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

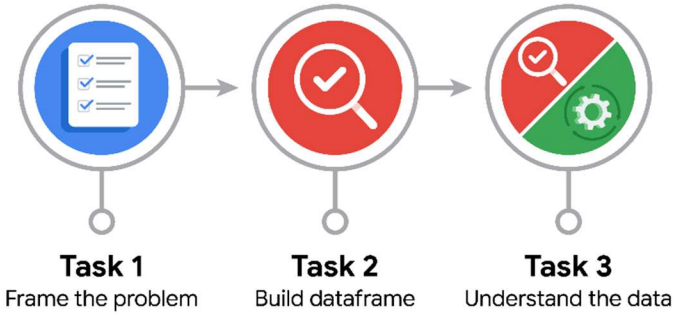
Relevant Interview Questions

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?

Start by noting down the type of information I am looking to extract from this data. (Clearly outline the goal of my work)

Then I would read the data as a panda dataframe such that I can visualize what type of information is being provided to me.

Then I would look into segregating the information which I believe is of interest to me and our goal and start looking at the distribution of the data and find important statistics about it.

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

I believe the most important notebook is the “pandas dataframe” notebook as this one introduces all the important methods and attributes I can use from the pandas module to understand and analyze data.

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

Writing pseudo-code is always helpful as this activity helps outline a road map structure and modularize my code.

PACE: Analyze Stage

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

I believe I have all the necessary tools and skills to achieve the goals outlined in this task

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

First I would look at the .info() method to look at the type of data I am provided and from that I can see what type of report/summary I can build. After that I would focus on the segments of the data I believe are of interest to our goal and start looking at some statistics to describe the data.

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

I notice that averages for the videos categorized as claims had significantly higher averages in views, likes, shares and comments than opinion videos. Another interesting point to outline is that not all the videos in the database provided to me were categorized into “Claims” nor “Opinion”, there were 298 videos that are yet to be categorized. Another feature important to outline is that we have about the same number of claim videos as opinion videos which will not introduce a sample bias. Another important point I noticed in this initial investigation is that most of the distributions of views, comments and shares are symmetric because the mean and the median were very similar.

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?

We would need to investigate the reason why most claim video authors are banned from the platform. Even after normalizing the number of banned authors by the sample size there is still a significantly big gap in the number of ban authors for claim videos than opinion videos.

It is clear that claim videos are more successful than opinion videos solely based on our metric for engagement, however, we see much controversy arise in this subject because of the high number of bans. It could be a tendency that authors for these types of videos are stepping over boundaries for fame.

- What data initially presents as containing anomalies?

Another pressing issue I saw in the data is that “claim video” authors who have been banned from the platform have the highest median share counts. Which means this type of content is very viral and could explain why so many people are doing it.

- What additional types of data could strengthen this dataset?

If there was a genre for classifying videos, or if we could see the main demographic of people that watch a specific author could significantly help us understand the bigger picture.