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

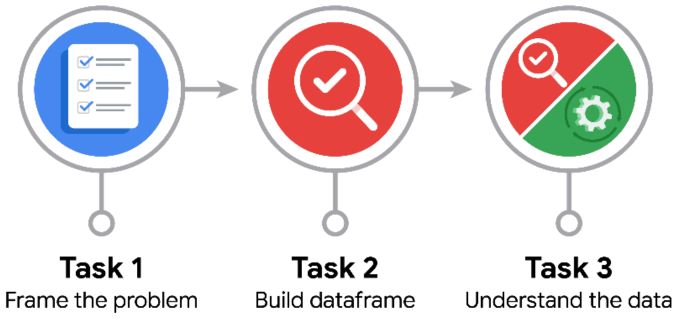
Relevant Interview Questions

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?
  - Begin by thoroughly reviewing the project's objectives and the provided data dictionary.
  - Create a mental or written outline of the steps needed to inspect and analyze the data.
  - Load the data into a Pandas DataFrame and use functions like `df.head()`, `df.info()`, and `df.describe()` to get an initial understanding of the data's structure and content.
  - Identify the data types of each column and check for missing values.

- What follow-along and self-review codebooks will help you perform this work?
  - Reviewing Pandas documentation for data loading, inspection, and summary statistics will be crucial.
  - Refer to any previously completed notebooks that demonstrate similar data inspection and cleaning techniques.
  - Utilize online resources and tutorials for specific Pandas functions as needed.

- What are some additional activities a resourceful learner would perform before starting to code?
  - Research the context of the data (e.g., Waze user data) to understand potential variables and their significance.
  - Consider potential biases or limitations in the data collection process.
  - Brainstorm potential questions or hypotheses that could be explored during the analysis.
  - Plan how the executive summary will be structured, and what information will be the most useful to the stakeholders.

PACE: Analyze Stage

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

  - Initially, the data appears sufficient for preliminary analysis. However, further investigation into the meaning of certain variables and potential data limitations may be necessary.
  - The presence of missing values in the 'label' column requires careful consideration and may impact the analysis.
  - The high amount of kilometers driven by users, could be an indicator of a non typical user base.

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

  - Use `df.describe()` to generate summary statistics, including count, mean, standard deviation, minimum, and maximum values.
  - Examine the minimum and maximum values of each numerical column to identify potential outliers or anomalies.
  - Use `df.median()` to find the median of the data, to compare to the mean, and determine if outliers are effecting the data.

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

  - Yes, the averages of `driven_km_drives` and `duration_minutes_drives` seem high, suggesting a population of users that drive very frequently, and for long distances.
  - The interval data, such as the amount of days between when a user signed up, and the data was collected, can show the general age of the user base.

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?
    - Investigate the reasons for the missing values in the 'label' column and determine if they can be imputed or if those rows should be removed.
    - Gather more information about the data collection process and the definition of key variables (e.g., sessions, favored navigation locations).
    - Clarify the user base, and if it is a random sampling of the general public.

- What data initially presents as containing anomalies?
    - The high maximum values in `driven_km_drives` and `duration_minutes_drives` suggest potential outliers.
    - The significant differences in driving behavior between churned and retained users may also indicate anomalies or specific user segments.

- What additional types of data could strengthen this dataset?
    - Demographic data (e.g., age, location, occupation) could provide valuable insights into user behavior.
    - Data on user feedback or app usage patterns (e.g., feature usage, error logs) could help identify potential pain points.
    - Information on the type of driving the user is doing, such as commercial, or personal.