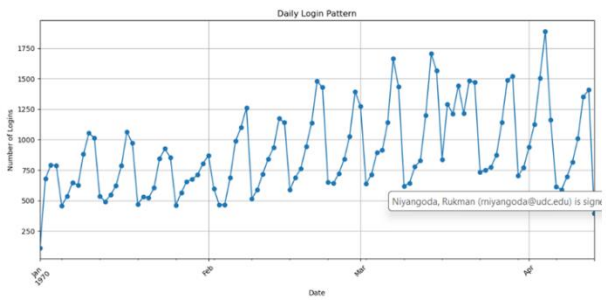
Data Analysis Interview challenge

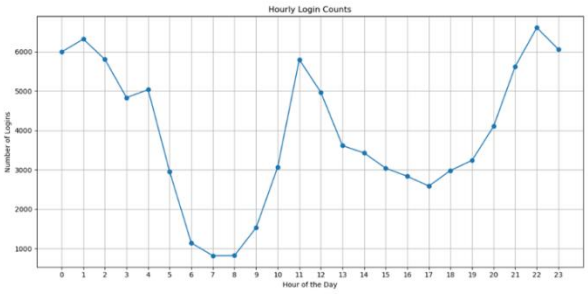
Part 1 – Exploratory Data Analysis

Logins.json Analysis
Preparation for analysis
The file was converted to a CSV file and initial exploration was done in a Jupyter Notebook

Overview of the data
Total number of columns: 1 – Login Time (a Date/Time field)
Total number of rows: 93142 – total count of Logins

Created an initial visualization in Jupyter Notebook and an overview of the data distribution over the months. A quick overview of the aggregated logins during each hour of the day

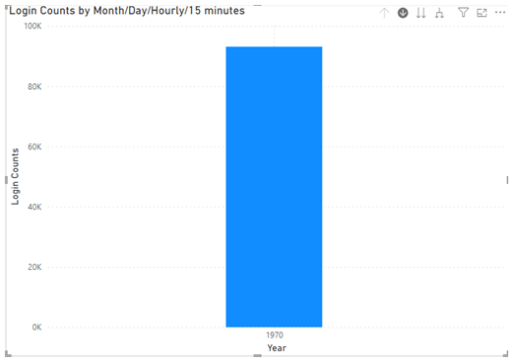


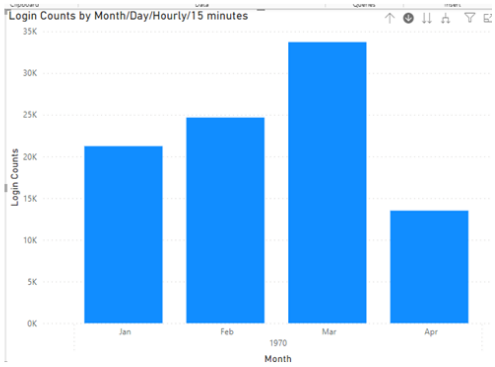


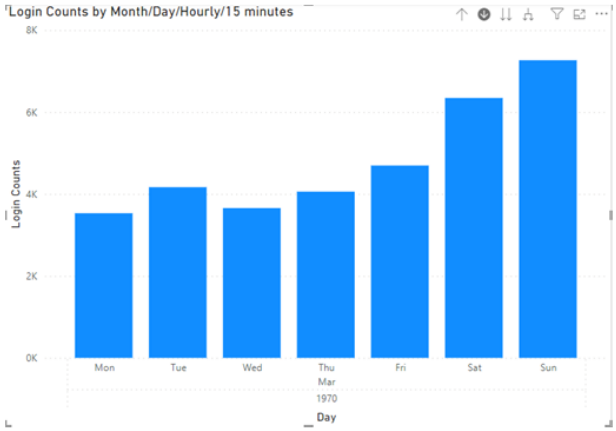
Further analysis was performed in Power BI

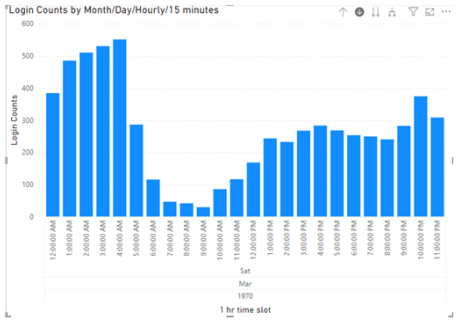
Summary of the analysis

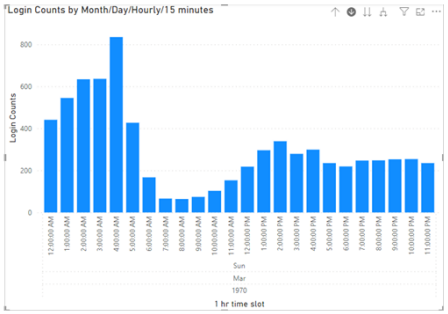
- The logins represented for a period of time in the year 1970 over four (4) months.
- There was a peak in the month of March.
- The peak logins were recorded during after hours of the day. Either very early morning hours or late into the nights. Although the aggregated peak times over the 4 months period, over the 24 hours seem to be between 10PM - 1AM over the weekend.
- The peak days happened to be the week-ends starting from Friday and most logins were recorded on Sunday.
- However, when considering the logins during the month of March, Sunday seemed to days that shows the highest logins, and the highest login times were between 4:30 and 4:45 early morning, rather between 4-5 AM.



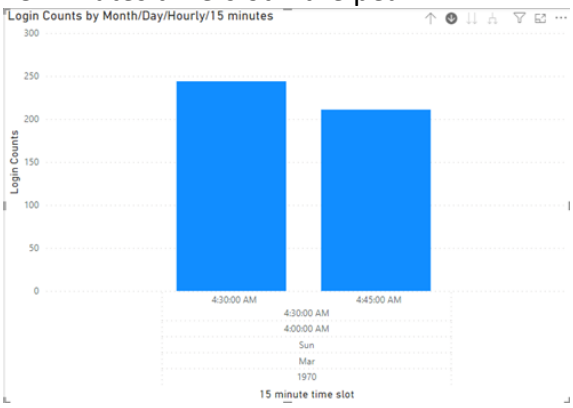








15 minutes time-slot – the peak



Potential issues

- The data only covered Jan – Apr months only.
- All login times were reported as Date/Time. However as there was no other data provided, I cannot make judgement on the consistency of the data, except that there was a consistent pattern of Weekends and after hours logins, for the period that was accounted for.
- One thing to note is that when tested for duplicated login times, I found 868 login times that are duplicated. This may need further investigation as to find out how certain login times are duplicated. Could that be a glitch in the login application? Could there integration issues? This needs further investigation and analysis

Part 2 – Experiment and Metrics Design

1. Measure of success

The key measure of success that propose is the increase in cross-city trips.
This measure directly reports the effectiveness of the experiment along the incentivizing re-imbursement. Significant increase in cross-city trips would indicate that drivers are willing to travel between cities.

2. The experiment

The experiment is to track the re-imbursement of drivers' toll-fees

2.1. The implementation

Set up the experiment.

a. Gather the historical data of cross-city trips
b. Set up the re-imbursement policy and procedures – the registration of drivers, setup agreements, rules and regulations, system to track re-imbursements.
c. Determine the observation period – like for 3 months.
d. Mechanism to gather participants' feedback.

Operation

At the end of the experiment period, analyze the re-imbursement data and the trips made between the cities.

Comparison

Compare the trips made by drivers who registered for the re-imbursement program against the inter-city trips by non-registered drivers.
Assess the two different groups to determine the validity and the success rate of experiment.

2.2. Statistical test to verify the significance of the observation

The null hypothesis : there is no association

Chi-square Test of Independence can be used determine if there is a significant association between the toll-reimbursement program vs non-reimbursement program.

Compare the observed number of inter-city trips, and the frequencies between two groups to see if there is a significant difference.

2.3. Interpretation

If the chi-square test indicates a significant association between toll reimbursement and cross-city trips, it suggests that the reimbursement policy has had an impact on driver behavior. In this case, it would be recommended to continue or expand the toll reimbursement policy to further encourage cross-city trips.

Part 3 – Predictive Modeling

Question 1

Creating the target variable: active column

The new column was added based on the criteria as the driver took a trip within 30 days since the sign-up. A binary data type created for (1) active and (0) inactive.

Cleaning, exploratory data analysis and visualization

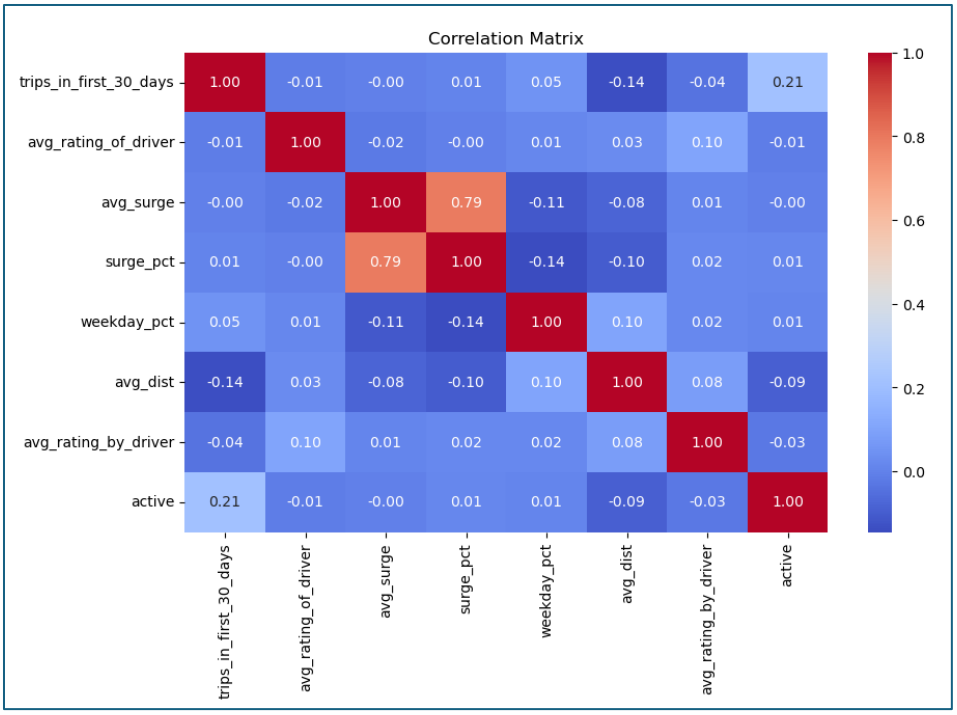
I checked for missing values and duplicate rows. 16% of the "avg_rating_of_driver" was missing and those missing values were replaced with the mean value.

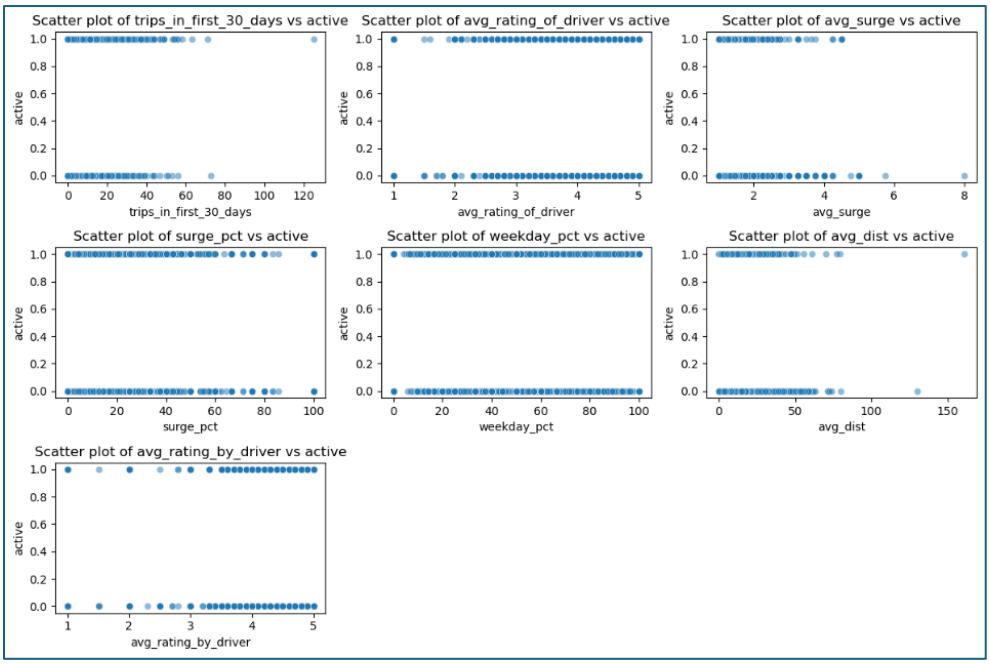
Also, "avg_rating_by_driver had NaN (missing values) and those were replaced with zero as that value was should have been provided by the driver, and I assume it was not provided. Therefore, the missing values were replaced with (0) zero.

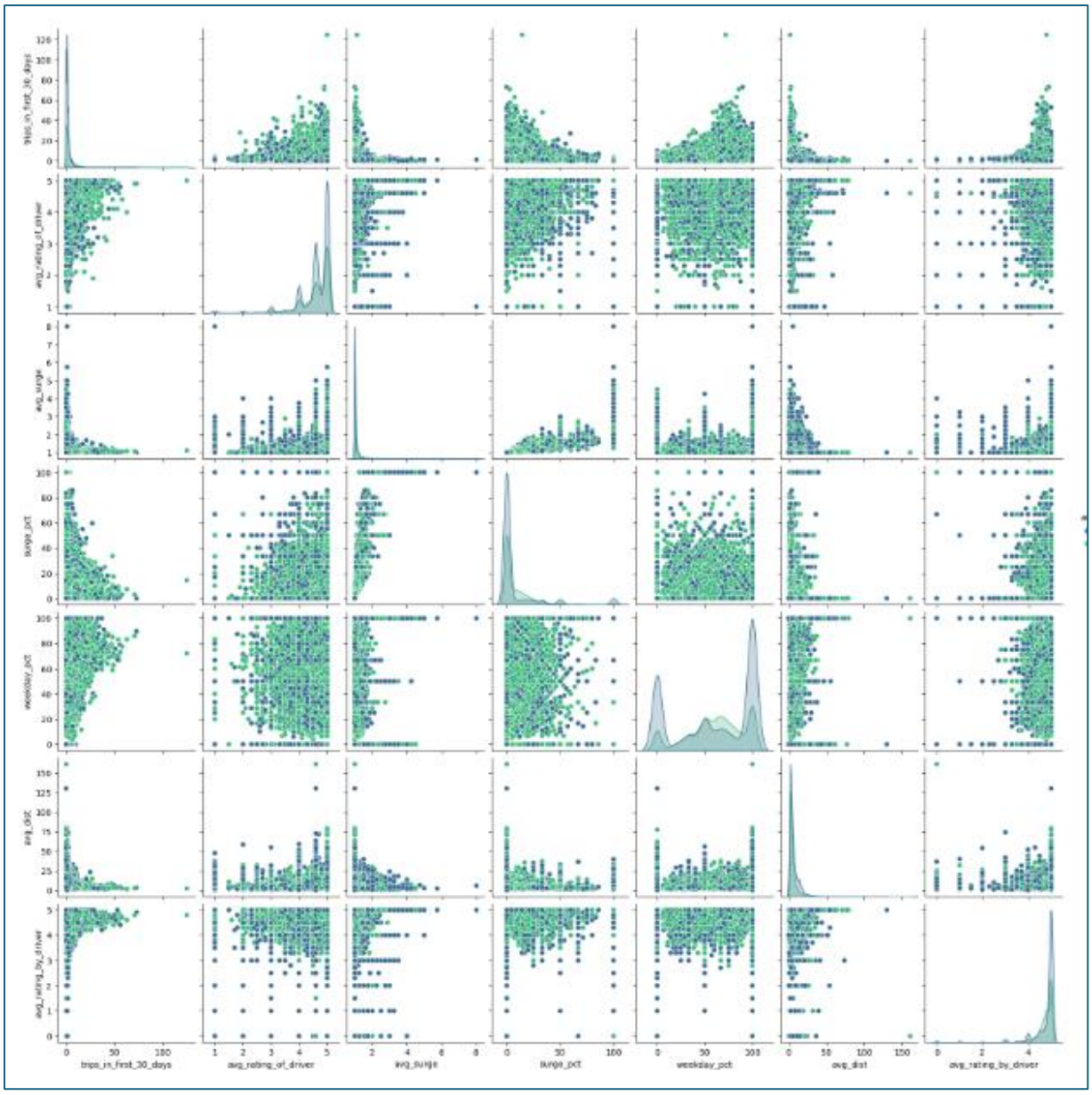
EDA and Visualization

Correlation

```
Pearson Correlation between trips_in_first_30_days and active: 0.21046322511130566
P-value: 0.0
There is a significant association between trips_in_first_30_days and active.

Pearson Correlation between avg_rating_of_driver and active: -0.010828732323239887
P-value: 0.015461338741197721
There is a significant association between avg_rating_of_driver and active.

Pearson Correlation between avg_surge and active: -0.0033330973902226608
P-value: 0.4560984370405743
There is no significant association between avg_surge and active.

Pearson Correlation between surge_pct and active: 0.011796748066780178
P-value: 0.008343129091357811
There is a significant association between surge_pct and active.

Pearson Correlation between weekday_pct and active: 0.009692972843291882
P-value: 0.030203503149421634
There is a significant association between weekday_pct and active.

Pearson Correlation between avg_dist and active: -0.09277986324064298
P-value: 5.270976361496232e-96
There is a significant association between avg_dist and active.

Pearson Correlation between avg_rating_by_driver and active: -0.007634675166147961
P-value: 0.08779464706814882
There is no significant association between avg_rating_by_driver and active.
```


Correlation Matrix
trips_in_first_30_days
avg_rating_of_driver
avg_surge
surge_pct
weekday_pct
avg_dist
avg_rating_by_driver
active
1.00 -0.01 -0.00 0.01 0.05 -0.14 -0.04 0.21
-0.01 1.00 -0.02 -0.00 0.01 0.03 0.10 -0.01
-0.00 -0.02 1.00 0.79 -0.11 -0.08 0.01 -0.00
0.01 -0.00 0.79 1.00 -0.14 -0.10 0.02 0.01
0.05 0.01 -0.11 -0.14 1.00 0.10 0.02 0.01
-0.14 0.03 -0.08 -0.10 0.10 1.00 0.08 -0.09
-0.04 0.10 0.01 0.02 0.02 0.08 1.00 -0.03
0.21 -0.01 -0.00 0.01 0.01 -0.09 -0.03 1.00
1.0
0.8
0.6
0.4
0.2
0.0


Scatter plot of trips_in_first_30_days vs active
Scatter plot of avg_rating_of_driver vs active
Scatter plot of avg_surge vs active
Scatter plot of surge_pct vs active
Scatter plot of weekday_pct vs active
Scatter plot of avg_dist vs active
Scatter plot of avg_rating_by_driver vs active
active
trips_in_first_30_days
avg_rating_of_driver
avg_surge
surge_pct
weekday_pct
avg_dist
avg_rating_by_driver



Fraction (%) of the retained drivers (active for up to 30 days after sign-up)
Num retained 18804 (active = 1)
total_users 49992 (total rows)
Fraction of retained users: 0.38 = 38%

Question 2

This project is about predicting a binary variable – Active or not. As a starting point for a classification model, I decided on a logistic regression model. The rationale for my selection:

- Well suited for a binary classification problem.
- Often can be modeled as a baseline model as it provides a reasonable starting point.
- Efficiency, where it is said be scaling well with large data sets like this one.

For this type of binary classification problem, the key performance indicators are:

1. evaluation metrics - Accuracy, Precision and F1-Score
2. Confusion matrix, where the model returns the True positives and True negatives, the performance of the classification model. This is an important performance indicator as the model must report insights into true positives, false positives, true negatives, and false negatives so the company can make better decisions and judgement. -.

The way to measure the validity of the model is by using common techniques such as evaluation metrics, cross validations and of course using the expert knowledge. The senior management must be part of the evaluation process, so that they are able to provide input to evaluate the model output.

Question 3

Ways the company can leverage the insights gained from the model to improve its long-term rider retention.

1. Detect the features that have strong correlations, such as frequent riders, usage, distance travelled, rider feedback and average ratings, so that the company can implement ways to improve those areas.
2. Once these key indicators are identified, carry out targeted marketing campaigns that address and enhance those areas.
3. Continue to enhance customer relationships, and support initiatives. Focus on riders who are at risk of churn. Create customized or personalized effort on those riders.