# USER CHURN ANALYSIS

EDA AND MACHINE LEARNING MODELLING

# PROJECT OVERVIEW AND GOALS

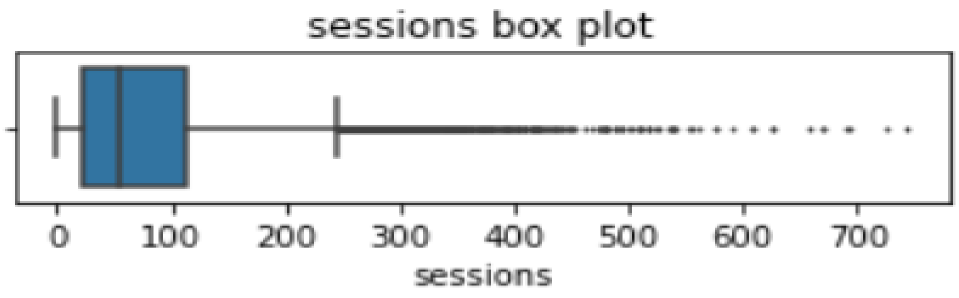
- Waze leadership has asked the data team to build a machine learning model to predict user churn. The model is based on data collected from users of the Waze app.
- We will achieve this through a series of milestones:
  - EDA and Data Visualizations
  - Computing descriptive statistics and conducting hypothesis testing
  - Building a regression model(for comparison) and evaluating that model
  - Building a machine learning model

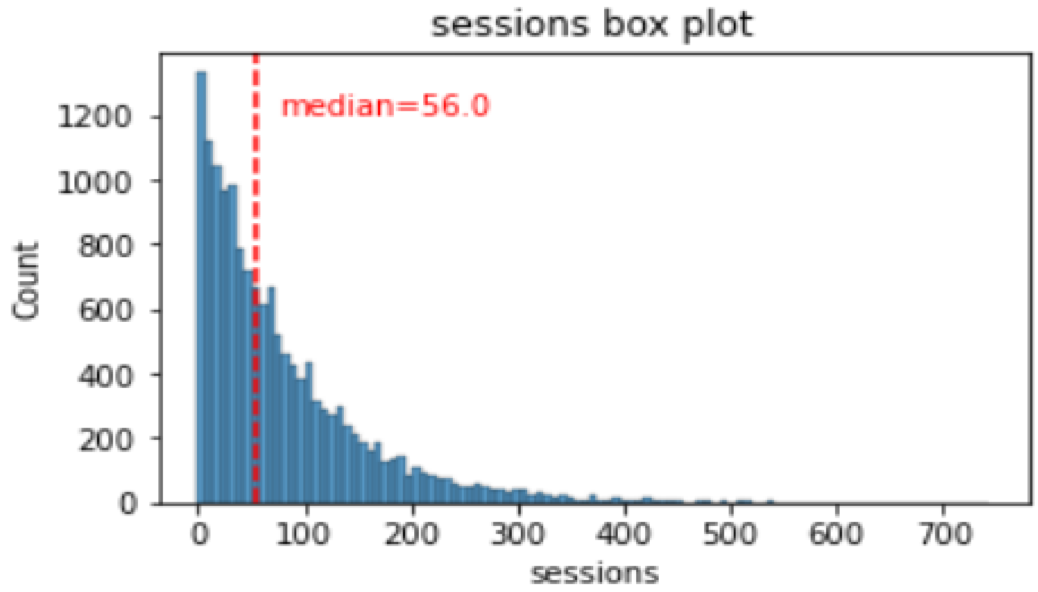
- Based on the data, communicate final insights and any recommendations

# METHODOLOGY AND TECHNOLOGY

- **Data Sources:**
  - Waze User Data(one-month) via waze_dataset.csv
- **Data Cleaning:**
  - Dataset was cleaned using Python *pandas* and *numpy*
- **Exploratory Data Analysis:**
  - EDA performed using Python *pandas, numpy, pyplot,* and *seaborn*
- **Hypothesis Testing:**
  - Hypothesis testing performed with Python *pandas* and *scipy stats*
- **Model Building and Evaluation:**
  - Models built using Python *sklearn.linear_model, RandomForestClassifier, XGBClassifier*

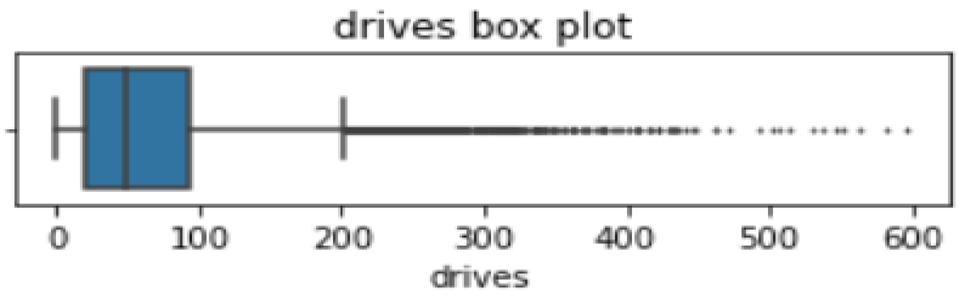
# SESSIONS

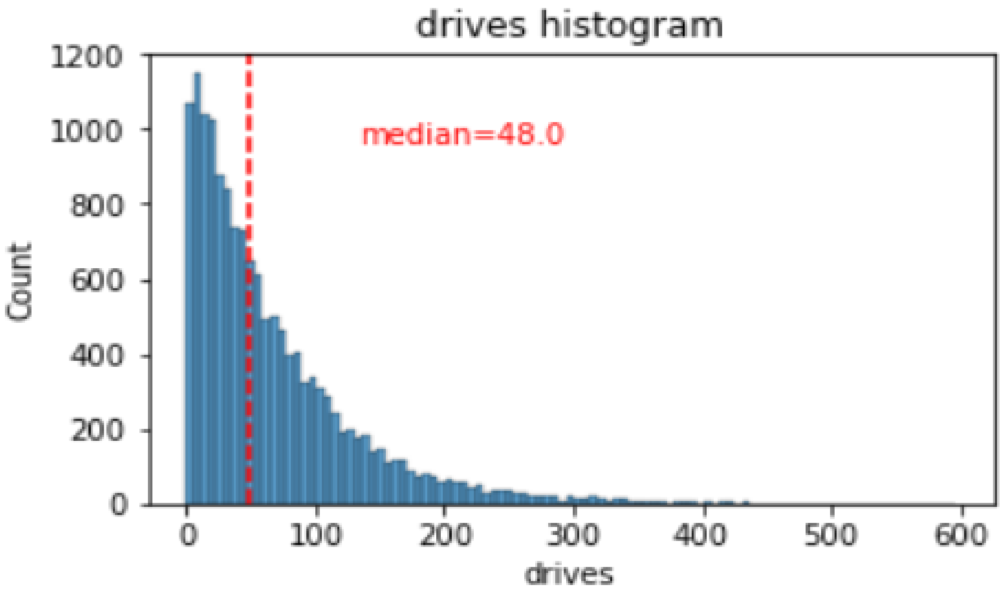



- The boxplot reveals that a **subset of users** has **more than 700 sessions**.
- The **median** number of session is 56.
- The sessions variable exhibits a **skewed distribution to the right**, where approximately **50%** of the **observations consist of 56 sessions or fewer.**

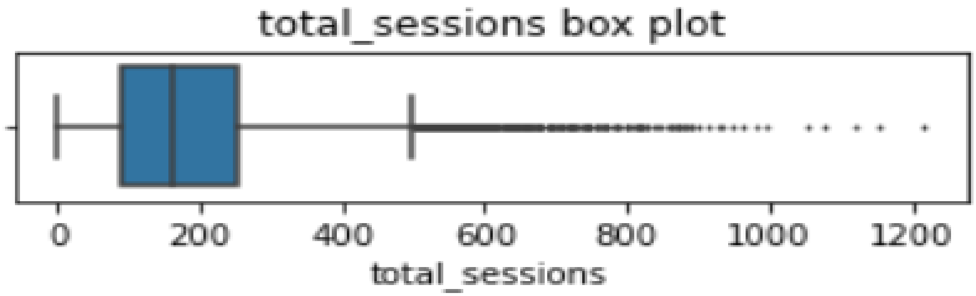
# DRIVES

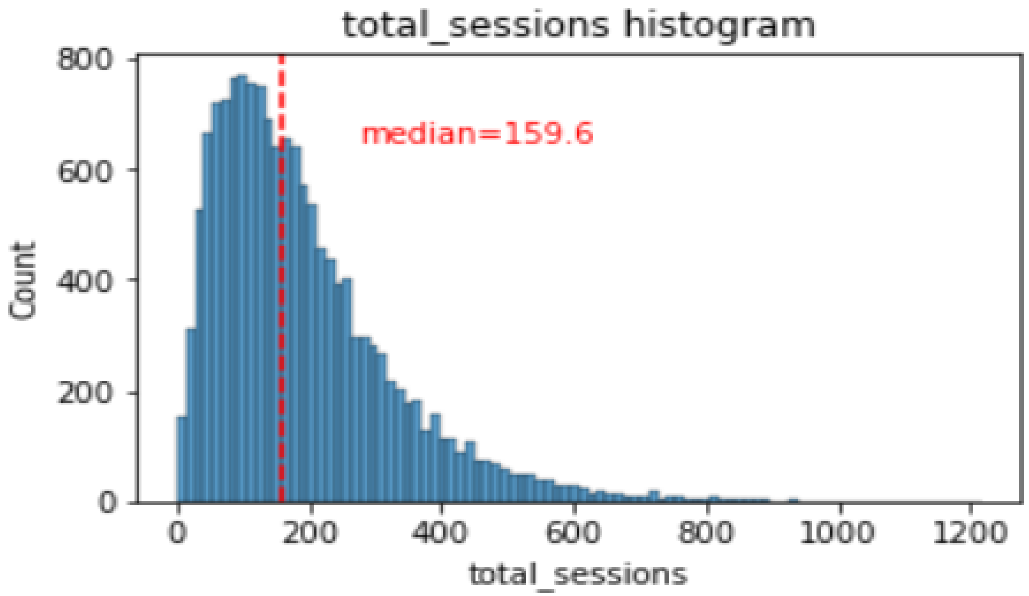



- The drives data exhibits a distribution resembling that of the 'sessions' variable.
- It is **right-skewed**, resembles a **log-normal distribution**, with a **median** of 48 **drives.**
- However, a **subset of drivers** recorded **over 400 drives in the last month.**

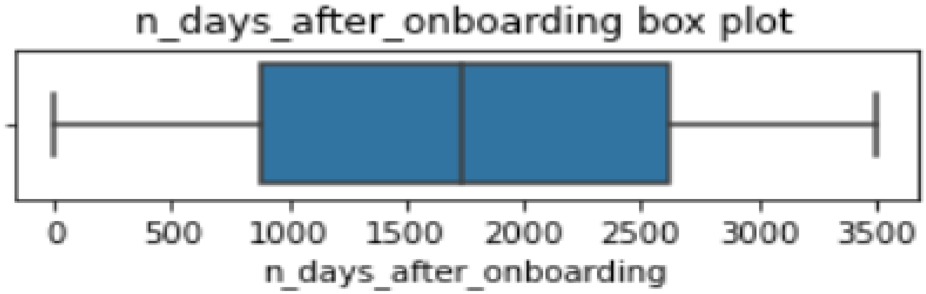
# TOTAL SESSIONS



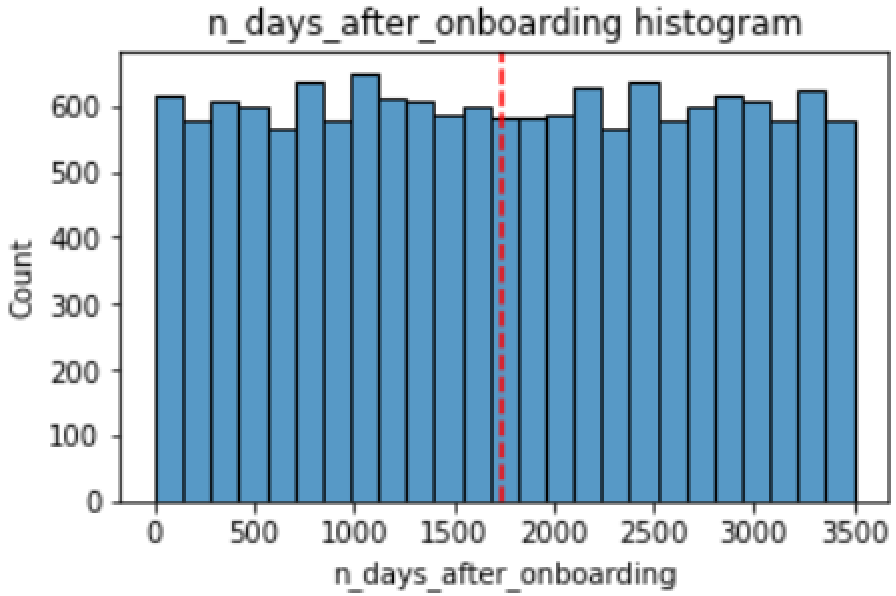




- The distribution of total_sessions is **right-skewed,** appearing closer to a normal distribution compared to the previous variables.
- The **median** total number of sessions is approximately **159.6.**
- If the median number of sessions in the last month was 48 and the median total sessions was around 160, it suggests that a **significant proportion of a user's overall sessions possibly occurred within the last month.**

# NUMBER OF DAYS AFTER ONBOARDING

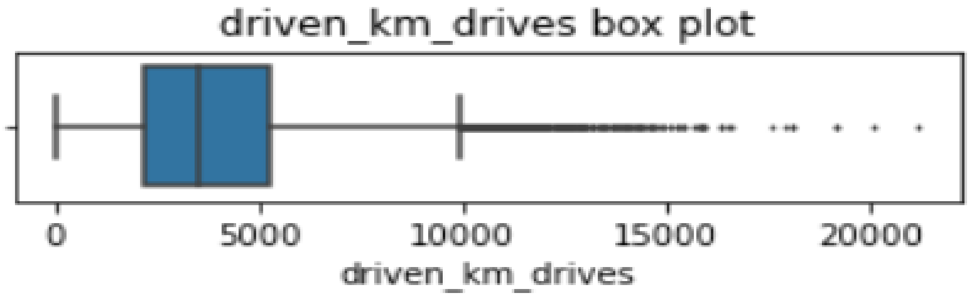


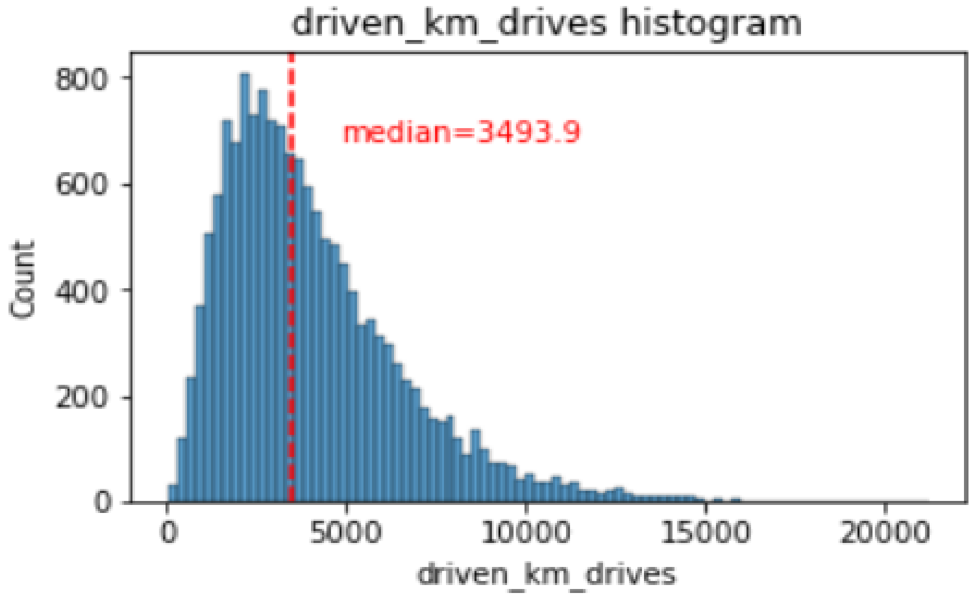

```
Median: 1741.0
```




- The total user tenure is a **uniform distribution** with values rangin from near-zero to ~3500 days, or roughly **9.5 years.**
- The **median** number of days since a user signed up for the app is 1741 days, or roughly **4.8 years.**

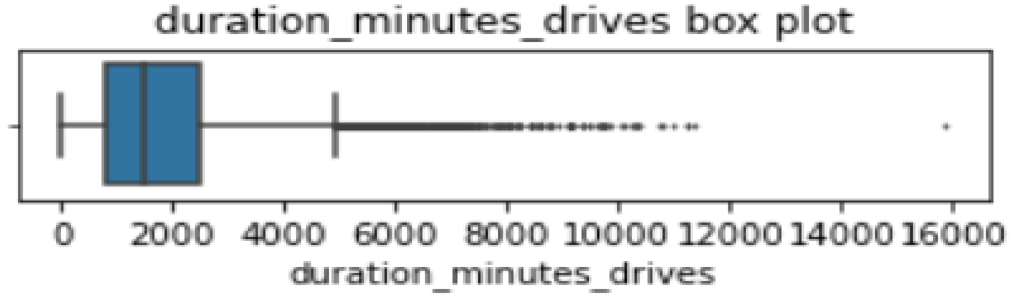
# TOTAL KM DRIVEN DURING THE MONTH

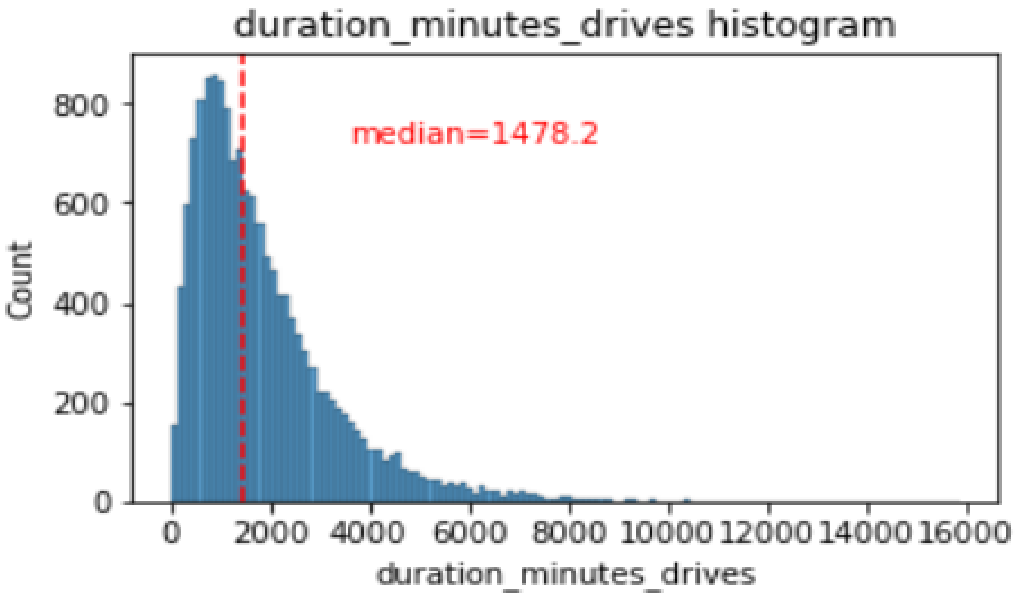



- The distribution of drives completed by each user in the last month exhibits **right-skewed normal distribution.**
- Roughly **50% of users drove fewer than 3,495 kilometers** during that period.
- The **median** number of total kilometers driven during the month 3494 **km.**

# TOTAL DURATION DRIVEN DURING THE MONTH

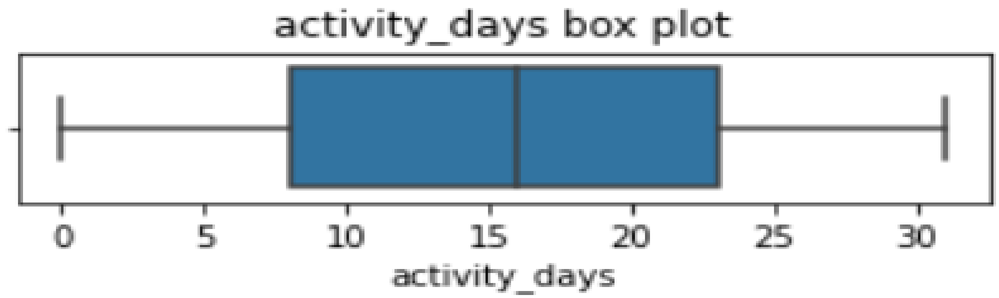


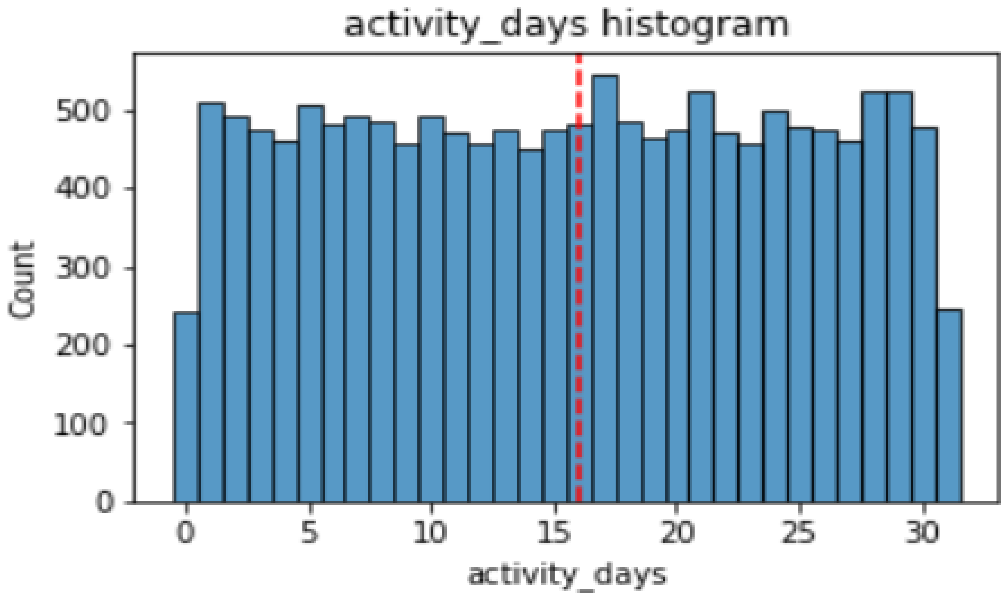

- The duration_minutes_drives variable has a **normalish distribution** with a heavily **skewed right tail.**

- Around **50%** of the users had a driving duration of **less than** the **median** of **1,478 minutes** (equivalent to about 25 hours), while **certain users recorded over 250 hours** of driving time throughout the month.

# ACTIVITY DAYS

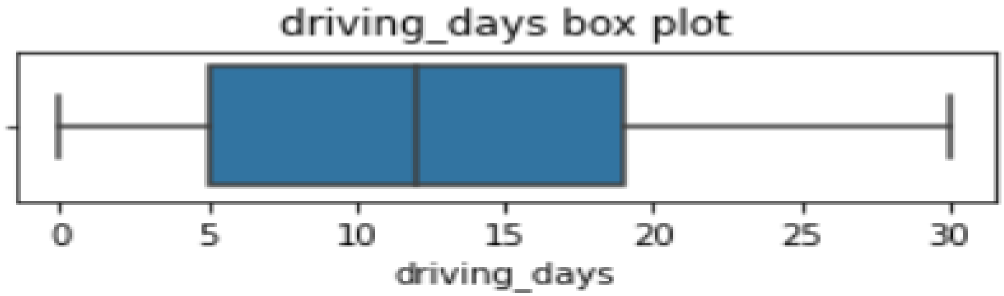



Median: 16.0

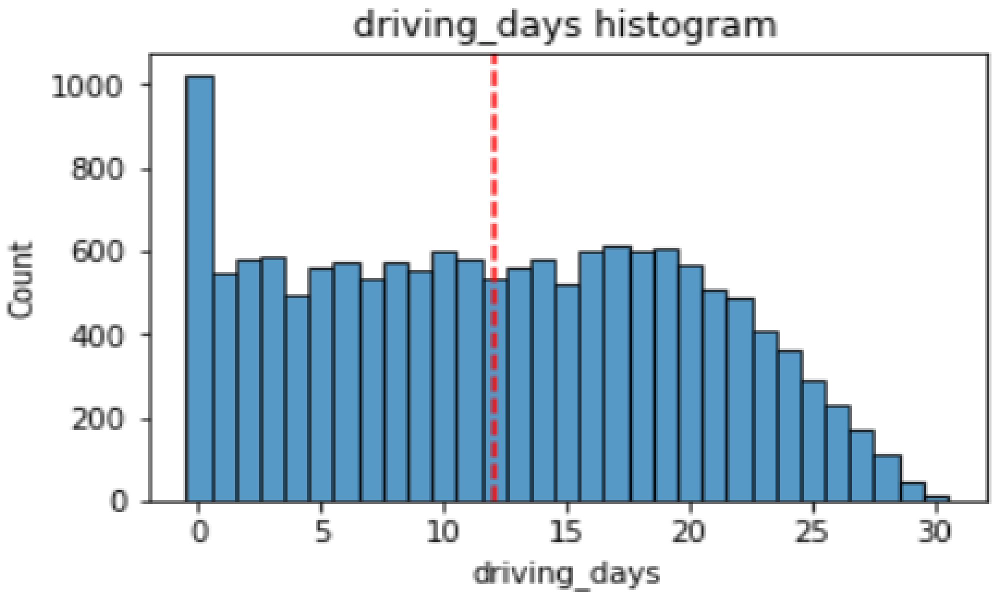



- In the past month, users had a **median of 16 app openings.**
- The box plot displays a **distribution that is centered**.
- The histogram indicates a **relatively uniform pattern** with approximately **500 individuals opening the app on each day.**
- However, there are approximately **250 users who did not open the app at all**, while **another 250 users opened it every day** throughout the month.

# DRIVING DAYS

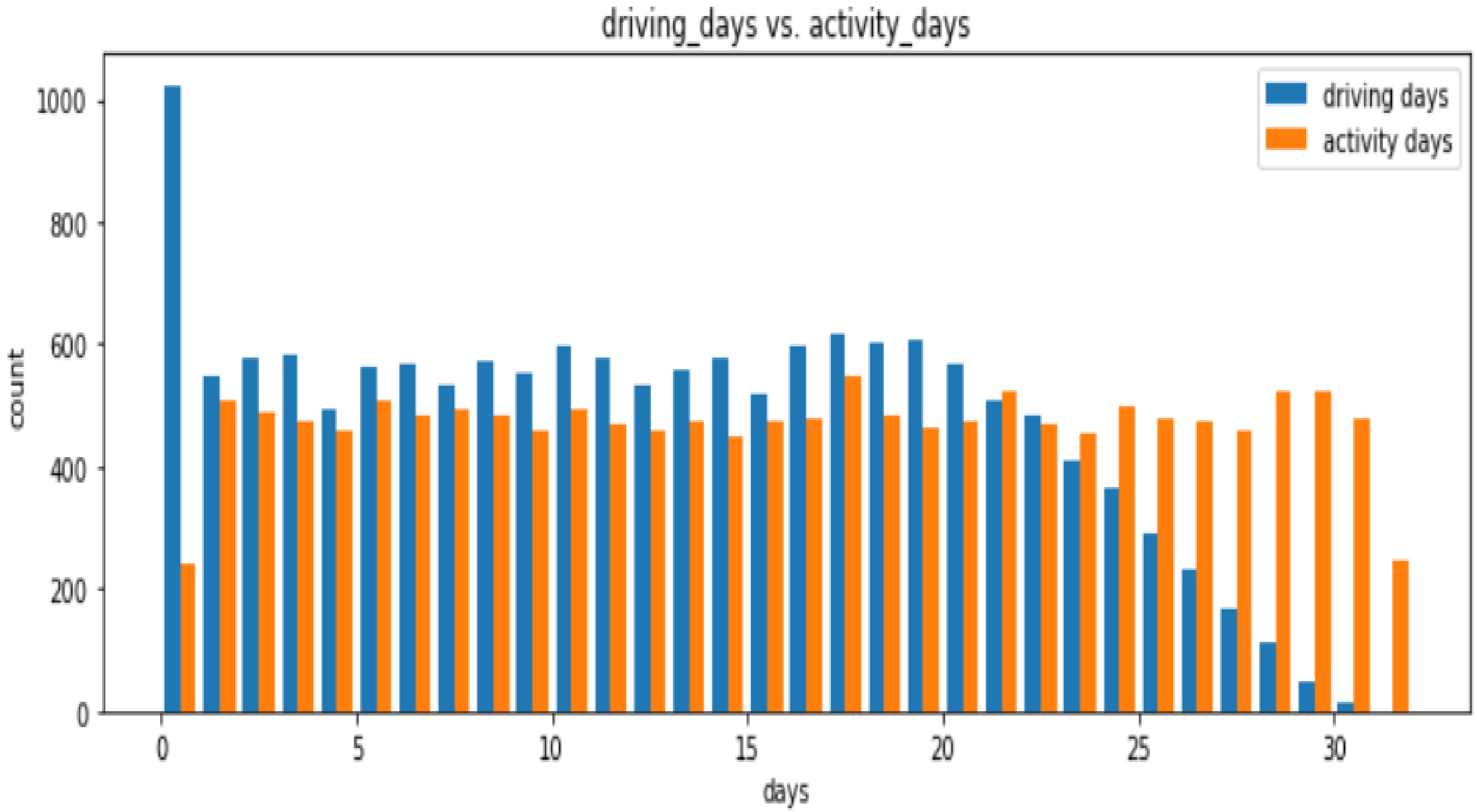



Median: 12.0



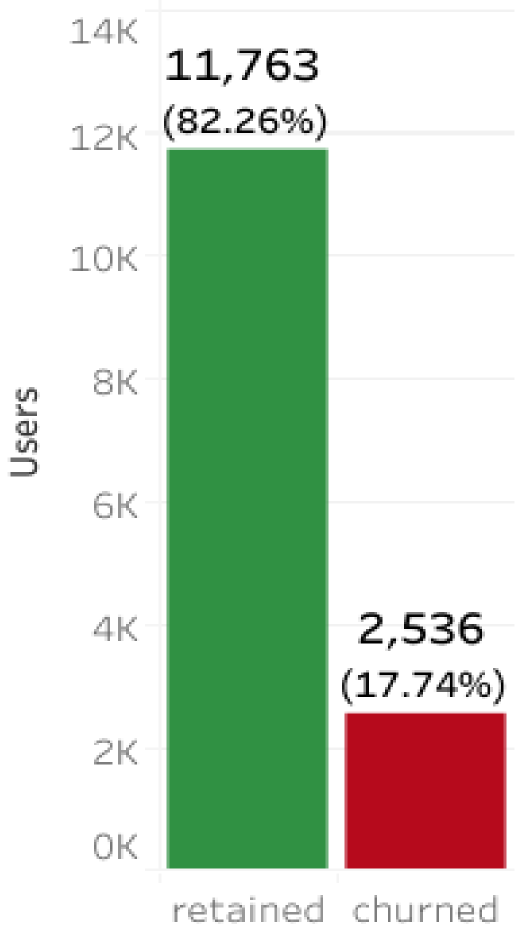

- The **median** number of days the users drove in the last month is **12 days**.
- The frequency of users driving each month shows a **relatively uniform pattern**, closely aligned with the number of days they accessed the app within the same period.
- The **distribution** of driving_days **skews towards lower values.**
- Interestingly, there were nearly **twice as many users** (~1,000 versus ~550) who **didn't engage in any driving** activity throughout the month..

# DRIVING DAYS VS. ACTIVITY DAYS



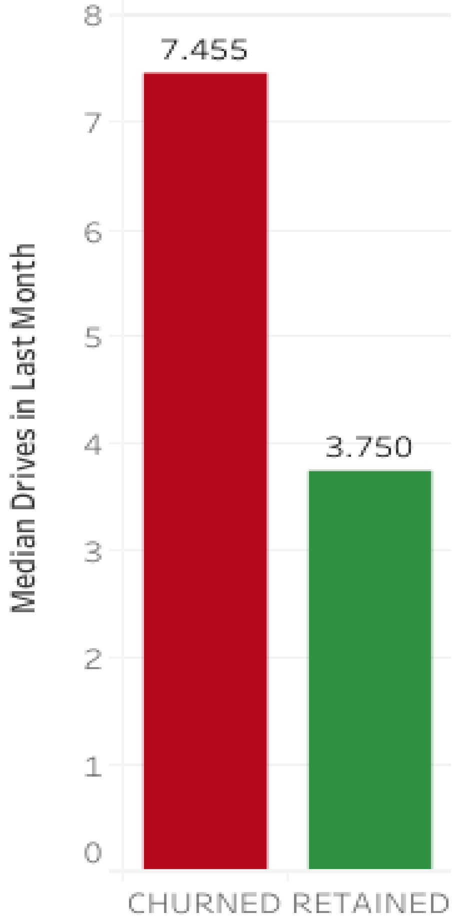

- Initially, more users had an increase in driving_days.
- The two variables stayed fairly consistent until around day 21.
- After day 21, driving_days steadily declined, while activity_days remained near its previous levels.
- This would suggest that though users weren't driving as much, they were still opening and using the app.

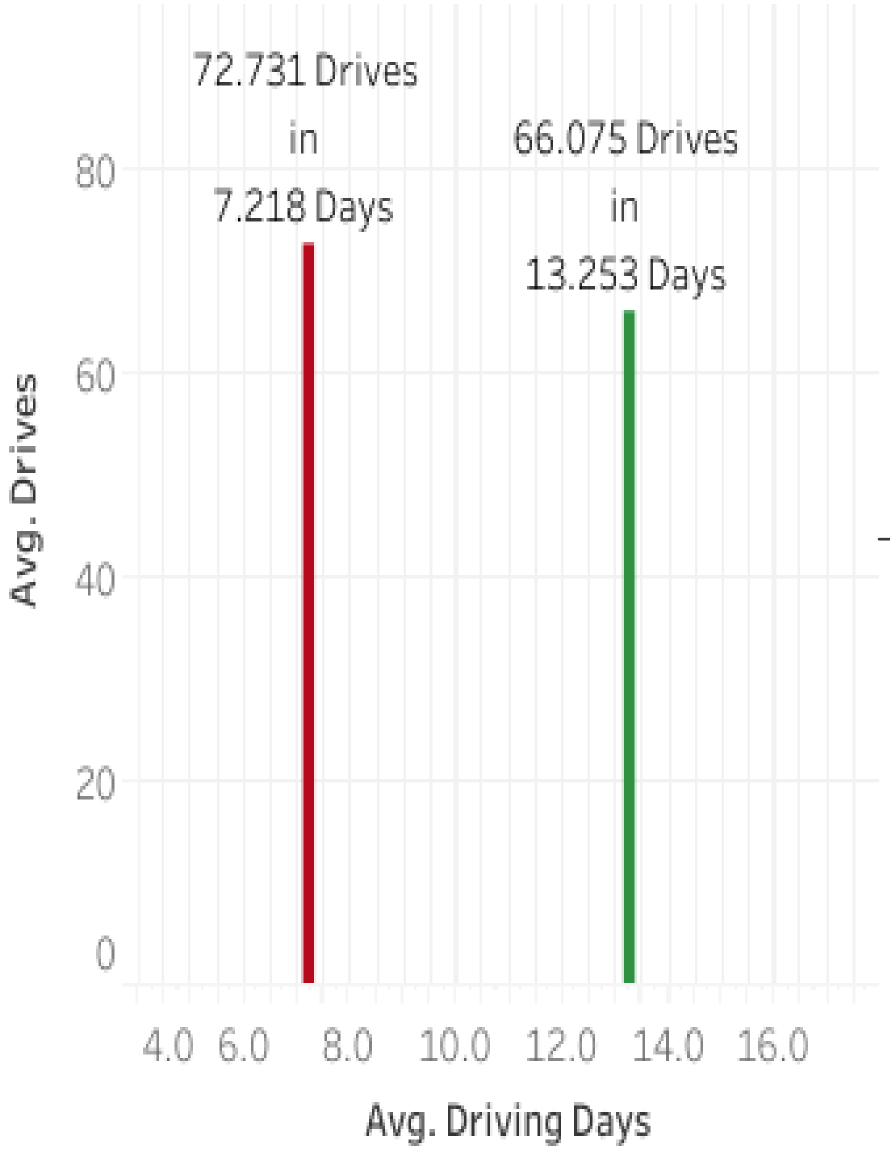
# CHURN VS. RETAINED USERS




This dataset contains **82% retained users** and **18% churned users**.

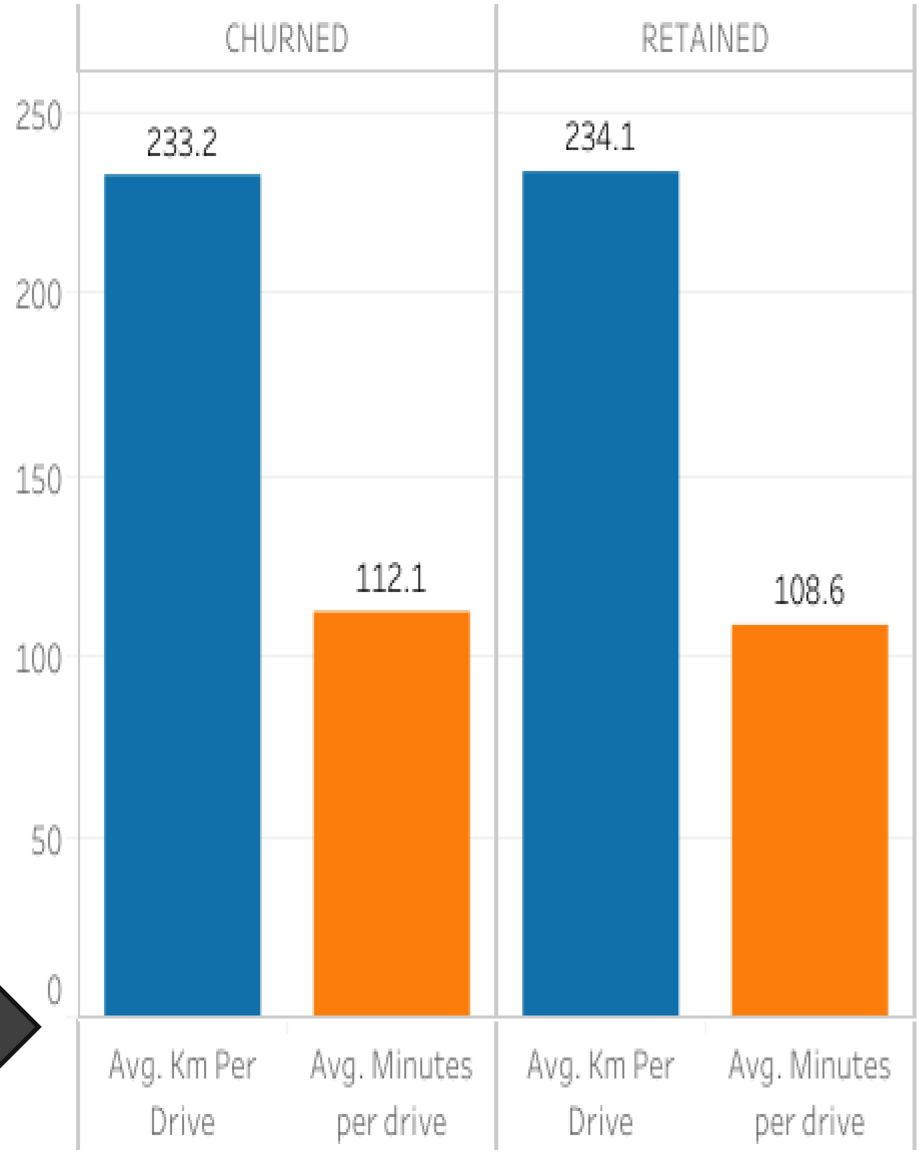
**Churned users averaged ~3 more drives** in the last month than retained users.



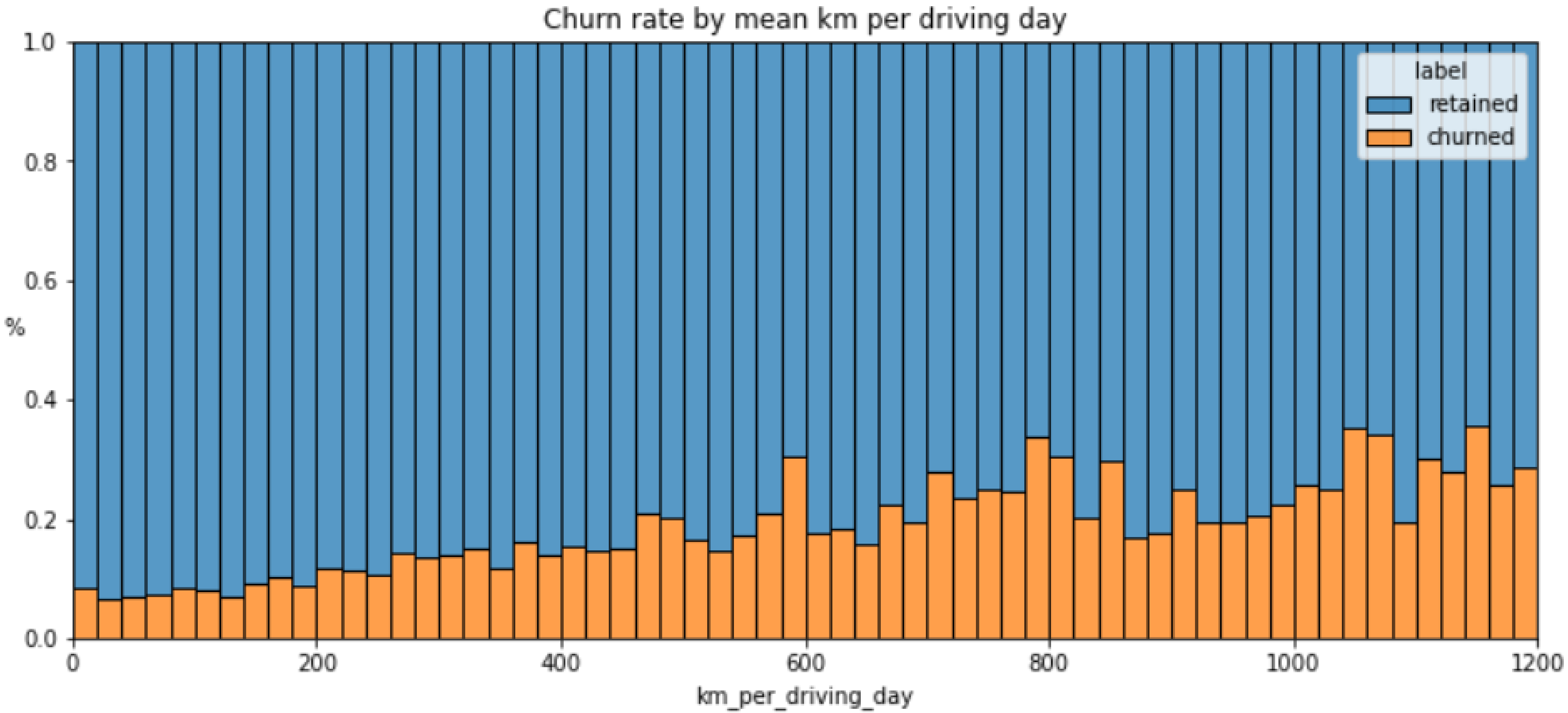



Churned users had **more drives** in **fewer days.**

Churned users trips were similar in length but slightly **longer in duration.**

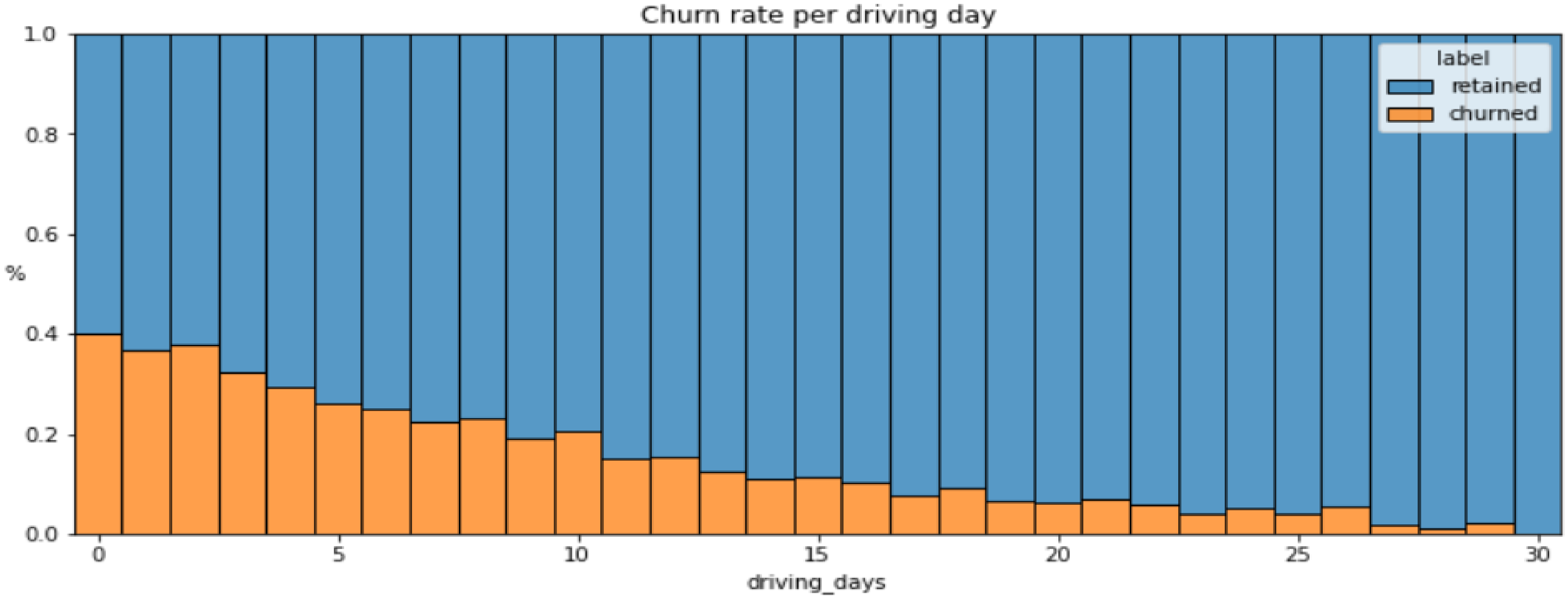


# RETENTION BY KM DRIVEN PER DRIVING DAY



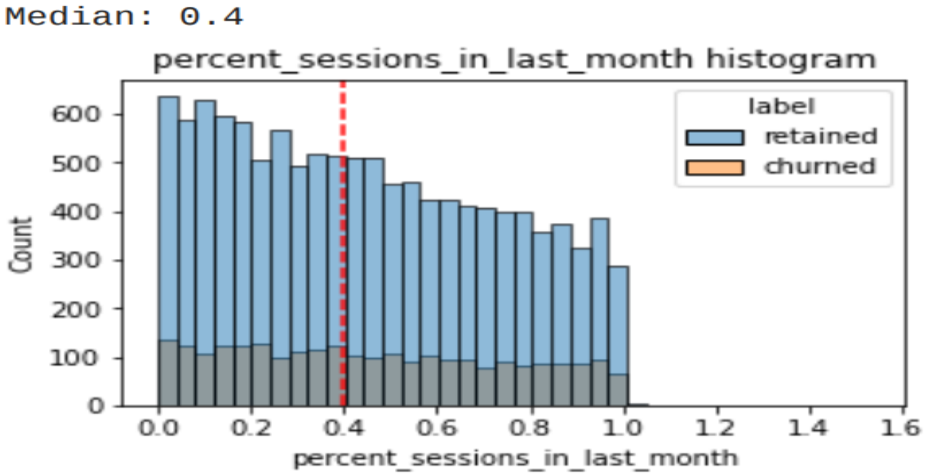

As the average daily distance driven increases, the churn rate also tends to rise.

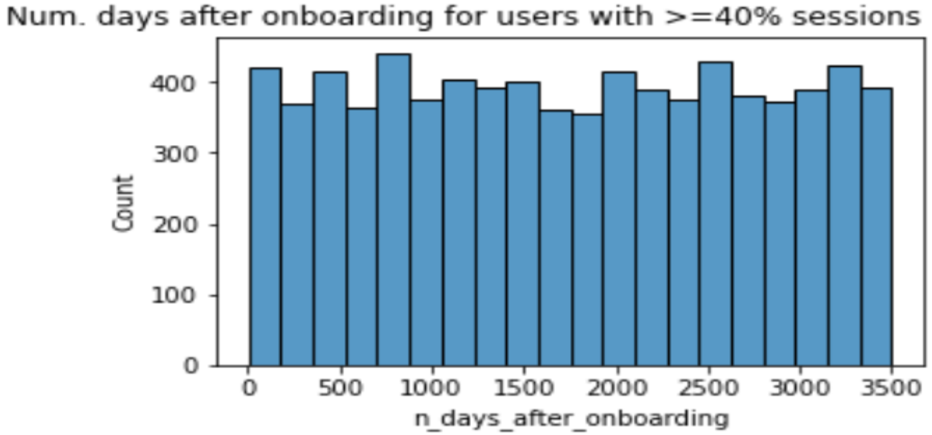
# CHURN RATE PER NUMBER OF DRIVING DAYS




The likelihood of **churn decreased as the frequency of app usage increased**. Among users who did not use the app at all in the last month, 40% churned, whereas **none of the users who used the app for 30 days experienced churn.**

# SESSIONS PROPORTIONS AND SURGE IN ACTIVITY FOR LONGSTANDING USERS



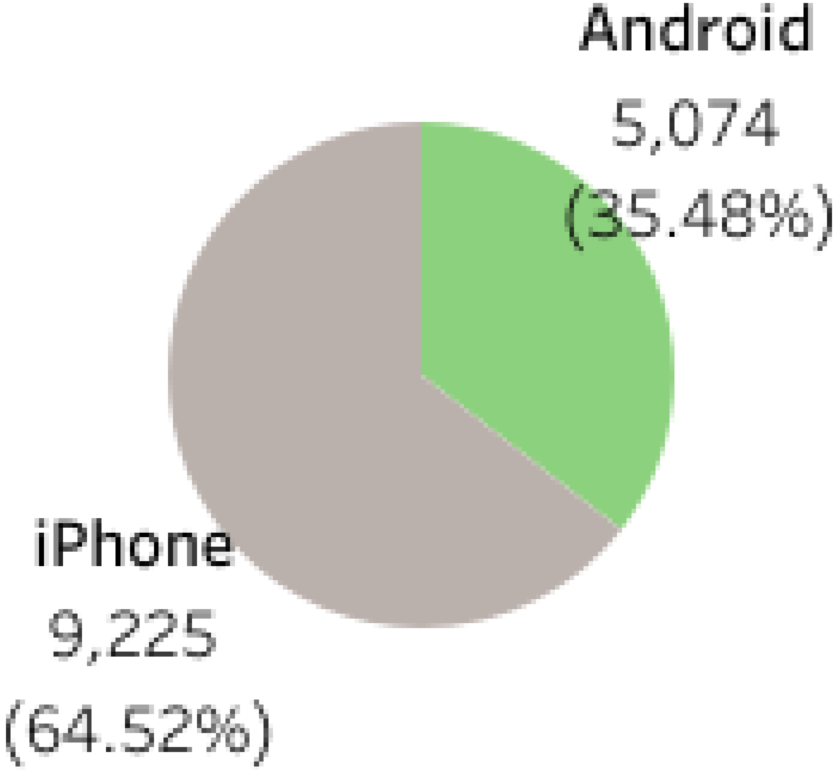




- Around **half of the users** included in the dataset had **40% or more of their sessions** concentrated solely **in the last month.**

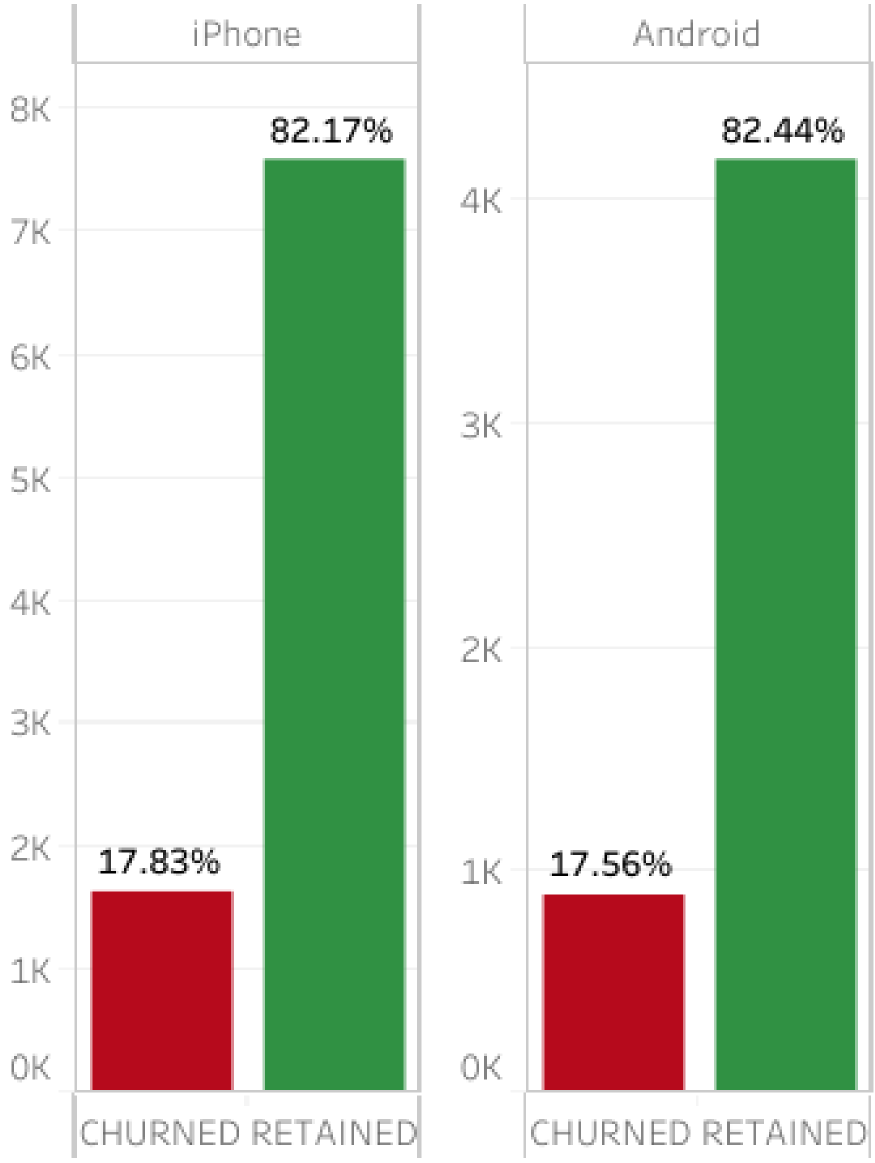
- The number of days since users onboarded, who have experienced 40% or more of their total sessions within the last month, conforms to a **uniform distribution**.

- **Why the sudden surge in app usage by these longstanding users during the recent month?**

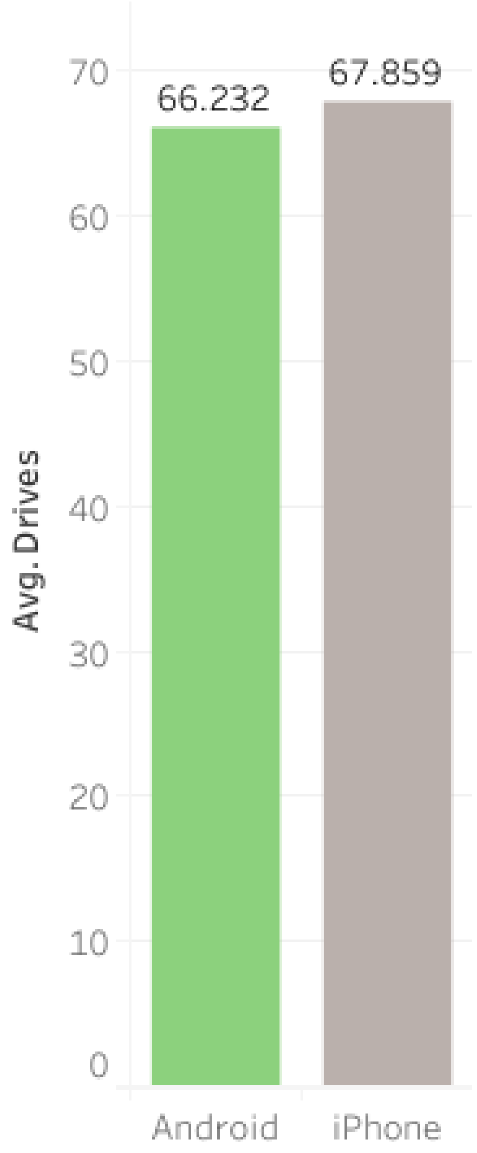
# DEVICES: ANDROID VS. IPHONE




- **iPhone devices** make up a **majority** of the users in this dataset.
- **Android devices** account for roughly **a third** of all users.




- The **proportion** of iPhone users to Android users remains **consistent** within both the churned and retained user groups.
- There is **no indication of any correlation** between device type and churn.




- Given the displayed averages, it seems that iPhone device users tend to have a higher average number of drives when using the application.
- However, it's important to consider that this disparity may be a result of random sampling rather than an actual difference in the number of drives.
- To determine if the distinction is statistically significant, I performed a hypothesis test.

# DEVICE HYPOTHESIS TESTING

Hypotheses:

- $H0$ : There is no difference in average number of drives between drivers who use iPhone devices and drivers who use Androids.

- $HA$ : There is a difference in average number of drives between drivers who use iPhone devices and drivers who use Androids.

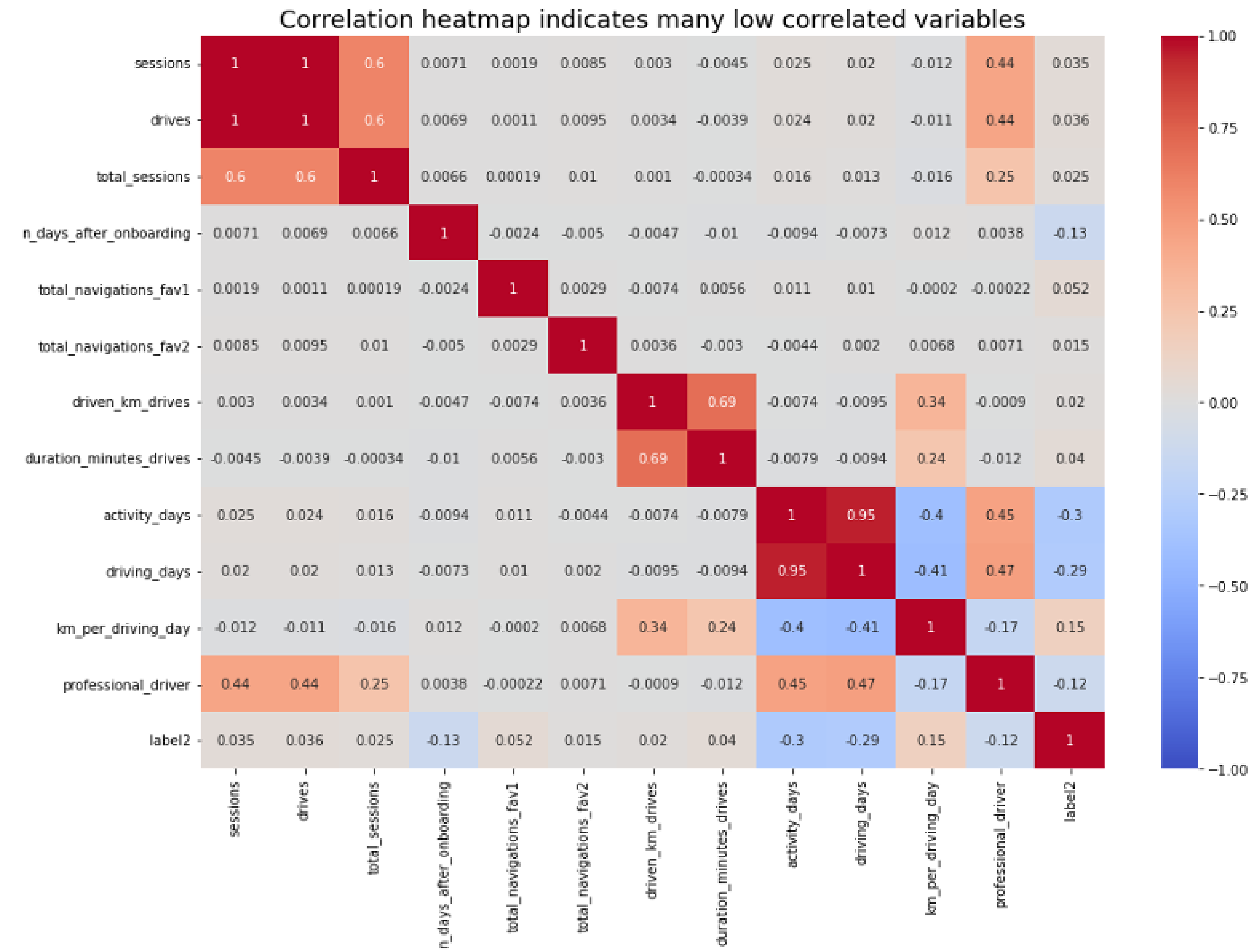
## Two-sample test with 5% as the significance level with a two-sample t-test.

```
# 1. Isolate the `drives` column for iPhone users.
iPhone = df[df['device_type'] == 1]['drives']

# 2. Isolate the `drives` column for Android users.
Android = df[df['device_type'] == 2]['drives']

# 3. Perform the t-test
stats.ttest_ind(a=iPhone, b=Android, equal_var=False)
```

```
Ttest_indResult(statistic=1.4635232068852353, pvalue=0.1433519726802059)
```

## p Value = 0.143...

As the p-value exceeds the selected significance level of 5%, we fail to reject the null hypothesis. This indicates that there is **no statistically significant distinction in the average number of drives between iPhone users and Android users.**




# Collinearity

As title suggests, the correlation heatmap indicates many low correlated variables.

Variables that are multicollinear with each other:

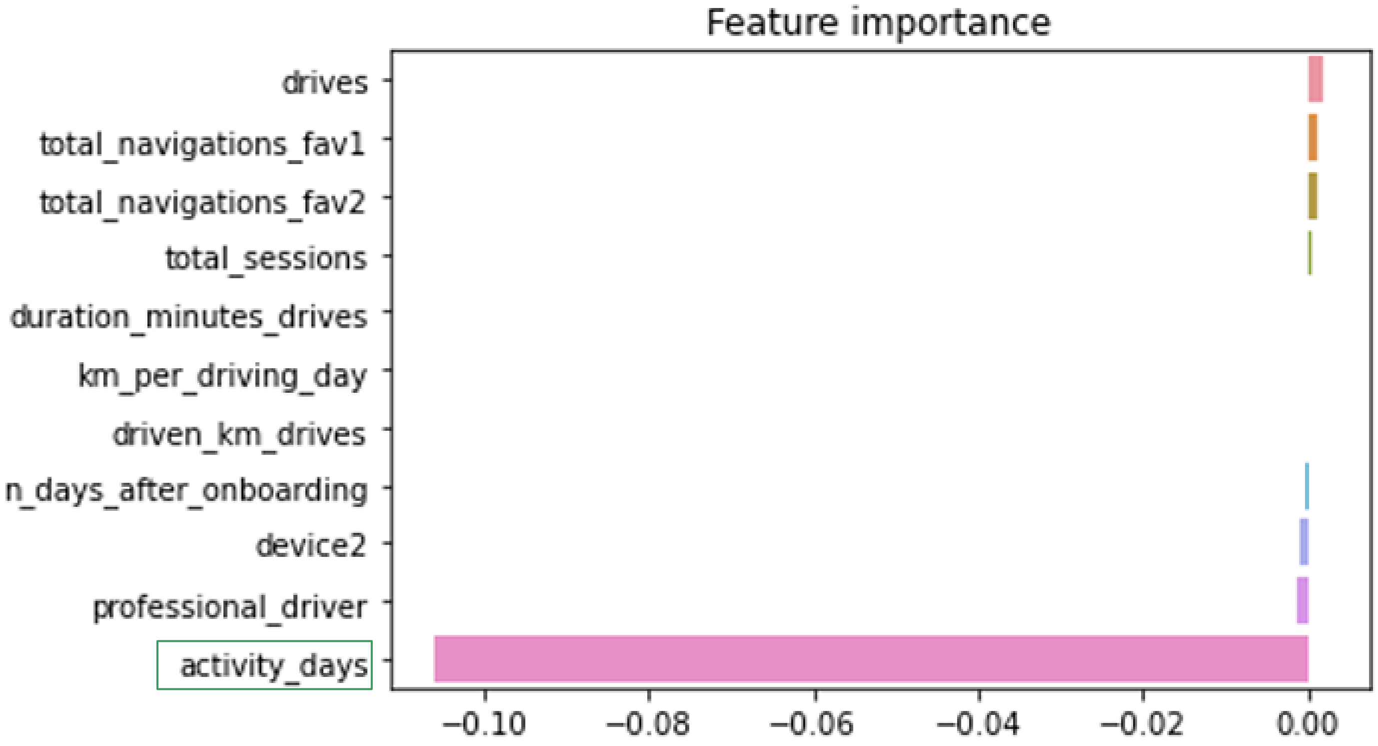
- sessions and drives: 1.0
- driving_days and activity_days: 0.95

# LOGISTIC REGRESSION MODEL

```
drives                       0.001913
total_sessions               0.000327
n_days_after_onboarding     -0.000406
total_navigations_fav1       0.001232
total_navigations_fav2       0.000931
driven_km_drives            -0.000015
duration_minutes_drives      0.000109
activity_days               -0.106032
km_per_driving_day           0.000018
professional_driver         -0.001529
device2                     -0.001041
dtype: float64
```

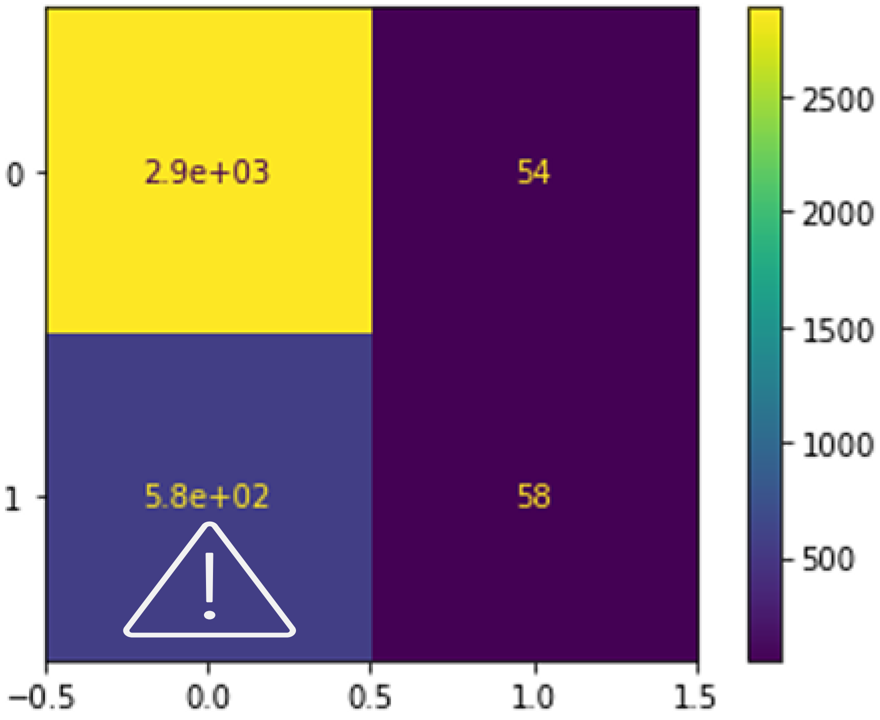
```
model.intercept_
```

```
array([-0.00170675])
```




Among all the features in the model, "activity_days" emerged as the most significant one, exhibiting a negative correlation with user churn.

# LOGISTIC REGRESSION MODEL




| | precision | recall | f1-score |
|---|---|---|---|
| retained | 0.83 | 0.98 | 0.90 |
| churned | 0.52 | $\boxed{0.09}$ | 0.16 |
| accuracy | | | 0.82 |
| macro avg | 0.68 | 0.54 | 0.53 |
| weighted avg | 0.78 | 0.82 | 0.77 |

Although the model demonstrates reasonable precision, its recall is extremely low, indicating a **high number of false negative predictions.**

Consequently, **it fails to identify and capture users who are likely to churn.**

# LOGISTIC REGRESSION MODEL INSIGHTS

- **“Activity_days" emerged as the most significant feature**, exhibiting a negative correlation with user churn.
  - This finding is not unexpected since "activity_days" is highly correlated with "driving_days," which was already identified to have a negative correlation with churn.
- During EDA, the user churn rate rose in conjunction with increasing values in **"km_per_driving_day."**
  - The correlation heatmap confirmed this observation, indicating that this variable exhibited the highest positive correlation with churn among all the predictor variables.
  - Surprisingly, **in the model, "km_per_driving_day" ranked as the second-least important variable.**

# LOGISTIC REGRESSION MODEL IMPROVEMENTS

- By leveraging domain knowledge, it is possible to engineer new features aimed at improving predictive signal.
  - In the context of this model, one of the engineered features, namely "professional_driver," emerged as the third-most influential predictor.
  - Scaling the predictor variables and reconstructing the model using different combinations of predictors can be beneficial in minimizing noise stemming from unpromising features.

- Possessing drive-level specifics for individual users, such as drive times and geographic locations would be beneficial.

- Obtaining more detailed information regarding how users engage with the app would likely provide valuable insights.

- Having knowledge of the monthly count of distinct starting and ending locations inputted by each driver could offer valuable additional information.

# LOGISTIC REGRESSION MODEL RECOMMENDATION

The usefulness of the model depends on its intended purpose.

- If the model is employed to inform critical business decisions, its performance may not be sufficiently strong, particularly evident from its low recall score.

- If the model is primarily utilized to guide further exploratory efforts and provide insights, it can still offer value in that context.

# MACHINE LEARNING MODEL
## RANDOMFOREST VS. XGBOOST

| | model | precision | recall | F1 | accuracy |
|---|---|---|---|---|---|
| 0 | RF cv | 0.458198 | 0.126782 | 0.198534 | 0.818626 |
| 0 | XGB cv | 0.442586 | 0.173468 | 0.248972 | 0.814780 |

- The XGBoost model not only outperformed the random forest model in terms of data fitting, but it also achieved a recall score that is nearly twice as high as the recall score obtained by the logistic regression model.

- It also demonstrates an improvement of almost 50% in recall compared to the random forest model, while maintaining similar levels of accuracy and precision.

# MACHINE LEARNING MODEL

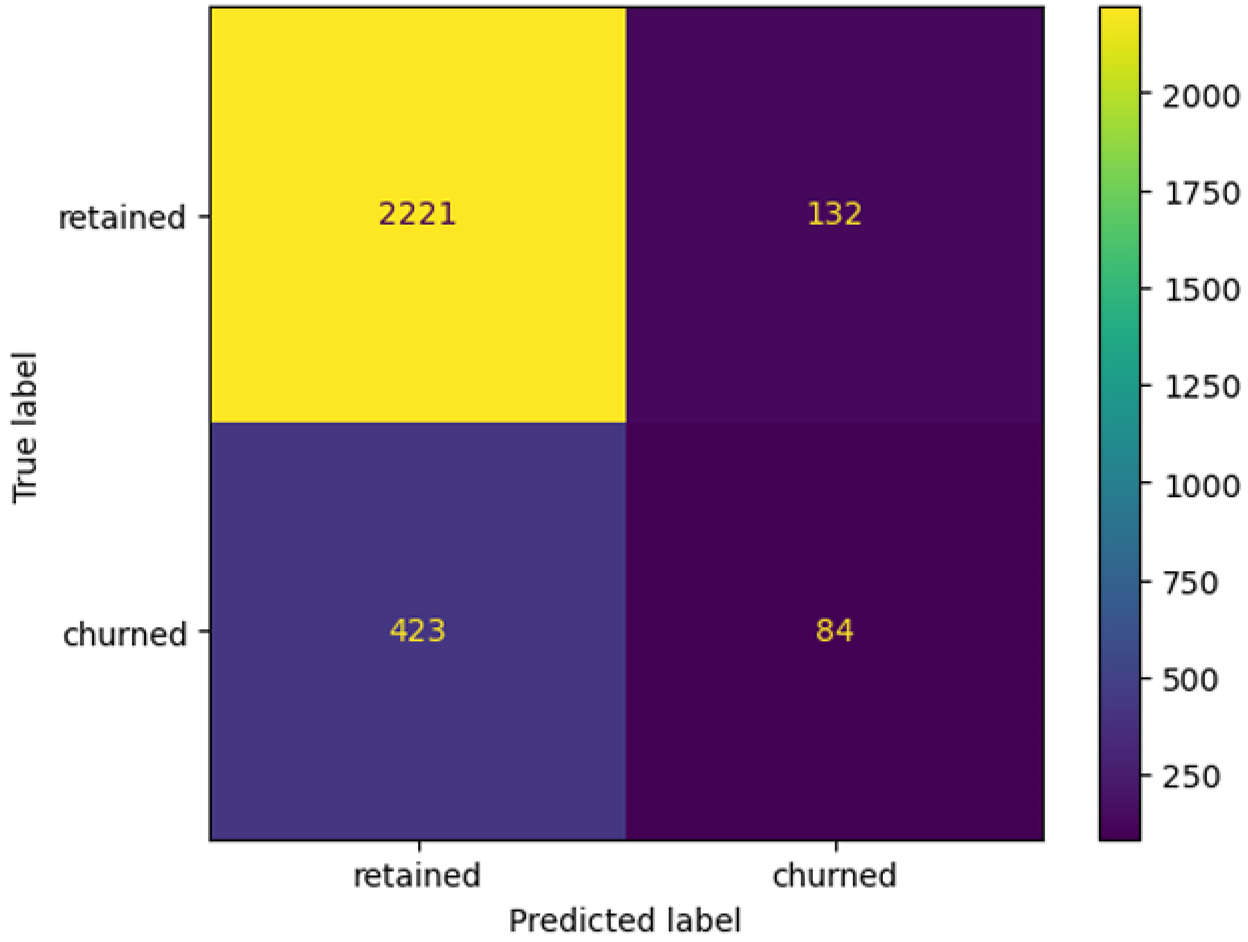
## VALIDATION AND TEST

| | model | precision | recall | F1 | accuracy |
|---|---|---|---|---|---|
| **0** | RF cv | 0.458198 | 0.126782 | 0.198534 | 0.818626 |
| **0** | XGB cv | 0.442586 | 0.173468 | 0.248972 | 0.814780 |
| **0** | RF val | 0.445255 | 0.120316 | 0.189441 | 0.817483 |
| **0** | XGB val | 0.430769 | 0.165680 | 0.239316 | 0.813287 |
| **0** | XGB test | 0.388889 | 0.165680 | 0.232365 | 0.805944 |

- The recall remained unchanged from the validation data, while the precision experienced a significant decline, resulting in a slight drop in all other scores.
- Nevertheless, these variations fall within an acceptable range for performance disparities between validation and test scores.

# MACHINE LEARNING MODEL

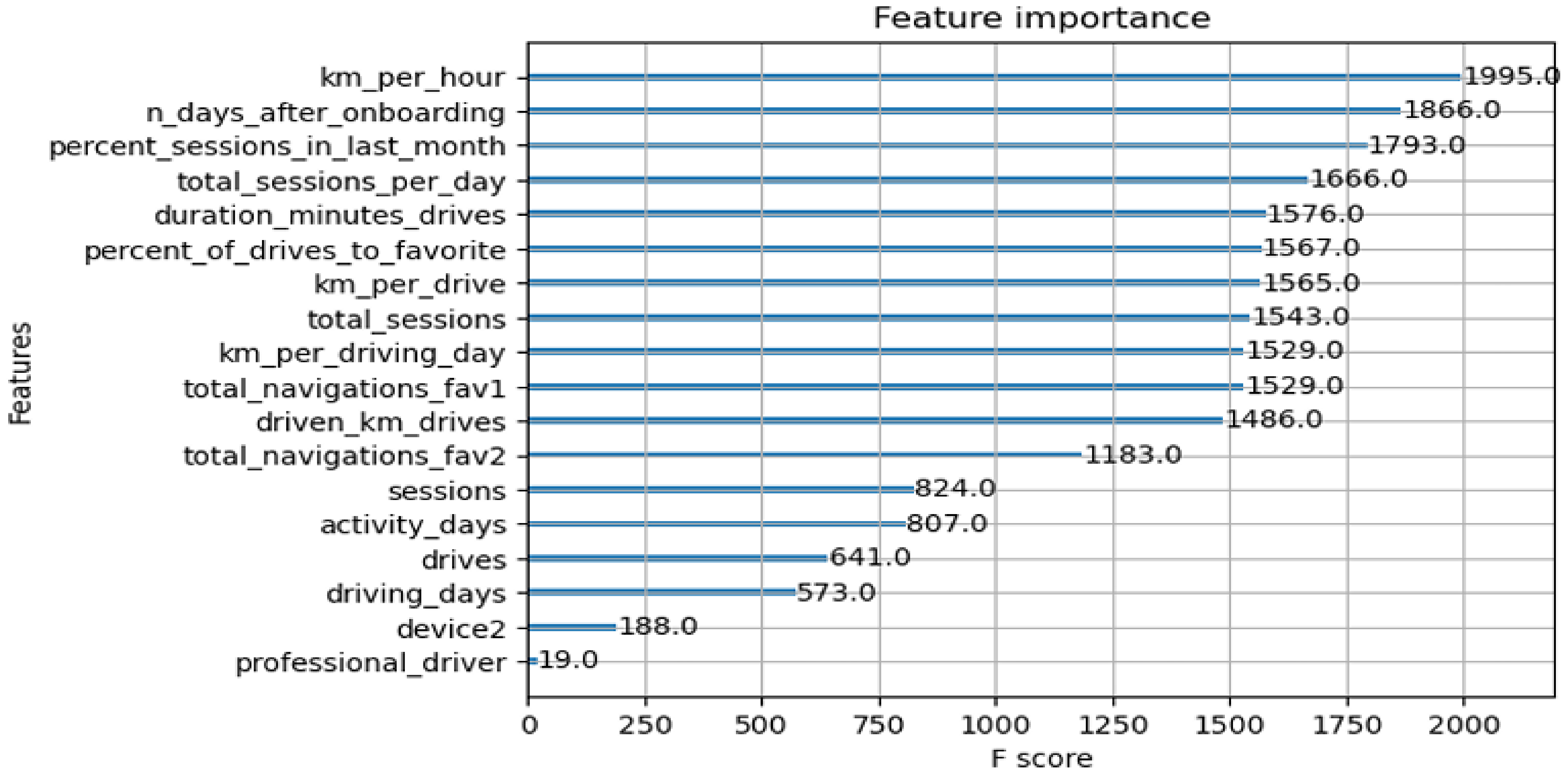
## VALIDATION AND TEST




- The model's false negatives outnumbered false positives by a factor of three.
- It accurately identified only 16.6% of the users who churned.

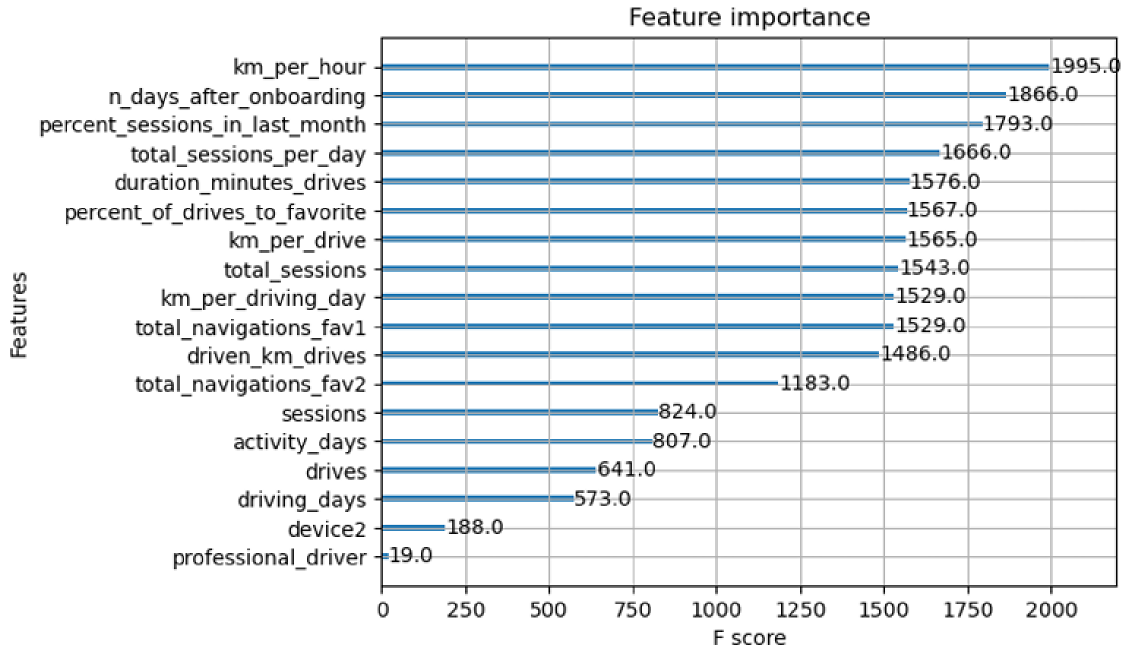
# MACHINE LEARNING MODEL

## FEATURE IMPORTANCE



**Top Five Most Important Features That Impact Churn:**

1. km_per_hour
2. n_days_after_onboarding
3. percent_sessions_in_last_month
4. total_sessions_per_day
5. duration_minutes_drives




- The XGBoost model utilized a greater number of features compared to the logistic regression model.
- Engineered features comprised six out of the top 10 features, including three out of the top five.
- It is worth noting that the selection of important features can vary between different models due to the complexity involved in feature selection.

# MACHINE LEARNING MODEL
## IMPROVEMENTS THAT CAN BE MADE

- Introducing new features could enhance the model's predictive capabilities, particularly with better domain knowledge.

- In the case of this model, engineered features accounted for over half of the top 10 most-predictive features employed by the model.

- Reconstructing the model using different combinations of predictor variables can help reduce noise originating from non-predictive features.

# MACHINE LEARNING MODEL
## ADDITIONAL FEATURES THAT COULD HELP IMPROVE THE MODEL

- Having drive-level information for each user, such as drive times and geographic locations, would be beneficial.
- More detailed data providing insights into user interactions with the app would be valuable.
- Knowing the monthly count of unique starting and ending locations provided by each driver could offer further assistance.

# FINAL RECOMMENDATION

- If the model is to be utilized for significant business decisions, then it falls short in being an ideal predictor, as evidenced by its low recall score.

- If the model is solely employed to guide exploratory efforts, it can provide value.

- The model could be more predictive if we gather more drive level data as mentioned previously, as well as exploring different engineered features.