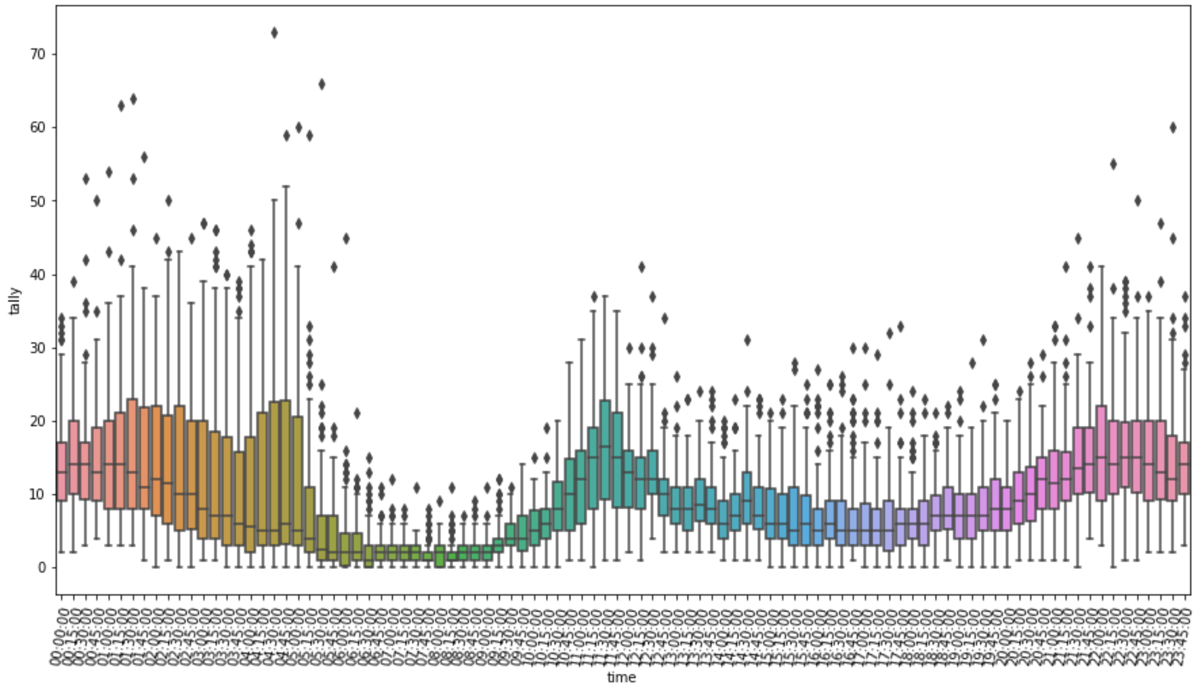
Ultimate Challenge

See jupyter notebooks in github folder for more details.

Part 1 - Exploratory data analysis



From the above box plot we can see that there is a spike in logins from 10:45-13:00 as well as a night spike beginning around 20:30, continuing until 01:00, whereby it begins to decrease until 05:30. The lowest login period is from 05:30-09:00 and there is very little variability in that time as well.

Part 2 - Experiment and metrics design

A number of possible metrics could be used to determine the success of the experiment including:
-number of toll reimbursements
-time between toll payments
-number of tolls into Gotham at night and number of tolls into Metropolis during day
-increased number of reimbursements during week compared to weekends

The most effective metric to use however, would be the following:
-Change in the mean number of rides completed in each city (non-dominant city should go up for each driver and dominant city should stay the same or slightly decrease if the experiment is a success)

To implement the experiment I would tell drivers that the reimbursement program is going to be introduced on a trial basis for 1 month. After the trial period I would perform a one sided t-test on the mean number of rides completed in each city for each driver to determine if there was a significant increase in drives in the drivers' non-dominant city.

Factors to consider:
-Increased number of time available for certain drivers leading to increased means
-Drivers using the program to travel across the toll bridge for free and not complete rides
-Decreased number of rides in non-dominant city on weekends
-Certain drivers only being able to work either during the day or at night not leading to increased rides

Part 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?

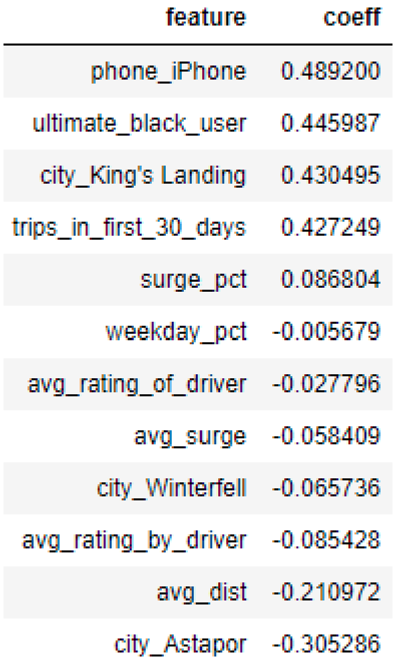
37.61%

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 built two models to determine user retention. I used the features shown in the table below and scaled them before building any models. The first model was a logistic regression model that had 70.9% accuracy, 0.69 precision and 0.55 recall. The second model was a K-nearest neighbors model that had 73.0% accuracy, 0.68 precision and 0.66 recall. Both models could definitely be improved with some hyperparameter tuning as the models used were simply out-of-the-box. While the K-nearest neighbors model performed better, it is less useful for determining which features are the most beneficial and which are the most detrimental to user retention. Therefore the coefficients of the logistic regression model were used to pull valuable insights.

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 can use the feature coefficients from the logistic regression model to determine the most and least beneficial features for user retention. The coefficients with negative values are detrimental and those with positive values are beneficial. For example, the top coefficient is the Iphone feature, meaning Iphone users are much more likely to be retained. This could suggest that the Android app is not as optimized and is leading to less retention among Android users. Next Ultimate could look at the coefficients for the 3 cities. It is clear that King's landing residents are much more likely to be retained and those from Astapor are much less likely. Perhaps Ultimate should consider looking into what reasons Astaporians are not being retained. The other coefficients can be thought about similarly and used to make changes that would hopefully increase retention.



feature	coeff
phone_iPhone	0.489200
ultimate_black_user	0.445987
city_King's Landing	0.430495
trips_in_first_30_days	0.427249
surge_pct	0.086804
weekday_pct	-0.005679
avg_rating_of_driver	-0.027796
avg_surge	-0.058409
city_Winterfell	-0.065736
avg_rating_by_driver	-0.085428
avg_dist	-0.210972
city_Astapor	-0.305286