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

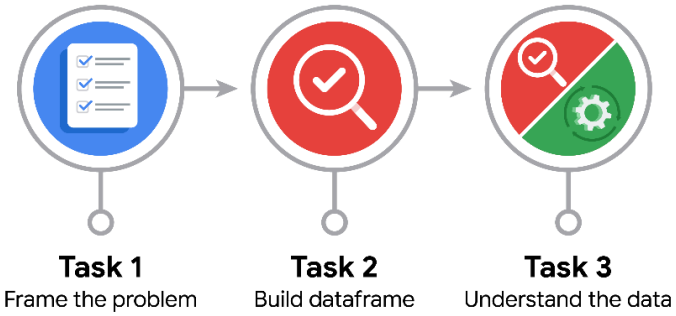
Relevant Interview Questions

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?

Structure the data in a pandas DataFrame to run preliminary data analysis.

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

Reading the pandas/numpy documentation will help with sorting and cleaning the data. A data summary of the dataset column fields will guide data exploration.

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

Think about the roles of the stakeholders involved, and determine how to share insights with technical and non-technical visualizations.

PACE: Analyze Stage

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

There needs to be more clarification on missing values and negative values found on the datasets. These discrepancies may poorly impact the model performance on test data.

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

Dataframe statistics can be accessed with the describe() method. The parameter “include=all” will display statistics for categorical data.

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

The average tip for cash users is 0, which is unusual. The averages are mostly balanced, but they may be impacted by severe outliers.

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 a clarification on when the data was obtained and if it was edited/merged with another dataset. Additional information needs to be provided on the nature of the dataset lacking null values.

- What data initially presents as containing anomalies?

The payment_method is supposed to have 6 variations, but only 4 types can be found. The trip distance contains 0 values, meaning the trip did not occur. The total amount is negative for some values.

- What additional types of data could strengthen this dataset?

Trip_distance and total_amount are the main categories affecting the cab fare. Other data types to include would be location demographics instead of ids for streamlined aggregation opportunities.