

The Cooper Union for the Advancement of Science and Art

City Harvest Group 2: Identifying and Predicting Food Need

Nicholas Gao, Ross Kaplan, Anushree Sreedhar, Zhenni Jane Zhu

Data Science for Social Good
Spring 2018
Professor Sam Keene and Will Shapiro

1. Objective

City Harvest is a NYC based organization that seeks to redistribute food over the five boroughs. The organization rescues 150,000 lbs of food each day, with more than 50% being fresh produce, and serves all kinds of facilities including soup kitchens and food pantries. The goal is to assist City Harvest in their mission to redistribute and rescue food in order to feed New York City's hungry. This includes identifying areas of need, highlighting City Harvest's and FeedNYC's successes in targeting areas of need, and predicting the potential of new locations.

2. Background

City Harvest provided data from 2001 until 2017 on how many people were served in a City Harvest location, different violations that the facility was charged with, and the type of facility. After conferring with members from City Harvest themselves, additional NTA (Neighborhood Tabulation Area) information was delivered that focused on food insecurity and meal gap data. This data focuses on food need, or lack of available financial resources for food at the level of the household. Initially, maps were generated using software like Tableau and GIS in order to visually see problem areas, especially in relation to average income data for an area. In addition, homelessness was another factor that was taken into consideration, along with other factors we believed could contribute to trends in the data.

The City Harvest-provided data on each location's monthly food served/provided resources, combined with external data sources such as that from FeedNYC, showed that trends can be observed on certain features. These features can then be fed into a custom made machine learning algorithm, where inputs are taken to deliver an output of potential number of people served in a given area. This model would help City Harvest determine how successful a certain location would be given a set of features trained to the given area, or NTA.

3. Initial Exploration

The first meaningful analysis conducted was combining the “Number Served by Location” data from City Harvest with public MTA subway stop data. Figure 1 shows the service locations being plotted on top of the locations of all the MTA subway stops.

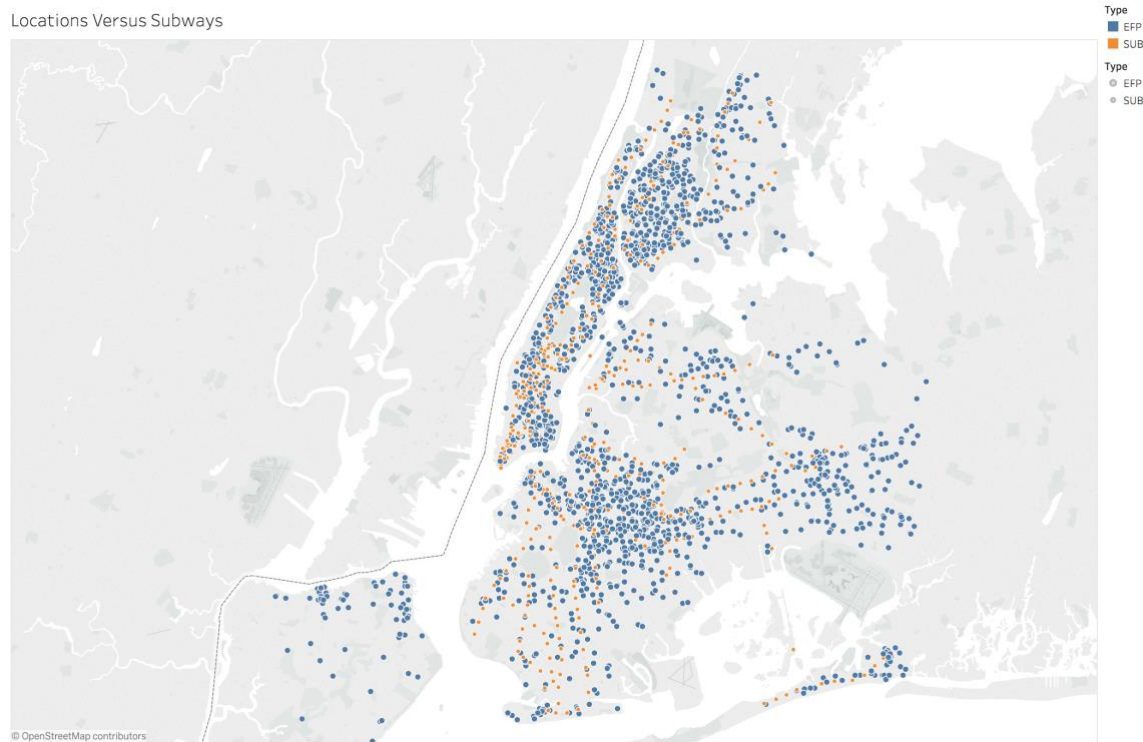


Figure 1: Plotting the subway locations with the locations that City Harvest services [Link 1]

We decided that accessibility may be a factor in determining the amount of people that visit each location. Thus we used the geopy module in Python to calculate the distances from each location to subway stops. We then recorded the distance to the closest subway stop and used this as the metric for accessibility. What we found after plotting number served versus distance to the subway was that when you view the data up close, there is no correlation and the data is scattered (Figure 2). However when you zoom out, the plot begins to show an upper envelope that follows an inverse relationship. This confirmed our intuition that less accessibility, given by a further distance to the a subway stop, leads to lower performance or number of people served. In all, this was the first indication that trends exist within the data. After this point, we sought to identify other potential factors in the number served by each location. After creating the list of factors, we had to think of ways they could be quantified and which public data resources would contain that data. Together the factor data can become feature data. Following a similar intuition as Figure 2, we would try to fit a model or curve to the data. Lastly, since we had multiple sets of factors or features, we needed to use machine learning in order to handle it. This process is detailed below.

Distance vs. People Served: 2013-Present

(Food Pantries ONLY)

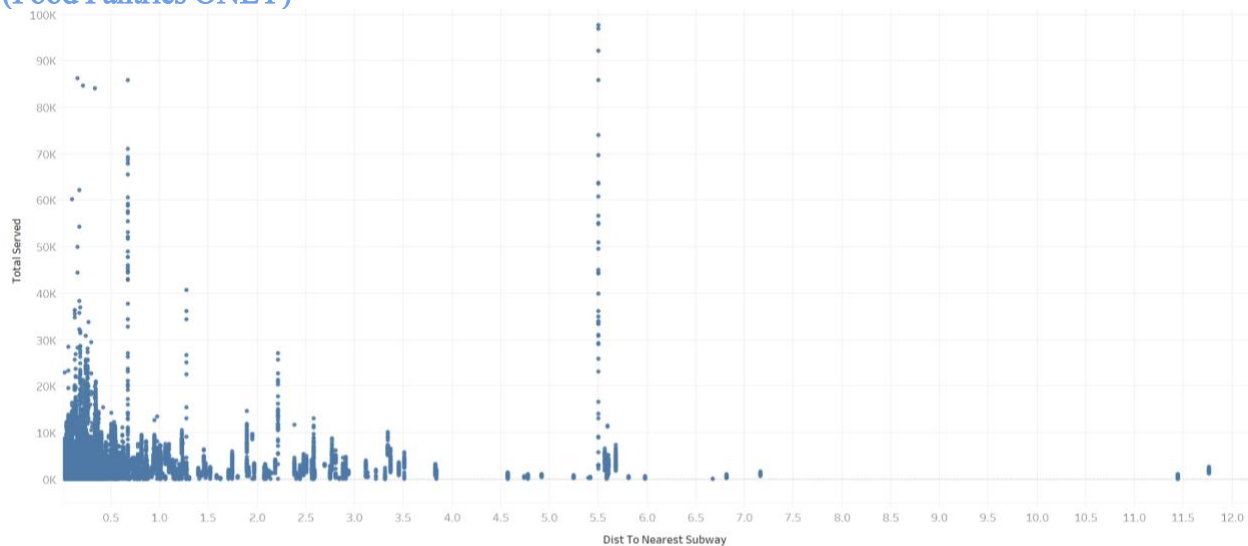


Figure 2: Plot showing the number served by each locations versus the locations distance to the closest subway stop.

4. Results and Discussion: Machine Learning Model

Machine Learning takes in a series of inputs in order to “train” a model on certain data, in order to predict an output. This output isn’t perfect as it only can “learn” how to “respond” to inputs it’s been provided. However, machine learning is a best guess option for prediction. Essentially, machine learning takes in these inputs and “learns” how to turn a particular input and turn it into the output. For the machine learning model used in this map, it is essentially learning how to plot these points and find a line of best fit, as shown in Figure 3. One important point here is that it’s trying to minimize error, so the line is best off if it goes through as many points as possible without too much curving of the line.

One major problem machine learning can have is shown in Figure 4 where you can see several data points are given by the red dot. As you allow the line of best fit to curve more and more, you allow the model to learn incorrectly, as the best line of fit is shown as the black line going through the middle. This error is called overfitting, and it makes machine learning predictions for unknown data (like the map will have to predict new locations) significantly less accurate. To avoid this problem, we implemented something called regularization, which penalizes how curvy the line of best fit gets. That penalty would grow very, very large for the red line shown in Figure 4, and would thus be reduced to something like the line in the middle that is much straighter.

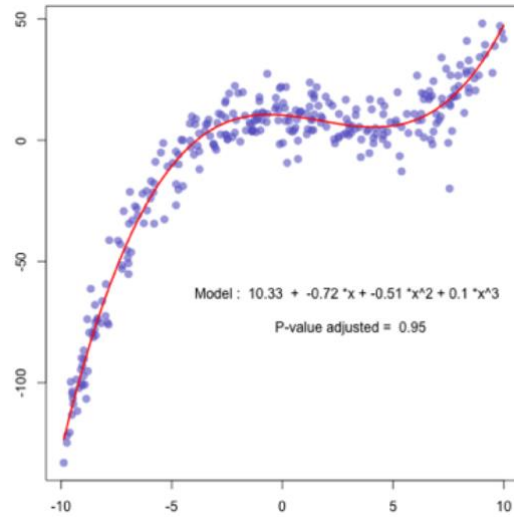


Figure 3: Example of Line of Best Fit

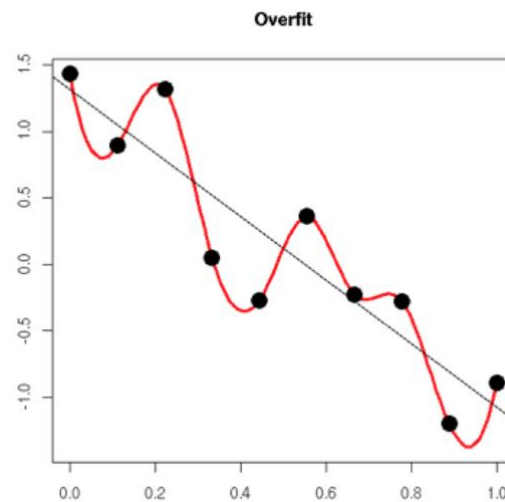


Figure 4: Dangers of Overfitting

The machine learning model created was based on the following features:

- Food Insecure Population
- Number Served in the NTA
- Number of other locations in NTA
- Facility Type
- Distance to Nearest Subway
- Median Household Income in NTA

The machine learning model was trying to predict the following labels:

- Total number served in that location in a year

The goal was to create a model that would use these features to predict how successful, or how many people would be served in a given location. The results are displayed in a user friendly interface on: [Link 2] (Figure 5). Figure 5 shows an example of how the model can predict number of people served.

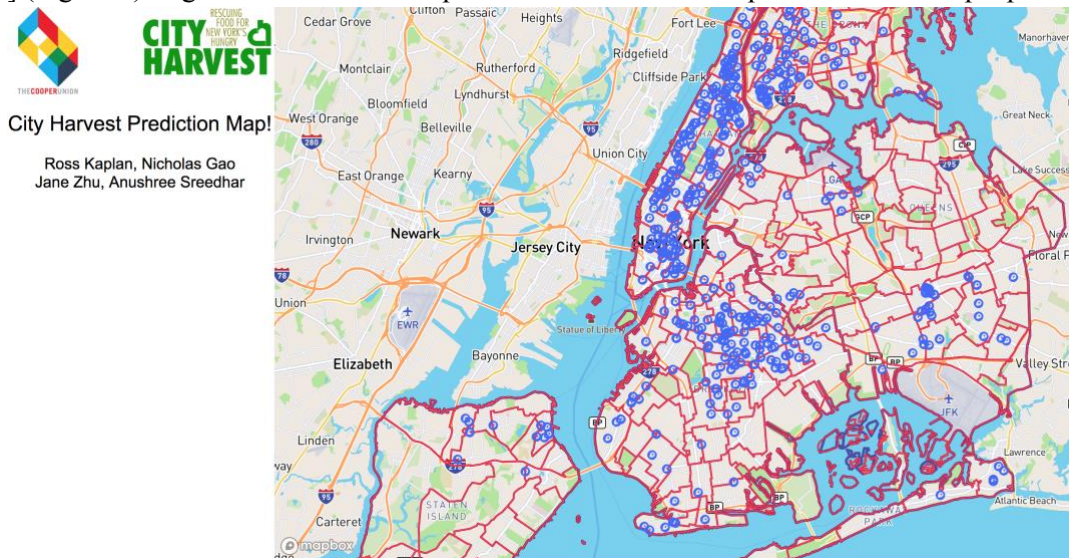


Figure 5: User Interface of Website

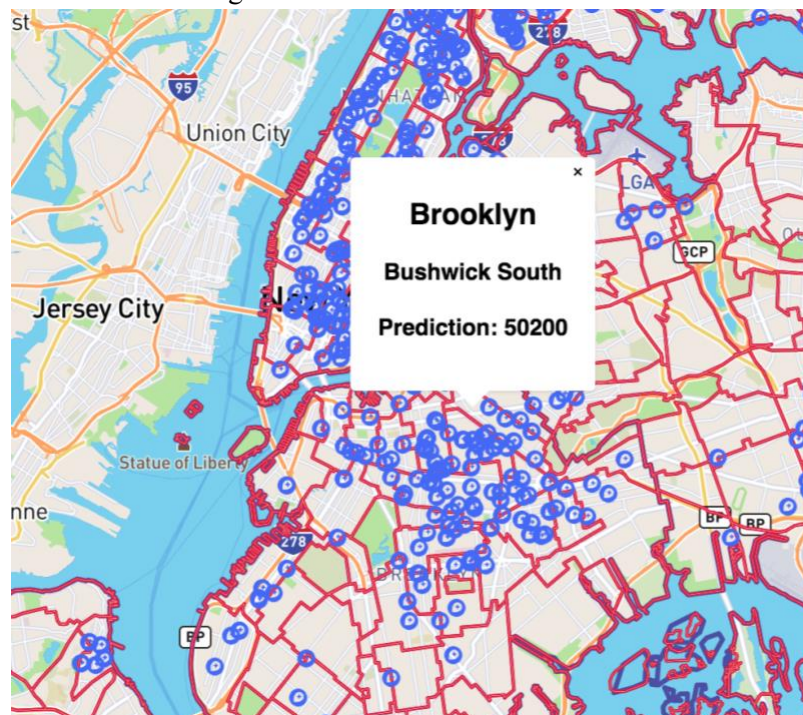


Figure 6: Example Prediction

The mean absolute percentage error is just below 10 percent. It's fairly complicated how this is computed, but the easiest way to understand that is to break it down.

At random, 20% of the data was set aside, and the machine learning model was only allowed to see the remaining 80% of the data. Once the machine learning model "learned" what it had to learn on the 80% of the data, we checked to see how well it performed on the 20% of the data it had never seen

before. We then computed how accurately it was able to predict the number of people served in that 20% to compute the error of the predictions.

If that was confusing, don't worry. Here's the tl;dr. The model is able to compute the predictions fairly accurately near where the pins are. Where there aren't any pins (known values), it's impossible to say how well the model predicts. However, near where there are known values, we can say with confidence that the values are fairly accurate.

One other note is that you'll see that, occasionally, far from some of the pins, there are values that are below 0. This is one of the downsides to regression, in which it can generate a negative value. This should only occur far from the pins, since there's no known data there, but just know this is where most of the error is concentrated.

Some conclusions we found were that the further a certain location was from a subway stop, the fewer the number served. This is in accordance with preliminary analysis done comparing distance from subway to number served. However, if the distance is great from one single subway stop, chances are there is a closer subway stop to that location, and the number increases once again. Another factor that was important when determining the outcome of the model was how close the immediate City Harvest centers were. The more saturated the City Harvest locations were in a given area, the less number was served per area.

5. Results and Discussion: Mapping Features

Meanwhile, over the course of the semester, we also made multiple visualizations of the City Harvest's data, as a standalone and its connection with other datasets. The graphs and machine learning algorithms can often become too abstract to understand, so the visualizations help ground the ideas we are trying to quantify.

One of the most striking visualizations was a dot density map showing the food insecure populations throughout NYC (Figure 7) [Link 3]. It was created from the NTA meal gap data from FeedNYC and using the GIS program arcMap.

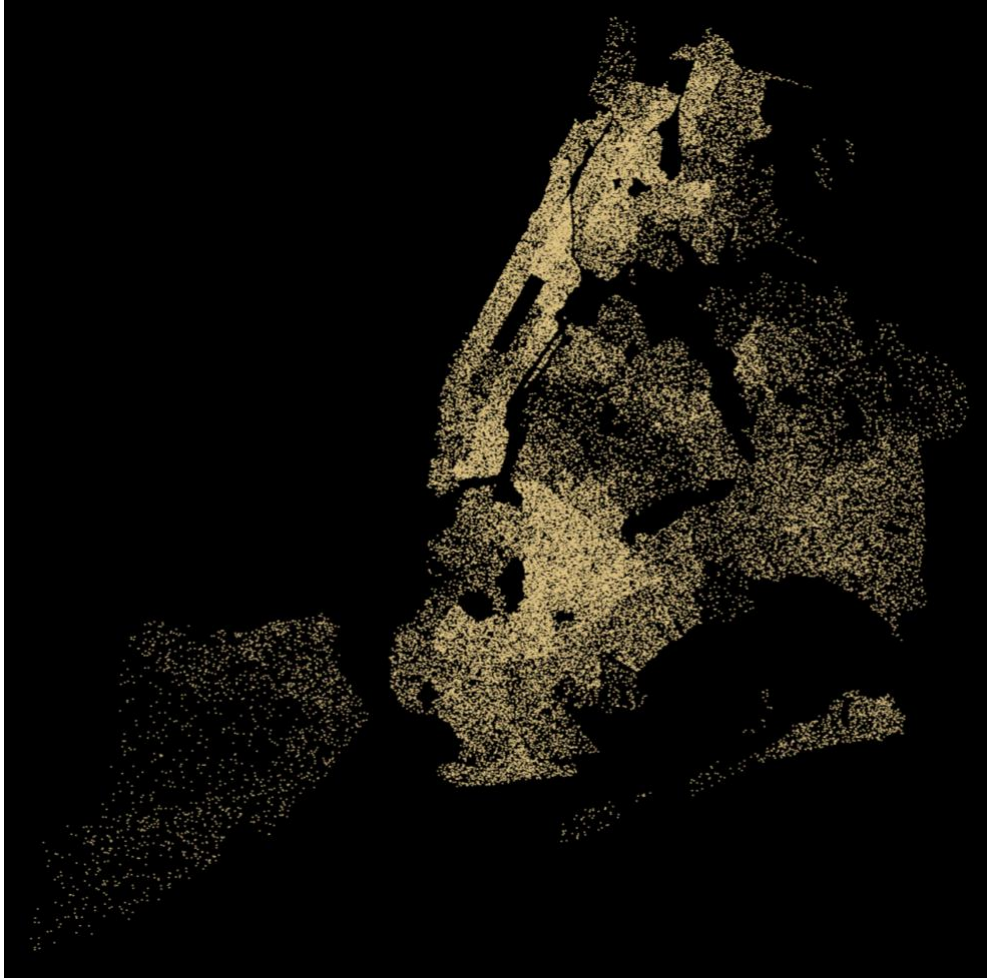


Figure 7: Dot Density Map showing the Food Insecure Population within each NTA. Every dot represents 100 people missing food [Link 3].

We then used this data and combined it with the number served data in order to compare the number of people who need food to the number of people who were served (Figure 8) [Link 4]. This was done on a per-NTA basis. We then plotted the discrepancies, focusing on the areas that were underserved in relation to the other NTAs. We visualized this through downward extrusions that were produced using Rhino on the exported map from arcMap.

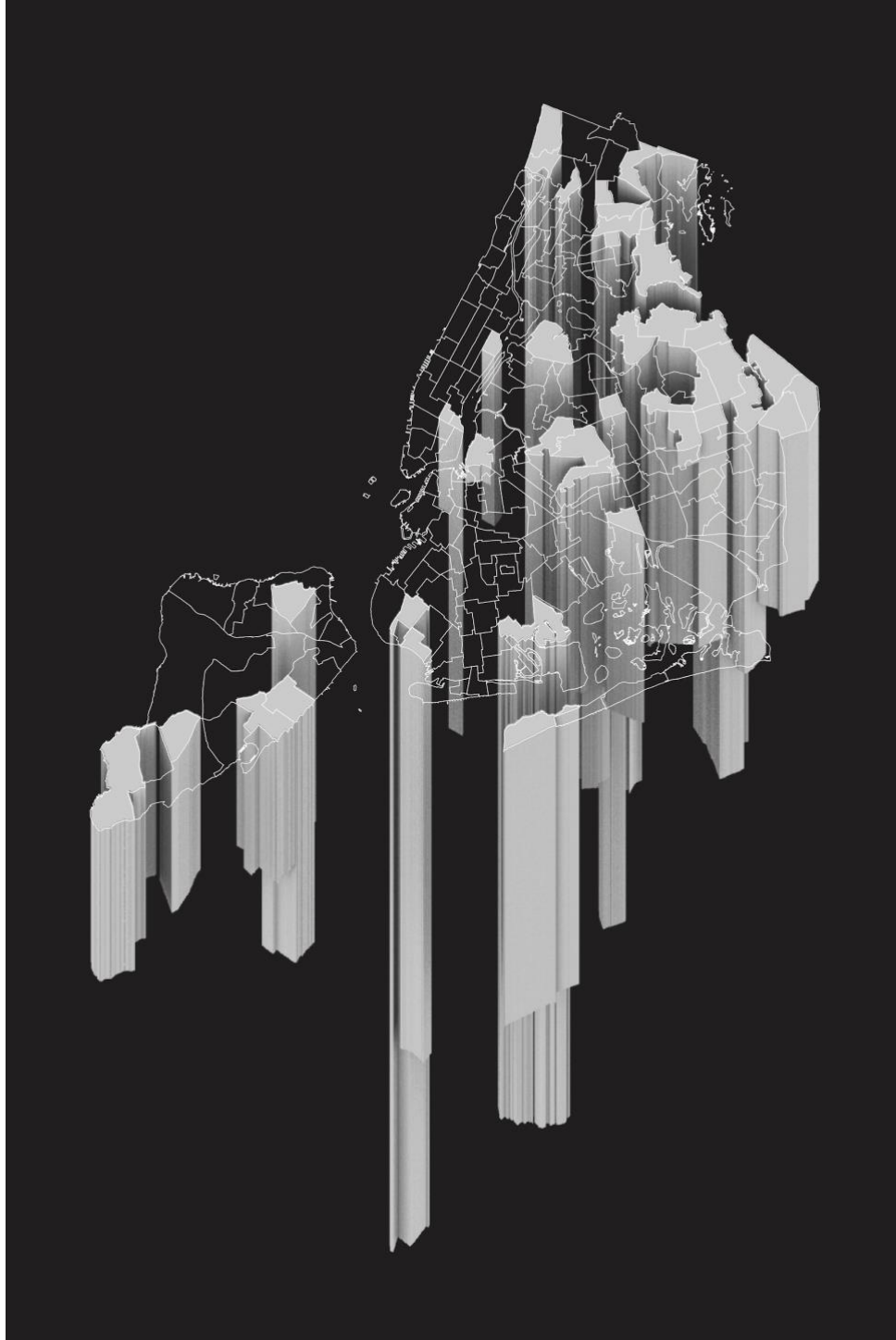


Figure 8: Extrusion map showing the discrepancies between the number of people who need food and the number of people served on a per-NTA basis. Only the underserved areas are shown [Link 4].

Next we looked at median household income. This data came from public census data on NYC (Figure 9) [Link 5]. Again we used arcGIS and viewed not only how median household income (MHI)

data varied across NTAs, but how service locations were located in relation to areas of low income. We also plotted the year-to-year change in MHI from 2013 to 2016, which can be viewed here [Link 7].

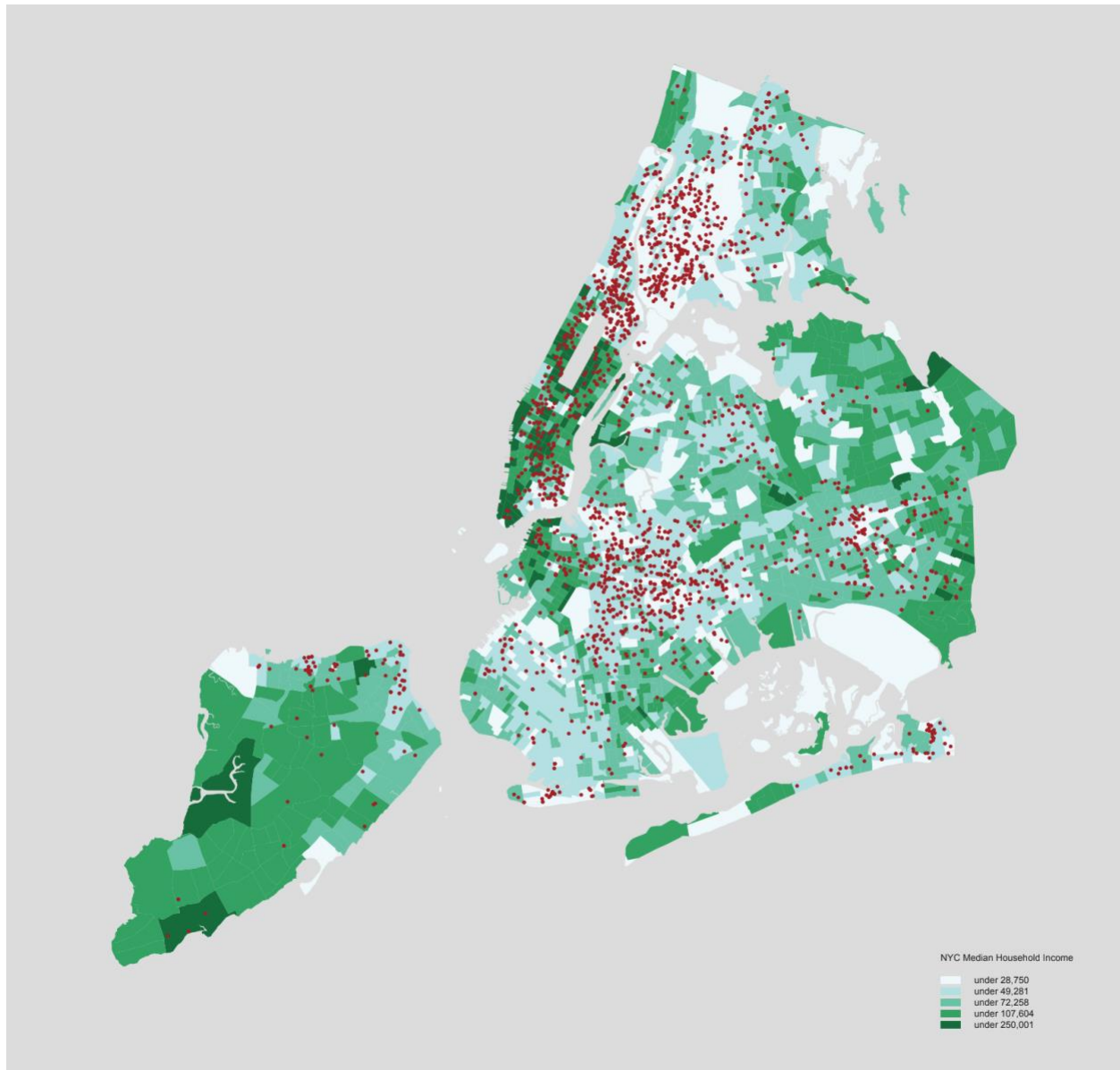


Figure 9: Median household income plotted per NTA coupled with the City Harvest-affiliated locations that provide food aid [Link 5].

Lastly, we used Tableau Public to create an interactive map that shows the number served by each location over time, from 2009 to the most recent data in 2017 (Figure 10)[Link 6]. This helped to measure how the food aid scene changed over time and how some locations grew or shrunk.

Number Served Over Time - January 2009

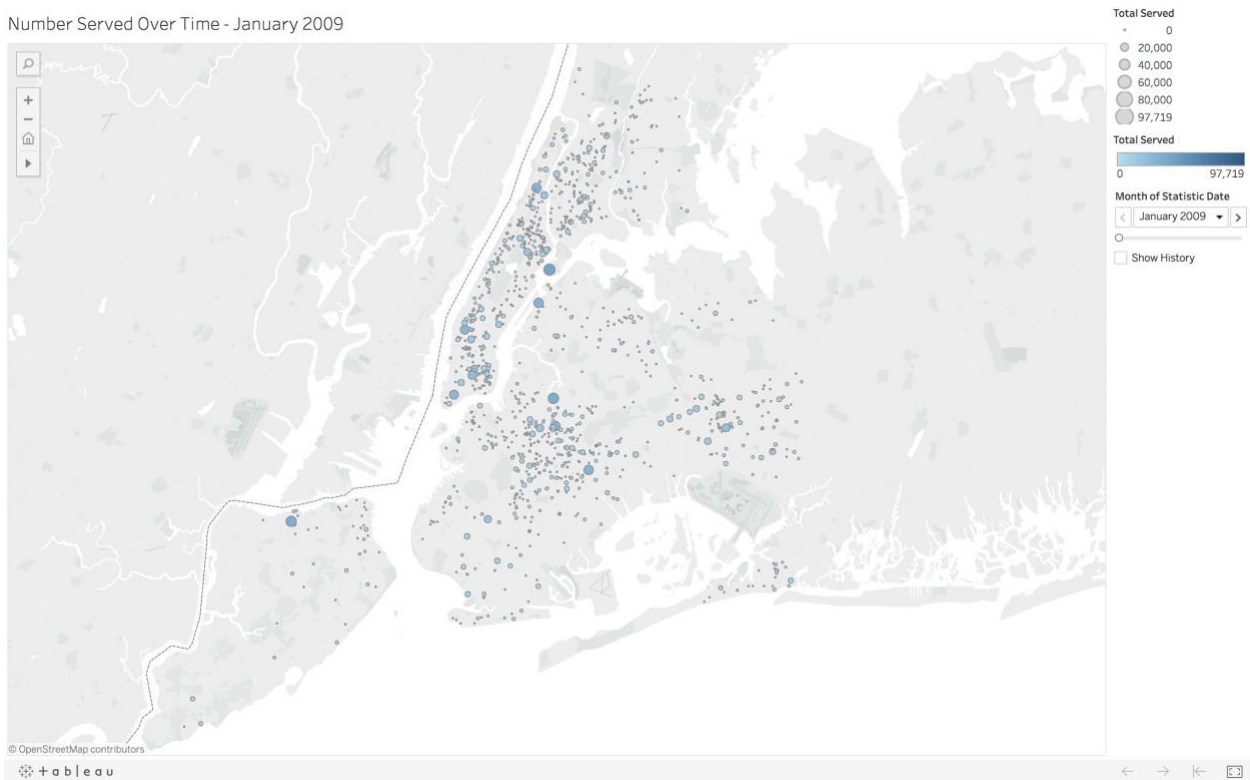


Figure 10: Timelapse Map showing the number served per location per month from 2009 to 2017 [Link 6].

6. Conclusion and Next Steps

Overall, by using City Harvest's past data and coalescing it with other datasets, we are able to save City Harvest time and money. We achieve this by creating datasets and visualizations that highlight their current successes and locate new areas of need. In addition, we developed a machine learning model that predicting the turnout of new potential locations prior to their inception. This helps City Harvest gauge performance before they begin the engagement and become invested in the location. In all, we are helping to aid and expedite City Harvest's current functions in order to save them the time and resources that are vital to any charitable organization.

7. Links

This section provides links to the different visualizations and tools mentioned in this paper that we've created throughout the semester.

- [1] [Locations vs. Subway Map](#)
- [2] [Interactive Prediction Map](#)
- [3] [Food Insecure Population - Dot Density Map](#)
- [4] [Food Need vs. Food Served - Extrusion Map](#)
- [5] [Locations vs. Median Household Income Map](#)
- [6] [Interactive Timelapse of Number Served by Each Location](#)
- [7] [Change in MHI Y2Y 2013-2016](#)