Part 1

Demand spikes daily at noon and midnight, with heavier demand on Friday and Saturday nights.

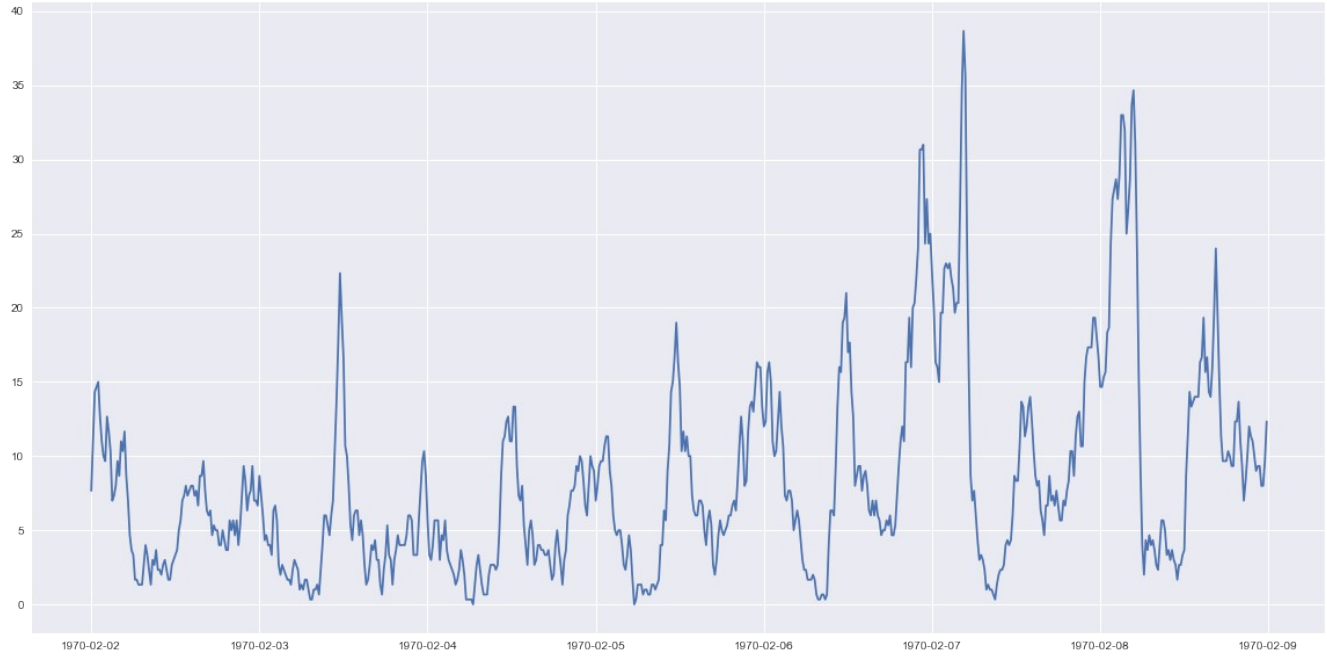



*Illustration 1: A sample Mon-Sun week of logins*

The data all claims to be from 1970. Given that the company was founded in 2010, that seems unlikely. I did not make any changes, since I don't know when the data was actually collected, but that should be addressed in the system that logs user activity.

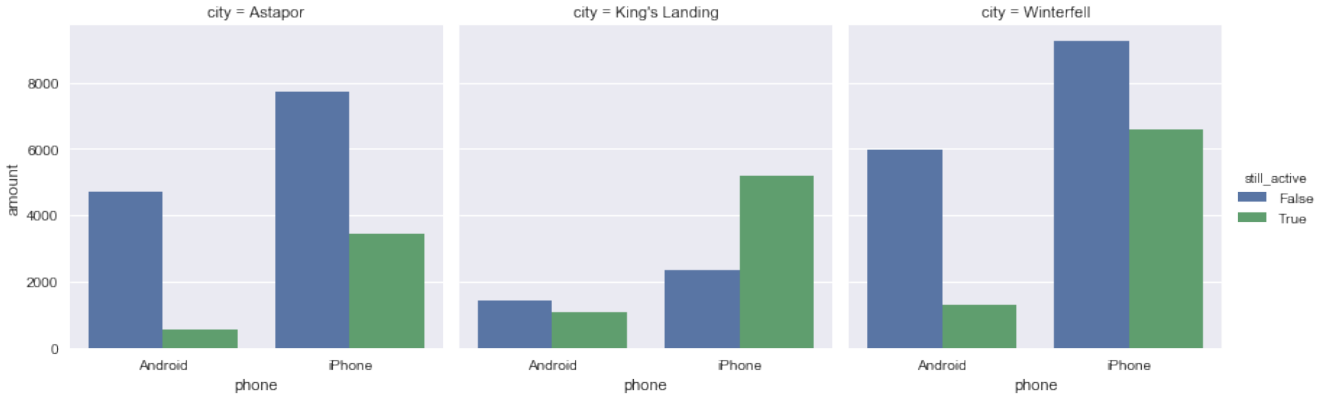
Part 2

1. I would measure success by the number of trips made over the trial period. The ultimate goal is to make sure that any existing demand is filled, and if facilitating drivers working both ciries will make sure that every user who wants a ride is able to get one. It's tempting to measure the participation by the number of reimbursements made, but since the two cities have a day/night cycle of peak activity, it could very well be the case that one reimbursement that moves the driver to the other city results in multiple customers serviced.
2. I would take a random sample of drivers and enroll them in the trial program, which should run for at least a week to cover both weekend and weekday activity. The trial group can then be compared to the control group at the end of the experiment to see what the effect size was of the change. Since we likely have lots of data about the usual population mean and standard deviation of the number of trips taken, a standard z-statistic should be fine here. I would interpret the results based on the cost of a reimbursement compared to the marginal gain in profit from providing it, checking to see if the project is more or less cost effective on weekdays/weekends.

Part 3

1. I made a separate column for 'still_active' that contains a boolean variable based on if last_trip_date > 2014-06-01. 36.6% of users were still active within the last month. I checked the dataframe for null values, and found the bulk of the missing data was in avg_rating_of_driver. For use in a model, those missing values would have to be imputed, but first I checked to see if a user deciding not to rate the driver had an impact on retention, and found that users missing that value only had a 19.3% retention rate.

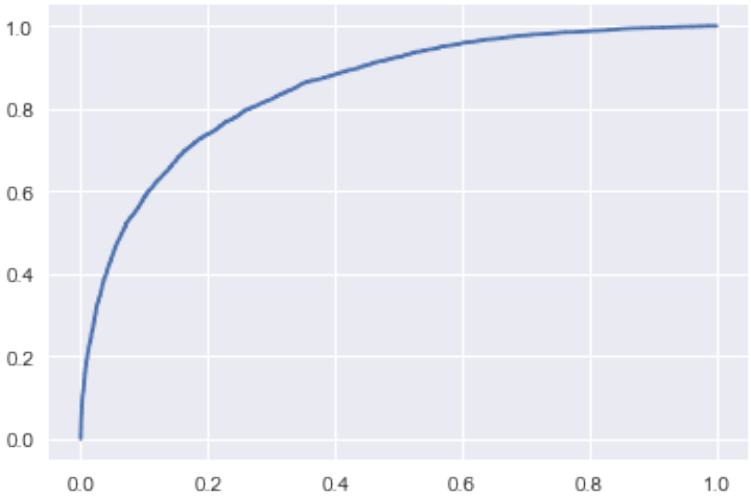
I checked some other variables to see if there was a similar behavior, and plotted it:




Two things jump out here: the retention rate is much higher in King's Landing than the other two cities, and Android users are far less likely to still be active than iPhone users.

At this point I prepped the data for the model by using KNN imputation to fill in missing values.

2. I used a random forest classifier for its ability to handle a mix of categorical and quantitative data, and its ability to give non-linear decision boundaries, which gives it some advantages over logistic regression or a SVM classifier. The classifier achieved a test set accuracy of 78.5%, and generated this ROC curve.



The feature importance generated by the model is as follows:

```
avg_dist                  0.071124
avg_rating_by_driver      0.171290
avg_rating_of_driver      0.048673
avg_surge                 0.095864
surge_pct                 0.126550
trips_in_first_30_days    0.063725
ultimate_black_user       0.061258
weekday_pct               0.112778
Astapor                   0.033399
King's Landing            0.108504
Winterfell                0.022432
Android                   0.046039
iPhone                    0.038364
```

- The difference between Android users and iPhone users is concerning. It might be worth checking the android app to see if there's something wrong that's turning away customers
- The user's rating by the drivers appears to be important. This might not be usable for improving retention rates, but it's the best indicator of whether a customer will stick around
- Kings Landing has much better retention rates than other cities. What's different about that city, and what of our business model can be adapted in the other cities to better fit the market there?