Presentation On
Netflix Original Films and IMDB Scores Dataset

Team 1:-

Dipti K C
Sujana Joshi
Bhumika Bista

9 June, 2025

Introduction

Brief about dataset:

- Contains information about Netflix Original Films
- Includes features like Title, Genre, IMDB Score, Premiere Date, Runtime, Language, etc.

Goal:

- Understand the data
- Clean it
- Analyze patterns
- Gain insights from visuals

Data Profiling

- Total Rows and Columns: df.shape
- Data types and non-null values: df.info()
- Sample data: df.head()
- Statistical summary: df.describe()
- Categorical distribution: df['Language'].value_counts()

Data Quality

Bullet Points:

- Missing Values: df.isnull().sum()
  - Example: Missing in Language or Runtime
- Fixes:
  - Filled missing numerical with median
  - Filled categorical with mode
- Duplicates found and removed: df.duplicated().sum()
- Data type conversion: Strings to categorical where needed

Visualizations

- **Univariate:**
  - Histogram of IMDB Scores
  - Count plot of Ratings
- **Bivariate:**
  - Boxplot of IMDB Scores by Rating
- **Correlation:**
  - Heatmap showing correlation between Runtime and IMDB Score

Feature Engineering

Bullet Points:

- Created a new feature: Rating Category (High if IMDB ≥ 7, else Low)
- Helped analyze what kinds of movies get higher scores
- Visualization: Countplot of Rating Category

Key Findings

Example insights:

- Most films have an IMDB score between 5 and 7
- English is the dominant language
- High IMDB scores tend to be in *Documentary* and *Drama* genres
- A few extreme runtimes (outliers) were found and handled
- Correlation between Runtime and IMDB Score is weak

Conclusion

- Summarized the EDA process
- Cleaned and explored key patterns
- Identified helpful visual and statistical insights
- Challenge faced: Unicode errors in CSV and handling missing categorical data

THANK YOU