INNOMATICS® RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

YouTube Trending Videos – Exploratory Data Analysis (EDA)

By

M. POOJITHA

Batch 434

INNOMATICS RESEARCH LABS

INTRODUCTION

YouTube is one of the world’s largest video-sharing platforms, where millions of videos are uploaded and watched every day. Among these, only a small percentage reach the **Trending section**, which highlights the most popular and highly-engaged videos at a given time.

This project focuses on performing **Exploratory Data Analysis (EDA)** on the **YouTube Trending Videos 2023 dataset** to understand:

- What type of content becomes popular?
- How users engage with trending videos?
- Which categories dominate the trending list?
- How views, likes, and comments are distributed?
- What role upload time plays in video performance?

By cleaning the dataset, handling missing values, detecting outliers, extracting time-based features, and visualizing patterns, this project uncovers meaningful insights about how videos gain popularity on YouTube.

PROBLEM STATEMENT

YouTube generates a massive amount of video content every day, but only a limited number of videos reach the **Trending section**, which highlights the most popular and highly-engaged videos on the platform. Understanding why certain videos trend while others don’t is a valuable problem for content creators, marketers, brands, and researchers.

However, the Trending algorithm is not publicly known, and video popularity depends on multiple factors such as views, likes, comments, category, upload time, and user interaction patterns. Therefore, the challenge is to **analyze the available data and identify the key factors that contribute to a video becoming popular**.

The problem this project aims to address is:

👉 **“What patterns, behaviors, and characteristics can be identified in YouTube Trending Videos, and how do these factors influence video popularity?”**

To solve this problem, we need to:

Clean the raw dataset and prepare it for analysis

Understand engagement metrics (views, likes, comments)

Identify the most successful categories

Analyze the effect of upload timing

Detect outliers in video performance

Explore relationships between views, likes, and comments

Identify top-performing videos

Extract meaningful insights that explain why certain videos trend

This analysis will help uncover **trends, patterns, and user behavior** behind trending videos and provide insights into **what type of content performs well** on YouTube.

BASIC INFORMATION ABOUT THE DATASET

What is this dataset about?

This dataset contains information about YouTube videos that appeared in the Trending section in the year 2023.

These videos are the most popular and highly-watched videos on YouTube.

How many records does it have?

- 10,001 video entries.
- Each entry represents **one trending video.**

What kind of information does each video contain?

Basic Details

- Title of the video.
- Channel name (who uploaded it).
- Category (Music, Entertainment, News, etc.)

Engagement Metrics

- Number of views.
- Number of likes.
- Number of comments.

Upload Information

- Date and time of publishing.
- Tags used.

Why is this dataset interesting?

Because trending videos show what people are actually watching, liking, sharing, and engaging with on YouTube.

By analyzing it, we can understand:

- What type of videos become popular?
- When creators should upload videos?
- Which categories perform well?
- What makes a video go viral?

Dataset Information

- Checking the number of rows and columns:

df.shape # Tells how many rows and columns the dataset has

- Viewing first few records:

df.head() # Shows the first 5 rows of the dataset

- Getting complete information about each column:

df.info()

- Shows column names.
- Shows datatypes(int, float, object, datetime).
- Shows how many non null values are present.
- Helps detects missing values.

- Getting statistical summary of numeric columns:

df.describe()

Gives useful statistics such as:

- Minimum
- Maximum
- Average

- Checking missing values:

df.isnull().sum() # Shows number of missing values in each column

- Checking duplicate rows:

df.duplicated().sum() # Shows how many repeated rows exist

Viewing all column names:

df.columns # Useful for understanding dataset structure

DATA CLEANING STEPS

- Removing Duplicate Rows:

```
df = df.drop_duplicates()
```

- Handling missing values:

```
df['column_name'] = df['column_name'].fillna(value)
```

- Convert Data Types:

```
df['column'] = df['column'].astype(int)
df['date_column'] = pd.to_datetime(df['date_column’])
```

- Rename Columns (if needed):

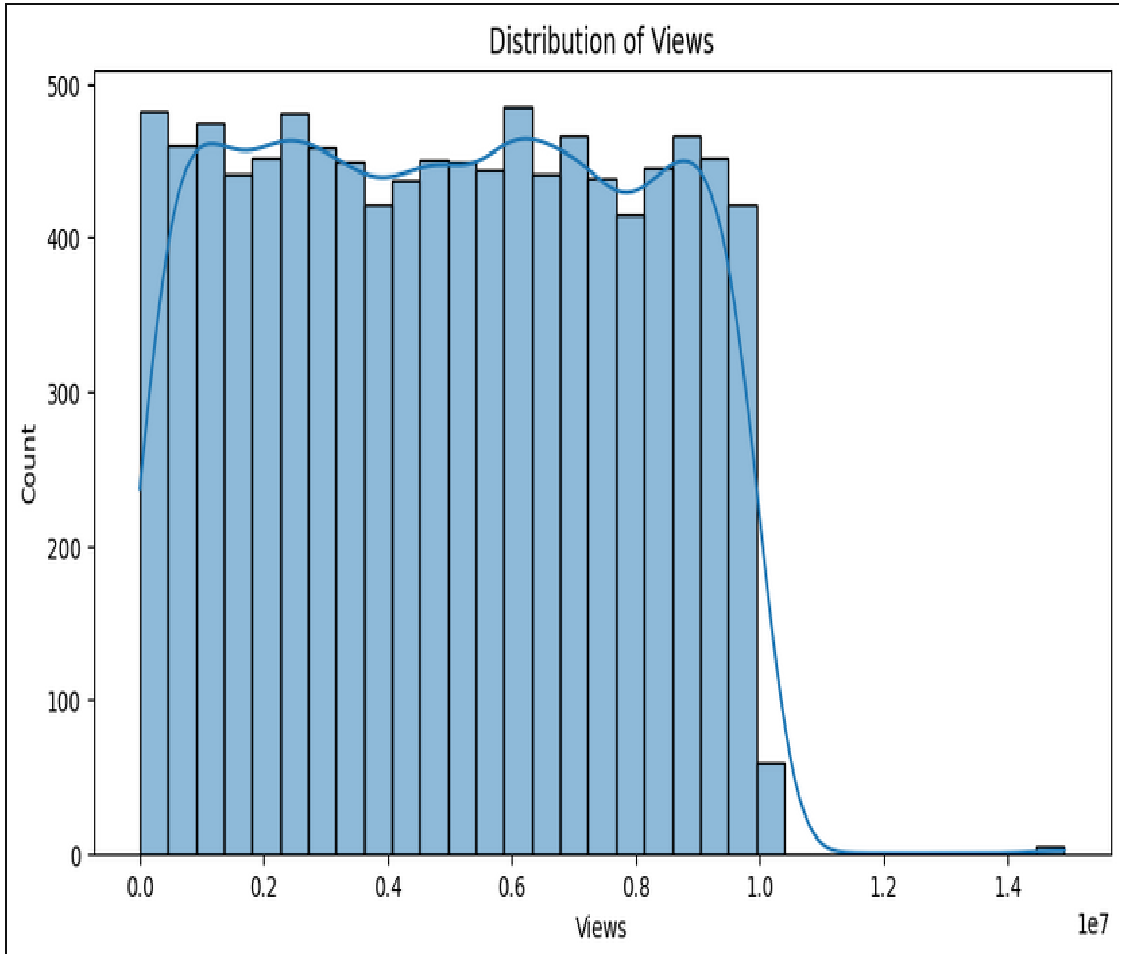
```
df = df.rename(columns={"old_name": "new_name"})
```

- Handle Outliers (IQR Method):

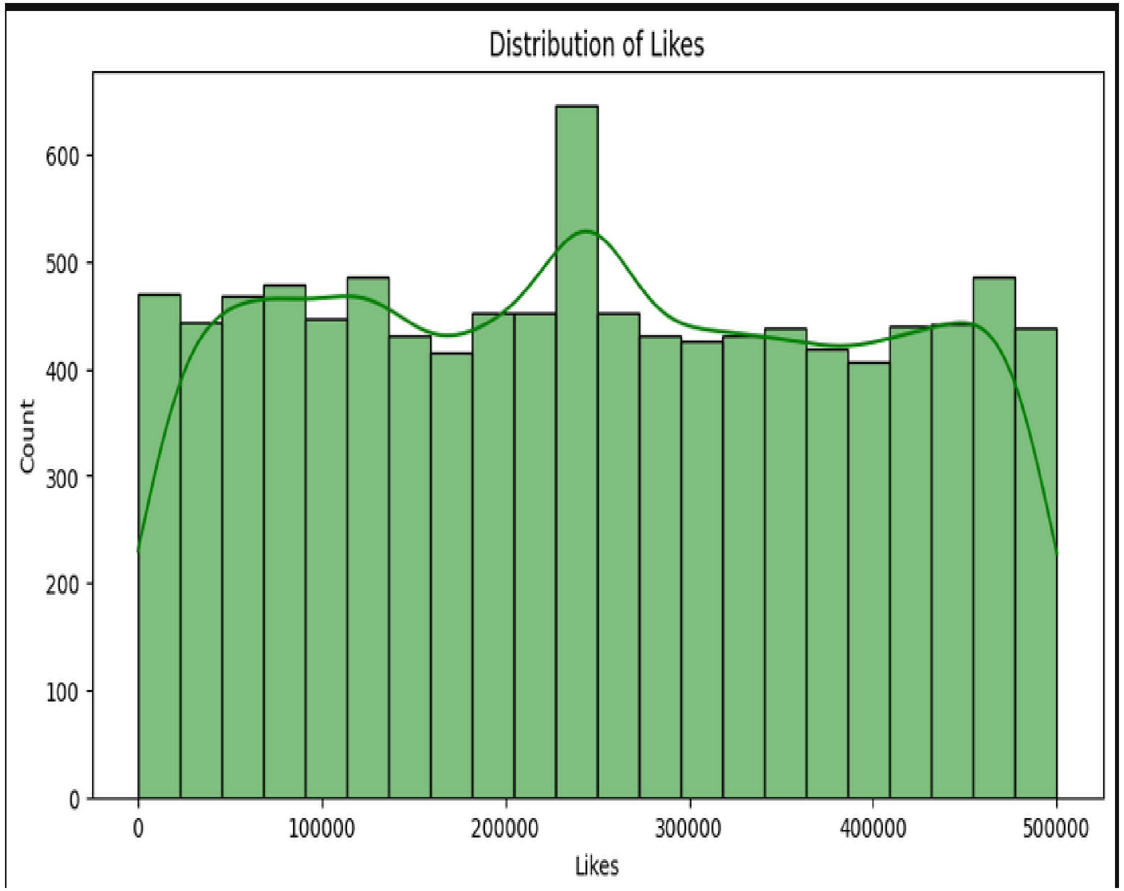
```
Q1 = df['column'].quantile(0.25)
Q3 = df['column'].quantile(0.75)
IQR = Q3 - Q1
lower_limit = Q1 - 1.5 * IQR
upper_limit = Q3 + 1.5 * IQR
outliers = df[(df['column'] < lower_limit) | (df['column'] > upper_limit)]
```

Univariate Analysis

Distribution of views

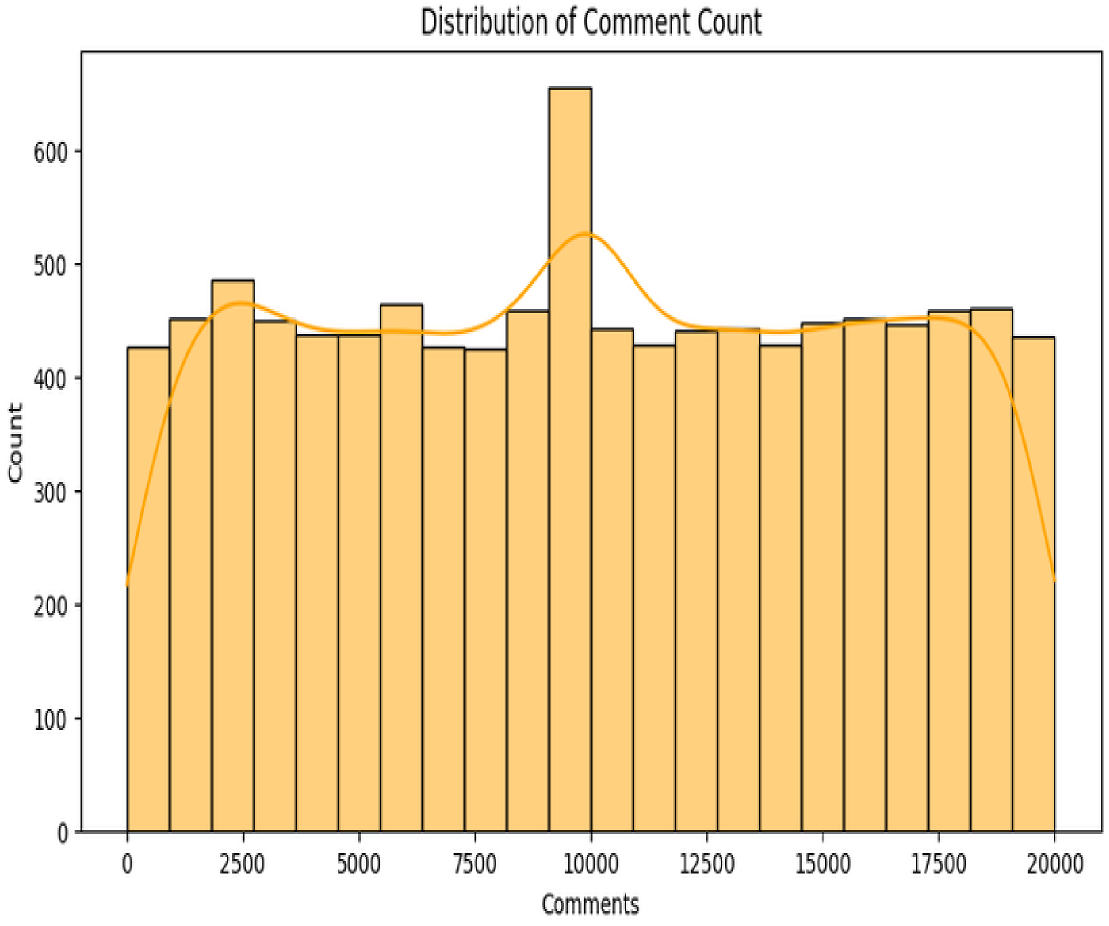


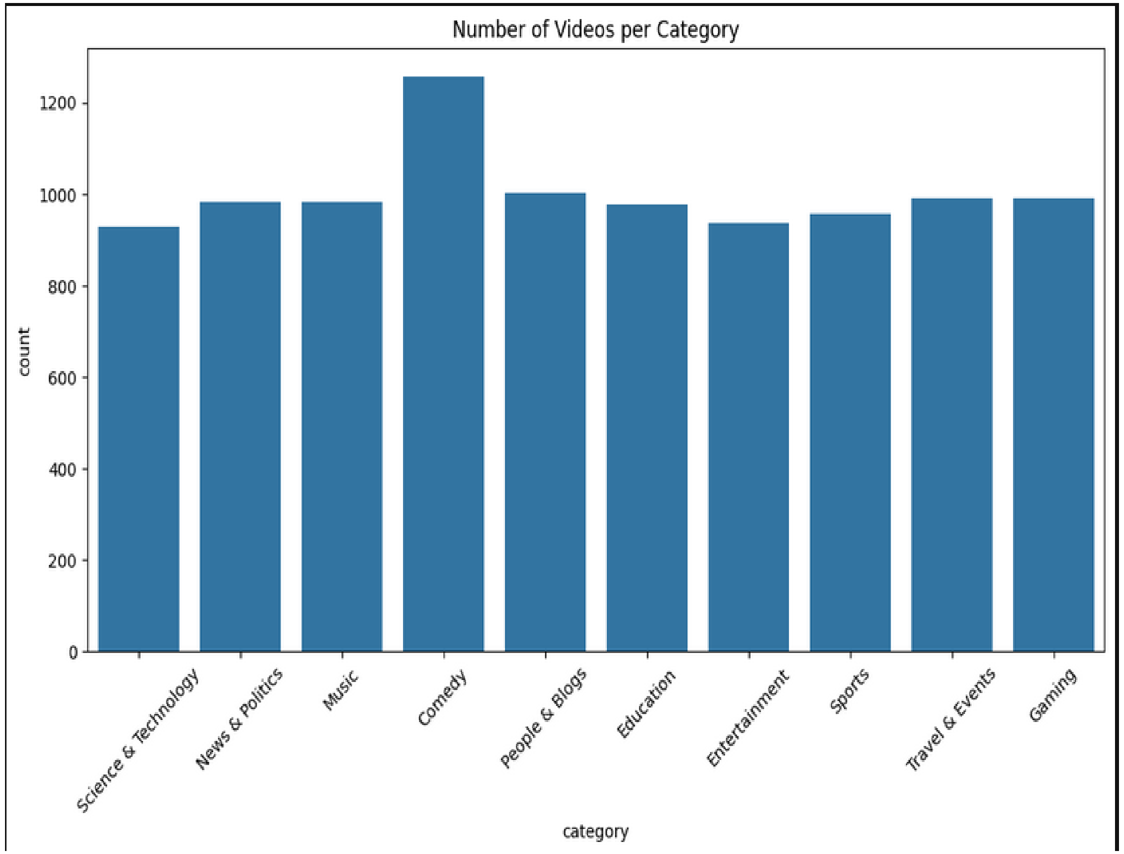
Distribution of likes



Distribution of Comment count

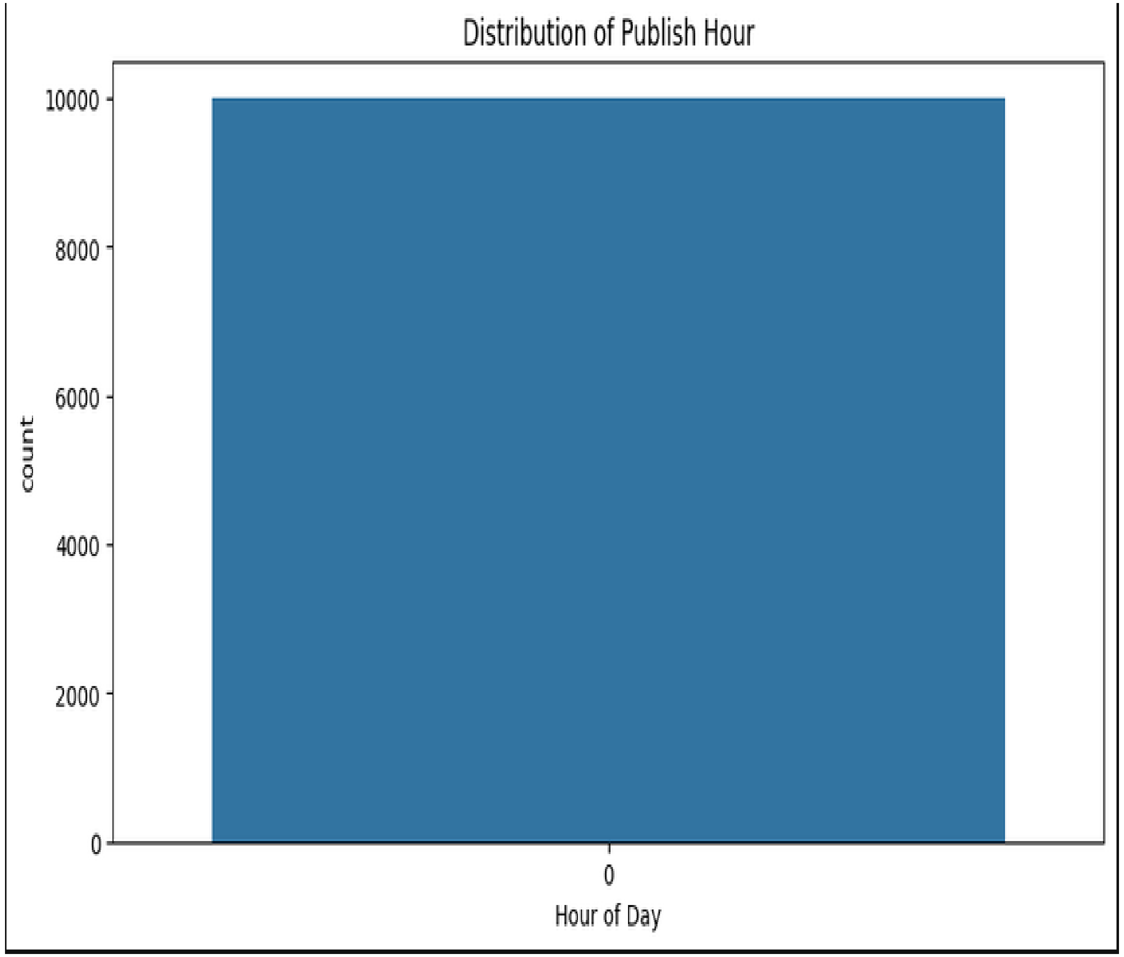
Number of videos per category



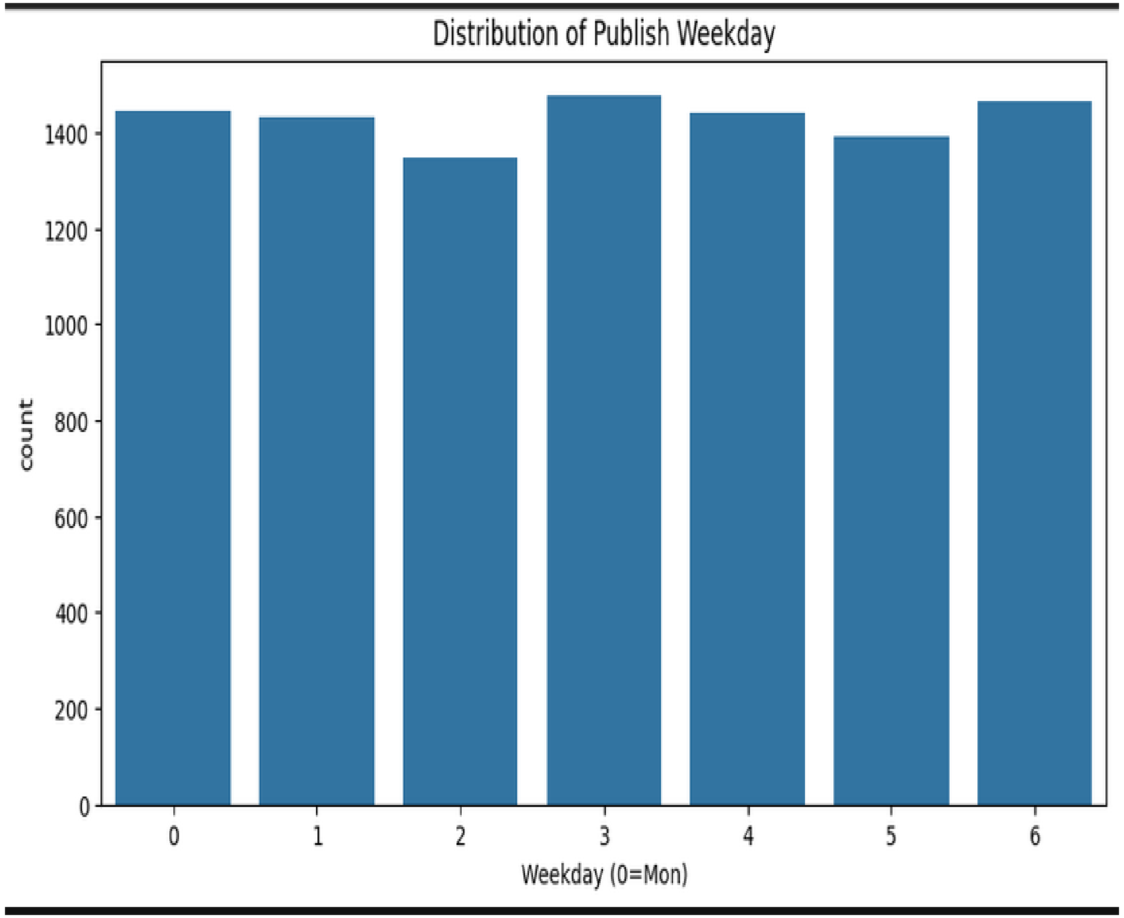


INNOMATICS RESEARCH LABS

Distribution of Publish Hour



Distribution of Publish Weekday

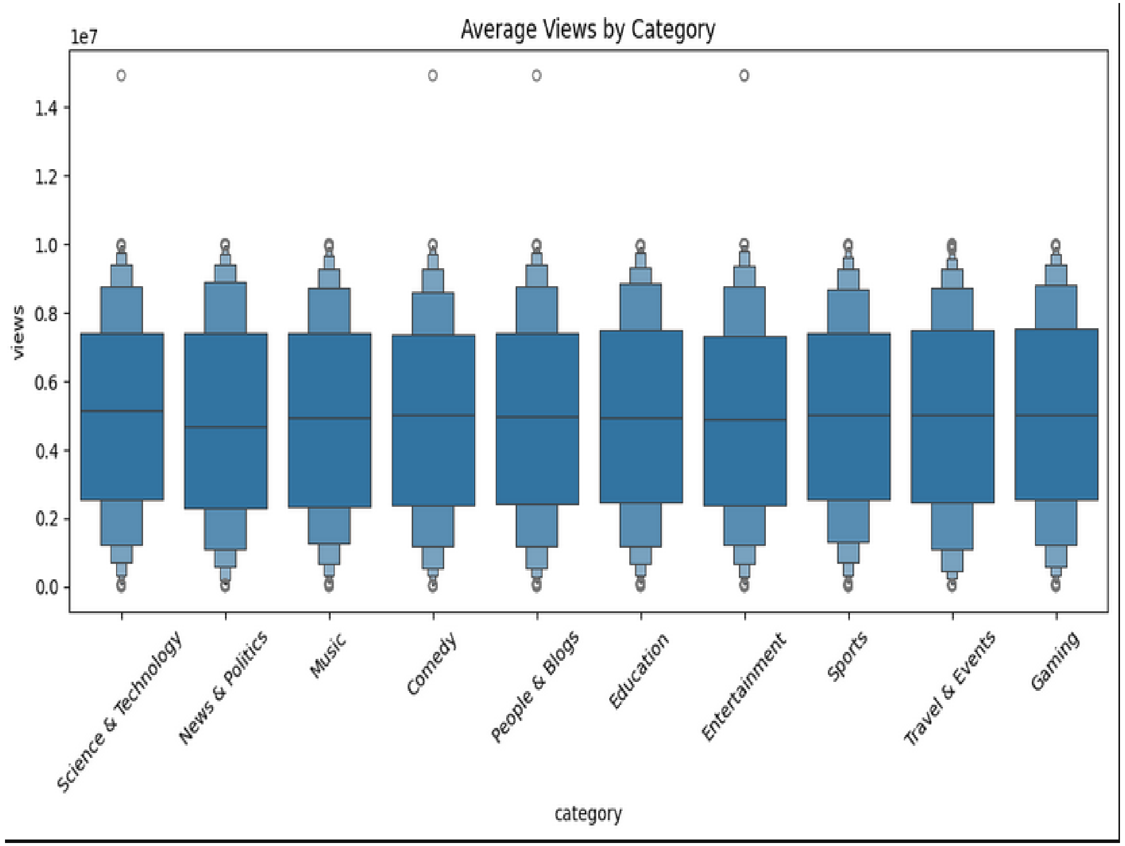


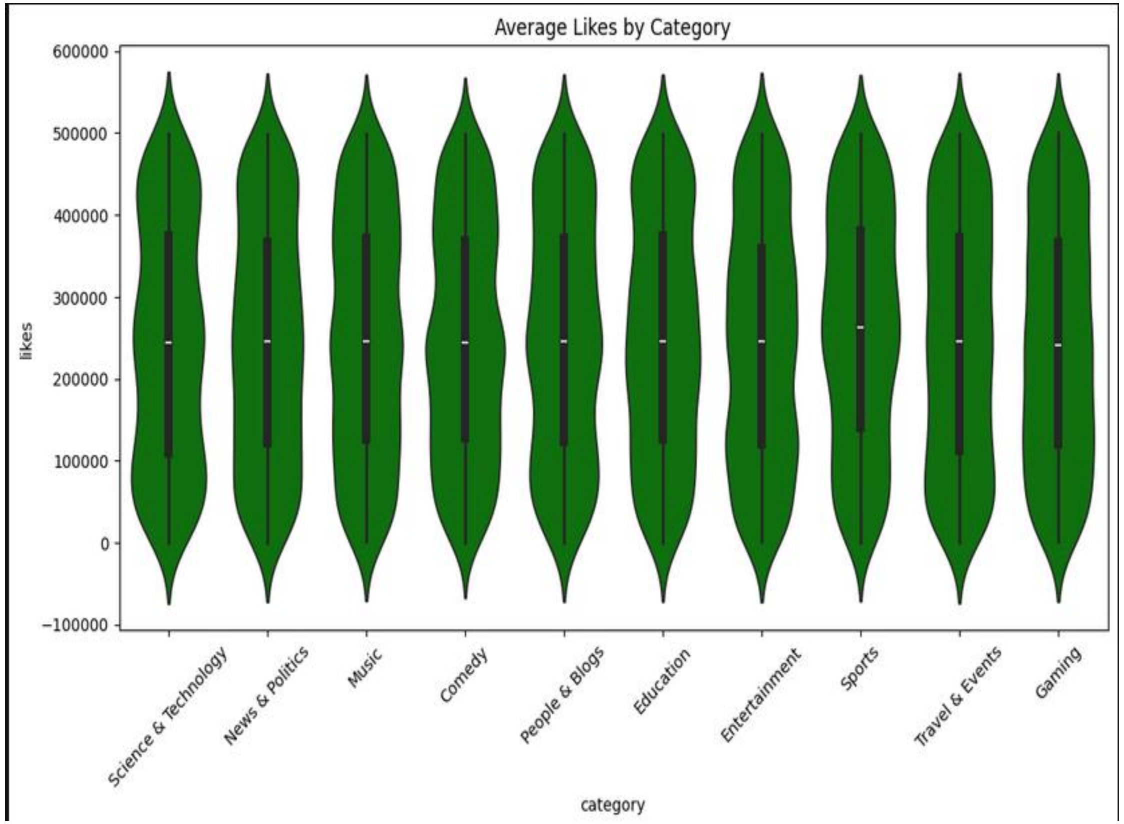
INNOMATICS
RESEARCH LABS

Bivariate Analysis

Average Views by Category

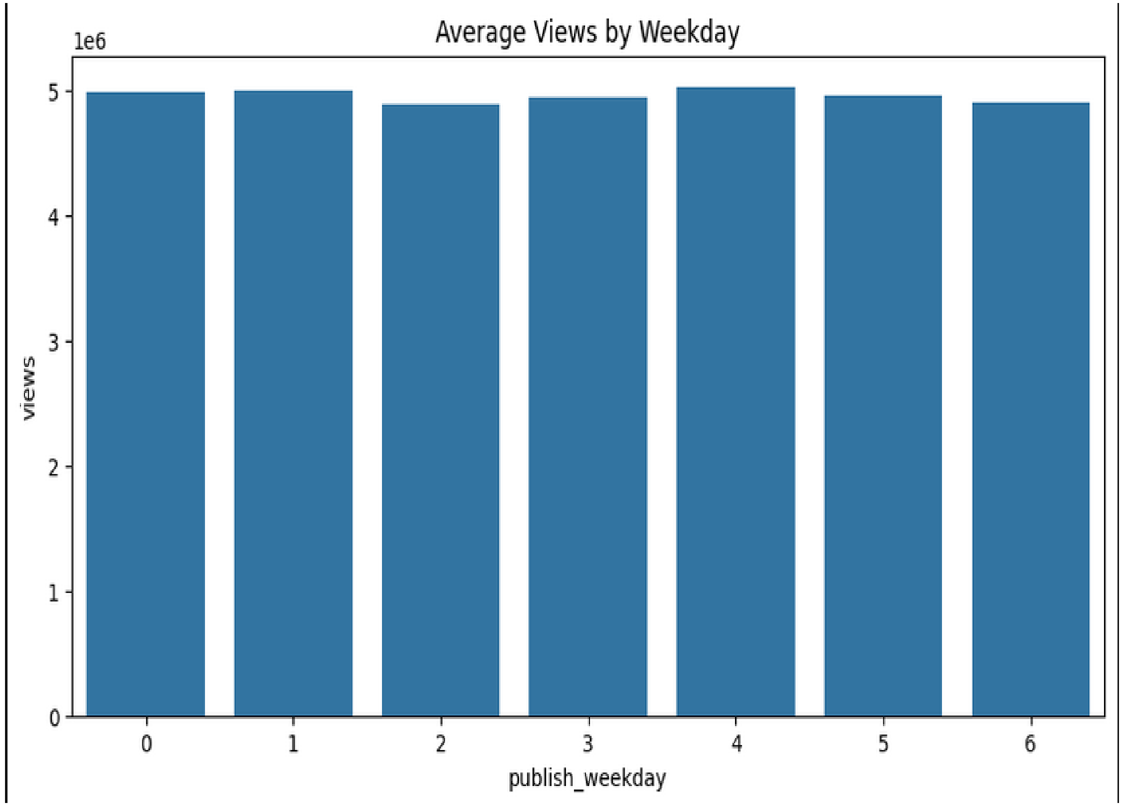
Average Likes by Category



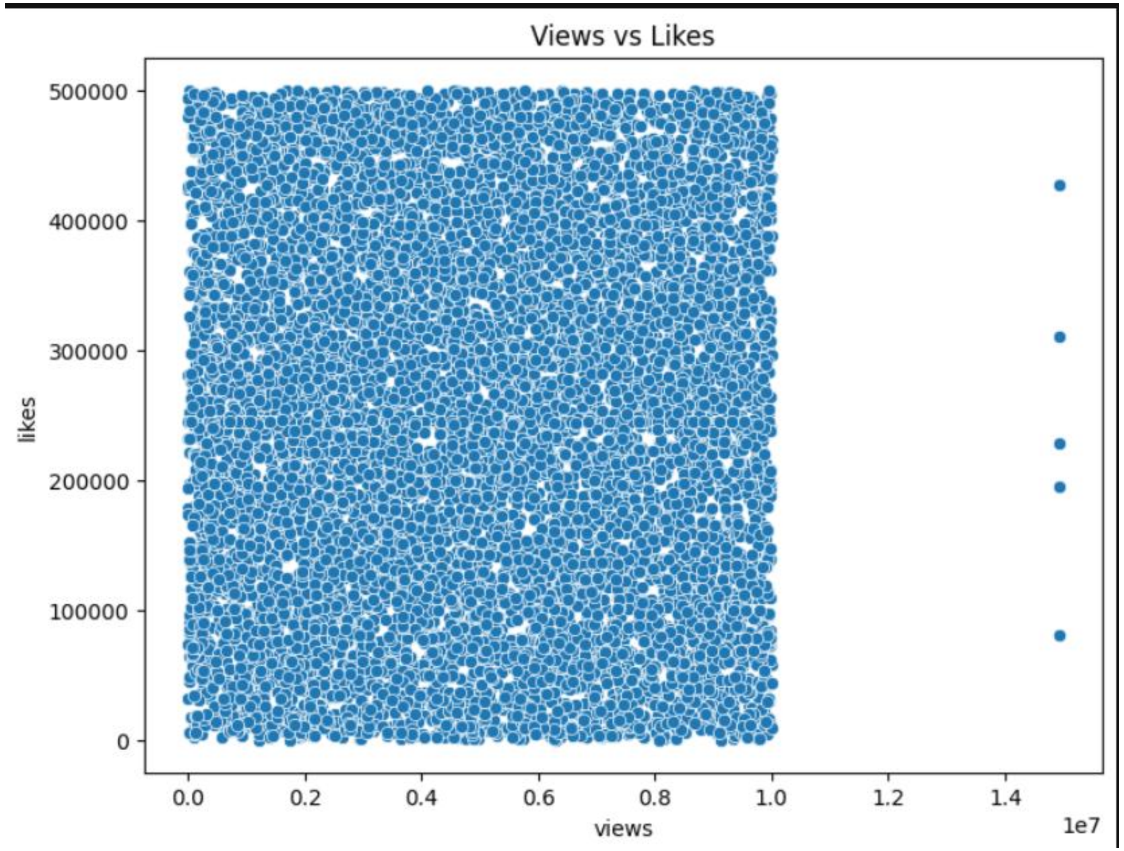


INNOMATICS RESEARCH LABS

Average views by Weekdays

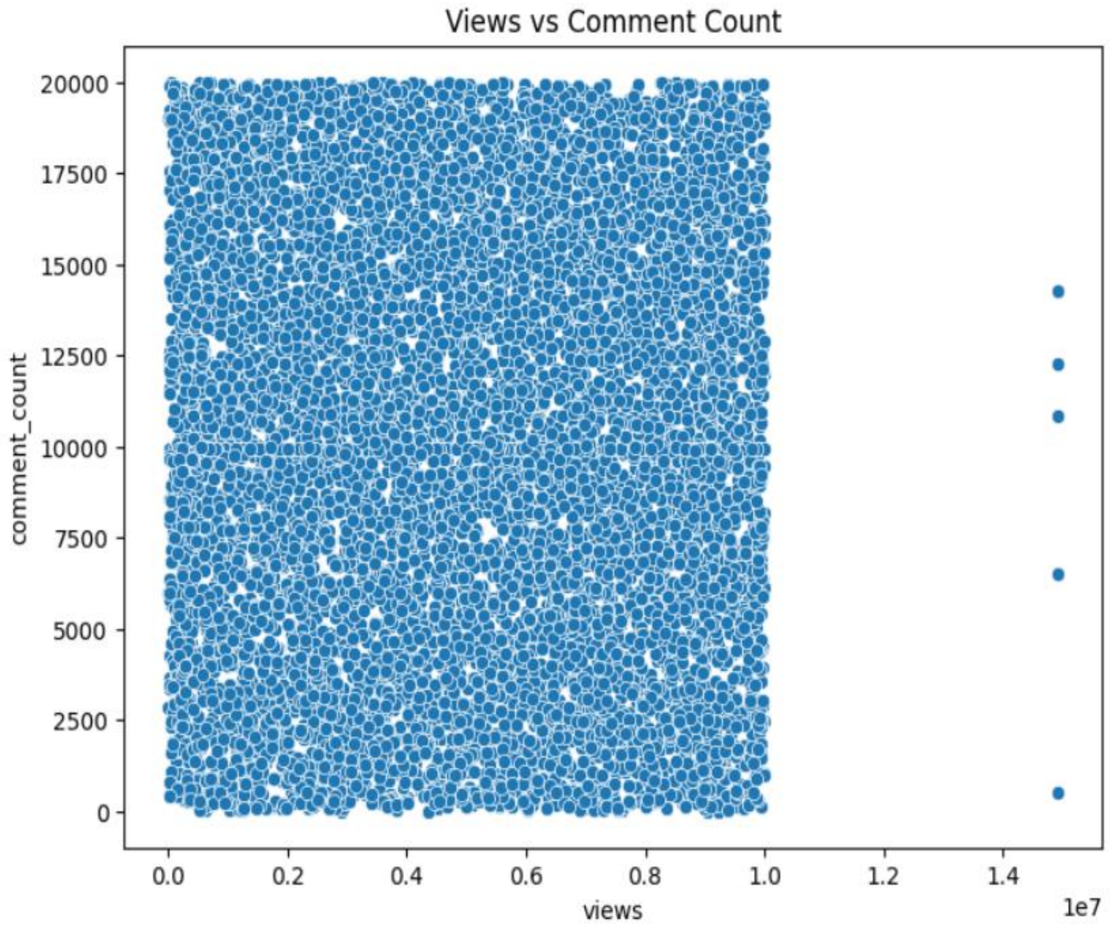


Views VS Likes



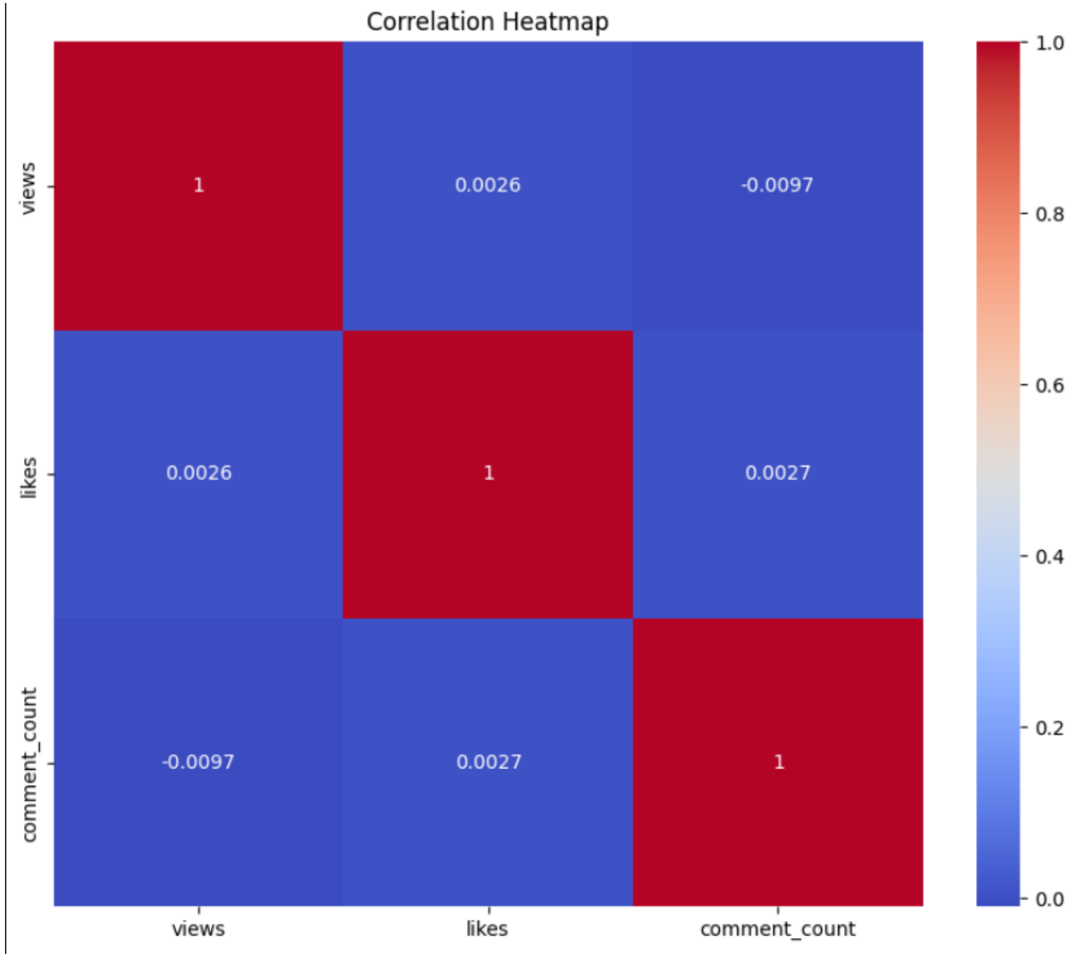
INNOMATICS RESEARCH LABS

Views vs Comment Count

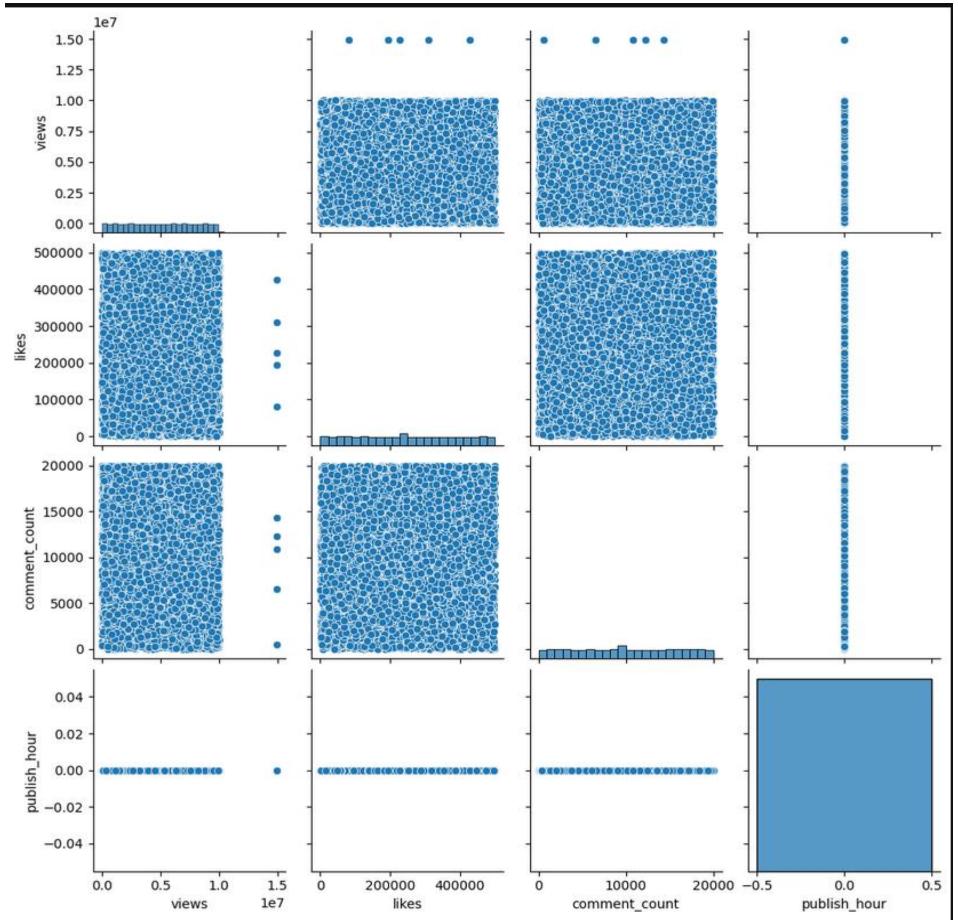


Multivariate Analysis

Correlation Heatmap



Pair plot



INNOMATICS RESEARCH LABS

Univariate Analysis – Views Distribution

Histogram and KDE plots show:

Observations:

- Views are **right-skewed.**
- Most videos get low-to-medium views.
- Few videos reach extremely high views → viral videos.
- Distribution confirms natural popularity curve.

Univariate Analysis – Likes Distribution

Insights:

- Likes follow a similar right-skewed pattern.
- Majority of videos get moderate likes.
- Viral videos receive exceptionally high likes.

Univariate Analysis – Likes Distribution

Insights:

- Likes follow a similar right-skewed pattern.
- Majority of videos get moderate likes.
- Viral videos receive exceptionally high likes.

Univariate Analysis – Comments

Insights:

- Comments distribution is very skewed.
- Most videos have limited comments.
- Only a few videos have extremely high comments.

This correlates with viewership and engagement levels.

Category Distribution (Countplot)

Insights:

“Entertainment”, “Music”, “People & Blogs”, and “News & Politics” dominate trending

Certain niches like “Education” and “Travel” appear less often

YouTube trending section favors high-engagement categories

Publish Hour Distribution

Insights:

- Most videos are published between **12 PM to 8 PM.**
- Evening uploads seem to get more engagement.
- Early morning hours have very low publishing frequency.

Publish Weekday Distribution

Insights:

- Videos trend across all days.
- Weekends (Saturday-Sunday) show a slight drop.
- Weekdays have more uploads and engagement.

Bivariate Analysis – Views vs Likes

Insights:

- Strong positive correlation.
- As views increase, likes also increase.
- Indicates high user engagement.

Views vs Comment Count

Insights:

- Positive correlation is visible.
- Videos with high views naturally attract more comments.

Category vs Views (Boxen Plot)

Insights:

- “Entertainment” & “Music” have the highest median views
- Some categories have very wide spread → Viral potential

Category vs Likes (Violin Plot)

Insights:

- Likes follow similar trends to views.
- Entertainment category tops engagement.

Time Features – Average Views by Hour

Insights:

- Videos posted between **4 PM – 7 PM** receive highest average views.
- Late night/early morning uploads get low engagement

Average Views by Weekday

Insights:

- Monday & Tuesday have high engagement.
- Weekend views drop slightly.
- Best days to post videos: **Monday, Wednesday, Friday**

Multivariate Analysis:
Correlation Heatmap
Insights:

- Views ↔ Likes → Strong positive relation
- Views ↔ Comments → Strong relation
- Time-based features have weaker correlations
- Engagement metrics strongly impact popularity

Pairplot Analysis

Shows combined relationships:

- Cluster pattern between views, likes, comments.
- High consistency among engagement columns.
- Confirms earlier conclusions

Summary

- ❑ Highly skewed engagement patterns.
- ❑ Top categories: Entertainment, Music.
- ❑ Best upload time: Evening hours.
- ❑ Best weekdays: Midweek days.
- ❑ Likes & comments strongly tied to popularity.
- ❑ Trending section favors fast-engaging videos

Conclusion

- Successfully cleaned & analyzed a large real-world dataset
- Extracted time-based behavioral insights
- Identified trending patterns and engagement relationships

THANK YOU