DSA210
Term Project

Analysis of Youtube Data

Alp Çetintaş - 32550

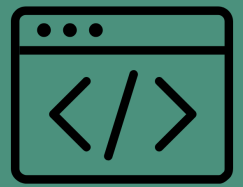
YouTube

Index & Content

Motivation - Why?

- Most used social media platform
- Reflects my interests w/ videos
- Uncover my habits

- What type of content do I watch
- Percentage of likes over videos viewed
- Is there any preferred/favored content creators
- Daily/Weekly watching habits
- My subscriptions
- Search history

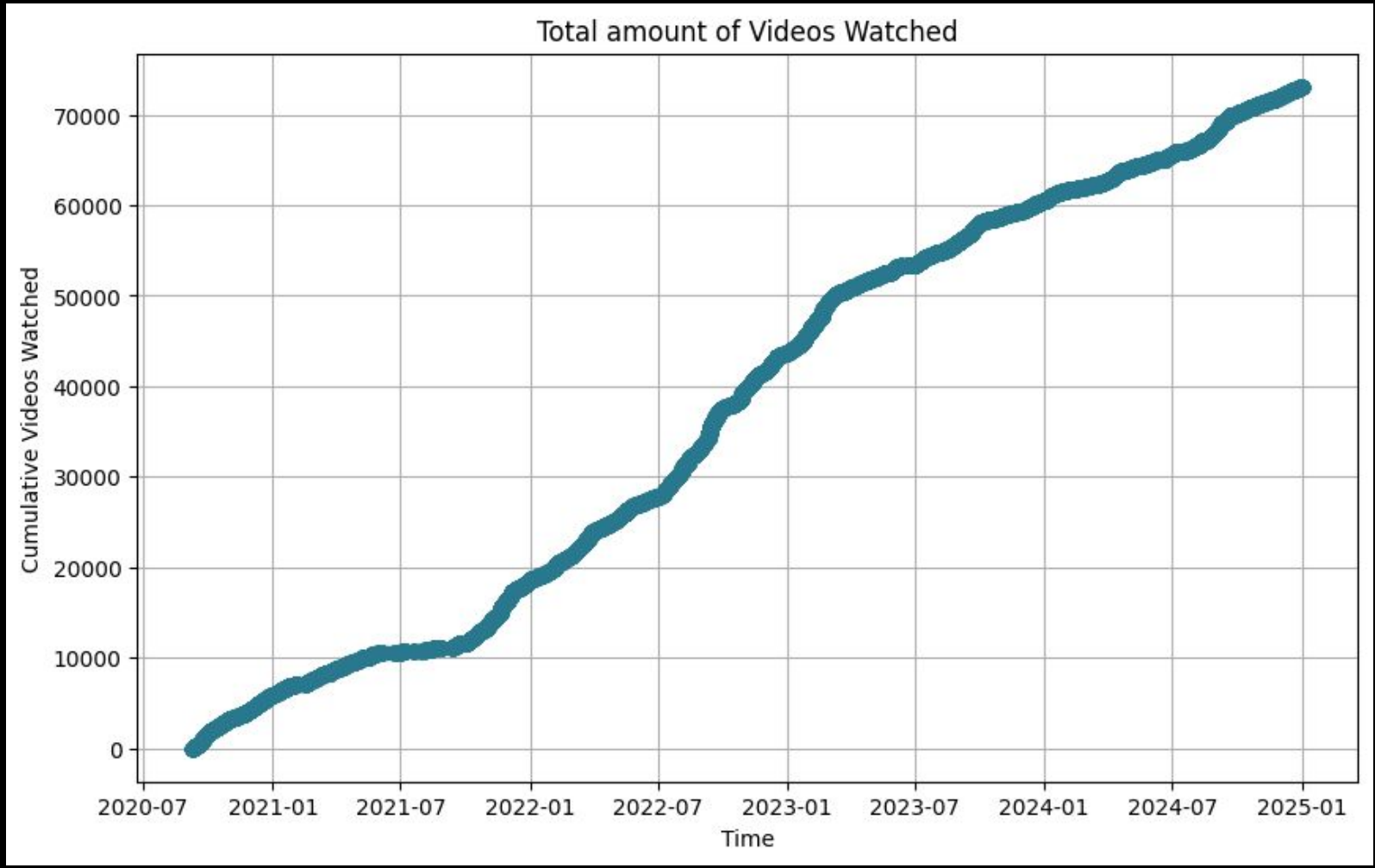


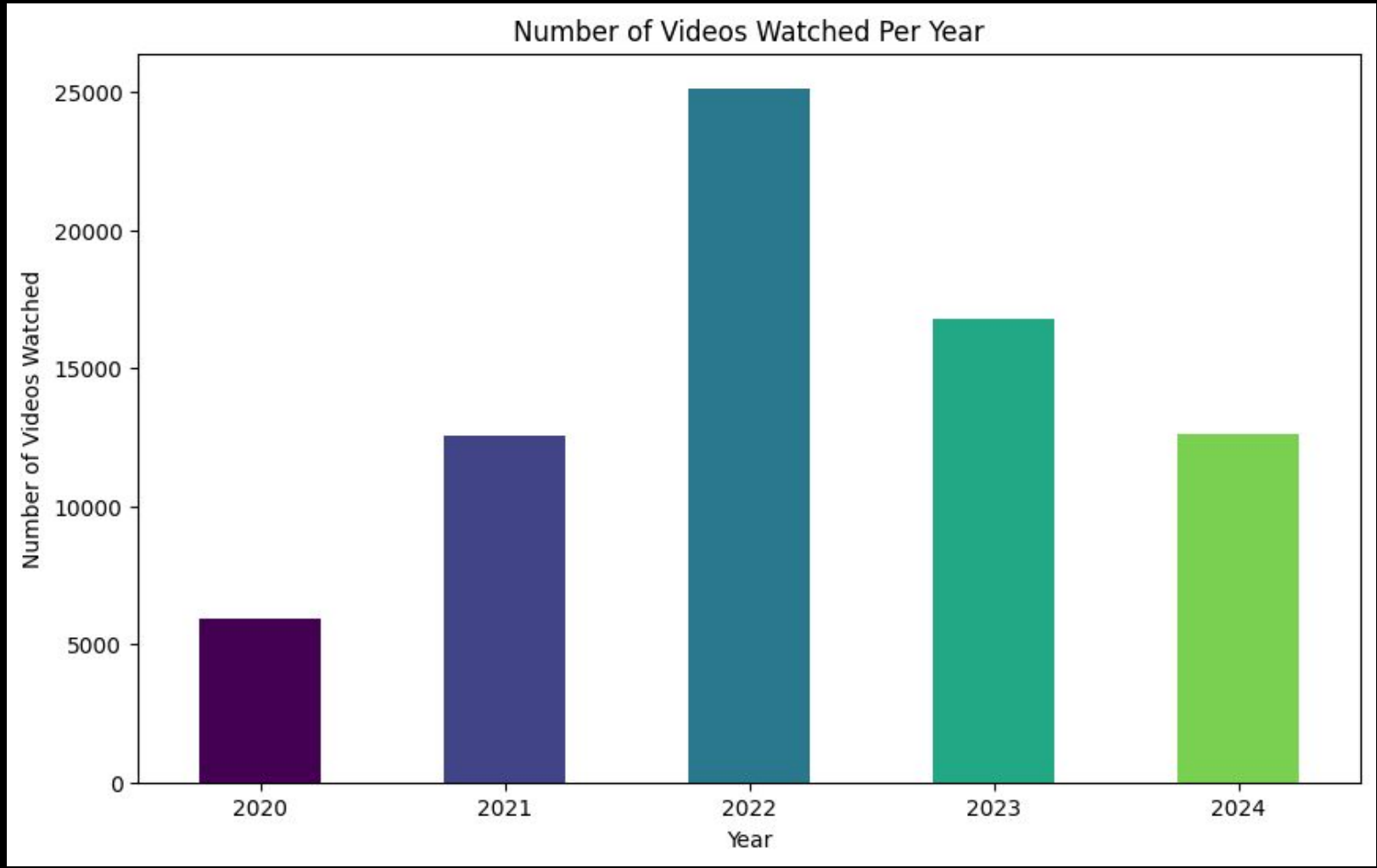
Data Source - Preprocessing

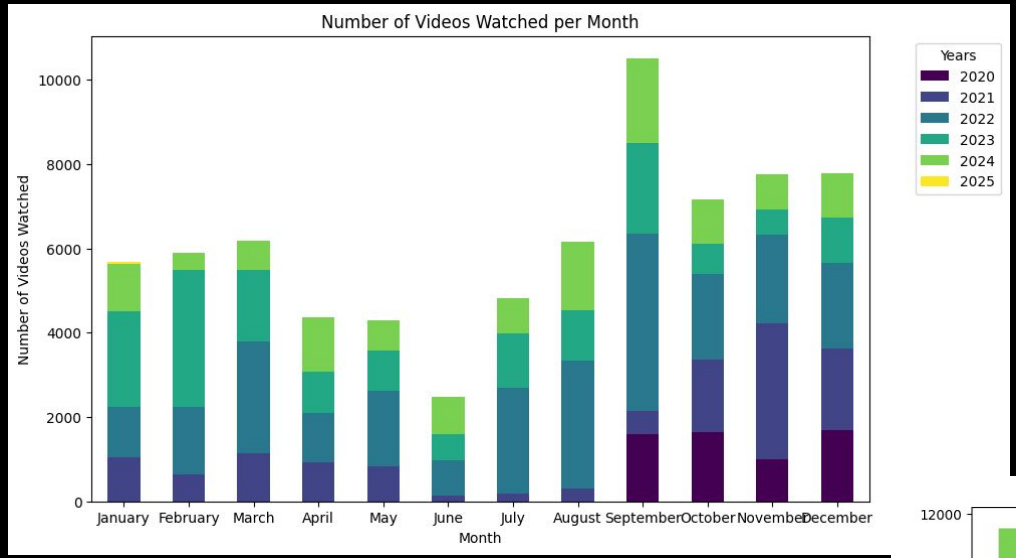
- Google Takeout
  - Watch History
  - Search History
  - Subscriptions
  - Comments
  - Playlists

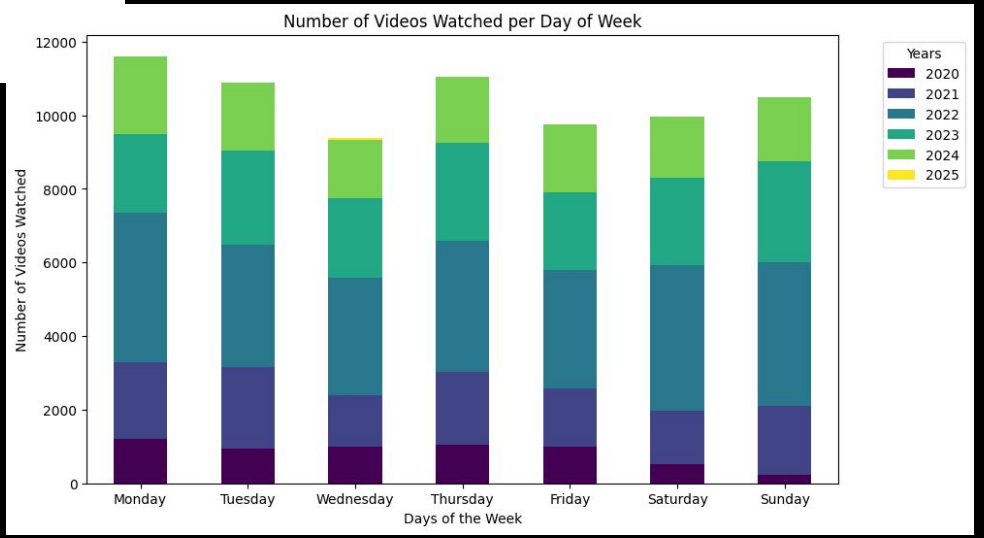
- Preprocessing
  - Merging two accounts
  - Unavailable Title Names
  - Converting time column - date_time
  - Sorting in time order

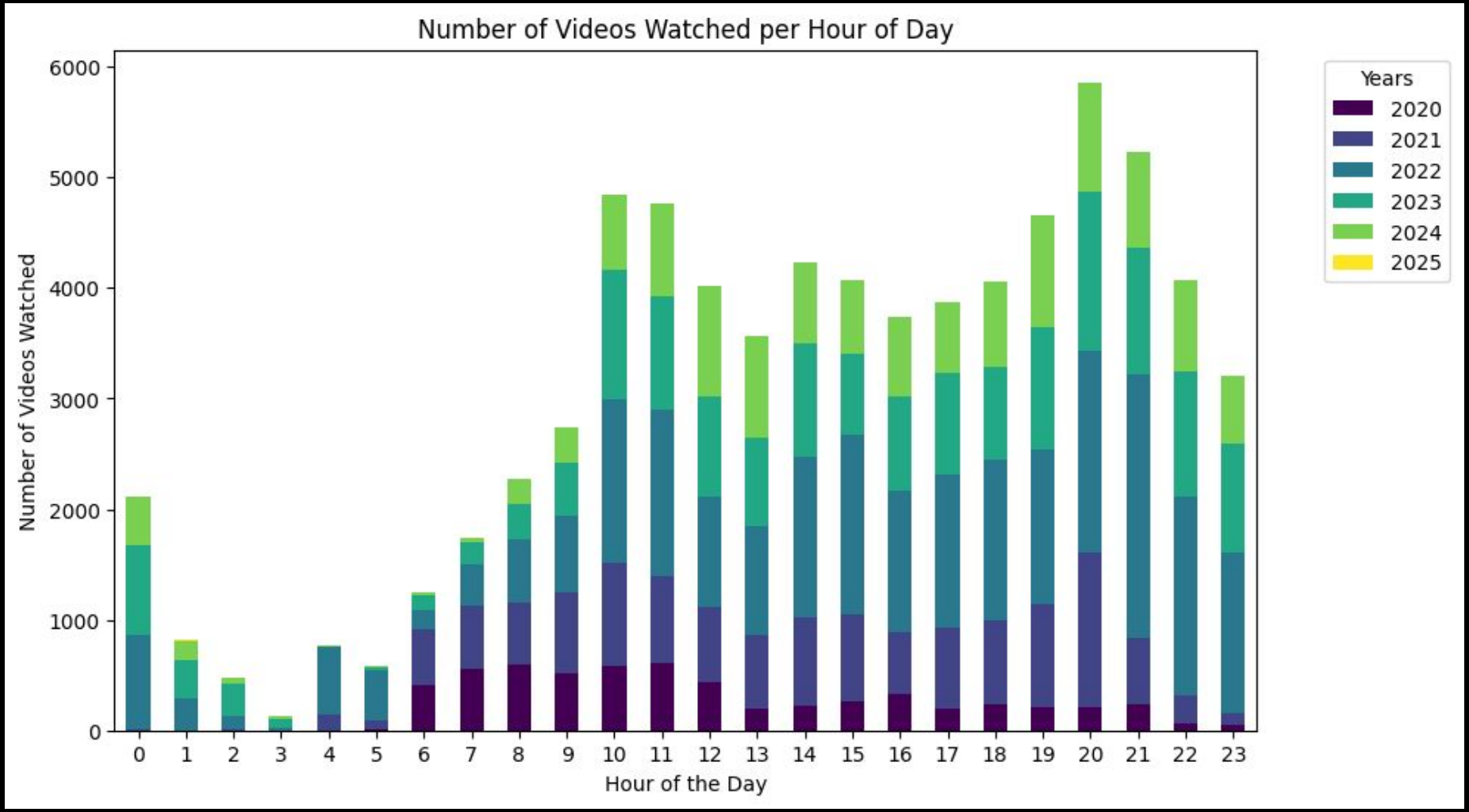

```
{
  "header": "YouTube",
  "title": "WW1 and Self-Inflicted Wounds adlı videoyu izlediniz",
  "titleUrl": "https://www.youtube.com/watch?v\u003dcf9eDFlawL0",
  "subtitles": [{
    "name": "Johnny Johnson",
    "url":
"https://www.youtube.com/channel/UCg7Q08KKOdjSrSnXcUL00Jw"
  }]
```

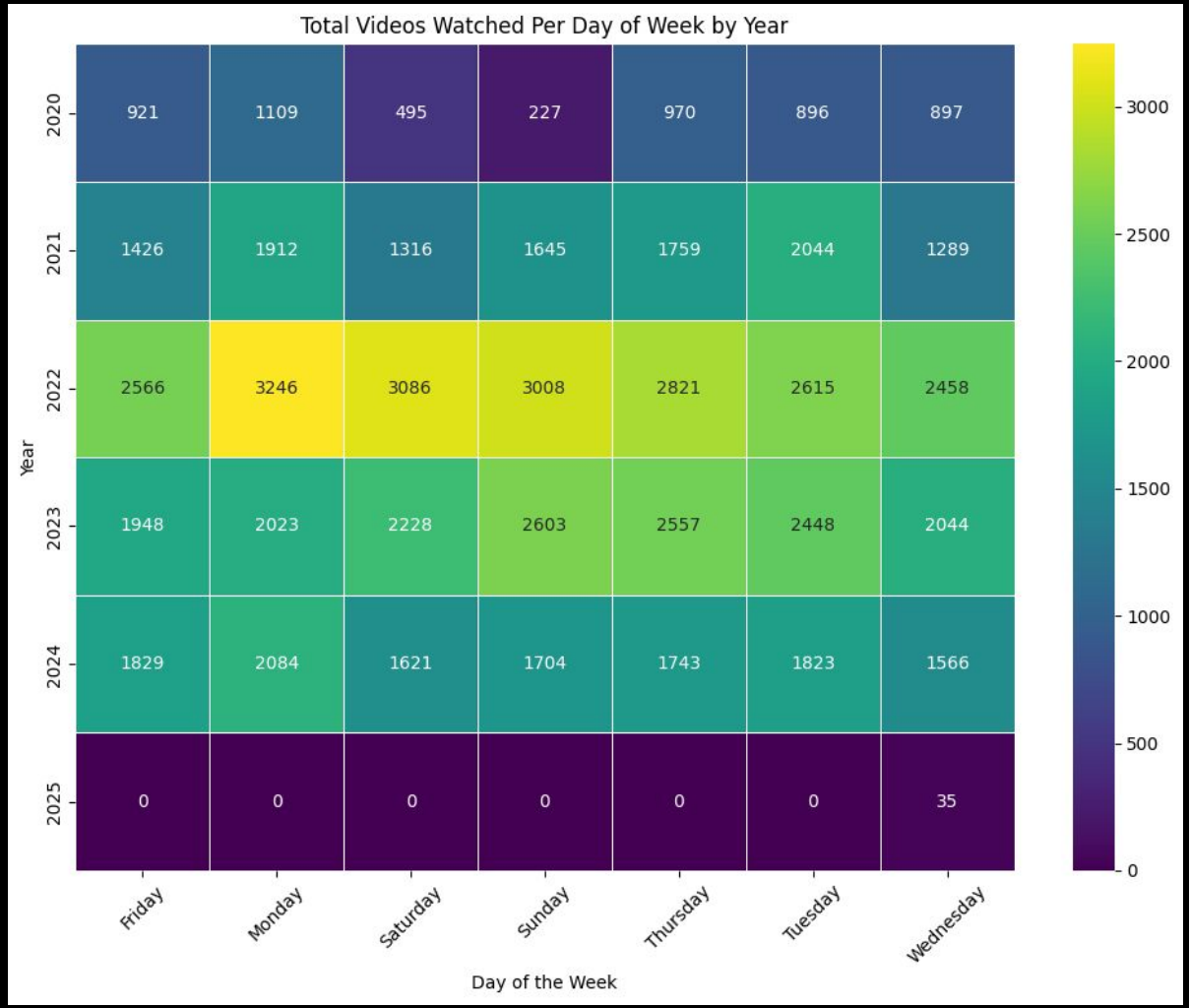

Total amount of Videos Watched
Cumulative Videos Watched
0
10000
20000
30000
40000
50000
60000
70000
2020-07
2021-01
2021-07
2022-01
2022-07
2023-01
2023-07
2024-01
2024-07
2025-01
Time


Number of Videos Watched Per Year
Number of Videos Watched
25000
20000
15000
10000
5000
0
2020
2021
2022
2023
2024
Year


Number of Videos Watched per Month
Number of Videos Watched
Month
Years
2020
2021
2022
2023
2024
2025
0
2000
4000
6000
8000
10000
January
February
March
April
May
June
July
August
September
October
November
December


Number of Videos Watched per Day of Week
Number of Videos Watched
Days of the Week
Years
2020
2021
2022
2023
2024
2025
0
2000
4000
6000
8000
10000
12000
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday


Number of Videos Watched per Hour of Day
Number of Videos Watched
Hour of the Day
Years
2020
2021
2022
2023
2024
2025
0
1000
2000
3000
4000
5000
6000
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23


Total Videos Watched Per Day of Week by Year
Year
2020
2021
2022
2023
2024
2025
921 1109 495 227 970 896 897
1426 1912 1316 1645 1759 2044 1289
2566 3246 3086 3008 2821 2615 2458
1948 2023 2228 2603 2557 2448 2044
1829 2084 1621 1704 1743 1823 1566
0 0 0 0 0 0 35
Friday
Monday
Saturday
Sunday
Thursday
Tuesday
Wednesday
Day of the Week
3000
2500
2000
1500
1000
500
0

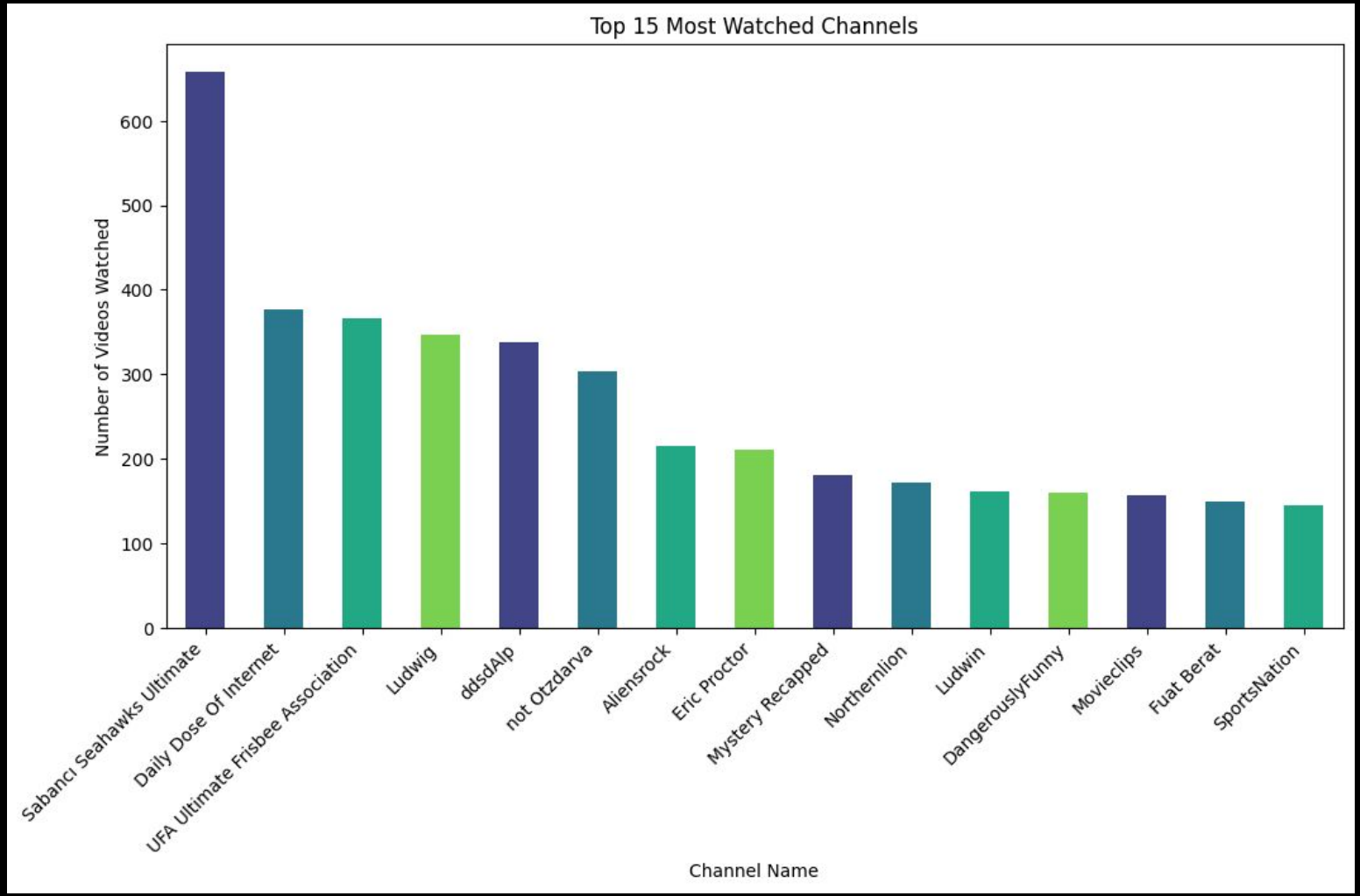
Hypothesis Testing

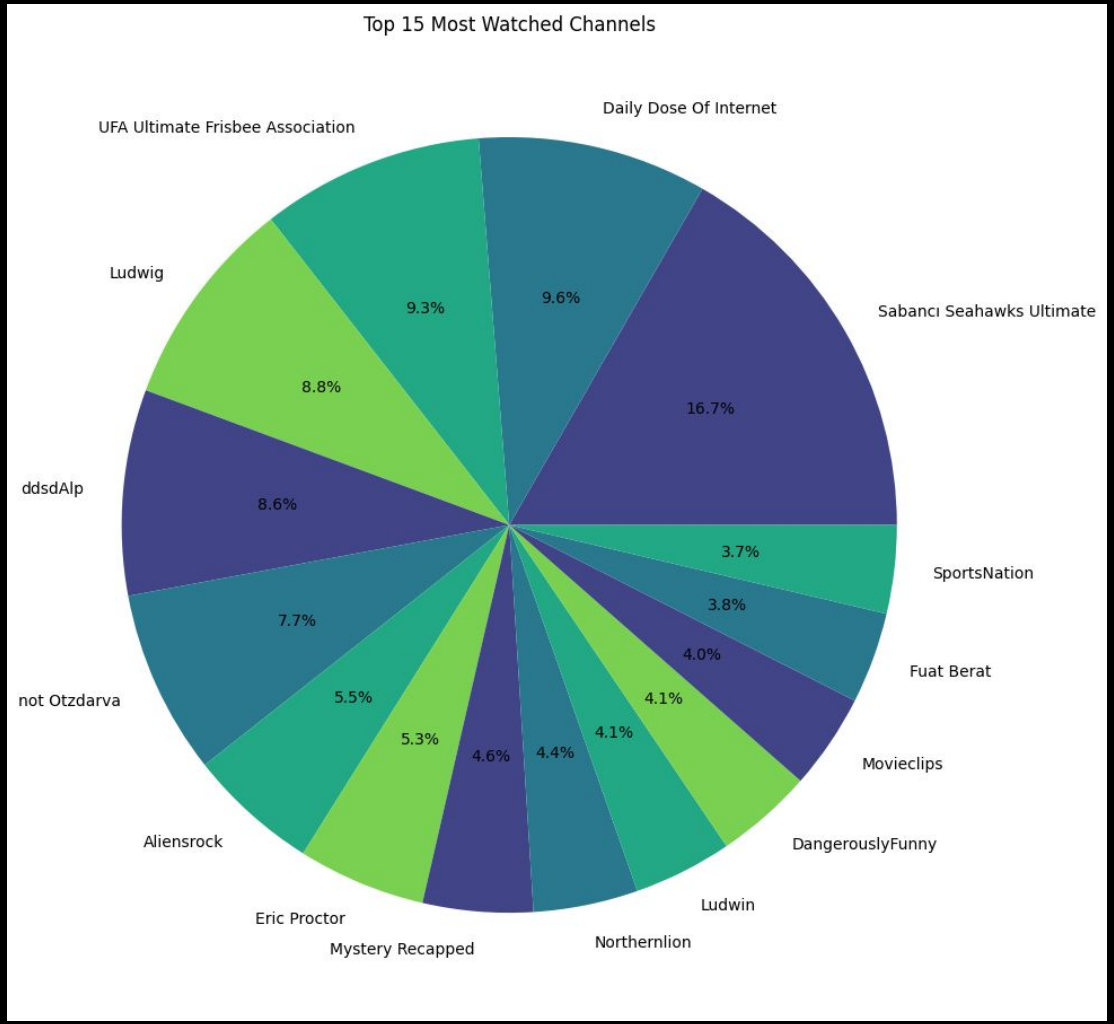
Hypothesis: The top 5 channels account for more than 10% of the watch time

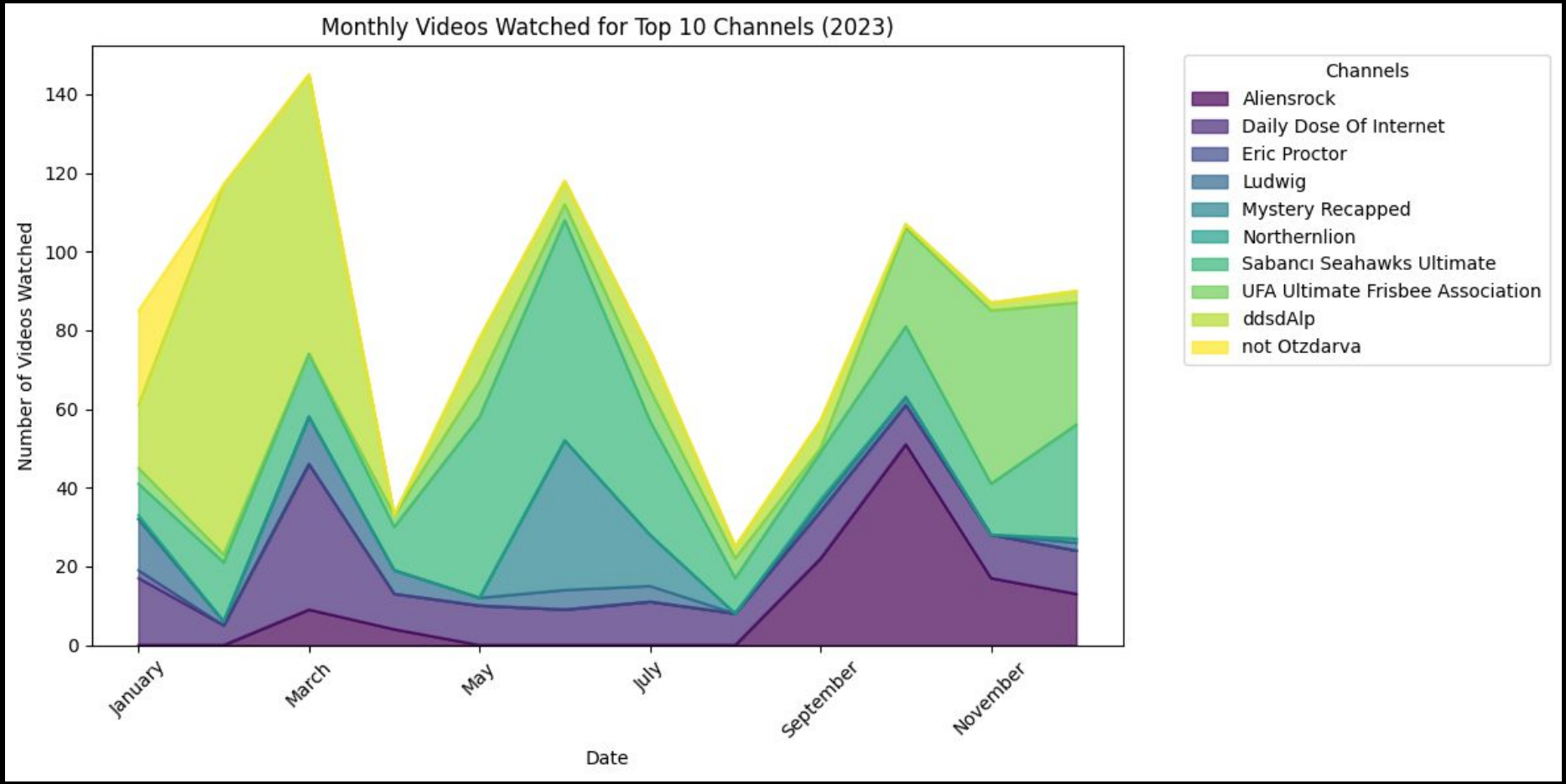
H0: The top 5 channels make up 10% or less of the videos watched
HA: The top 5 channels make up more than 10% of videos watched

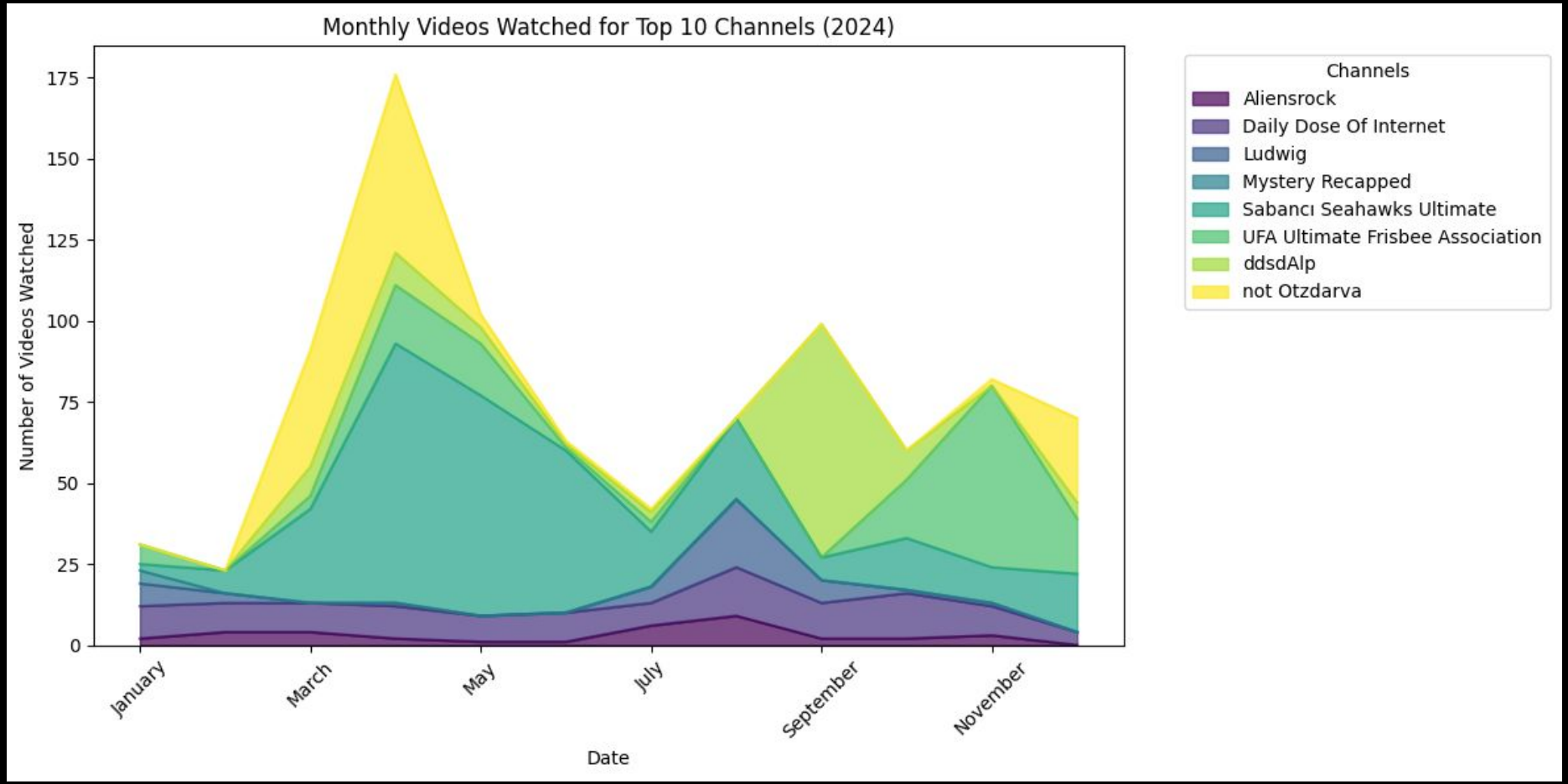
$$H_0 : p \leq 0.10\%$$
$$H_a : p > 0.10\%$$

(The value was 50% at first but I realized it was too much and lowered it down)


Top 15 Most Watched Channels
Number of Videos Watched
0
100
200
300
400
500
600
Sabancı Seahawks Ultimate
Daily Dose Of Internet
UFA Ultimate Frisbee Association
Ludwig
ddsdAlp
not Otzdarva
Aliensrock
Eric Proctor
Mystery Recapped
Northernlion
Ludwin
DangerouslyFunny
Movieclips
Fuat Berat
SportsNation
Channel Name


Top 15 Most Watched Channels
Daily Dose Of Internet
UFA Ultimate Frisbee Association
Ludwig
Sabancı Seahawks Ultimate
ddsdAlp
SportsNation
Fuat Berat
not Otzdarva
Movieclips
Aliensrock
DangerouslyFunny
Eric Proctor
Ludwin
Mystery Recapped
Northernlion
9.6%
9.3%
16.7%
8.8%
8.6%
3.7%
3.8%
7.7%
4.0%
5.5%
4.1%
5.3%
4.1%
4.6%
4.4%

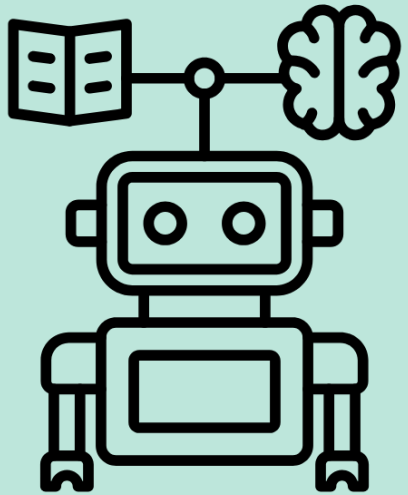

Monthly Videos Watched for Top 10 Channels (2023)
Number of Videos Watched
0
20
40
60
80
100
120
140
January
March
May
July
September
November
Date
Channels
Aliensrock
Daily Dose Of Internet
Eric Proctor
Ludwig
Mystery Recapped
Northernlion
Sabancı Seahawks Ultimate
UFA Ultimate Frisbee Association
ddsdAlp
not Otzdarva


Monthly Videos Watched for Top 10 Channels (2024)
Number of Videos Watched
0
25
50
75
100
125
150
175
January
March
May
July
September
November
Date
Channels
Aliensrock
Daily Dose Of Internet
Ludwig
Mystery Recapped
Sabancı Seahawks Ultimate
UFA Ultimate Frisbee Association
ddsdAlp
not Otzdarva

Machine Learning Model

- Creating a ML model to predict how many videos I'll watch in a day

- Features of the ML model
  - day_of_week
  - month
  - is_weekend
  - lag



Findings

- Almost never comment
- Quick change of interests and its correlation with my consumed content
- Not sticking to one/a group of channels
- My own channel is in my top 15
- The year 2022
- Active days, heatmap, active months

Limitations and Future Work

- I couldn’t get the packets for the textual analysis to work
- The dataset not having the watchtime of the video
- The confusion and problem this creates
- Future work if watch time was available
  - Avg. watch time of a video
  - Avg. percentage of the videos watched
  - Time spent on categories (gaming, education, coding, cars)
  - Time spent on channels