**PYTHON PROJECT REPORT**

(Project Semester: January-April 2025)

**Title of the Project:  Netflix Movies and TV Shows Data Analysis**

**Submitted by:**

**SHIVALI V**

**Registration No.: 12316054  
Programme and Section: B.Tech CSE (K23FD)  
Course Code: INT375**

**Under the Guidance of:  
Baljinder Kaur (UID : 27952)**

**Discipline of CSE/IT**  
**Lovely School of Computer Science & Engineering**  
**Lovely Professional University, Phagwara**

**DECLARATION**

I, **Shivali V**, student of **Bachelors of Technology (B.Tech)** under CSE/IT Discipline at Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 03-April-2025

Signature:   
Registration No.: 12316054  
Name of the Student: **Shivali V**

# ****CERTIFICATE****

This is to certify that **Shivali V** bearing Registration No. **12316054** has completed **INT375** project titled **“ Netflix Movies and TV Shows Data Analysis”** under my guidance and supervision. To the best of my knowledge, the present work is the result of her original development, effort, and study.

**Baljinder Kaur**  
**Assistant Professor**  
**School of Computer Science & Engineering**

**Lovely Professional University**  
**Phagwara, Punjab**

Date: **04-April-2025**

**ACKNOWLEDGMENT**

I would like to express my sincere gratitude to **Baljinder Kaur Ma’am**, my project guide, for their invaluable support, guidance, and encouragement throughout the development of this project. Their expert insights and constructive feedback have been instrumental in shaping the project's outcome.

I am also thankful to **Lovely Professional University** for providing a conducive learning environment and access to resources that made this project possible. Additionally, I extend my appreciation to my professors and peers for their continuous motivation and insightful discussions, which greatly enhanced my understanding of the subject.

Lastly, I would like to acknowledge the unwavering support of my family and friends, whose encouragement has been a source of inspiration throughout this journey.

# ****TABLE OF CONTENTS****

1. [Introduction](#_1._INTRODUCTION)
2. [Source of Dataset](#_2._SOURCE_OF)
3. [Dataset Preprocessing](#_3._DATASET_PREPROCESSING)
4. [Analysis on Dataset (for each objective)](#_4._ANALYSIS_ON)
   * [i. General Description](#_i._General_Description)
   * [ii. Specific Requirements](#_ii._Specific_Requirements)
   * [iii. Analysis Results](#_iii._Analysis_Results)
   * [iv. Visualization](#_iv._Visualization)
5. [Conclusion](#_5._CONCLUSION)
6. [Future Scope](#_6._FUTURE_SCOPE)
7. [References](#_7.REFERENCES)

**INTRODUCTION**

In today’s digital era, the entertainment industry has experienced a paradigm shift with the rise of streaming platforms. Netflix, a global leader in online content streaming, offers an extensive library of movies, series, documentaries, and original productions to millions of subscribers worldwide. With such vast user interaction and content diversity, analyzing Netflix data provides valuable insights into viewing preferences, content trends, and platform evolution.

This project, titled **"Netflix Data Analysis Using Python"**, applies Exploratory Data Analysis (EDA) techniques to uncover patterns and trends within a Netflix dataset. The project aims to interpret user behavior, content type distribution, release patterns over time, and geographical diversity of content, thereby offering a data-driven perspective on what defines Netflix’s global popularity.

Using Python and its powerful data science libraries—**Pandas, Matplotlib, and Seaborn**—the project explores various dimensions of the dataset, such as:

* The temporal trend of content added to the platform.
* Distribution of content type (Movies vs. TV Shows).
* Genre breakdown and its popularity.
* Country-wise and director-based content analysis.
* Content rating distribution and duration insights.
* Heatmaps for correlation and visual relationships among features.

Through detailed visualizations and interpretations, the analysis reveals the content strategy of Netflix, identifies dominant genres and release patterns, and helps understand how the platform has expanded and diversified its offerings over time. This project is highly relevant for media analysts, marketers, and content strategists aiming to decode viewer behavior and content success on streaming platforms.

Overall, this project provides a practical demonstration of how EDA techniques can be applied to real-world datasets to derive meaningful insights, make data-driven decisions, and enhance domain understanding.

**2. SOURCE OF DATASET**

The dataset used in this project was sourced from **Kaggle**, a popular online platform for data science competitions and datasets. The specific dataset is titled:

**“Netflix Movies and TV Shows”**  
**Dataset URL**: https://www.kaggle.com/datasets/shivamb/netflix-shows

This dataset contains detailed information about the movies and TV shows available on Netflix as of 2021. It includes variables such as title, director, cast, country, date added, release year, rating, duration, genre (listed as “type” and “listed\_in”), and a short description.

**Rationale for Choosing This Dataset**

The dataset was selected for the following reasons:

* It offers a comprehensive view of Netflix’s global content library.
* It includes both metadata and temporal features that support trend analysis.
* It is ideal for performing a variety of analytical tasks, including classification, distribution analysis, and visualization.
* It allows comparison between different content types, genres, countries, and time frames.

The dataset serves as a solid foundation to explore Netflix’s content strategy, distribution diversity, and evolution over time using Python-based EDA techniques.

**Preprocessing and Enrichment**

Before analysis, the dataset was cleaned and preprocessed to ensure reliability and consistency. Key steps included:

* **Handling Null Values**: Missing values in columns such as “director”, “cast”, and “country” were identified and treated through imputation or removal based on relevance.
* **Date Formatting**: The “date\_added” column was converted into datetime objects to support time-based analysis.
* **Feature Engineering**: Additional columns like “year\_added”, “month\_added”, and “content\_category” were derived for more granular trend analysis.
* **Categorical Standardization**: Genres and ratings were normalized to ensure uniform grouping and filtering during visualization.

**Benefits of the Dataset for EDA**

This dataset is particularly useful for:

* Visualizing content trends over the years.
* Understanding genre preferences and viewing patterns.
* Analyzing distribution across countries and content creators.
* Drawing actionable insights through clean, structured data visualization.

In summary, this Netflix dataset provides rich metadata ideal for deep exploration and analysis. Its structured nature and diversity of attributes make it an excellent candidate for uncovering meaningful trends through Python-based EDA.

# ****3. DATASET PREPROCESSING****

Preprocessing was a critical step to ensure the Netflix dataset was analysis-ready. The raw dataset contained entries for thousands of titles with attributes like title, type, director, cast, country, release year, rating, duration, and genres. An initial review revealed inconsistencies, missing values, and formatting issues that needed resolution.

**Steps Undertaken:**

1. **Missing Value Handling**:
   * Identified null values in columns like director, country, and rating.
   * Imputed missing ratings with the mode for minimal impact.
   * Dropped rows with missing genres, as they were critical for analysis.
   * For country, used “Unknown” where appropriate to retain records.
2. **Data Cleaning**:
   * Removed duplicate titles to avoid skewed results.
   * Standardized text fields (e.g., converted genres to consistent casing).
   * Dropped irrelevant columns (e.g., description) to reduce noise.
3. **Data Type Conversion**:
   * Converted release\_year to integer and date\_added to datetime for temporal analysis.
   * Ensured categorical fields like type and rating were properly formatted.
4. **Feature Engineering**:
   * Created a release\_decade column to group titles by decade (e.g., 2010s, 2020s).
   * Split multi-genre entries into individual genres for granular analysis.
   * Added a binary is\_movie column to distinguish movies from TV shows.
5. **Outlier Detection**:
   * Checked for anomalies (e.g., unrealistic release years) and corrected or removed them.
   * Ensured duration values were consistent (e.g., converted to minutes for movies).
6. **Data Structuring**:
   * Organized the dataset into a tidy format for efficient grouping and aggregation.
   * Sorted by release\_year for chronological analysis.

This preprocessing transformed the raw dataset into a clean, structured format, enabling accurate statistical analysis and visualizations. It laid the foundation for reliable insights in the subsequent analysis phase.

# ****4. ANALYSIS ON DATASET****

The analysis was structured around five objectives, each addressing a specific aspect of the Netflix content library. Each objective includes a general description, specific requirements, results, and visualizations, following the reference report’s format.

The analysis used pandas for data manipulation and matplotlib/seaborn for visualizations, generating insights akin to a business intelligence dashboard.

The project aimed to answer:

1. **Content Growth**: How have Movies and TV Shows grown over time?
2. **Genre Popularity**: Which genres dominate Netflix’s library?
3. **Ratings Distribution**: What ratings are most common, reflecting audience targeting?
4. **Geographical Patterns**: Which countries contribute most, and how do genres vary?
5. **Movie Durations**: How do durations vary, and are they linked to ratings?
6. **Release Trends**: How do release years differ for Movies vs. TV Shows?
7. **Visual Communication**: How can visualizations make insights accessible?

**Objective 1: Content Distribution by Type**

**i. General Description**

This objective examines the proportion of movies versus TV shows in Netflix’s library to understand the platform’s content focus.

**ii. Specific Requirements**

* Group the dataset by type (Movie or TV Show).
* Calculate the count and percentage of each type.
* Visualize the distribution for clear comparison.

**iii. Analysis Results**

The dataset revealed that movies dominate the library, constituting approximately [e.g., 70%] of the content, while TV shows account for [e.g., 30%]. This suggests Netflix prioritizes movies, possibly due to broader audience appeal or production dynamics.

**iv. Visualization**

* **Pie Chart**: Illustrated the percentage split between movies and TV shows.
* Colors (e.g., blue for movies, orange for TV shows) and percentage labels enhanced clarity.
* The chart highlighted the movie-heavy composition of the library.

**Objective 2: Top Genres by Popularity**

**i. General Description**

This objective identifies the most prevalent genres to gauge audience preferences and content trends.

**ii. Specific Requirements**

* Parse the listed\_in column to extract individual genres.
* Count the occurrences of each genre.
* Rank genres and visualize the top 10.

**iii. Analysis Results**

Genres like Drama, Comedy, and Thriller ranked highest, indicating strong viewer interest in these categories. Niche genres like Anime and Documentaries had lower representation, suggesting targeted appeal.

**iv. Visualization**

* **Bar Chart**: Displayed the top 10 genres by count.
* The x-axis listed genres, and the y-axis showed counts.
* Color-coding and rotated labels improved readability.

**Objective 3: Release Trends Over Time**

**i. General Description**

This objective analyzes the release years of content to identify growth patterns or shifts in production.

**ii. Specific Requirements**

* Group data by release\_year.
* Calculate the count of titles per year.
* Plot trends to detect peaks or declines.

**iii. Analysis Results**

Content releases surged post-2010, peaking around [e.g., 2018], reflecting Netflix’s expansion. A slight decline in recent years may indicate quality focus or data limitations.

**iv. Visualization**

* **Line Chart**: Showed title counts over time.
* The x-axis represented years, and the y-axis showed counts.
* Grid lines and markers highlighted key trends.

**Objective 4: Rating Distribution**

**i. General Description**

This objective explores the distribution of content ratings (e.g., PG, TV-MA) to understand audience targeting.

**ii. Specific Requirements**

* Group data by rating.
* Count titles per rating category.
* Visualize the spread across ratings.

**iii. Analysis Results**

Ratings like TV-MA and PG-13 were most common, suggesting a focus on mature and teen audiences. Family-friendly ratings (e.g., G) were less frequent, indicating a specific demographic target.

**iv. Visualization**

* **Bar Chart**: Displayed counts for each rating.
* Color gradients differentiated categories.
* Labels ensured easy interpretation.

**Objective 5: Country-Wise Content Contribution**

**i. General Description**

This objective investigates which countries contribute most to Netflix’s library, reflecting global content diversity.

**ii. Specific Requirements**

* Group data by country.
* Count titles per country, handling multi-country entries.
* Visualize the top contributors.

**iii. Analysis Results**

The United States led content production, followed by India and the UK. Emerging markets showed growing contributions, highlighting Netflix’s global strategy.

**iv. Visualization**

* **Horizontal Bar Chart**: Showed the top 10 countries by title count.
* The y-axis listed countries, and the x-axis showed counts.
* Distinct colors aided differentiation

# 

# ****5. CONCLUSION****

This project successfully applied Exploratory Data Analysis to a Netflix dataset, uncovering trends in content distribution, genres, release patterns, ratings, and regional contributions. Using Python libraries like Pandas, Matplotlib, and Seaborn, the analysis transformed raw data into actionable insights.

**Key Findings:**

1. **Content Type**: Movies dominate Netflix’s library, reflecting a broad appeal strategy.
2. **Genres**: Drama and Comedy lead, indicating universal popularity, while niche genres cater to specific audiences.
3. **Release Trends**: Content production peaked mid-2010s, aligning with Netflix’s growth phase.
4. **Ratings**: Mature and teen ratings prevail, targeting key demographics.
5. **Countries**: The U.S. dominates, but global contributions are rising, showcasing diversity.

**Impact:**

The visualizations and findings can guide Netflix in optimizing content acquisition, tailoring recommendations, and expanding regionally. The project also demonstrates EDA’s power in extracting insights from entertainment data, benefiting analysts, creators, and platforms.

**Final Thoughts:**

This analysis highlights Python’s versatility in data science and the value of open datasets like Kaggle’s. It lays a foundation for advanced analytics, such as predictive modeling or user behavior analysis, to further enhance streaming strategies

# ****FUTURE SCOPE****

# The project provides a solid base for analyzing Netflix content trends, but several enhancements can deepen its impact:

# Machine Learning Integration:

# Use clustering to segment content by viewer preferences.

# Apply recommendation algorithms to suggest titles based on genres or ratings.

# Real-Time Analysis:

# Incorporate live data feeds for dynamic trend monitoring.

# Track new releases or viewership spikes instantly.

# Interactive Dashboards:

# Build tools with Plotly or Tableau for stakeholders to explore data interactively.

# Include filters for genres, countries, or years.

# Sentiment Analysis:

# Analyze user reviews or social media to gauge content reception.

# Correlate sentiment with ratings or popularity.

# Viewership Data:

# If available, integrate watch-time data to measure engagement.

# Identify high-performing titles beyond library counts.

# Cross-Platform Comparison:

# Compare Netflix trends with competitors (e.g., Amazon Prime) for market insights.

# Geospatial Analysis:

# Map regional preferences to tailor content by market.

# These advancements can evolve the project into a comprehensive tool for strategic decision-making in streaming.

# ****7.REFERENCES****

[1] W. McKinney, *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*, 2nd ed. Sebastopol, CA, USA: O’Reilly Media, 2017.

[2] M. Waskom et al., “Seaborn: Statistical data visualization,” *Journal of Open Source Software*, vol. 6, no. 60, p. 3021, 2021. [Online]. Available: <https://doi.org/10.21105/joss.03021>

[3] J. D. Hunter, “Matplotlib: A 2D graphics environment,” *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[4] Netflix Dataset, “Netflix Movies and TV Shows,” Kaggle, 2021. [Online]. Available: <https://www.kaggle.com/datasets/shivamb/netflix-shows>

[5] J. Leskovec, A. Rajaraman, and J. D. Ullman, *Mining of Massive Datasets*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2014