Executive Summary

Phase-2 of the TikTok Claims Classification Project

ISSUE / PROBLEM

The TikTok data team is building a machine learning model to classify user submissions as claims or opinions, improving moderation. This summary covers initial data preprocessing by Saswat Seth, preparing the dataset for analysis and model development.

RESPONSE

The data team analyzed the claims classification dataset to identify key relationships and assess claim-opinion distribution. Rising user reports have led to moderation backlogs, making efficient classification essential. To address this, the team preprocessed the dataset, preparing it for deeper analysis and machine learning model development.

IMPACT

This analysis establishes key factors for future predictive modeling by identifying critical variables like video_duration and video_view_count. These insights will refine classification models, enhancing TikTok's content moderation efficiency and improving the accuracy of claim classification.

UNDERSTANDING THE DATA

The claim_status variable is crucial for classification, with a near-equal distribution of claims and opinions ensuring balanced modeling. However, null values and potential outliers require further cleaning before EDA.

```
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

Viewer engagement was analyzed through video view counts across claim and opinion categories. Controversial claims generate higher engagement per post despite fewer total views, likes, and shares, indicating they spark more discussions and reactions.

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

- The **claim_status** variable is well-balanced (9,608 claims, 9,476 opinions), but **298 null values** must be addressed before model training.
- **Claims** attract higher engagement despite fewer views and are more frequently flagged for **policy violations**. Potential **outliers** in the data require attention during **EDA**.

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