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

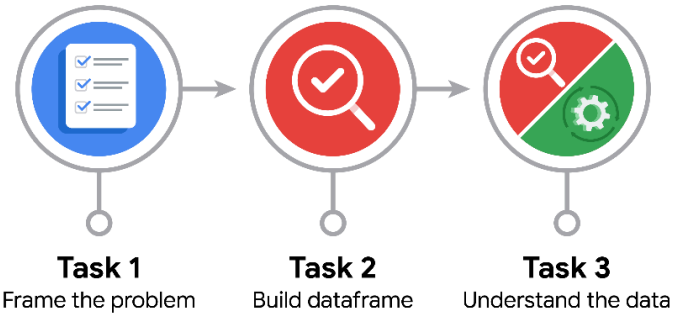
Relevant Interview Questions

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?

Review the data dictionary (if available) to understand the meaning of each column. Get familiar with the context of the data (e.g., NYC TLC regulations, typical taxi operations). Plan the initial steps for data inspection (e.g., loading data, checking data types, basic summary statistics).

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

Pandas documentation for data loading, inspection, and summary statistics. NumPy documentation for numerical operations. Tutorials on data cleaning and initial data exploration. Any provided course materials related to data analysis.

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

Research the data source (NYC TLC) and related datasets. Brainstorm potential questions and hypotheses to explore. Sketch out a plan for data cleaning and exploration. Review common data anomalies that occur within datasets.

PACE: Analyze Stage

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

Initially, the data provides a good starting point, but further investigation is needed due to anomalies. The datetimes, trip distance, and total amounts will provide valuable information.

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

Use df.info() to check data types and null values. Use df.describe() to get summary statistics (mean, min, max, quartiles). Use df.sort_values() to examine extreme values.

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

The average total_amount and trip_distance should be compared to expected values. The existence of negative values within the money columns is highly unusual. Interval data (e.g., trip_distance, total_amount) can be described by its range, quartiles, and distribution shape.

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 the negative values in fare_amount, extra, mta_tax, and total_amount. Examine the outliers in trip_distance and RatecodeID. Convert the datetime columns to datetime objects. Investigate the reason behind the passenger counts of 0.

- What data initially presents as containing anomalies?

Negative values in monetary columns. Zero values in passenger_count and trip_distance. Outliers in trip_distance and RatecodeID.

- What additional types of data could strengthen this dataset?

Weather data (to see how weather affects taxi usage). Geographical data (maps, borough information) for better location analysis. Event data (concerts, sports events) to understand peak demand. Information about taxi driver shifts, or driver information. Information about traffic conditions.