Motivation:

Boston is a busy city where many rely on public transportation. Unfortunately, we cannot expect to have all of our trains to run on time -- in which case we have to take matters into our own hands if we want to be punctual. You could call a cab, a Lyft, or an Uber instead of waiting. But why not bike to where you need to go instead of contributing to traffic?

Objective:

The goal of this research is to find the optimal placement of Hubway bike stations, considering which stations consistently run behind schedule. This will maximize the number of options commuters have available to them, given that a train is running late. Moreover, it will decrease traffic as well as promote eco-friendly transportation

Methods:

Our project utilizes k-means (specifically Lloyd's algorithm, which we used the SciKit KMeans library for) to identify where trains are most often late and place a Hubway station there. We have the following data sets (and the algorithms associated with them):

1. MBTA Green Line stations and their coordinates, assembled manually with collected coordinates from erikdemaine.org/maps/mbta/ (getGreenLineCoords.py)

2. Hubway stations and their coordinates, from Hubway (getHubway.py)

3. MBTA Green Line stations and their reliability in 2017 (a metric that measures how many people who have waited longer than the expected time for their train, calculated by MBTA), from the MBTA's Fiscal & Management Control Board (getReliability.py)

Here is a walk-through of what we did (kCoords.py):

1. Create a dictionary called lateTrains where keys are stations and values are reliability (rounded down to nearest integer).

2. Create a lateTrainsCoords list. For each stop x in lateTrains, add the value corresponding to x entry -- the latitude and longitude of x, which we get from getGreenLineCoords.py. For example, if lateTrains looks like this:

   ```
   {"Hynes Convention Center": 2, "Babcock Street": 4}
   ```

   and we know from that Hynes Convention Center is located at (42.0123, -71.456) and that Babcock Street is located at (42.0789, -71.000), then step 2 will create a list that looks like

   ```
   lateTrainsCoords = [ (42.0123, -71.456), (42.0123, -71.456),
   (42.0789, -71.000), (42.0789, -71.000), (42.0789, -71.000),
   (42.0789, -71.000) ]
   ```

3. Run k-means (kCoords.py) on lateTrainsCoords. The output will give us an idea of where to place a new Hubway station (or where to move a new one).

In our experiment we set k = 40, because that is roughly the number of existing stations surrounding the Green Line. We then found the average distance from each of the 40 means we found to the closest existing Hubway station (meanDistances.py).

Limitations:

There are three major limitations to our methods that we have identified. The first being that we narrowed our analysis and placement of Hubway stations to just the Green Line, which ignores potential poor placement in many other neighborhoods of Boston. The second limitation we identified is computational cost. While we scaled down our data to make it manageable, our methods still require searching through an overwhelmingly large number of data points, essentially a data point per person that has experienced a late train in a year. The third limitation we identified is that by confining our placement to just MBTA stops, we have ignored station-placement issues with busses or other high traffic pedestrian areas.

Results:

The first figure (Fig.1) below displays the ideal placement of Hubway stations based on the latest trains on the MBTA Green Line. The second figure (Fig. 2) expands on Fig. 1 by mapping it with the current location of Hubway stations.

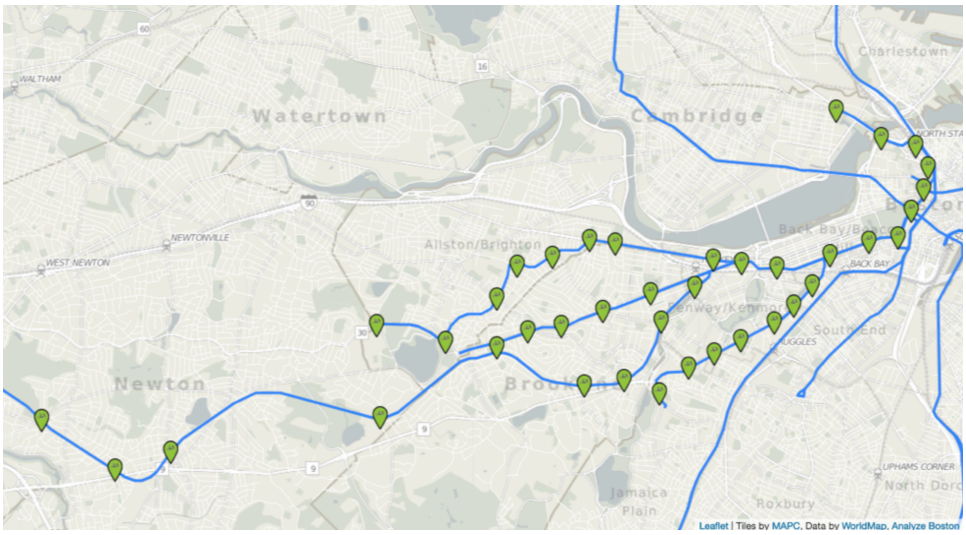

Fig.1

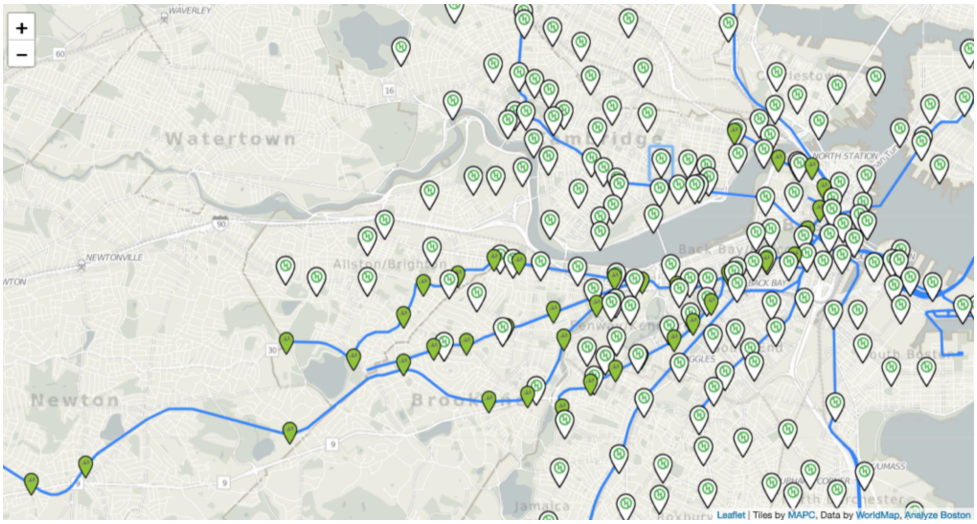

Fig. 2

Results:

The average distance found between existing hubway stations and the optimal location calculated by k-means was 640 feet. This is roughly 1.2x the distance between the BU East stop and the BU Central stop on the B-line of the MBTA Green Line. Walking versus biking this distance could be the difference between being on time and late.

It should be noted that the stations with the most frequent delays are in downtown Boston, as shown below. (Fig. 3). It should also be noted that many of these stations have Hubway stations very close. When running our k-means algorithm on just the downtown stops (Science Park, North Station, Haymarket, Government Center, Park Street, Boylston Street, Arlington, Copley, Hynes Convention Center, Kenmore), the average distance found between existing hubway stations and the optimal location was 98 feet. This likely because the density of Hubway stations is concentrated in the downtown area. (Fig. 4).

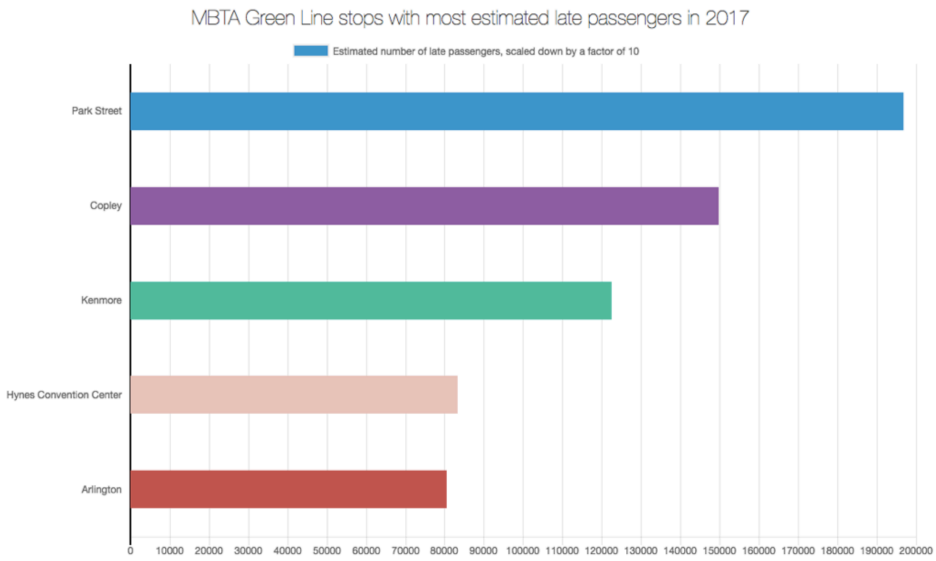

Fig. 3

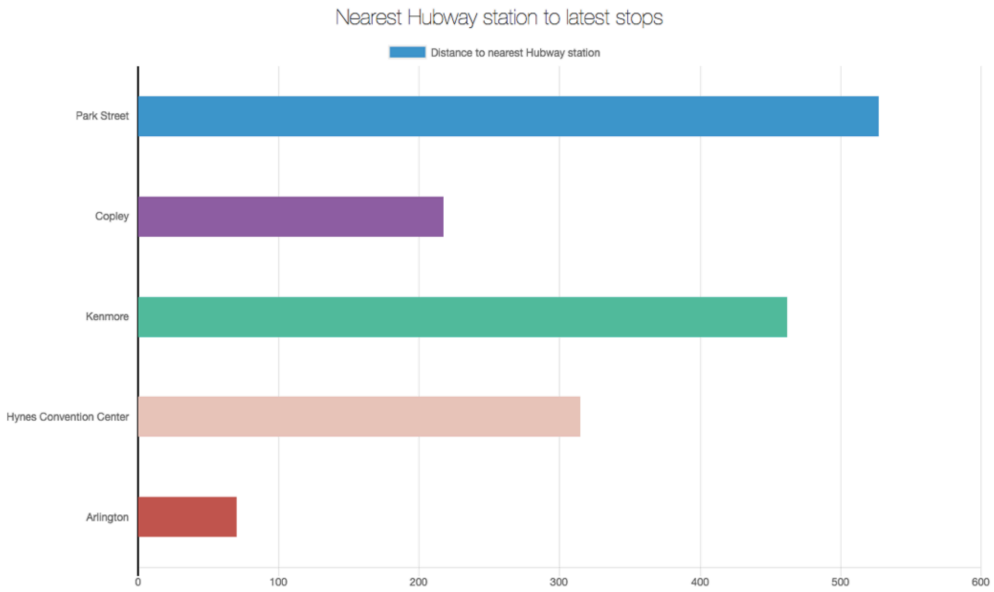

Fig. 4

Our k-means algorithm also included suburbs outside of Hubway's current regions of service. In the interest of increasing punctuality and reducing traffic across the Boston metro area, perhaps Hubway should consider expanding their regions of service.

Future Research:

Going forward we would like to see both an expansion of this analysis to the rest of the MBTA lines as well as the distance between MBTA stations and their nearest Hubway station. We would also like to see other factors, such as bus stops/reliability and underserved communities placed into the consideration of where an ideal Hubway station should be placed.

Running Instructions:

run execute.py If you need to run in trial mode run execute.py --trial