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

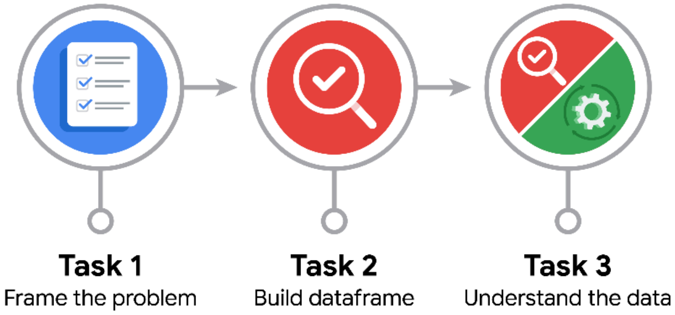
Relevant Interview Questions

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?

To prepare for understanding and organizing the provided taxi cab dataset, I will begin by reviewing the project goals, assigned activities, and expected deliverables, and then load the CSV file into a Jupyter Notebook to inspect its structure while aligning with the PACE framework. Next, I'll load the dataset into a pandas DataFrame and perform initial data profiling—including checking the shape, data types, and missing values—followed by compiling summary statistics to identify any trends or data quality issues. Finally, I will examine specific variables such as trip distance and pickup time, investigate distributions and outliers, and use these insights to plan for deeper exploratory data analysis (EDA).

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

Follow-along codebooks could be tutorial notebooks, guides, or resources that you follow while working on the lab and tasks related to the 2017_Yellow_Taxi_Trip_Data.csv.

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

A resourceful learner would carefully read through the project instructions and objectives, explore any available documentation or data dictionaries, and research the domain (in this case, taxi trip data) to understand real-world context. They might sketch out a workflow or data pipeline, prepare a list of questions to guide their analysis, review similar case studies, ensure all necessary tools and libraries are installed, and organize their project folders. Additionally, they may identify relevant code snippets or notebooks to reference and set up version control (e.g., using Git) before writing any code.

PACE: Analyze Stage

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

Yes, based on the initial exploration of the dataset, the available information appears largely sufficient to achieve the project's goal of analyzing taxi trip behavior and generating insights. The dataset includes key variables such as trip distance, pickup/dropoff timestamps, passenger count, fare details, and location IDs—all essential for understanding trip patterns and customer behavior. However, a few data quality issues (e.g., zero passengers, negative fares, outliers) will need to be addressed during cleaning. Assuming those are resolved, the dataset provides a strong foundation for effective analysis and visualization.

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

To build summary statistics, I first imported the pandas library and loaded the CSV data into a DataFrame called df. I used df.info() to quickly check the structure of the dataset, including column names, data types, and whether there are any missing values. Then, I ran df.describe() to get an overview of the numeric columns — including count, mean, standard deviation, min, max, and quartiles. If I want to isolate the minimum or maximum for specific columns, I can also use df['column_name'].min() and df['column_name'].max().

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

Some of the **averages do look a bit unusual**. For example:

- passenger_count has an average of **1.64**, which makes sense since most rides probably have 1–2 passengers, but the presence of non-integer averages hints that some values might not be accurate (like rides with 0 passengers).
- fare_amount averages **$13.03**, but the minimum value is **-120**, which is clearly an error or outlier. This could affect the reliability of the average.
- Similarly, the tip_amount average is **$1.83**, but the maximum is **$200**, which might be valid but still extreme. This suggests a right-skewed distribution.

Interval Data:

Many of the numeric variables are **interval data**, meaning they are continuous and ordered, but without a true zero (e.g., temperature-like). In this dataset:

- Columns like trip_distance, fare_amount, tip_amount, and total_amount are **interval** (or more precisely, **ratio** scale, because they have a true zero and allow meaningful multiplication/division).
- They can be analyzed using statistical methods like mean, standard deviation, histograms, and boxplots.

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?

Before diving into exploratory data analysis (EDA), I would recommend investigating the following key areas:

- **Data Quality Issues**: As noted, there are potential data quality problems such as negative fares, zero passenger counts, and outlier values for tip amounts and fare amounts. These need to be addressed first to ensure that the analysis yields reliable insights.
- **Datetime Handling**: The tpep_pickup_datetime and tpep_dropoff_datetime columns are currently in object format, and they should be converted to datetime format for accurate time-based analysis. It's essential to validate the consistency and completeness of the date and time fields.
- **Payment Types and Data Consistency**: Further investigation into the payment_type and its distribution would help identify any unusual payment patterns or discrepancies, especially for cases like "No charge" or "Voided trip."

- **Missing or Inconsistent Data**: Although there are no missing values, some inconsistencies (e.g., outliers in monetary values) might still affect the analysis. It would be good to review the data completeness and any edge cases that might not be apparent at first glance.

- What data initially presents as containing anomalies?

Anomalies in Data:

- **Negative Fare and Tip Amounts**: The fare_amount, tip_amount, and total_amount columns have negative values, which are clearly incorrect for the dataset.
  - **Example**: fare_amount has a minimum of -120 and tip_amount has values like -0.5, both of which are unrealistic for a taxi ride.
- **Zero Passenger Count**: The passenger_count column contains a minimum value of 0, which is illogical for a legitimate taxi ride. This could indicate canceled trips or erroneous entries.
- **Trip Distance**: The trip_distance column has values starting from 0, which could represent no trips, or it could be a data entry issue.
- **Extreme Outliers**:
  - total_amount and fare_amount have maximum values in the hundreds (e.g., fare amount up to 999.99 and total amount up to 1200.29), which may be valid but warrant investigation to check for rare events or erroneous entries.

- What additional types of data could strengthen this dataset?

Additional Data to Strengthen the Dataset:

- **Geospatial Data**: Adding latitude and longitude information for both pickup and drop-off locations would enhance the dataset, making it possible to analyze ride patterns geographically (e.g., popular routes, traffic analysis).
- **Weather Data**: Integrating weather data for the dates and times of the rides could provide insights into how weather conditions (e.g., rain, snow) affect ride volume, trip distance, or tips.
- **Driver Data**: Information on the driver (e.g., experience, ratings) could offer more context for analyzing the performance of different drivers, and potentially explain variations in fare amounts or customer satisfaction.
- **Time of Day**: While the dataset includes datetime columns, breaking this down further into more

granular time segments (e.g., day part: morning, afternoon, evening) could allow for deeper insights into patterns in demand, trip duration, and tips.

- **Customer Data**: If available, adding customer-related information (e.g., loyalty programs, return riders) would improve the model's ability to predict fare amounts or tip behavior, and uncover patterns in customer behaviors.