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

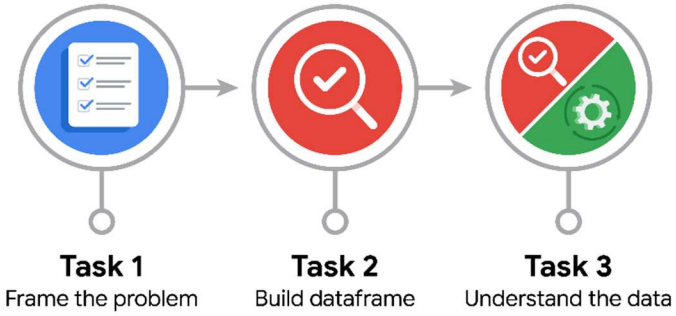
Relevant Interview Questions

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?

Complete an initial investigation of the data by reviewing the data dictionary. This provides a high-level overview of the types of variables included in the dataset and their specific types. Then, conduct a more in-depth investigation of the dataset and generate descriptive statistics.

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

The methods and techniques I used in previous workbooks for this course will be directly applicable to this project.

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

Thoroughly review the data dictionary and dataset. Outline a project plan, and before conducting any EDA, gather supporting resources such as previous workbooks or code snippet cheat sheets.

PACE: Analyze Stage

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

The available data is sufficient to begin preliminary analysis and engineer potentially useful variables based on domain knowledge.

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

The describe() function is generally sufficient for determining the minimum and maximum values of the dataset's variables.

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

The averages of some variables may be inflated by extreme outliers. Further investigation is needed before drawing conclusions.

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?

  Given the apparently higher engagement with videos by banned users, I would like to obtain more granular data for further analysis.

- What data initially presents as containing anomalies?

  Although some variables have missing values, the quantity is minimal. Several variables exhibit high right skewness and contain significant outliers.

- What additional types of data could strengthen this dataset?

  Additional data pertaining to banned authors is required, including, where possible, timestamps of posting and video duration.