Introduction/Business Problem

So far, in my lifetime, I have moved over 20 times for education and for jobs. I now live in the USA after having lived in India and in the Middle East. Within the USA, I have lived in 10 different cities - not just for a casual visit, but actually established residence and lived - for a minimum of 6 months each. When I was young and single the selection of the neighborhood to live in was straightforward.

Now, with a family and school going kids the process is much more elaborate. While proximity to work / school was an important factor, we also relied on inputs from coworkers. In addition, the decision process relied on our family members' subjective opinions based on observations during a Saturday morning apartment hunting trip:

that school building looks new
this strip mall has decent shops and appears to be safe
the rent in this area is within our budget
etc.

Many times, our gut feeling turned out to be OK, but not always.

I would like use Data Science techniques to make the selection process less subjective and more quantitative, by comparing different neighborhoods based on socio-economic factors, crime data and Foursquare venues data.

Data section

I realize that the in-depth analysis, identifying correlations and clusters, drawing conclusions are heavily dependent on the available data. The availability of *relevant* data is of utmost importance to analyzing and solving any problem.

In a real-life scenario businesses will have inhouse historical data. To augment inhouse data business routinely purchase data from external (syndication / industry) sources.

For my Capstone project, I will rely on data available from

- Foursquare location data (as required by the assignment)
- Publicly made available demographic / crime / education data from municipal / state agencies; Examples such as
    - Chicago Crime Data (from a previous Coursera Course)
    - Chicago Public School Data (from a previous Coursera Course)
    - Data from websites such as Wikipedia

Since these data sources are likely to have disparate data, I will apply data preparation, data cleansing, data filtering and data aggregation techniques to make the data useful for my analysis.