

New York city neighborhoods were chosen as the observation target due to the following reasons:

- The availability of real estate prices. Though very limited.
- The diversity of prices between neighborhoods. For example, a 2-bedrooms condo in Central Park West, Upper West Side can cost \$4.91 million on average; while in In wood, Upper Manhattan, just 30 minutes away, it's only \$498 thousands.
- The availability of geo data which can be used to visualize the dataset onto a map. The type of real estate to be considered is 2-bedroom condo, which is common for most normal nuclear families.

The dataset will be composed from the following two main sources:

- CityRealty which provides the neighborhoods average prices. <https://www.cityrealty.com/nyc/market-insight/features/get-know/average-nyc-condo-prices-neighborhood-june-2018/18804>
- FourSquare API which provides the surrounding venues of a given coordinates. The process of collecting and clean data:
  - Scrap the CityRealty webpage for a list of New York city neighborhoods and their corresponding 2-bedroom condo average price.
  - Find the geographic data of the neighborhoods. Both their center coordinates and their border.
  - For each neighborhood, pass the obtained coordinates to FourSquare API. The “explore” endpoint will return a list of surrounding venues in a pre-defined radius.
  - Count the occurrence of each venue type in a neighborhood. Then apply one hot encoding to turn each venue type into a column with their occurrence as the value.
  - Standardize the average price by removing the mean and scaling to unit variance.

The result dataset is a 2 dimensions data frame.

- Each row represents a neighborhood.
- Each column, except the last one, is the occurrence of a venue type. The last column will be the standardized average price. The dataset has 50 samples and more than 300 features. The number of features may vary for different runs due to FourSquare API may returns different recommended venues at different points in time. The number of features is much bigger than the number of samples. This will cause problem for the analysis process.