**Course Three**

# Go Beyond the Numbers: Translate Data into Insights



# Instructions

Use this PACE strategy document to record decisions and reflections as you work through this end-of-course project. You can use this document as a guide to consider your responses and reflections at different stages of the data analytical process. Additionally, the PACE strategy documents can be used as a resource when working on future projects.

# Course Project Recap

Regardless of which track you have chosen to complete, your goals for this project are:

* Complete the questions in the Course 3 PACE strategy document
* Answer the questions in the Jupyter notebook project file
* Clean your data, perform exploratory data analysis (EDA)
* Create data visualizations
* Create an executive summary to share your results

# Relevant Interview Questions

Completing the end-of-course project will help you respond these types of questions that are often asked during the interview process:

* How would you explain the difference between qualitative and quantitative data sources?
* Describe the difference between structured and unstructured data.
* Why is it important to do exploratory data analysis?
* How would you perform EDA on a given dataset?
* How do you create or alter a visualization based on different audiences?
* How do you avoid bias and ensure accessibility in a data visualization?
* How does data visualization inform your EDA?

**Reference Guide**

This project has six tasks; the visual below identifies how the stages of PACE are incorporated across those tasks.



**Data Project Questions & Considerations**

**PACE: Plan Stage**

* What are the data columns and variables and which ones are most relevant to your deliverable?

The dataset includes claim status, engagement metrics (e.g., video\_view\_count, video\_like\_count, video\_comment\_count), author ban status, and verification status. The most relevant variables are claim status, engagement metrics, and author ban status as these will help assess how these factors impact video performance.

* What units are your variables in?

Engagement metrics are in numeric counts (e.g., views, likes, comments). Claim status and author ban status are categorical (Claim or Opinion, Active or Banned).

* What are your initial presumptions about the data that can inform your EDA, knowing you will need to confirm or deny with your future findings?

Claim videos will likely have higher engagement than opinion videos. Non-active authors may have lower engagement due to reduced visibility. There might be outliers in engagement metrics because of viral content.

* Is there any missing or incomplete data?

I need to check for missing values in variables like claim\_status and engagement metrics. Missing data will be handled through removal or imputation.

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

I will verify that all numeric variables are in the correct format (integers or floats) and ensure that categorical variables have consistent naming conventions (e.g., no typos in claim\_status).

* Which EDA practices will be required to begin this project?

I will need to check summary statistics (e.g., mean, median) and visualize distributions (e.g., box plots, histograms) to identify outliers. Scatter plots and correlation matrices will help understand relationships between variables.

**PACE: Analyze Stage**

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

Clean the data, handle missing values, and use descriptive statistics to understand distributions. Create visualizations to detect outliers and explore relationships between engagement metrics and claim status.

* Do you need to add more data using the EDA practice of joining? What type of structuring needs to be done to this dataset, such as filtering, sorting, etc.?

The dataset is sufficient for analysis. I might need to filter out videos with missing or incomplete engagement data but do not need to join additional data.

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

Bar plots for comparing engagement metrics across claim status and author ban status. Box plots and histograms to explore distributions, and scatter plots for relationships between engagement metrics like views and likes.

**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?

I will create histograms, box plots, bar plots, scatter plots, and pie charts. Machine learning algorithms aren’t necessary for the current goal, as we’re focusing on exploratory data analysis.

* What processes need to be performed in order to build the necessary data visualizations?

Prepare and clean the data by addressing missing values and formatting. Create visualizations using Python libraries (e.g., seaborn and matplotlib) with appropriate labeling, legends, and titles for clarity.

* Which variables are most applicable for the visualizations in this data project?

Claim status, engagement metrics, and author ban status are the key variables for analysis and visualization.

* Going back to the Plan stage, how do you plan to deal with the missing data (if any)?

I will check the amount of missing data and either remove rows with missing values or fill missing data using the median or another imputation method, depending on the proportion of missing data.

******PACE: Execute Stage**

* What key insights emerged from your EDA and visualizations(s)?

Claim videos tend to have higher engagement than opinion videos, likely due to their greater visibility. Non-active authors have lower engagement, indicating that restrictions limit their content’s reach. Outliers exist, particularly in views and likes, which reflects viral content.

* What business and/or organizational recommendations do you propose based on the visualization(s) built?

The platform could work on strategies to boost engagement for opinion videos. Non-active authors could benefit from re-engagement strategies or changes in moderation practices. The platform might also focus on content strategies that balance engagement across both claim and opinion videos.

* Given what you know about the data and the visualizations you were using, what other questions could you research for the team?

I could explore how video categories (e.g., genre) influence engagement or whether video length or posting time has a significant impact on video performance for claim vs. opinion videos.

* How might you share these visualizations with different audiences?

For executives or stakeholders, I would provide high-level summaries and visualizations showing overall trends (e.g., total views for claim vs. opinion videos). For data analysts, I would present detailed charts and statistical analyses. Interactive visualizations in Tableau or another visualization tool could allow deeper exploration for different departments.