Netflix Data Analysis

Understanding Viewer Preferences and Performance

By - Aman Sharma

Content Table

Introduction

This project focuses on analyzing Netflix data to uncover viewer behaviors and content trends.
By refining data types, resolving inconsistencies, and leveraging SQL queries, the analysis aims to provide actionable insights for optimizing content strategies and enhancing user experience on the platform.

Objectives

- Enhance Data Quality and Consistency
- Deepen Understanding of Viewer Preferences
- Optimize Operational Efficiency
- Drive Strategic Decision-Making

Scope

- Enhance Data Quality and Consistency Deepen Understanding of Viewer Preferences Optimize Operational Efficiency Drive Strategic Decision-Making

Workflow

- Data Collection and Loading
- Data Cleaning and Transformation
- Data Analysis and Insights
- Data Storage and Management

Data Collection and Loading

Data Sources

- Netflix dataset containing information about movies, TV shows, directors, cast, genres, countries, duration, and ratings.

Data Loading Process

- Data was initially loaded into Python for preprocessing and then imported into the raw data layer of SQL Server.

DATA CLEANING

1.Enhancements in Netflix Data Table

Changed Data Type of Title Column

- Converted ‘title’ column from ‘varchar’ to ‘nvarchar’ to support foreign characters.
- Ensured visibility and correct representation of all characters in titles.

Adjusted Column Lengths

- Determined optimal column lengths using Python to improve loading speed.
- Changed from max length to required length for efficient data handling.

Set Primary Key for show_id

- Identified ‘show_id’ as unique across the dataset.
- Established ‘show_id’ as the primary key to ensure data integrity and uniqueness.

DATA CLEANING

```
create TABLE [dbo].[netflix](
  [show_id] [varchar](max) primary key,
  [type] [varchar](max) NULL,
  [title] [varchar](max) NULL,
  [director] [varchar](max) NULL,
  [cast] [varchar](max) NULL,
  [country] [varchar](max) NULL,
  [date_added] [varchar](max) NULL,
  [release_year] [int] NULL,
  [rating] [varchar](max) NULL,
  [duration] [varchar](max) NULL,
  [listed_in] [varchar](max) NULL,
  [description] [varchar](max) NULL
)
```

```
create TABLE [dbo].[netflix](
  [show_id] [varchar](10) primary key,
  [type] [varchar](10) NULL,
  [title] [nvarchar](200) NULL,
  [director] [varchar](250) NULL,
  [cast] [varchar](1000) NULL,
  [country] [varchar](150) NULL,
  [date_added] [varchar](20) NULL,
  [release_year] [int] NULL,
  [rating] [varchar](10) NULL,
  [duration] [varchar](10) NULL,
  [listed_in] [varchar](100) NULL,
  [description] [varchar](500) NULL
)
```

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, thereby ensuring duplicates are handled appropriately.

```sql
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
```

3. Refactoring Columns with Multiple Values into Separate Tables

```
select show_id, trim(value) as director
into netflix_directors
from netflix
cross apply string_split(director, ',')
```

To improve data structure and normalization, columns containing multiple values separated by commas were refactored into separate tables. This enhances data integrity and allows for more efficient querying and analysis of individual values.

netf1îx_gen

neïflix

(listed in,)

netf1îx_cas

netf1ix

(cast,)

show_îd, ( ) count:r'y

netftix countries

netflix

(country,)

4. 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
```

5. Populating missing values in Country Column

Null values in the "Country" column were filled by analyzing the "Director" column, assuming that directors typically create content in their own country.
This approach improves data completeness and enhances geographical insights within the dataset.

```
insert into netflix_countries
select show_id, m.country
from netflix as n
inner join (
select director, country
from netflix_countries as nc
inner join netflix_directors as nd
on nc.show_id = nd.show_id
group by director, country
) as m
on n.director = m.director
where n.country is null
```

6. Populating missing values in Duration Column

An issue was identified where null values in the "Duration" column were incorrectly placed in the "Rating" column. This discrepancy was resolved using a ‘CASE’ statement, ensuring that each column contains the appropriate data without compromising data integrity or analysis accuracy.

```sql
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,
case when duration is null
then rating
else duration
end as duration,
description
into netflix_f
from cte
```

DATA ANALYSIS

1. For each director count the number of movies and tv shows created by them in separate columns for directors who have created tv shows and movies both

```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
```

Top 6 rows of the 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

2. Which country has highest number of comedy movies

```sql
select  top 1 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=nc.show_id
where ng.genre='Comedies' and n.type='Movie'
group by  nc.country
order by no_of_movies desc
```

Country	No. of Movies
United States	685

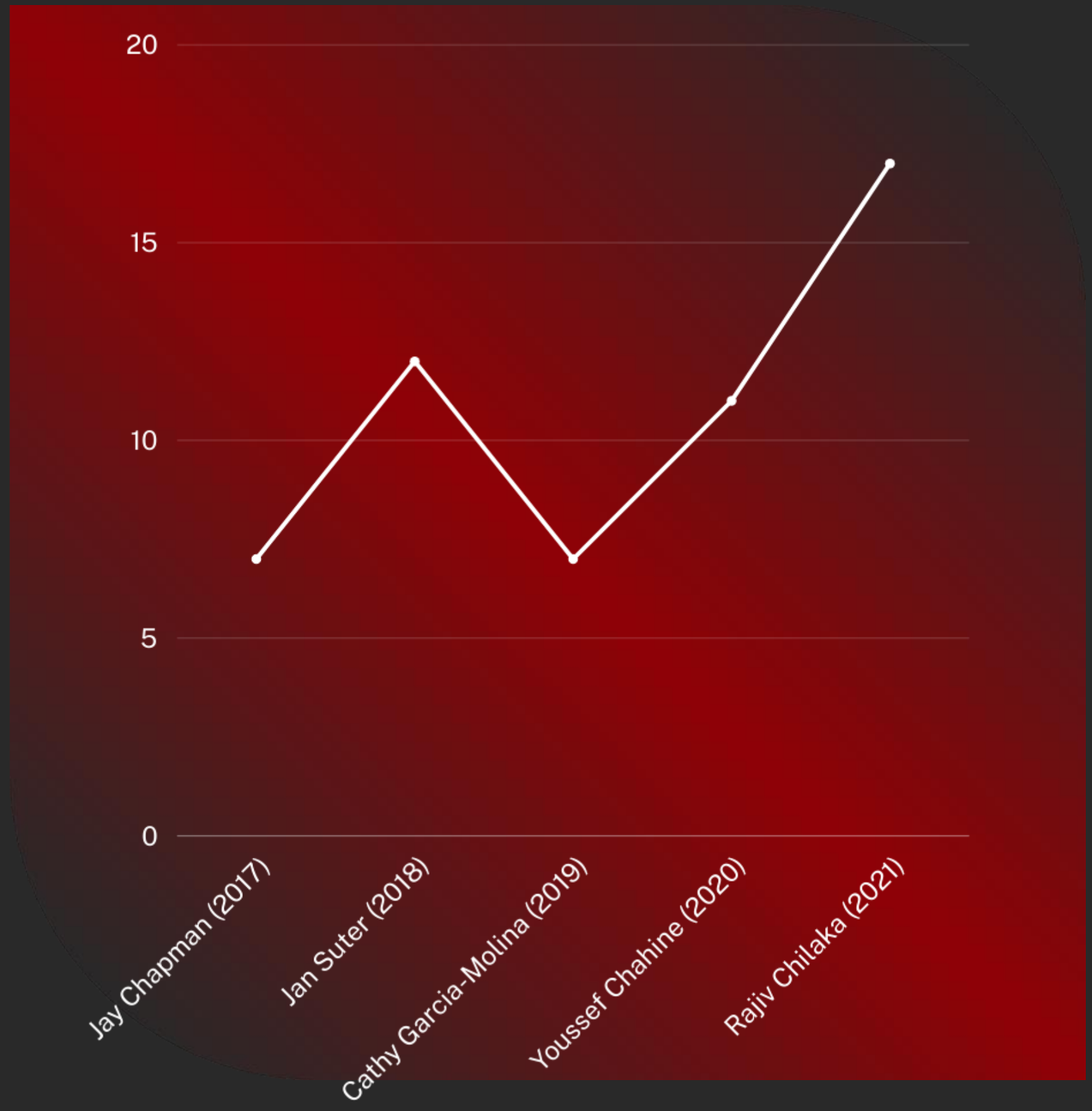
3. For each year (as per date added to netflix), which director has maximum number of movies released

```sql
with cte as (
select nd.director,YEAR(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 type='Movie'
group by nd.director,YEAR(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
```

Last 6 rows of the 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

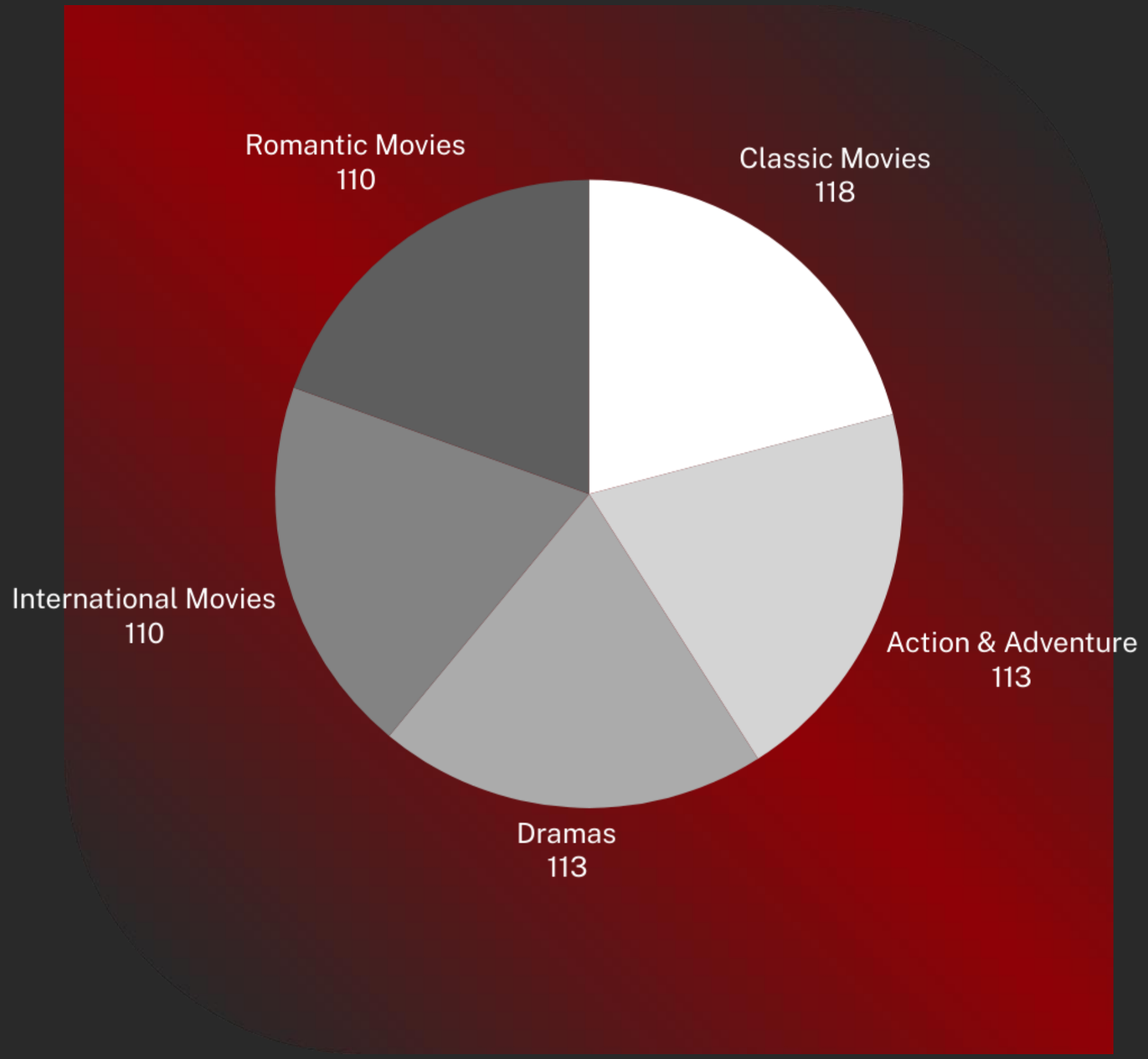
Directors with highest number of movies in respective year



4. What is the average duration of movies in each genre

```sql
select ng.genre , avg(cast(REPLACE(duration,' min','')
 AS int)) as avg_duration
from netflix n
inner join netflix_genre ng
on n.show_id=ng.show_id
where type='Movie'
group by ng.genre
order by avg_duration desc
```

Top 5 movies with highest average duration time in minutes



5. Find the list of directors who have created horror and comedy movies both.

Display director names along with number of comedy and horror movies directed by them.

```
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 type='Movie' and ng.genre in
('Comedies','Horror Movies')
group by nd.director
having COUNT(distinct ng.genre)=2;
```

6 random rows from the 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 and Results

Key Findings

- **Content Preferences:** Highlighted the popularity of comedy movies in the United States, guiding localized content acquisition strategies.
- **Directorial Excellence:** Recognized prolific directors and their impact on Netflix’s content diversity and viewer engagement.
- **Genre Analysis:** Provided insights into viewer preferences based on genre and average duration, informing content curation strategies.

Business Impact

1. Enhanced Content Relevance:
   - Targeting popular genres like comedy in the United States improves viewer engagement.

2. Strategic Partnerships
   - Collaborating with top directors diversifies content and attracts broader audiences.

3. Operational Efficiency
   - Optimizing content length enhances viewer satisfaction and retention.

4. Informed Decision-Making
   - Data-driven insights enable precise resource allocation and competitive positioning.

For exploring this presentation on the Netflix data analysis project. Your interest and engagement are greatly appreciated as we continue to uncover valuable insights from the data.

By - Aman Sharma