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

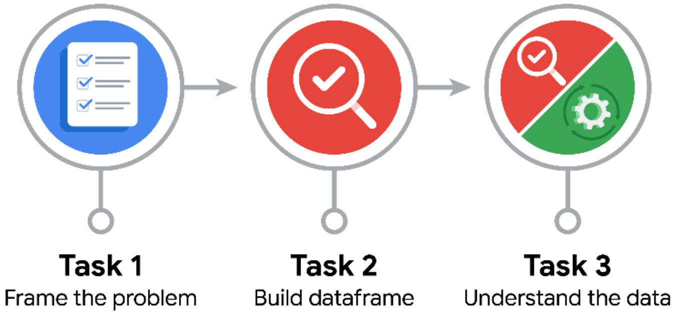
Relevant Interview Questions

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 data and preparing it for EDA.

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

The ones from the course videos.

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

Watching videos on the programming language and following along with the video.

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 much more data exploration and EDA is needed.

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

Using the .describe() method.

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

Yes. Several averages appear unusually high due to extreme outliers in driving behavior. Variables such as kilometers driven and driving duration show strong right skew, with maximum values far exceeding typical usage. The interval data (quartiles and medians) indicate that most users have much lower activity levels, meaning the mean is not representative of a typical user. As a result, the median provides a more reliable summary of central tendency for this dataset.

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?

I would recommend first validating the presence and impact of extreme outliers in driving-related variables (e.g., kilometers driven, driving days, and drives per day) and confirming whether these represent genuine high-intensity users or data quality issues. Additionally, the missing values in the churn label should be reviewed to confirm they are missing at random and do not bias analysis.

- What data initially presents as containing anomalies?

Driving-intensity variables such as total kilometers driven, duration of drives, kilometers per driving day, and drives per driving day appear anomalous due to extremely high values. These suggest a small subset of users with unusually intense driving behavior that may distort averages and require special handling during analysis.

- What additional types of data could strengthen this dataset?

The dataset would benefit from user segmentation data (e.g., commuter vs. professional driver), geographic context, trip purpose, time-of-day usage, and engagement features such as navigation errors or reroutes. Feedback or satisfaction metrics could also help explain churn behaviour beyond raw usage intensity.