**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 I think I need to have a look at the Data Dictionary to have the understanding of the dataset and its contents one by one. Then learn about the data by basics data exploration such as using describe() method, count\_values...

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

The Examplar responses

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

Reading provided Data Dictionary to fully understand the dataset

**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 dataset contains more than 22,000 rows with non-null values and there are 18 variables. So I think they are enough for us to achieve the goal.

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

I use the describe() method

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

Firstly, the trip\_distance is quite unusual with range from 0 to 33.96 miles while average is 2.9 miles.

Secondly, there are some categorical variables but listed as numerical variables such as VendorID, RateCodeID and Payment\_type so the distribution of these variables doesn't make sense in numerical

Finally, total\_amount sees the usual min of -120 and max of 1200 while mean is around 16.31.

**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 will recommend the manager to investigate about the trips with 0 trip\_distance but high total\_amount (over 100), the trip with negative total\_amount and the two largest total\_amount of 1200 and 450.

* What data initially presents as containing anomalies?

Based on the information provided, it seems that the anomalies in the data are related to the fare amounts. Specifically, there are instances where the fare amount is unusually high (e.g., 1200 and 450) despite the trip distance being relatively short (2.6 miles and 0 miles, respectively). This raises questions about the accuracy of the fare amounts and suggests the presence of anomalies or errors in the data.

* What additional types of data could strengthen this dataset?

I think this dataset is enough for analysis