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

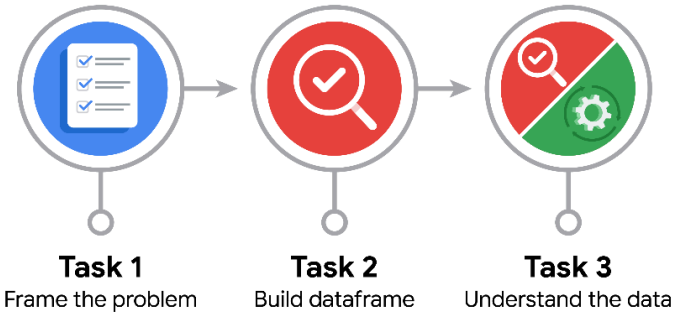
Relevant Interview Questions

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?

Get to know that product and sort of information that will most likely be present.

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

Codebooks describing numpy, pandas and jupyter notebook.

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

Understand the problem and plan a solution before jumping into 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 think the information is mostly sufficient given the transcript and type of claim given. It can be made easier by giving information of the claimant which can inform us of their habits and indicate if they are abusing the system.

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

Use pandas for dataframe and use info and describe to access all necessary information

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

The means are much higher than the median across the board meaning that super popular videos are skewing the results indicating a long right tailed distribution. Most interval data is right skewed with exception of video length.

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?

Missing entries must be dealt with and data needs to be normalised to be more useful due to very large difference in interval data between median and max. Columns such as claim info and author ban status can be encoded.

- What data initially presents as containing anomalies?

Missing values in data and heavily skewed data suggests some anomalies in most columns.

- What additional types of data could strengthen this dataset?

Claimant info, account history, captions, time stamps and video watch rate.