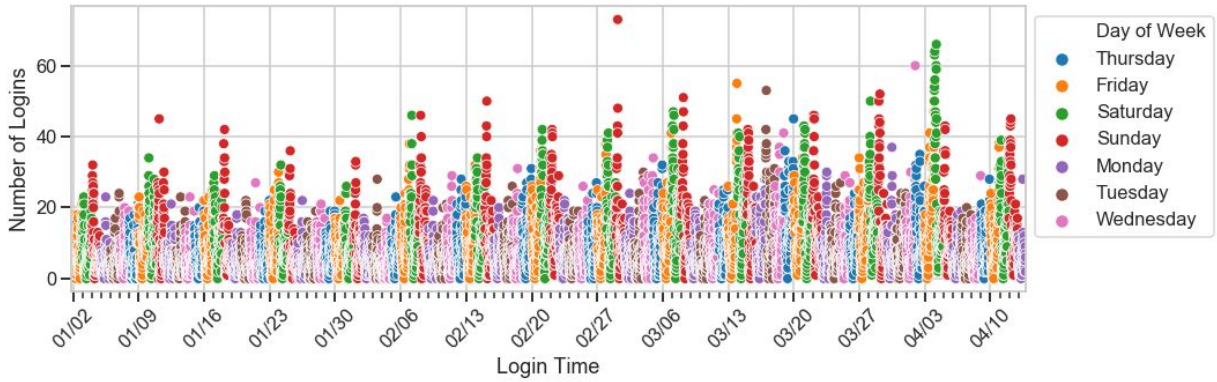
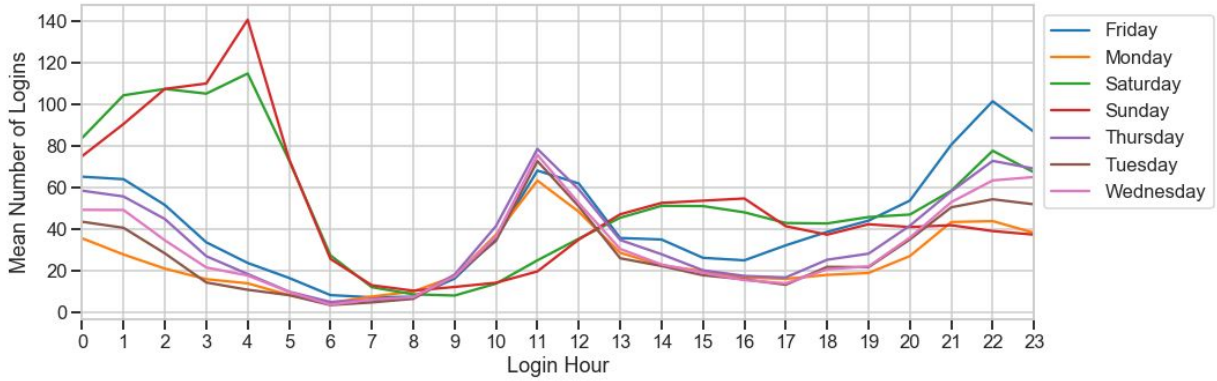
Part I

There are clear daily and hourly trends in login times. The first clear trend is spikes in logins over the weekend, where the highest numbers of logins per 15 minutes occur on Saturdays and Sundays.



Additionally, we can see clear hourly trends that differ by weekday v. weekend day:



On weekends, there are clear spikes between ~9 PM and 4 AM, followed by an increase in the late afternoon; however, weekdays look very different, with a spike around 11 AM followed by a small spike around 10 PM.

Part II

1. One way to measure whether waiving the toll fees encourages drivers to travel between the two cities would be driver pick-ups and drop-offs in different cities, i.e., % of pick-ups that are not the same city as drop-offs.
2.
    a. Information for each ride, via an app (I'd guess), can be collected into a dataset over the months following the toll change

b. We could look at % of cross-city rides a number of ways, e.g., mean % by day, and evaluate significant differences between days, weeks, or even months using a t-test or bootstrap. Additionally, it would be interesting to examine trends by driver and to convert this metric into % of drivers who complete at least 1 cross-city trip per day - this metric might be less dependent on users requiring cross-city trips and more indicative of drivers being willing to make them. Again, I would look for a significant increase in the mean % of drivers per day using a t-test or a bootstrap.
c. I think a 4-week stretch showing significantly higher % of cross-city rides than before the toll changes would be strong evidence that the toll change was impactful; similarly, the % of drivers who complete at least 1 cross-city trip/day could be a secondary metric to further support this change.

Part III

1. 69% of the customers in this dataset were retained.
2. To build a model, I started with a few data cleaning steps, including converting Boolean columns into numerical 1/0s and categorical data into 1/0s. I split the data into a training and test set to allow for model accuracy determination. From there, I tested a few classification models, including Logistic Regression and Random Forest, before determining that a Gradient Boosting Classifier was the most accurate. I started with a Logistic Regression model that includes all available features and found an accuracy (% categorized correctly) of ~75.6%; however, the accuracy if all data were classified as "not retained" is ~74%, so this isn't a great model. I also evaluated Area Under the ROC Curve (AUC), which was only 55% for the initial Logistic Regression, indicating the model only had a 55% chance of distinguishing between the two classes. I next explored feature selection using coefficients from the Logistic Regression to retrain a new Logistic Regression without success - reducing the features to those with the highest coefficients resulted in both a lower accuracy and AUC. After this, I decided to explore two other classification models, including Gradient Boosting and Random Forest. The Gradient Boosting model had a higher accuracy of ~78% and AUC of 64%; the Random Forest model had an accuracy of 75% and AUC of 63%. From these metrics, I chose to move forward with tuning the Gradient Boosting Classifier. I chose to tune a few metrics at once based on what would impact the model the most. I started with a learning rate of 0.1, the default, to keep the model training from being too computationally expensive. I then used GridSearchCV to determine number of boosting stages to perform; max depth (maximum nodes on the tree) and minimum number of samples to split a node; minimum number of samples to define a leaf and maximum features to consider when splitting a node; and the fraction of observations to be selected for each tree (random sampling). The final accuracy was 78.5%, with an AUC of 66%.
3. One of the most useful features for interpreting the model is the importance of the various features in making predictions - this can help determine what influences whether

someone is retained or not. The top 5 features were % of trips taken with surge pricing, average rating by driver, % of trips taken on weekdays, average surge multiplier, and whether the user signed up in King's Landing. Based on some density plots of these features, it looks like retained customers have <25% of their trips as surge priced, an average rating of 4-5 stars, a more even distribution of trips throughout the week (50-75% of trips) as opposed to all weekday or all weekend, lower surge pricing multipliers, and are less likely to be from King's Landing.