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

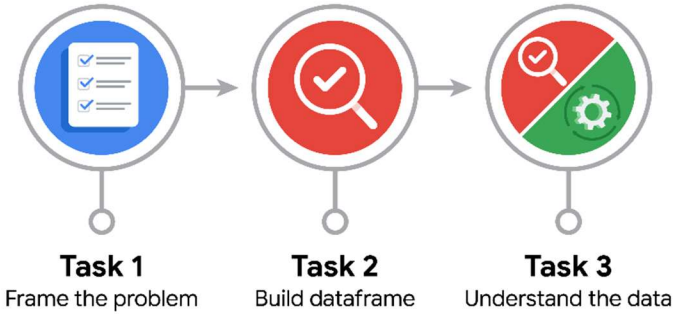
Relevant Interview Questions

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?

Utilize a framework to outline the necessary steps for the analysis. The PACE framework provides a solid foundation for structuring the entire analysis from beginning to end. Once EDA begins, analyzing the data at a high level through descriptive statistics and summarization methods is a good starting point to understand the data's characteristics and guide further actions.

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

The information in the project codebook, as well as previous codebooks completed throughout the previous weeks, will be invaluable for completing the necessary tasks. Additionally, project codebooks from the Automatidata analysis will also be helpful.

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

Before starting to code, it's useful to have any relevant prior codebooks on hand that contain functions, formulas, or scripts that can be utilized. Additionally, online resources where you can quickly search for and find solutions to encountered problems can be very helpful, as it's almost guaranteed that someone else has faced a similar challenge.

PACE: Analyze Stage

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

For this stage of the analysis, the data is sufficient. However, after completing an initial analysis, it became clear that additional data is needed to gain a better overall understanding of the population being analyzed.

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

By using the .describe() function, you can obtain the dataframe's statistics, including the minimum, maximum, mean, and standard deviation.

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

The average driving statistics originally collected and engineered, such as 'km_per_driving_day' and 'drives_per_driving_day', are significantly higher than what would be expected for the average driver. This could suggest that the data is not representative of the general population or is biased towards a particular group of drivers, such as long-haul truckers or long-distance commuters.

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?

Based on the statistics generated from existing and engineered variables, additional data is needed, ideally data that is more representative of the general population. Additionally, understanding how the data was collected would be beneficial.

- What data initially presents as containing anomalies?

The "label" column appears to have missing values. However, after further investigation, these missing values seem to have little or no impact and appear to be random. The driving statistics for churned and retained users are also very high, especially for churned users, which might indicate that the data is not representative of the general population.

- What additional types of data could strengthen this dataset?

First and foremost, it's essential to obtain data that is random and representative of the population. Information on the days of the week and times of day the app was used, as well as how the app was used (general use, long-distance commute businesses, long-haul trucking businesses, etc.), would be helpful.