# Using Alteryx to Explore Data: Bixi Residential Score

## Purpose

In this assignment I attempted to create a new scoring system for areas where Bixi is available. The purpose of this score is to quantify the “residential-ness” of a neighborhood/Bixi station. The theory behind this score is the farther users travel Monday-Friday between the hours of 5AM and 9AM the more likely that station exists in a very residential neighborhood (and being primarily used as a tool to get those users to work). Since long trips indicate Bixi being a mode of transport rather than simply a tool for tourists, I figured my logic was sound. The feature in question for this residential score is to find the median distance a bixi travels from a given station between the hours of 5AM and 9AM Monday through Friday for a year’s worth of Bixi data.

## Workflow

To begin my workflow I needed to import all monthly trip datasets into one table. I accomplished this using the Directory tool in combination with the Dynamic Input tool to merge all monthly data. Once this was accomplished I needed to import the single-file containing the station data. With everything imported I could join the data on the start\_station\_code of each ride. With this join in place I now have a table of all rides during the year joined with station information on the starting location. After a column rename I performed another join on the end\_station\_code so that for each trip during the year I had the start and end station information. After (another) column rename I was able to calculate the quantity that comprises the meat of this score: the distance, for each ride, between the start and end stations. To get this value I used the following formula to calculate the straight-line distance between the start and end stations:

Although this distance doesn’t incorporate information such as the exact route taken or the distance travelled via street – I believe it to be a reasonable representation for the distance travelled during the trip. Continuing with the workflow I removed columns that weren’t relevant for the analysis and began to filter the data. Since I am examining rides taken during morning rush hour from Monday-Friday, I needed to remove all trips occurring on Saturday and Sunday, on top of any trips that did not start between 5AM and 9AM. With this information tabulated I was able to perform a groupby function the starting station and return the median straightline distance travelled for each station (this median obviously only including data which passed my filter). I used the median because I didn’t want outlier distance values to affect the result of the score – and the median is less likely to be affected by these cases than the average. Since this groupby removed some information about the station I wanted to keep I performed another join function on the station data by start\_name to get the longitude/latitude information back for each station. With this table I sorted each station by distance travelled in descending order and ouput it to .csv.

## Challenges Faced