Data section Covid 19 phase two: Italy case

This is the data section of the coursera capstone project.

There are three data points:

- province population density. The reason behind population density is that the higher it is, the more likely people will get in contact with each other. Of course, this is an oversimplification of the problem, but the general idea stands true and it is highly applicable to the user case. The data about the density is freely available on wikipedia;
- province infection rate: we will get the numbers of infected people from the health ministry website (check this for reference), and then we will need to get the percentages of infected out of the total population, to get an estimation of how widely the virus is spread. Some data wrangling will be needed to get the names of the provinces match, between the two data sets, as there may be some slight variations in name. We will solve the problem in our project;
- the third data point will come from Foursquare, as it is required by the exercise. We will check the number of “aggregation activities/venues” available in the province, to understand how many risk locations are out there. It is a little bit too far stretched, since many of these activities/venues are currently closed, but you get the point.

Once the data is collected, we will proceed to create some clusters (between 4 and 10, depending on the variety of the data points) to decide where it is safer to open more activity and loosen the restrictions a bit, and where it is absolutely not safe to lift any restrictions, due to high infection rate.