**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 to 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 data columns include 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, with the most relevant variables to the deliverable being label, sessions, drives, total\_sessions, n\_days\_after\_onboarding, driven\_km\_drives, and duration\_minutes\_drives.

* What units are your variables in?

The variables are in units such as counts (e.g., sessions, drives), days (e.g., n\_days\_after\_onboarding, activity\_days, driving\_days), kilometers (e.g., driven\_km\_drives), and minutes (e.g., duration\_minutes\_drives).

* 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?

My initial presumptions are that higher engagement (more sessions and drives) and longer app usage (more days since onboarding) will correlate with lower churn rates, and that certain patterns in usage frequency and intensity will help identify users at risk of churning.

* Is there any missing or incomplete data?

There are missing values for the label data

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

Yes, all pieces of this dataset are in the same format, with columns containing data types of either integers, floats, or objects (strings).

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

Initial EDA practices will include summarizing statistics, visualizing distributions, checking for missing values, identifying and handling outliers, and examining correlations between variables to understand patterns and relationships in the data.

**PACE: Analyze Stage**

* What steps need to be taken to perform EDA in the most effective way to achieve the project goal?
* To perform EDA effectively, I need to clean the data by addressing missing values and outliers, visualize key variables and their relationships, perform statistical summaries, and analyze correlations to identify patterns that impact user churn.
* 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.?

Joining additional data may not be necessary if the current dataset is comprehensive, but structuring tasks such as filtering out irrelevant rows, sorting by key metrics, and aggregating data for meaningful insights will be essential for effective EDA.

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

Initial assumptions are that visualizations such as histograms, scatter plots, and heatmaps will be most effective for the intended audience to clearly illustrate distributions, relationships, and correlations in the data related to user engagement and churn.

**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?

To complete the project goals, data visualizations like histograms, scatter plots, and heatmaps will be needed, along with machine learning algorithms for classification and prediction, such as logistic regression or decision trees, to model user churn and identify key factors influencing it.

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

To build the necessary data visualizations, processes include data cleaning, selecting appropriate visual formats, generating plots with libraries like Matplotlib or Seaborn, and ensuring clarity through proper labeling, scaling, and color use.

* Which variables are most applicable for the visualizations in this data project?
* The most applicable variables for the visualizations are sessions, drives, total\_sessions, n\_days\_after\_onboarding, driven\_km\_drives, duration\_minutes\_drives, and percent\_sessions\_in\_last\_month, as they directly relate to user engagement and churn patterns.
* Going back to the Plan stage, how do you plan to deal with the missing data (if any)?

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

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

In the Plan stage, missing data will be addressed by imputing values where appropriate, removing rows with excessive missing values, and ensuring that the analysis accounts for any gaps without introducing bias.

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

Based on the visualizations, I recommend focusing on improving user engagement strategies for those showing high activity in the last month but at risk of churn, and enhancing features to retain users with low engagement and high churn risk.

* 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 could include identifying the specific factors driving high engagement in the last month, understanding seasonal or promotional effects on user activity, and exploring strategies to improve retention for users with low engagement and high churn risk.

* How might you share these visualizations with different audiences?

I would share these visualizations with different audiences by tailoring the presentation format: using detailed charts and graphs for technical teams, concise summaries and key insights for executives, and interactive dashboards for stakeholders needing ongoing access to data trends.