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:

PICTURE GOES HERE