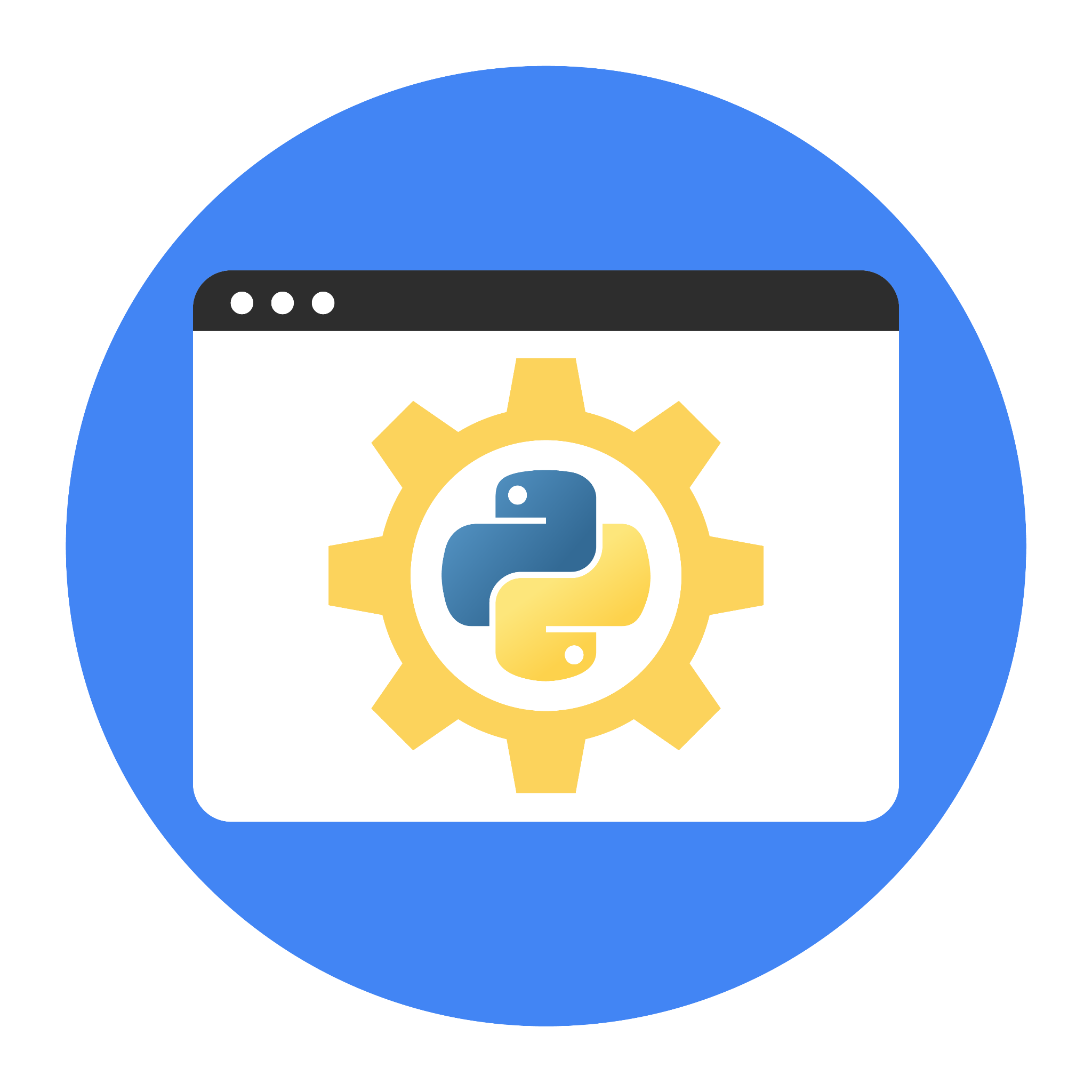
**Course Two**

# 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

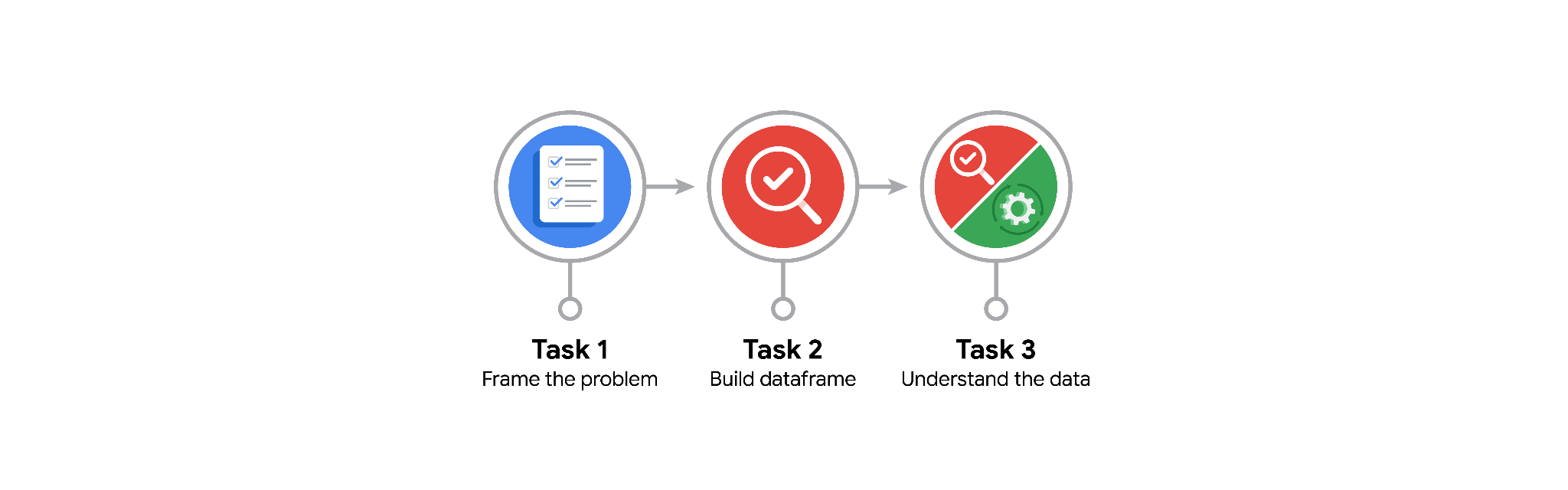
# Relevant Interview Questions

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?

Start by exploring the dataset using tools like `.info()` and `.head()` in Pandas to inspect data types and column names.

Create a data dictionary and note key variables (e.g., `fare\_amount`, `trip\_distance`, `pickup\_datetime`, `passenger\_count`).

Understand which columns are most relevant for predicting fares.

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

- Pandas for data cleaning and exploration

- NumPy for handling numerical operations

- Matplotlib & Seaborn for data visualization

- Scikit-learn for building predictive models

- Documentation from NYC TLC for understanding domain-specific features

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

- Research NYC taxi fare rules and surcharge structures

- Explore sample ride data using Tableau or another visualization tool

- Identify missing values and outliers

- Create a plan for how to clean and prepare the data

- Note down assumptions and potential business rules

**PACE: Analyze Stage**

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

Yes. Variables like `pickup\_datetime`, `trip\_distance`, `fare\_amount`, and location coordinates are sufficient to develop a fare prediction model.

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

Use `df.describe()` to summarize statistics. Use `df.min()` and `df.max()` or visualizations like histograms and boxplots to explore data distributions and spot potential anomalies.

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

Averages of `trip\_distance` or `fare\_amount` might look skewed due to outliers (e.g., extremely high fares or long trips). These columns are continuous interval data. Variables like `passenger\_count` are discrete and should be checked for unrealistic values (e.g., >6 or =0).

**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 data quality for:

- Pickup and drop-off coordinates

- `fare\_amount` outliers

- Trips with zero or negative `trip\_distance`

- Inconsistent or missing values in key fields

* What data initially presents as containing anomalies?

- Negative or zero `trip\_distance`, `fare\_amount`, or ride duration

- Locations outside the NYC boundary

- Unusual `passenger\_count` values (e.g., 0 or above 6)

- Rides with excessively high fares ($500+)

* What additional types of data could strengthen this dataset?

- Weather conditions (rain, snow, etc.)

- Public holidays and day of the week

- Traffic conditions or zone congestion data

- Toll and surcharge data based on time or location