



My Youtube History

DSA 210 2024-2025 FALL

Yeşim Tosun

32205



Motivation

- YouTube has been the application I've used consistently for as long as I can remember. For this project, I chose YouTube because it reflects my habits and interests more accurately than any other platform, as it is one of my most frequently used applications.
 - I decided to analyze this **7 year dataset** to test my hypotheses and discover how my YouTube behaviors have evolved over time.
-

Data Collection

Requested my Youtube watch history from Google Takeout.



Parsed the video-ids of the watch-history elements



Fetches relative meta data of the corresponding video-ids using Youtube API v3.





Data Processing

- Fetched data: Video category, video watch date, video duration.
- Created JSON Files for each year (e.g. 2017, 2018, ..., 2024)
- Each file consisting of :
 1. total duration watched,
 2. category counts,
 3. category watch durations for 12 months in a year.

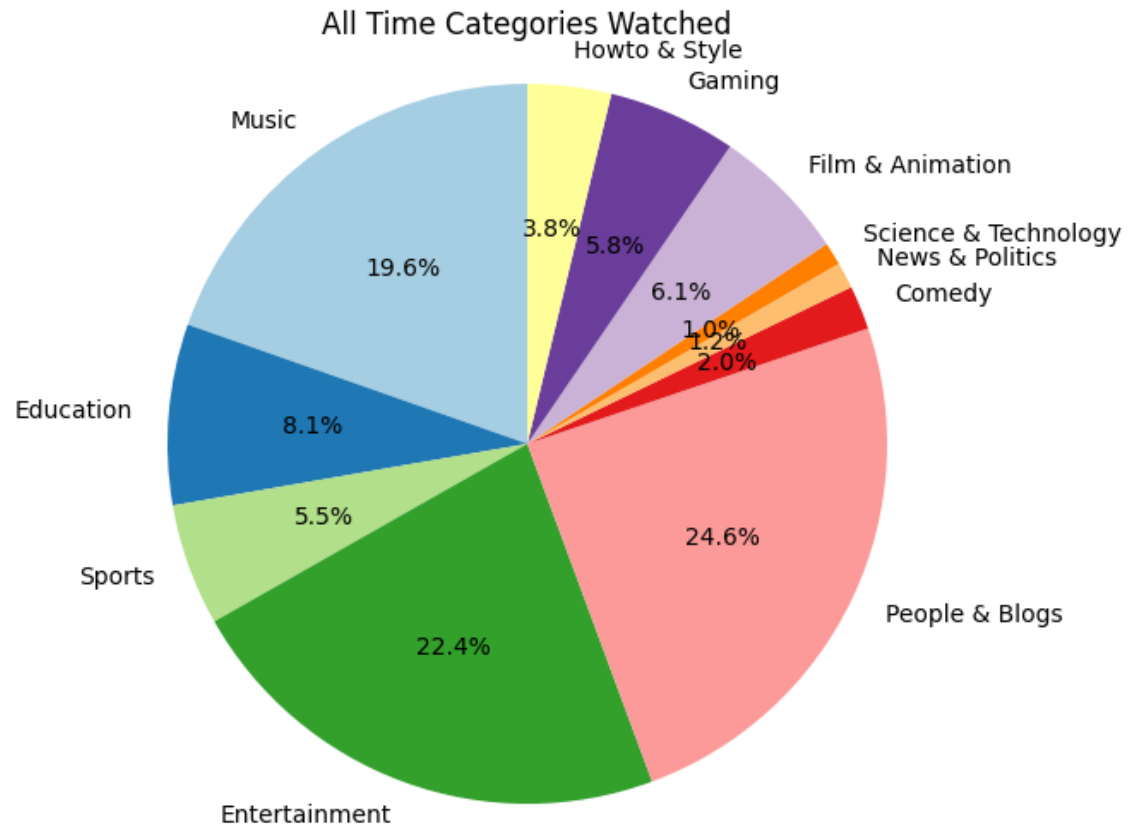


Data Analysis

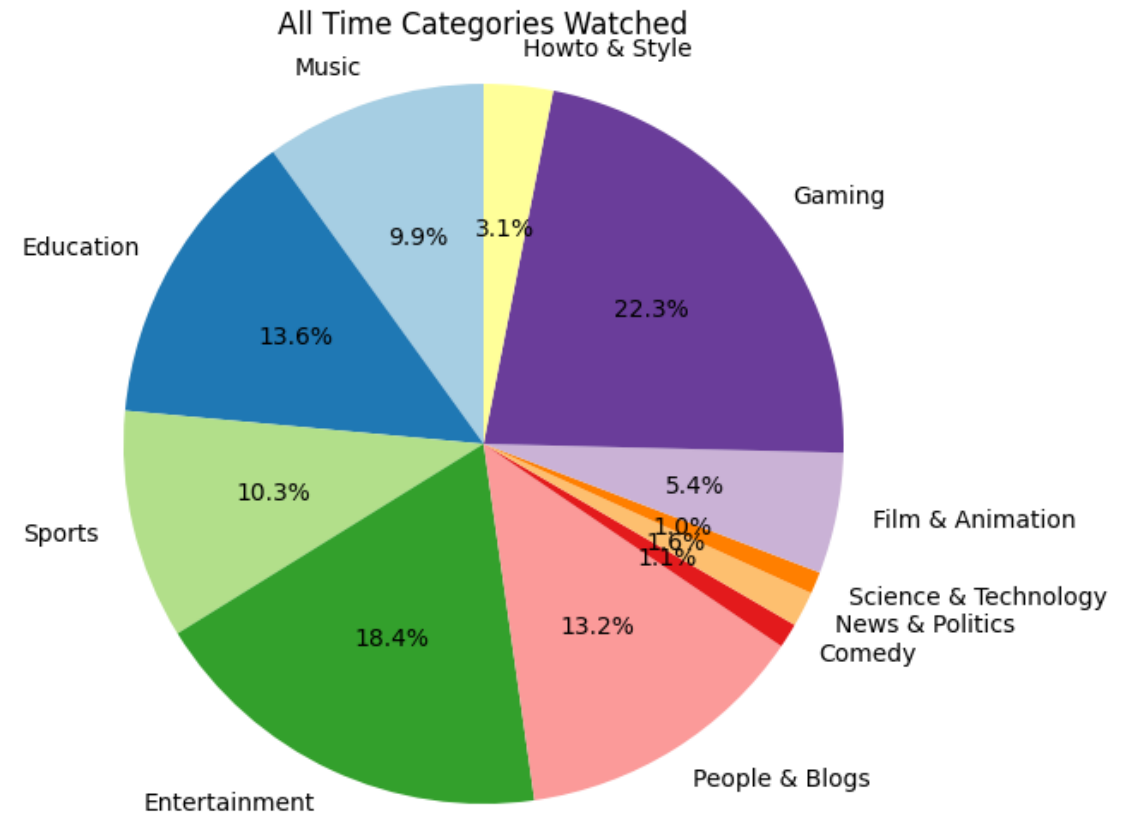
- 
- Starting from general to specific, I analyzed the data processed.

All Time Categories Watched:

- Watch quantity or duration?

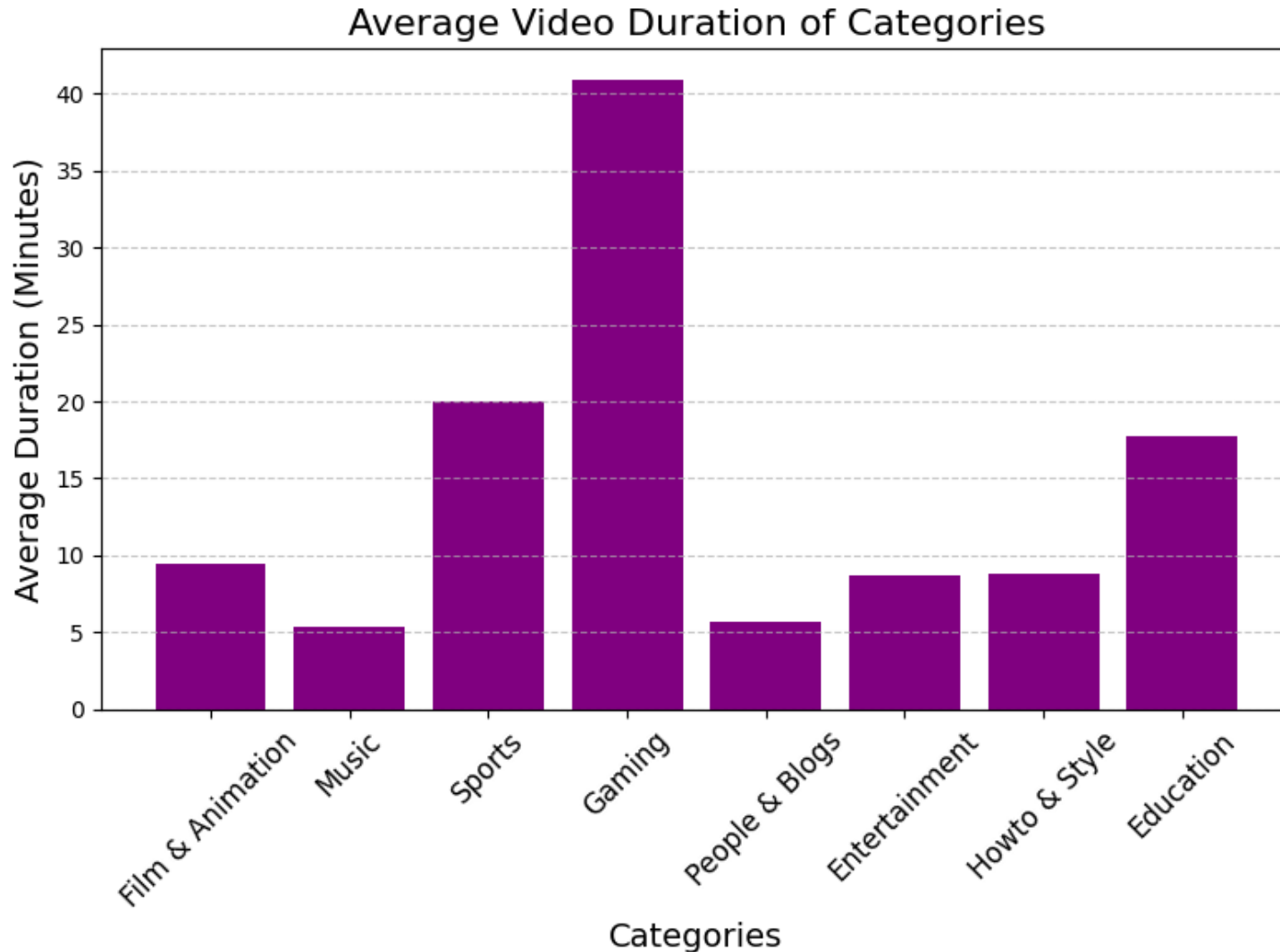


All Time Watch Time(Quantity) of Categories

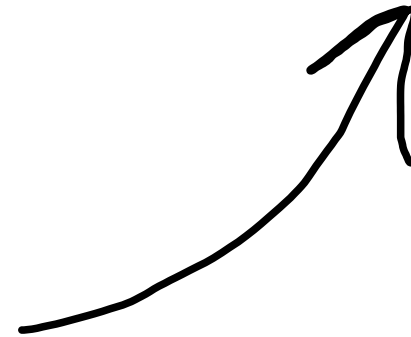


All Time Watch Durations of Categories

- As the percentages vary, I wanted to calculate the average video length of my watched category using my own data.



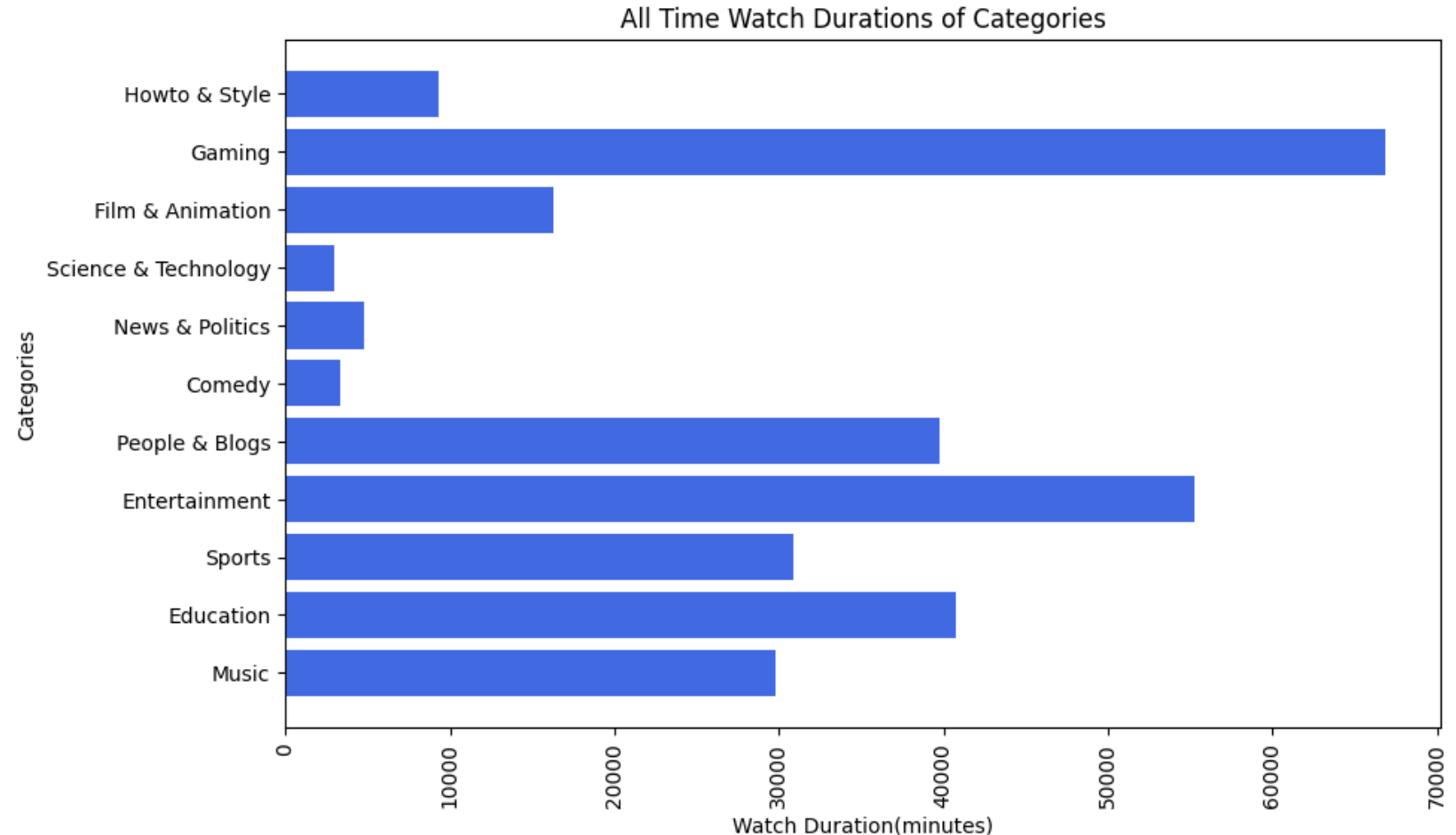
Observed in the plot, the longest videos watched are in the category "Gaming" with average length of 42 minutes.



- As the data for category watch duration and category watch times differ from each other, I will focus on using "category watch durations" for my further analysis.

Analysis of All Time Data

Longest watched categories:
Gaming and Entertainment



Hypothesis Testing

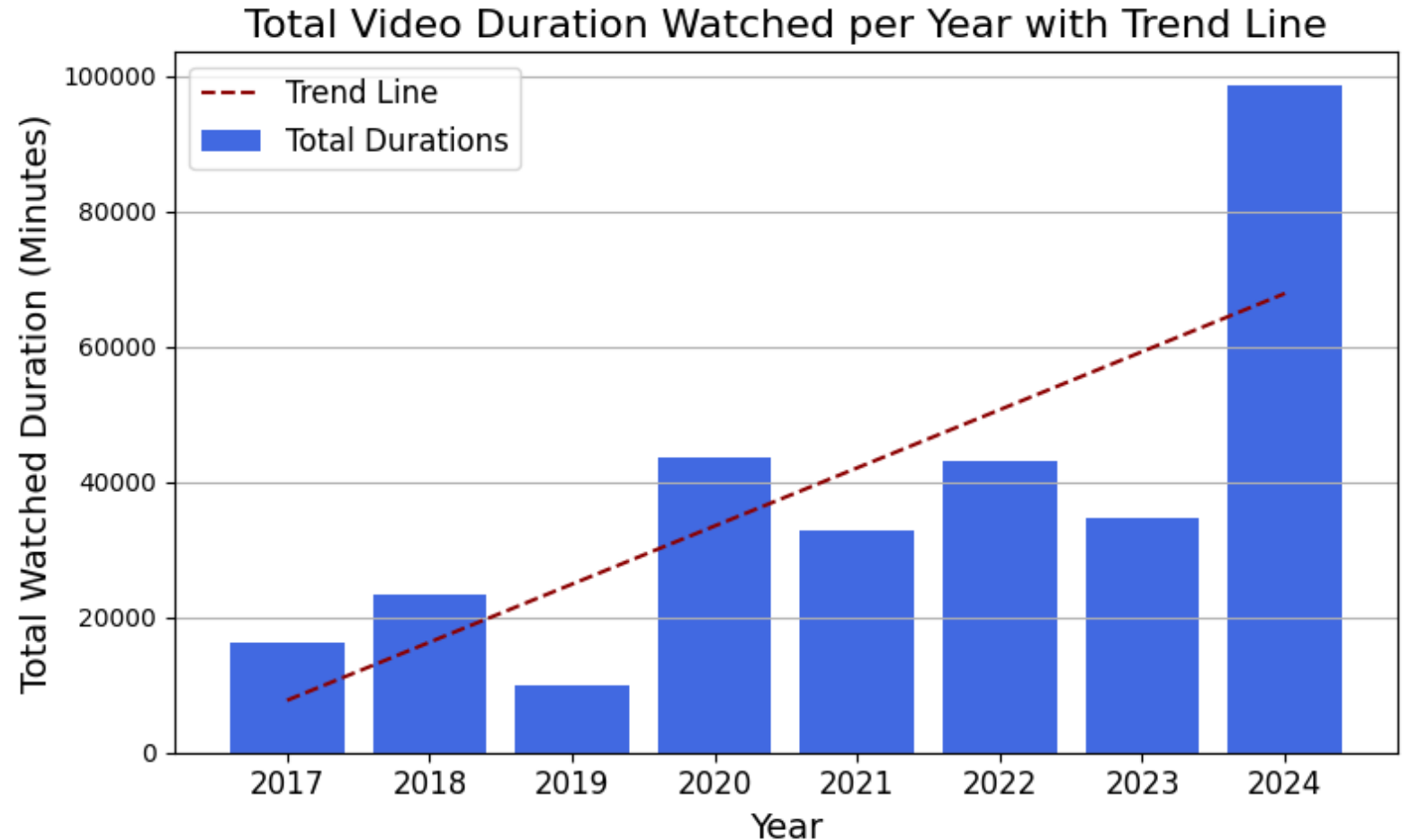
"My watch time has increased significantly over the years."


-Analysis by using Linear Regression Model

- **Null Hypothesis:** No significant upward trend in video watch durations.
- **Alternative Hypothesis:** There is a significant upward trend in video watch durations over the years.

Hypothesis proven right.


P-value: 0.0257
Reject the null
hypothesis: There is
a significant upward
trend in video watch
durations over the
years.





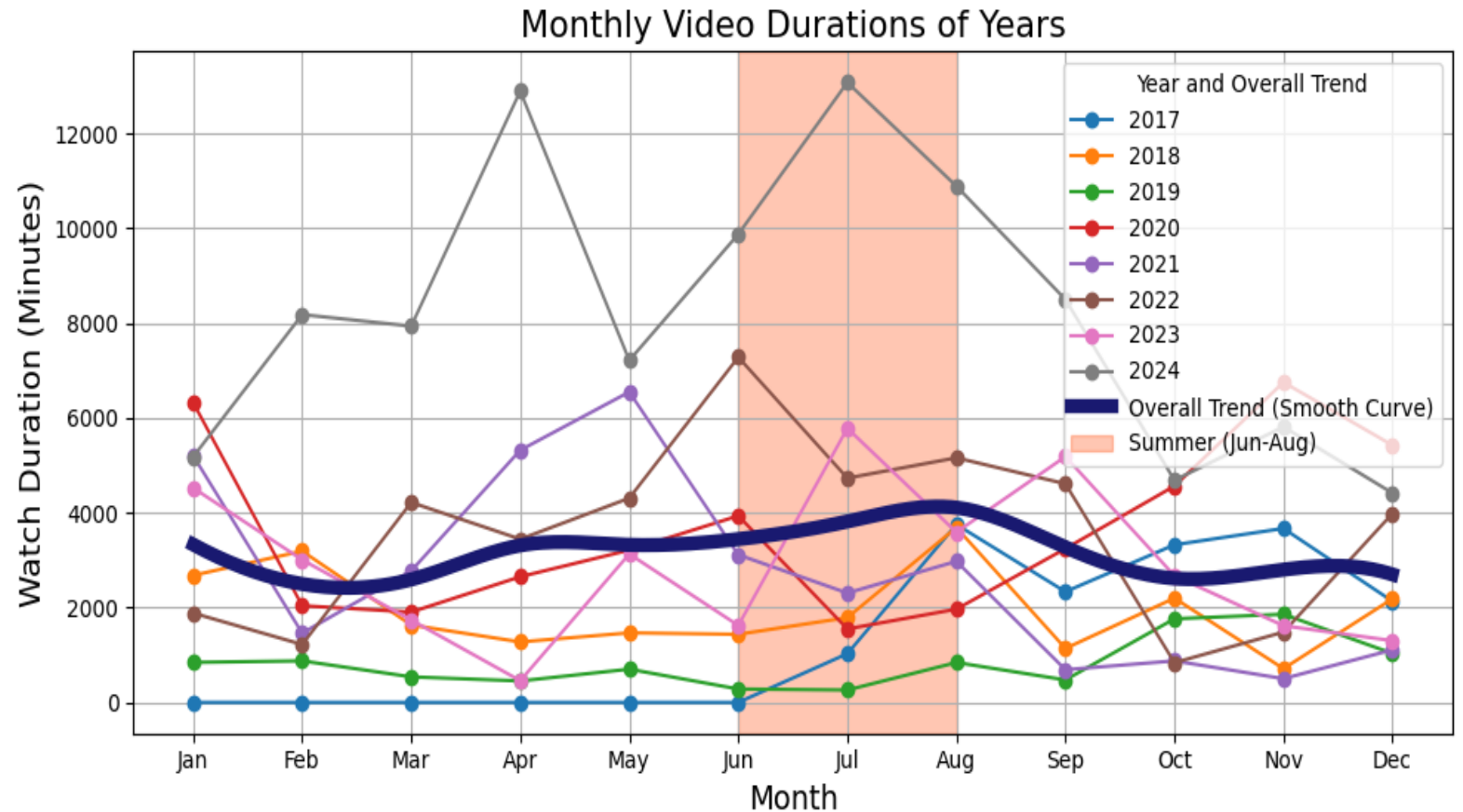
"My watch durations are significantly higher in summer."


-Using ANOVA F-statistic

- **Null Hypothesis:** No significant difference in summer watch durations.
 - **Alternative Hypothesis:** Watch durations are significantly higher in summer.
- 

Hypothesis proven wrong.

P-value: 0.4158 Fail to
reject the null
hypothesis: No
significant difference
in summer watch
durations.



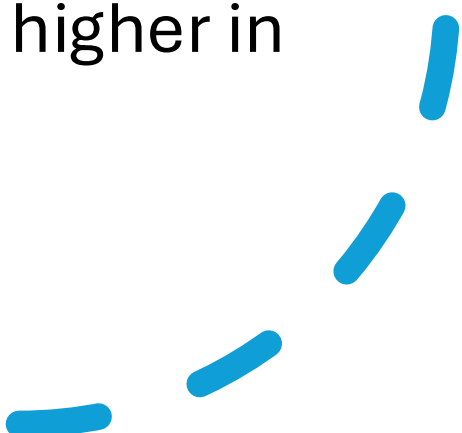


"I watched significantly more educational videos in 2022 while preparing for my university entry exam."

-Using t-test

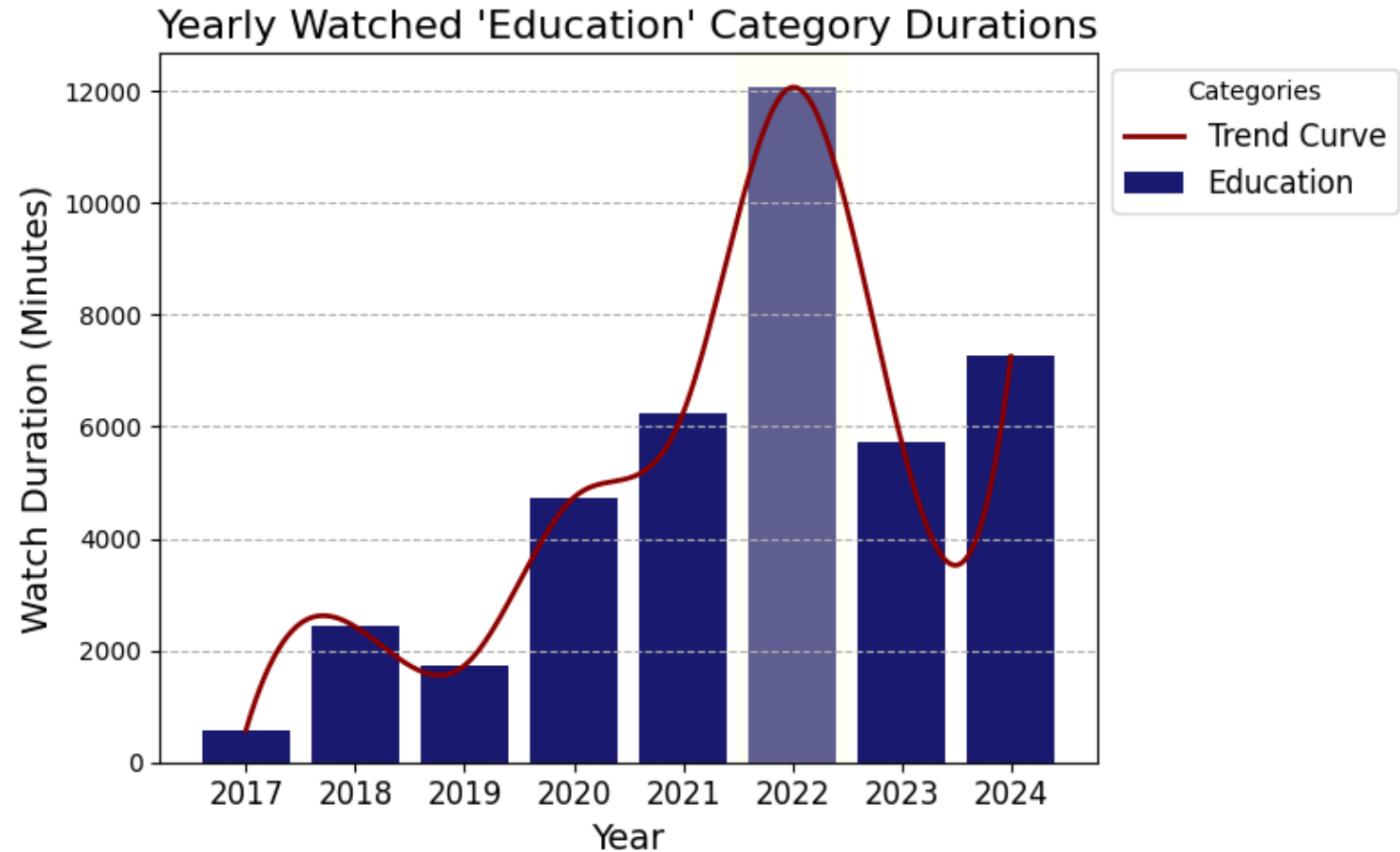
Null Hypothesis: No significant difference in educational video watch durations in 2022.

Alternative Hypothesis: Educational video watch durations are significantly higher in 2022.



Hypothesis proven right.

P-Value: 0.0002 Reject
the null hypothesis:
Watch durations of
'Education' videos are
significantly higher in
2022.



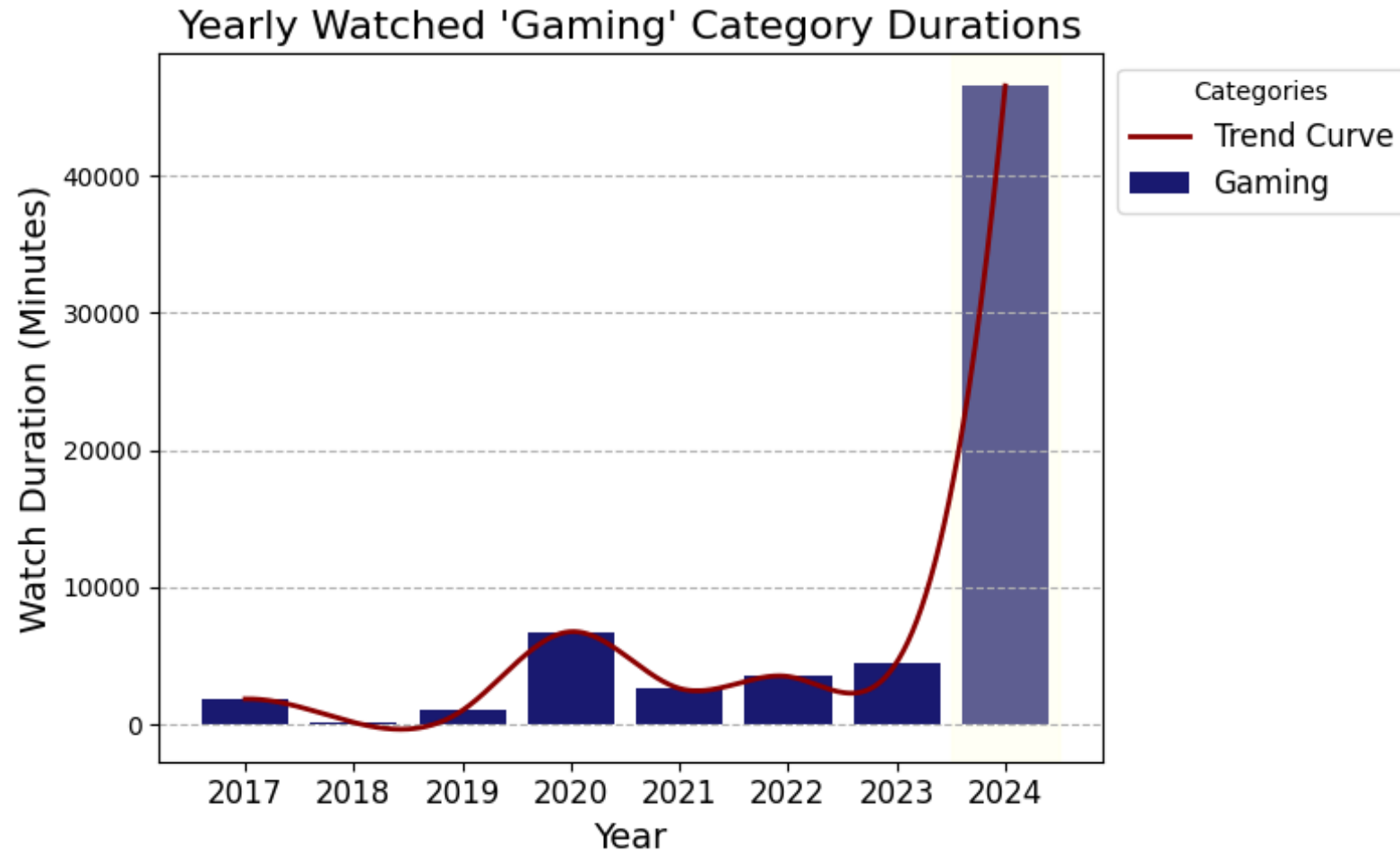
Testing if "Gaming" category has been watched significantly more than other years in 2024.

Null Hypothesis: No significant change in watch duration of "Gaming" category in 2024.

Alternative Hypothesis: There is a significant increase in watch duration of "Gaming" category in 2024.

P-Value: 0.0010 Reject the null hypothesis: Watch durations of 'Gaming' videos are significantly higher in 2024.

Hypothesis proven right.



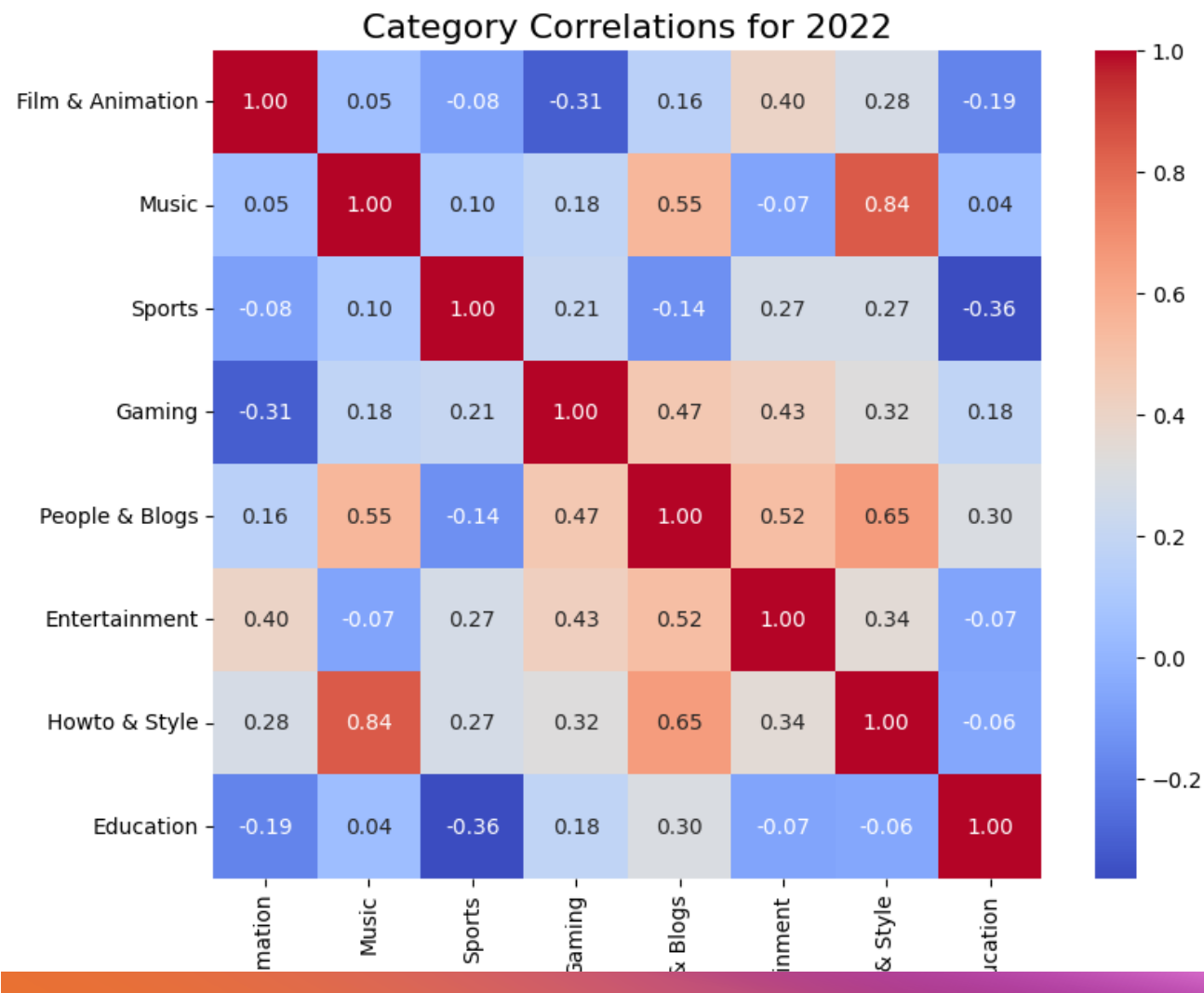
Then I analyzed all year category combinations if there was any significant change in one category in a year:

Output:

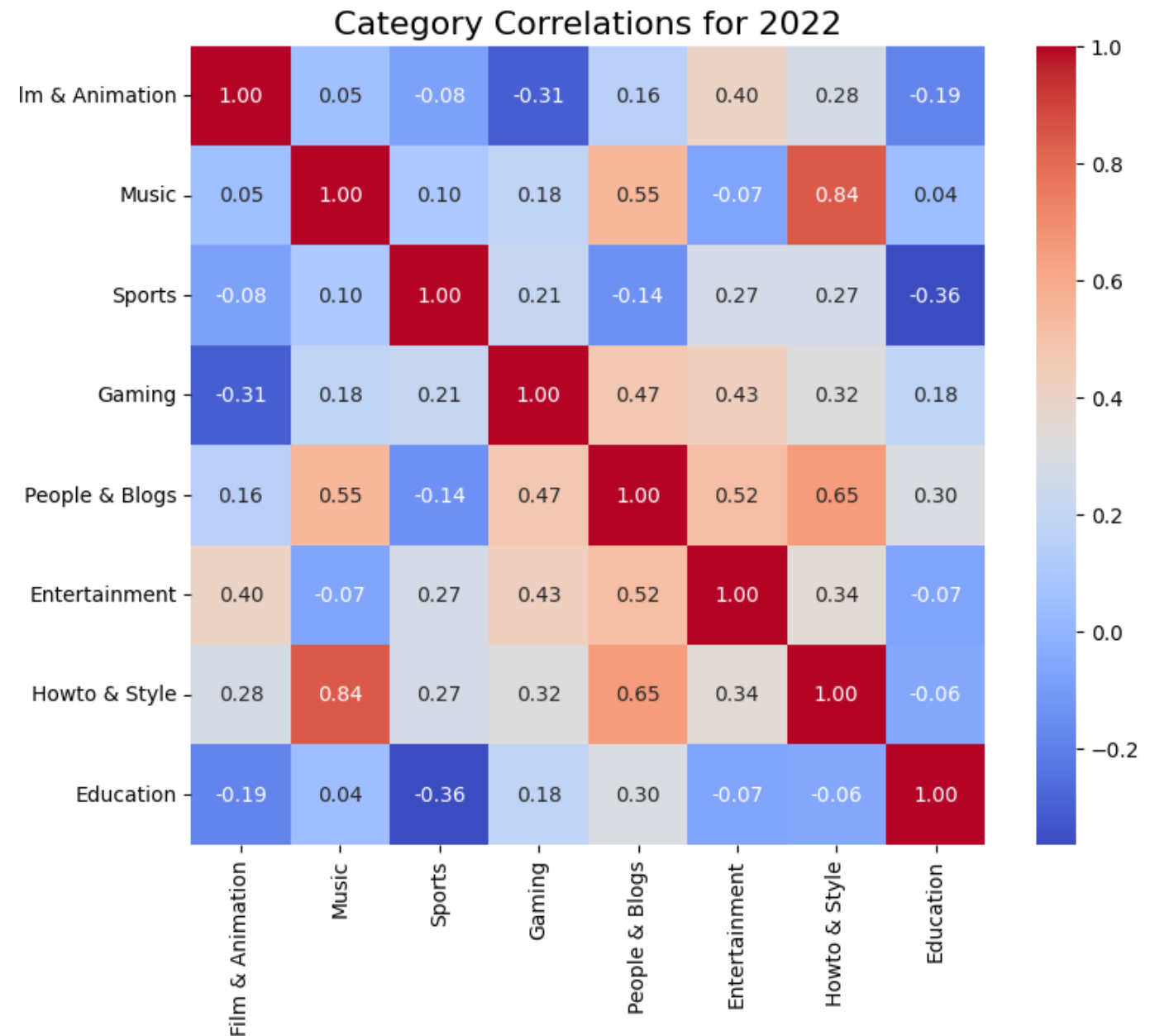
Watch durations of Howto & Style videos are significantly higher in 2017.
Watch durations of Music videos are significantly higher in 2018.
Watch durations of People & Blogs videos are significantly higher in 2019.
Watch durations of Entertainment videos are significantly higher in 2019.
Watch durations of Film & Animation videos are significantly higher in 2022.
Watch durations of Education videos are significantly higher in 2022.
Watch durations of Sports videos are significantly higher in 2024.
Watch durations of Gaming videos are significantly higher in 2024.

Analyzing Correlations Between Video Categories using Correlation Matrix

- Analyzing yearly data of categories' monthly watched times and creating correlation matrices

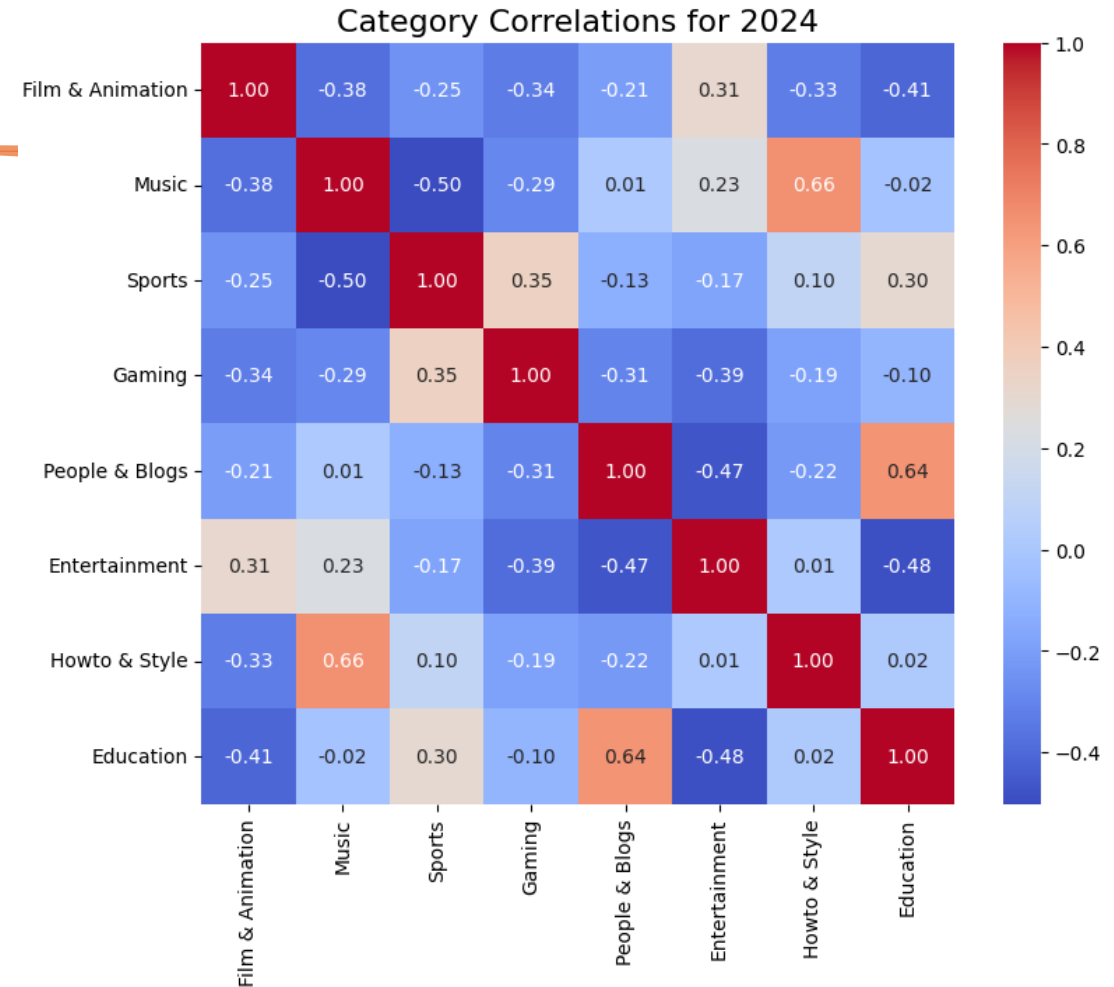


- The categories generally do not correlate strongly with each other. But the strongest correlation is between categories "*Music*" and "*Howto & Style*" with correlation **0.84** which is a *positive correlation*.
- The reason behind this correlation might be because Howto & Style videos that i watch are mostly tutorial videos. I watch a lot of music tutorials (e.g. guitar tabs, piano notes, drum covers etc.)
- Moreover, Howto & Style and People & Blogs categories seem to be in a moderate positive correlation of 0.65.

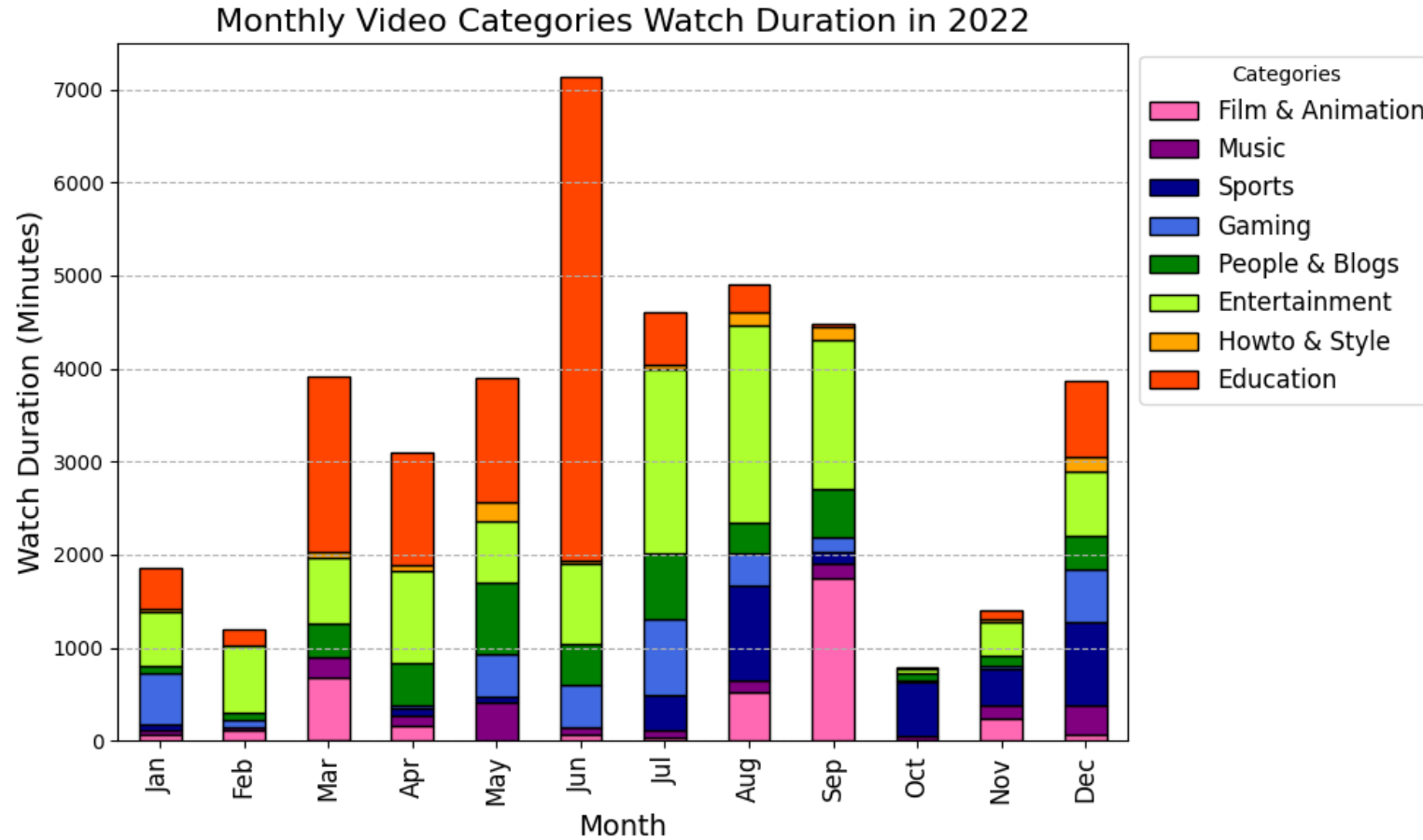


Year "2024" Analysis

- There appear to be more ***negative correlations*** than positive ones. It seems like I don't diversify my watching habits much, as I tend to focus on one category at a time. However, since these correlations are weak or moderate, we cannot draw any definitive conclusions from them.
- **Music & Sports , Education & Entertainment** video categories are the ones with higher negative correlations which can be predictable as I won't be watching much entertainment videos when for example I'm studying for exams or music and sports are whole different areas to be watching in the same time period.

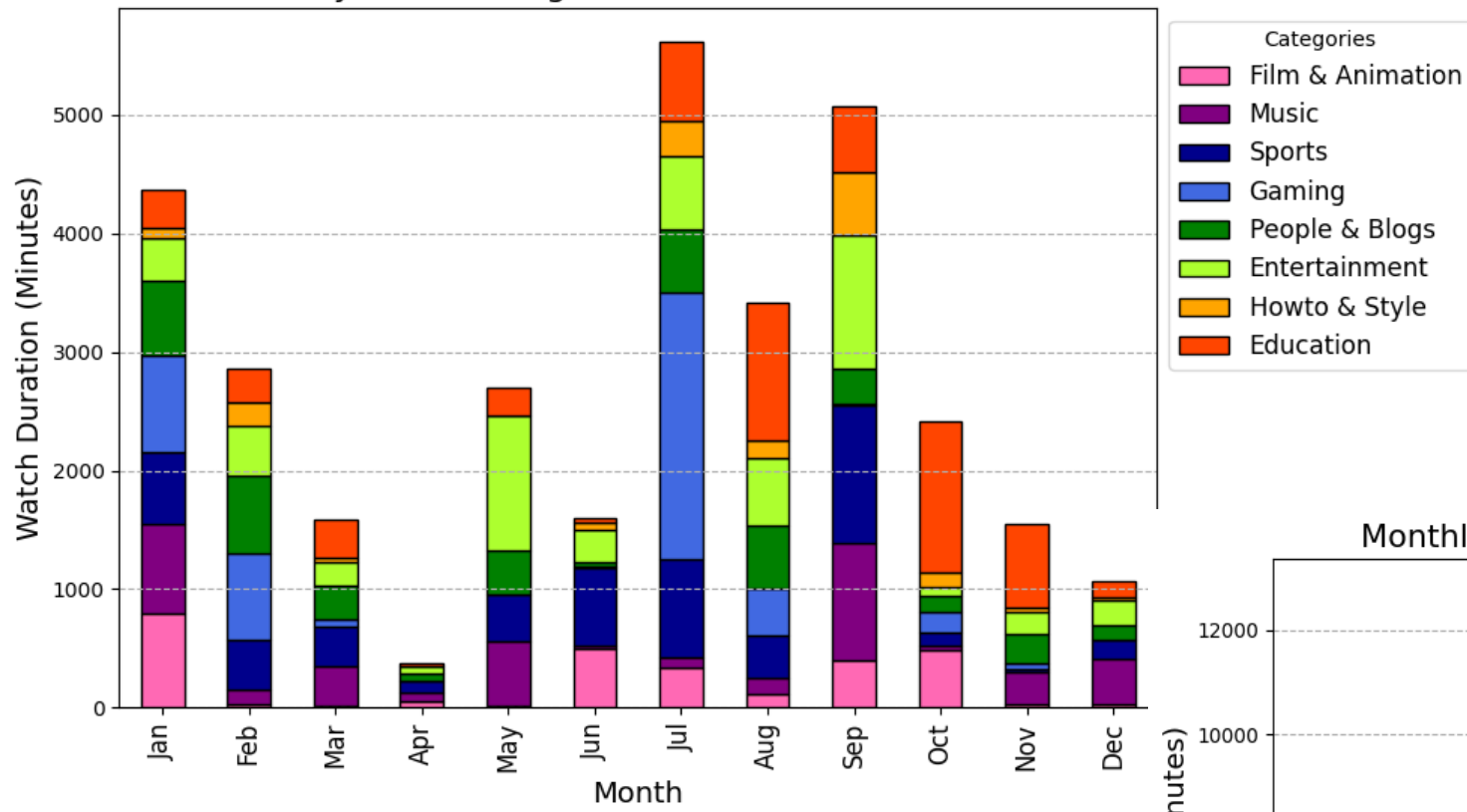


Visualizing Monthly Categories of the Years



As previously analyzed, "2022" stands out as the year when the **"Education" category was watched significantly more than in other years**, likely due to it being my *University Entry Exam year*. The plot clearly shows that this category not only differs from others but also dominates during specific months such as **March, April, May, and June**.

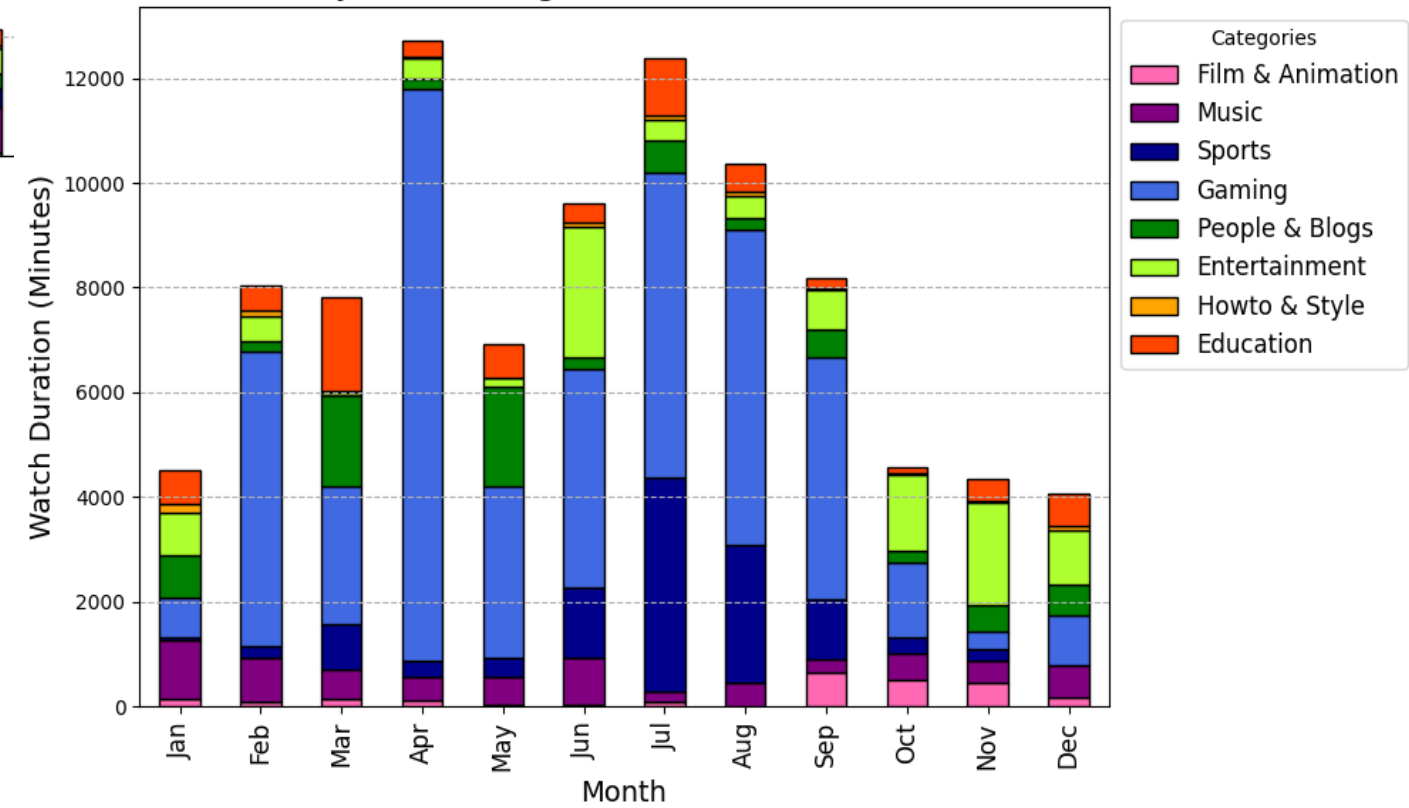
Monthly Video Categories Watch Duration in 2023



YEAR 2023

YEAR 2024

Monthly Video Categories Watch Duration in 2024



Compared Analysis of "2023" and "2024"



1. Comparing year "2023" and "2024" we can see on the y axis the watch duration levels are approximately 2 times of the the year "2023" in "2024". (2024 watch duration > 2023 watch duration x2)



2. In year "2023" we see diversification of categories along the year but in year "2024" gaming category dominates other video categories.



3. In year "2023" the educational videos' proportion is higher than "2024". Showing my reason of using Youtube evolved more into entertainment purpose.

Summary of Findings:

The total watched videos vary in terms of quantity and duration.


Most watched video categories by quantity (number of times):

- People & Blogs - 24.6%
- Entertainment - 22.4%
- Music - 19.6%





Most watched video categories by duration:

- Gaming - 22.3%
- Entertainment - 18.4%
- Education - 13.6%

Completely
different results!

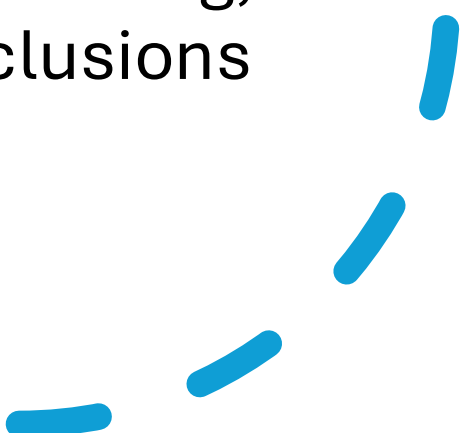
- 
- In my all-time watched categories, the "**Gaming**" category has the longest average video duration at **42** minutes. It is followed by "**Sports**" with an average of **20** minutes and "**Education**" with an average of **18** minutes.

By hypothesis testing I found out:

1. "My watch time has increased significantly over the years."
 2. "My watch durations are not significantly higher in summer."
 3. "I watched significantly more educational videos in 2022 while preparing for my university entry exam."
 4. "I watched gaming videos much more than ever in 2024."
- 
- 
- 
- 



Correlation Matrix Results:

- In year 2022, the categories generally do not correlate well with each other. But there is a strong correlation between categories "*Music*" and "*Howto & Style*" with **positive correlation of 0.84**.
 - In 2024, there appear to be more ***negative correlations*** than positive ones. However, since these correlations are not strong, we cannot draw any definitive conclusions from them.
- 

Limitations:

There were significant shortcomings in the YouTube data.

1. YouTube's inability to categorize videos with the right accuracy.

-Upon examining my raw data and the metadata of the videos, I noticed that the categorization was not entirely accurate. For instance, a concert video of a rock band might be categorized under *"Entertainment" rather than "Music."* Similarly, a guitar cover of a song is often categorized under *"People & Blogs" or "How-to & Style"* instead of *"Music"*.


However the data is huge thus only thing I could do was to accept the results of the Youtube API data.

- ***2. The video durations retrieved through the YouTube API reflect the total length of the original videos, not the actual time spent watching them.***
- -The issue arising from this limitation is that even if a user watches only 1 minute of an hour-long video, the data records the entire hour. Thus, **I assumed for this project that I watched a significant proportion of each video I have watched.**




Future Work:

Data can reveal an infinite amount of information about an individual. The possibilities for further analysis and hypothesis testing on this YouTube data are limitless.



At this point, it is difficult to stop discovering, as numerous scenarios and hypotheses come to mind for testing the data. In conclusion this project can evolve into different aspects of youtube usage and show many more interesting results in the future.





In this project, I used "All Time," "Yearly," and "Monthly" representations of my data due to its large size, spanning 7 years, which I wanted to explore comprehensively.



In the future, I could expand the analysis to include more granular forms, such as daily or hourly watch times.



With more detailed data, I would be able to conduct deeper analysis and draw more specific conclusions about my viewing behavior over time.

Thank you for your time!
