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.

Task 1
Frame the problem

Task 2
Build dataframe

Task 3
Understand the data
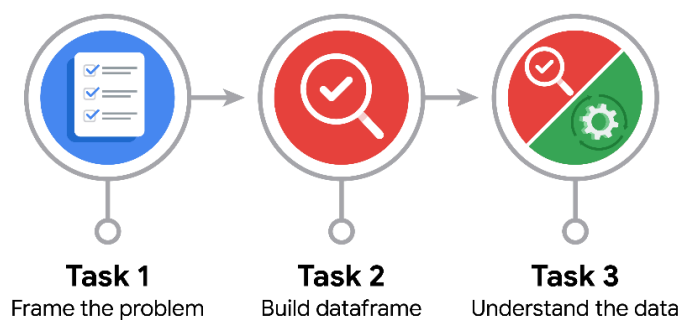
Data Project Questions & Considerations

PACE: Plan Stage

- How can you best prepare to understand and organize the provided information?

    Start by reviewing the NumPy and pandas libraries, and the various methods and attributes it offers. Making sure to use the appropriate functionalities of these libraries at appropriate places.

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

    Jupyter

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

    He/She may revise his/her knowledge about the two most important libraries, NumPy and pandas.

PACE: Analyze Stage

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

> Yes.

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

> I would use the functions present in pandas library such as head(), info(), describe(), etc.

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

> No, the averages seem normal.

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 building the model on total_amount and trip_distance variables.

- What data initially presents as containing anomalies?

  Minimum of fare_amount, extra, mta_tax, improvement_subcharge and total_amount is negative. But since we won't be using them for modeling, they don't pose an issue.

- What additional types of data could strengthen this dataset?

  A customer rating might improve the dataset.