

**Shri Ramdeobaba College of Engineering and**

**Management,Nagpur**

Department of Electronics Engineering

ENP361-3: Machine Learning Lab

Date: 19 April 2022

Micro Project

**NETFLIX CASE STUDY**

# Abstract -

A. To ensure retention what all things they need to build while content creation that will earn the best in terms of revenue, popularity and acclaim. B. To come up with characteristics of movies that will help them achieve their business goal

Let us take a look at some very important models of Netflix data to understand what’s best for their business. Some of the most important tasks that we can analyze from Netflix data are:

* understand what content is available
* understand the similarities between the content
* understand the network between actors and directors
* what exactly Netflix is focusing on
* and sentiment analysis of content available on Netflix.

**Introduction-** Netflix, Inc. is an American technology and media services provider and production company headquartered. Netflix was founded in 1997 by Reed Hastings and Marc Randolph in Scotts Valley, California.

The company’s primary business is its subscription-based streaming service, which offers online streaming of a library of films and television series, including those produced in-house.Netflix is a popular entertainment service used by people around the world.

This EDA will explore the Netflix dataset through visualizations and graphs using python libraries, matplotlib, and seaborn.We used TV Shows and Movies listed on the Netflix dataset from Kaggle. The dataset consists of TV Shows and Movies available on Netflix as of 2019. The dataset is collected from Flixable, which third-party Netflix search engine.

**Dataset and Features:** We used TV Shows and Movies listed on the

Netflix dataset from Kaggle. This dataset was taken from kaggle this dataset contains details about the director and country of shows and has a size of

3.24MB

The dataset contains the following main features

Id= show ID

Type = movie or a TV show

Title = the name of the show

Director = name of director of the corresponding shows

Cast = actors and actress of the show

Country= country which producing the show

In this dataset there are significant amount of null values as well which will get dropped in our code.

**Method -** The project will analyse a data-set on Netflix movies and Tv shows. The data set has been derived from keras. I am going to analyse the data to collect and visualize information on the movies and Tv shows. The project is using python, pandas, numpy, matplotlib as the tools to do the analysis

**1.Data visualisation and cleaning -**Data Profiling & Cleaning

Data Cleaning means the process of identifying incorrect, incomplete, inaccurate, irrelevant, or missing pieces of data and then modifying, replacing, or deleting them as needed. Data Cleansing is considered as the basic element of Data Science.

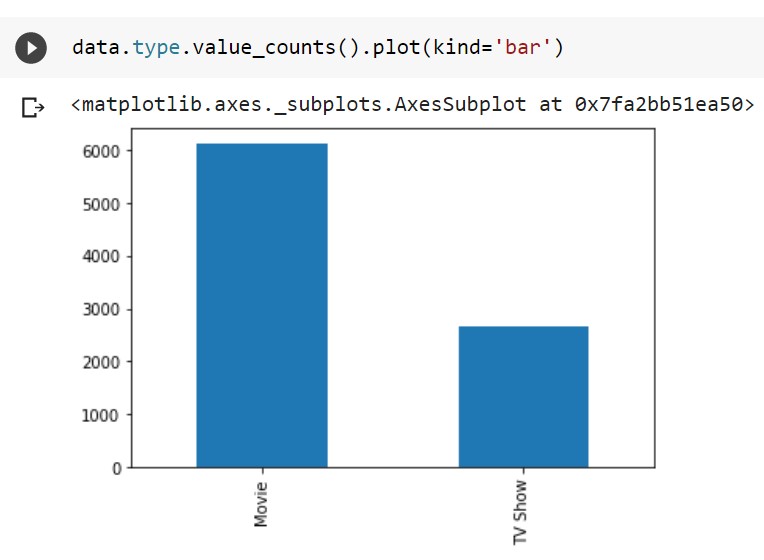
As we checked the data info we found the missing values in some coloumns like director and country, etc, hence we checked for the duplicated values in the entire column to hence clean the data and then fill the blank spaces with a variable ‘Unknown’ using the function fillna()

After filling the spaces we need to drop the unwanted columns from our dataset, we further will simplify our dataWe have also used DateTimeIndex function to give the universal date-time format

**2. EDA -**Exploratory Data Analysis

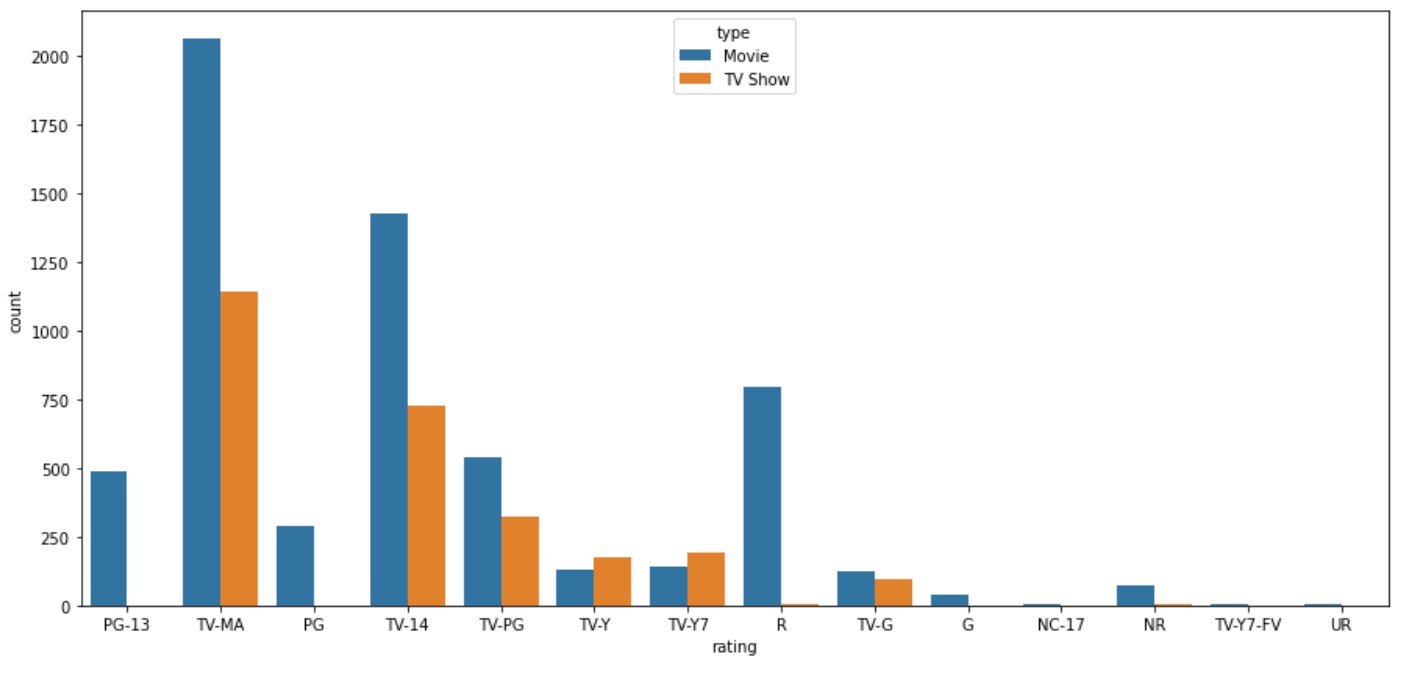
* **OVERALL ANALYSIS OF MOVIES AND TV SHOWS** -

We have first divided the data into two variables movies and shows for our ease and further done the deep analysis of the data Analysis entire Netflix dataset consisting of both movies and shows. Let’s compare the total number of movies and shows in this dataset to know which one is the majority.

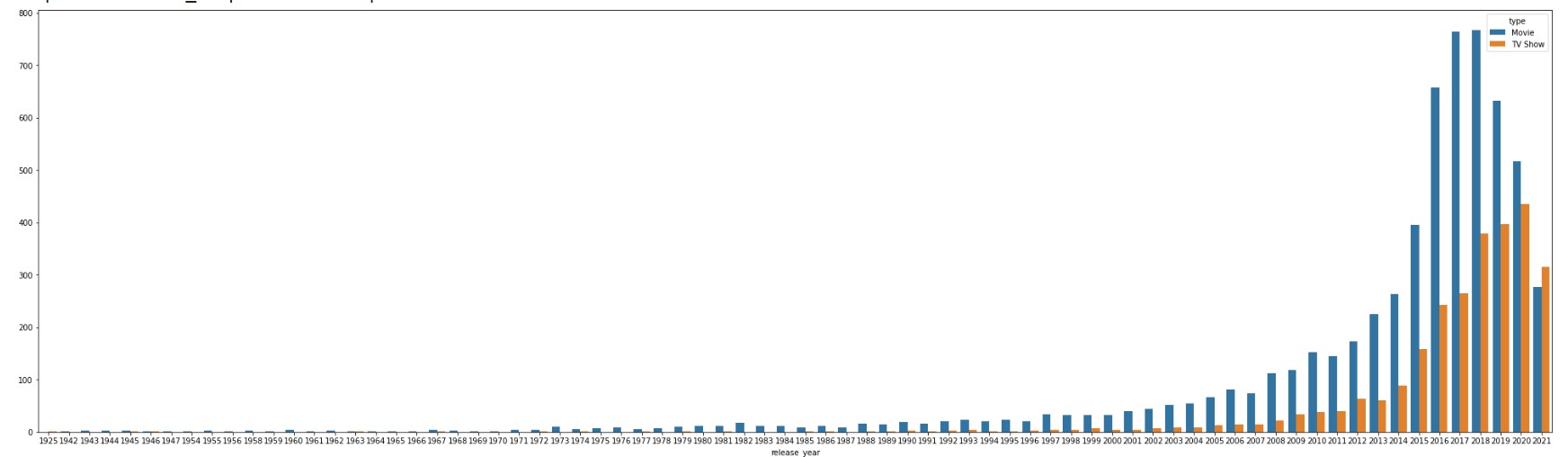


* **MOVIE RATING ANALYSIS** - The largest count of Netflix content is made with a “TV-14” rating. “TV-14” contains material that parents or adult guardians may find unsuitable for children under the age of 14. But the largest count of TV shows is made with a “TV-MA” rating.

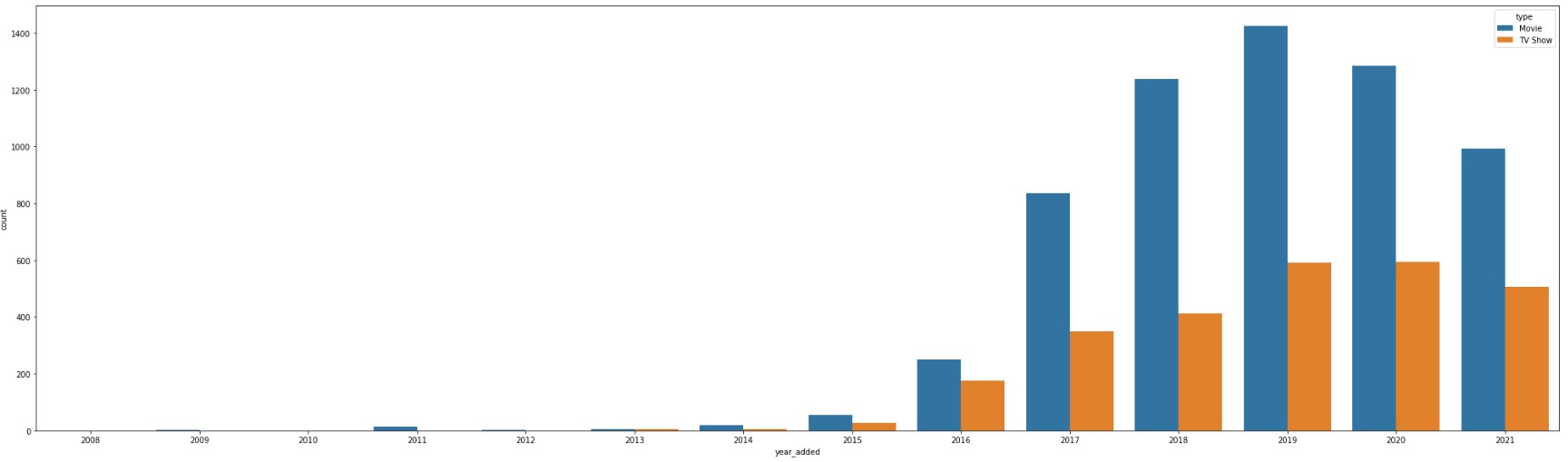
“TV-MA” is a rating assigned by the TV Parental Guidelines to a television program designed for mature audiences only.



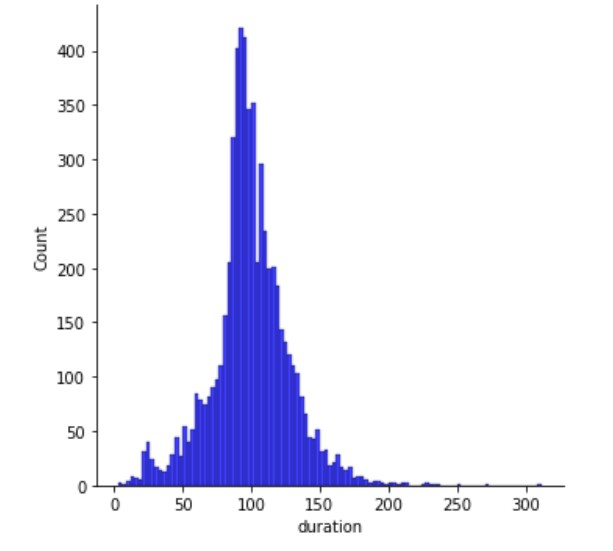
* **RELEASE YEAR:** Analyzing in what year the movies are released



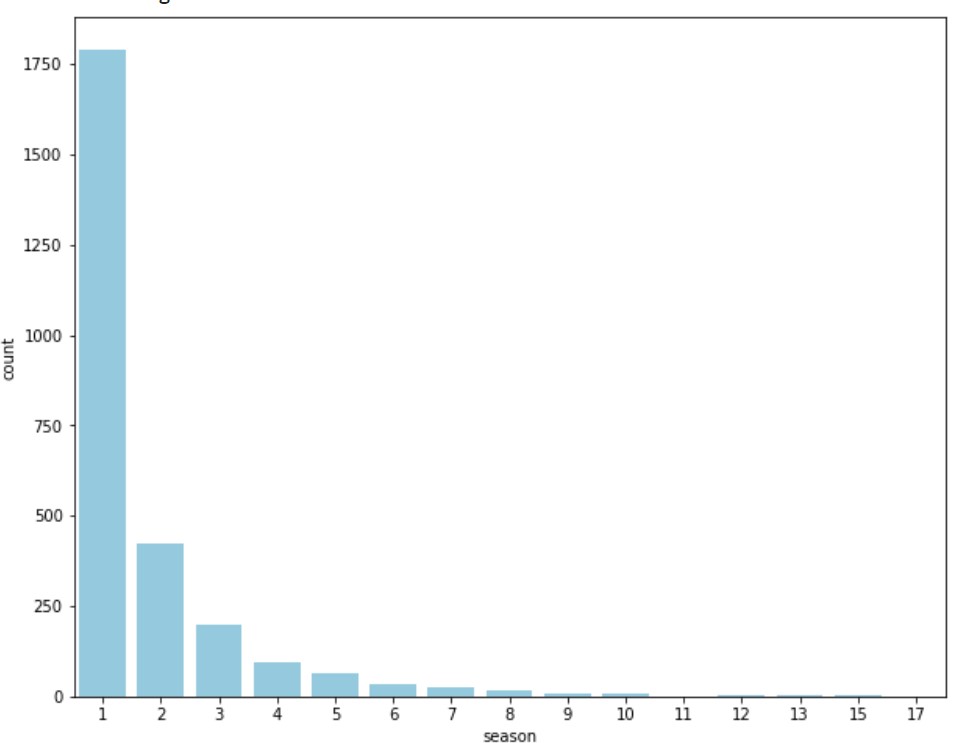
* **YEAR ADDED**: Analyzing when the movie is added to the platform



* **DURATION OF THE MOVIES:** Easily understood by the topic to see what duration the movies hold in our dataset

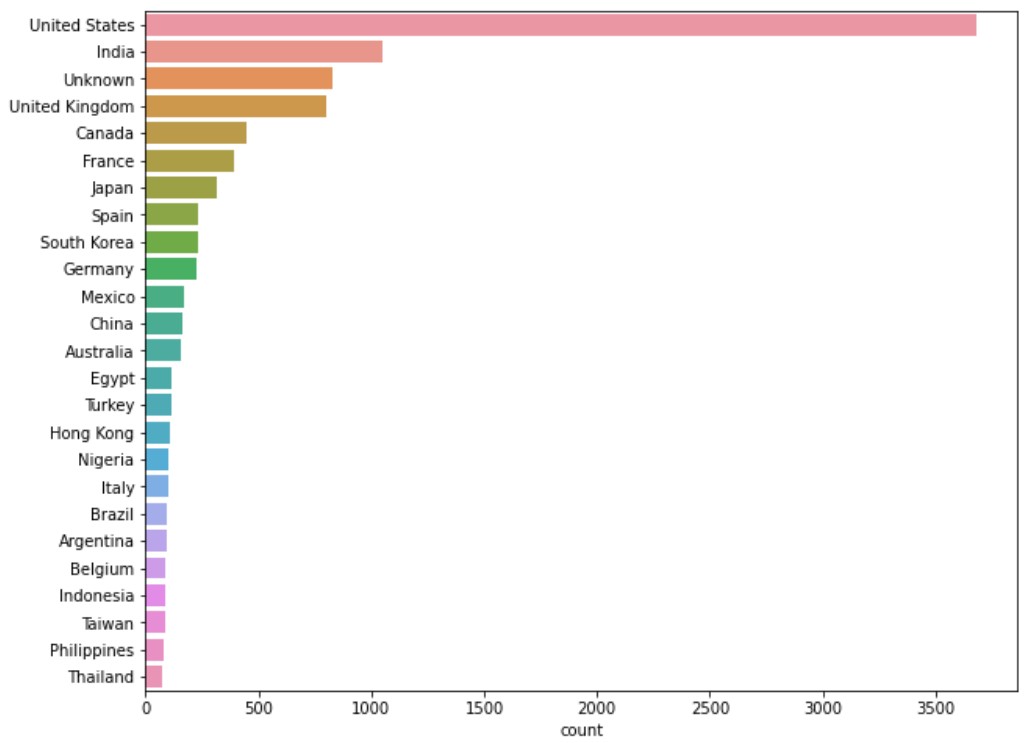


* **NO. OF SHOWS:** As for the movies we can classify shows into number of seasons



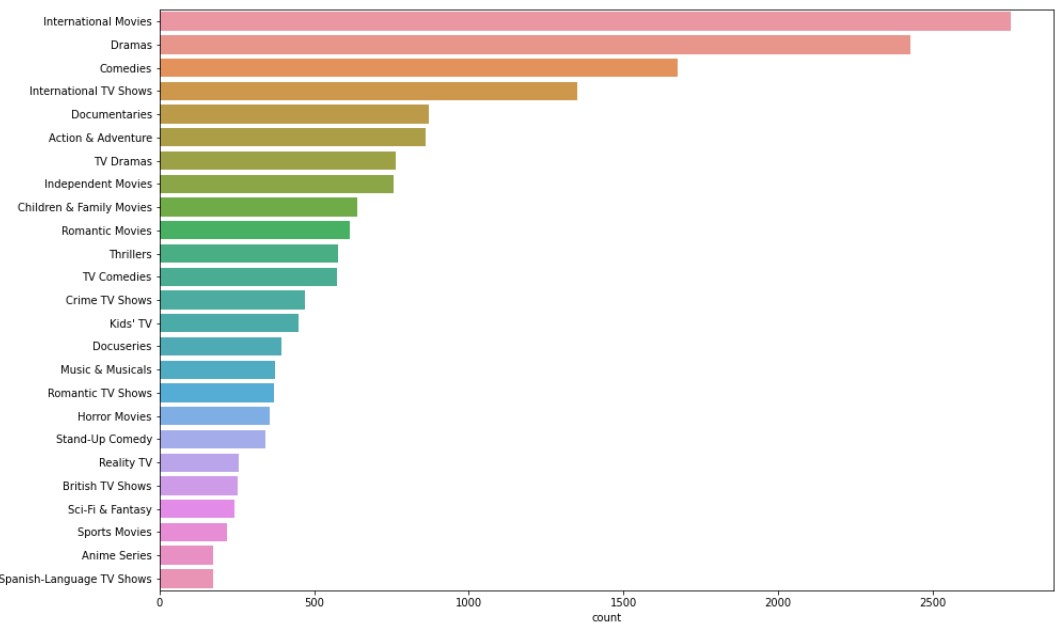
* **COUNTRIES AND CONTENT**- Next is exploring the countries by the amount of the produces content of Netflix. We need to separate all countries within a film before analyzing it, then removing titles with

no countries available.

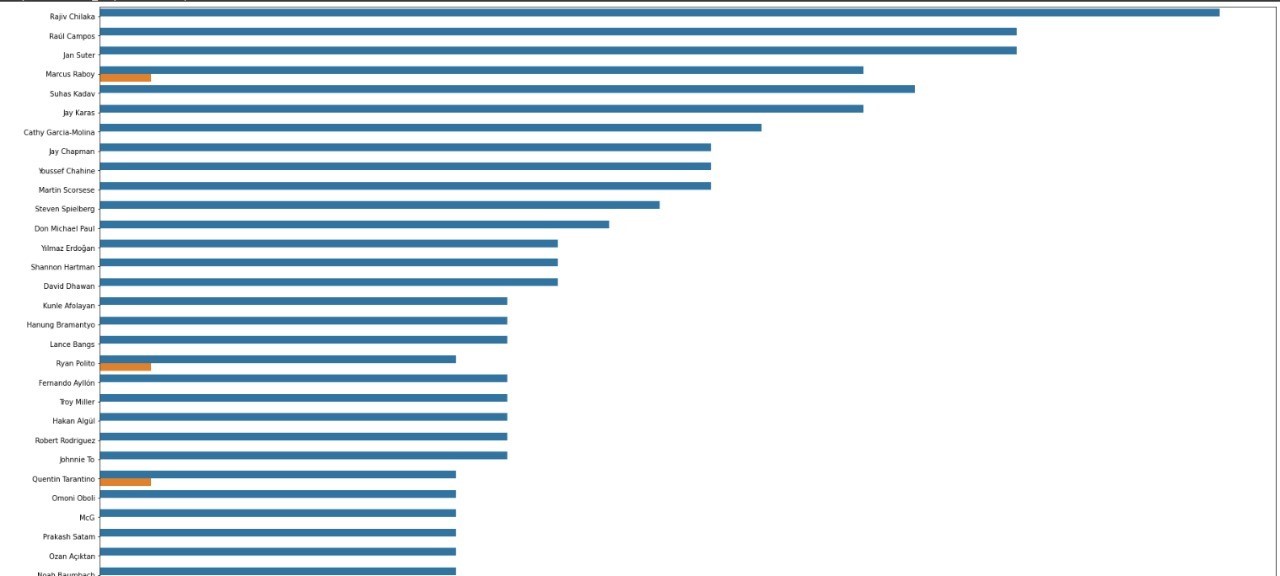


* **CATEGORIZATION:** Now as we can see the movies and shows can be classified into crime documentaries or sitcoms and similiar so we

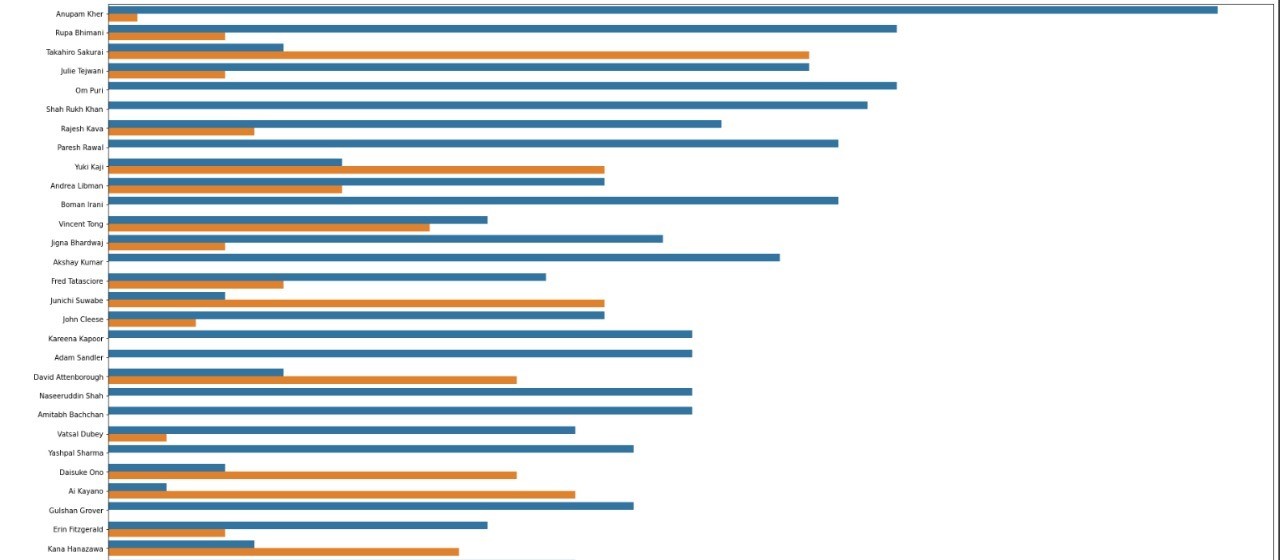
can plot the graph and see the most popular category



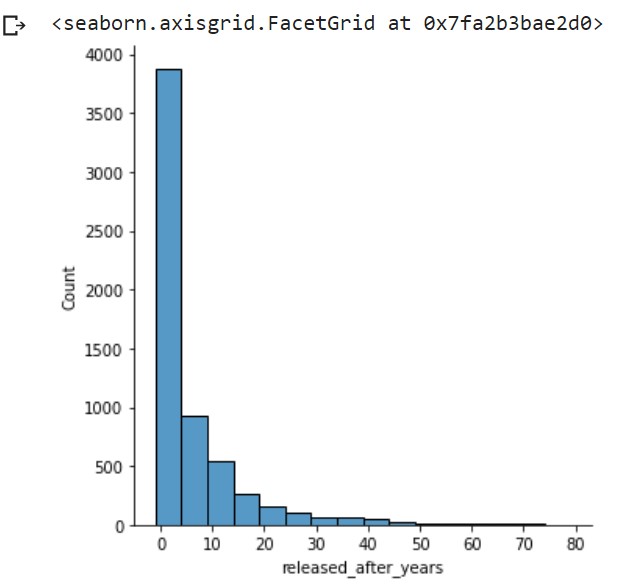
* **DIRECTOR AND CONTENT-** To know the most popular director, we can visualize it.



* **ACTORS AND CONTENT:** Like director we can do the similar for Actor



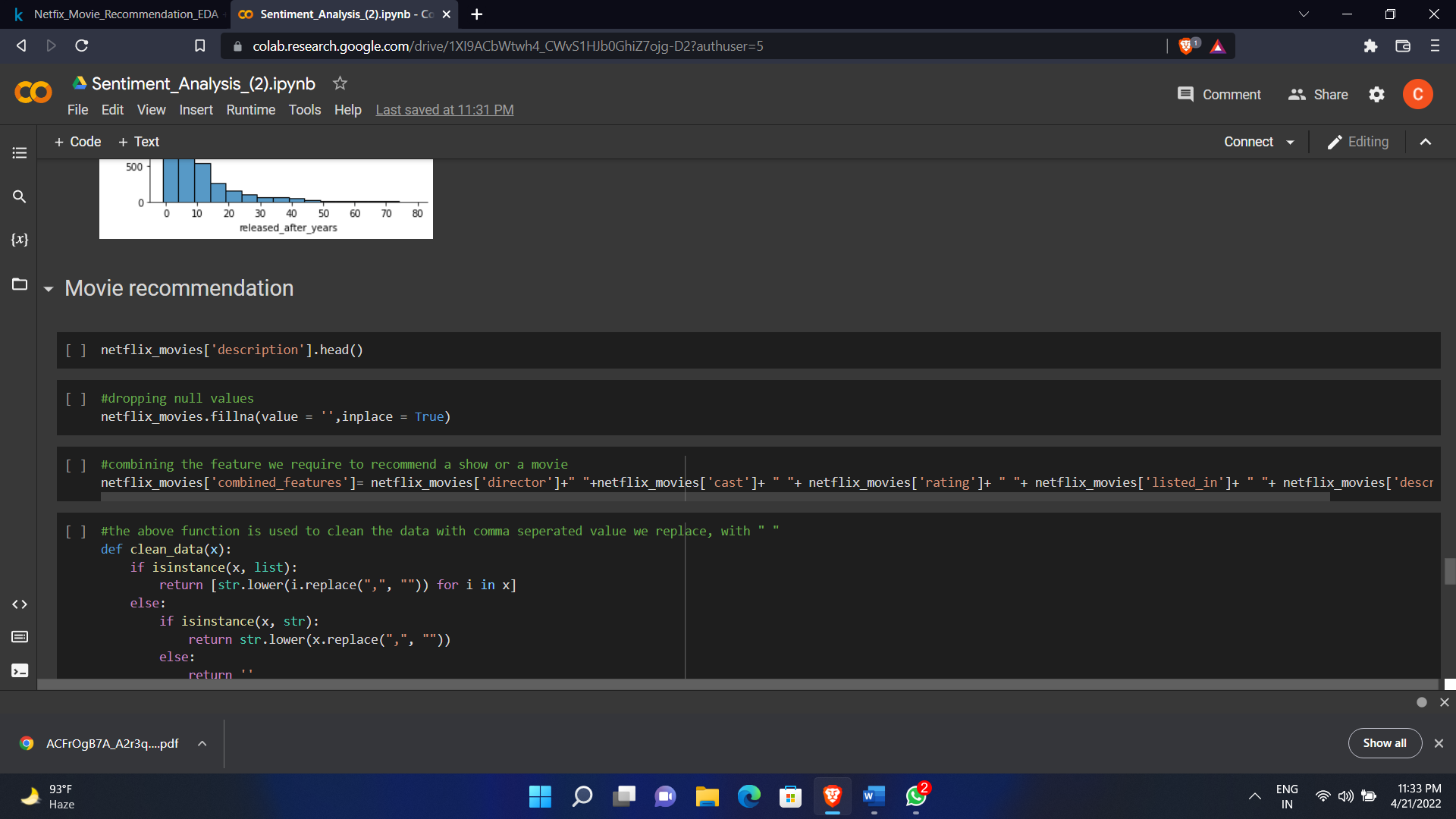
* **RELEASE AND CREATION-**  Finally we visualize after how many years our OTT has added to their platform after their release



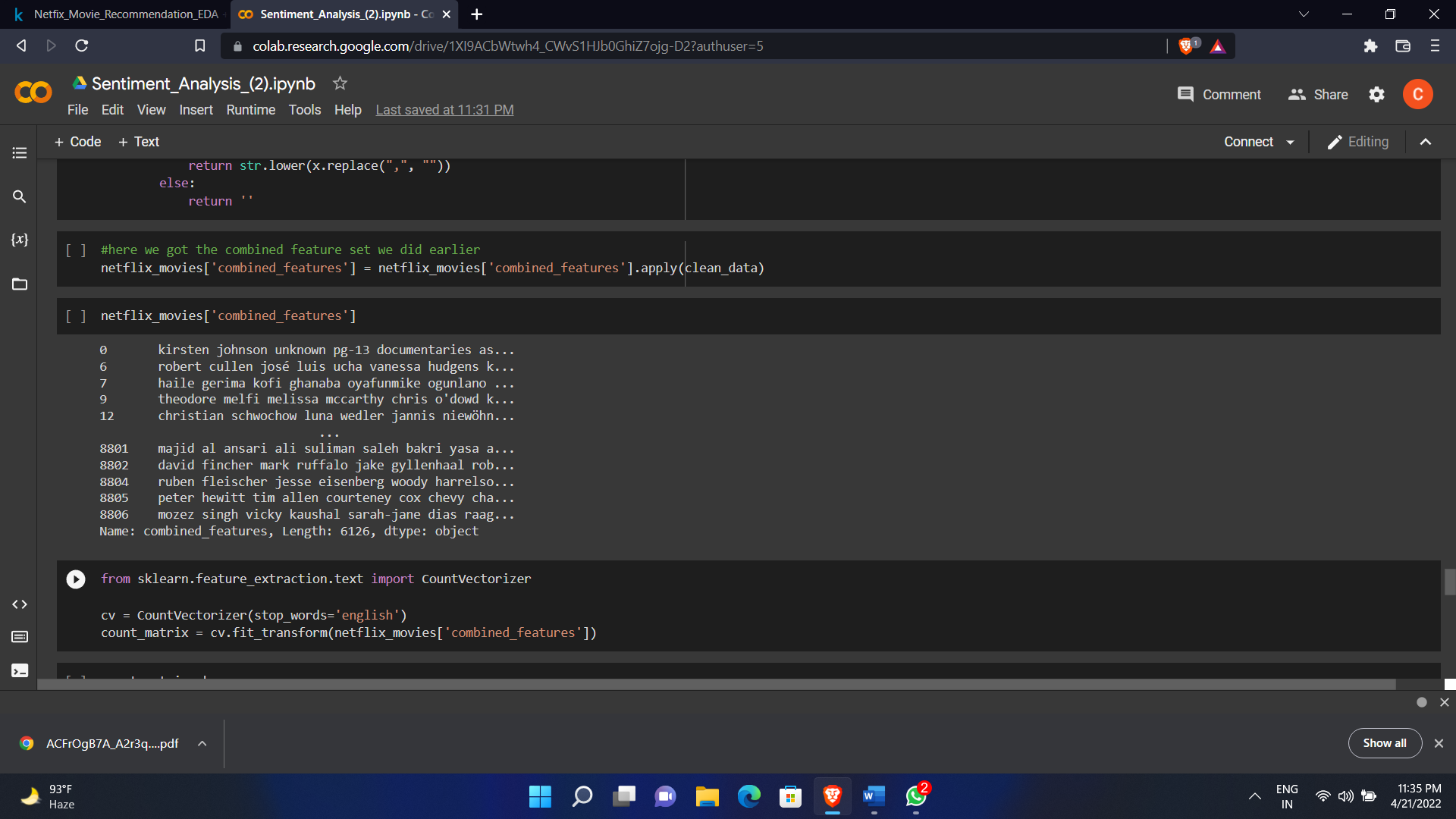
Next, we will explore the amount of content Netflix has added throughout the previous years. Since we are interested in when Netflix added the title onto their platform, we will add a “year\_added” column to show the date from the “date\_added” columns.

**Movie Recommendation System:** Until now we have explored our dataset and visualize important things in our Netflix case study. Now the next part of the segment is to recommend user a movie based on his/her interest.

The first part of our segment will be to once again verify shape and contents in our dataset. We then proceed to clean our dataset by replacing the NaN or the null values with a “ ” or simply with a blank. Next step will be to create a feature set which will contain all the necessary contents we need to predict a movie according to the user input. The new combiner feature set will be called as “combined\_features”.



To get all the features as a single entity we will create a function which will clean the data and combine our features. The following will be our new combined feature set.



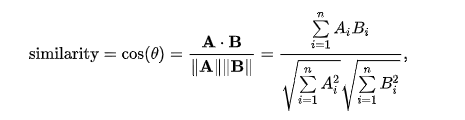
Machines cannot understand characters and words. So, when dealing with text data we need to represent it in numbers to be understood by the machine. Countvectorizer is a method to convert text to numerical data. It makes it easy for text data to be used directly in machine learning and deep learning models such as text classification. Therefore, with the help of countvectorizer we have assigned a numerical value to each movie title.

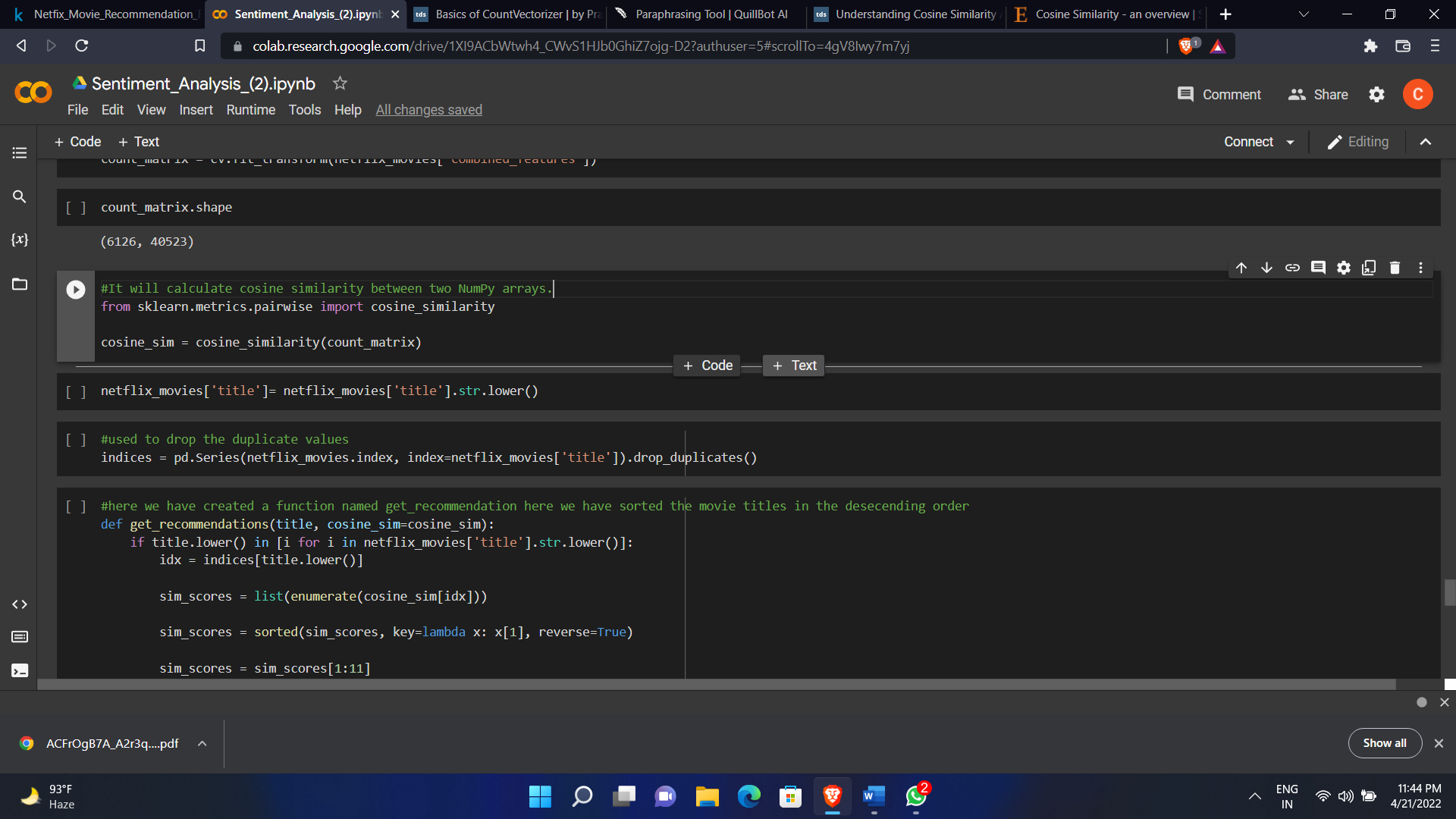
Now to recommend a movie we first need to find out the movies which are similar to each other be it with respect to genre watch time title etc. Thus, to do this we will use Cosine similarity

**Cosine Similarity is a measurement that quantifies the similarity between two or more vectors**. The cosine similarity is the cosine of the angle between vectors. The vectors are typically non-zero and are within an inner product space.

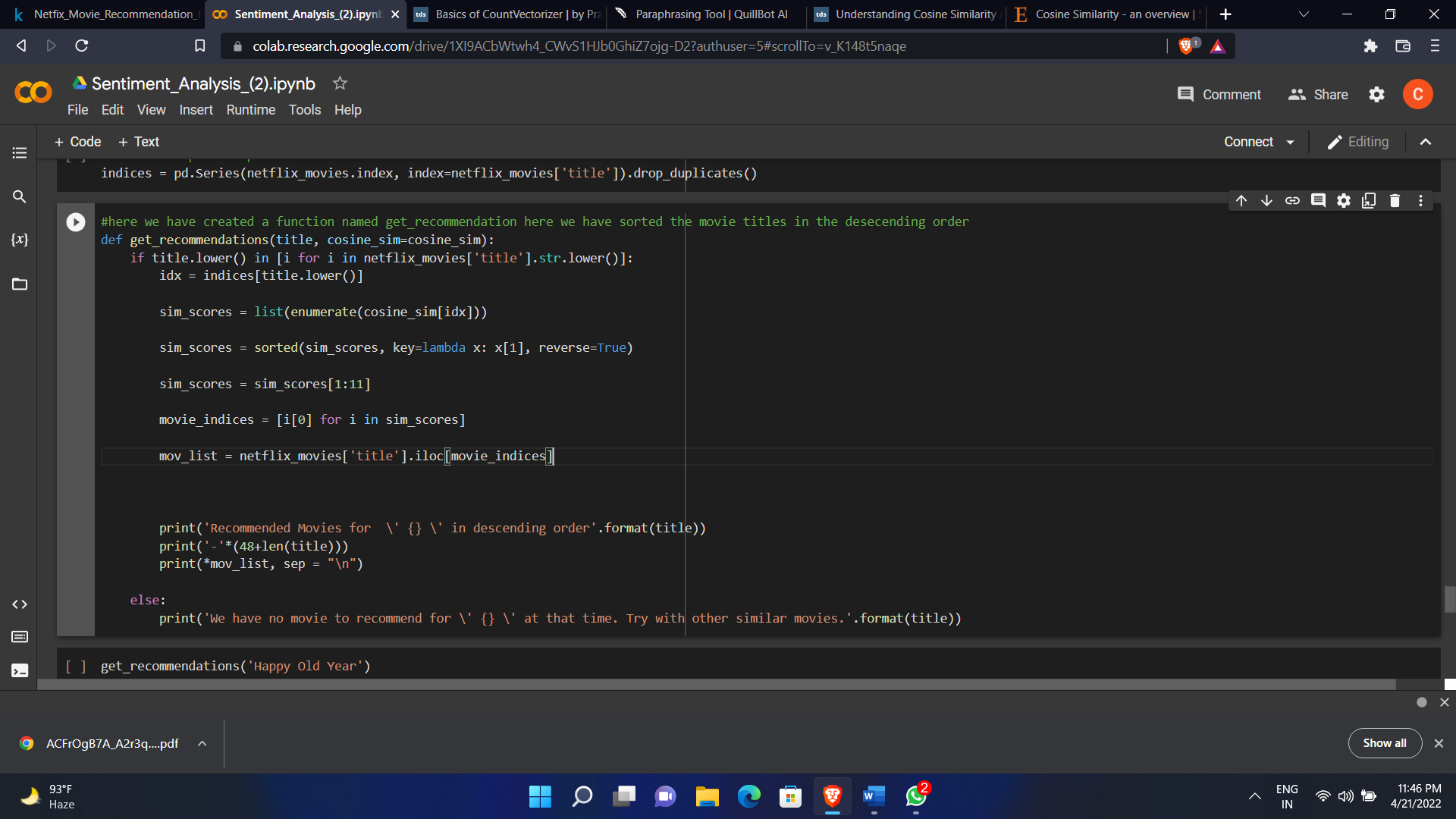
The cosine similarity is described mathematically as the division between the dot product of vectors and the product of the euclidean norms or magnitude of each vector.

Below is the formula for the cosine similarity:

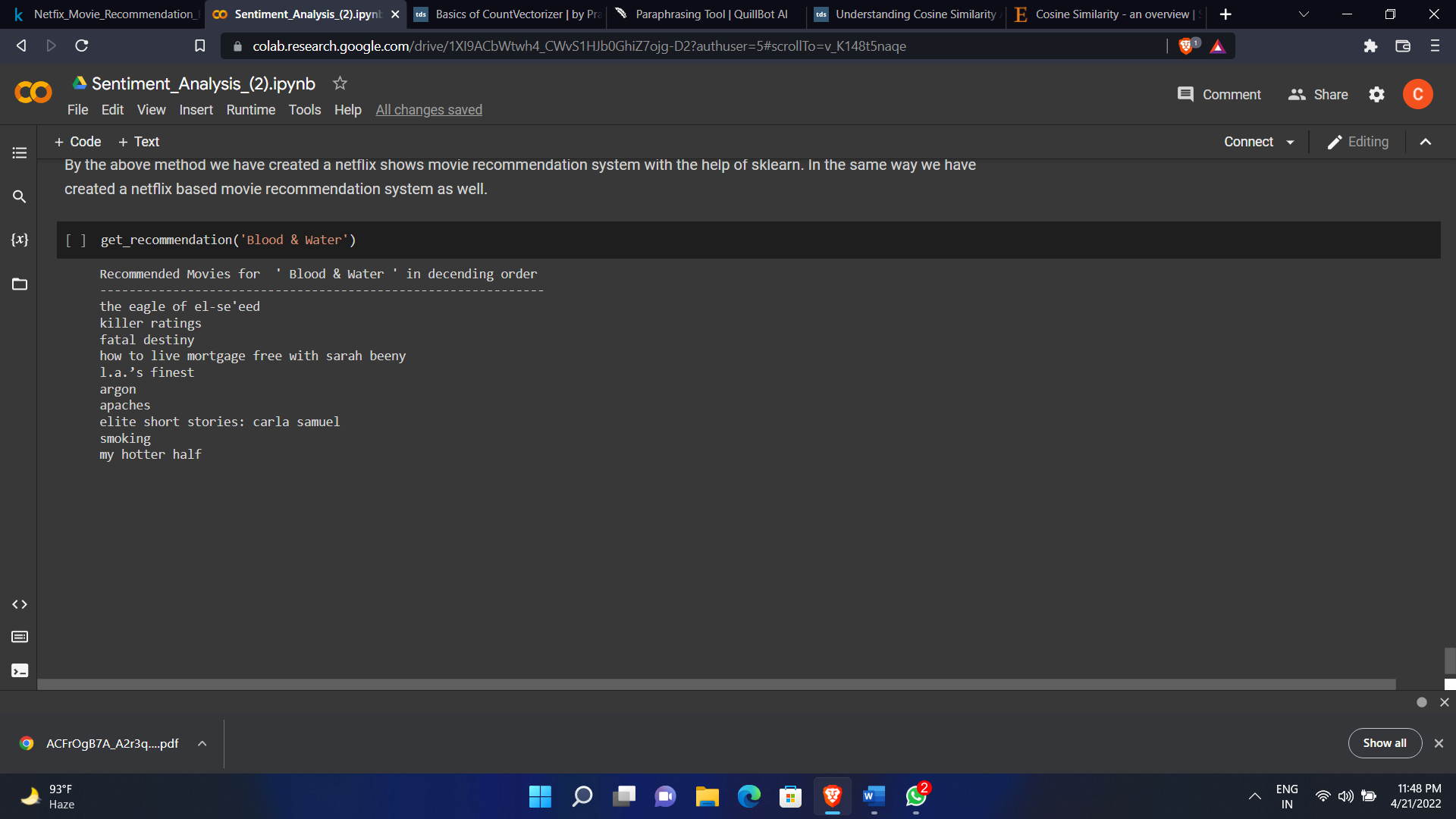
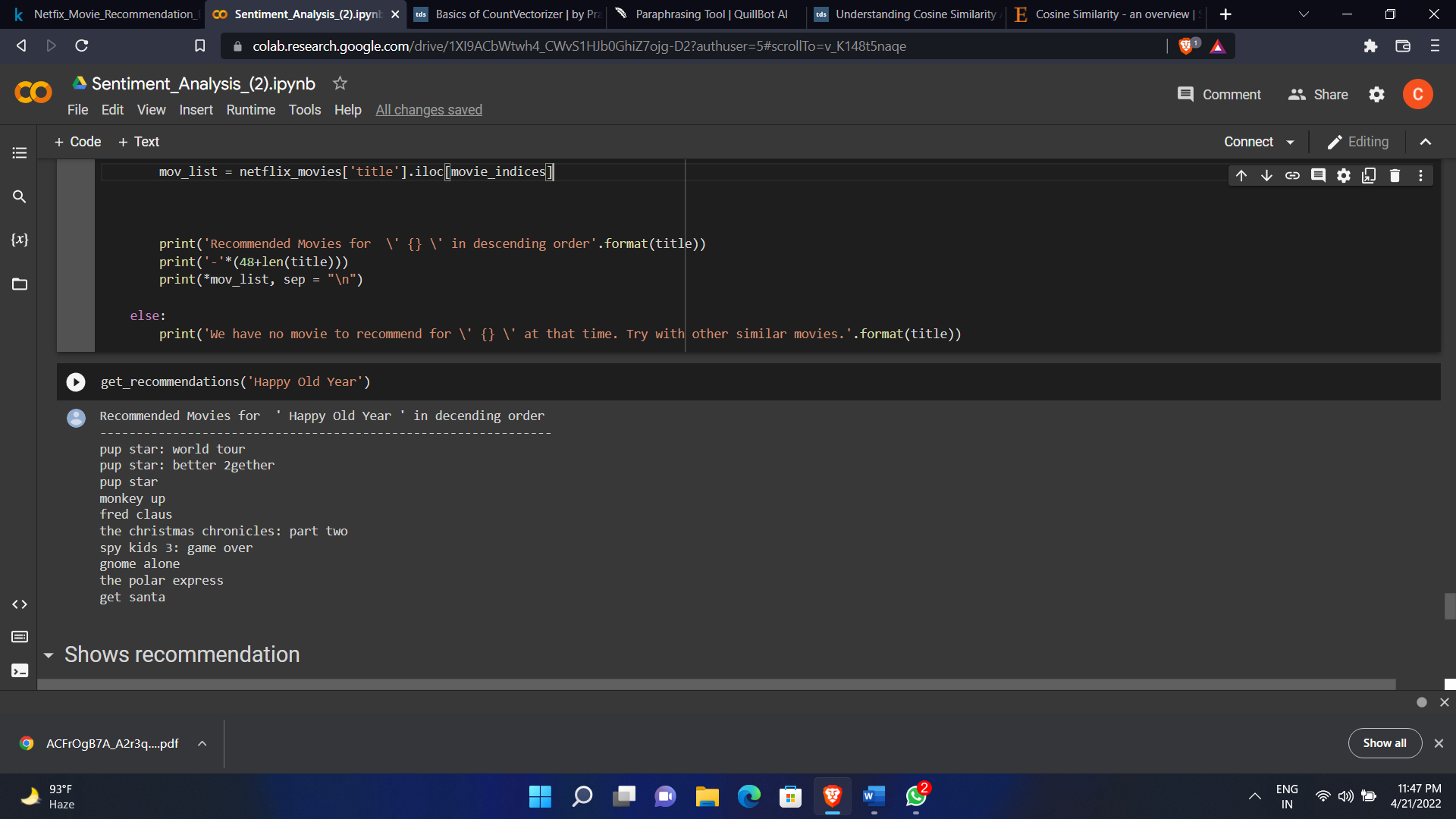




Then finally with the help of a function called “get\_recommendations” we have modified our combined feature set into descending order of their respective index.



We have used the same method for Netflix tv shows recommendation as well.

**Experiments/Results/Discussion:** We have explored the amount of content Netflix has added throughout the previous years and how useful the Case Study is going to be for the Data Analyst to work with their improvement of Netflix. The following are the outputs of our movies and tv shows recommendation.

**Model summary:** In this project we'll take a look at some very important models of Netflix data to understand what’s best for their business. Some of the most important tasks that we can analyze from Netflix data are:

* 1. understand what content is available
  2. understand the similarities between the content
  3. understand the network between actors and directors
  4. what exactly Netflix is focusing on
  5. and sentiment analysis of content available on Netflix.
  6. Recommended what movies / shows an user can watch on the basis of his/her input.

**Conclusion/Future Work:** We have drawn many interesting

inferences from the dataset Netflix titles; here’s a summary of the few of them:

The most content type on Netflix is movies,

The popular streaming platform started gaining traction after 2014. Since then, the amount of content added has been increasing significantly, The country by the amount of the produces content is the United States, The most popular director on Netflix, with the most titles, is Jan Suter.

International Movies is a genre that is mostly in Netflix,

We can also use our model to find the good recommendations of the movie and tv shows

**Contributions:**

**SWARALI JAROLE(C-19):** Wrote the code and made the whole model **VAISHNAVI DUBEY(C-27):** Researched about the dataset and brought out following conclusions

**ANUSHRI RATHI(C-53):** Constructed and properly maintained the

timeline of the project and worked on the report