**Course Three**

# Go Beyond the Numbers: Translate Data into Insights



# 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 3 PACE strategy document
* Answer the questions in the Jupyter notebook project file
* Clean your data, perform exploratory data analysis (EDA)
* Create data visualizations
* Create an executive summary to share your results

# 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:

* How would you explain the difference between qualitative and quantitative data sources?
* Describe the difference between structured and unstructured data.
* Why is it important to do exploratory data analysis?
* How would you perform EDA on a given dataset?
* How do you create or alter a visualization based on different audiences?
* How do you avoid bias and ensure accessibility in a data visualization?
* How does data visualization inform your EDA?

**Reference Guide**

This project has six tasks; the visual below identifies how the stages of PACE are incorporated across those tasks.



**Data Project Questions & Considerations**

**PACE: Plan Stage**

* What are the data columns and variables and which ones are most relevant to your deliverable?

The dataset contains 13 columns including 'ID', 'label', 'sessions', 'drives', 'device', 'total\_sessions', 'n\_days\_after\_onboarding', 'total\_navigations\_fav1', 'total\_navigations\_fav2', 'driven\_km\_drives', 'duration\_minutes\_drives', 'activity\_days', and 'driving\_days'. The most relevant columns for this analysis are 'sessions', 'drives', 'total\_sessions', and 'label'.

* What units are your variables in?

'sessions' and 'total\_sessions' are counts of app openings, 'drives' and 'driving\_days' are counts of driving occurrences, 'driven\_km\_drives' is in kilometers, 'duration\_minutes\_drives' is in minutes, and 'n\_days\_after\_onboarding' is in days.

* What are your initial presumptions about the data that can inform your EDA, knowing you will need to confirm or deny with your future findings?

Initial presumptions include that higher engagement with the app (more sessions and drives) is correlated with lower churn rates. Users with fewer sessions and drives are more likely to churn. Additionally, it is expected that the distributions of these engagement metrics will be right-skewed.

* Is there any missing or incomplete data?

Yes, the 'label' column has 700 missing values, which should be addressed during EDA.

* Are all pieces of this dataset in the same format?

Yes, the dataset appears to be consistently formatted, with appropriate data types for each column.

* Which EDA practices will be required to begin this project?

Initial data inspection, handling missing values (such as the missing 'label' values), descriptive statistics, and visualizations such as box plots and histograms to understand data distribution. Additionally, creating new columns for deeper insights (like km\_per\_driving\_day), handling outliers, and converting infinite values to zero will also be necessary.

**PACE: Analyze Stage**

* What steps need to be taken to perform EDA in the most effective way to achieve the project goal?

Steps include loading the dataset, checking for missing values, performing descriptive statistics, and creating visualizations to understand the data distribution. Specifically, exploring right-skewed distributions and handling outliers.

* Do you need to add more data using the EDA practice of joining? What type of structuring needs to be done to this dataset, such as filtering, sorting, etc.?

No additional data joining is necessary. The dataset should be filtered to remove any potential outliers and sorted for specific analysis if needed.

* What initial assumptions do you have about the types of visualizations that might best be suited for the intended audience?

Initial assumptions include that box plots and histograms will help visualize the distribution of 'sessions' and 'drives'. A scatter plot may be useful for identifying relationships between variables.

**PACE: Construct Stage**

* What data visualizations, machine learning algorithms, or other data outputs will need to be built in order to complete the project goals?

Data visualizations needed include box plots for 'sessions' and 'drives', histograms for the same variables, and scatter plots to explore relationships between variables.

* What processes need to be performed in order to build the necessary data visualizations?

Processes include data cleaning, handling missing values, calculating descriptive statistics, and using matplotlib and seaborn to create the visualizations.

* Which variables are most applicable for the visualizations in this data project?

'sessions', 'drives', 'total\_sessions', and 'label' are the most applicable variables for visualizations.

* Going back to the Plan stage, how do you plan to deal with the missing data (if any)?

If missing data is identified, it will be handled by either filling in with appropriate values (e.g., mean or median) or removing rows with missing values.

******PACE: Execute Stage**

* What key insights emerged from your EDA and visualizations(s)?

Key insights include the distribution and central tendency of user sessions and drives, potential outliers, and relationships between engagement metrics and churn rates. The sessions and drives variables are right-skewed, with a significant number of high-engagement users.

* What business and/or organizational recommendations do you propose based on the visualization(s) built?

Recommendations include targeted interventions for users with low engagement to prevent churn, such as personalized offers or notifications. Additionally, monitoring key metrics like sessions and drives can help identify at-risk users. Investigating why high engagement in the last month correlates with churn may provide further insights.

* Given what you know about the data and the visualizations you were using, what other questions could you research for the team?

Other questions to research include identifying specific features of the app that influence user engagement, seasonal trends in user behavior, and the impact of external factors like traffic conditions on app usage. Exploring the reasons behind the discrepancy in usage and driving behavior could be valuable.

* How might you share these visualizations with different audiences?

Visualizations can be shared through reports, presentations, or interactive dashboards tailored to different audiences, such as technical team members, business stakeholders, or executives.