**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?

|  |
| --- |
| * Clarify Objectives:   + Review the project goals to ensure you understand the desired outcomes (e.g., building a regression model to estimate taxi fares).   + Identify what success looks like for the client (e.g., model accuracy, ease of integration). * Categorize Information:   + Data: Available datasets and external data sources.   + Stakeholders: Roles, expectations, and how they will interact with deliverables.   + Deliverables: Required outputs, such as EDA reports, regression models, and final presentations. * Document Assumptions and Open Questions: Identify gaps in information or ambiguities, such as the quality of the TLC dataset or specific performance metrics desired by the client. |

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

|  |
| --- |
| * Data Dictionaries and Documentation: Obtain and review any codebooks or metadata provided with the TLC dataset to understand the structure, variable definitions, and data types. Examples include:   + Definitions for variables such as trip distance, fare amount, pickup/drop-off locations, and time of day.   + Codes for categorical variables (e.g., payment methods, zones). * Statistical References: Keep a reference for statistical methods, such as:   + Hypothesis testing and regression analysis techniques.   + Resources like "Introduction to Statistical Learning" or online platforms like Khan Academy or StatQuest. * Coding Standards: Review guidelines or best practices for data analysis in the coding language/tool being used (e.g., Python or R). |

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

|  |
| --- |
| * **Review Project Background**:   + Research the business context, such as the NYC TLC’s operations, and how fare estimation impacts customers and operations.   + Understand external factors that could influence the data, like traffic patterns, weather, or peak hours. * **Familiarize with Tools**:   + Ensure proficiency in the tools to be used for the project (e.g., Python, Jupyter Notebooks, RStudio, Tableau).   + Set up and test the development environment to ensure all libraries and dependencies are installed. * **Explore Similar Projects**:   + Review case studies or examples of regression modeling projects, particularly in transportation or similar industries, to gain insights and best practices. * **Plan for Collaboration**:   + Set up shared workspaces or version control systems (e.g., GitHub, Google Drive) to facilitate teamwork and track changes. * **Anticipate Challenges**:   + Identify potential roadblocks, such as missing data or ambiguous variable definitions, and plan mitigation strategies. * **Draft a Data Strategy**:   + Define a clear process for data handling, including cleaning, transformation, and validation steps.   + Outline data privacy considerations and compliance with any relevant regulations. |

**PACE: Analyze Stage**

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

|  |
| --- |
|  |

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

|  |
| --- |
|  |

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

|  |
| --- |
|  |

**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?

|  |
| --- |
|  |

* What data initially presents as containing anomalies?

|  |
| --- |
|  |

* What additional types of data could strengthen this dataset?

|  |
| --- |
|  |