Executive Summary

Preliminary Data Analysis

ISSUE / PROBLEM

Too many worker hours are exerted on reviewing problematic content. 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

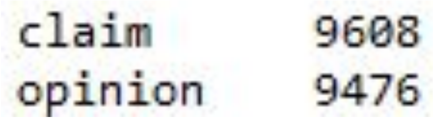
Our first step is to analyze the dataset to get a preliminary understanding of the relationships between the variables. The dataset has a column called claim_status, and we have printed the counts of each of the 2 classes for this variable to understand the workload of the algorithm.

IMPACT

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

Specific Actions:

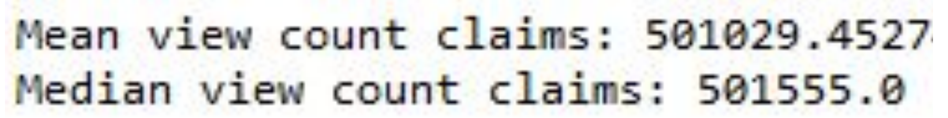
We grouped the data by the claim_status variable and used the count() function to learn about the proportion of tiktok types.



```
claim      9608
opinion    9476
```

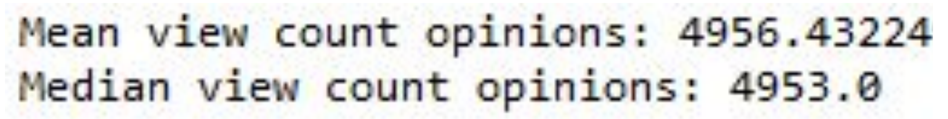
On the data, grouped by claim_status, we decided to print the mean and median of the view_count variable for claims and opinions. This gives us an idea of how these different tiktok types impact engagement.

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 is a roughly equal number of claim and opinion tiktoks. If it was the case that one category had a significantly larger count, we potentially would've had to adjust the data to ensure optimal model classification. If one class has too few data points, that could skew the model's performance.

Additionally, we found that claims have much higher engagement, in particular for video_view_count, with TikTok's audience. This is understandable from a human interest point of view, and also indicates that these video engagement features, like view/like/share counts, might be the most important for our model to use.