**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 includes 13 columns:

User behavior: sessions, drives, total\_sessions, driven\_km\_drives, duration\_minutes\_drives, activity\_days, driving\_days, total\_navigations\_fav1, total\_navigations\_fav2

User profile: device, n\_days\_after\_onboarding, label (target variable for churn)

User ID: ID (not used in analysis)

Relevant variables for churn analysis include:

sessions, drives, total\_sessions, driven\_km\_drives, duration\_minutes\_drives, activity\_days, driving\_days, n\_days\_after\_onboarding, label

* What units are your variables in?

sessions, drives, and total\_sessions: count

driven\_km\_drives: kilometers

duration\_minutes\_drives: minutes

activity\_days, driving\_days: count of days

n\_days\_after\_onboarding: number of days since onboarding

* 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?

Power users will drive high values in engagement metrics.

Users with recent and frequent activity are less likely to churn.

Long-tenured users might be more loyal, but this needs confirmation.

* Is there any missing or incomplete data?

Yes. The label column (churned/retained) has 700 missing values out of 14,999 rows. These will need to be dropped or imputed depending on modeling needs.

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

Yes. The dataset is clean with consistent data types (int64, float64, object). Basic formatting and data type conversions are not required.

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

Use of .describe() and .info() for summary statistics

Outlier detection via box plots

Distribution analysis via histograms

Correlation analysis

Pie chart for churn vs. retention

Feature engineering (e.g., percent\_sessions\_in\_last\_month, capping outliers)

**PACE: Analyze Stage**

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

Clean missing data in the label column

Visualize distributions to detect skewness

Identify and cap outliers (95th percentile)

Engineer new features for behavioral insights

Analyze relationships between activity metrics and 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.?

Not for this project. The dataset is self-contained.

Filtering out rows with missing label values

Creating capped versions of key metrics

Sorting and grouping by churn status for comparison

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

Pie charts for high-level churn overview

Box plots and histograms for behavior comparison

Bar charts for categorical features (like device type)

Line plots or scatterplots for time-based insights (e.g., n\_days\_after\_onboarding vs. activity)

**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?

Visuals: Box plots, histograms, pie charts

Engineered features: km\_per\_driving\_day, percent\_sessions\_in\_last\_month

Outlier-capped variables

Later: Random Forest and XGBoost models (in Course 6)

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

Use of matplotlib and seaborn for plotting

Data preprocessing to clean, transform, and engineer variables

Labeling charts for clarity and accessibility

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

sessions, drives, duration\_minutes\_drives, driven\_km\_drives, n\_days\_after\_onboarding, label, device

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

Drop rows where label is missing (700 rows), as the label is critical for classification

Confirm no other columns have NA values

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

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

~17.7% of users churned

High recent activity (≥ 40% sessions in last month) strongly predicts retention

Power users exist and were capped to reduce skew

Long-tenure doesn’t guarantee retention — recent activity matters more

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

Develop targeted campaigns for users with declining recent activity

Investigate what causes spikes in activity prior to churn

Incentivize continued usage after onboarding through nudges or rewards

Use activity metrics to create churn risk scores

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

What app features correlate with retention?

Are certain regions or demographics more likely to churn?

How do engagement patterns differ across devices?

Could in-app feedback predict churn before it happens?

* How might you share these visualizations with different audiences?

Technical team: Detailed plots with statistical annotations

Executives: High-level charts (pie, bar) with clear takeaways and visuals

Cross-functional teams: Visual dashboards or slide decks with business implications