**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 understanding and organizing the TLC data, I'll start by reviewing the project’s goal — predicting taxi fares using real-world trip data. Focus on what variables are available and relevant, then outline the steps I’ll take: loading the data, inspecting its structure, cleaning it, and preparing it for analysis.

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

The most useful resources will be the course’s Jupyter notebook, any sample code provided, the pandas and Numpy reference guides, and the data dictionary. These will guide me through loading, inspecting, and working with the data efficiently.

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

Before coding, it’s helpful to explore the dataset structure using .head(), .info() to spot any issues like missing values or incorrect types. I should also probably think about what features are likely to impact fare amounts and consider reviewing similar examples of fare prediction to get ideas for how to approach this analysis.

**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 available data includes all the key details needed to estimate fares ahead of time. With pickup and drop-off times, trip distance, and fare amount, we have enough context to identify patterns and make informed predictions for future trips. Some cleaning may be needed to improve accuracy, but the core data is in place.

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

You can use the .describe() method in pandas to generate a summary of each numeric column — including count, mean, min, max, and quartiles. You could also use .min() and max() for specific variables you want to inspect, or spot outliers or incorrect entries.

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

Some average values may look unusual — for example, if the average fare or distance is unexpectedly low or high, or even negative, it could point to invalid or outlier trips. There isn’t really any interval data in this dataset. Most of the numbers — like trip distance, fare amount, and trip duration (which I derived from pick-up time and drop-off time) — all have a clear starting point of zero and can be compared directly. For example, a 10-mile trip is twice as long as a 5-mile trip. That kind of data is more specific than interval data and is often called ratio data, which is better for doing calculations like averages or comparisons.

**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 checking for trips that have suspicious values — like zero or negative fare amounts, zero passengers, or trips with very short or very long durations. These could be errors, and they may affect any future analysis. Also, I’d suggest reviewing whether pickup and drop-off locations are coded properly and if we need to map them to actual neighborhoods for clearer insights.

* What data initially presents as containing anomalies?

Some examples of data that look unusual:

* Fare Amounts: Some trips may have a negative fare, 0 fare or extremely high fares (1200) that seem unrealistic.
* Trip Distance: Values of 0 miles with a valid fare amount could be mistakes.
* Trip Duration: Trips lasting just a few seconds or several hours (23) may not be valid, especially in a typical city (NYC) environment.
* Passenger Count: Some trips show 0 passengers, which doesn't make sense for a taxi ride.

These kinds of values may be outliers or data entry errors and may need to be filtered or handled before doing deeper analysis.

* What additional types of data could strengthen this dataset?

A few extra types of information could improve the dataset:

* Traffic conditions at the time of the trip — this could explain longer durations or higher fares.
* Surge pricing or fare rules — if available, these would help explain fare variations.
* Weather data — this can a bit excess maybe, but rainy or snowy days might affect trip time and demand.

Adding this kind of context would make fare estimation more accurate and insightful.