Tartan Data Science Cup Episode 1: Ride to Glory

Maksim Horowitz, Devin Cortese, Steven Silverman

February 21, 2016

Introduction

With funding allocated to open two new Citi Bike Stands in New York City, our group was tasked with finding optimal locations for the stands to increase ridership, especially among female riders. After intitial data exploration and analysis we realized that in order to understand which areas to target, we needed to understand where potential riders are located. To do so we downloaded data from the US Census Bureau containing information at the census tract level on population and demographics in New York City. We merged the census data with the given dataset to improve the accuracy and rigor of our methods resulting in our final suggestions.

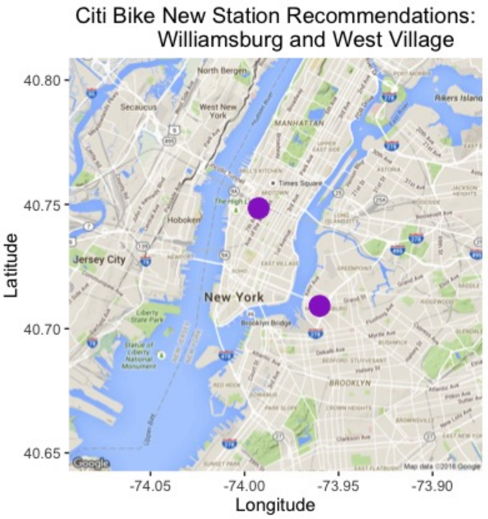
Methods

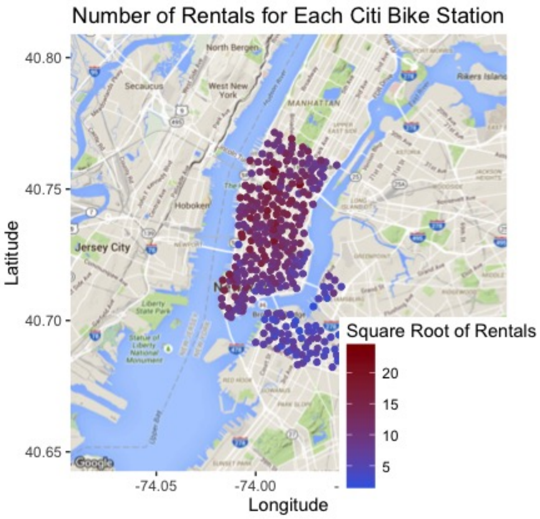
We omitted tract data corresponding to Staten Island and The Bronx due to their relative geographic remotness (none would have affected the analysis anyway). We initially had considered finding the pairwise distances between every station and every census tract to see which stations were surrounded by the most people, but rejected that as being too computationally intensive.

Instead, we approached the problem from two angles: current ridership and current population density. We decided to use lay our own grid of points over Manhattan and Brooklyn (roughly following the actual grid when possible), using major intersections as our breaking points. From there, we found the five closest current bike stations to each point, as well as the five closest census tracts. This is almost a hidden k-nearest neighbors regression, although there's no real model that we're attempting to fit.

We also found the number of riders who used each station (simply by counting), as well as the population density of every census tract (with the `rgeos` and `sp` packages, discussed below). Next, we averaged the five neighbor values for each of these attributes for every point on our overlaid grid, resulting in a ranking in both population density and rider density.

We decided to keep the final location selection simple. Rather than attempting to find a weighting system and determine which of ridership and population density is more important, we just selected the grid point with the highest neighborhood population density and the grid point with the highest neighborhood bike ridership, resulting in the following locations for new bike rental stations:


Citi Bike New Station Recommendations:
Williamsburg and West Village
Latitude
40.80
40.75
40.70
40.65
Longitude
-74.05
-74.00
-73.95
-73.90



Discussion

The most interesting aspect of our analysis by far was integrating the census data. We had two separate data sets: one with coordinates denoting the borders of each tract (as polygons), plus identifying information, and another with population information and identifiers. We used the `sp` package plus some string manipulation to convert the coordinates to a usable object, then took advantage of the `gDistance` and `gArea` functions in the `rgeos` package to calculate distances between our points and the enforced points. We also had to merge the two data sets to get neighborhood population estimates for each grid point.

The choices of the grid points posed another interesting problem. We decided to use points along every avenue (major north-south roads) and at every five cross streets in Manhattan proper, along with hand-selected points in Brooklyn (chosen thanks to one team member's status as a native New Yorker who knows the area quite well).

We are confident in our chosen locations being good decisions. The station in West Village was surrounded by stations that had extremely high ridership numbers, signalling that they are overworked and customers may be attempting to rent bikes but arriving at empty stations. An additional dock will allow for a large increase in ridership. The second point, in the Williamsburg area of Brooklyn, is in an area with very high population density but not many bike stations, meaning an additional one will again bring a large increase in riders. Williamsburg is also a hip, up-an-coming neighborhood which may attract more young (and female) clientele than other areas.