

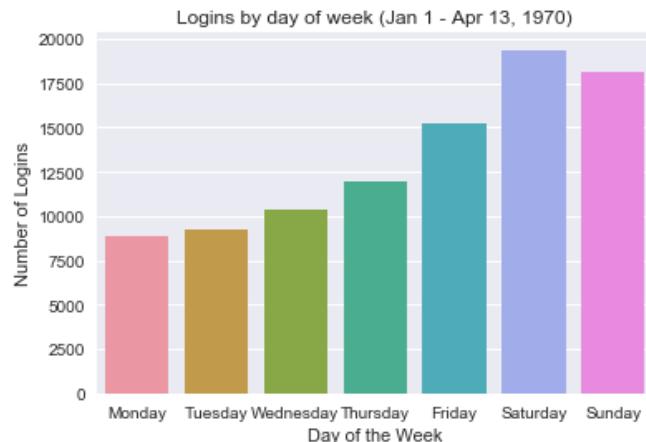
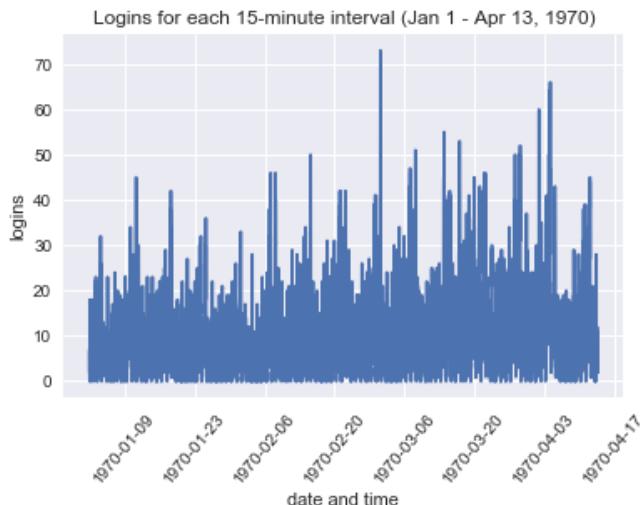
Ultimate Take Home Challenge

Step 1: EDA

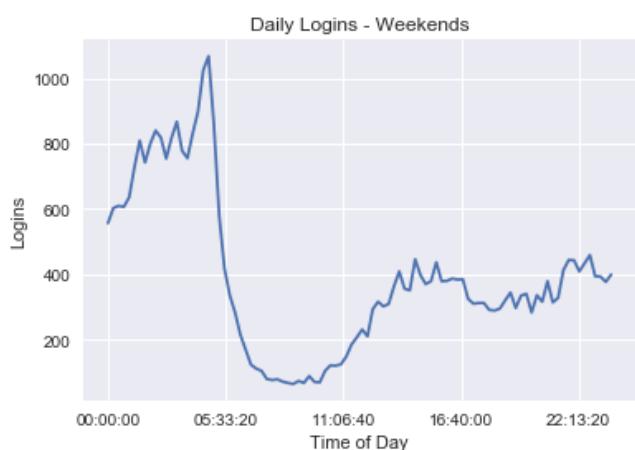
The logins.json file contains (simulated) timestamps of user logins in a particular geographic location. Aggregate these login counts based on 15-minute time intervals, and visualize and describe the resulting time series of login counts in ways that best characterize the underlying patterns of the demand. Please report/illustrate important features of the demand, such as daily cycles. If there are data quality issues, please report them.

From this data we can gain the following insights on user demand:

1. There are more logins on the weekend than on weekdays. Logins are lowest on Mondays and steadily climb to a peak, and then dip down again (but only slightly) on Sunday.



2. On weekends, there are a high number of logins from 12AM to 4AM, with a dip in the early hours of the morning, and are fairly steady from 1PM onward, whereas on weekdays logins are high around lunchtime and in the evening (~8PM to midnight).



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. What would you choose as the key measure of success of this experiment in encouraging driver partners to serve both cities, and why would you choose this metric?
2. Describe a practical experiment you would design to compare the effectiveness of the proposed change in relation to the key measure of success. Please provide details on:
 - a. how you will implement the experiment
 - b. what statistical test(s) you will conduct to verify the significance of the observation
 - c. how you would interpret the results and provide recommendations to the city operations team along with any caveats.

Answers:

1. Ultimately, we are looking to see if the company will make more money if they open up accessibility to both cities. The company might make more money as a result of this change because there will be more driver availability and lower customer wait time. Therefore, the key metric of success that I would choose to measure in this experiment is the total profit of all drivers after the reimbursement of their toll costs. It is important to measure profit after the reimbursement, because if the drivers make about the same amount of money as before, the company will be making less money than before due to the increased cost of drivers who travelled between cities.
2. I would test whether or not total profits (after toll reimbursement) would rise or fall following the change in company policy to reimburse drivers for the toll costs. In order to do so, I would conduct an A/B test, where one group of drivers was reimbursed for their travel costs and another group of drivers would continue as normal. After several months of tracking profits in both groups, I would use a t-test to test for significance at 95% confidence ($p = 0.05$) a null hypothesis that the total profits (after the cost of the toll reimbursement for one group) of the two groups of drivers is equal. If I get a $p \leq 0.05$, I would reject this hypothesis, and find that one group is performing better than the other. If the group that is getting reimbursed for tolls is performing better, I would advise Ultimate to make this their standard policy. However, if the group that carried on with business as usual is performing better, or if I find a p-value greater than 0.05 in my t-test, I would advise the company not to change anything.

Part 3: Predictive Modeling

Ultimate is interested in predicting rider retention. To help explore this question, we have provided a sample dataset of a cohort of users who signed up for an Ultimate account in January 2014. The data was pulled several months later; we consider a user retained if they were “active” (i.e. took a trip) in the preceding 30 days. We would like you to use this data set to help understand what factors are the best predictors for retention, and offer suggestions to operationalize those insights to help Ultimate. The data is in the attached file `ultimate_data_challenge.json`. See below for a detailed description of the dataset. Please include any code you wrote for the analysis and delete the dataset when you have finished with the challenge.

1. Perform any cleaning, exploratory analysis, and/or visualizations to use the provided data for this analysis (a few sentences/plots describing your approach will suffice). What fraction of the observed users were retained?
2. Build a predictive model to help Ultimate determine whether or not a user will be active in their 6th month on the system. Discuss why you chose your approach, what alternatives you considered, and any concerns you have. How valid is your model? Include any key indicators of model performance.
3. Briefly discuss how Ultimate might leverage the insights gained from the model to improve its longterm rider retention (again, a few sentences will suffice).

Answers:

1. 37.6% of observed users were retained. Calculated by finding the most recent ride and subtracting thirty days, finding how many users had taken a ride in that timeframe, and dividing by the total number of users.
 - a. *Data Cleaning:* There were null values for phone type, average rating of driver, and average rating for driver. I saw that iPhones were by far the most prevalent phone use type, so I filled this in for missing phone values. I imputed average ratings by gender and of the driver with the mean values of their respective columns.
 - b. *Data Visualizations:* I plotted percent of customers retained against various data metrics, including phone type, city, and if they are an ultimate black user.



2. I started with an SVC classifier, since there are relatively few data points. This had 76% accuracy. There are a relatively low number of false positives, which will be helpful in not over-predicting retention.
 - a. True Positives: 2315
 - True Negatives: 5337
 - False Positives: 869
 - False Negatives: 1479
3. I would make the following recommendations:
 - a. While the most rides take place in Winterfell, King's Landing City has the highest retention rate. I would suggest attempting to increase operations in King's Landing City and/or investigating what features (qualities in drivers or customers) contribute to retention to see if adjustments can be made in the other two cities.
 - b. Retention rates are way higher among iPhone users. I recommend investigating Android app design to see if bugs/poor app design are contributing to lower retention rates.
 - c. Ultimate Black Users have way higher retention rates. I would continue to push the program or add more perks to this program to keep people using the service.