Executive Summary

Milestore 2 of the TikTok Claims Classification Project

ISSUE

TikTok would like to streamline the claim processing time by implementing a machine learning model. The original dataset needs to be prepared and organized for exploratory data analysis

RESPONSE

Data Analytics has conducted initial investigation of the claims data to uncover the relationship between important variables

User claims need to be classified, the team reviewed counts for both claims and opinions and the types of video content.

IMPACT

There are two variables identified in this preliminary investigation that will be of importance as we move onto building the prediction model: video_duration (secs) and video_count c

UNDERSTANDING THE DATA

The variable claim_status is the initial clasification point for our project, the following snippet shows the values it takes on the provided dataset, this helps us understand our data:

```
[23]: # What are the different values for claim status and how many of
      value_counts = data['claim_status'].value_counts()
      print(value_counts)
```

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

Note the balance between claims and opinions for videos

ENGAGEMENT TRENDS

By working with view count, our team considered viewer engagement for both categories, claims and opinions. Reviewing the mean and media of view count reveals the particular association between content (claims or opinions) and video views

```
[24]: # What is the average view count of videos with "claim" status?
      mask = data['claim_status'] == 'claim'
      average_view_count = data[mask]['video_view_count'].mean()
      median_view_count= data[mask]['video_view_count'].median()
      print (average_view_count)
      print (median_view_count)
```

```
501029.4527477102
501555.0
```

```
[25]: # What is the average view count of videos with "opinion" status?
      mask = data['claim_status'] == 'opinion'
      average_view_count = data[mask]['video_view_count'].mean()
      median_view_count= data[mask]['video_view_count'].median()
      print (average_view_count)
      print (median_view_count)
```

```
4956.43224989447
4953.0
```

KEY INSIGHTS

- ❑ The number of claim videos vs claims is nearly equal. We will proceed with this consideration for our future analysis
- ❑ The initial investigation has provided us with the key variables from the dataset. We are ready to proceed with exploratory data analysis