**Data Analysis Interview Challenge**

Part 1 ‑ Exploratory data analysis:

Some interesting conclusions on the logins data set was that the dataset goes from January 1 to April 13, 1970 and all of the days of the week are represented but with Saturday represented the most at 19,377 logins and Monday represented the least at 8,823 logins. The shape of the dataset consists of 93,142 rows and one column that is a pandas datetime series object. The time-series plot showed that March had the highest frequency of 15 minutes time interval login counts. Plotting out the dataset showed an overall increasing trend of login counts from January 1970 to April 1970.

Part 2 ‑ Experiment and metrics design:

This could be a regression problem. Success would mean that the null hypothesis which is that driving in both cities favors Ultimate drivers is true and the alternate hypothesis would be that adding both cities does not favor drivers. Another hypothesis could be taking a difference of means between adding both cities vs only one = 0 and the alternate hypothesis would be that there is a statistical difference between driving in one city vs driving in both. Further data would be needed to implement frequentist testing. Implementation thus could be a gradient boosting regressor with best features selected. Data would need to be cleaned and pre-processed prior to beginning machine learning.[¶](http://localhost:8888/notebooks/Desktop/take_home_challenges/take_home_challenge1.ipynb#This-could-be-a-regression-problem.--Success-would-mean-that-the-null-hypothesis-which-is-that-driving-in-both-cities-favors-Ultimate-drivers-is-true-and-the-alternate-hypothesis-would-be-that-adding-both-cities-does-not-favor-drivers.--Another-hypothesis-could-be-taking-a-difference-of-means-between-adding-both-cities-vs-only-one-=-0-and-the-alternate-hypothesis-would-be-that-there-is-a-statistical-difference-between-driving-in-one-city-vs-driving-in-both.-Further-data-would-be-needed-to-implement-frequentist-testing.--Implementation-thus-could-be-a-gradient-boosting-regressor-with-best-features-selected.--Data-would-need-to-be-cleaned-and-pre-processed-prior-to-beginning-machine-learning.)

 60% of Ultimate business occurs during weekdays which implies 40% occurs on the weekends. This indicates that more folks use Ultimate on weekends then on weekdays since there is an average daily usage of 12% on weekdays and an average daily usage of 20% on weekends. The other interesting finding is that whereas the max distance is 160 miles, the average Ultimate ride is only 5.79 miles. Since revenue is not part of this dataset, distance will be used as a metric since driving more distance would increase revenue for the Ultimate drivers. The last interesting finding is that the ratings for ultimate drivers is relatively high and thus one concern the Ultimate drivers may have is if increasing distance travelled would affect their ratings. Finally, only an average of 2 trips are taken within the first 30 days of signing up with Ultimate thus perhaps another incentive would be that if there were no toll costs and drivers could thus travel easily from Gotham to Metropolis, the average number of trips taken in the first 30 days could potentially increase.

All sign-ups occurred in the month of January. Also, the last trip was taken on July 1, 2014 while first trip was taken on Jan 1, 2014. Most sign-ups occurred on Sundays and the least sign-ups occurred on Mondays. Also the majority of Ultimate users live in Winterfell and finally, 38% of users are Ultimate\_Black\_Users(UBU).

Although the machine learning model was very weak the feature selections were valuable and the conclusions that were arrived at were that an increase in weekday usage, an increase in UBU, and an increase in trips taken in the first 30 days will also increase the average distance the ultimate driver drives and thereby increase the driver's total revenue. In addition, the features of good driver ratings and the driver's surge percentage will also greatly influence the average distance an ultimate driver drives. Further data would be needed in this dataset to show that reimbursing tolls would increase weekday usage and thus be an incentive to ultimate drivers.

Part 3 ‑ Predictive modeling:

This is a classification problem and the trends seen are:1.) Heatmap shows correlation between average surge and surge percentage. 2.) Heatmap shows negative correlation between surge and weekday percent which indicates surge occurs on weekends in accordance to supply and demand. 2.) Heatmap shows a correlation between retained users and UBU. 3.) Heatmap shows riders taking trips on weekdays are more likely to be retained users. 4.) Heatmap shows a negative correlation between average distance and trips in first 30 days which indicates retained users that are taking more ultimate trips are taking short distance trips. This last trend may indicate why the previous machine learning model was weak since increasing average distance does not correlate with more retained users.

The majority of sign-ups took place on 1/18 and 1/25. Those riders who had trips in the first 30 days (retained users), were riders who had their last trip in March and in July. This shows that there are many retained users since the majority of last trip dates are in July. What this indicates is that there are many users that were lost in March and thus only were Ultimate riders for January and February. 30% of all users are not retained users. And the majority of those not retained users fall in the category of users whose rides took place in January and February but ended in March.

The plot shows that retained users are correlated with UBU and that retained users are correlated with weekday usage.

There is also a negative correlation with distance indicating that most retained users are taking short trips. Some theories on this could be that retained users are taking short trips to and from work on weekdays.

There is also a negative correlation with surge pricing indicating that there are less retained users when surge prices are high. In our earlier data exploration above we saw that the majority of rides take place on the weekends (20% per day on weekends vs 12% per day on weekdays). We also saw that surge prices are higher on weekends.

It seems that since the majority of retained users are on weekdays, reimbursing drivers for toll fees could potentially increase the weekday users even further. By reimbursing the toll fees for Ultimate drivers, we may also see weekday customers riding longer distances instead of only attracting those employees traveling short distances. Since there is a negative correlation between surge prices and retained users and a positive correlation between surge prices and the weekend as well as a positive correlation between UBU and retained users, one incentive that could increase weekend retained users would be to give discounts on surge prices to UBU.

In comparing the random forest classifier with the deep learning keras algorithm in determining retained users, the random forest classifier was the best model at 89% test accuracy, 89% precision, recall and F1 score.