Site selection for a new Peruvian restaurant

2. Data acquisition and cleaning

To identify potential areas for a new Peruvian restaurant, this project used tabular and geospatial data from different sources. First, tabular data was downloaded from the new website to explore census data at census tract level, our unit of analysis, within Cook county in the state of Illinois. These tables were: 2018 estimated population, 2018 estimated median income household of the last 12 months, 2018 estimated unemployment rate and 2018 estimated Latino population. These variables would be useful for exploring and analyzing them later. The reason of choice of these variables can be seen in table 1.

Table 1. Tabular data

Data	Description	Source
2018 estimated population	Total population in each census tract. This variable was selected to build other variables such as: density population, Latino population rate and crime rate which will be used for modeling.	https://data.census.gov/cedsci/advanced
2018 estimated median income household of the last 12 months	This variable was selected due to its importance to distinguish areas with high income and low income. Entrepreneurs will always be interested in opening a new business where households have high income	
2018 estimated unemployment rate	Entrepreneurs look for areas with low rates of unemployment. This is an indicator of stability and a high likelihood that employed people might visit its restaurant	
2018 estimated Latino population	Although it is not a key variable as the 3 above, a restaurant located in a Latino area could accept the Peruvian cuisine faster in comparison to other customs from Europe, Asian, and so on.	

On the other hand, geospatial data as census tracts boundaries were downloaded from The United States Census Bureau. Centers of population by census tract were identified from the same website to extract its coordinates. The advantage of using this kind of centers is that population live in those centers and they are not the common geographical centroid of census tracts where might not have people living. Another spatial data was CBD and obtained from Chicago Data Portal in order to select census tracts as our study area. Finally, crime locations represented as point data were extracted from the same website. More details about the use of each data in table 2.

Table 2. Geospatial data

Data	Description	Source
Census tracts	Census tracts boundaries in the state of Illinois. This data was downloaded as shapefile and then exported as 'json' file for mapping purposes	https://www.census.gov/geographies/mapping-files/time-series/geo/carto-boundary-file.html
Centers of Population by Census Tract	Real centroids with geographical coordinates where people live in each census tract in the state of Illinois. It provides a link to be read in a Python environment. The coordinates were used to identify food venues around each coordinate using Foursquare API	https://www.census.gov/geographies/reference-files/time-series/geo/centers-population.html
Central Business District	Central Business District boundary was helpful to identify census tracts within it and those census tracts located to 5 km. from the CBD. All these census tracts would form our study area	https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Central-Business-District/tksj-nvsw
Crime	Crime data that contains the following categories: assault, battery, burglary, robbery and theft which took place within and outside of restaurants. This data was processed to build a crime rate by census tract. The higher the rate, the lower the chance of opening a restaurant	https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present-Map/c4ep-ee5m