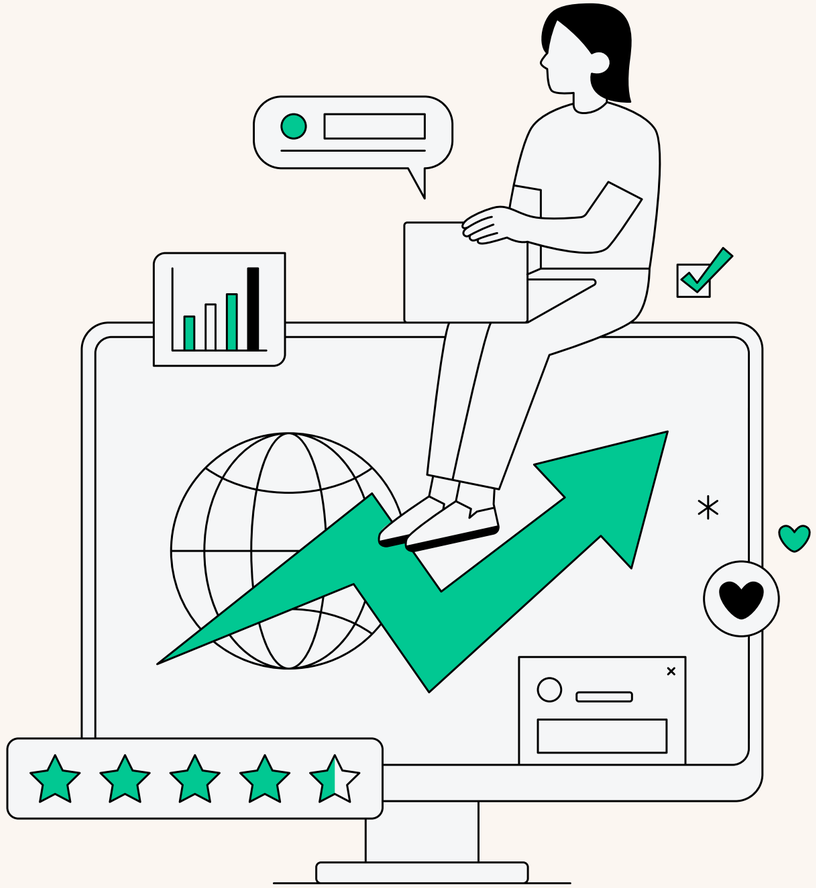
Badve Malhar
Suyog Arlimar
Akash Mestha
Chinmay Barwe
Surabhi Bangad

Data Science

Mini Project Synopsis

Div 1: Batch T3

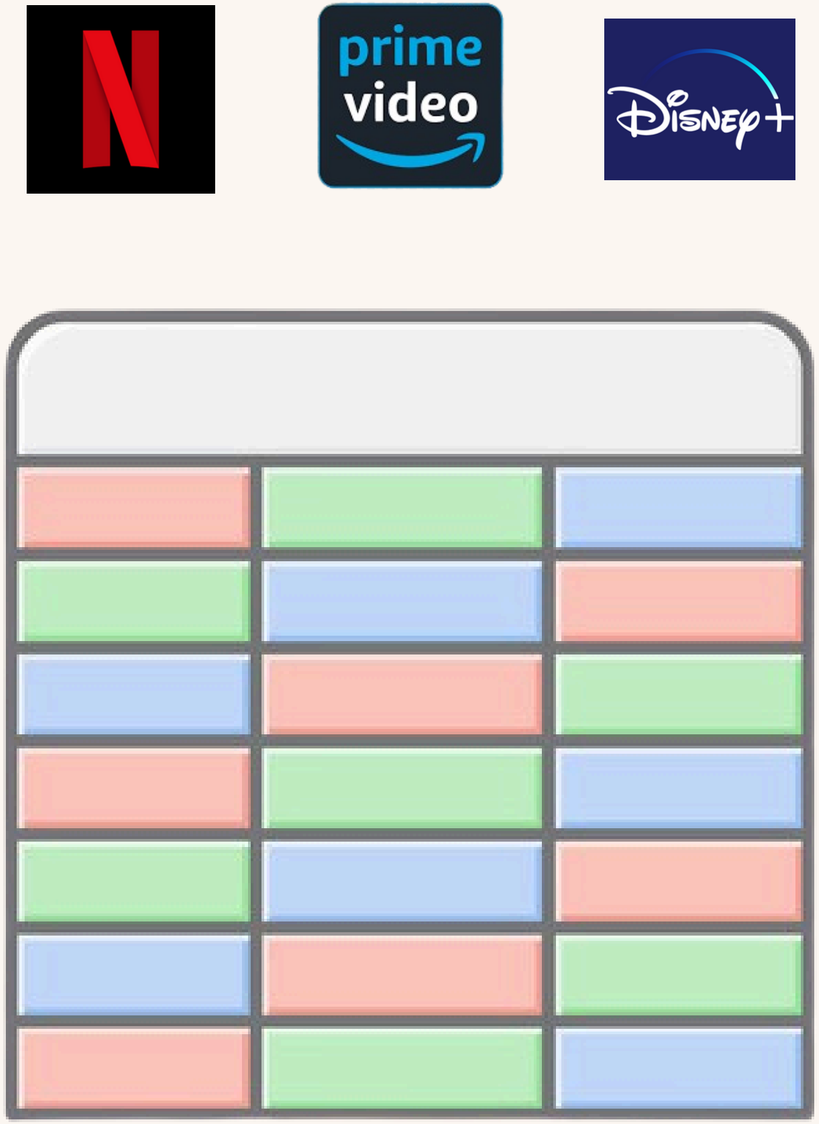


Problem Statement

Analyzing OTT Platform Viewership Data

With the rise of OTT platforms, understanding viewership trends is crucial for content creators, producers, and marketers. This study aims to analyze Netflix and other OTT platform data to derive meaningful insights on viewing habits, genre preferences, and peak watch times.

Dataset Gathering



Source: Kaggle, GitHub, netflix.com

Official viewership and content data gathered from respective OTT platform websites like **Netflix**, **Amazon Prime** and **Disney+**:

- Movies and TV Shows on the Platform
- Viewership data per each country
- Popularity Count of different movies

The data contains information like IMDB ratings of different movies, their runtime, viewer count, rankings, etc.

Dataset Gathering

Example Dataset:
Source: Official Netflix Viewership Database

Netflix is one of the most popular media and video streaming platforms. They have over 8000 movies or tv shows available on their platform, as of mid-2021, they have over 200M Subscribers globally.

Information on the site starts from June 28, 2021 and any lists published before June 20, 2023 are ranked by hours viewed.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5160 entries, 0 to 5159
Data columns (total 11 columns):
 #   Column                      Non-Null Count
---  ------                      --------------
 0   week                        5160 non-null
 1   category                    5160 non-null
 2   weekly_rank                 5160 non-null
 3   show_title                  5160 non-null
 4   season_title                2498 non-null
 5   weekly_hours_viewed         5160 non-null
 6   runtime                     1080 non-null
 7   weekly_views                1080 non-null
 8   cumulative_weeks_in_top_10  5160 non-null
 9   is_staggered_launch         5160 non-null
 10  episode_launch_details      40 non-null
dtypes: bool(1), float64(2), int64(3), object(5)
memory usage: 408.3+ KB
```

Data Preprocessing

This will ensure the data is clean, structured, and ready for analysis!

01.

Handling Missing Values

- Drop missing values (if the missing data is insignificant)
- Fill missing values with mean/median/mode (numerical columns)
- Use forward/backward fill (for time-series data)

02.

Normalizing Data

- Convert timestamps to datetime format
- Normalize numeric columns (e.g., watch time, ratings)
- Extract useful features (like hour, day of the week, month)

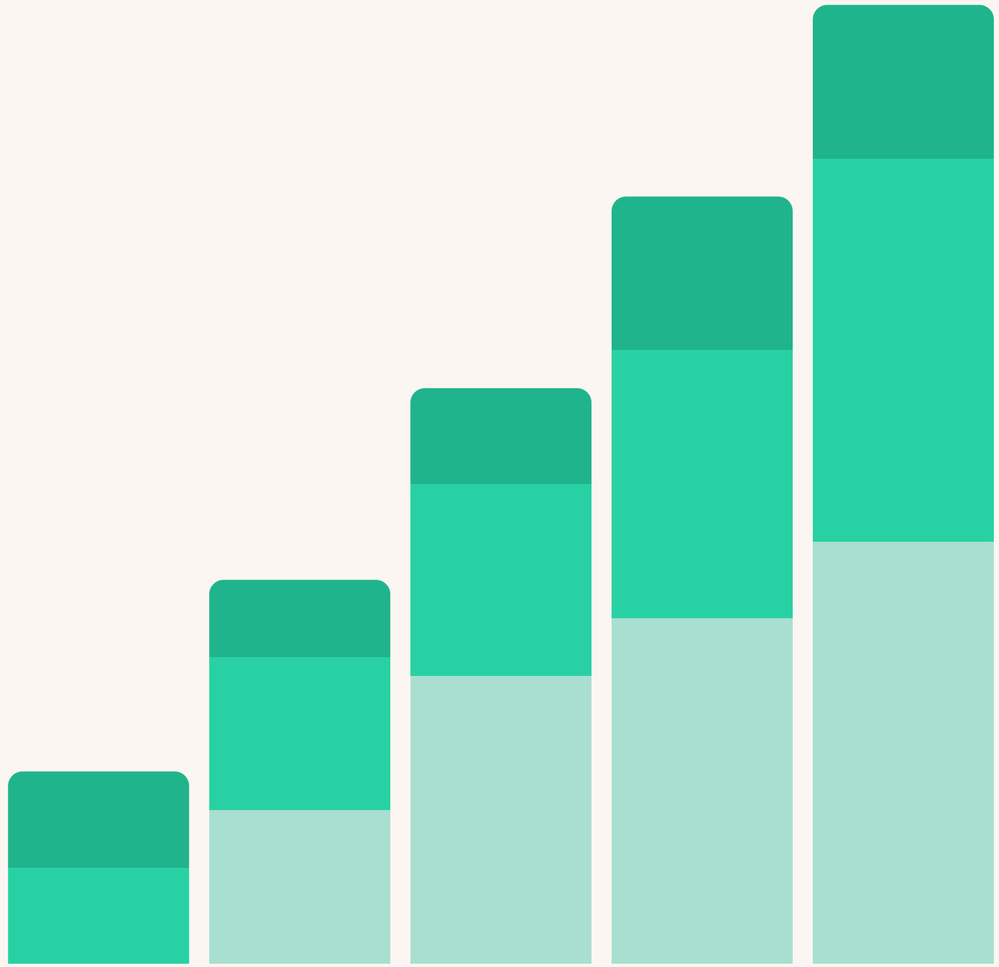
03.

Encoding Categorical Variables

- Machine learning models can't process text directly
- One-Hot Encoding (for categories with few unique values)
- Label Encoding (if the order matters)

Data Exploration

- Descriptive Statistics:
  - Mean, median, mode of key attributes
- Advanced Analysis:
  - Correlation Analysis
  - Time Series Analysis
  - Audience Segmentation
- Visualizations:
  - Bar Chart: Top genres watched
  - Line Graph: Viewership trends over time
  - Pie Chart: Distribution of content by OTT platform
  - Use Matplotlib/Seaborn visualizations



Data Modeling

Modeling helps in deriving meaningful insights and making data-driven decisions, such as segmenting users, predicting trends.

- Clustering: K-Means for User Segmentation
- Regression: Predicting Watch Time

Once models are built, we need to evaluate them to determine their effectiveness.

Model	Metric	Purpose
K-Means Clustering	Inertia (WCSS)	Cluster compactness
Linear Regression	RMSE	Prediction error measurement
Recommendation System	Precision & Recall	Relevance of recommendations

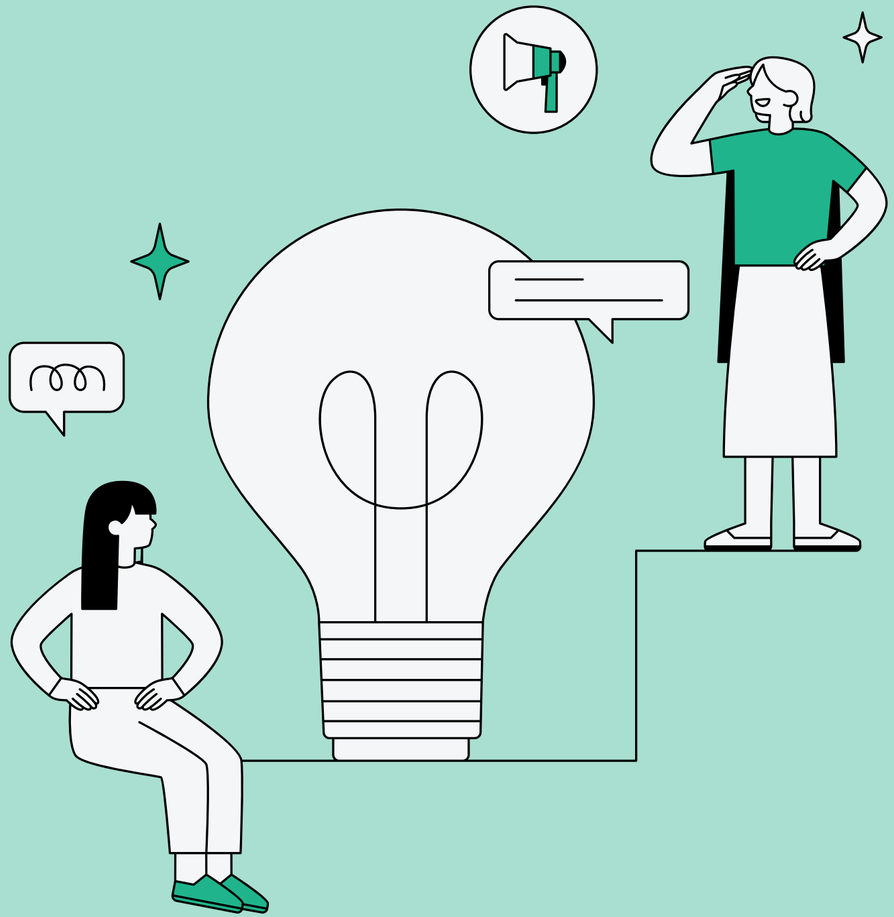
Insights & Conclusion

01. Certain genres (e.g., thrillers, dramas) might have higher engagement than others. Platforms can focus on acquiring or producing more content in high-engagement genres.

02. Personalized recommendations may lead to higher watch time and retention. Enhancing AI-driven recommendations could improve user experience and reduce churn.

03. Certain genres or languages may be more popular in specific regions. OTT platforms can localize content, offer better dubbing, or create region-specific recommendations.

04. Viewership trends may show peak hours in the evening or during weekends. Scheduling new releases and ad placements can be optimized for maximum engagement.



Future Scope

Expanding the Dataset

- Incorporating data from additional platforms like Disney+, HBO Max, or regional OTT services.
- Also including other kinds of content like sports, news, music, etc.

Incorporating Real-Time Data

- Monitor social media trends (Twitter, Reddit) to track audience sentiment and hype around upcoming shows.
- Dynamic dashboards that update live insights on what's trending.

Thank you very much!