Executive Summary

Milestone 3

ISSUE / PROBLEM

Tiktok seeks to use Machine learning model to pickout ideas within tik tok videos and comments and identify them either as claim or opinion. The team had already done some preliminary analysis and has now proceeded to milestone 3

RESPONSE

The data team performed EDA. From preliminary analysis and basic analysis of data data team formed following hypothesis

1. For videos with higher view and like count, claim_status is claim.
2. Mostly verified users posts lesser claims than not verified
3. Users with claims are more likely to be banned and less likely to be active as compared to those with opinion
4. Video with more duration is more likely to have claim

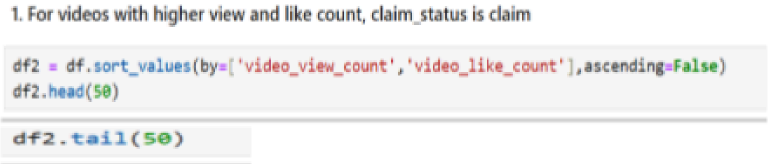
IMPACT

The result of this analysis will be helpful to prove hypothesis as well as gaining valuable insights about outliers and propose solution to deal with outliers

ANALYSIS

Data team performed 6 step exploratory data analysis which involved cleaning of dataset.
The data tema during analysis tested all four hypothesis to derive insights from the data

Hypothesis # 01

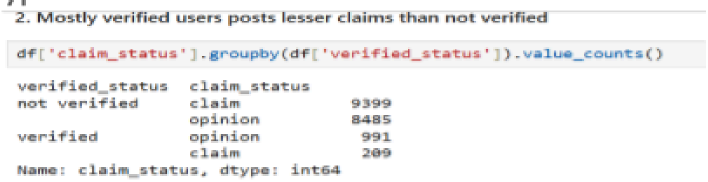


1. For videos with higher view and like count, claim_status is claim

```
df2 = df.sort_values(by=['video_view_count','video_like_count'],ascending=False)
df2.head(50)
```

```
df2.tail(50)
```

Hypothesis # 02

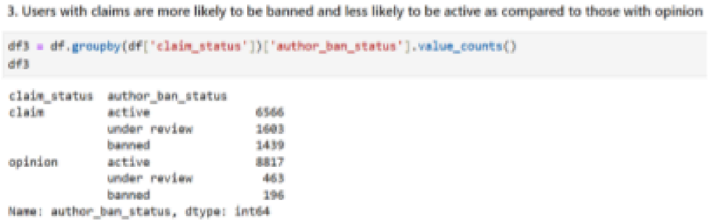


2. Mostly verified users posts lesser claims than not verified

```
df['claim_status'].groupby(df['verified_status']).value_counts()
```

```
verified_status  claim_status
not verified     claim           9399
                 opinion         8485
verified         opinion          991
                 claim            209
Name: claim_status, dtype: int64
```

Hypothesis # 03

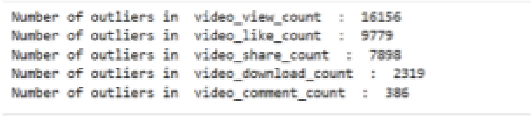


3. Users with claims are more likely to be banned and less likely to be active as compared to those with opinion

```
df3 = df.groupby(df['claim_status'])['author_ban_status'].value_counts()
df3
```

```
claim_status  author_ban_status
claim         active               6566
              under review         1603
              banned               1439
opinion       active               8817
              under review          463
              banned                196
Name: author_ban_status, dtype: int64
```

Hypothesis # 04 (Video with more duration is more likely to have claim) was proved to be false.
Additionally data team conducted analysis to determine outliers



```
Number of outliers in  video_view_count  :  16156
Number of outliers in  video_like_count  :  9779
Number of outliers in  video_share_count  :  7898
Number of outliers in  video_download_count  :  2319
Number of outliers in  video_comment_count  :  386
```

KEY INSIGHTS

1. During cleaning phase following conclusions were derived about the data,
   a. There were 298 missing values that were dropped during the process
   b. The data is rightly skewed
   c. There are right outlier in dataset
2. Following conclusions were derived from the data analysis
   a. For videos with higher view and like count, claim_status is claim.
   b. Mostly verified users posts lesser claims than not verified.
   c. Users with claims are more likely to be banned and less likely to be active as compared to those with opinion
3. The data was presented as tableau story
   Tableau Story Tik Tok