NETFLIX DATA: CLEANING, ANALYSIS AND VISUALIZATION

DOMAIN: DATA SCIENCE

ABOUT DATASET

- Netflix is a leading streaming service with a vast catalog of movies and TV shows.
- Dataset contains titles from 1925 to 2021.
- Filtered to: 2008–2021 for modern trend analysis
- Cleaned data used for analysis, visualized with Python.
- Goal: Explore trends, genres, and content strategy insights.

TOOLS & TECHNOLOGIES USED

- Programming Language: Python
- Libraries: Pandas, NumPy, Matplotlib, Seaborn, WordCloud, Scikit-learn, NLTK
- Visualization: Plotly, Tableau
- Machine Learning: Random Forest Classifier
- NLP: TF-IDF, LDA Topic Modeling, VADER Sentiment

DATA CLEANING STEPS

- Treated Nulls in important columns (director, country, date_added).
- Dropped duplicates.
- Converted date_added to datetime.
- Removed irrelevant or incomplete rows.
- Created new features for analysis and ML.

EXPLORATORY DATA ANALYSIS (EDA)

- Distribution of Movies vs. TV Shows.
- Top countries with most content.
- Common genres using word cloud and bar plot.
- Year-wise content addition trend.
- Top 10 directors by number of titles.

FEATURE ENGINEERING

- Extracted duration in minutes for movies.
- Counted number of genres per title.
- Used MultiLabelBinarizer for genres.
- Encoded ratings using LabelEncoder.
- Prepared data for classification model.

MACHINE LEARNING MODEL

- Used Random Forest Classifier.
- Target: Type (Movie=1, TV Show=0).
- Features: Genres, Rating, Duration.
- Evaluated using Accuracy and Classification Report.

INTERACTIVE VISUALIZATION

- Created a scatter plot using Plotly.
- X-axis: Duration (min), Y-axis: Number of Genres.
- Colored by Content Type (Movie or TV Show).
- Hover tool shows title and rating.

NLP-BASED CONTENT ANALYSIS

- TF-IDF: Extracted high-importance keywords from content descriptions to identify common themes.
- Word Cloud: Visualizes frequently used words in titles and genres for quick thematic understanding.
- LDA (Topic Modeling): Identified major topic clusters across content using Latent Dirichlet Allocation.
- Sentiment Analysis: Analyzed emotional tone of titles, revealing most were neutral with some positive or negative.

BUSINESS APPLICATIONS

- Optimize recommendations using topic modeling
- Improve search experience with TF-IDF-based metadata
- Guide content investment using genre and sentiment trends

CONCLUSION AND INSIGHTS

- Cleaned and analyzed Netflix titles dataset.
- Identified key trends in genres, years, and directors.
- Built a basic ML classifier to distinguish content type.
- Laid foundation for recommendations and deeper analysis.

FUTURE WORK

- Enhance ML model with more features.
- Use NLP for content description analysis.
- Integrate with external APIs for richer metadata.
- Build a full recommendation system.

THANKYOU !

VISMAYA VT