**Project : Data cleaning and Analysis for Netflix using SQL**

Netflix, a leading global streaming platform, possesses a dataset containing information about its shows. However, the dataset requires cleaning and analysis to derive valuable insights for business decision-making. As a data analyst with SQL expertise, your objective is to perform data cleaning and analysis on the Netflix dataset to help the company gain insights into their content offerings.

*At first, I loaded the Netflix dataset into mySQL workbench and with the SQL queries I perform Data Analysis on the Netflix dataset. ‘Netflix’ database was created and the data was populated in the Netflix\_data table.*

Segment 1: Database - Tables, Columns, Relationships

* Identify the tables in the dataset and their respective columns.
* Determine the number of rows in each table within the schema.
* Identify and handle any missing values in the dataset.

*Netflix\_data table was created using SQL query and then the data was inserted using insert statement.*

-- Creates the table netflix\_data with 11 columns in it

CREATE TABLE netflix\_data (

id int,

show\_id varchar(255),

type varchar(255),

title varchar(255),

director varchar(255),

country varchar(255),

date\_added varchar(255),

release\_year varchar(255),

rating varchar(255),

duration varchar(255),

listed\_in varchar(255)

);

*There are 8790 rows of data in the Netflix\_data table under Netflix database*

*-- displays the number of rows present in the netflix\_data table*

*select count(\*) from netflix\_data; -- 8790 rows, Duplicates:0 Warnings:0*

*There are no missing values found in the dataset. When I insert data into Netflix\_data table, it shows there are Duplicates:0 and Warnings:0*

*-- describes the table fields, data types, Null values and keys if any*

*DESCRIBE netflix\_data;*

A screenshot of a computer

Description automatically generated

Segment 2: Content Analysis

* Analyse the distribution of content types (movies vs. TV shows) in the dataset.
* Determine the top 10 countries with the highest number of productions on Netflix.
* Investigate the trend of content additions over the years.
* Analyse the relationship between content duration and release year.
* Identify the directors with the most content on Netflix.

*The content types are distributed such that there are about 70% movies versus 30% TV Shows in the dataset.*

SELECT type, count(\*) AS count

FROM netflix\_data

GROUP BY type;

|  |  |
| --- | --- |
| Movie | 6126 |
| TV Show | 2664 |

*Here are the top 10 countries with the highest number of productions on Netflix.*

SELECT country, count(\*) AS production\_count

FROM netflix\_data

GROUP BY country

ORDER BY production\_count DESC

LIMIT 10;

|  |  |
| --- | --- |
| United States | 3240 |
| United Kingdom | 638 |
| Spain | 182 |
| South Korea | 214 |
| Pakistan | 421 |
| Not Given | 287 |
| Japan | 259 |
| India | 1057 |
| France | 213 |
| Canada | 271 |

*The trend of content additions over the years looks like it starts with a single digit from the year 1925 and goes until year 2021 where there are about 500+ content additions.*

SELECT release\_year, count(\*) as content\_additions from netflix\_data

group by release\_year order by release\_year;

*The release year starts from 1925 and ends at 2021 and the content duration related to it has number of hours /seasons present in each TV Show and Movies.*

*The directors with most content are as shown below:*

select director, count(type) as content

from netflix\_data

group by director

order by content desc

limit 10;

|  |  |
| --- | --- |
| Rajiv Chilaka | 20 |
| Alastair Fothergill | 18 |
| Raúl Campos, Jan Suter | 18 |
| Marcus Raboy | 16 |
| Suhas Kadav | 16 |
| Jay Karas | 14 |
| Cathy Garcia-Molina | 13 |
| Martin Scorsese | 12 |
| Youssef Chahine | 12 |

A screenshot of a computer

Description automatically generated

Segment 3: Genre and Category Analysis

* Determine the unique genres and categories present in the dataset.
* Calculate the percentage of movies and TV shows in each genre.
* Identify the most popular genres/categories based on the number of productions.
* Calculate the cumulative sum of content duration within each genre.

*The unique genres in this dataset are ‘TV Show’ or ‘Movies’ that have various categories present as shown in the screenshot below.*

SELECT TYPE as genres,listed\_in as categories

FROM netflix\_data

GROUP BY type,listed\_in; -- 513 rows returned

*The percentage of Movie are about 70% and TV Show are the remaining 30% as shown below.*

SELECT TYPE,COUNT(show\_id)\*100/(select count(show\_id) from netflix\_data) as percentage

FROM netflix\_data

GROUP BY type; -- Movie 70% and TV show 30%

|  |  |
| --- | --- |
| Movie | 69.6928 |
| TV Show | 30.3072 |

*The most popular genre based on the number of productions are retrieved as shown in the screenshot below.*

SELECT listed\_in AS genre\_category, count(\*) AS production\_count

FROM netflix\_data

GROUP BY listed\_in

ORDER BY production\_count DESC;

A screenshot of a computer

Description automatically generated

*The cumulative sum of content duration within each genre is as listed below:*

*SELECT type,sum(duration) AS cumulative\_sum*

*FROM netflix\_data*

*GROUP BY type;*

|  |  |
| --- | --- |
| Movie | 610057 |
| TV Show | 4667 |

A screenshot of a computer

Description automatically generated

Segment 4: Release Date Analysis

* Determine the distribution of content releases by month and year.
* Analyse the seasonal patterns in content releases.
* Identify the months and years with the highest number of releases.

*The distribution of content releases by month and year is shown in the screenshot below.*

SELECT

MONTH(STR\_TO\_DATE(date\_added, '%m/%d/%Y')) AS month,

YEAR(STR\_TO\_DATE(date\_added, '%m/%d/%Y')) AS year,

COUNT(\*) AS count

FROM Netflix\_data

GROUP BY month, year;

A screenshot of a computer

Description automatically generated

*/\*Analyse the seasonal patterns in content releases. \*/*

SELECT

MONTH(STR\_TO\_DATE(date\_added, '%m/%d/%Y')) AS month,

COUNT(\*) AS content\_count

FROM Netflix\_data

GROUP BY month

ORDER BY content\_count DESC;

|  |  |
| --- | --- |
| 7 | 827 |
| 12 | 812 |
| 9 | 769 |
| 4 | 763 |
| 10 | 760 |
| 8 | 754 |
| 3 | 741 |
| 1 | 737 |
| 6 | 728 |
| 11 | 705 |
| 5 | 632 |
| 2 | 562 |

*/\*Identify the months and years with the highest number of releases.\*/*

SELECT

MONTH(STR\_TO\_DATE(date\_added, '%m/%d/%Y')) AS month,

YEAR(STR\_TO\_DATE(date\_added, '%m/%d/%Y')) AS year,

COUNT(\*) AS Highest\_release

FROM Netflix\_data

GROUP BY month, year

ORDER BY Highest\_release DESC;

Segment 5: Rating Analysis

* Investigate the distribution of ratings across different genres.
* Analyse the relationship between ratings and content duration.

*-- Investigate the distribution of ratings across different genres:*

*This query will provide a list of genres, their corresponding ratings, and the count of each rating within each genre.*

*This way, you can observe how different ratings are distributed across various genres on Netflix.*

SELECT listed\_in,

rating,

COUNT(\*) AS rating\_count

FROM netflix\_data

GROUP BY listed\_in, rating

ORDER BY rating\_count desc;

*-- Analyse the relationship between ratings and content duration.*

*Breakdown of ratings for each content duration.*

*You can use this information to see how ratings are distributed across different content durations.*

*In this situation, exploring visualizations (graphs) will be helpful to better understand the distribution of ratings and their relationship with content duration.*

SELECT duration, listed\_in,

rating,

COUNT(\*) AS rating\_count

FROM netflix\_data

GROUP BY listed\_in ,duration, rating

ORDER BY listed\_in , duration, rating;

A screenshot of a computer

Description automatically generated

Segment 6: Co-occurrence Analysis

* Identify the most common pairs of genres/categories that occur together in content.
* Analyse the relationship between genres/categories and content duration.

*Identifies the most common pairs of genres/categories that occur together in content.*

SELECT \* FROM (

SELECT DISTINCT type, listed\_in, count(\*) OVER (PARTITION BY type, listed\_in) AS cnt

FROM netflix\_data

)t

WHERE cnt > 1

ORDER BY type, listed\_in;

*Analysing the relationship between genres/categories and content duration.*

SELECT \* FROM

(

SELECT distinct type,listed\_in,duration,

count(\*) over (partition by type,listed\_in,duration) as co\_occurence

FROM netflix\_data

) t

where co\_occurence>1

ORDER BY type,listed\_in,duration;

SELECT \* FROM

(

SELECT Distinct type, listed\_in,

count(\*) OVER (PARTITION BY type, listed\_in) AS co\_occurence

FROM netflix\_data

) t

WHERE co\_occurence > 1

ORDER BY type, listed\_in;

A screenshot of a computer

Description automatically generated

Segment 7: International Expansion Analysis

* Identify the countries where Netflix has expanded its content offerings.
* Analyse the distribution of content types in different countries.
* Investigate the relationship between content duration and country of production.

*-- Identify the countries where Netflix has expanded its content offerings.*

SELECT DISTINCT country

FROM netflix\_data

WHERE date\_added IS NOT NULL;

*-- Analyse the distribution of content types in different countries.*

SELECT country,

COUNT(CASE WHEN type ='Movie' THEN 1 END) AS num\_movies,

COUNT(CASE WHEN type = 'TV Show' THEN 1 END) AS num\_tv\_shows

FROM netflix\_data

GROUP BY country;

*-- Investigate the relationship between content duration and country of production.*

SELECT country, AVG(CASE WHEN type = 'Movie' THEN duration\_minutes END) AS avg\_movie\_duration,

AVG(CASE WHEN type ='TV Show' THEN duration\_seasons END) AS avg\_tv\_show\_duration

FROM (

SELECT \*, CASE WHEN type = 'Movie' THEN CAST(SUBSTRING\_INDEX(duration, '', 1) AS UNSIGNED) END AS duration\_minutes,

CASE WHEN type = 'TV Show' THEN CAST(SUBSTRING\_INDEX(duration, '', 1) AS UNSIGNED) END AS duration\_seasons

FROM netflix\_data ) AS data

GROUP BY country;

A screenshot of a computer

Description automatically generated

Segment 8: Recommendations for Content Strategy

* Based on the analysis, provide recommendations for the types of content Netflix should focus on producing.

*Recommendations for Content Strategy Based on the analysis, provide recommendations for the types of content Netflix should focus on producing.*

*Dramas, International movies and Documenteries have the most content compared to Thriller, Horror and Crime TV Shows*

SELECT listed\_in, COUNT(\*) AS Content\_count

FROM Netflix\_data

GROUP BY listed\_in

ORDER BY Content\_count ;

*Identify potential areas for expansion and growth based on the analysis of the dataset.*

*Potential areas for expansion could be focused in European regions based on the dataset analysis.*

SELECT country, COUNT(\*) AS count

FROM Netflix\_data

WHERE country IS NOT NULL

GROUP BY country

ORDER BY count DESC

LIMIT 10;

|  |  |
| --- | --- |
| United States | 3240 |
| India | 1057 |
| United Kingdom | 638 |
| Pakistan | 421 |
| Not Given | 287 |
| Canada | 271 |
| Japan | 259 |
| South Korea | 214 |
| France | 213 |
| Spain | 182 |

A screenshot of a computer

Description automatically generated