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

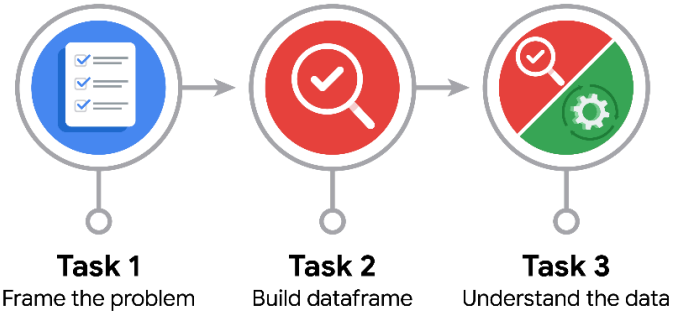
Relevant Interview Questions

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?

As stated in the notebook, I would first review the data dictionary (if available) to understand the meaning of each column. Then, I would explore the dataset using `data.head()`, `data.info()`, and `data.describe()` to get an initial overview of the data types, null values, and basic statistics. I would also consider researching common TikTok video metrics and trends to provide context.

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

○ Pandas documentation for data loading, inspection, and summary statistics. ○ Previous notebooks or code examples that demonstrate data exploration and summary statistics. ○ Any relevant tutorials or documentation on handling missing data and exploring categorical variables.

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

○ Research TikTok's content policies and community guidelines to understand potential reasons for author bans or flagged content. ○ Consider the types of claims or opinions likely to be shared on TikTok and how they might impact engagement metrics. ○ Develop initial hypotheses about relationships between variables (e.g., banned authors and claim status).

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 available information is sufficient for preliminary analysis and to identify key trends. However, further investigation into the context of the videos and potential biases in the data might be necessary for a more comprehensive analysis.

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

○ Use `data.describe()` to generate summary statistics, including count, mean, standard deviation, minimum, and maximum values. ○ Examine the minimum and maximum values of each numerical column to identify potential outliers or anomalies. ○ Use `data.info()` to see the data types, and null value counts.

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

○ The wide ranges and high standard deviations in the engagement metrics (views, likes, shares, comments) suggest the presence of outliers and variability in video popularity. ○ The interval data, such as video duration, can show the range of video lengths that are present on the platform.

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 the reasons for missing values in the `claim_status` and engagement metric columns.
  - Explore the content of videos associated with banned authors to understand the nature of their violations.
  - Further analyze the "super-engaged" videos (high views, likes, shares) to identify common characteristics.

- What data initially presents as containing anomalies?

  - The significant differences in engagement metrics between banned and active authors.
  - The wide ranges and high standard deviations in the numerical columns.
  - The high engagement rates of claim videos compared to opinion videos.

- What additional types of data could strengthen this dataset?

  - Video content metadata (e.g., tags, categories, audio features).
  - User demographic information.
  - Time-series data on video performance.
  - Information on the type of violations that banned authors committed.