Data Analysis Interview Challenge

Part 1 – Exploratory Data Analysis

The login times data had no null values in it and were between 8:12 PM11/1/1970 through 6:57PM 4/13/1970. I created 15-minute intervals running from 8 PM on the starting date through 7 PM on the ending date of the data. The monthly data showed the average logins per day for each month increase from January to March with a slight dip in April, but the last day counted in April ended early at 7 PM, skewing Aprils average lower.

Hourly average logins display a clear split in login behavior on the weekends compared to weekdays. Weekdays typically started with more logins after midnight and tailing off until 6 AM when login activity picked up again. It would then peak at around 11 AM before falling off until 4-5 PM where it would climb again until 11 PM. Weekends typically climbed in activity right after midnight before logins plummeted at 6 AM. The average logins per hour would then increase until around 2 pm when it would stay at a steady pace throughout the day.

https://github.com/Crit-Data-Science/Springboard/blob/24614607d42729c9612c6d6b8b430a7df81c6417/Take-home%20Challenge/Ultimate%20Technologies%20Take-Home/Logins%20EDA.ipynb

Part 2 – Experiment and Metrics Design

1. The first metric that came to mind was measuring the number of times driver partners cross the bridge. After considering that further, it became apparent that was one step removed from a more important metric of how many driver partners service both cities. This metric makes the most sense because the behavior the experiment intends to increase is driver partners doing rides for customers in both cities.
2. Practical Experiment:
    a. I would create select a representative group of driver partners that have been randomly selected and divide them into two groups. One group that has their tolls reimbursed and one that does not. The group that is having their tolls reimbursed would be notified and ideally, they would not know that they are part of an experiment. That may be hard to do if they are potentially in communication with drivers who are not getting reimbursement. I would run this experiment for 1-2 months but more or less time may be required depending on sample sizes and market trends.
    b. The statistical test I would go forward with in this case would be chi-squared testing a null hypothesis that states there will be no difference in the proportion of drivers servicing both cities.

c. The recommendation to proceed with full implementation of the toll reimbursement be determined by whether a statistically significant increase was observed in the group receiving the reimbursements. If that increase is seen the city operations managers could feel more confident in proceeding with the plan. If no increase is seen, they may have to consider other factors that may prevent divers from servicing both cities. The caveats to this experiment include factors such as already existing trends or other events that could be influencing the results.

Part 3 – Predictive Modeling

https://github.com/Crit-Data-Science/Springboard/blob/24614607d42729c9612c6d6b8b430a7df81c6417/Take-home%20Challenge/Ultimate%20Technologies%20Take-Home/Ultimate%20Data%20Challenge%20Dytko.ipynb