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~~

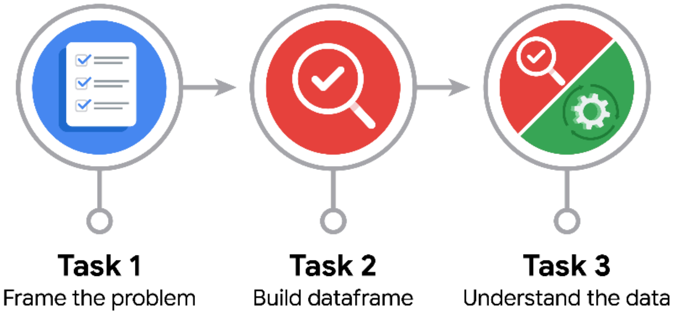
Relevant Interview Questions

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?

Familiarize with the Dataset:

Understand the structure, types of variables, and overall context of the dataset.
Review the metadata and documentation provided with the dataset.
Define Objectives:

Clarify the goals of the analysis and what insights are needed.
Identify key questions to be answered through the analysis.
Identify Key Variables:

Determine which variables are most relevant to the analysis objectives.
Focus on variables that have a significant impact on the outcomes of interest.

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

NYC TLC Dataset Documentation: Provides detailed descriptions of each variable. Explains data collection methods and any preprocessing steps. Data Science Codebooks: Pandas Documentation: For data manipulation and analysis. NumPy Documentation: For numerical operations. Matplotlib/Seaborn Documentation: For data visualization. Scikit-Learn Documentation: For machine learning model building and evaluation.

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

Literature Review: Research previous studies or analyses performed on similar datasets. Understand common methodologies and best practices. Data Exploration: Perform initial data exploration to get a sense of the data distribution and potential issues. Identify missing values, outliers, and data inconsistencies. Tool Setup: Ensure that the necessary tools and libraries are installed and configured. Set up a conducive environment for data analysis, such as Jupyter Notebooks.

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 initial exploration, evaluate if the available variables cover the necessary aspects to achieve the analysis goals.

Consider if additional data might be needed to fill any gaps.

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

Summary Statistics:

Python code:

```
summary_stats = df.describe()

print(summary_stats)
```

This provides count, mean, standard deviation, min, 25th percentile, median, 75th percentile, and max for numerical variables.

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

Assessing Min and Max Range:

Python code:

```
min_max = df.agg(['min', 'max'])

print(min_max)
```

Evaluating Averages and Describing Interval Data:

Unusual Averages:

Compare means with medians to identify skewness.

Analyze if any averages are outliers or seem inconsistent with the data context.

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?

Data Quality:

Investigate missing values, duplicate entries, and outliers.

Ensure data consistency and correctness.

Key Variables:

Focus on variables with high relevance to the analysis objectives, such as trip distance and fare amount.

- What data initially presents as containing anomalies?

Outliers:

Unusually high or low fare amounts or trip distances.

Inconsistent or incorrect passenger counts.

Missing Values:

Variables with a high percentage of missing data that might impact the analysis.

- What additional types of data could strengthen this dataset?

Weather Data:

Weather conditions can influence taxi demand and trip characteristics.

Event Data:

Information on city events or holidays that might affect taxi usage.

Demographic Data:

Demographic information about the areas served can provide additional insights.