

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

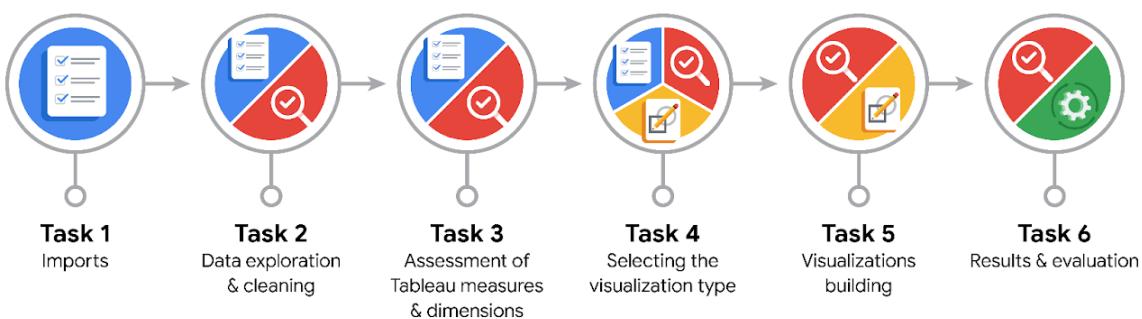
Completing the end-of-course project will help you respond these types of questions that are often asked during the interview process:

- 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?

ID, label, sessions, drives, device, total_sessions, n_days_after_onboarding, total_navigations_fav1/2, driven_km_drives, duration_minutes_drives, activity_days, driving_days. The variables that are most relevant to the deliverable are: label, sessions, drives, total_sessions, n_days_after_onboarding, driven_km_drives, duration_minutes_drives, activity_days and driving_days.

- What units are your variables in?

Integers, floats and categorical variables.

- 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?

That consistency of use across days will be more strongly associated with retention than raw usage intensity, and that newer users are expected to exhibit higher churn rates than longer-tenured users.



The device type of the user is presumed to have a minimal impact on user churn behaviour, and the user engagement metrics will be right-skewed with a small subset of highly active users driving the extreme values.

- Is there any missing or incomplete data?

Yes, the label column has 700 missing values.

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

Yes, the dataset is mostly consistent with the numeric variables stored as integers or floats and categorical variables stored as objects. Some derived fields required minor cleaning, but the overall data is uniformly formatted and suitable for analysis.

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

Discovering and Structuring.



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 in the most effective way, the steps needed are to address the missing values, identify and deal with outliers, analyze and visualize the distribution of the variables, explore relationships between features and user churn and compute summary statistics.



- 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 is needed as the dataset is sufficient for EDA. However, some structuring is needed such as filtering for outliers, handling missing and infinite values and creating derived features. Sorting and grouping by churn status and tenure are also useful to support comparison and interpretation.

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

The types of visualizations that might be best suited for the intended audience are box plots, histograms, scatter plots and bar charts.



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?

Box plots, histograms, scatter plots, pie charts and bar charts.

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

Clean and validate the data, choose the appropriate chart types and aggregate where needed.

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

The variables that are most applicable for the visualizations are label, sessions, drives, total_sessions, n_days_after_onboarding, driven_km_drives, duration_minutes_drives, activity_days and driving_days.



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

If the missing data is missing completely at random, then I would proceed with the analysis by removing the rows with missing values. If the data is not missing completely at random, then I would investigate the root cause of the missingness and ensure it would not interfere with the statistical inference and modelling.



PACE: Execute Stage

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

The key insights that emerged from the EDA visualizations are that the user engagement variables are right-skewed with significant outliers, churn is more strongly associated with inconsistent usage patterns than with high usage intensity and that the device type shows no meaningful relationship with user churn. The newer users in the dataset are much more prevalent and showcase a higher churn risk.

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

The recommendations I propose based on the visualizations built are to focus the retention efforts on increasing usage consistency than the driving intensity, especially for the newer and high-intensity users who display a higher churn risk. Another recommendation is to implement targeted engagement nudges, onboarding improvements and early-warning indicators to identify and support users with declining activity patterns.

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

Other questions I could research for the team are whether the changes in usage patterns precede user churn? How does onboarding experience and tenure influence retention? Whether high-intensity users represent distinct segments with different needs?



- How might you share these visualizations with different audiences?

I would ensure that these visualizations are easily interpretable to non-technical audiences and highlight the clear data stories of user churn rate, as well as ensuring design thinking is enforced for those with disabilities.