Executive Summary

Understand The Data

Stage Overview

This stage aims to investigate and understand the data provided, construct a data frame in Python, perform a cursory inspection of the provided dataset, and inform TikTok data team members of your findings.

Key Findings

- The data includes **19,383 rows** and **12 columns**
- The data quality issues:

  - All *_count variables are in wrong data types. They should be integers

  - # variable isn't useful

  - There are 298 rows with NULL values in the data

- The counts of each claim status are quite balanced. There are 9,608 claims and 9,476 opinions.
- After variables investigation, I see that the author_ban_status and the engagement rates (likes/comments/shares per view) are the most important variables for learning our classification model. See the visual details

Visual Details

Impact of high engagement rates on the probability of claim status

		likes_per_view		comments_per_view		shares_per_view	
		median	mean	median	mean	median	mean
author_ban_status	claim_status						
active	claim	0.326538	0.329542	0.000776	0.001393	0.049279	0.065456
	opinion	0.218330	0.219744	0.000252	0.000517	0.032405	0.043729
banned	claim	0.358909	0.345071	0.000746	0.001377	0.051606	0.067893
	opinion	0.198483	0.206868	0.000193	0.000434	0.030728	0.040531
under review	claim	0.320867	0.327997	0.000789	0.001367	0.049967	0.065733
	opinion	0.228051	0.226394	0.000293	0.000536	0.035027	0.044472

Video of banned / under review authors have a high probability of being claim

```
df.groupby(["author_ban_status", "claim_status"]).size()
```

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

Next Steps

- Modify data types of *_count variables to the appropriate integers.
- Create the engagement rate variables to the data provided
- Fix data quality issues
- Perform more depth EDA on the data.