TikTok Claims Classification Project

Exploratory Data Analysis (EDA) - Executive Summary

ISSUE / PROBLEM

The TikTok data team aims to develop a machine learning model to accurately classify user-submitted videos as either 'claims' or 'opinions.' Prior to model development, a thorough exploratory data analysis (EDA) is essential to understand the dataset's characteristics, identify potential issues, and inform the modeling process. This includes understanding user engagement patterns and addressing data quality concerns.

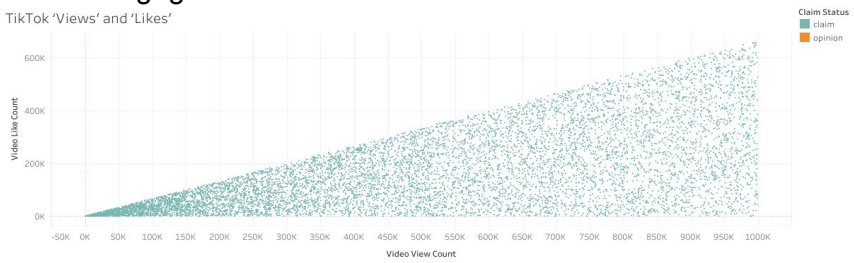
RESPONSE

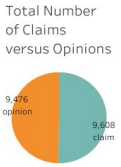
The TikTok data team conducted a comprehensive EDA, focusing on key variables that reflect user engagement: video view count, like count, and comment count. This analysis involved data cleaning, visualization, and statistical exploration to reveal underlying patterns and potential challenges.

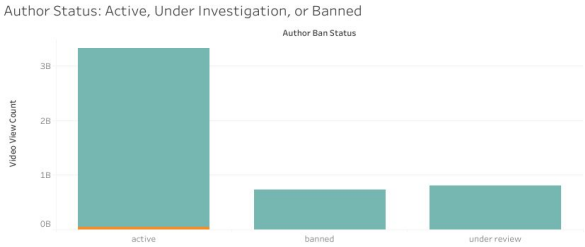
IMPACT

The EDA revealed significant insights into the distribution of user engagement metrics, highlighting the need to address data skewness and missing values. These findings will directly impact the design and performance of the future claims classification model, ensuring its robustness and accuracy.

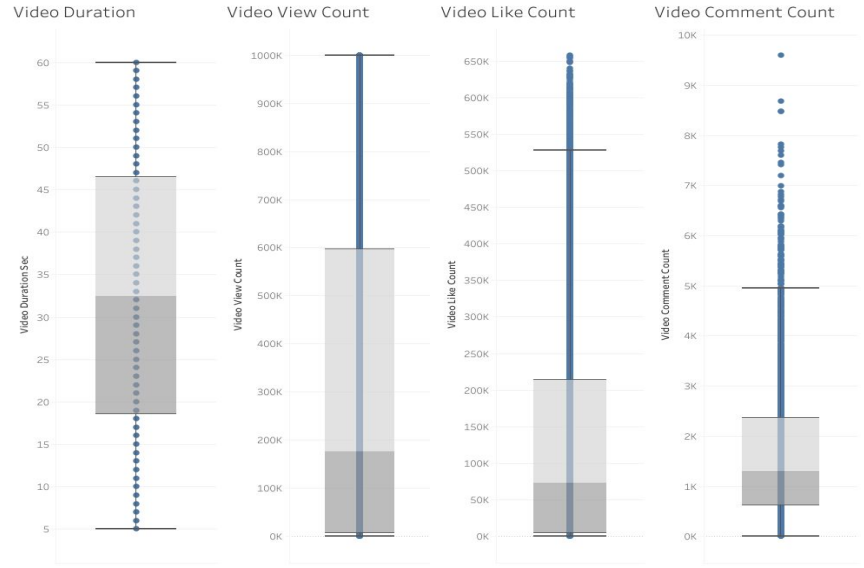
- **Views/Likes:** "Strong positive correlation."
- **Claims/Opinions:** "Balanced distribution."
- **Author/Views:** "Active authors: highest views, sole opinion posters, low engagement.







Boxplots show average video engagement: 32s duration, 190k views, 75k likes, 1.2k comments, providing a baseline.



KEY INSIGHTS

Skewed Data Distribution:

- Video view, like, and comment counts exhibit a strong right-skewed distribution, with the majority of videos concentrated at the lower end of the range. Specifically, over half of the videos received fewer than 100,000 views, and most videos had fewer than 100 comments. This skewness will necessitate careful consideration of model selection and data transformation techniques.
- The distribution of views above 100,000 is relatively uniform, while the distribution of likes tapers off, showing many videos with very high like counts.

Missing Values:

Over 200 null values were identified across seven columns. This necessitates a strategic approach to handling missing data during model development to avoid biased results. Further investigation is required to determine the root cause of these null values and their potential impact on subsequent analyses.

Imbalanced Opinion Video Counts:

The model will need to account for the imbalance in opinion video counts by incorporating them into the model parameters.