Data Mining Exploration of Netflix Movies and TV Shows

Jingrui Gan jingrui@umich.edu

Yu Yan kuminia@umich.edu

Background

In the era of digital streaming, the vast landscape of content available on platforms like Netflix presents an unprecedented opportunity for exploration and analysis. This data mining project embarks on a comprehensive investigation into the Netflix Movies and TV shows dataset, aiming to uncover valuable insights through the lenses of time series analysis, network structures, and a recommendation system. As viewership patterns evolve and user preferences become increasingly diverse, understanding the intricate dynamics within the Netflix ecosystem becomes pivotal.

Goal

In this data mining project, we hope to use data to predict the number of movies added by Netflix each month. The accurate prediction of Netflix's monthly increase in movie numbers is not only a key factor in business decisions, but also directly related to the improvement of market competitiveness. By predicting future movie launch trends, Netflix can formulate smarter business strategies and adjust resource investment to meet the growing expectations of users. In terms of market competition, predicting the increase in the number of movies allows Netflix to flexibly adjust advertising and marketing strategies to better meet user needs and maintain its leading position in the streaming media market.

In addition, because Netflix has many of its own original series and funded movies, choosing the right directors and actors becomes crucial. We will analyze the connections between directors and actors of existing films on Netflix. The popularity of the director and actors directly affects the commercial and artistic value of the film. This kind of influence can help attract more viewers, improve reputation, increase investment, form a brand, and establish a strong influence in the market, creating favorable conditions for the success of the film.

Besides, we will write a simple video recommendation system. Recommendation systems can make it easier for users to discover new content of interest, thereby increasing user retention and reducing churn rates. With accurate recommendations, users are more likely to spend more time watching content on Netflix, increasing platform activity and frequency of use.

Dataset

The data for our data mining project comes from the Kaggle website, the link is https://www.kaggle.com/datasets/shivamb/netflix-shows/data. This tabular dataset consists of listings

of all the movies and tv shows available on Netflix , and contains various information related to these movies and TV series, including title, director, actors, country, release date, rating, duration, description and other features. The data contains a total of 12 features and 8807 movies and TV Shows. We will use these features to analyze and mine Netflix's movie and TV series data.

```
print(df.info())
✓ 0.0s
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8807 entries, 0 to 8806
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   show_id       8807 non-null   object
 1   type          8807 non-null   object
 2   title         8807 non-null   object
 3   director      6173 non-null   object
 4   cast          7982 non-null   object
 5   country       7976 non-null   object
 6   date_added    8797 non-null   object
 7   release_year  8807 non-null   int64
 8   rating        8803 non-null   object
 9   duration      8804 non-null   object
 10  listed_in     8807 non-null   object
 11  description   8807 non-null   object
dtypes: int64(1), object(11)
memory usage: 825.8+ KB
None
```

Methods

In this data mining project, we will use data mining methods such as data processing, EDA, time series analysis, Network Analysis, and making recommendation systems through CountVectorizer and Cosine Similarity.

Exploratory data analysis (EDA) is a critical step in the data analysis process. It involves the initial examination and exploration of a data set to summarize its main characteristics, often with the help of statistical graphics and visualization techniques. The main goal of EDA is to discover patterns, identify anomalies, and gain insight into the underlying structure of the data

In time series analysis, we will perform data visualization analysis to make trends more intuitive and visible, helping users quickly understand the trend of data over time, which is an important means of observing trends and patterns. By observing time series graphs, outliers or outliers in the data can be easily discovered, helping to quickly identify problems and take appropriate measures. We also decompose the time series into trends, seasonal and residual decompositions to help gain a deeper understanding of the components of the time series and extract key information. We also used the Augmented Dickey-Fuller (ADF) test to perform stationary tests on the time series to ensure the validity of our time series model. At the same time, we also made graphs of the autocorrelation function (ACF) and the partial autocorrelation function (PACF). Observing ACF plot and PACF plot

can help us determine the appropriate time series orders p and q. We chose SARIMA (Seasonal AutoRegressive Integrated Moving Average) as our time series model. It is an extension of the ARIMA (AutoRegressive Integrated Moving Average) model and is used to process time series data with seasonal changes. The SARIMA model includes autoregressive (AR), integral (I), moving average (MA) and seasonal components. It is a powerful time series modeling tool. We train and visualize our model through the training set and test set, and finally put the model into Use the entire data set to predict the data and obtain the confidence interval of the prediction.

We used Network Analysis to analyze the relevance and popularity of directors and actors in Netflix's US movies. Network analysis can reveal the relationship structure in the data, helping to discover the interrelationships between entities, connection patterns, and underlying social network structures. This is very important for understanding the correlation behind the data. We can clearly understand the cooperation between the director and the actors through the Network graph. We calculated the Degree Centrality, Betweenness Centrality, and Eigenvector Centrality of the data. Degree Centrality is a simple measure that counts the number of edges (connections) a node has. Betweenness Centrality measures the extent to which a node lies on the shortest paths between other nodes in the network. Eigenvector Centrality measures a node's importance based on the quality of its connections, giving more weight to connections with highly central nodes. Different calculation metrics can bring us different aspects of data analysis results.

We used CountVectorizer and Cosine Similarity to create a recommendation system. CountVectorizer is a commonly used text vectorization tool in natural language processing. Its main function is to convert text data into a word frequency matrix. It converts each word (or vocabulary) in the text into a feature and counts the number of times each word appears in the text. In natural language processing, cosine similarity is often used to compare the similarity between texts. Representing the text as word vectors and calculating the cosine similarity between them can measure the semantic closeness of the text.

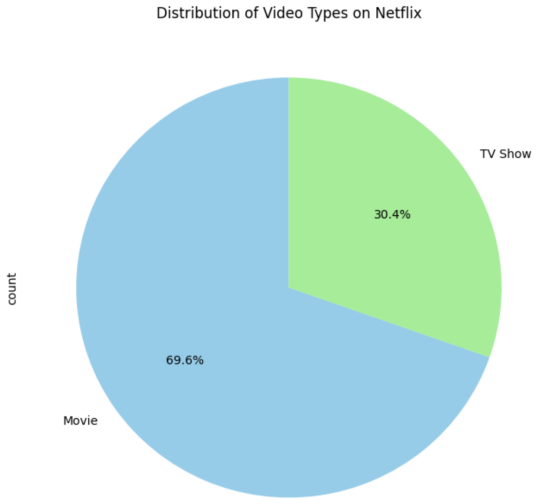
Data Processing

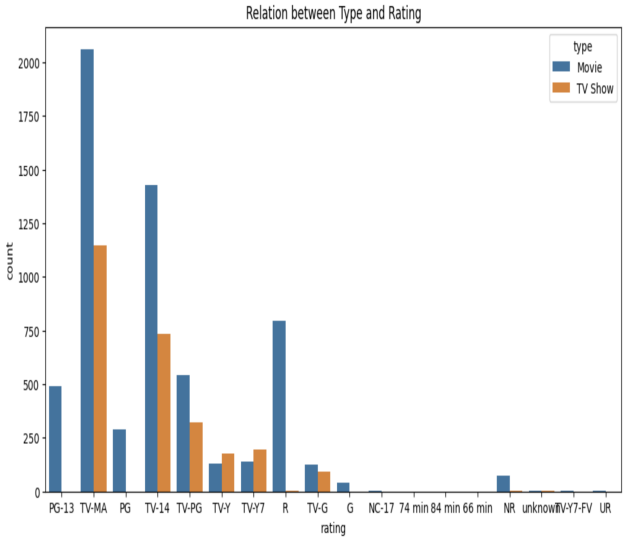
Before digging deep into the data, we process the data first. This ensures that the data used in the analysis process is of high quality, reliable and suitable for the required tasks, thereby increasing the effectiveness and credibility of data analysis and mining. First, we found that there are many missing values in the data, especially in the three features of movies, actors, and countries. They have 2634, 825, and 831 missing values respectively. We did not directly choose to delete this data, but replaced these missing values with "unknown" to indicate that we do not have this information, which can make our data content more reasonable. Then, when we browsed the data, we found that in the two features of director and actor, many times there was not only one director and one actor. Therefore, in order to facilitate the processing and calculation of the number of directors and actors in the

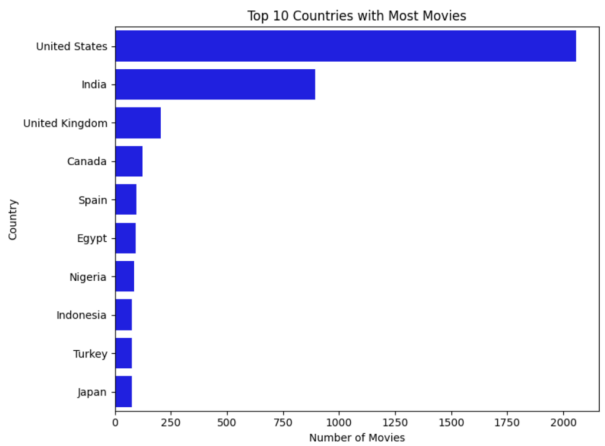
subsequent data mining process, we used the apply function to create two new columns to store them in the list type.

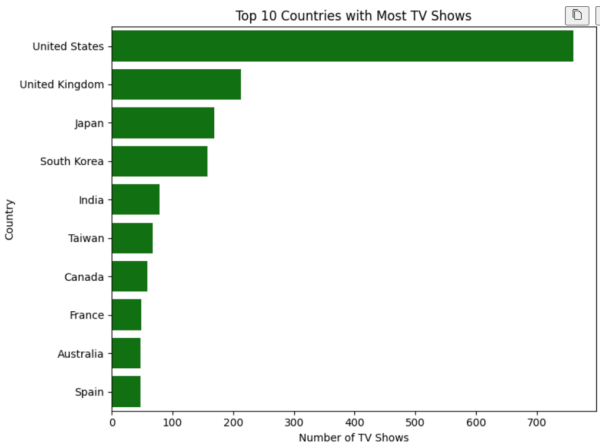
Exploratory Data Analysis

In the EDA part, we first made a pie chart to compare the number of movies and TV Shows in Netflix data. We can clearly find that the ratio of movies to TV Shows in Netflix's library is approximately 7:3. From the bar chart of movies and TV Shows, we can find that in Netflix's resource library, the top three countries providing movies are the United States, India, and the United Kingdom, and the top three countries providing TV Shows are the United States, the United Kingdom, and Japan. We can find that the vast majority of Netflix resources come from the United States. I think the reason why Japan ranks third in the number of TV Shows may be because Japanese anime has an exclusive category on Netflix.In addition, we also made a barplot to analyze the proportion of videos with different ratings on Netflix. We can find that the number of TV/MA rated movies and TV Shows are both ranked first. Programs with this rating are generally not suitable for people or individuals under the age of 17 (some sources may say 18). Therefore, we can infer that most videos on Netflix are available to adult users, so it is important to have an age verification system.





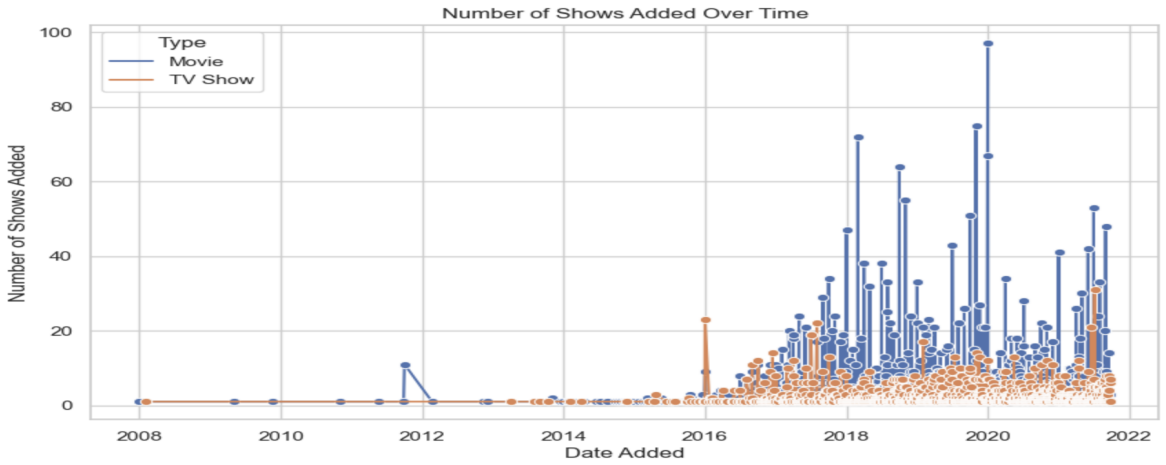


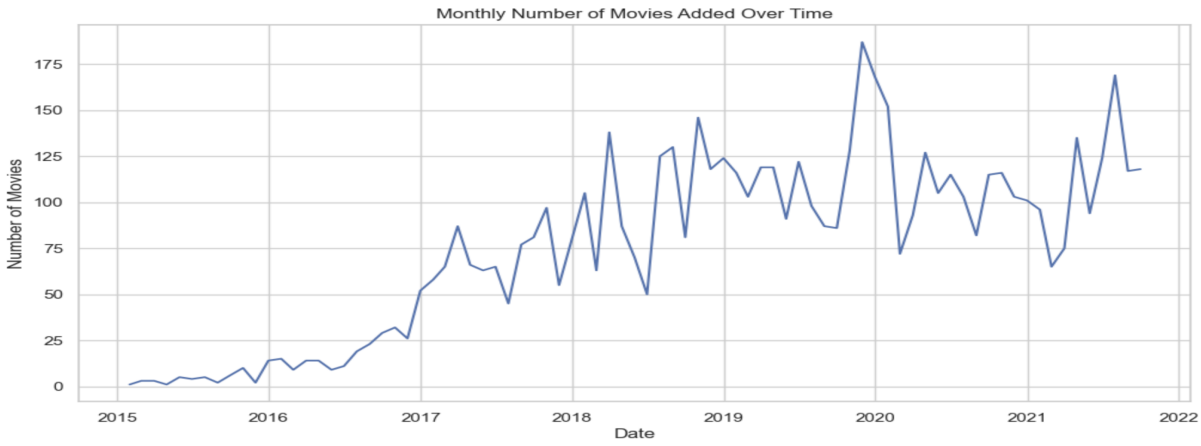


Time Series Analysis

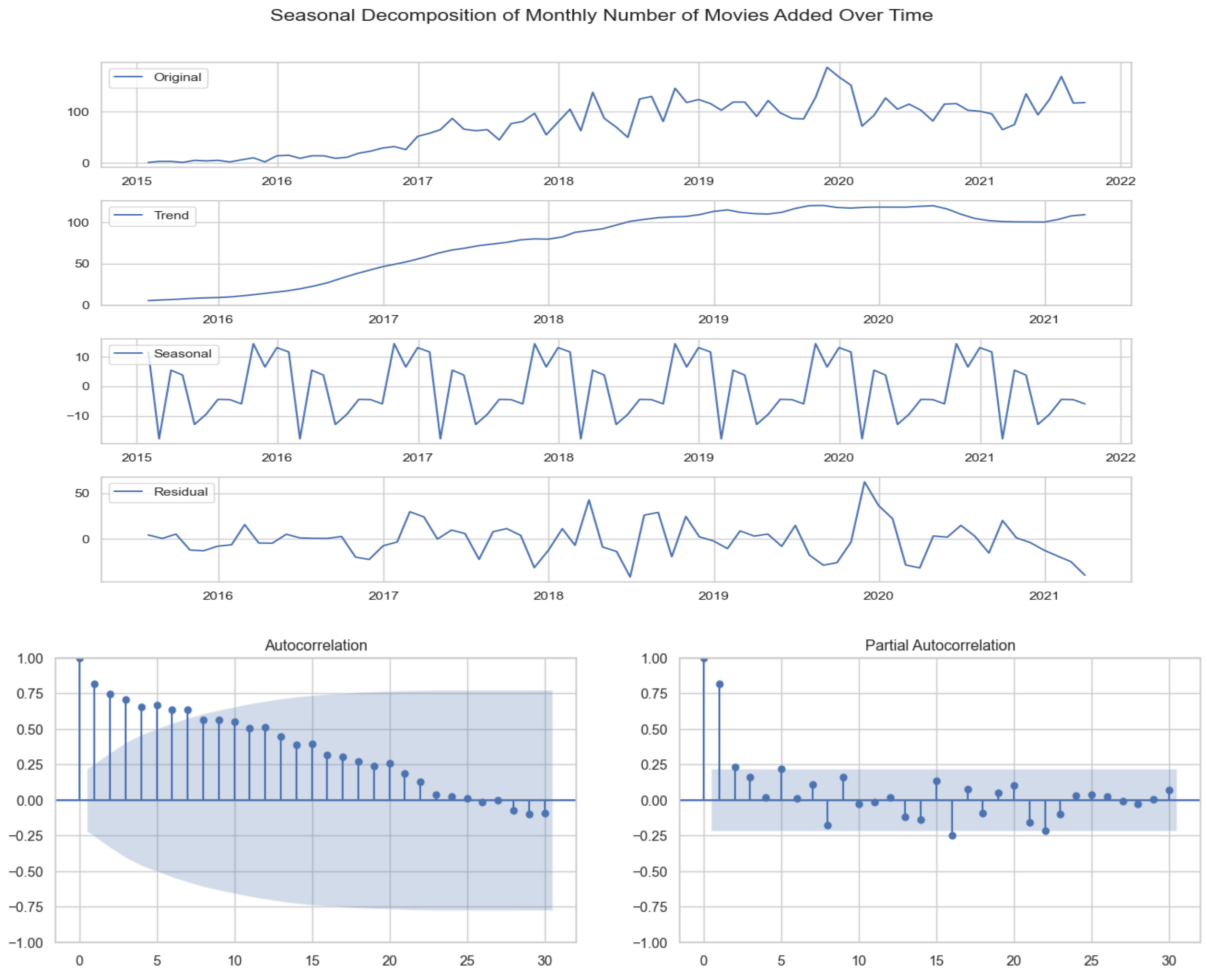
In this part, we use time series analysis for data analysis and mining. First, we use a visualization to reflect the number of new movies and TV Shows added by Netflix in recent years. From the picture, we can find that Netflix has been focusing heavily on movies in the last few years. We can also find that Netflix's movie increase reached its peak in early 2020, but then plummeted due to the Covid-19 epidemic. TV Shows haven't shown as much of a change as movies, so we decided to predict Netflix's future movie additions.

First, we added up the number of movies per month to calculate the number of movies added to Netflix each month, and then drew a time series model of the number of movies added.



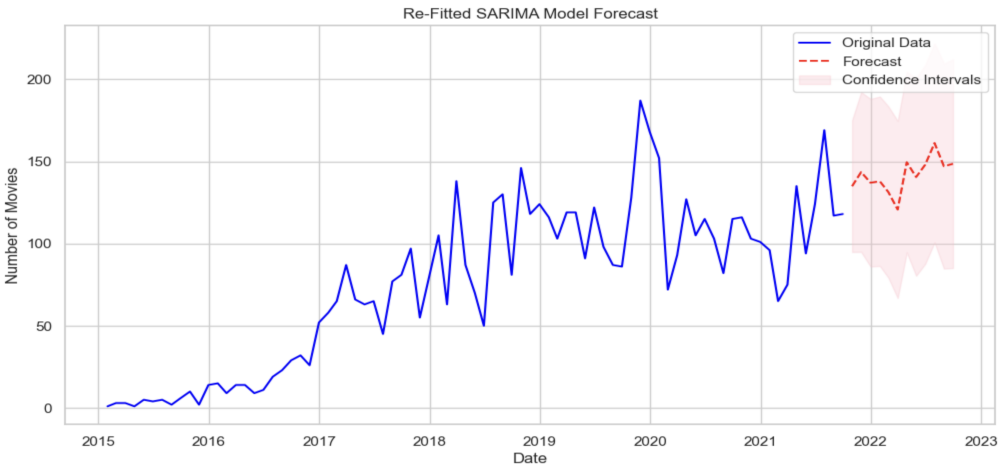


Then we started to dig deeper into the time series data. We first decomposed the data. We can clearly find that the data has obvious increasing trends and seasonal patterns. Before building the model, we need to analyze whether the time series is stationary. Using the Augmented Dickey-Fuller (ADF) test, we can see that the p-value is approximately 0.5419, which is significantly larger than 0.05, so we reject the null hypothesis and conclude that the data is not stationary. Also, from the ACF plot and PACF plot, the ACF decreases gradually and the PACF has a sharp drop, so we need a first-order differencing.



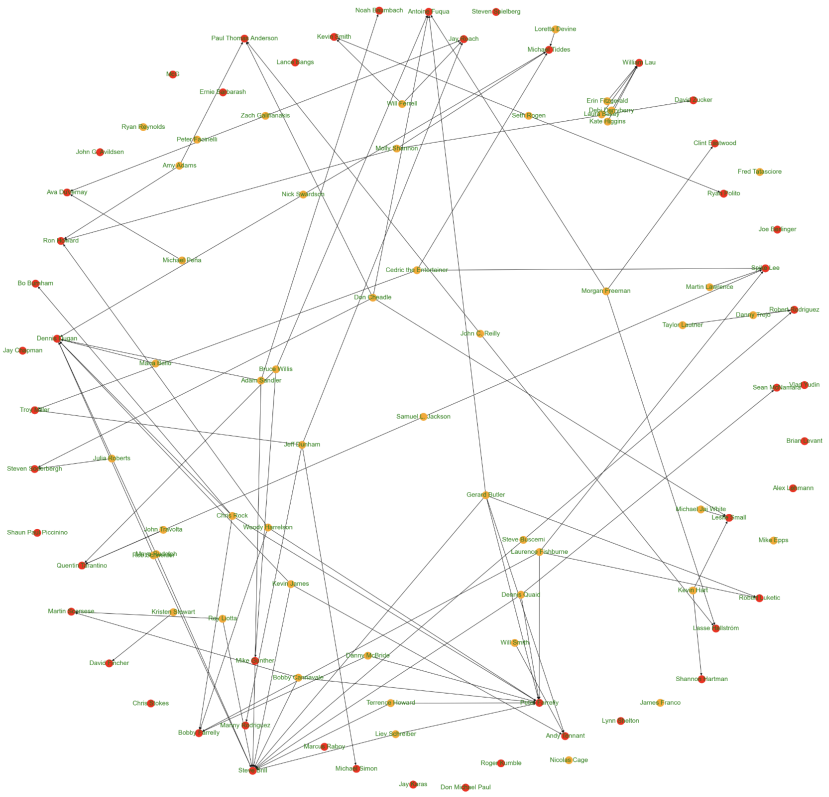
We divided the data into a training set and a test set based on around 2021 to analyze the effect of our model. We chose SARIMA as our time series model. Based on ACF plot and PACF plot, we obtained the values of parameters p and q, which are 2 and 8 respectively. In addition, our data has a periodicity of one year, so s is 12. We also need a first-order differencing, so d equals 1. So we get the final SARIMA model. After fitting the model on the training set, we used the model on the test set to predict and visualize the monthly increase in movies after 2021. We can find that almost all of our prediction results have very high similarity trends, and our real results are almost always within the 95% confidence interval of our prediction results.

We applied the SARIMA model to the entire movie data set, and finally obtained a prediction of movie growth in the next 12 months. Due to various environmental and resource uncertainties, although our forecast may not be 100% accurate, the forecast results can still help Netflix conduct more effective resource planning. Understanding future content growth trends can help optimize server and bandwidth resources to meet user demand for new content.



Network Analysis

In the Network section, we are going to study the partnership between directors and actors in American films on Netflix. We extract all the actors and directors in the data set and the number of times they appear, and select the top 50 (the number can be adjusted by modifying the code parameters) to become the nodes of our Network Graph and use red to represent directors and yellow to represent actors. If there has been a collaborative relationship between the director and the actor, they will be connected by edge. In addition, we also used degree centrality to find the node with the largest number of edges in the Network graph. The result was director Steve Brill, who has collaborated with 11 actors who appear in the top 50.



In addition, we also used degree centrality, betweenness centrality, and eigenvector centrality to analyze the popularity of directors and actors. Degree centrality provides information about the importance or influence of each node (actor or director) in the network based on the number of connections it has. Therefore, degree centrality can tell us which actors are most popular among directors and which directors are most popular among actors, which is Samuel L. Jackson and Steve Brill. High betweenness centrality for an actor indicates that they act as bridges or intermediaries between different groups of actors or directors in the collaboration network. Directors with high betweenness centrality contribute to the overall cohesion of the network by directing movies involving diverse sets of actors. The result is Samuel L. Jackson and Steve Brill. High eigenvector centrality for an actor indicates not only a large number of collaborations but also collaborations with other well-connected actors. High eigenvector centrality for a director suggests that they direct movies involving well-connected actors. The result is Terrence Howard and Elizabeth Banks.

Top 10 Actors by Collaborations using Degree Centrality:

	Name	Degree Centrality
0	Samuel L. Jackson	0.020467
1	Adam Sandler	0.019547
2	Dennis Quaid	0.018512
3	Terrence Howard	0.016327
4	Liev Schreiber	0.015753
5	Maya Rudolph	0.014718
6	Bobby Cannavale	0.014258
7	James Franco	0.013798
8	Fred Armisen	0.013683
9	Nicolas Cage	0.013568

Top 10 Directors by Collaborations using Degree Centrality:

	Name	Degree Centrality
0	Steve Brill	0.008936
1	Martin Scorsese	0.007954
2	Peter Farrelly	0.006481
3	Steve Carr	0.006088
4	Elizabeth Banks	0.005499
5	Quentin Tarantino	0.005401
6	Steven Spielberg	0.005302
7	Clint Eastwood	0.005302
8	David Fincher	0.005008
9	Robert Rodriguez	0.005008

Top 10 Actors by Collaborations using Betweenness Centrality:

	Name	Betweenness Centrality
0	Samuel L. Jackson	0.027149
1	Dennis Quaid	0.018353
2	Terrence Howard	0.015828
3	Adam Sandler	0.013262
4	Fred Tatasciore	0.012922
5	Nicolas Cage	0.012026
6	John Travolta	0.011812
7	Liev Schreiber	0.011687
8	Ray Liotta	0.011018
9	Chris Rock	0.010276

Top 10 Directors by Collaborations using Betweenness Centrality:

	Name	Betweenness Centrality
0	Steve Brill	0.054100
1	Quentin Tarantino	0.031102
2	Dennis Quaid	0.025990
3	Peter Farrelly	0.023792
4	Garry Marshall	0.022859
5	Robert Rodriguez	0.020163
6	Antoine Fuqua	0.019875
7	Martin Scorsese	0.019256
8	Elizabeth Banks	0.018084
9	Steve Carr	0.017915

Top 10 Actors by Collaborations using Eigenvector Centrality:

	Name	Eigenvector Centrality
0	Terrence Howard	0.142241
1	Emma Stone	0.121160
2	Jack McBrayer	0.116736
3	Liev Schreiber	0.115964
4	Bobby Cannavale	0.114865
5	Patrick Warburton	0.113965
6	Elizabeth Banks	0.112464
7	Dennis Quaid	0.108843
8	Gerard Butler	0.108226
9	Maya Rudolph	0.107010

Top 10 Directors by Collaborations using Eigenvector Centrality:

	Name	Eigenvector Centrality
0	Elizabeth Banks	0.292807
1	Steve Brill	0.217448
2	Peter Farrelly	0.205766
3	Steve Carr	0.204265
4	Brett Ratner	0.197621
5	Rusty Cundieff	0.197237
6	Griffin Dunne	0.195302
7	Will Graham	0.194150
8	James Duffy	0.194150
9	Jonathan van Tulleken	0.194150

Recommendation System

Everyone must have experienced relevant recommendations from film and television platforms. For example, when you search for or open a movie or TV series, the system will recommend your search results or works with similar content that you open. These similar works can have the same director, the same actors, similar genres and similar plots. By recommending content that matches users' preferences, it can effectively stimulate users' interests, improve user experience, and help users more easily discover and appreciate film and television works that match their interests and tastes. So we made a Netflix recommendation system using the information provided in the data set.

Our recommendation system adopts a content-based filtering method, using the video title, category, director, actor, country, rating, theme, and introduction data in the Netflix dataset. Through data processing, we concatenate these data to create The "text_content" feature, which helps us calculate the similarity score. Additionally, we used CountVectorizer and Cosine Similarity to process "text_content" and calculate the similarity score. CountVectorizer is a text processing tool that converts a collection of text documents into a matrix of token counts. With our text data transformed into a meaningful representation, we proceed to compute cosine similarity scores. This measure quantifies the similarity between pairs of records based on their text content. Next, we create a function as our recommendation function. This function takes the title of the movie or TV series as input, searches the corresponding index, calculates the similarity score and sorts, and finally returns the top 10 recommended movies or TVs. show. If the input title is not in our data set, the recommendation function will output 'Title was not found in our resource library'. Below are several examples of recommendation results from our recommendation system.

```
recommendations('Stranger Things')
✓ 0.0s
```

	Recommendation
0	Beyond Stranger Things
1	The Umbrella Academy
2	The Twilight Zone (Original Series)
3	Motown Magic
4	Manifest
5	The Messengers
6	The 4400
7	Mystery Science Theater 3000: The Return
8	Nightflyers
9	The Vampire Diaries

```
recommendations('Jaws')
✓ 0.0s
```

	Recommendation
0	Jaws 2
1	Poseidon
2	Jaws: The Revenge
3	Indiana Jones and the Temple of Doom
4	Mutiny on the Bounty
5	Rocky
6	Indiana Jones and the Raiders of the Lost Ark
7	Cleopatra Jones
8	Indiana Jones and the Last Crusade
9	Adrift

```
recommendations('Avengers: Endgame')
✓ 0.0s
'Title was not found in our resource library'
```

When we input a Netflix original TV series "Stranger Things" into the recommendation function, the highest similarity we got was "Beyond Stranger Things", which is very similar to our input name. It is a retrospective show about Stranger Things, discussing the inspiration for the series, sharing behind-the-scenes stories and analyzing key aspects of the storyline. In the top ten list, we can also find TV series with similar themes to Stranger Things such as The Twilight Zone.
When we enter a classic movie into the recommendation function: Jaws. Not only will we get the sequels to the series, Jaws 2 and Jaws: The Revenge, but we'll also get other Steve Spielberg-directed films like Indiana Jones and the Temple of Doom and Indiana Jones and the Raiders of the Lost Ark. This shows that our recommendation results can well cover different recommendation directions.
When we enter a movie "Avengers: Endgame" that is not in the Netflix resource library into the recommendation function, the recommendation system will output "Title was not found in our resource library". The user will then know that the movie is not in the library and start the next search.

Conclusion

In this data mining project, we conducted multi-directional analysis and mining of Netflix's movie and TV Shows resource library data. We used the SARIMA time series model to predict the increase in the number of Netflix movies. We use Network to explore the popularity and connections between directors and actors in Netflix's US movies. We also use cos similarity to create a recommendation system to recommend similar resources to users. It is very interesting to use various data mining techniques to deeply analyze the trends and internal relationships of the data. Mining Netflix data can bring various benefits to the company and users. We can use the information and conclusions obtained from data mining to realize the value of data.

Statement of Work

Section	Jingrui Gan's Contribution	Yu Yan's Contribution
Data Source	50%	50%
Data Processing	20%	80%
EDA	35%	65%
Time Series Analysis	80%	20%
Network Analysis	65%	35%
Recommendation System	30%	70%
Writing Report	50%	50%
Making Poster	70%	30%