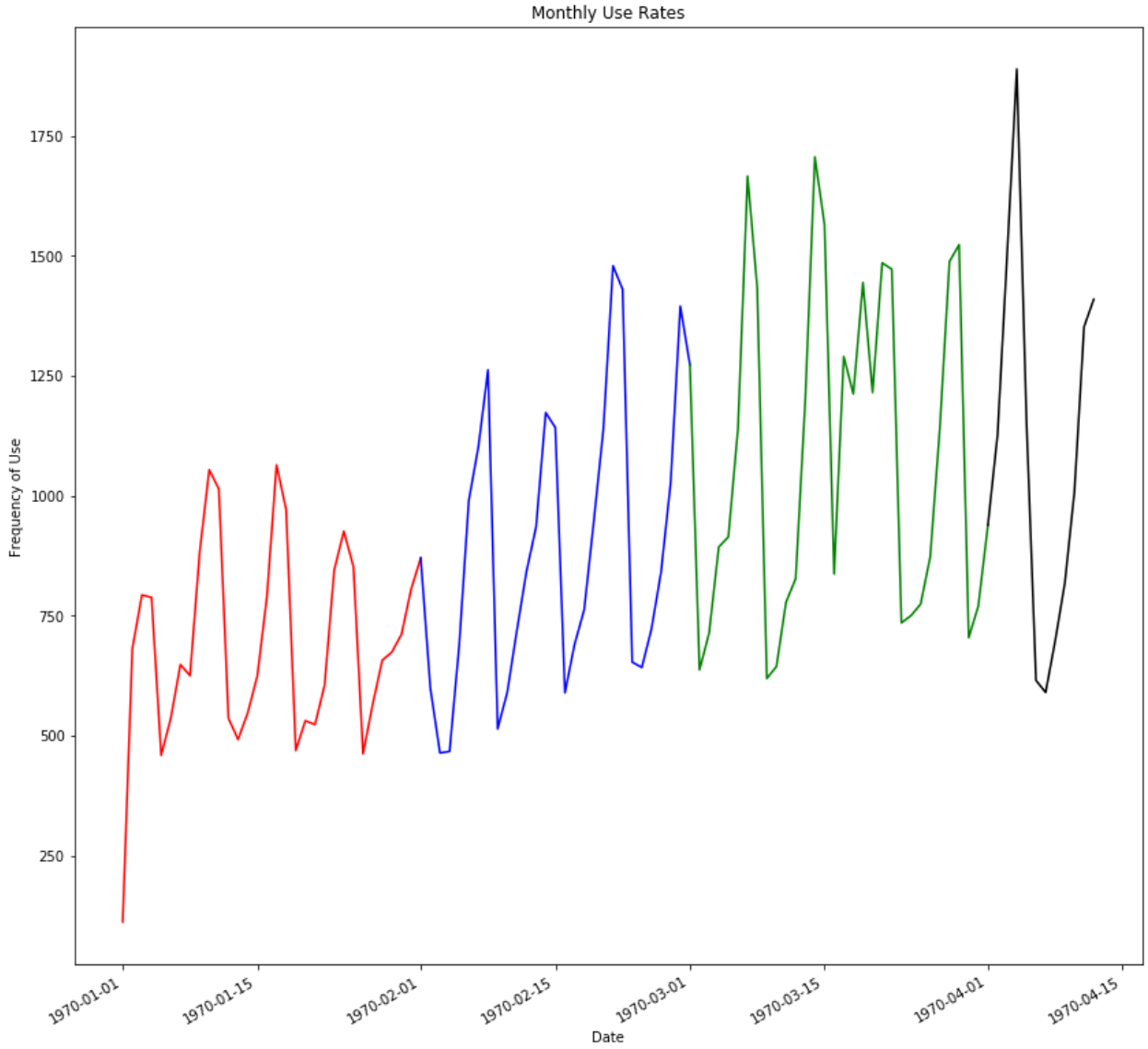
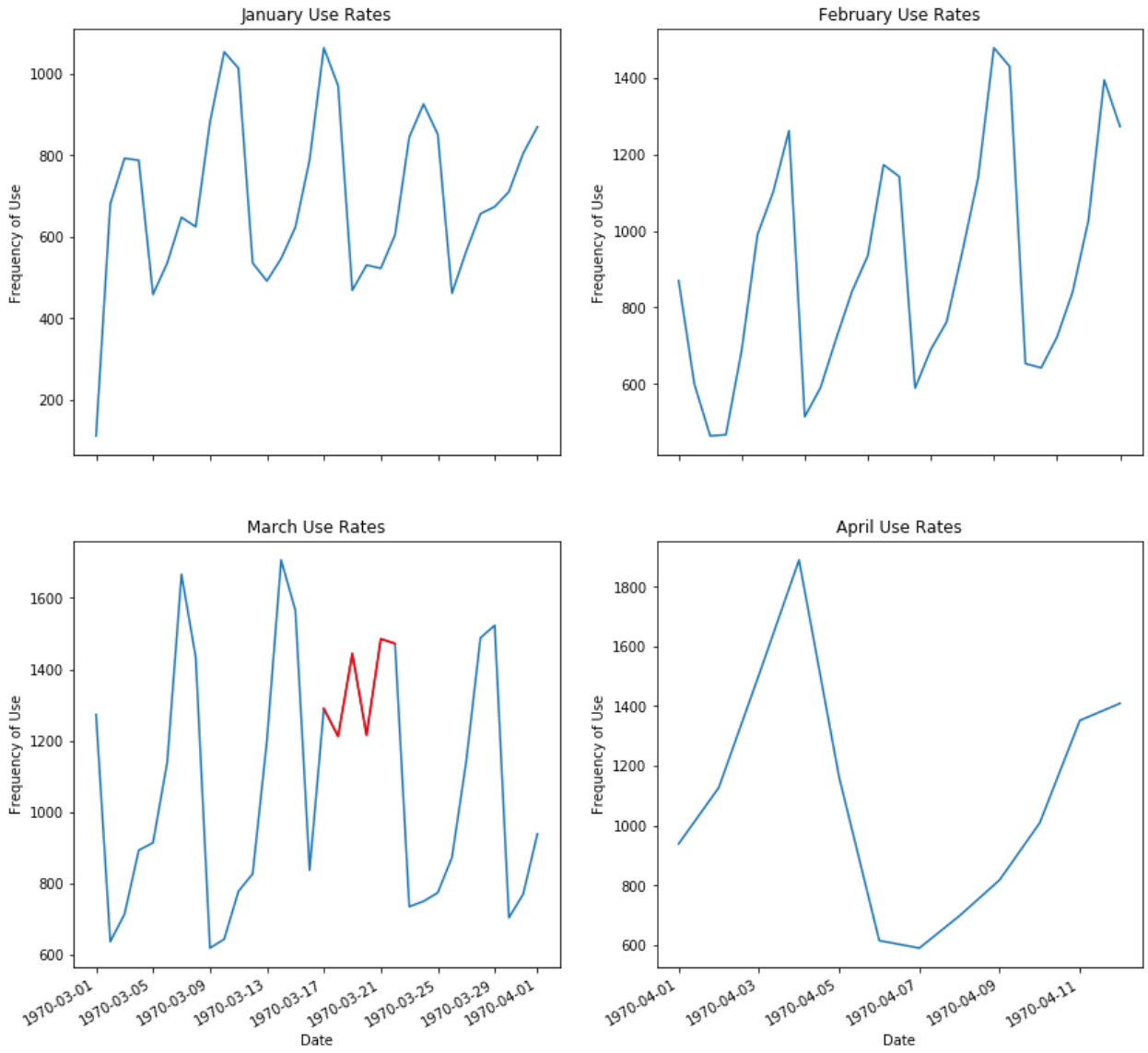
Analyzing the Movement of People in Taxis

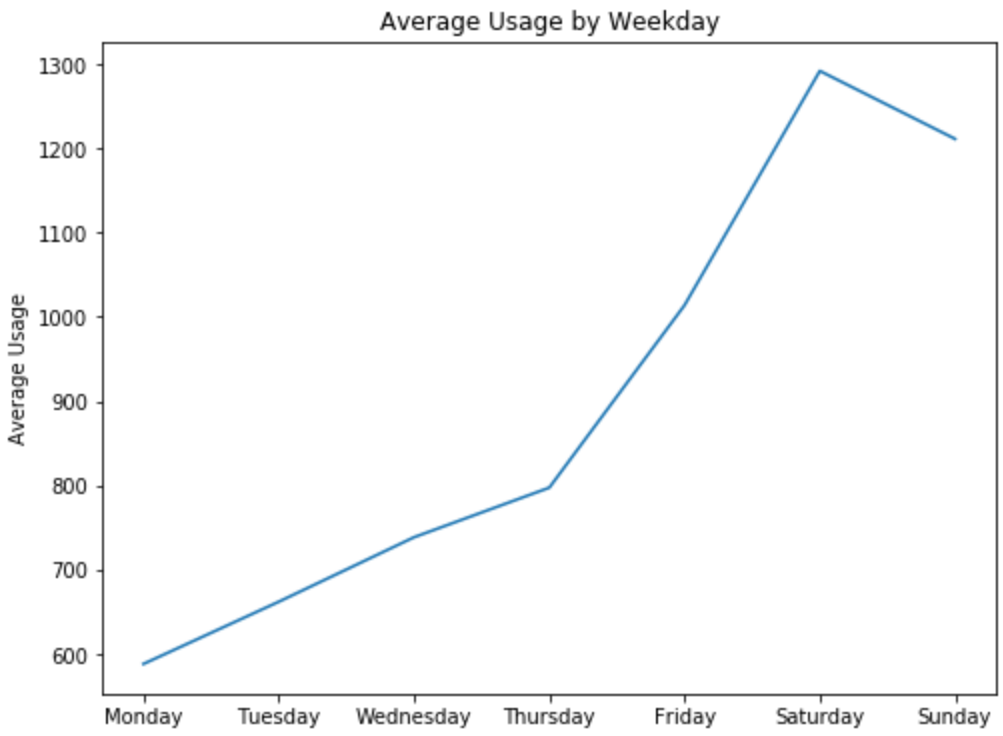
Part One: EDA



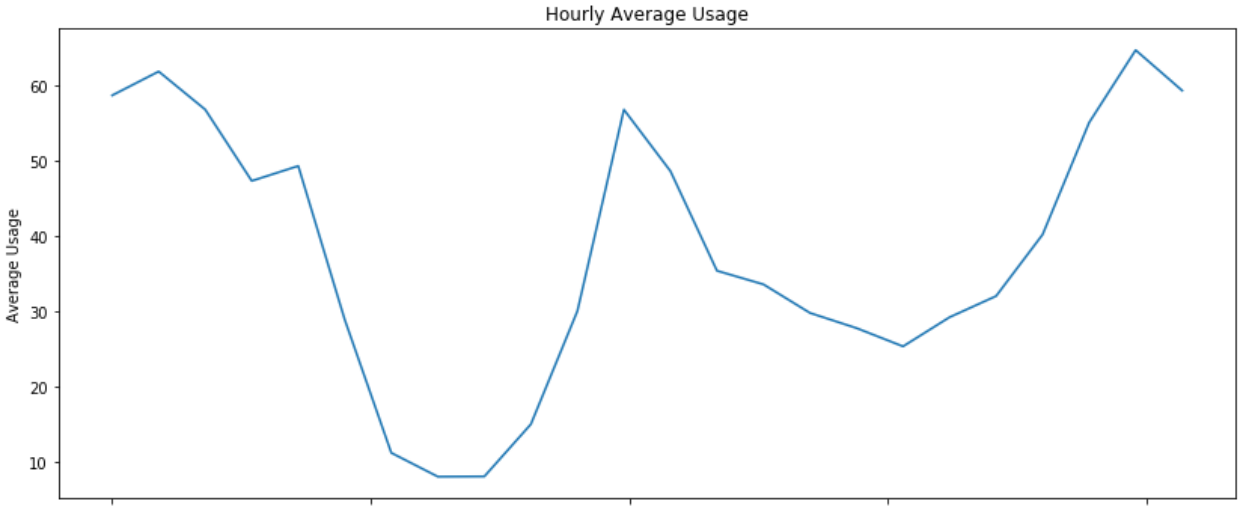
Notice that over the course of the first 4 months of the year, the data appears to trends upward, in that use at the given location of these logins has increased in these 4 months.

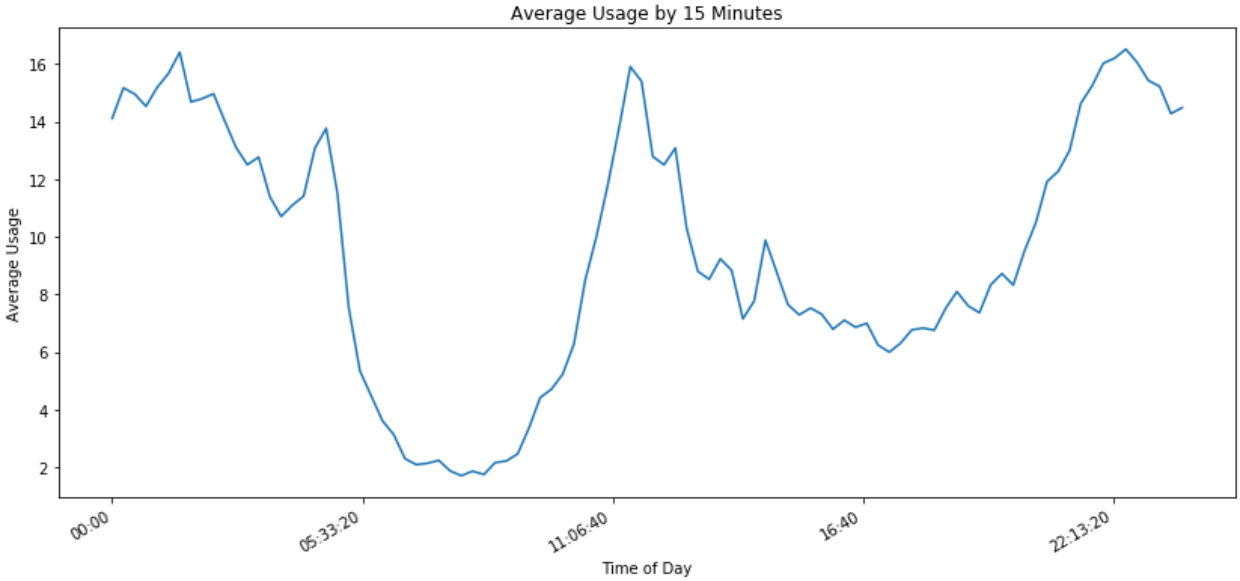


When looking at just one month, it can be seen that the use peaks on weekends, where the peaks on the graph exist. This continues in all 4 months in the dataset. Notice that in the case of March, there is one week that bucks the trend (see red). This is in the third week of March, and may have been due to St. Patricks day, which occured on the 17th, and represents the first spot where the graph hikes up that week. As the holiday is often celebrated with excessive drinking, it would make sense that the use of taxis that day would increase. It is also a holiday that can sometimes last over the course of the week, potentially causing the odd use pattern within that week in that month.

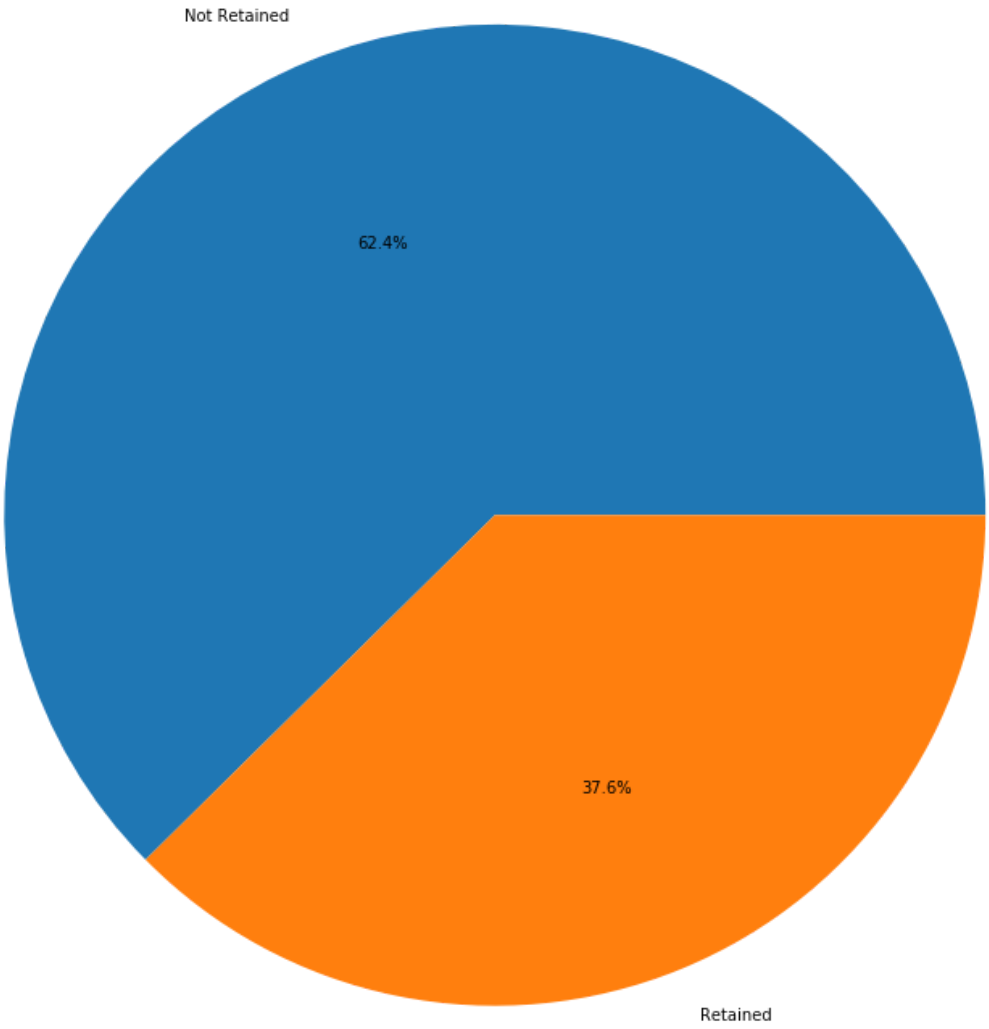


Based on the above graph, we can see that the highest usage occurs on the three weekend days of Friday, Saturday and Sunday. The earlier days in the week have much less use, though they do has a slight upward trend. Sunday has slightly less use than Saturday, but still significantly more than the weekdays.

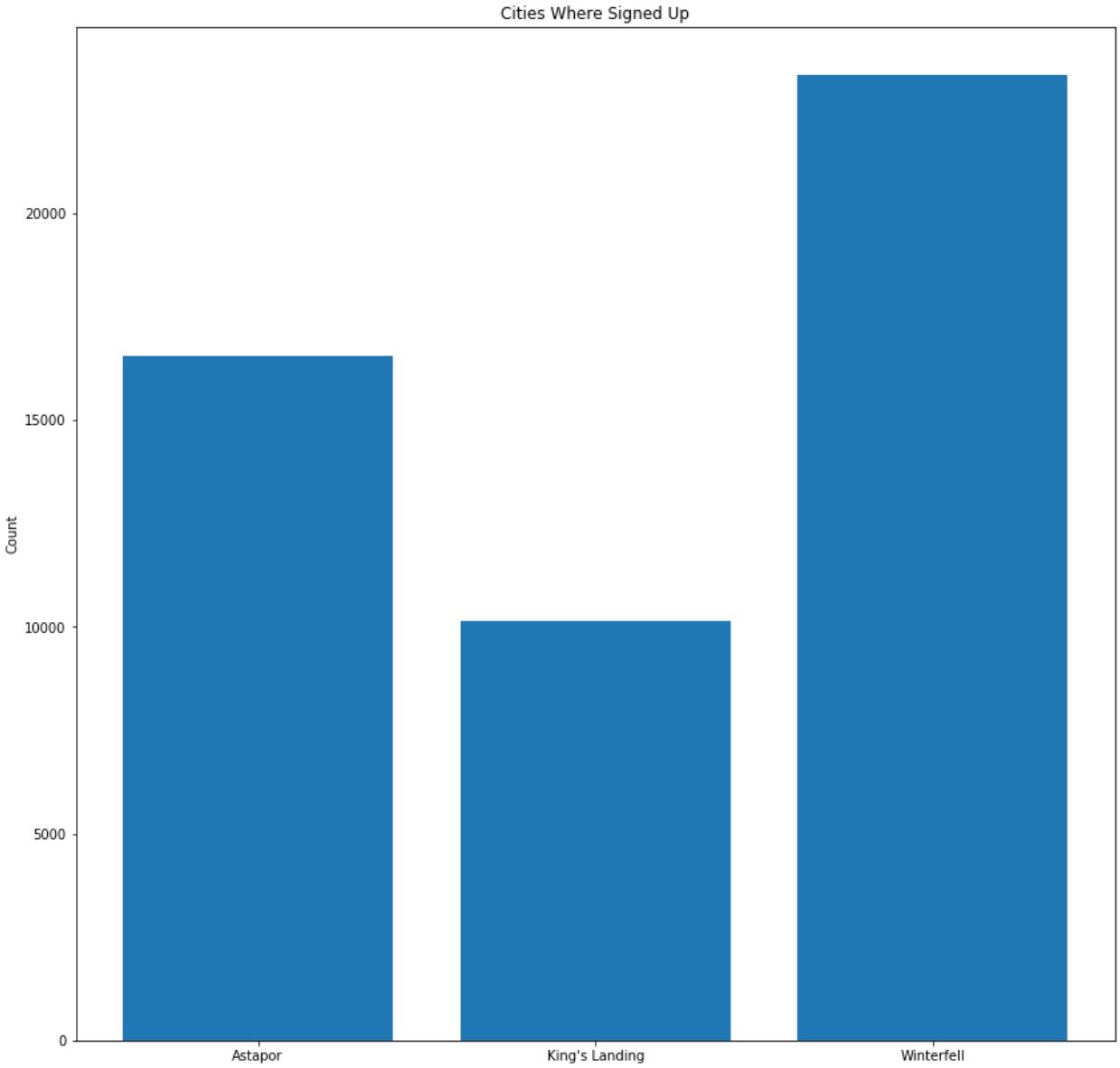




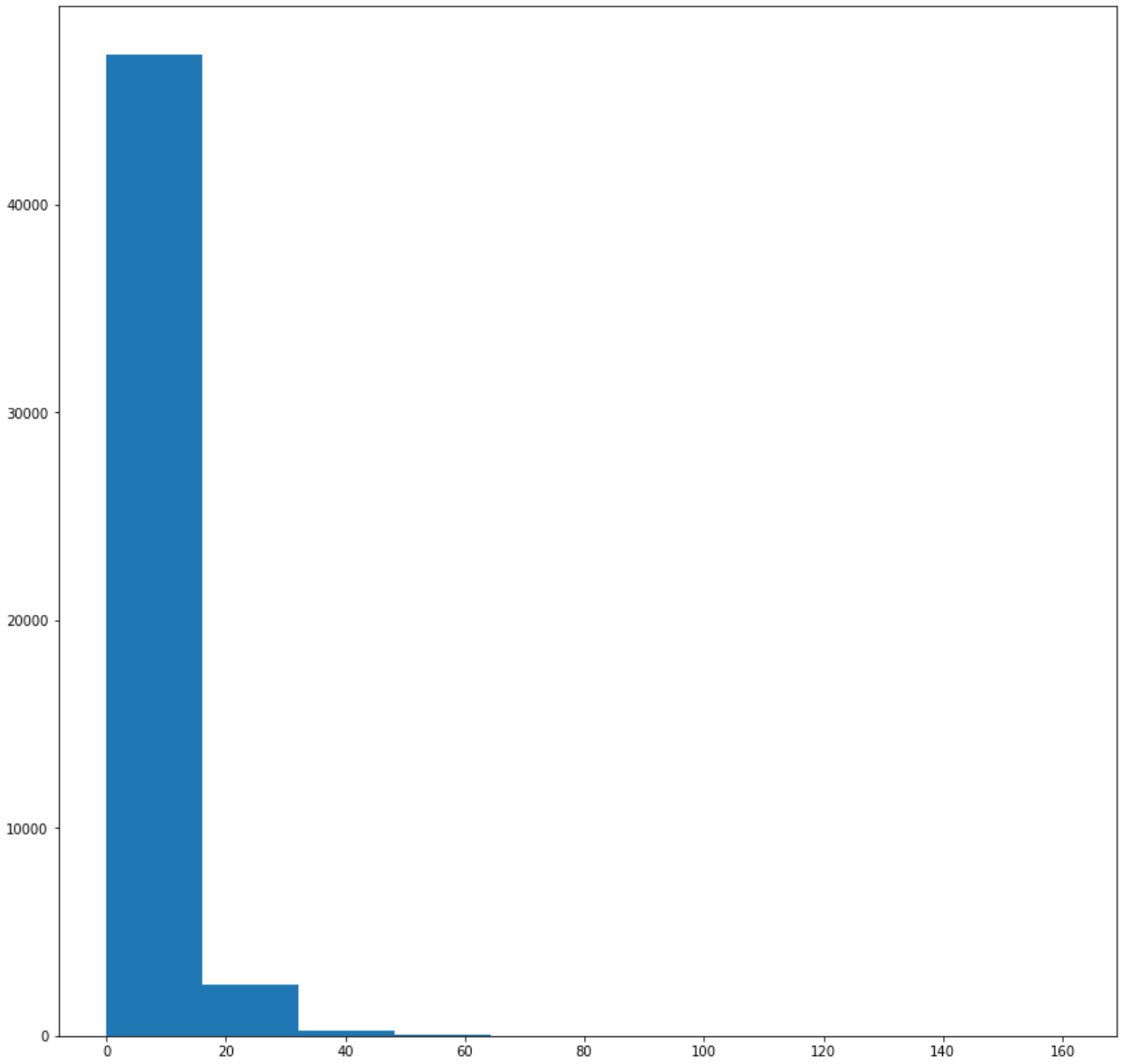
Note that the above graphs show similar trends, with the bottom graph showing more detail, as it is by 15 minute intervals, rather than hourly intervals. Both graphs start at midnight, and finish at midnight the next day. It would appear that there are two relative spikes in use over the course of the day. The first, shown at the beginning and end of the day, peaking just before midnight, could be from people getting home from after work activities. The second spike appears at around lunch time, and most likely is due to people taking a taxi to lunch.



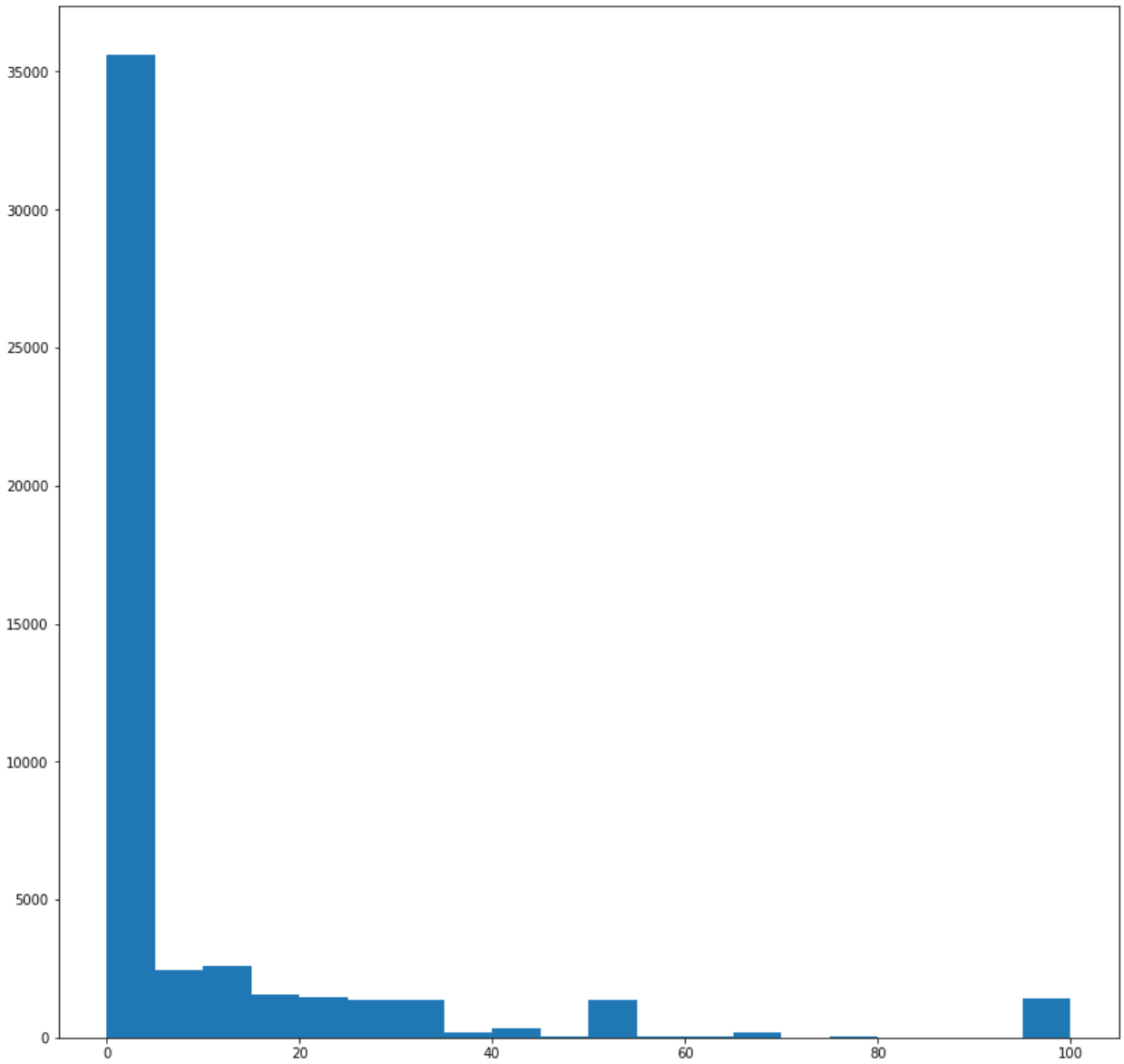
Notice that the number of retained users is only a third of the entire dataset, suggesting that continued use of the service is not what the majority of users do.



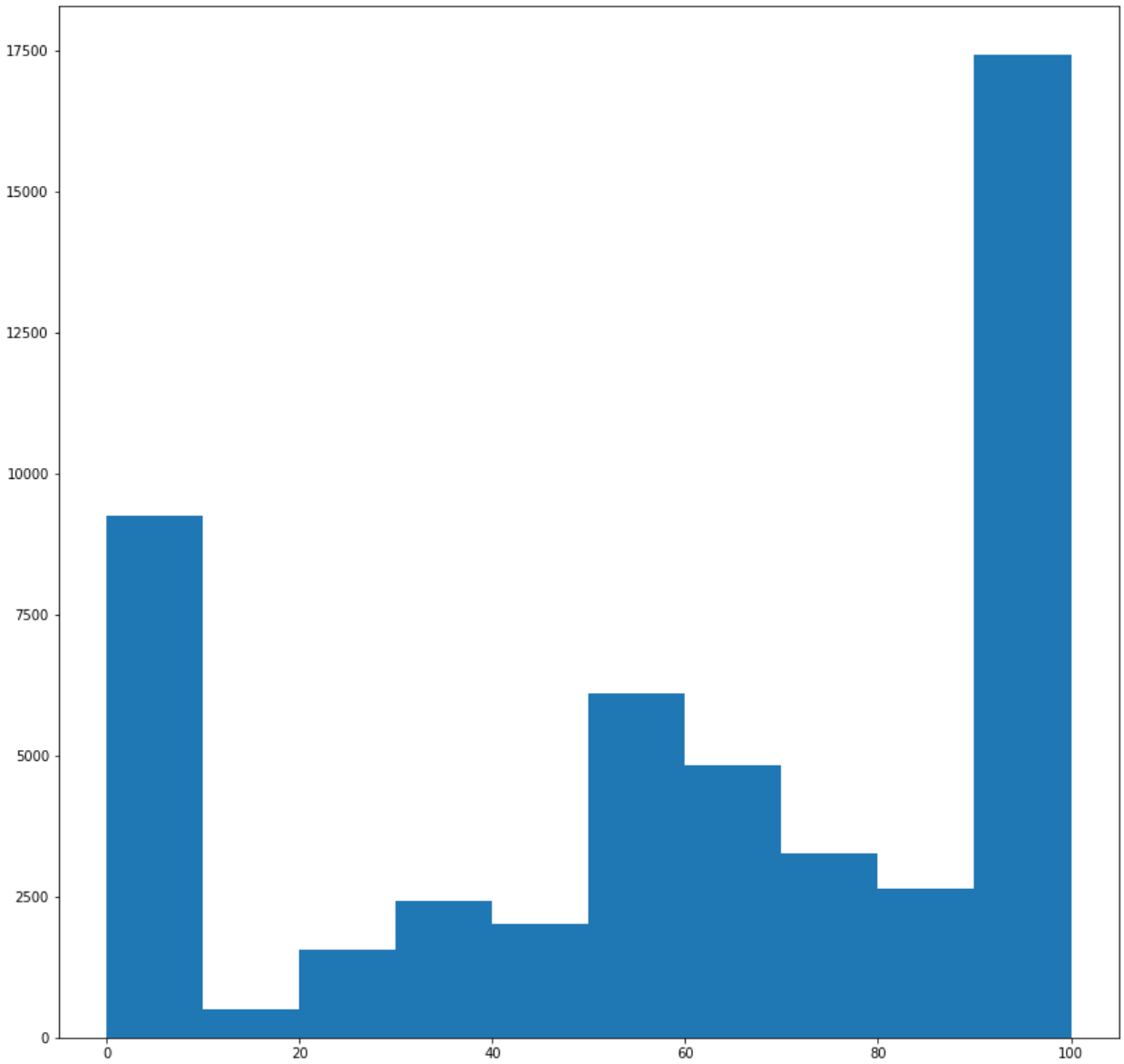
While Winterfell had the largest number of signups, all of the cities have large numbers of people who signed up in their areas.



Most users had fairly short trips when they first signed up, as the data is heavily left skewed.



While many kept out of surge times,there were still small peaks at 50%, and just under 90%, representing travellers who travelled at peak times 50% of the time, and nearly 100% of the time respectively.



This distribution shows a few different spikes. The first occurs near 0, representing the users who signed up and either never used the service on weekdays, or only did once or twice. In the middle there is a wider bump, representing those who are just as likely to use the service on a weekday as a weekend. Finally, the largest spike on the far right shows those who use the service mainly on weekdays. These likely represent those who use the service for commuting, or for lunches at work.

While there appears to be no data quality issues in this portion of the data set, all of the values are either summed or averaged to help decrease the effects of outliers. By doing such a grouping, and also looking at the trends of the data at multiple levels of time grouping, it is possible to make conclusions even if there are chunks with poor data.

Part 2: Experiment and Metrics Design

1. I would use pickup location along with total trips to assess the sucess of this experiment. By grouping the trips by city, and then marking the percentage of each city per driver, it can be seen where they habitually provide service. I would use this calculated measure to determine success as the current goal of the cities is not to increase the use of Ultimate on the whole, but to increase the cross pollenation of drivers between the two cities. This percentage value can then be used in the experimental design described below.
2.
    a. . The experiement will be done over the course of 2 months. In the first month, a random sample of drivers would be taken from the entire driver pool. This sample size can be determined by the constraints of budget, as a t-test will be used later, which allows for smaller sample sizes. These drivers would have their pickup locations recorded over the course of the month, and then the proportion of those pickups in each city recorded. They would then be divided into two separate goroups based on those proportions: those who service Gotham more, and those who service Metropolis more. Then for the second month of the experiement, all of the sampled drivers would be reimbursed for their toll fees, and over the course of the month, all pickup locations would be recorded. Then, at the end of the second month, each driver would get a new proportion value based on the new data. Note that a driver who was found to service Gotham more in the first month would keep their Gotham proportion as their value for both of the months while a driver found to service Metropolis more would keep their Metropolis proportion for both of the months.
    b. In order to determine any statistical significance from the above experiment, 2 metrics would be calculated, one on each of the sub groups. In both cases, the test would be a paired t-test. In the case of the group who started servicing Gotham more, the null hypothesis would be that the mean difference in their proportion of serving Gotham before and after is zero. We would perform a lower tailed test to give an alternative hypothesis that the mean difference in their proprtion of serving Gotham is smaller than 0, thus meaning that they served Gotham less in the second month, and must have served Metropolis. The same logic is then applied to a lower-tailed paired t-test with the Metropolis group.
    c. There are 4 different possibilities for this experiment and the paired t-test outcomes. First, both of the t-tests could not reject the null hypothesis. In this case, there is no statistical evidence that the mean difference is zero or less than zero, so the test is inconclusive. A new experiment would have to be designed, or the previous one analysed in a different way to deal with not rejecting the null hypothesis. The second case is that the null hypothesis is rejected for the Gotham group and not the Metropolis group. That means that the Gotham drivers did service more Metropolis users in the second month, but we can make no conclusion about the Metropolis drivers. The third case is the inverse of the second case, and would conclude that Metropolis drivers serviced more Gotham users, but we can make no conclusion about the Gotham drivers. Finally, if both t-tests reject the null hypothesis, the experiement shows the program is fully successful, and that both sets of drivers serviced the other city more often than before.\

Experimental drawbacks: There are few drawbacks to the experiemental design about. First, it has not way of factoring in if drivers change their frequency of trips under the new or old system. So, a driver who makes 1 trip in Metropolis and 10 trips in Gotham the first month, and then 1 and 5 respectively in the second month will still tend towards rejecting the null hypothesis, when the case may be that they just had fewer trips total. Another is that it doesn't account for cross-city trips, as only the pickup location is counted. So, if a driver took a client from Gotham to Metropolis, and then drove back to Gotham to pick up another client, the driver would never show being in Metropolis. While this may not occur many times, it is not accounted for in this experiemental design.

Part 3: Predictive Modeling

The data was modelled via logisitic regression using 2 different versions of the dataset, built based on the handling of N/A values. In the first case, the dataset dropped all rows, while in the second, it drops columns that include N/A values.

I chose to use logistic regression, as the output is a binary variable, so it fits into logistic regression well. Also, logistic regression gives clear values for outputs that allow for fairly quick analysis and understanding. This could also be modelled with any number of other algorithms, including random forest or naive bayes. Logistic regression will allow for some colinearity between variables that may occur in this dataset, so it is also a solid fit in this case.

Above are the two summaries of the two logistic regression models for the two different methods of handling null values in the dataset. They both produce very similar results, with the only difference coming in the size of the coefficients of the variables. Those changes are small, and since the two models produce good results with an f1 score of more that 0.92, either model could be used.
Features with an effect: Note variables are only included if their coefficient is significantly different from zero.

Positive Effects: avg_surge: The more often someone uses the service at surge times, the slightly more likely they are to continue using the service. trips_in_first_30_days; Users who use the app in the first 30 days after signing up are more likely to continue to use the service. ultimate_black_user: Users who use the ultimate black service in the first month are more likely to be retained. time_from_signup: Someone who has had the service for longer is more likely to continue using the service

Negative Effects: weekday_pct: The more a user uses the service during the week, the less likely they are to continue using hte service. time_from_last: The longer it has been since a user has used the service, the less likely they are to use it. This has the largest effect of any of the features in the set.

Cities: Users in Asataphor and Winterfell are slightly less likely to continue using the service, while those in King's Landing are more likely to use it.
Possible Improvements/Focuses for future retention:

1. Remind users that the service exists, as those who have used the app more recently are more likely to use it again.
2. Increase marketing in Asatpor and Winterfell, where users are less likely to be retained.
3. Give users a credit for ultimate black to give them a chance to use that part of the service and be more likely to be retained.