Title

Subtitle

ISSUE / PROBLEM

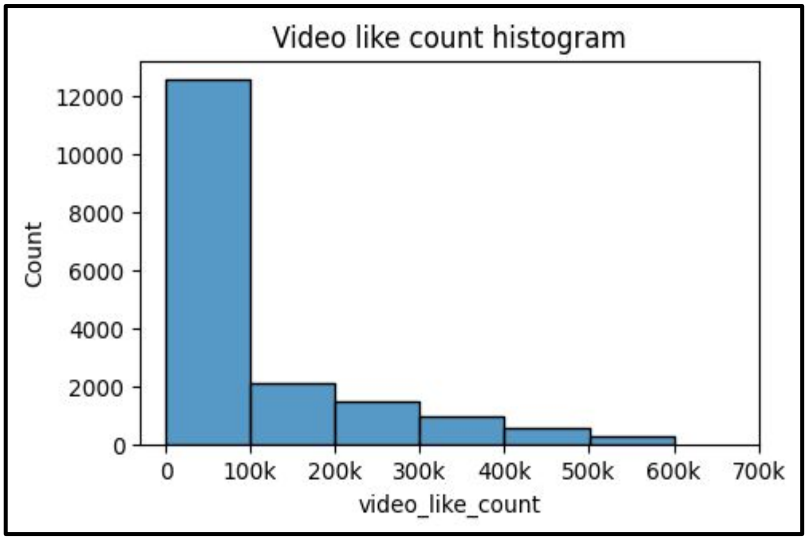
TikTok leadership has tasked the data team with developing a machine learning model to classify user content. This summary highlights Saswat Seth's preprocessing efforts, including continued EDA with Python and visualizations in Tableau, which lay the foundation for the subsequent model development.

RESPONSE

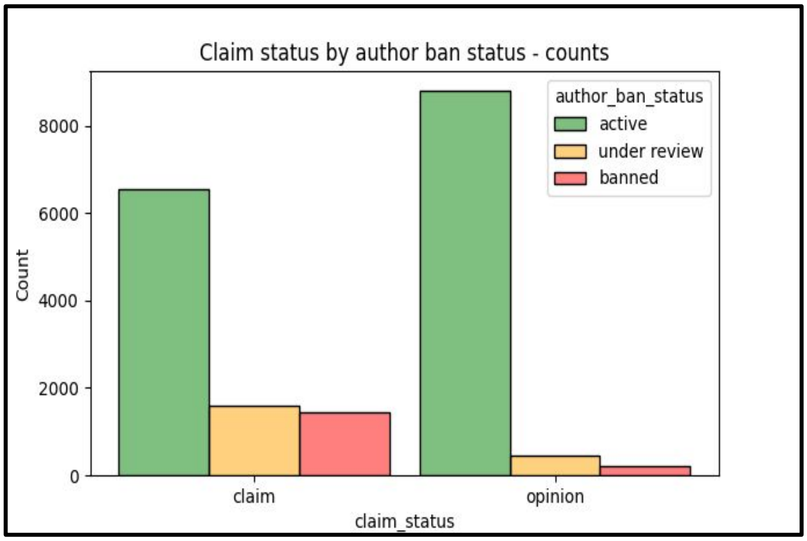
During the discovery phase, summary statistics were calculated, and the data was structured through sorting, grouping, and other techniques to uncover meaningful insights. Iterative validation steps ensured the integrity of the data. Visualizations were created to detect outliers, examine distributions, and derive key insights.

IMPACT

The dataset has been thoroughly analyzed, with key statistics and measures identified to guide machine learning model selection. Validation has been performed and will continue iteratively to ensure data accuracy. Visualizations have uncovered patterns, trends, and outliers, providing critical insights and preparing the data for model development.



This histogram shows video engagement distribution, with video count (Y-axis) against engagement levels (X-axis). The metrics are right-skewed, tapering towards higher values, and like view count, most videos have fewer than 100,000 likes.



This bar chart shows the distribution of content classified as either a claim or an opinion, grouped by user ban status. It highlights that banned users are more prevalent in the "claim" category, suggesting a potential link between user status and content type.

KEY INSIGHTS

- A strong correlation exists between claim status and video view count, suggesting that claim videos generally attract more views, which requires further exploration.
- Similar patterns are observed between claim status and other engagement metrics, including likes and shares, indicating that claim videos generate higher interaction.
- Claim-status videos are predominantly posted by banned users, while opinion-status videos show a different user behavior.
- Engagement metrics, as visualized through histograms, exhibit a right-skewed distribution. Only a small fraction of videos garner substantial engagement, while the majority consistently fall within lower engagement ranges.