# 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?

I will use Python to analyze data in a faster, more efficient and more powerful way because it optimizes every phase of the data workflow, from exploring, cleaning and visualizing data to building machine learning models.

For understanding the provided information, I will view and inspect the summary information about the dataframe and check for relevant variables and data types that help store and organize the data.

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

Jupyter notebook, an open-source web application will help perform this work because it creates and shares documents containing live code, mathematical formulas, visualizations and text. You can add comments, annotations, and explanations using markdown syntax. A docstring will also help perform this work.

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

Firstly, a resourceful learner would understand and frame the problem. Have communication with stakeholders and your teammates. This could be asking questions about your data, gathering additional sources of information, and active listening.

Do online searches, networking, mentorships. Set a project workflow, and time management.

**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 information looks sufficient to achieve the goal based on my intuition and the analysis of the variables.

The two most important variables that will prove to be helpful for building a predictive model are the trip\_distance and fare\_amount/total\_amount. Building a predictive model based on these two variables will help the users predict their taxi fare in advance of their ride.

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

To build the summary dataframe statistics, the describe() function will be useful. describe() function will return a table of summary statistics for the dataframe which will also help to access 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?

The average tip amount for each passenger count (credit card payments only), for passenger\_count = 0, the tip amount has been calculated. This value looks 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?

After a comprehensive investigation of the dataset, several anomalies and questionable values have surfaced, prompting the need for further investigation before embarking on exploratory data analysis.

1. Zero values: Noteworthy instances of minimum values, specifically zero, have been identified in key variables such as passenger\_count, trip\_distance, and fare amount.
2. Inconsistencies in Time Duration: In some cases, despite a recorded trip distance of 0, the total amount has been calculated. This discrepancy is particularly noticeable when the time period between pick-up and drop-off is only a few seconds.
3. Fare Amount Discrepancies: Anomalies are apparent in cases where the fare amount does not align proportionally with the trip distance. There are instances where rides with shorter distances exhibit considerably higher fare amounts compared to rides with greater distances, and vice versa.
4. Negative values: The dataset contains observations with negative values, even when trip distances are positive.

These anomalies require thorough investigation as negative values in critical variables suggest potential errors in data entry or calculations.

* What data initially presents as containing anomalies?

1. Zero Values: Instances of minimum values, particularly zeros, in passenger count, trip distance, and fare amount variables suggest potential inaccuracies or anomalies.
2. Inconsistencies in Time Duration: Notable instances where a trip distance of 0 is recorded, yet the total amount is calculated, indicating discrepancies in the time duration between pickup and drop-off.
3. Fare Amount Discrepancies: Anomalies are observed when the fare amount does not align proportionally with the trip distance, leading to discrepancies in fare calculations.
4. Negative Values: The presence of negative values, even with positive trip distances, suggests potential errors or inconsistencies in data entry or computation.

* What additional types of data could strengthen this dataset?

The two most crucial variables for constructing an effective predictive model are the "trip\_distance" and "total\_amount". Building a predictive model based on these key variables holds significant promise in providing users with the ability to anticipate their taxi fare ahead of their ride. This model would empower users to make informed decisions about their transportation costs, enhancing the overall user experience and convenience.