# PROJECT MILESTONE 3

## PACE: Plan Stage

- What are the data columns and variables and which ones are most relevant to my deliverable?

The dataset contains 12 columns of the following dtypes: float64 (5), int64 (3), object (4).

The most relevant variables for my deliverable are the following data columns: “video_duration_sec,” “video_like_count,” “video_comment_count,”, “video_view_count”, “video_download_count”, “video_share_count”, “author ban status”, “verified_status” and "claim_status".

- What units are the variables in?

“video_duration_sec” stores the duration of the video in seconds, all the 5 count variables count the different user video engagements, and the 3 status variables are categorical string variables.

- What are my initial presumptions about the data that can inform my EDA, knowing I will need to confirm or deny with my future findings?

My initial presumptions are that all these 8 variables can predict the claim status of the video.

- Is there any missing or incomplete data?

From the basic information about the dataset, I can see 298 missing data in all count variables, in “claim_status” and in “video_transcription_text” variables.

- Are all pieces of this dataset in the same format?

No, 5 columns are of float64, 3 columns are of int64, and 4 columns are of object dtypes.

- Which EDA practices will be required to begin this project milestone?

1. Get the basic information about the dataset: first few rows, size and shape.

2. Get the descriptive statistics.

## PACE: Analyze Stage

- What steps need to be taken to perform EDA in the most effective way to achieve the project goal?

1. Data loading and inspection
2. Visualization
3. Correlation analysis
4. Feature engineering
5. Outlier detection
6. Data transformation

- What initial assumptions do I have about the types of visualizations that might best be suited for the intended audience?

1. To identify outliers, I'll use box plots.
2. To visualize the data distributions - histograms.
3. To compare categorical variables – count plots.
4. To compare different variable counts - bar plots.
5. To depict proportions - pie graphs.
6. To visualize the trends, patterns and outliers – scatter plots.

## PACE: Construct Stage

- What data visualizations, machine learning algorithms, or other data outputs will need to be built in order to complete the project goals?

Box plots, histograms, count plots, bar plots, pie graphs, scatter plots.

Box plots will determine outliers and where the bulk of the data points reside.
Histograms will show the distributions of “video_duration_sec” and count variables.
Count plots will show the comparison of categorical variables “claim_status”, “verified_status” and “author_ban_status”.
Bar plots will show the comparison of median view counts between the three different author ban statuses.
Pie graphs will depict the proportions of total views for claim videos and total views for opinion videos.
Scatter plots to visualize “video_view_count” versus “video_like_count” according to 'claim_status', and “video_view_count” versus “video_like_count” for opinions only.

## PACE: Execute Stage

- What key insights emerged from the EDA and visualizations(s)?

The conducted EDA analysis revealed many considerations for the classification model, including missing values, “claims” to “opinions” balance, and overall distribution of data variables. The two key insights from this analysis were:

1. Null values

2. Skewed data distribution

- Given what I know about the data and the visualizations I was using, what other questions could I research for the client?

I want to further investigate distinctive characteristics that apply only to claims or only to opinions. Also, I want to consider other variables that might be helpful in understanding the data.

- How might I share these visualizations with different audiences?

I prepared all the data visualizations to be clean, easily understandable, accessible, and also considering color, contrast, emphasis, and labeling.