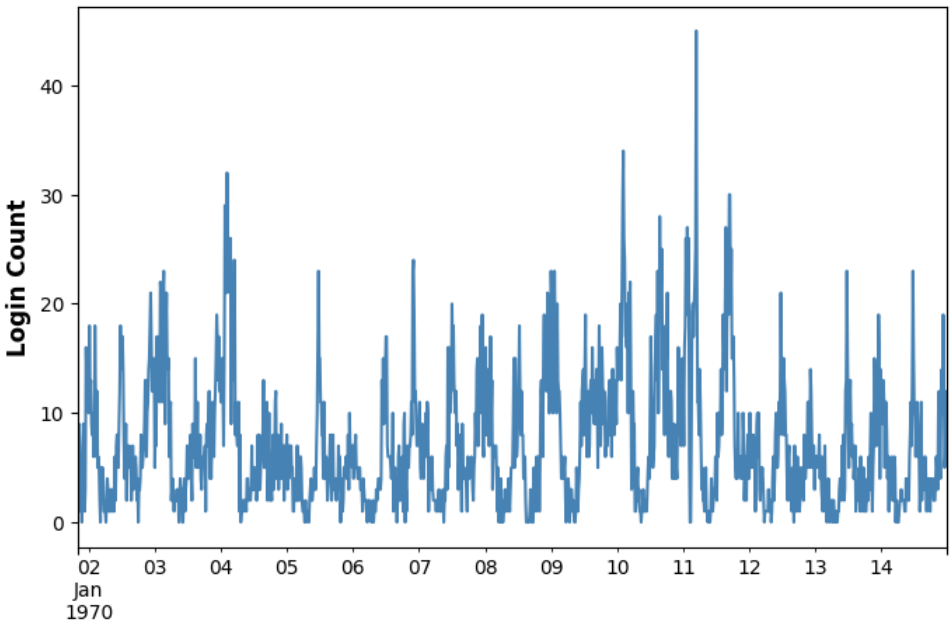
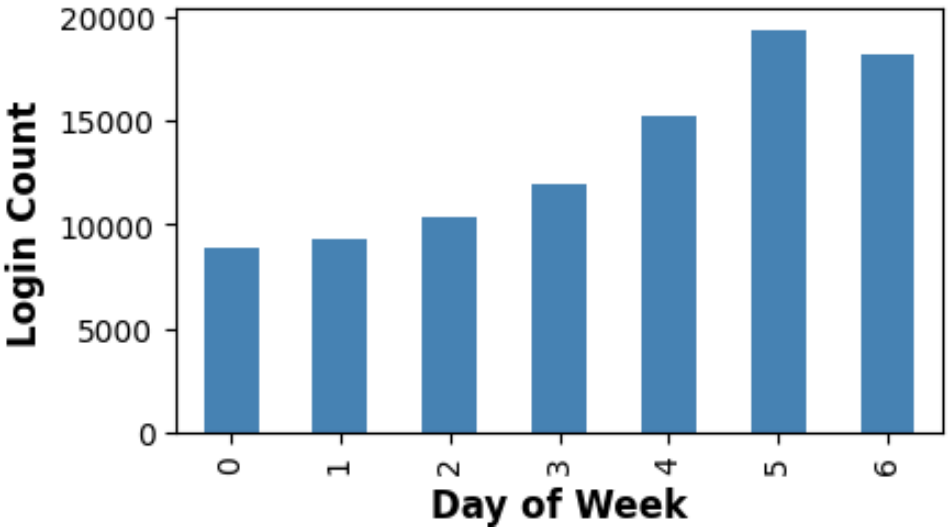
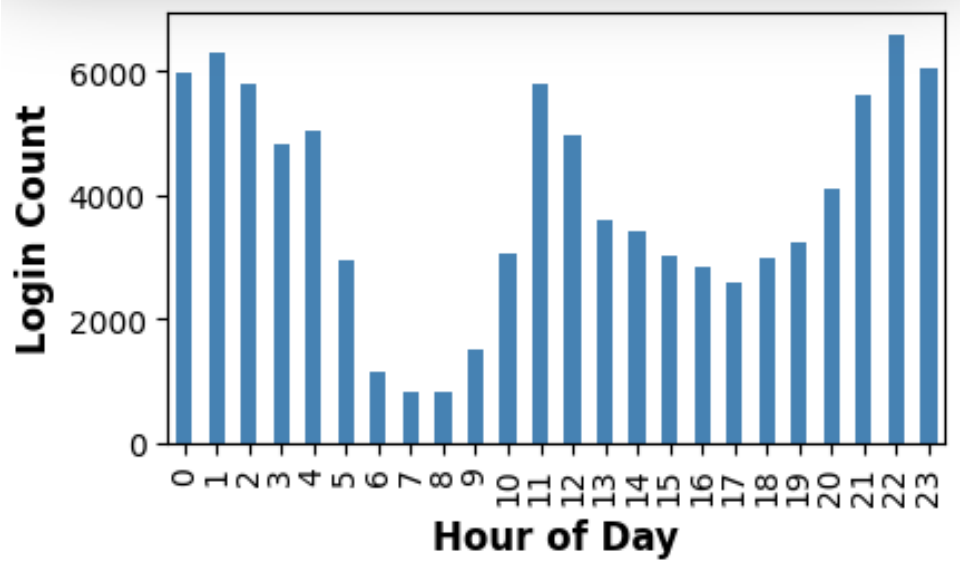
ULTIMATE TECHNOLOGIES - CASE STUDY
A supporting Jupyter notebook is in this same folder (Ultimate_Technologies.ipynb).

1 EXPLORATORY DATA ANALYSIS - LOGINS.CSV
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.

The login data set contains 93,142 login timestamps with values ranging from from Jan 1, 1970 to mid-April 1970 (04/13/1970). The data has some periodicity with higher counts on weekend days of the week (1 = Mon, 2 = Tues, etc) and higher counts overnight and near mid-day and lower counts on weekdays during earlier morning and mid-afternoon hours.







2 EXPERIMENT AND METRICS DESIGN

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?

Let's abbreviate Gotham with 'G' and Metropolis with 'M' and let (X,Y) denote a fare that starts in city X and ends in city Y. Hence, there are 4 kinds of fares: { (G,G), (G,M), (M,M), (M,G) }. If the goal is to encourage drivers to serve both cities (versus another goal such as to explicitly maximize revenue), I would measure the percent of total fares that are of each type for drivers that have tolls reimbursed versus drivers that do not have tolls reimbursed. My expectation is that drivers with tolls reimbursed would have a more even distribution of fares across types whereas drivers that do not have fares reimbursed would have fares predominately of type (G,G) or (M,M) (but not both) and fewer fares of type (G,M) and (M,G).

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.

Building on my answer to (1), an experiment might be conducted as follows. Randomly select a set of drivers to participate in the experiment and randomly divide the drivers into 'control' and 'treatment' groups in which the control group is not reimbursed for tolls and the treatment group is reimbursed for tolls. Then for several weeks measure the quantities of fares of each of the 4 types described in (1).

A statistical test could be constructed as follows. Define the metric p_mixed as the proportion of fares of types (G,M) or (M,G).
Null Hypothesis: p_mixed is not different between control and treatment groups
Alternative Hypothesis: p_mixed is larger for the treatment group when compared to the control group
A permutation test could be conducted using the difference between p_mixed for control and treatment groups as a test statistic.

An increase in fares of types (G,M) or (M,G) would indicate an increase in drivers serving both cities. One issue is that drivers, with tolls not reimbursed, might predominantly have fares of type (G,G) or (M,M), but not both. If tolls are then reimbursed, these drivers might then have fares with both types (G,G) and (M,M), but still relatively few fares of mixed type, (G,M) or (M,G). A solution to this is, for each driver participating in the experiment, for a period of time before the experiment, measure fare types to assess whether (G,G) or (M,M) is predominant. Then implement the experiment and compare the percent of fares of each type before and after the onset of the experiment.

3 PREDICTIVE MODELING

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?

I loaded the data, set data types for dates and boolean variables, imputed missing data for the average ratings features using the median of available data, one-hot encoded phone and city variables, and I calculated the target variable for retention, retain_flg, as 1 for last_trip_date in prior 30 days (Jun 2, 2014 to Jul 1, 2014).

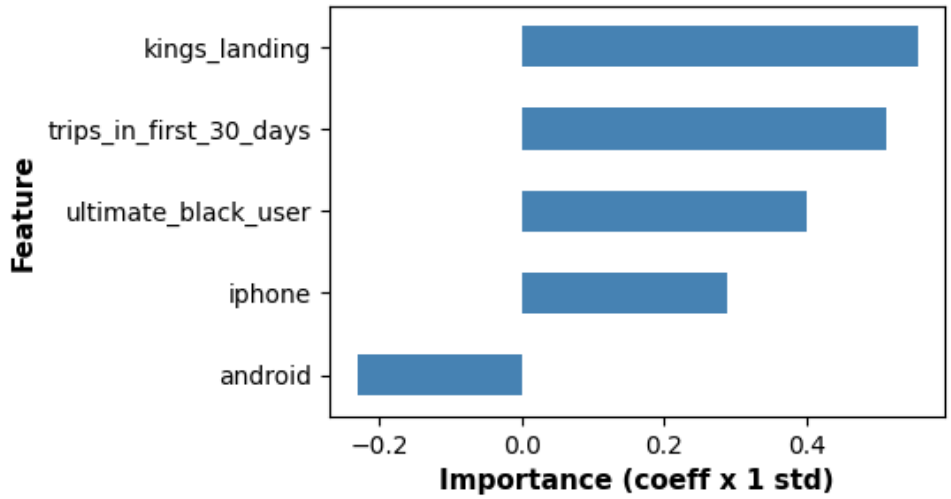
The mean retention rate over all users is 36.6%

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.

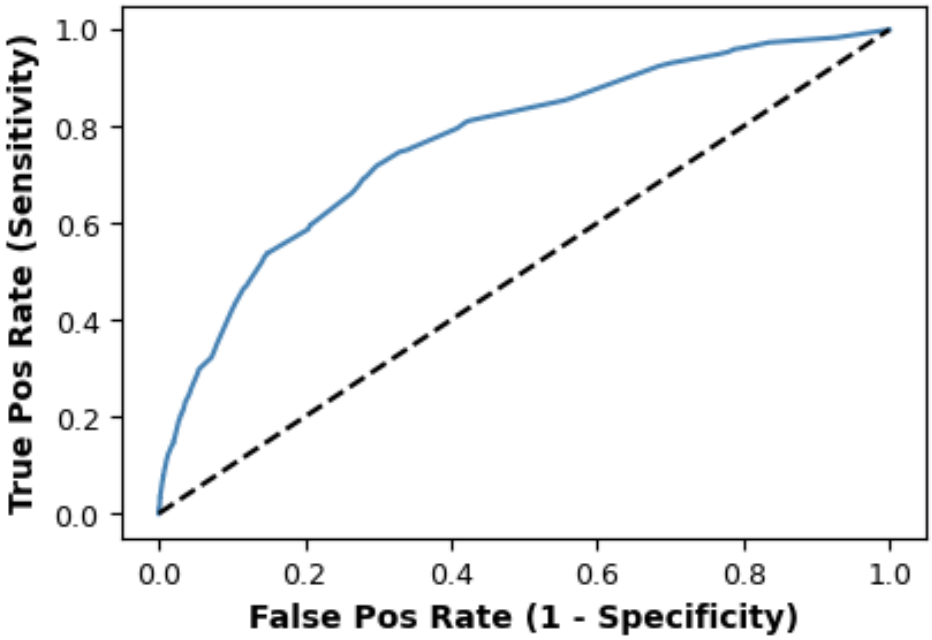
I first explored the data by calculating the mean retention rate by group within each feature variable. I displayed the retention rate in tables for categorical features and in graphs for numerical features. Some of the features exhibit nonlinear relationships with retention rate, hence for these features I added the square of the feature value as a quadratic term and additional feature. I modeled the data using logistic regression because it is a relatively simple model and it appeared to be sufficient for these data.

I performed cross validation with training data on the logistic model using the SelectKBest function in sklearn which identifies the k best features in the data. A model with k = 5 has an accuracy of 72.7% (on the training data) which is essentially the same

as the best model, a 6 parameter model with an accuracy of 72.7% (after rounding). The k = 5 model has accuracy of 71.9% on the test data. The five best features are trips_in_first_30_days, ultimate_black_user, android, iphone, and kings_landing.



The k = 5 model has an AUC score of 75.2% for its ROC curve on the test data, indicating the logistic regression model does better than a random classifier.



Because there are some nonlinearities in the data noted during my exploratory analysis, a decision tree model might yield improved results and this would be an alternative model to consider in further analyses.

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).

Ultimate might target their marketing budget on iPhone users in King's Landing and perhaps incentivize the most positive experience possible in each user's first 30 days perhaps also including the Ultimate Black service during this initial period, maybe as promotion. A decision tree model might also be considered to assess any nonlinear effects of certain features, which might identify further ruse attributes that are important for retention. For both the logistic model and any subsequent models, a review of the importance of different features based on the known modeling results should be discussed with the appropriate business teams as a check on whether the top features match or disagree with expectations from experts in the business.