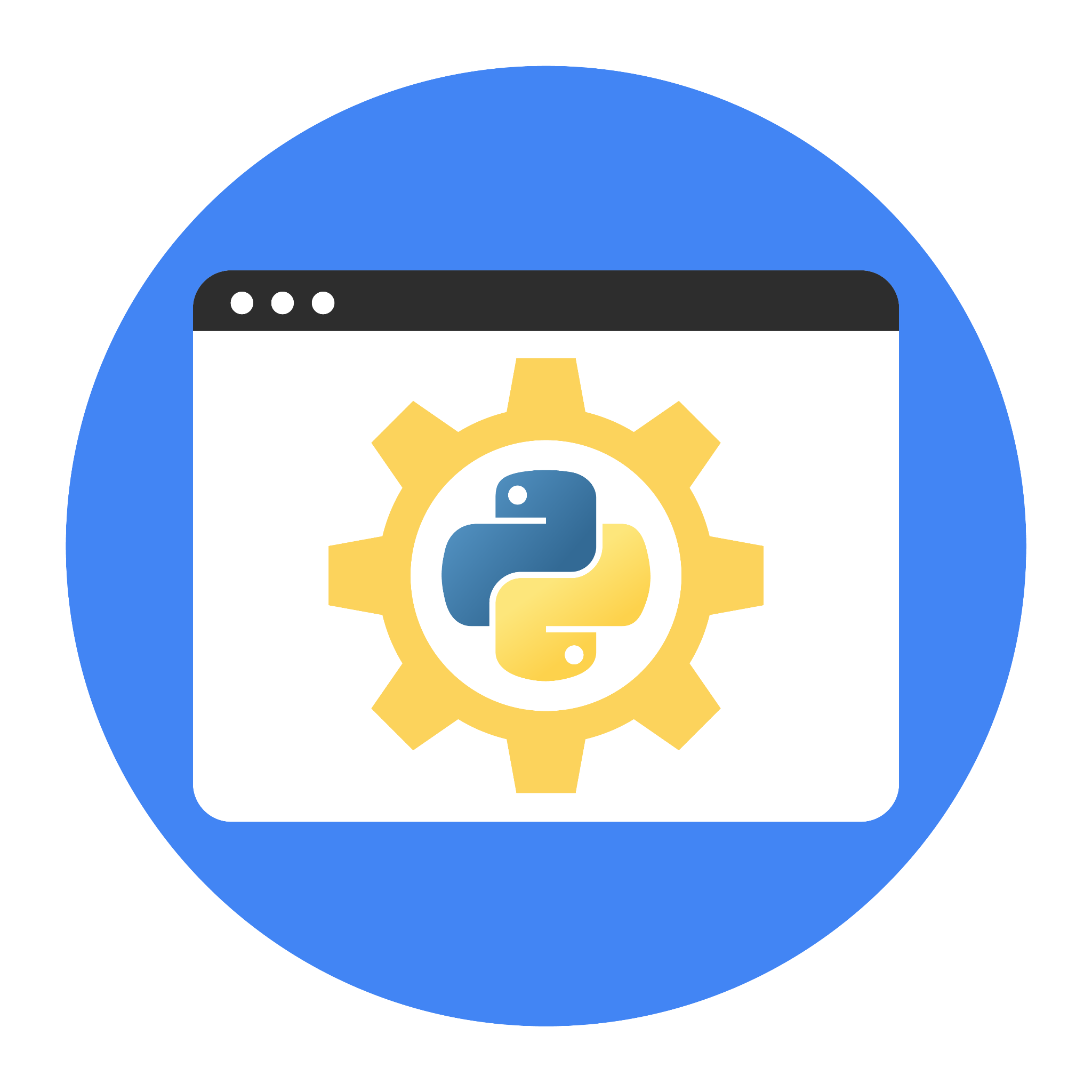
**Course Two**

# 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~~

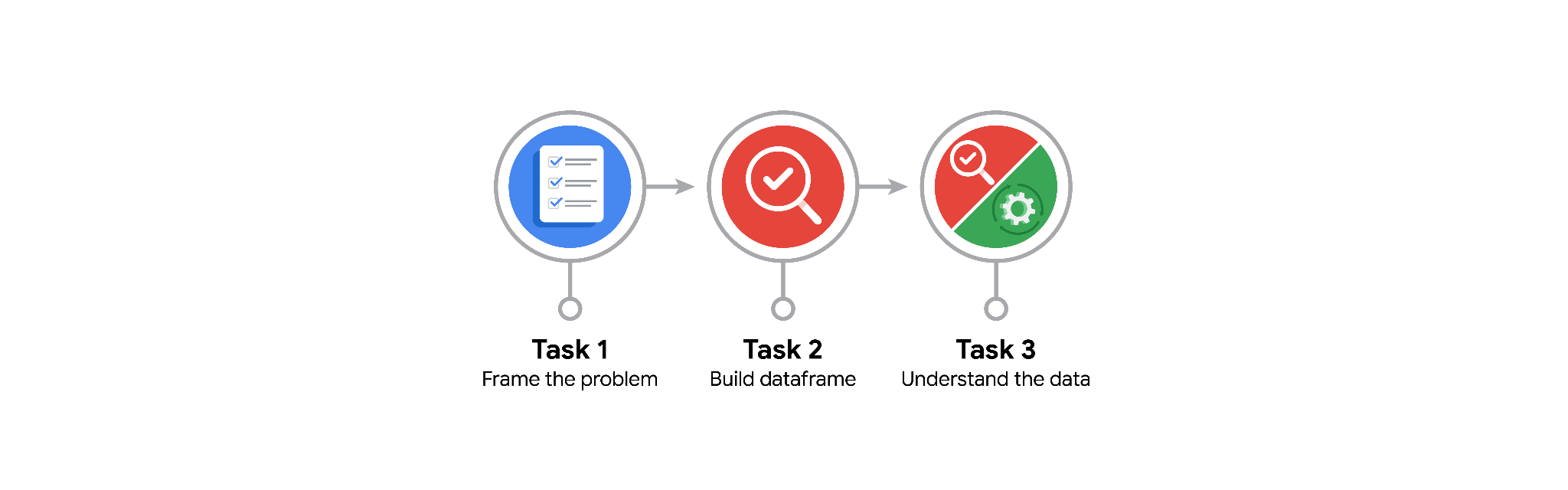
# Relevant Interview Questions

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?

**Preparation**: Review column descriptions, dataset structure (df.info()), and variable types.

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

**Resources**: Use the course's follow-along code notebooks and video walkthroughs as references.

* What are some additional activities a resourceful learner would perform before starting to code?
* **Additional activities**: Research common issues with NYC taxi data (e.g., outliers in fare/tip amounts, null values, duplicate entries).

**PACE: Analyze Stage**

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

**Data Sufficiency**: Yes, the dataset is rich in information for passenger behavior, location, and trip economics.

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

**Summary Stats**: Use df.describe() and df.info() to examine value ranges and detect anomalies.  
**Unusual Averages**: Example – unusually high trip\_distance or total\_amount may indicate data errors.

* Do the averages of any of the data variables look unusual? Can you describe the interval data?
* **Interval Data**: Variables like trip\_distance, fare\_amount, tip\_amount, and total\_amount are continuous numerical values.

**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?

Investigate trips with total\_amount = 0 or negative.

Investigate trips with trip\_distance = 0 but positive fare.

* What data initially presents as containing anomalies?

Negative or zero values for total\_amount and trip\_distance.

tip\_amount for cash payments (should be 0 or NaN).

* What additional types of data could strengthen this dataset?

Weather data or public event schedules for NYC on those dates.

Driver ID (to assess tipping patterns or fraud detection).

Vehicle type or service type (standard, accessible, etc.).