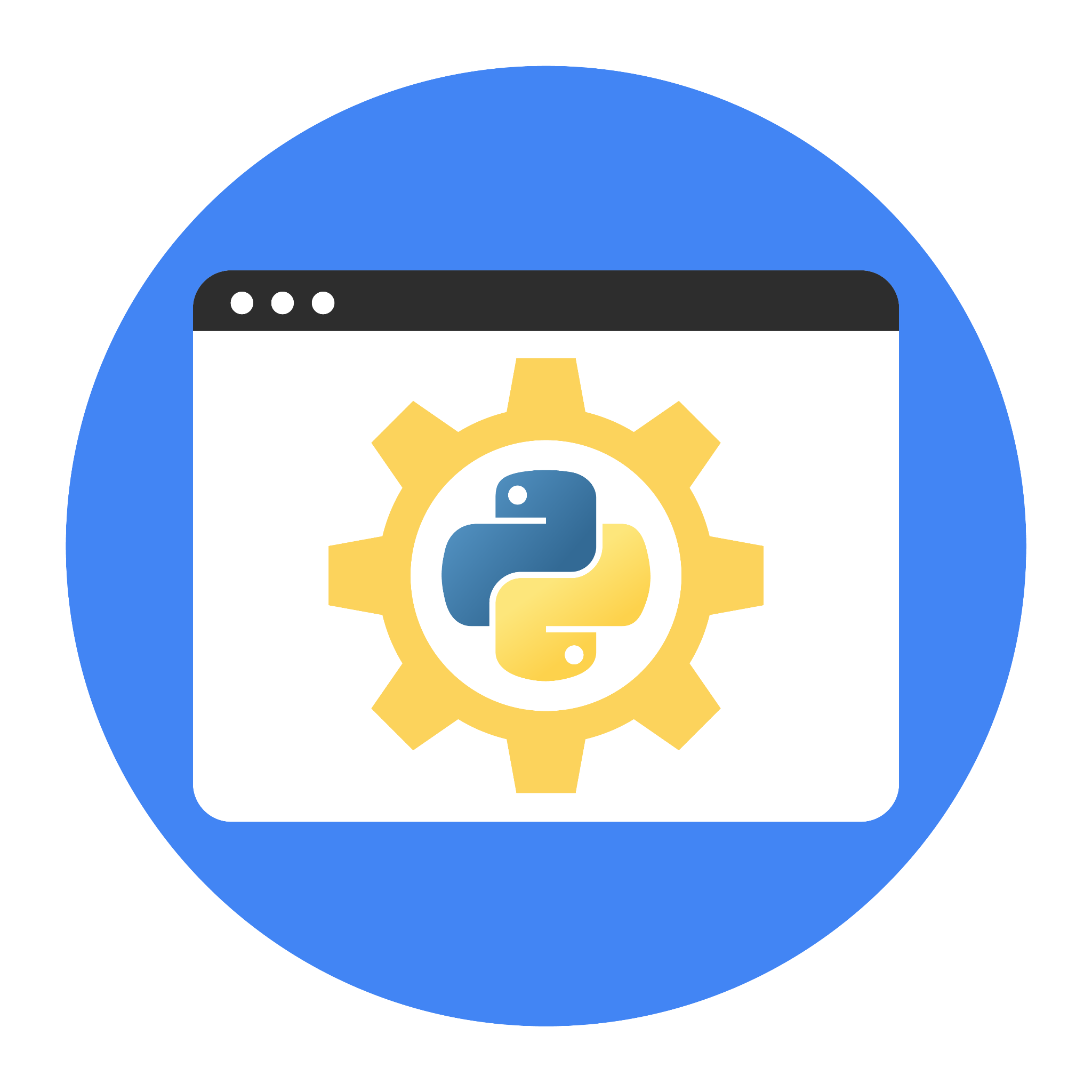
**Course Two**

# 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

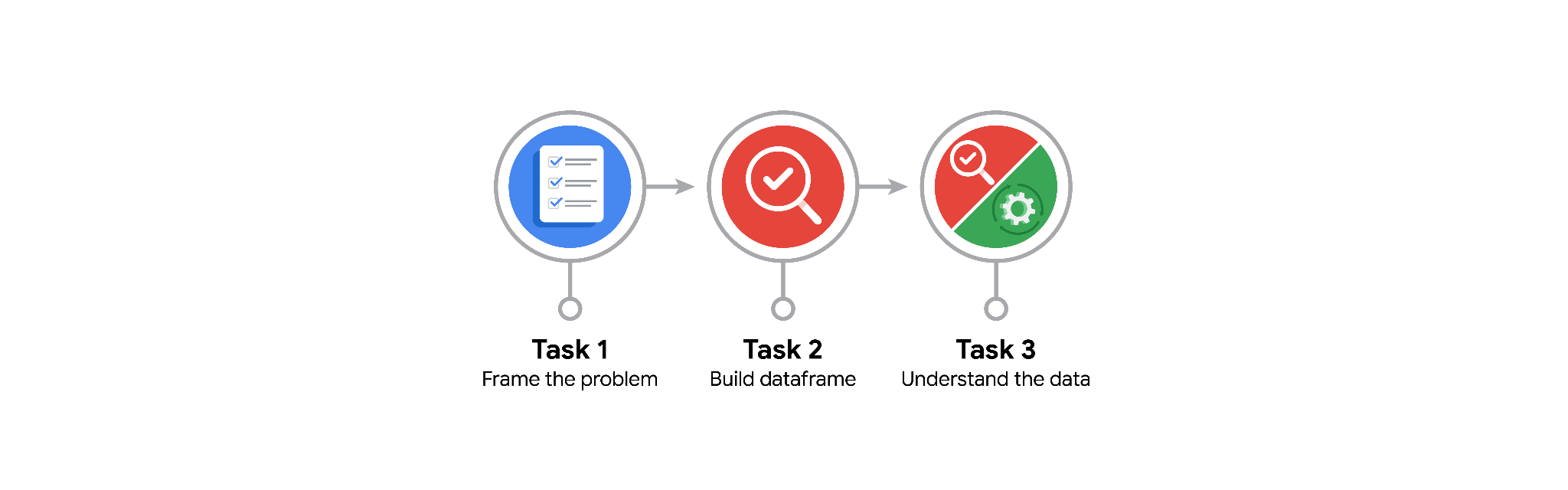
# Relevant Interview Questions

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?

I should first examine the “Data Dictionary” which describes the kind of data stored in the table. After that I should examine general descriptive statistics, such as mean, max, min, and standard deviation to understand the shape of the data.

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

Again, the “Data Dictionary” will be helpful here. I can also check my work against the provided exemplar code to verify that I have performed the task correctly before moving on.

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

I should email the person who provided the data to get more information about how the data was collected, when it was collected, and ask about if the data is clean.

**PACE: Analyze Stage**

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

Yes. This is all the data provided by the course, so I assume it will be enough. I was also able to find trends that canl be used to teach the machine learning model.

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

I can use the following code:  
data.describe()

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

There is clearly a large variance in video performance. The poor performing videos do very poorly and the highest performers are incredibly successful.

**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 anticipate that the claim status will be related to the author ban status and the verified status. If an author is banned, then they likely make claims rather than simply state opinions and unverified authors may be more likely to engage in this behavior because they can easily make another account if banned.

* What data initially presents as containing anomalies?

There seem to be 19382 rows, but some columns only have 19084 rows with non-null values. This indicates that some columns are missing data.

* What additional types of data could strengthen this dataset?

I believe that exploring upload frequency and follower count could provide other useful variables for developing the machine learning model.