Executive Summary

TikTok Claims Classification Project

ISSUE / PROBLEM

The TikTok data team is working on creating a machine learning model to help categorize claims from user submissions. In this phase of the project, it is essential to examine, explore, tidy up, and organize the data before proceeding with the development of the model.

RESPONSE

The data team conducted an initial examination of the claims classification dataset, aiming to uncover significant relationships between variables

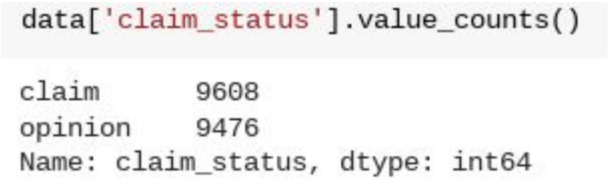
In response to the request for classifying user claims, the team examined the frequencies of claims and opinions to gain insights into the distribution of each type of video content.

IMPACT

The influence of this initial analysis will become apparent in the subsequent stages. To gauge the impact of user videos, the data team pinpointed two crucial variables for consideration: video duration (in seconds) and video view count. Both of these variables are significant factors to take into account for future prediction models.

UNDERSTANDING THE DATA

Upon reviewing the supplied dataset, the variable "claim_status" emerged as notably valuable for the client's proposed project. The subsequent screenshots depict critical points of analysis necessary to comprehend the "claim_status" variable.



```
data['claim_status'].value_counts()
```

```
claim      9608
opinion    9476
Name: claim_status, dtype: int64
```

Note: The distribution of each claim status is relatively even, with 9,608 instances of claims and 9,476 instances of opinions.

ENGAGEMENT TRENDS

The data team examined viewer engagement with videos in both the claim and opinion categories. To assess viewer engagement, the team focused on the view count. The mean and median view counts were analyzed to gauge the impact of each video category. Specifically, the mean and median view counts for both claim and opinion categories reveal the correlation between content type (claim or opinion) and the number of video views.

Claims:

```
Mean view count claims: 501029.45
Median view count claims: 501555.0
```

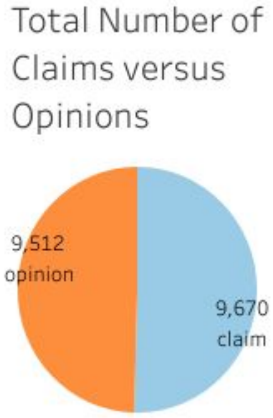
Opinions:

```
Mean view count opinions: 4956.43
Median view count opinions: 4953.0
```

KEY INSIGHTS

- There is a nearly equal distribution between opinions and claims. With this awareness, we can advance our future analysis confidently, recognizing that the dataset contains a relatively even number of both claims and opinions in the included videos
- Having identified the key variables and conducted the initial examination of the claims classification dataset, the exploratory data analysis process is ready to begin.

Pie chart visualizes the comparison of the count of claims and opinions


Total Number of Claims versus Opinions
opinion
claim