Executive Summary

ISSUE / PROBLEM

The TikTok data team is in the early stages of developing a machine learning model aimed at classifying claims made in user-submitted videos. Before moving forward with model development, the raw dataset must be thoroughly organized and prepared to ensure it is suitable for effective exploratory data analysis (EDA).

RESPONSE

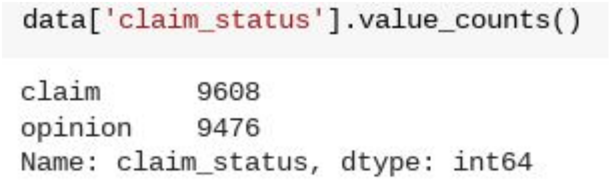
The data team conducted a preliminary analysis of the claims classification dataset to identify key relationships between variables. Given the objective of classifying user-submitted claims, the team first examined the distribution of claim types by evaluating the counts of both “claim” and “opinion” labels. This provided an initial understanding of the dataset’s composition and the relative frequency of each content type.

IMPACT

The impact of this preliminary analysis lays the foundation for the next phase of the project. To better understand user engagement and content impact, the data team identified two key variables for further exploration: `video_duration_sec` and `video_view_count`. These variables are likely to play a significant role in shaping future predictive models, especially in assessing how video characteristics influence viewership and classification outcomes.

UNDERSTANDING THE DATA

Upon reviewing the dataset, the variable `claim_status` emerged as particularly valuable for the project’s objective. Since the goal is to classify whether content includes a claim or an opinion, this variable will likely serve as the target for future predictive modeling. The following screenshots highlight key analyses conducted to better understand the distribution and characteristics of the `claim_status` variable.



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

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

Note: The counts of each claim status are quite balanced. There are 9,608 claims and 9,476 opinions.

ENGAGEMENT TRENDS

To evaluate viewer engagement with the two types of video content—claims and opinions—the data team analyzed the `video_view_count` variable. Both the mean and median view counts reveal a significant disparity between the two categories:

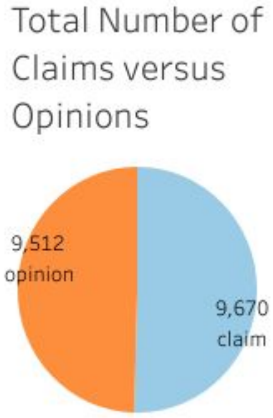
- **Claim Videos:**
  - *Mean view count:* 501,029
  - *Median view count:* 501,555
- **Opinion Videos:**
  - *Mean view count:* 4,956
  - *Median view count:* 4,953

This substantial difference suggests that videos categorized as claims generate significantly more engagement than opinions. These findings will help guide the development of a predictive model and support TikTok’s efforts to prioritize content moderation more effectively.

KEY INSIGHTS

The dataset contains a nearly equal distribution of opinions and claims—9,476 opinions and 9,608 claims. This balanced split ensures a solid foundation for unbiased model development. With essential variables like `claim_status`, `video_view_count`, and `video_duration_sec` identified, the team is well-positioned to move forward with exploratory data analysis (EDA) to uncover deeper patterns and inform the design of the future classification model.

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


Total Number of Claims versus Opinions
opinion
claim