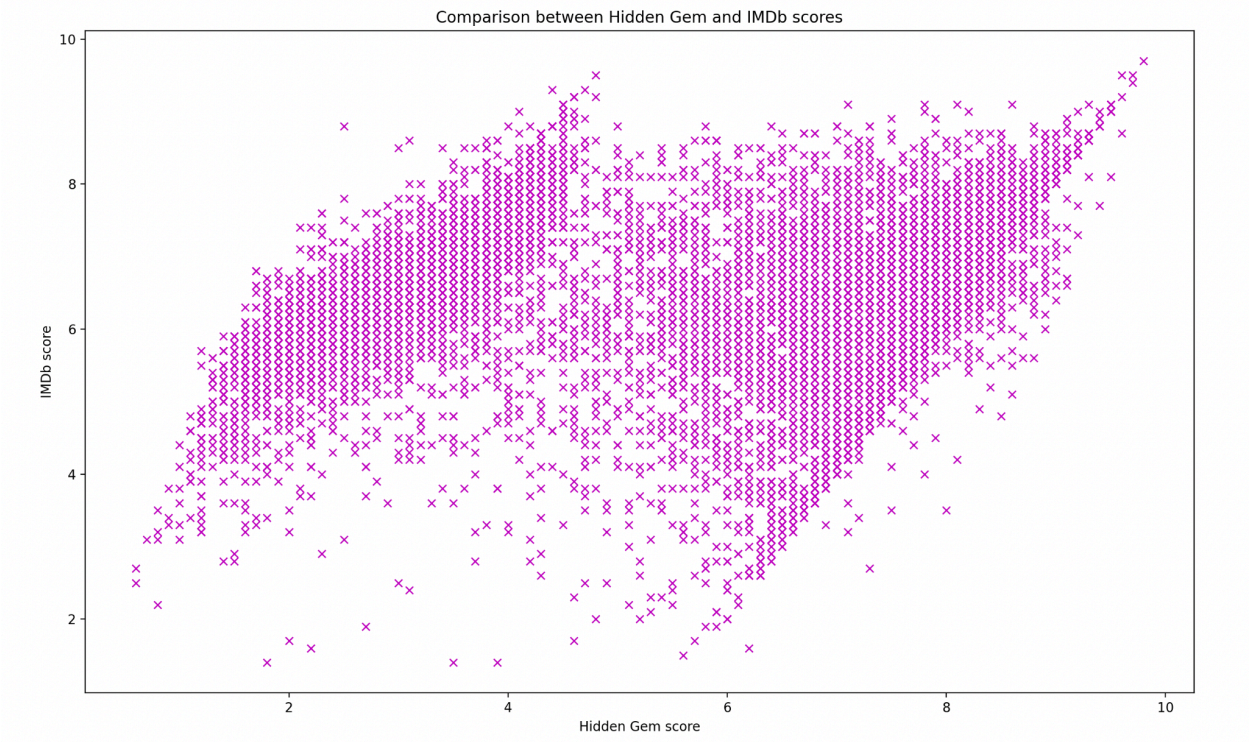
This final project showcases the skills developed in CSCI 51p by applying data analysis techniques to explore trends within Netflix movies. The analysis focuses on three main areas: the correlation between IMDb and Hidden Gem scores, the diversity of languages in which content is produced, and shifts in viewership ratings over time. Using a dataset sourced from Kaggle, we formulated research questions aimed at uncovering meaningful insights into Netflix's content distribution and user preferences.

Our approach involved selecting the dataset, defining our research questions, and implementing Python code to generate both numerical data and visualizations. Specifically, we utilized matplotlib for visual representation, enabling us to parse, process, and interpret the data effectively. Key findings were presented in a brief report summarizing our methods, results, and insights. Additionally, an ethical reflection was completed to address data biases and their implications for analysis. This project demonstrates a practical application of programming and data science skills to solve real-world problems in the entertainment domain.

Final Project: Write-up

Part 1: Is a movie's IMBd score positively correlated with its Hidden Gem score? Or do ratings for the same movies diverge significantly between ratings?

In this section, we explore the relationship between a movie's IMDb score and its Hidden Gem score. Specifically, we examine whether these scores are positively correlated or if the ratings for the same movies show significant divergence between the two scores. The graph presents a scatter plot of the Hidden Gem and IMDb scores for each Netflix movie in our dataset. As we can see, there is a modest positive trend, as indicated by a generally positive diagonal alignment of the data points, though they are dispersed broadly across the plot. This suggests a slight positive correlation between the two scores. To quantify this relationship, we calculated the Pearson correlation coefficient and obtained a value of approximately 0.123. This indicates that while there is a positive correlation, it is relatively weak. This outcome aligns with our visual assessment of the scatter plot, where it is evident that for any given IMDb score, there is a broad range of corresponding Hidden Gem scores. It is important to note that the Hidden Gem score is compiled from a variety of websites and prioritizes films that receive high ratings but have fewer reviews, potentially affecting the objectivity of the score. In contrast, IMDb functions as a single platform where a diverse user base submits reviews, which could explain the observed discrepancies between the two scores and the resultant weak correlation.



Relevant code:

To analyze the correlation between IMDb and Hidden Gem scores, we used three key functions:

- `ratings()`: Extracts IMDb and Hidden Gem scores from the dataset.
- `correlation()`: Calculates the Pearson correlation coefficient to quantify the relationship between the two scores.
- `plot_ratings()`: Generates a scatterplot to visualize the correlation.

Key Code Snippets:

- Core Loop in `ratings()`:

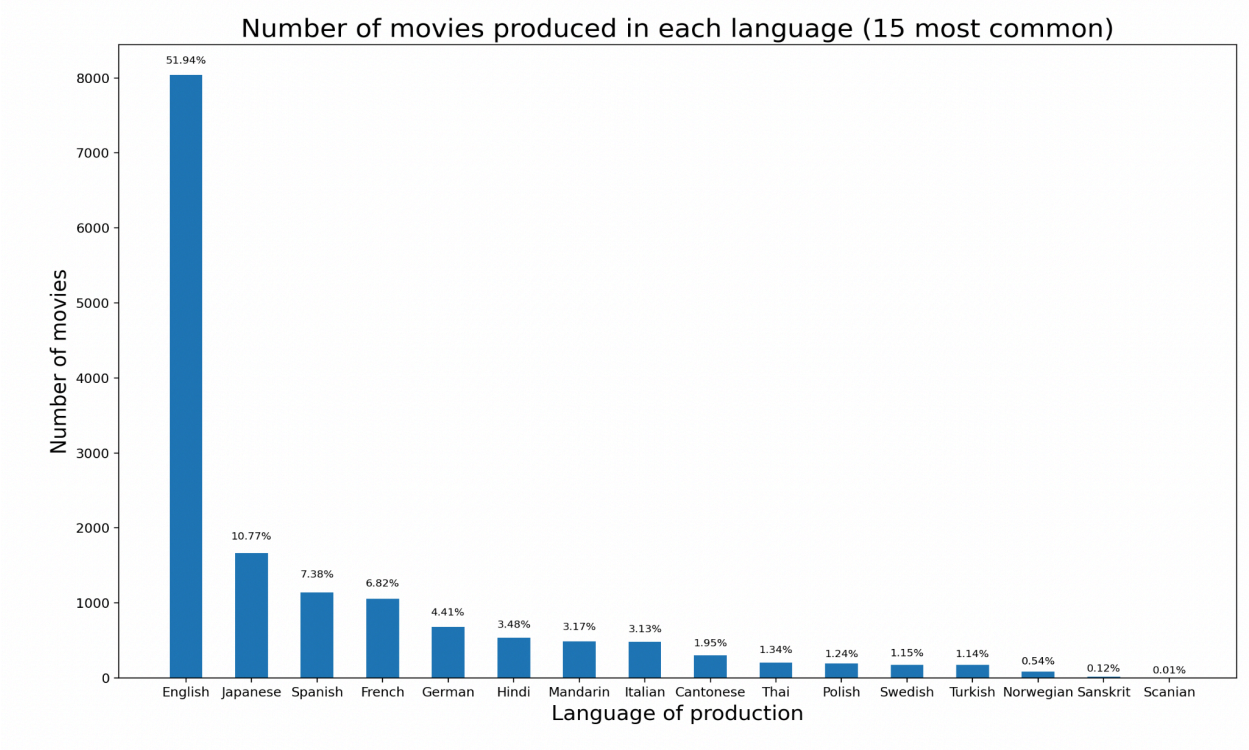
```
# Core loop for extracting scores in ratings()
for row in rows:
    if row[5] != "" and row[12] != "":
        hid_gem[row[0]] = float(row[5])
        imdb[row[0]] = float(row[12])
```

- Correlation Calculation in `correlation()`:

```
# Calculate Pearson correlation in correlation()

x = np.array(list(dict1.values()))
y = np.array(list(dict2.values()))
r = np.corrcoef(x, y)
```

Part 2: Which languages do contents mainly originate from and how many movies originate from each language?



This graph displays the 15 most widely used languages of production, defined as the original language in which the movie was produced. As we can observe graphically, the largest number of movies (about 8000) were produced in English, which is slightly above 50% of the content. This should be expected since Netflix is an American streaming platform, and the United States generally produces more entertainment content than other countries. We observe a relatively broader gap between the number and proportion of movies produced in English than in other languages: Japanese, the second most common language of production, has less than 2000 movies and about 10% of the total content; Spanish and French each represent about 1000 movies and around 7% of the total content; finally, the 11 other languages each represent a proportion of less than 5%. This reflects a relatively unequal representation of production contents from different countries on Netflix, since even when we selected the 15 most commonly used languages in content production, most of these languages occupy an infinitesimal portion of the total number of movies, which could originate from the fact that a great number of foreign movie producers do not have contracts with Netflix, so content produced in that language is not broadcasted via Netflix.

Note: The percentages displayed on top of each bar corresponds to the number of movies produced in that language divided by the total number of movies. Therefore, the percentages will

exceed 100% if one takes the sum, since some contents are produced in more than one language and have more than one language cited in their “Languages column”, so will be counted

Relevant code:

In this section, we analyzed the distribution of movies produced in different languages on Netflix. We focused on the top 15 languages by frequency:

- `lang_collector()`: Collects the number of movies produced in each language, returning a dictionary of counts for the most common languages.
- `lang_perc()`: Calculates the percentage of movies produced in each language relative to the total content.
- `plot_lang()`: Creates a bar chart to visually represent the number of movies produced per language, annotated with percentage values.

Key Code Snippets:

- Core Loop in `lang_collector()`:

```
# Collect movie counts by language in lang_collector()
for row in lang_reader:
    if row:
        row_list = row[3].split(",")  # Splits the 'Languages' entry
        for lang in row_list:
            lang = lang.strip()
            if lang != "":
                if lang not in lang_occ.keys():
                    if counter <= 15:
                        lang_occ[lang] = 1
                        counter += 1
                else:
                    lang_occ[lang] += 1
```

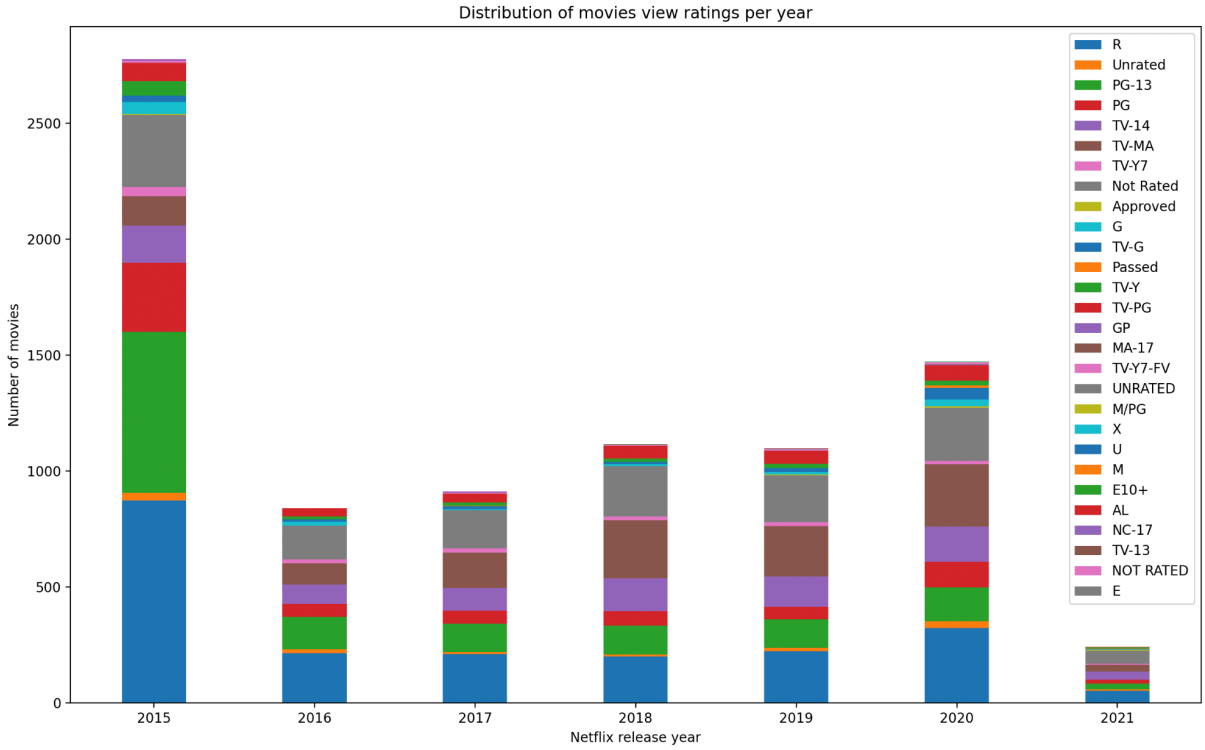
- Percentage Calculation in `lang_perc()`:

```
# Calculate language percentages in lang_perc()
for num in dct.values():
    perc = num / len(rows) * 100  # Calculates language percentage
    percent.append(round(perc, 2))
```

For the `plot_lang()` function, we used matplotlib to generate a bar chart visualizing the language distribution, annotated with percentages.

Part 3: How has the proportion of different view ratings (e.g. R-rated, PG-13, PG, etc) changed through time?

This stacked bar chart represents Netflix's annual movie releases from 2015 to 2021, segmented by viewership ratings. Notably, 2015 saw the highest total number of releases, whereas 2021 recorded the lowest. The data reveals that the TV-G category (represented first in blue) consistently holds the largest share across the years. Following TV-G, the PG-13 rating (shown in green) emerges as the second most prevalent category. There is a close contest between TV-14 and TV-MA ratings, depicted in purple and brown, respectively. While the proportions of R, PG-13, unrated, and TV-14 ratings remain relatively stable over the years—the variability in the proportion occupied by other ratings like PG (first red) is more pronounced. The distribution among the R, PG-13, PG, TV-14, TV-MA, and E ratings is fairly balanced. These observations suggest some variations in Netflix's target demographic over the years, although the distribution of movie ratings remains largely consistent with previous patterns.



Relevant code:

This analysis examines how the proportion of different viewership ratings (e.g., R-rated, PG-13, PG, etc.) has changed for Netflix releases from 2015 to 2021. We focused on identifying trends in content ratings over time:

- `view_rate()`: Constructs a nested dictionary that tracks the number of movies per year for each view rating, which serves as the basis for visualization.
- `view_plot()`: Plots a stacked bar chart to display the annual distribution of viewership ratings, allowing for easy comparison across years.

Key Code Snippets:

- Data Aggregation in `view_rate()`:

```
# Collecting movie counts by rating per year in view_rate()
for row in rows:
    year = row[19][:4]  # Extract year from release date
    rating = row[11]  # Extract view rating
    if rating in rate_dict and year in rate_dict[rating]:
        rate_dict[rating][year] += 1
```

- Plotting with `view_plot()`:

```
# Stacked bar plot for view ratings over time
bottom = np.zeros(len(years))
for rating, rate_counts in data.items():
    ax.bar(years, rate_counts, width, label=rating, bottom=bottom)
    bottom += rate_counts
```

Ethical reflection:

The main ethical issue our data set possibly evokes is the bias when it comes to data collection and interpretation. This manifests mainly through our first two questions. In our first graph, on one hand, the Hidden Gem score is calculated by combining ratings from different platforms (including IMDb but also other platforms such as Rotten Tomatoes) and takes into account a variety of factors such as viewership numbers and critical reviews. In particular, Hidden Gem scores have the aim of highlighting underrated and underappreciated. By contrast, the IMDb score is calculated by the IMDb platform itself from its own users with its own criteria. Therefore, the IMDb scores and Hidden Gem scores do not necessarily correlate, as shown by our first graph. This shows the importance of consulting diverse sources when exploring data and the importance of understanding the underlying criteria or methods used by each data source to draw accurate and reasonable conclusions. In our case, it could be easy to say overly negative or overly positive things about a movie by looking only at one scoring system and not the other.

Data bias could also manifest itself through our second graph on languages of production. As demonstrated, a very small number of languages are truly represented on Netflix, and content produced in English clearly dominates over content originally produced in a foreign language. This means that a great number of quality content from overseas will not be found on Netflix, which might significantly decrease the number of views they get and the revenue they make, since globally speaking, Netflix is possibly the most popular streaming platform. Therefore, if a study tried to use Netflix views to determine the popularity and quality of movies, it should consider the fact that not Netflix only streams a limited number of content even if it captures a large number of viewers because this would make English (especially American-produced content) much more popular among national and international viewers.

In today’s technology landscape, understanding and mitigating data biases is critical for developing fair and inclusive systems. As data increasingly influences decision-making, ethical considerations help ensure that technology serves diverse communities responsibly. This awareness aligns with my commitment to ethical standards in tech and resonates with Microsoft's mission to empower every individual and organization globally.