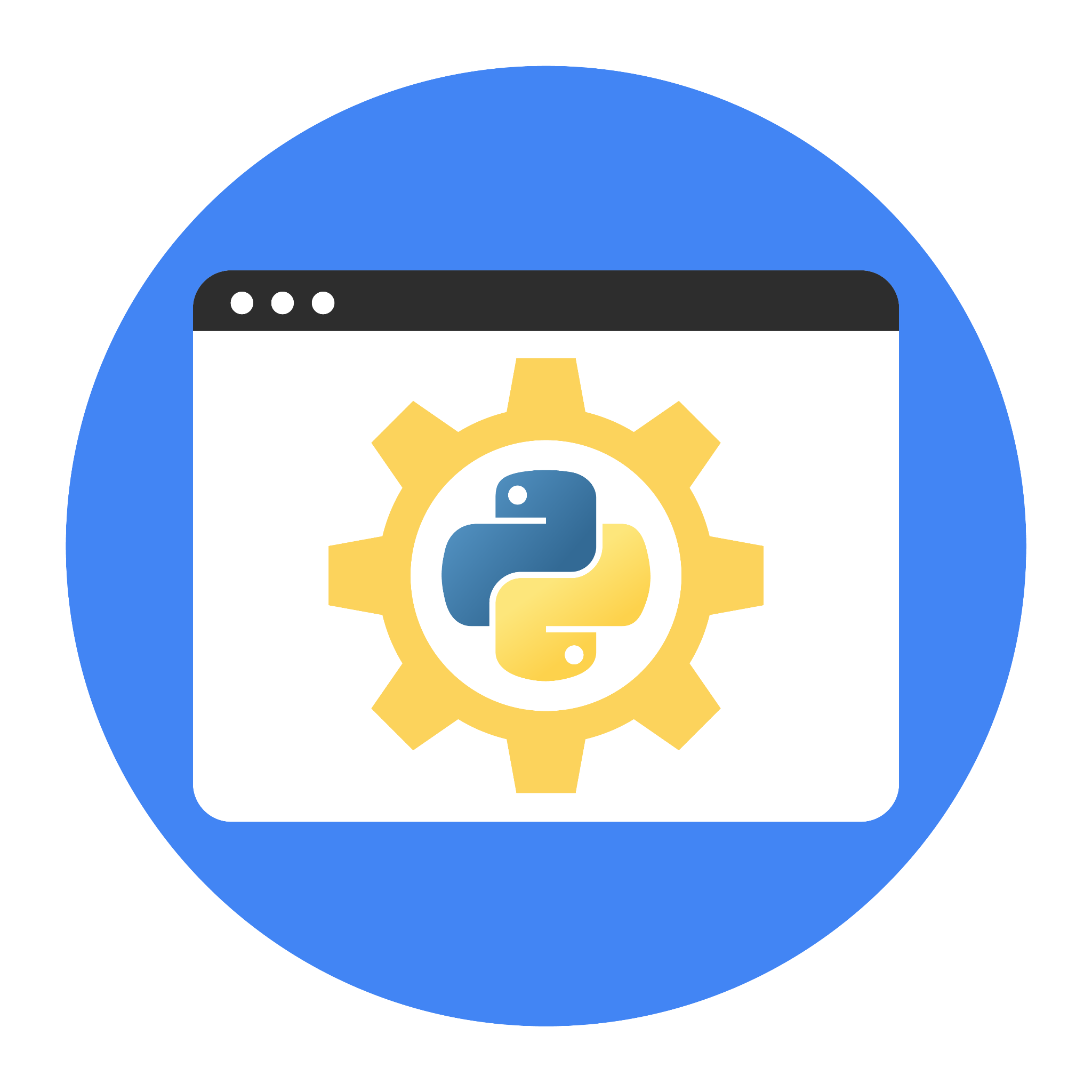
**Course Two**

# 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

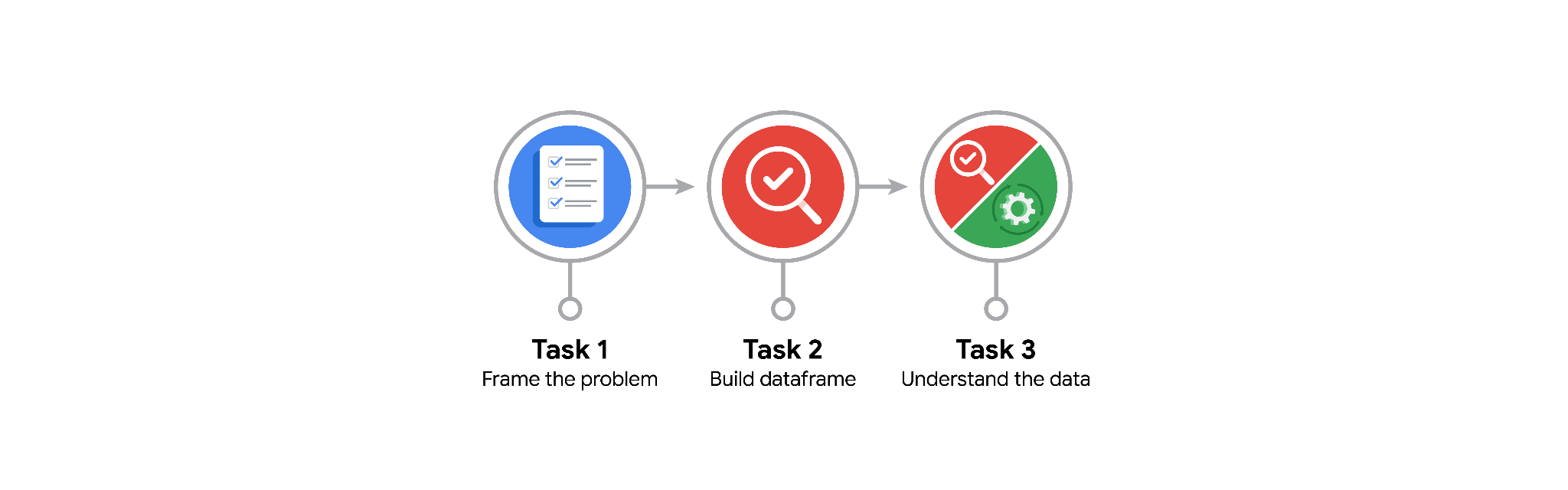
# Relevant Interview Questions

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 prepare, I began by:

Importing the necessary libraries (pandas and numpy) to handle and analyze the dataset.

Exploring the structure of the dataset using .head(), .info(), and .describe() to understand the column names, data types, missing values, and statistical distributions.

Reviewing the project’s data dictionary and proposal to align with the classification objective: identify whether a video is a "claim" or an "opinion".

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

The Course 2 practice notebooks (particularly those focusing on data loading, cleaning, aggregation, and Boolean masking).

Documentation for pandas functions like .groupby(), .agg(), .isnull(), and .fillna().

Personal notes from lessons on exploratory data analysis (EDA) and summary statistics.

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

Check for missing or null values and understand the proportion of missing data.

Investigate data types and convert them if necessary (e.g., ensuring numerical columns are correctly typed).

Research domain knowledge (e.g., understanding what “claim” vs “opinion” might mean in TikTok context).

Examine unique values for categorical features like claim\_status, author\_ban\_status, and verified\_status.

**PACE: Analyze Stage**

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

Yes. The dataset contains a claim\_status column (the target variable), as well as multiple features that reflect engagement and account status (views, likes, shares, verified status, etc.), which are useful for classification.

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

I used .describe() to compute:

* Count, mean, std, min, max, and percentiles for all numeric columns.
* This helped identify potential outliers and understand engagement metrics across the dataset.
* Do the averages of any of the data variables look unusual? Can you describe the interval data?

Yes:

The average video view count (~255K) is much higher than the median (~10K), indicating positive skew due to outliers (some videos going viral).

Similarly, engagement columns like likes, shares, and downloads show significant disparity between mean and median, pointing to a few highly viral videos driving up the average.

**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?

Explore the relationship between claim status and engagement metrics (likes, shares, comments).

Analyze the role of author ban status — banned users tend to have higher engagement, which might affect how content spreads.

Investigate whether there is bias or inconsistency in how “claims” are flagged and how banning decisions are made.

* What data initially presents as containing anomalies?

There are some videos with view counts as low as 20, and others nearing 1 million, indicating potential outliers.

Some videos have 0 comments/likes/shares, which might be legitimate or a sign of data collection gaps.

Engagement rates (likes per view, shares per view) vary drastically across different videos and ban statuses.

* What additional types of data could strengthen this dataset?

Timestamp data (e.g., upload date) to analyze trends over time.

Topic or category of the video (e.g., politics, health, entertainment) to understand content type.

User demographics (age, location) to explore audience engagement patterns.

Follow count of video authors to normalize engagement (e.g., likes per follower).

Text sentiment scores from video transcripts to correlate with claim/opinion status.