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

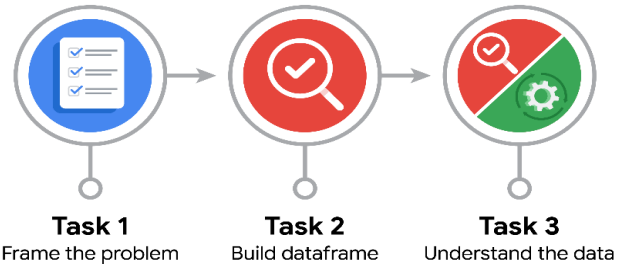
Relevant Interview Questions

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?

Before diving into the data, it's essential to review the project proposal and dataset documentation thoroughly. Familiarize yourself with the column names, descriptions, and any potential data quality issues mentioned. Additionally, brainstorming key questions or hypotheses related to the project goals can help guide your analysis.

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

Reviewing Python documentation and tutorials on data manipulation, such as pandas and numpy, would be beneficial for performing data organization and analysis tasks. Additionally, exploring sample code notebooks or tutorials specifically focused on exploratory data analysis (EDA) can provide insights into common techniques and best practices.

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

A resourceful learner might engage in discussions with team members to gain insights into the project's context and objectives. They could also conduct preliminary research on similar projects or datasets to understand potential challenges and approaches for analysis. Moreover, creating a detailed plan outlining the steps and milestones for data exploration and analysis can help streamline the process and ensure thoroughness.

PACE: Analyze Stage

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

Based on the dataset description and variables provided, it seems likely that the available information will be sufficient to achieve the project goal of analyzing taxi trip data.

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

To build summary statistics, I would use the describe() function in pandas to generate statistics such as mean, standard deviation, min, and max for numerical columns. To assess the min and max range of the data, I would directly use the min() and max() functions in pandas on specific columns.

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

I would examine the averages of numerical variables like trip distance, fare amount, and tip amount to identify any unusual patterns or outliers. Interval data refers to numerical data where the difference between any two values is meaningful and consistent across the entire range. In this dataset, variables like trip distance, fare amount, and tip amount represent interval data as the numerical differences between values have consistent meanings.

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.

In the context of this project, the Construct stage may not be directly applicable as it typically involves creating new data elements, features, or structures based on the insights gained from the Analyze stage. However, depending on the specific project requirements and goals, there may be opportunities to perform additional data manipulation, feature engineering, or model development tasks. Adaptation of the PACE framework ensures flexibility in addressing the unique needs of different projects and allows for a tailored approach to achieving project objectives.

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), it would be prudent to investigate the quality and completeness of the dataset. Specifically, I would recommend:

1. Conducting data validation checks to ensure the integrity of the data.

2. Identifying and handling missing values appropriately.

3. Checking for data outliers or anomalies that may skew the analysis results.

4. Verifying the consistency and accuracy of data across different columns and records.

- What data initially presents as containing anomalies?

1. Fare_amount: Any negative fare amounts or unusually high values may indicate data entry errors or outliers.

2. Trip_distance: Similarly, negative distances or extremely long trip distances could be indicative of anomalies.

3. Tip_amount: Anomalies may be present in the tip amount column, such as negative tips or disproportionately large tip values compared to the fare amount.

- What additional types of data could strengthen this dataset?

To enhance the dataset, consider incorporating weather, traffic, demographic, and event data.