

Waze PACE Strategy Document III

Exploratory Data Analysis

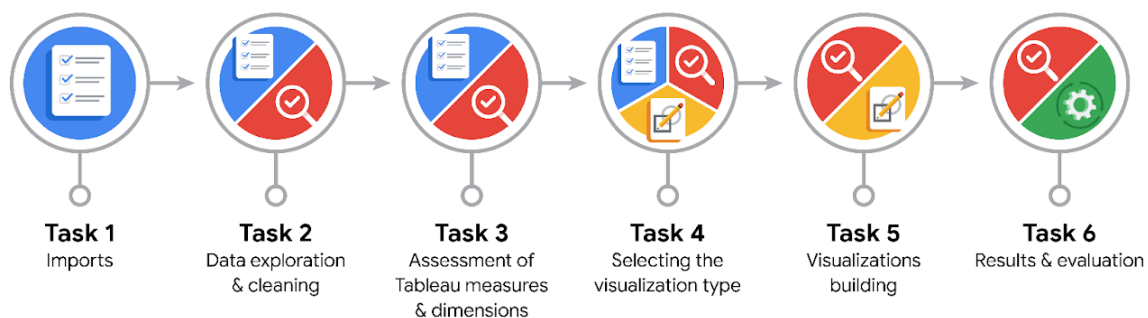
Introduction

PACE stands for Plan, Analyze, Construct and Execute. It is a framework that illustrates the foundation and structure for data analysis projects and each letter represents an actionable stage in a project. The stage “Plan” involves the definition of the project scope, the research of business data and the workflow development. The stage “Analyze” involves data scrubbing, data conversion and database formatting. The stage “Construct” involves building models and machine learning algorithms and selecting a modeling approach. The stage “Execute” involves the presentation of results to decision-makers, stakeholders and others in order to receive feedback. This framework is built upon an iterative cycle where each stage may reveal new insights, requiring the return to earlier stages. A PACE strategy document is used to record decisions and reflections at different stages of the data analytical process. It typically includes the definitions of roles and actions to ensure clarity and accountability.



Purpose

Waze’s free navigation app makes it easier for drivers around the world to get to where they want to go. We want to analyze user data and develop a machine learning model that predicts user churn on the Waze app. Churn quantifies the number of users who have uninstalled the Waze app or stopped using the app. The project focuses on monthly user churn. Developing a churn prediction model will help prevent churn, improve user retention and grow Waze’s business. An accurate model can also help identify specific factors that contribute to churn and answer questions such as “Who are the users most likely to churn?”, “Why do users churn?” and “When do users churn?”. For this stage of the project, we identify six main tasks that are presented in the following visual.





Considerations



PACE: Plan Stage

- What methods are best for identifying outliers?

We use *NumPy* to calculate the mean and median and understand the data range. We create boxplots and histograms to visualize and observe how the data is distributed. We also create scatter plots which allow us to observe trends, patterns and outliers between variables.

- How do we decide whether to keep or remove outliers from future models?

There are three main ways to handle outliers: retaining them, removing them or reassigning values. The choice depends on the nature of the outliers and the model's assumptions. If the outliers are clearly errors like typos and the data will be used in modeling or machine learning, removing them is often best (though this is the least common option). If the dataset is small or used for modeling, reassigning values might be necessary. If the data will be used for EDA or the model is resistant to outliers, it is usually best to keep them.

- Which columns and variables in the dataset are most critical for our analysis and final deliverable?

Relevant variables include the label, the number of sessions, the number of completed drives, the number of days a user drove, the number of days a user opened the app, the distance driven and the device type.

- What are the units of measurement for each variable?

Sessions, drives, driving days and activity days are discrete variables that are measured as counts (number of occurrences). The distance driven is measured in kilometers, the duration in minutes and the label is a categorical variable for the churn status.

- What are our initial assumptions about the data that will guide our exploratory data analysis and what do we expect to confirm or adjust through our findings?

It is assumed that users who have higher app engagement (e.g. more sessions, more drives) are less likely to churn. The dataset may have some outliers or anomalies (e.g. extremely high distances driven). Variables like distance driven and driving days may have a positive correlation with churn (users who drive long distances more frequently may be more likely to churn). These assumptions will need verification through further analysis.

- Is there any missing or incomplete data in the dataset?

There is missing data, particularly in the label (churn) column, with around 700 rows missing values. All other variables appear to be complete.

- Is the dataset consistent in terms of format?

For the most part, the dataset is structured consistently. However, there are some discrepancies, such as mismatched maximum values between driving days and activity days (31 vs. 30), which may indicate inconsistent data collection periods.

- What exploratory data analysis techniques will we need to use to start this activity?

Key EDA practices include summarizing the dataset with measures like mean, median, standard deviation and interquartile ranges, data cleaning, outlier detection, distribution analysis and correlation analysis.




PACE: Analyze Stage

- What actions should be taken during EDA to ensure the project reaches its objectives?

We should Impute or remove rows with missing churn labels, identify outliers using box plots and create visualizations like histograms and scatter plots to understand distributions and relationships between variables.

- Is there a need to combine more datasets as part of the EDA process and what structuring tasks (such as filtering or sorting) are necessary for the current data?

There is no immediate need to combine more datasets unless new information (e.g. demographics, external user behavior) is available. Structuring tasks include filtering irrelevant variables (like user ID) and sorting based on key variables, such as engagement levels.

- 
- What initial ideas do we have about which types of visualizations might be most effective for our audience?

Histograms will be ideal for illustrating the distribution of sessions and drives, box plots for highlighting outliers and spread in variables like distance driven, scatter plots for showing relationships between churn and variables like driving days or sessions and bar charts for categorical data comparisons (e.g. iPhone vs. Android users).



PACE: Construct Stage

- What visualizations, machine learning models or other outputs will be essential to achieve the project's goals?

Box plots, histograms, scatter plots, bar charts and possibly a classification model to predict user churn.

- What steps are needed to create the necessary visualizations?

Data cleaning, filtering outliers and generating visualizations like box plots, histograms, scatter plots and bar charts.

- Which variables should be prioritized in the data visualizations for this project?

The number of sessions, drives, distance driven and activity days should be the priority. These variables likely have the strongest relationship with churn.

- Returning to the planning stage, how will we address any missing data that is identified?

For the label column, we can either impute values using statistical methods and machine learning techniques or remove rows with missing churn labels if the missingness is random.



PACE: Execute Stage

- What key insights have we gathered from your EDA and visualizations?

Users with more app sessions or drives are less likely to churn. Long-distance drivers tend to churn more frequently. The number of sessions and drives have right-skewed distributions, indicating that most users have lower engagement. Additionally, anomalies exist in some variables (e.g. extreme distance values), requiring further investigation.

- What recommendations would we provide to the organization based on our findings?

The focus should be on retaining long-distance drivers by addressing any pain points that may be leading them to churn. Users with very high engagement in the last month should be investigated to understand any shifts in behavior. The potential impact of iPhone against Android users should be further analyzed, as their churn rates are consistent but usage patterns may differ.

- Based on our current knowledge of the data and visualizations, what additional research questions could be investigated for the team?

Key research questions include:

- What factors contribute to the high churn rate among long-distance drivers?
- Why do certain users have spikes in engagement (sessions, drives) in the last month?
- How does churn behavior differ based on geographic location or driving patterns?

- How do we plan to present these visualizations to different stakeholders or audiences?

Dashboards like Tableau provide a clear summary of key trends and patterns, tailoring the presentation for managers or technical audiences. In general, for non-technical audiences, we should use simplified bar charts and accessible Tableau dashboards with clear explanations. For data-savvy stakeholders, we can present detailed boxplots, histograms and correlation analyses to explain deeper insights into data patterns.