Part 1:

We present the general trend of logins between the months of January 1970 and April of 1970. Between midnight and 6 AM, we see incredible variance between the logins. Assuming that Ultimate is still a ride sharing company or taxi service, this could be explained by weekend, evening, or holiday outings such as St. Patrick’s Day trips to the bar or a concert. We then notice this dip to nearly 0 between the hours of 6 and 9 AM. This is most likely due to the fact that most people are either sleeping or starting their days and getting to work. We then notice an increase as we approach noon, most likely due to people going out for lunch. After about noon, we notice slight decline as people are steadily using the service throughout the afternoon and into the early evening. We then see a slight increase as people are looking to go out for the evening.

Part 2:

1. The simplest and most powerful variable that we could use to study the success of any initiative is simply by counting the number of times we reimburse a driver for crossing the given tolls. This might look like developing a piece of software that if the driver crosses a particular threshold, we can automatically reimburse the driver as long as they have a current fare and we can count the number of times we reimburse the drive. We can then examine the average number of reimbursements for drivers. From here, any initiative or incentive we begin, we should see this number increase as drivers travel between cities. The other variable we should also see increase is the average distance the average driver commutes. As it stands currently, it seems that maybe only a select few of the drivers are taking on fares that travel to the opposite city. As more drivers take these longer distance fares, we might expect to see this number increase as well
2. For any experiment, especially something that has to deal with reimbursement, it would seem best to perform a more longitudinal study, as opposed to any cross-sectional study. We don’t want drivers upset that some of them are getting reimbursed while others aren’t. In addition, the longitudinal format has the benefit of satisfying particular conditions to be able to derive a causal explanation.
   1. Assuming this is an app, we could send out a notification to the drivers that for the next month, we will implement this reimbursement policy. Essentially, if drivers cross this particular demarcation line while they have an ongoing fare, they will be reimbursed. Depending upon the company’s stance, we might think to reimburse the toll amount multiplied by the surge rate at a given time. If not, we could always save that idea for a future experiment if we need to further increase drivers traveling between these cities. After the month, we can them examine the data
   2. First, we can compare the percentage of drivers that took on fares that traveled into the opposite cities. Second, we could compare the average distance per driver between the three groups (a: never took fares that cross, b: took fares that crossed after reimbursement c: has always taken fares that cross), this should give us insight as to whether or not that metric of average distance increases as drivers travel into the other city. Finally, we can compare the average number of fares that cross into the city before and after the start of the initiative. For these, we would want to use hypothesis testing for comparing proportions and comparing distributions. We could also use Chi-squared test for the three groups. If we did decide to multiply the reimbursement by the surge multiplier, we could even see if a correlation exists between the amount being reimbursed and the number of trips taken into the opposing city depending upon the distance between Gotham and Metropolis. We would also want to compare results from drivers between cities.
   3. If we notice that drives are more inclined to take on fares if they are being reimbursed for the toll, then we would see statistically significant results for after the initiative began. We want the city operations team to know that as we continue to collect and process additional information, the findings may show a different result or paint a larger picture, so the policy may change such as drivers being to only take on fares that cross tolls to be reimbursed especially if we are multiplying the reimbursement by the surge amount.

Part 3:

1. We first divided the users into active or passive users if they had not taken a ride since June 1. From there, we can explore a little more as the mechanisms underlying retention. First, we see that overall, ultimate has 37.6% retention rate among its users. The biggest differences we seem to be able to discern between active and passive users seems to be that on average active users took 2 more trips in the first 30 days and they are about 2/3 more like to be an ultimate black user. In addition, we see that only 21% of Android users are active as opposed to about 45% of iPhone users who are active. Furthermore, we see the cities of Astapor and Winterfell with 25.5% and 35% of active users respectively, while King’s Landing has about 63% active users.
2. For this model, we decided to construct a Random Forest Classifier seeing as they are one of the strongest classifiers, we have in our tool box. We thought about possibly using a logistic regression, but opted for the random forest since we figured the results would be more robust. We first cleaned up our features, we even converted the categorical features into numeric values in case we wished to pursue the logistic regression to compare. We initialized our random forest model, fit and scored it. Even thought we did split the data into training and test sets, we have evidence of overfitting our model. We see a very high score for our training data but only a decent score for the test data. We should be concerned about not tuning parameters but deploying it anyways. We see that our model is roughly 75% accurate on the test set. We believe that we can do better, but this score for an initial score is fairly decent. We could continue examining the precision and recall, but we may wish to save that for after hyperparameter tuning to get a full picture of how good our model is.
3. Based upon the important features of our model (weekday percent and average distance), ultimate should focus on making sure that directions are the shortest distances and attempt to figure out how to increase individuals using the app during the weekday. To expand further, ultimate should focus on maximizing the algorithm (the bread and butter of the application) used to map out the course for riders. If this is inefficient, people may no longer use the app. It also seems that people who rely on the app for their commutes during the weekday are more likely to stay active users. This means that Ultimate may seek incentive or campaign for users to rely more on the app for their weekday commuting needs. Other insights, we may want to review the user interface for Android users and seek to improve functionality for that platform to enhance user experience. We may wish to also examine if there are any further complications in Astapor or Winterfell. This may require for us to examine what the competition is doing in those cities and figure how ultimate can do it better.