Deliverable 3, Transit Equity Team 1

Goal

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. For Deliverable 1, we completed Steps 1 and 2. For Deliverable 2, we have completed Steps 3 and 4 as well and discuss our findings at the end of this document.

However, for the reasons outlined below, in this Deliverable we adopt an alternative approach to the plan described in the original Project Description:

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 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 deliverables that adopt the same fundamental approach: a baseline analysis and a "reach" analysis.

For the baseline analysis, we assigned “population” (derived by summing census data for non-car commuters) and income levels (discrete values from 1 - 4 based on median household income) to stops based on the tract that they are located in. For example, if some Stop X is located in Tract Y with a population of 100 and an income level of 1, then Stop X’s corresponding population and income level would also be 100 and 1, respectively. This would provide us with a sufficient analysis to fulfill the client’s request.

Discussions with Spark! also resulted in a “reach” analysis which takes a weighted average of nearby tracts’ populations and incomes to determine the values assigned to a particular bus stop. 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 they are closest to, regardless of whether they live in a different tract than the one the stop is located in. Additionally, many tracts do not have a corresponding stop; the commuters who would use this stop are not reflected in the previously mentioned approach. Though we reasoned that the results would likely be very similar to the baseline result, this approach provides a more accurate depiction of commuters served by each stop. Both approaches are shown here to depict all the steps taken to achieve the end result.

Data Aggregation

We were still able to use the income level calculations for each stop; the process for these calculations is described in Deliverable 1 and was further completed for Steps 3 and 4. Here is the code and a ReadMe explaining the income level attribute.

To begin the new population calculations, we used the US Census 2018 ACS 5-Year Estimates. Background information can be found here, with the necessary dataset linked here. This dataset has the 2018 ACS 5-Year Estimates filtered on Income (Households, Family, Individuals) and grouped by tract in Massachusetts. The csv file we worked with can be found here.

Data Manipulation

Baseline result

The code for generating tract-level population data can be found here. This tract level data was then used to assign population and income levels to stops, and the code can be found here. This stop data was used to determine which routes would most positively impact low-income riders if made free. The code for this can be found here.

0.5-mile radius result

The code and corresponding explanation can be found under Step 5 here. Now we need to do more on geometric attributes, so the library *GeoPandas* helps a lot. The *GeoDataFrame* is analogous to the DataFrame of library *Pandas*, but with a ‘geometry’ column, which includes GeoSeries objects (Point, Polygon, Multipolygon). Then we can take advantage of its built-in functions to read shapefiles and manipulate geometric data.

To calculate the population impacted by each stop, we drew a circle around each stop using the function *geopandas.buffer()*(its distance unit is meter, so we transferred miles into meters). Then we found overlappings between each circle and tracts using *GeoSeries.overlaps()* and weighed the median income and impacted population by the area proportion intersected. By the way, all data manipulated above were under the *'EPSG:26986' Coordinate Reference System (CRS)*, because geometry data under different coordinates could not work together.

Results

Baseline result

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 1 for more details. For a spreadsheet view of the final stop data, see here, and see here for the final route data.

Figure 2 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. The stop-level data was primarily used for the visualization; further work will need to be done to make route-level data more accurate.

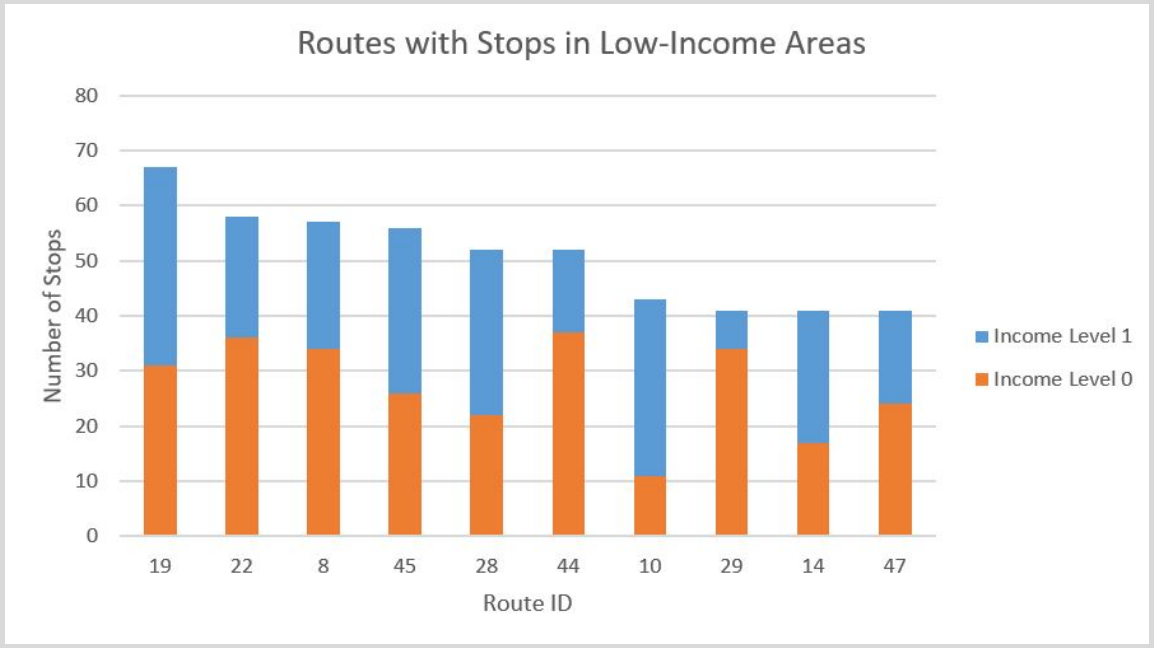

Figure 1. The top 10 routes that pass through the greatest amount of stops in low-income areas using the baseline calculations. Income level 0 represents lowest-income, and income level 1 represents lower-middle income. The income levels are defined according to the Pew Research Center.

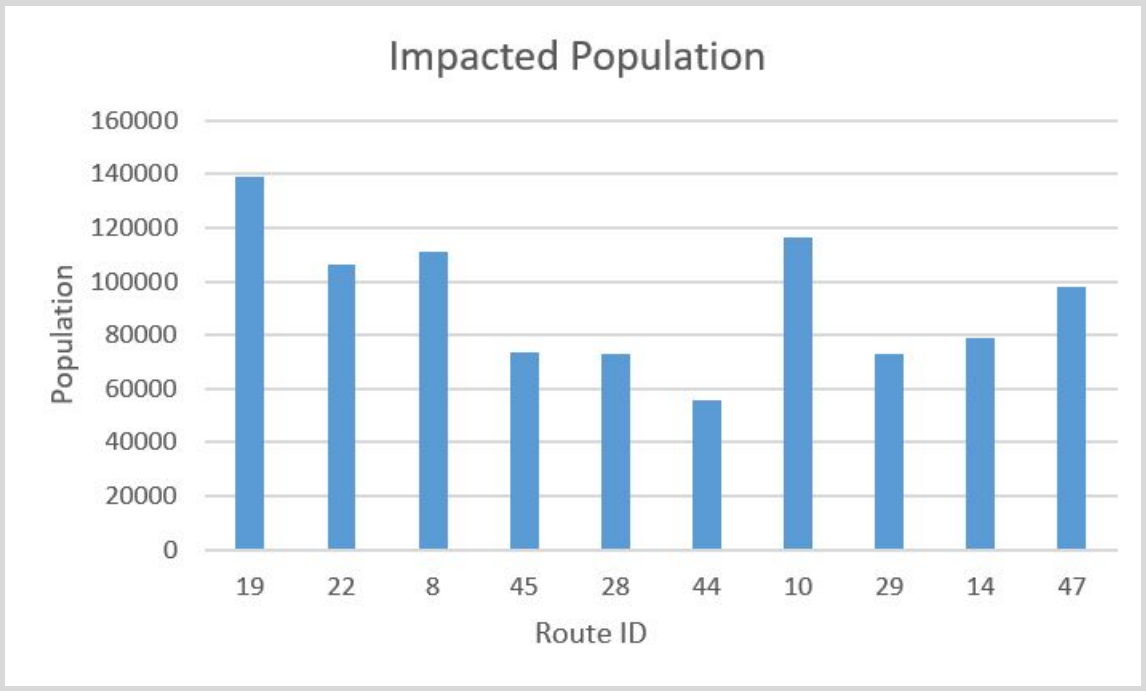

Figure 2. The population that would be impacted if that route were made free, using the baseline calculations. Note that this is calculated using the data for each stop within a route, so this is an estimate.

The best way to view the visualization is to download Tableau Reader and view the Tableau file there (on the GitHub page, click Raw > right-click on the raw data page > Save as). Otherwise, you can view it at this link. It is interactive; you can zoom in/out and click on stops and routes.

0.5-mile radius result

Using the 0.5-mile radius method to calculate income and impacted population for each stop, the top 5 routes that pass through the most stops in low-income areas are Routes 45, 19, 28, 22, 14, and 44 (Routes 24 and 14 tie for #4). See Figure 3 for more details. For a spreadsheet view of the final stop data, see here, and see here for the final route data.

Figure 4 provides a rough estimate of the impacted population broken down by route. However, because radii for stops on the same route may overlap, the population at those intersections have been counted more than once. We estimate that the impact of this is negligible. As with the baseline results, the impacted population here was calculated at the stop-level for visualization purposes; the calculation will need to be further refined for data at the route-level.

Figure 3. The top 10 routes that pass through the greatest amount of stops in low-income areas, using the 0.5-mile radius calculations. Income level 0 represents lowest-income, and income level 1 represents lower-middle income. The income levels are defined according to the Pew Research Center.

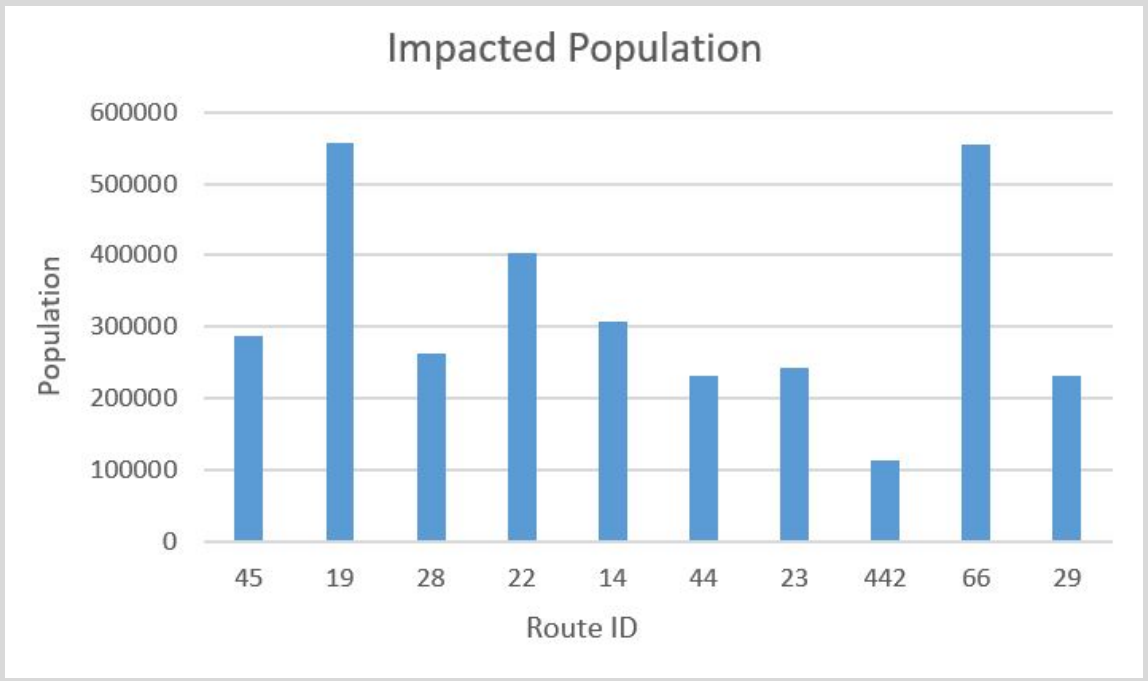

Figure 4. The population that would be impacted if that route were made free, using the 0.5-mile radius calculations. Note that this is calculated using the data for each stop within a route, so this is an estimate.

As with the baseline visualization, the best way to view this second visualization (on the GitHub page, click Raw > right-click on the raw data page > Save as) is through Tableau Reader. Otherwise, you can see the interactive map here.

One of the suggestions to further offer a more decisive impacted population number by stop is to somehow calculate route-wise the total impacted population number by doing counting population minus population overlap. The problem with this suggestion is that given our data for the surrounding polygon and its census population size, when there is an overlap between two intersections of the polygon, the two different polygons are most likely to have different population to area ratio. This is an inconsistency that makes this suggestion infeasible, the problem lies in the simple fact that we don't know the decisive number of the overlapping population. If given two areas with different population counts, let's say a dense population in a small area and a low population count in a big area, if they overlap in a certain area. The dense populated area for that proportion of the overlapping area will argue that area has a lot more population than if the low population density big area will argue for. This is already a hard problem to figure out with just two areas arguing about the true value of the overlapping population but now imagine that with multiple areas overlapping, it is practically impossible to decisively conclude the impacted population if we do this by bus stop population - the ratio of impacted population by overlapping areas.

Next Steps

The results provided here meet the client's requirements as described in the introduction. Final steps include making route-level population data more accurate to avoid double-counting people and tweaking the visualizations per the client's and Spark!'s input.

Previous Goal

Steps 3 and 4 of the Project Description

Previously, Step 3 was to calculate bus ridership for the MBTA (though listed as Step 4, this actually needed to be done before calculating the average revenue of each stop). This was done using data produced for Deliverable 1 and data found here. The number of yearly onboardings for each stop was calculated as described in this Jupyter Notebook.

Step 4 was calculating the average revenue earned by each stop. We initially did a weighted calculation as described in the above Notebook, using the average percentages of fares for all bus routes here and the fares listed on the MBTA website here. However, after discussion with Spark!, we decided to do a baseline calculation as well using only the full fare and yearly onboardings calculated in Step 3. That process is outlined here and calculation is below.

$$revenue_i = \sum_{stop\_id=i} (average\_ons_i) \times fare_i$$

Although we are no longer using the generated data moving forward, it could be useful in the future for further financial analysis.