**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?

Make sure there are no missing values. Box plots to help identify outliers. Histograms can also be helpful.

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

I have my own guides and templates.

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

I always find it helpful to look at the observations that are at the top and bottom when sorting each column. This is a good way of identifying potentially corrupt data.

**PACE: Analyze Stage**

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

The data should be sufficient to predict fares.

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

I would use Pandas and Numpy tools.

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

The averages all look fine. I don’t think interval data will be very helpful in this case.

**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 would recommend looking into the non-existance of tips when payment type is cash. I would also advise them to tell the TLC to investigate large fare and large tip trips with very low or zero distances.

* What data initially presents as containing anomalies?

High fare, low distance trips. Cash payments with zero tips.

* What additional types of data could strengthen this dataset?

Since we are only targeting the fare, there really isn’t much to this project. This is heavily regulated and easily calculated.