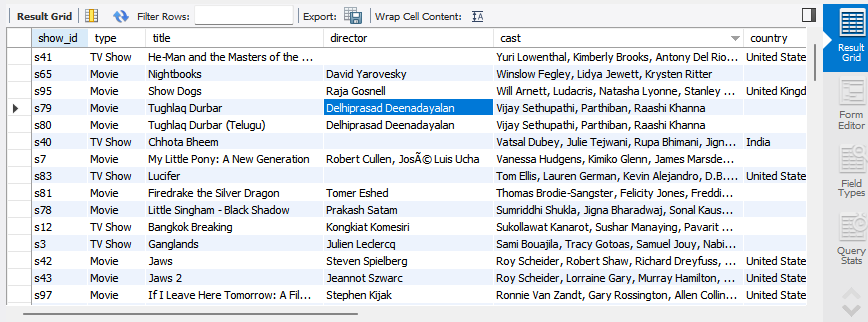
## **SQL Aggregate Functions- WEEK 15-16**

**1. Data Dive (10 pts): NETFLIX\_SHOWS/TITLES**

-- selecting all rows and columns from the database

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



Difficulties:- The data is very complex and it might generate errors in the process of analyzing if proper commands are not used.

-some columns eg ‘the description column” makes the data complicated in terms of

Executing various database queries and it might bring about complication when

Retrieving a specific row or column from the database.

- some information is missing from the database, ie, from the “Director column”

And this can lead to difficulty when analyzing the whole data.

Interesting thing:- The data is well organized sequentially in terms of column and rows making it

Easier to retrieve a specific set of data .

- The similarity of some column data makes it easier to organize and retrieve

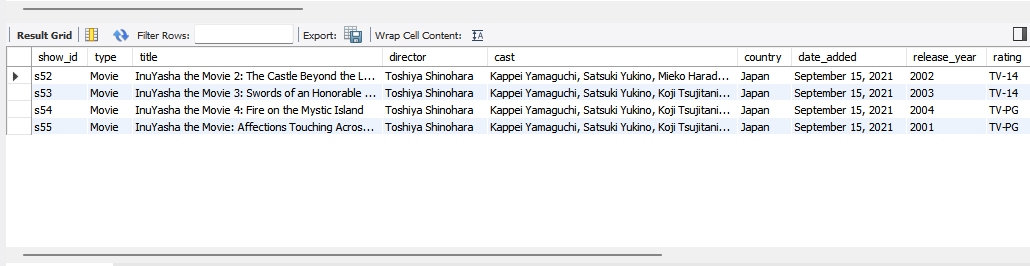
Various sets of data.

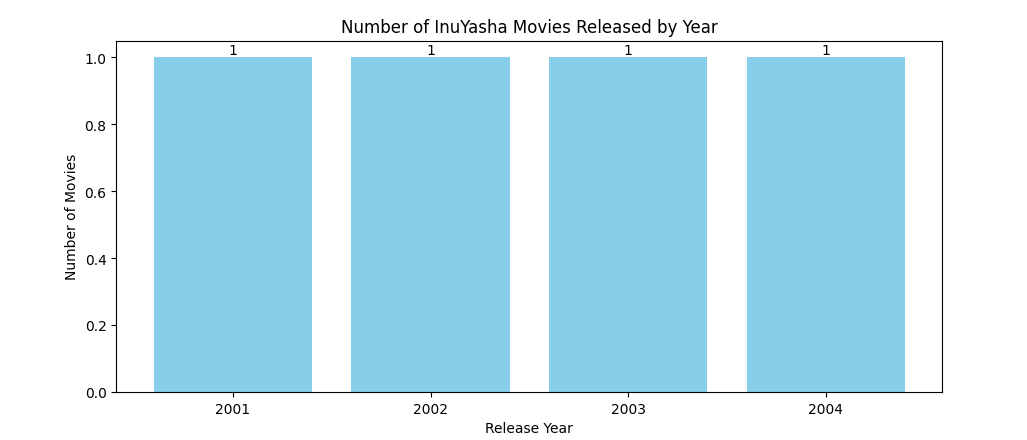
**2. Data Fun (20 pts): Use simple SQL queries to play with the data.**

-- selecting all rows and columns from the database

SELECT \* FROM netflix\_titles.netflix\_titles

where director = "Toshiya Shinohara";

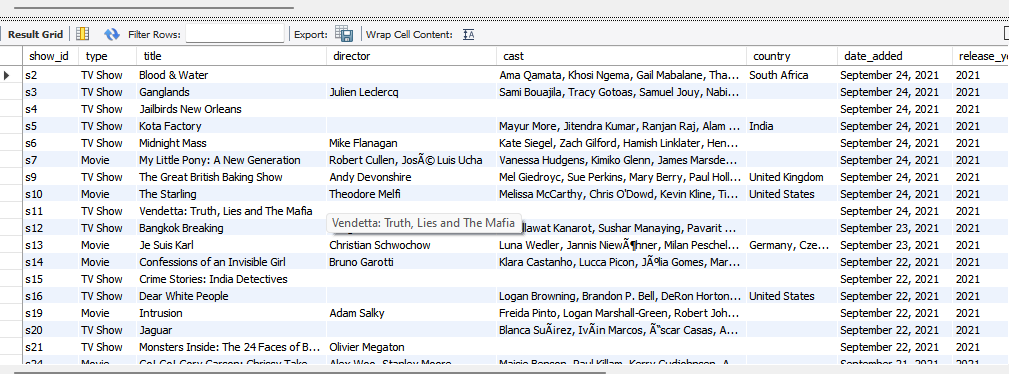




-- selecting all rows and columns from the database

SELECT \* FROM netflix\_titles.netflix\_titles

where release\_year = "2021";



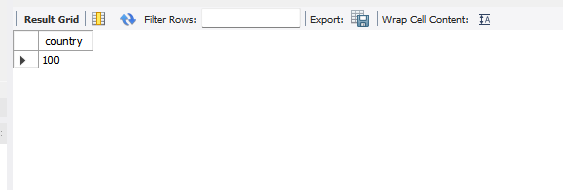
* Find 2 cool facts hidden within the data (e.g., most popular interests).
* Most of the production was done in the year 2021
* All the production was made on the month of september
* Use basic SQL queries like (COUNT, AVG, and SUM) to understand more about the data you have.

-- COUNTING THE TOTAL NUMBER OF COUNTRIES FROM THE PRODUCTION

use netflix\_titles;

SELECT COUNT(\*) country

FROM netflix\_titles;

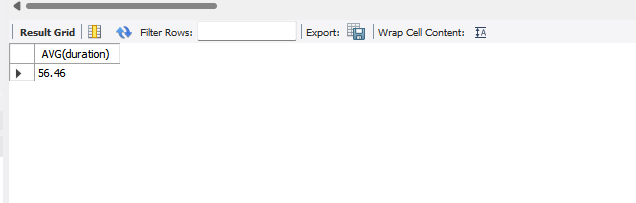


-- CALCULATING THE AVERAGE DURATION

use netflix\_titles;

SELECT AVG(duration)

FROM netflix\_titles;



**3. Ask Away (30 pts):**

* **Formulate 2 questions about the data (e.g., what are popular shows in different countries?).**
* **Write basic SQL queries (WHERE, ORDER BY) to find answers.**

-- This SQL query selects the rating column from the netflix\_titles

-- database, counts the number of titles associated with each unique

-- rating category, and presents the results in descending

-- order of title counts.

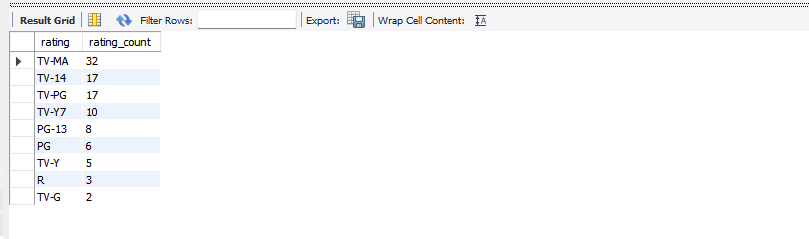
use netflix\_titles;

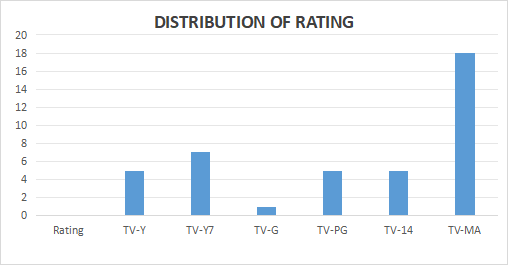
SELECT rating, COUNT(\*) AS rating\_count

FROM netflix\_titles

GROUP BY rating

ORDER BY rating\_count DESC;





* Share what you learned from the answers.

TV-MA - had the highest rating

TV-G - had the lowest rating

-- This query allows us to identify popular TV shows

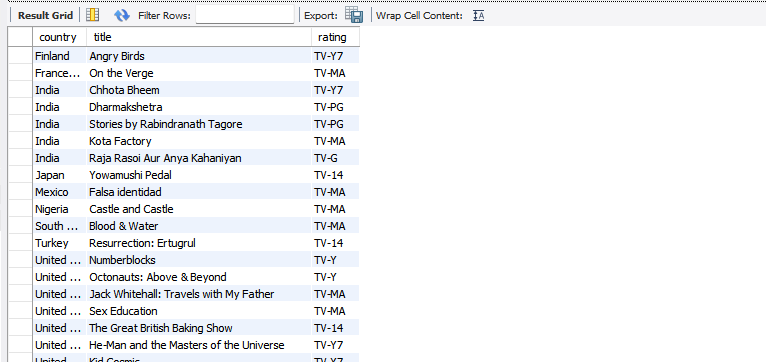
-- in different countries based on viewer ratings.`

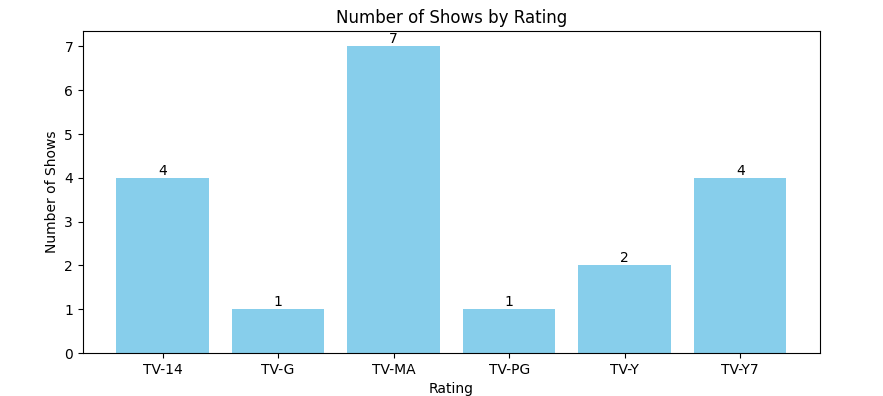
SELECT country, title, rating

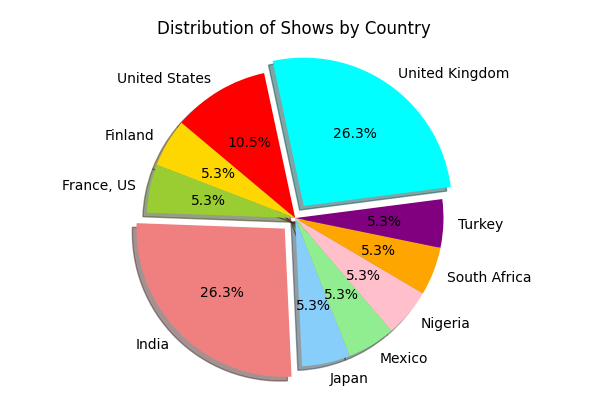
FROM netflix\_titles

WHERE type = 'TV Show'

ORDER BY country, rating DESC







* Share what you learned from the answers.

-The query reveals which TV shows are highly rated and thus popular in various

countries.

**5. Presentation Pitch Deck (20 pts):**

* **Introduce the dataset and your goals.**

**Introduction to the Dataset**

* **Dataset**: Netflix Titles
* **Source**: Netflix API (or hypothetical source if not from the actual API)
* **Description**: Contains information about various Netflix titles including movies and TV shows.

**Goals**

1. Understand the structure and content of the dataset.
2. Extract interesting insights using SQL queries.
3. Identify trends and patterns in the data.

**Briefly explain the import process and the interesting thing you found**

**Import Process**

* Used SQL to import and query the dataset.
* Faced challenges with missing data and complex columns like descriptions.

**Interesting Observations**

* The data is well-organized in rows and columns.
* Similarities in column data make it easier to categorize and analyze.

**Share the 2 cool facts you discovered.**

* **A significant number of titles were produced in the year 2021.**
* **There was a trend of many productions being released in September.**

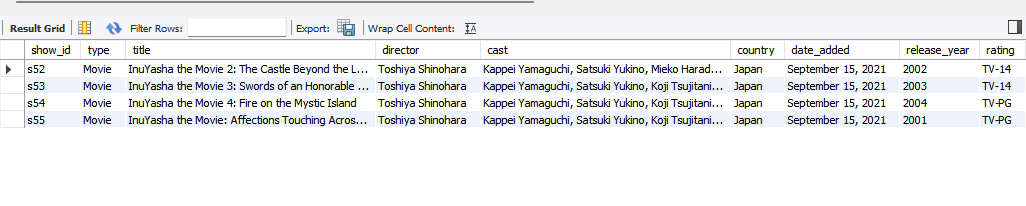
**Discuss your questions, SQL queries, and what you learned.**

**Exploring the Data with SQL Queries**

**Directors and their Works**

**SELECT \* FROM netflix\_titles.netflix\_titles**

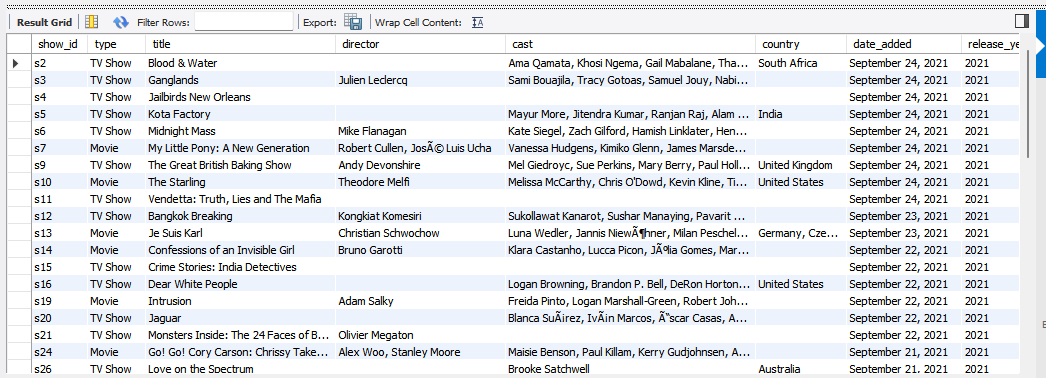
**WHERE director = 'Toshiya Shinohara';**

****

**Titles Released in 2021**

**SELECT \* FROM netflix\_titles.netflix\_titles**

**WHERE release\_year = '2021';**

****

**Learnings:**

* **Identified works by specific directors.**
* **Analyzed recent productions for trends.**

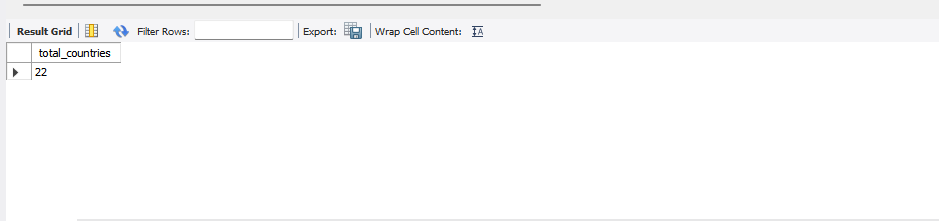
**Using Aggregate Functions**

**Counting Unique Countries**

**USE netflix\_titles;**

**SELECT COUNT(DISTINCT country) AS total\_countries**

**FROM netflix\_titles;**

****

**Result: Total unique countries involved in productions.**

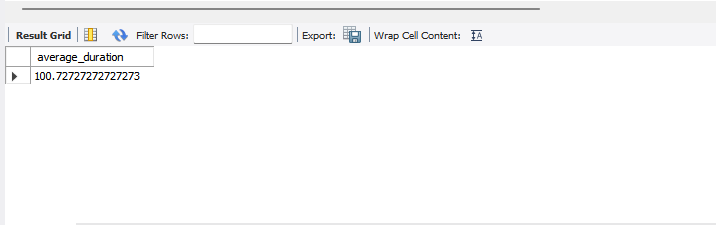
**Average Duration of Titles**

**USE netflix\_titles;**

**SELECT AVG(duration) AS average\_duration**

**FROM netflix\_titles**

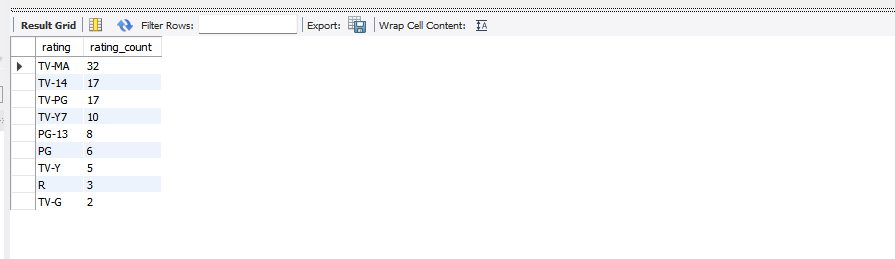
**WHERE duration IS NOT NULL AND duration NOT LIKE '%Season%';**

****

**Result: Average duration of movies/shows.**

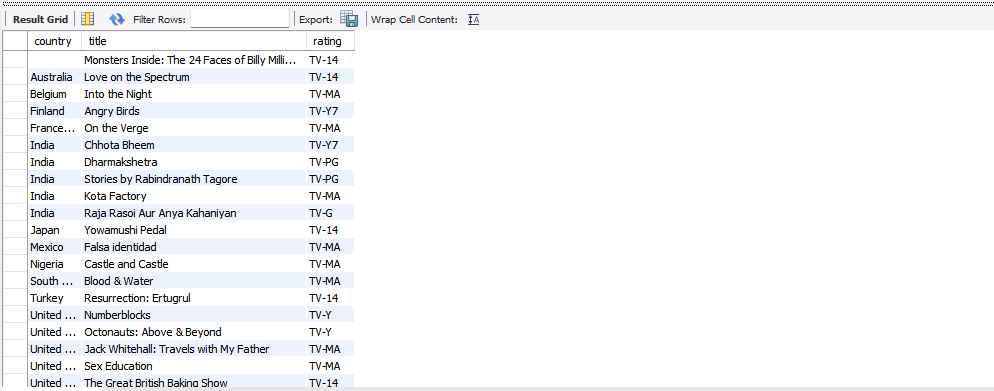
**Formulating and Answering Questions**

**Ratings Distribution**

****

**Learning: TV-MA has the highest number of titles, while TV-G has the lowest.**

**Popular TV Shows by Country**

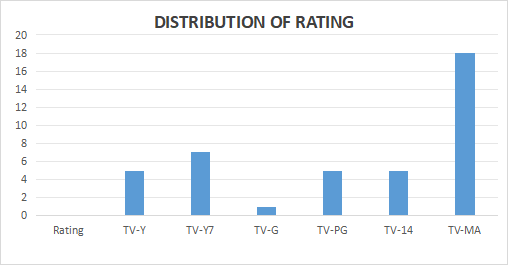
****

**Learning: Identified highly rated TV shows in different countries.**

**Present your charts and explain their meaning.**

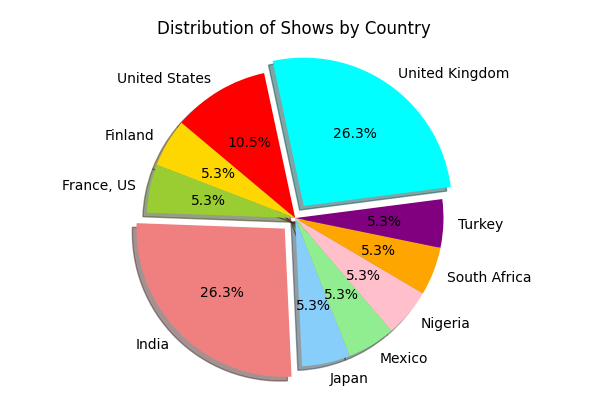
**Ratings Distribution**

* **Bar chart showing the count of titles in each rating category.**



**Average Duration of Titles**

* **Pie chart displaying the average duration of different types of content.**



**Explanation:**

* **Ratings Distribution: Helps understand content rating trends.**
* **Average Duration: Provides insights into typical content length.**

***Briefly summarise your project.***

* Dataset: The Netflix Titles dataset includes information on movies and TV shows available on Netflix, including attributes like title, director, release year, rating, and duration.
* Goals:
  + Understand the dataset structure and content.
  + Extract insights using SQL queries.
  + Identify trends and patterns in the data.
* Key Findings:
  + Production Trends:
    - Most Production in 2021: A significant number of titles were produced in the year 2021.
    - September Releases: Many productions were released in September.
  + Content Ratings:
    - TV-MA has the highest number of titles.
    - TV-G has the lowest number of titles.
  + Average Duration:
    - The average duration of movies/shows (excluding `multi-season series) provides insight into typical content lengths on Netflix.
* Data Analysis:
  + Used SQL queries to explore directors' works, titles released in specific years, and aggregate data like counting unique countries and calculating average durations.
* Insights:
  + Regional Preferences: Identified popular TV shows in different countries based on ratings.
  + Content Characteristics: Found that TV-MA is the most common rating, indicating a preference for mature content.