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 on the platform. In this stage, raw data will be explored and prepared for the next stages.

RESPONSE

A preliminary investigation was performed with regard to the claims classification dataset with the aim of knowing the primary state of the data, understanding how the data is structured, and having a first impression of it. Some of the steps are to check for missing values, check how the data is balanced in some variables, like the claim status and the verified status, and seek other relationships.

IMPACT

The impact of this preliminary analysis will be evident in the next steps. 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

As a result of the analysis, it was concluded that the "claim_status" variable will be of high importance in the next stages. The dataset is balanced with respect to this variable in the next way:

```
claim_status
claim      50.345839
opinion    49.654161
Name: proportion, dtype: float64

Number of NaN values: 298
```

Note: The count is balanced, which is good news because the dataset has enough data of both types.

ENGAGEMENT TRENDS

Let's consider viewer engagement with each video in the claim and opinion categories. In order to understand viewer engagement, descriptive statistics were performed. The mean and median view counts show the impact of each status; specifically, the mean and median view counts for both categories show the association between content and video views.

Claims:

```
Mean view count claims: 501029.5
Median view count claims: 501555.0
```

Opinions:

```
Mean view count opinions: 4956.4
Median view count opinions: 4953.0
```

KEY INSIGHTS

- The column 'claim_status' contains interval data, and it was necessary to clean some null values in this column.
- There are more claim videos than opinions, specifically 1.4% more, which is not a significant difference.
- Even though the number of active authors' content is higher, content by banned authors has a higher engagement.

Pie chart visualizes the comparison of the count of verified and not verified