Executive Summary

Milestone 2 of TikTok Claim Classification Project

ISSUE / PROBLEM

The TikTok data team want to make a automatic claim classification system. To begin, the team need to organize the raw dataset and prepare it for further Exploratory Data Analysis (EDA).

RESPONSE

The data team want to perform a preliminary investigation of the claim classification of the dataset with the aim of learning important relationship between the variable.

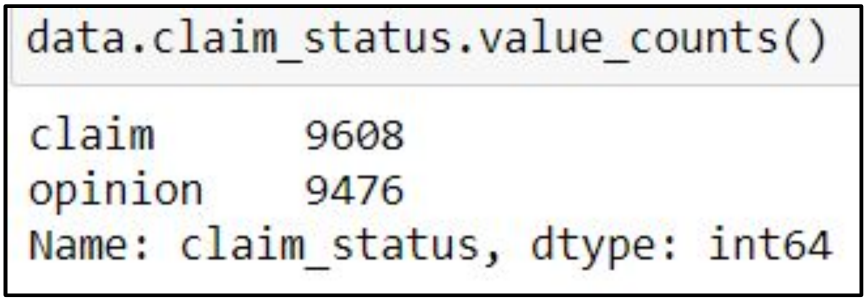
Given the ask of a classification of users claim, the data team looked at the counts of claim and opinion in order to understand the count of each type of video content.

IMPACT

The impact of the preliminary analysis are will be evident in the next step. To get the impact of users video, the data team identified two variable: video duration, and view counts. These two variable are consider for future prediction model.

UNDERSTAND THE DATA

After reviewing the data, the feature claim_status seemed particularly useful. The following screenshot show the summary statistics of claim status variable.



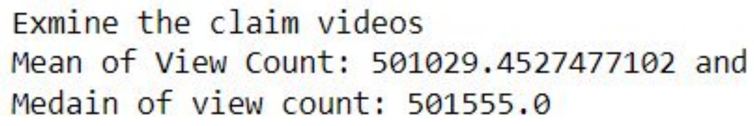
```
data.claim_status.value_counts()

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

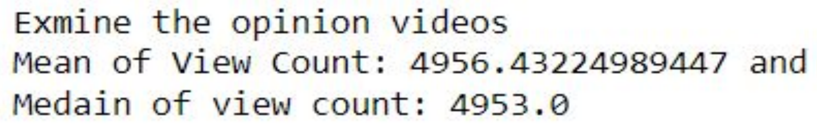
Note: In the dataset the variable claim status is quite balanced. There are total 9608 claim video, and 9476 opinion videos.

ENGAGEMENT TRENDS

The data team considered viewer engagement with each video in the claim and opinion categories. In order to understand the viewer engagement, the data team considered the view count. The mean and median view count show the impact of each category of video.



```
Exmine the claim videos
Mean of View Count: 501029.4527477102 and
Medain of view count: 501555.0
```

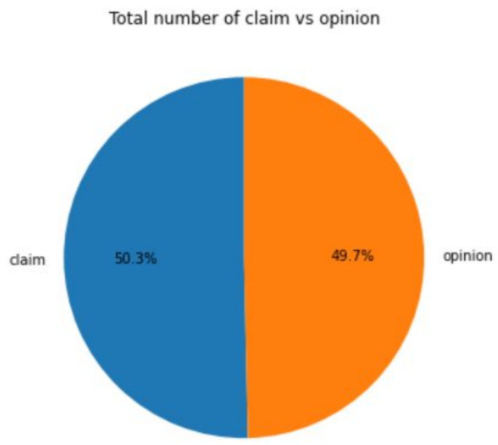


```
Exmine the opinion videos
Mean of View Count: 4956.43224989447 and
Medain of view count: 4953.0
```

The mean and median of view counts between each category is significantly different.

KEY INSIGHTS

- Total number of claim and opinion is approximately similar, which indicate the dataset is balanced.
- With the key variable identified and initial investigation of the claims classification, the process of exploratory data analysis can begin.


Total number of claim vs opinion
claim
opinion