# TikTok Claims Classification Project

## Results of Exploratory Data Analysis (EDA)

*All code at https://github.com/izsolnay/TikTok_Python*

## ISSUE / PROBLEM

TikTok aims to efficiently prioritize claim reports

**Objective**: develop a reliable machine learning model which effectively classifies claim reports to streamline their processing

**Steps**: organize, analyze, explore, and structure data for model building

## RESPONSE

After preparing the data, in order to understand statuses on user engagement, the team:

- Focused on the key variable **`‘claim_status’`**, targeting relationships between its two labels: Claims and Opinions and the variable **`‘verified_status’`**
- Quantified users’ downloads, likes, views, shares, and comments by both **`‘claim_status’`**, and the variable **`‘author_ban_status’`**

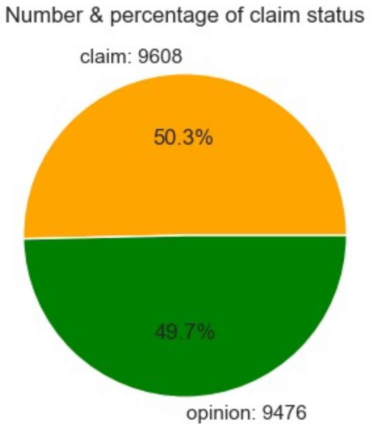
## IMPACT

Results of the EDA indicate a severe imbalance in user engagement with Claims vs Opinion videos. Next steps for model building:

- Address outliers
- Address the unbalanced data set
- Investigate text of comments

## UNDERSTANDING THE AUTHOR

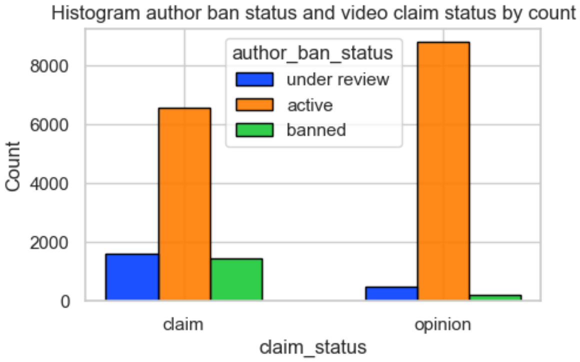
Videos labelled Claims or Opinions are evenly divided in the data set.



However, through investigation of the variables **`‘verified_status’`** and **`‘author_ban_status’`** the data team uncovered that of all video authors:

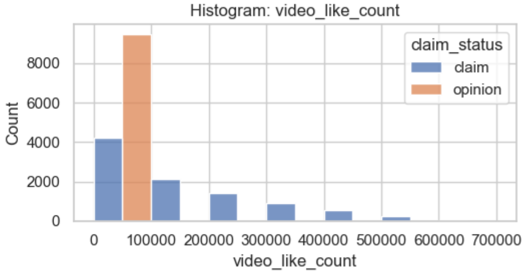
- ~94% are Not verified
- ~11% are Under review
- ~9% are Banned

- More Opinion videos are posted by authors with Active status (~1500 more)
- Claims videos are ~90% more likely to be posted by Banned authors and ~80% more likely to be posted by Under review authors



## UNDERSTANDING THE USER

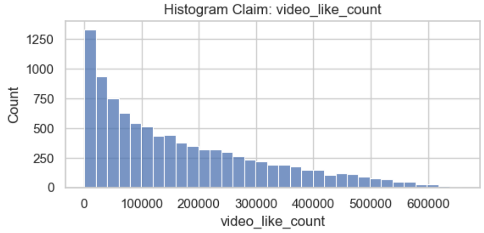
For all author statuses, engagement with Claim videos account for nearly 100% of all calculated medians of engagement across all categories of engagement. Example:

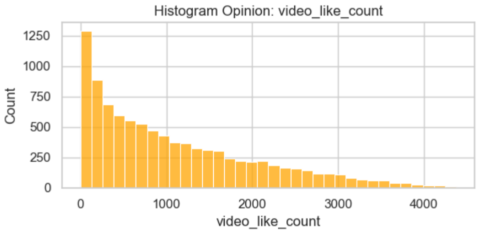


## KEY INSIGHTS

The EDA performed by the data team revealed a heavily skewed data set with Claims videos receiving far more engagement than Opinion videos.

The team also discovered wide discrepancies in the engagement levels between videos, regardless of Claim or Opinion status. These distributions were similar between Claim and Opinion videos:


Histogram Claim: video_like_count
Count
video_like_count


Histogram Opinion: video_like_count
Count
video_like_count