

# Exploratory Data Analysis of Retention data

Charly Moreno - Data Analyst (Waze's Data Analysis Team)

## Overview

The **Data Analysis Team** is making progress on their **project to develop a machine learning model to predict user churn**. A users data snippet was generate to perform initial EDA and understand current trends of the relevant variables that will be used to develop the final model and tackle Waze's retention problem.

## Objective

The purpose of this project is to **conduct exploratory data analysis (EDA) on the provided dataset**.

The goal is to examine the data that, adding relevant visualizations that help communicate the story that the data tells, and finding any data problem that may be behind it.

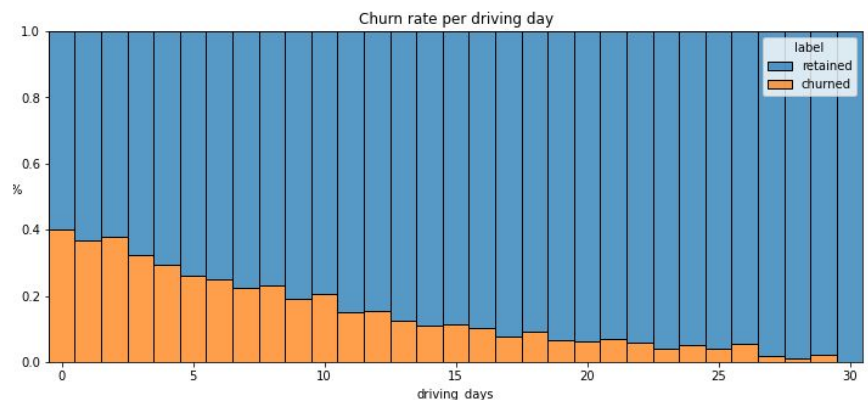
## Results

We have **17.8% of users who churned**, against 82.2% who don't.

**Six variables** were identified with **relevant correlation** with the **churners**: 'activity\_days' (correlation -0.3 ), less 'driving\_days' (-0.29), 'km\_per\_driving\_day' (0.15), 'n\_days\_after\_onboarding' (-0.13), 'sessions\_per\_activity\_day' (0.19) and 'drives\_per\_driving\_day' (0.17). Device Type is not relevant for explaining the Churn.

**Two common distribution types** were found in the variables: **right-skewed** ('duration\_minutes\_drives', 'drives\_km\_drives', 'sessions', 'total\_sessions', 'drives') and **Uniform/SemiUniform** ('n\_days\_after\_onboarding', 'driving\_days', 'activity\_days').

Some **questions require further research**: Why 60% of long-time users suddenly used the app in the last month? Why do long-distance users churn? Why 700 records with null values



## Next Steps

1. **Confirm that the monthly variables were collected during the same month**, given the fact that some have max values of 30 days while others have 31 days.
2. **Consider the variables identified as possible candidates in a future churn model**.