**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?

**Data columns:** label, sessions, drives, device, total\_sessions, n\_days\_after\_onboarding, total\_navigations\_fav1, total\_navigations\_fav2, driven\_km\_drives, duration\_minutes\_drives, activity\_days, driving\_days.

**Relevant Variables:**  label, sessions, drives, device, total\_sessions, n\_days\_after\_onboarding, total\_navigations\_fav1, driven\_km\_drives, duration\_minutes\_drives, activity\_days, driving\_days.

* What units are your variables in?

driven\_km\_drives: km, duration\_minutes\_drives: minutes, and n\_days\_after\_onboarding, activity\_days and driving\_days are 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?

Most people that churn most likely have been dissatisfied with their user experience.

* Is there any missing or incomplete data?

There is almost no missing data except for the label variable: 700 missing rows.

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

Yes, they all are.

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

Discovery and presenting

**PACE: Analyze Stage**

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

Discovery, presenting, structuring, cleaning, validation, cleaning, validation, presenting.

* 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, I think this dataset will be largely sufficient. The type of structuring that needs to be done include: sorting, extracting, filtering and grouping.

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

Histogram, boxplot and scatter plots.

**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?

Histogram or density plots for numerical variables like essions, drives, total\_sessions, n\_days\_after\_onboarding, driven\_km\_drives, duration\_minutes\_drives, activity\_days, and driving\_days. Bar charts for categorical data like devices, scatter plots or line graphs to show correlation between two variables.

Classification algorithms would be appropriate to predict user churn: logistic regression, decision tree, gradient boosting machine, or neural networks.

The data output would be a predictive model, performance metrics, feature importance, visualizations, and insights and recommendations.

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

Preparation of data by cleaning the data - handling missing values, removing duplicates. Next, we conduct exploratory data analysis, and then choose the appropriate visualization type for each variable.

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

label, sessions, drives, device, total\_sessions, n\_days\_after\_onboarding, total\_navigations\_fav1, driven\_km\_drives, duration\_minutes\_drives, activity\_days, driving\_days.

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

Delete containing rows.

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

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

Frequent use of the Waze app was associated with lower churn rates, suggesting that satisfied users are less likely to stop using the app.

Long-distance drivers were more likely to churn from the app, possibly due to dissatisfaction with the service or lack of need for the app.

The number of drives made within a session was positively correlated with user retention, meaning users who make more drives within a session are less likely to leave the app.

Users who use the app less frequently were more likely to churn, indicating that regular use may contribute to user retention.

Professional drivers, who initiate more than 60 drives per month, churned at a much lower rate than average users, suggesting that these users are more committed to using the app.

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

Target marketing efforts towards users who already use the app regularly to encourage frequent use and lower churn rates.

Improve the service specifically for long-distance drivers to enhance navigation features, improve traffic updates, and offer discounts for long-distance travel.

Implement features that encourage users to make more drives within a session to reduce churn such as incentives for completing multiple trips.

Promote the benefits of regular use to encourage more frequent usage through reminder notifications, push notifications, and a referral program.

Offer special services or benefits for professional drivers who initiate more than 60 drives per month to increase their commitment to the app.

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

What factors contribute to a user's decision to churn after using the app for a significant amount of time?

Why do users who don't use the app frequently tend to churn more often compared to those who use the app regularly?

What is the correlation between the number of days a user uses the app in a month and the number of kilometers they drive?

* How might you share these visualizations with different audiences?

Use simple language and clear labels.

Choose appropriate visual types for the data.

Use a consistent and easy-to-understand color scheme.

Make element sizes proportional to the data they represent.

Consider making your visualizations interactive for deeper exploration.

Provide context for the data, including its significance and any limitations.

Share in multiple formats to cater to different preferences.