Executive Summary

Preliminary Data Analysis

ISSUE / PROBLEM

Tiktok has a problem, that too many resources are exerted in censoring content worth censoring. Our goal is to develop a machine learning algorithm that will classify tiktoks as either claims or opinions to make that process more efficient.

RESPONSE

Our first step is to analyze the dataset to get a preliminary understanding of the relationships between the variables. The dataset has a column called claim_classification, and we have printed the counts of each of these variables to understand the workload of the algorithm.

IMPACT

By analyzing the relationships between variables like video_duration, video_view_count, and video_like_count, we would be able to establish which variables to focus for the predictive model.

Specific Actions:

We grouped the data by the claim_status variable and used the count() function to learn about the proportion of tiktok types.

```
claim      9608
opinion    9476
```

On the data grouped by claim_status, we decided to print the mean and median of the view_count variable for claims and opinions. This gives us an idea of how these different tiktok types impact engagement.

Claims

```
Mean view count claims: 501029.4527
Median view count claims: 501555.0
```

Opinions:

```
Mean view count opinions: 4956.43224
Median view count opinions: 4953.0
```

We found that claims generated a lot more engagement than opinions did.

KEY INSIGHTS

For future analysis it is important to register there are an equal number of claim and opinion tiktoks. If it is the case that one category has a significantly larger need for censorship, we will be able to more efficiently write the algorithm to target this type.

Additionally, we found that claims have much higher engagement with Tiktok's audience. This could be indicative of a more clickbait type content implying a more efficient use of our resources would target the claim tiktok type.