**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 to 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 data includes columns such as #, claim\_status, video\_id, verified\_status, author\_ban\_status, video\_duration\_sec, video\_view\_count, video\_like\_count, video\_share\_count, and video\_comment\_count.

* What units are your variables in?

The variable video\_duration\_sec is measured in seconds. Other variables do not have applicable units.

* 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?

Given that the data is first-party, it should require less cleaning and have been properly maintained. However, these assumptions should be confirmed or denied through future findings.

* Is there any missing or incomplete data?

There is missing data in seven columns: claim\_status, video\_transcription\_text, video\_view\_count, video\_like\_count, video\_share\_count, video\_download\_count, and video\_comment\_count.

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

All pieces of this dataset appear to be in the same format.

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

Cleaning. Discovery and structuring has been done.

**PACE: Analyze Stage**

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

Discovery, structuring, cleaning, presenting, cleaning, validation, joining, presenting.

* 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.?

Extraction to select most relevant data, sorting, filtering, and grouping.

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

I think a stacked bar plot would be effective in making comparisons between the videos comment, likes, shares and views count.

**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?

To achieve the project goals, we need to build histograms, bar charts, and heatmaps for data visualization; implement text classification models like Naive Bayes, SVM, or BERT for data classification; and use NLP techniques for feature extraction. Other potential outputs include predictive models for future predictions and anomaly detection for unusual activity.

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

To create bar charts, pie charts, and scatter plots for the TikTok dataset, we need to prepare the data by cleaning the data - handling missing values, removing duplicates. Next, we conduct exploratory data analysis, and then choose the appropriate visualization type for each variable we are interested in.

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

For the TikTok dataset project, the most applicable variables for visualizations could include video view count, like count, share count, download count, comment count, video duration, verification status, ban status, and claim status.

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

Deleting them.

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

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

From the Exploratory Data Analysis (EDA) and visualizations, it was observed that TikTok videos are generally short in duration, with a uniform distribution of lengths between 5-60 seconds. The view count, like count, share count, download count, and comment count are all highly skewed towards lower values, indicating that most videos receive fewer views, likes, shares, and comments in keeping with social media states. There are significantly more unverified users than verified ones, and verified users are more likely to post opinions. Claim videos tend to have higher view counts than opinion videos, suggesting that claim status might be an indicator of popularity.

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

Consider focusing on shorter content, encouraging user verification, promoting claim videos, and encouraging active authors to post more frequently to increase engagement and view counts.

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

To better understand user behavior on TikTok, the team could analyze user interactions, track video performance, monitor hashtag and sound trends, and observe follower growth.

* How might you share these visualizations with different audiences?

Visualizations can be shared with different audiences through executive summaries, presentations, reports, dashboards, social media, collaborative tools, emails, websites or blog posts, and infographics, depending on the audience's needs and the complexity of the data.