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

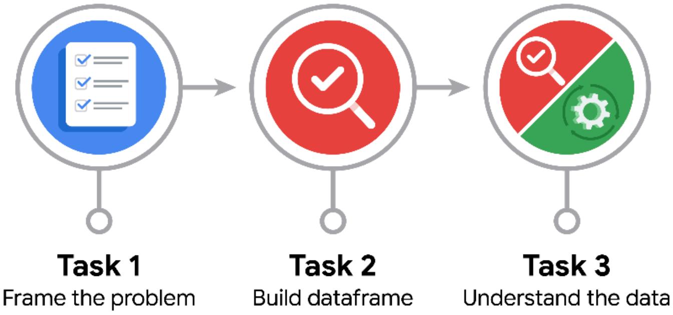
Relevant Interview Questions

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?

By understanding the project objectives.

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

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

PACE: Analyze Stage

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

For a first attempt in defining the fare amount, the current information seems enough.

For instance, correlating trip time (starting and end time) and distance we can infer if there was a traffic jam and how this will affect the fare.

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

By using pandas.describe()

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

Some short trips have high fare amount

We can notice some values are off.

For instance, trips with distance equals to zero; fare amount with a negative number (most likely due to a dispute).

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?

Investigate negative values for total amount (if they are all related to payment disputes)

Investigate trips with zero distance, zero passengers.

- What data initially presents as containing anomalies?

Data related to payment amounts (fare, extra, mta extra, surcharge and total).

Distance and passenger quantity.

- What additional types of data could strengthen this dataset?

Fleet availability, meaning, how many other vehicles are available at the time.