Executive Summary

Milestone 3 of the TikTok Claims Classification Project

ISSUE / PROBLEM

TikTok's data team is building an ML model to classify user-generated claims. Current focus is on performing EDA to identify distinguishing features between videos with claims and those with opinions.

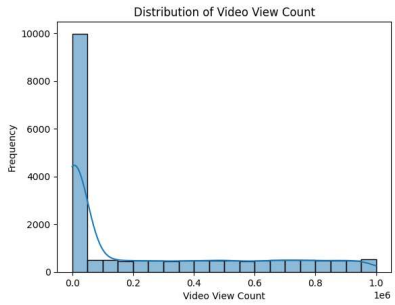
RESPONSE

This stage involved exploratory data analysis (EDA) to understand the impact of TikTok videos on users. User engagement metrics were investigated, including likes, shares, comments, and views.

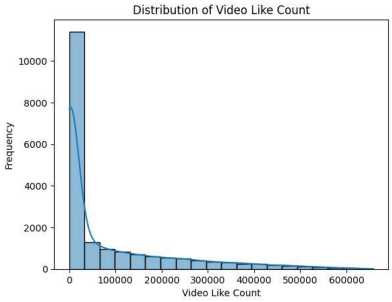
IMPACT

Significant class imbalance between opinions and claims in the video dataset will need to be investigated and addressed for the future claims classification model development.

Visualizing the data is a key component, and in the following Histograms it can be seen that the majority of videos are grouped at the lower-end of the range of values for the four video count variables.

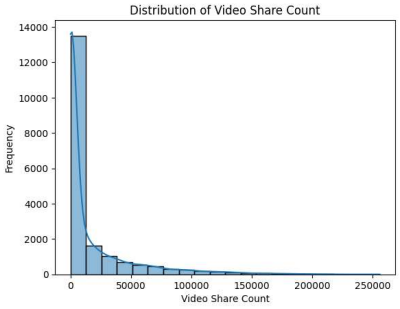


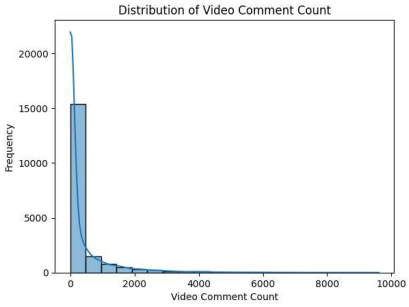
The majority of videos have received less than 100,000 views. The distribution over 100,000 views, however, is uniform.



Similar to views, videos with less than 100,000 likes make up the majority of videos in the data.

Shares and comments also follow a similar distribution, being highly right-skewed.





KEY INSIGHTS

Going forward, the data team will need to address the significant right skew of the count variables, the class imbalance between opinions and claims, and the missing values.

Missing Values: ~200 missing values in 7 columns were identified. These will have to be considered for future modeling as well as investigation into the cause of the missing values for statistical analysis and model development.

Skewed Data Distribution: The video count values are very right-skewed, and this will need to be taken into account during future model development.