Part 1 - Exploratory Data Analysis

The original json file had to be converted to DateTimeIndex in order to invoke the time series analysis capabilities of Pandas. I was then able to resample it in many different ways to gain insights on the behaviours of users.

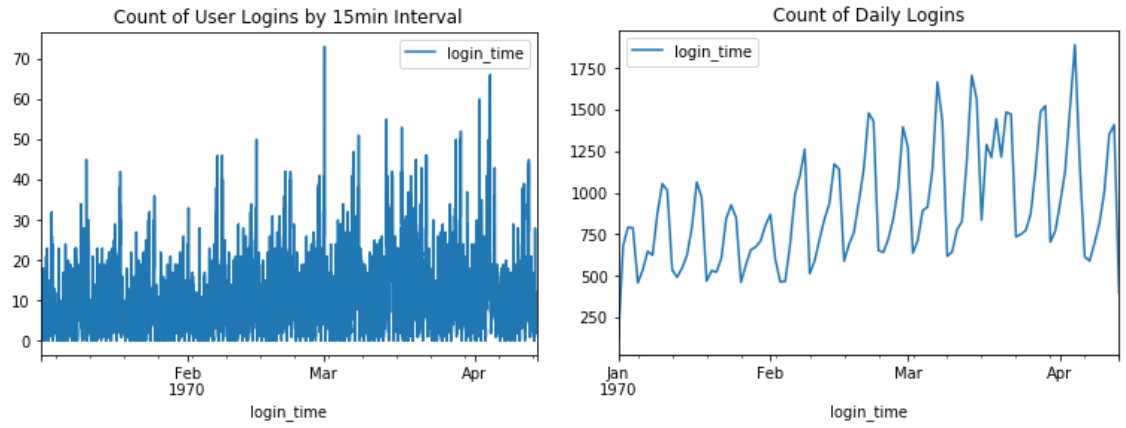

Fig 1 - Resampling for Meaning

Initial downsampling to 15min intervals was not meaningful. I then opted to look at it on a daily, weekly, monthly basis for an interpretation. We see that generally usership is increasing over the history of the data.

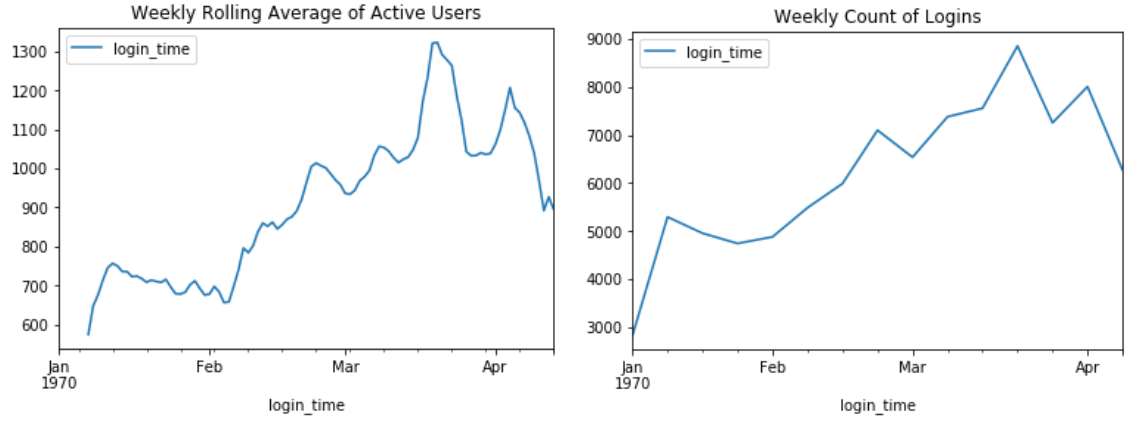

Fig 2 - Weekly Behaviours

Whether we review the data from a weekly rolling average or weekly count we see the increasing usership and potentially that more people ride in the spring versus winter. It is unfortunate that the data only runs for four months, otherwise we could do more seasonality analysis.

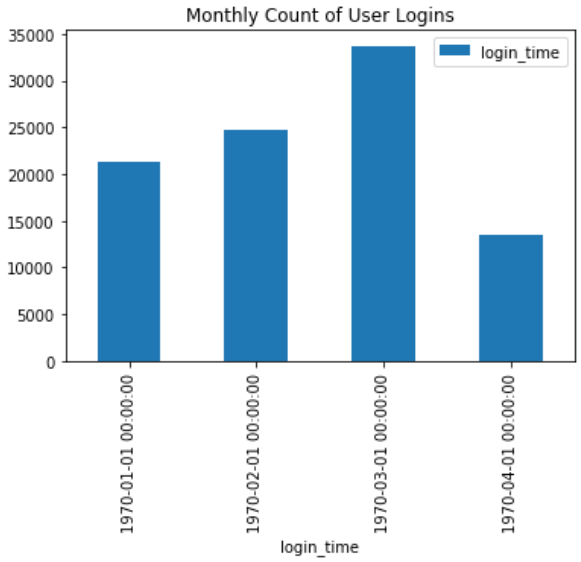

Fig 3 - Monthly Activity

We see the bar chart above as an interpretation of monthly active logins. Users are logging in more often with the passing of time. Again, short term data doesn't allow for much in the way of conclusions.

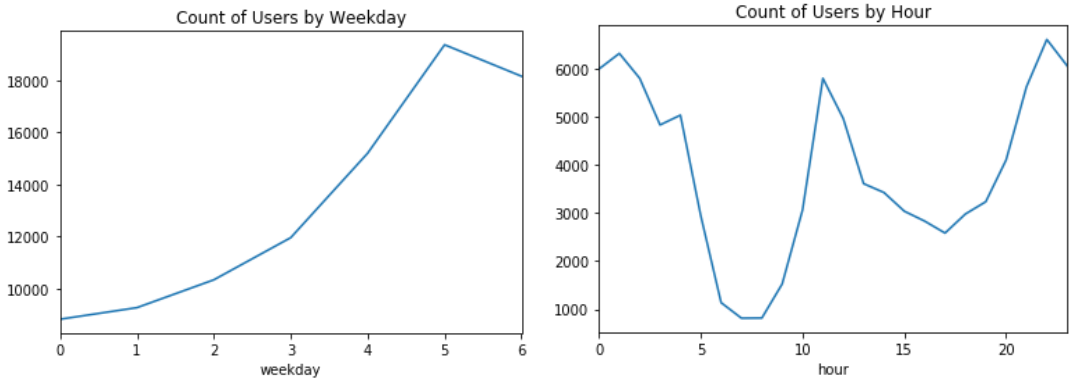

Fig 4 - Highlights of the EDA

Two major findings are that usership is much higher on weekends and there are two busy times during the day, one midday and one in the evening.

Part 2 - Experiment and Metrics Design

Neighboring cities of Gotham and Metropolis have circadian rhythms that are out of sync with each other. This means the ridership will be offset by roughly twelve hours. The geography is such that between these two cities there is a toll bridge. The bridge is believed to limit cross town traffic because the drivers are forced to pay the toll out of pocket. An experiment is to be designed with the goal of increased driver availability in both cities.

The key metrics determining success of this experiment would involve:

- the quantity paid out to the drivers on weekends

- pickup/dropoff location geography

I would choose this metric because the circadian rhythm effect may influence the test. Natural demand changes in the different locations during specific times of the day/week. For this reason I would focus the metric on the toll during the weekend since it is representative of true cross town traffic when both markets are active.

I would include the latitude and longitude of pickups to study proximity to the bridge. I would aim to uncover if a driver is performing quick trips from one side of the bridge to the other versus making longer trips from the opposite side of town.

Statistical test for significance would involve the two samples t test based upon when the test is running and when it isn't. The sample size would be greater than 50 rides to ensure we have enough data to satisfy the Central Limit Theorem. I would test the effect of compensating the drivers for their payments to the toll. This would allow us to see if the average number of payments made go up or down when the company is paying for it versus the drivers.

I would interpret the results using a hypothesis test and a confidence interval of 95%. My test would consider the t-statistic as we would not be sampling the entire population of rides which would include the weekdays. Ensuring that my null hypothesis - test has no impact - is rejected would require a one tail test meaning the statistic value would have to be greater than the alpha value.

Part 3 - Predictive Modeling

The data provided required some cleaning in order to successfully model. The following steps were taken to order and organize the data appropriately.

- Converted dates to datetime
- Engineered the 'retained' field by testing if most recent ride occurred in last 30days
- Manipulated the datetime fields for meaningful features like weekday and month
- Engineered the weekend field by testing if weekday was > 4
- Created dummy variables for city, phone type and ultimate black user (yes/no)
- Tested for null values and filled with adjacent
- Dropped fields that could not be used in the Random Forest model, i.e. text

Ultimate is currently tracking a retention factor of 36% on the observed users. This was calculated by first testing if the last trip was within 30 days of the most recent date in the dataset.

The predictive model that I used was a Random Forest. This model is considerably adaptable to all types of data in classification problems. It performed reasonably well for a first pass.

I tested the correlation between fields by using the pandas .corr() function. It revealed that there were some instances of high correlation that might yield issues of multicollinearity. For example, surge percentage and average surge showed a correlation of 0.79. This would typically be considered very high and present some complications but the fields are obviously dependent on each other. As such it was left as is in order to retain their relationship with the response variable.

Conversely, the 'month' field was used to calculate if a user had been retained. Therefore it was too highly correlated with the response variable to provide any real meaning and was dropped from the model.

The feature importance of the model is as follows:

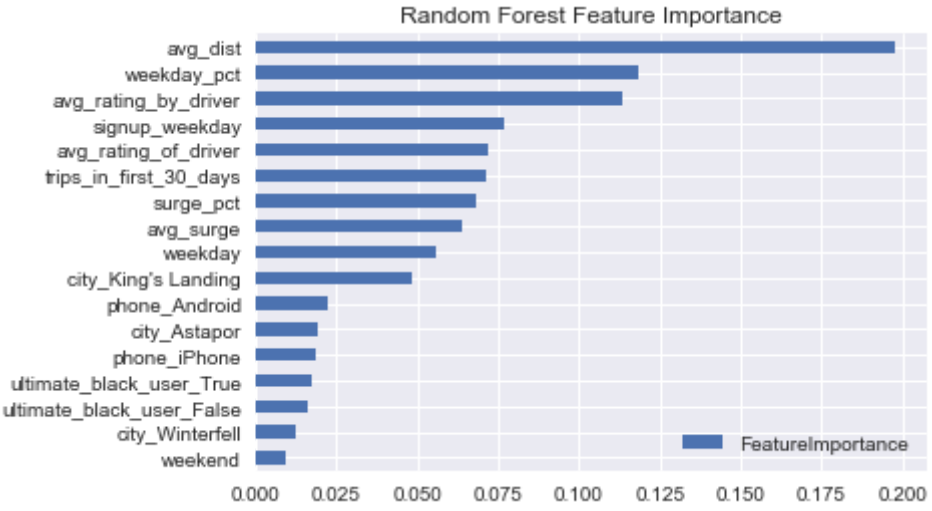

Fig 5 - Feature Importance of Random Forest Model

Intriguing that longer trips and weekday riders correspond to higher rates of retention. It is also telling that if the user is rated well by the driver they are more likely to be retained. Not surprisingly if the user enjoys their experience with the drive they are also more likely to reuse the service. Some fields can safely be ignored such as 'signup weekday' however we cannot ignore that there seems to be significant retention if the user makes use of the service within 30 days of signing up.

The model performed with the following accuracy:

	Precision	Recall	F1 - Score	Support
0	0.79	0.85	0.82	9523
1	0.70	0.60	0.64	5477
avg/total	0.75	0.76	0.75	15000

While it is not supremely accurate it does provide us with a good start in analysing areas of further investigation. Some concerns with the model are whether or not it is adequately fit given the small amount of data. This model may have a hard time generalizing unseen data and may benefit from K-Folds optimization. As a first pass, the benefits of random forest ensemble learning were considered adequate for the purposes of this challenge.

Ultimate might choose to follow up with the following insights:

- What can be done to improve the user experience, particularly the ratings given by the drivers? Users receiving good feedback are more likely to ride again.
- The first 30 days are crucial, can Ultimate provide incentives to ride early in a subscription? I.e. effect of free rides/credits
- Can anything like special rates for short distance trips be implemented? It appears that users who don't travel far are less likely to return to the service.

Ultimate might also consider using a Bayesian model that would predict the likelihood of retention given a user's characteristics.