Covid-19 phase two: Italy case

Introduction

In this project, we're going to understand how we can use a data driven approach to the beginning of phase two.
Actually, Italy has been hit hard by covid-19, in such a way that for eight consecutive weeks the country has been put in a strict lockdown, where people were forbidden to get outside, unless it is for strict necessity or for emergencies.
Right now, Italy is beginning its phase two, meaning that some restrictions have been lifted and some categories can get back to work, people can go visit relatives, and so on.
Actually, the restrictions are the same for all the countries, even though some regions have decided to take a more restrict or more open approach to it.

Business Case

The question is: can we leverage the data we currently possess to understand where restrictions should be applied more firmly and where we can lose them up a little bit more? in other words, is it possible to tell the difference between a province where the risk of infection is low, compared to those where the risk is much higher?

Target Audience

The category that would be interested in such a study would be the political one, since politicians are the ones required to take decisions, balancing the need to reopen companies and other stuff and the need to preserve people's lives.

Considering this is just an exercise, we will not take into account all possible factors, but we will set the basis for further investigations which will include other factors.

Data and Sources

The data we will be using is:

- provinces' population density. The higher the density, the riskier to reopen activities, since it will help the virus spread faster. Data about provinces' density can be freely found on wikipedia (https://it.wikipedia.org/wiki/Province_d%27Italia);
- provinces number of infected people. The more people in a province are infected, the more the virus in that province is spread. Information about infected per province is provided by the health ministrywebsite, which releases daily reports with updated numbers, such as the one on this page (http://www.salute.gov.it/imgs/C_17_notizie_4702_1_file.pdf). The language is italian, but the reports are basically just returning numbers, so it is fairly understandable;

- **only to include some foursquare data**, as requested by the exercise, but actually making it just more complex than it is supposed to be, we will find on foursquare data about aggregation venues, such as gardens, beaches and parks. The higher the density, the riskier it is to reopen, since it will incentivize people to aggregate in such places. This is of course an approximation, but it is needed to comply with the exercise.

Methodology

Methodology involves a lot of data wrangling, since the data presented in wikipedia is quite messy: supposedly numeric fields are object fields, numbers have spaces in between them, provinces' names don't match between different sources, etc…
So, the first part of the task is to wrangle the data and make it match between different sources.
Another insurmountable limitation is the limited foursquare api calls: since we're examining all Italy, we cannot expect to retrieve *all* information about *all* relevant venues in Italy. Therefore, the exercise will remain theoretical, meaning that it will not display reliable results, **but** the methodology would be effective if such limitations didn't exist.
Gathering all the relevant data includes:

- density
- total population
- total infected (updated to 07/05/2020)
- total risky venues

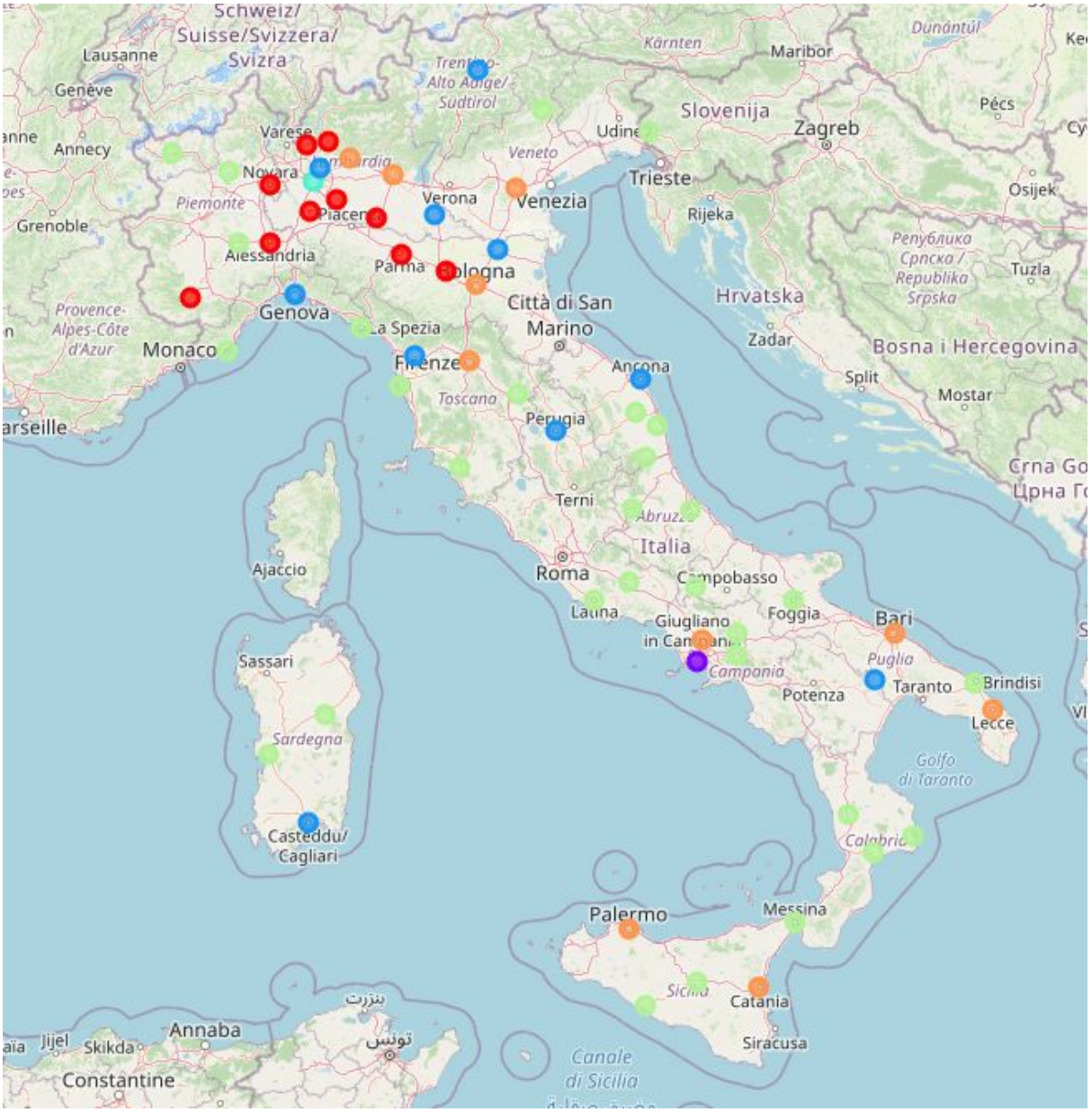
Once we collect all of the above, we proceed to cluster the provinces based on those criteria, using the k-means clustering method, in an effort to find the safest cluster possible to lift some movement restrictions. We will create 6 clusters.
Here is the link to the github notebook:
https://github.com/EmanueleLanzani/Coursera_Capstone/blob/master/Covid%20-%2019%20Phase%20Two.ipynb

Results

In our analysis, we finally find out the following:

- cluster 0: the infected are in the thousands, while venues, density and population are average
- cluster 1: infected in the thousands, but outlier density
- cluster 2: high number of venues, compared to other clusters
- cluster 3: high density, high population, high venues and extremely high infected
- **cluster 4: mid low density, mid low infected, mid low venues**
- cluster 5: high population, average density, mid high venues

By looking at this, we notice cluster 4 is the safest to undertake some risks with, since the chances for people aggregation seem to be the lowest.
Therefore, we suggest our stakeholders to undertake some risks with cluster 4 and to possibly increase restrictions in cluster 3.



Cluster 4 is the light green colored, cluster 3 is just Milan, the turquoise circle in the north-west.