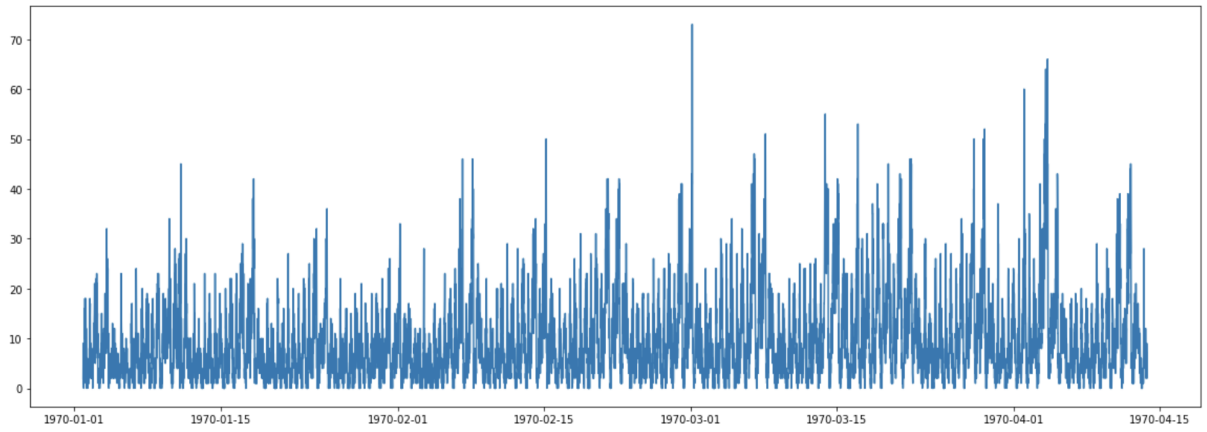
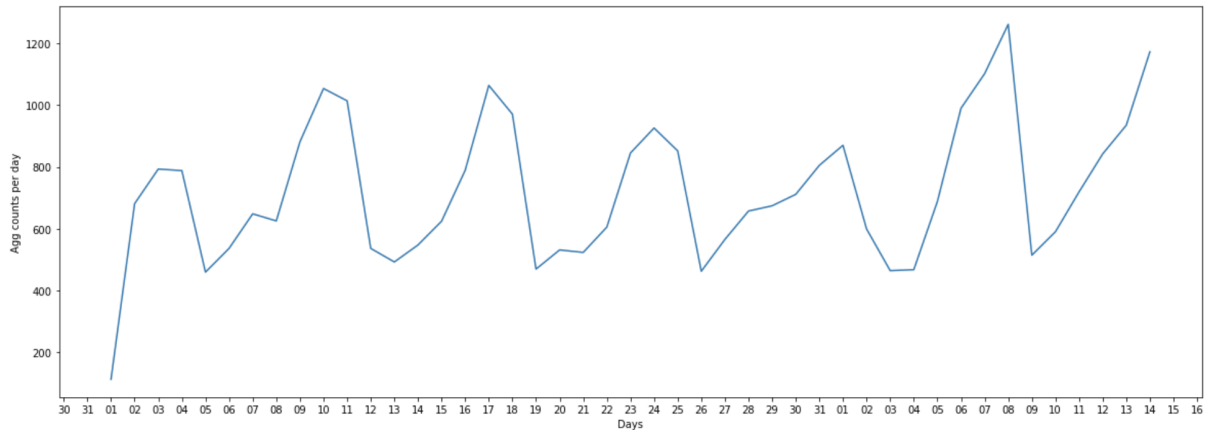
Take Home Challenge

Part 1

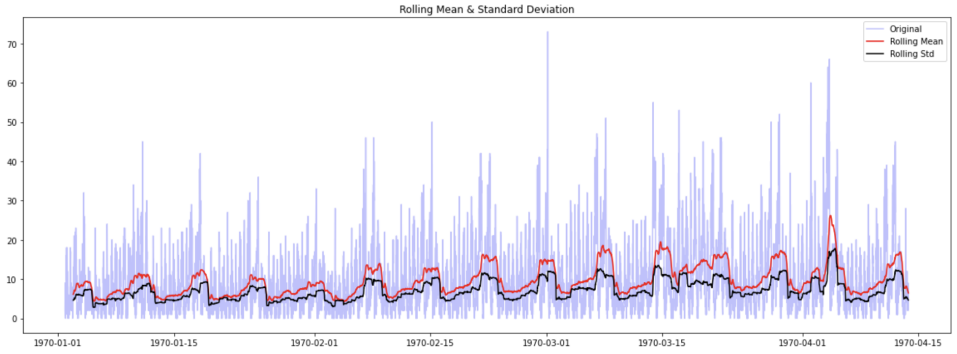
When aggregated into 15-minute intervals, the logins over time can look pretty confusing:



However, when aggregated a rolling mean over hours is produced, the graph seems to take on a pattern:

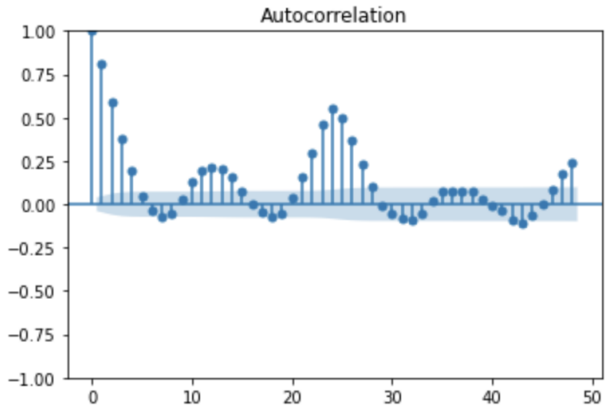


When passed to a Dickey-Fuller test, the results indicated the data is not stationary:

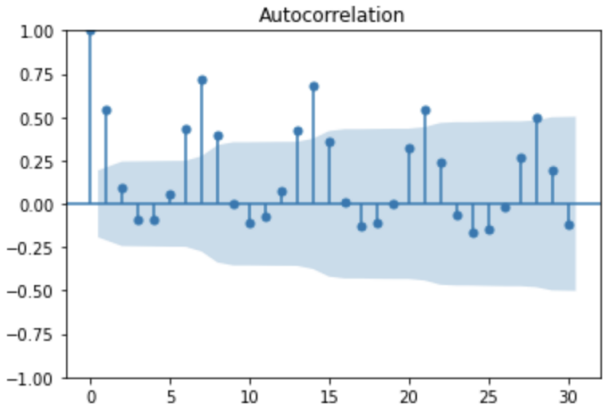



```
Results of Dickey-Fuller Test:
Test statistic:  -10.340737765830102
Critical value:  -3.431021011779871
Non-stationary
if the Test Statistic: -10.340737765830102 is greater than the Critical Value: -3.431021011779871 than the time ser
ies is stationary
```

Using autocorrelation, there does seem to be hourly patterns. There are increases in requests at even multiples of 6 (12, 24, 36, 48…) and decreases in requests at odd intervals (18, 30, 42..).



When it comes to weekly patterns, Monday and Tuesday have the lowest requests while Saturday and Sunday have the highest.



There are no distinct monthly patterns.

Part 2

1. Assuming data on location is provided to the city, the difference in the percentage of drivers serving both cities could be used. If there is an increase in this over time, it would mean that drivers are distributing themselves across the city. This metric could be collected via GPS location.

2. a) The first step to build a suitable experiment would be to randomly select half of the drivers in each city. One half of the drivers in each city would be given a reimbursement for the toll cost while the other half, the control group, would continue as they normally would. If the metric, the number of drivers in each others' cities is higher than an agreed upon critical value, the experiment would be considered a success. Gotham's experiment would run in the day while Metropolis's experiment would run in the night due to their complementary demand.

b) The following could be used:

P = (# of drivers of city A operating in city B / # of total drivers in city A) x 100%

$P_{Goth,\ day} = P_{Goth} - P_{Goth,\ control}$
$P_{Met,\ night} = P_{Met} - P_{Met,\ control}$

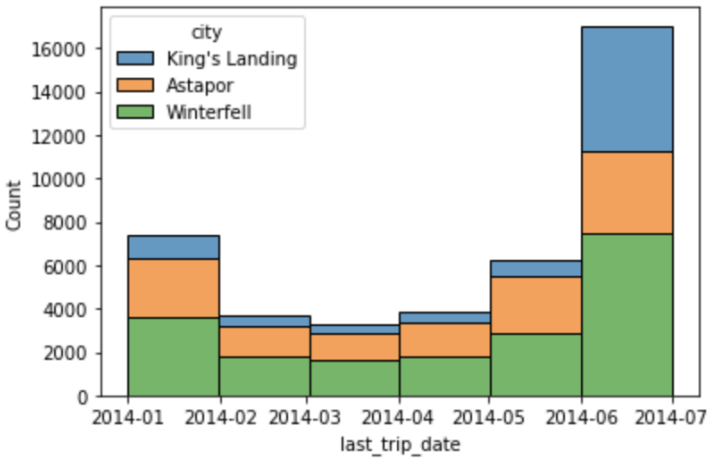
H_0 : $P_{Goth,\ day} >$ Critical value
$P_{Met,\ night} >$ Critical value
H_a : $P_{Goth,\ day} <=$ Critical value
$P_{Met,\ night} <=$ Critical value

b) The experiment would run for let's say 30 days. Snapshots of GPS location every hour could be used to determine where drivers are for 24 hours, creating a percentage metric for each hour. A probability distribution of this metric for a night-cycle (8PM - 5AM) for Metropolis and a day-cycle (6AM - 7PM) for Gotham could be aggregated using the hourly snapshot metrics. Two one-sided t-tests could determine whether the increase in mean of driver distribution are higher than the critical value and are statistically significant for each of the two tests.

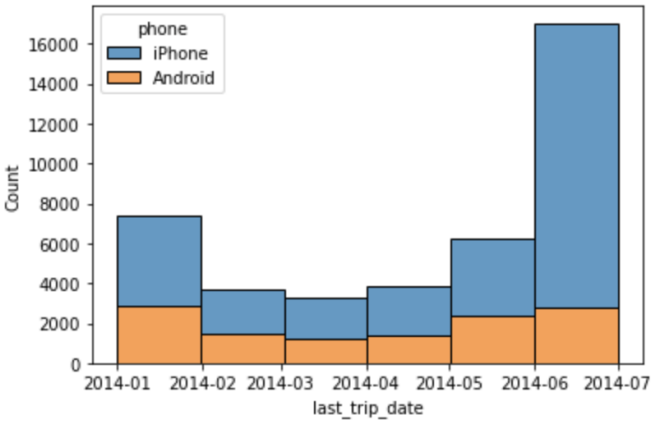
c) If both t-tests confirm the null-hypothesis, the alleviation of the toll would be a success. If not, it would mean that driver concentration is not evenly distributed and further tests would need to be done to determine the cause. Since the alternative hypothesis states that concentrations aren't equal, it would be unknown whether This does not mean that the alleviation of the toll was an error.

Part 3

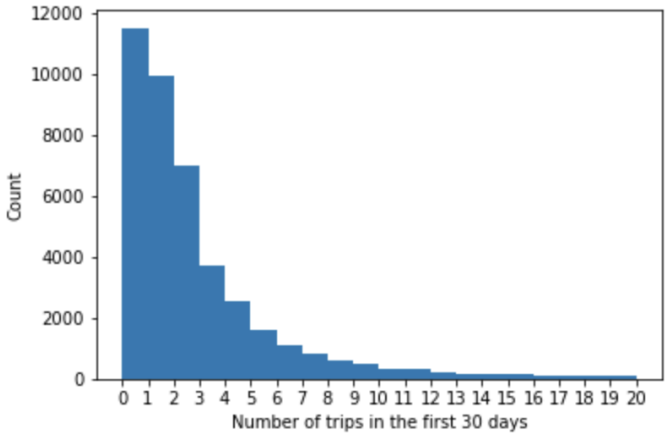
1. There were many null values in the avg_rating_by_driver column which were removed. Since these null values were evenly distributed across time, it shouldn't interfere with the effect time has in the analysis. The last recorder date for trips was 2015-07-01.



It seems that many of the last trips of users were in June itself. Many of these users were from King's Landing.

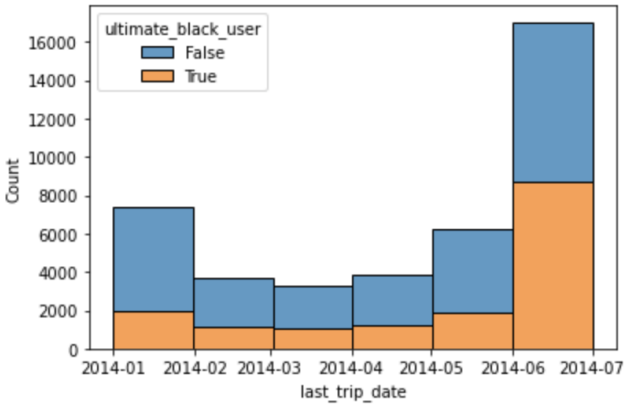


Most of the frequent users of the app were iPhone users.



The distribution of the number of trips of each user in the last 30 days was heavily skewed from 0-3 as seen in the histogram above. Meaning that many users didn't use the service very frequently.

It seems as though and equal number of Ultimate black users and regular users have used the service recently.

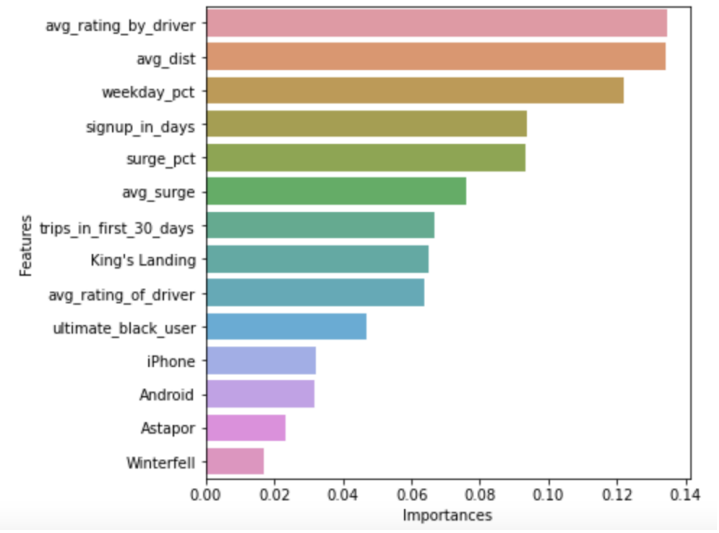


40% of users were retained.

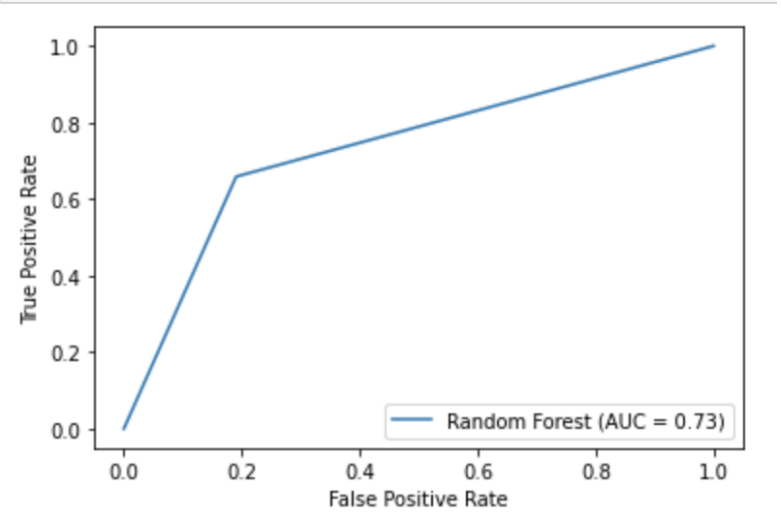
2. 3 models were fit to predict whether users were retained. The problem was that of binary classification; whether users used the service in the last 30 days or not. Using last_trip_date, a metric: 'in30' was calculated and served as the target variable and last_trip_date was dropped. sign_up_date was converted into the number of days since the date the first customer signed up. 3 models were built using this data, a K Nearest Neighbours Classifier, a Random Forest Model and a Logistic Regressor. Recall was chosen as the metric since it is of utmost importance to accurately determine which users are actually retained. The numeric values were scaled using a Standard Scaler prior to being split into train and test sets. The following are the recall results of each of the classifiers:

	Test	CV
Default KNN	0.669985	0.650109
Tuned KNN	0.698490	0.663657
Default Random Forest	1.000000	0.679793
Tuned Random Forest	0.902746	0.693100
Default Logistic Regression	0.674535	0.541848
Tuned Logistic Regression	0.674535	0.541848

The Tuned Random Forest was chosen as the best model since it performs the best during cross-validation. The feature importances can be seen below.



The ROC-curve can also be seen below:



All in all, the Random Forest model does seem to do a good job, however it is surprising that customers with an iPhone are listed so low in terms of feature importance since a large fraction of the most recent users did use iPhones. I also expected users coming from King's Landing to be higher up in importance for the same reason.

The recall and AUC score aren't the best and they do leave a lot of room for error. More features of each customer could be considered to create a more accurate model. Additionally, a more robust model itself could be built.

3. The average rating by the driver, average distance travelled in the first 30 days and the percentage of trips taken during the weekday were the most important values. The average rating of the customer by the driver is a metric that is subjective and can't be measured, however the other features definitely can be disseminated. Users who travel longer could be wealthier and thus continue to use the service. Customers travelling to work during the weekday would be wealthy enough to afford trips to and from work, while weekend users could be ones that use the service to get around town during the night.

Ultimate should consider focusing on the core users of the service to increase retention. The core users definitely seem wealthier however, a premium service like the Ultimate Black service don't seem to drive retention in a significant way.