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

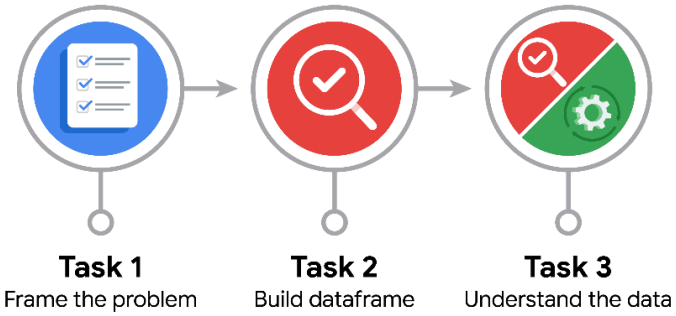
Relevant Interview Questions

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?

We'll start by looking at where the data is stored and what format it is stored in. Given that information, we'll then import the data into a Pandas DataFrame using a DF constructor. Then we can use the .head() and .info() methods to gather essential insights on the data's shape, types, names, and formats. Reviewing (or creating) a Data Dictionary will clarify what each column of data represents.

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

Python for Data Analysis, 3rd Ed. Google DA Course Notebooks (Course 2, Module 4)

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

Prepare notes from learning about NumPy and Pandas, bring up summary pages with links to relevant documentation, set up notepage for insights gathered while coding.

PACE: Analyze Stage

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

For the goals outlined above, yes.

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

The first and easiest step is to use the .describe() function to get a table of summary statistics for each numeric column in the dataset. Then we can take a look at non-numeric columns by grouping and aggregating their information to get different views of the data to provide insight on their correlations.

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

The video_id column has strange statistical values since it is not quantitative data but a unique identifier. Video duration has a fairly typical mean and interval. The remaining numeric columns contain data with a higher standard deviation than mean. This implies that data in those columns has a large amount of variance, and many datapoints with relatively low numbers clustered together compared to datapoints above the average with very high values spread out more.

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?

A larger dataset is always better for statistical analysis. See if the 298 rows with missing data can be filled in. Notify of the correlation between banned users and higher views/engagement.

- What data initially presents as containing anomalies?

The video_id column should be left out of statistical analysis given it is not quantitative data.

- What additional types of data could strengthen this dataset?

Engagement ratios such as likes or comments per view could prove a useful metric.