**TOPIC: Netflix Data: Cleaning, Analysis and Visualization**

**Prepared by: Olympia Devi Gurumayum**

**PROJECT REPORT**

**Tools Used:** Python, SQL, Excel and Power BI

**Executive Summary**

This project explores Netflix content metadata to uncover trends in content type, production geography, release timing, genres, ratings and durations. Analysis was performed using Python, SQL/Power Query and Power BI for interactive dashboards.

* Movies dominate Netflix’s catalogue (70%).
* United States and India are top contributors of Netflix content.
* Surge in content additions after 2015, peaking in 2018.- Drama, Comedy, Documentaries dominate genres.
* Majority of content targets mature audiences

**Objectives**

1. Clean and preprocess the dataset for analysis.

2. Conduct exploratory data analysis (EDA).

3. Visualize insights with Python libraries and Power BI.

4. Provide recommendations for content strategy.

1. **Data Description Primary dataset**

Netflix content metadata (CSV) with fields such as type, title, director, cast, country, date\_added, release\_year, rating, duration, listed\_in, description. Key features: Mix of Movies (~70%) and TV Shows (~30%), with strong growth in content after 2015.

1. **Data Cleaning**

**Tools used:** SQL, Power Query Editor and Pandas

* 1. Ensured unique (no duplicates) and Removed nulls for the **“show\_id”** column.
  2. Normalized the data into **1NF** where comma separated values in the column **“listed\_in”** are separated into multiple columns.
  3. Changed the format of **“Date\_Added”** column from text to Date and to the format DD/MM/YYY.
  4. Left Outer Join in Power Query: Mass Replacing the wrong values with the corrected ones

a) Created a new table which contain two columns of **“old\_names”** and **“new\_names”** which contains the old wrong spellings and new corrected spellings of the “director” column.

b) Merged the two tables with left join from the main table to the new table which contains the new and old names.

c) Inserted the null values which comes in the new table with the corresponding values in the director column using M code.

* 1. Added two conditional columns viz **“Movie\_Duration\_Min”** and **“Season\_Duration”** where the numeric durations of the movies and TV shows are mentioned separately.

1. **Exploratory Data Analysis (EDA)**

1. Content type distribution: Movies constitute 70% and TV Shows constitute 30% of the total contents.

2. Content additions over time: sharp increase post-2015.

3. Top producing countries: United States, India and United Kingdom.

4. Genre distribution: International Movie, Drama, Comedy, Documentaries dominate.

5. Ratings: Majority targeted at mature audiences.

1. **Recommendations**

1. Invest in Drama, Comedies, International Movies and Documentaries.

2. Strengthen mature-audience strategies with curated collections.

3. Monitor post-2018 content strategy for quality vs quantity.

4. Push local-content production, especially in India.

5. Use duration-aware recommendations for personalization.