Executive Summary

Milestone 2 of the TikTok Claims Classification Project

ISSUE / PROBLEM

The TikTok data team is undertaking the development of a machine learning model to facilitate the classification of user-submitted claims. Initial efforts will focus on preparing the raw dataset for exploratory data analysis (EDA).

RESPONSE

An initial data analysis was conducted to identify relationships between variables.

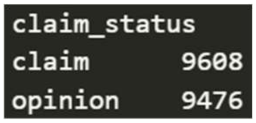
To understand the distribution of video content types, the team analyzed claim and opinion counts.

IMPACT

video_duration (seconds) and video_view_count were identified as potentially important variables for inclusion in future predictive models.

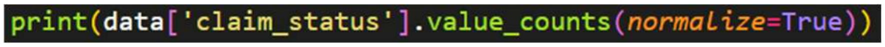
Understanding The Data

The claim_status variable effectively distinguishes between videos categorized as claims and those categorized as opinions.



```
claim_status
claim       9608
opinion     9476
```

Note: There is very little class imbalance within the claim_status variable.

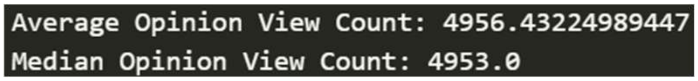


```
print(data['claim_status'].value_counts(normalize=True))
```

Engagement Trends

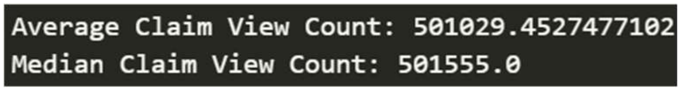
Analysis of video view counts by category reveals that claims have significantly higher median and mean views than opinions.

Opinions



```
Average Opinion View Count: 4956.43224989447
Median Opinion View Count: 4953.0
```

Claims

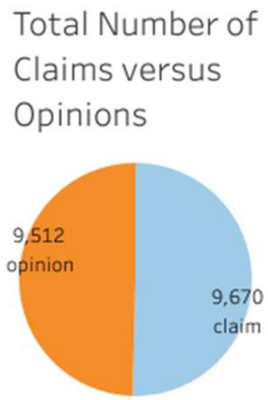


```
Average Claim View Count: 501029.4527477102
Median Claim View Count: 501555.0
```

KEY INSIGHTS

The claim_status variable demonstrates negligible imbalance, enabling subsequent analysis without risk of detrimental impact.

Having identified key variables and completed an initial dataset investigation, exploratory data analysis (EDA) may now commence.


Total Number of Claims versus Opinions
opinion
claim