Contents

Netflix Viewer Behavior Analytics

1.0 Introduction

1.1 Project Background

OTT (Over-The-Top) platforms have revolutionized content consumption, with Netflix being one of the most popular globally. Understanding viewer behavior and content trends is critical for decision-making, content strategy, and improving user engagement.

1.2Project Objective

The objective of this Capstone project is to analyze Netflix's content library to derive insights into:

- Content type distribution (Movies vs TV Shows)
- Genre popularity
- Country-wise production trends
- Temporal trends in content addition (month/year)
- Rating distributions

The project integrates **Python for data cleaning**, **SQL for structured querying**, and **Power BI for interactive dashboard visualization**.

1.3 Dataset Description

The dataset is obtained from Kaggle, containing **8,807 rows** and the following key attributes:

Column	Description	Example
show_id	Unique ID for each title	s8807
type	Content type (Movie/TV Show)	Movie
title	Title name	Zubaan
director	Director of the title	Rajiv Chilaka
cast	Lead actors	David Attenborough
country	Country of production	United States
date_added	Date added to Netflix	January 1, 2020

Column	Description	Example
release_year	Year of release	2017
rating	Content rating	TV-MA
duration	Runtime or number of seasons	90 min / 1 Season
listed_in	Genres/categories	Dramas, International Movies
description	Short description of the content	Paranormal activity at a lush, abandoned property

---

2.0 Data Preprocessing and Cleaning

2.1Loading the Data

- The raw CSV dataset was imported into **Google Colab**.
- Initial exploration using head(), info(), and describe() to understand structure, null values, and duplicates.

2.2 Handling Missing Values

- Columns like director, cast, and country contained missing values.
- Decision: For key analysis, missing values were kept for non-critical columns; others were cleaned where necessary.

2.3 Removing Duplicates

- Checked for duplicate rows — **none were found**, ensuring unique data integrity.

2.4 Creating Helper Columns

- Added Int_Sum = 1 to facilitate counting in SQL and Power BI.
- Converted date columns to proper formats for time-based analysis.

2.5Exploratory Analysis

- Value counts for type, listed_in, country, and rating were examined.
- Initial insights: Movies outnumber TV Shows, USA dominates content production, and Drama is the most frequent genre.

3.0 SQL Analysis

3.1Connecting to SQL

- SQLAlchemy in Python was used to connect the cleaned dataset to a SQL environment.
- Created temporary tables and executed queries for analysis.

3.2 Queries for Analysis

Sample SQL queries include:

1. **Count of Movies vs TV Shows**

```
SELECT type, COUNT(*) AS total
FROM netflix_master
GROUP BY type;
```

2. **Top 10 Countries by Content**

```
SELECT country, COUNT(*) AS total_titles
FROM netflix_master
GROUP BY country
ORDER BY total_titles DESC
LIMIT 10;
```

3. **Most Popular Genres**

```
SELECT listed_in, COUNT(*) AS total_titles
FROM netflix_master
GROUP BY listed_in
ORDER BY total_titles DESC
LIMIT 10;
```

4. **Titles Added Per Month**

```
SELECT month_added, COUNT(*) AS total_titles
FROM netflix_master
GROUP BY month_added
```

ORDER BY FIELD (month_added,
'January','February','March','April','May','June','July','August','September','October','November','December');

- A total of **20 analytical queries** were formulated covering all major insights.

4.0 Power BI Dashboard Development

4.1Loading Data into Power BI

- Imported the **cleaned CSV** and SQL query outputs.
- Verified column types and ensured all measures (Int_Sum, counts) were correctly mapped.

4.2 KPI Cards

KPI	**Measure/Field**	**Description**
Total Movies Total TV Shows	Int_Sum where Type = Movie Int_Sum where Type = TV Show	Total number of movies Total TV Shows
Co-Productions	Int_Sum for multiple countries	Total co-produced titles
Most Popular Genre	DAX measure on Listed_in	Genre with maximum titles

4.3 Charts and Visuals

1. **Distribution of Movies vs TV Shows** – Clustered Bar Chart
    - X-axis: Type, Y-axis: Int_Sum → Count
    - Title: “Distribution of Movies and TV Shows on Netflix”
2. **Titles Released Per Year** – Line Chart
    - X-axis: Release Year Sum, Y-axis: Count
    - Shows content release trends over years.
3. **Titles Added Per Month** – Column Chart
    - X-axis: Month Added, Y-axis: Count
    - It shows peak months for new content.
4. **Top 10 Countries by Titles** – Clustered Bar Chart
    - Axis: Country, Values: Count

    - Highlights top content-producing countries.

5. **Distribution by Rating** – Donut Chart
    - Legend: Rating, Values: Count
    - Shows proportion of TV-MA, PG-13, etc.
6. **Most Popular Genres** – Bar chart
    - Category: Listed_in, Values: Count
    - Shows genre dominance visually.

4.4Filters and Interactivity

- Added **Slicers** for: Type, Country, Release Year
- Enables dynamic filtering across visuals for interactive insights.

5.0Insights and Observations

- Netflix library contains more Movies than TV Shows (6,131 Movies vs …).
- United States produces the majority of content (2,818 titles).
- Drama and International Movies dominate genres.
- TV-MA and TV-14 are the most common ratings.
- Peak months for content additions are January and July.
- Content release trend shows a steady increase over the last decade.

6.0 Conclusion

This Capstone project demonstrates **end-to-end analytics** from raw data to actionable insights:

1. **Data Cleaning** in Python to handle missing values and duplicates.
2. **SQL Analysis** to answer structured analytical questions.
3. **Power BI Dashboard** to visualize patterns, trends, and KPIs.

The dashboard provides a comprehensive view of Netflix's content strategy and viewer behavior insights.

7.0 Files and Deliverables

File	Purpose
Netflix_Capstone.ipynb	Python Notebook for cleaning & SQL queries
Netflix_Master.csv	Cleaned dataset
Netflix_Capstone.pbix	Power BI dashboard
PDF/Word report	Documentation with screenshots