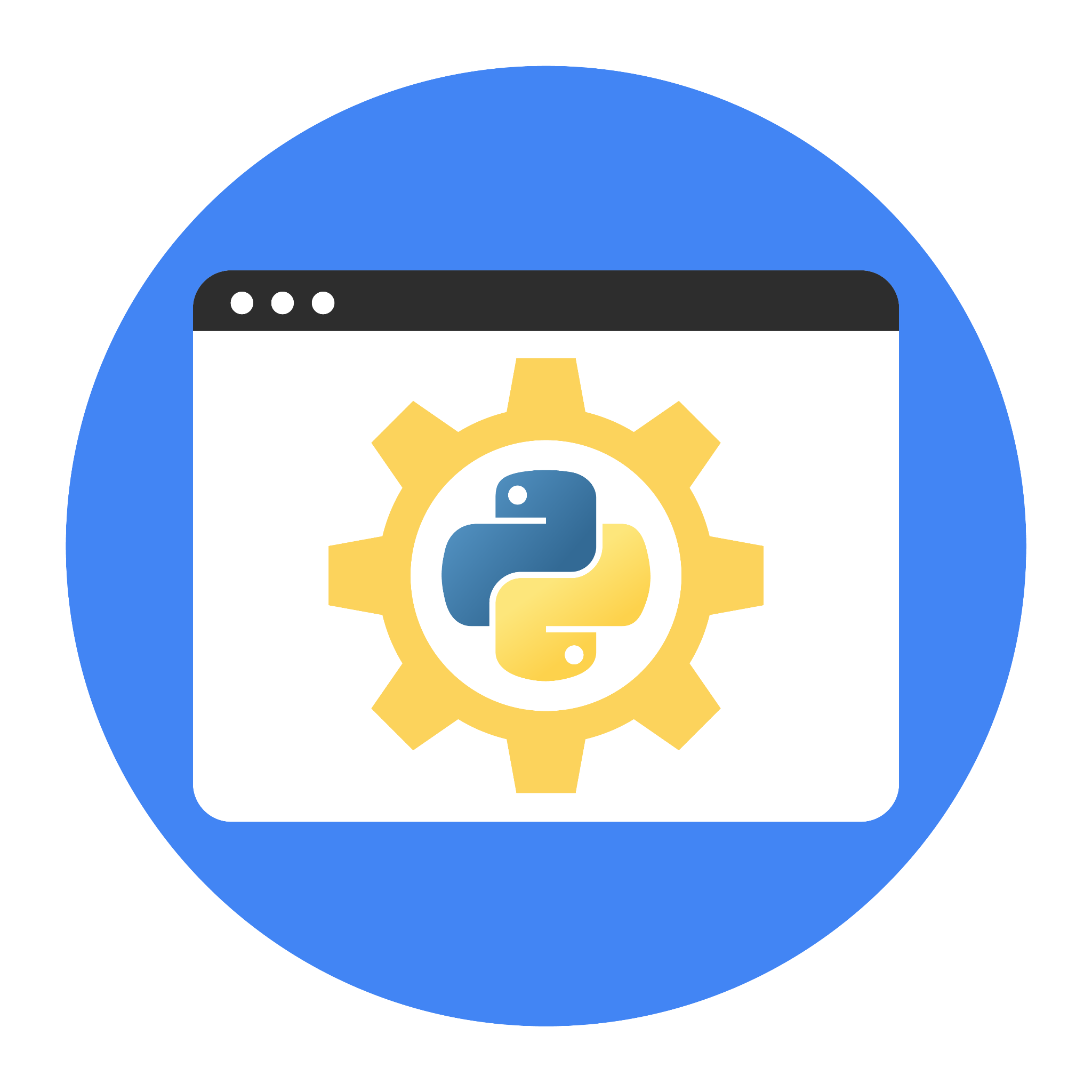
**Course Two**

# 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

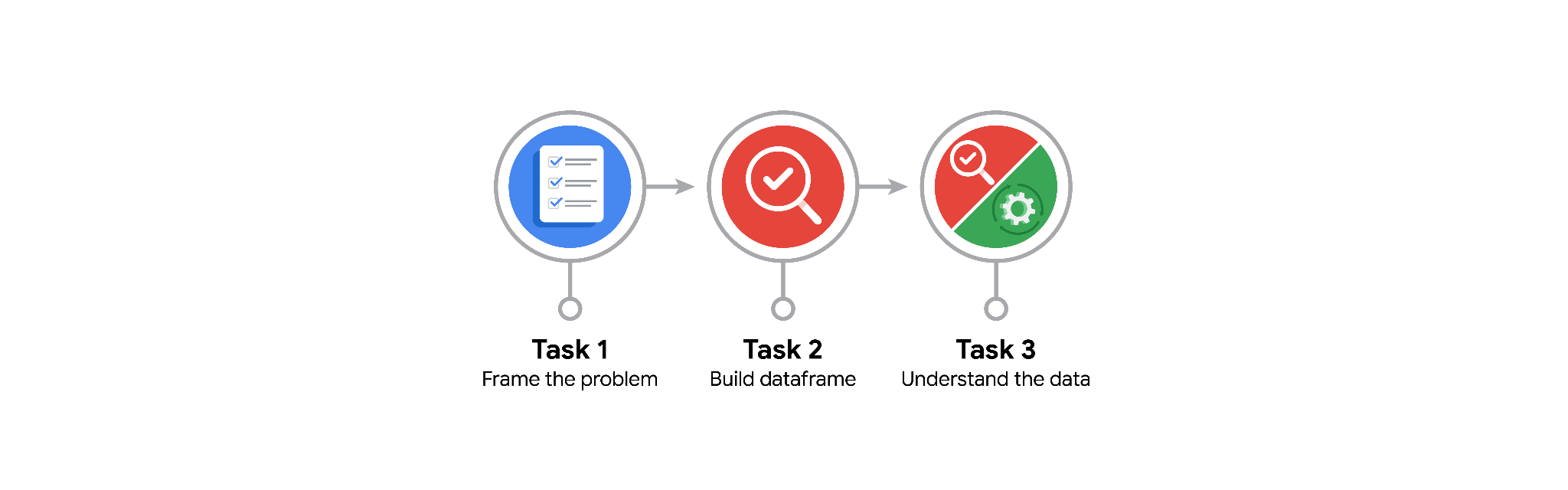
# Relevant Interview Questions

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?

Review the structure of the dataset, including column names and their types (dtypes).

Understand the meaning of key variables such as drives, driven\_km\_drives, driving\_days, label, and device.

Identify which variables are numerical, categorical, or potentially missing.

Review project goals: investigate user churn behavior based on app usage patterns.

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

Follow-along notebooks from the “Get Started with Python” course, especially those on:

Data types and conversions

Missing value handling

Groupby operations

Descriptive statistics (.describe(), .median(), .value\_counts())

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

Scan for outliers and anomalies in numerical variables (e.g., very large values in driven\_km\_drives)

Review distribution of key categorical values (e.g., check value\_counts() for label and device)

Sketch out possible questions or hypotheses (e.g., do heavy users tend to churn more?)

Review example projects and real-world case studies on churn prediction for inspiration.

**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 includes key indicators of user behavior (e.g., drive count, distance, duration) and device type. These are relevant for understanding user engagement and churn.

However, additional variables like time since account creation, in-app feedback, or user occupation could strengthen the analysis.

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

Use .describe() to quickly get summary statistics for numerical columns.

Use .min(), .max(), and .median() by group (df.groupby('label')) to assess variation.

Plot histograms or box plots to visualize distributions and detect skewness or outliers.

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

Yes. Some variables like driven\_km\_drives have extremely high maximum values (e.g., over 21,000 km), which suggests outliers.

Median values are preferred over means here, as they provide a more robust summary for interval data.

The data shows users with high engagement (many drives, long distances), likely a skewed subset of the general driving population.

**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 if users with extremely high activity levels (e.g., many drives in few days) are a specific user group (e.g., truck drivers).

Check if any behavioral patterns differ between users who churned and those who were retained (e.g., drives per day, distance per drive).

Consider enriching the data with external information like user type, trip purpose, or feedback ratings.

* What data initially presents as containing anomalies?

driven\_km\_drives has a maximum over 21,000 km, which is unusually high.

Users with very few driving days but very high number of drives and kilometers suggest intensive driving periods that may not reflect typical behavior.

* What additional types of data could strengthen this dataset?

User metadata: sign-up date, location, occupation (e.g., professional drivers).

Trip-level data: start/end times, trip purpose.

In-app engagement: frequency of using features like navigation, traffic reports.

Qualitative data: user feedback, satisfaction scores.