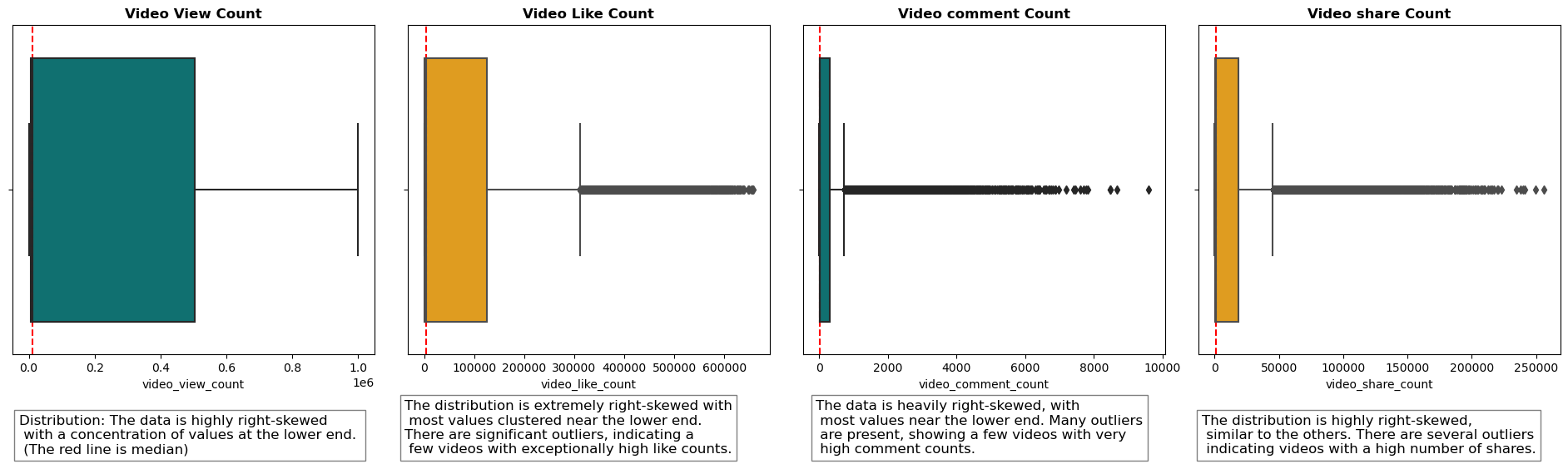
# Executive Summary (Task 2)

## Overview

The TikTok data team seeks to develop a machine learning model to assist in the classification of claims for user submissions. To begin, the data team needs to organize the raw dataset and prepare it for future exploratory data analysis. In this preliminary investigation, I aim to summarize column Data types, data value nonnull counts, relevant and irrelevant columns and understand the data distribution.

## Understand the Data

* The provided dataset is in long format, with **19382 rows and 12 columns**
* Most of the numeric columns are in **integer or float datatypes**; however, claim\_status, verified\_status, and author\_ban\_status are currently in object types and they should be boolean masked in order to build our predictive model.
* The dataset contains **298** rows with null values. I consider to drop these as the null values appeared simultaneously to many key columns and our sample size is large enough to yield an accurate result. However, further investigation is necessary to find the reasons for these nulls.
* There are **no duplicated values** found in the dataset.
* The counts of each claim status are quite balanced. There are **9,608 claims and 9,476 opinions**.
* View, like, comment, share count are heavily right skewed, suggesting there are some significant outlier videos with significant audience engagement. Therefore, the data distribution is **right skewed**, which will inform the types of models that we will build.



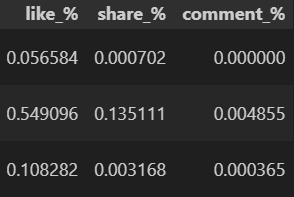
## Preliminary findings

This section lays the foundation understanding of the relationship between variables.

1. Claim videos have more views: By grouping video claim status on their mean and median video view count, I discovered that the mean and median number of views for 'claim' is higher than 'opinion'. In other words, people watch more videos identified as 'claim' rather than 'opinion'.

2. Variables like author ban status, verified status, and video view count are important for our predictive model

3. I created three extra columns to calculate proportion of like, comment, share per view. These new columns may help to further derive deeper insights.



## The next step

In summary, I believe that video view count, author ban status, share count may be some key independent variables for our predictive model. However, we might discover more relevant variables as we progression to later stages of the project. With the key variables identified and the initial investigation of the claims classification dataset, the process of exploratory data analysis can begin.