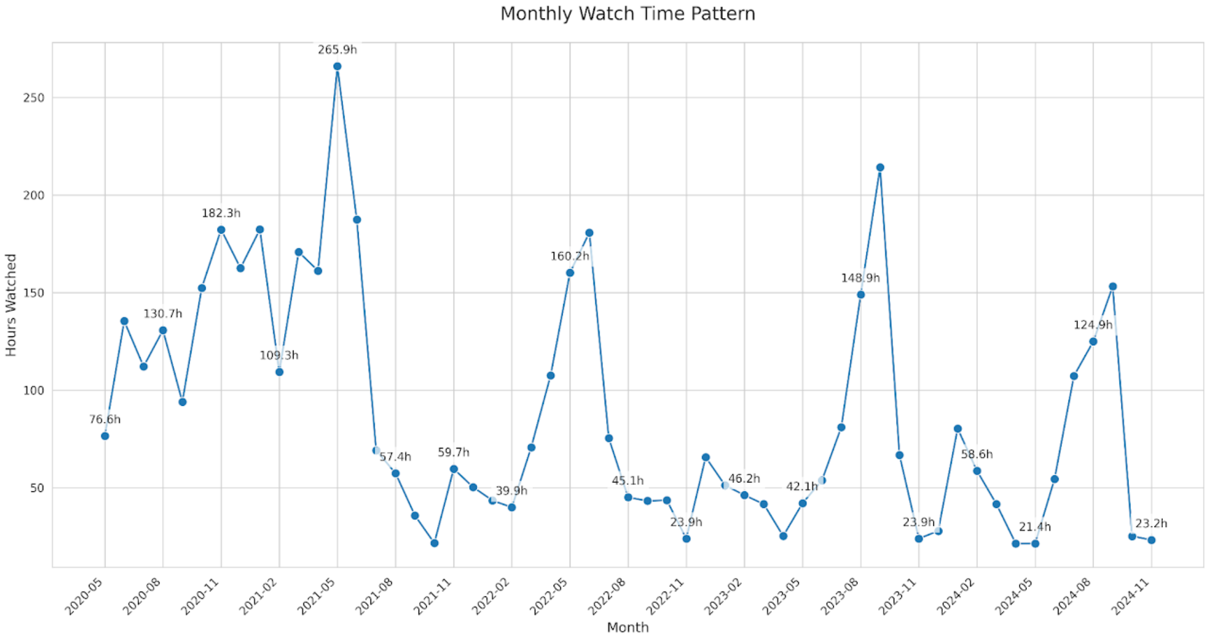
Enes Kafa - YouTube Usage Analysis

About the Project:

Fall 2024-2025 DSA210 Introduction to Data Science Term Project: Analyzing my watching behavior from various aspects.

Temporal Analysis

When I started the project, understanding how much time I spent on YouTube was an important goal for me. This way, I could learn how much time I had spent on this social media platform in the past and find out how much of my day YouTube had taken up. Therefore, starting the project by conducting a temporal analysis of my YouTube viewing seemed like a good idea. And I reached some interesting findings.



My actively used YouTube account contained watch history dating back to 2020, so I had 4.5 years of data at my disposal. Initially, before creating this visualization, I assumed that the time I spent on YouTube didn’t fluctuate much, but this assumption turned out to be incorrect. The chart on the left shows how much time I spent watching videos on YouTube on a month-by-month basis since 2020.

I used hypothesis testing to analyze this finding in a more tangible way. For this, using the Kruskal-Wallis test was a good idea because my data varied month by month and had a non-normal distribution. Additionally, the Kruskal-Wallis test handles outliers better and does not make any assumptions about the distribution of the data. I formulated the necessary hypotheses for the test:

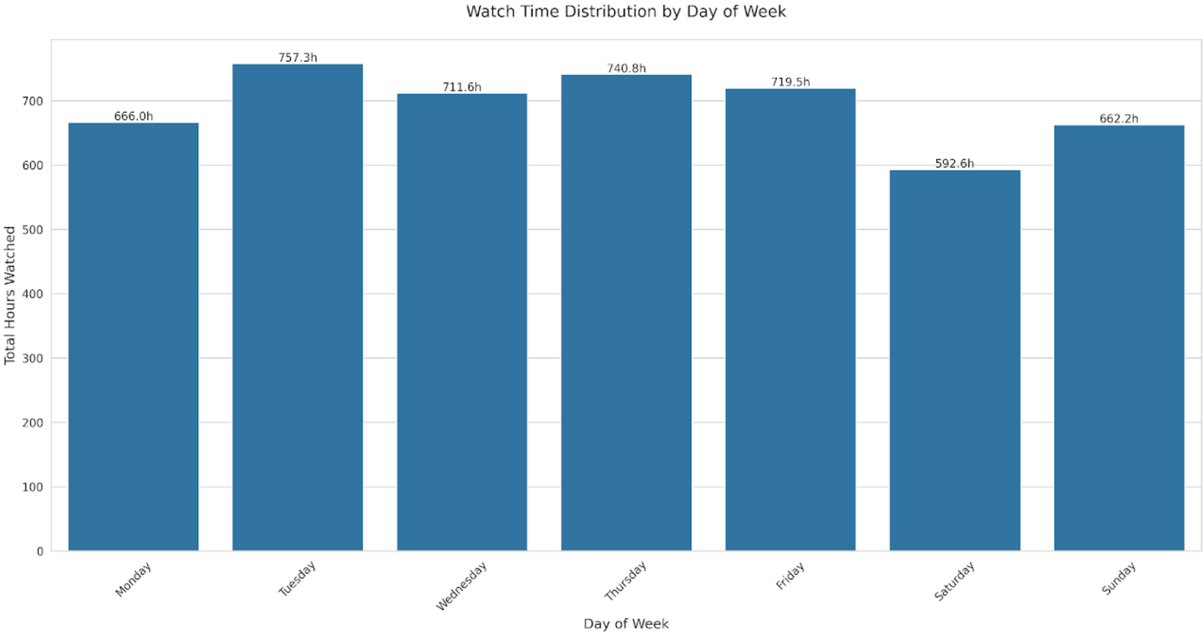
- **Null Hypothesis**: The distributions of video-watching durations are the same across all months.
- **Alternative Hypothesis**: The distributions of video-watching durations differ across months.

I set my significance level at 0.05 and applied the test. The results were as follows:

- **H-statistic**: 2253.8504
- **P-value**: $p < 0.001$

As expected, p-value was extremely low, so I rejected the null hypothesis and concluded that there are significant differences in video-watching durations across months.

In the next step, I aimed to analyze the distribution of my viewing times across the days of the week. Initially, I predicted that my viewing times would be concentrated on weekends because, since 2020, I have been attending school and university on weekdays, and have more free time on the weekends.



In the resulting chart, I noticed that my viewing times were evenly distributed across the days of the week. To deepen my analysis, I formulated the following hypotheses:

- **Null Hypothesis**: There is no difference in daily watch time between weekdays and weekends.
- **Alternative Hypothesis**: There is a difference in daily watch time between weekdays and weekends.

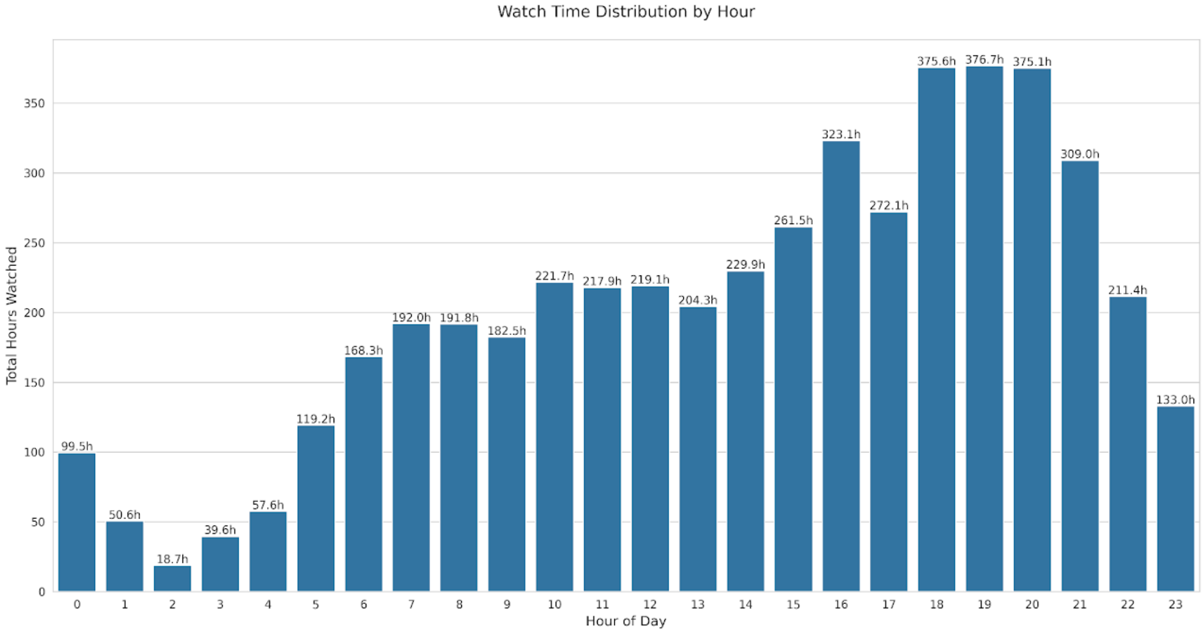
Using a t-test for this analysis was a logical choice because it is appropriate for comparing means between two independent groups (weekday vs. weekend viewing times).

I set my significance level at 0.05 again and tested the hypothesis. The results were as follows:

- **T-statistic**: 1.8306
- **P-value**: 0.0673

Since the p-value was greater than 0.05, I **failed to reject** the null hypothesis, meaning there is no significant difference in watch time between weekdays and weekends.

So far, I have conducted analyses on monthly and weekly periods. Now it's time to narrow down the period and examine how much I watch during different hours of the day.



The results came out as shown in the chart above. It was no surprise to me that my viewing times peaked around 6-7-8 PM, as one of my favorite activities after dinner has always been lying in bed and watching videos on YouTube. What did surprise me, however, was the considerable amount of viewing time in the very early morning hours (4-5-6-7 AM). For this data, I would assume that on some days, I fell asleep while watching videos, leaving YouTube to play videos on its own. Alternatively, on some days when I stayed up all night, I might have chosen to watch videos in the morning. However, I can't say for certain.

To take the analysis a step further, I decided to use the Kruskal-Wallis test again to evaluate the watch time across different periods of the day. However, while doing this, I excluded some data from the night hours (00:00–05:59), as mentioned earlier, since the data concentrated in these hours could be misleading. I formed the following hypotheses:

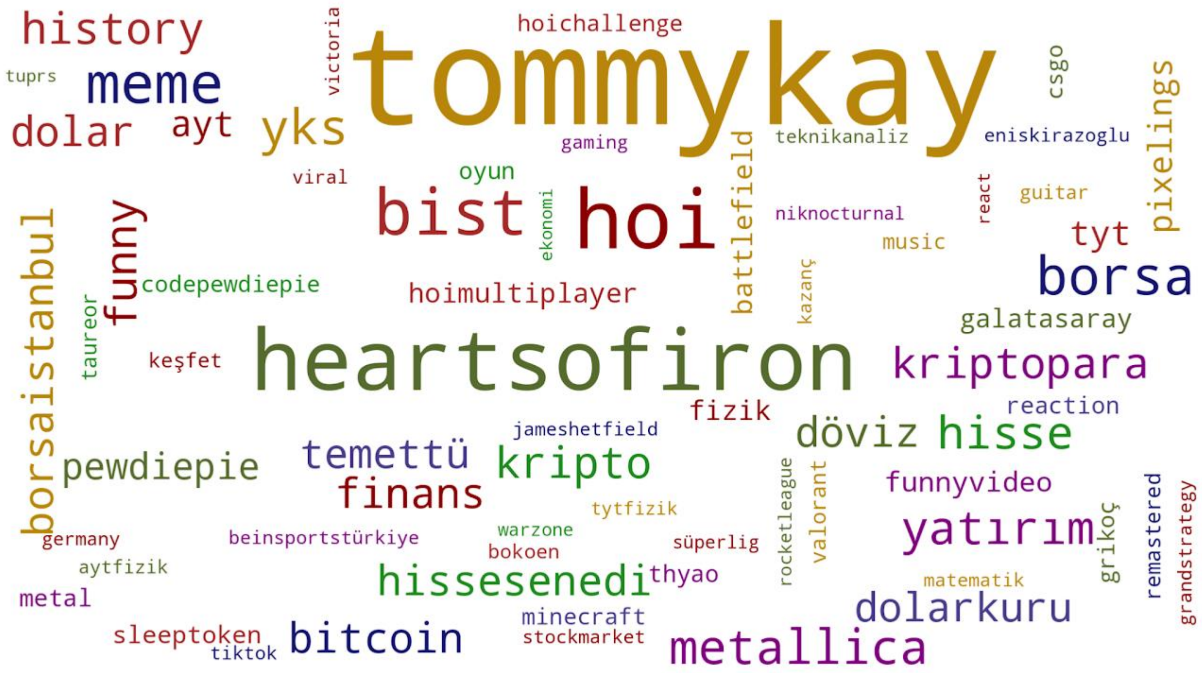
- **Null Hypothesis**: The distributions of watch time are the same across morning (06:00–11:59), afternoon (12:00–17:59), and evening (18:00–23:59) periods.
- **Alternative Hypothesis**: At least one time period has a different distribution of watch time.

I set the significance level at 0.05, and got the following results:

- **H-statistic:** 69.0793
- **P-value:** $p < 0.001$

With these results, i concluded that there are significant differences in viewing duration distributions across time periods.

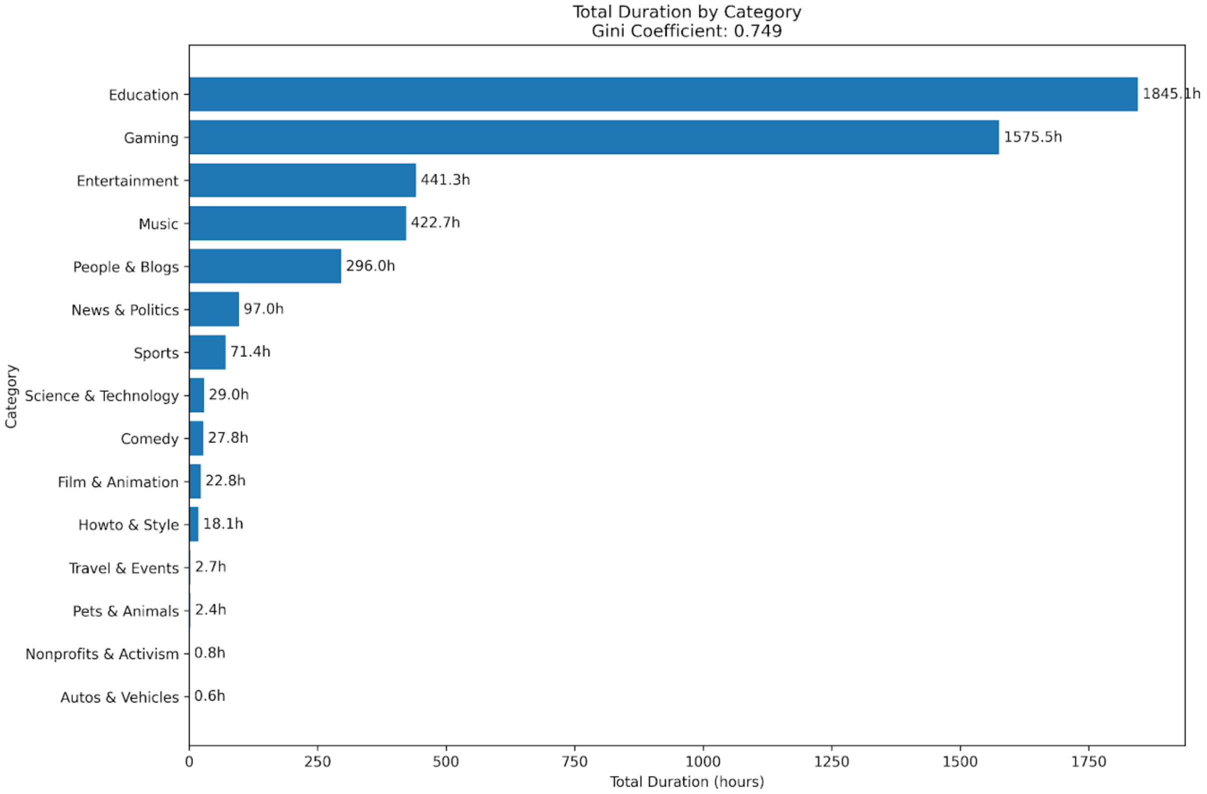
Categorical Analysis



In addition to conducting a temporal analysis of my YouTube watch history, it was also important for me to analyze how my viewing preferences have changed since 2020. The chart above shows the most frequently used hashtags in the videos I watched. This word cloud provides a general overview of what I enjoy watching.

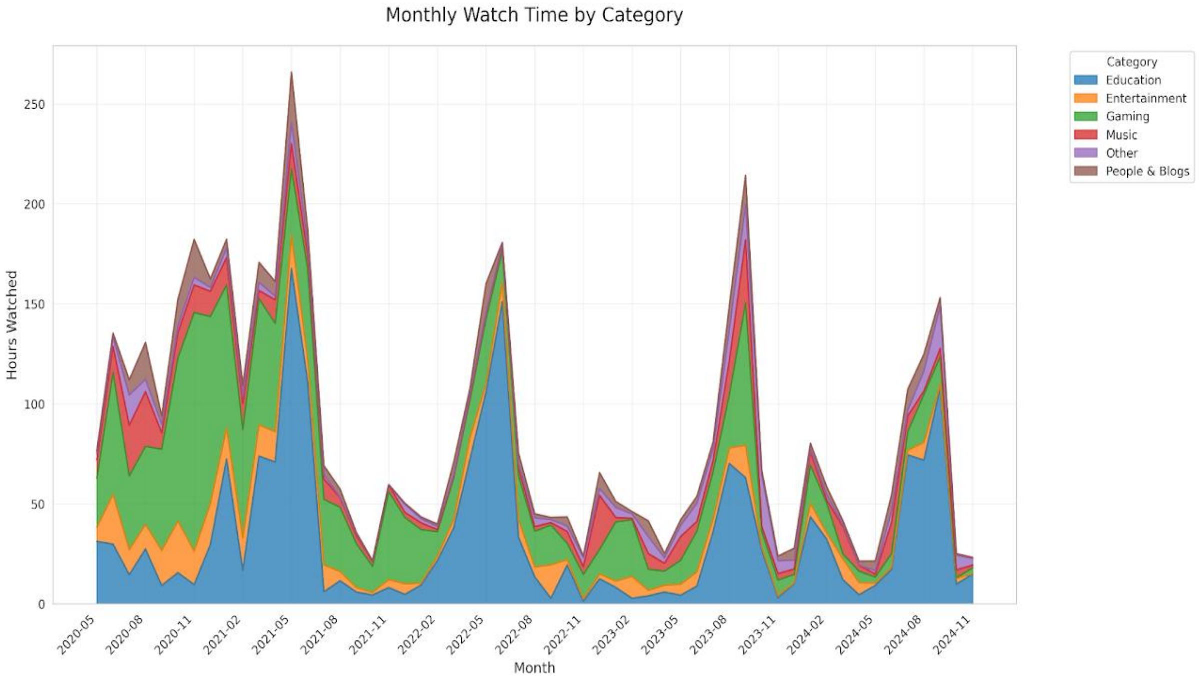
For instance, I used to love playing and watching content related to the game *Hearts of Iron IV*, so terms like *TommyKay* (a content creator for that game) and *Hearts of Iron (hoi)* appear in this visualization. In addition, there are some hashtags related to the financial markets, which I have been actively involved with for the past two years. Watching live concert recordings of Metallica was also a favorite activity of mine. Moreover, I noticed some keywords related to the Turkish university entrance exam (*YKS*).

The goal of this analysis is to understand the relationship between the words I see here and the categories of videos I frequently watched since 2020. Let's get started.



The chart above shows the distribution of the time I spent on YouTube by category. This chart has a highly skewed and uneven distribution, with the *Education* and *Gaming* categories, which have the highest viewing hours, significantly outpacing the other categories. To statistically validate this observation, I calculated the Gini coefficient and found it to be 0.749.

A value of 0.00 represents perfect equality, while 1.00 represents perfect inequality. A value of 0.76 indicates that a few categories dominate the others, with many categories having low viewing durations, highlighting the imbalance across categories.

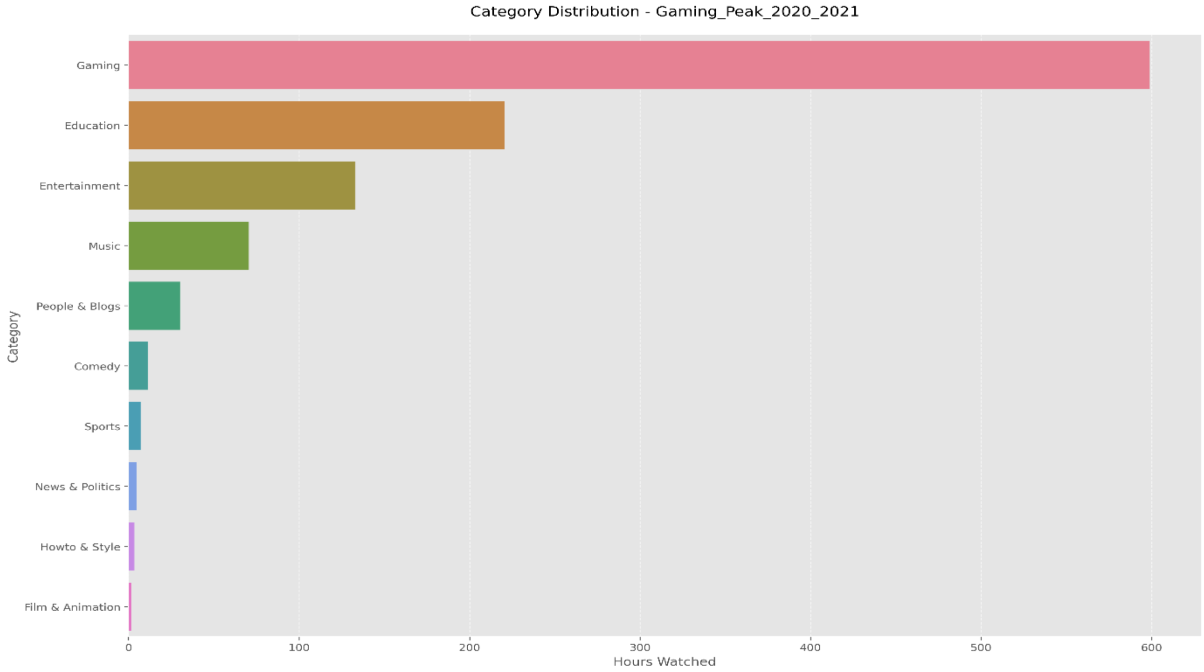


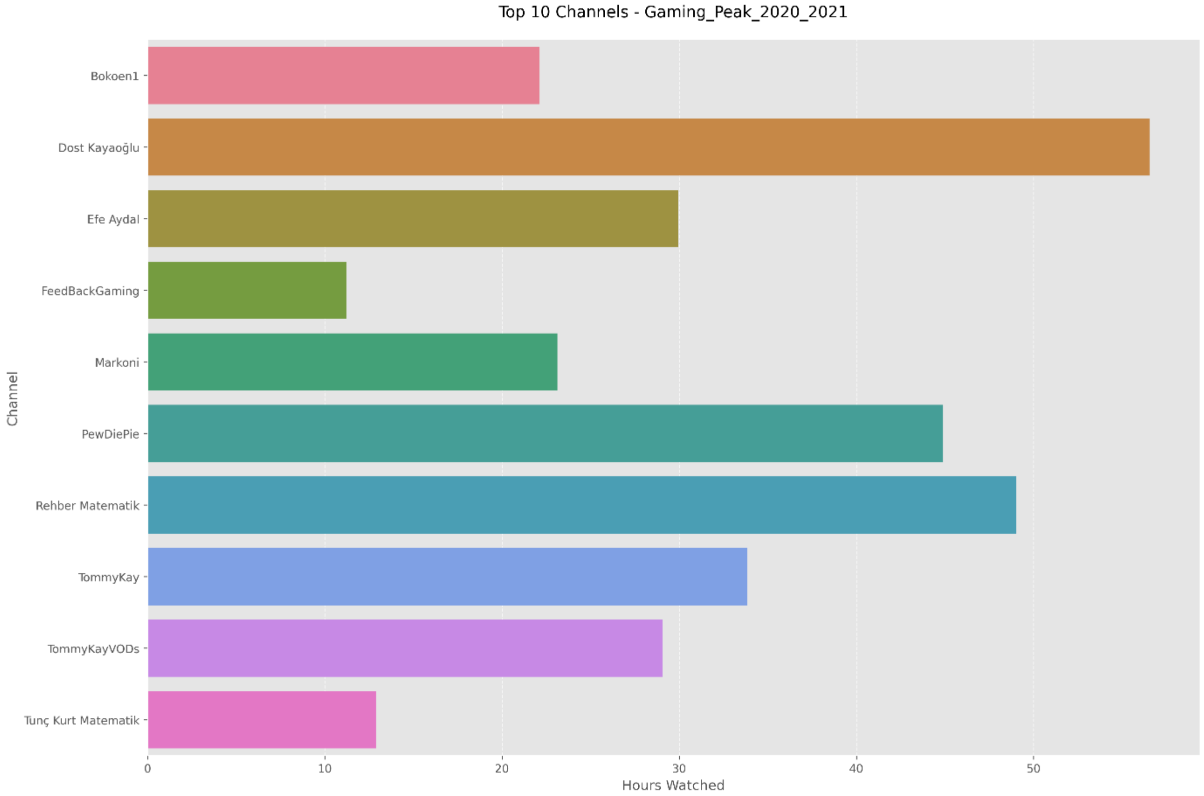
The chart above shows the distribution of my viewing times by category from 2020 to 2024. A few points caught my attention in this chart:

1. I noticed that the *Gaming* category had a strong presence between 2020 and 2021. I attribute this to the global pandemic during those years. Since we were forced to stay home, playing games and watching gaming videos became a good option.
2. I observed that my chart peaked in early 2021, mid-2022, late 2023, and late 2024.

Then, I decided to explore which channels I watched the most during these periods and analyze the viewing preferences that caused these peaks.

2020-05 to 2021-02 Period





The image carousel above shows the distribution of my viewing time during the pandemic period by category and channel. As mentioned earlier, the *Gaming* category dominates the others, with some viewership also observed in the *Education* category. This can be attributed to my preparation for the YKS (university entrance exam) during that time, as YKS-focused channels like *Rehber Matematik* and *Tunç Kurt Matematik* were among the top channels I watched.

For this period, I used the Mann-Whitney U test and formulated the following hypotheses:

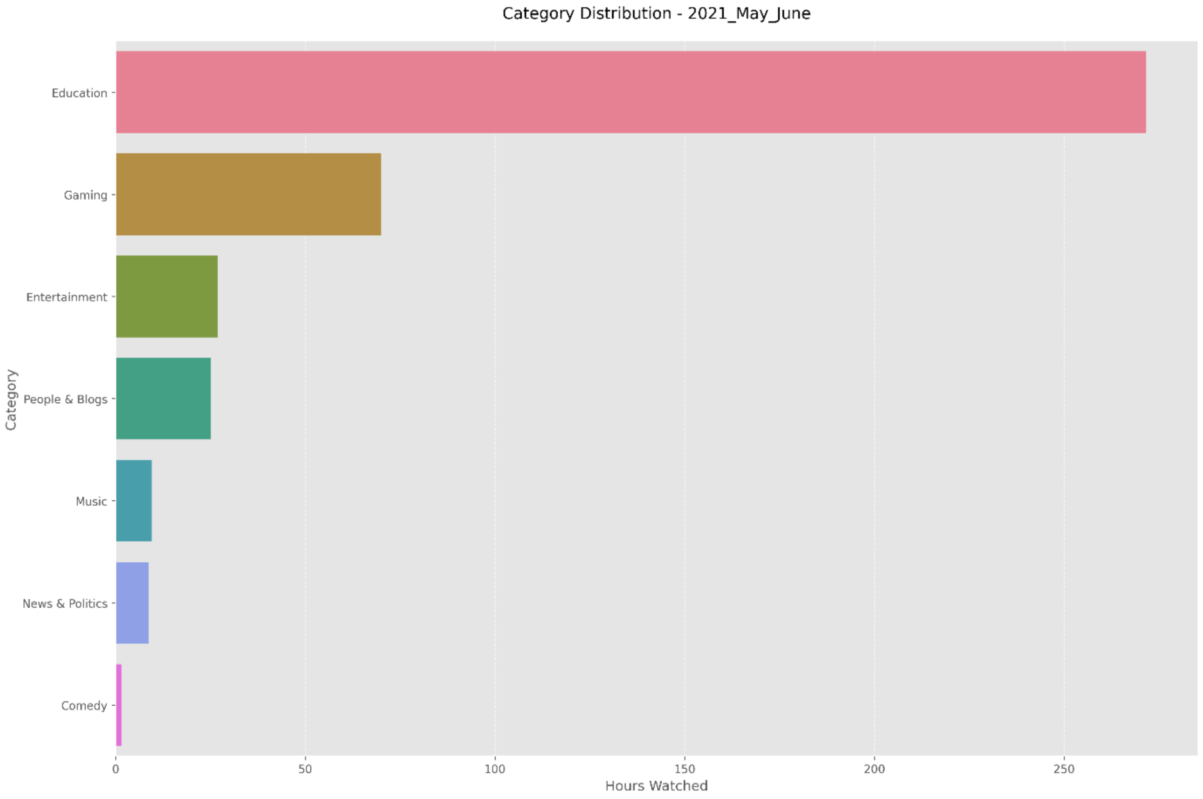
- **Null Hypothesis (H_0)**: No difference in watch time between gaming and non-gaming content
- **Alternative Hypothesis (H_1)**: Gaming content has higher watch time than non-gaming content

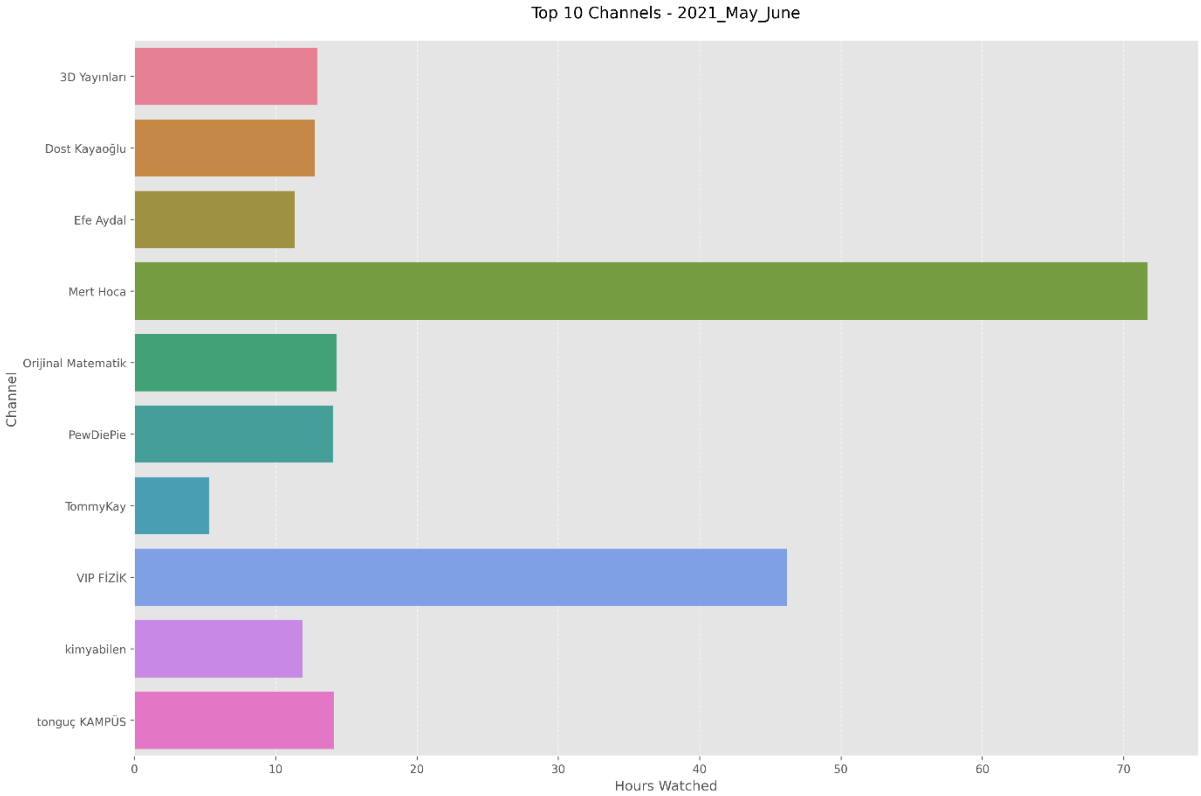
I set my significance level at 0.05, and the results were as follows:

- **U-statistic**: 4,281,761.5
- **p-value**: < 0.00001

The p-value was extremely low, so I rejected the null hypothesis and concluded that gaming content has higher watch time than non-gaming content.

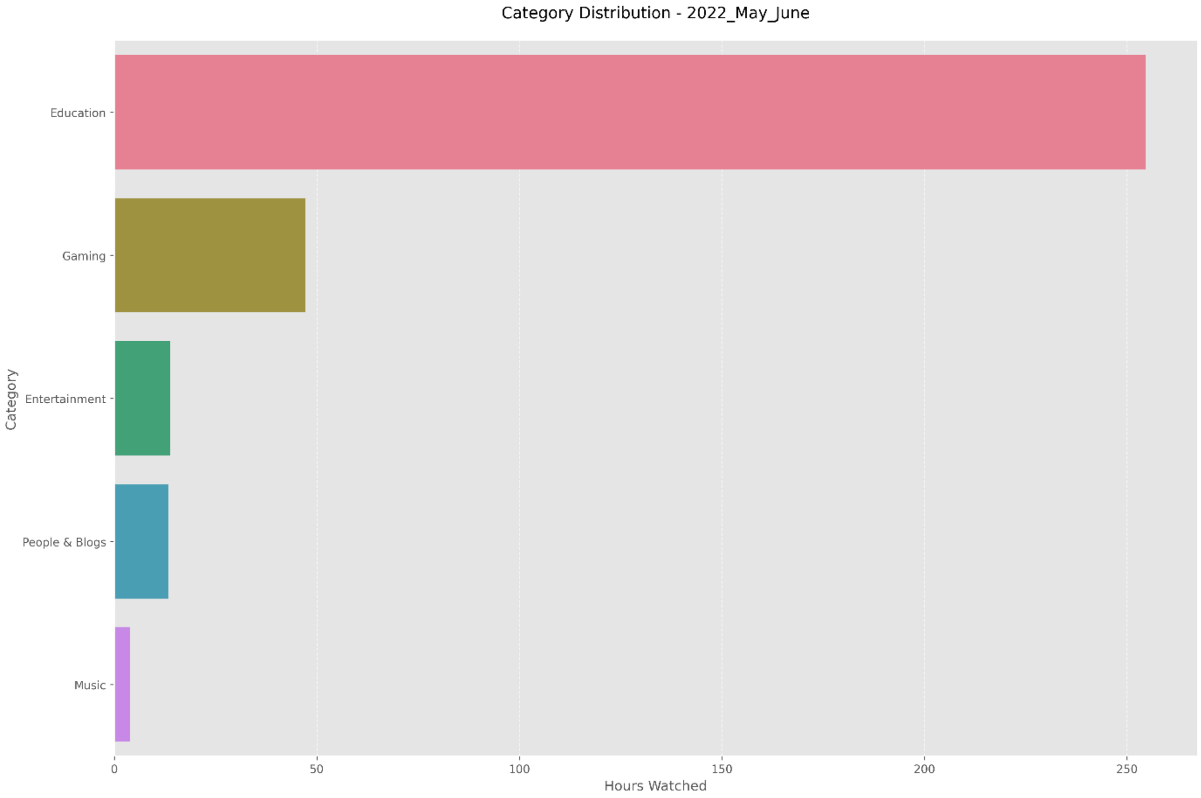
2021 May - June Period

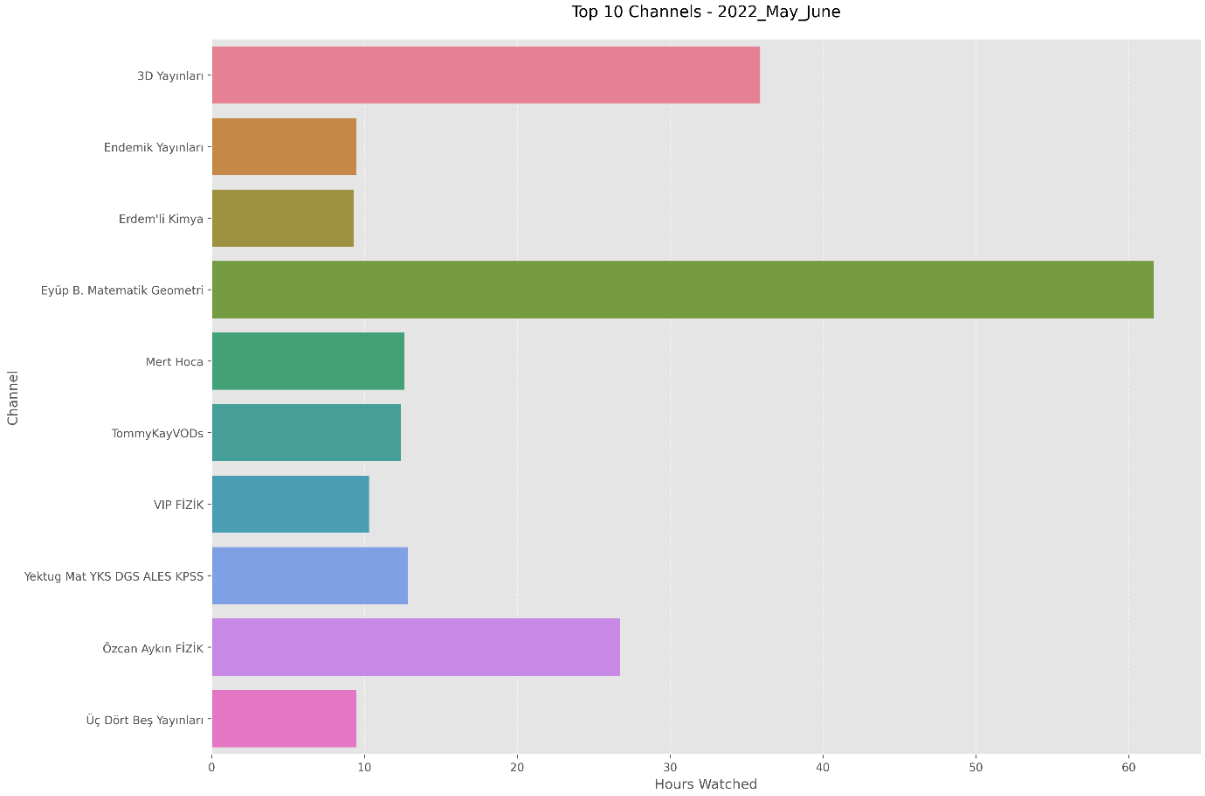




When we move to the May-June 2021 period, we see that the *Education* category takes the lead. This makes sense because in May - June 2021, I was preparing for my first university entrance exam. As a result, education-related content made up the majority of my YouTube viewing. This is further supported by the fact that 6 out of the top 10 channels I watched during this period were YKS-focused channels, with *Mert Hoca* leading the list.

2022 May - June Period





In the May-June 2022 period, a distribution very similar to the May-June 2021 period can be observed. The reason for this is that I was dissatisfied with the results of the exam I took in my first year and decided to study for the YKS for another year. This time, instead of *Mert Hoca*, the channel *Eyüp B. Matematik Geometri* took the lead. I attribute this to my perception that *Eyüp B.* produced better content.

After observing the distributions of these two periods, I asked myself the following question: *"Were the proportions of videos I watched in the Education category the same in 2021 and 2022?"* This way, I could determine in which year I focused more on the Education category. To answer this, I decided to use the two-proportion Z-test and formulated the following hypotheses:

- **Null Hypothesis (H_0)**: The proportion of educational content views remained the same.
- **Alternative Hypothesis (H_1)**: There was a significant change in the proportion of educational content views.

I set my significance level at 0.05, and the results were as follows:

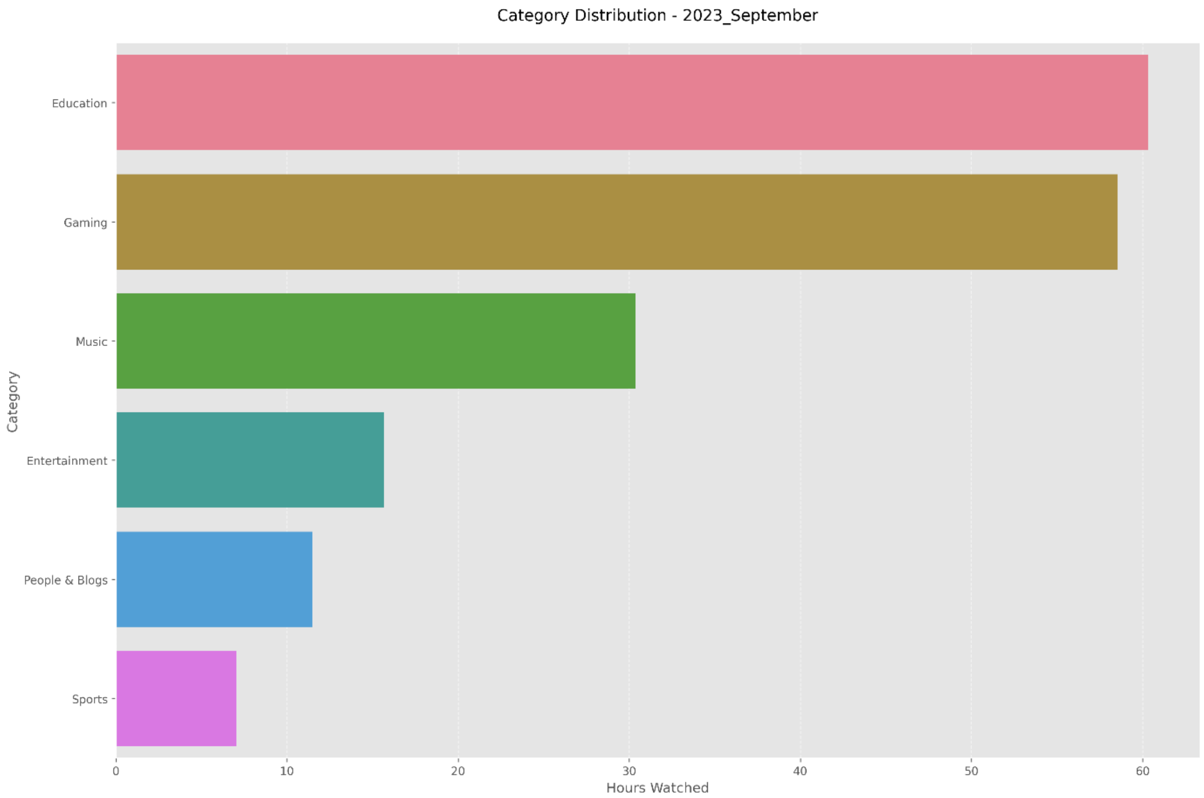
- **2021, Education Proportion**: 28.0%
- **2022, Education Proportion**: 46.4%

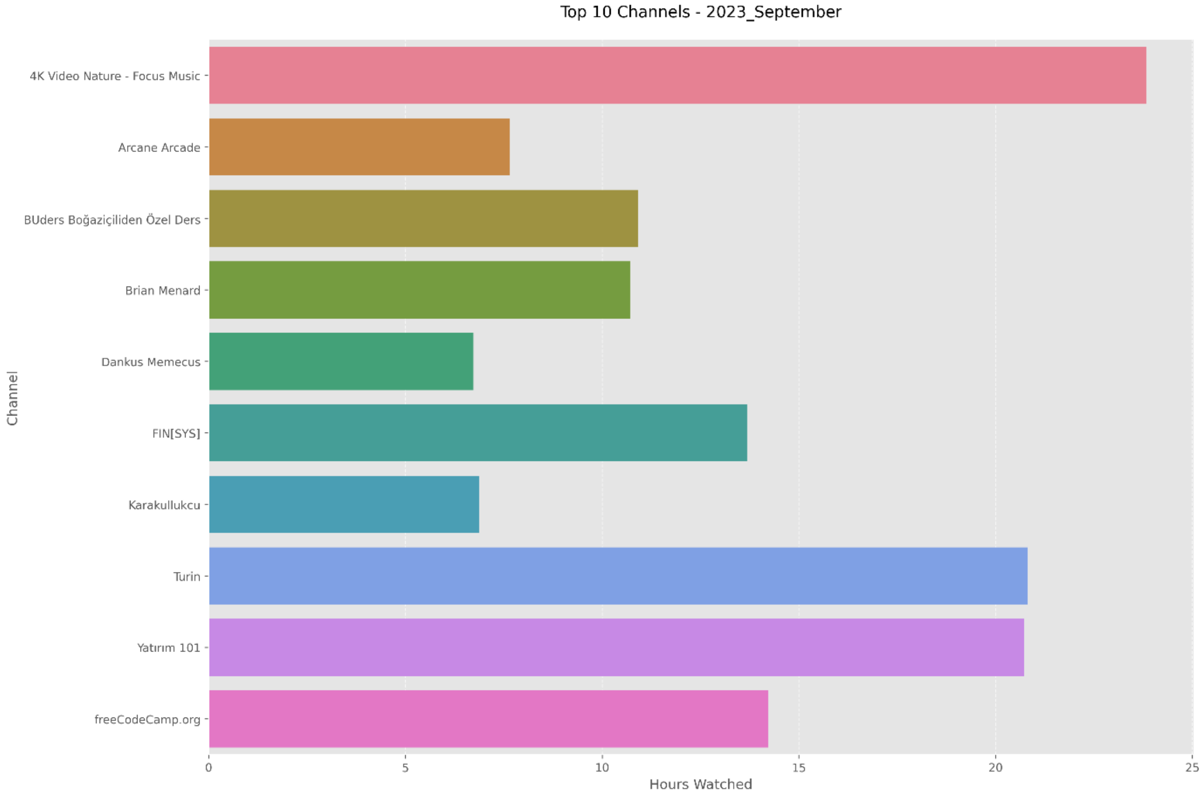
Z-Test Results:

- z-statistic: 9.5178
- p-value: $p < 0.000001$

Since the p-value was extremely low, I rejected the null hypothesis and concluded that there is strong statistical evidence of a change in the proportion of educational content views between 2021 and 2022.

2023 September Period





This period shows a more balanced distribution compared to previous ones, despite the dominance of the Education and Gaming categories. During this time, I had started watching stock market-related educational videos (e.g., FIN SYS, Yatırım 101) because the stock market was in the midst of a rally at the time, and I wanted to learn how to navigate it.

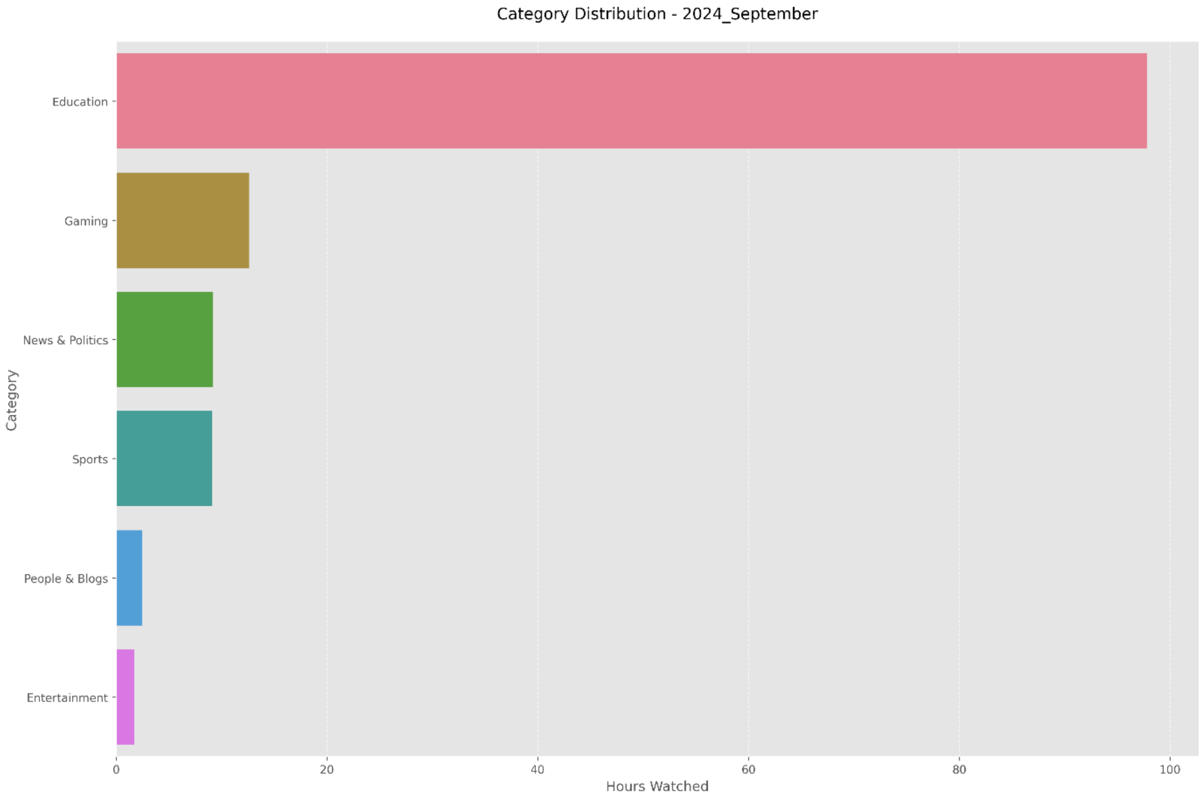
However, in addition to this, something else caught my attention, particularly in the graph showing my most-watched channels. Some channels (Arcane Arcade, Brian Menard, Turin) felt very unfamiliar to me. When reviewing my watch history, I realized that I had either watched or partially watched some videos from these channels. This might have been due to accidentally clicking on these videos or leaving YouTube's autoplay feature enabled. Another limitation of the YouTube Data API V3 became apparent here: even if I didn't watch some videos in their entirety, they were still recorded in my watch history. However, I couldn't access information on how much of the video I had actually watched, which could lead to inaccuracies in my data analyses in certain cases.

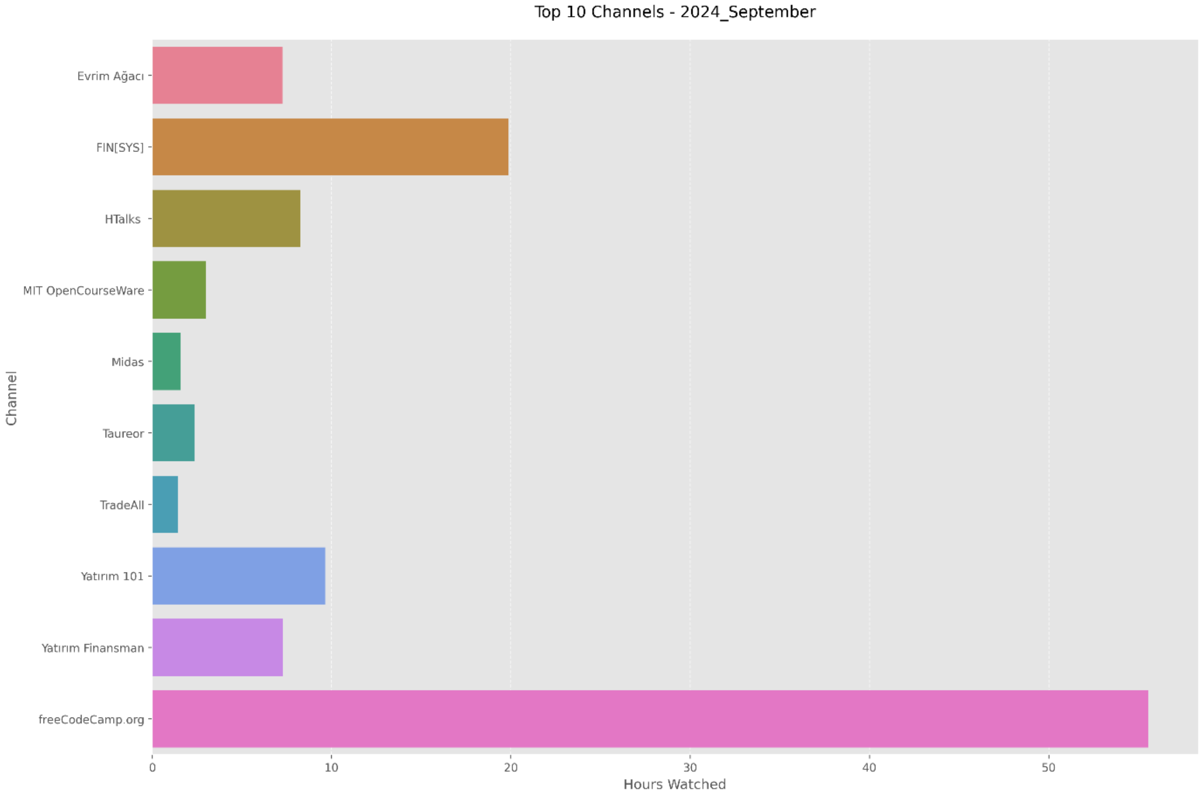
For this reason, I wanted to conduct a separate analysis here. I manually gathered the XU100 index values for June, July, August, and September and compared them with the time I spent watching stock market-related videos during the same months. I decided to use **Pearson's Correlation Coefficient** to calculate the correlation between these two datasets. The data for these months and the resulting outcome were as follows:

```
Monthly Statistics:

==================================================

2023-06:

Viewing Duration: 1.40 hours

XU100 Value: 5000

2023-07:

Viewing Duration: 30.43 hours

XU100 Value: 5500

2023-08:

Viewing Duration: 47.27 hours

XU100 Value: 6500

2023-09:

Viewing Duration: 36.25 hours

XU100 Value: 7000

Correlation Analysis:

==================================================

Pearson Correlation Coefficient: 0.8069
```

With this result, I noticed how the market rally aligned with the changes in my viewing preferences, as a Pearson coefficient of 0.8 indicated a strong correlation.

2024 September Period





In this period, you can once again see that the Education category stands out significantly compared to other categories. This time, the context of the videos in the Education category shifts, with a focus on programming, as can be inferred from the significant lead of the "freeCodeCamp.org" channel over the others.

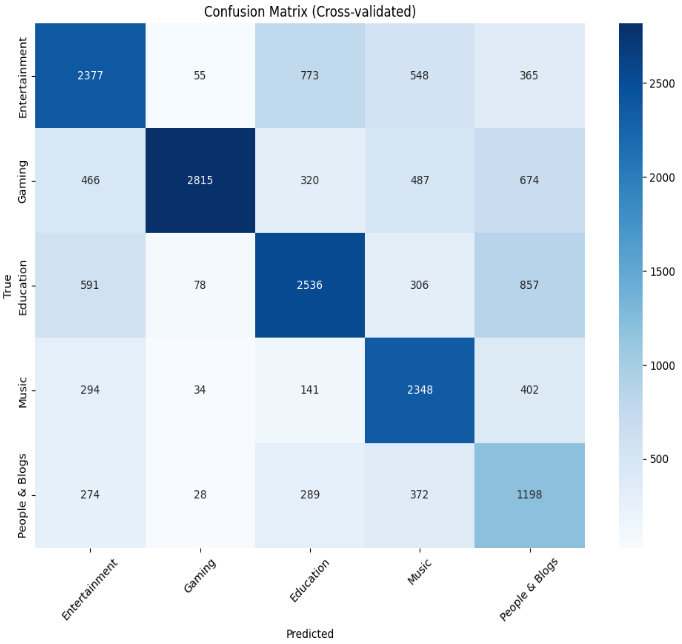
Finance-related channels (i.e., Yatırım Finansman, Yatırım 101, TradeAll, Midas, FIN[SYS]) have also found a significant place for themselves during this period.

Machine Learning Experiment

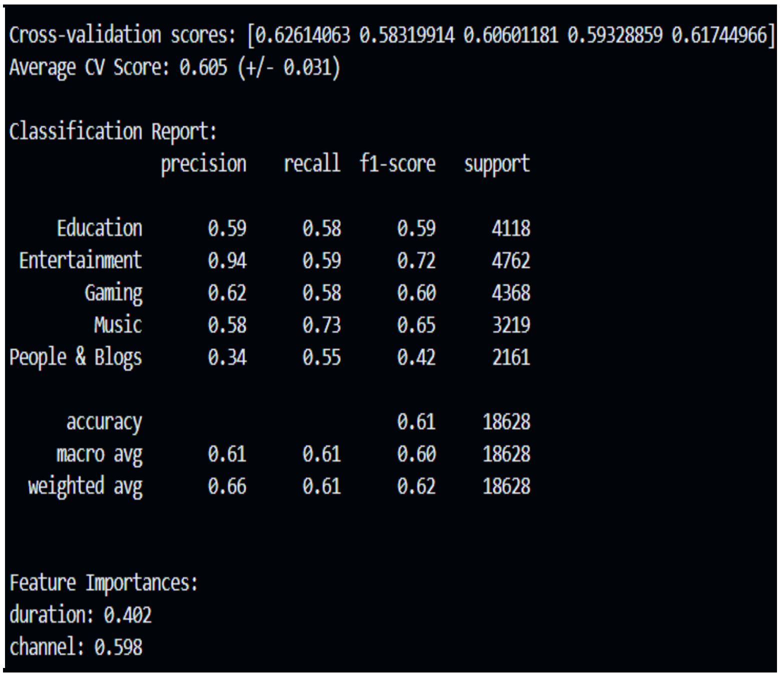
In this part of my project, I wanted to try something experimental using a machine learning method, and what I aimed to do was a classification task. For this, I decided to use a Decision Tree because decision trees can handle both numerical and categorical features. At the same time, they can capture non-linear relationships. What I wanted to do was predict the categories of videos using the *duration* and *channel name* features. I also had to encode the channel names because machine learning models can only work with numerical data. Therefore, each channel name was encoded with a unique number.

To improve the model's accuracy, I used the k-fold cross-validation technique, setting *k* to 5. This meant that each fold would use 80% of the data for training and 20% for testing, and the positioning of these percentages would change in each iteration. My dataset consisted of videos from the top 5 most-watched categories.

The results were as follows:



True \ Predicted	Entertainment	Gaming	Education	Music	People & Blogs
Entertainment	2377	55	773	548	365
Gaming	466	2815	320	487	674
Education	591	78	2536	306	857
Music	294	34	141	2348	402
People & Blogs	274	28	289	372	1198



```
Cross-validation scores: [0.62614063 0.58319914 0.60601181 0.59328859 0.61744966]
Average CV Score: 0.605 (+/- 0.031)

Classification Report:
               precision    recall  f1-score   support

    Education       0.59      0.58      0.59      4118
Entertainment       0.94      0.59      0.72      4762
       Gaming       0.62      0.58      0.60      4368
        Music       0.58      0.73      0.65      3219
People & Blogs      0.34      0.55      0.42      2161

     accuracy                           0.61     18628
    macro avg       0.61      0.61      0.60     18628
 weighted avg       0.66      0.61      0.62     18628


Feature Importances:
duration: 0.402
channel: 0.598
```

The model performed best in the Entertainment category, achieving 94% precision, 59% recall, and an F1-score of 0.72. It performed worst in the People & Blogs category, likely due to the more diverse nature of this category.

When looking at feature importance, we see that the *Channel* feature scored higher than the *Duration* feature. This makes sense because certain channels tend to focus on specific types of content.

Additionally, the model performed better than random guessing (20% for 5 categories).

It's also worth pointing out one important detail: YouTube channels typically stick to specific categories. The model might simply be memorizing which channel posts which category, leading to higher accuracy.