**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 provided Data Dictionary lists the columns as #, claim\_status, video\_id, video\_duration\_sec, video\_transcription\_text, verified\_status, author\_ban\_status, video\_view\_count, video\_like\_count, video\_share\_count, video\_download\_count, video\_comment\_count. The most important columns are claim\_status and the engagement metrics like video\_like\_count.

* What units are your variables in?

It depends on the variable. Some are categories such as claim\_status while others are counts, and video duration is provided in seconds.

* 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?

I am assuming that claim videos receive more engagement.

* Is there any missing or incomplete data?

There are some NA entries but they are minor and few in number so they can be ignored without compromising data integrity.

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

No. Some like claim\_status are objects, others like video\_id are integers, and others like video\_like\_count are float.

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

I’ll want to review descriptive statistics for each engagement metric to see what outliers might exist.

**PACE: Analyze Stage**

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

I will want to separate or filter the engagement metrics by claim status to compare engagement between the two categories.

* 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.?

As I mentioned, filtering out NA values and then structuring the data so that claim status can be compared will be important.

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

Pie charts to compare claim status and box plots to show outliers.

**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?

The goal of the project is to create a machine learning algorithm so visualizations that reveal trends which can be used to construct the learning model will be important.

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

I will need to perform EDA to understand what story I want the visualizations to tell. I need to find the data trends that I want to highlight.

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

claim\_status, video\_like\_count, video\_comment\_count, and other engagement metrics.

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

I’ll need to assess the extent of the missing data and then decide if I can cut the data.

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

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

Claim videos are more likely to receive more engagement even though they are only around 50% of videos.

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

We need to monitor videos that receive high levels of engagement since those are most likely to be claim videos.

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

There may be cultural differences that correspond to the likelihood of a user making a claim or stating an opinion. Could geographic location be provided to test this?

* How might you share these visualizations with different audiences?

I would adjust the language used to be sure that it can be understood by a non-technical audience and I would also verbally describe the key insight for each graph.