# Netflix Dataset -Documentation

## Netflix Dataset Cleaning:

The Netflix dataset cleaning phase focused on preparing the raw titles data for accurate and efficient analysis. Initially, duplicate rows were identified and removed to eliminate redundancy and prevent biased results in statistical summaries or machine learning models. Missing values in important columns such as director, cast, country, rating, and duration were handled by replacing them with appropriate placeholders like “Not Available”, “Unknown”, or “0”, ensuring data completeness without discarding useful entries. Further, data standardization steps were applied — removing leading and trailing spaces from categorical fields like type, rating, and country using string operations, and eliminating unwanted special characters through regular expressions to retain only alphabetic values. Text-based columns such as director were converted to lowercase for uniformity across the dataset. After all these cleaning operations, the refined dataset was saved as cleaned\_netflix\_titles.csv, which served as the foundation for further feature transformation and modeling tasks.

## Feature Normalization and Encoding:

Following the cleaning process, the dataset underwent a series of normalization and encoding steps to make it suitable for machine learning algorithms. The rating column was first label-encoded to convert textual categories such as “TV-MA”, “PG”, and others into numeric representations that models can interpret easily. Additionally, frequency encoding was applied to categorical features like country, rating, and primary\_genre, transforming each category into a numeric value based on its occurrence frequency. This allowed the model to understand the relative importance and distribution of these features without arbitrary assignments. For the listed\_in column, which contained multiple genres per title, MultiLabelBinarization was used to split genres into individual binary columns — assigning 1 if a genre was present and 0 otherwise — thereby converting multi-genre information into a structured numerical format. The final processed data, containing both frequency and one-hot encoded features, was saved as freq\_encoded\_netflix\_titles.csv for advanced analysis and predictive modeling. These normalization techniques enhanced data consistency, improved feature representation, and enabled machine learning models to achieve better accuracy, scalability, and interpretability.

## Exploratory Data Analysis (EDA):

Exploratory Data Analysis (EDA) was performed on the cleaned Netflix dataset to gain meaningful insights and prepare the data for further analysis and model development. The dataset was first loaded and examined to understand the structure, types of columns, and general content information such as titles, type (Movie or TV Show), release year, genre, rating, and country of origin. A detailed analysis of the listed\_in column, which contains multiple genres for each title, was conducted by splitting and expanding the values to identify the most frequent genres across the platform. Bar charts were plotted to visualize both overall and top 10 genre distributions, revealing Netflix’s content preferences and viewer trends. Similarly, the rating column was analyzed to understand content maturity levels, highlighting how different audience categories (like PG, R, and TV-MA) are represented. The type column was explored to determine the ratio between movies and TV shows, helping to understand the nature of content Netflix focuses on. Additionally, the country column was split and analyzed to identify the top countries contributing the most content, providing a geographical perspective of Netflix’s content library.

## Feature Engineering:

In the feature engineering phase, several data transformation steps were performed to improve data quality and model readiness. Multi-valued columns such as listed\_in and country were cleaned and converted into structured categorical features to enhance interpretability. Missing or inconsistent values were handled appropriately to ensure accuracy and reliability in subsequent analysis. These engineered features allow for more granular insights and serve as useful predictors in recommendation systems or content success prediction models. Overall, the EDA and feature engineering processes provided a deeper understanding of the dataset, identified important trends, and refined the data into a more meaningful form. These improvements significantly enhance the potential performance of machine learning models by ensuring cleaner input data, more relevant features, and stronger relationships between variables such as genre, rating, and country of production.

**Overall Conclusion:**

In this project, the Netflix dataset was carefully cleaned, normalized, and analysed to improve its quality and usefulness for further machine learning and data analysis tasks. Data cleaning helped remove duplicates, fill missing values, and standardize text formats, ensuring consistency. Normalization and encoding converted text-based information into numerical values that models can easily understand. Through Exploratory Data Analysis (EDA), important patterns related to genres, ratings, and countries were identified, giving insights into Netflix’s content trends. Feature engineering further improved the dataset by creating meaningful variables for better predictions. Overall, these steps made the dataset more structured, reliable, and ready for building accurate and efficient recommendation or prediction models.