Final Deliverable, Transit Equity Team 1

Jamie Dela Cruz, Kenneth Lin, Jun Li, Jun Hao Lei

Goal

Background

Our client is State Representative Nika Elugardo. She represents the 15th Suffolk District in the Massachusetts House of Representatives, encompassing the towns of Boston and Brookline. She would like to explore the feasibility of making certain bus lines free for both the Massachusetts Bay Transportation Authority (MBTA) and other regional transit authority (RTA) lines in Massachusetts. Our analysis supports and provides a basis for legislation that she is drafting, the goal of which is to improve transit equity in Massachusetts. Transit equity involves removing barriers to usage of public transportation for low-income communities and communities of color (Gardner, 2020; Vaccaro, 2020). Along with socioeconomic benefits, achieving transit equity may also help government agencies increase the efficiency of their public transit and support the local economy (Worcester Regional Research Bureau, 2019).

Steps

As per the [Project Description], the initial goal was to examine ridership data to identify which bus routes would have the greatest positive effect on low income riders if admission were made free. In order to determine which routes to make free, we needed to first examine the stops that make up those routes. Specifically, we needed to calculate the population and average income of commuters that would benefit from making an individual stop free, then connect the dots to get route-level information. The steps in the description are as follows:

1. Collect data - create a spreadsheet of all the different bus stops in Massachusetts including MBTA, Regional Transit Authorities, and City/Town buses.
2. Assign an income level to each stop based on the census tract data.
3. Determine average fare for each transit stop based on fares.
4. Calculate bus ridership for each transit authority.
5. Identify which bus routes, stops, or zones would have the most positive effect on low income riders if free. Identify which towns would be impacted?
6. Generate visualizations: TBD with client using software such as ArcGIS or tableau as a final deliverable along with the list data.

Based on initial conversations with BU Spark!, both teams decided to focus our analyses on the MBTA as it had a robust developer portal with public data suitable for our project. The initial plan was to work to complete an analysis of the MBTA before moving on to subsequent RTA's; however, we quickly realized that this approach was impeding progress and split the transit authorities up. Our team focused exclusively on analysis of the MBTA while Team 2 worked on the other Massachusetts RTA's.

For Deliverable 1 and Deliverable 2, we completed Steps 1-5 per our original approach following the plan in the Project Description. This process is described in Appendices I and III.

However, for the reasons outlined below, we adopted an alternative approach focused on Steps 1, 2, 5, and 6 for [Deliverable 3] and this Final Deliverable:

We requested that the client provide an example of a headline that they would like to make at the end of our analysis. Our client provided the following response:

> *"The resulting map shows dot density of population with color coding by income level. When you overlay this with transit lines and bus lines you can see the massive barriers to ridership. There are at least three problems with the current system. One is the poorest people are the users of buses that are less frequent and less maintained (data available on this) two is poor people can't afford the buses that are near them (as per above analysis). There are a lot of poor people who do not live near buses including those that live near trains and can't afford those.*
>
> *These would be some of the basis for argument that we should have expanded bus lines and make them free so everyone uses them the way we do with libraries and public parks."*

Previously, we had been analyzing ridership data using the "average ons" for each stop (the number of people that board a particular stop at a given time) to calculate the number of riders that would be impacted by making fares free. However, we realized that there were two main issues with this approach:

1) A single rider could be double-counted in the data if they board at different stops throughout the day.
2) Analysis of current ridership data doesn't capture the true effect of making one or more bus routes free. Ridership data only reflects people that are presently able to afford the bus system and excludes those who are unable to afford it.

During meetings with our BU Spark! PM, we extensively discussed the impact that free bus routes would have on low-income workers. Based on these conversations, we decided to analyze the total number of people who use non-car methods of transportation to commute to work (i.e. public transportation, walking, and other means). They would benefit the most from making bus routes free, and benefits to commuters provide a convincing argument in favor of transit equity legislation. In collaboration with BU Spark!, we agreed on two possible final results that adopt the same fundamental approach: a baseline analysis and a "reach" analysis.

For the baseline analysis, we assigned population and income levels to stops based on the tract that they are located in. Further elaboration and the results of this analysis are provided in Appendix I.

For the "reach" analysis, we assigned population and income level to each stop by calculating a weighted average of the surrounding tracts' populations and incomes. In this approach, a 0.5 mile buffer is drawn around each stop to reflect the approximate area that a particular stop serves. This calculation accounts for the fact that people likely use the stop that they live closest to, even if they live in a different tract than the one the stop is located in. Additionally, many tracts do not have a corresponding stop, so the commuters living in those tracts are not reflected in the baseline approach. Overall, this approach provides a more accurate depiction of commuters served by each stop. The main portion of this deliverable focuses on this approach.

Data Collection

To achieve our goal, we needed data on the median household income level and population of areas in Massachusetts as well as information about MBTA bus stops and routes. This information was scattered throughout various shapefiles and CSV datasets from the United States Census Bureau, MassGIS (Bureau of Geographic Information), and the MBTA Open Data Portal. The data collected for our analysis is described below.

1. MassGIS Data: Datalayers from the 2010 U.S. Census ("tract shapefile"): This shapefile was used to obtain the geographical data of Massachusetts tracts. Tracts are represented as polygon and multipolygon shapes in the spatial data.
2. MassGIS Data: MBTA Bus Routes and Stops ("bus stops shapefile"): This shapefile was used to obtain the geographical data of MBTA bus stops and bus routes. Bus stops are represented as points in the spatial data, whereas bus routes are represented as lines.
3. MBTA Bus Ridership by Trip, Season, Route/Line, and Stop ("ridership CSV"): This CSV file of past ridership data was used to determine which routes pass through each stop.
4. US Census 2018 ACS 5-Year Estimates ("census CSV"): This CSV was used to obtain population data by commuter category (public transportation, walking, other non-car). Background information can be found here. The linked dataset has the 2018 ACS 5-Year Estimates filtered on Income (Households, Family, Individuals) and grouped by tract in Massachusetts.
5. MassGIS Data: Massachusetts House Legislative Districts ("House Representatives shapefile"): This shapefile was used to obtain the geographical and State House representative data of Massachusetts districts. Districts are represented as polygons and multipolygon shapes in the spatial data.

Analysis

Our analysis sought to answer these strategic questions from the client:

1. What bus routes and stops, if made free, would most benefit low income riders in Massachusetts?
2. Which towns (and districts) would most benefit by a policy change to the fare change to these routes?

When referring to “population,” we are referring to the total number of workers 16 and over commuting to work via public transportation, walking, and other non-car methods living in each tract. These numbers were found from the census CSV.

The following explanation is a walkthrough of the code found here.

Stop-level data

Our first step was to draw a circle of 0.5-mile radius around each bus stop to find the core area that each bus stop serves (as per transit-friendly definitions provided by our BU Spark! contacts). Because our manipulations involved geometric attributes, we utilized the Python library GeoPandas. A GeoDataFrame is analogous to the DataFrame of Pandas but with a ‘geometry’ column, which includes GeoSeries objects (Point, Polygon, Multipolygon). It has the added utility of built-in functions and methods that can read shapefiles and manipulate geometric data. For example, we drew a circle around each stop using the function *geopandas.buffer()*. Something to note is that GeoPandas measures distance in meters by default, so we converted miles into meters. Then, we found the intersections of each circle and its surrounding tracts using *GeoSeries.overlaps()* and weighed the median income and impacted population by the proportion of area intersected.

When finding the proportion of intersection for median household income calculations, the proportion of intersection for a tract *i* that intersected with the circle of stop *j* was:

$$proportion_{ij} = \frac{overlapping\_area_{ij}}{area_{circle}}$$

In other words, if Tract *i* is the only tract that Stop *j* intersects, then the median household income of Stop *j* is that of Tract *i*. If Tract *i* only makes up 50% of Stop *j*’s circle and Tract X makes up the other 50%, then the median household income of Stop *j* is:

$$income_j = 0.5 * income_{Tract\ i} + 0.5 * income_{Tract\ X}$$

The calculation of weighted impacted population differs a little:

$$proportion_{ij} = \frac{overlapping\_area_{ij}}{area_{tract\ i}}$$

In other words, if Tract *i* has a population of 100 people, and if Stop X’s circle covered 50% of that tract, then the population served by Stop X is 50 people.

To summarize: given that the circle for a stop intersects with n tracts with proportions of area $\{p_1, p_2, ..., p_n\}$, median income $\{in_1, in_2, ..., in_n\}$, and population $\{pop_1, pop_2, ..., pop_n\}$, then the weighted median household income and impacted population will be calculated as below:

$$income_{weighted} = \frac{\Sigma^n{}_{i=1}\ in_i \times p_i}{\Sigma^n{}_{i=1}\ p_i}$$

$$impacted\_pop_{weighted} = \sum_{i=1}^{n} pop_i \times p_i$$

An example is shown in Figure 1. When calculating population, we assume that people are uniformly distributed within a tract.

If a tract had a null value for median household income, we assigned its corresponding proportion to 0 so it would not contribute to the calculation. Because we divide the calculated income by the sum of proportions, we avoid skewing the data if a null value occurs. If we didn't remove null values from the denominator of this calculation, our weighted average would be biased towards 0. This only applies to the calculation of weighted median income because there are no null population values in the 2018 census data.

Once we calculated the income for each stop, we assigned it an income level according to this data from [Pew Research](), which stratifies income levels into 5 groups:

LEVEL	**INCOME GROUP**	**INCOME/\$**
0	Lowest income	31,000 or less
1	Lower-middle income	31,000 - 42,000
2	Middle-income	42,000 - 126,000
3	Upper-middle income	126,000 - 188,000
4	Higher-income	188,000 or more

All the data manipulated above uses the *'EPSG:26986' Coordinate Reference System (CRS)* because geometry data under different coordinate systems is incompatible. The resulting data was exported as a shapefile named stops.shp.

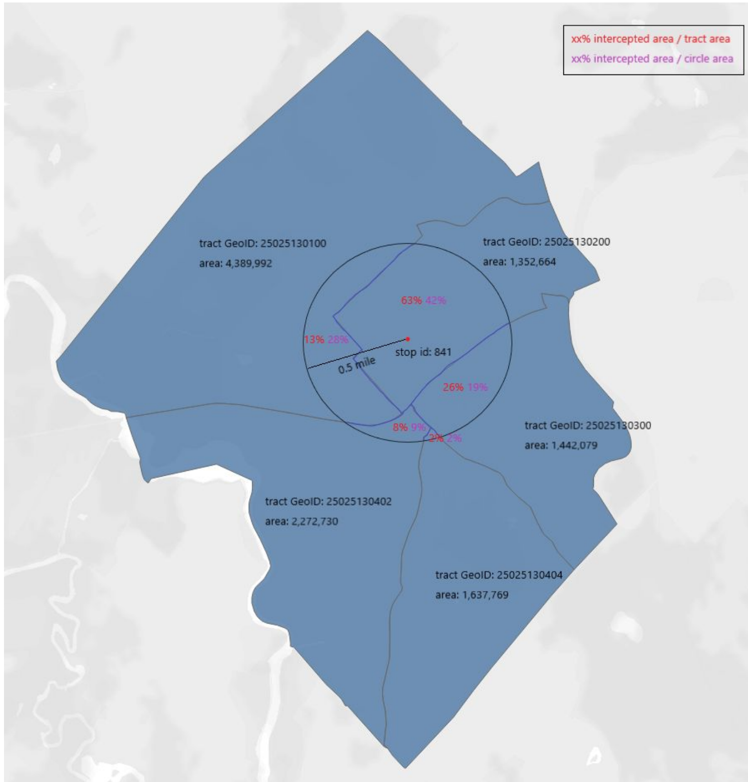

Figure 1. For stop 841, a 0.5 mile buffer is drawn to reflect the approximate area that the stop serves. The ‘serving area’ overlaps 5 tracts: 25025130402, 25025130300, 25025130200, 25025130100, and 25025130404. The percentage overlapped of each tract is: 8%, 26%, 63%, 13%, and 2%. These overlaps comprise these percentages of the circle’s area: 9%, 19%, 42%, 28%, 2%. The impacted population and income level can be calculated using this, under the assumption that population density is uniformly distributed within each tract.

Route-level data

Now that we calculated the population and median household income level for each individual stop, we needed to determine which routes these stops are in. Though we were unable to find a source that gave a perfect matching of stops to routes, our next-best solution was to use the ridership CSV which had data on “trips”: the path a particular rider took along a particular route, starting from a particular stop. We were able to collect all the stops for which there was ridership data as well as their corresponding routes from this dataset to generate an adequate mapping of stops to routes. Though this resulted in some stops having no routes attributed to them, this had a negligible impact on our analysis. With this information, we added a column in the previously-generated dataset that lists all routes that each stop passes through.

For the rest of our analysis, we focused on stops that had income level 0 and 1, i.e. those with weighted median household incomes in the lowest and lower-middle income groups. We used several functions to determine the routes with the highest counts of low-income stops.

District-level data

Next, we wanted to find out what districts had the highest number of low-income stops. We imported the House Representatives shapefile and our generated stops shapefile into the visualization software Tableau. We could then easily use Tableau to find where stops and districts overlap and export this data for analysis (for an example of using Python's GeoPandas to find the location of stops inside geographic polygons, see the function *tract_for_stop* here). Repeating a similar process as the route-level data, we found the number of low-income stops within each district as well as the state representative for these impacted districts.

Town-level data
Lastly, we created several functions to calculate the number of low-income stops within each impacted town. Because the data was on a much larger scale for this level of analysis, we also decided to calculate the percentage of low-income stops that each town has.

Results

Routes

The top 5 routes that pass through the most stops in low-income areas are Routes 19, 45, 28, 22, and 14. See Figure 2 for more details. For a spreadsheet view of the final stop data, see here, and see here for the final route data.

Figure 2. The top 10 routes that pass through the greatest number of stops in low-income areas. Routes 442 and 441 tie for #9, and routes 66 and 8 tie for #10. Income level 0 and 1 represent the lowest and second lowest income levels, respectively.

Figure 3 provides an estimate of the impacted population broken down by route. However, because stops along the same route may have radii that overlap, the population at those intersections have been counted more than once.

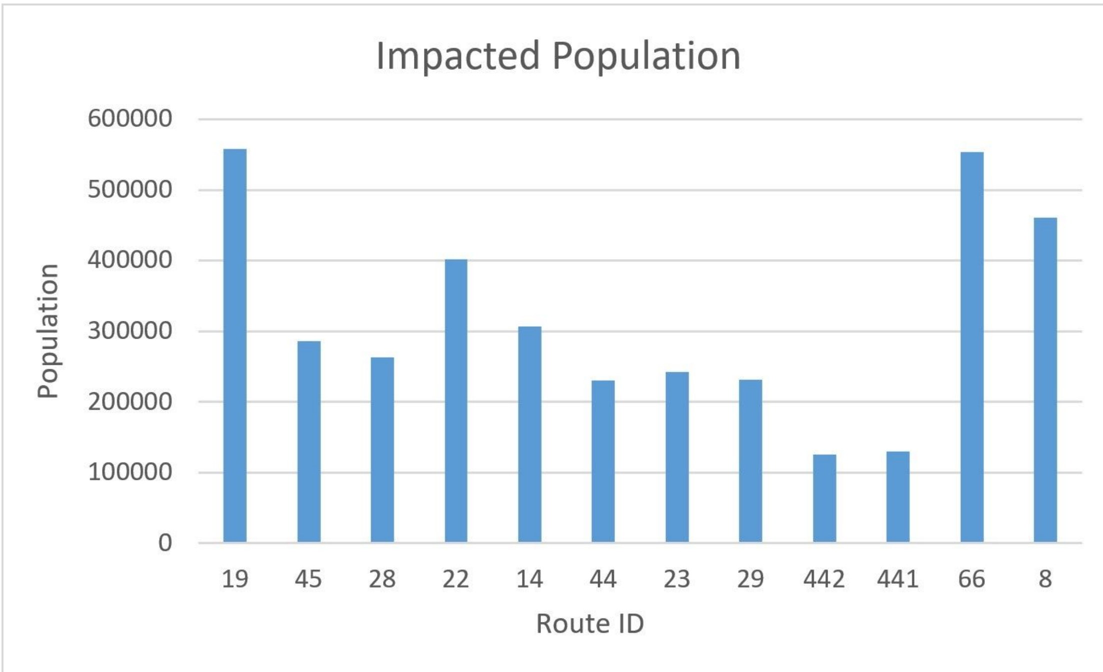

Figure 3. The population that would be impacted if that route were made free, calculated for the top 10 routes that pass through the greatest number of stops in low-income areas. Note that this is estimated using the data for each stop within a route.

We utilized Tableau to create a map of our data that fulfills the visualization requirements described in the *Goals* section, with an interactive version here. Each dot represents a stop, color coded by income level and sized according to its impacted population. Clicking on a stop highlights its information and shows all other stops with the same income level. Search bars on the side allow one to locate a given route or highlight stops in a certain district. Figure 4 shows the top 5 routes highlighted.

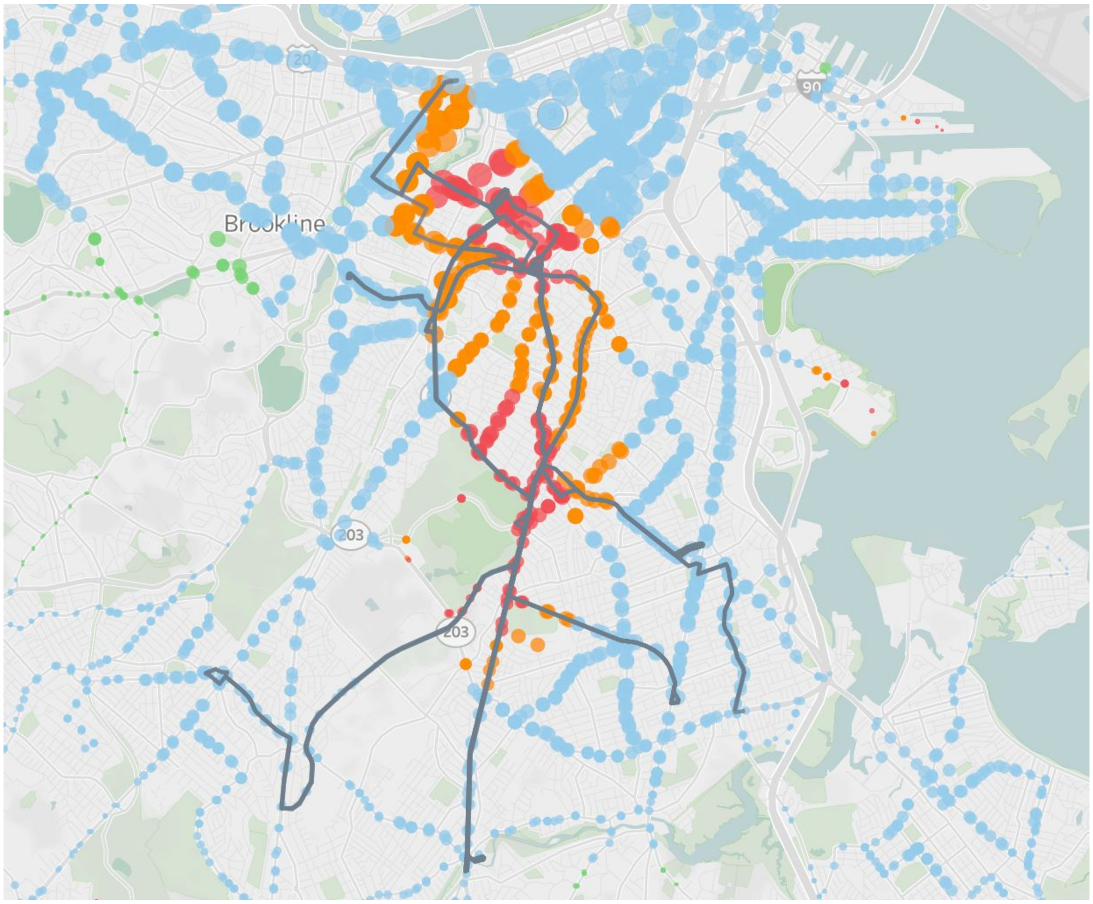

Figure 4. Image showing the locations of the top 5 routes listed above. An interactive form of this image is here.

Districts

The top 5 districts with the most low-income stops are 7th Suffolk, 11th Suffolk, 11th Essex, 6th Suffolk, and 5th Suffolk. See Figure 5 for more details. Visit here for an interactive map that overlays all stops over a map of the districts. Clicking on a stop will highlight the district it is located in along with all other stops located within that district.

Figure 5 lists all impacted districts, and Figure 6 highlights the stops located within the top 5 impacted districts.

district	level_0	level_1	total	rep
7th Suffolk	52	57	109	Chynah Tyler (D)
11th Suffolk	42	30	72	Elizabeth A. Malia (D)
11th Essex	5	43	48	Peter L. Capano (D)
6th Suffolk	20	20	40	Russell E. Holmes (D)
5th Suffolk	3	35	38	Liz Miranda (D)
9th Suffolk	9	17	26	Jon Santiago (D)
10th Essex	0	18	18	Daniel H. Cahill (D)
2nd Norfolk	7	9	16	Tackey Chan (D)
19th Suffolk	0	16	16	Robert A. DeLeo (D)
1st Suffolk	3	10	13	Adrian C. Madaro (D)
4th Suffolk	5	6	11	David Biele (D)
15th Suffolk	0	11	11	Nika Elugardo (D)
14th Suffolk	4	5	9	Angelo M. Scaccia (D)
8th Essex	0	8	8	Lori A. Ehrlich (D)
16th Suffolk	3	2	5	RoseLee Vincent (D)
13th Suffolk	1	1	2	Daniel J. Hunt (D)

Figure 5. Table showing all districts with at least 1 low-income stop. Columns ‘level_0’ and ‘level_1’ show the count of income level 0 and income level 1 stops within each district, respectively. The ‘total’ column sums the numbers in ‘level_0’ and ‘level_1’. The ‘rep’ column shows the full name and party affiliation of that district’s representative.

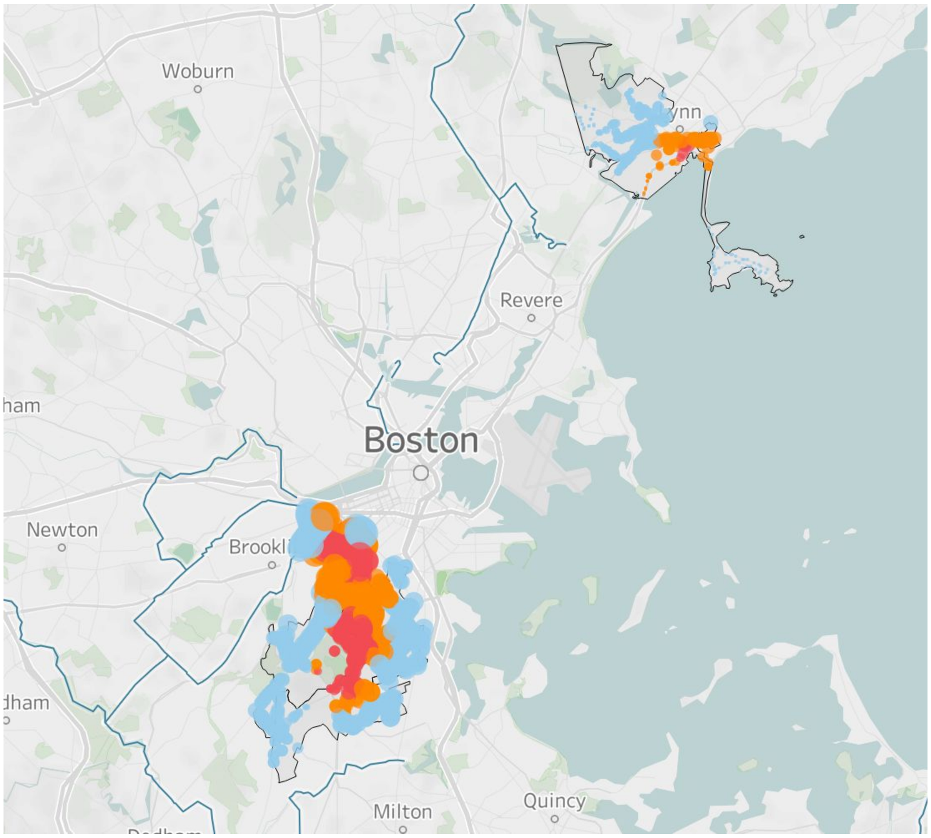

Figure 6. Image showing the stops located in the top 5 impacted districts. An interactive version of the image can be found here.

Towns

The top 3 impacted towns are Boston, Lynn, and Revere. See Figure 7 for more details.

town	level_0	level_1	total	percentage
BOSTON	139	192	331	16.708733
LYNN	5	67	72	17.307692
REVERE	3	16	19	10.734463
QUINCY	7	9	16	3.245436
WINTHROP	0	2	2	2.197802
NAHANT	0	2	2	5.128205

Figure 7. Table showing all towns with at least 1 low-income stop, sorted by most to least number of low-income stops. Columns 'level_0' and 'level_1' show the count of income level 0 and income level 1 stops within each town, respectively. The 'total' column sums the numbers in

‘level_0’ and ‘level_1’. The ‘percentage’ column shows what percentage of all bus stops in a given town are low-income.

Challenges

1. **Package Installation:** We encountered significant issues installing and using the Python package Geopandas, which was required for reading, analyzing, and manipulating geospatial data (e.g. shapefiles). Geopandas required numerous manual installations of dependencies, which often had dependencies of their own or threw errors. Even after a “successful installation,” we often ran into errors when trying to do specific tasks that required a complete reinstallation of Geopandas (and its dependencies) as well.

2. **Sparsity of data:** Although we are certain that a true mapping of MBTA bus stops to routes exists (for internal use by the MBTA), BU Spark! and our team could not find any such dataset publicly available on the Internet. As a workaround, we used the approach detailed in the analysis section above. However, this resulted in 140 stops which were not mapped to any routes. We hypothesize that ridership information was not collected for these particular routes or stops. Luckily, the vast majority (138 of 140) of affected stops were assigned income levels >= 2 by our calculations, so we believe that the absence of this data has a negligible impact on our final results.

3. **Non-unique ID’s**: Throughout the project, we operated under the assumption that tract IDs, 6-digit numeric codes assigned to each tract, uniquely identified individual tracts. We later discovered that this field was not unique and 16 tracts in Massachusetts were identified by only 7 unique IDs, leading us to improperly assign income and population data for many stops. To resolve this, we revised all of our code that previously used tract ID as an identifier, instead merging or performing other operations based on a unique census identifier called a Geo ID (an 11-digit code composed of a tract ID prepended by a county and state code).

4. **Null Values:** Census data pertaining to certain tracts’ median household incomes was missing, even for years prior to 2018 (the most recent available census data). To remedy this, when finding the weighted average income and impacted population of tracts affected by a particular stop, we specifically excluded null values from our calculations (the specific change is described in greater detail in the analysis section).

Next Steps

1. **Removing intersections between overlapping radii:** One issue with our current calculations is that, in route level calculations, it isn’t entirely accurate to state that the total number of people impacted by making a particular route free is equal to the sum of the impacted population of each stop along the route. Since we draw a 0.5 mile radius around each stop to calculate the weighted average affected population, if bus stops along a route are closer than 0.5 miles from each other, these radii may overlap. In these

instances, we'd actually end up *double-counting* the people represented by the intersections of the radii and overstate the total number of affected people. In the future, we'd like to identify each such intersection and divide their affected populations in half when finding route-level data.

2. **Request a true mapping of stops to routes:** Given more time, we would have liked to directly request the mapping of stops to routes from the MBTA's developers. We only realized that our workaround was imperfect mid-way through the semester, at which point BU Spark! advised us against reaching out so that we could complete our analysis on time, but it would be helpful for future calculations to use the complete mapping.

3. **Determine the financial costs of making a particular route free:** We began an analysis of the revenue lost by making a particular stop free, which would then have led into an analysis for each route. However, after feedback from the client, we didn't complete this analysis in favor of other goals that the client placed greater emphasis on within our short time frame. Future examination of financial feasibility will be essential for the client's inevitable push to pass transit equity legislation.

Works Cited

Gardner, B. (2020, August 13). *Transit Equity: Research and Advocacy Inspires Government Action. Data-Smart City Solutions.* https://datasmart.ash.harvard.edu/news/article/transit-equity-research-and-advocacy-inspires-government-action

Vaccaro, A. (2020, January 6). *The wild idea of making MBTA buses free is gaining traction*. BostonGlobe.Com. https://www.bostonglobe.com/metro/2020/01/06/the-wild-idea-making-mbta-buses-free-gaining-traction/A5mw1bOfykqS9JdY6vXNIJ/story.html

Worcester Regional Research Bureau. (2019, April). *The Implications of a Fare-Free WRTA*. http://www.wrrb.org/wp-content/uploads/2019/05/WRRB-FareFree-Transit-Report.pdf

Appendix I - Baseline Analysis

Goals

Before our final results detailed in this deliverable, we went through several iterations of our analysis. As an initial proof of concept, we assigned population and income levels to stops based on the tract that they are located in. For example, let's say that Tract Y has a population of 100 and an income level of 1. If some Stop X is located in Tract Y, then Stop X's corresponding population and income level would also be 100 and 1, respectively. Had we been unable to do the 0.5-mile radius calculation, this stop-to-tract analysis would provide us with sufficient information to fulfill the client's request.

Data Collection

We used the same sources as listed in the main Data Collection section previously. However, instead of directly downloading the census data from the US Census 2018 ACS 5-Year Estimates website, BU Spark! provided us with a function that gathered the tract IDs and median household incomes of each tract ("tract income data") using an API call, although we later confirmed that this income data was identical to the data in the 2018 census.

Analysis

This baseline analysis has 2 key differences from the final analysis previously described:

1. We did not draw a circle around each stop to determine the impacted area.
2. All analyses used tract IDs instead of Geo IDs. As described in the *Challenges* section, this approach led to several errors in this baseline analysis.

For each stop in our data, we added a column indicating which tract it was located in. We used GeoPandas' *.within()* function to find which stops were within each tract. This allowed us to assign a median household income to each stop by matching to the tract income data on tract ID. We then assigned an income level to each stop using the same Pew Research income levels in the main analysis.

To assign populations to stops, we first used data from the census CSV to sum the number of workers 16 and over commuting to work via public transportation, walking, and other non-car methods living in each tract. We then assigned each stop the calculated population of its tract by matching on tract ID.

Now that we had stop-level data, we needed to determine which routes these stops are in. We used the same method as our main analysis to create a column that lists all routes that each stop passes through. We estimated the impacted population of each route by summing the impacted population of each stop along that route.

For the rest of our analysis, we focused on stops that had income level 0 and 1. We used several functions to determine the routes with the highest counts of low-income stops.

To view the code used to assign stops to tracts and incomes to stops, see here. To see how we found the impacted population for each stop, see here.

Results

The top 5 routes that pass through the most stops in low-income areas are Routes 19, 22, 8, 45, 28, and 44 (Routes 28 and 44 tie for #5). See Figure S1 for more details. For a spreadsheet view of the final stop data, see here, and see here for the final route data.

Figure S2 provides a rough estimate of the impacted population broken down by route. However, because stops for a given route may belong in the same tract, a tract's population may have been counted more than once.

Figure S1. The top 10 routes that pass through the greatest amount of stops in low-income areas, using the baseline calculations. Income level 0 represents lowest-income, and income level 1 represents lower-middle income.

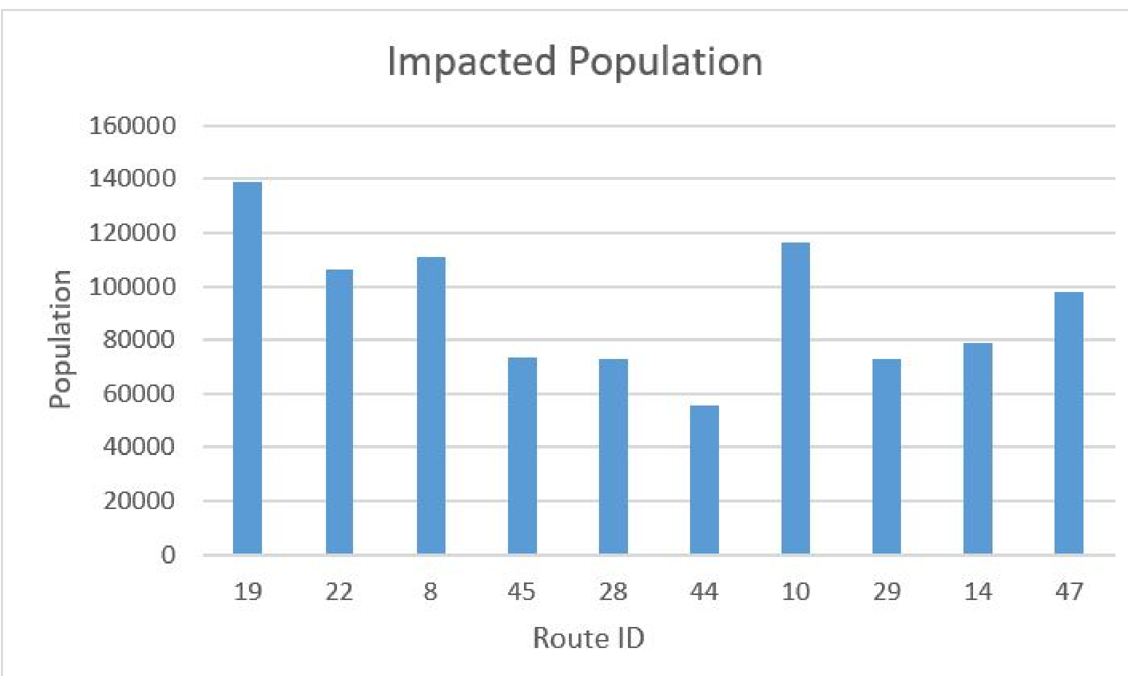

Figure S2. The population that would be impacted if that route were made free, using the baseline calculations. Note that this is calculated using the data for each stop along a route, so this is an estimate.

Again, we used Tableau to create a map of our data that fulfills the visualization requirements described in the *Goals* section, with an interactive version here. Each dot represents a stop, color coded by income level and sized according to the impacted population. Some income values are Null or -1 because some tracts have missing income values, causing their stops to have missing income values. However, the impact of this on our analysis was negligible.

Appendix II - Revenue Calculations

Goals

The initial Steps 3 and 4 were to calculate bus ridership for the MBTA and the average revenue earned by each stop. This would answer the following strategic question posed by the client: What would the cost be to the MBTA and regional transit authorities for each proposed free bus route/stop/zones (based on ridership and fare costs)?

Data Collection

In addition to the sources in the main Data Collection section, revenue calculations required that we use additional data:

1. routes ("routes CSV") = This CSV file was used to obtain the fare class of each route
2. 2015-17 MBTA SYSTEMWIDE PASSENGER SURVEY ("demographic survey") = This demographic survey was used to determine the percentage of riders that use monthly passes or pay per ride.
3. Bus Fares = This website lists the full and reduced fares for all types of MBTA buses.

Analysis

First, we needed to determine the fare prices for each route, focusing only on bus routes that fell under the category of Local Bus, Inner Express, Outer Express, and Free. For each route in the routes CSV, we added a column for its full fare and its reduced fare.

After listing the routes that pass through each stop (in the same manner as described previously), we needed to calculate ridership data. Extrapolating data from Fall 2019, we calculated the number of passengers that boarded each route at each stop in one year, assuming that ridership trends throughout the year are uniform.

Now that we had the number of people that board a stop every year, we calculated the average revenue earned by each stop. We initially did a baseline calculation using only the full fare for each type of route multiplied by the number yearly onboardings at that stop as calculated in Step 3. That process is outlined here and the calculation is represented below.

$$revenue_i = \Sigma_{stop\_id=i} average\_ons_i \times fare_i$$

We also decided to add nuance to the calculations by doing a weighted average. We categorized the number of people that board each stop in a given year into monthly pass riders and pay-per-ride passengers for all MBTA bus routes. According to the demographic survey, 70% of riders used monthly passes and 22% paid per ride. The remaining 8% either used 1-day LinkPasses, 7-day LinkPasses, or another method; this population was ignored in our analysis because a large portion of those people were likely to have been tourists using the LinkPasses for one-time visits. For the two main categories, monthly passes and pay-per-ride, they were further broken down into the percentage of people using the different payment methods within each category as described in the demographic survey. After calculating the number of yearly riders belonging to each category, we multiplied the number of people in each category by the fare rate of that group to get the revenue generated from that population at each stop for each route. In this weighted calculation, we assumed that this revenue composition is fixed for each route, that a rider uses a monthly pass twice every weekday, and that a month has 22 weekdays.

Due to the shift in priorities mentioned in the beginning, we no longer used the generated data in our final analysis, and the results are not in the final deliverable repository. However, if you would like to view the code used to generate the calculations, see here. To view the resulting CSVs, see here and here.

Appendix III - Initial Weighted Analysis

For Deliverable 2 and Deliverable 3, we showed results for our 0.5-mile radius calculations. However, the results in those deliverables differ from the final results presented here for two reasons:

1. The proportion used to calculate both population *and* income for a given stop was the proportion of a tract overlapped by the stop's circle divided by the total area of the tract:
$$proportion_{ij} = \frac{overlapping\_area_{ij}}{area_{tract\ i}}$$
As highlighted in the main analysis, this works perfectly for calculating population — if Tract Y has a population of 100 people, and if Stop X's circle covered 50% of that tract, then the population served by Stop X is 50 people. However, this does not work for calculating income — if imaginary Tract Y has a median household income of \$100, and if Stop X's entire circle covered 50% of that tract, then we would mistakenly assign Stop X a median household income of \$50. However, unlike population, income does not scale with the area of land covered. In this case, we should be assigning Stop X a median household income of \$100.

2. We assumed that tract IDs were unique. However, as described in the Challenges, some tract IDs actually repeated for distinctly different sections of land. As a result, we were sometimes pulling population and income data from the wrong tracts if a stop overlapped with these repeats.

Given that this approach was not used to generate the results in this Final Deliverable, the related code was not added to the final deliverable repository. However, it can be found on our Deliverable 3 branch.