Executive Summary

ISSUE / PROBLEM

The TikTok data team aims to build a machine learning model to classify videos as either claims or opinions. To begin, the team needs to clean and organize the dataset before starting exploratory data analysis.

RESPONSE

The team conducted an initial analysis of the dataset, focusing on understanding how user-generated content is labeled as either a claim or opinion. The analysis included a review of engagement metrics to support future modeling efforts.

IMPACT

The analysis revealed that **claim videos have much higher engagement levels** than opinions. Videos from **banned or under-review users** also showed very high share counts

UNDERSTANDING THE DATA

After reviewing the dataset, the variable claim_status was central to the project. The dataset is well balanced:

```
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 better understand viewer interaction, the team created new metrics: likes per view, comments per view, and shares per view. These engagement ratios helped identify how users respond to different types of content.
The analysis showed that **claim videos consistently received higher engagement** than opinion videos across all metrics. This suggests that claim content may be more attention-grabbing or controversial, encouraging more 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 dataset is balanced (~50% claims, ~50% opinions), which is ideal for classification.
- Claim videos are more viral and generate more interaction than opinions.
- Videos with high engagement are more likely to come from users who were banned or under review.
- The dataset is ready for deeper exploratory analysis and feature selection for modeling.

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