**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?

Reviewing the dataset structure and documentation to understand the variables and their relationships. Identifying key stakeholders and understanding their expectations and requirements.

Setting clear goals and milestones for data preparation and analysis tasks.

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

Python documentation and tutorials for data manipulation and analysis (e.g., Pandas, NumPy).

Previous course materials on data cleaning and EDA.

Jupyter notebooks with examples of similar projects.

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

Conducting a preliminary analysis of the dataset to identify potential issues.

Creating a data dictionary to document variable definitions and data types.

Setting up the development environment with necessary libraries and tools.

**PACE: Analyze Stage**

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

Initial inspection suggests that the dataset contains relevant variables to analyze user behavior and churn patterns. However, the data seems biased in users that drive more than a typical driver. Further analysis is needed to confirm its sufficiency and/or more data of different datasets needs to be compared to speculate possible bias.

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

Using Python libraries like pandas to calculate summary statistics (mean, median, mode, min, max).

Visualizing data distributions to identify outliers and anomalies.

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

Both driven\_km\_drives and duration\_minutes\_drives.

Interval data such as sessions and drives can provide insights into user activity levels.

**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 any missing values or anomalies in the data.

Identify the distribution of key variables and their correlation with the target variable (churn).

Further analyze the user profiles of those with extreme values in driven kilometers and session durations.

* What data initially presents as containing anomalies?

The label column has missing values.

Unusually high or low values in sessions, drives, driven\_km\_drives, and duration\_minutes\_drives.

* What additional types of data could strengthen this dataset?

User demographics (age, gender, location).

External factors such as weather conditions or traffic patterns.

User feedback and satisfaction ratings.