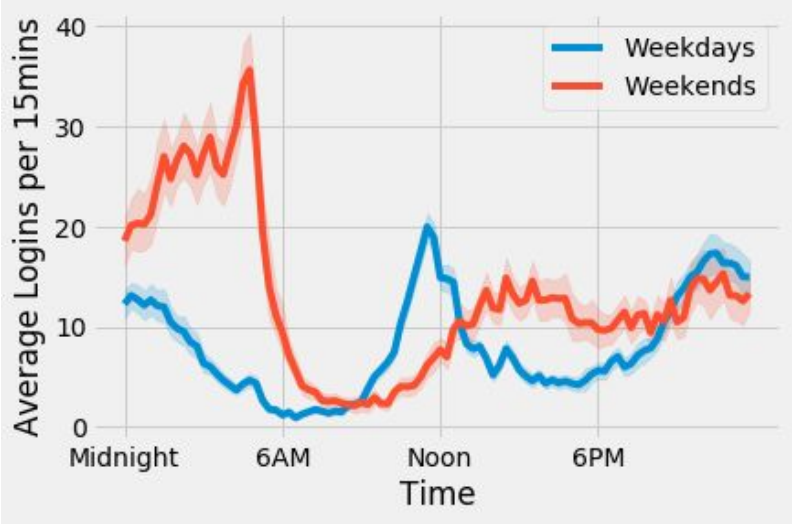
Part 1 Exploratory data analysis

There were 93,142 total logins from 103 days in the dataset. Daily logins on weekends were significantly higher than those on weekdays (t-test, $p<0.001$).

Weekdays and weekends also have different daily patterns. On weekdays, the numbers of logins were highest around 11AM and 10PM. On weekends, the numbers of logins were highest from midnight to 5AM.



There were no quality issue with the dataset, other than that the dates were in 1970. I assume that the data were modified to this form for confidentiality.

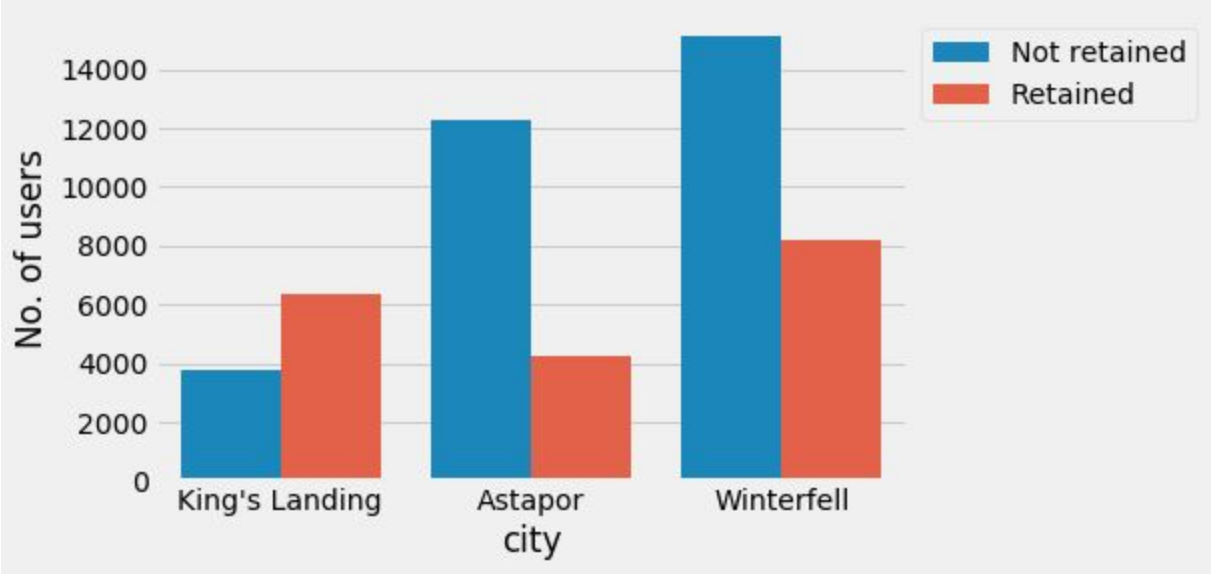
Part 2 Experiment and metrics design

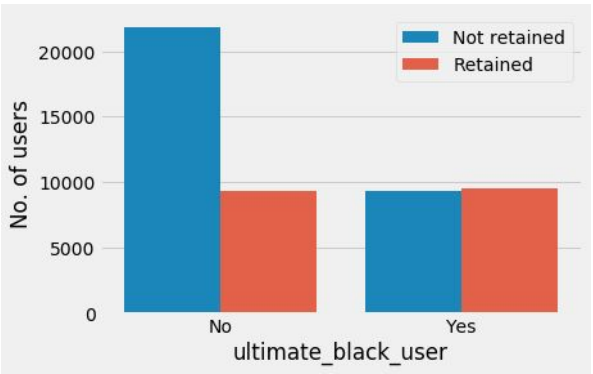
1. The metric of success would depend on company goals at this moment. Reimbursing toll costs might encourage drives to take on more inter-city rides, but might incur a financial cost to the company. If larger market share is company's main concern, I would choose the total number of rides in test period as the metric. On the other hand, if profit is the main concern, then I would choose the profit in the test period as the metric. If both factors are important, I would choose a metric that reflects both, such as the number of rides gained per dollar lost.

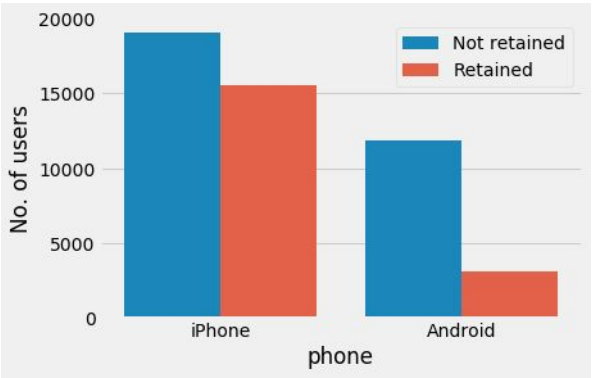
2. Randomly select a group of (at least a few hundred) drivers, send them the offer of toll reimbursement. Conduct the experiment for a few weeks. At the end of the experiment, collect data on the number of trips they have operated, the start and end of their trips, company revenue and cost on their trips. Conduct A/B testing between these drivers and those who did not receive the offer to answer the following questions:
(1) Did drivers who received toll reimbursement operate more trips? How much? (z-test)
(2) Did drivers who received toll reimbursement have a higher share of cross-city trips? (chi-square test)
(3) Did drivers who received toll reimbursement bring more or less profit? How much? (z-test)
(4) How many more trips were gained at the cost of 1 dollar? (z-test)
(5) At what time of the day did the cross-city trips occur most frequently?
For each of these questions, answer them separately for weekdays and weekends.
Based on the results, I can inform operations team the potential of market share gains and associated costs of the toll reimbursement plan if they were extend the plan to all drivers. I can also report at what days and times the plan is most effective. In terms of caveats, I will examine whether there were any significant events during the test period that might have affected the test outcome.

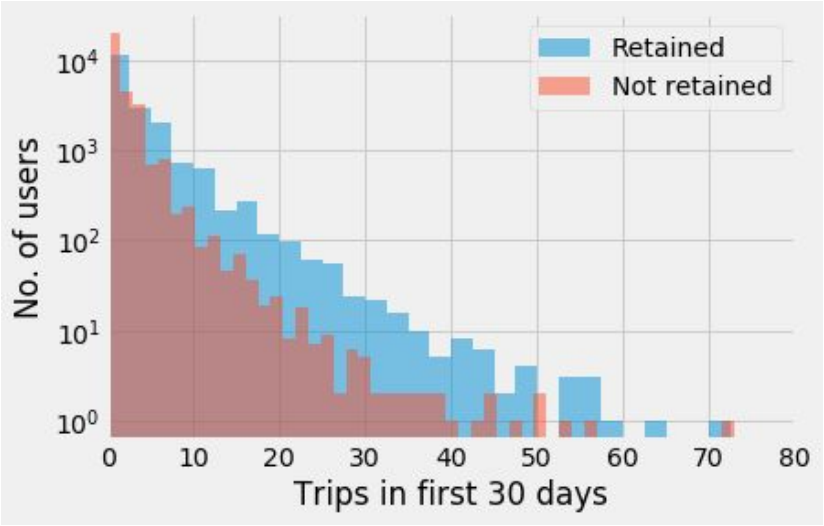
Part 3 Predictive Modeling

User retention data were inspected and out of 50000 users, 18804 were still active in their 6th month. Through exploratory data analysis, I found that users in King's Landing were more likely to become retained than those in other cities; iphone users more likely than android users, Ultimate Black users more likely than those who are not. In addition, higher number of trips in the first 30 days were associated with higher retention rate in users.









I performed data cleaning and dealt with missing data. I built a logistic regression model that achieved ~72% accuracy in predicting whether a user is retained or not. The four features mentioned above had the largest coefficients in the model, which meant they were the most important factors deciding user's outcome.

Based on the results, I have 4 suggestions for the business team:
(1) Look further into the reasons why retention rate is higher in King's Landing and see if there was any successful strategy the local team was carrying out, and if they can be applied to other cities.
(2) Promote the use of Ultimate Black through marketing, especially try to encourage users who have never used Ultimate Black to try it out.
(3) Further investigate why Android users have lower retention rate. Maybe the Android app is not as well designed, maybe iPhone users have more economic resources, maybe marketing on the Android platform is not working as well.
(4) Promote the use of Ultimate in a user's first 30 days through various marketing strategies.

Last but not least, there may be caveats in this analysis in that the data was from users who signed in one month and three cities. Consumer trends may change over time, and may vary from city to city. Further study and monitoring is required to update the results.