1. Introduction

This project analyzes my personal YouTube watching habits by exploring category preferences, engagement metrics, and temporal patterns. I used my watch history dataset enriched with additional metadata fetched via the YouTube API to perform statistical analyses. The main objective was to uncover trends and test hypotheses about changes in viewing behavior over time and different parts of the day.

2. Data Collection and Preparation

1. **Data Collection:**

   - I requested my YouTube watch history from YouTube, which provided a CSV file containing information such as video IDs, watch timestamps, and basic metadata.

   - Using the YouTube API, I queried the videoId field to fetch additional information, including:

     - Video categories (category_id),

     - Descriptions,

     - Engagement metrics (likes, comments, views).

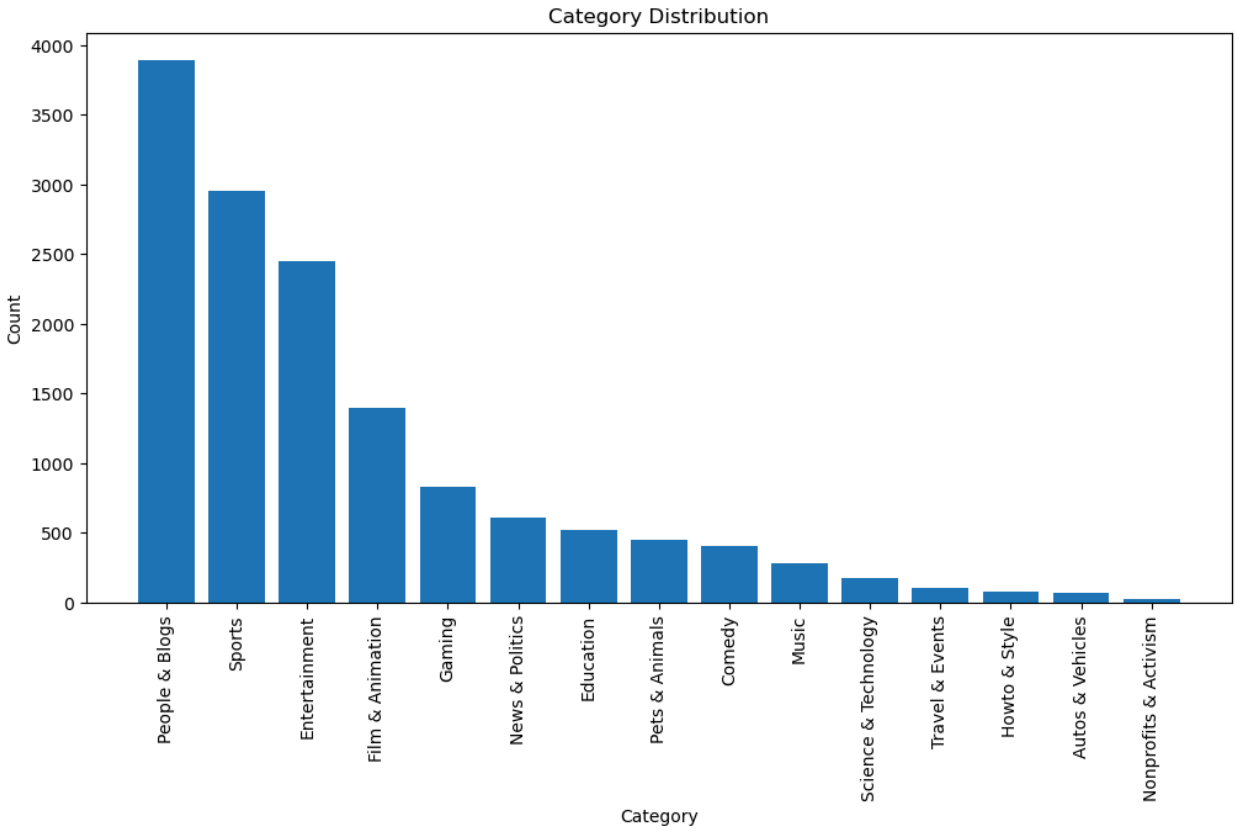
2. **Data Transformation:**

   - Converted the raw data into a Pandas DataFrame for easier manipulation.

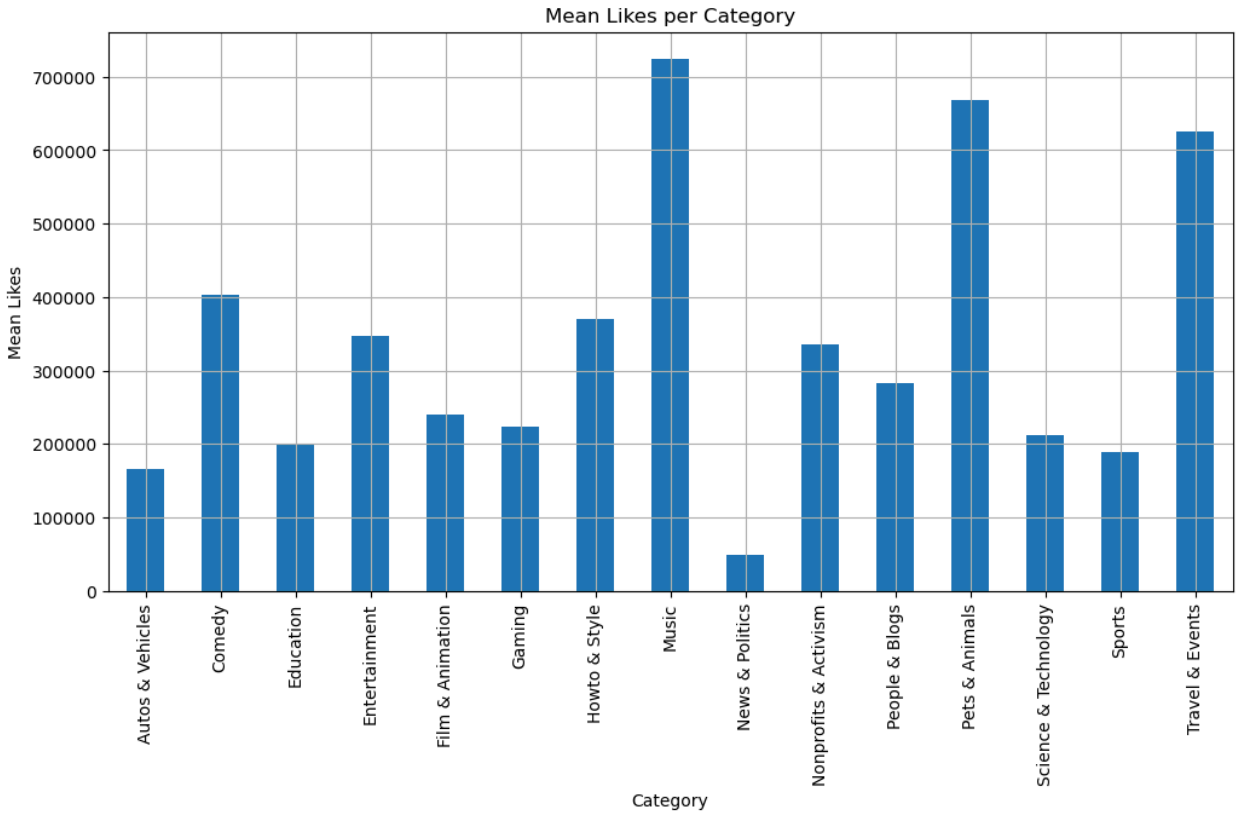
	video_id	title	description	published_at	tags	category_id	defaultAudioLanguage	duration	channel	views	lik
0	pILKU5fB6Jk	George R. R. Martin Regrets: Game of Thrones F...	#georgerrmartin #gameofthrones #houseofthedrag...	2024-08-14T19:30:06Z	['Game of Thrones', 'George R. R. Martin', 'Go...	24	zxx	PT58S	Clips Theory	109276	43:
1	5G4Mg58vEYo	Rachel and Ross who hurt each other #friends #...	NaN	2024-10-31T01:54:13Z	[]	1	en-US	PT1M	Jacob Evie	8161020	3362
2	jYlAlEpiuEE	#beyazfutbol #beşiktaş #galatasaray #fenerbahç...	NaN	2024-12-01T04:17:32Z	[]	17	tr	PT49S	90 Dakikka	5022	(
3	SeBr75boB_c	5 Year Old Saves Mother In Court 😱	NaN	2024-11-19T22:00:16Z	[]	22	NaN	PT1M1S	courtshorts	21423646	12288
4	6hB2RLVqRl8	Alex Pereira KNOCKING OUT Jamahal Hill 🤭 #noco...	Order UFC PPV on ESPN+ ➡️ https://ufc.ac/3NKBv...	2024-11-30T14:00:05Z	['ufc', 'mma', 'ultimate fighting championship']	17	en	PT52S	UFC	756576	337

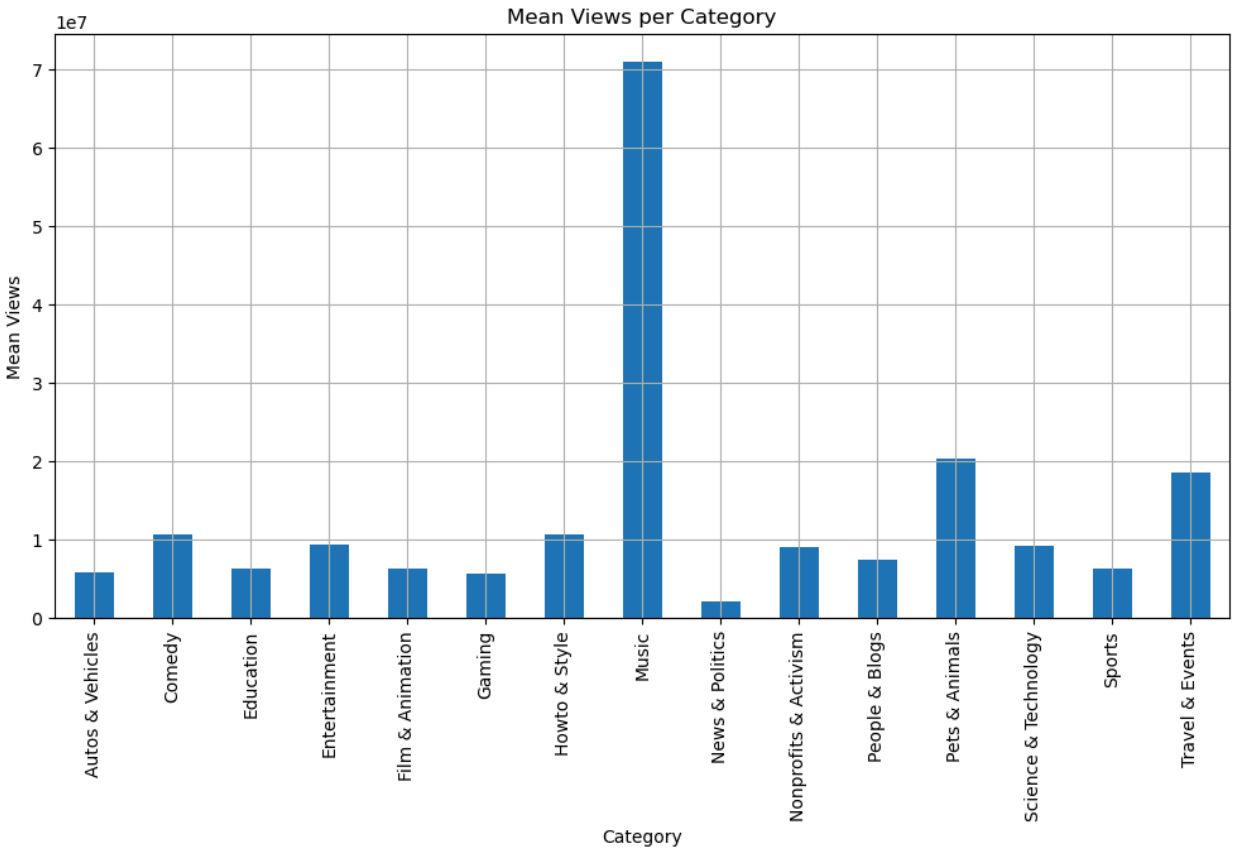
defaultAudioLanguage	duration	channel	views	likes	comments	engage	caption	favorite
zxx	PT58S	Clips Theory	109276	4334	140	4.334001e+03	False	0
en-US	PT1M	Jacob Evie	8161020	336246	541	3.362460e+05	False	0
tr	PT49S	90 Dakikka	5022	64	0	6.400000e+01	False	0
NaN	PT1M1S	courtshorts	21423646	1228864	2698	1.228864e+06	False	0
en	PT52S	UFC	756576	33730	402	3.373000e+04	False	0

- Mapped the category_id field to human-readable category_name.
- Conducted exploratory data analysis (EDA) to identify:
    - My most-watched categories,

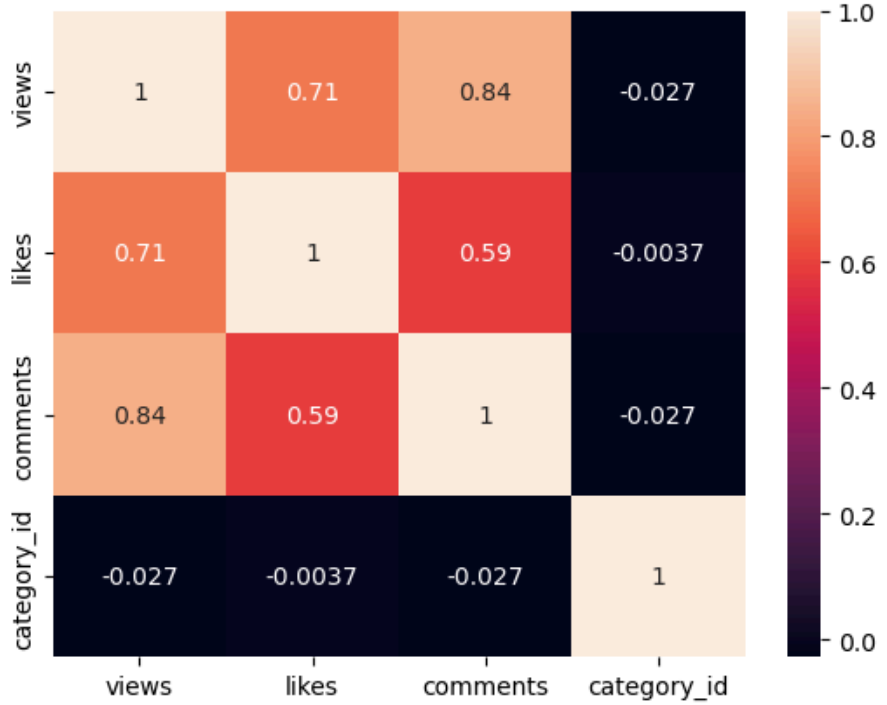


- Mean likes and views per category,





- Correlation matrix of variables



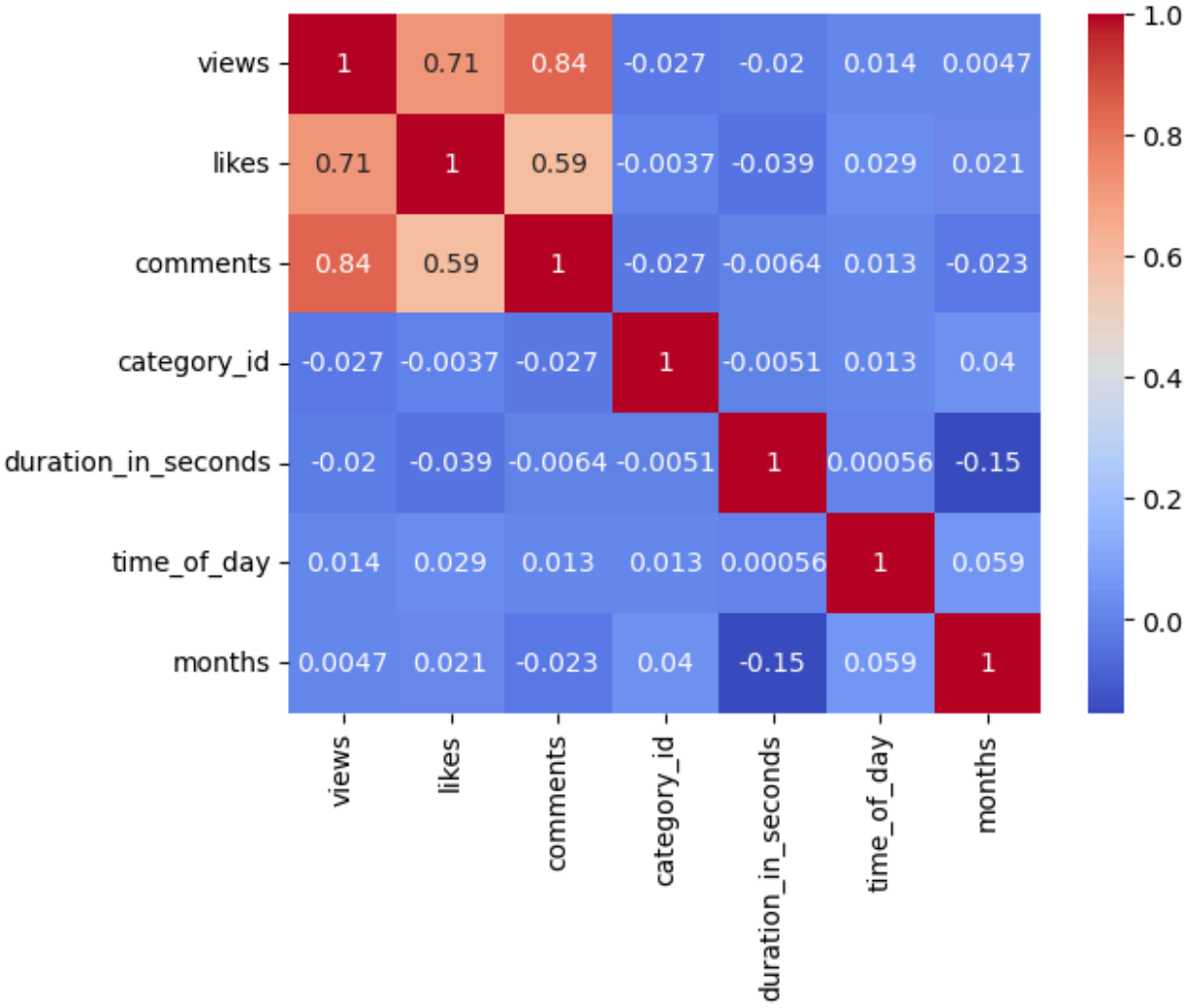
It is seen that there is no strong relationship between category_id and engagement metrics. Even though mean likes and views of music category is significantly higher than other categories, the other variables are not varied a lot in terms of engagement metrics.

3. **Time Segmentation:**

    - I divided the whole time into quarter periods and month periods. Added new columns ‘time_segment’ for quarters and ‘time_segmen_in_months’ for month periods’
    - Added a new column time_segment_in_day to classify each video’s watch time into one of these segments.
    - I divided the day into four segments:
        - Night (12 AM–6 AM),
        - Morning (6 AM–12 PM),
        - Afternoon (12 PM–6 PM),
        - Evening (6 PM–12 AM).
    - Added a new column time_segment_in_day to classify each video’s watch time into one of these segments.

4. **duration_in_seconds variable:**

    - I have added duration_in_seconds variable which is converted version of duration variable to second units. This variable will be used to see if the total watch time varies over time with the assumption that I watched all the videos until the end.



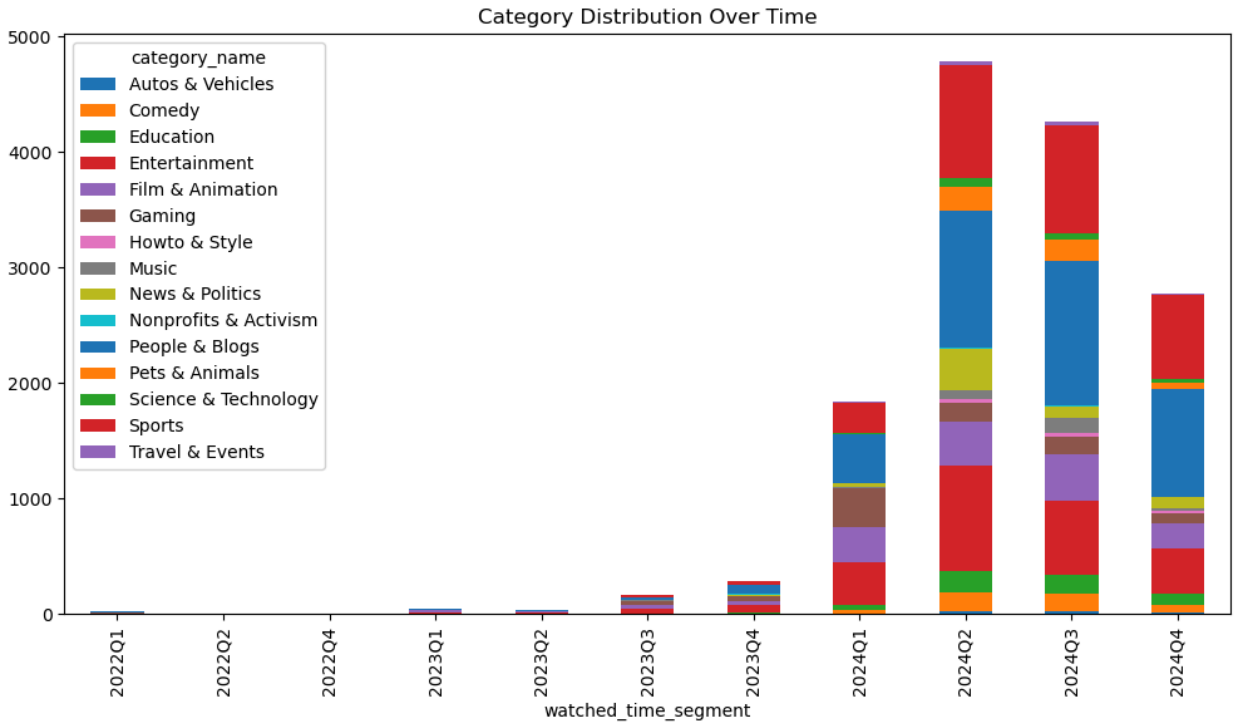
Correlation matrix between new variables still cannot imply a strong relationship between category_id and other variables. However I applied chi-squared tests to see if categories vary over time.

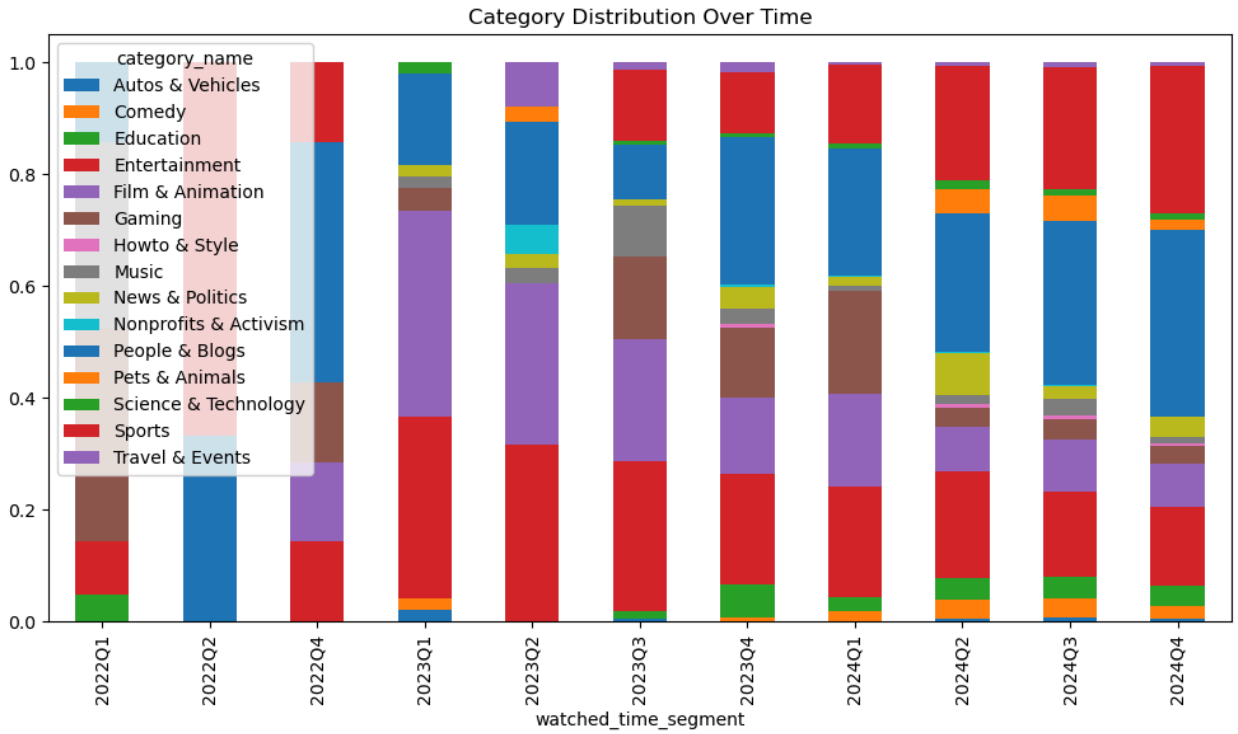
3.1. Categories Over Time

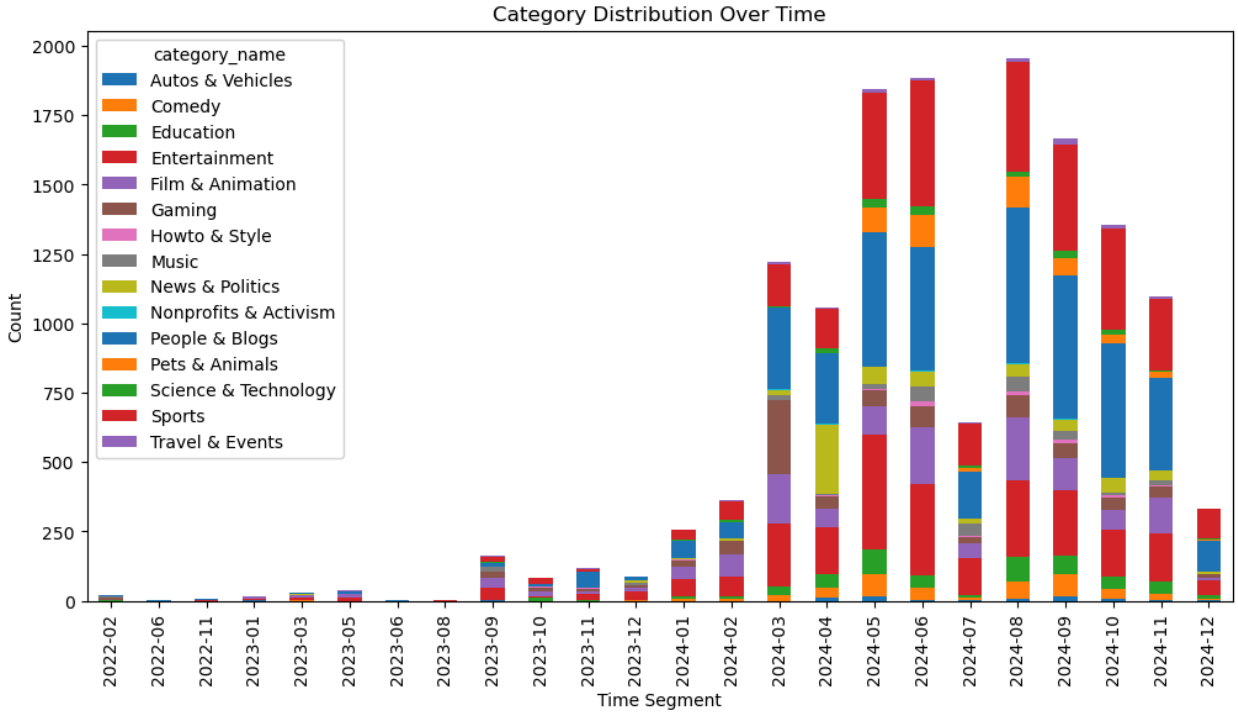
- **Hypothesis:**
  - Null Hypothesis (H_0): The categories I prefer do not change significantly over time.
  - Alternative Hypothesis (H_1): The categories I prefer change significantly over time.
- **Method:**
  - Applied a **Chi-Square Test of Independence** to evaluate category distributions over:
    - Quarters,

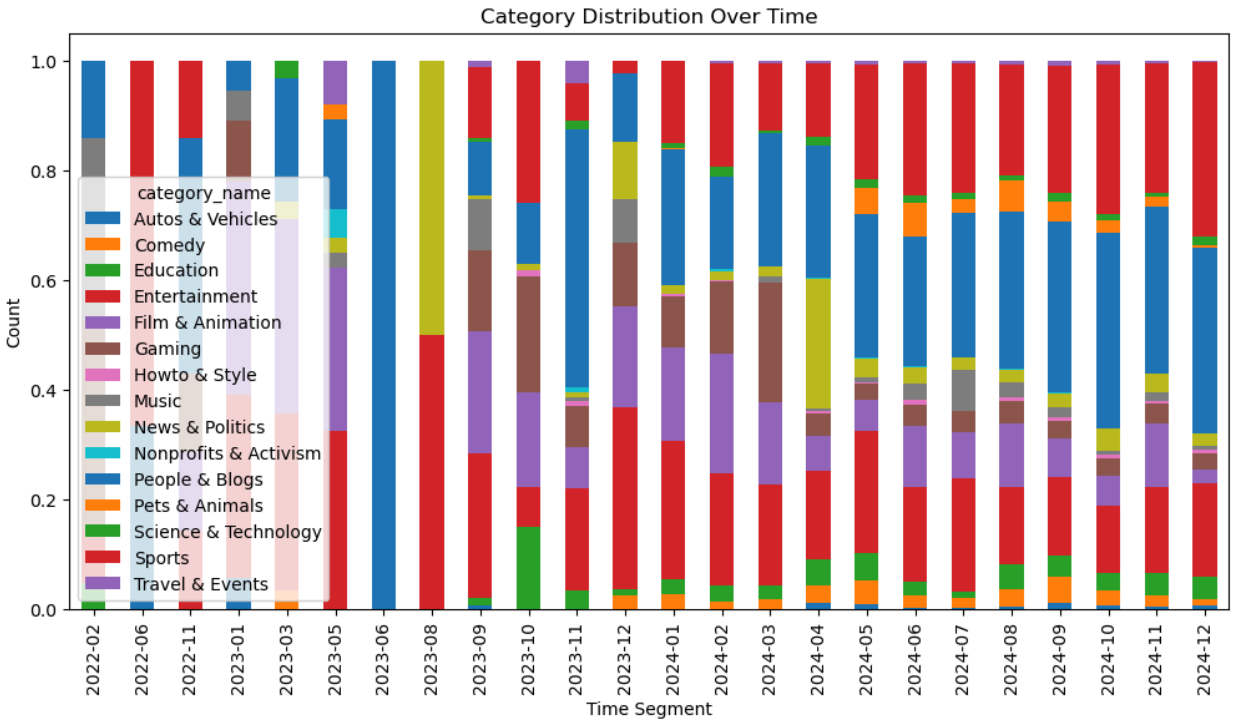
- Months.

- **Findings:**
  - Using **raw counts**, I found significant changes in category preferences over time.
  - Using **proportions**, there were no significant changes, highlighting stability in relative preferences despite variations in overall volume.




Category Distribution Over Time
category_name
Autos & Vehicles
Comedy
Education
Entertainment
Film & Animation
Gaming
Howto & Style
Music
News & Politics
Nonprofits & Activism
People & Blogs
Pets & Animals
Science & Technology
Sports
Travel & Events
1.0
0.8
0.6
0.4
0.2
0.0
2022Q1
2022Q2
2022Q4
2023Q1
2023Q2
2023Q3
2023Q4
2024Q1
2024Q2
2024Q3
2024Q4
watched_time_segment


Category Distribution Over Time
category_name
Autos & Vehicles
Comedy
Education
Entertainment
Film & Animation
Gaming
Howto & Style
Music
News & Politics
Nonprofits & Activism
People & Blogs
Pets & Animals
Science & Technology
Sports
Travel & Events
Count
2000
1750
1500
1250
1000
750
500
250
0
2022-02
2022-06
2022-11
2023-01
2023-03
2023-05
2023-06
2023-08
2023-09
2023-10
2023-11
2023-12
2024-01
2024-02
2024-03
2024-04
2024-05
2024-06
2024-07
2024-08
2024-09
2024-10
2024-11
2024-12
Time Segment

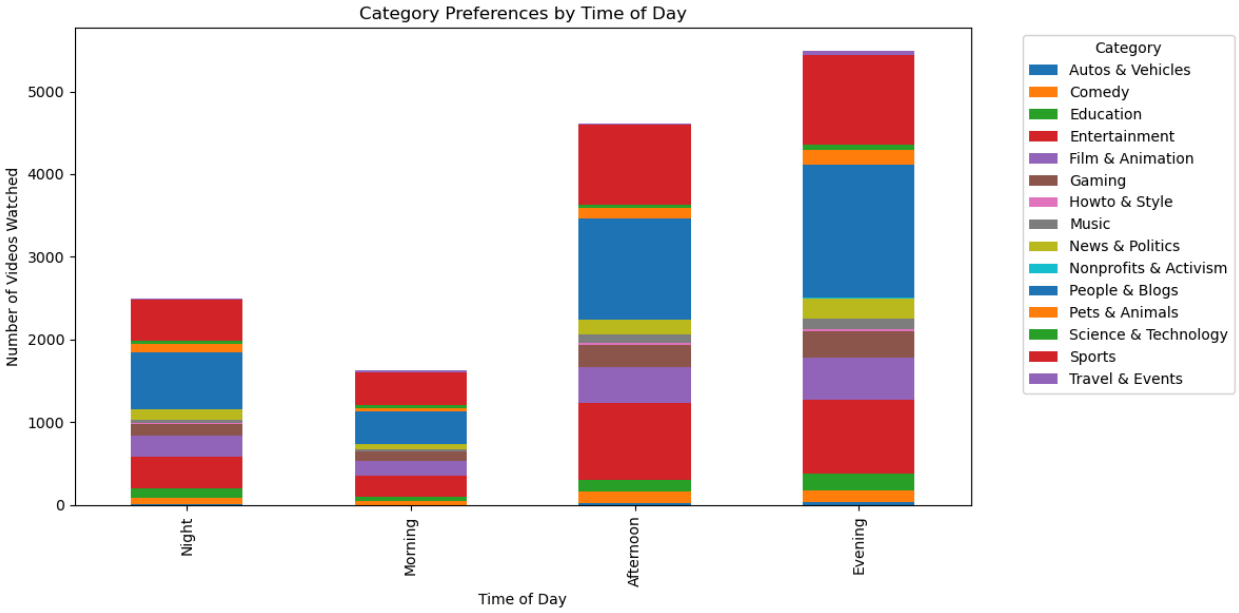


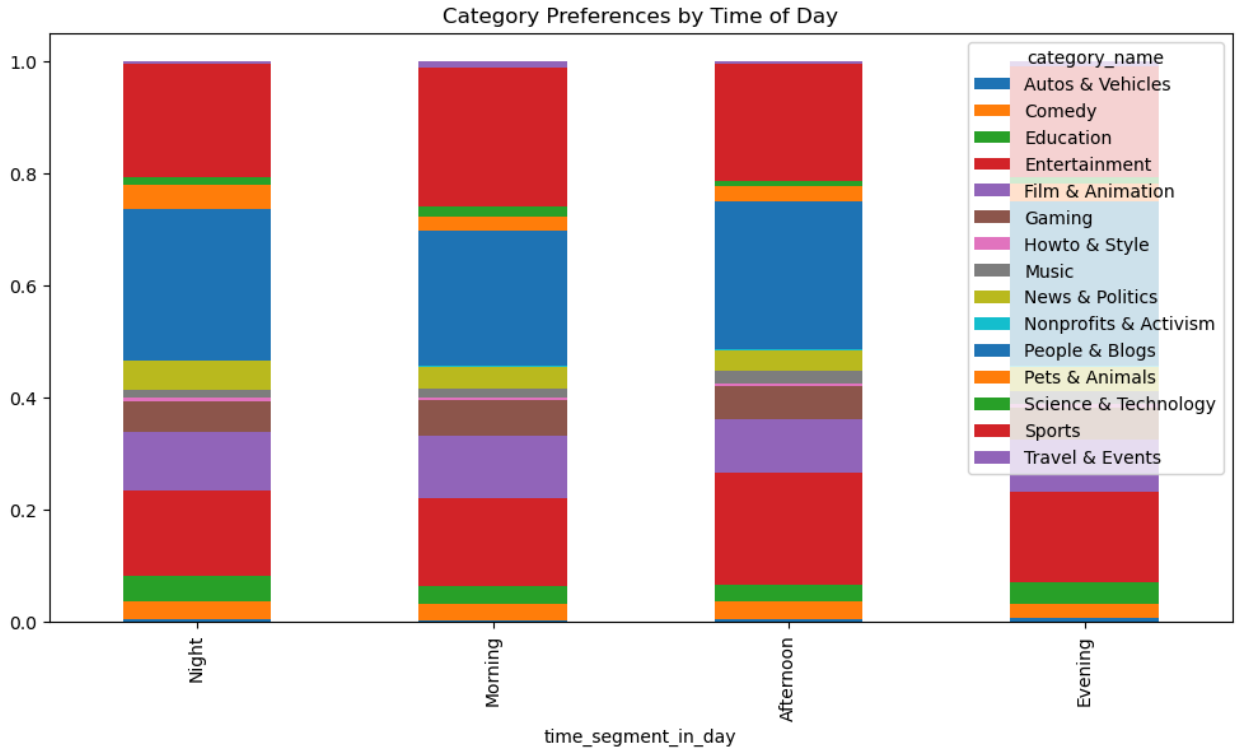
As it can be seen from the figures, the different results from two approaches which are taking raw counts and proportions stems from the variety of the number of videos watched over time.

These results shows that there is no shift in preferences in my match history.

3.2. Categories by Time of Day

- **Hypothesis:**
    - Null Hypothesis (H_0): The categories I prefer do not change significantly by time of day.
    - Alternative Hypothesis (H_1): The categories I prefer change significantly by time of day.
- **Method:**
    - Divided the day into four segments (Night, Morning, Afternoon, Evening) and applied a **Chi-Square Test**.
- **Findings:**
    - We have the same result with overall time that there is no significant difference when we use proportions as input and there is significant difference when we use raw counts.



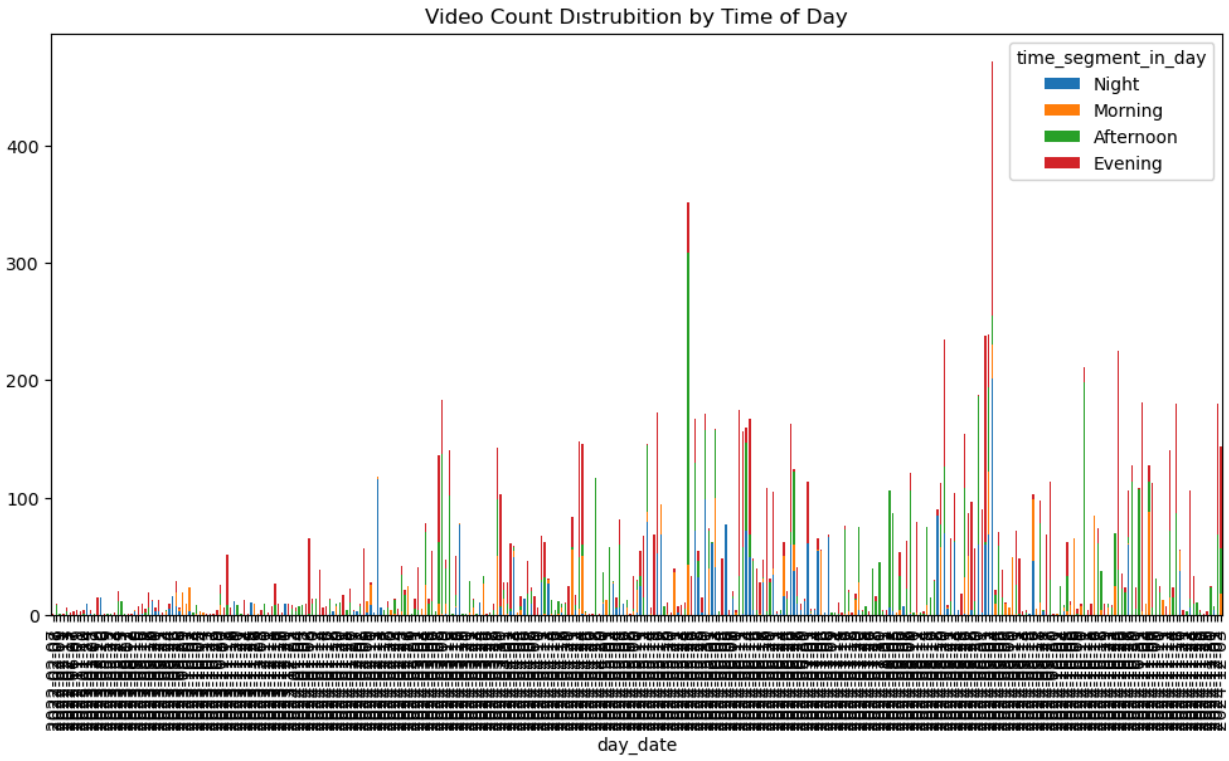


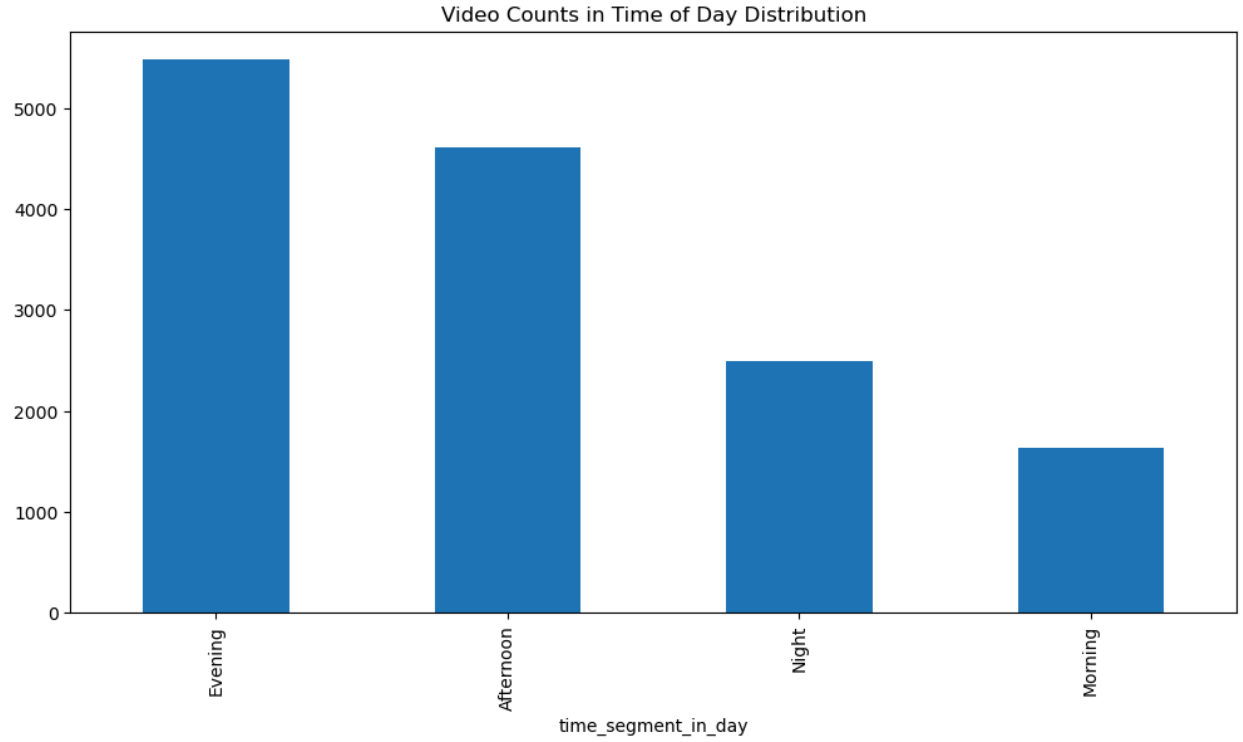
3.3. Total Videos and Duration by Time of Day

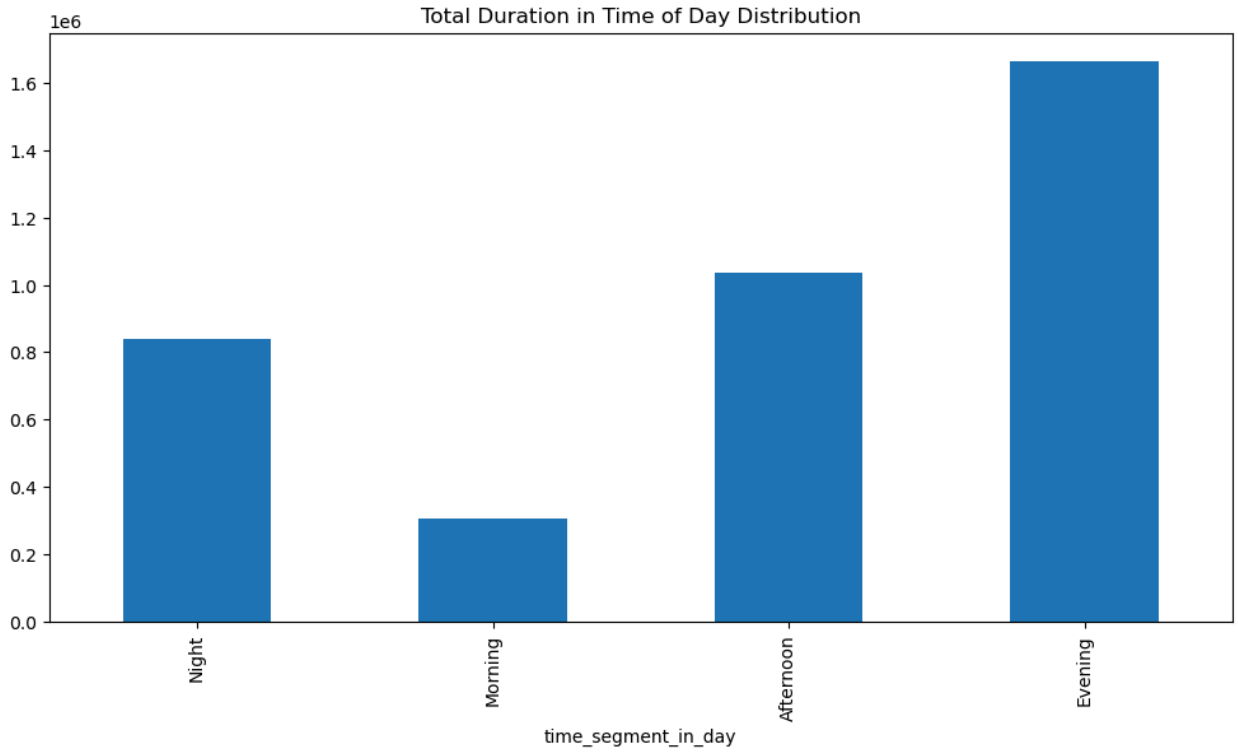
- **Hypothesis:**

- Null Hypothesis (H_0): The total number of videos or their total duration does not change significantly by time of day.
- Alternative Hypothesis (H_1): The total number of videos or their total duration changes significantly by time of day.

- **Method:**
  - Reshaped the DataFrame to aggregate counts and total durations by day and time segment.


Video Count Distrubition by Time of Day
time_segment_in_day
Night
Morning
Afternoon
Evening
400
300
200
100
0
day_date


Video Counts in Time of Day Distribution
5000
4000
3000
2000
1000
0
Evening
Afternoon
Night
Morning
time_segment_in_day



It can be seen that both total duration and number of videos changes according to time in day. I mostly watch videos in evening. However we will still apply tests to see if this difference can be proved.

- Applied:
    - **ANOVA:** Found significant differences in both metrics across time segments.
    - **Levene's Test:** Variances were unequal, violating ANOVA assumptions.
    - **Kruskal-Wallis Test:** A non-parametric test confirmed significant differences across time segments.

4. Key Findings

1. **Category Preferences:**
    - Even though we saw that categories change by time with the first. This difference occurred since the variety of number of videos. In overall, the preferences of categories are not changed significantly, which we were able to see by testing proportions.
2. **Video Counts and Durations:**
    - Both the number of videos and their total durations varied significantly by time segments (Night, Morning, Afternoon, Evening).

- Despite unequal variances (Levene's test), the Kruskal-Wallis test confirmed significant differences.

5. Limitations

1. The dataset reflects my personal watching habits, limiting the generalizability of the findings.
2. External factors (e.g., recommendations, mood) influencing watching habits were not accounted for.
3. Levene's test indicated unequal variances, which violates ANOVA assumptions. However Kruskal-Wallis test is applied later.
4. The assumption that I have watched all the videos until the end can easily be violated.

6. Conclusion

This project revealed significant temporal patterns in my YouTube watching habits:

- Category preferences and engagement metrics are influenced by both time of day and time period.
- Absolute video counts and durations differ significantly across time segments.
- While relative preferences remain stable, overall behavior varies due to volume differences.

Future work could focus on exploring recommendation impacts or building a predictive model for viewing habits.