**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?

First loading the data, reviewing the Data Dictionary, understanding what each column represents, deriving a summary and understanding the specifics of the data contained. All of this data can be organized in jupyter notebook cells and the findings can then be shared in an executive summary.

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

I can go back to my notes, or review the course data in the analytics program. I can always refer to Pandas documentation for more information.

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

Understanding the problem by researching, familiarizing with libraries like Pandas or Matplotlib.

Exploring the dataset by reviewing structure and a sample portion.

Setting up the environment, ensuring tools and packages are ready.

**PACE: Analyze Stage**

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

If the dataset is complete and of good quality and the variables within are directly relevant to the project problem, then it’s safe to say that the available information will be sufficient.

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

Using the describe() method will get us key statistics such as count, mean, standard deviation, min, and max for each numerical column.

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

The maximum fare amount is a much larger value ($1000) than the 25-75 percent range of values. Also, it's questionable how there are negative values for fare amounts. Averages seem to be normal but the min and max values for fare amount are unusual.

**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?

Check for any missing or null data points across columns. Investigate potential outliers that may distort analysis, especially in numerical data. Confirm that the available variables align with the business goals or questions we aim to answer.

* What data initially presents as containing anomalies?

Extremely long or short trip durations that deviate significantly from the average. ntries where the fare amount is negative or zero may suggest data entry mistakes or could indicate rides that were not completed or were canceled.

* What additional types of data could strengthen this dataset?

Adding weather data can help analyze how weather affects taxi demand, trip durations, and passenger behavior. Data on local events, concerts, festivals, and public holidays can help understand spikes in demand and adjust service levels accordingly. Information about drivers, including their experience, ratings, and history, can help analyze the impact of driver performance on ride quality and passenger satisfaction.