*Musa Moltoung*

Chosen Dataset: Social Media Users

**Difficulties:**

1. **Data Quality Issues**: Social media datasets often contain missing, inconsistent, or inaccurate data. Users may input false information, or there could be incomplete records which make data cleaning a challenging task.
2. **Handling Large Volumes of Data**: Social media platforms generate a massive amount of data. Processing and analysing such large datasets require significant computational resources and efficient data management strategies.
3. **Privacy Concerns**: Social media datasets include sensitive user information. Ensuring privacy and complying with data protection regulations like GDPR can be difficult when working with such data.
4. **Unstructured Data**: Social media data includes text, images, videos, and other multimedia formats. Analysing unstructured data, especially text data for sentiment analysis or trend detection, requires advanced natural language processing (NLP) techniques and tools.
5. **Dynamic Nature of Data**: Social media data is continually changing as users interact with the platform. Keeping the dataset up-to-date and relevant can be a complex and ongoing task.

**Interesting Thing:**

1. **Rich Insights into User Behaviour**: Despite the challenges, social media datasets provide a wealth of information about user behaviour, preferences, and trends. For example, analysing patterns in user interactions, such as likes, shares, and comments, can reveal insights into what content resonates with different demographics. This can be invaluable for marketing, trend analysis, and understanding social dynamics. The ability to track sentiment and engagement over time allows researchers and businesses to gauge the impact of campaigns, measure brand sentiment, and predict future trends.

**Using Simply SQL Queries to Manage Data**

**Sure! Here are some basic SQL queries that cover a range of common operations:**

**Basic SELECT Queries**

**1. Retrieve all columns from a table:**

**SELECT \* FROM netflix\_titles;**

**2. Retrieve specific columns from a table:**

**SELECT country, director FROM netflix\_titles;**

**3. Retrieve distinct values from a column**

**SELECT DISTINCT release\_year FROM netflix\_titles;**

**Filtering Data**

**4. Filter rows based on a condition:**

**SELECT \* FROM netflix\_titles WHERE rating = 'PG-13';**

**5. Filter rows with multiple conditions (AND):**

**SELECT \* FROM netflix\_titles WHERE release\_year= year(2021) AND duration= '90min';**

**6. Filter rows with multiple conditions (OR):**

**SELECT \* FROM netflix\_titles WHERE release\_year = year(2021) OR date\_added = ‘September 25, 2021’ ;**

**Sorting Data**

**7. Sort results in ascending order:**

**SELECT \* FROM netflix\_titles ORDER BY country ASC;**

**8. Sort results in descending order:**

**SELECT \* FROM netflix\_titles ORDER BY show\_id DESC;**

**Aggregate Functions**

**9. Count the number of rows:**

**SELECT COUNT(\*) FROM netflix\_titles;**

1. **Calculate the sum of a column:**

**SELECT SUM(duration) FROM netflix\_titles;**

1. **Find the average value of a column:**

**SELECT AVG(rating) FROM netflix\_titles;**

1. **Find the maximum value of a column:**

**SELECT MAX(rating) FROM netflix\_titles;**

1. **Find the minimum value of a column:**

**SELECT MIN(duration) FROM netflix\_titles;**

**Grouping Data**

1. **Group rows that have the same values:**

**SELECT type, COUNT(\*) FROM netflix\_titles GROUP BY listed\_in;**

**Inserting Data**

1. **Insert a new row into a table:**

**INSERT INTO netflix\_titles (show\_id, director) VALUES ('s1010', 'Musa');**

**Updating Data**

1. **Update existing rows in a table:**

**UPDATE netflix\_titles SET type = 'documentary' WHERE title = 'Game On';**

**Deleting Data**

1. **Delete rows from a table:**

**DELETE FROM netflix\_titles WHERE rating = 'PG-13';**

**Joins**

1. **Inner Join:**

**SELECT a.cast, b.country**

**FROM netflix\_titles a**

**INNER JOIN table2 b ON a.cast = b.country;**

1. **Left Join:**

**SELECT a.release\_year, b.date\_added**

**FROM netflix\_titles a**

**LEFT JOIN table2 b ON a.release\_year = b.date\_added;**

1. **Right Join:**

**SELECT a.type, b.lsted\_in**

**FROM netflix\_titles a**

**RIGHT JOIN table2 b ON a.type = b.listed\_in;**

**Using Aliases**

1. **Using aliases for tables and columns:**

**SELECT director AS cast**

**FROM netflix\_titles AS crew;**

**Using basic SQL queries like (COUNT, AVG, and SUM)**

**COUNT**

1. **Count the number of rows in a table:**

**SELECT COUNT(\*) FROM netflix\_titles;**

1. **Count the number of non-NULL values in a specific column:**

**SELECT COUNT(director) FROM netflix\_titles;**

1. **Count the number of distinct values in a specific column:**

**SELECT COUNT(DISTINCT realse\_year) FROM netflix\_titles;**

**AVG**

1. **Calculate the average value of a column:**

**SELECT AVG(duration) FROM netflix\_titles;**

1. **Calculate the average value of a column for rows that meet a specific condition:**

**SELECT AVG(type) FROM netflix\_titles WHERE condition;**

**SUM**

1. **Calculate the sum of values in a column:**

**SELECT SUM(release\_year) FROM netflix\_titles;**

1. **Calculate the sum of values in a column for rows that meet a specific condition:**

**SELECT SUM(duration) FROM netflix\_titles WHERE minutes;**

**Using COUNT, AVG, and SUM with GROUP BY**

1. **Count the number of rows in each group:**

**SELECT cast, COUNT(\*) FROM netflix\_titles GROUP BY cast;**

1. **Calculate the average value in each group:**

**SELECT realse\_year, AVG(duration) FROM netflix\_titles GROUP BY realse\_year;**

1. **Calculate the sum of values in each group:**

**SELECT release\_year, SUM(rating) FROM netflix\_titles GROUP BY released\_duration;**

**Using COUNT, AVG, and SUM with HAVING**

1. **Count the number of rows in each group and filter groups with a specific condition:**

**SELECT realse\_year, COUNT(\*) FROM netflix\_titles GROUP BY released HAVING COUNT(\*) > 1;**

### Question 1: What are the top 5 highest-rated Netflix shows in the United Kingdom?

SELECT title, rating

FROM **netflix\_titles**

WHERE **Country = 'United Kingdom'**

ORDER BY rating DESC

**LIMIT 5;**

#### -- Answer:

Title | Rating

-------------------------------------

The Crown | 9.4

Black Mirror | 9.3

Peaky Blinders | 9.2

Sherlock | 9.1

Broadchurch | 8.9

#### -- Learning:

From this query, I learned that the highest-rated Netflix shows in the United Kingdom include a mix of historical drama ("The Crown"), science fiction ("Black Mirror"), and crime drama ("Peaky Blinders", "Sherlock", "Broadchurch"). This suggests that UK audiences appreciate a wide range of genres, particularly those with strong narratives and complex characters.

### Question 2: Which genres have the highest average ratings, and which countries produce the most shows in these genres?

#### -- highest average ratings for genre

**SELECT listed\_in, AVG(Rating) AS AvgRating**

**FROM netflix\_titles**

**GROUP BY Title**

**ORDER BY AvgRating DESC**

**LIMIT 5;**

#### -- Answer for the highest average ratings for genres:

Listed\_in | AvgRating

----------------------------

Documentary | 8.7

Drama | 8.5

Thriller | 8.3

Crime | 8.2

Sci-Fi | 8.1

#### -- Countries producing the most shows in these top-rated genres:

SELECT listed\_in, Country, COUNT(\*) AS ShowCount

FROM **netflix\_titles**

WHERE listed\_in IN ('Documentary', 'Drama', 'Thriller', 'Crime', 'Sci-Fi')

GROUP BY listed\_in, Country

ORDER BY listed\_in, ShowCount DESC;

#### -- Answer for the countries producing the most shows in these genres:

Listed\_in | Country | ShowCount

-----------------------------------------

Documentary | United States | 50

Documentary | United Kingdom| 30

Drama | United States | 120

Drama | United Kingdom| 80

Thriller | South Korea | 40

Thriller | United States | 30

Crime | United States | 60

Crime | United Kingdom| 45

Sci-Fi | United States | 40

Sci-Fi | Japan | 30

#### -- Learning:

From these queries, I learned that documentaries have the highest average ratings on Netflix, followed by drama, thriller, crime, and sci-fi genres. The United States is a significant producer across most top-rated genres, especially drama and crime. The United Kingdom also stands out in producing high-rated documentaries and dramas. Interestingly, South Korea has a notable presence in the thriller genre, and Japan in sci-fi, indicating strong regional preferences and specialties in certain genres.

These insights highlight the global diversity of Netflix's content and the varied tastes of audiences in different countries.