Applied Data Science Project Proposal
Team: Benjamin Miller, Christopher Streich, Richard Vecsler, Geoff Perrin

Question: What are the determinates of Citi Bike commuting?

Abstract: We wish to determine what are the individual factors which make the Citi Bike program a viable option for Manhattan commuters. We specifically would like to identify the primary reasons a person would chose to utilize Citi Bike as a commuter. We will explore and quantify which of the following factors are the greatest determinant of a person's choice; distance to work, income level, work industry, distance from subway stations/bus lines, and age.

Motivation: Bike share is a rapidly growing part of the NYC's transit network. Citi Bike clocked in over 10 million trips in 2015[1] and has already surpassed that mark in 2016. In one early September week, Citi Bike twice broke its record for daily ridership, eclipsing 60,000 trips each day[2]. As Citi Bike expands throughout NYC and more riders depend on it to fit into the greater transit network, it is increasingly important to understand demand for the service. Further, in order for the system to function reliably, Citi Bike must work to ensure that bikes and docks are available where and when they are needed.

Methodology: To start, we intend to assess the impact of commute distance on Citi Bike ridership. To do this, we plan to map block-level Census Longitudinal Employer-Household Dynamics (LEHD) to Citi Bike data. From the LEHD Origin-Destination (OD) data we can find aggregate commute distances for each Citi Bike station which we will use as an independent variable in explaining the number of rides originating at each station. After building this simple model, we plan to see what other factors impact ridership. We hope to look at weather, income, employment, race, age, and whatever else we can find. If possible, we may also attempt to match Citi Bike rides 1:1 to LEHD unique commutes. We plan to limit our analysis to Manhattan where the street grid makes our distance assumption reasonable and where there is a high degree Citi Bike station density.

Constraints: The data within the LEHD utilizes census blocks for location. We will utilize a straight line system for measuring distance from worker's home census block to their work census block. Due to the limitations of utilizing census blocks as opposed to firm address locations, as well as locations of Citi Bike stations being fairly limited outside of Manhattan, we choose to limit our analysis to the island of Manhattan. We do this for two reasons; as an island with a regular street grid system we have no major physical impediments that could significantly skew a straight line distance measurement; Manhattan with a high population density tends to have census blocks which are small in size allowing us a level of granularity.

Future Work: We would like to test our results with other bike share systems.

Datasets: Citibike data, Census LEHD, weather data.

[1] Office of the Mayor, NYC. Madeline Kaye on Dec 21, 2015
[2] **Gothamist** by Jen Chung in News on Sep 9, 2016;