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~~

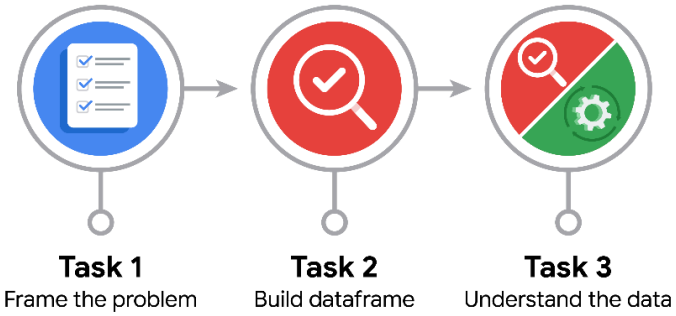
Relevant Interview Questions

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?

I began by reviewing the dataset structure using `df.info()` and `df.describe()` to understand column types, missing values, and ranges. I also consulted documentation on NYC TLC data fields to understand each variable.

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

I used Jupyter Notebook walkthroughs, pandas documentation, and reference code from earlier Python exercises to build cleaning, transformation, and summary logic.

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

I explored sample outputs, read through the column names for domain context, searched for definitions of `payment_type`, and compared data rows to expected real-world taxi behavior.

PACE: Analyze Stage

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

Yes, the data contains key variables like tip amount, payment type, trip distance, and passenger count. These are relevant and sufficient for understanding and predicting tipping behavior.

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

I used `df.describe()` and `.value_counts()` to compute means, medians, and outliers, and sorted columns like `trip_distance` and `total_amount` to find extremes.

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

Yes, the average tip is ~$2.73 for credit cards but $0.00 for cash, indicating missing data. Also, `trip_duration_min` has a negative value, likely a data error. Outliers exist in `fare_per_mile` and `total_amount`.

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 recommend deeper investigation into outliers in `fare_per_mile`, negative trip durations, and filtering for valid payment types before modeling tips.

- What data initially presents as containing anomalies?

Negative fares and total amounts, 0-mile trips, trips over 1000 minutes, and fare-per-mile values over $100 are strong anomalies.

- What additional types of data could strengthen this dataset?

Adding categorical variables like time-of-day (rush hour), weather conditions, driver ID, and drop-off neighborhood zones could significantly enhance prediction accuracy.