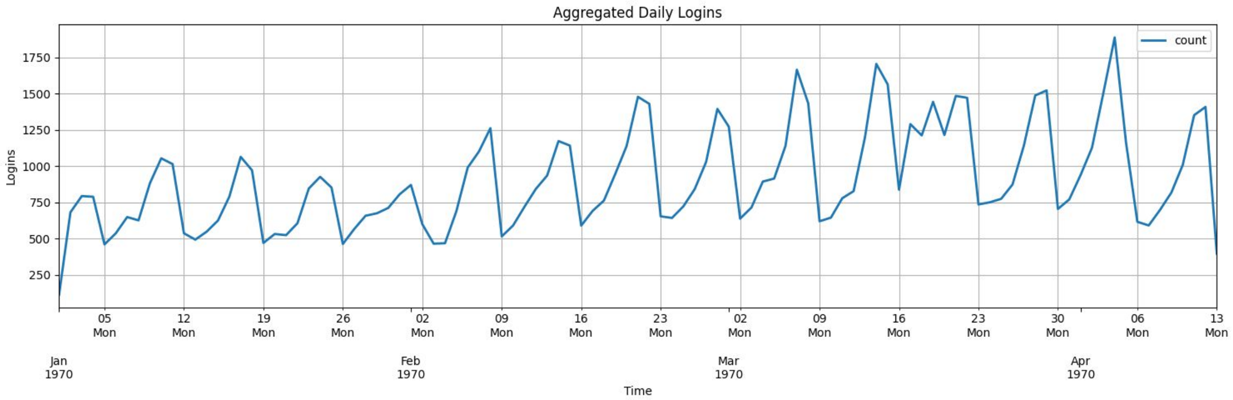
Ultimate Data Analysis Challenge

Report

Part 1 - Exploratory Data Analysis

The following is a daily aggregate plot of the data:



- As shown, the most number of logins happen on the weekends and the least number of logins seem to occur on Mondays.
- The trend was increasing, but the last week's numbers seem lower than normal.
- It's also worth noting there are ~3.5 months of data, so not enough to confidently predict seasonality.

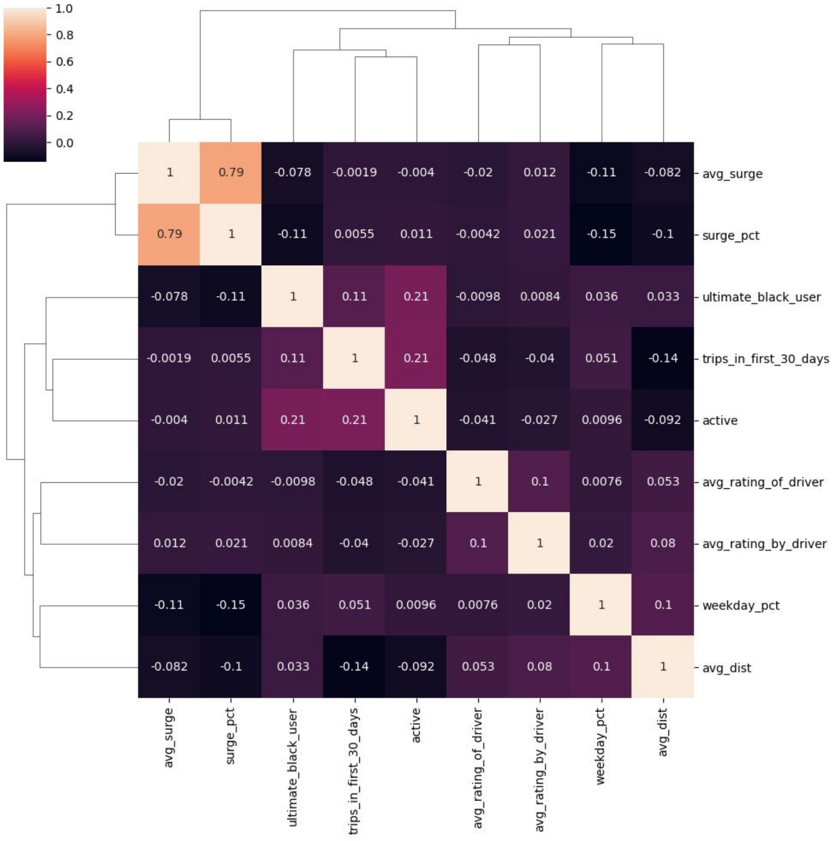
Part 2 - Experiment and Metrics Design

1. The key metric I would choose would be **total trips between cities**. This metric compared with the total intercity trips outside of the experiment will determine if there is a statistically significant increase.

2. Experiment design for reimbursing toll costs affect on inter-city trips.
    a. We will want to measure the number of trips between the cities for a week or more before the experiment to get the control group. We can then promote the reimbursement program and measure how many trips occur between cities.
    b. We'll use a paired t-test to measure the statistical significance between the trips between cities before and during the experiment.
    c. The null hypothesis is that the reimbursement incentive will have no effect. If there is a statistical significance determined from the paired t-test, then we can reject the null hypothesis and determine that the incentive did have an effect.

Part 3 - Predictive Modeling

1. *Perform any cleaning, exploratory analysis, and/or visualizations to use the provided data for this analysis (a few sentences/plots describing your approach will suffice). What fraction of the observed users were retained?*
    a. Most of the data were without null values except for the avg_rating_of_driver, phone, and avg_rating_by_driver. I dropped all of the rows which had a null phone value and used the median for the avg rating columns.
    b. The positive correlation between avg_surge and surge_pct is to be expected and there were only slight other correlations between active and some other features.
    c. I kept 99.2% of the observed data.



	avg_surge	surge_pct	ultimate_black_user	trips_in_first_30_days	active	avg_rating_of_driver	avg_rating_by_driver	weekday_pct	avg_dist
avg_surge	1	0.79	-0.078	-0.0019	-0.004	-0.02	0.012	-0.11	-0.082
surge_pct	0.79	1	-0.11	0.0055	0.011	-0.0042	0.021	-0.15	-0.1
ultimate_black_user	-0.078	-0.11	1	0.11	0.21	-0.0098	0.0084	0.036	0.033
trips_in_first_30_days	-0.0019	0.0055	0.11	1	0.21	-0.048	-0.04	0.051	-0.14
active	-0.004	0.011	0.21	0.21	1	-0.041	-0.027	0.0096	-0.092
avg_rating_of_driver	-0.02	-0.0042	-0.0098	-0.048	-0.041	1	0.1	0.0076	0.053
avg_rating_by_driver	0.012	0.021	0.0084	-0.04	-0.027	0.1	1	0.02	0.08
weekday_pct	-0.11	-0.15	0.036	0.051	0.0096	0.0076	0.02	1	0.1
avg_dist	-0.082	-0.1	0.033	-0.14	-0.092	0.053	0.08	0.1	1

2. *Build a predictive model to help Ultimate determine whether or not a user will be active in their 6th month on the system. Discuss why you chose your approach, what alternatives you considered, and any concerns you have. How valid is your model? Include any key indicators of model performance.*
    a. Created an ‘active’ column that used the last_trip_date column to determine which users were still active in the last 30 days (37.64% active users).
    b. Transformed the categorical categories using label encoding

   c. Dropped the date columns, because they were only necessary to determine active users label.
   d. Classification models - Logistic Regression, Decision Tree, Random Forest, Extra Trees, and XGBoost.
   e. XGBoost did best with an F1 Score of 0.706.
   f. Evaluation using F1 Score because we can get a general sense of how recall and precision are behaving before hyper-tuning.

3. *Briefly discuss how Ultimate might leverage the insights gained from the model to improve its long term rider retention (again, a few sentences will suffice).*
   a. Ultimate can use a model to determine which customers might churn after a 30-day period and use this insight to better serve the customer based on further data through surveys or interviews.