# Course Two

## Get Started with Python

### 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 2 PACE strategy document
- □ Answer the questions in the Jupyter notebook project file
- □ Complete coding prep work on project’s Jupyter notebook
- □ Summarize the column Dtypes
- □ Communicate important findings in the form of an executive summary

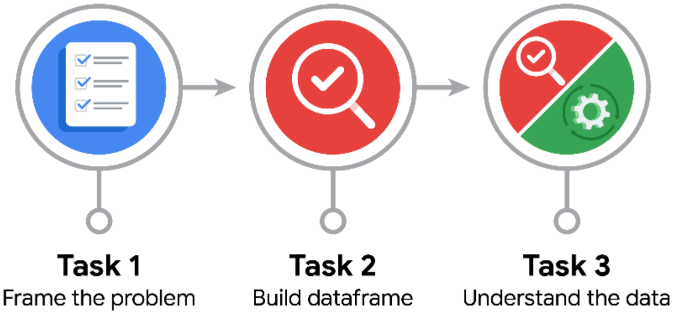
### Relevant Interview Questions

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

- Describe the steps you would take to clean and transform an unstructured data set.
- What specific things might you look for as part of your cleaning process?
- What are some of the outliers, anomalies, or unusual things you might look for in the data cleaning process that might impact analyses or ability to create insights?

## Reference Guide

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



## Data Project Questions & Considerations

### PACE: Plan Stage

- How can you best prepare to understand and organize the provided information?

To best prepare, we must understand the questions that we will be answering and the deliverables that are expected as the outcome.

- What follow-along and self-review codebooks will help you perform this work?

Jupyter Notebook

- What are some additional activities a resourceful learner would perform before starting to code?

Ask for clarification on any aspects that are unclear, inspect the raw data for structural insights, and review all documentation already in-hand.

## PACE: Analyze Stage

- Will the available information be sufficient to achieve the goal based on your intuition and the analysis of the variables?

The available information will be sufficient to achieve the goal after analysis and the application of machine learning algorithms. However, the use of only intuition and analysis of the variables would likely not be enough to complete the goals of the project.

- How would you build summary dataframe statistics and assess the min and max range of the data?

After importing and storing the data in a pandas dataframe, we could use the pd.datafram.describe() method to show summary statistics.

- Do the averages of any of the data variables look unusual? Can you describe the interval data?

There is a distinct difference between the averages of the banned vs active accounts. The interval data was based on the engagement with the video, for example, views, shares, comments, and likes.

## PACE: Construct Stage

**Note**: The Construct stage does not apply to this workflow. The PACE framework can be adapted to fit the specific requirements of any project.

## PACE: Execute Stage

- Given your current knowledge of the data, what would you initially recommend to your manager to investigate further prior to performing exploratory data analysis?

I would recommend that we further explore statistical testing to confirm correlations that are suspected. Additionally, I would recommend clearly defining key points that would determine whether a submission is a claim or opinion as declared by a human moderator.

- What data initially presents as containing anomalies?

View, like, share, comment, and download all have the same number of values null, which might indicate a block of submissions are anomalous and should be ignored.

Also of note, the fact that video posted by banned authors have much higher engagement numbers, which feels like an anomaly, although my gut tells me that this is due to inflammatory nature and or inappropriate content.

- What additional types of data could strengthen this dataset?

Additionally, data that only includes active and banned users, for better training and testing purposes.