Team members: Declan Halbert, Maxime Gavronsky, Tom Corcoran

Background:

The subway is an essential mode of transportation in New York City. Life in NYC has gotten far more expensive, and the subway is no exception; a ride fare has increased 37.5% (from $2.00 to $2.75) in the last 16 years. To help make life more affordable, we thought it could be worthwhile to rethink subway fares. Taking the London Underground zoning system as an example, we thought that maybe NYC subways could benefit from fare zones. In London the cost of a subway ride depends on the distance of your travel (how many zones you cross). We thought that we could rezone the subway system, but instead of creating zones solely based on location, we could add economic data from each NYC neighborhood (NTA) into the equation. By factoring in average neighborhood income, we are making the subway more affordable for the underserved communities within NYC.

How was the data obtained?

Data USA provided information about the average income in 2016 per NYC census tract. We utilized NYC OpenData for geographical information about each census tract in order to insert census data into each tract's relevant (each census tract belongs to exactly one neighborhood). NYC OpenData also provided data on neighborhood population from the year 2000. Ny.gov provided data on 2013 subway exit/entrance location (before the expansion of the Q train to Lenox Hill and Yorkville). Lastly, we obtained commuter information from NYC Planning which was originally obtained from the 2012-2016 American Community Survey.

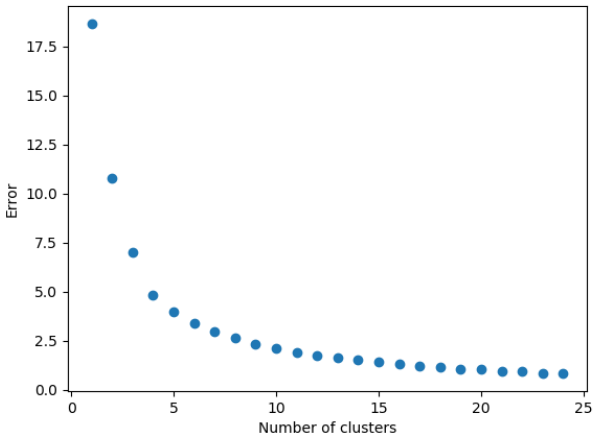
How was the data used?

Neighborhood population, location, and commuter information was combined based on matching NTA code. Subway stations were merged into their respective neighborhoods using an algorithm that determined whether or not a single point (subway station) was inside of a polygon (neighborhood). The two census data sets were merged together, omitting all tracts that had a population of 0. Lastly, all of above census information was merged into each tract's respective neighborhood. During this merge, average income per neighborhood was also calculated by summing the census income information, and then dividing by the number of census tracts.

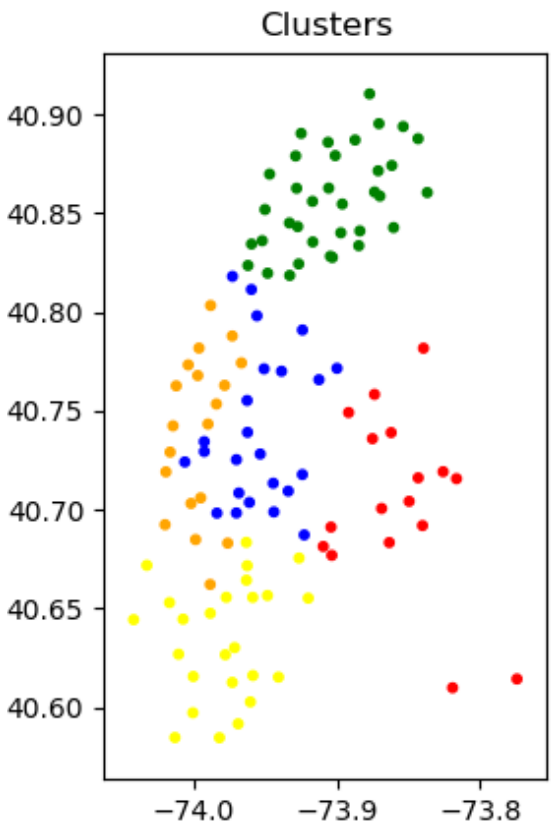
Determining Zones:

We utilized the k-means algorithm for the creation of each zone for the new fare system. The algorithm used the latitude, longitude, and income of each neighborhood. To be able to run k-means with these variables, we first had to scale them all into numbers in the range between 0-1 using the MinMaxScaler. After running k-means on this scaled data, we plotted the error as a function of *k* to see which value of *k* minimized the error. (See Figure 1)

The curve begins to flatten at $x = 5$, so we determined that the optimal number of zones, without having too many, would be 5. Then, we plotted the neighborhoods by their original un-scaled latitude and longitude points, color-coded with their respective zones. Using latitude and longitude in addition to income for k-means gives a better grouping for the zones instead of basing it solely off of income. (See Figure 2)



(Figure 1)



(Figure 2)

Limitations in Determining Zones:

In order to run k-means, each neighborhood is represented by one income, one latitude, and one longitude. Each neighborhood contains a "multi-polygon" consisting of a collection of points roughly outlining the border of each neighborhood. In the optimization we arbitrarily chose the maximum (latitude, longitude) point for each neighborhood, and then proceeded to run the algorithm.

Determining Fares:

Instead of finding a flat fare for each zone, we decided to calculate a fare for every possible route combination. For example, if there were 3 zones, we would determine pricing for the following routes: {zone 1 ↔ zone 1, zone 1 ↔ zone 2, zone 1 ↔ zone 3, zone 2 ↔ zone 2, zone 2 ↔ 3 zone, zone 3 ↔ zone 3}. It is important to note that the zones are ordered by descending average income. In order to find the new route fares, we created a constraint satisfaction problem, and utilized a z3-solver to solve said problem. There are four main types of constraints we created:

(1)

The first constraint calculates the overall revenue that each route produces by multiplying an (NTA's population) * (the percentage of commuters that use public transportation to get to work every day) * (2.75)[1]. We ensured that the new prices for each route, when multiplied by the same population and commuter percentages, would still produce the same revenue as the current subway fare.

(2)

To ensure a balance in revenue between each route, created a constraint in which the new fares cannot create a route that accounts for more than *x* percent of daily MTA revenue. Similarly, we created the opposite constraint to ensure that no one route accounts for less than *y* percent of daily revenue. With a system of *k*=5 zones, we determined that x should be between 15% and 30%. Likewise, it is impossible to satisfy

[1] $2.75 Current price of a single subway ride.

constraints with y > 7%[2] and y should be > .5%. This was constraint was created so that the MTA would need to rely too heavily on one particular route for its revenue.

(3)

We guaranteed that a route with a lesser sum (zone 1 ↔ zone 2; sum = 3) must have a more expensive fare than a route of a higher sum (zone 4 ↔ zone 3; sum = 4).

(4)

Lastly, we concluded that the fare for the most expensive route (zone 1 ↔ zone 1) may be only at most *x* times as large as the cheapest route (zone *k* ↔ zone *k*). This constraint in combination with constraint (3) would create fares that hover closely to one another. In our experience, ideally $1.25 < x < 1.75$.

Limitations in Determining Fares:

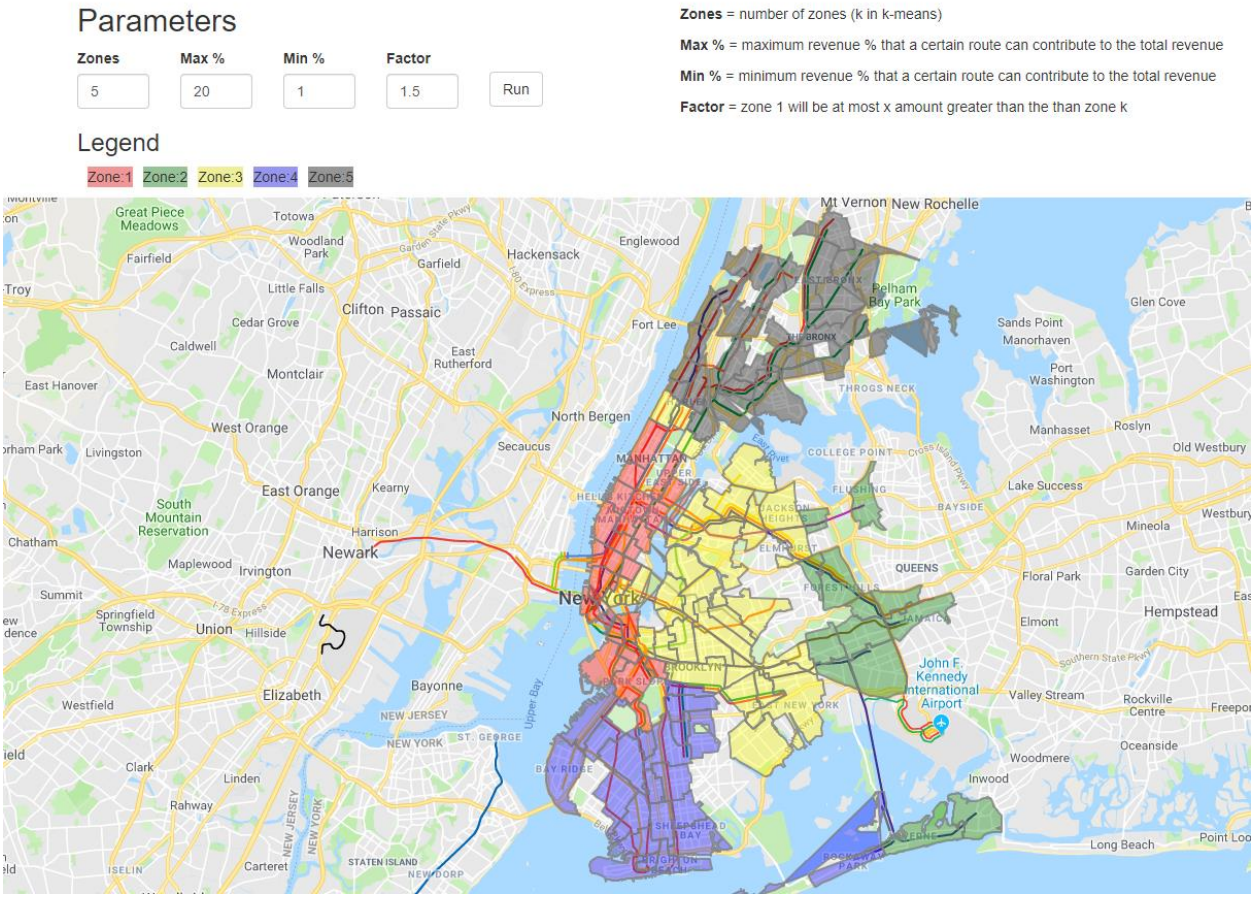
There are a few data limitations in determining the route fares. We could not find subway passenger data that was detailed enough, so we had to create our own assumption on revenue of the MTA. While average daily subway revenue does exist for the entire city, it does not exist on a per-neighborhood level (which is what we were looking for). In calculating the revenue for each neighborhood, we multiplied the number of public transit riders (including busses) of each neighborhood by population. While the percentage of New York citizens commuting on a bus is not very significant, we do not account for it since the data we found was not detailed enough. Lastly in calculating revenue we only account for commuters, which obviously excludes weekend subway riders, and tourists, etc.

Conclusions:

In summary, we created a service that allows users to customize a *k* (number of zones) as well as the variable values for constraints (2) and (4). If the user entered constraints satisfy the problem, a color-coded map is displayed, with each neighborhood belonging to a certain zone (See Figure 3). The user can click on each neighborhood, and a small text box will pop up containing the neighborhoods average income, the average income of the zone to which the neighborhood belongs, and a list of fares for every possible route including the NTA's zone (See

[2] *K =5* creates 15 route combinations, 100/15 = 6.67. Therefore, the routes can never all contribute more than 6.67% percent to the overall revenue.

Figures 4 & 5). If the user enters a combination of constraints that are unsatisfiable, the user will be notified, and will be allowed to re-enter a new set of constraints.



(Figure 3)

We successfully achieved our goal of rezoning NYC's subway system and making a more affordable transit system for lower income neighborhoods. Depending on the parameters that the user inputs you can receive a more or less realistic zoning system. In the final product of this project, we decided to omit Staten Island due to its low percentage of subway ridership, since most residents of Staten Island have cars. In addition, the Staten Island Railway is seen as a separate entity from the subway system. The uncolored portions of the map in Figure 3 are all neighborhoods that do not contain a single subway station, and are therefore ignored in the final product of this project.

Parameters

Zones	Max %	Min %	Factor	
5	20	1	1.5	Run

Zones = number of zones (k in k-means)

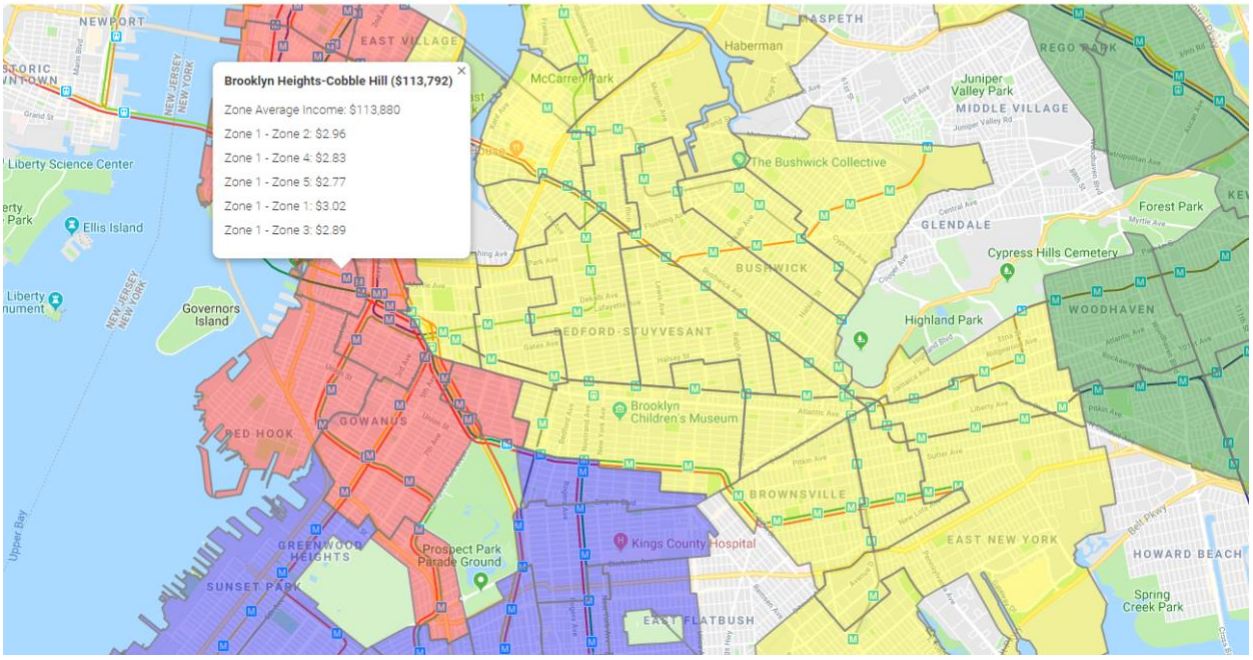
Max % = maximum revenue % that a certain route can contribute to the total revenue

Min % = minimum revenue % that a certain route can contribute to the total revenue

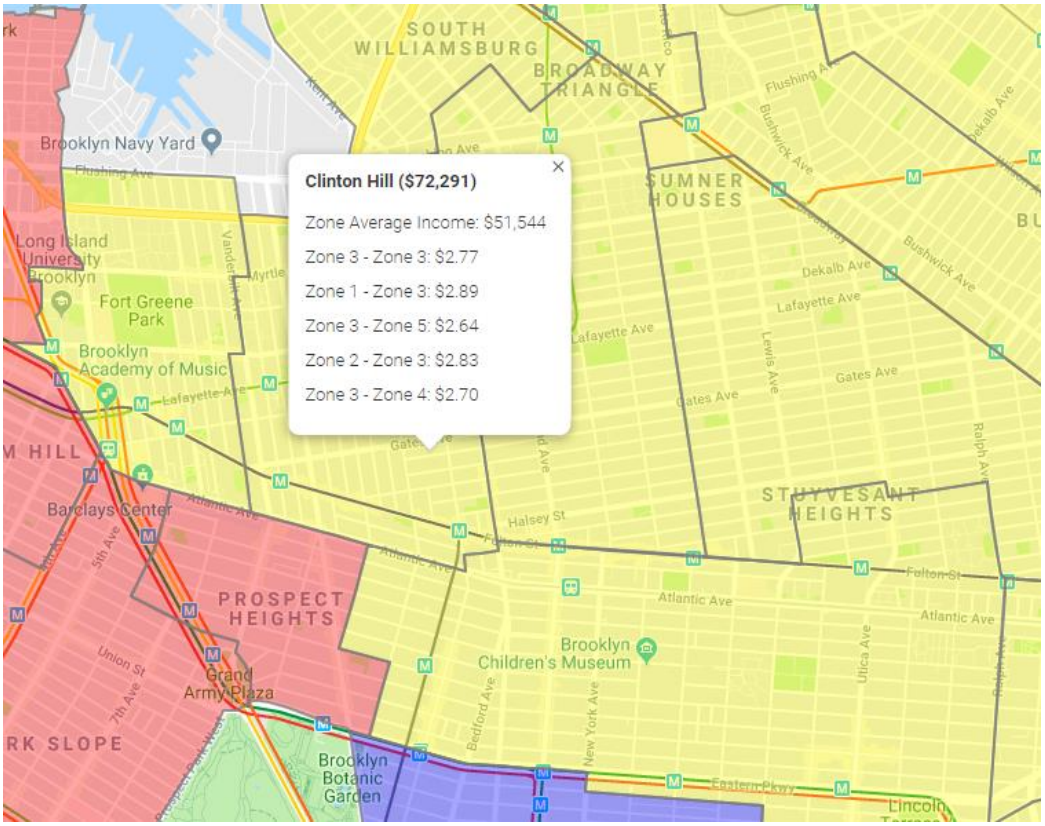
Factor = zone 1 will be at most x amount greater than the than zone k

Legend

Zone:1 Zone:2 Zone:3 Zone:4 Zone:5



(Figure 4)



(Figure 5)

Future Work:

More data (and time)! The pricing for routes could be even more precise with more time to search for data. If more precise transit data where to become available, we could use it to better service the citizens of New York. We could find riding patterns, and popular transit routes and factor that into our pricing. As previously stated, with more data on ridership within specific neighborhoods that could be factored into the creation of zones and their pricing. We could also look into a deeper volume of type of travel (busses, taxis, Ubers, etc ...) to help determine which neighborhoods need more support from the cheaper subway fares. In addition, we could run k-means with more dimensions (e.g. different demographics or age ranges in each neighborhood) to determine better zones for reduced fares.