This project references the code from the following notebook: https://github.com/clahonta/SB_other_assignments/blob/master/Ultimate%20Challenge/ultimate_challenge/Ultimate%20Technologies.ipynb

Part 1 - Exploratory Data Analysis

Timestamps were first ordered into 15-minute intervals as directed. The pattern of demand was simple and pretty consistent. The weekend days (Fri-Sun) saw the highest login count, with Saturday leading the charge (*n* = 19377). The weekday days began with the lowest usage on Monday (*n* = 8823) and gradually increased each day until Thursday (*n* = 11960). This pattern remained stable over the course of the 3+ months of the data set, with slight variations week over week. Also, there seemed to be a gradual increase in demand for rides, with the monthly total slowly increasing from January (*n* = 21239) to March (*n* = 33723). Though this dataset ends on April 13, the month was on pace to beat March's login total.

Assuming these data are from the United States, the primary outlier was the week of St. Patrick's Day. This year, St. Patrick's Day was on a Tuesday, a Tuesday that had more logins than any day in January and all but three days in February. The weekend before St. Patrick's Day, and the Wed-Fri after all had higher-than-average totals as well. Outliers like this particular week are good to keep in mind as Ultimate seeks to tactically provide more frequent service.

As far as daily cycles, Mon-Fri saw the highest usage midday, around noon. Throughout the week, the evening and night usage would increase with Friday having the highest usage at these times. For Saturday and Sunday, highest usage is actually in the early morning, perhaps indicating usage from the previous night's revelers.

Part 2 - Experiment and metrics design

First, the problem at hand is based on an assumption I am not comfortable making. Since I have no other information but their names, I assume Gotham and Metropolis are both major, densely-populated cities where ridesharing is a common means to get around town. It is reasonable to expect, therefore, that drivers in each of these cities have adapted their lifestyles to the ride-hailing behavior of the riders. I would be curious if, in the event an effective toll reimbursement program was implemented, that such a program would affect the availability of rides in a way that justified the costs of the tolls. It may turn out that driver exclusivity to each city is not a problem.

I will be making the assumption for this project that the only tolls drivers are responsible for are tolls to collect fares from the opposite city. Tolls incurred during the fare are typically charged to riders, and as such a toll compensation designed to change driver behavior should not apply to those. The structure of the test would be a simple A/B test, where the experimental group would be drivers with reimbursed tolls, and the control would be the group without reimbursed tolls.

For this experiment, the key measure of success would be the number of fares in one city that are serviced by drivers from the opposite city, versus the control group. Even without toll reimbursements, drivers from one city could theoretically still service the other city, at personal cost to themselves. Therefore, the effect of covering this cost would best be observed

by measuring this difference. It would also be interesting to see if the number varies at specific times of the day (for example, drivers from Gotham covering more fares in Metropolis during the day, when Metropolis has higher rates of fares). To examine this difference, a *t*-test would be sufficient to determine significance.

If we observed a statistically significant difference in the number of fares served by opposite-city drivers, my next step would be to put that difference in terms of cost/benefit. Whether or not the added fares from reimbursing toll costs generates enough revenue to make up for the costs of the tolls will be the decision maker for the city operations team, so it will be necessary to provide them that context. I would also like to present any differences in rides during surge times, driver/rider ratings, and other metrics as well in case the difference is negligible from a financial perspective, but provides other benefits to Ultimate's business.

Part 3 - Predictive Modeling

Before beginning my explanation, I want to call out an ambiguous description in the assignment. The assignment says: "The data was pulled several months later; we consider a user retained if they were "active" (i.e. took a trip) in the preceding 30 days." This doesn't make sense without knowing the "date" that the information was pulled. Since I do not know the date of this exercise, I will be interpreting a retained user as a user who uses the service more than 30 days after signing up for it.

The first step was to add a column denoting a "retained" user, or a user who used Ultimate after the 30 days following sign up. This was done by subtracting the last trip date from the signup date, and labelling all users as "retained" if their result was greater than 30 days. This was labelled as `1` for "retained" and `0` for "not retained". After performing this transformation, more of the users were retained (73.9%) than were not retained (26.1%). I also performed another transformation where I identified users who had been active in both January and June (25.4%) versus those who dropped off by then (74.6%). These two criteria were created for use in all of the analyses and EDA.

The first interesting finding was that, on average, ratings were lower when they involved "retained" users. That is, retained users had lower average ratings of their drivers, and drivers had lower average ratings of users. While not a necessarily useful metric for predicting retention, it does provide an avenue for future projects. Also, although there seems to be a small difference in average surge multiplier, retained users tend to take a higher percentage of their trips within surge pricing timeslots. A higher percentage of retained users also took an Ultimate Black within their first 30 days.

For modelling, I chose to go with a Logistic Regression algorithm, as it will predict probabilities that a user will be retained. Probabilities will be useful for guiding business decisions for targeting users with a potential to become "retained," and gives the business the ability to change the decision threshold based on the business case. Users with different probabilities of retention could be targeted by different strategies or initiatives. The logistic regression turned out to be an effective algorithm, and investigation into which features have the strongest influence on retention can be used to drive business decisions around retaining users.