**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?

Label, sessions, drives, driven\_km\_drives, driving\_days, device and km\_per driving day are the most relevant data columns as far as the project’s deliverables are concerned.

* What units are your variables in?

Count variables(sessions, drives, total\_sessions),

Distance(driven\_km\_drives, km\_per\_driving\_day)

Duration(duratiom\_minutes\_drives)

Days(n\_days\_after\_onboarding, activity\_days, driving\_days)

Categorical(label, device)

* 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?

Higher Engagement lower churn rate

Longer a user uses the app, lower the churn rate

* Is there any missing or incomplete data?

Yes, there are 700 missing values in the label column

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

Yes

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

Discovering, Structuring and Cleaning

**PACE: Analyze Stage**

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

Ensure that the data is effectively cleaned and structured, checked for outliers and missing values so that the resulting dataset is in its optimal condition as far as the project’s requirements are concerned.

* 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.?

There is no immediate need for adding more data.

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

Histograms and Box Plots to show distributions of numerical variables and identify outliers.

Bar Charts and Pie Charts to compare user churn rates across various categories.

Scatter plots to identify relationships between variables.

Line charts to illustrate trends over time.

**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 Visualisations**

Histograms and Box Plots to show distributions of numerical variables and identify outliers.

Bar Charts and Pie Charts to compare user churn rates across various categories.

Scatter plots to identify relationships between variables.

Line charts to illustrate trends over time.

**Machine Learning Algorithms**

Logistic Regression

Clustering

**Other Data Outputs**

Descriptive Statistics

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

Data selection, data cleaning, data transformation, choosing the right visualization, matching the data type to the visualization.

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

Label, sessions, drives, driven\_km\_drives, driving\_days, device and km\_per driving day

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

Based on the reason, extent and nature of the missing data, we can consider:

Data Removal

Data Imputation

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

* What key insights emerged from your EDA and visualizations(s)?
* **Overall Churn Rate:** Approximately 17%.
* **Device Type:** Churn rate is consistent across iPhone and Android users.
* **High-Mileage Drivers:** Users who drive very long distances are more likely to churn.
* **Engaged Users:** Frequent app usage (sessions, drives) correlates with lower churn.
* **Data Anomalies:** Maximum driving days < 31; long-time users with sudden usage spikes.
* **Churn Correlations:** Higher mean daily driving distance and infrequent usage are linked to higher churn.
* **User Tenure:** No clear overrepresentation of newer users compared to long-time users.
* What business and/or organizational recommendations do you propose based on the visualization(s) built?

**Addressing Long-Distance Drivers:**

* Investigate pain points and unmet needs.
* Develop features for extended journeys (e.g., route optimization, real-time traffic).
* Personalize communication with relevant tips and information.

**Enhancing Engagement:**

* Incentivize frequent usage with rewards, challenges, or gamification.
* Provide personalized recommendations (routes, destinations).
* Use proactive notifications to prompt app usage.

**Data Quality and Integrity:**

* Investigate and address identified data anomalies.
* Improve data collection processes.

**User Segmentation and Targeting:**

* Identify high-risk users based on behavior and usage patterns.
* Implement personalized interventions for at-risk users.

**Continuous Improvement:**

* Monitor churn metrics and adapt strategies accordingly.
* Adopt an iterative approach for ongoing optimization
* Given what you know about the data and the visualizations you were using, what other questions could you research for the team?
* **Long-Distance Drivers:** What specific pain points and feature preferences do they have?
* **User Engagement:** How can personalized recommendations and incentives be optimized to improve engagement and habit formation?
* **Data Anomalies:** What caused the limitation on driving days and the usage spikes among long-time users?
* **User Demographics:** Are there specific demographic groups with higher churn risk?
* **App Features:** Which features are most important for retaining users, and are there any that contribute to frustration?
* How might you share these visualizations with different audiences?

**Executive Audience (e.g., Harriet Hadzic, Director of Data Analysis):**

* **Focus:** High-level insights, key trends, and actionable recommendations.
* **Format:** Concise presentation or report with clear visuals and minimal technical details.
* **Delivery:** In-person presentation, executive summary document, or interactive dashboard.
* **Emphasis:** Business impact of churn, potential solutions, and data-driven decision-making.

**2. Product and Engineering Teams:**

* **Focus:** Detailed insights into user behavior, feature usage, and potential areas for improvement.
* **Format:** Interactive dashboards, detailed reports with supporting data, and visualizations highlighting specific features or user flows.
* **Delivery:** Team meetings, online dashboards accessible to team members, or shared documents with collaborative features.
* **Emphasis:** User experience, feature development priorities, and data-driven product enhancements.

**3. Marketing and Communications Teams:**

* **Focus:** User segmentation, messaging strategies, and campaign targeting opportunities.
* **Format:** Visualizations highlighting user segments, churn drivers, and potential engagement tactics. Reports with actionable insights for marketing campaigns.
* **Delivery:** Presentations, marketing reports, or data visualizations embedded in marketing materials.
* **Emphasis:** Understanding user preferences, targeted messaging, and campaign optimization.

**4. Data Science and Analytics Team:**

* **Focus:** Detailed analysis, methodology, and technical insights.
* **Format:** Comprehensive reports with detailed visualizations, code snippets, and statistical analysis. Interactive notebooks or dashboards for exploring the data.
* **Delivery:** Team meetings, code repositories, shared notebooks, or internal wikis.
* **Emphasis:** Data exploration, model development, and in-depth analysis of user behavior.

**5. External Stakeholders (e.g., investors, partners):**

* **Focus:** High-level overview of churn trends, business impact, and strategies for improvement.
* **Format:** Public reports, presentations with clear and accessible visualizations, or interactive dashboards with limited data sensitivity.
* **Delivery:** Public website, investor presentations, or partner meetings.
* **Emphasis:** Transparency, business performance, and future outlook.