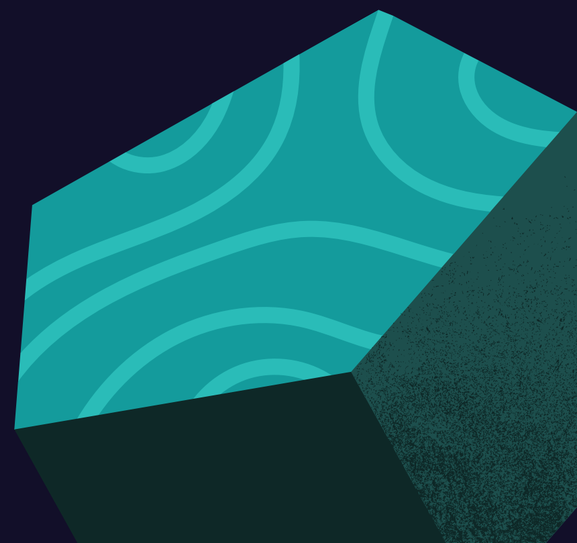
CS210 Project: YouTube Watch History

Emir Memis
30882

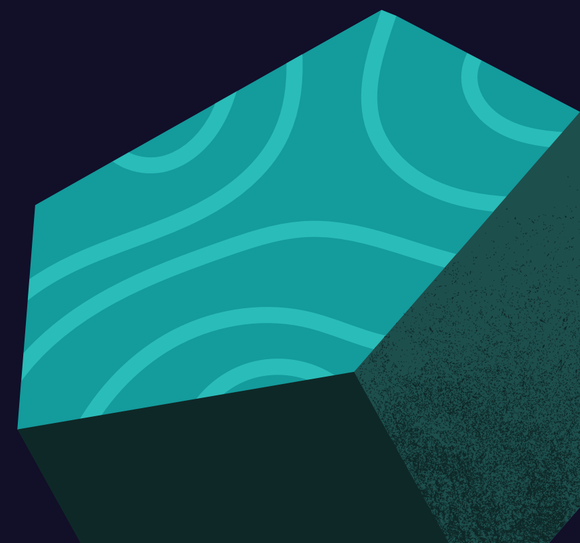
Motivation

The motivation behind this project stems from my recognition of YouTube as a pivotal component of daily digital interaction. As the application I engage with most frequently across my smart devices, it presents a unique opportunity for introspection and data-driven insight into my content consumption patterns.



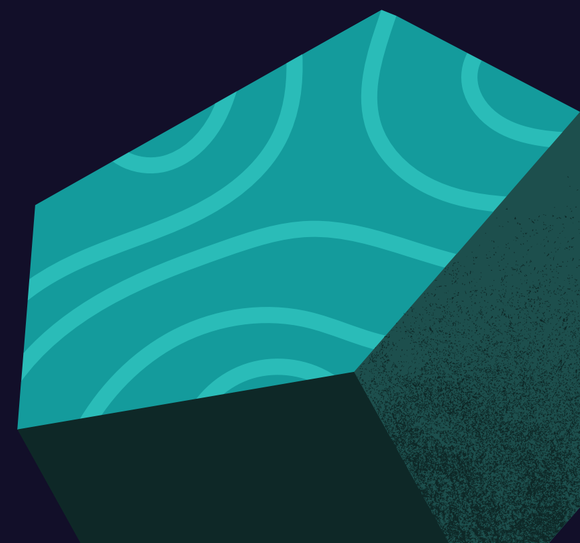
Data Source

The dataset for this project was obtained directly from Google. It was compiled through a formal request to Google's data takeout service, which allows users to access the data that Google has accumulated based on their activity across various services, including YouTube. This service provides a comprehensive archive of a user's interactions, engagements, and preferences, offering a rich resource for personal data analysis.



Data Analysis Overview of YouTube Viewing History

The project commenced with the extraction of YouTube viewing data, directly requested from Google's data takeout service. The data was initially in an HTML format, necessitating the use of BeautifulSoup for parsing and extraction. Key information such as video titles, channel names, and timestamps were meticulously extracted.



Data Extraction and Transformation

- Used BeautifulSoup to parse the HTML file.
- Extracted relevant data: video titles, channel names, and viewing dates and times.
- Customized date parsing to accommodate Turkish date formats using datetime and locale.
- Transformed the parsed data into a structured pandas DataFrame.

Data Analysis

Initial Data Cleaning

Employed pandas for preliminary data cleaning, including handling missing values and resetting indices after data filtering.

Feature Engineering

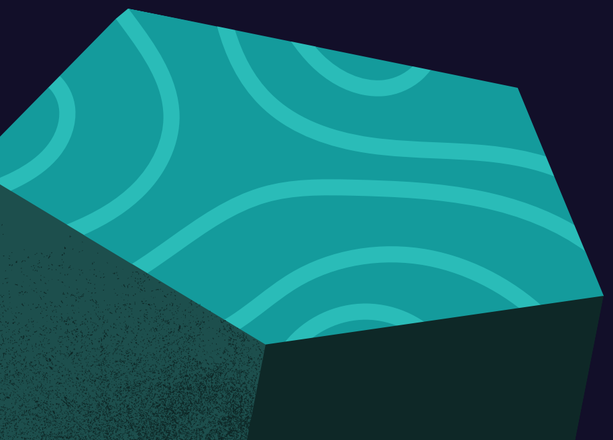
Enhanced the dataset by extracting additional features from the 'Date/Time' column, such as month, hour of day, and day of the week, to facilitate more detailed analysis.

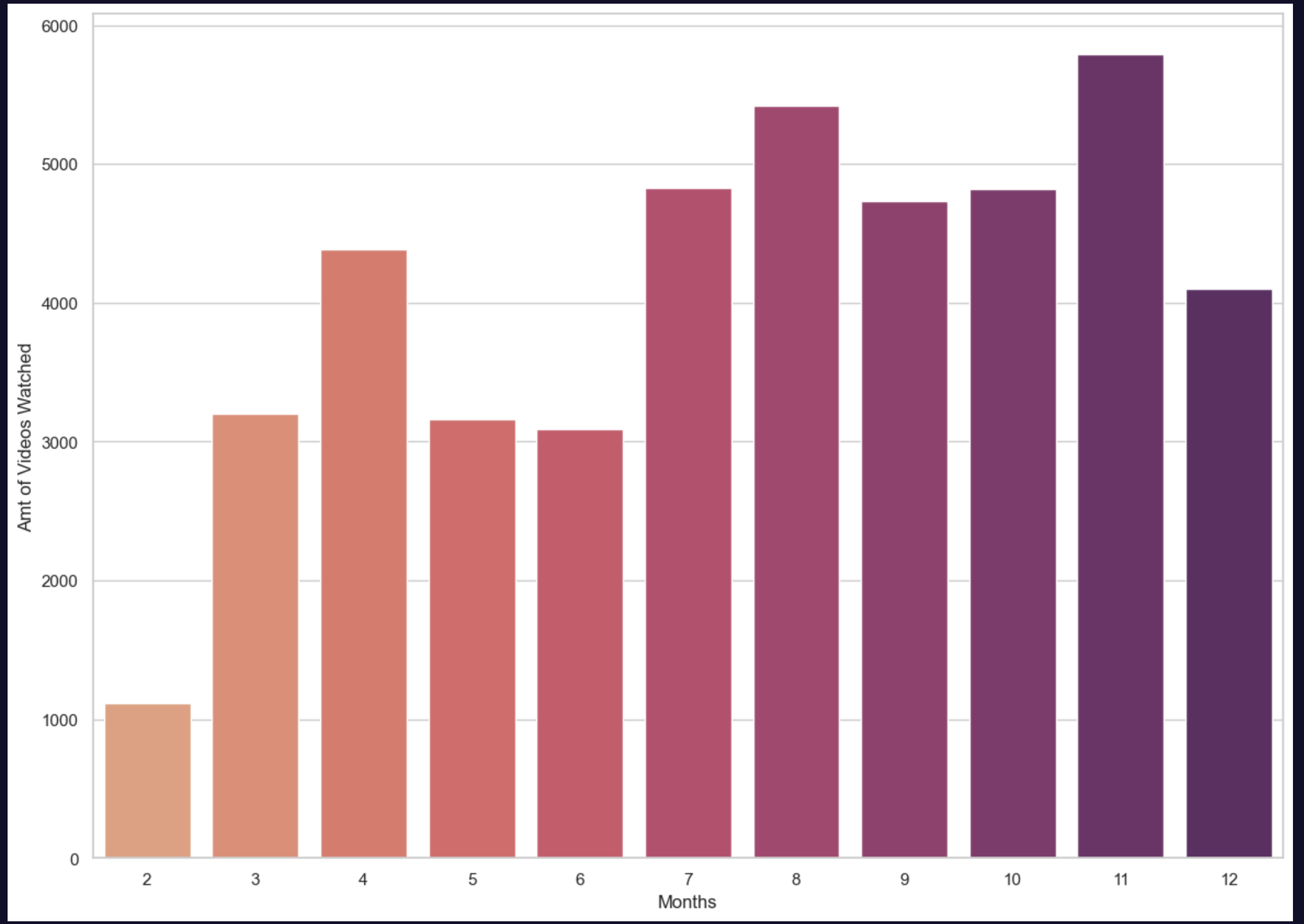
Exploratory Data Analysis

- Conducted a thorough exploratory analysis using pandas functions like describe() and custom summary functions to understand data distributions and identify unique values.
- Investigated specific patterns, like channel preferences and viewing habits over time.

Visual Data Analysis

- Utilized seaborn and matplotlib for visual exploration, creating bar plots and heatmaps to reveal trends and patterns in monthly and hourly video consumption.
- Grouped data by various time dimensions (e.g., year, month, day of the week) for more granular insights.




6000
5000
4000
3000
2000
1000
0
Amt of Videos Watched
2
3
4
5
6
7
8
9
10
11
12
Months

Data Analysis - Continued

Specialized Analysis

- Conducted targeted analysis on specific video categories like Python-related content.
- Implemented text analysis techniques, including word frequency analysis and word clouds, to explore common themes in video titles.

Political Ideology Prediction

- Developed a custom function to categorize videos based on political keywords, enabling an estimation of political leanings from the video titles.
- Visualized the distribution of political ideologies in the viewing history.

Technological Stack

- Python for data processing and analysis, with key libraries including pandas for data manipulation, seaborn and matplotlib for visualization, and BeautifulSoup for HTML parsing.
- Jupyter Notebook as the interactive environment for coding, analysis, and visualization.


#shorts
Bölüm
Arşivi
Sungur
Souls
Elden
Pokemon
Marşı
#keşfet
Atatürk
WIRED
Belgeseli
GÜNLÜK
#viral
#edit
Mayıs
Erdoğan
HNGKA
Piece
Pokémon
32.Gün
Armoni
Nasıl
Mustafa
Final
Deniz
Dragon
Matematik
Kenan
Metal
İsmet
Eylül
Persona
Emrah
Kemal
#onepiece
Theme
ELDEN
Fight
#tarih
Opening
@MertGunhan
Türkçe
Golden
İlişkideki
Mızıkası
Cover
miyim?
(Official
Tayyip
Ecevit
#anime
#reels
BÖLÜM
İzmir
Armored
Gürkan
Beyblade
Fantasy
Walkthrough
#Shorts
#short
Olasılık
BAKIŞ
Bülent
Evren
İnönü
YATAY
#pokemon
Atatürk'ün
Türkiye
Ranking
Celal
Sekiro
Turkish
DİYON
Neden
#starwars
Every
Ayrık
Remastered
#luffy
Kurtuluş

Findings

Preferred Content and Channels:

- Analysis revealed a clear preference for specific content types and channels. For instance, a significant portion of my viewing history was dedicated to Python programming tutorials and technology-related content, indicating a strong interest in these areas.
- Certain channels appeared more frequently in the top-viewed list, suggesting loyalty to these content creators or a particular affinity for their style or subject matter.

Viewing Patterns Over Time

- The temporal analysis of viewing history, broken down by month, year, and hour of the day, uncovered distinct patterns. For example, there was a noticeable increase in video consumption during specific periods, possibly correlating with leisure time or professional learning phases.
- Peak hours of YouTube activity could suggest my preferred times for consuming content, whether for relaxation or education.

Shifts in Interests

A year-wise breakdown indicated shifts in interests over time. For instance, an increased number of videos related to a particular subject in a specific year might reflect a newfound interest or a professional requirement during that time.

Political Leanings

The categorization of videos based on political keywords provided an intriguing glimpse into the political nature of the content consumed, although this should be interpreted cautiously given the subjective nature of such categorizations.

Word Cloud Insights

The word cloud generated from video titles offered a visual representation of the most frequent terms, highlighting dominant themes and topics that resonated with me.

Behavioral Insights

The exploration of viewing habits, such as the tendency to watch certain types of videos at specific times, offered a deeper understanding of how I engage with digital content, potentially influenced by daily routines, mood, and personal interests.

Content Evolution

Observations of how my content preferences have evolved over time, possibly indicating changes in my professional interests, personal hobbies, or lifestyle.

Limitations and Future Work

1

Data Scope

The dataset mainly includes titles and timestamps, missing detailed metadata like video descriptions or engagement metrics. This limits the depth of analysis.

2

Categorization Subjectivity

The use of predetermined keywords for categorizing content could introduce bias or inaccuracies.

3

Lack of Comparative Context

The analysis is limited to my data, without comparing it to broader trends or other users.

Future Plans

Advanced Analytics

Implement machine learning for more nuanced categorization and sentiment analysis.

Extended Temporal Analysis

Analyze over a longer period to observe evolving preferences and habits.

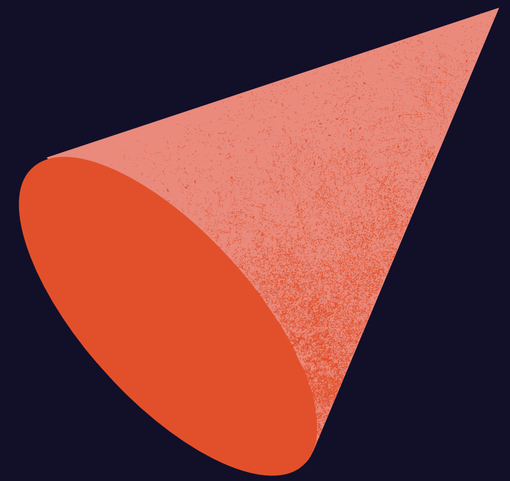
Cross-Platform Analysis

Include data from other platforms for a comprehensive view of digital consumption behaviors.

Ethical Considerations

Ensure privacy and ethical standards in all future analyses.

"The past informs the present and inspires the future."



-Unknown source.