Netflix study page from students

<https://community.scaler.com/t/business-case-netflix-data-exploration-and-visualisation/17562/12>

**Mindset**

1. Evaluation will be kept lenient, so make sure you attempt this case study.
2. Read the question carefully and try to understand what exactly is being asked.
3. Brainstorm a little. If you’re getting an error, remember that Google is your best friend.
4. You can watch the lecture recordings or go through your lecture notes once again if you feel like you’re getting confused over some specific topics.
5. Discuss your problems with your peers. Make use of the Slack channel and WhatsApp group.
6. Only if you think that there’s a major issue, you can reach out to your Instructor via Slack or Email.
7. There is no right or wrong answer. We have to get used to dealing with uncertainty in business. This is exactly the skill we want to develop.

**About NETFLIX**

Netflix is one of the most popular media and video streaming platforms. They have over 10000 movies or tv shows available on their platform, as of mid-2021, they have over 222M Subscribers globally. This tabular dataset consists of listings of all the movies and tv shows available on Netflix, along with details such as - cast, directors, ratings, release year, duration, etc.

**Business Problem**

Analyze the data and generate insights that could help Netflix in deciding which type of shows/movies to produce and how they can grow the business in different countries

**Dataset**

Link: [Dataset\_link](https://d2beiqkhq929f0.cloudfront.net/public_assets/assets/000/000/940/original/netflix.csv" \t "_blank)

*(After clicking on the above link, you can download the files by right-clicking on the page and clicking on "Save As", then naming the file as per your wish, with .csv as the extension.)*

The dataset provided to you consists of a list of all the TV shows/movies available on Netflix:

**Show\_id:** Unique ID for every Movie / Tv Show  
**Type:** Identifier - A Movie or TV Show  
**Title:** Title of the Movie / Tv Show  
**Director:** Director of the Movie  
**Cast:** Actors involved in the movie/show  
**Country:** Country where the movie/show was produced  
**Date\_added:** Date it was added on Netflix  
**Release\_year:** Actual Release year of the movie/show  
**Rating:** TV Rating of the movie/show  
**Duration:** Total Duration - in minutes or number of seasons  
**Listed\_in:** Genre  
**Description:** The summary description

**Hints**

1. The exploration should have a goal. As you explore the data, keep in mind that you want to answer which type of shows to produce and how to grow the business.
2. Ensure each recommendation is backed by data. The company is looking for data-driven insights, not personal opinions or anecdotes.
3. Assume that you are presenting your findings to business executives who have only a basic understanding of data science. Avoid unnecessary technical jargon.
4. Start by exploring a few questions: What type of content is available in different countries?
   1. How has the number of movies released per year changed over the last 20-30 years?
   2. Comparison of tv shows vs. movies.
   3. What is the best time to launch a TV show?
   4. Analysis of actors/directors of different types of shows/movies.
   5. Does Netflix has more focus on TV Shows than movies in recent years
   6. Understanding what content is available in different countries

**Evaluation Criteria (100 Points):**

1. Defining Problem Statement and Analysing basic metrics **(10 Points)**

2. Observations on the shape of data, data types of all the attributes, conversion of categorical attributes to 'category' (If required), missing value detection, statistical summary **(10 Points)**

3. Non-Graphical Analysis: Value counts and unique attributes ​​**(10 Points)**

4. Visual Analysis - Univariate, Bivariate after pre-processing of the data

Note: Pre-processing involves unnesting of the data in columns like Actor, Director, Country

4.1 For continuous variable(s): Distplot, countplot, histogram for univariate analysis **(10 Points)**

4.2 For categorical variable(s): Boxplot **(10 Points)**

4.3 For correlation: Heatmaps, Pairplots **(10 Points)**

5. Missing Value & Outlier check (Treatment optional) **(10 Points)**

6. Insights based on Non-Graphical and Visual Analysis **(10 Points)**

6.1 Comments on the range of attributes

6.2 Comments on the distribution of the variables and relationship between them

6.3 Comments for each univariate and bivariate plot

7. Business Insights **(10 Points)** - Should include patterns observed in the data along with what you can infer from it

8. Recommendations **(10 Points)** - Actionable items for business. No technical jargon. No complications. Simple action items that everyone can understand

**Submission Process:**

1. Type your insights and recommendations in the rich-text editor.
2. Convert your jupyter notebook into PDF (Save as PDF using Chrome browser’s Print command), upload it in your Google Drive (set the permission to **allow public access**), and paste that link in the text editor.
3. Alternatively, you can directly submit your PDF on the portal.
4. Optionally, you may add images/graphs in the text editor by taking screenshots or saving matplotlib graphs using plt.savefig(...).
5. After submitting, you will not be allowed to edit your submission.

# Brainstorming - 11jan

Can I use google to find out

* which are the most populous countries
* what is the avg viewing time in these countries.

No, at this point, I will restrict myself to the data provided. But I will mention that the most important data for the required decision could have been overall run time of each movie/show….how many minutes ppl have spent on it

Another extremely important metric would be the distribution of these minutes..like if we can have minutes spent on each entry per day, we can know that when a movie comes, how enthused ppl are to see it asap, and when a popular tv showcomes, then what happens..anecdotal evidence can be given here to prove this data is super critical for the result that has been sought in this study. This information can greatly enhance this study.

Based on Netflix dataset find out.

* In each country, what sells more, movies or tv shows.
* How to gauge popularity of tv shows versus movies- the historical data itself is proof which ones were more popular and thus we can assume netflis executitivs would have calibrated tthe different types of offerings accordingly.
* Find avg revenue per country…what is the mix of tv shows and movies there.

Problem Statement

1. Help Netflix in deciding which type of shows/movies to produce
2. How they can grow the business in different countries

Lets look at the columns closely and let me put my comments

|  |  |
| --- | --- |
| **Show\_id:** Unique ID for every Movie / Tv Show | Can be useful if we split the data per country |
| **Type:** Identifier - A Movie or TV Show | Important |
| **Title:** Title of the Movie / Tv Show | Not much use |
| **Director:** Director of the Movie | Can be useful once we zero in on type |
| **Cast:** Actors involved in the movie/show | Can be useful once we zero in on type |
| **Country:** Country where the movie/show was produced | Important |
| **Date\_added:** Date it was added on Netflix | Important |
| **Release\_year:** Actual Release year of the movie/show | Similar to date has only year. |
| **Rating:** TV Rating of the movie/show | Important |
| **Duration:** Total Duration - in minutes or number of seasons | Important |
| **Listed\_in:** Genre | Important |
| **Description:** The summary description | LLM analysis? |

Which country has max movies and

|  |
| --- |
|  |
| 1. Defining Problem Statement and Analysing basic metrics (10 Points) |
| 1. Help Netflix in deciding which type of shows/movies to produce 2. How they can grow the business in different countries |
| 2. Observations on the shape of data, data types of all the attributes, conversion of categorical attributes to 'category' (If required), missing value detection, statistical summary (10 Points)  The data contains 8807 rows and 12 columns   |  |  |  | | --- | --- | --- | | Column | Observation | Significance | | Show\_id: Unique ID for every Movie / Tv Show | Categorical data  It is a string with s<number> format. A unique id for every show | Can be useful if we split the data per country | | Type: Identifier - A Movie or TV Show | categorical, nominal , two values | Important | | Title: Title of the Movie / Tv Show | categorical, nominal, name of show | Not much use | | Director: Director of the Movie | categorical, nominal,  comma separated, has nulls | Can be useful once we zero in on type | | Cast: Actors involved in the movie/show | categorical, nominal,  comma separated, has nulls | Can be useful once we zero in on type | | Country: Country where the movie/show was produced | categorical, nominal,  comma separated, has nulls | Important | | Date\_added: Date it was added on Netflix | Numerical, Interval | Important  Range analysis | | Release\_year: Actual Release year of the movie/show | Numerical | Similar to date has only year. | | Rating: TV Rating of the movie/show | ordinal, 14 possible values | Important | | Duration: Total Duration - in minutes or number of seasons | Numerical, ratio  contains duration in minutes and seasons, needs to be normalized | Important | | Listed In | Categorical, nominal  Comma separated | Important | | description | categorical, nominal | Important if we could do sentiment analysis on it. | |
| Some columns have null values. Here are the details  Column: director, NA Values: 2634  Column: cast, NA Values: 825  Column: country, NA Values: 831  Column: date\_added, NA Values: 10  Column: rating, NA Values: 4  Column: duration, NA Values: 3 |

|  |
| --- |
| 3. **Non-Graphical Analysis: Value counts and unique attributes (10 Points)**   * Analysis on **Type** column   + There are two type of shows – Movies and TV Show   + There are 6131 movies and 2676 TV Shows * Analysis from Director   + The field is a comma separated list of values   + Of the total 8807 shows, directors for 2634 shows’ director details are missing * Analysis from Cast   + The field is a comma separated list of values   + For 825 shows, cast details are not available in the data * Analysis from **countries** field   + The field is a comma separated list of values   + Netflix is present in 124 countries out of approx. 197 countries present   + For 831 shows, country information is not provided. We shall be assuming that these shows are released in all countries. * Analysis of **Year** fields   + The Date Added field shows when the show was added to Netflix, while the year field tells when the show was actually produced. * Analysis on **Ratings** * There are 14 Ratings - 'PG-13', 'TV-MA', 'PG', 'TV-14', 'TV-PG', 'TV-Y', 'TV-Y7', 'R',   'TV-G', 'G', 'NC-17', 'NR', 'TV-Y7-FV', 'UR'   * TV-MA is the most common rating for both movies and TV Show * Analysis of **duration** field   + For movies, the value is given in minutes, for TV shows the value is given in number of seasons.   + Comparison, aggregation will require normalization. * Analysis from **listed\_in** field   + The field is a comma separated list of values   + It contains the following 42 unique values - 'Documentaries', 'International TV Shows', 'TV Dramas', 'TV Mysteries', 'Crime TV Shows', 'TV Action & Adventure', 'Docuseries', 'Reality TV', 'Romantic TV Shows', 'TV Comedies','TV Horror', 'Children & Family Movies', 'Dramas', 'Independent Movies', 'International Movies', 'British TV Shows', 'Comedies', 'Spanish-Language TV Shows', 'Thrillers', 'Romantic Movies', 'Music & Musicals', 'Horror Movies', 'Sci-Fi & Fantasy', 'TV Thrillers', "Kids' TV", 'Action & Adventure', 'TV Sci-Fi & Fantasy', 'Classic Movies', 'Anime Features', 'Sports Movies', 'Anime Series', 'Korean TV Shows', 'Science & Nature TV', 'Teen TV Shows', 'Cult Movies', 'TV Shows', 'Faith & Spirituality', 'LGBTQ Movies','Stand-Up Comedy', 'Movies', 'Stand-Up Comedy & Talk Shows', 'Classic & Cult TV' |
| 4. **Visual Analysis - Univariate, Bivariate after pre-processing of the data**   * 50% of shows are 5yr or less old * 80# are 9 yr old or less * 4994 director details are available * As of 2021, 36440 actors have some show of theirs listed on Netflix |
| Note: Pre-processing involves unnesting of the data in columns like Actor, Director, Country |
| 4.1 For continuous variable(s): Distplot, countplot, histogram for univariate analysis (10 Points)   |  | | --- | |  | |  | |  | | Univariate analysis in duration variable | | |  |  | | --- | --- | |  | We can see that the durations are clustered around two points – 100 min and 480 minutes.  The bivariate plots done later will reveal that the graph around 100 min is for movies and graph bar at 480 min is for TV Shows | |  | Plot reveals that 67% of TV shows did only one season  Thus we infer that 33 percent of shows went on to make a new season after season 1. A show will only go for a new season if it was successful in the first one. It can be construed that 33 percent of shows are successful | | | Univariate analysis on Release year | | |  |  | | --- | --- | |  | * Netflix has a movie as old as 1925. | |  | * Most shows on Netflix are not more than 10 yrs old * Max number of shows are 4 yrs old. | |  | * 50% of shows are 5yr or less old * 80% are 9 yr old or less | |  | * Max number of shows are released in July and December. July is the start of the season maybe due to US independence day. * Feb has the least number of releases | | | Univariate analysis on Rating column | | |  |  | | --- | --- | |  | * Only 10% of shows out of 8807 are kids-friendly * 36% of shows on netflix are for mature audience, and overall 90% of shows are for adults only. | |  |  | | | Univariate analysis of country variable | | |  |  | | --- | --- | |  | * Out of total 8807 shows, 42% were released in United States.  Next biggest country in terms of shows is India, UK and Canada. * 9% of shows are not marked with country, so we can assume they were released in all countries. | | |  | |  | | Bivariate analysis on Year and Type field | | |  |  | | --- | --- | |  | * Maximum number of shows were added to Netflix in 2019 * The number of movies is typically only double the number of TV Shows. Considering that each TV Show may consume 480 minutes, while movie is only 100 minutes, here is another confirmation that TV Shows comprise of 4 times the viewing minutes compared to movies. | | | Bivariate analysis with type and duration variables | | |  |  | | --- | --- | |  | Assuming that each season has 12 episodes and each episode is 40 min on avg, we see that while number of TV shows are less than half of movies, they contain 3.7 times more ‘viewing minutes’ compared to movies.  In other words, out of every 10 minutes of material on Netflix, 8 minutes is TV Show  Recommendation: Netflix should focus more on TV Shows |  |  |  | | --- | --- | |  | The mean duration for movies is 100 minutes  The mean duration for TV Shows is 820 minutes | |  | This plot brings out that while mean of TV Show is 820 minutes, actually some shows went on till even 8000 minutes. The range for duration for TV shows is 480 to 8000 minutes. Many shows did till 7 seasons | | | Bivariate analysis of Listed\_in and type | | |  | | --- | |  | | * Out of 6131 movies, 45% are international movies * Out of 2676 TV Shows, 50% are international TV Shows * Drama is the most popular category for both TV Shows and Movies, every 2 out of 5 movies is a drama, every 1 out of 3 TV Shows is a drama * Every 4th movie is a comedy * Every 5th tv show is a comedy | |  | |  | | | Bivariate analysis of Listed In and year\_added | | |  | | --- | |  | | * International movies, Drama and Comedies are the three categories that saw maximum percentage rise in recent years. | | | Bivariate analysis of Rating and Release year | | |  |  | | --- | --- | |  | * TV14 and TV-PG rated shows are the most consistent over the years. It means these type of shows are “evergreen” * Kid movies have lesser “shelf life”. Nobody wants to see old kid movies. This can explain why Netflix has less number of kid friendly movies. They don’t have longevity. | | | Bivariate analysis of Type and Release year | | |  |  | | --- | --- | |  | Very few TV shows of timeline older than 1980 are present when compared to movies. | | | Bivariate analysis of Country and Listed\_In = International | | Reducing to first 20 countries     * India has the maximum number of internation shows listed * Japan and South Korea are Asian countries with more shows than China, despite having significantly less population than China | | Trivariate analysis of Population, Number of shows and Countries for ListedIn=International | | We bring in third party data of population of each country to identify any possibility of a recommendation from this extra data | | We calculate ratio of Population to Number of shows to see if there is any country which has large population but less number of shows    Bigger plot, only first 20 countries with higher ratio     * On basis of ratio of population to number of local language show, Ethiopia, Bangladesh, Sudan and Uganda are countries to target. However one would need more data in terms of per capita income.  How many people can afford Netflix in these countries | |  | | Trivariate analysis of Country and Type and Show Count | | |  |  | | --- | --- | |  | Among countries where Netflix has 400+ shows, South korea has highest ratio of TV Show versus Movies.  There are many more TV shows released in S Korea as compared to Movies | | |  | |  | |  | |  | |  | |  | |  | |  | |  | |
|  |
|  |
|  |
| 4.2 For categorical variable(s): Boxplot (10 Points) |
| 4.3 For correlation: Heatmaps, Pairplots (10 Points) |
| 5. **Missing Value & Outlier check (Treatment optional) (10 Points)**  The missing values in following columns were fixed   |  |  | | --- | --- | | Variable | Treatment of missing value | | director | Missing values were filled with the string ‘other’ | | cast | Missing values were filled with the string ‘other’ | | country | Missing values were filled with the string ‘’ALL’. The reason is the assumption that if some row is missing country, maybe that show was released in all 124 countries, which may be difficult to store, hence they may have left the cell empty | | date\_added | Used the release year to fill these values. Last day of that year was assumed to be the day the show got listed on Netflix. For ex, if release year is 2020, then 31-Dec-2020 was filled. | | rating | TV-MA is the most common rating (mode) for both Movie and TV Show. So imputed the missing values with this value | | duration | The duration values were wrongly put in rating column. So just copied the values from there | |
| 6. Insights based on Non-Graphical and Visual Analysis (10 Points) |
| 6.1 Comments on the range of attributes |
| 6.2 Comments on the distribution of the variables and relationship between them |
| **6.3 Comments for each univariate and bivariate plot**  I have given this alongside each plot as its easier to read. |
| 7. Business Insights (10 Points) - Should include patterns observed in the data along with what you can infer from it   * Business insight from type field   + While the number of movies is 70% of the overall shows, and the number of movies that get added every year is TWICE that of TV Shows, it is the TV Shows that comprise 80% of viewing minutes on Netflix. TV shows are much more popular and engaging that movies. If we have had data on the actual duration of these shows consumed, it can be used to corroborate which amongst the two types has higher shelf life.   + **Recommendation 1:** Netflix should focus more on TV Shows * Business insight from **duration** field   + Out of every 10 minutes of material on Netflix, 8 minutes is TV Show   + 32 percent of TV Shows went on to create more than one season. A show will only go for a new season if **it was** successful in the first one. It can be construed that 32 percent of shows are successful (see Duration.2 plot)   + **Recommendation 2:** Netflix should encourage producers of TV Shows that have done more than one season to do more seasons * Business insight from **Ratings** Field   + 36% of shows on Netflix are for mature audience, and overall 90% of shows are for adults only. This indicates that mostly the subscribers of Netflix may be the audience that prefers this genre of material.   + **Recommendation 3:** Most people using Netflix are looking for adult content. So quality content in these genres is always going to sell * Business insight from **Year** fields   + Maximum number of shows were added to Netflix in 2019   + Netflix has a movie as old as 1925.( see Chart Year 2)   + Most shows on Netflix are not more than 10 yrs old (see Chart Year 3)   + Max number of shows are 4 yrs old. (see Chart Year 3)   + 50% of shows are 5yr or less old(see Chart Year 3)   + Only 6% of shows on Netflix are older than the Year 2000   + It confirms that most shows are of recent production and customers prefer not to see very old movies or shows.  Maybe only hit or classic shows will make it to Netflix which are more than 10 yrs old. * Business insight from **listed\_in** field   + 55% movies and 50% TV Shows are in English (considering Netflix data that mentions *International* pertains to non-English shows). While English is surely the most common language understood in the world, Chinese, Hindi, Spanish, Arabic and French speakers combined are twice those of English speakers. While many of these people may be bilingual and understand English too, there is surely a case for larger share of International shows in Netflix’s repertoire.   + **Recommendation** **4**: International shows (non English) currently comprise of 50% of overall shows, and this count can be inched up   + Drama is the most popular category for both TV Shows and Movies, every 2 out of 5 movies is a drama, every 1 out of 3 TV Shows is a drama(see chart ListedCategory1)   + Every 4th movie is a comedy(see chart ListedCategory1)   + Every 5th tv show is a comedy(see chart ListedCategory1)   + International movies, Drama and Comedies are the three categories that saw maximum percentage rise in recent years. (see chart ListedCategory2)   + **Recommendation 5:** International movies, Drama and Comedies are the show categories that should be encouraged more. * Business insight from **countries** field   + Netflix is present in 124 countries out of approx. 197 countries present   + Netflix is not yet present in quite a few countries in the African continent   + In the countries Netflix is not present, top 10 countries in terms of population is Tanzania,Myanmar,Uzbekistan,Madagascar,Niger,Mali,Zambia,Chad,Rwanda,Guinea. The maximum population is in Tanzania of around 50 million.   + Largest number of local language movies are released in India   + Among countries where Netflix has 400+ shows, South korea has highest ratio of TV Show versus Movies.  There are many more TV shows released in S Korea as compared to Movies   + On basis of ratio of population to number of local language show, Ethiopia, Bangladesh, Sudan and Uganda are countries to target.  But this needs more analysis in terms of per capita income.  How many people can afford Netflix in these countries. However psychological studies have shown that even people with relatively lesser per capita income can become Netflix subscribers. Here is a summary of that study  |  | | --- | | The desire for people in lower-income countries to subscribe to a service like Netflix can be understood through several psychological and sociocultural lenses:   1. **Aspiration for a Better Lifestyle**: This urge might stem from the desire to experience aspects of a lifestyle they perceive as more affluent or desirable. 2. **Social Inclusion**: Access to popular global content can create a sense of connection with wider global cultures and trends, reducing feelings of exclusion. 3. **Escapism and Entertainment**: People might seek entertainment and escape from daily challenges. Streaming services offer a broad range of content that provides a mental getaway. 4. **Access to Information**: Educational content and documentaries can provide valuable information and learning opportunities that might not be readily available otherwise. 5. **Status Symbol**: Subscribing to an international service can serve as a status symbol, reflecting modernity and technological savvy. 6. **Psychological Comfort**: Having access to the same entertainment options as those in wealthier countries can offer psychological comfort and a sense of equity. 7. **Brand Loyalty and Trust**: Recognizable global brands like Netflix often garner trust and loyalty, making people inclined to subscribe.   These motivations can be collectively referred to as **aspirational consumption**—the pursuit of goods or services that represent a higher social status or improved quality of life. |  * + **Recommendation 6**: Netflix can look into how to reach out to some of the populous countries in Africa like Ethiopia, Sudan and Uganda, and Bangladesh   + Out of total 8807 shows, 42% were released in United States.  Next biggest countries in terms of shows are India, UK and Canada with 12%, 9% and 5% share respectively   + 9% of shows are not marked with country, so we can assume they were released in all countries.   + **Recommendation 7**: In China, the world’s second most populous country, only 2% of shows were released out of total 8807 shows. The reason for this can be investigated further * Business insight from Director   + Plil Sgriccia is the most prolific TV Show director with more than 15 seasons   + **Recommendation** **8**: 60 TV directors out of 299 have achieved the distinction of creating more than one season.  Any shows by these directors should be prioritized * Business insight from Cast   + 456 actors have done more than 10 shows on Netflix with 30 doing more than 20 shows.   + **Recommendation 9**: Any movie by these 456 actors should be released on Netflix |
| **8. Recommendations (10 Points) - Actionable items for business. No technical jargon. No complications. Simple action items that everyone can understand** |
| **Recommendation 1:** Netflix should focus more on TV Shows |
| **Recommendation 2:** Netflix should encourage producers of TV Shows that have done more than one season to do more seasons |
| **Recommendation 3:** Most people using Netflix are looking for adult content. So quality content in these genres is always going to sell |
| However this does not mean that they would not prefer non-adult contents as many a time people are watching with entire family. Netflix can look at adding more content that is kids’ friendly. |
| **Recommendation 4:** International shows (non English) currently comprise of 50% of overall shows, and this count can be inched up |
| **Recommendation 5:** International movies, Drama and Comedies are the show categories that should be encouraged more. Data related to language will help to further refine this decision |
| **Recommendation 6**: Netflix can look into how to reach out to some of the populous countries in Africa like Ethiopia, Sudan and Uganda, and Bangladesh. |
| **Recommendation 7**: In China, the world’s second most populous country, only 2% of shows were released out of total 8807 shows. The reason for this can be investigated further |
| **Recommendation** **8**: 60 TV directors out of 299 have achieved the distinction of creating more than one season.  Any shows by these directors should be prioritized |
| **Recommendation 9**: Any movie by these 456 actors should be released on Netflix |