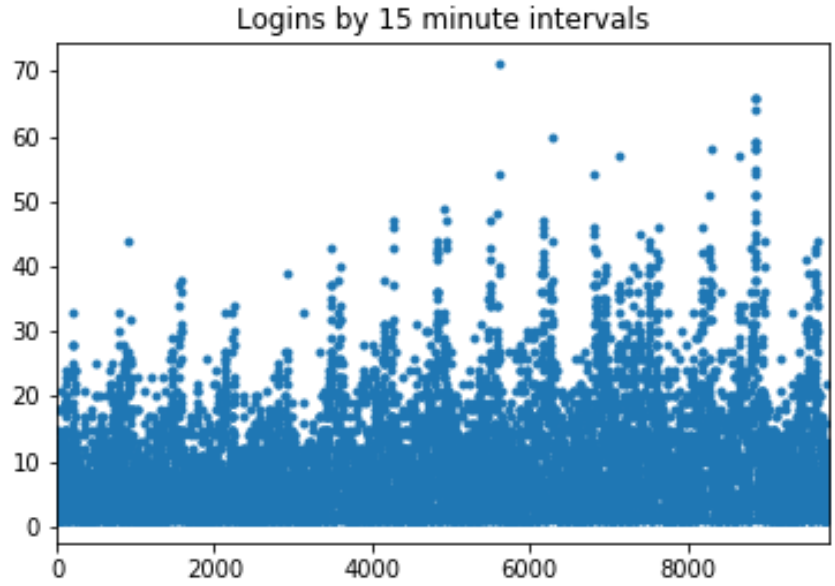
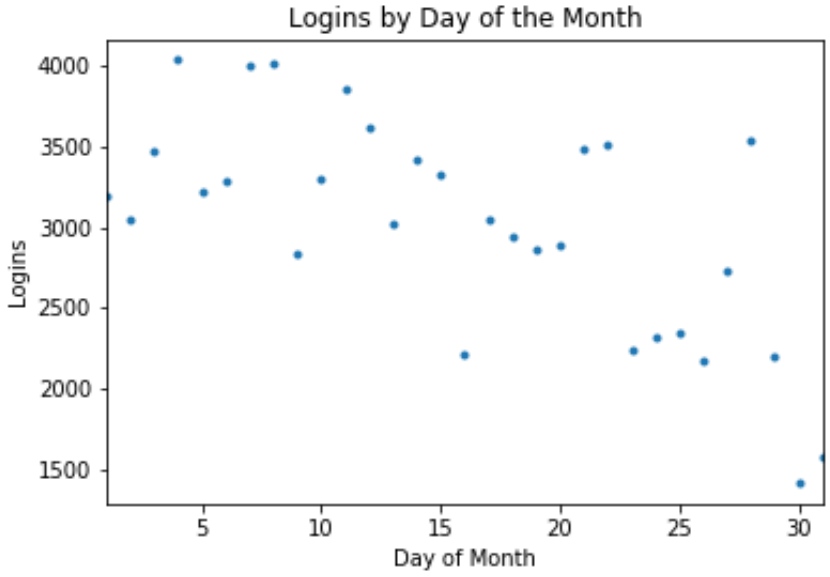
Data Analysis Interview Challenge

Part 1: Exploratory Data Analysis

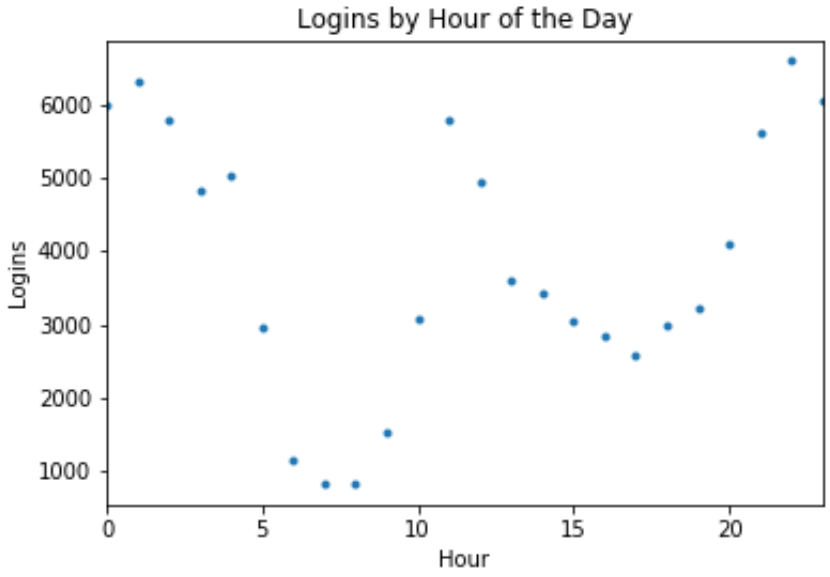
To look for patterns in the demand, I first aggregated the login counts based on 15 minute time intervals, then plotted the resulting time series. The plot is shown below.



This plot is difficult to read because of the volume of data, so I tried several other approaches to find patterns in the data. First, I grouped the data by day of the month, and calculated the number of logins per day. The resulting plot is shown below, and shows a clear decreasing trend in the number of logins. So it is highest at the beginning of the month, and decreased toward the end of the month.



I then grouped the data by hour of the day, and plotted the results. We can see that the highest number of logins occurs around midnight and around noon.



Part 2: Experiment and Metrics Design

To measure the success of the experiment, I would use a sample of data collected prior to the toll reimbursement program, and compare it to data collected after the toll reimbursement is implemented. If possible, I would collect data on each driver, including how many times they crossed the toll bridge. I would then calculate the average number of toll bridge crossings per driver for each week. This would be the key measure of success of the experiment.

I would then conduct a hypothesis test to determine whether the average number of bridge crossings after the toll reimbursement is significantly higher than the average number of bridge crossings before the toll reimbursement. If the difference is statistically significant, I would conclude that the program is effective. I would consider the difference significant if the amount of increased revenue was enough to offset the amount of toll reimbursed.

I would also split the data into groups according to time and day, and look at the difference for night trips, day trips, weekday trips, and weekend trips. It is possible that the program would be more effective only at night, or only on the weekends. This would help the company to maximize profits, because if the program is effective on weekdays but not on weekends, they should only reimburse toll for weekday trips.

I would present the results to the company, and recommend that they reimburse the toll if the experiment showed that the program was effective. If the results showed that it was effective only on certain days or at certain times, I would recommend that they reimburse the toll only for those days or times.

Part 3: Predictive Modeling

After inspecting the data, I noticed that the dates were formatted as strings, so I converted these to datetime objects so that they could be used in the model. I converted all other values to type integer or float as needed. I determined the proportion of users retained by extracting the data where the most recent trip was within the last 30 days (month 6), and dividing this by the total number of users. Just over 36% of the observed users were retained.

To build a predictive model, I imported the data into a dataframe, and added a column of ones and zeros indicating whether the user was active in their sixth month on the system. The dataset had some missing values, so I decided to use the LightGBM machine learning algorithm, since it can handle null values. I first calculated the correlation between all variables, and found that there was a strong correlation between surge_pct and avg_surge, so I used PCA to reduce these two features into one.

I split the data into training and test sets and fit the model. I tested the model on the test data, and it predicted rider retention correctly about 77% of the time. To ensure that the model was not overfitting, I computed the accuracy score on the training data, and it was almost the same as on the test data.

It is likely that the model could be improved by collecting more data, especially over a longer period of time. However, this gives a good indication of what the rider retention will be. It may also be useful to investigate the individual features more, and look at which ones are the best indicators of retention. This would allow Ultimate to devise strategies to improve rider retention and increase their profits.