
🔹 Week 3: Data Cleaning and Exploratory Data Analysis (EDA) on Netflix Dataset Using Power BI (Power Query)

🔸 Objective

The objective of Week 3 was to perform **data cleaning and Exploratory Data Analysis (EDA)** on the **Netflix Movies and TV Shows dataset** using **Power BI Power Query Editor**.
This week focused on:

- Cleaning real-world data using Power Query
- Handling missing and inconsistent values
- Understanding data patterns using descriptive analysis
- Preparing clean data for effective visualization

🔸 Tools & Technologies Used

- Power BI Desktop
- Power Query Editor
- Netflix Movies and TV Shows Dataset

🔸 Dataset Overview (Netflix Dataset)

The Netflix dataset contains information about movies and TV shows available on the Netflix platform.
Key columns include:

- Show ID
- Type (Movie / TV Show)
- Title
- Director
- Cast
- Country
- Date Added
- Release Year
- Rating
- Duration

- Genre (Listed In)
- Description

This dataset represents **real-world streaming platform data**, which contains missing values and inconsistent formatting.

🔸 **Step 1: Importing Dataset into Power BI**

- Imported the Netflix dataset into **Power BI Desktop**
- Opened the dataset in **Power Query Editor**
- Analyzed:
    - Column names
    - Data types
    - Null values
    - Data distribution

🔸 **Step 2: Data Cleaning Performed in Power Query**

1️⃣ **Removing Unnecessary Columns**

- Removed columns that were not required for analysis
- Improved dataset clarity and performance

2️⃣ **Handling Missing Values**

- Identified null values in:
    - Director
    - Cast
    - Country
- Replaced missing values with:
    - "Unknown" for categorical fields
- In some cases, rows with excessive missing values were filtered out

---

3️⃣ **Data Type Correction**

- Converted:
    - Date Added → Date format

- Release Year → Whole number
- Ensured correct data types for accurate analysis and visualization

---

4️⃣ Cleaning Text Columns

- Removed extra spaces using **Trim** and **Clean** options
- Standardized text formatting in:
  - Country
  - Genre
  - Rating

5️⃣ Splitting and Transforming Columns

- Split **Duration** column into:
  - Duration Value
  - Duration Type (Minutes / Seasons)
- This helped separate Movies and TV Shows logically

6️⃣ Standardizing Categorical Values

- Standardized inconsistent values in:
  - Rating
  - Type (Movie / TV Show)
- Ensured uniform category names across the dataset

7️⃣ Removing Duplicate Records

- Identified duplicate titles
- Removed duplicate rows to maintain data accuracy

🔸 Step 3: Exploratory Data Analysis (EDA) Using Power BI

After cleaning, EDA was performed directly using Power BI visuals.

1️⃣ Content Type Analysis (Movies vs TV Shows)

- Created a bar chart for content type distribution

Insight:

- Movies dominate Netflix's content library
- Indicates Netflix's stronger focus on movie releases

2️⃣ Release Year Trend Analysis

- Analyzed content growth over years using line charts

Insight:

- Significant increase in content after 2015
- Rapid expansion of Netflix content in recent years

3️⃣ Rating-wise Content Distribution

- Created a bar chart for ratings

Insight:

- TV-MA and TV-14 ratings are most common
- Netflix primarily targets mature and teenage audiences

4️⃣ Country-wise Content Analysis

- Used bar charts to analyze top producing countries

Insight:

- United States leads content production
- Strong presence of international content such as India and UK

5️⃣ Genre Analysis

- Analyzed most common genres using count visuals

Insight:

- Dramas and International Movies are most popular
- Reflects Netflix's global storytelling approach

6️⃣ Duration Analysis

- Compared:
    - Movie durations (minutes)

- TV shows by number of seasons

Insight:

- Movies typically range between 90–120 minutes
- TV Shows mostly have 1–3 seasons

🔸 **Step 4: Key Insights from EDA**

- Movies form the majority of Netflix content
- Netflix content grew rapidly after 2015
- Mature-rated content dominates the platform
- US is the largest producer, but international content is rising
- Dramas and international genres are highly preferred

🔸 **Outcome of Week 3**

By the end of Week 3:

- Successfully cleaned Netflix dataset using Power Query
- Performed structured EDA using Power BI visuals
- Understood content trends and platform strategies
- Strengthened data preparation and analytical skills

🔸 **Key Learning**

“Power Query simplifies real-world data cleaning, while EDA in Power BI helps convert raw data into meaningful business insights.”