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

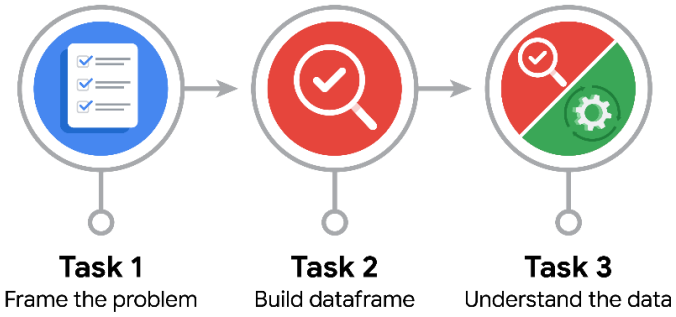
Relevant Interview Questions

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?

By exploring the dataset and inspecting summary information and descriptive statistics.

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

The Python notebook that has been formatted for me by Orion. In other words, the Python lab activity Notebook.

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

Python or Python Packages documentation, such as pandas Documentation or NumPy Documentation.

PACE: Analyze Stage

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

No as there is 298 null values in 7 Variables, but if these can be accounted for then yes this will be sufficient to achieve the goal.

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

I would use the describe() method, this method will produce a dataframe with the count, mean, standard deviation, min, median and max statistics for all the variables in the dataframe.

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

The summary statistics reveal that many of the variables, especially the engagement metrics like view, like and comment count are highly right-skewed with large maximum values compared to their means and quartiles, indicating the presence of outliers and possibly questionable values such as zeros, that may represent missing or suppressed data.

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?

Before starting exploratory data analysis, I recommend investigating missing values, extreme outliers in engagement metrics and the impact of the author ban status.

- What data initially presents as containing anomalies?

The engagement metrics, such as video view, like, share and comment count, initially present as containing anomalies due to their extremely high maximum values compared to the mean and median, suggesting the presence of outliers.

- What additional types of data could strengthen this dataset?

Additional data that could strengthen this dataset are: video category or topic, follower count of the author and the geographic location.