Capstone Project 1: Data Wrangling

- After importing both data frames and looking of their structures, I noticed in particular one column of each data frame ('healthcares' and 'municipalities') that had a lot of null data (>100,000 and > 35,000 respectively). I decided to remove it (the values were not necessary for the project).

- The 'zip code' columns, common in both data frames are very important because they could be useful to merge them later. Then, I decided to remove all the null values in 'zip code' columns in the two data frames.

- The France is composed of a continental territory and five overseas departments. The data frames include data for all these territories. In this study, we will focus only on metropolitan France. Thus, I deleted the data concerning the five overseas departments in both data frames.

- I noticed that the type of 'zip code' column of 'municipalities' dataset is still object despite the elimination of the null cells. To understand why, I sorted the data frame by the 'zip code' column in descending order. And I found ten rows that not contain a zip code but a string value. I deleted these ten rows.

- I noticed that some name of region in 'municipalities' dataset are false so I corrected them.

- I created 4 new columns that could be useful later in 'municipalities' dataset.
  - The first one is composed of the object values: 'north' and 'south' according to the location of the regions. There are 5 regions in the south and 8 regions in the north of the France.
  - Sometime, one zip code can be the same for several municipalities. In prediction to the merge of the two datasets, the population numbers were grouped by each zip code and then this new dataset was merged with the Municipalities dataset. This second new column is the population by zip code.
  - The third and fourth column are the latitude and longitude for each municipality present in the same column in the dataset.

- Then, the important step here was to merge both data frames. We have the 'zip code' column common in each dataset. The problem with the 'zip code' column is that each zip code could match with several different municipality and after the merge with a left join, we obtain a repetition of several rows for this type of zip code. The population was calculated by zip code, so the repeated rows were deleted (the first one was kept) using the column with the name of each doctors.