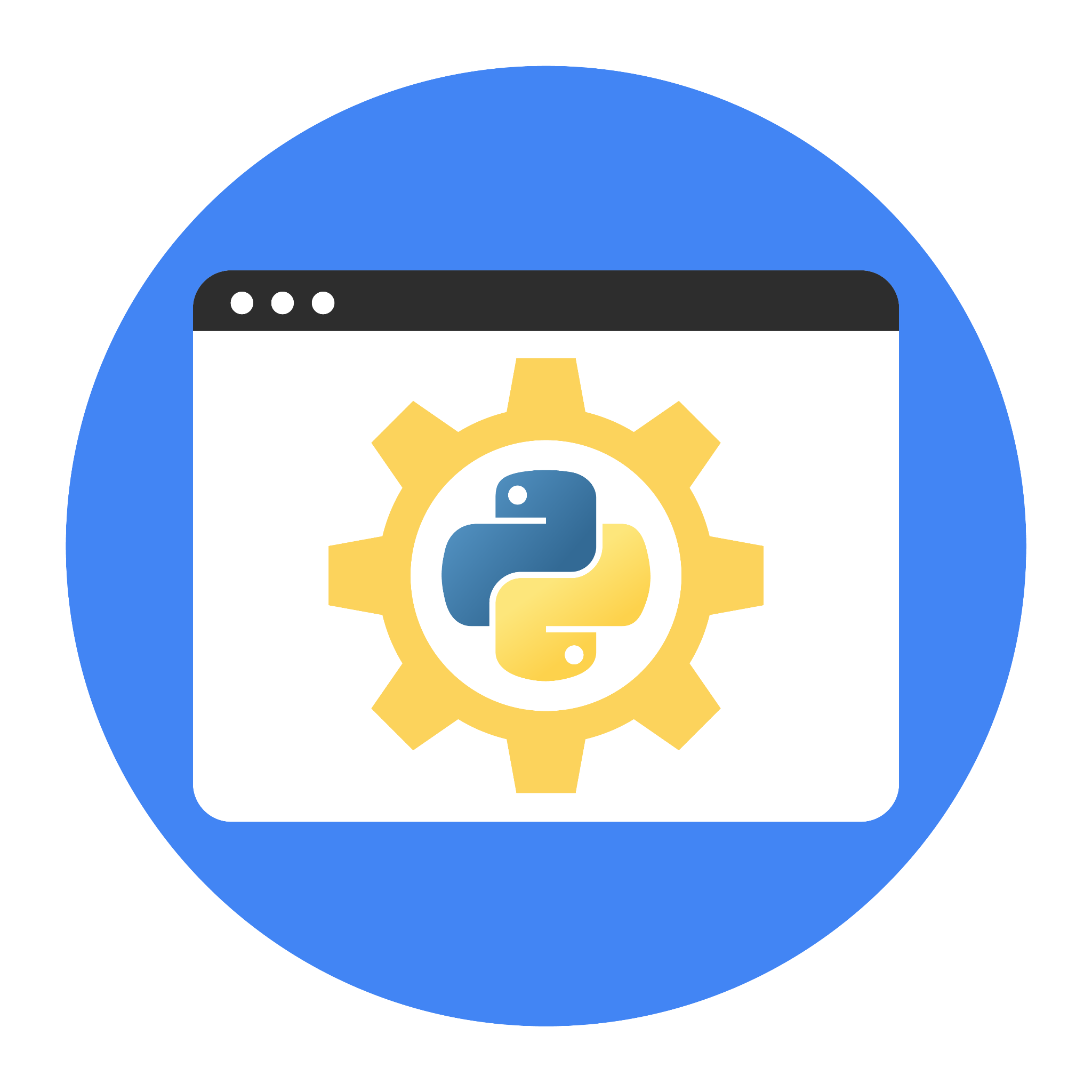
**Course Two**

# 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

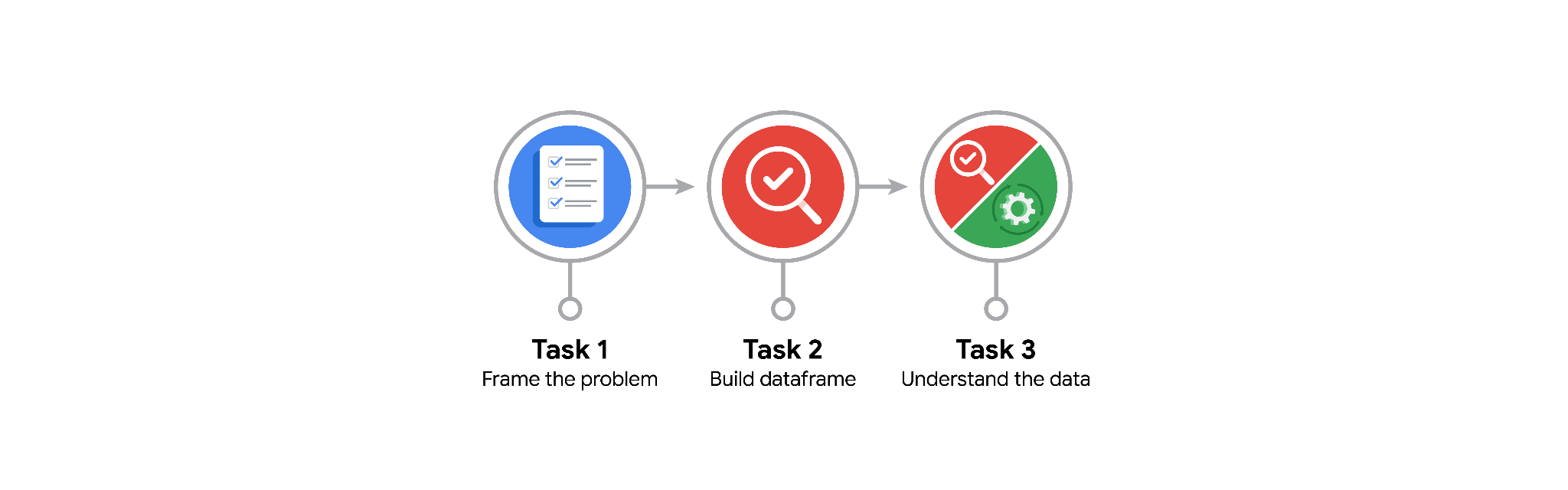
# Relevant Interview Questions

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 effectively prepare for understanding and organizing the information, a structured approach is best: begin by internalizing the project goals and specific Course 2 tasks via provided documents and emails, then dedicate time to thoroughly comprehending the data variables using the Data Dictionary, followed by establishing the necessary Python environment with required libraries, using the PACE framework itself as an overall guide for structuring the subsequent data inspection and preparation tasks.

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

Performing this work effectively involves using the guided "Course 2 Automatidata project lab" Jupyter Notebook as the main follow-along codebook, while simultaneously utilizing course content like instructional videos and readings on pandas operations for self-review, and potentially referencing external pandas documentation for clarifying specific syntax or options when needed.

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

Beyond reviewing provided materials, a resourceful learner could proactively enhance preparation before coding by formulating initial hypotheses about potential data patterns or issues (like outliers or data types needing conversion based on the dictionary), sketching a high-level structure for the analysis notebook to organize the workflow, and seeking out complementary resources such as the NYC Taxi Zone lookup table likely needed for interpreting location IDs later

**PACE: Analyze Stage**

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

While a full assessment requires deeper Exploratory Data Analysis, the initial review of the available variables strongly suggests sufficiency for tackling the fare prediction goal; intuition aligns with the data containing critical potential predictors like travel distance, time/duration factors, location identifiers, and fare/toll amounts, and the preliminary analysis confirms these are populated without missing values, giving confidence that we have the necessary raw materials to build and test initial models, acknowledging further refinement and potential data quality handling will be necessary.

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

Building summary statistics and assessing the data's range would centrally rely on using the .describe() method in pandas, which automatically calculates key statistical measures including the minimum and maximum values needed to determine the range for numerical columns; this would typically be complemented by using .info() to understand data types and non-null counts, providing a fuller initial statistical picture of the dataframe.

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

Examining the averages (mean values) from the initial summary statistics might reveal potentially unusual results, particularly for trip\_distance, fare\_amount, and calculated duration, especially when considered alongside their minimum values (like 0.0 for distance) and very large maximum values indicating outliers; this suggests the mean might be skewed and perhaps less representative than the median. Regarding the data scales, the key quantitative variables used in analysis (distance, duration, fares, counts) are best described as ratio-scale data, which is a type of interval data possessing a true zero point, allowing for meaningful comparisons of both differences and ratios, while identifiers like LocationIDs or Rate Codes are categorical.

**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?

Given this preliminary review, I would initially recommend further investigation into two key areas before the main EDA begins: first, confirming the precise definitions and ensuring the plausibility of key variables, particularly the target fare variable and the calculated duration; second, addressing the immediate data integrity flags, specifically the trips showing zero distance and the potential influence of the extreme upper-range values observed for several core numerical metrics.

* What data initially presents as containing anomalies?

To generate comprehensive summary statistics and evaluate the minimum/maximum range for the dataset, the primary method involves applying the pandas .describe() function directly to the DataFrame; this efficiently computes essential statistics like count, mean, standard deviation, minimum, maximum, and quartile values for all numerical columns simultaneously, with the 'min' and 'max' rows in the output explicitly providing the data range for assessment.

* What additional types of data could strengthen this dataset?

A review of the variable averages generated by .describe() might flag certain fields as unusual, primarily because the presence of known data anomalies (e.g., trip\_distance minimum of 0.0) and expected significant outliers (in fares, distance, duration) can heavily distort the mean value away from the typical central tendency indicated by the median. In terms of describing the interval data, the essential numerical variables for this project—distance, fares/tolls/tips, passenger count, and calculated duration—all function as ratio-scale data, meaning they have meaningful intervals between values and a true, non-arbitrary zero point, distinguishing them from the categorical codes used for IDs and types.