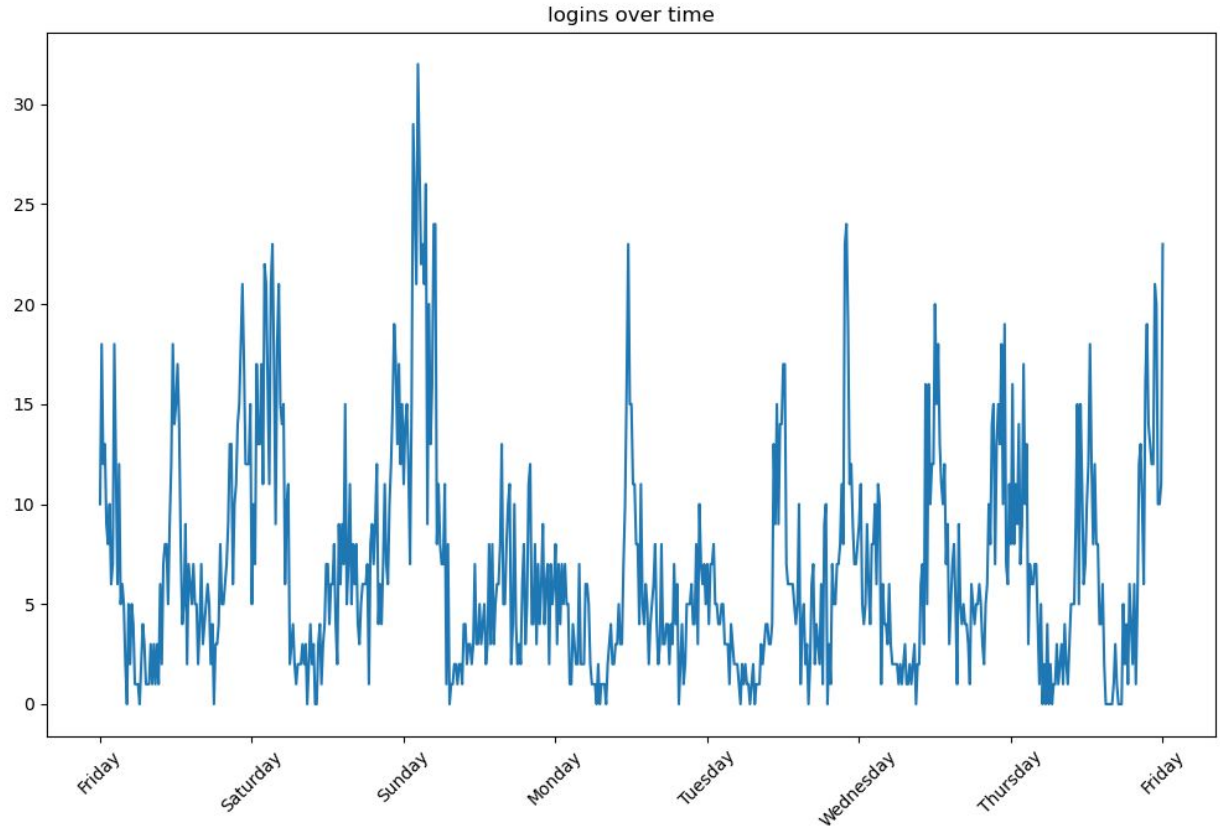
Part 1.

It's easiest to view the patterns when we focus on a particular week in the dataset:



The most pronounced pattern is that there appears to be peaks in the middle of the day and also around midnight. Additionally Weekends have a higher level of logins compared to weekdays.

Part 2.

1. I would choose the percentage of all drivers that serve both cities to measure success. I would define serving both cities as a minimum of their time spent (something like 10%) being in both cities. I chose this metric to make sure the absolute number of drivers is not a factor. That being said the analysis will still have to take into account seasonality of all types (weekends/weekdays/holidays/weather etc.) to ensure they are not confounding factors.
2. a) Assuming it is legally viable, the best experiment would be to split the drivers into two experimental groups, one which will be given the reimbursement and one which will not be. The two groups will have to be randomly assigned and large enough to ensure that they are roughly the same distribution with similar characteristics. The

percentage discussed above will then be defined as a percentage of that entire group rather than all drivers.

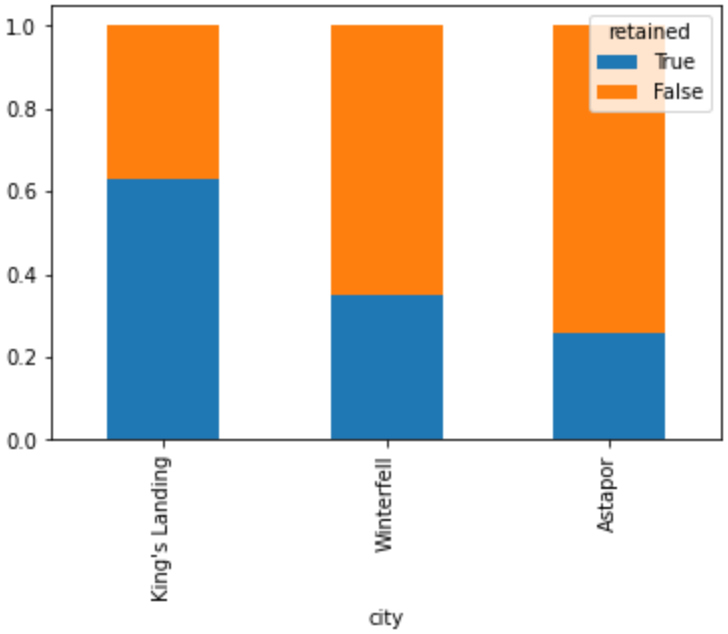
b) I would use the two proportion z-test to see if the two groups differ significantly in the proportion of drivers who service both cities

c) Because the experiment is close to ideal, I would trust the results pretty strongly. However, I would question the stationary of the results depending on how long the experiment is ran for. There are many factors that may impact how the drivers react to a reimbursed toll which may actually change in a longer time frame. I would also question the benefits of having drivers service both cities, as reimbursing the drivers will obviously cost money. Does the increased service actually contribute well to longer term profits.

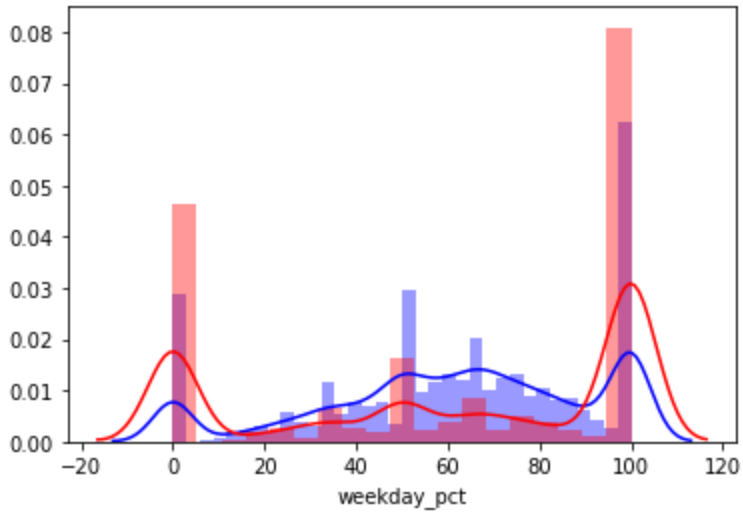
Part 3.

Please view the jupyter notebook in conjunction with these answers.

1. Checking how many missing data there is, it only seems to be affecting “avg_rating_of_driver”, “phone”, and “avg_rating_by_driver”. For the ratings, encoding missing to the arithmetic average seems the most neutral. For phone, I will leave missing as an additional class. For low cardinality categorical features we can view how it varies



with the target:
Here we see that just being in King’s Landing has a much higher retention rate than the other two cities.



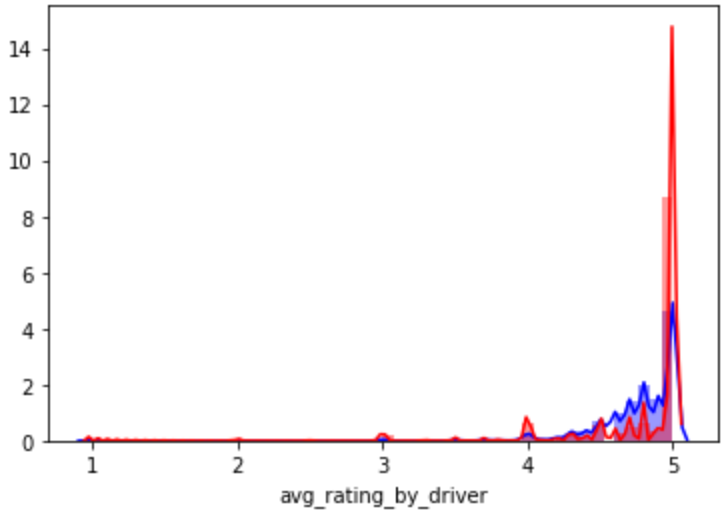
For continuous features, I like to view retained (blue) / not retained (red) distributions for that feature. Here we see the main difference is that retained drivers have more varied weekday_pct. Non-retained users matches either 0 or 100% much more. This is because they ride less so there's a higher chance that they either hit 0, 100, or 50%. If one looks at the average "trips_in_first_30_days" for retained vs. non-retained we can see that clearly: 3.31 vs 1.66. The fraction of overall retained users is 37.6%.

2. It is clear we need to use a classification model. The typical models we would use for this would be Logistic Regression or Decision Tree Ensembles. Since I cannot tell in advance which one would be most successful, I will try them all with a good validation metric. I decided to use ROC-AUC as the positive class is not imbalanced at 37.6%. I will also use 5-fold cross validation to evaluate the metric. My results were as follows:

Model	Mean ROC-AUC
Logistic Regression	.760
Random Forest	.821
Scikit-Learn Gradient Boosted Tree	.852
XGBoost	.851
CatBoost	.860

As a 1% difference is significant, especially as we get closer to a perfect 1.0 AUC, I will go ahead and use CatBoost for this model.

3. The top 4 feature importances are: “city”, “trips_in_first_30_days”, “weekday_pct”, and “avg_rating_by_driver”. The first 3 we discussed briefly earlier, here’s how the distribution looks for average rating by driver:



Here we see a similar phenomenon to weekday_pct in that there are less trips by non-retained users so they tend to clump up more in the integer ratings vs. a mixed rating. So 3 of these features are really a proxy with earlier frequent use. This suggests if Ultimate wants to increase the percentage of retained users, they should heavily market early on in the user’s lifecycle. Perhaps use more reduced fares and discounts for new users. For the city, they are likely to gain higher ROI gaining new users in King’s landing vs. Astapor. However this is dependent on other data not currently available, such as the profitability of each user and how saturated each market is.