# 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

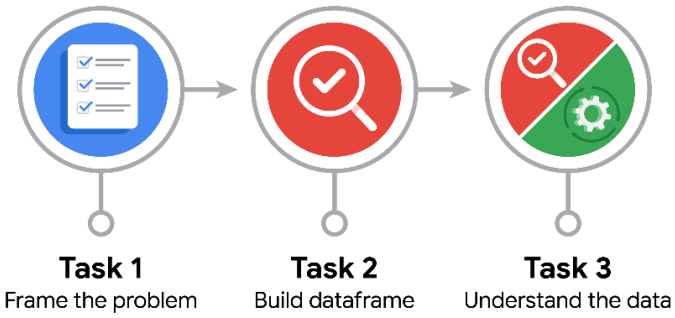
### Relevant Interview Questions

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 reading the project instructions and understanding the dataset structure. |
| • Identify key variables, their data types, and how they relate to the analysis goals. |
| • Create a data dictionary to document column names, data types, and descriptions. |
| • Review similar datasets or projects to gain insights into potential challenges. |
| • Outline a step-by-step plan for cleaning, analyzing, and interpreting the data. |
| • Review emails from May Santner and Chidi Ga to align with project expectations. |

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

| |
|---|
| • Python documentation for Pandas, NumPy, and Matplotlib for data manipulation and visualization. |
| • Google Data Analytics course materials and example Jupyter notebooks. |
| • Previous datasets from Kaggle or similar platforms to compare best practices. |
| • Online coding tutorials or Stack Overflow for troubleshooting. |
| • A checklist for common data cleaning and transformation steps. |
| • Review the provided Waze dataset documentation for key insights. |

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

| |
|---|
| • Python documentation for Pandas, NumPy, and Matplotlib for data manipulation and visualization. |
| • Google Data Analytics course materials and example Jupyter notebooks. |
| • Previous datasets from Kaggle or similar platforms to compare best practices. |
| • Online coding tutorials or Stack Overflow for troubleshooting. |
| • A checklist for common data cleaning and transformation steps. |
| • Review the provided Waze dataset documentation for key insights. |

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

| |
|---|
| • Conduct exploratory data analysis (EDA) to understand trends and potential issues. |
| • Perform a preliminary check for missing values and inconsistencies. |
| • Research domain-specific knowledge relevant to the dataset. |
| • Sketch out a rough workflow for cleaning and analysis. |
| • Review similar case studies to understand expected outcomes. |
| • Familiarize yourself with Waze's business model and how user churn impacts it. |

## PACE: Analyze Stage

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

| |
|---|
| • A preliminary review of the dataset should determine if it contains the necessary variables. |
| • Checking for completeness (missing values, duplicates) ensures reliability. |
| • If critical information is missing, consider external data sources or adjusting the scope of the analysis. |
| • Review the dataset for inconsistencies between expected values and the provided data. |
| |

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

| |
|---|
| • Use Pandas functions such as `.describe()` to get summary statistics. |
| • Apply `.min()` and `.max()` to find the range of numerical columns. |
| • Check for inconsistencies such as negative values where they shouldn't exist. |
| • Create histograms or boxplots to visualize data distribution. |
| • Summarize the column data types to verify correct formatting. |

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

| |
|---|
| • Calculate mean, median, and mode to detect skewness. |
| • Compare calculated statistics to expected industry or domain benchmarks. |
| • Investigate extreme values or outliers using standard deviation and interquartile range (IQR). |
| • If the dataset includes time-series data, check for seasonality or trends. |
| • Evaluate if user activity metrics (sessions, drives) correlate logically. |

## 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?

| |
|---|
| • Identify any discrepancies between expected and actual data distributions. |
| • Look into missing values and determine if they need imputation or removal. |
| • Verify consistency in categorical data entries (e.g., uniform spelling and formatting). |
| • Ensure that numerical values align with business logic (e.g., no negative revenue figures). |
| • Recommend focusing on key predictive features like `sessions`, `drives`, and `n_days_after_onboarding`. |

- What data initially presents as containing anomalies?

| |
|---|
| • Outliers in numerical columns (extreme high or low values). |
| • Categorical values that do not fit expected categories. |
| • Unexpected null or blank entries in key columns. |
| • Duplicate records that may affect analysis outcomes. |
| • Anomalous values in `driven_km_drives` or `duration_minutes_drives` that don't match session counts. |

- What additional types of data could strengthen this dataset?

| |
|---|
| • External demographic or economic data to provide context. |
| • Additional timestamps or location data for deeper trend analysis. |
| • Data from similar datasets for cross-validation. |
| • Industry benchmarks for comparison to identify anomalies more effectively. |
| • Behavioral insights on user engagement with app features. |
| • Traffic condition data to analyze its effect on driving habits. |