Youtube Data Presentation 1

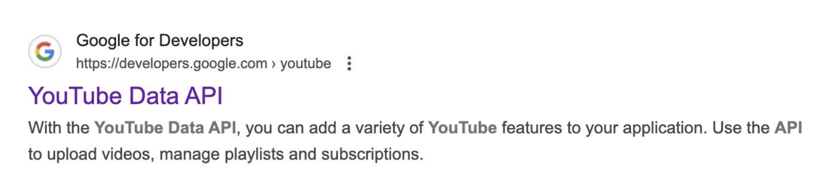
Muneera Aldoseri - 23229578
Maliha Moshrafa Hossain - 23229026
Arwen Sakura Wise - 23229617

OVERVIEW

OUR DATA	Youtube v3 API
Research Question 1	Are music videos getting higher view counts than educational videos?
Research Question 2	Are fitness videos getting higher like counts than vlogging videos?
Research Question 3	Does upload timing affect the number of likes a video receives?
Conclusion	Brief restatement of the Hypothesis 1, 2 and 3
References	https://developers.google.com/youtube/v3

How We Collected Our Data

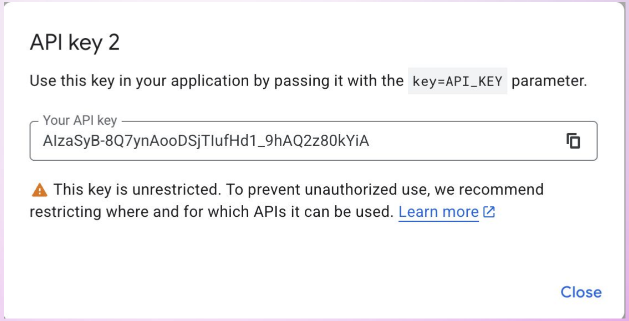
- Used YouTube Data API v3 via Google Cloud Console
- Generated API key and connected it to Python
- Collected data from YouTube for selected topics related to our hypotheses
- Extracted:
  - Video title, published date
  - View count, like count, comment count
  - Duration and other metadata
- Stored results in Pandas DataFrames for analysis



Google for Developers
https://developers.google.com › youtube

YouTube Data API

With the **YouTube Data API**, you can add a variety of **YouTube** features to your application. Use the **API** to upload videos, manage playlists and subscriptions.



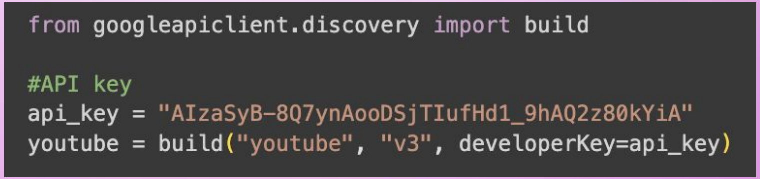
API key 2

Use this key in your application by passing it with the `key=API_KEY` parameter.

Your API key
AIzaSyB-8Q7ynAooDSjTIufHd1_9hAQ2z80kYiA

⚠ This key is unrestricted. To prevent unauthorized use, we recommend restricting where and for which APIs it can be used. Learn more

Close



```
from googleapiclient.discovery import build

#API key
api_key = "AIzaSyB-8Q7ynAooDSjTIufHd1_9hAQ2z80kYiA"
youtube = build("youtube", "v3", developerKey=api_key)
```

Research Question & Hypothesis 1

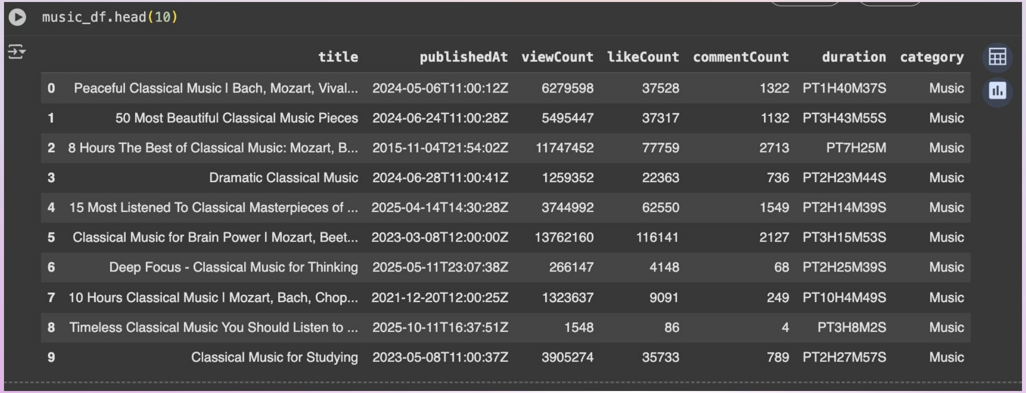
Research Question: Do music videos get higher view counts than educational videos on YouTube?

Hypothesis: Classical music videos have higher average view counts than math tutorials.

Null Hypothesis: There is no significant difference between the two categories.

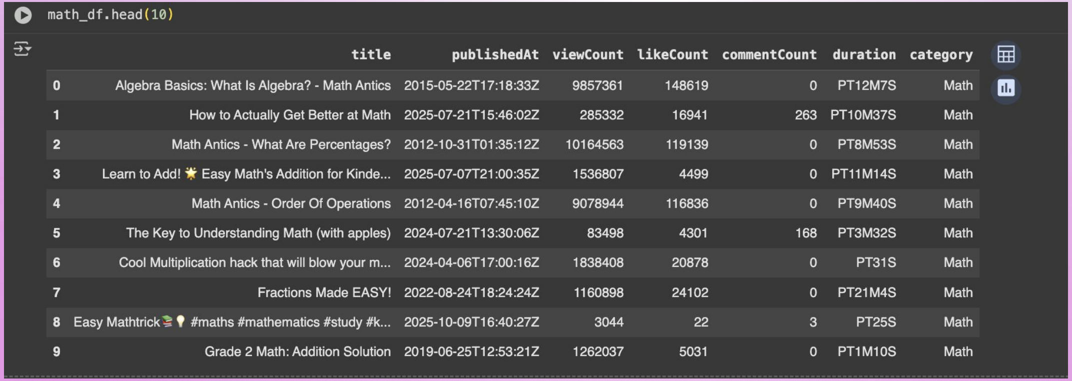
Data Preparation

- Retrieved data on two topics: “classical music” and “math tutorial”
- Fetched 400 videos (200 each)
- Labeled each video with a Category column (Music / Math)
- Stored results in Pandas DataFrames



```
music_df.head(10)
```

	title	publishedAt	viewCount	likeCount	commentCount	duration	category
0	Peaceful Classical Music I Bach, Mozart, Vival...	2024-05-06T11:00:12Z	6279598	37528	1322	PT1H40M37S	Music
1	50 Most Beautiful Classical Music Pieces	2024-06-24T11:00:28Z	5495447	37317	1132	PT3H43M55S	Music
2	8 Hours The Best of Classical Music: Mozart, B...	2015-11-04T21:54:02Z	11747452	77759	2713	PT7H25M	Music
3	Dramatic Classical Music	2024-06-28T11:00:41Z	1259352	22363	736	PT2H23M44S	Music
4	15 Most Listened To Classical Masterpieces of ...	2025-04-14T14:30:28Z	3744992	62550	1549	PT2H14M39S	Music
5	Classical Music for Brain Power I Mozart, Beet...	2023-03-08T12:00:00Z	13762160	116141	2127	PT3H15M53S	Music
6	Deep Focus - Classical Music for Thinking	2025-05-11T23:07:38Z	266147	4148	68	PT2H25M39S	Music
7	10 Hours Classical Music I Mozart, Bach, Chop...	2021-12-20T12:00:25Z	1323637	9091	249	PT10H4M49S	Music
8	Timeless Classical Music You Should Listen to ...	2025-10-11T16:37:51Z	1548	86	4	PT3H8M2S	Music
9	Classical Music for Studying	2023-05-08T11:00:37Z	3905274	35733	789	PT2H27M57S	Music



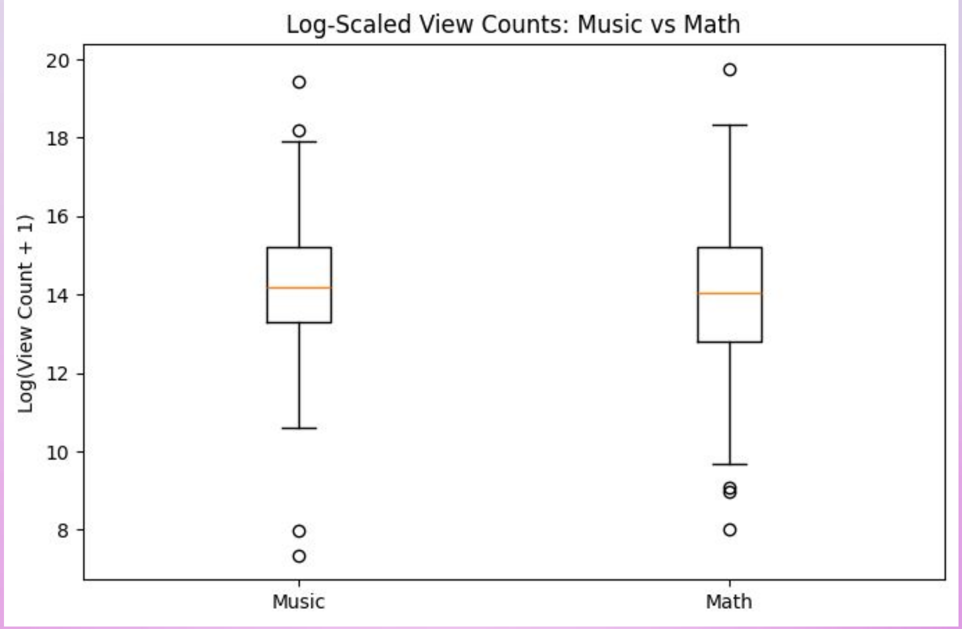
```
math_df.head(10)
```

	title	publishedAt	viewCount	likeCount	commentCount	duration	category
0	Algebra Basics: What Is Algebra? - Math Antics	2015-05-22T17:18:33Z	9857361	148619	0	PT12M7S	Math
1	How to Actually Get Better at Math	2025-07-21T15:46:02Z	285332	16941	263	PT10M37S	Math
2	Math Antics - What Are Percentages?	2012-10-31T01:35:12Z	10164563	119139	0	PT8M53S	Math
3	Learn to Add! 🌟 Easy Math's Addition for Kinde...	2025-07-07T21:00:35Z	1536807	4499	0	PT11M14S	Math
4	Math Antics - Order Of Operations	2012-04-16T07:45:10Z	9078944	116836	0	PT9M40S	Math
5	The Key to Understanding Math (with apples)	2024-07-21T13:30:06Z	83498	4301	168	PT3M32S	Math
6	Cool Multiplication hack that will blow your m...	2024-04-06T17:00:16Z	1838408	20878	0	PT31S	Math
7	Fractions Made EASY!	2022-08-24T18:24:24Z	1160898	24102	0	PT21M4S	Math
8	Easy Mathtrick📚💡 #maths #mathematics #study #k...	2025-10-09T16:40:27Z	3044	22	3	PT25S	Math
9	Grade 2 Math: Addition Solution	2019-06-25T12:53:21Z	1262037	5031	0	PT1M10S	Math

Visual Analysis

Boxplot:

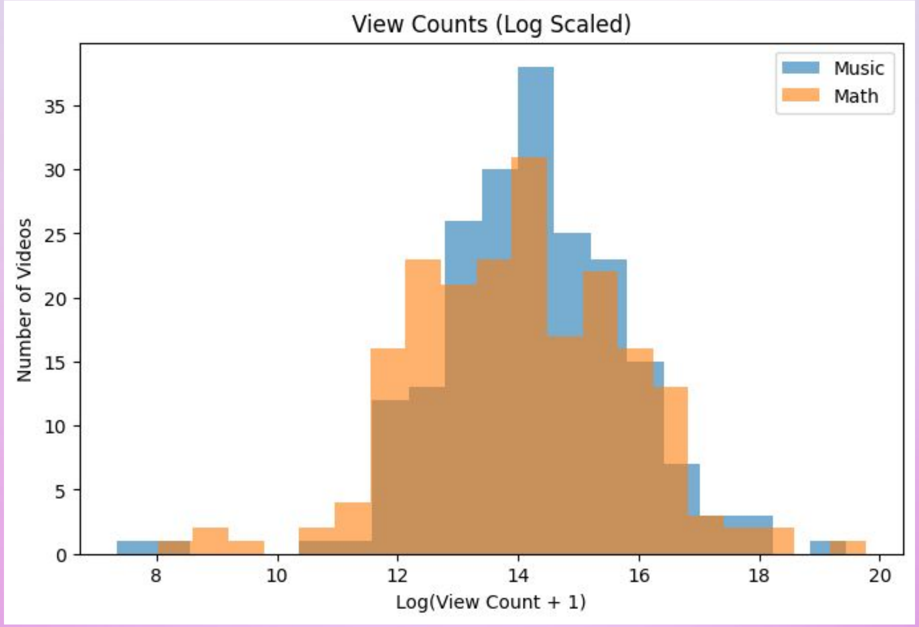
- Compared view count distributions for classical music and math tutorials
- Used logarithmic scale to handle large differences in view counts
- Similar medians → average view count levels are close
- Similar spread of views
- Some outliers on both



Visual Analysis

Histogram:

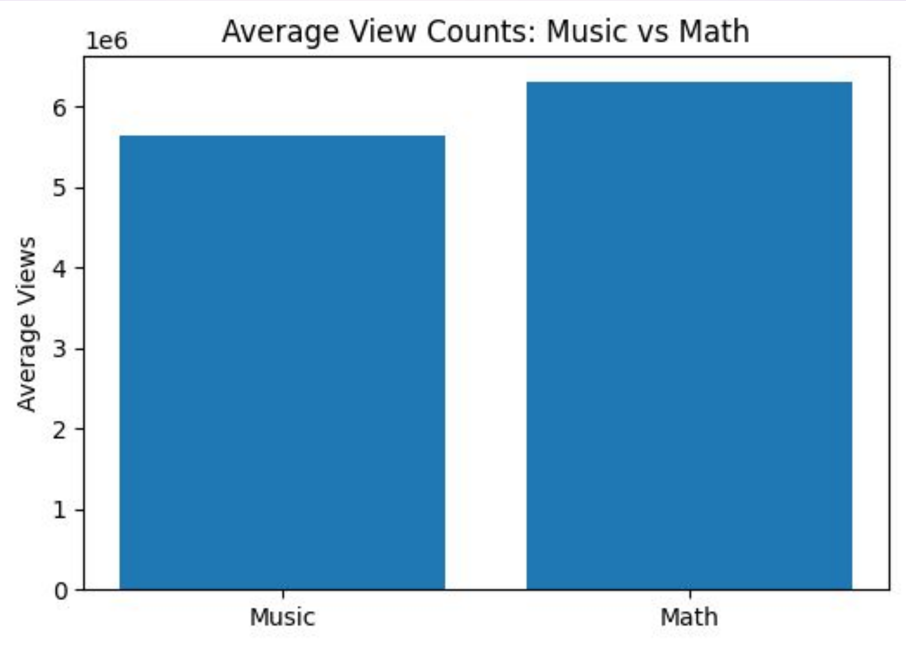
- Displays distribution of log scaled view counts for both topics
- Overlapping shapes → similar range of views
- Overall similar pattern, peaking at mid-range views
- Suggests that classical music videos generally attract more viewers, supporting the hypothesis



Visual Analysis

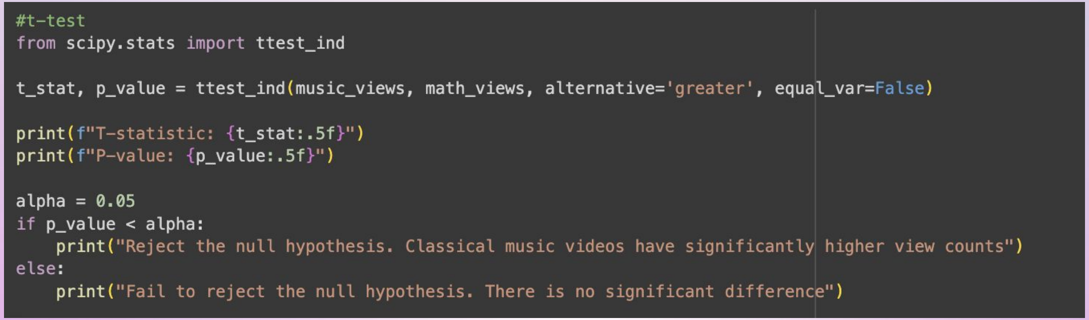
Bar chart:

- Compared average view counts between classical music and math tutorial videos
- Math videos slightly higher average views
- Small difference → perform similarly



Statistical Test & Results

- Used t-test to compare both groups
- T-statistic: −0.26
- P-value: 0.6028
- Significance level (α): 0.05
- Since p > α → Fail to reject the null hypothesis
- No significant difference between view counts

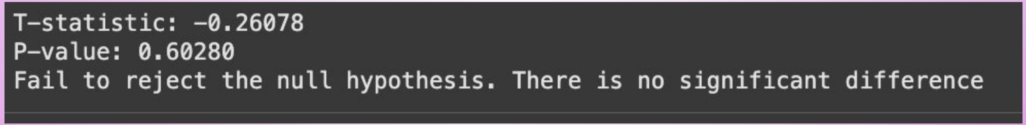


```
#t-test
from scipy.stats import ttest_ind

t_stat, p_value = ttest_ind(music_views, math_views, alternative='greater', equal_var=False)

print(f"T-statistic: {t_stat:.5f}")
print(f"P-value: {p_value:.5f}")

alpha = 0.05
if p_value < alpha:
    print("Reject the null hypothesis. Classical music videos have significantly higher view counts")
else:
    print("Fail to reject the null hypothesis. There is no significant difference")
```



```
T-statistic: -0.26078
P-value: 0.60280
Fail to reject the null hypothesis. There is no significant difference
```

Research Question & Hypothesis 2

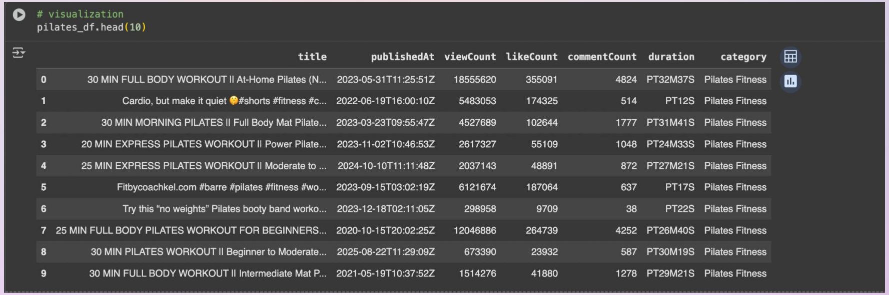
Research Question: Are fitness videos getting higher like counts than vlogging videos?

Hypothesis: Travel vlog videos get higher average like counts than Pilates Fitness videos.

Null Hypothesis: There is no significant difference between the two categories.

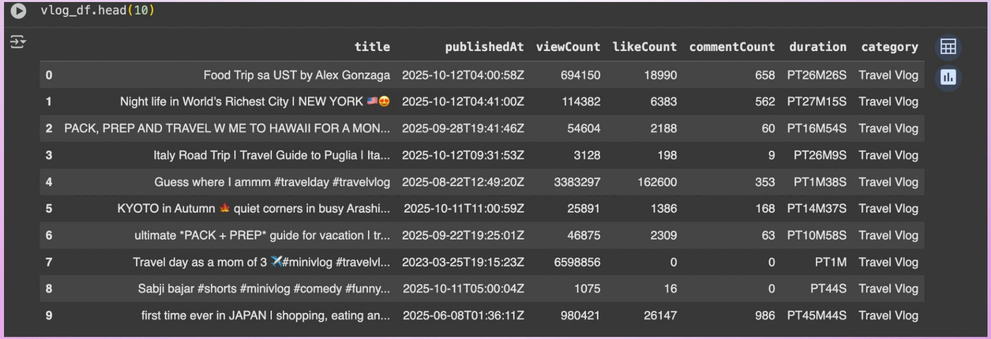
Data Preparation (2)

- Searched and fetched data for two topics -
  1. Travel vlog videos
  2. Pilates fitness videos
- Labeled categories and showed them in data frames for each categories.
- Visualised summary statistics for each categories.
- Stored the results.



```
# visualization
pilates_df.head(10)
```

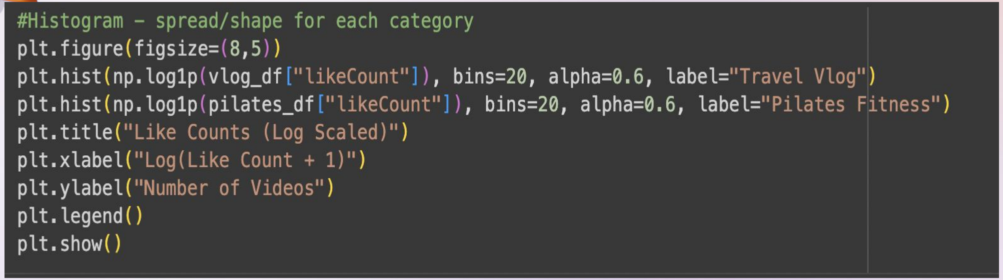
	title	publishedAt	viewCount	likeCount	commentCount	duration	category
0	30 MIN FULL BODY WORKOUT II At-Home Pilates (N...	2023-05-31T11:25:51Z	18555620	355091	4824	PT32M37S	Pilates Fitness
1	Cardio, but make it quiet 🤫#shorts #fitness #c...	2022-06-19T16:00:10Z	5483053	174325	514	PT12S	Pilates Fitness
2	30 MIN MORNING PILATES II Full Body Mat Pilate...	2023-03-23T09:55:47Z	4527689	102644	1777	PT31M41S	Pilates Fitness
3	20 MIN EXPRESS PILATES WORKOUT II Power Pilate...	2023-11-02T10:46:53Z	2617327	55109	1048	PT24M33S	Pilates Fitness
4	25 MIN EXPRESS PILATES WORKOUT II Moderate to ...	2024-10-10T11:11:48Z	2037143	48891	872	PT27M21S	Pilates Fitness
5	Fitbycoachkel.com #barre #pilates #fitness #wo...	2023-09-15T03:02:19Z	6121674	187064	637	PT17S	Pilates Fitness
6	Try this “no weights” Pilates booty band worko...	2023-12-18T02:11:05Z	298958	9709	38	PT22S	Pilates Fitness
7	25 MIN FULL BODY PILATES WORKOUT FOR BEGINNERS...	2020-10-15T20:02:25Z	12046886	264739	4252	PT26M40S	Pilates Fitness
8	30 MIN PILATES WORKOUT II Beginner to Moderate...	2025-08-22T11:29:09Z	673390	23932	587	PT30M19S	Pilates Fitness
9	30 MIN FULL BODY WORKOUT II Intermediate Mat P...	2021-05-19T10:37:52Z	1514276	41880	1278	PT29M21S	Pilates Fitness



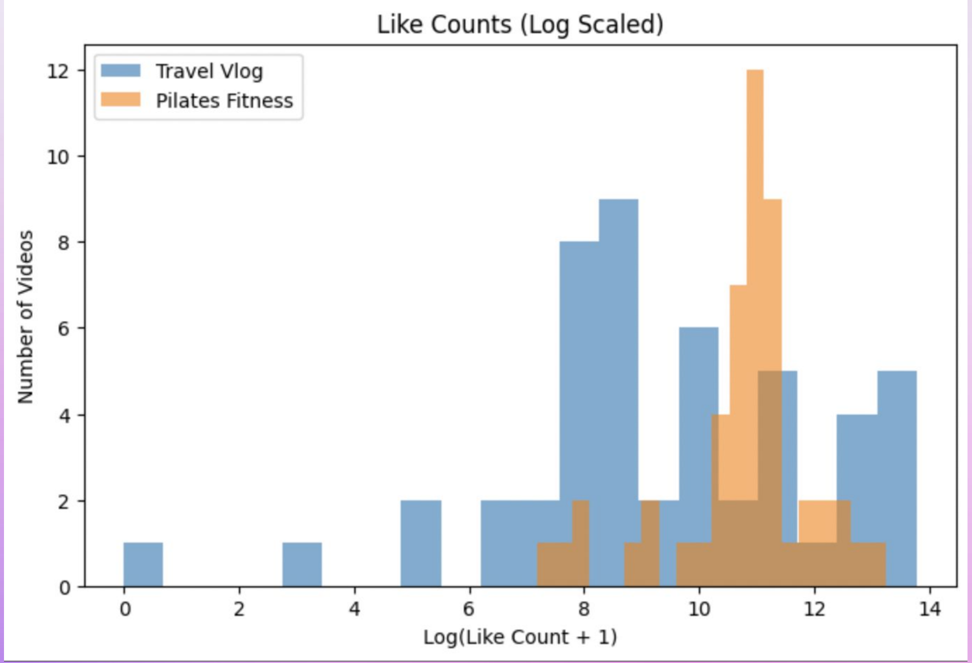
```
vlog_df.head(10)
```

	title	publishedAt	viewCount	likeCount	commentCount	duration	category
0	Food Trip sa UST by Alex Gonzaga	2025-10-12T04:00:58Z	694150	18990	658	PT26M26S	Travel Vlog
1	Night life in World’s Richest City I NEW YORK 🇺🇸😍	2025-10-12T04:41:00Z	114382	6383	562	PT27M15S	Travel Vlog
2	PACK, PREP AND TRAVEL W ME TO HAWAII FOR A MON...	2025-09-28T19:41:46Z	54604	2188	60	PT16M54S	Travel Vlog
3	Italy Road Trip I Travel Guide to Puglia I Ita...	2025-10-12T09:31:53Z	3128	198	9	PT26M9S	Travel Vlog
4	Guess where I ammm #travelday #travelvlog	2025-08-22T12:49:20Z	3383297	162600	353	PT1M38S	Travel Vlog
5	KYOTO in Autumn 🍁 quiet corners in busy Arashi...	2025-10-11T11:00:59Z	25891	1386	168	PT14M37S	Travel Vlog
6	ultimate *PACK + PREP* guide for vacation I tr...	2025-09-22T19:25:01Z	46875	2309	63	PT10M58S	Travel Vlog
7	Travel day as a mom of 3 ✈️#minivlog #travelvl...	2023-03-25T19:15:23Z	6598856	0	0	PT1M	Travel Vlog
8	Sabji bajar #shorts #minivlog #comedy #funny...	2025-10-11T05:00:04Z	1075	16	0	PT44S	Travel Vlog
9	first time ever in JAPAN I shopping, eating an...	2025-06-08T01:36:11Z	980421	26147	986	PT45M44S	Travel Vlog

Visual Analysis (2)



```
#Histogram - spread/shape for each category
plt.figure(figsize=(8,5))
plt.hist(np.log1p(vlog_df["likeCount"]), bins=20, alpha=0.6, label="Travel Vlog")
plt.hist(np.log1p(pilates_df["likeCount"]), bins=20, alpha=0.6, label="Pilates Fitness")
plt.title("Like Counts (Log Scaled)")
plt.xlabel("Log(Like Count + 1)")
plt.ylabel("Number of Videos")
plt.legend()
plt.show()
```

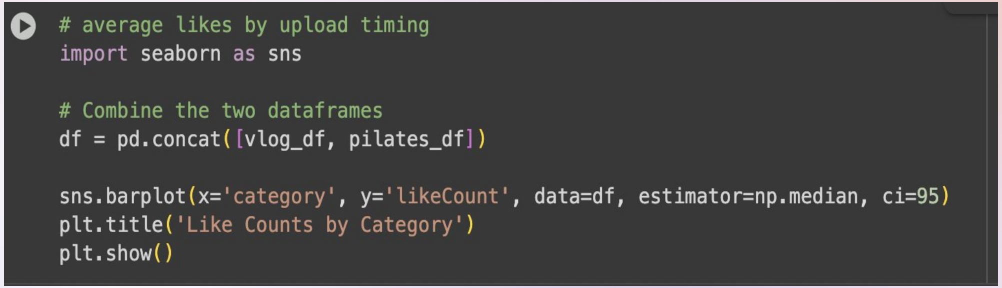


Histogram: Travel vlog videos have higher clusters(for no. of likes) while Pilates fitness videos are more spread out. Thus, travel vlog videos have more likes.

Pilates fitness videos have **taller bars for number of videos** which doesn't indicate **more videos have higher likes**.

Travel vlog videos are **more spread out** which indicates less number of videos but **has more likes than Pilates Fitness videos.**

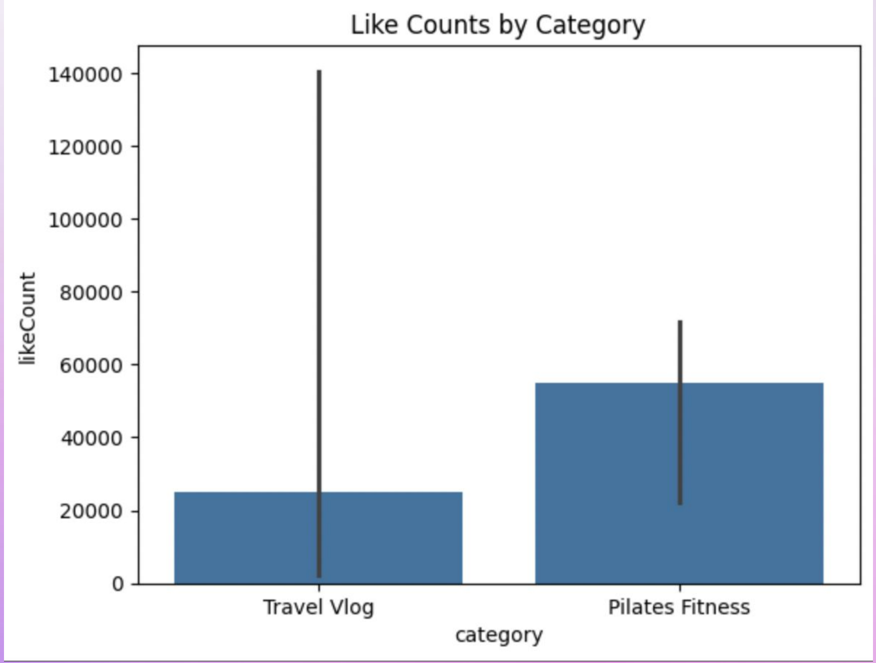
Visual Analysis (2)



```
# average likes by upload timing
import seaborn as sns

# Combine the two dataframes
df = pd.concat([vlog_df, pilates_df])

sns.barplot(x='category', y='likeCount', data=df, estimator=np.median, ci=95)
plt.title('Like Counts by Category')
plt.show()
```



Bar chart: The black line for travel vlog videos is much higher than that of pilates fitness videos.

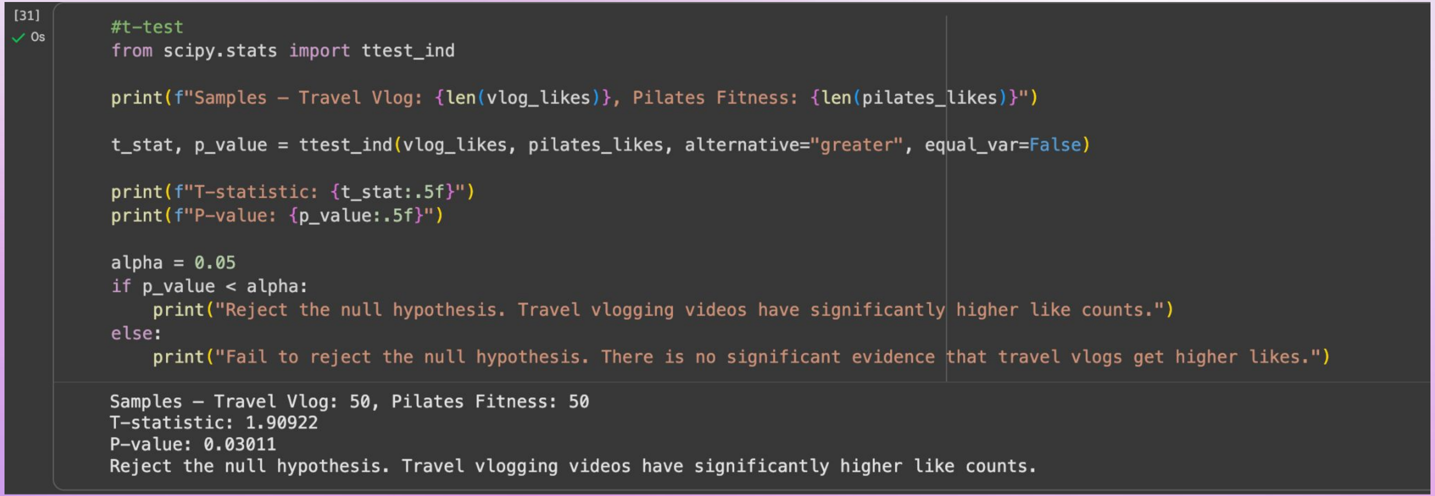
Travel vlog avg ~ 20,000
Pilates fitness avg ~ 58,000

- **Observations: Huge overlap**
- When **confidence intervals overlap**, it suggests that the **difference in medians might *not* be statistically significant**.
- Even though the **Pilates Fitness bar is higher** (≈ 58,000 than that of Travel vlog videos ≈ 20,000),
- The **large overlap** tells us that there's still a **high chance they're not significantly different** in reality.

Statistical Test & Results (2)

- T-test has been carried out -

  1. **P-value:** 0.03011
  2. **T-statistics:** 1.90922
- Significance level is 0.05 and since **P-value < 0.05** it rejects the null hypothesis
- So, we have enough evidence to support the hypothesis.



```python
#t-test
from scipy.stats import ttest_ind

print(f"Samples – Travel Vlog: {len(vlog_likes)}, Pilates Fitness: {len(pilates_likes)}")

t_stat, p_value = ttest_ind(vlog_likes, pilates_likes, alternative="greater", equal_var=False)

print(f"T-statistic: {t_stat:.5f}")
print(f"P-value: {p_value:.5f}")

alpha = 0.05
if p_value < alpha:
    print("Reject the null hypothesis. Travel vlogging videos have significantly higher like counts.")
else:
    print("Fail to reject the null hypothesis. There is no significant evidence that travel vlogs get higher likes.")
```

```
Samples – Travel Vlog: 50, Pilates Fitness: 50
T-statistic: 1.90922
P-value: 0.03011
Reject the null hypothesis. Travel vlogging videos have significantly higher like counts.
```

Research Question & Hypothesis 3

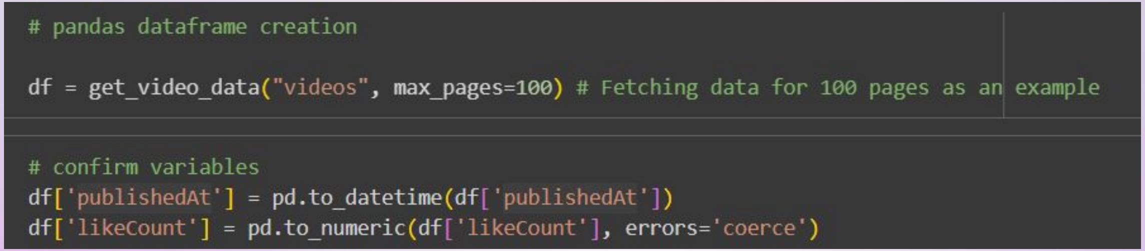
Research Question: Does upload timing affect the number of likes a video receives?

Hypothesis (null/H0): There is no difference in average like counts between videos uploaded on weekdays and weekends.

Hypothesis (alternative/H1): Videos uploaded during the weekend get higher average like counts in comparison to videos uploaded on weekdays.

Data Preparation (3)

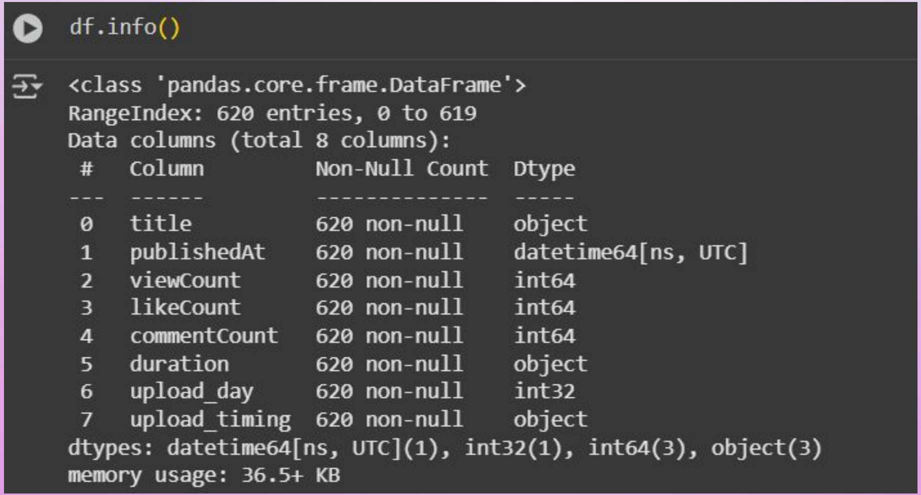
- Fetched 100 pages of data, and stored into a dataframe.
- Confirmed variables: ‘publishedAt’ and ‘likeCount’ to measure publish timing and interaction values.



```
# pandas dataframe creation

df = get_video_data("videos", max_pages=100) # Fetching data for 100 pages as an example

# confirm variables
df['publishedAt'] = pd.to_datetime(df['publishedAt'])
df['likeCount'] = pd.to_numeric(df['likeCount'], errors='coerce')
```



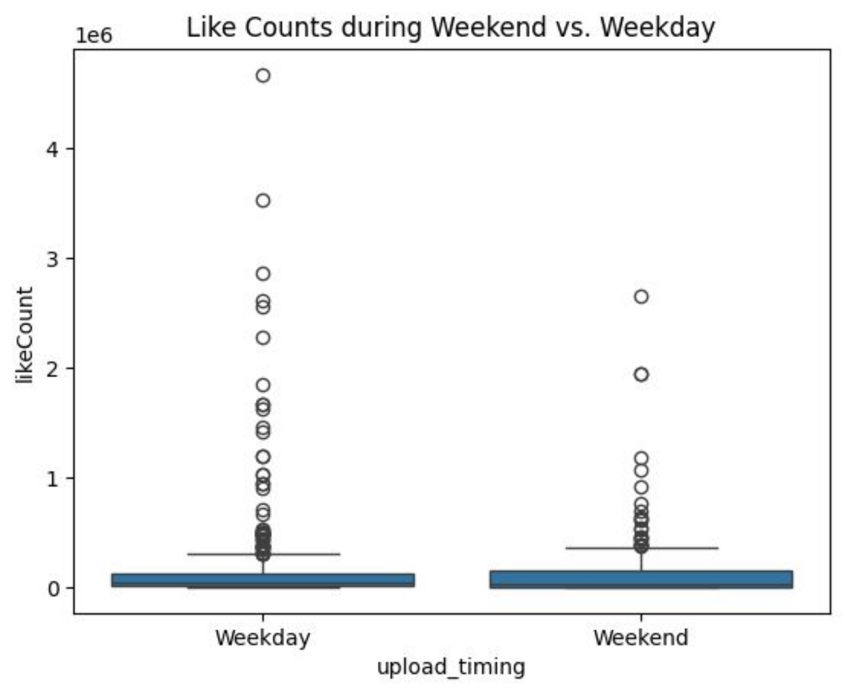
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 620 entries, 0 to 619
Data columns (total 8 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   title          620 non-null    object
 1   publishedAt    620 non-null    datetime64[ns, UTC]
 2   viewCount      620 non-null    int64
 3   likeCount      620 non-null    int64
 4   commentCount   620 non-null    int64
 5   duration       620 non-null    object
 6   upload_day     620 non-null    int32
 7   upload_timing  620 non-null    object
dtypes: datetime64[ns, UTC](1), int32(1), int64(3), object(3)
memory usage: 36.5+ KB
```

Visual Analysis (3)

Boxplot: Comparison of like counts and upload timing between the weekend and weekday.

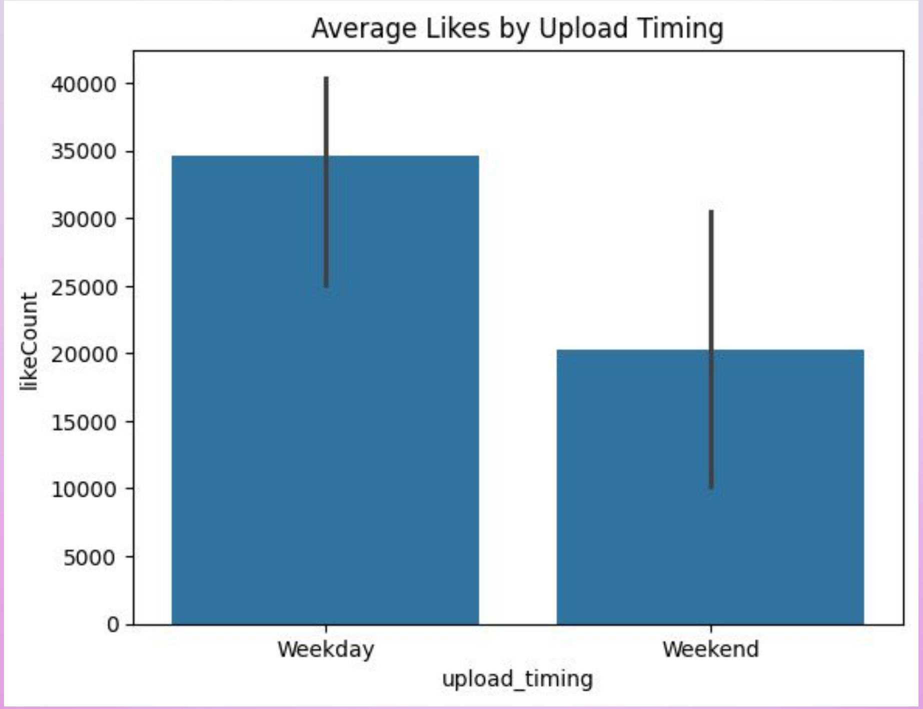
- **X-Axis:** upload timing (weekday vs. weekend)
- **Y-Axis:** like count (# likes per video)
- **Interpretation:**
  - Median like counts are about the same
  - Outliers are the “viral” videos
  - No significant differences between both groups
- **Conclusion:** Fail to reject the null hypothesis



Visual Analysis (3)

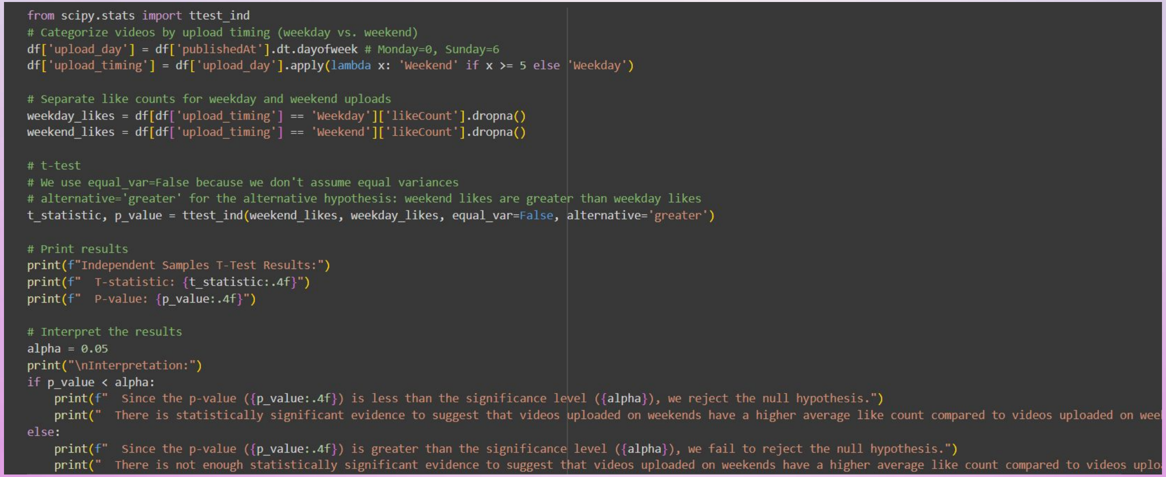
Barplot: comparing the average likes of videos uploaded on the weekday vs. the weekend.

- **X-Axis:** upload timing (weekday and weekend)
- **Y-Axis:** like count (avg. # of likes per video)
- **Weekday Avg:** ~ 35,000 likes
- **Weekend Avg:** ~ 20,000 likes
- **Interpretation:**
  - Black lines are 95% confidence intervals around mean (uncertainty)
  - Overlap of errors means the difference between weekend and weekday averages are not significantly different
  - Mean like count for weekday visibility higher, BUT visual evidence doesn't confirm significant difference.
  - Fail to reject null hypothesis



Statistical Test & Results (3)

- One-tailed t-test was executed
- **T-test =** -1.1089
- **P-value =** 0.8660
- Seeing as the p-value (0.8660) is greater than the alpha value of 0.05, meaning we fail to reject the null hypothesis.



```python
from scipy.stats import ttest_ind
# Categorize videos by upload timing (weekday vs. weekend)
df['upload_day'] = df['publishedAt'].dt.dayofweek # Monday=0, Sunday=6
df['upload_timing'] = df['upload_day'].apply(lambda x: 'Weekend' if x >= 5 else 'Weekday')

# Separate like counts for weekday and weekend uploads
weekday_likes = df[df['upload_timing'] == 'Weekday']['likeCount'].dropna()
weekend_likes = df[df['upload_timing'] == 'Weekend']['likeCount'].dropna()

# t-test
# We use equal_var=False because we don't assume equal variances
# alternative='greater' for the alternative hypothesis: weekend likes are greater than weekday likes
t_statistic, p_value = ttest_ind(weekend_likes, weekday_likes, equal_var=False, alternative='greater')

# Print results
print(f"Independent Samples T-Test Results:")
print(f"  T-statistic: {t_statistic:.4f}")
print(f"  P-value: {p_value:.4f}")

# Interpret the results
alpha = 0.05
print("\nInterpretation:")
if p_value < alpha:
    print(f"  Since the p-value ({p_value:.4f}) is less than the significance level ({alpha}), we reject the null hypothesis.")
    print("  There is statistically significant evidence to suggest that videos uploaded on weekends have a higher average like count compared to videos uploaded on wee
else:
    print(f"  Since the p-value ({p_value:.4f}) is greater than the significance level ({alpha}), we fail to reject the null hypothesis.")
    print("  There is not enough statistically significant evidence to suggest that videos uploaded on weekends have a higher average like count compared to videos uplo
```

Brief recap of Hypothesis 1, 2 and 3

Hypothesis 1: Classical music videos have higher average view counts than math tutorials.

Null Hypothesis: There is no significant difference between the two categories.

(Failed to reject the null hypothesis)

Hypothesis 2: Travel vlogging videos get higher average like counts than pilates fitness videos.

Null Hypothesis: There is no significant difference between the two categories

(rejects the null hypothesis)

Hypothesis 3: Videos uploaded during the weekend get higher average like counts in comparison to videos uploaded on weekdays.

Null Hypothesis : There is no difference in average like counts between videos uploaded on weekdays and weekends.

(Fail to reject the null hypothesis)

References

- Google. (n.d.). YouTube Data API. Google Developers. https://developers.google.com/youtube/v3