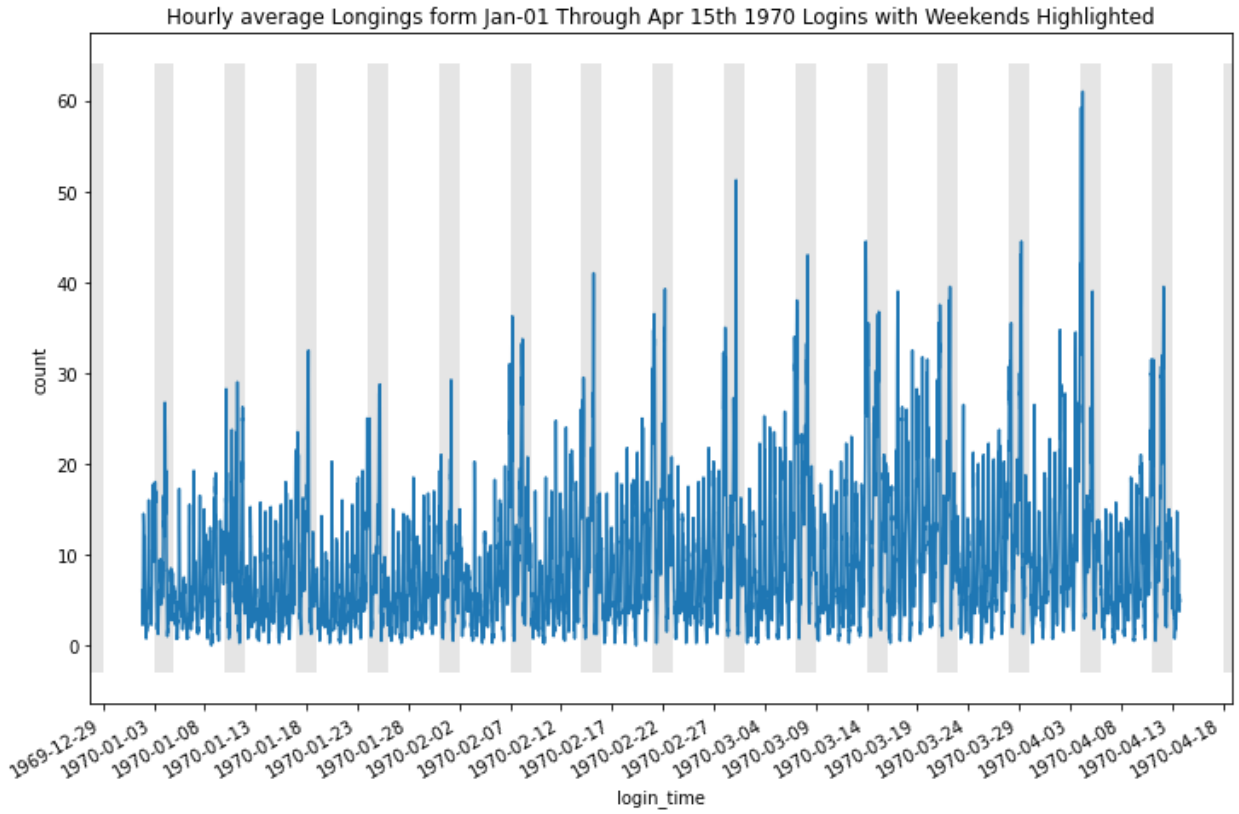
# Data Analysis Interview Challenge

## Part 1 - Exploratory data analysis¶

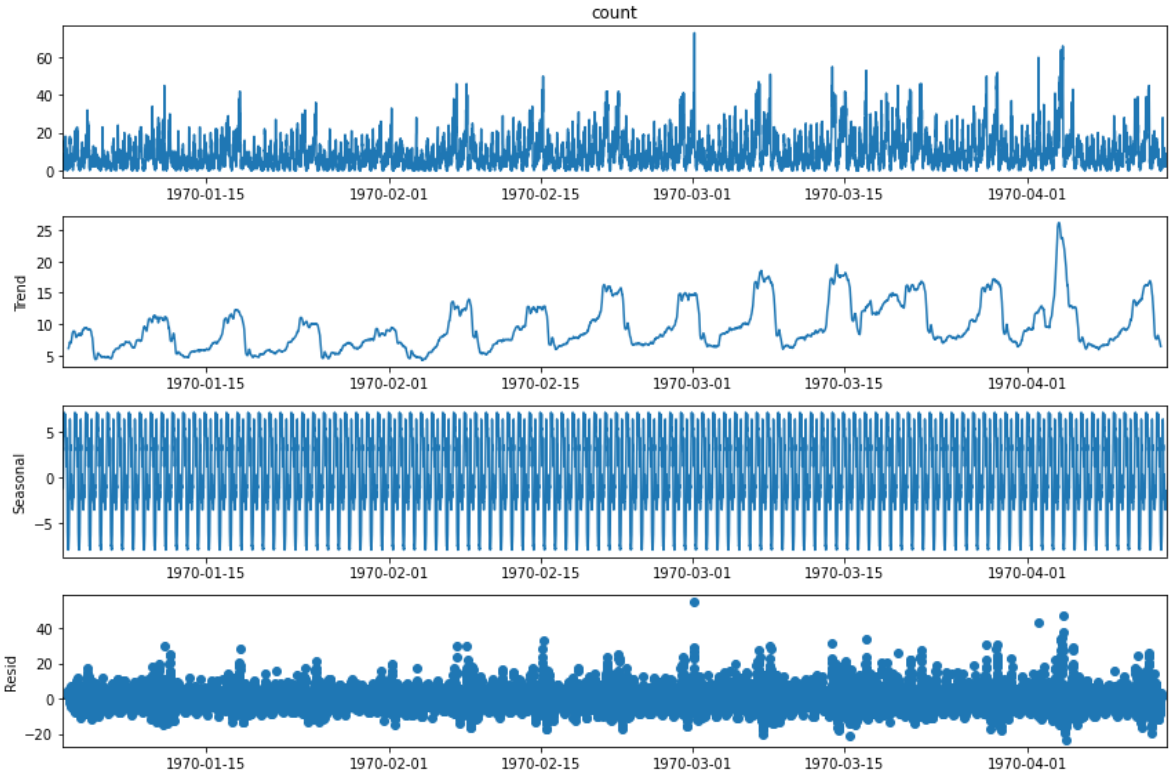
The attached logins.json file contains (simulated) timestamps of user logins in a particular geographic location. Aggregate these login counts based on 15-minute time intervals, and visualize and describe the resulting time series of login counts in ways that best characterize the underlying patterns of the demand. Please report/illustrate important features of the demand, such as daily cycles. If there are data quality issues, please report them.




## Part 1 Summary

The full aggregations and exploration can be viewed in the notebook: 'part1.ipynb'

- The user logins tend to spike on Friday and Saturday nights.
- The mean number of logins of each month goes up from month to month peeking in March. There is a daily ebb and surge with daily highs in the evenings.
- We see an upward trend that seems to drop the last week of the collected data.




## Part 2 - Experiment and metrics design

The neighboring cities of Gotham and Metropolis have complementary circadian rhythms: on weekdays, Ultimate Gotham is most active at night, and Ultimate Metropolis is most active during the day. On weekends, there is reasonable activity in both cities.

However, a toll bridge, with a twoway toll, between the two cities causes driver partners to tend to be exclusive to each city. The Ultimate managers of city operations for the two cities have proposed an experiment to encourage driver partners to be available in both cities, by reimbursing all toll costs.

1. What would you choose as the key measure of success of this experiment in encouraging driver partners to serve both cities, and why would you choose this metric?
2. Describe a practical experiment you would design to compare the effectiveness of the proposed change in relation to the key measure of success. Please provide details on:
    - how you will implement the experiment
    - what statistical test(s) you will conduct to verify the significance of the observation
    - how you would interpret the results and provide recommendations to the city operations team along with any caveats.

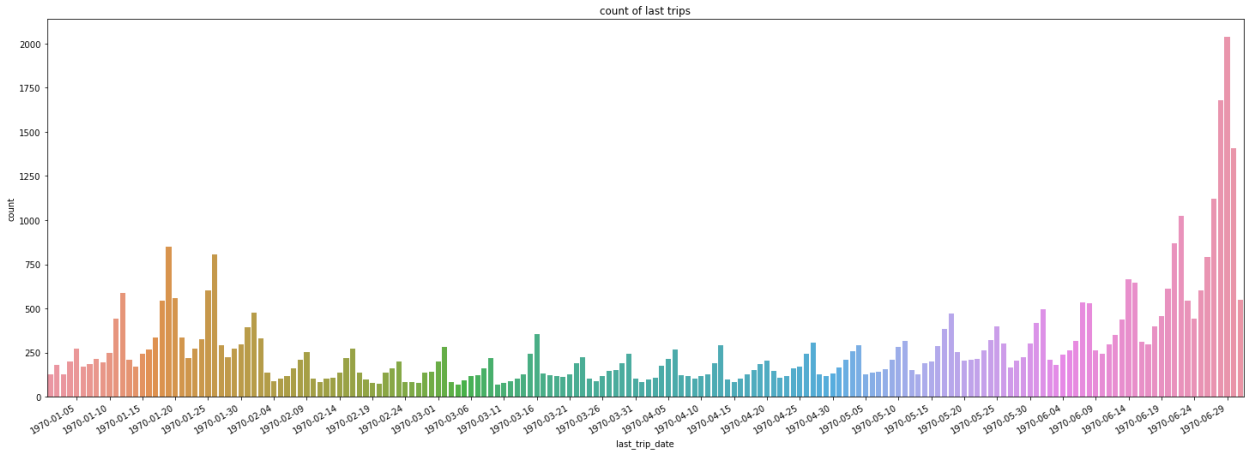
## Part 2 Proposal & Summary

Hopefully, there would be access to the previous years of toll records. First I would collect the data on toll reimbursements and the increase in tolls for a period of 4 months. I would compare the increase in tolls during the reimbursement period to the same months from prior years. As well as how closely the tolls increased aliens with reimbursed tolls. I would do a hypothesis test to test the significance of these observations. Specifically, I would be looking to see if the change in observe tolls with the null hypothesis test that there is no change. We would look for a p-value of less than .05 indicating that we could reject the null hypothesis and accept that the reimbursed tolls increased the driver's willingness to service both cities. If the p-value is above .05 then we find there is no significant change from our experiment. Should we find significant change and accept the alternative hypothesis, I would recommend that the city of operations should discontinue reimbursing toll costs and offer discounts to driver companies instead. Selling discounted passes to the driver companies takes the responsibility for the toll off of the drivers and the city will still make part of the toll from the company. It also reduces the paperwork and transactions of reimbursing individual tolls

## **Part 3** - **Predictive modeling**

Ultimate is interested in predicting rider retention. To help explore this question, we have provided a sample dataset of a cohort of users who signed up for an Ultimate account in January 2014. 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. We would like you to use this data set to help understand what factors are the best predictors for retention, and offer suggestions to operationalize those insights to help Ultimate. The data is in the attached file ultimate_data_challenge.json. See below for a detailed description of the dataset. Please include any code you wrote for the analysis and delete the dataset when you have finished with the challenge.

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?
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.
3. Briefly discuss how Ultimate might leverage the insights gained from the model to improve its longterm rider retention (again, a few sentences will suffice).




## Part 3 Summary

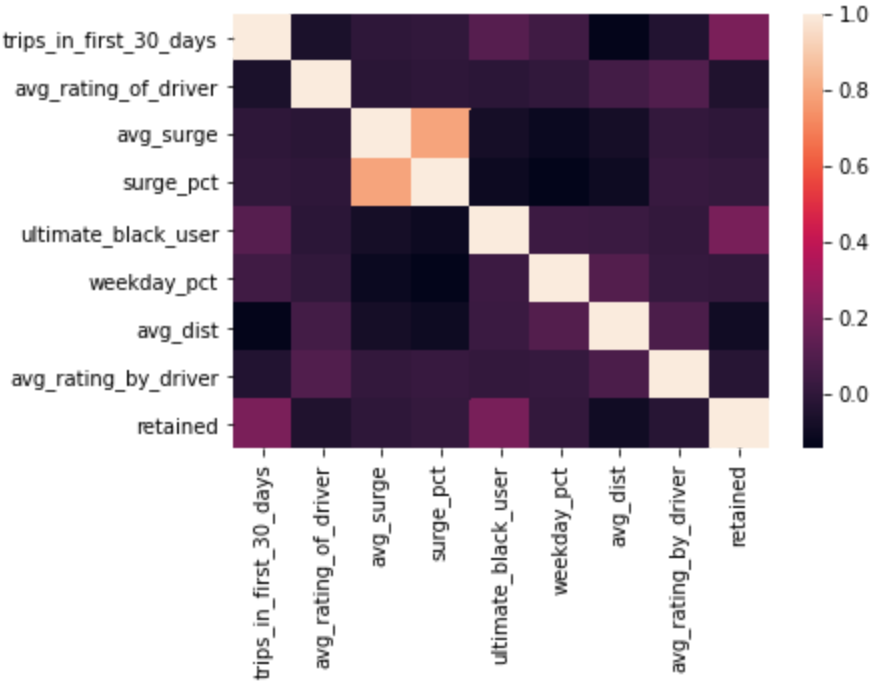
The full cleaning, exploratory analysis, preprocessing, and modeling can be viewed in the notebook: ‘part3.ipynb’

**Cleaning:**

- Columns with missing data: phone filled ‘unknown’, ‘avg_rating_of_driver’ and ‘avg_rating_by_driver’ filled mode
- Converted ‘last_trip_date’ to datetime
- Created target feature ‘retained’ as a binary feature 1 for active in the last month

**Data Exploration:**

- The data ranges from 2014-01-01 to 2014-07-01
- Almost 38% of observed users were retained
- The majority of lost users took their last trip within the first month of use
- More iPhone users are retained
- there are fewer users in King’s Landing but more of them seem to be retained
- Few features seem to be explanatory or correlate to the target retained



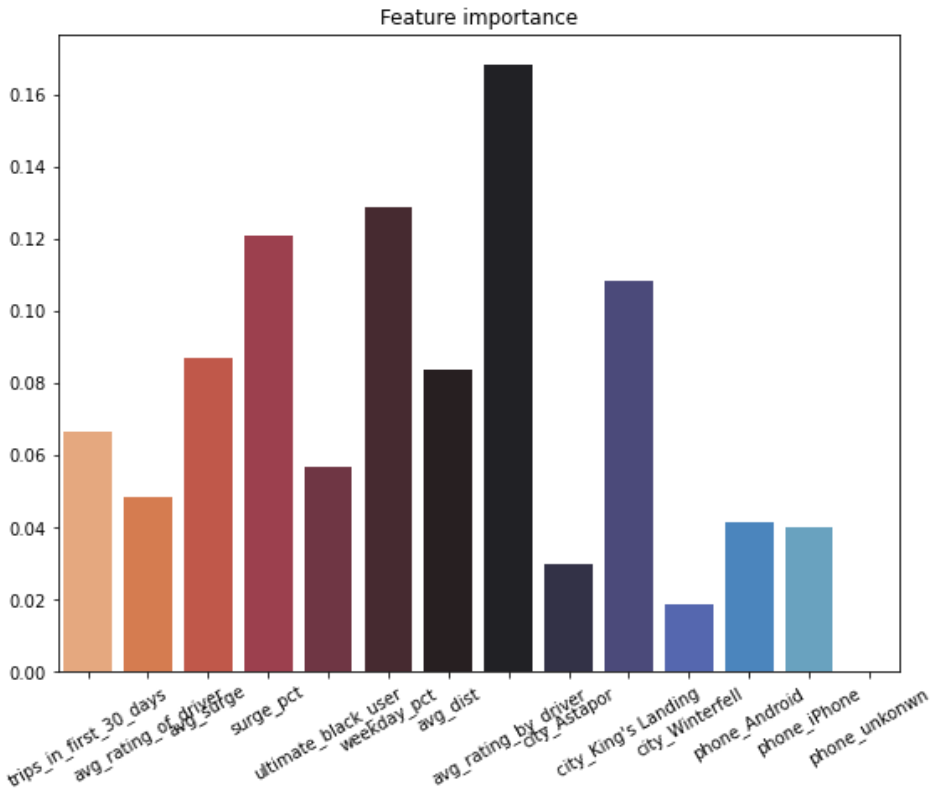

**Preprocessing:**

- With no clear linearity, or correlation Logistic Regression may struggle with this set
- With no clear clustering and the number of dimensions, KNN would struggle
- Random Forest is the simplest and easiest to interpret but computationally expensive
- Random Forest Classifier doesn't need data to be scaled
- One hot encode categorical features
- Split train and test sets

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.80 | 0.86 | 0.83 | 9346 |
| 1 | 0.74 | 0.65 | 0.69 | 5654 |
| accuracy | | | 0.78 | 15000 |
| macro avg | 0.77 | 0.76 | 0.76 | 15000 |
| weighted avg | 0.78 | 0.78 | 0.78 | 15000 |

**Modeling:**

- Gridsearch to tune hyperparameters
- The model predicts the retained users with %80 precision, and %86 recall
- The model only predicts slightly better on the train set but is not overfitting.
- The feature the model found of most importance is the 'avg_rating_by_driver' followed by 'surge_pct,' 'weekday_pct,' and 'city_King's Landing'




**Conclusions:**

- The Random Forest model can be used to predict user retention with 86% recall.
- With ‘avg_rating_by_driver’ being the most influential feature I would suggest surveying the drivers to find any patterns on who they are rating higher.
- The next most influencing features: percent of trips taken with surge and percent of trips taken on weekdays indicate that ultimate may benefit from running an experiment offering users who are likely to not retain a weekday promotion.
- The model may be improved by further features such as avg monthly trips, or the number of complaints the user has filed