**Introduction**

This project will examine a sector of the online social network industry, YouTube. Data has been collected on trending videos. According to YouTube, its trending videos are calculated using an algorithm that examines view count, rate of growth in views, the age of the video, the place the views come from, and what it sees as relevant and appealing to a broad audience. The data collected about these videos include statistics about those videos, including views, likes, both upload and trending dates, comments, and other relevant data points to the data. In this project, I will be examining how each of the independent variables has an effect on the dependent variable, which is amount of views. To narrow down the sample size and to avoid the language barriers of other countries, I have limited the data to the United States. In analyzing this data, it could further be used for marketing purposes in the future in the sense of: what do consumers like and/or have interest in? Then, businesses can demographics accordingly. More and more people are consuming online media, and how can businesses adapt to change our strategies, target markets, and influencer strategies? And how should a business create a presence online, including YouTube? Is there an opportunity there or an opportunity to market there? Descriptions of the Variables are as follows:

|  |  |
| --- | --- |
| **Variable** | **Description** |
| trending\_date | when the video went trending |
| title | what the video title was--does title matter? |
| channel\_title | what channel the video was on--does channel matter? |
| category\_name | video category--does category matter? |
| publish\_time | when the video was published  |
| views | how many views?---does view matter? |
| likes | How many likes and dislikes?, and does that play a role? |
| dislikes |
| comment\_count | how many comments does the video have?--do comments affect views? |
| like\_ratio | likes per dislike |

**Descriptive Statistics**

|  |  |  |
| --- | --- | --- |
| **Variable** | **Mean**  | **Standard Deviation**  |
| views | 1315324.473 | 5690513.255 |
| likes | 43513.799 | 162874.783 |
| dislikes | 3101.528 | 43712.702 |
| comment count | 5443.928 | 30013.460 |
| like ratio | 45.091 | 65.076 |

On average, a trending video has 1,315,324 views with 45,514 likes, 3102 dislikes, and 5444 comments. There are 45 likes per dislike. The standard deviations, showing how far away from the mean a value is on average, are also shown in the table above.

|  |
| --- |
| **Mode: Channel with Most Trending Videos** |
| Refinery29 | 16 |
| Vox | 16 |

Here are the two channels that are tied for the mode amount of trending videos. In the chart beside it, the number of trending videos for an account is summarized by the number of channels that have that amount of trending videos. As evidenced above, it is clear that there are many more unique channels than reoccurring channels, so if a brand is trying to work with influencers, it is best to build a larger network.

|  |  |
| --- | --- |
|  |  |
| Average Title Word Count | 8.82 |
| Maximum Title Word Count | 23 |
| Minimum Title Word Count | 2 |
| Mode  | 7 |
| Range | 21 |
| Standard Deviation  | 3.46 |

Above are the categories that have the most trending videos, as well as descriptive statistics about title word count. If a business were to market on YouTube, it would be wise to think about title word count and phrasing. Additionally, a company should align itself with influencers in the proper categories that resonate with the brand, but it should also note what kinds of videos have gone trending recently and see what it can do to market to that sector as well.



Every publishing time is unique, so we see a uniform distribution of data there.

**Hypothesis Testing**

For my one sample tests, I conducted tests for the like ratio and the word count for the title. The results were as follows:

The null hypothesis for the like ratio was that the population mean would be less than or equal to 45 likes per dislike. Likewise, that means the alternative hypothesis was that the population mean for the like ratio would be greater than 45. The p-value returned was 0.52, so the null hypothesis is rejected and it can be concluded that the average like ratio is at least 45. The alpha chosen for this was .05, which means that this can be said with 95% confidence. This would make sense for it to be a larger number, since a trending video should get good ratings overall.

The null hypothesis for the title word count was that the population mean would be at least 10 words. So, the alternative hypothesis was that the population mean would be less than 10 words. The p-value returned was 1, which is greater than the alpha of .05 so we fail to reject the null hypothesis and conclude with 95% confidence that the mean title word count for the population is at least 10 words.

For the two sample tests, likes and dislikes were compared, and view count and comments were compared.

 The null hypothesis for likes and dislikes was that mu1-mu2 would be less than or equal to zero, while the alternative was that that mu1-mu2 would be greater than zero. Mu1 denotes likes, while mu2 denotes dislikes. The p-value generated was 3.25x10^-20, less than the chosen alpha of .05. With this, it can be concluded with 95% confidence that the mean likes are greater than or equal to the mean dislikes. It can be inferred that if a video is trending, it is generally generating interest, and therefore should have a higher like count.

 The null hypothesis for views and comments was that mu1-mu2 would be less than or equal to zero, where mu1 is views and mu2 is comments. The alternative hypothesis, then, was that mu1-mu2 would be greater than zero. The p-value returned 8.99x10^-19, so that is less than the alpha of .05. With 95% confidence, then, we reject the null and conclude that the mean view count is greater than the mean comment count. This would make sense, because not everyone comments on a video.

For ANOVA, I compared my data on views across the corresponding categories.

 The null hypothesis was that the means for each category would be equal, while the alternative was that the means would not be equal. The p-value of 5.91x10^-7 is less than the chosen alpha of .05, so it can be concluded with 95% confidence that the mean views differ across categories.

The tests of association that I think would be helpful in this project would be testing the publishing time and the date as well as channels. I think that that would be a more effective way to test these variables than performing ANOVA tests or one or two sample hypothesis tests. In particular, I think that a test of independence would be helpful in this case. These variables are more categorical variables, so a test of independence would produce better results. It would need some research into proportions first, but then it would be easy to test the publishing time and the date published with a test of independence.

**Conclusion**

When analyzing opportunities to market on YouTube, be sure to go over these key factors. Look at who is putting out trending content, and extend offers to partner with them. Also observe who may be likely to put out such content in the future so that you can capitalize on their success. Influencer marketing and YouTube in general can be a great marketing strategy, if it is done correctly and with these aspects in mind.

**Resources**

https://www.kaggle.com/cloudchaoszero/youtubedatanov2017dec2018cleaned/data

http://www.sanfordbrown.edu/Student-Life/blog/February-2015/Streaming-Media-Industry-Trends-and-Opportunities

https://support.google.com/youtube/answer/7239739?hl=en

<https://www.birdsonganalytics.com/guides/youtube-title-description-length/>

**Variables Used**

Views

Comments

Like Ratio

Likes

Dislikes

Title Word Count

Tests of Association: trending date and time, channels