Rideshare Evaluation

The goal of this project is to create an algorithm that will examine a market's potential for ridesharing (consolidating multiple taxi trips into one). I then use this method to evaluate Manhattan's potential for ridesharing using a full week's worth of yellow cab (efficiency exploration notebook) and explore how things vary based on time and location (location and time exploration).

Code Files: **Efficiency Exploration.ipynb** and **Location and Time Exploration.ipynb**

Part 1: Aggregation

1. Assess Potential Efficiency

Algorithm Summary: I define an efficiently aggregated ride as one that reduces the amount of drivers and completes the rides without hurting user experience and revenue. My method of determining if a location has potential for producing such aggregated rides is as follows:

1. Simulate/aggregate rides that have the following points in common
    a. Going in similar direction (angle difference < 15 degrees)
    b. Pickup time difference < 5 minutes
    c. Similar Route distance
    d. Pickup location distance < .5 miles
2. For all aggregated rides, I analyze:
    a. Total rides aggregated
    b. Reduction in driver costs based on driver hourly wage
    c. Average increase in trip duration for passengers
    d. Total and average revenue increase or decrease

Assumptions:

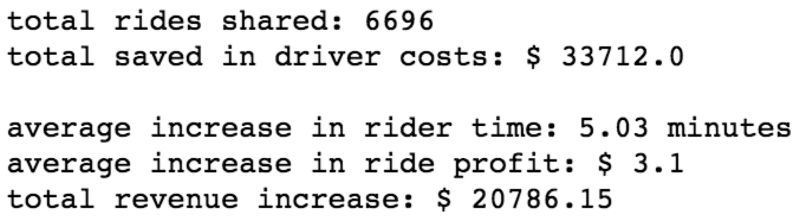
- $7 flat fee for each rideshare passenger, original fare for single ride comparison
- Trips to combine can be compared by straight line distance, not route/street distance
- Pickup/drop off of new rider takes 2 extra min
- The cost saved from removing a driver (aggregating 2 cars to one), is the duration of that driver's previous ride at their hourly wage ($20/hour)
- Rides cannot be matched together with more than 4 passengers

Goal: If a location is not losing revenue from the simulation of aggregated rides, and new customer trip times are not going up by more than 5-10 minutes on average, then the location likely has potential for producing efficiently aggregated rides.

Constraints and thresholds for matching rides to aggregate may be easily adjusted and would be valuable to examine in a sensitivity analysis as well.

2. Apply to Manhattan

After simulating and aggregating rides on a subset of the data for the first full week of June, we achieve the following results on 6696 new rides (aggregated from many more)



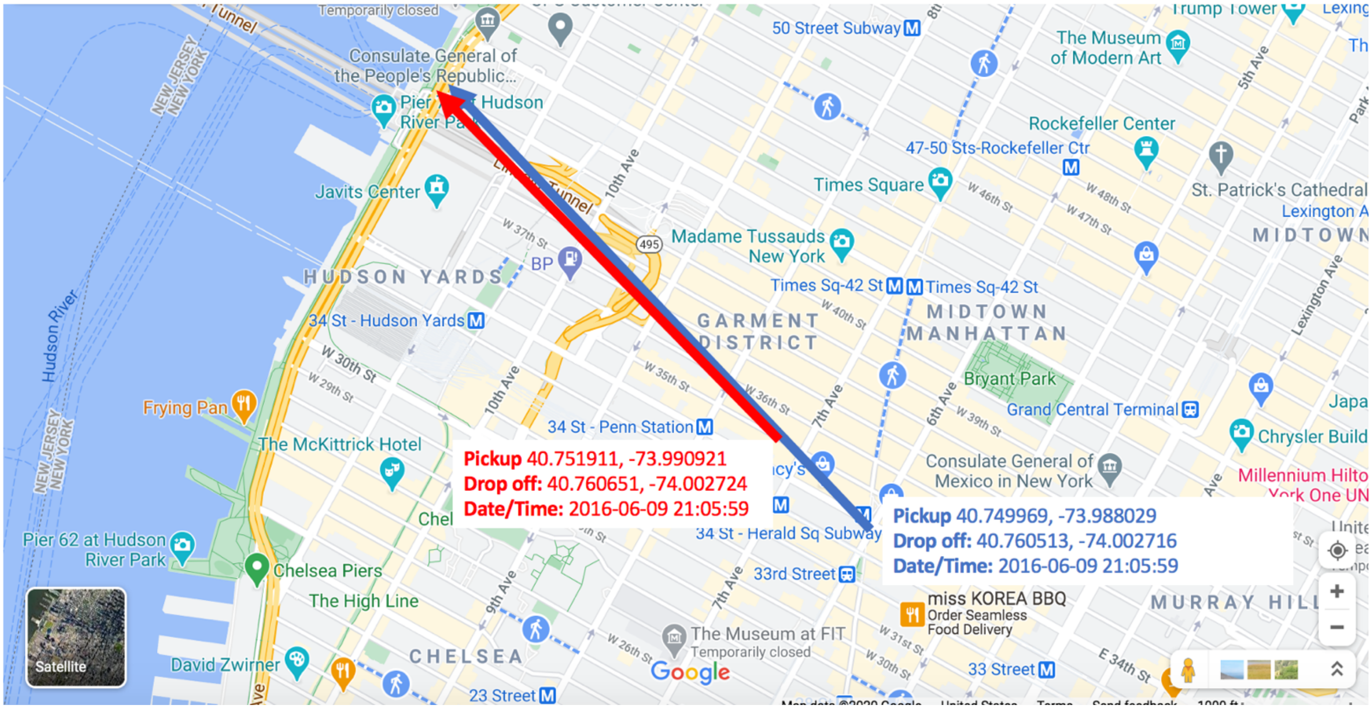
```
total rides shared: 6696
total saved in driver costs: $ 33712.0

average increase in rider time: 5.03 minutes
average increase in ride profit: $ 3.1
total revenue increase: $ 20786.15
```

Overall we see that the reduced driver costs and flat fare of passengers outweigh the original fare of the non-aggregated ride, for most rides in Manhattan using this simulation. Increase in passenger ride duration remained acceptable as well. Based on this, Manhattan looks to be a good candidate for further exploration of potential.

<u>Model Complexity:</u> My algorithm for simulating rides and evaluating efficiency will not grow exponentially complex. I sort the data based on pickup time and feed it in batches to the algorithm to aggregate, group based on matches, and then output into the evaluation.

So instead of comparing n to n rows, I am comparing each constant-sized subset to itself, n/x times, where n is the number of rides, and x is the batch size.

<u>Example of 2 Aggregated Rides</u>



Pickup 40.751911, -73.990921
Drop off: 40.760651, -74.002724
Date/Time: 2016-06-09 21:05:59

Pickup 40.749969, -73.988029
Drop off: 40.760513, -74.002716
Date/Time: 2016-06-09 21:05:59

3. Location and Time Analysis

Location Analysis

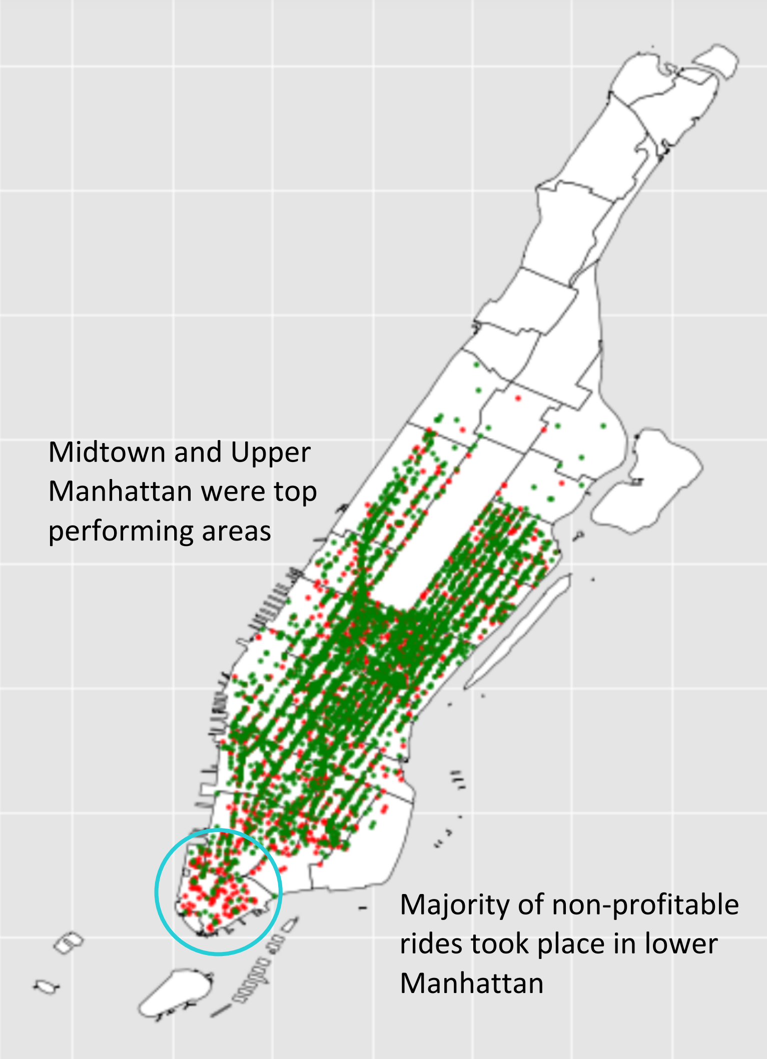
Midtown and Upper Manhattan proved to be efficient areas to aggregate rides in, showing higher revenue post-aggregation and acceptable duration change.

Lower Manhattan struggled to produce efficient rides even with our constraints.

Ride duration fluctuations were largely marginal based on location.

Rides with a higher duration relative to distance typically lost more money when they were aggregated, likely because the original cab fare exceeded the total flat fee of users. **Negative profit rides had an average duration 12 minutes higher than profitable rides.**

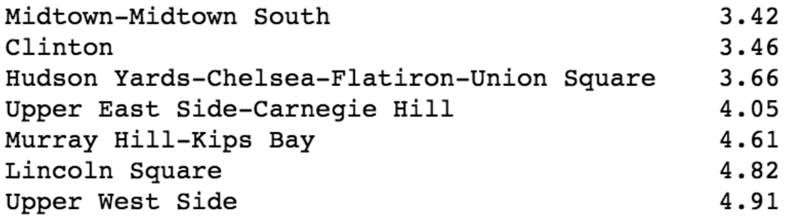
Low profit neighborhoods often had high duration rides originally, which may give some insight as to why they underperformed: More maneuverable grid-like structure of Midtown vs lower Manhattan, for example.



Midtown and Upper Manhattan were top performing areas

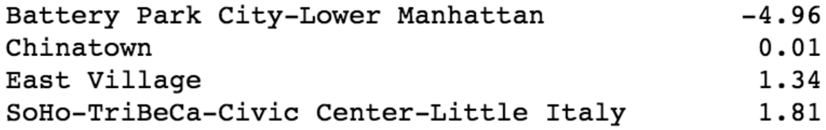
Majority of non-profitable rides took place in lower Manhattan

Avg revenue increase per ride in top performing neighborhoods (after aggregation)



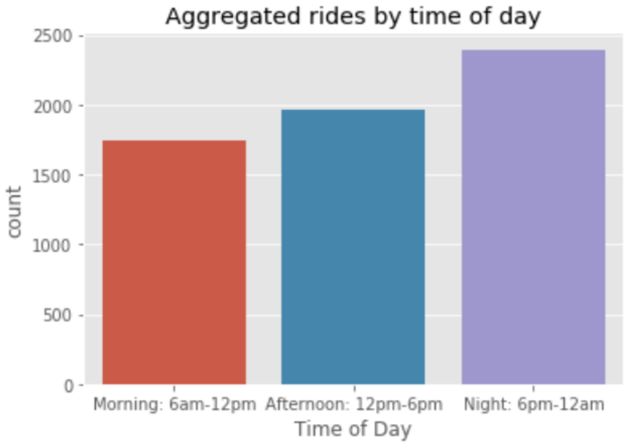
```
Midtown-Midtown South                           3.42
Clinton                                         3.46
Hudson Yards-Chelsea-Flatiron-Union Square      3.66
Upper East Side-Carnegie Hill                   4.05
Murray Hill-Kips Bay                            4.61
Lincoln Square                                  4.82
Upper West Side                                 4.91
```

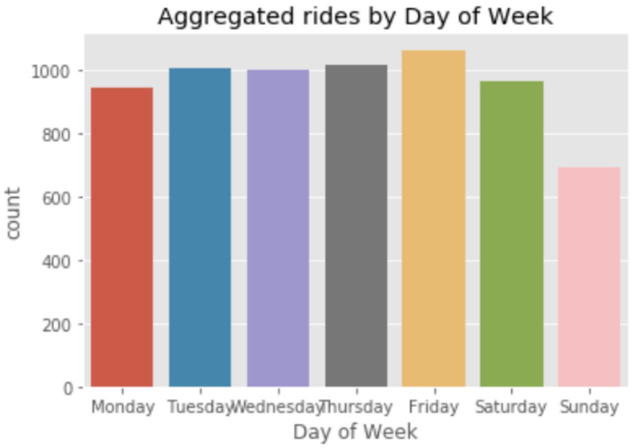
Avg revenue increase/decrease per ride in low performing neighborhoods (after aggregation)



```
Battery Park City-Lower Manhattan              -4.96
Chinatown                                       0.01
East Village                                    1.34
SoHo-TriBeCa-Civic Center-Little Italy          1.81
```

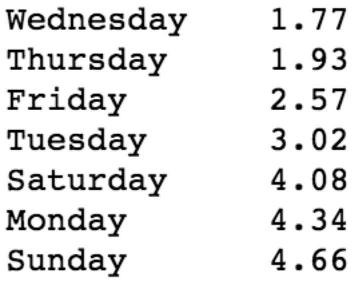
Time





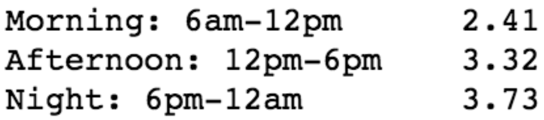
The algorithm aggregated more rides at night than during the day despite a balanced subset of data. My parameter constraints were more rigidly met during this time. In addition to this, aggregated rides were more profitable at night and on the weekend (particularly Saturday and Sunday). New duration changes were still marginal. Note: minimal data from 12am-6am

Avg revenue increase per ride by day of week



```
Wednesday    1.77
Thursday     1.93
Friday       2.57
Tuesday      3.02
Saturday     4.08
Monday       4.34
Sunday       4.66
```

Avg revenue increase per ride by time of day



```
Morning: 6am-12pm       2.41
Afternoon: 12pm-6pm     3.32
Night: 6pm-12am         3.73
```

Thoughts and Other Business Implications

One feature that stood out in our low-efficiency zones was the average duration of rides before aggregation. Longer rides proved to result in lower profit rides when aggregated. This is likely due to the original cab fare exceeding the total fixed flat fee structure in place in my algorithm. Perhaps the flat fee can be adjusted when expected ride duration/distance crosses a certain threshold to make sure we remain positive on our routes within these areas.

Our most efficient rides tend to take place on weekend nights in mid-upper Manhattan. Isolating efficient aggregation areas by neighborhood and only offering the feature to our high value zones (or marketing it more heavily in these zones) may also be a way to increase our performance on one specific area. By doing this, we can even potentially add value to cities where aggregation is inefficient on average, but worth it when we are selective.