2. Data acquisition and cleaning

2.1. Data sources

- Census data for Chicago between 2008-2012, obtained from the city's data portal.
- Crime data for the city between 2008-2012, obtained from Kaggle.
- Chicago public school data for the academic year 2011-2012, obtained from the city's data portal.
- The list of neighbourhoods by community area in Chicago, scraped from Wikipedia.
- The top 100 venues in a 500 m radius of each neighbourhood, obtained using the Foursquare API.

2.2. Data cleaning and feature selection

- The *census dataset* contained almost wholly relevant information. It contains the community area name and number, which are needed for cross-referencing data in the crime and school datasets. It also contains the target variable, the hardship index, which is given for each of Chicago's 77 communities, along with some key indicators, such as the percentage of households below the poverty line, the average per capita income, and the percentage of people living in crowded households. This dataset is *kept intact*.
- The *dataset of crimes* contains a number of fields which are not useful for this analysis: case ID, case number, date of reporting, and FBI code, for example. The *most important fields which are retained* are: the *community area number*, the *latitude and longitude of the location of the crime*, the primary *category of the crime*, and *whether the investigation eventually led to an arrest*. The *dataset was also trimmed* using this last piece of information. Since an arrest is a key indicator of a serious crime taking place, those crimes which were reported, but did not lead to an arrest, were removed from the dataset. This also helped in reducing the size of the dataset, which contained about 1.2 million rows prior to the trimming.
- Similar to the crime dataset, the *public school dataset* contains a number of features which are not connected to this analysis, such as the contact details of the school, school ID, and which network it belongs to. The *important retained fields* are: the *average student attendance*, the *number of students who end up attending college*, the *community area number and name*, and the *school's geolocation*.
- Scraping the *list of Chicago's neighbourhoods* by community name was straightforward. It was *done using the* `requests` *and* `BeautifulSoup` *libraries*. The neighbourhoods were grouped into their communities, since the target variable is given by community, not neighbourhood. The *latitudes and longitudes of the communities* were then *retrieved using the* `geocoder` *library*.

- Likewise, retrieving *venue data* using the Foursquare API was straightforward. For each venue, its *name, latitude, longitude, neighbourhood location, and category were retrieved*. Although venues could have been retrieved for each community, doing so for each neighbourhood within the community increases the number of venues, and hence the confidence in any statistical conclusion that is drawn.