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~~

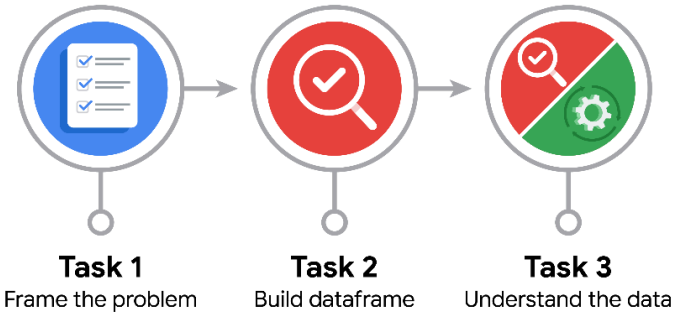
Relevant Interview Questions

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 project proposal, team emails, and dataset structure to identify key goals, stakeholders, and relevant columns. Plan how to approach data inspection and organization using Python tools and templates.

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

Python reference guides for pandas and NumPy, Jupyter notebook templates, and cheat sheets for data cleaning, summary statistics, and dataframe operations.

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

Inspect a sample of the dataset to detect formatting issues, sketch a plan for combining or modifying variables, and identify which columns to focus on during analysis.

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 dataset contains relevant columns for claims and opinions. Initial inspection and descriptive statistics will clarify which variables are meaningful for analysis.

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

Use df.describe() for numeric summaries, df.info() to check column types and non-null counts, and value_counts() for categorical distributions. Evaluate min and max values to detect outliers or unexpected ranges.

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

Identify columns where means or medians are extreme compared to expected ranges. Interval data (numeric, continuous) should be assessed for outliers, missing values, and consistency.

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 columns with missing or inconsistent values and identify which variables may need cleaning or transformation. Highlight columns suitable for creating new meaningful features.

- What data initially presents as containing anomalies?

  Columns with null values, unexpected zeros, negative values, or unusual formatting. Text columns may contain empty strings, symbols, or inconsistent entries.

- What additional types of data could strengthen this dataset?

  Include engagement metrics (likes, shares, comments), timestamps, content category labels, or source reliability indicators to improve context and predictive modeling potential.