

**Case Study 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 ijn 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)

*(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.

**Netflix: Problem statement**

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

**Analyzing Basic Metrics**

To analyze the basic metrics, we'll focus on the following aspects:

* Content Type Distribution: Determine the number of movies and TV shows.
* Genre Distribution: Analyze the genres listed and their frequencies.
* Rating Distribution: Examine the distribution of content ratings (e.g., PG-13, TV-MA).
* Duration Analysis: For movies, analyze the distribution of durations. For TV shows, analyze the number of seasons.
* Geographical Distribution: Identify the countries represented in the dataset and the number of entries from each country.
* Yearly Release Trend: Analyze the number of releases per year.
* Director and Cast Analysis: Determine the most frequent directors and cast members.
* Date Added: Explore the trend of when content was added to the platform.
* Listed Categories: Analyze the frequency of different categories (e.g., Documentaries, International TV Shows).
* Descriptive Statistics: Provide an overview of basic statistics like mean, median, mode for numerical columns.

1. 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)**
2. Observations on the Shape of Data:

* What are the dimensions of the dataset (number of rows and columns)?
* Are there any duplicated rows that need to be removed?

1. Data Types of All Attributes:

* What are the data types of each column in the dataset?
* Are the data types appropriate for the respective columns? For instance, is the release\_year column an integer, and is the date\_added column in datetime format

1. Conversion of Categorical Attributes to 'category':

* Which columns contain categorical data that can be converted to the 'category' dtype for memory efficiency and better performance?
* Are there any columns that should not be converted to 'category' despite containing text or string data? For instance, columns with a high number of unique values.

1. Missing Value Detection:

* Are there any missing values in the dataset?
* Which columns contain missing values, and what percentage of the data do they represent?
* What are the possible reasons for missing values in each column?
* How should missing values be handled (e.g., imputation, removal)?

1. Statistical Summary:

* What are the basic statistics (mean, median, mode, standard deviation, etc.) for numerical columns?
* What is the distribution of categorical data (frequency counts for each category)?
* Are there any outliers or anomalies in the data?

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

Value Counts:

* What are the most common and least common values in each categorical column?
* How many movies and TV shows are there?
* What are the most frequent genres?
* What are the most common ratings assigned to the content?
* How many entries are there for each country?

Unique Attributes:

* How many unique values exist in each categorical column?
* What are the unique values for key columns such as type, rating, country, listed\_in, etc.?
* Are there any columns with a large number of unique values that might indicate high variability?

1. 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)**

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

Preprocessing the Data

Preprocessing involves cleaning and preparing the data for analysis. Specifically, we'll "unnest" the columns that contain multiple values, such as actors, directors, and countries.

4.1 Univariate Analysis for Continuous Variables

We will create distribution plots (distplot), count plots, and histograms for univariate analysis.

4.2 Univariate Analysis for Categorical Variables

We will use boxplots to understand the distribution of continuous variables within different categories.

4.3 Correlation Analysis

We will use heatmaps and pairplots to analyze the correlation between variables.

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

Missing Value Detection

1. Identify Columns with Missing Values:
   * missing\_values: Counts the number of missing values in each column.
   * missing\_percentage: Calculates the percentage of missing values in each column.
2. Display Rows with Missing Values (Optional):
   * missing\_data: Displays rows that have any missing values for further inspection.

Outlier Detection

1. Function to Detect Outliers using IQR:
   * The detect\_outliers\_iqr function calculates the Interquartile Range (IQR) for each numerical column.
   * Identifies outliers as values outside the range [Q1−1.5×IQR,Q3+1.5×IQR][Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR][Q1−1.5×IQR,Q3+1.5×IQR].
2. Visualize Outliers using Boxplots:
   * Plots boxplots for each numerical column to visually inspect the presence of outliers.
3. 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

#### 6.1 Comments on the Range of Attributes

* **Release Year**:
  + The release\_year attribute ranges from older to more recent years, providing a wide temporal coverage of the dataset.
  + Most content appears to have been released in recent years, indicating a growing library of new content.
* **Duration**:
  + For movies, the duration ranges from short films to longer feature films. The majority of movie durations cluster around typical feature film lengths (90-120 minutes).
  + TV shows are categorized by the number of seasons rather than specific episode durations.
* **Rating**:
  + The rating attribute includes a variety of audience classifications, from "G" (General Audiences) to "NC-17" (Adults Only). This diversity indicates content suitable for different age groups and preferences.
* **Genres (**listed\_in**)**:
  + The listed\_in column contains a broad range of genres and categories, showcasing the diversity of content available on the platform.
* **Countries**:
  + The dataset includes content from a wide range of countries, emphasizing its international appeal.

#### 6.2 Comments on the Distribution of the Variables and Relationship Between Them

* **Distribution of Release Year**:
  + The release year distribution is right-skewed, with a noticeable increase in content production in recent years.
  + This trend might indicate an increasing investment in content creation by the platform.
* **Distribution of Duration**:
  + The duration of movies typically falls within the 90-120 minute range. Outliers include very short and very long films, suggesting a few documentaries or specials.
* **Distribution of Ratings**:
  + The countplot for ratings shows that TV-MA (Mature Audience) is the most common rating, reflecting a significant amount of content targeted at adults.
  + The platform also has a substantial amount of content rated PG-13, indicating a focus on teenage and adult audiences.
* **Genres Distribution**:
  + Certain genres like International TV Shows and Documentaries are highly represented, indicating the platform's focus on diverse and informative content.

#### 6.3 Comments for Each Univariate and Bivariate Plot

* **Univariate Analysis (Distplot for** release\_year**)**:
  + The distplot shows a significant increase in content releases in recent years, highlighting the platform's growing library.
  + The spike in recent years could be attributed to the platform's original content production strategy.
* **Univariate Analysis (Histogram for Movie** duration**)**:
  + The histogram reveals that most movies are around 90-120 minutes long, which is typical for feature films.
  + There are a few outliers with very short or very long durations, indicating some variance in movie lengths.
* **Univariate Analysis (Countplot for** rating**)**:
  + The countplot indicates that TV-MA is the most frequent rating, showing the platform's focus on mature content.
  + Ratings like G and PG have fewer entries, suggesting less content aimed solely at younger audiences.
* **Bivariate Analysis (Boxplot for Movie** duration **by** rating**)**:
  + The boxplot shows that movie durations do not vary significantly across different ratings, though some ratings like NC-17 have fewer entries.
  + This suggests that the rating does not strongly influence the length of a movie.
* **Correlation Analysis (Heatmap)**:
  + The heatmap reveals weak correlations among numerical variables, suggesting that variables like release\_year and duration are relatively independent of each other.
  + The weak correlations indicate that other factors (like genre or audience preference) might play a larger role in determining content characteristics.
* **Pairplot**:
  + The pairplot provides a visual representation of the relationships between variables, showing potential clusters and trends.
  + For example, it might reveal if newer releases are more concentrated in certain genres or if certain ratings are more prevalent in recent years.

1. Business Insights **(10 Points)** - Should include patterns observed in the data along with what you can infer from it
2. Recommendations **(10 Points)** - Actionable items for business. No technical jargon. No complications. Simple action items that everyone can understand

**KPI:  
Content Type Distribution:**

* **KPI:** Percentage of movies vs. TV shows in the dataset.
* **Calculation:** Count of movies and TV shows, and their respective percentages of the total dataset.

**Top Genres:**

* **KPI:** Most common genres listed.
* **Calculation:** Count of occurrences of each genre in the "listed\_in" column.

**Country Distribution:**

* **KPI:** Content count by country.
* **Calculation:** Number of titles produced in each country.

**Director and Cast Popularity:**

* **KPI:** Most frequently occurring directors and actors/actresses.
* **Calculation:** Count of occurrences of each director and cast member.

**Release Year Analysis:**

* **KPI:** Number of titles released each year.
* **Calculation:** Count of titles by "release\_year".

**Content Rating Distribution:**

* **KPI:** Distribution of content ratings (e.g., PG-13, TV-MA).
* **Calculation:** Count of titles by their rating.

**Average Duration of Movies:**

* **KPI:** Average duration of movies.
* **Calculation:** Average of the "duration" column for movies, converting "min" to numeric values.

**Recent Additions:**

* **KPI:** Number of titles added each month.
* **Calculation:** Count of titles by "date\_added".

**International vs. Domestic Content:**

* **KPI:** Ratio of international to domestic content.
* **Calculation:** Count of titles by country, distinguishing between domestic (e.g., United States) and international.

**Top Listed Titles:**

* **KPI:** Titles with the most seasons (for TV shows).
* **Calculation:** Count and list TV shows with the highest number of seasons.

From year 2012 movies/shows release has been increased per year with the inclining trend till 2018 and further again after that, it is down trending.

No of movies addition trend seems to be declining from the 2016 while the trend for the movie shows addition has increased, which is showing focused approach towards the addition of shows vs movies while that surpassed no of movies added vs TV shows in 2021.

If we study the movies/shows carefully we can see the clear focus on the TV-MA and TV-14 type of content targeting teenage and adults for the content consumptions. Also Netflix has very low penetration for the below 13 years age content which Netflix can add more considering the children as next targets.

General movies length is ideally between 90 to 120mins which has clearly indicating that to careful while producing movie in the range of 90mins to 120mins, which will not make it too short or too lengthy.

Netflix can also target to cater the content in the genre segments of international movies, dramas and comedies.

Last 30 years its evident that movies addition to tv shows addition ratio is consistently falling, with indicating the Netflix increased focus towards the shows over the movies.

If we see the till date released movies/shows July is most favored for the release of movies/shows both, as that is the summer holiday time also has the independence day for US in July which makes it most favored time to release movies to increase the views.