cct College Dublin
Computing • IT • Business

Strategic Thinking (HDip Data Analytics - Sept 2020 cohort)

A Study of IMDb and their review scores. Can we predict movie scores from their description or title alone? What impact can these predictions have on streaming sites such as Amazon Prime Video and Netflix?

Lurdes Teixeira Power Sba 19131

Jessica Cullen Sba 20157

Group ID DAP3

Table of Contents

1.0 -Problem Solution Report

1.1 -Table of Tasks

We first created a table with tasks for each of the two team members, so we knew what everyone was doing. We found this beneficial in phase 1, as it reduced overlap. We looked at our skills we each had and decided on which role suited us best, we felt from phase 1,2 and 3 we learned which roles we felt we were best at and suited our skills

Team Member	Roles	Responsibilities
Lurdes	Project manager	Liaise with Lecturers over any concerns. Managed and constructed the documents. Researched new datasets to help answer questions for the project. Bought datasets and questions to a team meeting to discuss. Look at various modelling techniques and weigh different models to find the best model to use to get the best results to answer the questions. Worked on the Jupyter code for the project for the modeling
Jessica	Communication Manager	Created meetings and kept basecamp updated. Created Gantt Chart and extracted the stages of the CRISP-DM. Looked at ways to improve the project from phase 1 and 2. Looked at best practices for phase 3 such as holding regular meetings and using basecamp more to update lecturers. Staying committed and focusing on team effort. Worked on the code for modeling.

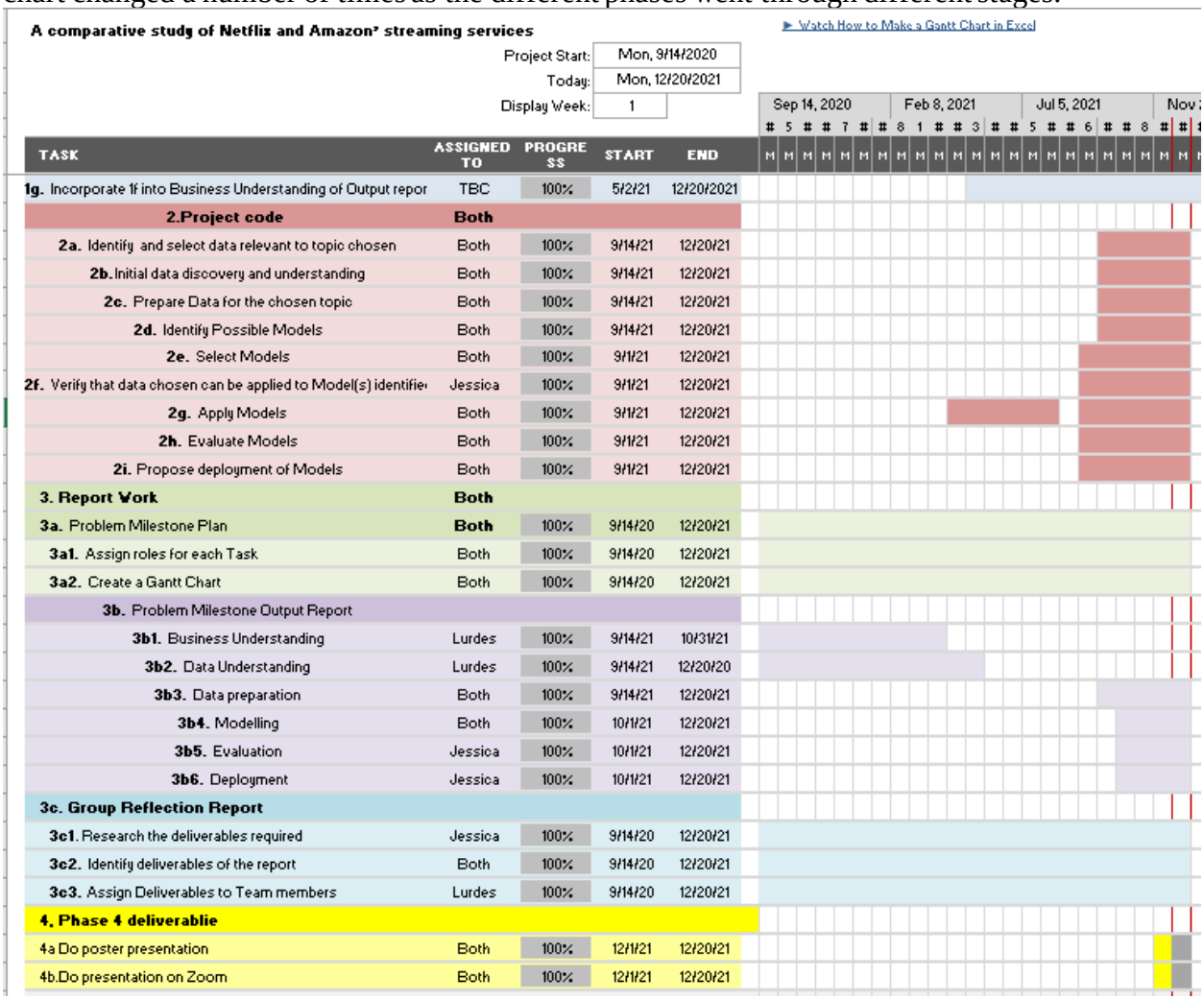
(Figure 1-Table of Tasks)

1.2-Gantt Chart

According to proof hub '*A Gantt chart is a type of bar chart that illustrates a project schedule and shows the dependency relationships between activities and current schedule status."*

A Gantt chart is useful for managers, as they need to know that these charts make the planning process easier. Since they are simple to create, use, and keep track of, they prove to be of great help for planning a product.

We used a Gantt chart as our project management tool. It was used to assist in the planning and scheduling of tasks to be done to complete this phase. The Gantt chart helped us view individual tasks and how long they would take. It helped to plan the duration of the project, so we were able to complete the project by the due date. It is very important in projects to stay on schedule. Since we are coming to the end of the project, we had to look at tasks we had completed we felt wasn't to a high standard to review as improve also.
The Gantt chart is also great to show other people at a quick glance how we as a team on progressing this was particularly beneficial during our sprint meeting. During the phases this Gantt chart changed a number of times as the different phases went through different stages.



A comparative study of Netflix and Amazon' streaming services

▶ Watch How to Make a Gantt Chart in Excel

Project Start: Mon, 9/14/2020
Today: Mon, 12/20/2021
Display Week: 1

Sep 14, 2020 | Feb 8, 2021 | Jul 5, 2021 | Nov

TASK	ASSIGNED TO	PROGRESS	START	END
1g. Incorporate 1f into Business Understanding of Output repor	TBC	100%	5/2/21	12/20/2021
2.Project code	**Both**			
2a. Identify and select data relevant to topic chosen	Both	100%	9/14/21	12/20/21
2b. Initial data discovery and understanding	Both	100%	9/14/21	12/20/21
2c. Prepare Data for the chosen topic	Both	100%	9/14/21	12/20/21
2d. Identify Possible Models	Both	100%	9/14/21	12/20/21
2e. Select Models	Both	100%	9/1/21	12/20/21
2f. Verify that data chosen can be applied to Model(s) identifie	Jessica	100%	9/1/21	12/20/21
2g. Apply Models	Both	100%	9/1/21	12/20/21
2h. Evaluate Models	Both	100%	9/1/21	12/20/21
2i. Propose deployment of Models	Both	100%	9/1/21	12/20/21
3. Report Work	**Both**			
3a. Problem Milestone Plan	**Both**	100%	9/14/20	12/20/21
3a1. Assign roles for each Task	Both	100%	9/14/20	12/20/21
3a2. Create a Gantt Chart	Both	100%	9/14/20	12/20/21
3b. Problem Milestone Output Report				
3b1. Business Understanding	Lurdes	100%	9/14/21	10/31/21
3b2. Data Understanding	Lurdes	100%	9/14/21	12/20/20
3b3. Data preparation	Both	100%	9/14/21	12/20/21
3b4. Modelling	Both	100%	10/1/21	12/20/21
3b5. Evaluation	Jessica	100%	10/1/21	12/20/21
3b6. Deployment	Jessica	100%	10/1/21	12/20/21
3c. Group Reflection Report				
3c1. Research the deliverables required	Jessica	100%	9/14/20	12/20/21
3c2. Identify deliverables of the report	Both	100%	9/14/20	12/20/21
3c3. Assign Deliverables to Team members	Lurdes	100%	9/14/20	12/20/21
4. Phase 4 deliverablie				
4a Do poster presentation	Both	100%	12/1/21	12/20/21
4b.Do presentation on Zoom	Both	100%	12/1/21	12/20/21

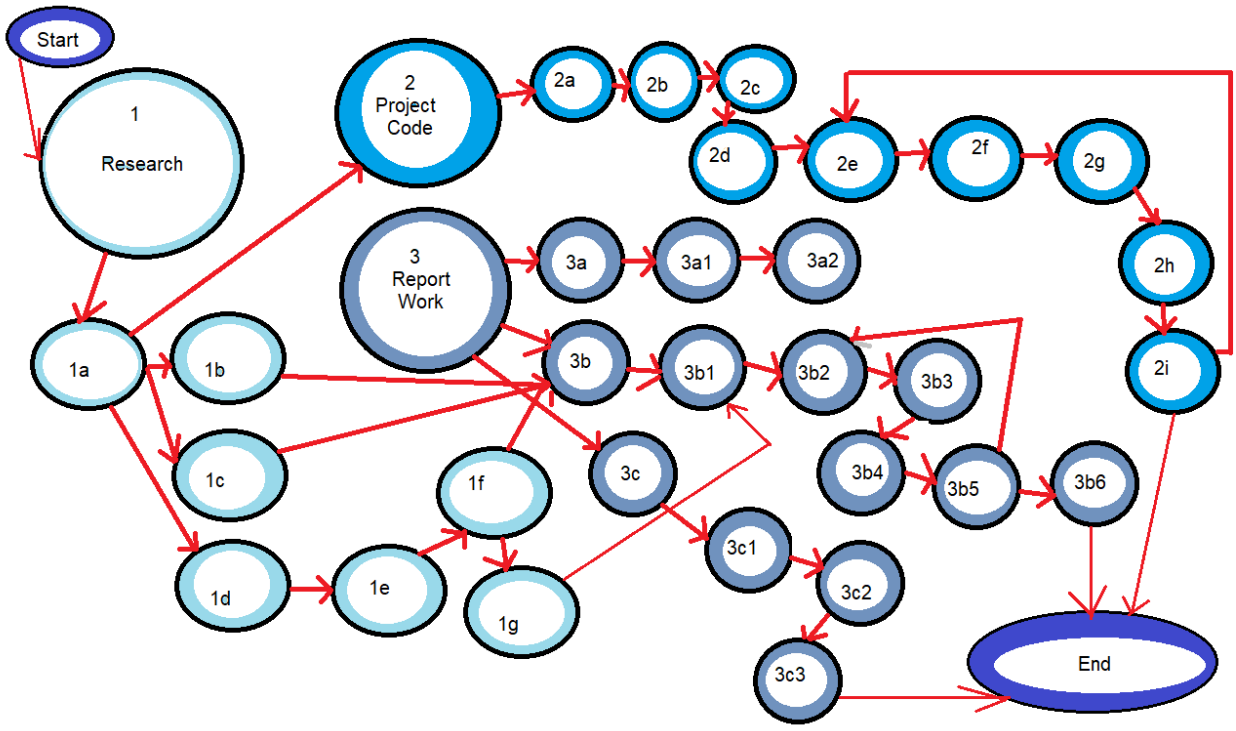
(Figure 2-Gantt Chart)

1.3 -The Program Evaluation Review Technique (PERT) chart

“A PERT chart is a project management tool that provides a graphical representation of a project's timeline” (Carol M. Kopp, Investopedia.com).
This chart can be beneficial, as a complement to a Gantt chart, as a graphical method to view the tasks that need to be performed. This chart also has the benefit of showing the various relationships between the tasks. This type of chart shows which tasks are truly independent, and which one depends on others being done first. So more than a timeline, this chart becomes a task relationship coordinator. We can clearly see which tasks are of most importance and why they need to be done first, as well as what the impact on other unfinished tasks is if some of those tasks are incomplete or overlooked.
This chart cannot however replace the Gantt chart as this type of chart can sometimes be confusing when used on its own. The chart is best used as a complementary method of the Gantt to manage and keep track of a project.

For our tasks we have decided to create and incorporate a Pert chart because we thought that the relationships between certain tasks of the project could affect the timeline, delivery, and completion of the overall project; and therefore, requires careful consideration and organization. We felt that the Gantt chart alone did not reflect those relationships between tasks or their impact on the overall project.

The chart below represents phase 3 and is relatively simple. For clarity purposes the tasks have been numbered the same as in the Gantt chart and does not show any timelines. It however shows the relationships between the tasks and clarifies the order in which things need to be done.



(Figure 3-PERT)

2.0 -Problem Solution Report

2.1-Introduction and Background

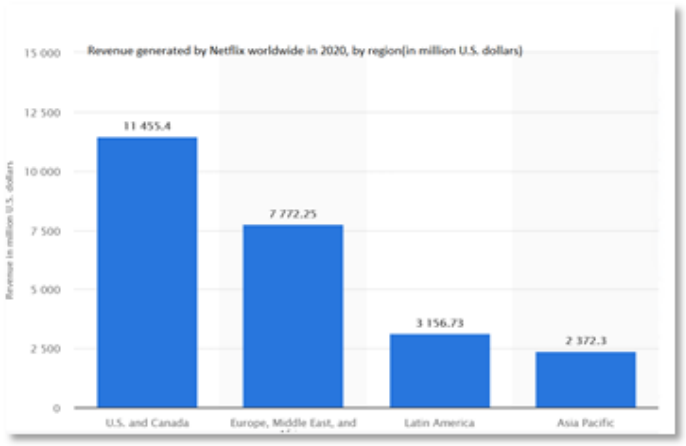
More than a simple scoring website, IMDB (originally known as Internet Movie Database) is a database with information related to films, televisions show, video games and content online. It provides information on them such as cast, plot, rating, and crew. It is a subsidiary of the parent company Amazon. It was founded in 1990 by Col Needham, now CEO of IMDB. Amazon bought the company in 1998. The company grew over the years to add more features about the film and television shows such as release date and box-office gross earnings.

Netflix Inc. was launched in 1997 by Reed Hastings and Marc Randolph. Netflix first started operating by offering a rental service for DVDs. The system was simple. The customer would buy a subscription and the subscription allowed the customer to rent as many movies as they wished for a calendar month. The movies were sent out and returned by post. In 2007, the business began offering an alternative service, unlimited movie streaming on demand.

Amazon Prime Video is a streaming subscription service, owned by Amazon.Inc and was established in September 2006, originally named as Amazon Unbox video. Much like Netflix, Amazon Prime Video offers the possibility to stream unlimited content for movies and tv shows all over the world, with some exceptions (notably China); while the majority of their subscribers are in the US, the statistics show that the users in Europe and Asia-Pacific regions are on the increase.

2.2 -Phase 1: Business Understanding

A good way to introduce any business is to look at their numbers. And the numbers here speak by themselves. The graphs below were borrowed from Statista which is a website dedicated to business statistics.

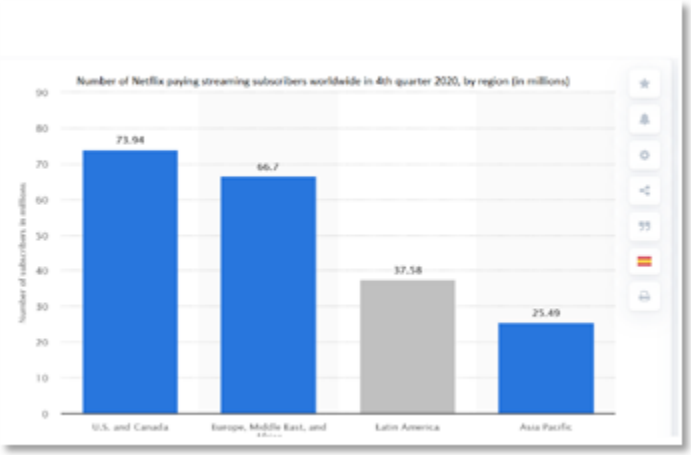


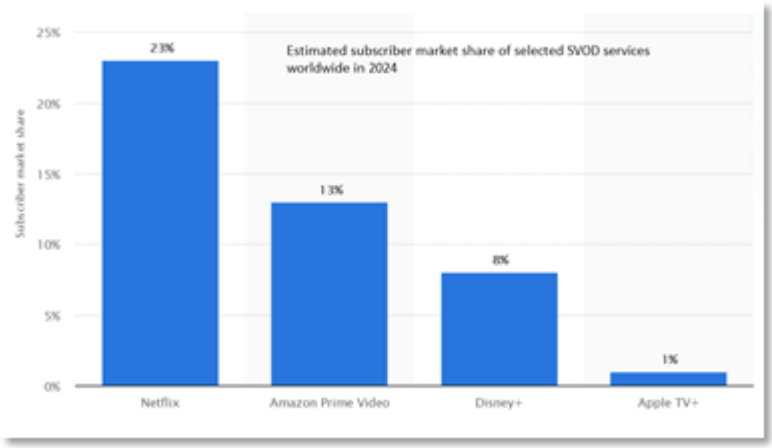
According to Statista.com Netflix's "revenue has grown from ...1.36 billion to around 15.8 billion in ... ten years"

(Graph 1-Revenue generated by Netflix)

"The number of Netflix subscribers has followed a similar trend, growing from less than 22 million in 2011 to nearly 150 million in 2019 (Statista.com)."

(Graph 2-Number of people paying for Netflix)





In 2020 almost "50% of adults in the USA had a Netflix's subscription", showing the market share that the entertainment giant possesses, and going further by estimating that by 2024, Netflix will lead the competition by about 10% (Statista.com)

(Graph 3-Subscriber market Share in 2024)

We now will take a look back at the origins of Netflix as a business, an overview of their business models and a deeper overview of the overall industry, as well as, documenting who their main rivals and threats are.

Netflix Inc. was launched in 1997 by Reed Hastings and Marc Randolph. The current headquarters of the company are now set up in Los Gatos -California, in the United States of America. Netflix first started operating by offering a rental service for DVDs. The system was simple. The customer would buy a subscription and the subscription allowed the customer to rent as many movies as they wished for a calendar month. The movies were sent out and returned by post. In 2007, the business began offering an alternative service, unlimited movie streaming on demand. A partnership was also created that enabled customers to watch Netflix movies via game consoles and Blu-Ray players. A streaming only plan was progressively introduced around the world, starting with the USA in 2010, Canada in 2011, UK and Ireland in 2012, reaching up to 190 countries by 2016.2013 saw the start of Netflix's own original content being broadcasted, progressively growing to over 1,000 original titles by 2018.By 2018, Netflix

had over 130 million subscribers for its streaming services while the DVDs rental was still there and doing well but not as profitable as the streaming side.

Wilmot Reed Hastings Jr went to university and studied for a master's degree in computer science. He became a computer developer and founded Pure Software in 1991, which he later sold for a profit. In 1997 he created Netflix with a partner Marc Randolph (Who then left in 2004). Reed Hasting becoming CEO later in 1997 and remains as CEO till this day. However, he now shares the CEO position with Ted Sarandos since 2020.

(Graph 4 to right and Graph 5 below. Pictures form Netflix)

The Mission and Vision Statements of the business have evolved over the years but have a common denominator: Entertainment. The current Mission Statement is to "Entertain the World", while the Vision Statement is "To continue being one of the leading firms of the internet entertainment era.", which aligns to the corporate ethos and culture.

Netflix's business plan and strategies involve the Platform and Pipeline business models. Netflix uses their site as a platform for other producers' content to reach audiences. The pipeline model applies to the side of the business that deals with the production of their own content, thus enabling them to rely less on production houses for their content. The Pipeline strategy provides control for the company while the Platform model provides a strategy to support their growth. Netflix's unlimited subscription strategy helps attract and retain customers, as well as providing a constant income source. Netflix's intensive growth strategies include the penetration of the international markets. The business applies cost leadership and differentiation

in order to stay competitive. Through minimized costs and selling prices, and a product that is a little bit different than its competitors (mainly via their own content productions), Netflix manages to retain the leadership in an industry where new entrants keep arriving and competition is close.

According to Investopedia.com, "as of March 2020, Netflix remains the largest subscription service with roughly 183 million paid subscribers worldwide". As the market leader, the threat of new entrants has a relatively low to medium impact. With its very broad customer base, and a very

high volume of subscriptions, the bargaining power of the customers is low. If one customer cancels a membership, this has little or no impact on the overall business. However, with the increase in popularity of on demand entertainment, newcomers are always arriving to the subscription market. Newcomers to the market may have a different pricing structure and a different content, or perhaps their own content, making Netflix's subscriptions plans and entertainment content possibly less desirable to potential customers.

Netflix has, now (in 2021), more than ever, a real potential threat of being substituted for another service that its customers could find more suitable. Amongst its direct competition, the streaming and on demand choices are vast and on the increase. Its main competitor Amazon Prime's price, whose price plan can sometimes be cheaper than Netflix's, is their biggest threat. Apple Tv is another competitor whose services tend to appeal to Apple's vast community of users. More recently Disney + and Discovery +, seem to have managed to enter the market successfully, managing successful partnerships with Sky TV in Ireland. There are now also specialized streaming apps (like Fit for example), that dedicate themselves to one particular customer, whose main focus is exercise, appealing to one very special type of streaming content user.

Traditional terrestrial channels like RTE are free (provided you have a Tv License), but tend to appeal to a different customer base, mainly traditional users. Cable and Satellite services like Virgin media and Sky TV also remain. These services have evolved over the years and are becoming more and more diversified and incorporate apps and access to streaming from their own platforms. By staying current with innovation and new technologies being used, other providers are bringing their services to today's customer's requirements. Most offer the possibility to have an all-in-one solution for your media, with a login screen within their platform for other streaming apps such as Netflix, Amazon Prime, or Disney+. This can however be a double-edged sword for Netflix, as more customers may access their services, but customers may discover other services too and may decide they no longer need Netflix. And in the larger scope of things, customers can also find entertainment via other methods such as cinemas, YouTube and via the use of their own DVD players at home (although the last one becoming less popular). Netflix can balance this with its innovation into their own shows however (example Orange is the New Black, Bridgeton, Queen's Gambit), attract specific customers, with specific interests. Those shows have proven to be very popular, which is good for Netflix.

Although they are not tied to any particular supplier, Netflix does depend on getting the content from the various producers, and according to Investopedia.com, the cost for content purchasing alone was expected to be in the region of 15 billion dollars for 2019. One company may decide to pull its content out of Netflix. The BBC did this with their Line of Duty series -causing uproar in social media amongst disgruntled Netflix customers- where, in anticipation of a new season of the series starting, the BBC decided to pull the rest of the seasons off Netflix in order to show the complete series on their channel and BBC player)

Netflix may also be at the mercy of certain companies who have interests in other streaming services. Walt Disney is more likely to show its content on the increasingly popular platform Disney + for example and less likely to provide content for Netflix.

Netflix has several direct competitors in Ireland, with some well-established ones like Amazon Prime and Apple TV, and newer entrants to the market Disney+ and Discovery + quickly gaining customers. Sky Tv for example has arrangements with several suppliers with Sky Q customers getting free trials of platforms. Discovery plus is one of those partnerships that Sky Q has with Sky TV offering free access to discovery plus for a whole year, for Sky Q subscribers. Sky Tv has

also realized that they needed to reward their existing customers by creating a specialized reward app. This app allows for free upgrades, free box office content, free access to certain sporting events and so on.

Netflix 's main competitor remains Amazon Prime Video.

Amazon Prime Video is a streaming subscription service, owned by Amazon.Inc and was established in September 2006, originally named as Amazon Unbox video, then as Amazon Video on Demand in 2008, and later rebranded as Amazon Instant Video in 2011.

According to Investopedia.com, "as of the fourth quarter 2019, Amazon prime Video had about 150 million subscribers - a number that's been growing at a fast pace over the past two years." According to Market.us, 56 million of those subscribers are in the US in 2020.

Amazon Prime also now has their own exclusive shows (such as "Fleabag" and "Homecoming" and have recently won exclusivity on certain movies such as "Coming to America 2" possibly steering customers away from Netflix. Their services are now available" through the Prime Video App, Smart TVs, game consoles, streaming media players and Amazon's Fire TV "(Investopedia.com), making them more accessible than ever.

Much like Netflix, Amazon Prime Video offers the possibility to stream unlimited content for movies and tv shows all over the world, with some exceptions (notably China)., while the majority of their subscribers are in the US, the statistics show that the users in Europe and Asia-Pacific regions are on the increase.

On their own, Amazon's pricing is competitive with Netflix, However, Amazon's Prime Video pricing plans are variable depending on whether you have Prime membership, which is different to Netflix's offerings. Those all-in-one Prime plans represent good value for money when bought as Amazon Prime membership, in conjunction with shipping benefits to Amazon 's purchases. Although a good bonus, most Amazon Prime customers originally buy Amazon Prime in order to avail of the free shipping on their purchases. On its own, Amazon Prime Video offers a free 7-day trial before the user has to sign up. Subscribers to Amazon Prime are expected to be in the region of 164 million for 2020. Market. Us estimates that "64% of households" have an Amazon Prime subscription and 25% of the households who have a Prime subscription have used or use the Amazon Prime Video service during the last 3 years.

Similarly, to Netflix, Amazon Prime Video, obtain the content they stream via various license agreements. It is estimated by Market.us that Amazon.inc would have spent about 5 billion dollars on content for Amazon Prime Video in 2018. In an effort to stay current and on trend, Amazon Prime video will have seen the introduction of a variety of new series in 2020 such as season 1 for Ice Road truckers, les Misérables, or Riviera. More movie content is also constantly added with classics such as "Raiman", "Top Gun", or Spiderman 3 entering the service.

One shouldn't also forget that Amazon Prime Video, being a sister company to Amazon, could be heavily subsidized or pumped with investment by the American giant.

A recent article in the Irish Times by journalist Laura Slattery, published on April 19th, 2021, mentions that the budget alone of Amazon Prime Video latest series to be developed, could cost as much as 388 million euro for the first series of the Lord of the Rings Prequel alone. The same article refers to a partnership with New Zealand, with the country subsidizing up to a massive 25%

of the cost of the series, for the privilege to be chosen as location of choice for the filming of the much-anticipated show.

(Graph 6- From Eir Website)

There are a variety of partnerships in place for Amazon Prime Video, notably in Ireland with Eir offering up to 12 months free membership for the streaming platform as part of Eir' subscription cost for their new and existing subscribers

Other substantial partnerships include the one with Liberty Global, a tv and broadband provider providing free access to Amazon Prime Video to its 4 million subscribers via Virgin Media customers in the UK.

The Irish Times article by journalist Laura Slattery, also references a letter to shareholders by Amazon founder Jeff Bezos, revealing that the current amount of Amazon Prime subscribers has now actually reached 200 million worldwide.

With Amazon Prime Video quickly catching up to their old-time rivals Netflix, one must wonder what sets them apart and what tools do both businesses possess in order to predict who has the best offerings in terms of their content. If there is a way to know what their content score is, then each company should be able to deduct from that, which movies or title works, but also, in a larger sense, which company needs to improve their variety of offerings, in order to attract potential customers.

More than a simple scoring website, IMDB (originally known as Internet Movie Database) is a database with information related to films, televisions show, video games and content online. It provides information on them such as cast, plot, rating, and crew. It is a subsidiary of the parent company Amazon. It was founded in 1990 by Col Needham. He is the founder and CEO of IMDB. Amazon bought the company in 1998. The company grew over the years to add more features about the film and television shows such as release date and box-office grosses. Over the years IMDB has even created a free service and fee-based service for industry professionals where they can post their Curriculum Vitae (CV) and productions can post jobs.

Now, how exactly can the businesses tell how the movies are rated? Well, websites such as Rotten Tomatoes or IMDb offer users the opportunity to rate movies. IMDb then applies their "magical" algorithm that produces a score, for each and every one of the titles on their website.

According to the Frequently Asked Questions section of their website, IMDb scores or "ratings" are "aggregated and summarized votes". Users on the platform can provide a vote for a movie or show from 1 to 10, those votes are then calculated via "weighted average mean" calculation before they are published. The "weighted average mean" is different to the standard "arithmetic" mean. IMDb FAQ page continues by explaining that not all votes have the same "impact" or "weight".

The page also mentions that some unusual votes or activity can change the calculations and that in order to keep the rating "effective", IMDb does not "disclose" the exact calculations of their ratings. The website seems therefore to keep a certain secrecy over their ratings and "calculations, which is why we were implying some "magical" algorithm earlier in this report. One could go further and say that by reading the FAQ section of the IMDb website, one clearly gets a feeling that the rating algorithms are often discussed and disapproved of, meaning that the rating system certainly seems to bring a sense of controversy amongst IMDb users.

Knowing how the ratings are created and where they are stored, why would giant companies such as Netflix and Amazon Prime be interested in IMDb scores and ratings? Well, Netflix can build better algorithms for recommendations based on users' reviews or associations such as if you liked "Breaking Bad" now you may like "Better Call Saul" for example. Amazon has its own reasons for using IMDb too, such as links directing users to Prime Video titles and advertisements for Prime's services or shows. Of course, one must remember that IMDb is now partly owned by Amazon and this purchase may not be entirely coincidental…

Given the above, **the team will attempt to study IMDb and their review scores. Can we predict movie scores from their description or title alone? What impact can these predictions have on streaming sites such as Amazon Prime Video and Netflix?**

2.2.1 -Assessing the Situation

2.2.1.1 -Available Resources to the project

The team consists of 2 members. Lurdes Teixeira Power will act as project manager and main researcher for all topics of this task. Jessica Cullen will act as the communication manager, managing Basecamp WhatsApp communication with other team members and lecturers.

Both team members will use their individual laptops. The hardware available for this task is Dell-Inspiron 13 2in1 Laptop, 7359, Intel Core i5-6200U Processor, 8GB RAM, 500GB 5400 rpm Hybrid Hard Drive + 8GB Embedded Flash Cache, 13.3-inch Full HD Touchscreen. The software used to achieve this project is Jupyter notebook, Anaconda3, Python 3, Microsoft Excel, and Microsoft Word, via Google Drive. Note that for sharing purposes Google drive and Google Colab were used for parts of this project.

2.2.2.1 -Requirements, assumptions, and constraints

Both team members to be provided full read and writing access to the datasets required for completing the project
Both team members assume the data provided to be complete and correct at the time of completion of this project

The team will face a time constraint. College and submission deadlines will be a major constraint and may limit the research and code aspects of the project.

Hardware limitations and coding errors will be another constraint that may limit the team. Should the team have any hardware coding errors linked to memory usage, the team may be constrained to change the code in order to bypass the errors.

2.2.3.1 -Risks and contingencies

Some of the risks associated with the project are:

The team could have insufficient or incomplete data to work with. The datasets, although very recent, cannot represent the constant updates that both streaming platforms experience. Without the required amount of data, the algorithm used the modelling phase may not run correctly or may not predict correctly. The incompleteness of the data may mean that the predictions and classifications offered are not current enough to satisfy the task on hand.
The team may have chosen the wrong algorithm to work with in the modelling phase and changing the algorithm may take some time and therefore could delay the project's completion.

Another risk discussed above is the possible impact or constraint due to hardware and coding requirements. Should the student 's machine hardware produce errors related to RAM, the team may need to change direction in modelling, perhaps change model choice, or perhaps run the code via Google Colab in order to bypass the errors.

Some contingencies have been thought of should those events occur.
The team could add the data from another dataset and increase the amount of data available. The team could add more columns to the datasets with more current data. By manually inputting current information, although this could delay the classification and forecasting tasks.
The team could use another algorithm for classification. Although not as easy to implement, there are other possibilities in the machine learning family of algorithms that the team can use. The team could also change the split in the training of the models, should the results not be entirely as expected, or perhaps use K fold clustering in order to improve the results.

2.2.4.1 -Defining your Data Mining Goals and Project Plan

The aim of this project is to study IMDb scores / ratings and check if those scores can be predicted. The team plans to use three different datasets in order to have a variety of input data, but also in a spirit for fairness for both Netflix and Amazon Prime platforms, to see which data yields the best predictions once our algorithms and models are applied to their relevant IMDb scores.

Before each dataset can be used for modelling, the data will need some thorough cleaning. This will involve looking for duplicate data or null or missing values. If identified, the values need to

be sorted and transformed so that all of the data is usable. Once clean, the data can then be visualized and correlations checked outliers and be identified, which may need to be addressed before the modelling.

Part one of the project code will reflect IMDb scores as taken directly from the website. The plan for this part of the project is to do a sentimental analysis on the title and description of the movie in order to view and classify words into positive and negative categories. Once vectorized and encoded the data is then used to predict the IMDb rating.

Part two of the project code will concentrate On Amazon Prime data and their IMDb scores. The plan here is to use the various columns of the data to classify the IMDb score based on the column names and dataset's most relevant features. This will hopefully be achieved via a decision tree model. The tree will be based and built around the dataset's best features as "branches" or nodes of the tree.

The third and final part of the project code will focus on Netflix data and their IMDb scores for this part of the analysis, the team will try a mixed or "hybrid" approach to the modelling. This means that, all going well, we will first classify the data and use a classification model like the decision tree used in part two; and then, we will use the similar data than phase 1- like title and description (from the Netflix dataset available)- in order to do a sentimental analysis on this data. The aim of this mixed approach is to evaluate which approach and model works better, given the same starting data.

The team will use supervised algorithms for this analysis Part two and three will focus on the decision tree model, which is a model used for classification purposes. The choice behind this comes from the variety of features in the datasets; the team thinks that those features, classified with a decision tree model, can be used to predict the IMDb score. Should the decision tree method not work, one could possibly modify this technique to build a random forest model, which is similar to decision trees, but where the trees are classified as a multitude of trees hence the "forest" name.
The other modelling technique, sentimental analysis is interesting because one could easily consider it as a supervised or as an unsupervised technique. Here the issue depends on labelling, whether or not, the positive, neutral, and negative feelings were labelled and trained (making it supervised); or whether other methods of labelling previously unlabeled data are going to be used, which in this case would make this technique unsupervised. Once the words have been labelled into positive neutral and negative categories, the models can then be applied. For this we may use a Naive Bayes approach, or the SVM (Support Vector Machines) model.

2.3 -Phase 2: Data understanding

2.3.1 -Gathering and describing the data

The team will use three datasets for this analysis.

The first one is called “IMDB_movie_reviews_detaills.csv”

” And was found at the following location

:https://www.kaggle.com/iarunava/imdb-movie-reviews-dataset

However, since downloading this file earlier in phase 3 of this project, the data file seems to have been strangely deleted from the Kaggle website. We do of course have a copy of this data, and the data will be shared in a zip file for the lecturers.

It is not the biggest of our datasets in terms of rows. It has 10 columns and 1000 rows. We will refer to this dataset as “movies” during our analysis, and in the enclosed code file.

This dataset was chosen because it provides a good overview of the titles available on the IMDb website. This overall data also seems relevant to our analysis and seems quite recent. Of most interest to the analysis team is the rating column, which represents the IMDb rating score awarded to each movie out of 10.

Note that this dataset capture below was taken before any of the cleaning exercise was done and represents the complete data from the dataset as when it is downloaded from the Kaggle site.

Note that for clarity, some of the columns of the dataset are not shown in this capture. The omitted columns are “ID”, and “Runtime”.

```
path= "/content/drive/MyDrive/CollabDatasets/IMDB_movie_reviews_details.csv"
```

```
movies=pd.read_csv(path)
movies.head()
```

	Unnamed: 0	name	year	runtime	genre	rating	metascore	timeline	votes	gross
0	0	The Shawshank Redemption	1994	142	Drama	9.3	80.0	Two imprisoned men bond over a number of years...	2,394,059	$28.34M
1	1	The Godfather	1972	175	Crime, Drama	9.2	100.0	An organized crime dynasty's aging patriarch t...	1,658,439	$134.97M
2	2	Soorarai Pottru	2020	153	Drama	9.1	NaN	Nedumaaran Rajangam "Maara" sets out to make t...	78,266	NaN
3	3	The Dark Knight	2008	152	Action, Crime, Drama	9.0	84.0	When the menace known as the Joker wreaks havo...	2,355,907	$534.86M
4	4	The Godfather: Part II	1974	202	Crime, Drama	9.0	90.0	The early life and career of Vito Corleone in ...	1,152,912	$57.30M

The second dataset is called “Prime_TV_Shows_Data_set.csv” and is available at the following location:https://www.kaggle.com/nilimajauhari/amazon-prime-tv-shows

. Below are the columns. It has 8 columns and 404 rows. We will refer to this dataset as “prime2” in our analysis and in the attached code file.

This dataset is possibly the smallest of the 3 but was chosen again because it was representative of Amazon Prime Video data. The data in this dataset is relatively current with titles from 2020 amongst the lists. It also includes the IMDb rating that we mentioned earlier, as well as a few other interesting columns such as age, language and possibly the season variables.

```
prime2=pd.read_csv("C:/Users/Lulu/Downloads/strategicThinking/Prime_TV_Shows_Data_set.csv")
```

```
prime2.head()
```

	S.no.	Name of the show	Year of release	No of seasons available	Language	Genre	IMDb rating	Age of viewers
0	1	Pataal Lok	2020.0	1.0	Hindi	Drama	7.5	18+
1	2	Upload	2020.0	1.0	English	Sci-fi comedy	8.1	16+
2	3	The Marvelous Mrs. Maisel	2017.0	3.0	English	Drama, Comedy	8.7	16+
3	4	Four More Shots Please	2019.0	2.0	Hindi	Drama, Comedy	5.3	18+
4	5	Fleabag	2016.0	2.0	English	Comedy	8.7	18+

The last dataset we will work with is called : "netflixData.csv" and is available at the following location : https://www.kaggle.com/satpreetmakhija/netflix-movies-and-tv-shows-2021?select=netflixData.csv

This dataset is the largest amongst the three used in this project and contains 5967 rows and 13 columns. This data was chosen for a variety of reasons. The first reason this data was chosen is that the code reflects Netflix data, and as we had a dataset for Amazon Prime, we wanted to be able to evaluate both streaming platforms separately. The second reason this dataset was picked was that it contained a good balance of columns, with columns we are very interested in such as the IMDb rating and columns representing the name and description of the show, which, as will be discussed further down in this document, will be needed for our analysis.

The head of the data is shown in the screenshot below. Note that this is the head of the data as, when it was downloaded, and before any of the cleaning exercises were performed.

```
path3 ="/content/drive/MyDrive/CollabDatasets/netflixData.csv"
```

```
netflix=pd.read_csv(path3)
netflix.head()
```

	Show Id	Title	Description	Director	Genres	Cast	Production Country	Release Date	Rating	Duration	Imdb Score	Content Type	Date Added
0	cc1b6ed9-cf9e-4057-8303-34577fb54477	(Un)Well	This docuseries takes a deep dive into the luc...	NaN	Reality TV	NaN	United States	2020.0	TV-MA	1 Season	6.6/10	TV Show	NaN
1	e2ef4e91-fb25-42ab-b485-be8e3b23dedb	#Alive	As a grisly virus rampages a city, a lone man ...	Cho Il	Horror Movies, International Movies, Thrillers	Yoo Ah-in, Park Shin-hye	South Korea	2020.0	TV-MA	99 min	6.2/10	Movie	September 8, 2020
2	b01b73b7-81f6-47a7-86d8-	#AnneFrank - Parallel Stories	Through her diary, Anne Frank's story	Sabina Fedeli, Anna	Documentaries, International	Helen Mirren, Gengher	Italy	2019.0	TV-14	95 min	6.4/10	Movie	July 1, 2020

2.3.2 -Verifying data quality

This part of Crisp dm focuses on examining the dataset or (datasets in our case) and checking if the data is usable and complete. If the data is not complete, we need to identify the incomplete parts in order to turn all the data into usable data.

2.3.2.1 -Missing Data

```
# A first search for zero values in the "movies" dataset,
movies.isnull().sum()
```

```
Unnamed: 0      0
name            0
year            0
runtime         0
genre           0
rating          0
metascore     159
timeline        0
votes           0
gross         171
dtype: int64
```

Next, we look for zero values in all 3 datasets: Starting with the "movies" dataset, shown on the left, we identify all of these zero values. We will need to decide what to do with some of these columns as some of these columns are missing over half of their data.

In the "prime2" dataset, we identify these zero values. Although this dataset is not very representative of our data, we would like to keep a maximum of IMDb rating values as these are the values we are hoping to use during the exploring and modelling phase.
We will therefore need to decide what to do with the IMDb column as over half of its data is missing.

```
missing_values_table(prime2)
```

```
Your selected dataframe has 8 columns.
There are 7 columns that have missing values.
```

	Missing Values	% of Total Values
IMDb rating	222	55.0
Name of the show	11	2.7
Year of release	11	2.7
No of seasons available	11	2.7
Language	11	2.7
Genre	11	2.7
Age of viewers	11	2.7

```
# A first search for zero values
netflix.isnull().sum()
```

```
Show Id                  0
Title                    0
Description              0
Director              2064
Genres                   0
Cast                   530
Production Country     559
Release Date             3
Rating                   4
Duration                 3
Imdb Score             608
Content Type             0
Date Added            1335
dtype: int64
```

In the "Netflix" dataset, we have 5 columns with a lot of missing values, while 3 more columns only have a few missing values. As the" IMDb Score" is a column we have a keen interest in, we will need to be careful how we handle and sort this data.

2.4 -Phase 3: Data Preparation

2.4.1 -Cleaning data

During the cleaning exercise, we hope to turn the available data into a dataset that is usable.
This means that during this stage of the crisp dm we need to remove what we do not want from the dataset and decide on what we want to keep.

2.4.2 -Drop unnecessary columns

In the "movies" dataset we found all the columns to have some value, so we have decided to keep them all. As this dataset is quite small, we felt it was necessary to keep it as complete as possible.

In the "prime2" dataset, we are also keeping all columns. this is the smallest of all datasets, so we want to keep it as complete as possible,

In the "Netflix" dataset we are going to drop the "Show Id "column as we do not feel we need it for our analysis at this point.

```
# We are also going to drop the column called "Show Id", as we do not see a need to keep this column
netflix.drop(['Show Id'],axis=1,inplace=True)
```

2.4.3 -Dealing with zeros

The columns with zero values and missing data were identified in a previous step.
We also know that in some cases, those values represent a very little percentage of the data but in other cases, they represent a substantial part of the data. This means that, at this point, we need to make choices.

We have decided that it would be best to drop the column called "metacore" and "gross" from the "movies" dataset as there are many values missing from those columns. "Metascore" refers to another method where scores are attributed to various "metacritics" and are also weighted scores. As we are working with IMDb scores and ratings which are assigned by users of the sites, we thought the "metascore" category may lead to confusion and have therefore decided to drop it.

```
# From the "movies" data set , we are going to drop "metascore" column and the "gross" column since so many values are missing
movies.drop(['metascore'],axis=1,inplace=True)

movies.drop(['gross'],axis=1,inplace=True)
```

In the "prime2" dataset, we want to keep the "IMDb rating" column. The reason behind the decision is simple. As the IMDb column will be used for forecasting and modelling, we need this data to be as complete as possible. Here, we have decided that filling the NaN values with zeros

was actually better than dropping otherwise good data. This is not an ideal solution, replacing NaNs with zeros will skew the correlations and models for this data. However, the team feels we have very little choice other than transforming those NaNs into zeros. The same is performed for the "number of seasons available" column.

```
prime2['IMDb rating']=pd.to_numeric(prime2['IMDb rating'],errors='coerce').fillna(0)
```

```
prime2['No of seasons available']=pd.to_numeric(prime2['No of seasons available'],errors='coerce').fillna(0)
```

Finally, we look at the "Netflix" dataset.
Here a decision was made to drop the "Director" and "Date Added" columns. The reasoning behind this is common in parts. The "Director" column has about 2000 values missing, with a dataset of about 5000 values, the missing data in that column represents a very high proportion. The Date Added column is missing about 1300 values which is about one fifth of the data, and for this particular column, we have no particular interest in keeping those dates for analysis.

```
# From the "netflix" data set , we are going to drop  the "Director", "Cast","Production Country" and "Date Added" column
netflix.drop(['Director'],axis=1,inplace=True)


netflix.drop(['Date Added'],axis=1,inplace=True)
```

2.4.4 -Dealing with Duplicates

Duplicates or zero values can skew the results of the analysis and the results of any forecasting, so we need to try to address this issue, first by identifying these values and later, during the cleaning exercise, we need to decide on either removing them or by replacing them with a suitable alternative.

We check for duplicates in all three datasets:
During this exercise, we find 0 duplicates.

```
# Looking for duplicates  in the "movies" dataset and removing them if necessary. we are keeping the last occurence.  0 duplicates found.
len(movies)-len(movies.drop_duplicates(keep='last'))
```

```
0
```

```
# Looking for duplicates in the prime2 dataset and removing them if necessary. we are keeping the last occurence.  0 duplicates found.
len(prime2)-len(prime2.drop_duplicates(keep='last'))
```

```
0
```

```
# Looking for duplicates  in the "netflix" dataset and removing them if necessary. we are keeping the last occurence.  0 duplicates found.
len(netflix)-len(netflix.drop_duplicates(keep='last'))
```

```
0
```

2.4.5 -Exploring data

2.4.5.1 -Correlations

The first task is to check correlations. Finding correlations is an important step in the data preparation because it can expose the relationships between various columns.
We look for correlations on each of the datasets:
The table below represents the correlations for the "movies" dataset

```
#to look for correlations within this data set
movies.corr()
```

	Unnamed: 0	year	runtime	rating
Unnamed: 0	1.000000	0.045961	-0.238318	-0.939927
year	0.045961	1.000000	0.190382	-0.123427
runtime	-0.238318	0.190382	1.000000	0.243063
rating	-0.939927	-0.123427	0.243063	1.000000

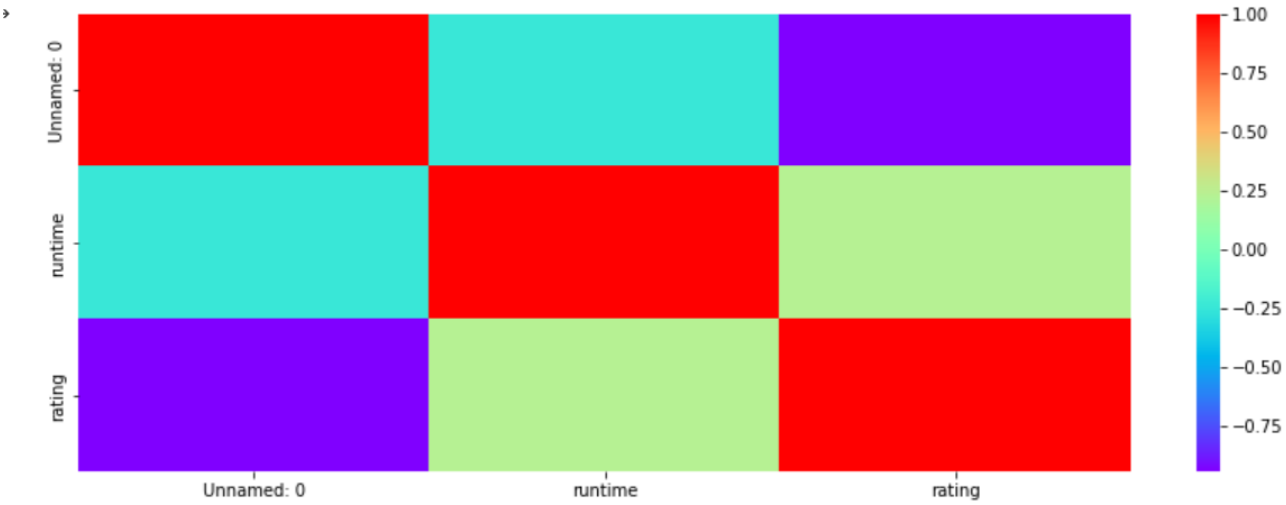
We then want to visualize these correlations on a heatmap as it is easier to interpret.
Below is the heatmap of correlations for the streaming dataset:
From the graph below we can see that the rating column is correlated to runtime, but the correlation is not very strong.

```
# Plotting correlations on a heatmap

# Figure size
plt.figure(figsize=(16,8))

# Heatmap
sns.heatmap(movies.corr(), cmap='rainbow');
```

(Code 1 movies. corr heatmap)



When looking at the "prime2" dataset, we also notice some correlations. We first check the correlations on a table (table below):

```
# To look at correlations of the "Prime2" dataset
# We look at the person correlation between variables first
# The closer we are to 1 the more correlated a value is
prime2.corr()
```

	S.no.	Year of release	No of seasons available	IMDb rating
S.no.	1.000000	-0.143054	-0.233798	-0.216746
Year of release	-0.143054	1.000000	-0.211353	0.182694
No of seasons available	-0.233798	-0.211353	1.000000	0.174057
IMDb rating	-0.216746	0.182694	0.174057	1.000000

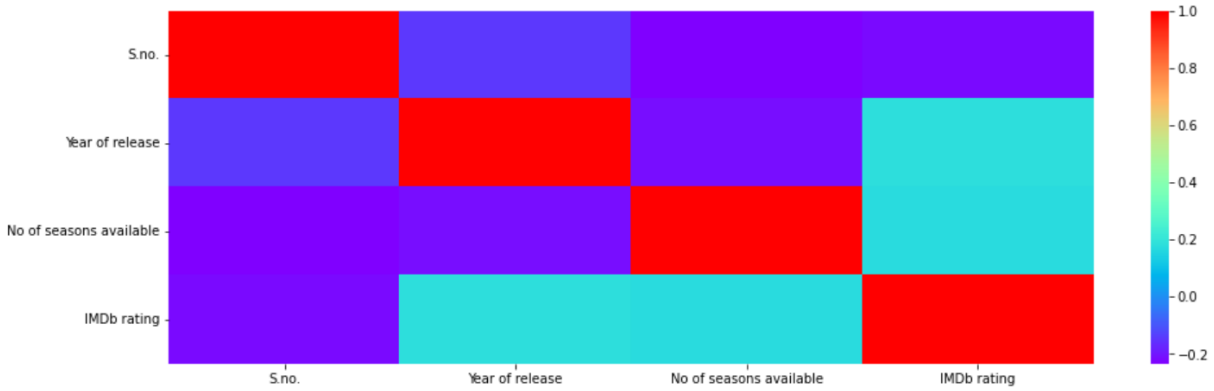
We then plot these on a heatmap to able to visualize these more easily:
We notice that we have a correlation between the "no of seasons available", "year of release" column and the "IMDb" rating column, but the correlation is quite low also.

```
# plotting correlations on a heatmap

# figure size
plt.figure(figsize=(16,8))

# heatmap
sns.heatmap(prime2.corr(), cmap='Blues');
```

(Code 2 prime2.corr heatmap)

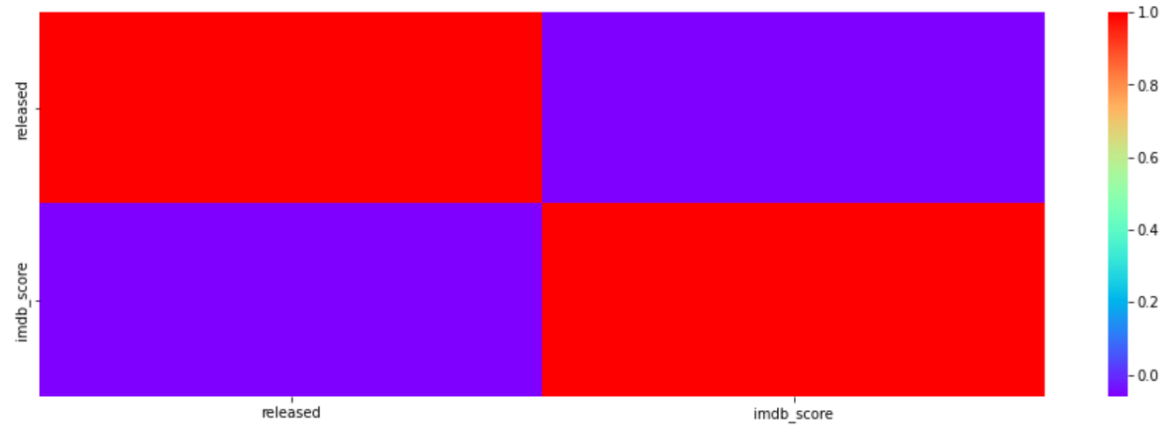


From the “Netflix” dataset, we seem to have a very low negative correlation with the “released “column

```
#looking at the correlations of the data.
netflix.corr()
```

	released	imdb_score
released	1.000000	-0.059626
imdb_score	-0.059626	1.000000

(Code 3 Netflix. corr heatmap)

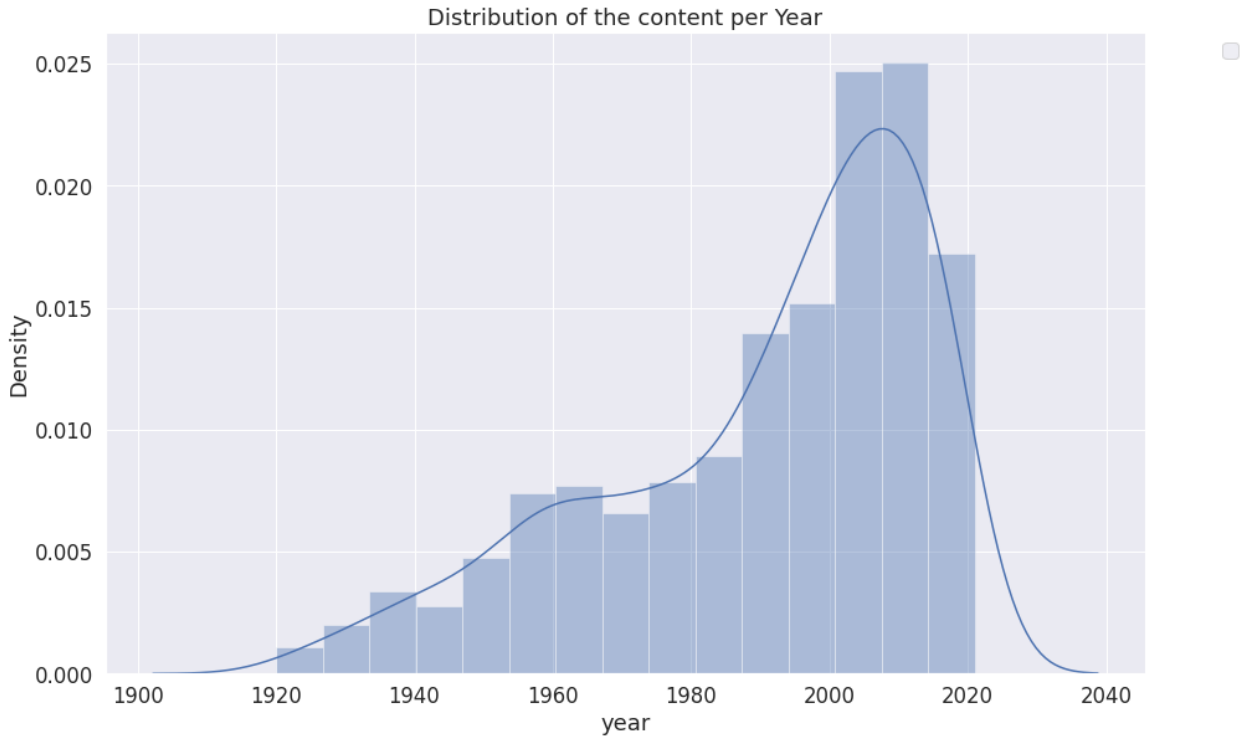


2.4.5.2 -Distributions and Outliers

We represent some of the distributions of the datasets to check their shape. The distribution can also be a good representation of the mean to see if the curve is evenly distributed or not. It can sometimes also show outliers, as we will see in some of the distributions represented below. We also implemented the color theory we had learned and decided to opt to show our visualizations in one color scheme in order to have a better flow.

We first start with the "movies" dataset.

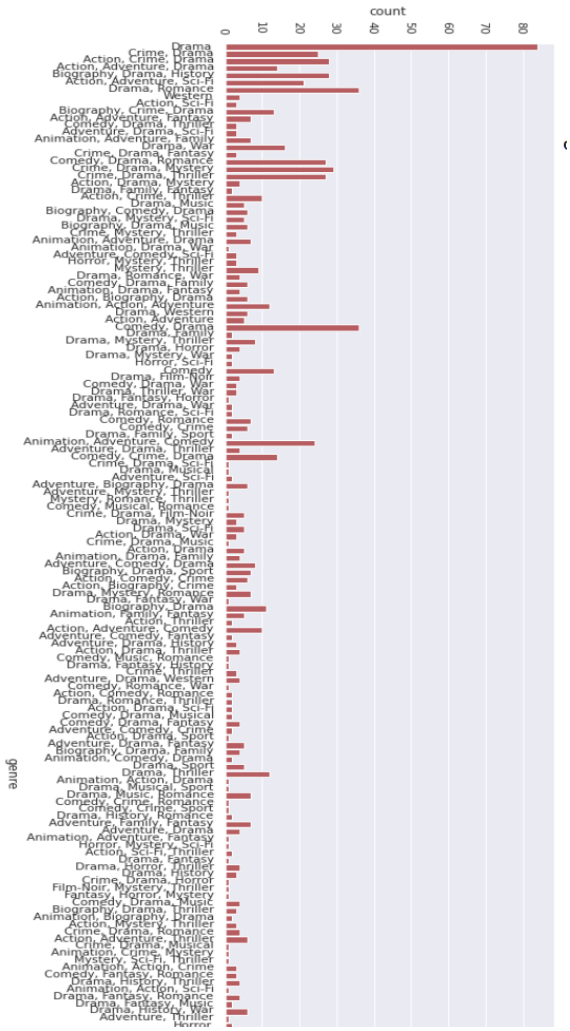
(Code 4: Movies Distribution of the content per year)



The first distribution below represents the year the content was created. We can see that this is not a normal distribution and shows a negative skew. It also possibly identifies some outliers for the years before 1920, for which there are very few movies or items uploaded

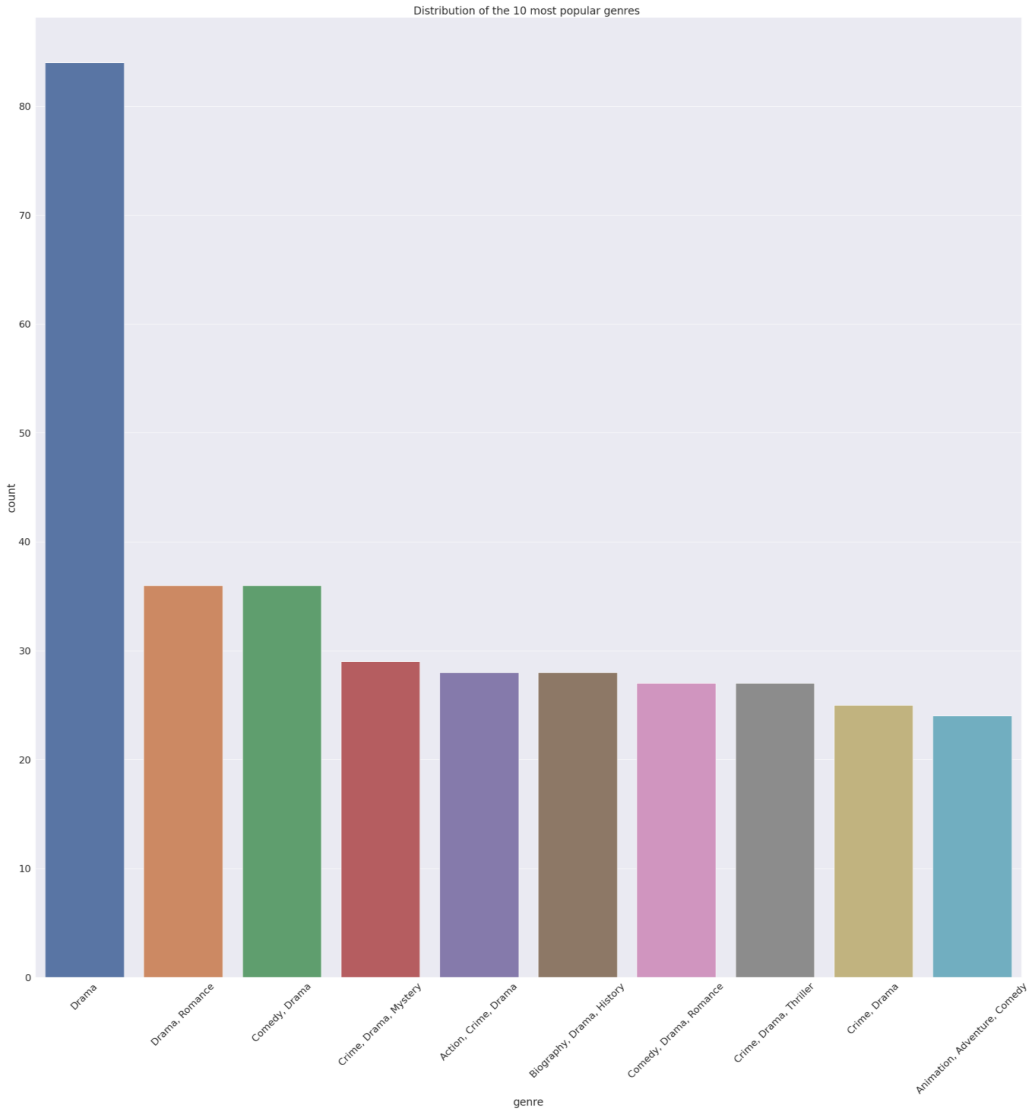
Next, we plot the Genre category, this is quite difficult to plot as movies can belong to different categories. However, from the distribution below we can notice a positive skew in the distribution, with the “drama” section significantly higher above the others.
(Note that the plot represented here does not represent the complete set of values. The plot had to be shortened for page esthetics

(Code 5: movies genres)



As the previous figure was not showing the results accurately for the movie genres for the complete dataset, we thought it would be better to show the top ten of the movie genres and have created the plot below.

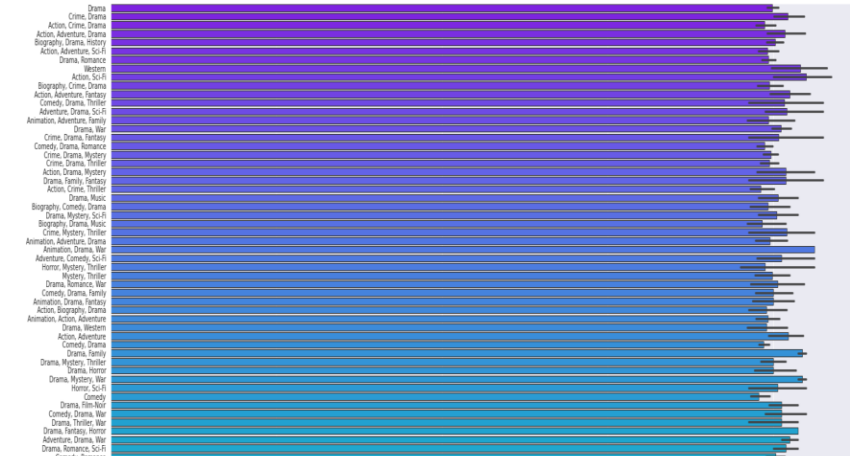
(Code 6 distribution of the 10 most popular movie genres)



As the “genre category seems important for the analysis, we have decided to compare it to the ratings, so we create this plot

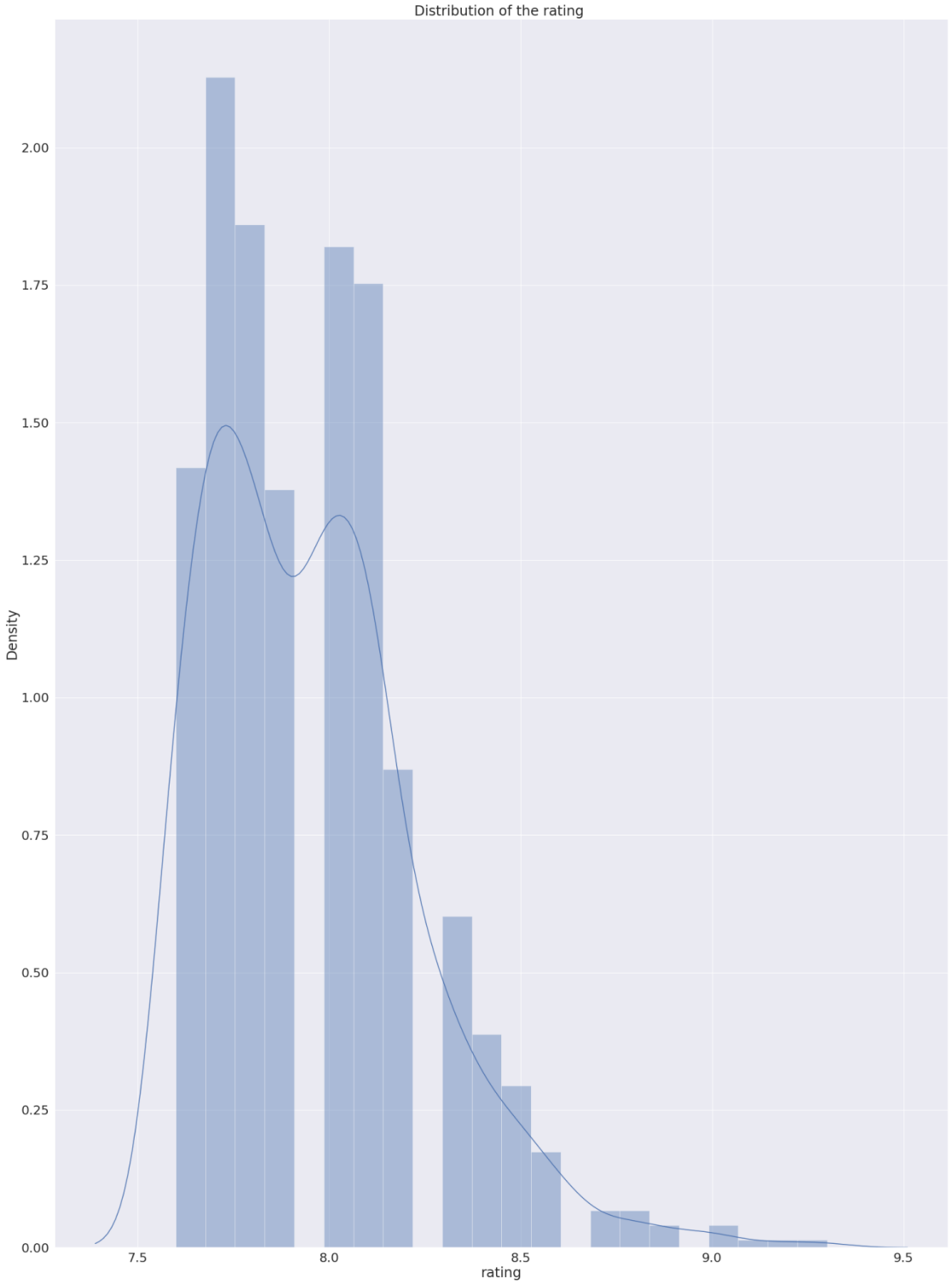
(Code 7 rating per genre)

```
sns.set(rc={'figure.figsize':(28,40)})
sns.barplot(x='rating',y='genre',data=movies,palette='rainbow', edgecolor=".2")
plt.xticks(rotation=90);
```



We look next, at the ratings, as this column is very valuable for our analysis

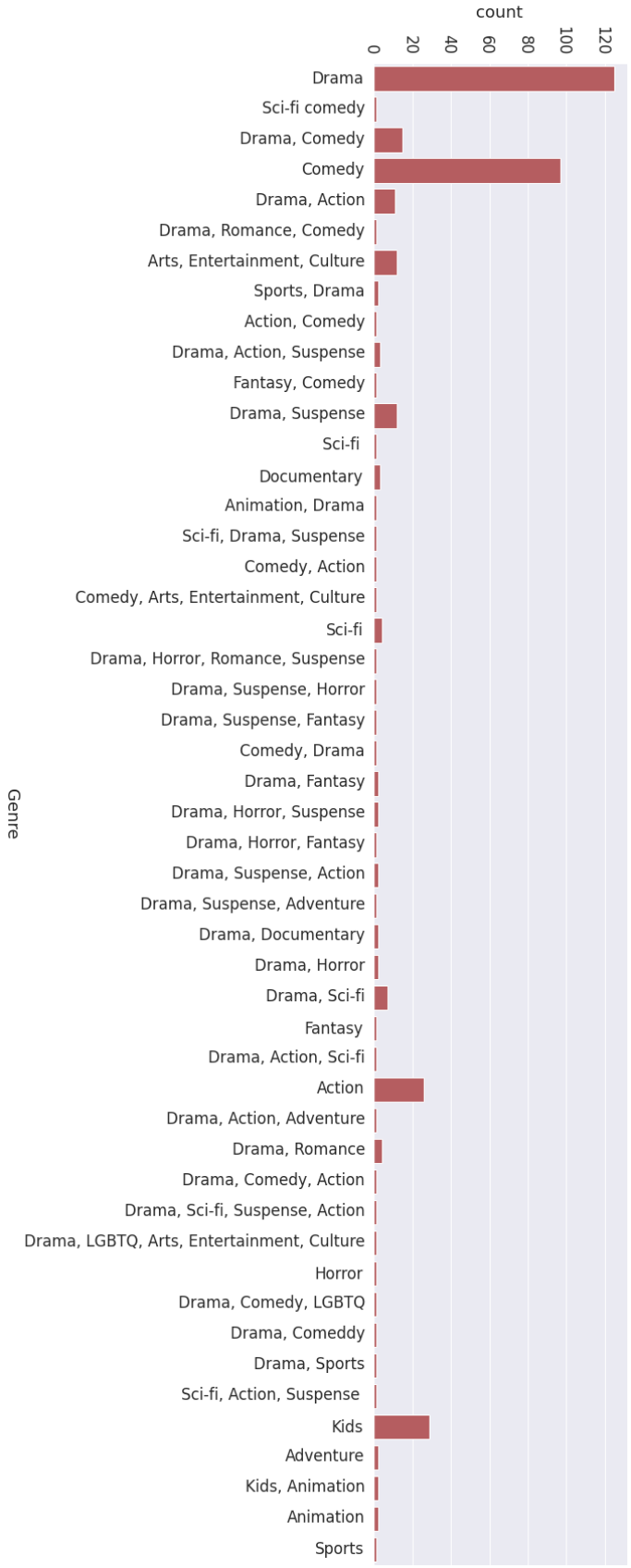
(Code 8 Distribution of the rating)



We then overview the “prime2” dataset’s distributions.

As this dataset also has a “genre” column, we decided to check its results. Again the “Drama” genre is well populated, followed by “Comedy”. According to this data, the gap does not seem as wide between the content available for “Drama” and “Comedy” genres.

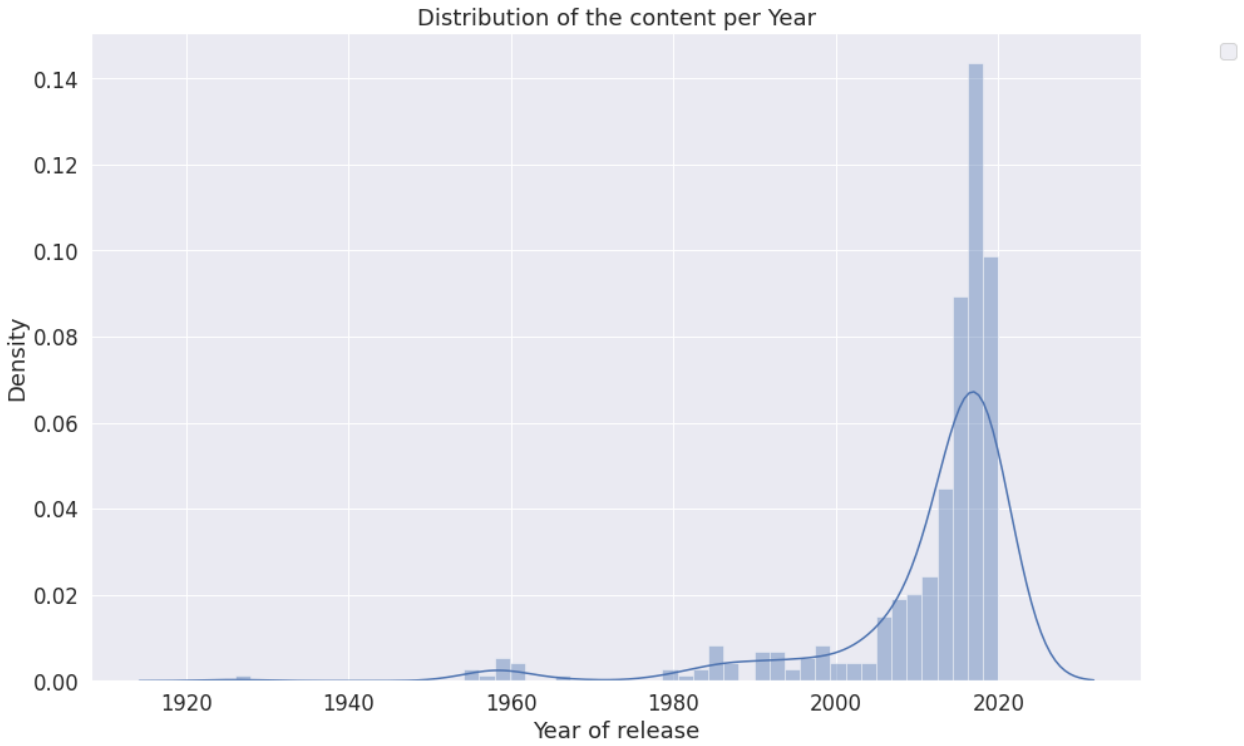
(Code 9 prime2 Movie genres)



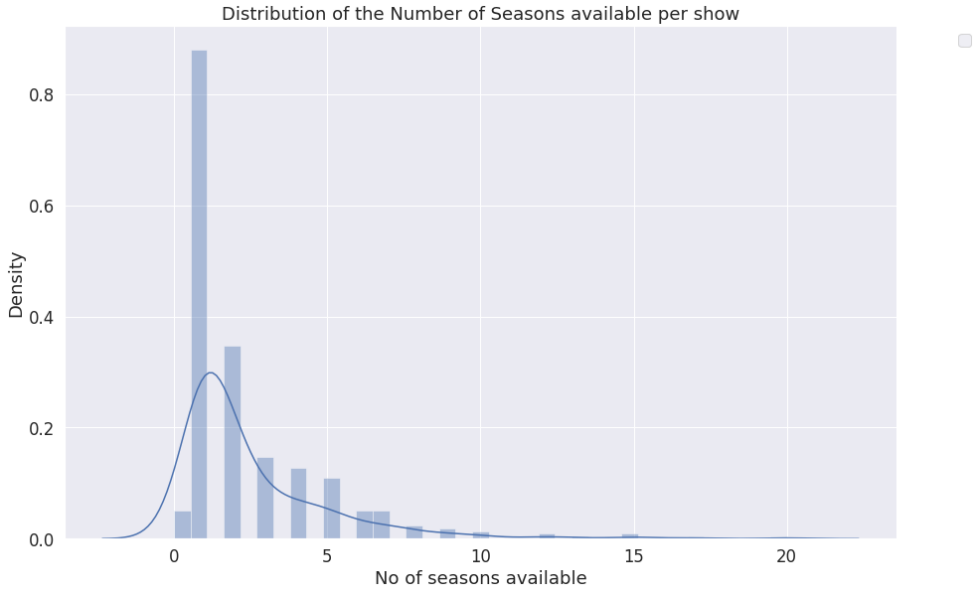
As this dataset also contains the “year” column, we represent the distribution of the Amazon Prime content across the years.

(Code 10 prime2: content per year)

Once again, the curve shows a distribution that is not normal and shows a negative skew. It also shows that most content is from 1980 onwards and possibly identifies outliers before this date.

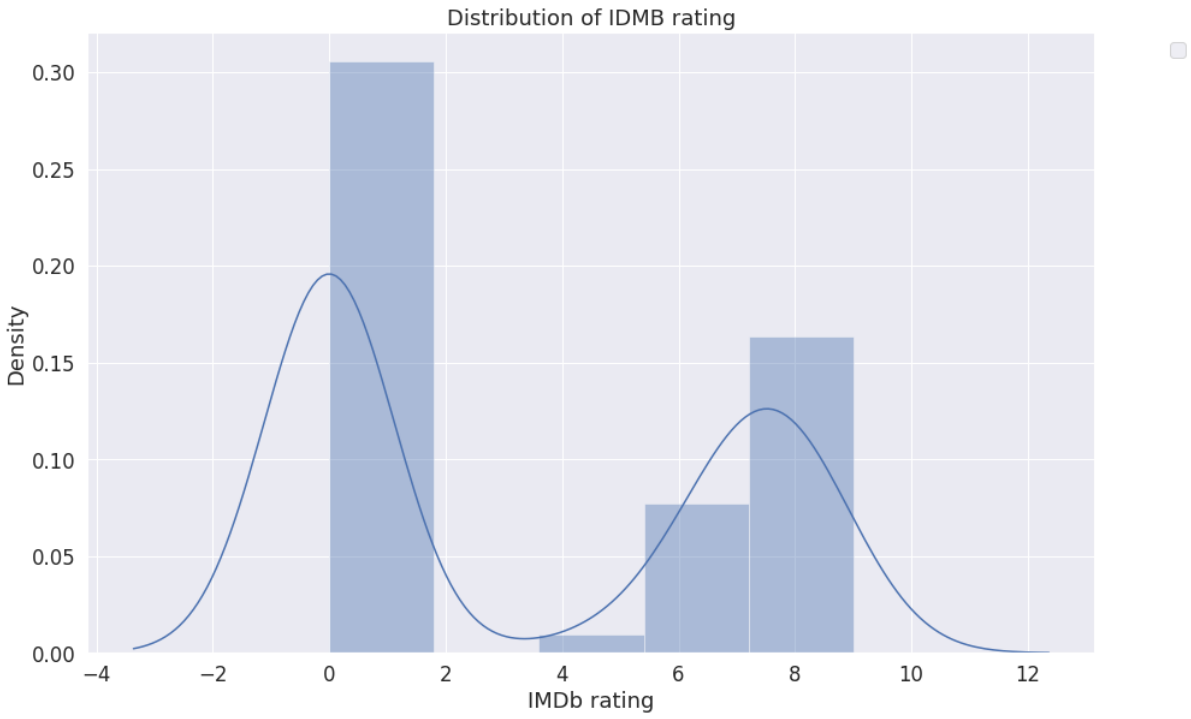


(Code 11: prime2: number of seasons available)



Next, we look at the IMDb rating, which is an uneven distribution. However, for this plot, one must remember that the zero values represent the missing data so the relevant plot data is only values outside four and up, which would show a negative skew. The plot is small but clearly indicate the missing or zeroed values and the ratings as the alternative right part of the plot

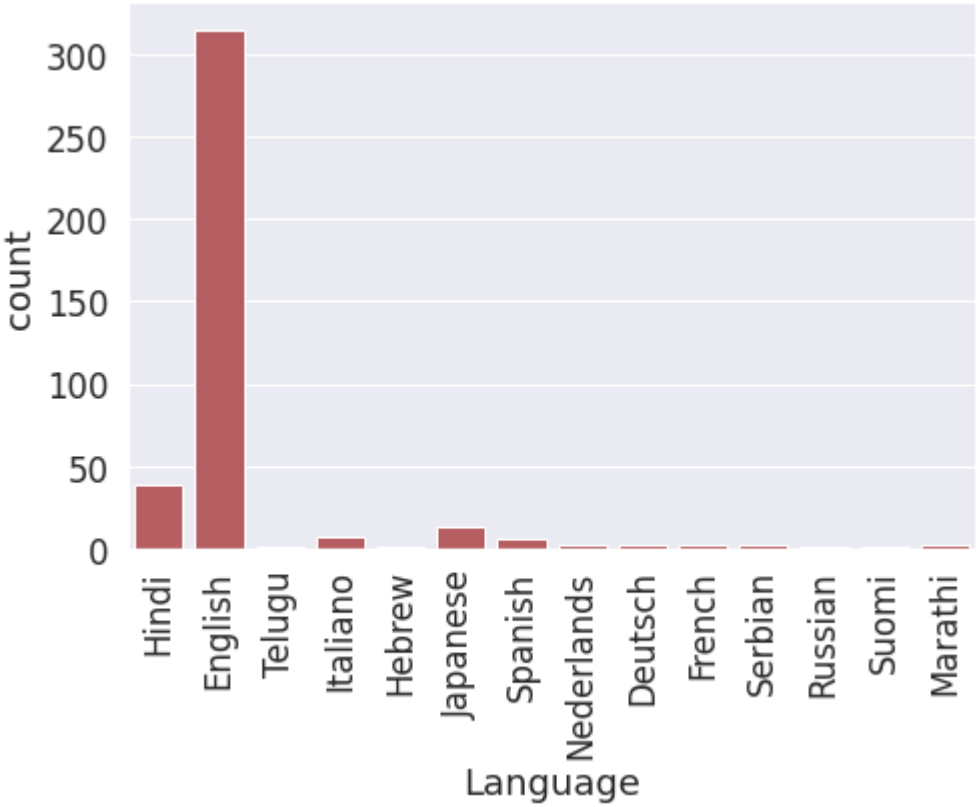
(Code 12 distribution of the IMDb rating)



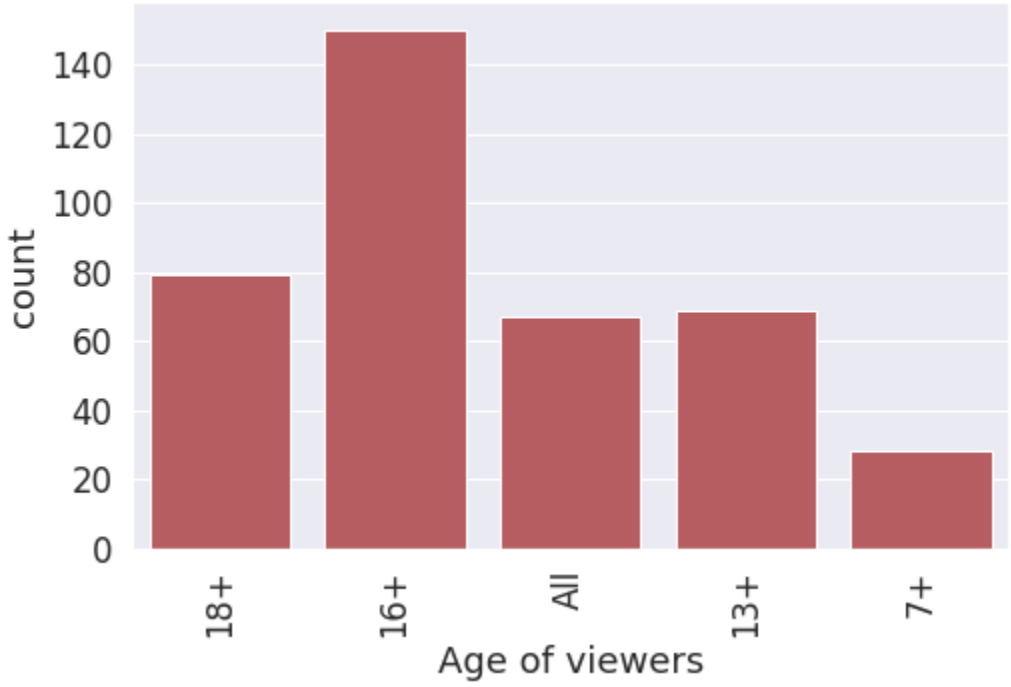
We then decide to represent a number of other graphs that are perhaps less relevant, but still a good representation of the dataset.

The first plot shows the number of seasons available, and the two other plots represent the language and the distribution of the ages of the viewers.

(Code 13 Prime2: Language)

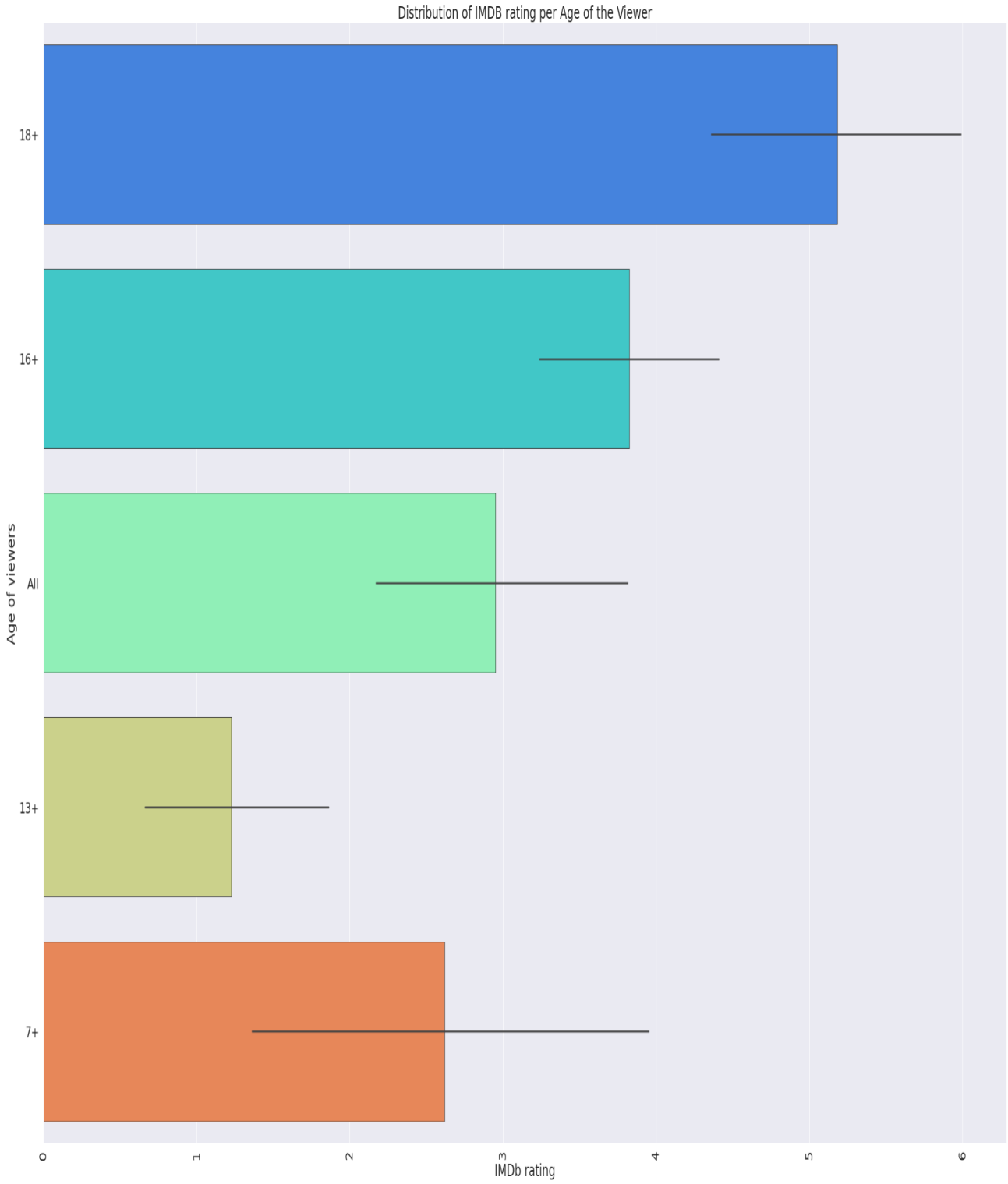


(Code 14 Prime2: Age of the viewers)



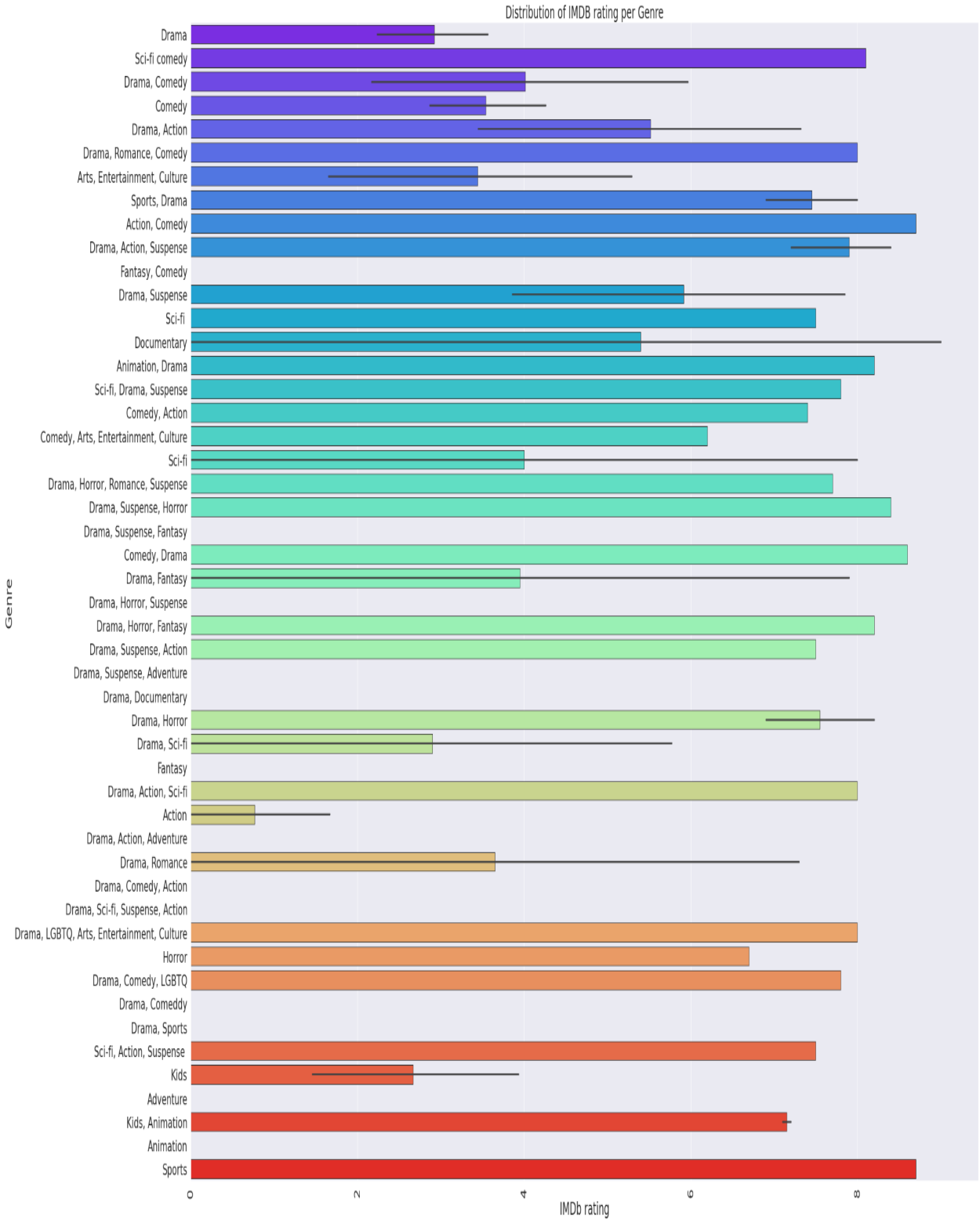
As the work we are planning concentrates on the IMDb scores, we thought the graphs below represented the IMDb scores well on a number of categories.
Below is the iMDb representation against the age of the viewers.

(Code 15: prime2 rating versus Age of the viewers)



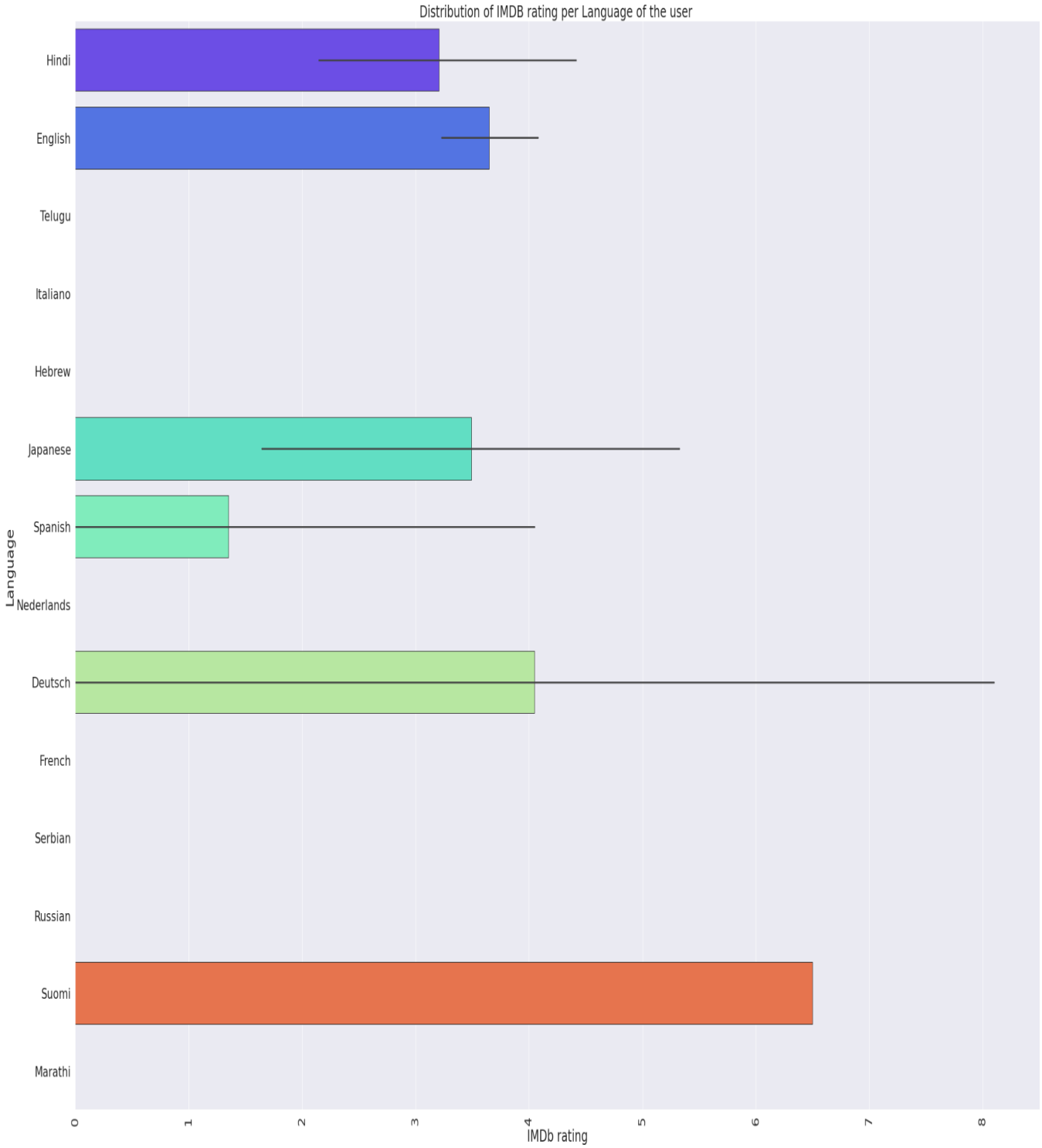
Next, we represent the rating per genre

(Code 16 prime 2: rating per genre)



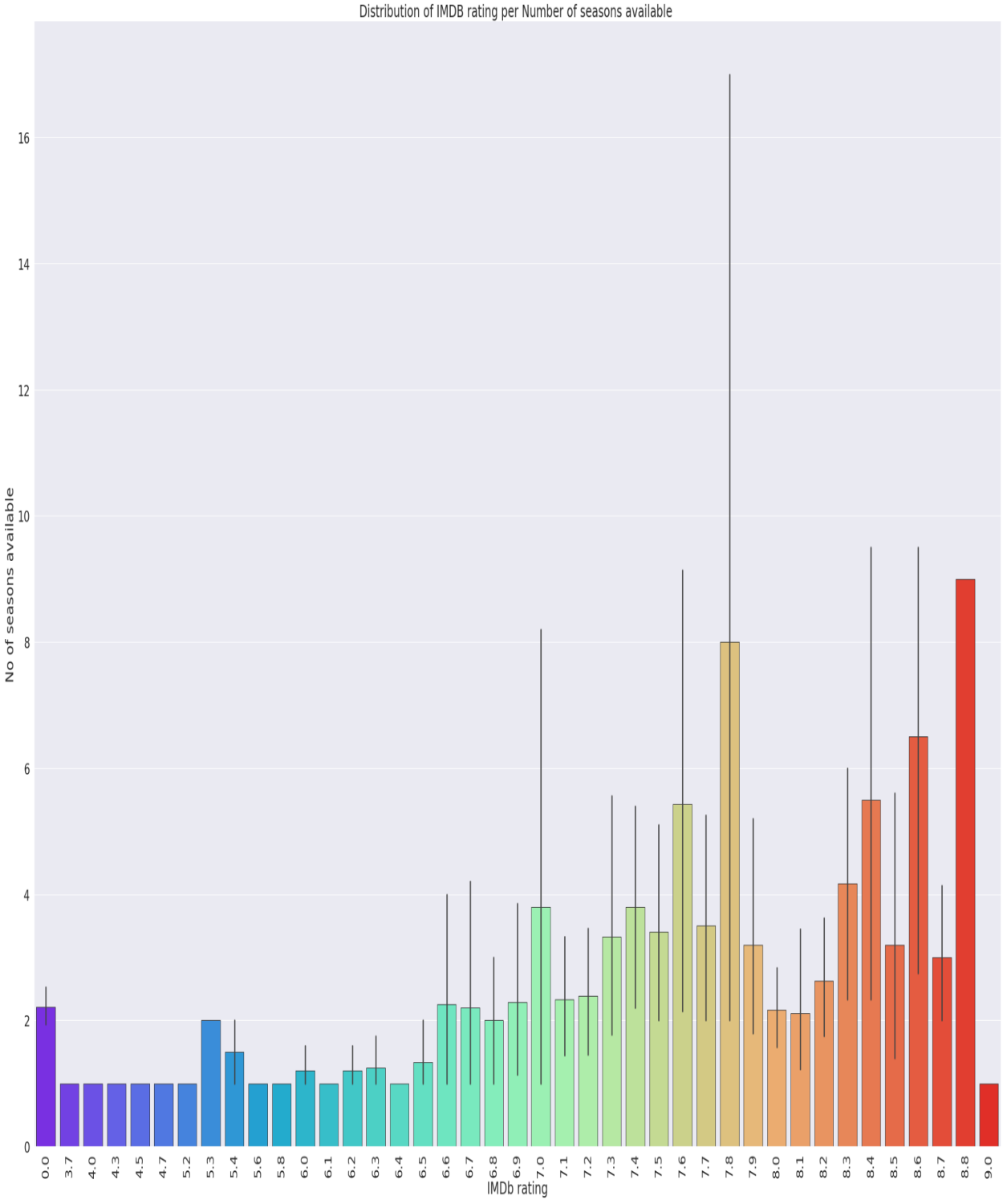
Next, we represent the rating per language. This one is a bit of a surprise with most ratings found for Suomi, which is Finnish, followed by Deutsch, which is German. English comes in third position, closely followed by Japanese.

(Code 17 prime2: rating per language of the viewer)



We plot the IMDb versus the number of seasons. This gives a good indication that the higher ratings come from the titles that have multiple seasons. However, it is difficult to visualize at this stage.

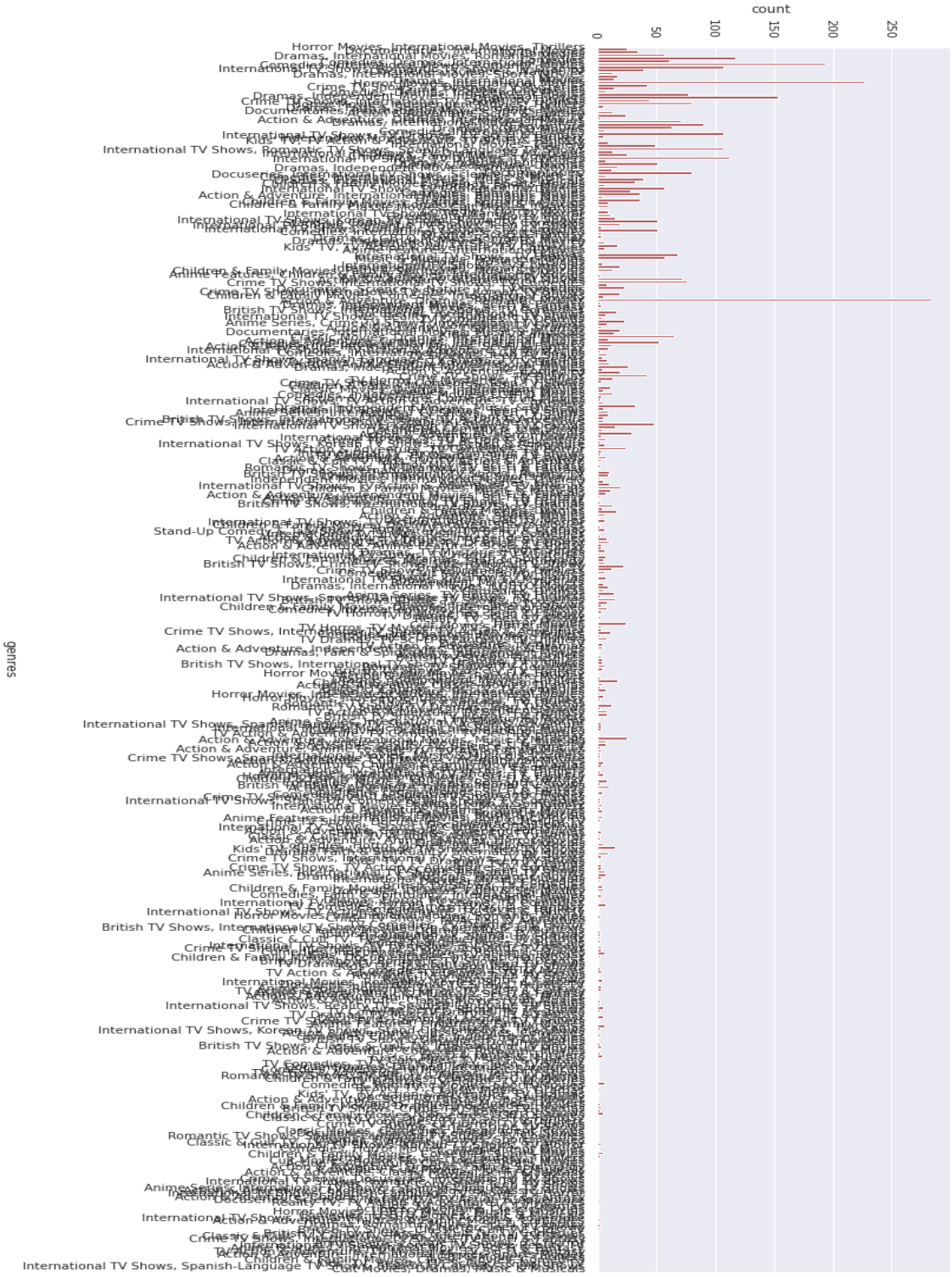
(Code 18 prime2: rating per season)



We now overview the distributions for our last dataset, the “Netflix” dataset,

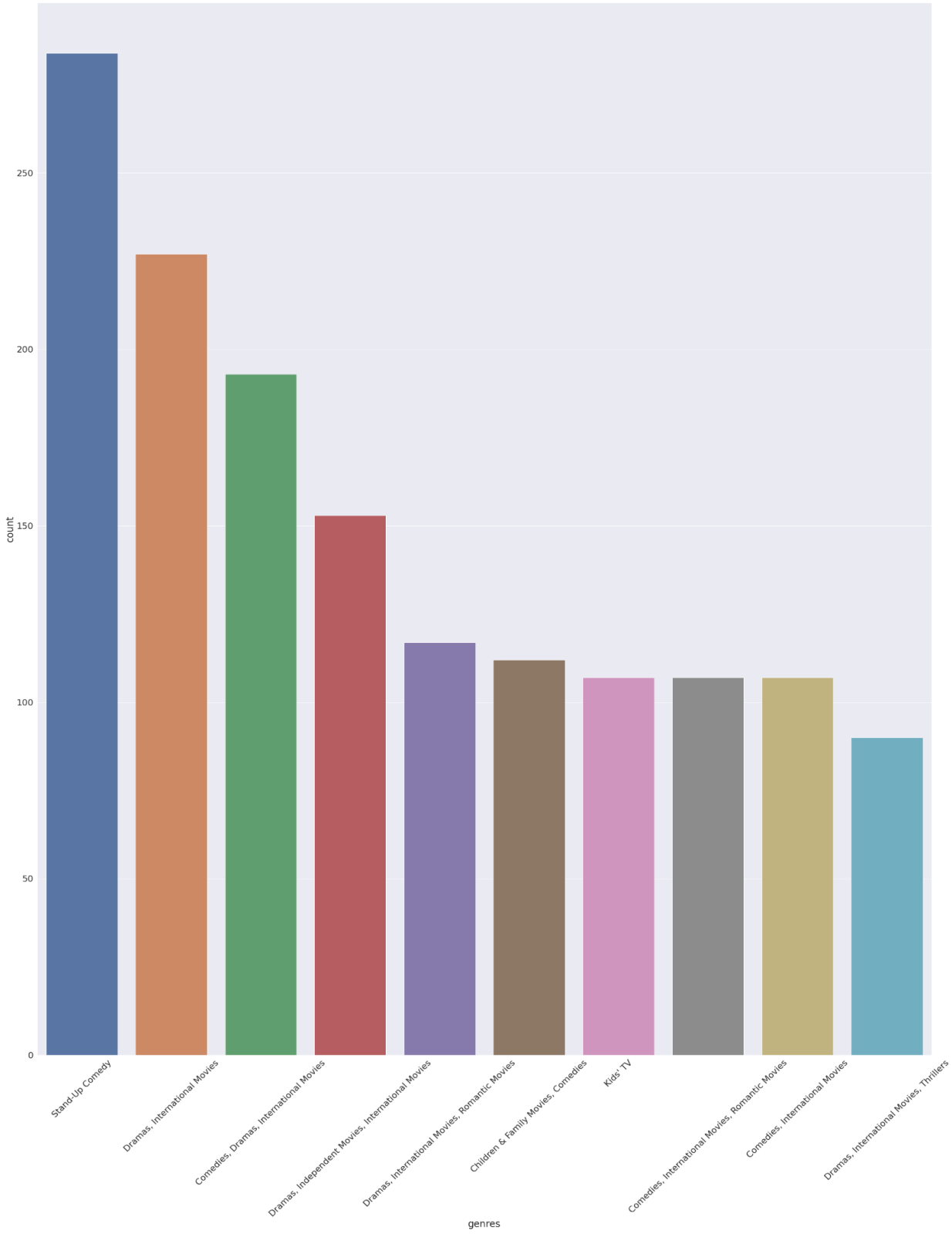
We first look at the “genres" column, like we did for the other datasets, however in this case, this plot is very difficult to read and interpret, because there are just too many categories listed.

(Code19 Netflix: genres)



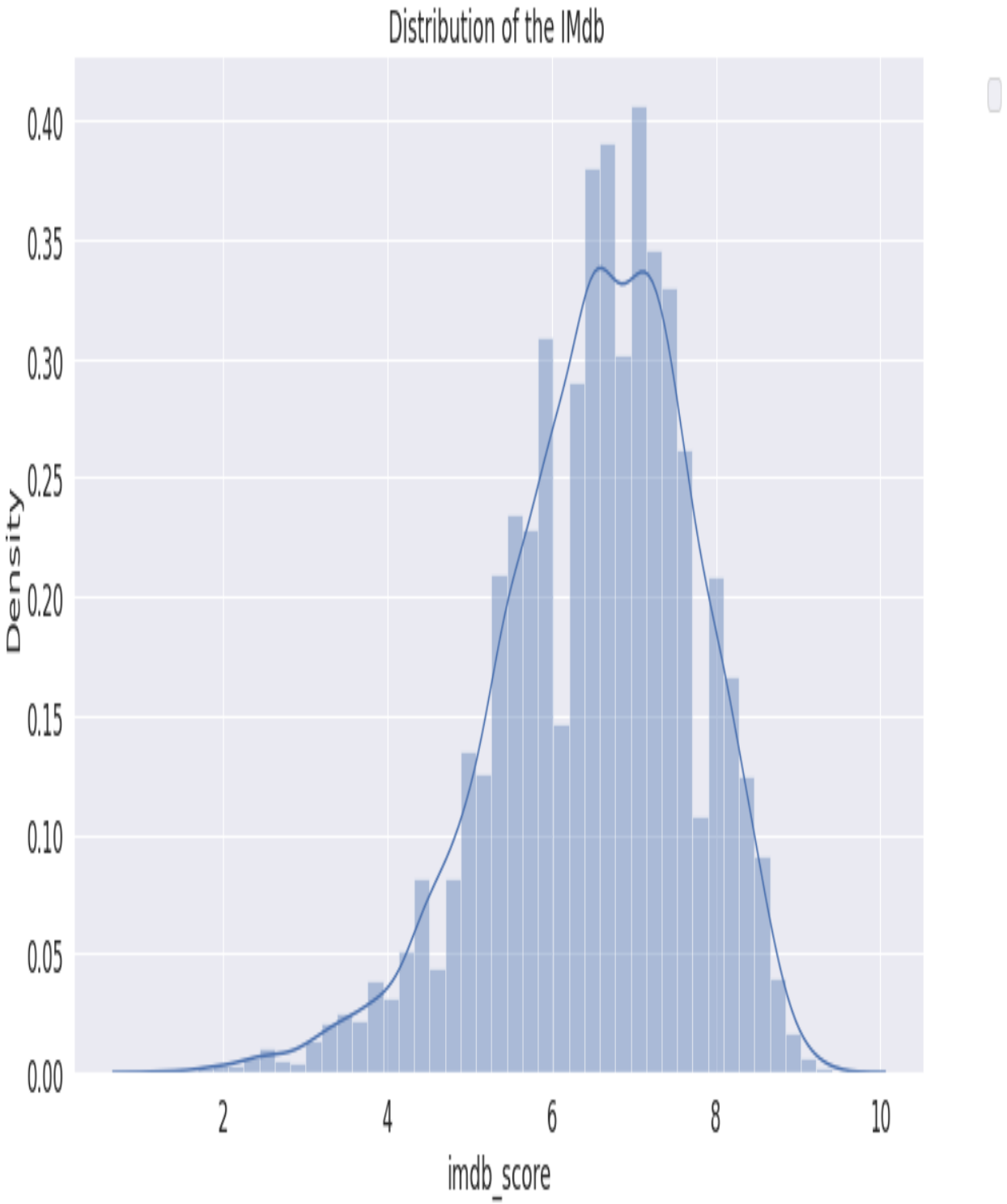
As this graph is not readable, the team thought that having a smaller graph representing only the top ten movie genres would be much better to represent this data, so we have created this plot.

(Code 20 Netflix top 10 genres)



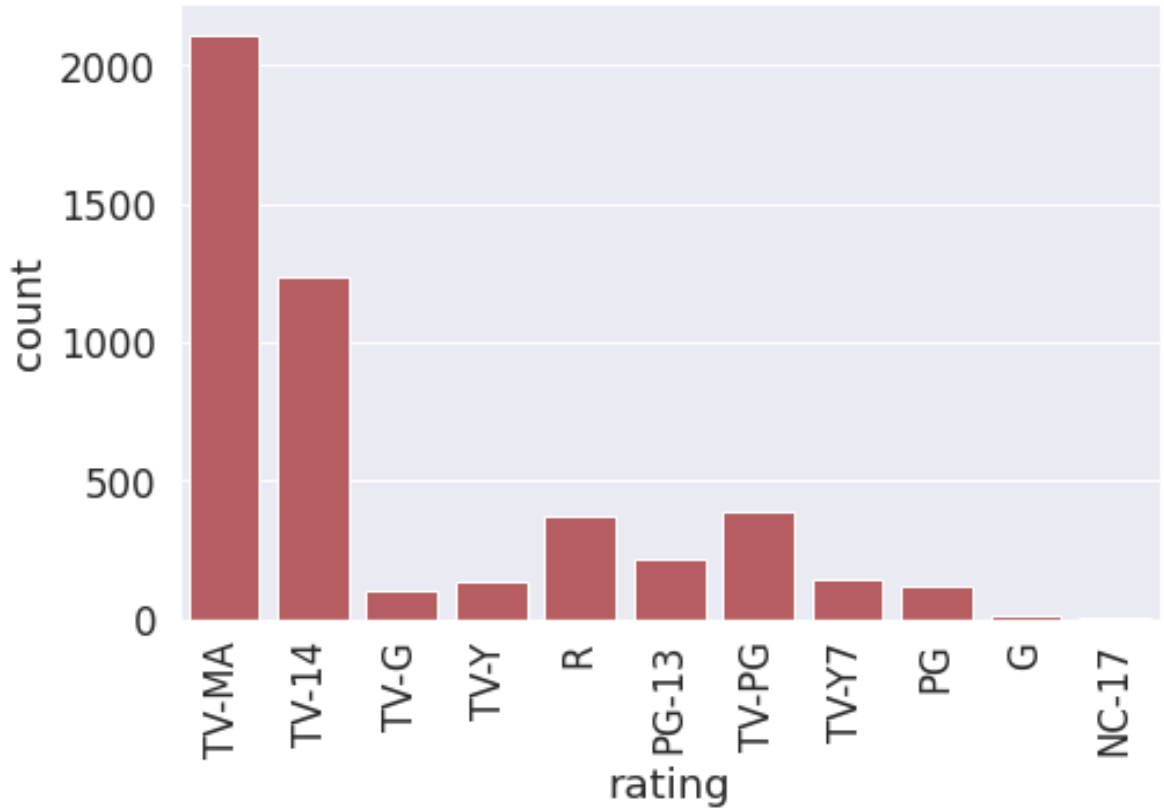
Next, we want to see the distribution for the IMDb score as it is important for this analysis. The distribution looks relatively even with perhaps a slight negative skew.

(Code 21 Netflix: Distribution of the IMDb)

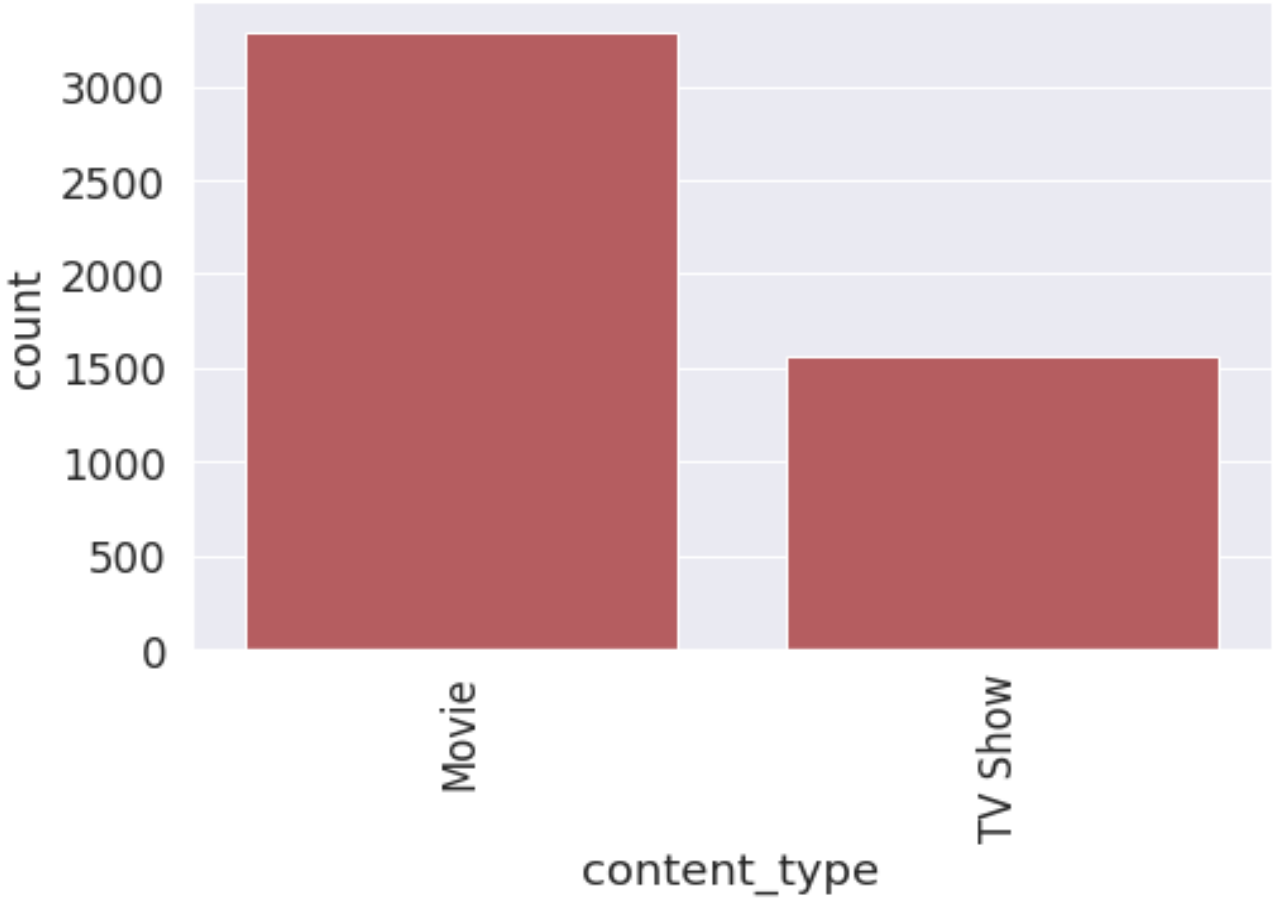


Now to represent some of the other categories and distributions, that still have a value for the analysis, but perhaps not as critical.

(Code 22: Netflix: movie age rating type)

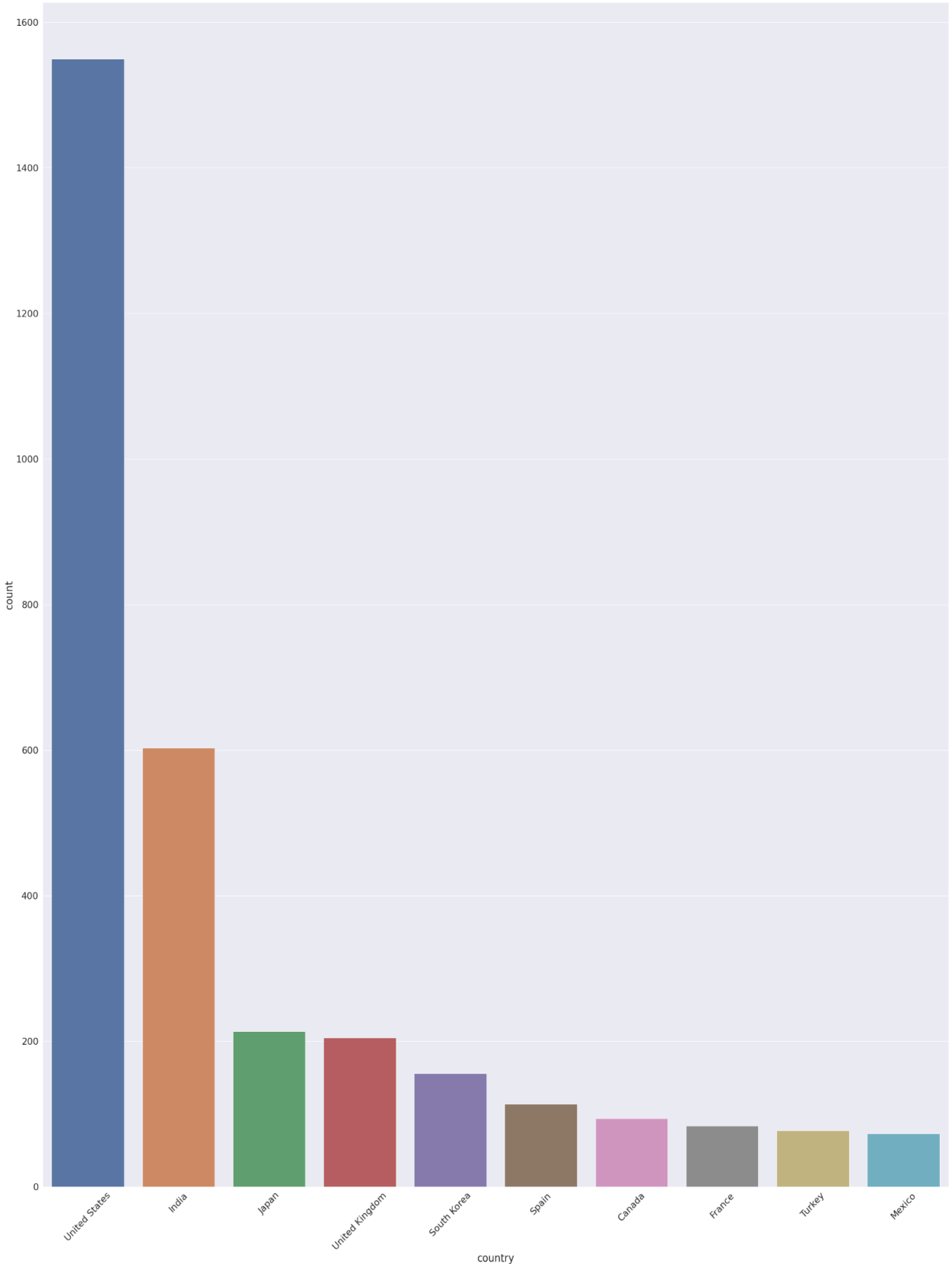


(Code 23: Netflix: content type)



We also represent the countries of production for the shows:

(Code 24: Netflix: Countries of production for the movies)



2.5 -Phase 4: Modelling

2.5.1 -Selecting modelling techniques

In phase 2 we had planned to look at different classification models to try and find the best solution to help answer our questions. We looked at KNN, Logistic regression and Decision Trees approach. When we looked further into these models and answering our questions, we looked at different models such as Sentiment analysis with the use of Classification with Countvectorizer, Classification with the use with TF-IDF features and decision Tree Classifier. For this method to work we must first clean our data in the three different datasets and generated a procedure to test the quality and validity of the model. We did this by splitting the data into training and testing sets. Then built the model on the training set and then looked at the quality of it on the test set.

The first modeling we did was Classification with Countvectorizer, Classification with the use of TF-IDF features. Since it allows us to do Sentiment analysis on the words we had in this dataset. Sentiment analysis allows us to determine the emotional tones behind the words used in the titles to understand and analyze the different options and emotions of them.

The next method we used was a Decision Trees approach. We had decided to use this model in phase 2.
A Decision Tree is as the name suggests a method to reach a decision in a tree like model. An example of a simple decision tree would be to think of ordering a drink for example. The first decision to make is whether you want a cold or hot drink. Once we know we want a hot drink, we need to know if it contains caffeine, yes or no. Assuming the answer is no, then finally we should arrive at a decision that the drink is hot chocolate. In a machine learning system, a decision tree is used both as a classification method and as a regression method. According to the sci-kit learn library on decision trees, its aim "is to create a model… based on simple decisions" learned from the data that we imputed. They are usually quite thorough.

2.5.2 -Model assumptions

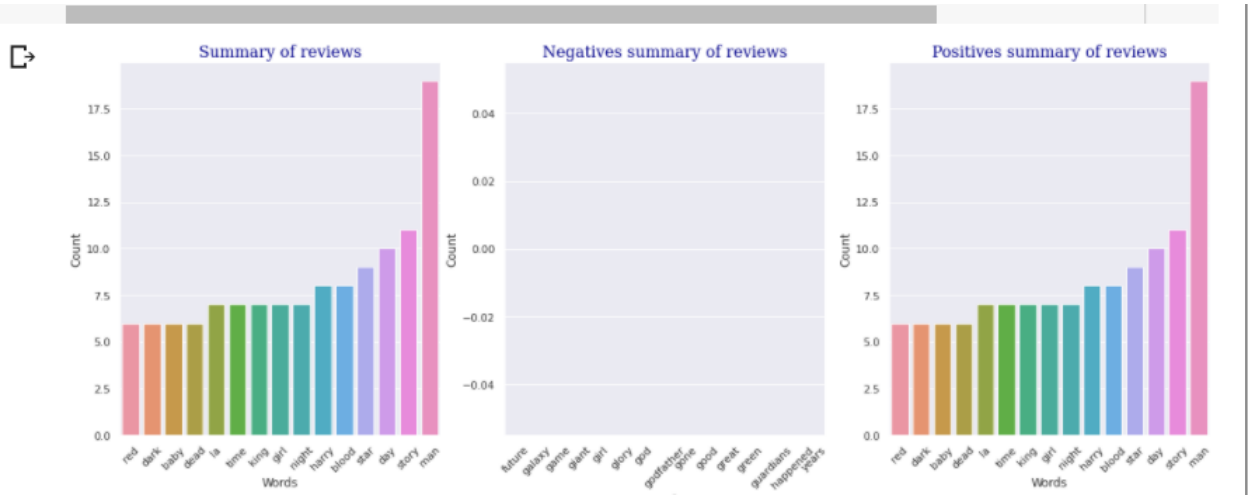
All models have different assumptions that we had to look at when picking the models, we will use to answer our question in the dataset. We must be aware of these assumptions when analyzing the data, we get from our models.
Sentimental analysis is done by TextBlob which looks at its own standard as to what is determined to be Positive, Negative and Neutral words.
Decision trees use all the data as their root, while features are usually categorical. This is why we felt it was the best model to help answer our questions

2.5.3 -Building test design(s)

To make sure we pick the best model to answer our question and get the best accuracy we will split the data into training and tests. We will try and do different splits to find the best split to give us the highest accuracy.

2.5.4 -Building model(s)

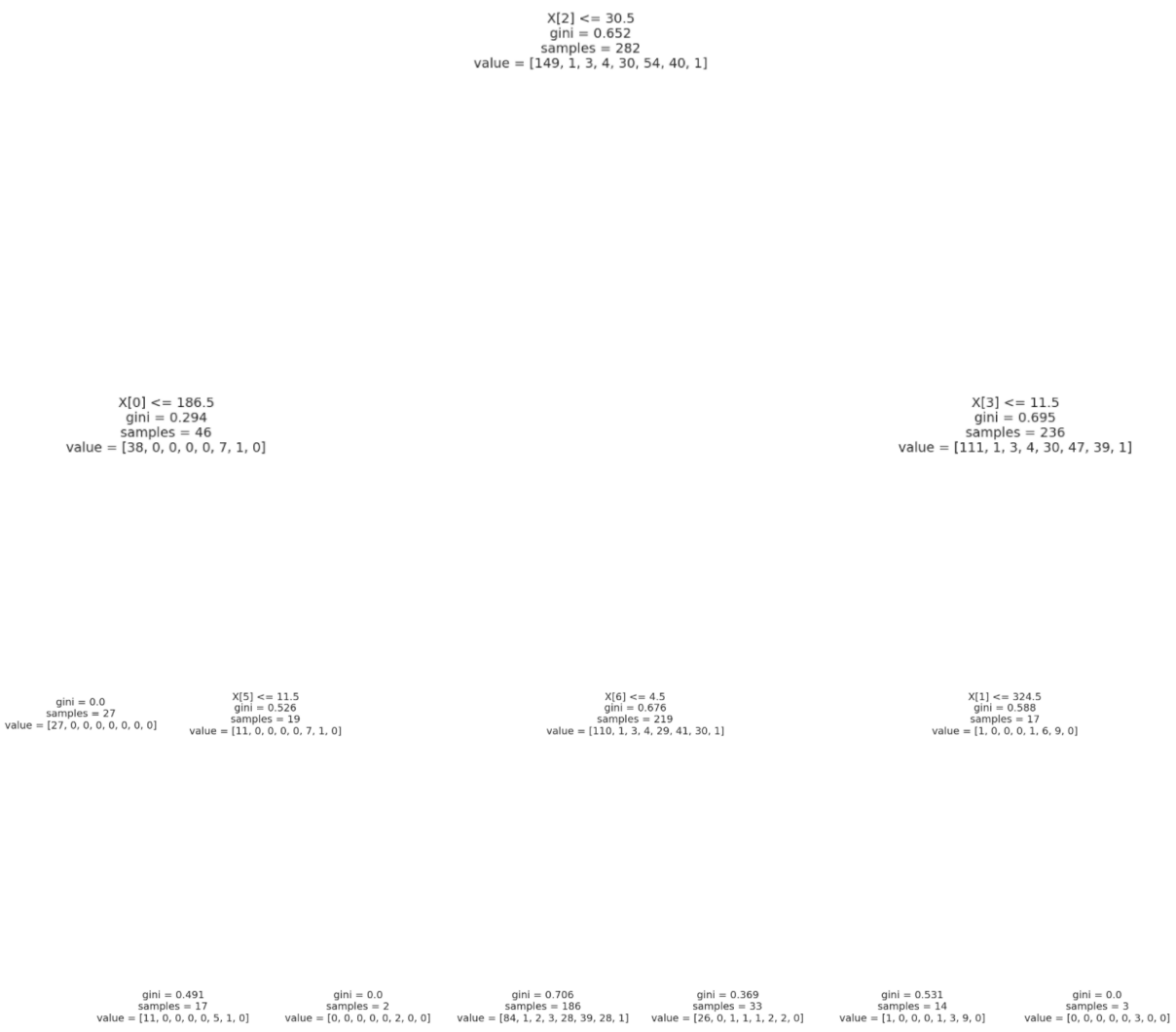
Our first dataset we looked at was the IDMB. This was the new dataset we bought into help answering our question we were asking from the previous datasets. We first cleanse the data by removing all the duplicates and zeros values. We also dropped any columns we did not feel added to the data. The CountVectorize helped us analyze the number of unique words in the titles of the films in the dataset. We got a result of 245 unique words. We wanted to know and see the difference between the positive and negative words used and if they resulted in the negative reviews for the dataset. We then looked at the reviews left in the dataset if they were positive or negative and got the below results. From the below result we can see that all the reviews had positive words rather than negative.



We looked at the titles of the movies in the dataset and saw that the movies all had positive titles. We had 805 unique words in the tiles of the movies listed

For the second dataset which was our Amazon prime we wanted to predict the IMDB reviews scores based on other features such as the Genre of the year of release of the movie. we first like

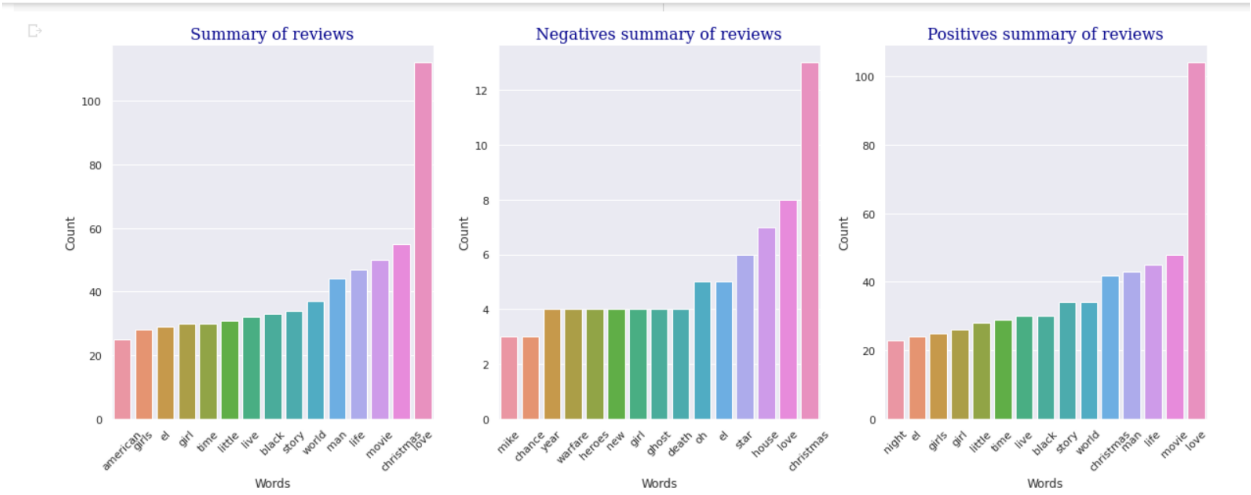
the other models we had to clean our dataset and checked we had data that was easy to work with in our modeling. We used a classified decision tree and we got accuracy of 57% on the test data using Gini. See our trees below, where we used different samples to see which gave us the best results.



We then used Decision Tree Classifier Entropy to see what below we would get from this. We got an evaluation of 57%. This is a low accuracy for a model

Based on the two decision tree models, we can say that a classification model works, but the predictions we got are not very good and would be hard to give good predictions with. The predictions based on Decision Tree classifier Gini are slightly better than ones based on Decision Tree Classifier Entropy.

For the third dataset on Netflix, we used both decision tree and sentimental analysis. We first started cleaning our data and splitting into testing and training. The decision tree accuracy on this was only 36%. This was too low to use to predict the information. We wanted to apply a similar model for predictions of the review scores based on the description (also known as plot) of the movie. We did both Decision Tree classifier Gin and Decision Tree Classifier Entropy. We looked at ways to improve this accuracy and looked at increasing the spilt, but this does not add much of a difference to the model. We then tried the sentimental analysis on this dataset to try and get better results. We got 1546 unique words in titles. We can see below the positive and negative review summaries we got from the analysis



We can see the negative and positive of the different words in that dataset. We then looked at the different scores we got in different models. We can see higher results than we got in the decision tree. It looks like the SVC gives the best results with the count vectorizer

```
[344] scores=accurac_model1(vocab,netflix.Bad_or_Good,class_models1)
      report1(scores)

      Titles :
      Naive Bayes : 0.8723421948161605
      SVC : 0.9012597122363516
      Complement Naive Bayes : 0.5600096802531365
      Perceptron : 0.8556096427261639
      Kneighbors : 0.8182018609966824
      ------------------------------
      Reviews :
      Naive Bayes : 0.8723421948161605
      SVC : 0.9012597122363516
      Complement Naive Bayes : 0.5600096802531365
      Perceptron : 0.8556096427261639
      Kneighbors : 0.8182018609966824
      ------------------------------
```

The classification with Tf-idf showed us that the Naive Bayes was the best model to use with the sentimental analysis.

```
Titles :
Naive Bayes : 0.9027062064495219
SVC : 0.899607032895803
Complement Naive Bayes : 0.5577365265375398
Perceptron : 0.8473458195805508
Kneighbors : 0.8204758675980178
------------------------------
Reviews :
Naive Bayes : 0.9027062064495219
SVC : 0.899607032895803
Complement Naive Bayes : 0.5577365265375398
Perceptron : 0.8473458195805508
Kneighbors : 0.8204758675980178
------------------------------
```

- Analytics Vidhya, 2021. Gaussian Distribution Transform

2.5.5 -Assessing model(s)

We look at each of the models we used in both sentimental analysis and decision tree and we can see that the sentimental analysis worked best in the Netflix dataset. Since our dataset had a lot of text this meant this model could analyze it and give us back the results.

The decision tree in both datasets gave us lower accuracy than we expected. We had hoped to get accuracy of 70% in these models to have good predictions.

2.6 -Phase 5: Evaluation

2.6.1 -Evaluating results

The Evaluation process determines how well a model is performing since we used supervised learning models, the data that was used for the training and testing was labelled. We cleaned each of our datasets before using them to make sure we got good predictions. We had to make sure not to overfit or underfit the model. We felt we got our best results when we used both sentimental analysis and decision tree and we can see that the sentimental analysis worked best in the Netflix dataset.

2.6.2 -Reviewing the process and determining the next steps

We have looked at all the models we used, and we felt the algorithm that worked the best was the sentimental analysis when we compared both models on our last dataset, we had hoped to receive better results with the sentimental analysis in the first dataset. This dataset was very fitted already and didn't give us the negative results we would have hoped to receive

We will present our findings and results during the deployment stage of the project. We now will look at the best ways to show our findings.

2.7 -Phase 6: Deployment

2.7.1 -Deployment materials used and rationale for selection

The road we have taken has brought us to choose three datasets and the team has decided to present the project as a whole, rather than separating and presenting the datasets one at a time.
This report is only part of the final delivery for this project.
A team has created a PowerPoint presentation that presents the facts in an orderly and simple manner. A voice over from both team members accompanies the presentation to make the presentation clear to understand from all sides of the table.
The presentation aims to be suited for and presentable to everyone from shareholders to employees, competitors, viewers and public alike, as well as the lecturers and the academic staff of the faculty.
A separate poster will accompany the PowerPoint presentation. The poster will include the best representations and graphs for the models selected for deployment. The poster also represents the 3 businesses, and the same three datasets that have been used throughout this project. The poster visualizations may, however, differ from this report. Although the representations and graphs in the presentation follow the poster, it may not follow the report's layout and detail, for space and clarity reasons.
The decision to make this material as accessible as possible was taken taking in consideration that the results of the projects may interest or possibly even influence the mindset of some of the viewers/users of the services.
.

2.7.2 -Rationale for color scheme for deployment

Visualizations play an important role in any project. In data analytics that role is amplified, someone viewing the visualizations must be able to understand what they convey, as the data or table they are based on May not. The team feels that the choices below were somewhat influenced by its graphical design interests, the color theory, and demographics that the team had learned on the course.
For the presentation, we opted for a vibrant and colorful design structure. This was purposely chosen because of its appeal to mostly young users, but also because the team wants this presentation to stand out and to be remembered, in a word, the team made this choice for its impact.
The team feels that the colorful and vibrant design in the presentation also balances the visualizations well. While some text was essential to present the backgrounds of each business, text was then limited, making space for the visualizations.
Based on the feedback from previous phases of the project, the graphs and visualizations were maximized where possible. A legend was added on the graphs in order to show exactly what was being represented, which would help all ages to understand what is being presented.

For the main Poster, the main color blue was kept as a background. in order for the poster to keep a certain symmetry, the poster was divided into three sections, the left and right sections being smaller and the middle main panel to be large as this panel was to contain most of the visualizations. The first panel holds the introduction and abstract, the data information as well as the two research questions. The main panel shows some of the results of the data exploration. For clarity purposes the graphs shown in the data exploration do not show the full extent of all the datasets. The team feels that similar results were produced for all datasets during the exploration stage so the full extent of the graphs would have been repetitive. Then the models chosen were also represented side by side, for the viewers to get a grasp of how each model works and looks. The last panel shows the results of the modeling stage and evaluates the results. The team kept the feel of fun and vibrant colors with a rainbow color scheme throughout the project code and throughout the visualizations used. Legends and figure numbers were added for clarity under all visualizations.

3.0 -Conclusions

At this stage of the project, the team feels that the machine learning aspects and general coding of the task have been completed. The end result has emerged from the code, although work could still be done to evaluate the results further, and to visualize the results better.

With the creation of the interactive presentation, diligently commented on by both team members and posters created for the presentation, the team now feels that the task is complete.

The aim of the project was to look at the IMDb score in order to present results; and the team believes that the results produced in the report answer some of our intended questions.

The team feels that question one was indeed answered successfully, and the models show just how well the titles and plots can predict the scores. We were not able however to answer question two. We found that some of the needed data for that question was not present in the datasets such as the revenues per movie versus score obtained would be beneficial. Some interesting information has however come out of the process such as the viewer age per score for example.

Both Netflix and Amazon as businesses should be interested in this project, as it has value for both companies. Each company should be able to find, within the details and data of this project, areas of weakness and areas of strength. IMDb being a subsidiary of Amazon, Amazon may have a bigger interest in the contents of this report. The team has an interest in digital marketing and project management (after studying for summer courses at CCT) and would also be able to discuss the possible potential of the research of this project with shareholders should it be required.

4.0 -Appendix

4.1 -Appendix 1 References

Amazon.co.uk. 2021. Amazon.co.uk: Amazon Prime. [online] Available at: <https://www.amazon.co.uk/amazonprime/261-0172314-3910223?_encoding=UTF8&primeCampaignId=prime_assoc_ft&tag=iefinder-21> [Accessed 3 November 2021].

Ayele, W., 2020. Adapting CRISP-DM for Idea Mining. International Journal of Advanced Computer Science and Applications, 11(6). [Accessed 3 November 2021]

S. Brown, M., 2020. Phase 1 Of The CRISP-DM Process Model: Business Understanding - Dummies. [online] dummies. Available at: <https://www.dummies.com/programming/big-data/phase-1-of-the-crisp-dm-process-model-business-understanding/> [Accessed 3 November 2021].

S. Brown, M., 2020. Phase 2 Of The CRISP-DM Process Model: Data Understanding - Dummies. [online] dummies. Available at: <https://www.dummies.com/programming/big-data/phase-2-of-the-crisp-dm-process-model-data-understanding/> [Accessed 3 November 2021].

S. Brown, M., 2020. Phase 3 Of The CRISP-DM Process Model: Data Preparation - Dummies. [online] dummies. Available at: <https://www.dummies.com/programming/big-data/phase-3-of-the-crisp-dm-process-model-data-preparation/> [Accessed 3 November 2021].

S. Brown, M., 2020. Phase 4 Of The CRISP-DM Process Model: Modeling - Dummies. [online] dummies. Available at: <https://www.dummies.com/programming/big-data/phase-4-of-the-crisp-dm-process-model-modeling/> [Accessed 3 November 2021].

S. Brown, M., 2020. Phase 5 Of The CRISP-DM Process Model: Evaluation - Dummies. [online] dummies. Available at: <https://www.dummies.com/programming/big-data/phase-5-of-the-crisp-dm-process-model-evaluation/> [Accessed 3 November 2021].

S. Brown, M., 2020. Phase 6 Of The CRISP-DM Process Model: Deployment - Dummies. [online] dummies. Available at: <https://www.dummies.com/programming/big-data/phase-6-of-the-crisp-dm-process-model-deployment/> [Accessed 3 November 2021].

Cionnaith, F., 2019. 'No way to avoid it': Everything you need to know about the new 'broadcasting charge'. [online] Irish Examiner. Available at: <https://www.irishexaminer.com/news/arid-30941563.html> [Accessed 4 November 2021].

Citizensinformation.ie. 2020. TV licenses. [online] Available at: <https://www.citizensinformation.ie/en/consumer/phone_internet_tv_and_postal_services/tv_licences.html#> [Accessed 4 November 2021].

Clarke, S. (2019) Netflix to open new EMEA headquarters in Amsterdam as part of European expansion, Variety.com. Variety. Available at: https://variety.com/2019/tv/news/netflix-new-emea-headquarters-amsterdam-european-expansion-1203386677/ [Accessed 3 November 2021].

Dan, M., 2016. Who Are Netflix's Main Competitors? [online] Investopedia. Available at: <https://www.investopedia.com/articles/markets/051215/who-are-netflixs-main-competitors-nflx.asp> [Accessed 3 November 2021].

DeepAI. 2021. Logistic Regression. [online] Available at: <https://deepai.org/machine-learning-glossary-and-terms/logistic-regression> [Accessed 12 October 2021].

Encyclopedia Britannica. 2021. IMDb | History, Features, & Facts. [online] Available at: <https://www.britannica.com/topic/IMDb> [Accessed 10 November 2021].

Enterprise Knowledge. 2021. Metadata Use Case: IMDB in Amazon Prime Video - Enterprise Knowledge. [online] Available at: <https://enterprise-knowledge.com/metadata-use-case-imdb-in-amazon-prime-video/> [Accessed 10 November 2021].

Global subscriber market share of SVOD services 2024 (no date) Statista.com. Available at: https://www.statista.com/statistics/1052803/global-svod-subs-market-share/ [Accessed 3 November 2021].

Hannam, L., 2020. Viewers midway through Line of Duty furious as it's pulled from Netflix without explanation. [online] Entertainment Daily. Available at: <https://www.entertainmentdaily.co.uk/tv/line-of-duty-taken-off-netflix-where-and-how-to-watch-it-now/#:~:text=Line%20of%20Duty%20has%20been,being%20taken%20off%20%40netflix. %E2%80%9D&text=%23lineofduty. %E2%80%9D, -Line%20Of%20Duty> [Accessed 3 November 2021].

Help.imdb.com. 2021. IMDb | Help. [online] Available at: <https://help.imdb.com/article/imdb/track-movies-tv/ratings-faq/G67Y87TFYYP6TWAV#> [Accessed 10 November 2021].

Hibler, J., 2012. Reed Hastings | Biography, Netflix, & Facts. [online] Encyclopedia Britannica. Available at: <https://www.britannica.com/biography/Reed-Hastings> [Accessed 4 November 2021].

Hosch, W., 2009. Netflix | Founders, History, Programming, & Facts. [online] Encyclopedia Britannica. Available at: <https://www.britannica.com/topic/Netflix-Inc> [Accessed 4 November 2021].

Hoff, T., 2017. Netflix: What Happens When You Press Play? - High Scalability -. [online] Highscalability.com. Available at: <http://highscalability.com/blog/2017/12/11/netflix-what-happens-when-you-press-play.html?currentPage=2#:~:text=Netflix%20to%20a%20completely%20different, it's%20own%20datacenters%20anymore%20either> [Accessed 24 October 2021].

Irish Examiner. 2021. eir teams up with Amazon Prime to offer customers access to the service as part of the TV package. [online] Available at: <https://www.irishexaminer.com/business/arid-30964002.html> [Accessed 14 October 2021].

Instagram.com. 2021. Login • Instagram. [online] Available at: <https://www.instagram.com/netflixuk/?hl=en> [Accessed 4 April 2021].

Linkedin.com. Available at: https://www.linkedin.com/company/netflix/ [Accessed 4 November 2021].

McKenna, K., 2021. Irish Media and Broadcasting Law Update. [online] Default. Available at: <https://www.matheson.com/insights/detail/irish-media-and-broadcasting-law-update---summer-2020> [Accessed 14 October 2021.

Medium. 2021. Unsupervised Sentiment Analysis. [online] Available at: <https://towardsdatascience.com/unsupervised-sentiment-analysis-a38bf1906483> [Accessed 10 November 2021].

Moore, A. (2019) Netflix's generic strategy, business model & intensive growth strategies, Rancord.org. Available at: https://www.rancord.org/netflix-business-model-generic-strategy-intensive-growth-strategies-competitive-advantage [Accessed 24 October 2021].

Netflix global revenue by region 2020 (no date) Statista.com. Available at: https://www.statista.com/statistics/1090098/netflix-global-revenue-by-region/ [Accessed 14 October 2021].

Netflix paid subscriber count by region 2020 (no date) Statista.com. Available at: https://www.statista.com/statistics/483112/netflix-subscribers/ [Accessed 4 April 2021].

Netflix Prize: Home (no date) Netflixprize.com. Available at: https://www.netflixprize.com/index.html [Accessed 14 October 2021].

Netflix prize: Review rules (no date) Netflixprize.com. Available at: https://www.netflixprize.com/rules.html [Accessed 24 October 2021].

Netflix - statistics & facts (no date) Statista.com. Available at: https://www.statista.com/topics/842/netflix/ [Accessed 24 October 2021].

Rivera, A. (2019) Netflix's mission statement & vision statement: A strategic analysis, Rancord.org. Available at: https://www.rancord.org/netflix-corporate-vision-statement-mission-statement-strategic-analysis [Accessed 24 October 2021].

Scikit-learn.org. 2021. 1.10. Decision Trees — scikit-learn 0.24.2 documentation. [online] Available at: <https://scikit-learn.org/stable/modules/tree.html> [Accessed 24 October 2021].

Siapera, E., Kirk, N. and Doyle, K., n.d. [online] Bai.ie. Available at: <https://www.bai.ie/en/media/sites/2/dlm_uploads/2019/12/Netflix-and-Binge-Final-Report.pdf> [Accessed 23 October 2021].

Sky. 2021. discovery+ help | Sky Help. [online] Available at: <https://www.sky.com/help/articles/discovery-plus> [Accessed 13 October l 2021].

Sky. 2021. What is Sky VIP? [online] Available at: <https://www.sky.com/pages/vip/what-is-sky-vip> [Accessed 3 November 2021].

Stewart, E., 2021. About Netflix - NetZero + Nature: Our Commitment to the Environment. [online] About Netflix. Available at: <https://about.netflix.com/en/news/net-zero-nature-our-climate-commitment> [Accessed 4 October 2021].

The Irish Times. 2021. Can Amazon Prime Video ever be lord of the streamers? [online] Available at: <https://www.irishtimes.com/business/media-and-marketing/can-amazon-prime-video-ever-be-lord-of-the-streamers-1.4541306> [Accessed 14 October 2021].

Tutorialspoint.com. 2021. KNN Algorithm - Finding Nearest Neighbors - Tutorialspoint. [online] Available at: <https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_knn_algorithm_finding_nearest_neighbors.htm> [Accessed 14 October 2021]

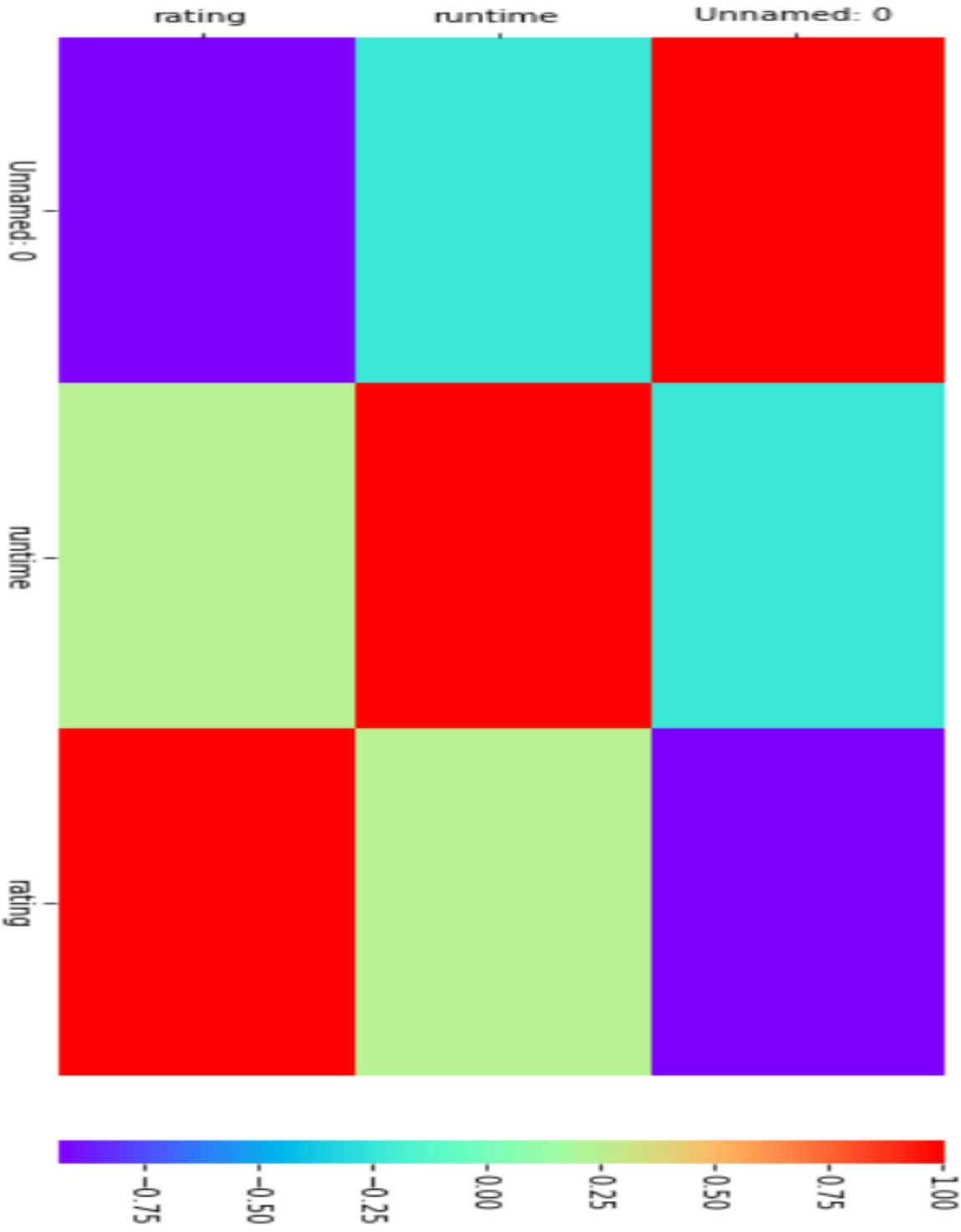
VentureBeat. 2021. [online] Available at: <https://venturebeat.com/2015/10/30/25-years-of-imdb-the-worlds-biggest-online-movie-database> [Accessed 10 November 2021].

Webdevelopersnotes.com. 2021. Amazon uses IMDb as an advertising platform. [online] Available at: <https://www.webdevelopersnotes.com/amazon-uses-imdb-as-advertising-platform> [Accessed 10 November 2021].

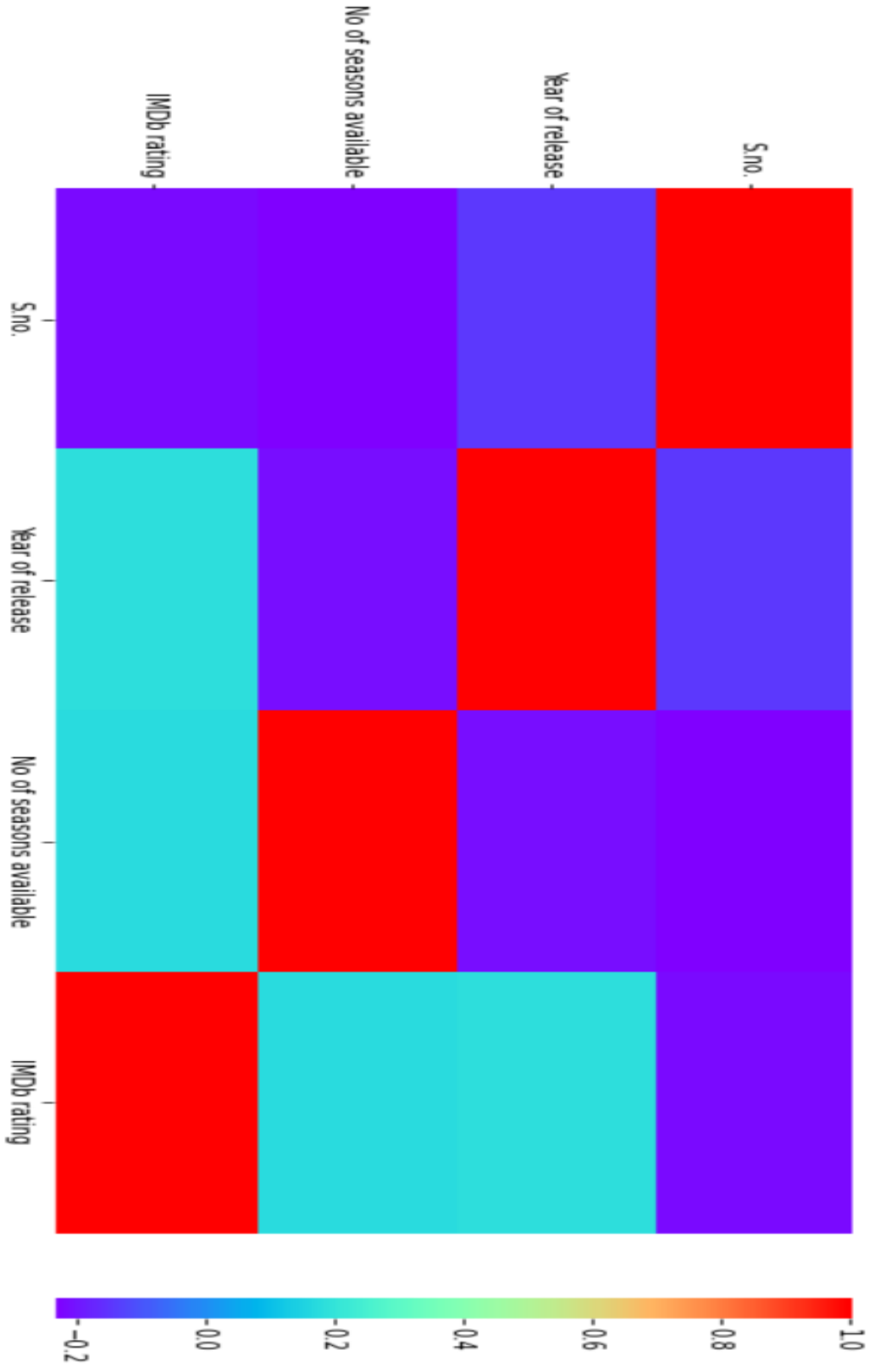
WIRE, B., 2021. Liberty Global Announces Prime Video Partnership with Amazon. [online] Businesswire.com. Available at: <https://www.businesswire.com/news/home/20190423005470/en/Liberty-Global-Announces-Prime-Video-Partnership-with-Amazon> [Accessed 14 May 2021].

4.2 -Appendix 2 List of Figures

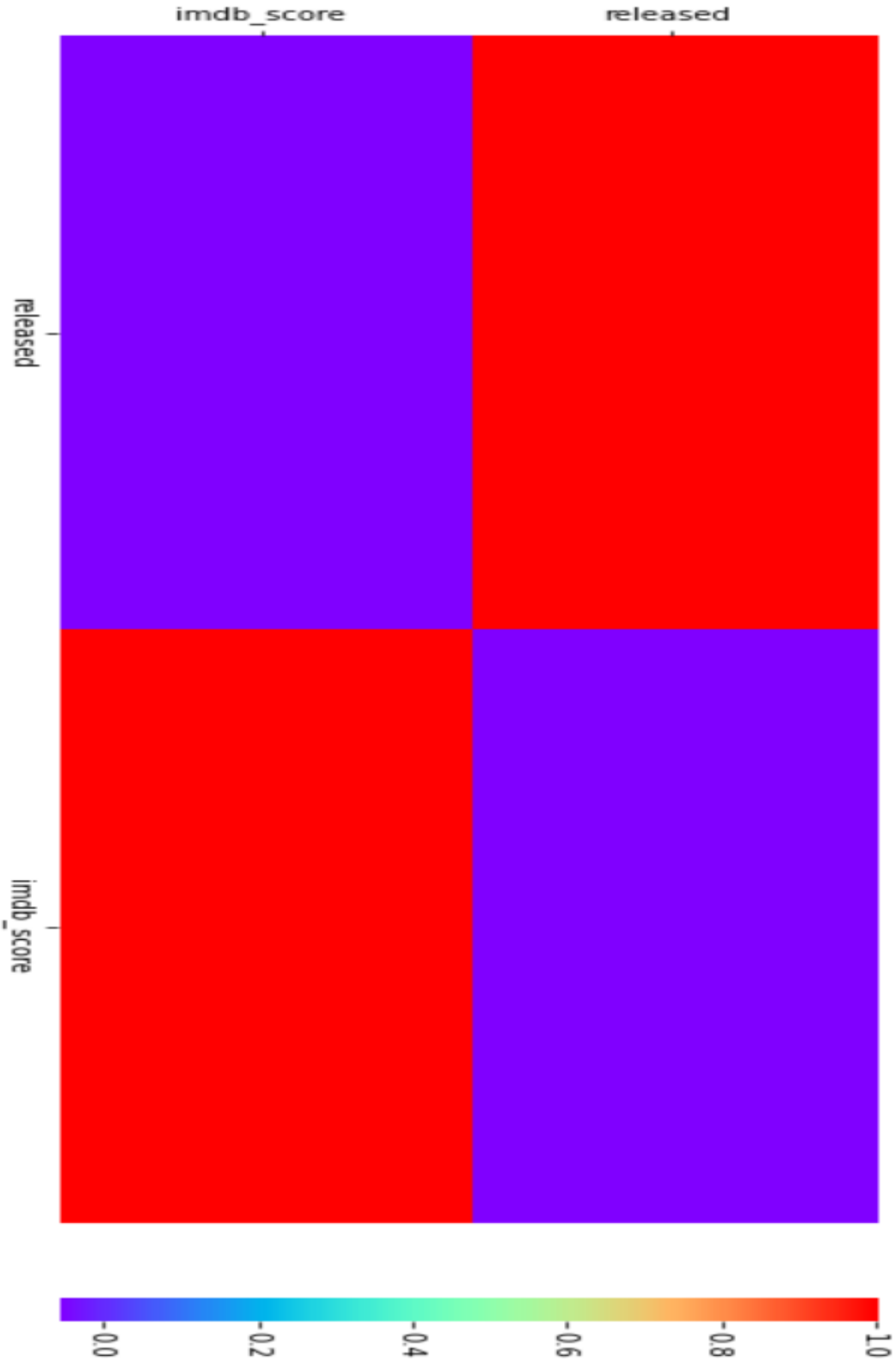
(Code 1 movie. corr heatmap)



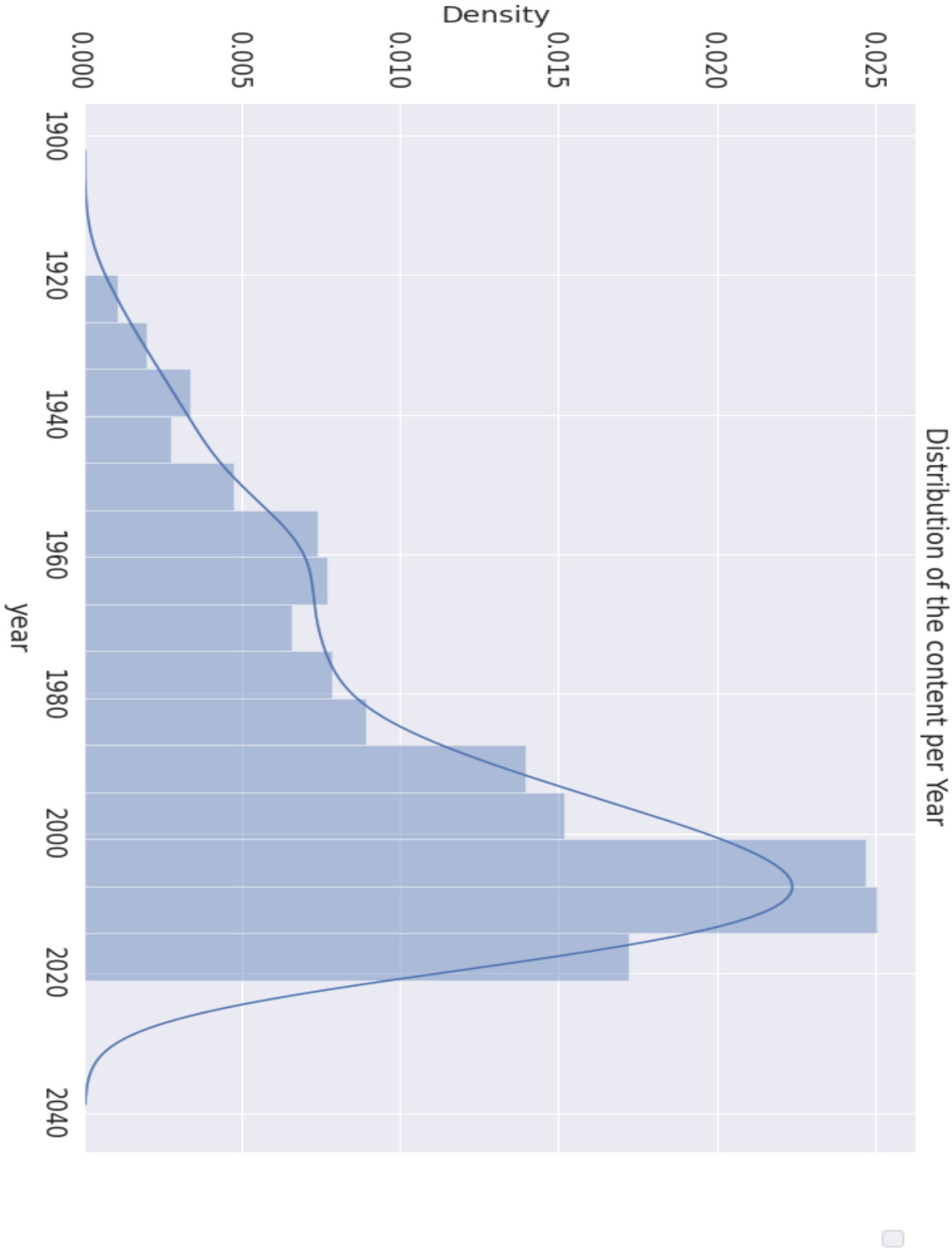
(Code 2 prime2.corr heatmap)



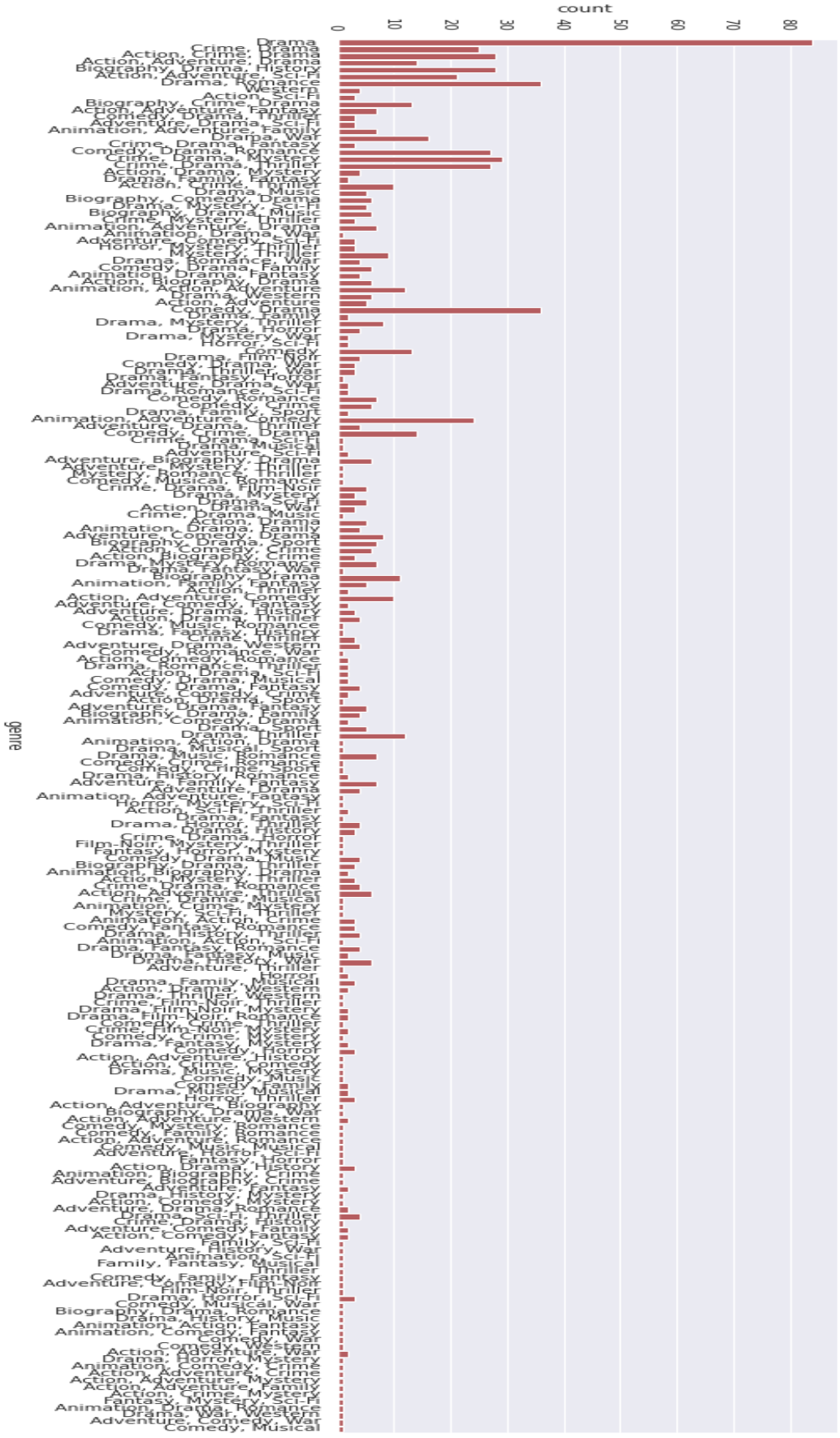
(Code 3 Netflix. corr heatmap)



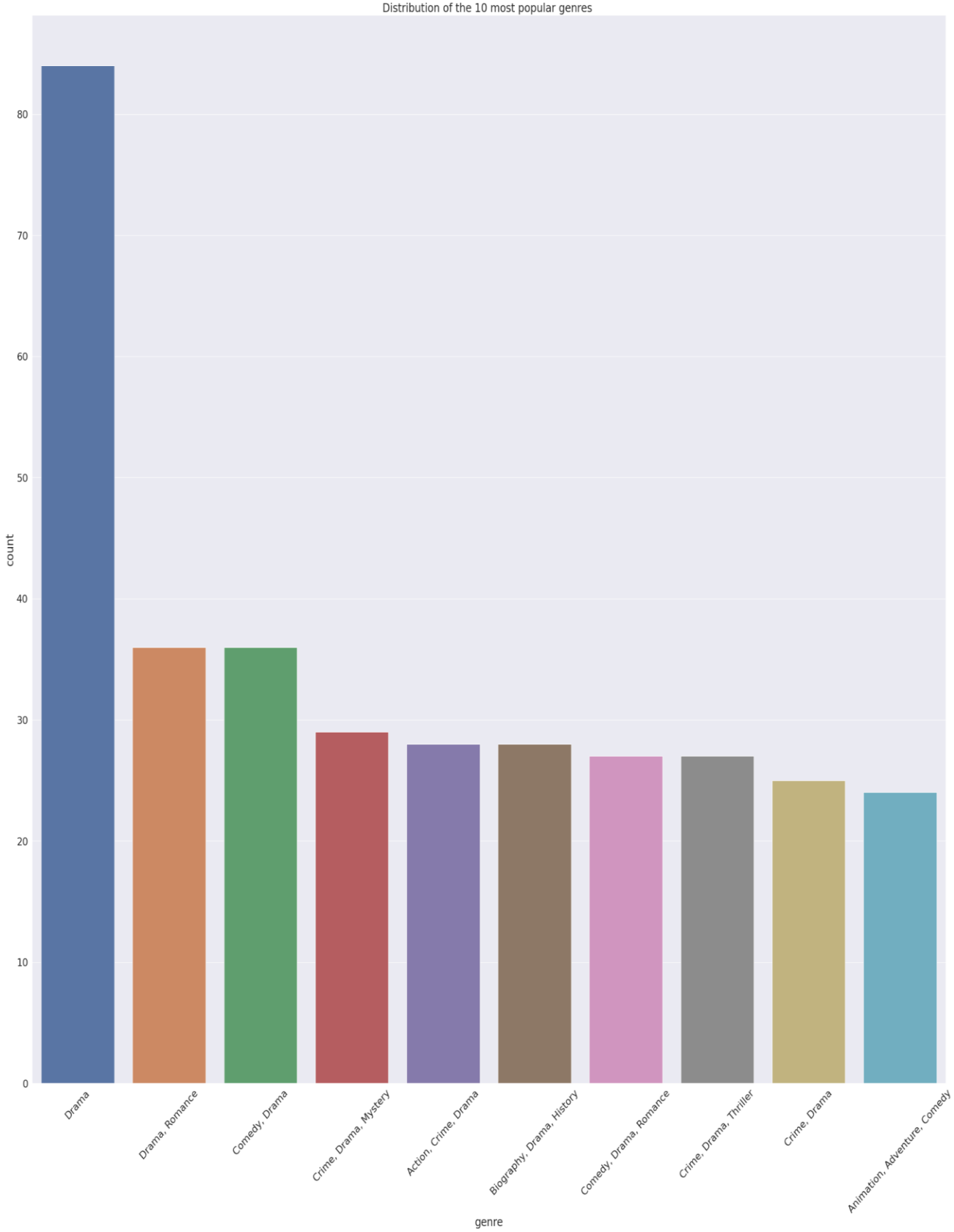
(Code 4 movies: Distribution of the content per year)



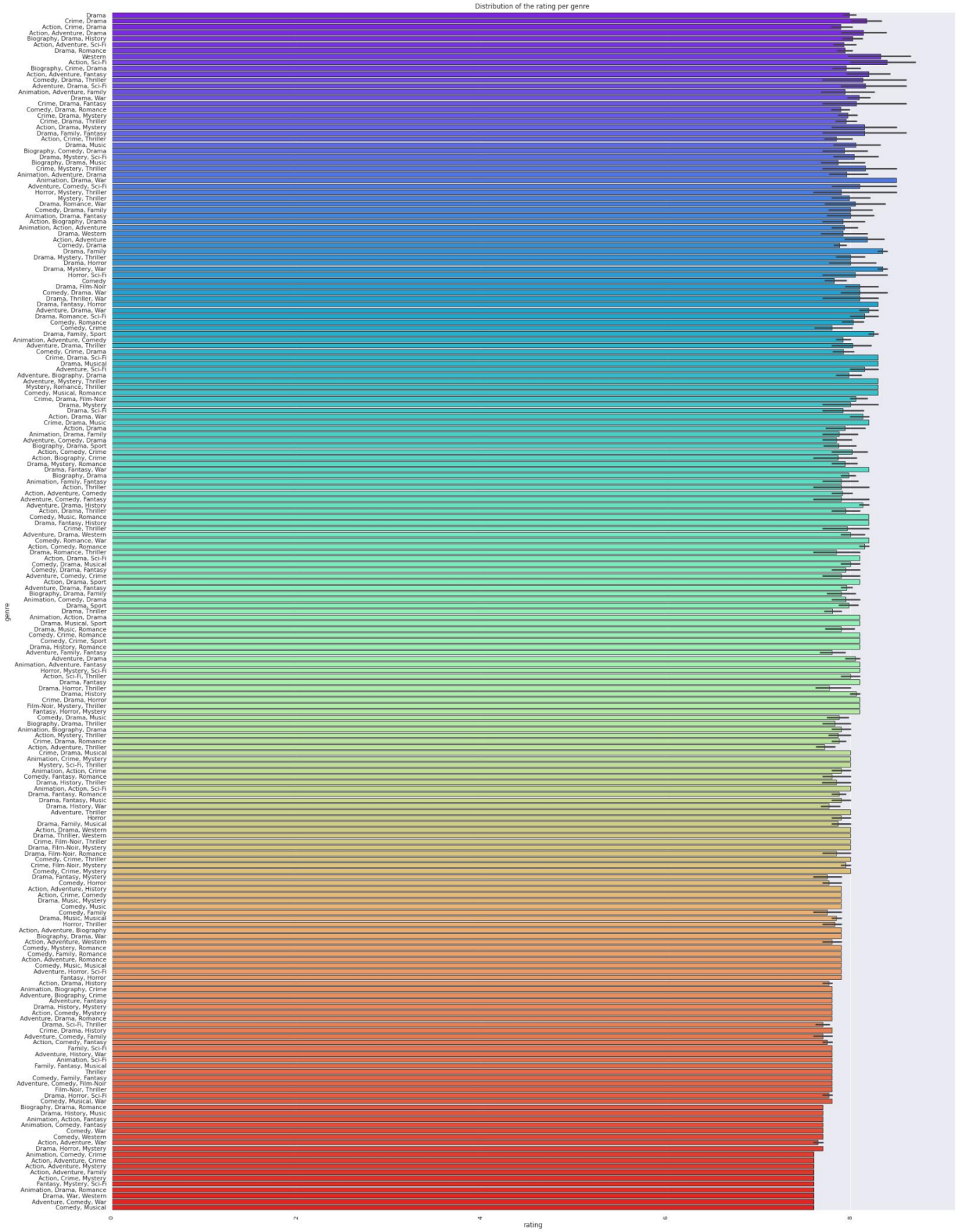
(Code 5 Movie genres)



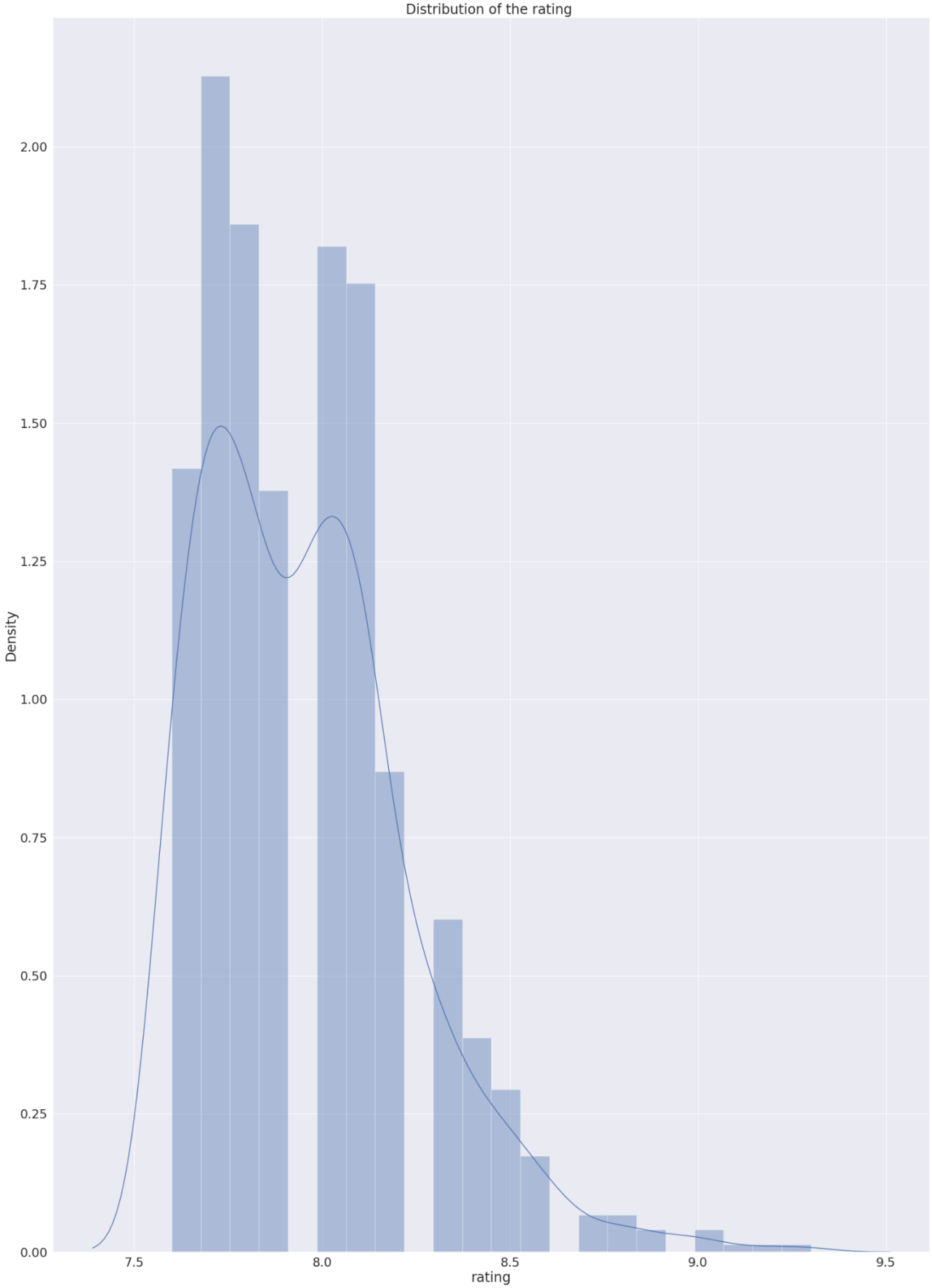
(Code 6: Distribution of the 10 most popular movie genres)



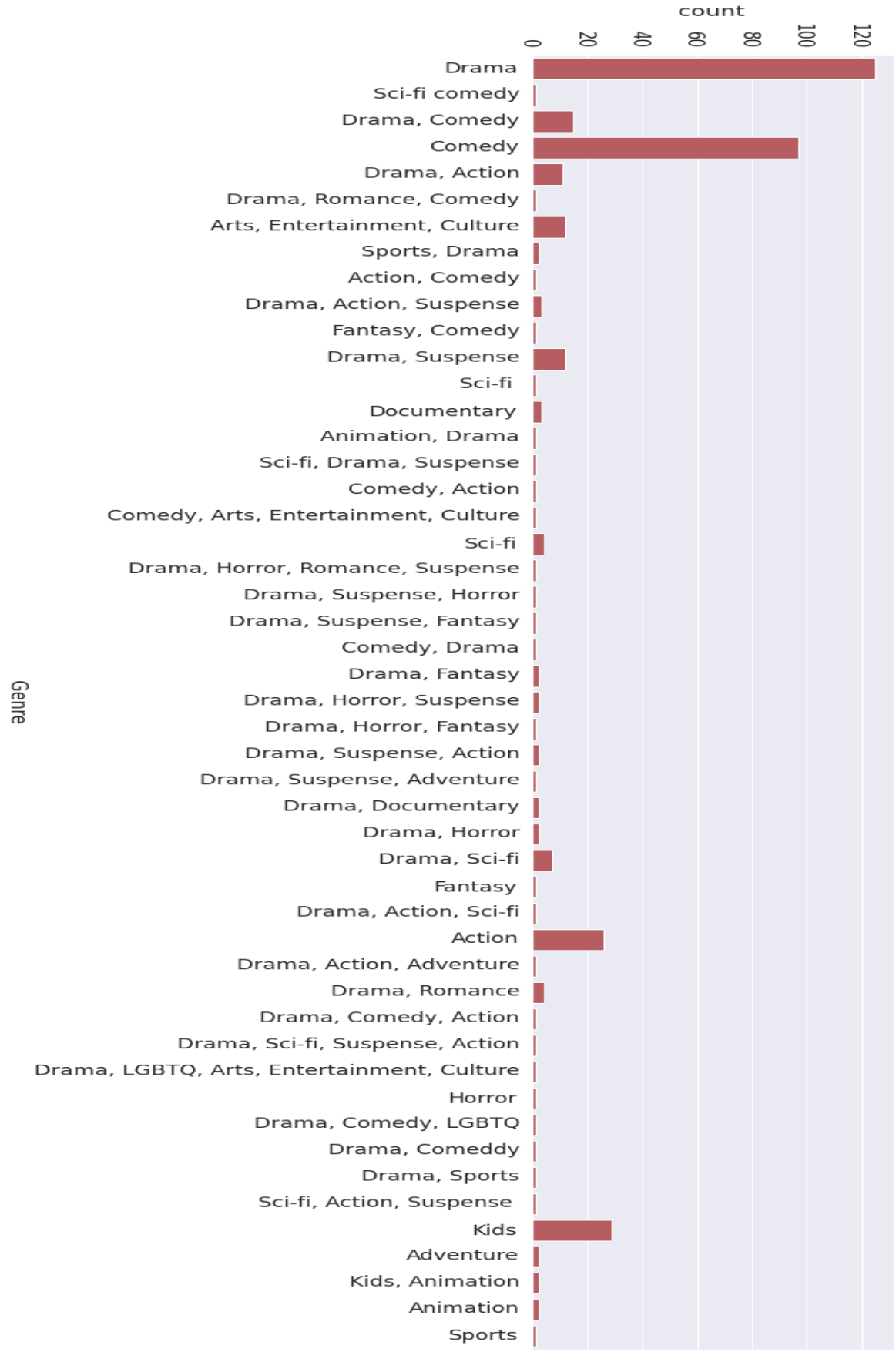
(Code 7 Rating per Genre)



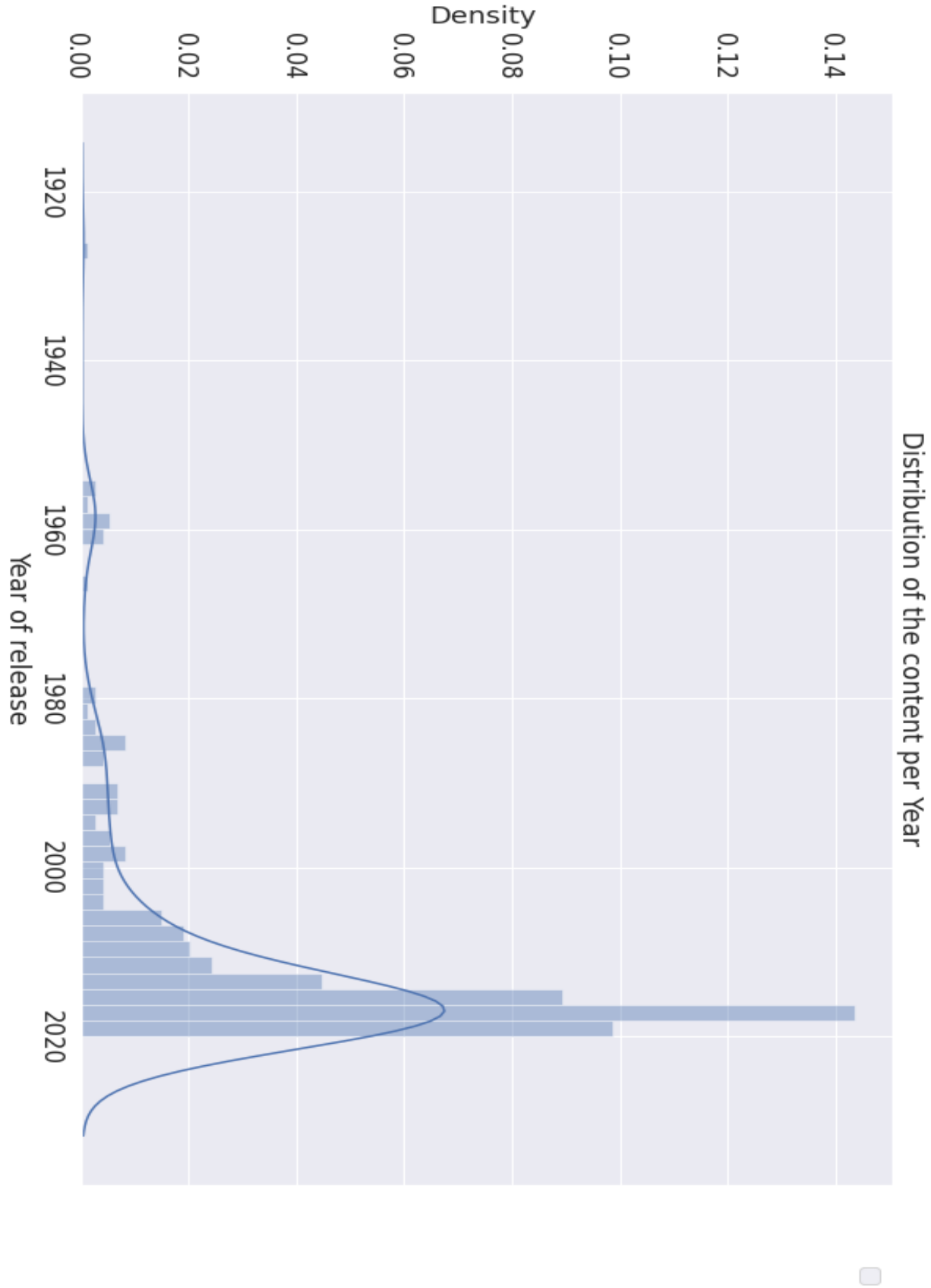
(Code 8; Distribution of the rating)



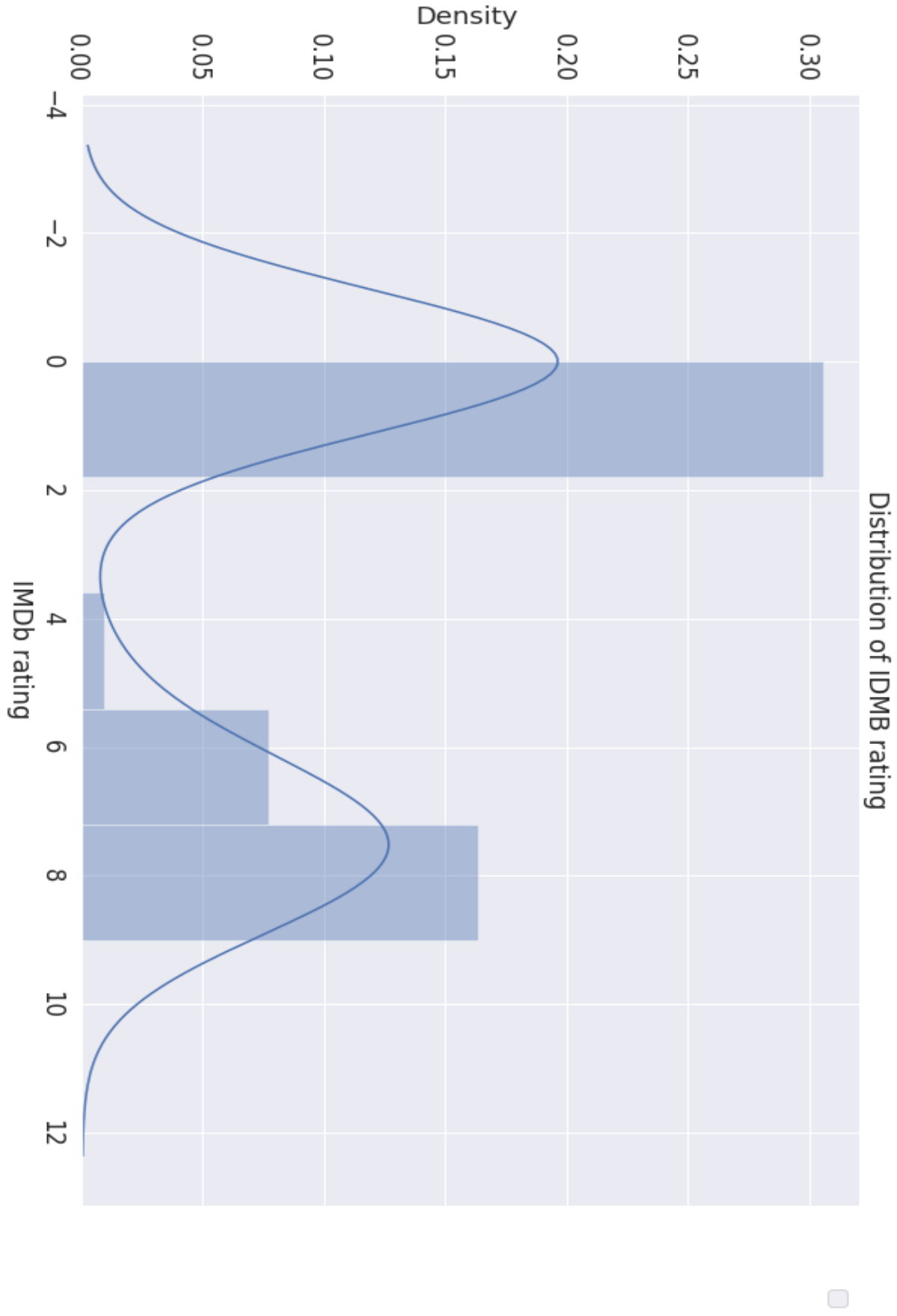
(Code 9 prime2: movie genres)



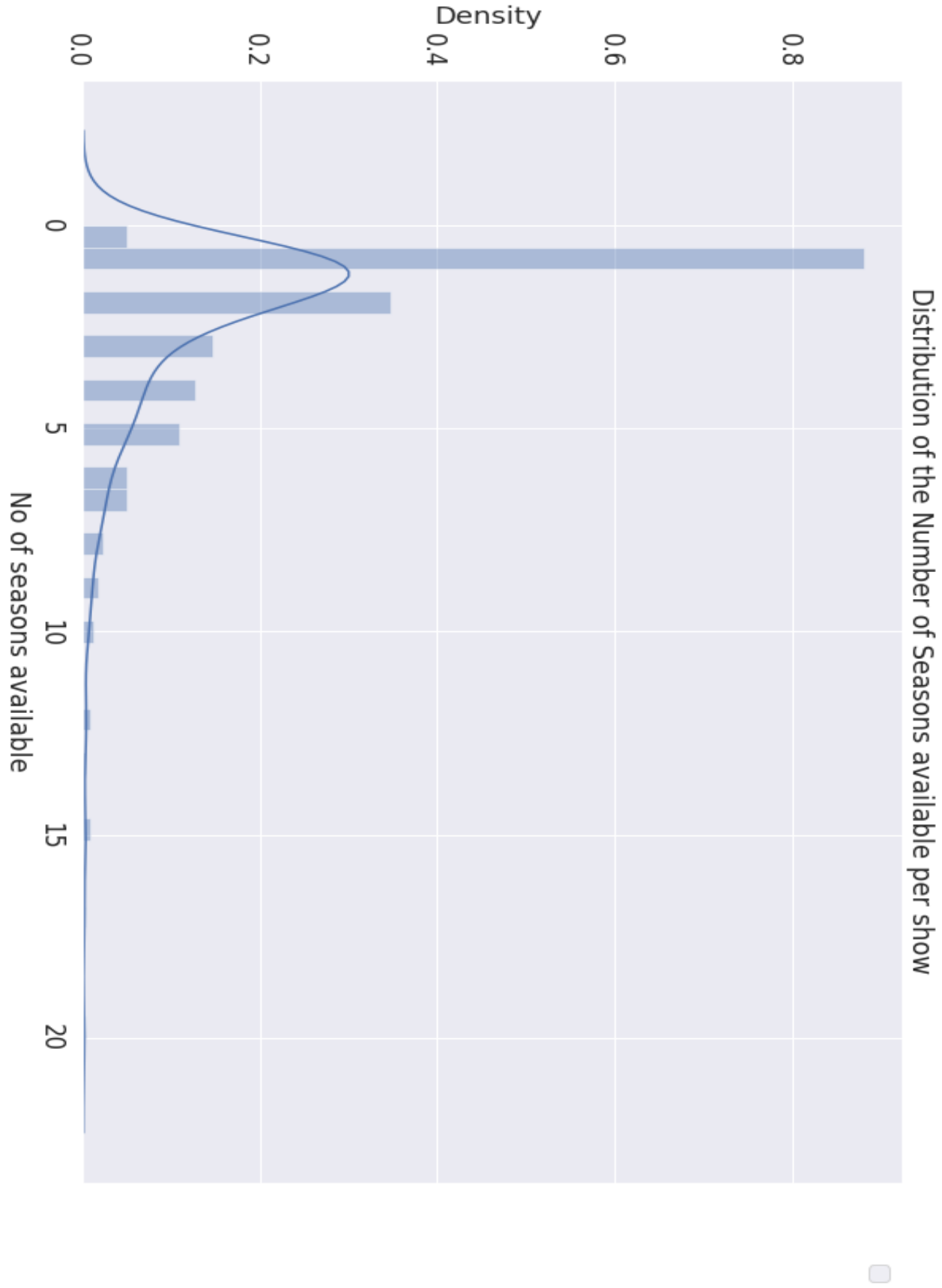
(Code 10 prime2: content per year)



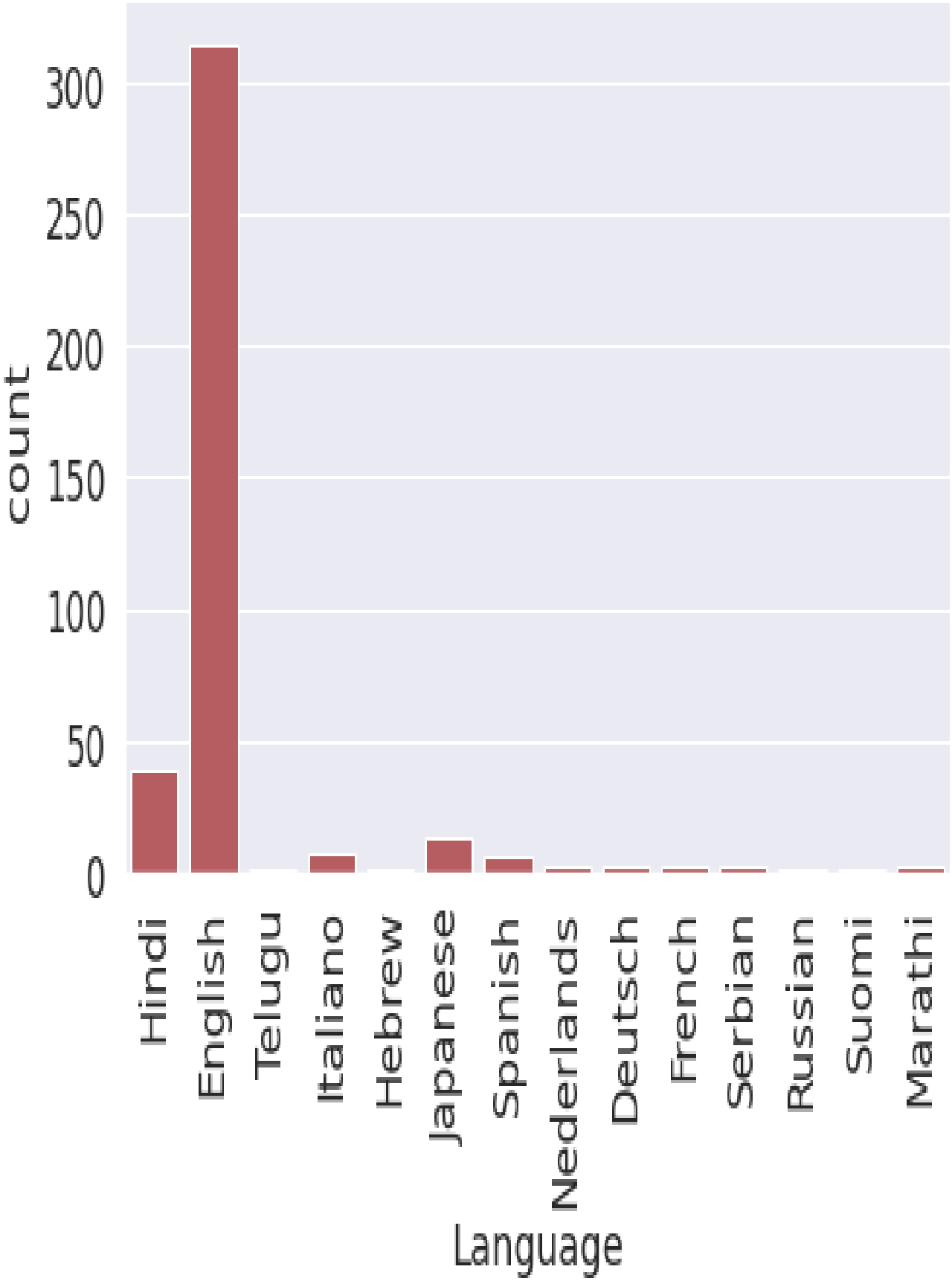
(Code 11 prime2: IMDb rating)



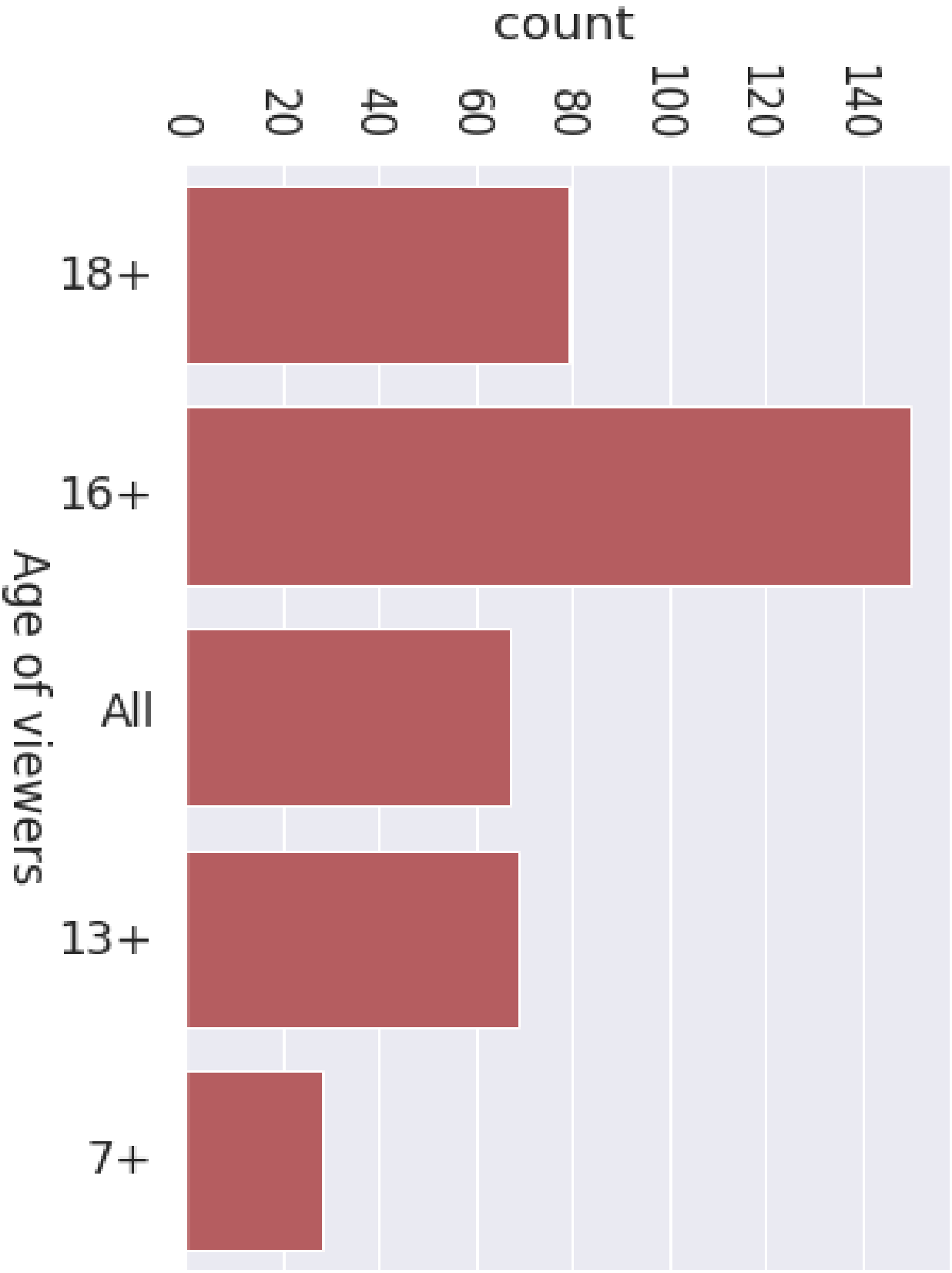
(Code 12 prime2: number of seasons available)



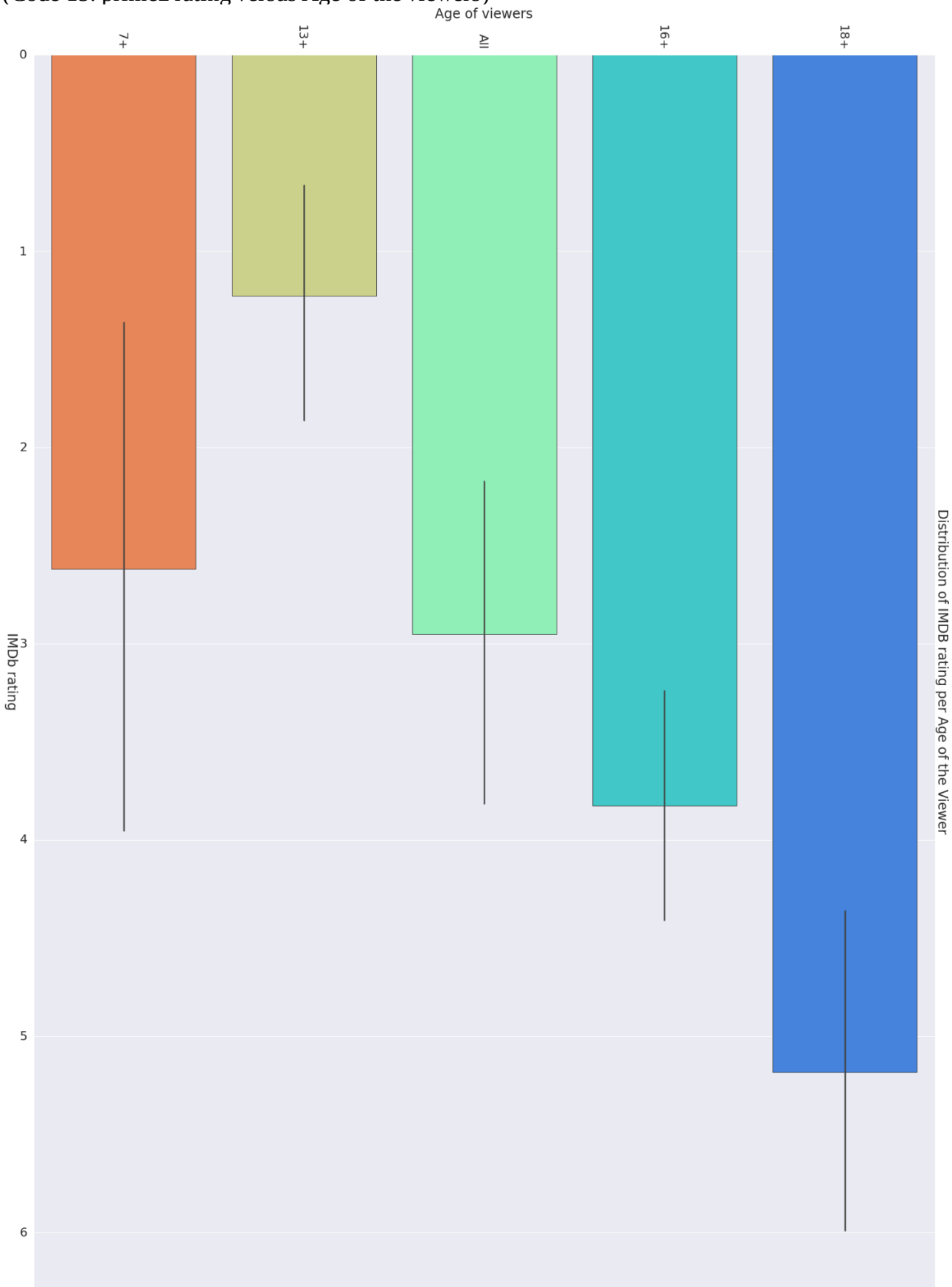
(Code 13 Prime2: Language)



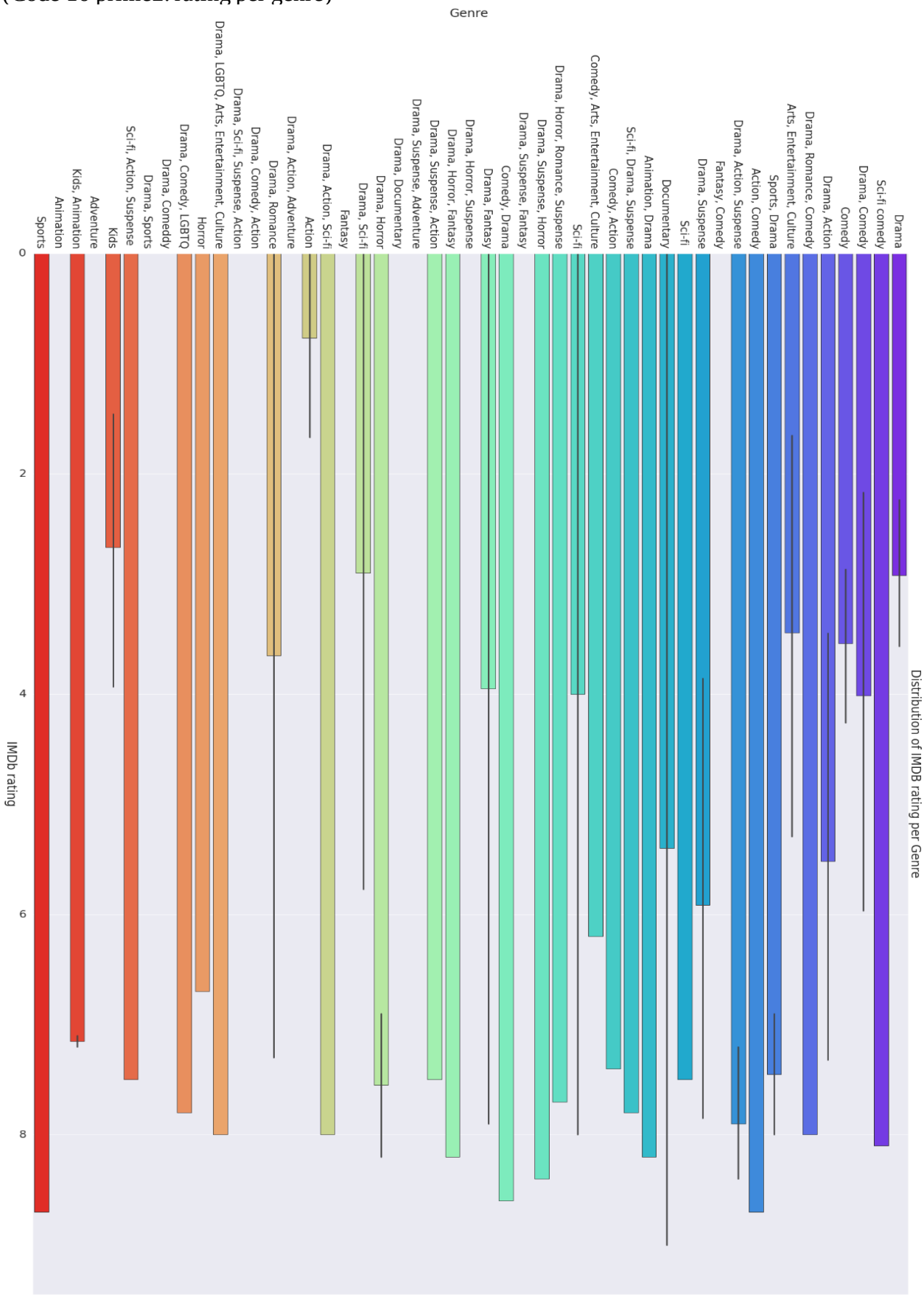
(Code 14 Prime2: Age of the viewers)



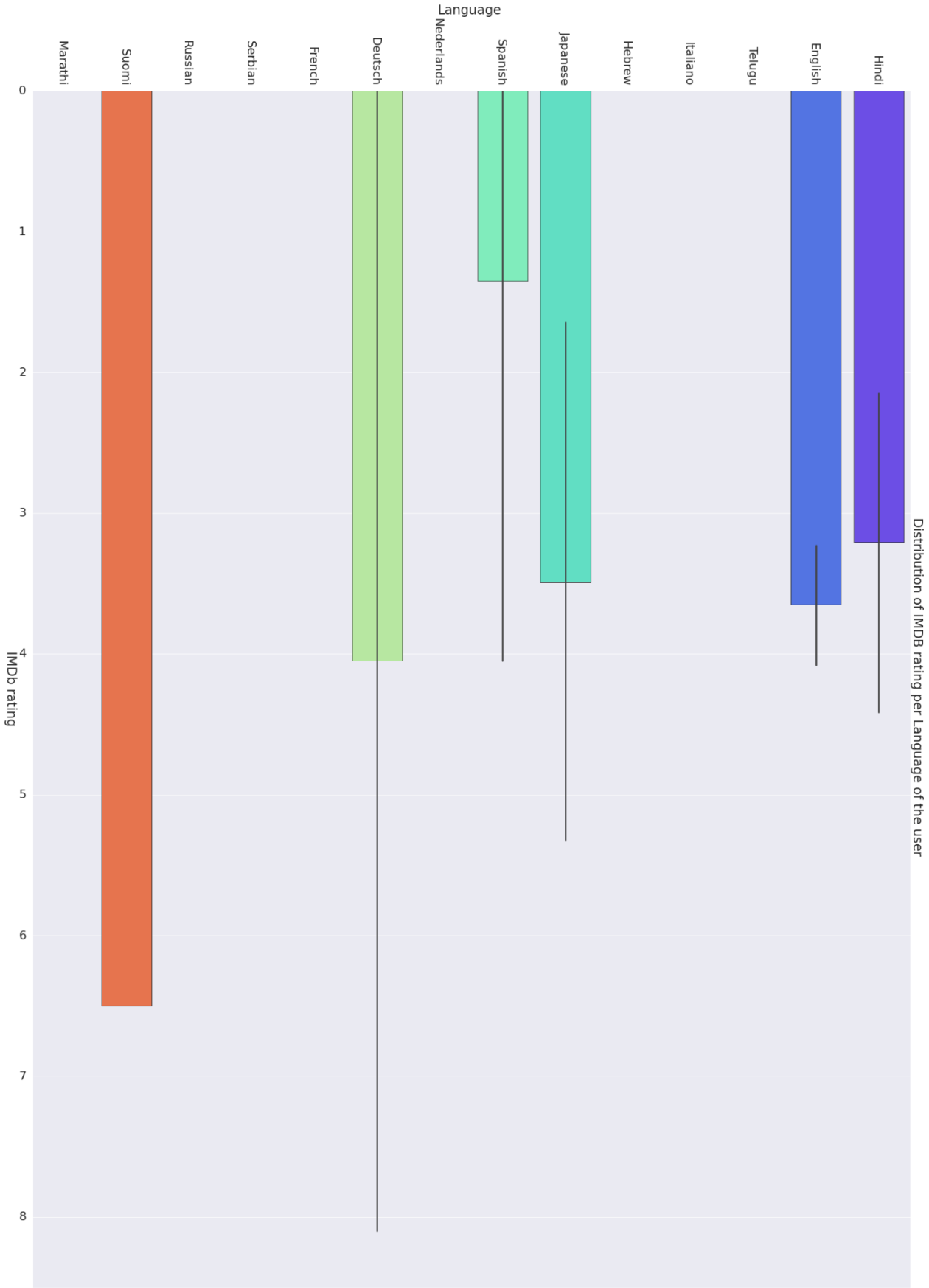
(Code 15: prime2 rating versus Age of the viewers)



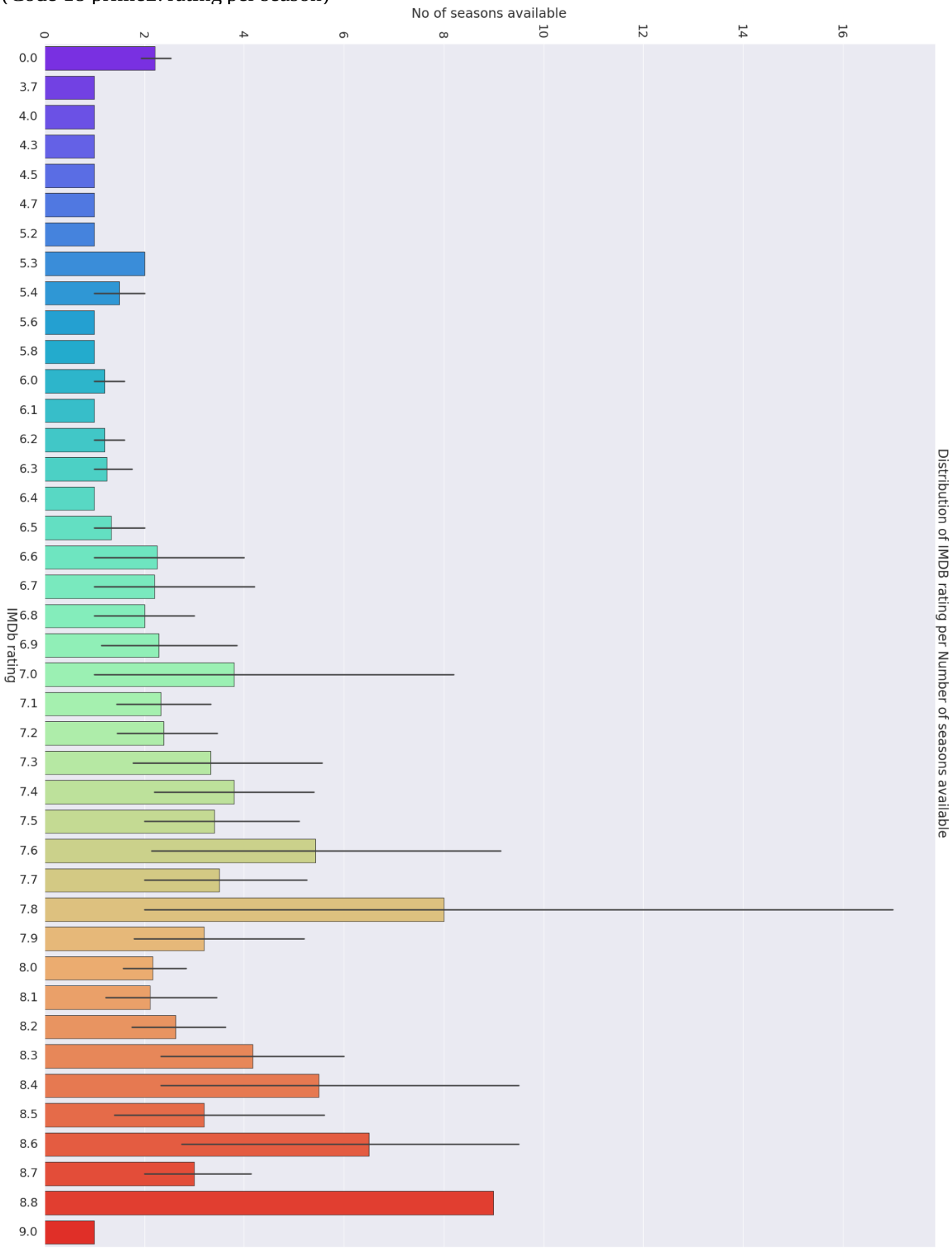
(Code 16 prime2: rating per genre)



(Code 17 prime2: rating per language of the viewer)

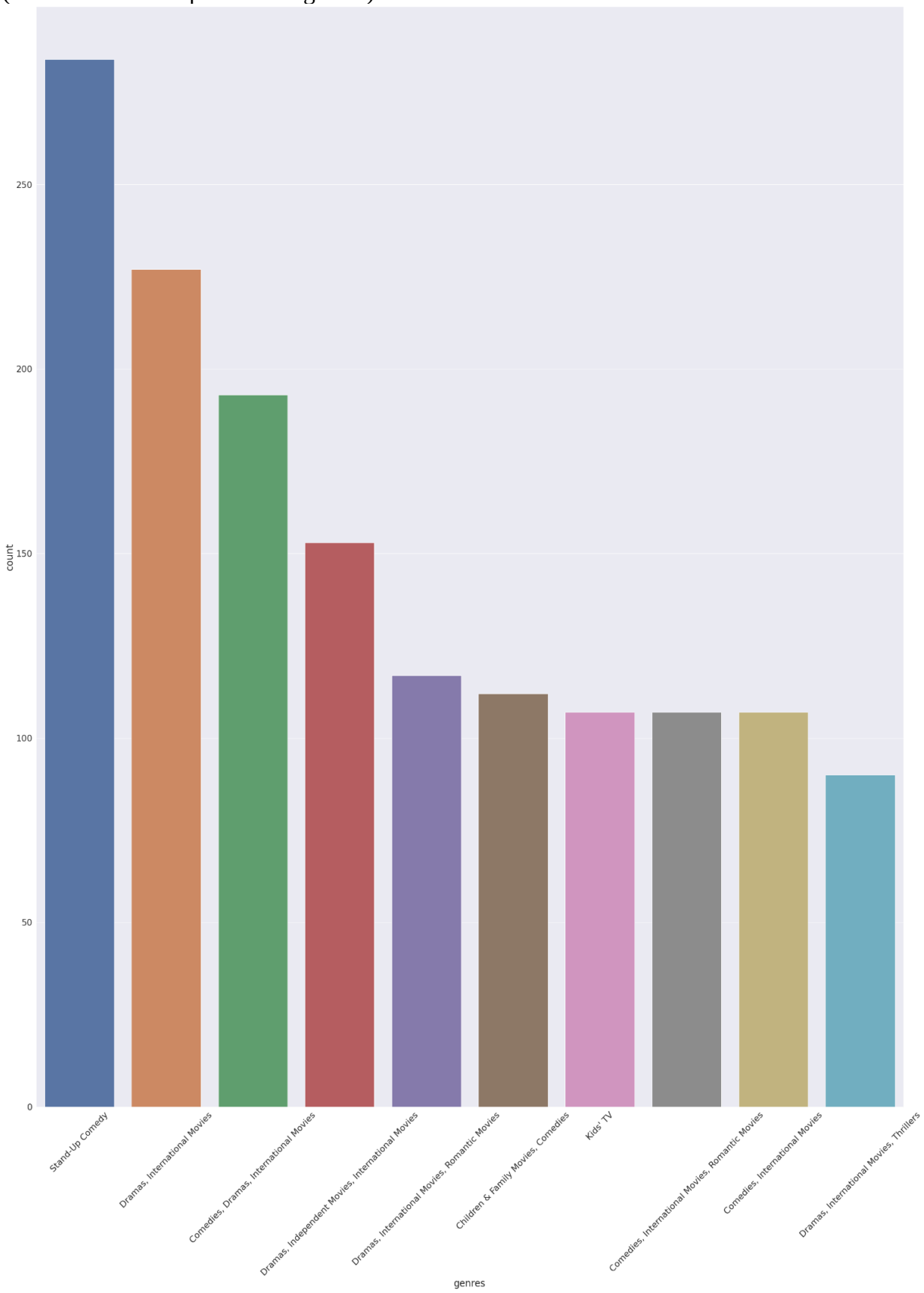


(Code 18 prime2: rating per season)

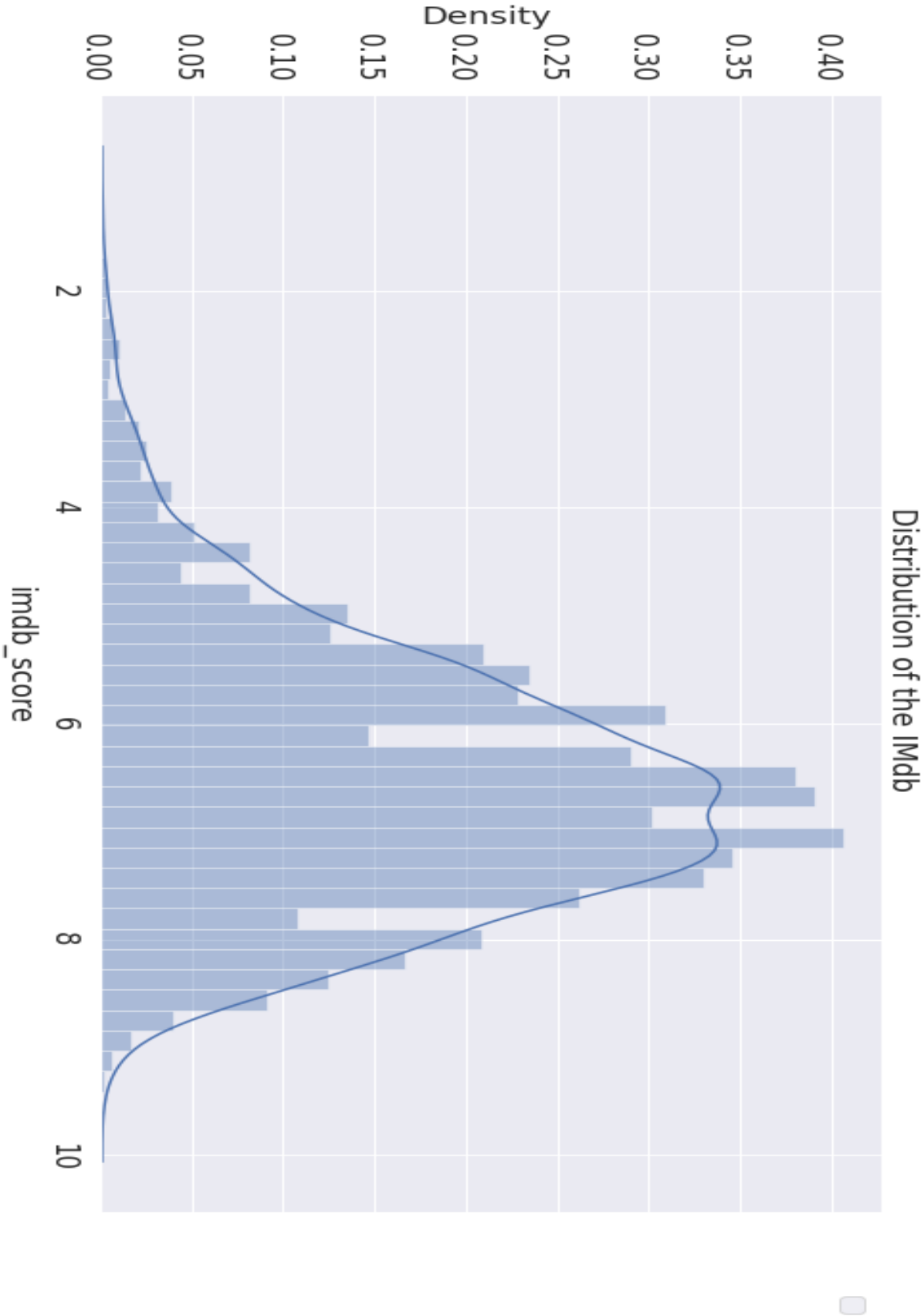


(Code 19 Netflix genres)

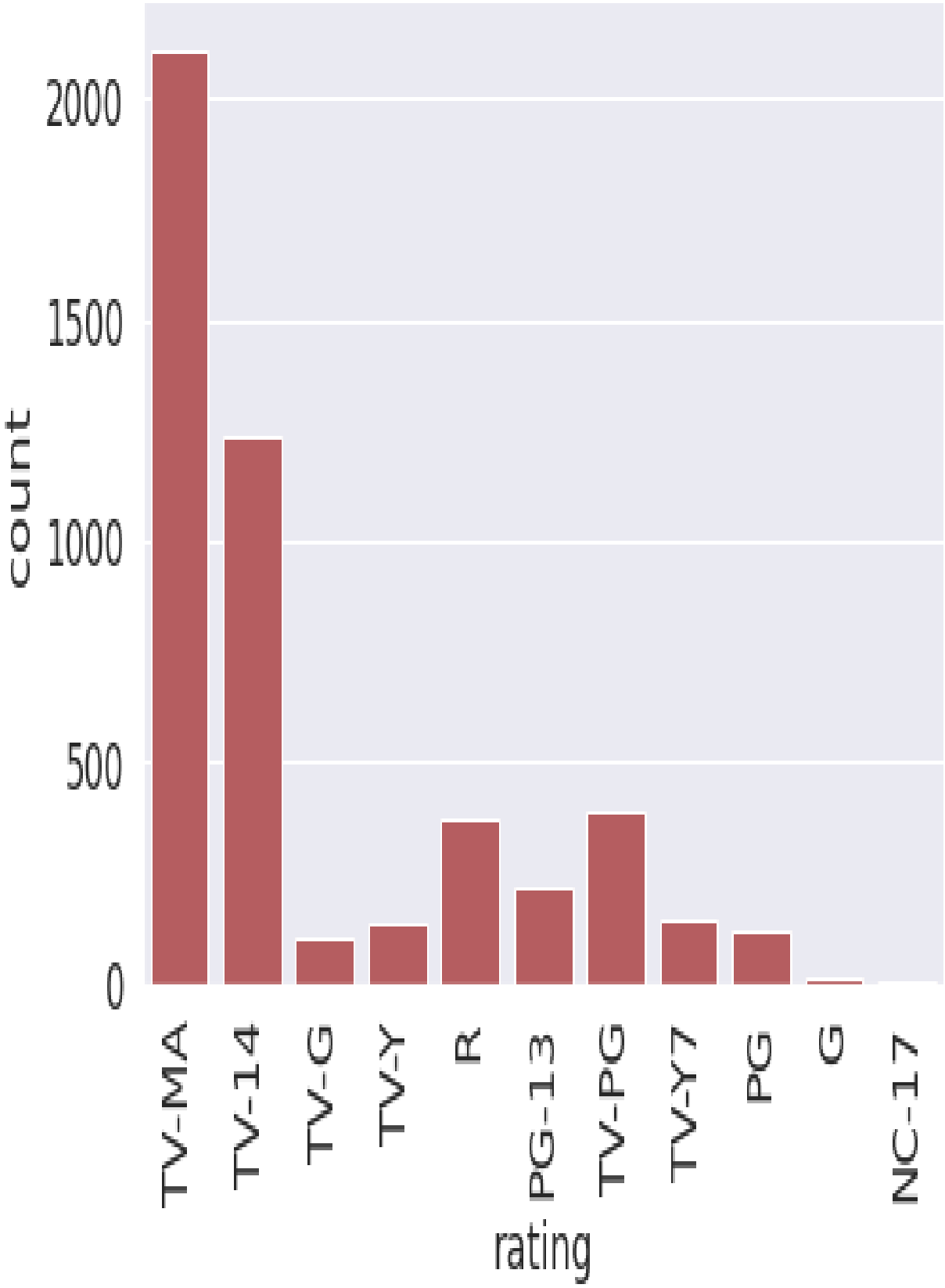
(Code 20 Netflix top 10 movie genres)



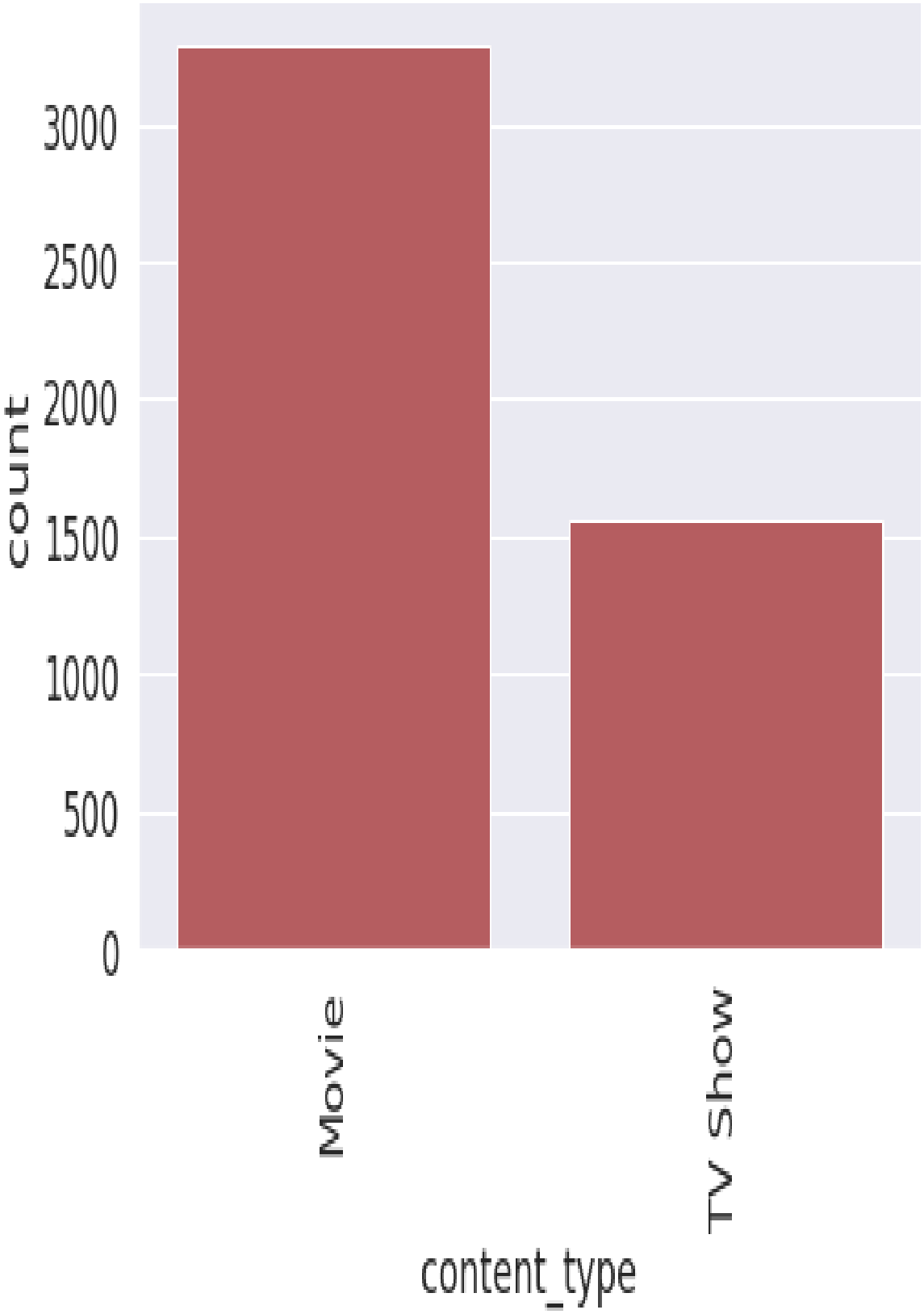
(Code 21Netflix: Distribution of the IMDb)



(Code 22 Netflix: movie age rating type)



(Code 23 Netflix: content type)



(Code 24 Netflix: Countries of production for the movies)