Title

Waze's Data Team Exploratory Data Analysis (EDA) Project

Overview

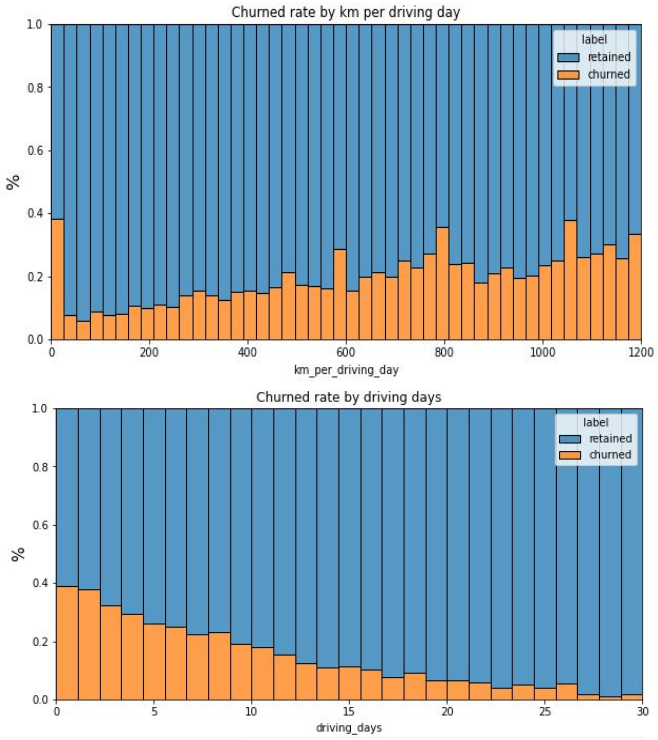
This project aims to perform the Exploratory Data Analysis (EDA) process after getting a general understanding of the given dataset.

Objective

The goal is to continue exploring the dataset by performing the EDA process through steps such as: cleaning, analyzing, generating, finding important key points and insights through visualizations, and presenting what we have found. After presenting the EDA process, we might choose to be ready for steps in building statistical models, and machine learning algorithms. If necessary, we might continue with further analysis of the data based on the specific needs.

Results

- Key variables related to the churn rate are activity days and driving days. These two are correlated with each other.
- The number of newer and longer tenures is quite evenly on the number of using the app last month.
- Mostly the visualizations (box plots and histograms) show there are extreme values and anomalies. Also, both the right-skew and uniform distributions are present for variables respectively, but mostly the right-skew.
- There are several users (~250) who didn't open the app at all last month, but there are ~1,000 users who don't drive at all, which might be counterintuitive.
- The percentage of retained users is ~82%, and churned users are ~18%.
- The rate is consistent when compared to Android and iPhone for both users, approximately 36% and 64% respectively.
- Users tend to churn when they drive on long-distance driving days, as well as long-duration driving time.
- They also tend to churn when they have fewer or no driving and active days during a month, with ~40% respectively.
- However, they are less likely to churn when they drive and are active more days in a month.



Next Steps

- Getting deeper analysis of the number of churned users who drive long-distance about why they choose to churn. Determining the demographics of these users.
- Further researching the factors that make users churn when they have less or no driving or active days.
- Performing validating process on variables that produce anomalies (driving_km_drives, duration_minutes_drives, activity_days, driving_days), which might be problematic or something relevant.
- Analyzing the activity days and driving days to check that the values in these variables are represented correctly.