# Kaggle Disney+ Dataset

## Dataset Selection and Structure

The dataset chosen for analysis is the "Disney Movies and TV Shows" collection from Kaggle. It comprises detailed information on titles available on the Disney+ streaming platform, including attributes like type, director, cast, country of origin, date added, release year, rating, and genre.

To optimize the dataset for SQL queries and gain deeper insights, we decomposed the data into several relational tables:

1. Titles: Contains core information about each show, such as the ID, title, type, release year, and date added to Disney+.

2. Directors: Lists directors with a unique ID assigned to each.

3. Casts: Lists cast members with a unique ID for each person.

4. Countries: Enumerates the countries where the titles were produced.

5. Genres: Catalogs the genres of the titles.

6. TitleDirector, TitleCast, TitleCountry, TitleGenre: These junction tables link titles to their respective directors, cast members, countries, and genres, facilitating many-to-many relationships.

## SQL Queries and Findings

Several SQL queries were crafted to extract meaningful insights:

* Content evolution over time, revealing trends in the platform's offerings.

A screenshot of a computer code

Description automatically generated

A screenshot of a computer

Description automatically generated

* Directorial impact, showing which directors' works are most prominent.

A screen shot of a computer code

Description automatically generated

A screenshot of a screen

Description automatically generated

* Geographical diversity, indicating the global footprint of the content.

A computer screen shot of text

Description automatically generated

A screenshot of a computer

Description automatically generated

* Genre popularity and variety, highlighting the most common genres.

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

* Top 10 directors who have the highest number of titles with a specific rating, say 'PG'.

A computer screen shot of a code

Description automatically generated

A screenshot of a computer

Description automatically generated

After running the queries, we learned, for instance, that the content on Disney+ includes a diverse range of genres and has expanded over the years to include a significant number of countries, indicating Disney's global content strategy.

## Learnings from the Activity

Through this exercise, we were able to practice data normalization, which is crucial for efficient SQL querying. We also observed the importance of understanding the data types within a dataset, such as the categorical nature of content ratings, which required us to adjust our SQL queries accordingly.

## Concluding Thoughts

The Disney+ dataset offered a rich opportunity to apply SQL for real-world data analysis, highlighting the platform's content strategy and diversity. The SQL analysis demonstrated the power of relational databases in dissecting and examining large datasets to uncover trends and patterns. This activity has reinforced the practical applications of data analytics in the media and entertainment industry, showcasing how data-driven insights can guide content creation and business strategies. The ability to transform raw data into actionable knowledge is a valuable skill in the era of big data, and the Disney+ dataset served as an excellent canvas for this endeavor.

[**Week 10 Assignment: SQL Query Assignment #2**](https://ulearn.jwu.edu/webapps/assignment/uploadAssignment?content_id=_8991918_1&course_id=_4192815_1&group_id=&mode=view)

**Instructions:** Perform Data Analysis using MySQL.

1. Retrieve a dataset of your choice from [Data.gov](https://catalog.data.gov/dataset?res_format=CSV) or another opensource data repository of your choice. Provide a comment in your assignment submission to document where the dataset came from. You will NOT be submitting the actual dataset when you submit your project source code.
2. Load the dataset into MySQL using the MySQL Workbench.
3. Break out the data and create a new table with a subset of the data from the original table.
4. Use SQL commands to extract some of the columns (at least three) from the table you created. Specifically, you need to write queries that include the following SQL features:
   * One query that performs a select on multiple rows and columns.One query that contains an aggregate function.
   * One query that contains sorting and grouping.
   * One query that contains a subquery.
   * One query to join two tables together.

Create a 1-2-page summary (using MS Word) that includes the following information:

* Which dataset did you select?
* How did you break out the data into multiple tables?
* What queries did you run?
* What information did you find about the data after performing your analysis?
* What did you learn from this activity?
* What are your concluding thoughts about the dataset you selected and the SQL analysis you were able to perform?

**This assignment is worth 10 points. Submit your summary document by Sunday at 11:59 PM ET. Do not submit the dataset.**