**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 understand and organize the provided taxi cab information, review the project proposal and client requirements, explore dataset documentation, communicate with team members, explore a sample of the data, identify key stakeholders, and define key metrics.

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

The follow-along guide is an annotated Jupyter notebook organized to match the content from each module of instructional videos. It contains the same code shown in the videos for that module, additional information to explain concepts, reasons for code structure, and tips for running the code.

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

Review pre-reading materials, Set up the development environment, Familiarize themselves with tools and IDEs, Quickly review documentation, Understand the code structure, Explore the dataset if applicable and Read through the follow-along guide.

**PACE: Analyze Stage**

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

Based on the information provided, it seems that the available information is sufficient to achieve the initial goals of the project. The project aims to investigate and understand the user data provided by NYC TLC Agency, preparing it for exploratory data analysis (EDA) and future statistical activities.

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

Building Summary DataFrame Statistics:

Use the describe() function to generate count, mean, std, min, 25%, 50%, 75%, max for numeric columns.

Include data types and missing values to provide a comprehensive overview of the dataset.

Assessing Min and Max Range:

Utilize agg(['min', 'max']) to obtain the minimum and maximum values for each numeric column.

Calculate the range by subtracting the minimum from the maximum to understand the spread or variability in the data.

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

Yes, the some values are significantly higher than the others, based on the results The most expensive rides are not necessarily the longest ones.

**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?

Examine extreme values in variables such as fare\_amount, trip\_distance, and total\_amount. The presence of outliers could significantly impact statistical measures and visualizations. Check for any inconsistencies or errors in the data, especially in columns like passenger\_count, where there are instances of a passenger count of 0.

* What data initially presents as containing anomalies?

There are instances where the fare\_amount variable has negative values, which is not typical for a taxi fare. Further investigation is needed to understand the reasons behind negative fare amounts and whether these entries should be treated as anomalies or errors.

There are instances where the passenger\_count variable has a value of 0. This is unusual as taxi rides typically involve at least one passenger.

* What additional types of data could strengthen this dataset?

Traffic Data:

Include data related to traffic conditions, road closures, and congestion. Understanding traffic patterns can provide insights into travel times, route choices, and potential delays affecting fares.

Holiday and Events Calendar:

Incorporate a calendar of holidays and major events in New York City. This information can be valuable for analyzing how special occasions impact ride demand, fare amounts, and overall travel behavior.