Data Mechanics
Boston Public School Transportation Challenge
Megan Horan, Ryan Chen, Victoria Thomson

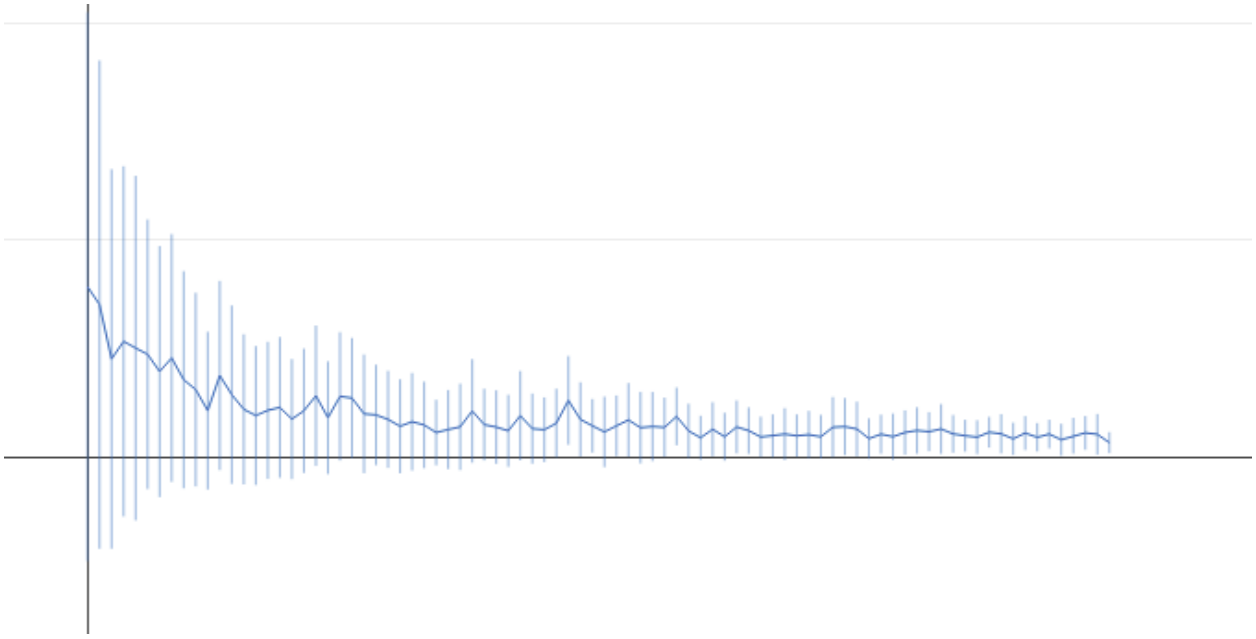
The problem our group chose to tackle was set forth by the Boston Public Schools, in an attempt to optimize school busing systems. This problem was interesting because it provided us with datasets that were accurate and easy to obtain, affording us the opportunity to obtain beneficial results for the lives of students and bus drivers. We chose to focus on the location of bus yards in relation to schools, and to get some useful numbers about the different average distances between students. While these two analyses do not optimize a bus route as a whole, they serve to bolster further research by providing useful representations of data for an ideal bus route.

The data sets we used to formulate our results were the randomly generated data sets that represent the students, the schools, and the bus yards. The ‘students’ data set provided information about the students pick-up location and the schools that they were attending, the ‘schools’ dataset gave the name and location of each school, and the ‘buses’ dataset provided its location and the number of buses in that yard. From these datasets we were able to programmatically extract only the data essential to our algorithms, and so we retrieved the latitude and longitude from the school’s dataset and the number of buses per yard from the bus’ dataset, and the students latitude and longitude of their home addresses. Armed with all of this data, we set out to optimize the bus yard locations and find the most meaningful average distances between students.

In order to analyze the bus yard locations, we used the k-means algorithm to find the hubs of where schools are located. We wanted to cluster the schools because if we could see where the schools were congregating, we would find out the locations for the bus yards that best served the largest amount of schools. If the bus yards were placed in their ideal areas, where they would be serving the most schools they possibly could, then the total travel time from start to finish for those buses would decrease drastically and shave off any unnecessary travel time. Given that motivation, we employed k-means clustering, which will partition n instances, in this case schools, into k number of means, with each school assigned to its closest mean. One question we wanted to find to further our analysis, was how many means is the ideal number? We strove to find the k value that minimized the distance between each school and its assigned mean, ie minimizing the cost for each point. The intuition behind k-means tells us that one or two means will give us a large cost, as many schools would be very far away from their respective means, and a large number of means will give us a small cost, as with more means there is less distance between school and its mean.

Building off that intuition, we wanted to try all the different means and find the ideal mean, which would be a k-value after which adding more means would not decrease the cost significantly. To find the distance between every point and its mean we used Vincenty distance,

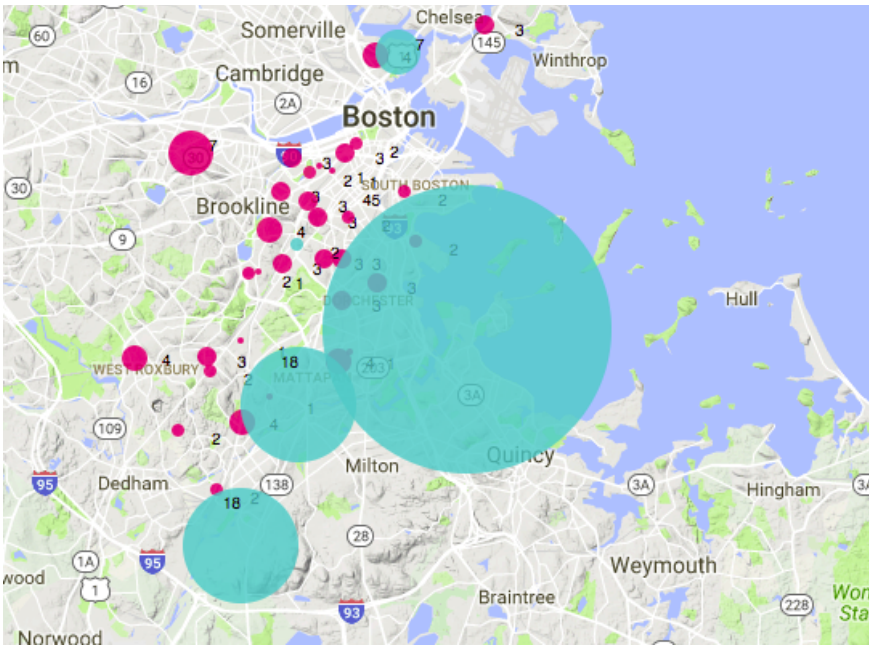
which calculates the distance between two points on the surface of the globe using geodesy. So from our data set we tested values of k from zero to the number of schools and found the overall cost, which was the average distance between each school and its mean divided by the number of schools, for each of those k-values. This graph below is the representation of data collected: on the x-axis is k values and on the y-axis is the average cost between all the schools and their



respective means. The light blue lines branching off the line are the standard deviations of the distance between the schools and their means. One can see that the results from our graph demonstrated what we had predicted: smaller k-values shows that there is greater cost with more points varying in distance from the mean , and a higher k-value shows that there is less cost and that schools are uniformly closer to the means. One can also see that at around halfway on the x-axis, the line begins to taper off, which means after that point adding more means does not reduce to cost significantly. And so the ideal number of means gives us a k-value of roughly half the total number of schools.

Now we can apply the ideal number of means to derive a meaningful visualization of where these ideal bus yards are located. Our k-means algorithm then returns a dataset that has the location of the bus yards and all the closest schools it would service. This graph below shows the visualization of that: the pink circles show where the ideal bus yard, where the radius of each circle is a representation of how many schools it would pick up students for. The blue circles are where the current bus yards are and how many schools they service, using the same function as the pink circles to calculate which schools are closest to which bus yard given the current configuration of bus yards.

This map gives us a good idea of how few bus yards there are currently and how many bus routes start in one area and have to travel to service many more schools than the ideal bus



yard means we found using our k-means algorithm. While it is not feasible to put bus yards where all the pink circles are located, but it gives a good idea of where another bus yard should be. For example, in the area by Northeastern University, given the number of pink circles in that area it might be a good idea to split up huge blue circle and start a new bus yard there to minimize the distance that some buses inevitably have to travel. Overall, this map gives a good indicator of where schools are congregating, and where the bus yards could be constructed to optimize the bus routes by moving the buses closer to the school they have to stop at.

The next analysis we did was to find some meaningful statistics about where students are located, calculating the average distance using a few different metrics. In this transformation we use an r-tree, which are tree data structures that are used for spatial access method, because it was the fastest way to group 80000 or so students in the school system. When calculating these averages we originally were comparing the average distances between all the students, but that took much too long and we realized that it wouldn't give us a good representation of what bus routes would actually look like, since buses generally pick up students in a given area. From there we switched to r-trees, which were more logical for our purposes and took a fraction of the time to compute than what we were originally trying to calculate.

We then found two different distance averaging metrics, with one grouping averaging the distance of the closest 10 students per student, and one grouping averaging the distance between students in a .5 mile radius. We found the distance using vincenty distance, as we did in the k-means algorithm. We found that the average distance between the 10 closest students is 0.439337627979 miles and the average distance between students within a 0.5 mile radius is 0.4736018743886687. This data would be useful for buses to see how far they would have to travel between stops on average, in order for them to envision timing. Also, it would show us where to place optimal bus stops for students, and the data is malleable to group students by different end and start times for schools.

We hope that in the future, other groups or interested parties will take this data and build from it to come closer to optimizing the bus routes. From the k-means algorithm, the bus yards

give us an idea of where to have buses begin and end their journey to minimize the distance buses have to travel on their journey. Our hope is that someone will factor in these yards to begin and find a route based off these ideal bus yard starting locations. Also, anyone considering the time it takes to pick up students in a given radius can use the average distances we found in their calculations to see how long it might take for a bus to pick up x many students in a given area or on an optimized route they found. Overall we were happy that the data was so readily accessible so that we could obtain fecund results to further future research. One can see that there are many paths that groups can go down in hopes of optimizing the bus routes in a way that builds off of the data.