# Data

## Data Sources

There are two files that the project will involve. The first file is a CSV file containing thousands of affordable units that are supported by City of Chicago. The list is updated periodically when construction is completed for new projects or when the compliance period for older projects expire, typically after 30 years. It does not include every City-assisted affordable housing unit that may be available for rent, nor does it include the hundreds of thousands of naturally occurring affordable housing units located throughout Chicago without City subsidies. It is available for download from data.gov. This file is required to identify the locations of the existing affordable housing properties and to uncover the common venues close to them using Foursquare location data.

The second file is a GEOJSON file containing the boundary data of all the community areas (neighbourhoods) in Chicago. This file is available from the Chicago Data Portal, a repository of data from the City of Chicago. This file is required to create a Choropleth map showing the number of rental housing units that are present in each community area.

## Data Cleaning

The first file contains the name of each property, the community area it belongs to, its property type, number of units as well as the coordinates of each affordable housing property, amongst other features. There is a total of 389 affordable housing properties (rows) grouped into 60 community areas and 19 columns, with each column providing information about the affordable property. The data is relatively clean, with all no missing values and all present values in a standardised format. Therefore, no additional data cleaning measure is required.

The second file contains the boundary information of all the community areas in Chicago. The data has not been tested yet, but it is highly likely to be clean given its source, so no additional data cleaning measure is required.

## Feature Selection

For the Choropleth map, the total number of units in each community area is chosen because 1) it can show immediately how popular each community area already is when it comes to housing development 2) the data is already present in the data set. While there are other data that are pertinent to choosing the best location such as crime rates and average household income, they will not be included explicitly in this project. This is because such the number of units in each community area can serve as a suitable proxy for these metrics; past applicants for affordable housing would have preferred safer and richer areas (that are still within their reach) and the city government and housing authorities would presumably have tried to meet those preferences.

Subsequently, for the sake of simplicity, only one community area will be chosen in populating the most common venues around the housing developments of the chosen. Thereafter, the areas most suited for new housing developments will be those that have many housing developments already close by (strong social network) and amenities such as schools and hospitals to support the lives of new residents.