Executive Summary

Milestone 2 of the TikTok Claims Classification Project

ISSUE / PROBLEM

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 will organise the raw data to prepare it for exploratory data analysis. (EDA)

RESPONSE

The data team performed a preliminary investigation of the claims classification dataset with the aim of learning important relationships between variables.

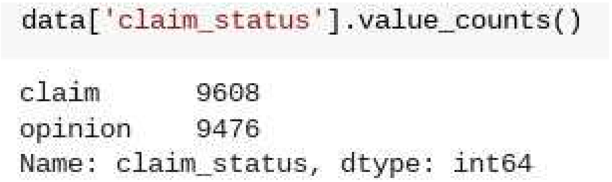
The data team looked at the counts of claims and opinions in order to understand the count of each type of video content.

IMPACT

The impact will be evident in the next stage. In order to understand the impact of user videos, the data team identified two important variables to consider. The variables video_duration (in seconds) and video_view_count are both important factors to consider for future prediction models.

UNDERSTANDING THE DATA

After reviewing the provided dataset, the variable claim_status seemed particularly useful, given the client's proposed project. The screenshots show important points of analysis required to understand the claim_status variable.



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

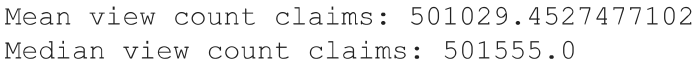
```
claim      9608
opinion    9476
Name: claim_status, dtype: int64
```

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

ENGAGEMENT TRENDS

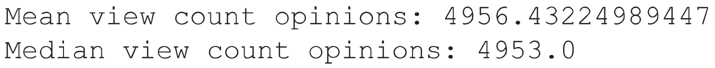
The data team considered viewer engagement with each video in the claim and opinion categories. In order to understand viewer engagement, the data team considered the view count. The mean and median view count show the impact of each category of video; specifically, the mean and median view counts for both categories show the association between content (claim or opinion) and the video views.

Claims:



```
Mean view count claims: 501029.4527477102
Median view count claims: 501555.0
```

Opinions:

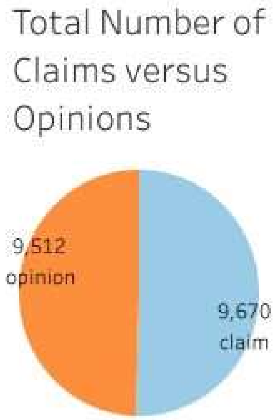


```
Mean view count opinions: 4956.43224989447
Median view count opinions: 4953.0
```

KEY INSIGHTS

- There is a near equal balance of opinions versus claims. With this understanding, we can proceed with our future analysis knowing that there is a fairly balanced amount of claims and opinions for the videos included within this dataset.
- With the initial investigation of the claims classification dataset and key variables identified, the process of exploratory data analysis can begin.

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


Total Number of Claims versus Opinions
opinion
claim