# 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

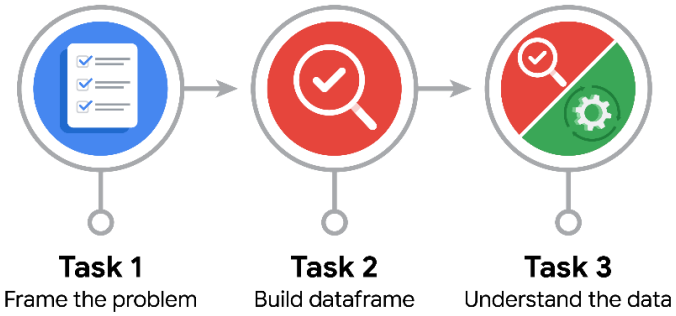
## Relevant Interview Questions

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?

  Study the information given in the data dictionary regarding the different columns of data available. Explore the dataset using the pandas package, compute descriptive statistics.

- 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?

  Refresh knowledge on the pandas package, understand what metrics/statistics are useful to produce an executive summary, read reports of similar projects if available.

## 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 is sufficient to build a predictive model that distinguishes between claims and opinions as the content views, shares, comments and likes are different between the two categories. They are also different when you consider both claim status and author ban status. |

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

| |
|---|
| Summary statistics can be obtained using the function describe(). Minimum and maximum values are shown and using the percentile values (25%, 50%, 75%) we can get a sense of the spread of data. |

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

| |
|---|
| The standard deviation of the variables describing the number of views, likes, comments, shares and downloads is greater than the mean. This indicates that the values of these variables span a large range and that mean is not the best variable to describe the information. Median is likely a better descriptor in this case.<br><br>The median value of the number of views, likes, comments, shares and downloads is very small compared to the maximum values of these variables. Also, values greater than the median span a larger range than that below it. |

## 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?

| |
|---|
| Further analysis of the data is required before deciding on the variables that have predictive value. Several columns of the dataset have null values, they need to be removed prior to EDA. |

- What data initially presents as containing anomalies?

| |
|---|
| There were no columns identified as having anomalies. |

- What additional types of data could strengthen this dataset?

| |
|---|
| A column identifying whether the video content caused the user ban would be a useful predictor, if available. |