Exploratory Data Analysis (EDA) Report

Exploratory Data Analysis (EDA) is an essential step in the data science workflow that helps in understanding the structure, patterns, and behavior of a dataset before applying any machine learning models. In this task, EDA was performed on the *Netflix Movies and TV Shows* dataset using Python libraries such as Pandas, Matplotlib, and Seaborn. The primary objective of this analysis was to explore numerical and categorical features, detect outliers, analyze relationships between variables, and identify features that may be important for prediction.

Dataset Overview

The Netflix dataset contains information about movies and TV shows available on the platform. It includes attributes such as content type, release year, country of production, duration, rating, and genre. The dataset consists of both numerical and categorical variables, making it suitable for a wide range of exploratory techniques. Initial inspection of the dataset showed that it is reasonably well-structured and provides meaningful attributes for understanding content trends on Netflix.

Distribution Analysis of Numerical Features

To understand the behavior of numerical variables, histograms were plotted for features such as the release year. The distribution of release years showed that the majority of Netflix content has been released after 2010, with a significant increase in titles after 2015. This trend reflects Netflix's rapid growth and increased investment in content production in recent years. The analysis indicates that Netflix primarily focuses on modern content rather than older releases.

Analysis of Categorical Features

Categorical variables were analyzed using count plots and bar charts. The comparison between content types revealed that movies dominate the Netflix catalog when compared to TV shows. This suggests that Netflix places a stronger emphasis on movies, although TV shows still play an important role in user engagement.

Country-wise analysis showed that the United States is the largest contributor of content on Netflix, followed by other countries such as India. This highlights Netflix's strong presence in the US market while also indicating its expansion into international markets. The distribution of content across

countries reflects Netflix's global strategy, though content production remains concentrated in a few regions.

Outlier Detection

Box plots were used to identify outliers in numerical features such as movie duration. The analysis showed that most movies have durations ranging between 80 and 120 minutes, which aligns with standard industry practices. However, a few movies had unusually long durations, which were identified as outliers. These outliers are relatively rare and do not significantly impact the overall data distribution, but their detection is important to ensure robust modeling in future steps.

Correlation Analysis

A correlation heatmap was plotted to understand relationships between numerical features such as release year and duration. The results indicated a weak correlation between these variables, suggesting that newer movies are not necessarily longer or shorter in duration. Additionally, no strong correlations were observed among numerical features, implying that there is no multicollinearity issue. This is beneficial for machine learning models, as highly correlated features can negatively affect model performance.

Feature Importance for Prediction

Based on the exploratory analysis, several features were identified as important for potential prediction tasks. The type of content (movie or TV show) plays a crucial role in understanding viewing patterns. Release year helps in identifying trends and popularity over time, while country provides insights into regional preferences. Duration influences user engagement, and rating is important for audience segmentation. These features can be effectively used in recommendation systems or content classification models.

Conclusion

The Exploratory Data Analysis provided valuable insights into the Netflix dataset by uncovering content trends, distributions, and feature relationships. The analysis revealed Netflix's strong focus on recent content, dominance of movies over TV shows, and concentration of content production in a few countries. Outlier detection and correlation analysis ensured a better understanding of

data quality and feature behavior. Overall, the dataset is suitable for further machine learning tasks, and this EDA step establishes a strong foundation for predictive modeling.

Learning Outcome

Through this task, I developed a clear understanding of how to explore real-world datasets using visualization techniques. I learned how to analyze numerical and categorical variables, detect outliers, interpret correlations, and derive meaningful insights that support data-driven decision-making.