**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 the 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?

We can best prepare to understand and organize the information provided by exploring the given dataset and reviewing the Data Dictionary for the stakeholder. We need to load the data first by importing the packages that import pandas as pd and import NumPy as np.

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

I have used Jupyter Notebook to perform this work efficiently.

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

First, read the data, view the data dictionary, and explore the data to identify key variables for the stakeholder.

**PACE: Analyze Stage**

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

The analyzing stage includes importing and data loading, understanding the data-inspecting the data, and investigating the variables. Information is sufficient because at first, we imported the packages pandas and NumPy prior to exploring the dataset. After importing packages, next up is to view the summary information and it can be done by coding the following data.head(), data.info(), and data.describe().

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

We build summary statistics by using simple code-named data. describe(). After using this function, it is shown that many of the count variables have outliers at the high end of the distribution. They have high maximum values, and the standard deviation is quite large.

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

Yes, the averages of data variables of opinion and claim status look unusual. There is a huge discrepancy between view counts for videos labeled as opinions and videos labeled as claims.

**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 suggest continuing to investigate engagement levels that will provide in-depth information prior to performing exploratory data analysis.

* What data initially presents as containing anomalies?

Initially, a few observations were that banned authors and those under review get far more views, shares, and likes than active authors. It also seems that the mean in most groups is greater than the median.

* What additional types of data could strengthen this dataset?

We need to create three new columns likes\_per\_view, comments\_per\_view, and shares\_per\_view to get a better understanding of engagement rates.