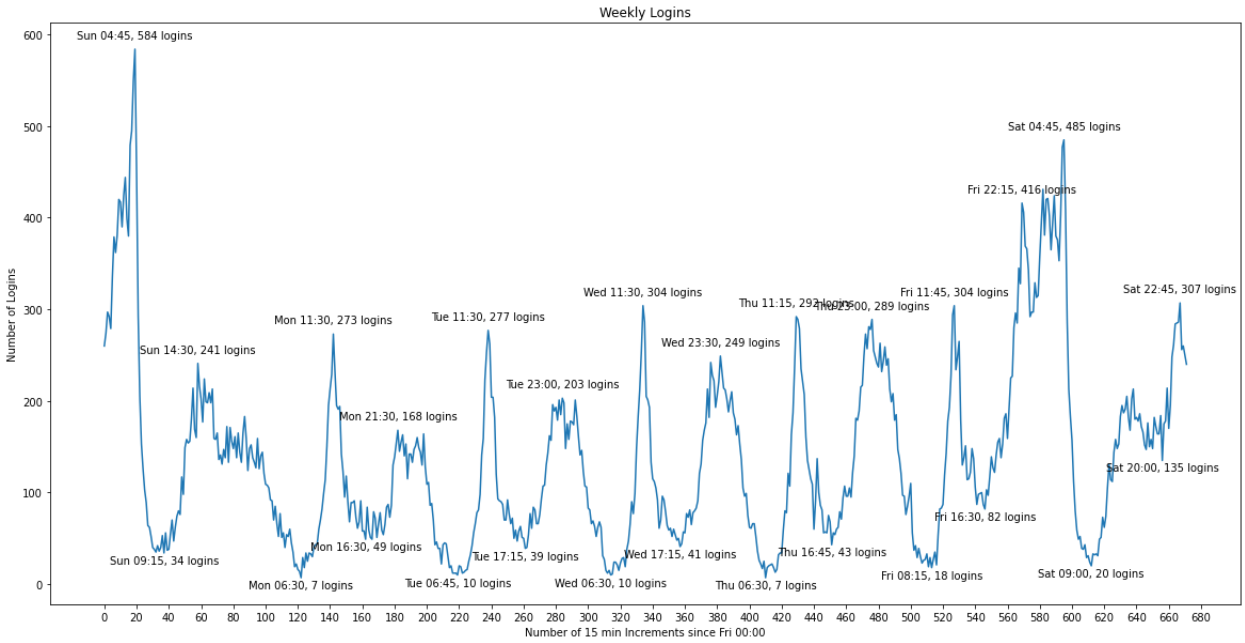
Take Home Challenge - Ultimate Tech Submission

Part 1 - Exploratory Data Analysis

After aggregating into 15 min increments and conducting EDA, the login time series data can be illustrated in the following plot.



During the week, the most logins occur around the beginning of lunchtime (11:30-11:45am) and in the evening (9pm-12am), while the least logins occur in the morning (6-7am) and late afternoon (4-5:30pm).

During the weekend, the login behavior changes since users sleep, (and as a result, wake-up) much later. More specifically, the most logins occur around 4-5am on Saturday and Sunday while the least logins occur approximately 8:30-9:30am on Saturday and Sunday before resetting for the week and repeating the weekday cycle.

Part 2 - Experiment and Metrics Design

1. I would choose Revenue as the key metric since, at least in my opinion, more profit is what would motivate driver partners to change driving behavior. Since toll costs will be reimbursed, driver expenses should decrease and drivers should also provide longer trips as they travel between cities. Overall, more money will be kept in their pockets.

2. I would develop a small experiment with a select few drivers that is designed to be scalable. First, candidate drivers would be screened based on the drivers' overall rating and revenue over a specified timeframe. Then, these candidates would be interviewed for driver interest, geographical knowledge/experience and, most importantly, acknowledgement/consent. After reaching (or failing to reach) the goal for an established revenue threshold for a specified timeframe, the experiment would end and the drivers should be rewarded with an incentive as a way of saying thank you for your participation. However, this experiment should account for caveats such as construction, natural disasters, local events (parades, festivals, farmers markets, etc.), and anything that might cause a reduction in transportation activity such as a pandemic.

   After gathering relevant experimental data (i.e. number of trips, miles driven, date/time of each trip, price per trip, etc.), testing for difference of means and ANOVA should be sufficient for this case since Ultimate Tech should determine if reimbursing tolls results in higher revenue (than revenue without reimbursing toll) and this increase is not due to random chance, respectively. Additionally, the caveats listed earlier will affect these results so analysts should seek what peculiar data points (or even outliers) are a direct result of these caveats. These will then be presented as recommendations to implement as company policy.

Part 3 - Predictive Modeling

1. After receiving the data, I conducted EDA by using "df.info()" and "df.describe()" to observe data integrity. New columns were created to assist with modelling categorical data and the percentage of retained users was found to be approx. 62%. Then, the data was cleaning by deciding to drop all null values since approx. 83% of the original data will still be retained. Histograms and time series plots (displayed in Jupyter notebook) were also generated to observe column behavior over time.
2. I used a Logistic Regression model simply because the category of interest is binary (i.e. either a user is considered "active" or not). This model's parameters were optimized through Randomized Search cross-validation and produced an impressive AUC score of 0.99. Other classification models such as Random Forests and Decision Trees are also applicable here and can prompt further modelling opportunities. Numerous unique combinations of columns were tested but overall the "length_of_membership" had the most predictive power and was decided to be the primary factor of active user retention. Using other features add a negligible amount of additional predictability and most importantly, by convention, the best models should be simple.
3. Since the chance of long-term rider retention is 0.18%, Ultimate Tech should consider strategies that can provide ongoing value to users such as flash discounts or a points-based rewards system for continuous usage. Other marketing strategies that emphasize the ongoing/long-term value of using the ride-share system should also be considered and implemented by leadership.