Description of the data

Data sources

To start the analysis it has been necessary to collect the information above:

- Madrid neighborhoods
    - Name and latitude and longitude coordinates were scraped from wikipedia.
    - Population of every neighborhood was found on the web of Madrid City Hall.
- Venues of every neighborhood
    - Foursquare API was used.

Links list:

➔ https://es.wikipedia.org/wiki/Anexo:Barrios_administrativos_de_Madrid
➔ http://www-2.munimadrid.es/TSE6/control/seleccionDatosBarrio

Data cleaning and wrangling

The **first step** was to scrape Wikipedia to extract the name of every neighborhood. I got a dataframe and I deleted some not important columns and translated the name of headers from spanish to english.

Then, it was necessary to collect all the links from the same web of every neighborhood in order to obtain the coordinates (latitude and longitude). After a small manual manipulation of the list with these links, I collected the coordinates and did a new dataframe. Then, I merged previous dataframes.

Finally, I created a dataset (csv file) with the population of every neighborhood from a website and then I imported this set and performed a new dataframe. The final issue was to merge this dataframe with the previous one (main dataframe).

The **second step** was to use Foursquare API to obtain the venues. Previously, I filtered the main dataframe and I selected only the neighborhoods in the north of Madrid. These neighborhoods belong to the following boroughs: “Fuencarral - El Pardo”, “Hortaleza” and “Barajas”. There are 19 neighborhoods.

According to the way that I worked in Labs (previous weeks), I used Foursquare to get the venues, check which neighborhood has a bakery and which does not. I also performed a map using the “Folium” library. This map has marks with the selected neighborhoods and a prominent mark in the neighborhoods with a bakery.