Waze User Churn Project: Results of EDA

Waze is a free navigation app that makes it easier for drivers around the world to get to where they want to go. Waze leadership wants to optimize the company's retention strategy, enhance user experience, and make data-driven decisions about product development. They would like an analysis of WAZE data to understand their users better and the development of a machine learning model that predicts user churn. (Churn is understood to be the number of users who have uninstalled the Waze app or stopped using it.)

This project is part of a larger effort at Waze to increase growth. It assumes that high retention rates indicate satisfied users who repeatedly employ the Waze app over time. Identifying and predicting which users are likely to churn will allow the WAZE team to target such individuals to induce their retention, thereby allowing Waze to grow its business.

Data

Waze's free navigation app makes it easier for drivers around the world to get to where they want to go. Waze's community of map editors, beta testers, translators, partners, and users help make each drive better and safer. Waze partners with cities, transportation authorities, broadcasters, businesses, and first responders to help as many people as possible travel more efficiently and safely. The data set is in-house from Waze for Cities (https://www.transportation.gov/office-policy/transportation-policy/faq-waze-data).

Deliverables

(Since this is an exercise, all models are predetermined. All Python code can be located at: https://github.com/izsolnay/WAZE_Python)

I. An analysis of WAZE data to understand their users better
II. The development of a machine learning model that predicts user churn
    a. A binomial logistic regression model
    b. A winning tree-based model

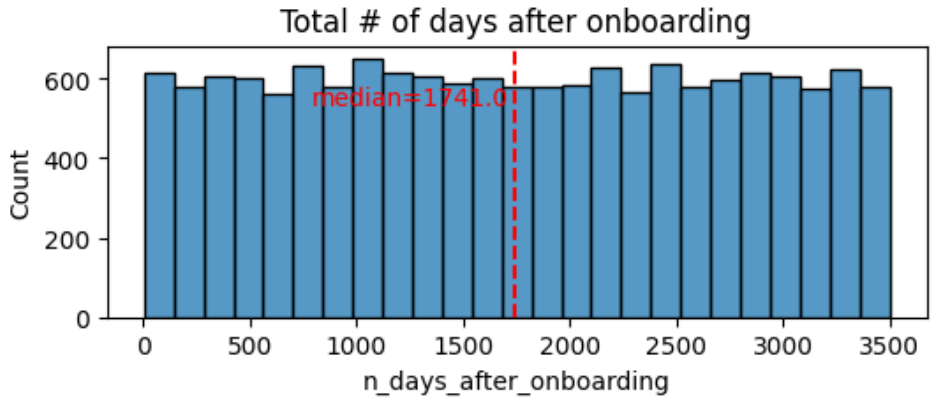
Appendix: A 2 sample t-test based on a sample of user data determining if there is a statistically significant difference in the mean number of rides between iPhone® users and Android™ users

Preliminary conclusions for part I

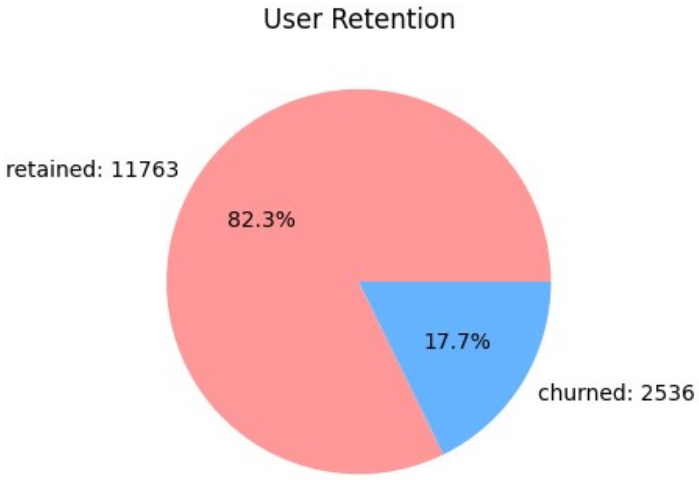
The data set contains the records of just shy of 15,000 users (14,999). The bulk of the data is from the (previous) month and contains information such as how long they've had the app, the total number of days a user drove, how many kilometers they drove during the month, how many minutes, and what type of device they used (iPhone or Android).

Basics

There is a uniform distribution in the number of days since a user on boarded; roughly 600 users for any given length of time:

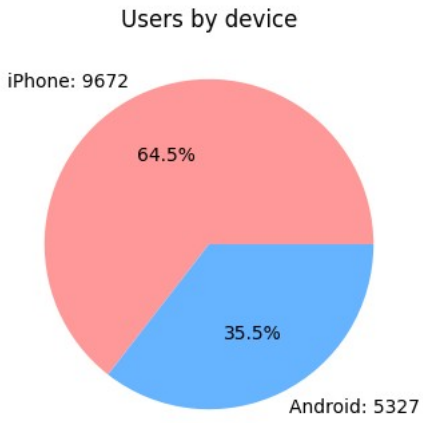


The median number of days since a user onboarded is 1741 or about 4.8 years. The longest a user has been had the app is ~3,500 days or about 9.5 years. Of these users, only about 18% churned (2536), while the remaining 82% (11763) remained loyal to the app:

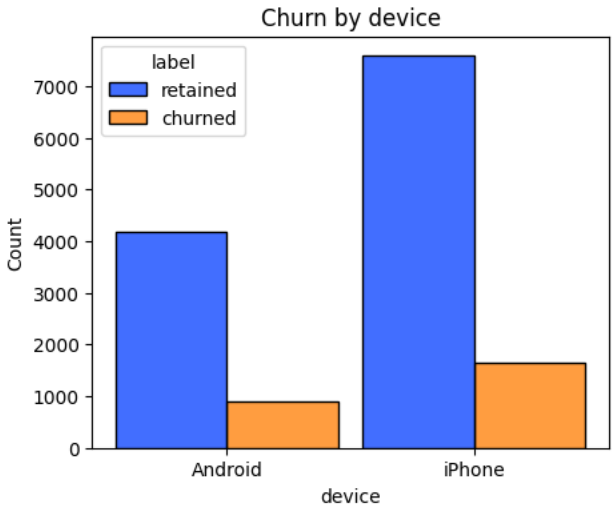


Type of device

About 64% of users opened the app with an iPhone and 36% with an Android model:



Of these, the rates of churn reflect the percentages of iPhone and Android subsets:



Of the 18% of users who churned (2536), about 65% (1645) were iPhone users, and 35% (891) Android users. Of those who were retained, about 65% were iPhone users (7580), and 35% (4183) Android users. Thus, the ratio of remaining and churning iPhone users to Android users is the same regardless of category:

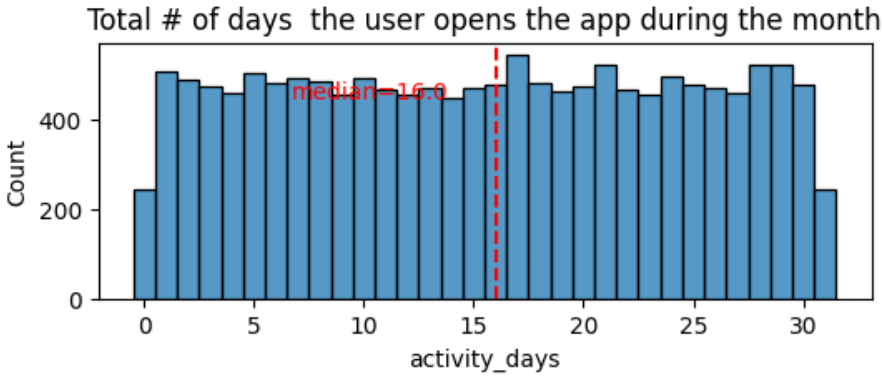
- ~65% of retained users and churned users use iPhones
- ~35% of retained users and churned users use Androids

(The results of the hypothesis testing (a 2 sample t-test) as to whether there is a statistically relevant correlation between device use and churn are in the appendix.)

ALL USERS

Use of app for any reason during the month

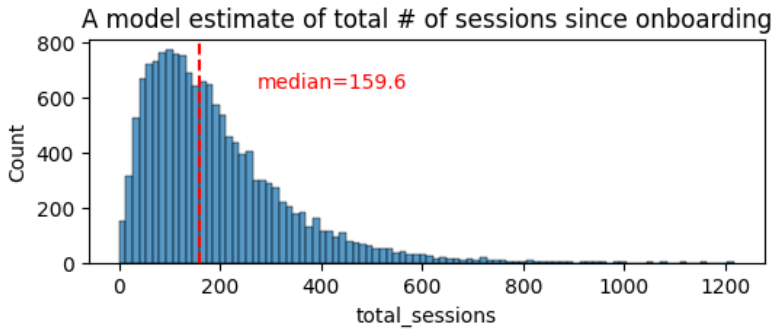
For the last month, there is a fairly uniform distribution of daily use. Roughly 550 users opened the app a given number of days in the month:



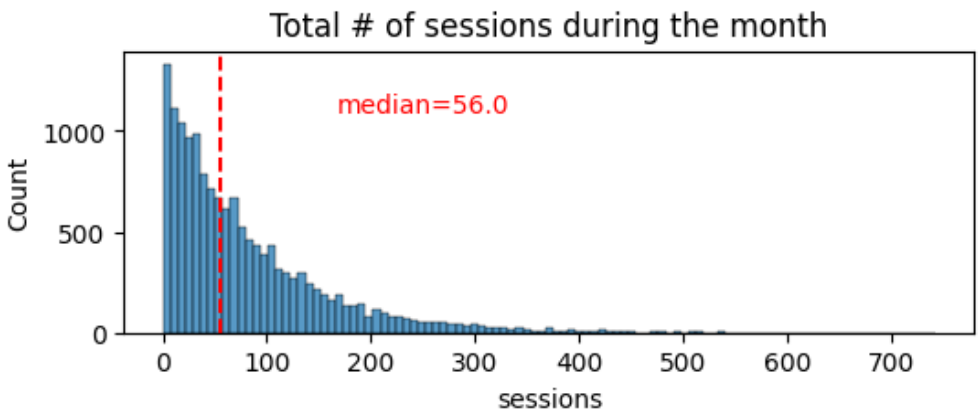
This is confirmed by the median number of days the user opened the app during the month at 16. With the spread going from 0 day to 31. Approximately 250 users never opened the app while about 250 opened it daily.

This said, the number of sessions over the month varies dramatically. The median number of occurrences of a user opening the app during the month was 56 (~2x a day), with this stretching from ~20 > 110 (or about 3.6x a day). The outliers have some people opening the app over 700x a month. This amounts of over 23x a day, which seems excessive, but aligns with the finding that the outliers have 19 drives per day.

However, this monthly usage does not align with the usage since onboarding. The median number of occurrences a user opened the app since onboarding was 159.6 and the median number of years since onboarding is ~4.8 years, which would mean that the app was opened on average only 34x a year (or about once a month).

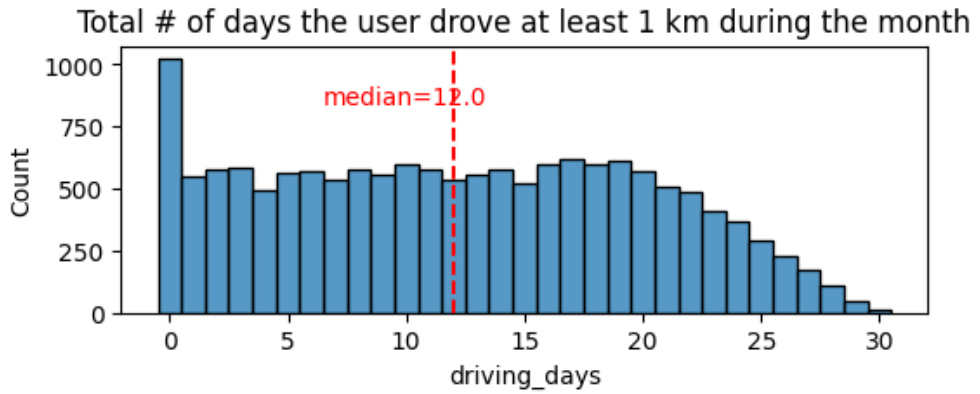


This also contrasts to the results of the variable 'sessions' which records the number of times a specific user opened the app during the last month for any reason. This result was twice a month.



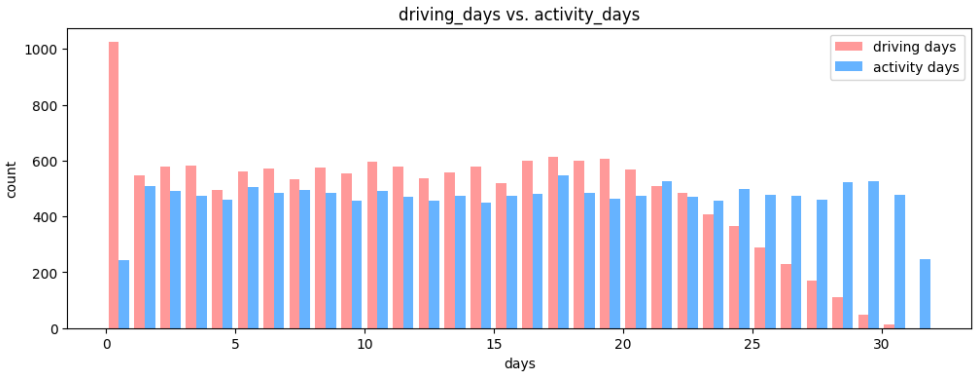
App for driving

As for driving habits, the median number of days a user drove at least 1km during the month is 12. The histogram shows a fairly uniform distribution up until about 20 days out of the month with roughly 550 users driving at least 1km per day:



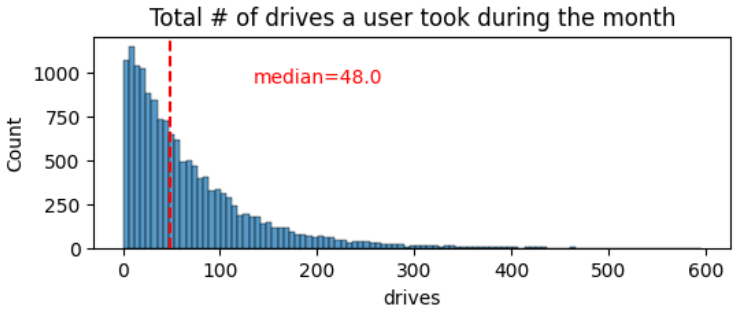
At the 20-day mark, the number of days a user drove at least 1km during the month tapers off to very few. This aligns with commuter driving habits if 20 days of the month are used as work days.

There were also about 1000 users who never drove anywhere, which when the 250 users who never opened the app at all and accidental openings are taken into account suggests that at least 750 users used the app predominantly for checking routes. This is navigator aspect is reflected in the misalignment of activity and driving days.



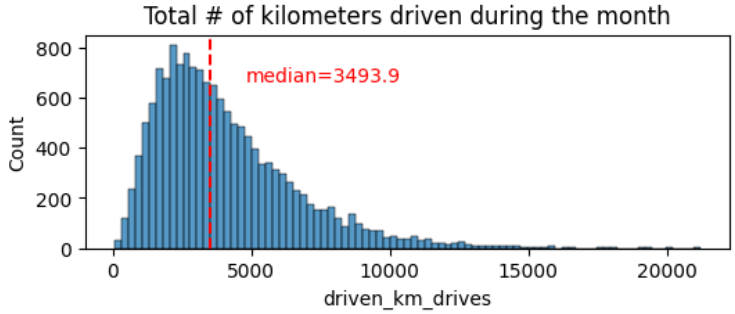
The median number of days the app was opened for any reason was 16 and the median number of days a user drove was 12.

The median total number of drives a user drove during the month was 48:



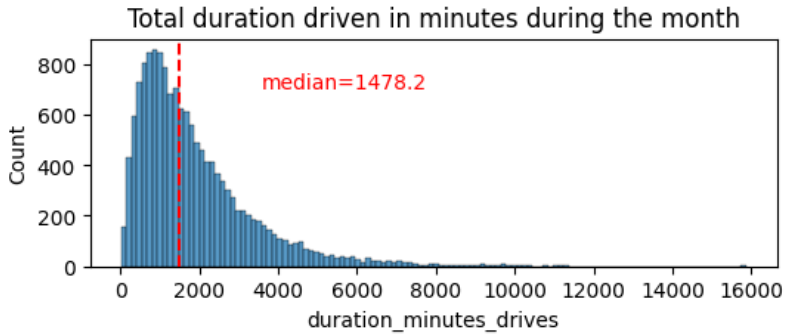
This also aligns with commuter driving habits, coming to about 1½ drives per day or 2 per day on weekdays, with an additional 2 on the weekends. With this said, the swing between number of drives is large. There is the group of 1000 users who never drove, but there is also a group of "super drivers" with over between 200-596 drives for the month. This works out to between 7 and 19 per day.

The median number of total kilometers driven during the month was roughly 3,495 kms or abut 2172 miles. When true driving days are taken into account, this works out to about 84 > 94 miles per day.



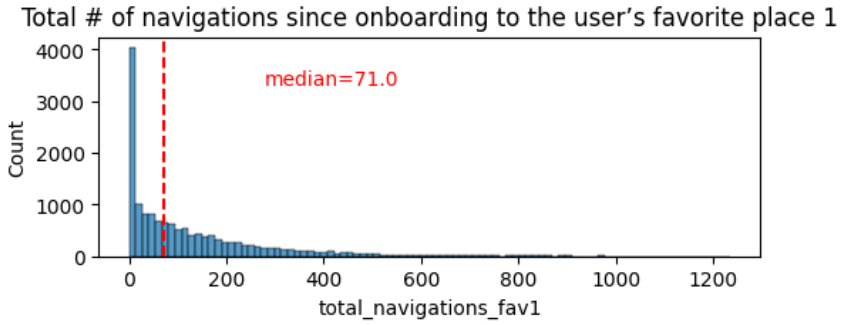
However, again a set of "super drivers" appear. These users are driving between 10000 > 25000 kms during this month. This far end works out to 424.5 miles per day on average.

The median number of total minutes spent driving during the month was 1478 or 25 hrs. If median total miles works out to about 90 and the days users drove tapers down at the 28 days mark, again the data points to commuter drivers.



However, again the "super drivers" appear in the outliers where some users are driving for over 11,500. This comes to 192 hrs for the month or about 8 hrs a day).

Finally, of the 1499 users of the app, 4,000 have never navigated to any favorite place and an additional ~2500 have never navigated to a 2nd favorite place.



On the other hand, the majority of users have navigated at least 100 times to their favorite places, with the outlier group going 1000 times. This latter group is not unreasonable if a user has had the app for at least 4 years and uses it for commuting.

ALL USERS SUMMARY

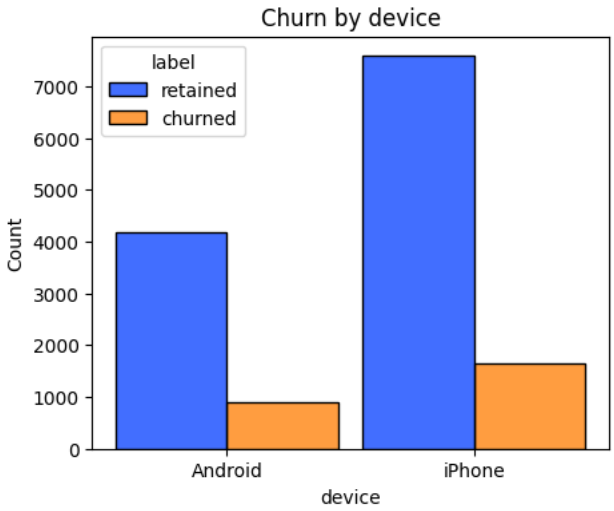
Based on the medians of the variables, the majority of users would seem to be commuters who drive ~90 miles a day, or about 2hrs, 20 days a week. Within this group are also likely commuters with shorter drives who travelled to new places on weekends. There is also a set of users who do not use the app for driving, but instead likely use it to do simple navigation or check travel times.

Statistical analysis also uncovered drivers who could be called super drivers who drive for very long and very far distances. These could be professional drivers.

Finally, there is a variable in the data set 'total_sessions', which is a model estimate of total number of sessions since onboarding. The data contained within it does not align with any of the other variables. The median number of occurrences a user opened the app since onboarding was 159.6. This simply does not align with the median number of users opening the app in the last month at 56 times.

CHURN INITIAL SUMMARY

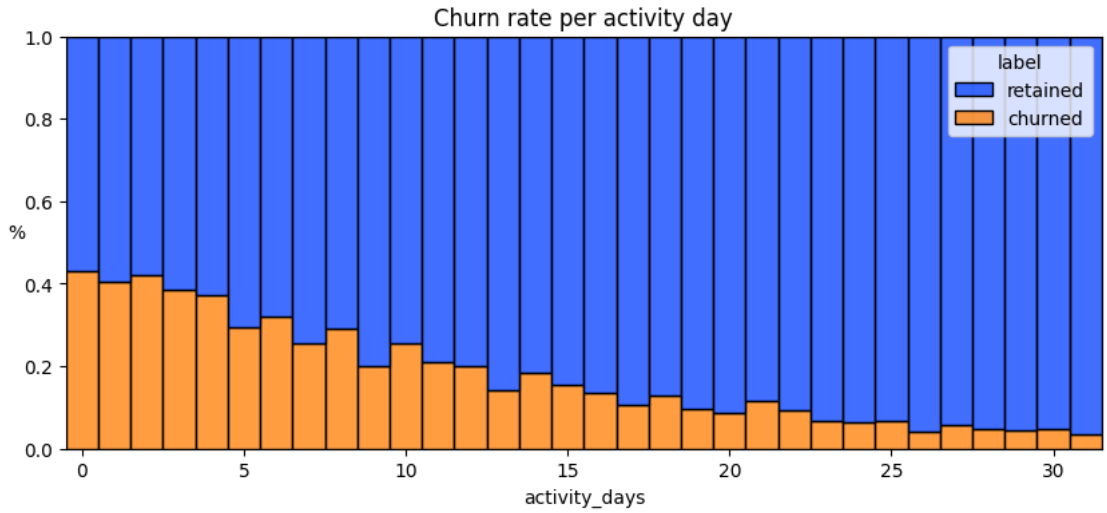
As stated in the beginning of this report, the data set is unbalanced with only 18% of users churning and 82% remaining. The ratio of remaining and churning iPhone users to Android users is the same regardless of category. About 65% of retained users and churned users use iPhones and 35% of retained users and churned users use Androids.



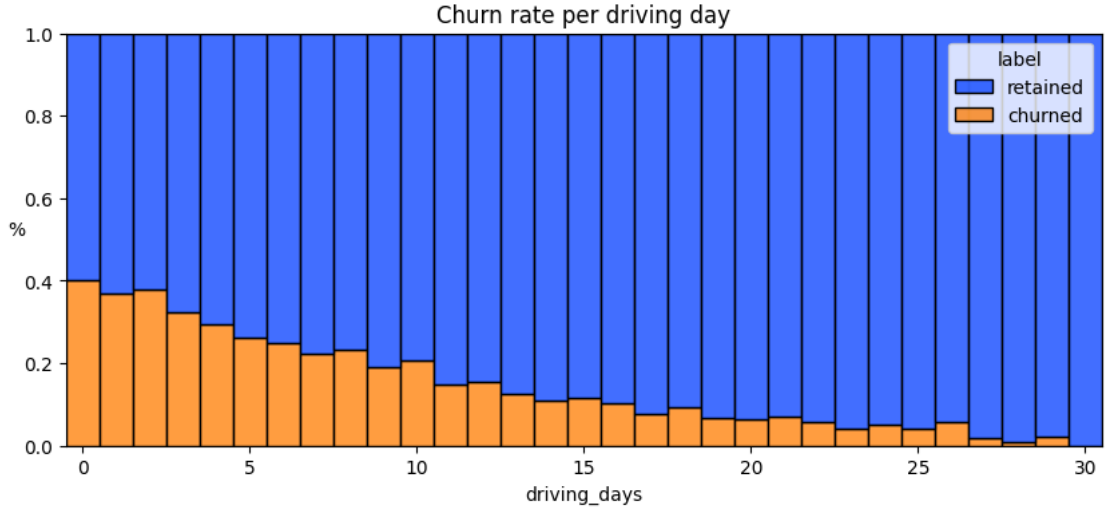
Examining the median churned user, regardless of device type, churned users used the app more for driving while they had it than remaining users. Although, all users drove the same distance per drive, and had almost an even number of drives, sessions, and total sessions since onboarding, churned users drove almost 200 kms more, took almost 2 more drives per day in the month, and for spent roughly 2.5 hours more on the road. Furthermore, they took more advantage of favorite navigations features.

The median churned user did this while having the app for a shorter length of time (about 3.5 years to 5), in less than ½ the number of activity days during the last month (8/17) and ½ the number of driving days (6/14) than retained users during the last month.

The histogram pegs a consistent ~40/60% ratio at about 5 days for the number of days churned users and retained users were active during the month. This ratio gradually widens until about the 20 days mark. At this point, users who churned only make up about 1% of those users who had 20-31 activity days.



A similar phenomenon occurs for days users drove, however tapering off of churned users begins earlier, at about the 3 days mark. Up until 3 days, there is again a ~40/60% churned users to retained users ratio. By the 23 days mark, users who churned are no longer driving with the app (< 1%).



In short, the median churned user:

- drove the same distance per drive, and had almost an even number of drives, sessions, and total sessions since onboarding as retained users

The churned user also:

- used the app more for driving while they had it
- drove for ~2.5 hours more
- took on average more than 2 more drives per day
- took more advantage of the favorite navigations features
- had fewer than ½ the number of activity days (8/17) and driving days (6/14) than retained users
- took more than 2 more drives per day
- had a median tenure of 1.5 years less than the 5 years of the retained user

The curious case of the super user

The data set revealed a "super user" category. The usage statistics of these users, where identifiable, were in the 90th percentile of usage. On the maximum end, in the last month, they opened the app 23x a day, had 19 drives of 1km or more per day, drove over 400 miles a day, and drove for over 8 hrs a day. These drivers are most likely professionals who may intersect with the churned users, since the churned users tended to use the app more than the retained users.

Further steps/considerations

- Create a professional driver feature to address the outlying "super users"
- Perform a hypothesis test
- Perform a logistical regression

Also, recommended

- A revisit of the EDA
- After a request a true total sessions accounting
- And a request for a true total drives taken for the entirety of having the app