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:

- [x] ~~Complete the questions in the Course 2 PACE strategy document~~
- [x] ~~Answer the questions in the Jupyter notebook project file~~
- [x] ~~Complete coding prep work on project's Jupyter notebook~~
- [x] ~~Summarize the column Dtypes~~
- [x] ~~Communicate important findings in the form of an executive summary~~

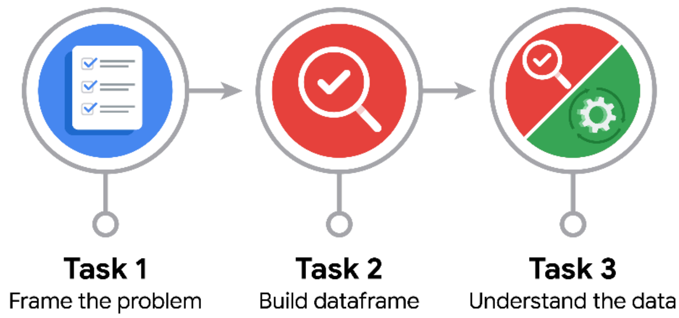
Relevant Interview Questions

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?

● Review the project description and dataset details. ● Examine the data dictionary to understand variable meanings and data types. ● Identify key objectives: user churn analysis and predictive modeling.

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

● Python documentation for `pandas`, `numpy`, and `matplotlib`. ● Previous course materials and sample Jupyter notebooks. ● Example datasets with similar churn analysis case studies.

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

● Research common patterns in churn analysis. ● Explore industry best practices in data cleaning and preprocessing. ● Identify potential data transformations needed for feature engineering.

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 dataset includes relevant attributes such as driving patterns, device type, and engagement metrics.
- Further data on user interactions (e.g., app feedback, session duration) could improve predictive modeling.

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

- Use `df.describe()` to obtain summary statistics.
- Use `df.info()` to check for missing values and data types.
- Create visualizations (histograms, box plots) to examine distribution and outliers.

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

- Outliers in `driven_km_drives` and `drives_per_driving_day` could indicate extreme usage patterns.
- Checking for skewness in distributions using `df.skew()`.
- Interval data (e.g., daily drive count) should be analyzed for patterns 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?

- Investigate high-usage drivers and their impact on churn.
- Examine possible correlation between churn and infrequent usage.
- Look for device-specific engagement differences.

- What data initially presents as containing anomalies?

- High `km_per_drive` values exceeding normal travel distances.
- Missing labels in the churn column.
- Extreme values in `driven_km_drives`, possibly representing long-haul drivers.

- What additional types of data could strengthen this dataset?

- User feedback or satisfaction scores.
- App usage frequency beyond driving behavior.
- External factors such as traffic conditions or fuel prices affecting usage.