CSE 4/587 **Spring 2023**
Jason Mai, Cameron Toy **Project Phase #1**
Due 3/6/23

Problem Statement

Exploring the relationship between User Ratings and Netflix.

As one of the largest streaming platforms in the world, Netflix and IMDB collect a vast amount of data about user preferences. This is a significant problem for Netflix when it comes to investment decisions for making TV Shows, Movies, and even getting rights to stream certain shows and movies. The potential of our project that can contribute to Netflix's problem can be great since it can prevent huge losses in revenue for Netflix. By analyzing user ratings, Netflix can learn and gain insights of the factors that can satisfy or attract consumers.

Our group gathered initially gathered information from (https://www.kaggle.com/datasets/luiscorter/netflix-original-films-imdb-scores) "Netflix Original Films & IMDB Scores" which contains most of Netflix's movies and shows with their respective IMDB and TMDB scores, but the data was not clean. Some data like age-certification, IMDB and TMDB ratings were missing. We ultimately decided to use "Netflix Movies and TV Shows" dataset (https://www.kaggle.com/datasets/shivamb/netflix-shows) which included casts and directors for the movies and shows so we can further see if casts or directors play a role in good ratings. Unfortunately, this dataset doesn't contain most missing data that we can use to fill in for the dirty data of our first dataset. Once we gathered our data be then began the cleaning process.

Data Cleaning

We approached data cleaning in multiple ways:

- Filtering Data
- Sorting Data
- Normalizing the Data
- Merging Data
- Imputation on Data
- Handling Duplicate Data
- Removing Data

Firstly our group **merged** two datasets, one that contained Netflix's shows and movies, and the other that contained the missing ratings. This step involves combining two or more datasets into a single dataset. In this case, the missing ratings were filled in by merging the two datasets. This was done using the merge() function in pandas, with the on argument set to the column to be merged.

Secondly we started with **filtering** the data. In this step, our group filtered the data to preserve only the relevant movies that were past the year 1999. This is a common step in data cleaning, as it helps to remove irrelevant or outdated data that can affect the analysis. This was

done using a filter method that selects only the rows that met the specific criteria of being after the year 1999.

Thirdly we **sorted** our dataset. Sorting was done by movie type, then by IMDB rating. When data is sorted, it is easier to determine the data's distribution and understand the data better. The sorting process also helps identify outliers and inconsistencies in the data. In pandas, the sort_values() function was used with the by argument set to the columns to sort.

Fourth, we began **normalizing** the data. This step involved normalizing the IMDB scores by rounding them to a whole number. This is done to reduce the complexity of the data and to make it easier to analyze. Normalizing the data also helps to reduce any errors or discrepancies that may arise due to decimal places. The round() function in pandas was used to round off the IMDB scores to a whole number.

Then we began the **imputation** which is the process of filling in missing data with values that are likely to be correct. In this case, our group filled in null data in the "seasons" column by placing an integer of 0 for all movies. This was done because not all movies have seasons. Imputation was done using the fillna() function in pandas, with the value set to 0 for all missing values.

After we begin to handle **duplicate** data because it can cause errors in the analysis, and it is important to remove it. My group removed all duplicated data resulting from merging the two datasets. This was done using the drop_duplicates() function in pandas, which removes all rows that are duplicates based on a specific column or set of columns.

Finally, our group **removed** all rows that didn't have age-certification. This is important because some movies may not be suitable for all audiences, and it is necessary to remove them from the analysis. This was done using the dropna() function in pandas, with the subset argument set to the column containing the age-certification. Our group was content with the clean data we had and then we began the next process which is exploratory data analysis.

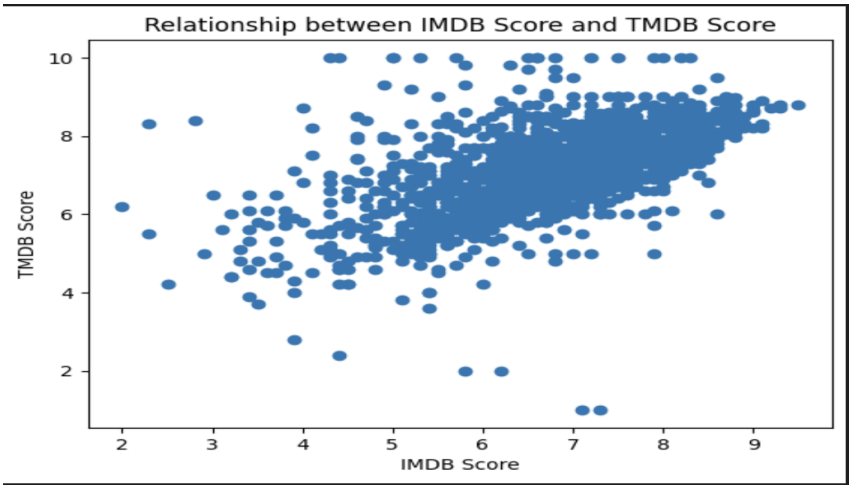
Exploratory Data Analysis

As the entertainment sector expands and changes, Netflix has emerged as one of the top streaming services, with these series' rankings being determined by well-known databases like IMDB and TMDB. Due to the fact that these businesses have millions of customers worldwide, they gather a lot of information regarding user preferences and behavior. This poses a huge difficulty when Netflix decides to make investments in things like buying streaming rights and generating original content. We are investigating what motivates contentment and attraction in order to address this difficulty. Our objective is to give Netflix useful information so that it can make informed decisions and avoid losing money. By promoting more effective and lucrative content generation, we can have a big impact on the entertainment sector.

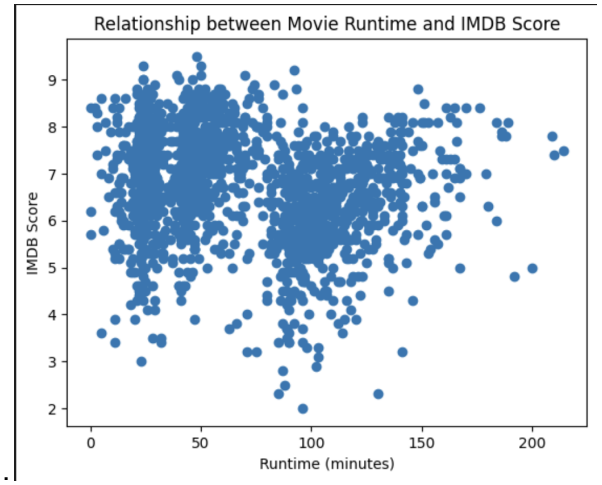
To arrive at our results, we conducted exploratory data analysis on the ratings of movies and shows on Netflix. We began by plotting the distribution of IMDB scores and TMDB scores, which allowed us to see the scores' range and the data's central tendency. We created these plots using Python's data visualization libraries, such as Matplotlib. We sparked our initial thinking by looking at the top 25 IMDB rated movies and shows Netflix has to offer. By looking at the data below we wanted to further explore and analyze what factors led these shows and movies to be rated highly.

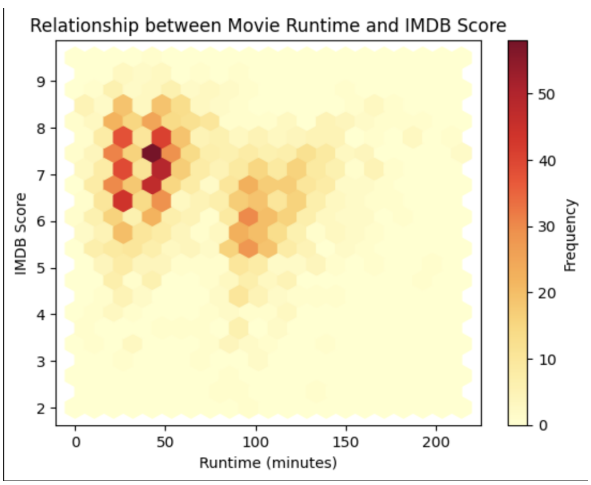
	title	type	imdb_score
3963	Breaking Bad	SHOW	9.5
2664	Our Planet	SHOW	9.3
1743	Avatar: The Last Airbender	SHOW	9.3
2088	Reply 1988	SHOW	9.2
4	Kota Factory	SHOW	9.1
440	The Last Dance	SHOW	9.1
1684	My Mister	SHOW	9.1
3832	DEATH NOTE	SHOW	9.0
282	Okupas	SHOW	9.0
1180	Leah Remini: Scientology and the Aftermath	SHOW	9.0
513	Attack on Titan	SHOW	9.0
2568	When They See Us	SHOW	8.9
979	Still Game	SHOW	8.9
1279	David Attenborough: A Life on Our Planet	MOVIE	8.9
3594	Narcos	SHOW	8.8
2557	Black Mirror	SHOW	8.8
1994	Better Call Saul	SHOW	8.8
864	Chappelle's Show	SHOW	8.8
2268	Vientos de agua	SHOW	8.8
444	Hospital Playlist	SHOW	8.8
2290	The Untamed	SHOW	8.8
2336	Peaky Blinders	SHOW	8.8
2279	BoJack Horseman	SHOW	8.8
222	Inception	MOVIE	8.8
...			
3037	The Haunting of Hill House	SHOW	8.6
21	Love on the Spectrum	SHOW	8.6
1817	Middleditch & Schwartz	SHOW	8.6
351	Shtisel	SHOW	8.6

Next, we compared the IMDB and TMDB scores by plotting them on a graph, which helped us determine the ratings' accuracy and consistency. We also plotted the average IMDB scores by release year and directors by IMDB score to understand the trends in ratings over time and identify the top and bottom-rated directors. Again, we used Python's data visualization libraries to create these plots.

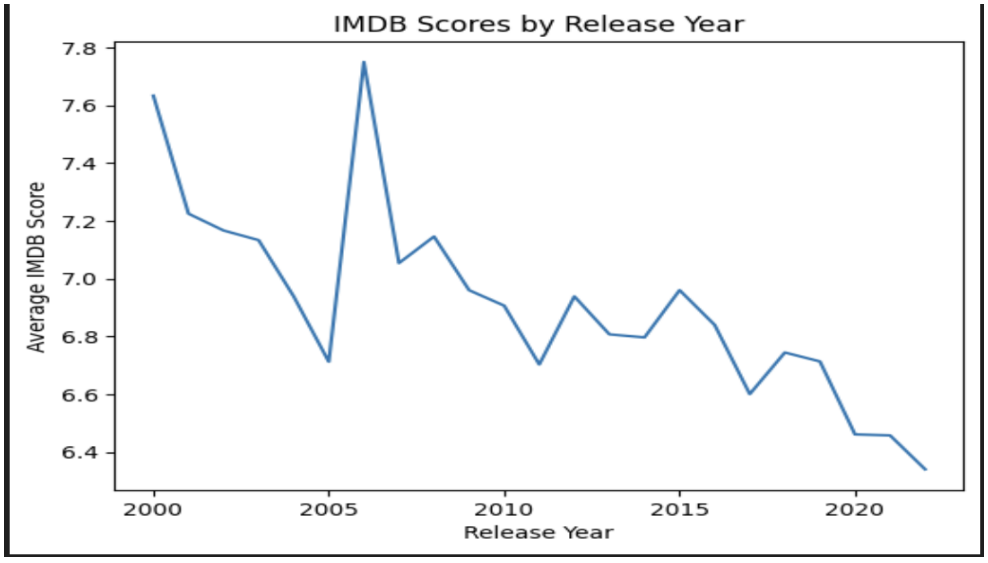


We plotted the data on a scatter plot and heat map to investigate the relationship between runtime and IMDB score. We used the Pandas library to manipulate and analyze the data and Matplotlib to create the visualizations

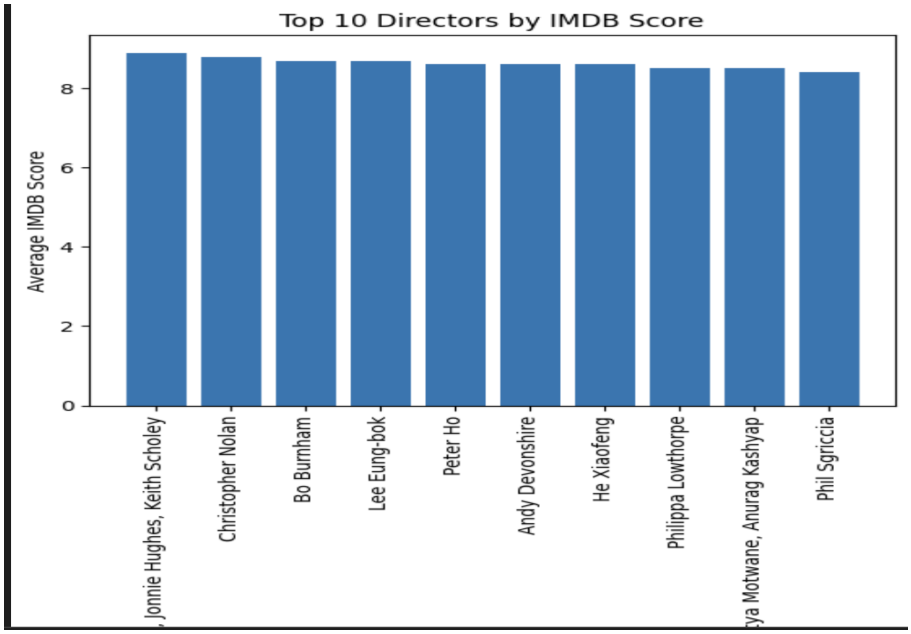




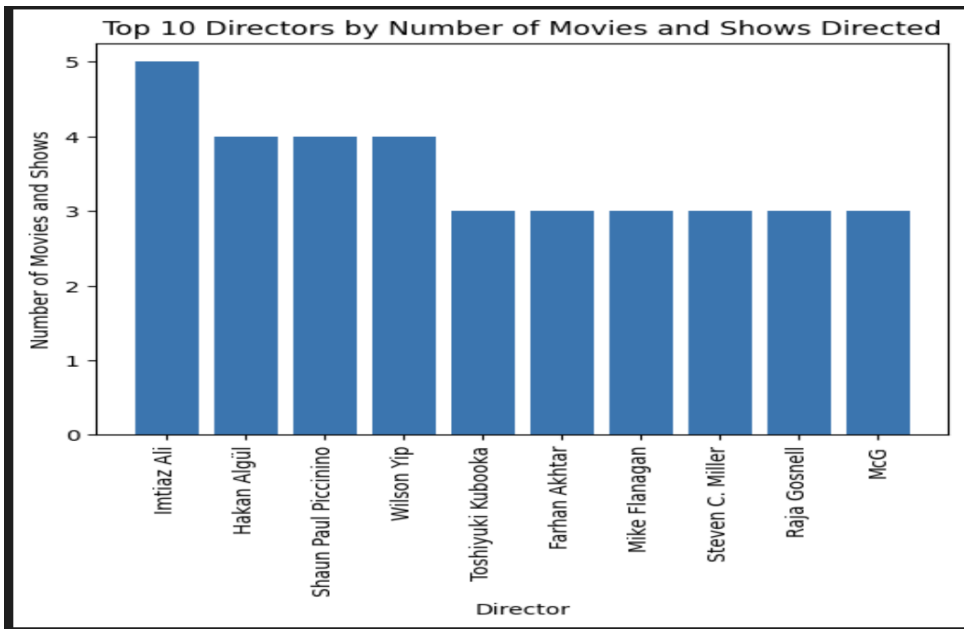
By plotting the Average IMDB Scores and the Release year, we gained insight into the overall trend of movie ratings over time. The highest-rated movies were released around 2007, with ratings gradually declining as we moved toward more recent releases. This information can help Netflix determine which movies and shows to invest in for streaming and prioritize those with high overall ratings.



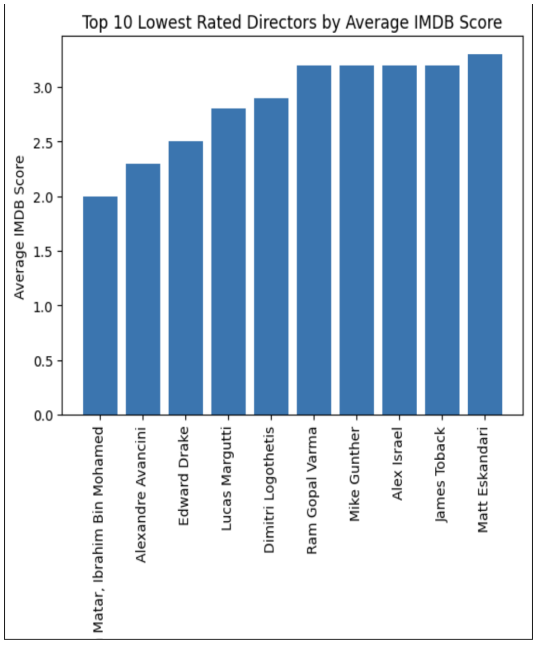
In addition to analyzing the overall trend in ratings, we investigated individual directors' impact on ratings by plotting the top 10 directors based on their IMDB scores. This allowed us to identify directors that Netflix should invest in when producing original content to ensure high ratings.

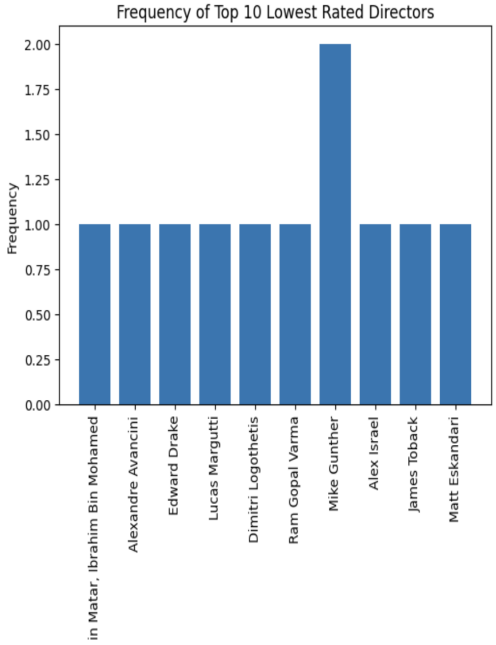


In addition we also investigated the top ten directors by the number of movies and shows they have directed to make sure that the analysis is not skewed due to having a high imdb score for just directing one movie or show.

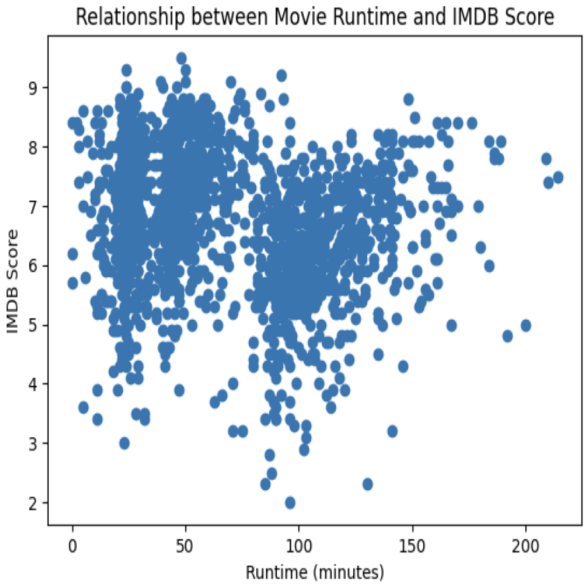


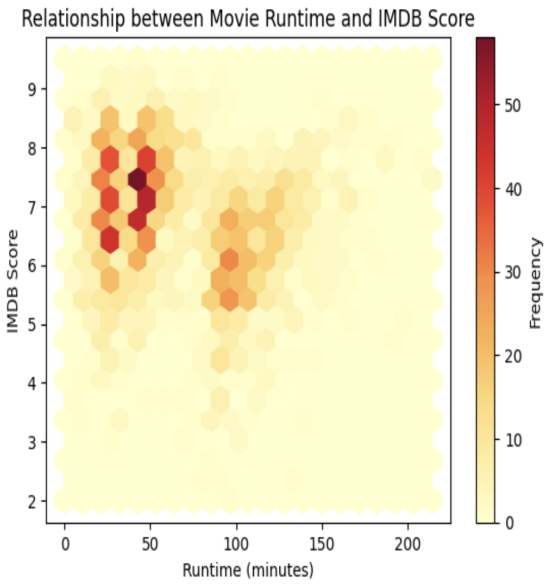
Conversely, we plotted the lowest-rated directors by average IMDB score to identify directors that Netflix should avoid when producing original content. To ensure that our data was not skewed, we also plotted the frequency of each director in our dataset to see how many shows or movies they took part in.



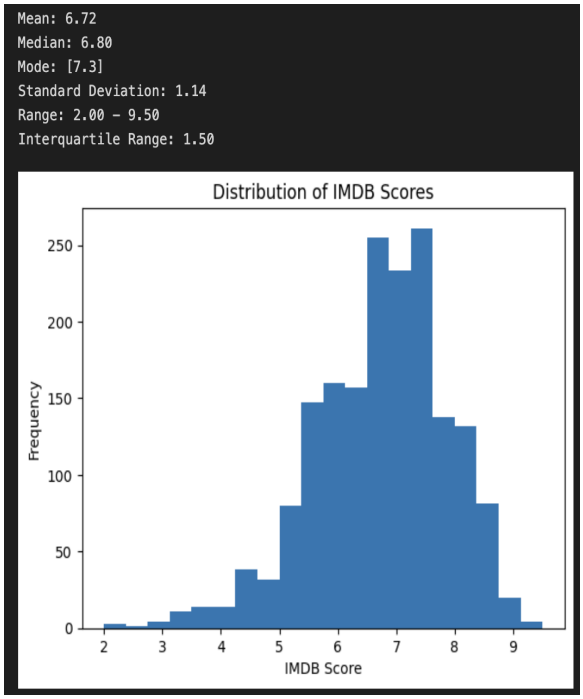


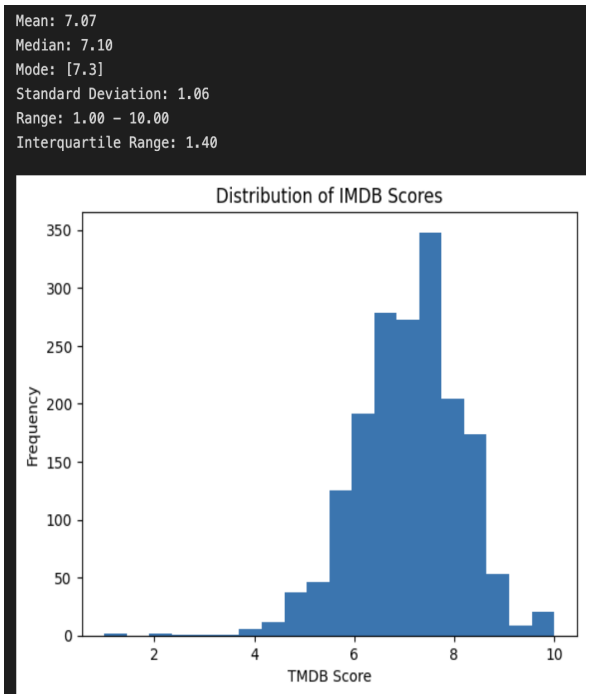
Lastly, we explored the relationship between movie runtime and IMDB score by plotting the data on a scatter plot and heat map. The mixed results indicated that a shorter runtime sometimes leads to higher ratings. However, we found that many movies with short runtimes of around 50 minutes had an average IMDB score of 6-8, whereas as the runtime increased, the IMDB score decreased slightly to approximately 5-7. This information can be used by Netflix to determine the optimal runtime for movies and shows to ensure high ratings.



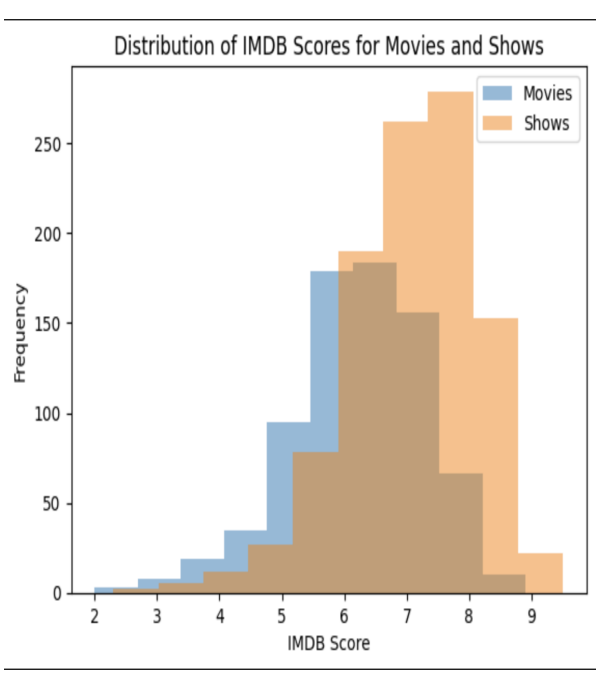


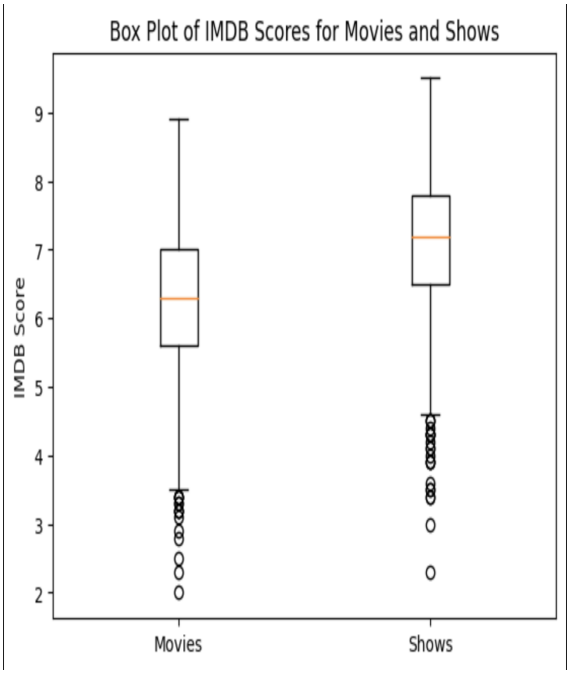
We ensured that our data was reliable and unbiased throughout our analysis by verifying our results and plotting the frequency of directors. In addition to using statistics to describe the data accurately, we used mean, median, mode, standard deviation, and interquartile range. This is shown in our initial figures below the distribution of IMDB Scores and TMDB scores.





One of our final exploratory data analysis we have done is analyzing the IMDB score between movies and shows. Our goal was to find if it is better for Netflix to invest in streaming/creating shows versus streaming/creating movies. This analysis shows that Shows average a higher IMDB score than Movies on Netflix's platform.





The ratings of movies and shows on Netflix were analyzed through various data visualization techniques and statistical measures. Our exploratory data analysis determined accuracy, rating trends over time, and the highest and lowest-rated directors. Using the information from our exploratory data analysis, Netflix can make data-driven decisions to improve its platform and provide a better user experience.