DSA210 Youtube Watch History Analysis

Motivation

- While trying to find a source that has easily accessible data and that I have spent an observable amount of time, I found YouTube. With my current data, I have about 4 years of data that can provide for the project and help me to reach to a conclusion about my questions about the data. The questions that I want to answer are:

- "Which categories or channels did I spend the most time watching?"
- "Are there seasonal patterns in my YouTube usage?"

Data Source

I got my data from two main sources:

Google Takeout:

It was very easy to get my watch history from google takeout, the request was given in approximately 1-2 hours after I created a request. The requested data was in JSON format, so I needed to fetch the wanted parts manually from the data in raw format. The data provided by google takeout includes my watch history with:

- -Duration of the video
- -Watched time
- -Published time
- -Name of the video
- -Tags (#gaming, #life etc.)

YouTube API v3:

This was the most crucial source as parsing the data into categories was not possible with google takeout only, the youtube data api helped to fetch categories much easier as they categorize the videos themselves, which was helpful in my case. The videos fetched with youtube api was more modifiable compared to google takeout, but its quota limit was a big drawback for the fetching progress as it was never enough to fetch in 1-2 days, the progress took longer than 5 days as I needed to fetch both my categories and trends at the time. The data fetched from YouTube API v3 includes:

- -Durations
- -categories
- -Sub categories (FIFA/FC, Minecraft etc. under gaming category)

In addition to categorizing my watch history, the API also helped fetch global trends for specific categories like Gaming. This allowed for meaningful comparisons between my viewing habits and broader YouTube trends.

Both sources were the core of data fetching as there is an easy access around both and the provided data was readable and totally writeable from start to end.

Data Analysis

Data Cleaning and Preprocessing

- **Google Takeout Data**:
- Extracted relevant fields (e.g., video title, watched time, published time, duration, tags) from the JSON files provided by Google Takeout.
- Converted the raw JSON data into a structured format. Using libraries like pandas
- Handled missing or incomplete data by:
    - Dropping entries without essential fields (e.g., missing title or watched time).
    - Converting timestamps to a standard format for easier analysis.
    - Some of the videos were deleted or made private, so I needed to remove them from the data set.

- **YouTube API Data**:

- Retrieved additional fields like video categories, subcategories, and durations using the YouTube Data API. This helped with categorizing the most watched channels and videos with minutes instead of just number of videos watched.
- Combined data fetched via the API with the cleaned Takeout data to enrich the dataset. By doing this saving the categorized data and plotting it became much easier.
- Implemented progress-saving techniques to avoid re-fetching and ensured data integrity.

This enriched dataset allowed for detailed analysis of watch habits and trends. While the API provided valuable insights, its daily quota limits required progress-saving mechanisms and delayed the fetching process.

Together, these sources offered complementary data that enabled comprehensive analysis.

Data Summarization

- Aggregated data to identify key trends and patterns:
- **Monthly Watch Time**:
    - Grouped data by month and category to calculate the total watch duration for each month. By this way I analyzed the data under 48 different sub-categories to get better results.
    - Analyzed seasonal trends to see how viewing habits changed over time.
- **Most-Watched Channels and Categories**:
    - Summarized the total watch time for individual channels and categories. I also wanted to fetch the most watched channels to check if the most watched games have a correlation with the most watched games.
    - Identified the most-watched channels and specific games, such as FIFA/FC, Minecraft, and Clash Royale. As the categorization should be done manually, I created the categories and put the data under its category.

Data Visualization and Plotting

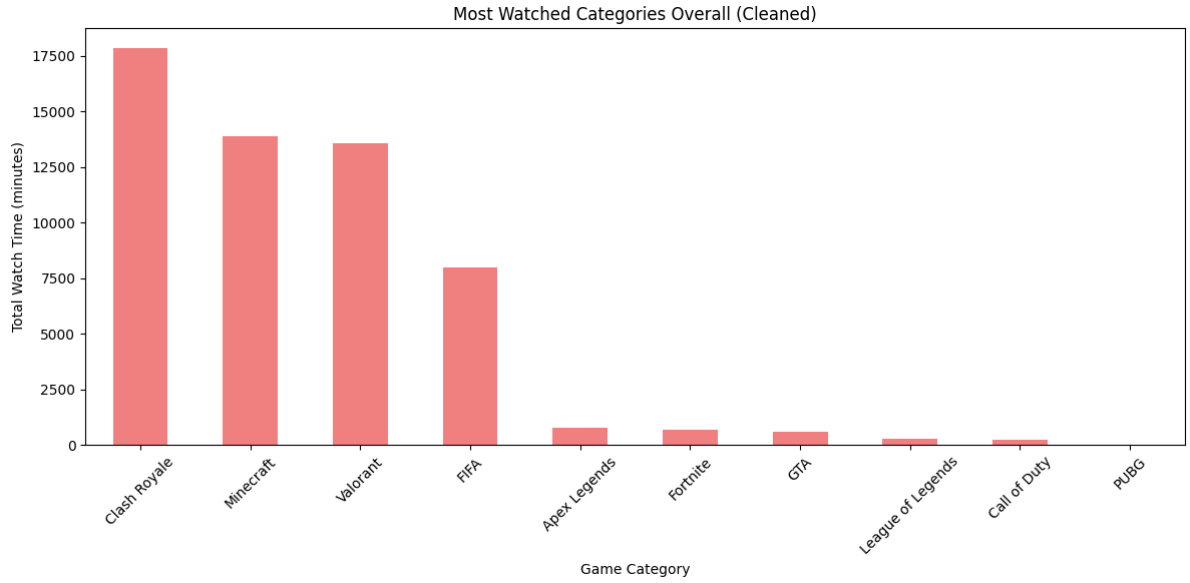
Visualizations were created to highlight monthly watch time, top categories, and channel-level analysis. These plots provided insights into personal viewing habits and their alignment with global trends. The categories were:

- Games
- Duration (in minutes)
- Watched date

By using these categories, I was able to compare my own data with the trends of that time and get a result about the correlation between my genre of watched videos and the gaming trends of the time being.

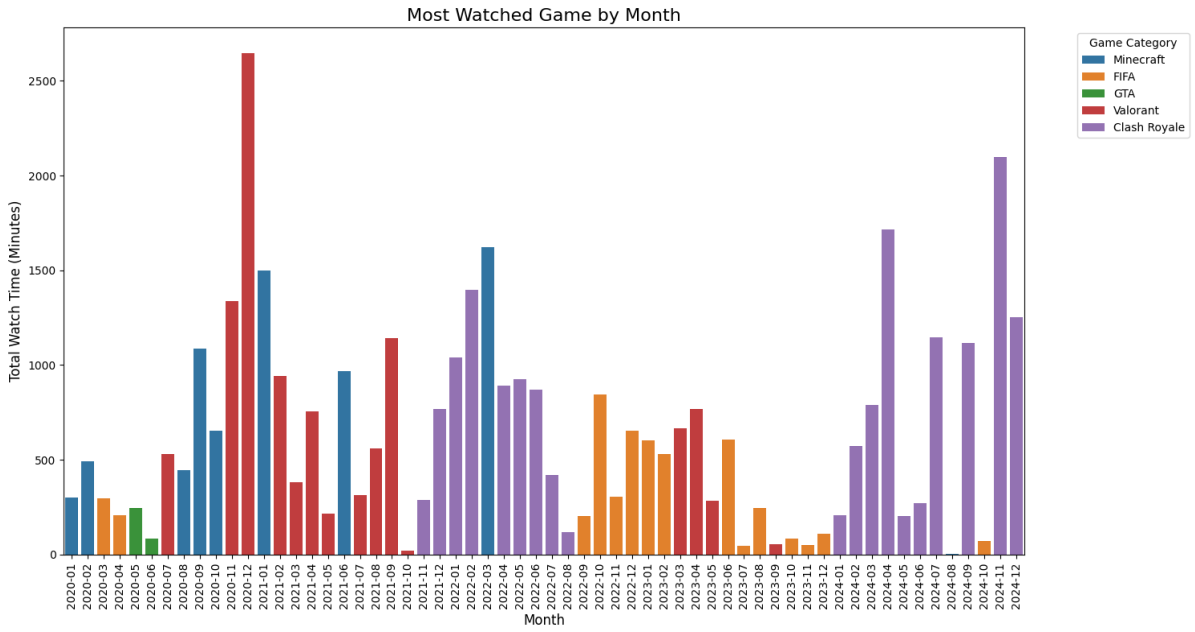
Findings

Firstly, I took my categorized data and plotted in month by month to see which channels and game genres I watched the most. The outcomes of my watch pattern did not surprise me at all, but the order of the list got me a little bit surprised. Here is the data plot of my 4 years of YouTube:



So, this is the cleaned version of my data, the uncleaned version had deleted videos and private videos that labeled as “other” category, which caused some issues on

watched minutes, so I tried to fix it by removing the unknown data from the set. The order of the data being like this surprised me as I have big blanks with watching clash royale. On the other hand, I watched and played FIFA most of my life, so I was expecting that it would be top 2 at least, which left me in confusion. But if we look at the general layout of the data, I would not say that I am surprised. As I was also interested in my monthly data, I also plotted it too and it looked like this:



As I said before, I watched and played FIFA all my life and this can be seen by looking at the graph, there is data of FIFA starting from 2020 to 2024 all the time. And also from the data, I can conclude that I had so much free time in the pandemic as it has the most watch time in the graph. Also, Valorant being the top watched game type in the 2020 is because that the game was launched in 2020. If we comment on every game:

Minecraft:

- Watched frequently until 2022, the reason that I stopped watching is that my favourite channel's owner sadly passed away.
- Second in the top watched genre list.

FIFA:

- It is in the data starting from 2020 all the way to 2024.
- Frequently watched, peaked in 2022.
- Top 4 in the most watched list.

GTA:

- Just topped for two months, then wasn't seen again.

- Has no place in the top 5 but still made to the most watched of a month, which quite surprised me.

Valorant:

- Peaked in 2020, which is the launch year of the game.
- Watched many times as it was the pandemic year.
- Top 3 on the most watched list.

Clash Royale:

- Most watched genre in the list, there are traces of it starting from 2021 all the way to 2024.
- The game became popular in 2021, then I did not really stop watching it.

HYPOTHESIS TEST

The tests are done on signifigance level 0.05.

I tested my hypothesis which was “My watch time increases in summer months” against the data and I got a p-value of 0.0166. This led me to reject the null hypothesis, this is also understandable by just looking at the data, on the contrary, the winter months seem to be more active than the other months, so I also put this to the test with the hypothesis “There is an increase in the watch time in winter months compared to the other months.”. This test’s p-value resulted as 0.1922, so we failed to reject the hypothesis and the data we have has shown that there is an increase in winter months compared to the other months.

To summarize:

Hypothesis #1:

- “My watch time increases in summer months”.
- p-value: 0.0166
- Result: Reject the null hypothesis

Hypothesis #2:

- “There is an increase in the watch time in winter months compared to the other months.”
- p-value: 0.1922

- Result: Fail to reject the null hypothesis.

Limitations and Future Work

While the analysis provided meaningful insights into my YouTube watch history and its alignment with global trends, several limitations were encountered during the project:

- **Quota Limit**: The YouTube Data API's daily quota limit required splitting the data-fetching process over multiple days, slowing progress and increasing complexity. This made the process a lot harder than it should be.
- **Manual Categorization**: Assigning videos and channels to specific subcategories required manual intervention, which introduced potential bias and limited scalability, which could result in more correlation between most watched games and trends.
- **Deleted or Private Videos**: Some videos were no longer accessible, resulting in incomplete records, particularly for older data. This made the categorizing process to hard as there was no information about the private and deleted videos. I needed to remove them from the dataset to make a better visualization.
- **Untitled or Incomplete Data**: Missing fields like video titles or durations led to the exclusion of certain entries, potentially affecting the overall analysis.
- **Trend Alignment**: The fetched trends may not fully represent global YouTube viewership patterns, as they were based on search results rather than actual viewing metrics. I tried to fetch the trends, but there was no way to do it as there is no API that could manage it.

To address these limitations and enhance the scope of the project, several areas for future work have been identified:

- **Automated Categorization**: Implementing a machine learning model to classify videos and channels more accurately and efficiently.
- **Data Completeness**: Exploring alternative approaches to infer missing data, such as using channel-level statistics.
- **Additional Features**: Incorporating metrics like view counts, likes, and comments to provide a richer context for analysis.
- **Trend Comparison**: Supplementing YouTube API trends with data from alternative sources to ensure more comprehensive comparisons.

- **Interactive Visualizations**: Developing dashboards to enable dynamic exploration of the data and trends for deeper insights.

These enhancements would not only address the current limitations but also expand the analytical capabilities of the project, leading the way for more meaningful conclusions.