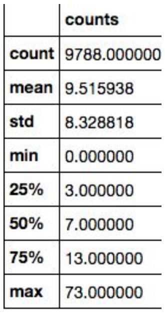
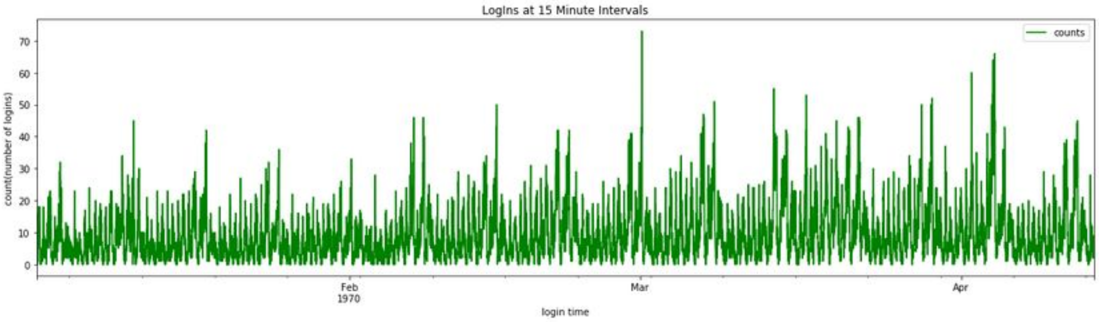
Part 1 - Exploratory data analysis

On average there are approximately 10 logins every 15 minutes. The median shows 7 logins every 15 minutes. There are a few outliers that are quite high around 70 logins. There can be exploration on these exact dates to look for causes on these extremes. Even those over 50 look to be more of outliers. Looking at the quartiles or the visual plot, it is apparent the more normal amount of users falls well below those extremes.



	counts
count	9788.000000
mean	9.515938
std	8.328818
min	0.000000
25%	3.000000
50%	7.000000
75%	13.000000
max	73.000000



Part 2 - Experiment and metrics design

The neighboring cities of Gotham and Metropolis have complementary circadian rhythms: on weekdays, Ultimate Gotham is most active at night, and Ultimate Metropolis is most active during the day. On weekends, there is reasonable activity in both cities. However, a toll bridge, with a two way toll, between the two cities causes driver partners to tend to be exclusive to each city. The Ultimate managers of city operations for the two cities have proposed an experiment to encourage driver partners to be available in both cities, by reimbursing all toll costs.

1. A key measure of success to encourage driver partners to serve both cities would be to see growth in the weeknight activity for Ultimate Metropolis, and growth in the weekday activity for Ultimate Gotham. It is reasonable to conclude that toll costs are affecting the weekday activity if the reimbursement shows significant change. Further, it will be interesting to see if this changes the weekend activity in cities. If that changes drastically it will lead to more questions to be explored outside of toll, which could be more about the cities themselves and their offerings.
2. Describe a practical experiment you would design to compare the effectiveness of the proposed change in relation to the key measure of success.
    a. To implement the experiment there would need to be an initial set of data being used for the initial statement about the differences. This set could be based on monetary values or potentially number of rides requested. Exact monetary costs per ride fluctuates, whereas the number of riders are set. There needs to be a dataset over a period of time showing number of rides in each city, broken down with dates and times. The experiment, once all tolls are getting reimbursed, should be done for the same amount of time with the same drivers. The drivers

would continue business as usual, with no changes to the way they advertise or connect to their service to get passengers.

b. To verify the significance of observation, a few tests could be used. Hypothesis testing based on their hypothesis that the toll charges are affecting the activity in each city. The z-test for this larger sample would work once established there is a normal distribution from the Shapiro-Wilk test. Further, finding the significance level and p=value. It would be good to use the Pearson test to check the linear correlation of tolls to activity.

c. The results could require running a Shapiro-Wilk test of normality on each dataset. This way the p-value can make a judgement about the normality of the data. If it is not showing to be so, the data can be evaluated for nuisance parameters and outliers, then adjusted to eliminate those causing the fluctuations. The z-test would be used over the t-test due to the large size of the sample, and determine whether to accept or reject the null hypothesis.
   It would be wise to use seaborn plots and visuals such as a pair plot to show the operations team the results from the initial data, compared to the data once the tolls were taken out of the equation. Recommendations to the city could be based on looking at the number of rides during times of day and days of the week, grouping them into day and night activity as mentioned in the design.

Part 3 - Predictive modeling