INTRODUCTION: Our project analyzes YouTube content to answer a key question: How do subscriber count and video characteristics impact video performance across different categories or watchers? Using data from the top 10 subscribed channels in categories like Music, Gaming, and Sports, we gathered insights from the 50 most popular videos in each channel via the YouTube API. By seeing how different categories of users engage with content, creators in those categories can find what drives a successful video. You can view the project details in our Google Slides presentation and access the full dataset and code on GitHub.

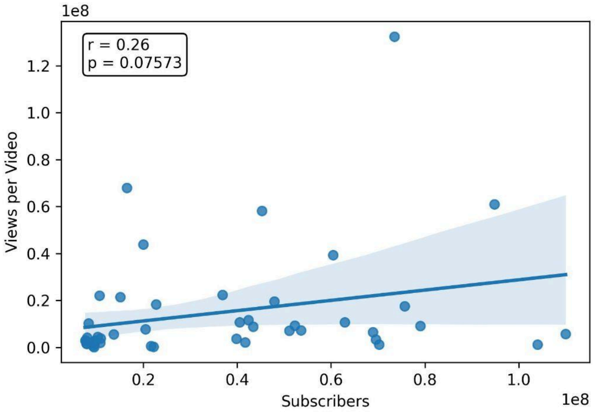
DATASET: The dataset for this analysis was built using a combination of SocialBlade and the YouTube Data API v3. We started by using SocialBlade to identify the top 10 subscribed channels in five categories: Music, Gaming, Auto & Vehicles, Sports, and Travel. For each of these channels, we searched their titles using the YouTube API to retrieve essential channel-level information, such as channel ID, total views, video count, subscriber count, and display name. Using the channel ID, we then fetched the 50 most popular videos from each channel, collecting detailed video-level data, including video ID, video title, likes, views, comment count, upload date, and video length. This structured dataset is well-suited for analysis as it contains both channel- and video-level attributes, enabling us to explore relationships between subscriber base, video performance, and engagement across categories. By leveraging the API, we ensured the dataset was up-to-date and comprehensive, making it ideal for gaining insights into YouTube content performance metrics.

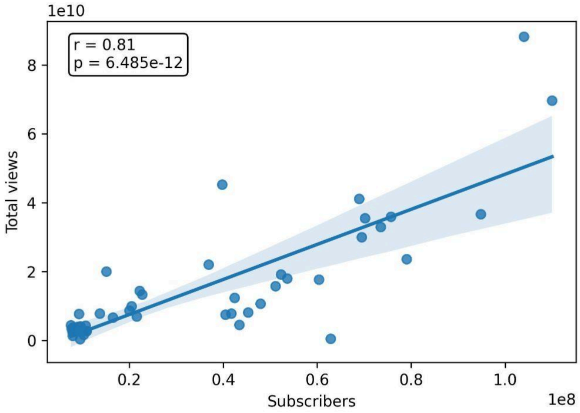
ANALYSIS TECHNIQUE: For this analysis, we employed bar plots to compare views per video and engagement (views per subscriber) across different categories, as these visualizations effectively highlight differences in performance between categories. Linear regression was used to explore the relationship between subscriber count and both views per video and total views. This technique was chosen because it helps quantify the strength of these relationships, providing insights into how channel size influences video performance. Regression is particularly suitable here since it allows us to determine if subscriber count can be a predictor of viewership trends, which is valuable for content creators looking to understand audience growth and engagement. (add about video-level insights as well)

RESULTS:

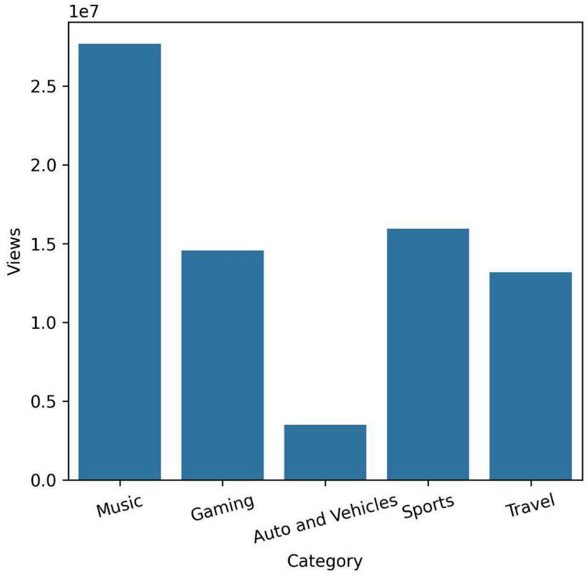
1. **Subscriber to Channel Performance:**
   The regression analysis results show different relationships between subscriber count and video performance. For views per video versus subscribers, the slope is 0.319, indicating a modest positive association, but the relationship is weak (r = 0.26) and not statistically significant (p = 0.076). The null hypothesis here states that there is no significant relationship between subscriber count and **views per video**, while the alternative hypothesis suggests that a significant relationship does exist. Since the p-value is greater than the common threshold of 0.05, we fail to reject the null hypothesis, meaning that we do not have enough evidence to conclude a statistically significant relationship between subscribers and views per video in this case. This suggests that having more subscribers doesn't always predict a significant increase in views per video. In contrast, for **total views** versus subscribers, the slope is

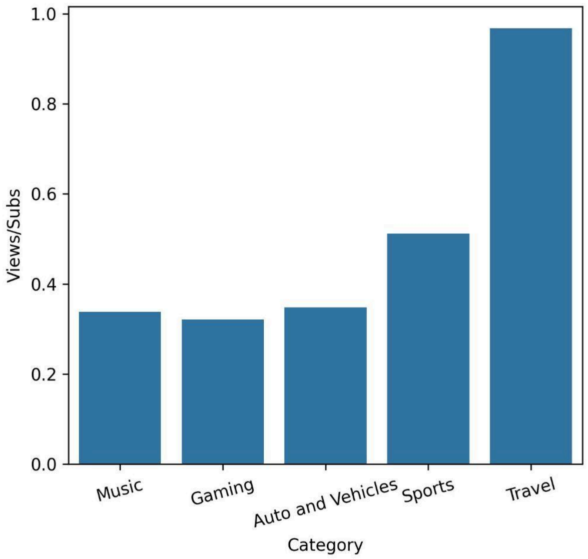
0.0013, and the relationship is much stronger (r = 0.81) with a highly significant p-value (p < 0.00001). This can be explained by algorithms being tuned to show what you will click on, and not what you are subscribed to. This means that people subscribe to channels that make videos they watch, they don’t watch channels they are subscribed to.





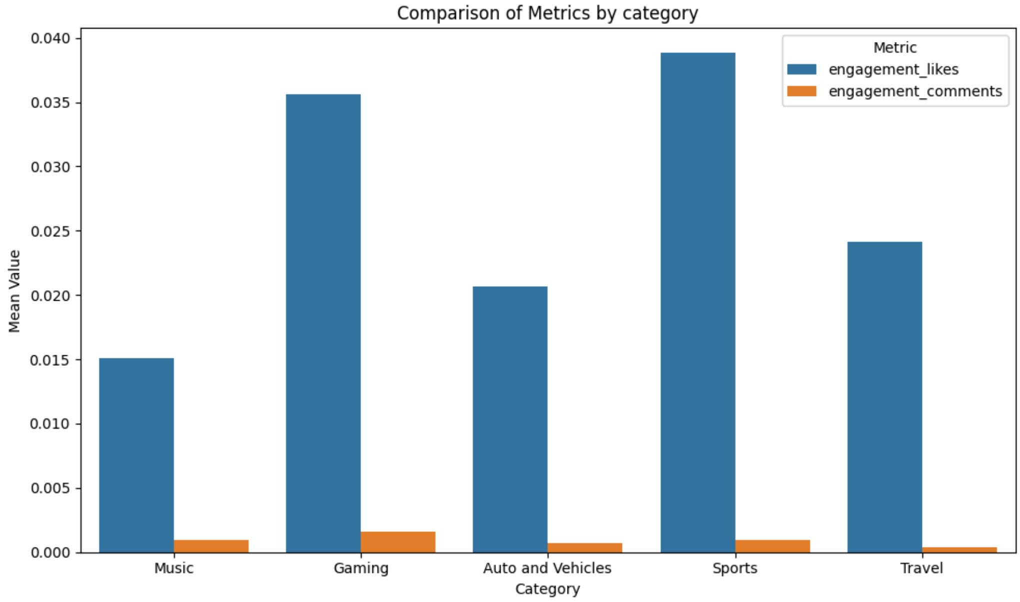
2. **YouTube Video Performance by Category**:
The analysis revealed significant differences in video performance across YouTube categories. For instance, the Music category recorded the highest average views per video, approximately 27.68 million. This can be attributed to the repeated plays of music videos, as listeners often return to their favorite tracks multiple times. In contrast, the Travel category exhibited the highest engagement ratio, with views per video per subscriber at 0.97. This indicates that a travel channel's subscriber count is a more important metric for how well a video will perform than for other categories.



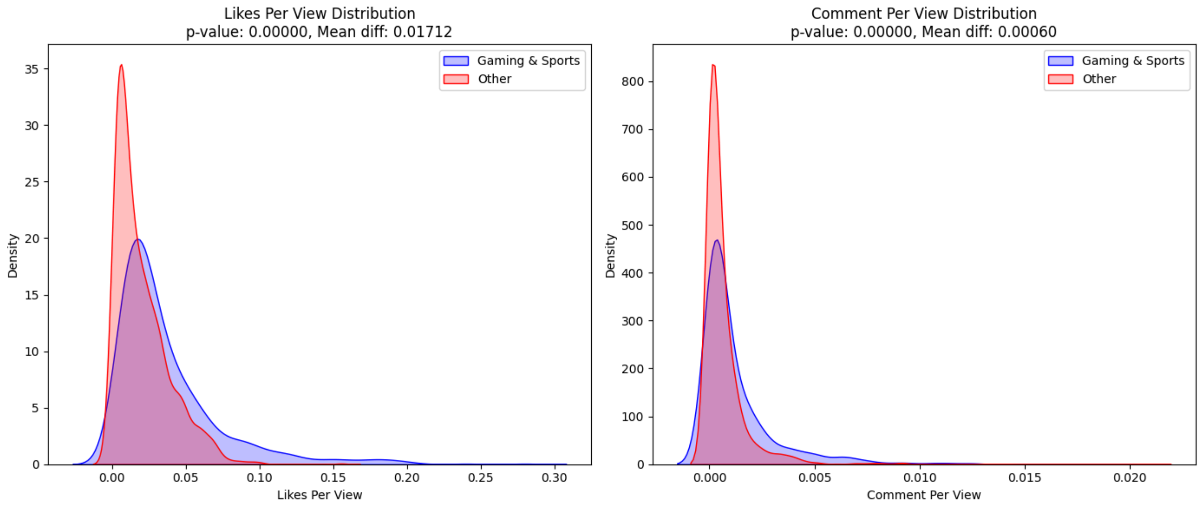


3. **Gaming and Sports Engagement:**
Plotting the comments per view and likes per view in a bar chart comparison shows that sports and gaming have a higher engagement per view then the other three

categories. To see if the difference between sports and gaming engagement and the other categories is truly significant a t-test was performed.

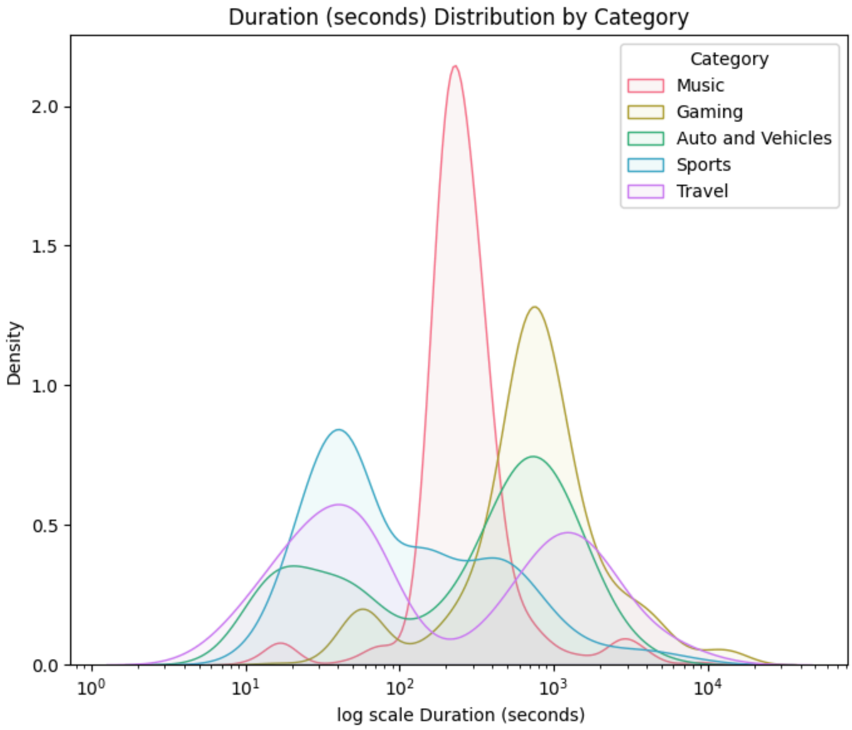


The t-test indicates that gaming and sports do have statistically significant but small changes in engagement. The small changes are across millions of views so the effect size is actually quite large even with small numbers, especially for likes. Interestingly travel does not seem to have as much engagement based on the high sub to view shown previously. Travel videos might be a more passive watching experience then gaming and sports.



4. **Categories and Video Duration:**
The most popular videos for each category duration reveals the lengths of videos users are watching for each category. The Music peak is between 2-4 minutes like an average song. Most other categories have two peaks where one is under a minute because of the short form content on youtube that is separated from the normal videos. Gaming seems to be the category that has the least amount of videos in the shorts category and the most in the 10-20 minute range. Most other categories have a similar peak in the same range. Sports does not have a peak in the 10-20 min range, only having a peak in short form content. This means despite gaming and

sports having similar high engagement because of stronger communities the style of content is very different



TECHNICAL:

The dataset required significant formatting and cleaning to prepare it for analysis. We removed outliers in view counts, handled missing data, and ensured data type consistency across all columns. One major challenge involved retrieving channel IDs using the `search().get_list()` method from the YouTube Data API. Fetching channel statistics required 100 tokens per search, with a daily limit of 10,000 tokens. This restriction made data collection time-consuming, as it took several days to aggregate the necessary data for our project. The costly nature of API usage made efficient token management critical throughout the process.

We employed linear regression due to its simplicity and effectiveness in modeling relationships between continuous variables, such as subscriber count and video performance metrics. This approach helped us evaluate how the size of a channel can affect video success. Bar plots were used to provide a clear visual comparison across different categories, highlighting differences in views per video and engagement. We also used t-tests and distribution plots to compare performance metrics between categories, helping us assess if any category's performance was significantly better than others. These techniques provided deeper insights into how certain content types engage audiences more effectively than others.

Initially, we explored simple correlations between channel size and views. However, total views were heavily skewed by a few large channels, leading us to focus on views per video instead. This adjustment allowed for a more balanced comparison. Regression analysis quantified the relationships between subscriber count and views. We also calculated audience engagement by measuring the liking and commenting rates (like_count/views and comment_count/views). We found that the

Sports and Gaming categories had higher engagement rates, particularly due to the interactive nature of live-streamed events and gaming content, where viewers are more likely to leave comments and likes during or after a live session. To further investigate this, we applied a t-test to determine if the engagement rates in these two categories were significantly higher than those in other categories, developing new analysis around audience interaction. We also evaluated the correlation between video length and views, but due to the high variability in views for videos of similar lengths, the correlation was weak, and no meaningful conclusions were drawn. Additionally, we examined how videos uploaded during the pandemic (2020-2021) performed compared to non-pandemic times. We observed significantly higher views and engagement during the pandemic, but since this analysis fell outside the project's main focus, it was ultimately excluded from the final results.