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~~

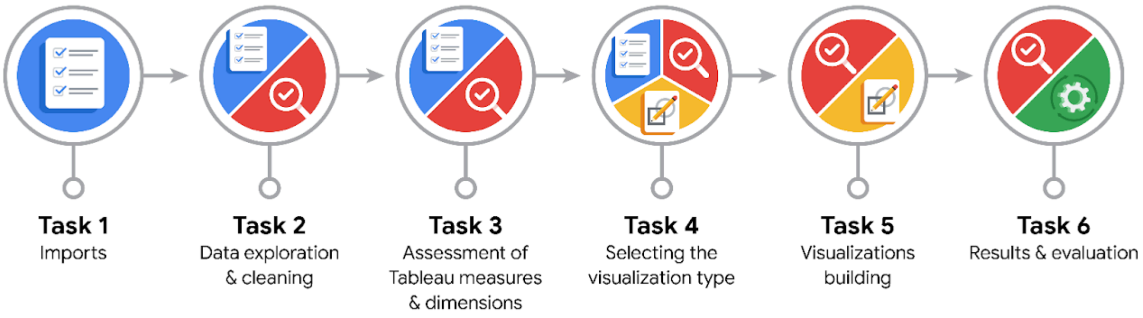
Relevant Interview Questions

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.



Task 1
Imports

Task 2
Data exploration & cleaning

Task 3
Assessment of Tableau measures & dimensions

Task 4
Selecting the visualization type

Task 5
Visualizations building

Task 6
Results & evaluation

Data Project Questions & Considerations

PACE: Plan Stage

- What are the data columns and variables and which ones are most relevant to your deliverable?

`ID`: Unique user identifier

`label`: Indicates whether the user churned or was retained

`sessions`: Number of app sessions during the month

`drives`: Number of driving occurrences

`total_sessions`: Estimated total app sessions since onboarding

`n_days_after_onboarding`: Days since the user joined Waze

`total_navigations_fav1` / `total_navigations_fav2`: Usage of favorite destinations `driven_km_drives`: Total kilometers driven

`duration_minutes_drives`: Total driving minutes

`activity_days`: Number of days the user opened the app

`driving_days`: Number of days the user drove

`device`: Type of device used (Android/iPhone)

The most relevant variables for churn analysis are **`label`**, **`sessions`**, **`drives`**, **`driving_days`**, and **`driven_km_drives`**, since these directly reflect user engagement and driving activity.

- What units are your variables in?

Sessions, drives, activity_days, driving_days: Counts (integer values)

Total sessions, n_days_after_onboarding: Counts or days

driven_km_drives: Kilometers

duration_minutes_drives: Minutes

Device and label: Categorical variables (e.g., Android/iPhone, retained/churned)

- 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?

Users who **drive or use the app more frequently** are less likely to churn.

Long-distance drivers may show higher churn rates due to fatigue or different travel behavior.

iPhone and Android users will have similar churn rates.

Engagement level (sessions and drives) will be the strongest predictor of retention.

- Is there any missing or incomplete data?

Most columns have complete data. However, the **`label`** column has a few missing entries that represent users with unclear churn status. Other variables such as `total_navigations_fav1` or `fav2` may contain zeros, which indicate no recorded activity rather than missing data.

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

Yes, all columns are in a consistent format — numerical columns are either `int64` or `float64`, and categorical columns like `label` and `device` are `object` type. No conversion issues were detected during inspection with `df.info()`.

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

Use `.info()` and `.describe()` to summarize structure and data distribution.

Identify missing values using `.isnull().sum()`.

Visualize distributions using **box plots** and **histograms** to detect outliers.

Apply **correlation analysis** and **scatter plots** to understand variable relationships.

Perform **data cleaning** (e.g., handling missing values, capping outliers at the 95th percentile).

PACE: Analyze Stage

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

Inspect the dataset using `.info()` and `.describe()` to understand data types, missing values, and basic statistics.

Clean and preprocess the data: handle missing or infinite values, remove or cap outliers (e.g., at the 95th percentile).

Explore distributions of key variables (`sessions`, `drives`, `activity_days`, `driving_days`, `driven_km_drives`) using histograms and box plots.

Analyze relationships between engagement metrics and churn (`label`) using scatter plots and correlation heatmaps.

Segment users by device type, activity level, and driving distance to identify behavioral patterns.

Summarize insights to guide retention strategy recommendations.

These steps ensure that EDA not only detects data issues but also supports identifying the strongest predictors of 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.?

No additional datasets are required; all relevant churn, activity, and demographic information is already included.

However, **structuring and filtering** are needed to improve clarity:

Filter out extreme outliers (e.g., users driving more than 1,200 km/day).

Sort and group users by churn label and device type to compare retention patterns.

Create derived columns, such as `km_per_driving_day`, to better understand user behavior.

Normalize values where appropriate to ensure fair comparisons between users.

These transformations make the data cleaner and more comparable, enabling accurate visualization and analysis.

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

Since the audience includes **Waze's data analysis team and business stakeholders**, visualizations should be clear, engaging, and easy to interpret:

Pie chart for churn vs. retained proportions.

Histograms for usage distribution across metrics like `activity_days` and `driving_days`.

Scatter plots for exploring relationships (e.g., distance vs. churn).

Bar charts to compare churn rates across devices or engagement levels.

Box plots to highlight outliers and variation in user activity.

Visuals will prioritize simplicity, color consistency, and interpretability to support data-driven decisions about user retention.

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?

This project focuses on **data visualizations**, not machine learning. The following visualizations will be built to explore churn and engagement:

Pie charts to show proportions of churned vs. retained users and device distribution.

Histograms for the distributions of sessions, activity days, and driving distances.

Scatter plots to identify correlations between activity metrics (e.g., driving_days vs. activity_days).

Box plots to identify outliers and compare engagement patterns.

Bar charts to display churn by device type and engagement level.

Histograms with multiple fill colors to show churn rate across continuous variables like kilometers driven or session count.

These visuals will help reveal user behavior patterns and factors influencing churn.

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

- **Prepare and clean** the data (handle missing values, replace infinite values, remove extreme outliers).
- **Transform** data by creating derived columns such as `km_per_driving_day`.
- **Group and aggregate** data (e.g., by churn label, device type, or activity range).
- **Use visualization libraries** like Matplotlib and Seaborn to build and format charts.
- **Refine visuals** for readability — add titles, legends, labels, and consistent color palettes.
- **Interpret patterns** and summarize findings to link data visuals to the churn insights.

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

Engagement metrics: `sessions`, `drives`, `activity_days`, `driving_days`, `total_sessions`

Distance metrics: `driven_km_drives`, `km_per_driving_day`, `duration_minutes_drives`

User type indicators: `device`, `label`

Tenure and history: `n_days_after_onboarding`

These variables capture user activity, driving behavior, and churn outcomes, making them central to visual analysis.

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

- Identify missing values using `.isnull().sum()`.
- For **numeric variables**, impute missing values with median or mean if appropriate.
- For **categorical variables** (like `device`), use the most frequent category.
- Treat zeros carefully — for instance, zero in `drives` may represent a valid case (non-driver), not missing data.
- Document and justify all handling decisions to maintain transparency.

PACE: Execute Stage

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

The **overall churn rate** is approximately **17%**, consistent between iPhone and Android users.

Driving frequency strongly correlates with retention — users who drive or open the app more often are less likely to churn.

Users who drive longer distances per day are slightly more likely to churn, possibly due to fatigue or changes in travel behavior.

Engagement patterns (sessions and activity days) reveal that active app users tend to maintain loyalty.

Some **outliers** in long-distance driving and session frequency required capping at the 95th percentile to ensure realistic analysis.

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

- **Improve retention among long-distance drivers** — provide features such as route-based rewards, fatigue alerts, or entertainment integrations.
- **Encourage frequent engagement** through gamification or achievement badges for regular app usage.
- **Personalize user notifications** — send targeted reminders to infrequent users who show early signs of inactivity.
- **Analyze regional behavior** to identify external factors (e.g., seasonal driving changes or commuting habits).
- **Regularly monitor churn trends** to identify shifts caused by app updates or policy changes.

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

- What external factors (weather, travel restrictions, or gas prices) influence churn behavior?
- Does user churn differ across **regions or countries**?
- How does **app update frequency or performance** impact user retention?
- Are there **specific driving distances or times of day** that correlate with disengagement?
- Could **survey feedback** explain why some frequent drivers still churn?

- How might you share these visualizations with different audiences?

For **executives**: create a **concise dashboard or slide deck** summarizing churn rate trends and actionable recommendations.

For the **data team**: share **detailed Jupyter notebooks** or Tableau dashboards with interactive filters.

For **marketing and operations teams**: present simplified charts emphasizing user behavior insights and engagement strategies.

For **external stakeholders or partners**: prepare a **brief report or infographic** highlighting retention metrics and progress over time.