Course One

Foundations of Data Science

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 1 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

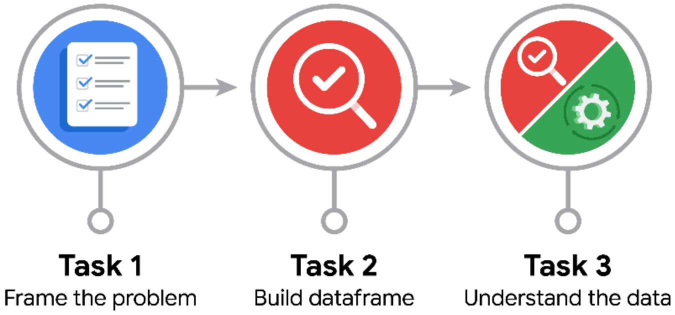
Relevant Interview Questions

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?

The best preparation is to understand how the Waze application works and how users interact with it. Then, review the dataset structure, variables, and definitions to ensure a full understanding of the data and identify any misleading or incomplete information that could affect churn analysis.

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

I would refer to the data dictionary, column descriptions, and any project documentation. I would also review previous examples of churn analysis and machine learning workflows to guide the coding and ensure alignment with stakeholder goals.

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

A resourceful learner would explore how the Waze app functions from a user perspective, review company goals related to retention, and research common reasons for user churn in similar apps. This background knowledge helps connect the data to real user behavior.

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 data looks enough to start building a churn prediction model because it shows clear user behavior like sessions, drives, and activity days, which help understand how users interact with the app.

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

i would use functions like `describe()` and `agg()` to see the main statistics, then check the minimum and maximum values to make sure there are no unrealistic numbers that could affect the analysis.

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

yes, some values like driven kilometers and drive duration look very high compared to normal users, which means there may be extreme users in the data. most of the variables are numerical and continuous, showing different levels of user activity over time.

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 looking deeper into the extreme users who drive very long distances and understanding if they represent a special group of drivers, because they might affect the model results.

- What data initially presents as containing anomalies?

the variables related to driven distance and driving duration show unusually large values, which might be outliers or special cases.

- What additional types of data could strengthen this dataset?

it would help to have more information about user behavior, like trip purpose, time of day usage, user location patterns, or feedback from users about why they stopped using the app.