Data Analysis Interview Challenge

This is your chance to wow us with creative and rigorous solutions! Please include your code at

the end of your submission, or in a separate file. We also accept incomplete solutions.

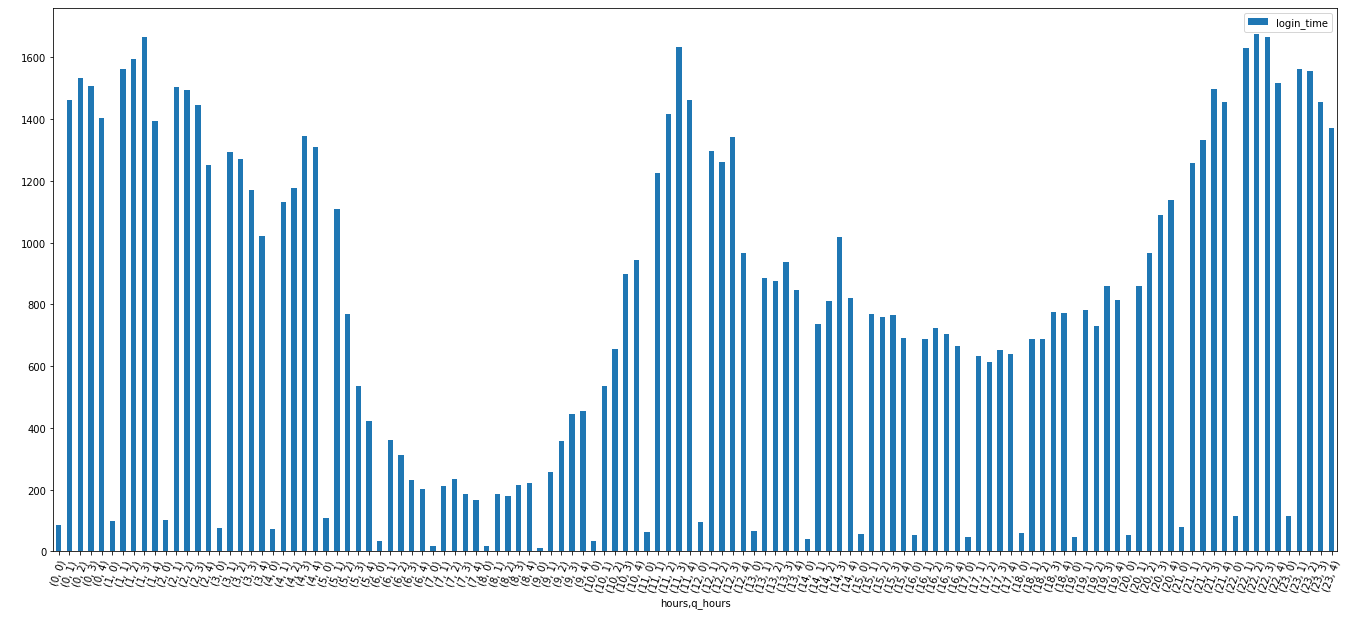
Part 1 ‑ Exploratory data analysis

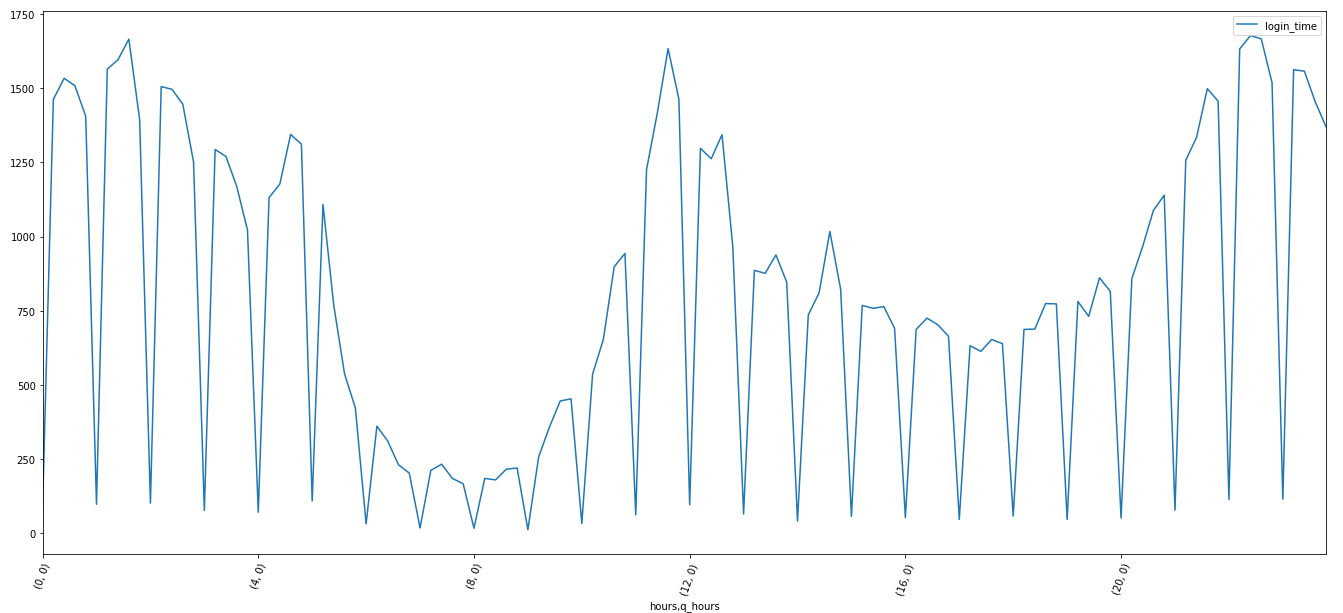
The attached *logins.json* file contains (simulated) timestamps of user logins in a particular

geographic location. Aggregate these login counts based on 15minute

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.

ANSWER:





The basis of a scientific research should be clear. The problem here is discernible time zone. Being asked to do a trend analysis without knowing the time zone is not going result in false findings. Here we can see aggregated login data, where the pertaining data belongs to rider or driver is unknown. However, it would be much more likely to have peaks of demand (and therefore supply – if data is for drivers) during rush hours. The charts have 3 peak points – midnight, noon and late evening- and a low plateau at 8 am. This contradicts natural flow of life. The problem, most probably is time zone, which could be UTC. It is mandatory to contact the information provider at this point and verify. The other problem is the first quarter of each hour and it is quite apparent that the root of the problem is the data itself. Distribution of login times (as a parameter of hour, minute) should have been evenly distributed. While an assertion about busy hours (rush hours, friday night etc) is acceptable, the low points at each first quarter of the hour is inexplicable. Also login times start with 1970, which points out an error.

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.

ANSWER:

There are two aspects of this experiment. First, I must make sure that the experiment results are free of errors, and overall even though not specified, profits should be increasing. I may get statistically correct results, and if the reimbursed toll rates are greater than the profit coming from the new method of Ultimate, it means the company is losing money. So, unless there is another aspect of changing the way we operate (such as having a plan for the future) we should be caring about both aforementioned principles.

1. Key measure of success could be:
   1. Increase in usage of toll bridge
   2. Increase in Gotham driver in Metropolis (or vice versa)
   3. Increase in Gotham driver having a Metropolis rider (or vice versa)
   4. Geolocation – start and end points of rides

etc.

1. Practical experiment can be done in many ways. We need to see probabilities distribution of before and after the experiment, or have a test and control group. For example, supposing we have all preset conditions for the central limit theorem, we select samples with replacement from rides (bootstrap), and see the probability of a ride to include bridge toll. Then we plot distribution charts (histograms). We can then for example take the mean of the second distribution and see where that value is represented in the first distribution chart. In order to see the statistical significance we can calculate the pvalue, or make a 2 sample z test with bootstrap.
2. Desired end state comes into play after the experiment. If the hypothesis is proven after the experiment and toll reimbursement decreases driver exclusion, we must then focus on desired end state (i.e. how does that reflect to customer satisfaction, profitability etc.), because the test results may be proving the test hypothesis without doubt, but it may not take the company anywhere. For example, hypothesis is toll reimbursement provided riders to be active at both sides, but if that did not result in increasing the number of rides etc, it may not mean much.

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 long term rider retention (again, a few sentences will suffice).

ANSWER:

1. Fraction of retained riders if about 37 percent.
2. Random Forest, Logistic regression and Catboost predictive models have been built. Through feature engineering some feature columns were created in order to increase model skill. Categorical features and dummies were created. The retention as described here have been formulated and projected in the dataset. So we can think, if a rider that joined 6 months ago is retained today and we create a prediction model based on this training data, I can predict if another rider signing up today will be active 6 months later or not. It is really hard to predict human behavior, but based on customer profile (phone model, city that he lives in, whether he is an Ultimate commuter, weekend user etc) we can predict retention. Accuracy is 78%, given the data. Feature importance in retention have also been plotted.
3. Astapor has low retention rate, whereas King’s Landing retain rate is three times as much. The riders that have high weekday percentage (daily commuters) are much likely to be retained. Surge percent increase also increases retention rate.