Netflix Data Exploration Business Case

- ❖ **Topic:** DAV Libraries
- ❖ **Duration:** 1 week

Why this case study?

From company's perspective:

- Netflix is one of the most popular media and video streaming platforms. They have over 8000 movies or tv shows available on their platform, as of mid-2021, they have over 200M Subscribers globally.
- The particular business case focuses on the Netflix show data and provides insightful information on 8807 shows
- Analyzing the data and generating insights helps Netflix decide which type of shows/movies to produce and how to grow the business.

From learner's perspective:

- Solving this business case holds immense importance for aspiring data analysts and scientists.
- Data analysis using Python libraries is widely popular among the Data Scientists and Data Analysts. By working through this case study, individuals gain hands-on experience and practical skills while attempting this case study.
- Additionally, it will enhance one's ability to communicate with the stakeholders involved in data-related projects and help the organization take better, data-driven decisions.

Dataset:

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.

Link : https://www.kaggle.com/shivamb/netflix-shows

The data is available in a single csv file :

- **Show ID** : The ID of the 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

____

How to get started?

To complete the case study, begin by downloading the CSV files from the provided link. Afterward, proceed to upload them onto Google Colab / Jupyter Notebook for further analysis.

If you are using Google Colab, you can directly start working on the notebook on Colab.

Install Anaconda using the link. Once Anaconda has been installed on your system, open Jupyter Notebook. Refer link.

Now, the netflix CSV file needs to be uploaded/imported in the Colab/Jupyter notebook respectively.

Once the file have been successfully uploaded/imported, you can conveniently access them within the notebook using the read_csv( ) method.

____

What is expected?

Assuming you are a data analyst/ scientist at Netflix, you have been assigned the task of analyzing the given dataset to extract valuable insights and provide actionable recommendations.

Submission Process:

- **Type your insights and recommendations in the text editor.**
- 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.
- Optionally, you may add images/graphs in the text editor by taking screenshots or saving matplotlib graphs using plt.savefig(...).
- After submitting, you will not be allowed to edit your submission.

General Guidelines:

1. Evaluation will be kept lenient, so make sure you attempt this case study.
2. It is understandable that you might struggle with getting started on this. Just brainstorm, discuss with peers, or get help from TAs.
3. 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.

Basic Analysis

1. **Un-nesting the columns**
    a. Un-nest the columns those have cells with multiple comma separated values by creating multiple rows

2. **Handling null values**
    a. For categorical variables with null values, update those rows as unknown_column_name.

    Example : Replace missing value with **Unknown Actor** for missing value in Actors column.

    b. Replace with 0 for continuous variables having null values.

What does ‘good’ look like?

---

1. **Find the counts of each categorical variable both using graphical and non-graphical analysis.**

    a. For Non-graphical Analysis:

    Hint : We want you to find the values counts of each category for the given column

    b. For graphical analysis:

    Hint : We can use a count plot to get the counts of each category

---

2. **Comparison of tv shows vs. movies.**

    a. Find the number of movies produced in each country and pick the top 10 countries.

    Hint : We want you to apply group by each country and find the count of unique titles of movies

    b. Find the number of Tv-Shows produced in each country and pick the top 10 countries.

    Hint : We want you to apply group by each country and find the count of unique titles of Tv-shows

---

3. **What is the best time to launch a TV show?**

    a. Find which is the best week to release the Tv-show or the movie. Do the analysis separately for Tv-shows and Movies

    Hint : We expect you to create a new column and group by each week and count the total number of movies/ tv shows.

    b. Find which is the best month to release the Tv-show or the movie. Do the analysis separately for Tv-shows and Movies

Hint : We expect you to create a new column and group by each month and count the total number of movies/ tv shows.

———

4. **Analysis of actors/directors of different types of shows/movies.**

    a. Identify the top 10 directors who have appeared in most movies or TV shows.

    Hint : We want you to group by each actor and find the count of unique titles of Tv-shows/movies

    b. Identify the top 10 directors who have appeared in most movies or TV shows.

    Hint : We want you to group by each director and find the count of unique titles of Tv-shows/movies

———

5. **Which genre movies are more popular or produced more**

    Hint : We want you to apply the **word cloud** on the **genre** columns to know which kind of genre is produced

———

6. **Find After how many days the movie will be added to Netflix after the release of the movie (you can consider the recent past data)**

    Hint : We want you to get the difference between the columns having date added information and release year information and get the mode of difference. This will give an insight into what will be the better time to add in Netflix

———

FAQs

Q. Which platform am I supposed to use?
You may use either Google Colab or Jupyter notebook.

Q. I am having issues setting up Jupyter notebook

Install Anaconda using the link. Once Anaconda has been installed on your system, open Jupyter Notebook. Refer link.