CS210 Term Project

Youtube Analysis – Bora Tosun 29014

Overview

In this project, you delve into my YouTube viewing history to uncover insights about my digital entertainment preferences. The aim is to analyze my viewing habits, including frequently watched video categories, favorite channels, preferred channels, viewing times, and any emerging patterns in my content consumption.

Data Collection

I first downloaded my YouTube Watching History from Google's website takeout.google.com, then I processed 'izleme geçmişi.html' with beautiful soup as it is in pure html format.

Data Cleaning and Preprocessing

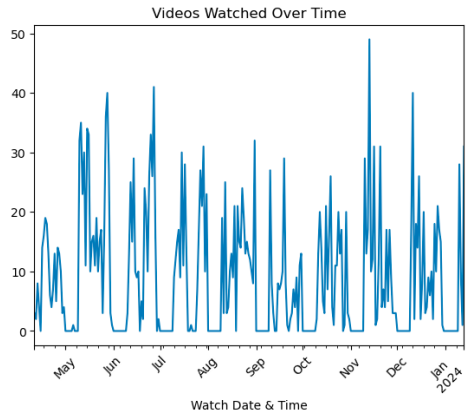
- `pandas` for data manipulation.

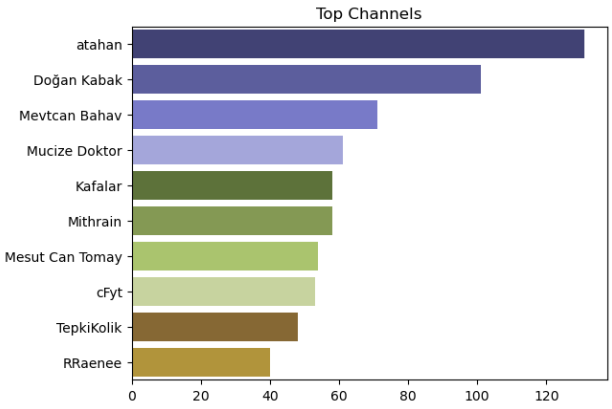
Converted date to datetime style, handled missing values and duplicates, created key columns.

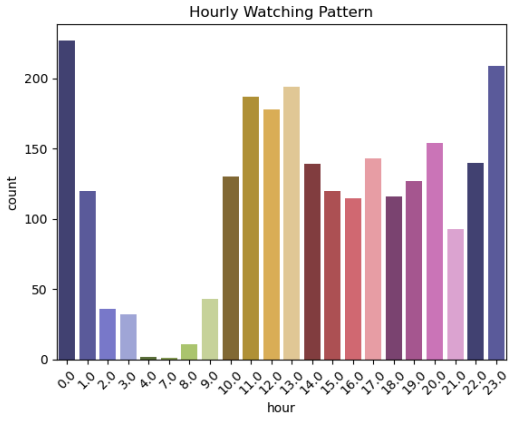
Exploratory Data Analysis (EDA)

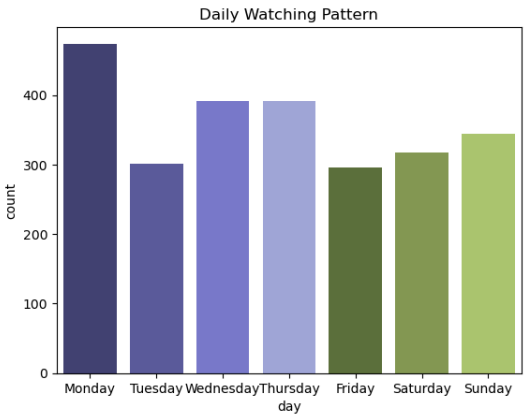
- `matplotlib` and `seaborn` for visualization.
  - **Category/Genre Analysis**: Identified the most frequently watched video categories or genres.
  - **Channel Analysis**: Analyzed which YouTube channels you watch the most.
  - **Viewing Time Analysis**: Explored patterns in the times of day you typically watch videos.
  - **Content Preference Trends**: Investigated how my content preferences may have changed over time.

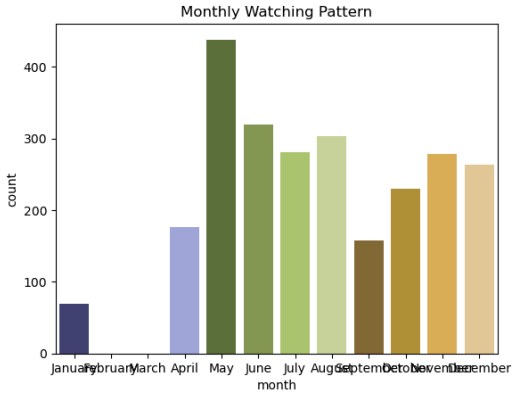
  All related visuals and more are attached below:

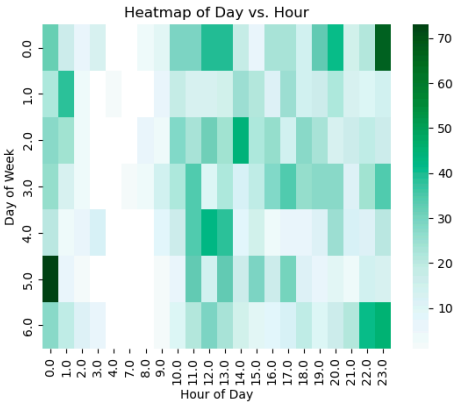


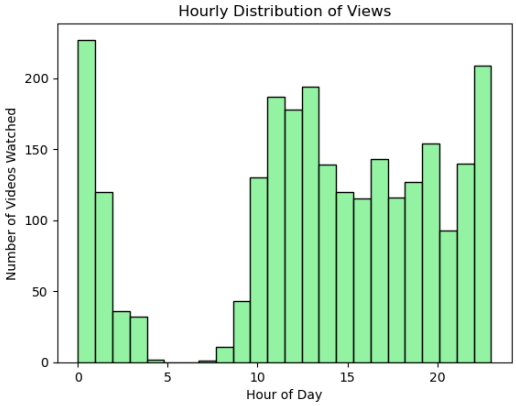


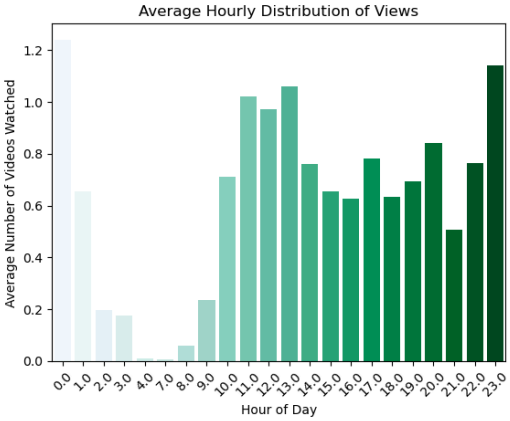

Hourly Watching Pattern
count
hour
0.0
1.0
2.0
3.0
4.0
7.0
8.0
9.0
10.0
11.0
12.0
13.0
14.0
15.0
16.0
17.0
18.0
19.0
20.0
21.0
22.0
23.0

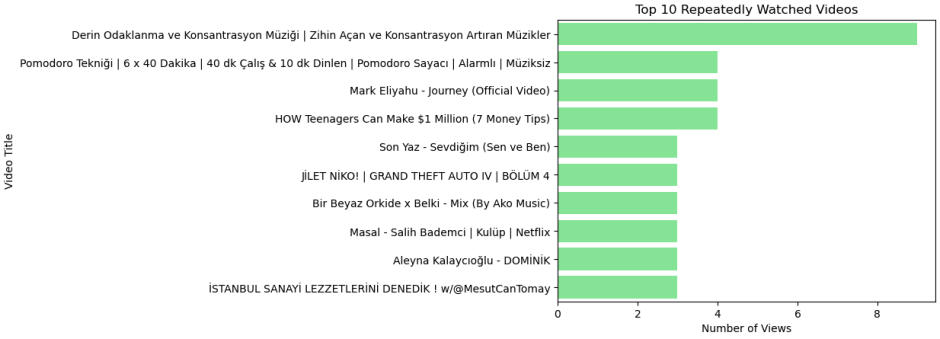

Daily Watching Pattern
count
day
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday


Monthly Watching Pattern
count
month
January
February
March
April
May
June
July
August
September
October
November
December


Heatmap of Day vs. Hour
Day of Week
Hour of Day


Hourly Distribution of Views
Number of Videos Watched
Hour of Day


Average Hourly Distribution of Views
Average Number of Videos Watched
Hour of Day


Top 10 Repeatedly Watched Videos
Video Title
Derin Odaklanma ve Konsantrasyon Müziği | Zihin Açan ve Konsantrasyon Artıran Müzikler
Pomodoro Tekniği | 6 x 40 Dakika | 40 dk Çalış & 10 dk Dinlen | Pomodoro Sayacı | Alarmlı | Müziksiz
Mark Eliyahu - Journey (Official Video)
HOW Teenagers Can Make $1 Million (7 Money Tips)
Son Yaz - Sevdiğim (Sen ve Ben)
JİLET NİKO! | GRAND THEFT AUTO IV | BÖLÜM 4
Bir Beyaz Orkide x Belki - Mix (By Ako Music)
Masal - Salih Bademci | Kulüp | Netflix
Aleyna Kalaycıoğlu - DOMİNİK
İSTANBUL SANAYİ LEZZETLERİNİ DENEDİK ! w/@MesutCanTomay
Number of Views

Machine Learning

- For implementing Machine Learning properly I created a wordcloud in order to find out the categories I watch most.

- Then I labelled my data with some of meaningful words:

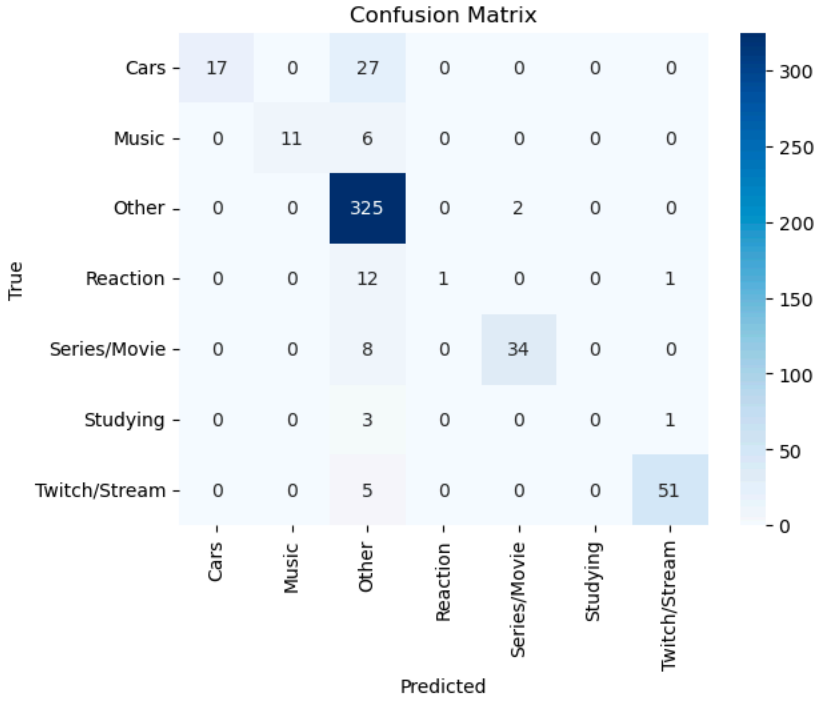
```
def categorize_title(title):
    title = title.lower()
    if any(keyword in title for keyword in ["music", "müzik", "müziği", "şarkısı", "şarkı"]):
        return "Music"
    if any(keyword in title for keyword in ["bölüm", "episode", "fragman", "trailer"]):
        return "Series/Movie"
    if any(keyword in title for keyword in ["drag", "bmw", "car", "araba", "otopark"]):
        return "Cars"
    if any(keyword in title for keyword in ["pomodoro", "çalış", "çalışmak", "final"]):
        return "Studying"
    if any(keyword in title for keyword in ["rraenee", "elraen"]):
        return "Twitch/Stream"
    if any(keyword in title for keyword in ["tepki", "vs"]):
        return "Reaction"
    return "Other"

df['Category'] = df['Video Title'].apply(categorize_title)
```

- After training the dataset with sklearn, I got these results:

	precision	recall	f1-score	support
Cars	1.00	0.39	0.56	44
Music	1.00	0.65	0.79	17
Other	0.84	0.99	0.91	327
Reaction	1.00	0.07	0.13	14
Series/Movie	0.94	0.81	0.87	42
Studying	0.00	0.00	0.00	4
Twitch/Stream	0.96	0.91	0.94	56
accuracy			0.87	504
macro avg	0.82	0.55	0.60	504
weighted avg	0.88	0.87	0.85	504

- Then to finish implementation I created a confusion matrix to check results:



Conclusion

This project is a comprehensive exploration of my YouTube viewing history. It utilizes data science methods to reveal patterns and preferences in my digital entertainment choices, providing insights into the types of content that captivate and engage me on YouTube.