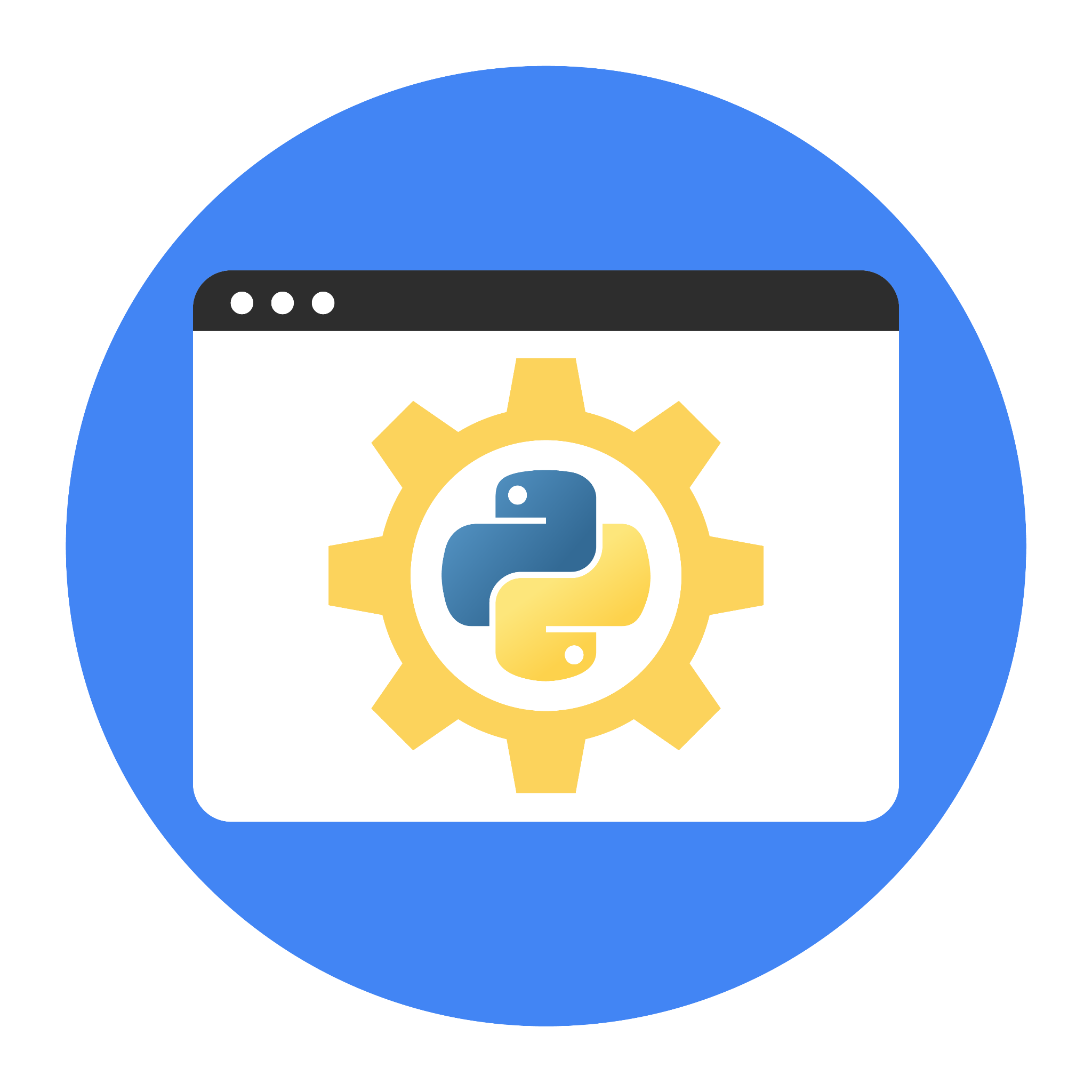
**Course Two**

# 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

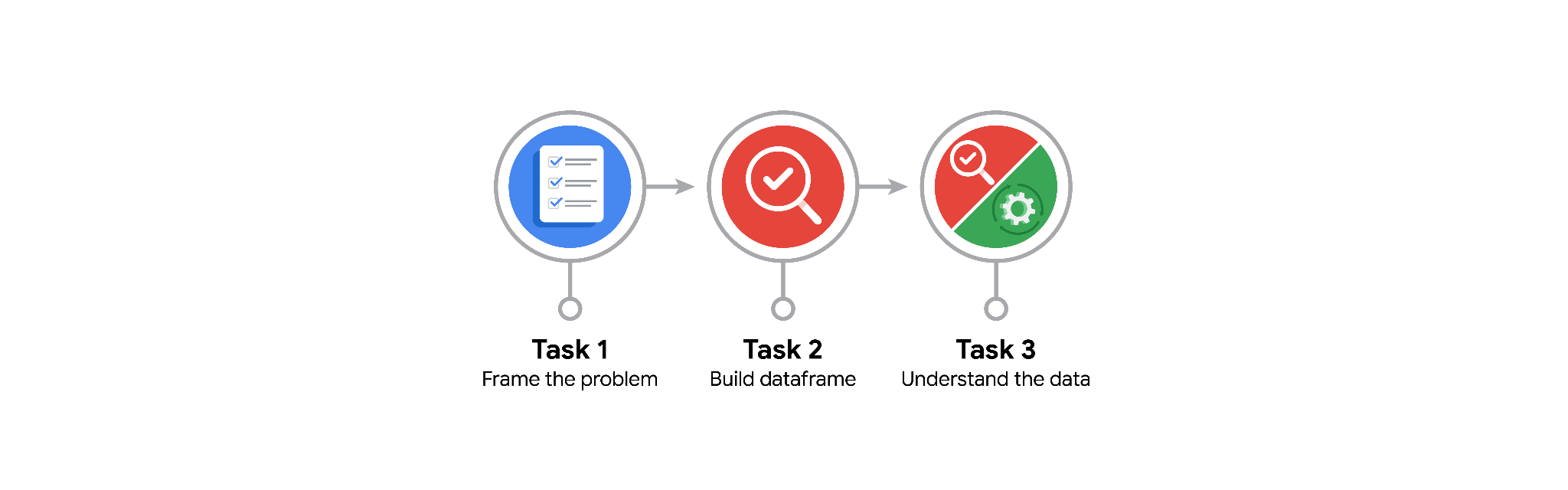
# Relevant Interview Questions

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 for analyzing the provided taxi dataset, I would start by understanding the structure of the data, which includes variables like trip distance, fare amount, passenger count, and timestamps. Reviewing the dataset's column names and definitions would help clarify the content and relevance of each column. I would then inspect the dataset for missing or inconsistent data and evaluate its completeness and format.

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

I would refer to pandas documentation for basic data manipulation, numpy for numerical operations, and matplotlib or seaborn for visualization, should I need to create plots during the analysis. Additionally, I would follow any codebooks that outline best practices for cleaning data, such as dealing with missing values and handling categorical variables.

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

Before coding, a resourceful learner might:

Examine the data dictionary or metadata (if available) to understand each column's meaning.

Explore the source of the dataset to understand how data is collected and any potential biases or issues.

Check if there are known anomalies or challenges with the dataset (e.g., from previous analyses).

Determine which tools or libraries might be most effective for data cleaning and analysis (e.g., pandas for manipulation, matplotlib for plotting).

**PACE: Analyze Stage**

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

The dataset includes important variables such as trip distance, fare amount, payment type, and timestamps, which seem sufficient for initial analysis. These variables can help us understand customer behavior, trip patterns, and the relationship between trip characteristics and fare amounts. However, further investigation is needed to understand any anomalies or potential gaps in the data.

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

To assess the data, I would use the df.describe() function to obtain summary statistics such as mean, median, standard deviation, min, and max values for numerical columns. I would also inspect the df.info() output to check for missing values or inconsistent data types.

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

After calculating summary statistics, I would look for outliers in variables like trip\_distance and total\_amount. For example, extremely high or low values might indicate data issues (e.g., negative trip distances or unusually high fares). The interval data (such as trip duration or fare amounts) should fall within reasonable ranges, and any outliers or anomalies should be flagged for further investigation.

**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 recommend investigating any missing or null values, especially in critical columns like trip\_distance and total\_amount. Additionally, any anomalies or outliers identified in variables such as trip\_distance, total\_amount, and tip\_amount should be examined to determine whether they result from data entry errors, special cases, or legitimate outliers. Data consistency across columns (e.g., ensuring fare amounts align with trip distances) should also be confirmed.

* What data initially presents as containing anomalies?

The trip\_distance column may contain extremely large values (e.g., 100 miles or more), which could indicate either data errors or unusual trips. Similarly, the total\_amount column might include very high or low values, suggesting possible issues such as voided transactions, discounts, or incorrectly recorded fares. Other columns like tip\_amount could also contain anomalies, especially if the values are negative or excessively large.

* What additional types of data could strengthen this dataset?

Additional data that could strengthen the dataset includes:

Weather data: Temperature, precipitation, or other weather conditions could impact trip demand and fare amounts.

Traffic data: Information about traffic conditions or routes taken would be helpful to analyze trip durations and fare amounts.

Customer demographics: Data on customer types (e.g., regular vs. occasional riders) might provide insights into trip behavior.

Location data: GPS data or pickup/dropoff locations could help analyze trip patterns across neighborhoods and refine route optimization.

By following the PACE framework, I will ensure that I systematically approach the analysis and make well-informed decisions as I prepare the data for more advanced tasks like exploratory data analysis (EDA), feature engineering, and model building.