Netflix Data Analysis

Understanding Viewer Preferences and Performance

By - Chirag

Content Table

Introduction

Welcome to an insightful journey into Netflix’s data universe! This project leverages data analytics to explore Netflix’s content library. By examining content trends, directorial influence, and viewer preferences, we aim to unveil data-driven strategies that enhance both user satisfaction and platform growth.

Objective

- To analyze Netflix’s content library for meaningful trends and patterns.
- To evaluate the impact of directors, genres, and content attributes on viewer engagement.
- To provide actionable insights for strategic decision-making.

Scope

- Focused analysis on Netflix’s movie and TV show metadata.
- Highlighting key metrics like genre preferences, directorial trends, and global content dis

Workflow

Data Collection Aggregated Netflix content metadata for structured analysis.

Enhance data quality, removed duplicates, and standardized data types. **Data Cleaning**

Analysis Execution Applied SQL queries to derive key insights.

Presented findings with actionable business recommendations. **Insight Compilation**

Data Collection and Loading

Data Sources

The Netflix dataset includes comprehensive metadata on **movies, TV shows, directors, cast, genres, countries, durations, and ratings**, providing a rich foundation for analysis.

Data Loading Process

- The dataset was initially imported into Python for preprocessing, where essential cleaning and transformation tasks were executed.
- After preprocessing, the refined data was migrated to the raw data layer in SQL Server, ensuring a robust and scalable environment for analysis.

DATA CLEANING

1. Enhancements in Netflix Data Table

1. Title Column Upgrade

- Conversion to Support Multilingual Titles: Transformed the title column from varchar to nvarchar, ensuring accurate representation of titles with foreign and special characters.
- Enhanced global compatibility for title data.

2. Optimized Column Lengths

- Efficiency in Data Storage: Analyzed and optimized column lengths using Python to fit actual data requirements, significantly reducing unnecessary storage overhead.
- Transitioned from default maximum lengths to calculated optimal lengths, resulting in faster data loading and processing.

2. Primary Key Assignment for Data Integrity

- Unique Identifier Implementation: Designated the show_id column as the primary key to guarantee data integrity and uniqueness across the dataset.
- Ensures accurate linkage and retrieval of records for robust analysis.

```sql
CREATE TABLE netflix (
    show_id VARCHAR(10) PRIMARY KEY,
    type VARCHAR(10),
    title NVARCHAR(200),
    director VARCHAR(250),
    cast VARCHAR(1000),
    country VARCHAR(150),
    date_added VARCHAR(20),
    release_year INT,
    rating VARCHAR(10),
    duration VARCHAR(10),
    listed_in VARCHAR(100),
    description VARCHAR(500)
);
```

```sql
SELECT * FROM netflix;
```

2. Remove Duplicates

This SQL query efficiently selects only unique rows for each title in the Netflix data table by assigning row numbers partitioned by title. This ensures that duplicates are identified and appropriately removed while preserving data integrity and reliability.

```
SELECT show_id, COUNT(*)
FROM netflix
GROUP BY show_id
HAVING COUNT(*) > 1;
)
SELECT *
FROM cte
WHERE rn = 1;
```

```
WITH cte AS (
    SELECT *,
        ROW_NUMBER() OVER (PARTITION BY title, type ORDER BY show_id) AS rn
    FROM netflix
)
SELECT *
FROM cte
WHERE rn = 1;
```

```sql
-- List records where the combination of `title` and `type` has duplicates.
SELECT * FROM netflix
WHERE CONCAT(title, type) IN (
    SELECT CONCAT(title, type)
    FROM netflix
    GROUP BY CONCAT(title, type)
    HAVING COUNT(*) > 1
)
ORDER BY title;
```

```sql
-- List unique combinations of `title` and `type`.
SELECT title, type
FROM netflix
GROUP BY title, type
HAVING COUNT(*) > 1;
```

```sql
-- Identify and list duplicate records based on `show_id`.
SELECT show_id, COUNT(*)
FROM netflix
GROUP BY show_id
HAVING COUNT(*) > 1;

-- Identify and list duplicate records based on the `title`.
SELECT title, COUNT(*)
FROM netflix
GROUP BY title
HAVING COUNT(*) > 1;
```

3. Data type conversions for ‘date added’

The data type of the "Date Added" column was changed from varchar to date to enhance data consistency and facilitate accurate date handling.

```
with cte as(
select *
ROW_NUMBER() OVER(PARTITION BY title,
type order by show_id) as rn
from netflix
)
select show_id, type, title,
cast (date_added as date) as date_added,
release_year, rating, duration, description
from cte
where rn = 1
```

DATA ANALYSIS

Directors: TV Shows vs. Movies

Identified directors who have created both TV shows and movies, with separate counts for each.

```sql
SELECT nd.director,
       COUNT(DISTINCT CASE WHEN nf.type = 'Movie' THEN nf.show_id END) AS no_of_movies,
       COUNT(DISTINCT CASE WHEN nf.type = 'TV Show' THEN nf.show_id END) AS no_of_tvshow
FROM netflix_directors AS nd
INNER JOIN netflix_f AS nf ON nd.show_id = nf.show_id
GROUP BY nd.director
HAVING COUNT(DISTINCT nf.type) > 1;
```

Output table

Director	No. of Movies	No. of TV Shows
Abhishek Chaubey	4	1
Alastair Fothergill	1	3
Alban Teurlai	1	1
Alessandro Angulo	1	1
Andrew Tan	1	1
Anurag Kashyap	8	1

Comedy Movies by Country

Determined the country producing the highest number of comedy movies.

```sql
SELECT nc.country,
       COUNT(DISTINCT ng.show_id) AS no_of_movies
FROM netflix_genre ng
INNER JOIN netflix_countries nc ON ng.show_id = nc.show_id
INNER JOIN netflix_f n ON ng.show_id = n.show_id
WHERE ng.genre = 'Comedies' AND n.type = 'Movie'
GROUP BY nc.country
ORDER BY no_of_movies DESC
LIMIT 1;
```

The United States of America has the highest numbers of comedy moive. (ie, 685)

Directors with Maximum Annual Releases

Analyzed directors with the highest number of movie releases annually, based on the "Date Added" column.

```sql
WITH cte AS (
    SELECT nd.director,
           YEAR(n.date_added) AS date_year,
           COUNT(n.show_id) AS no_of_movies
    FROM netflix n
    INNER JOIN netflix_directors nd ON n.show_id = nd.show_id
    WHERE n.type = 'Movie'
    GROUP BY nd.director, YEAR(n.date_added)
),
cte2 AS (
    SELECT *,
           ROW_NUMBER() OVER (PARTITION BY date_year ORDER BY no_of_movies DESC, director) AS rn
    FROM cte
)
SELECT *
FROM cte2
WHERE rn = 1;
```

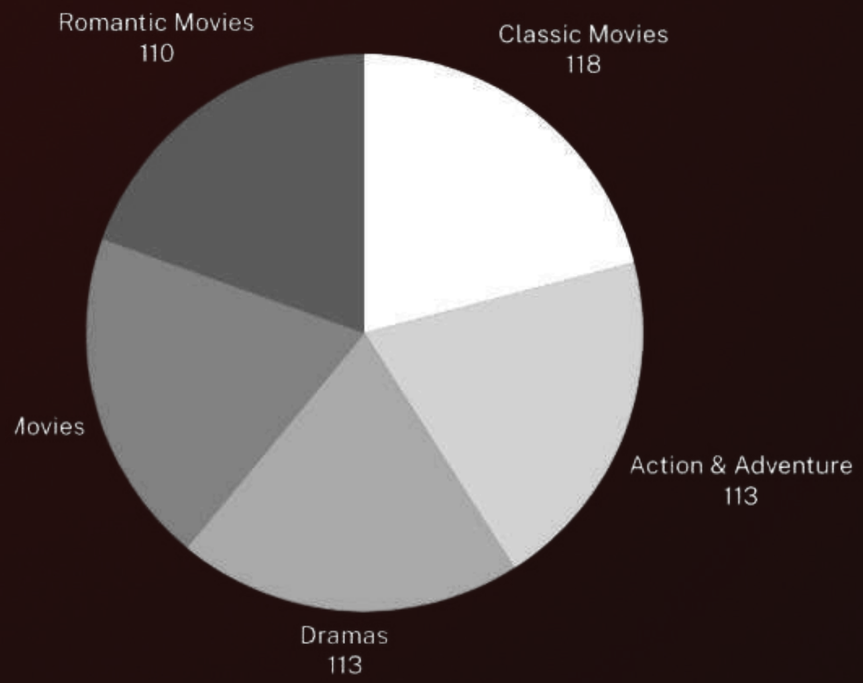
Output table

Director	Year	No. of Movies
Jan Suter	2016	4
Jay Chapman	2017	7
Jan Suter	2018	12
Cathy Garcia-Molina	2019	7
Youssef Chahine	2020	11
Rajiv Chilaka	2021	17

Average Movie Duration by Genre

Calculated the average duration of movies across various genres.

```sql
SELECT ng.genre,
       AVG(CAST(REPLACE(duration, ' min', '') AS UNSIGNED)) AS avg_duration
FROM netflix n
INNER JOIN netflix_genre ng ON n.show_id = ng.show_id
WHERE n.type = 'Movie' AND duration LIKE '% min'
GROUP BY ng.genre
ORDER BY avg_duration DESC;
```



Directors Overlapping in Comedy and Horror Genres

Listed directors creating both comedy and horror movies, along with their respective counts.

```sql
SELECT nd.director,
       COUNT(DISTINCT CASE WHEN ng.genre = 'Comedies' THEN n.show_id END) AS no_of_comedy,
       COUNT(DISTINCT CASE WHEN ng.genre = 'Horror Movies' THEN n.show_id END) AS no_of_horror
FROM netflix n
INNER JOIN netflix_genre ng ON n.show_id = ng.show_id
INNER JOIN netflix_directors nd ON n.show_id = nd.show_id
WHERE n.type = 'Movie' AND ng.genre IN ('Comedies', 'Horror Movies')
GROUP BY nd.director
HAVING COUNT(DISTINCT ng.genre) = 2;
```

Output table

Director	No. of Comedy	No. of TV Horror
Banjong Pisanthanakun	1	3
Don Michael Paul	3	3
Jeff Baena	2	1
Kevin Smith	5	3
Michael Tiddes	4	2
Poj Arnon	3	5

INSIGHTS & RESULTS

Key Findings

1. Content Preferences

Comedy movies dominate U.S. content preferences, suggesting opportunities for localized content acquisition.

2. Directorial Excellence

Identified top-performing directors, emphasizing their role in content diversity and audience engagement.

2. Genre Analysis

Provided a genre-wise breakdown of movie durations, aiding in tailoring content to viewer habits.

Business Impact

- Enhanced Content Relevance: Focusing on popular genres like comedy boosts viewer engagement.
- Strategic Partnerships: Collaborating with high-impact directors broadens audience reach.
- Operational Efficiency: Streamlining content lengths optimizes viewer satisfaction.
- Informed Decision-Making: Data-backed insights enable targeted investments and competitive positioning.

Thank You!

Thank you for exploring this presentation on Netflix data analysis. Your interest and engagement are invaluable as we continue to uncover the transformative potential of data-driven insights.

---

c.sharma251201@gmail.com +91 852-705-4320 www.linkedin.com/in/chirag2501/