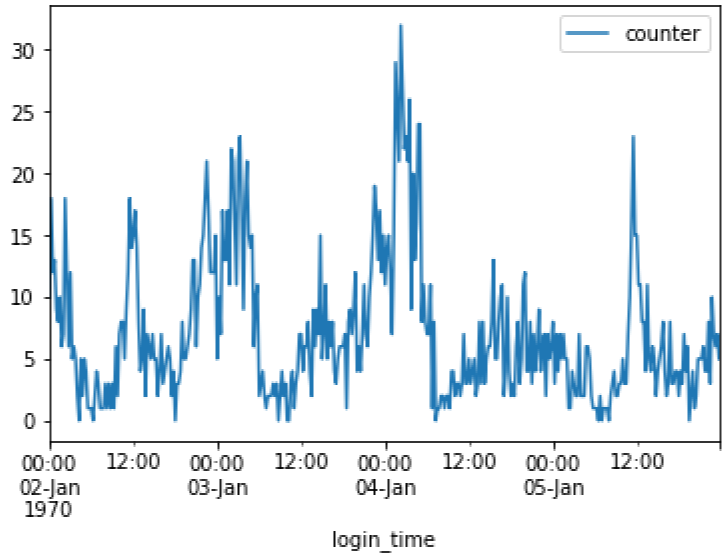
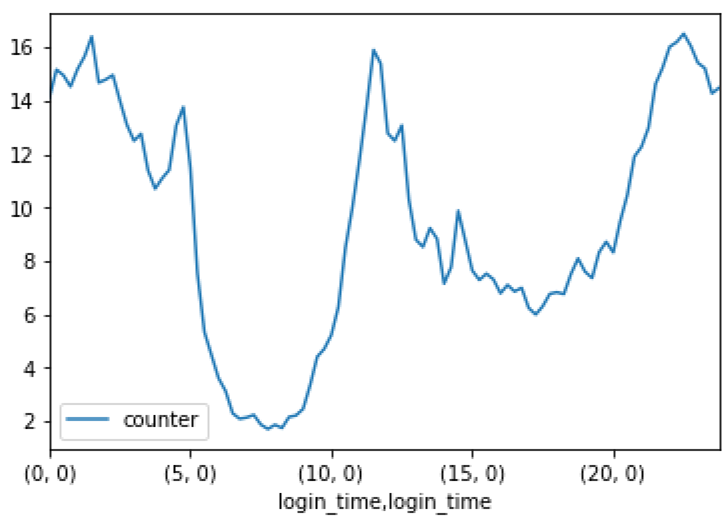
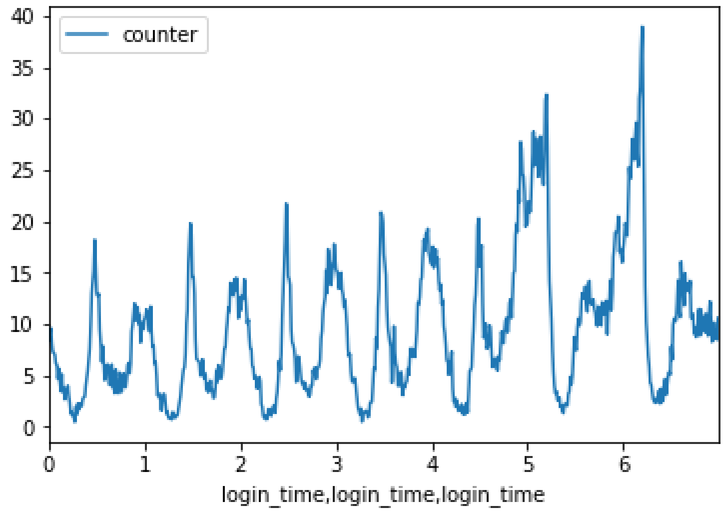
Part 1 - Exploratory Data Analysis

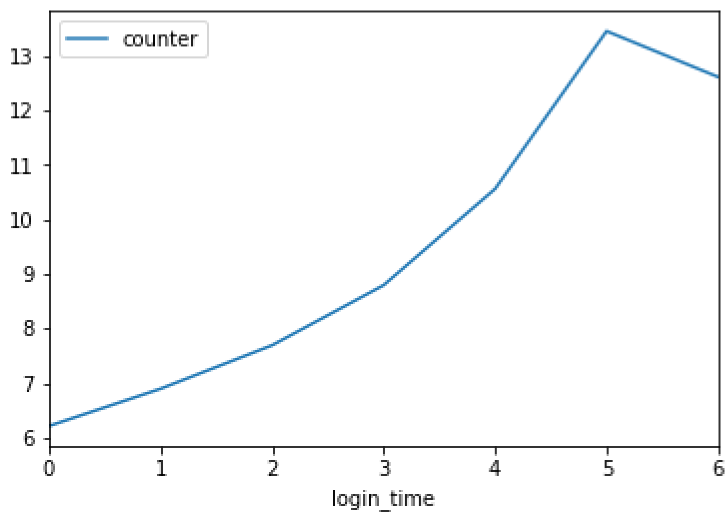


This is a sample of the first four full days of data. On three of these days there are two distinct peaks; one with a maximum near noon and one with a maximum near midnight. To see if this trend continues through the dataset, I will average all of the days grouped by hour and minute.



There is a consistent daily cycle with two distinct peaks. One of these is fairly narrow and the other is much broader. There may also be a weekly cycle in demand so I investigated that as well.





On Monday through Friday (0 - 4), the morning peaks all look fairly similar. Saturday and Sunday have shorter and broader peaks in the morning than the weekdays. The evening peak grows continuously from Monday to Saturday. Sunday has no evening peak.

Part 2 - Experiment and Metrics Design

The key metric I would use for this experiment is the mean number of crossings per week per hour worked for each driver. This is better than a daily average because weekdays and weekends have different activity patterns. On a weekday, I would not expect multiple crossings because both cities are not both busy at the same time, but only one when the demand changes from one city to the other. Since both cities are busy throughout the weekend, this is when drivers would be more likely to switch cities multiple times.

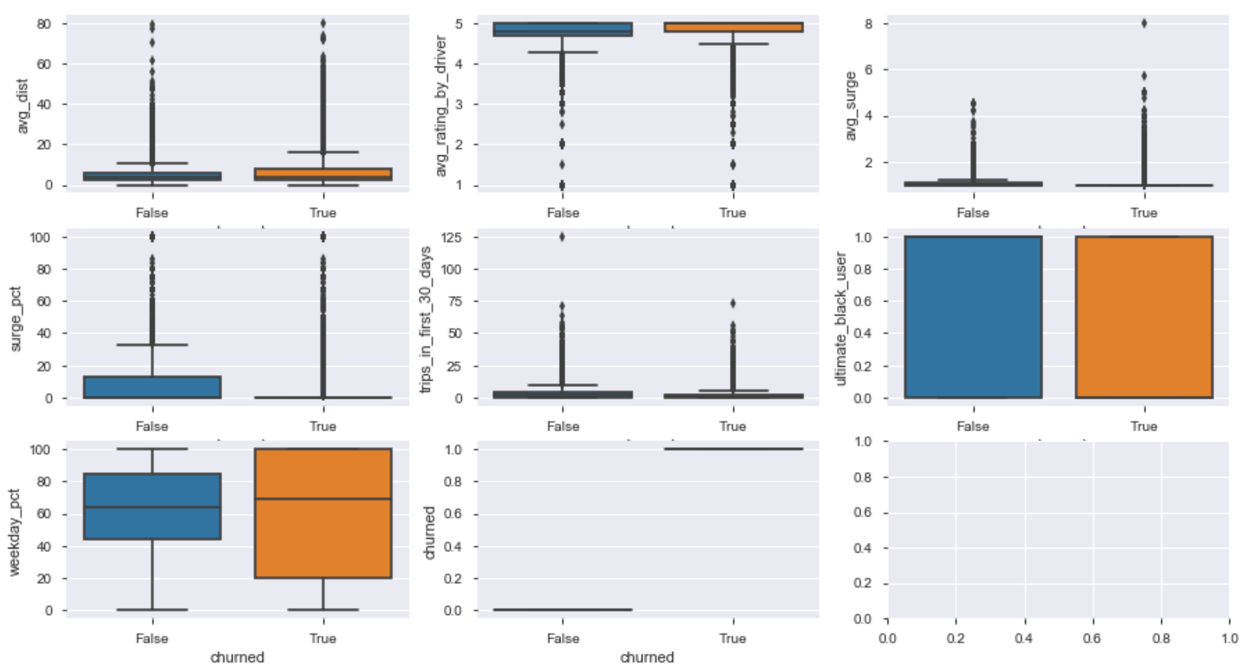
To perform this experiment, I would randomly select a group of drivers from each city and offer half of each group reimbursement for the cost of the toll. The other half would be left as a control. I would set a significance cutoff of $p < 0.01$ and work with the cities to determine what effect size would be needed to say that the program is a success.

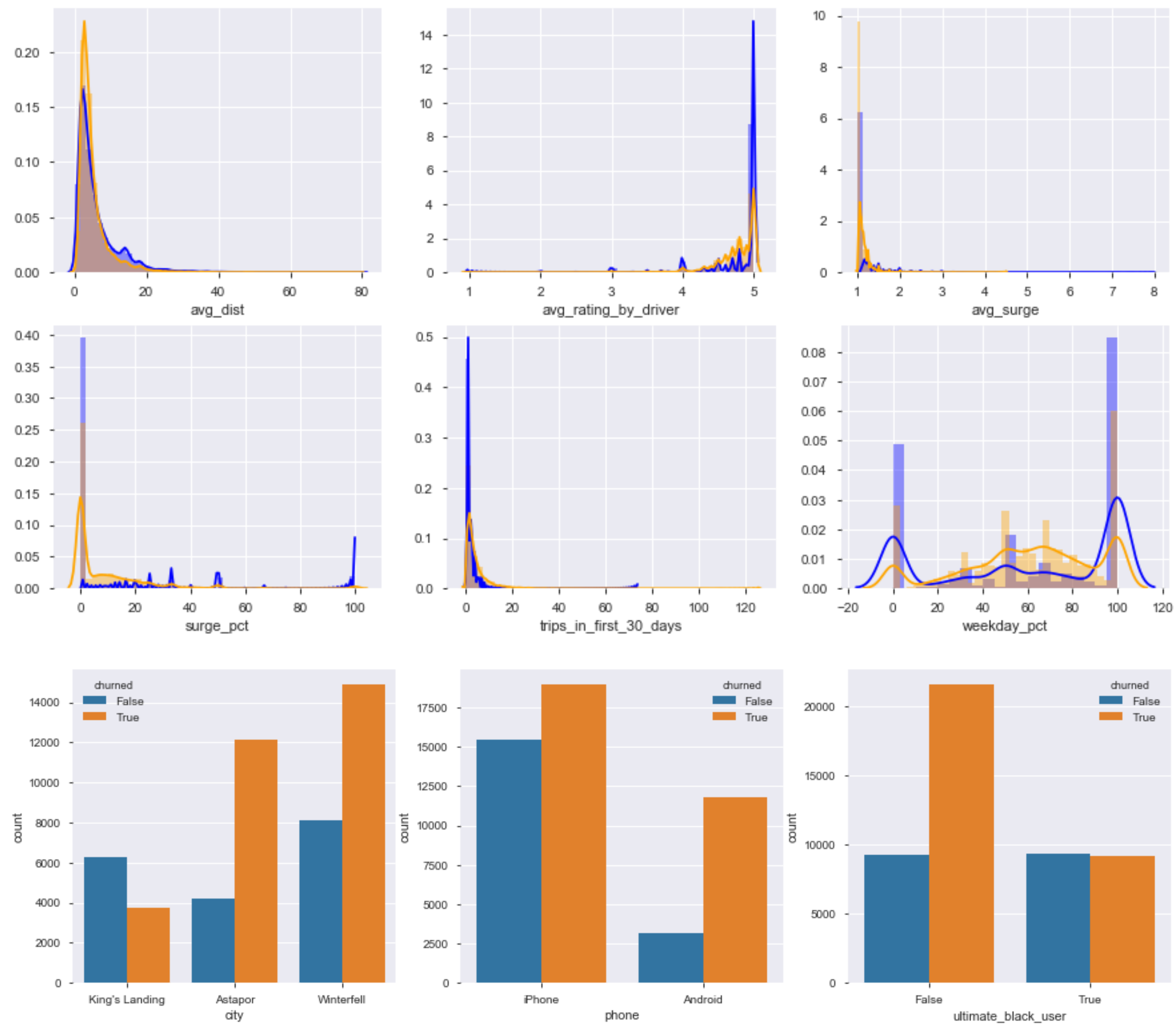
I would use two tests to determine if there is a change in the mean number of crossings per week per hour worked. The first I would use is the t-test to determine if there is a significant difference in the mean between the group that received reimbursements and the control. This is a commonly used test and would be easy to explain to the clients. I would also ‘re-randomize’ the drivers, compute the difference in means for the new groups, and repeat thousands of times. I can then determine how often a difference as big as the original occurred. A second test is useful for checking results, but this is not as easy to explain.

If both tests are statistically significant and the effect size is larger than the predetermined size, I would recommend implementing this program for all drivers. If one of the significance tests fail, I would say that there is only very weak evidence that and effect exists and is possible that there is no effect. If both tests fail, I would strongly recommend not implementing this change.

Part 3 - Predictive Modeling

I used the pandas profiling tool to create a report that shows basic statistical details about each variable with a histogram. It also finds duplicate rows, columns with missing values, and correlations between variables. I then looked at boxplots and distributions by churn status. 37.6% of customers who signed up in January were retained six months later.





For modelling, I decided to compare logistic regression with random forests. Logistic regression is the typical starting point for a classification problem. Random forests typically perform better even without tuning the hyperparameters and are still fairly simple to explain. I used precision, recall, and AUC to evaluate the models and found that, overall, the random forest classifier performed better than the logistic regression.

The two most important features in the random forest model were average distance and weekday percent. Those who were retained were more likely to have a shorter average trip distance. Customers who left Ultimate were more likely to either use the service only on weekdays or to never use it on weekdays. Incentivizing shorter trips and/or everyday use for early users may help retention.