1. Introduction

1.1 Problem

We are currently experiencing a global pandemic that truly influences our lives every day. It has been a year since the spread of the infection in my home country (Italy), but the situation does not seem to have changed so much. We still have lots of SARS-CoV-2 infected people and hospitals are in increasing difficulty to provide the optimum care to everyone.
No one can solve the problems we are experiencing in an easy way, but we can use the data we have to try and understand the reasons of infection and its geographical correlation. In this project, I will tempt to cluster my region's districts on the basis of the COVID-19 infection incidence. After that, I will try to identify the environmental reasons that bring to a higher frequency of contagion in an area more than another. This last issue will be addressed using Foursquare API.

1.2 Interest

Tuscany - my region - has a growing number of infected people everyday and my home city - Siena - is being under pressure in hospitals.
This report and data analysis would prove useful to Tuscany inhabitants and anyone interested on understanding the pandemic phenomenon better.

2. Data acquisition and cleaning

2.1 Data sources

I acquired several publicly available datasets for the purpose, published and frequently updated by Tuscany government at http://dati.toscana.it.

- http://dati.toscana.it/dataset/843000c5-8d28-4426-bed3-54703399be06/resource/c472c0cb-4105-43b9-ae38-d66b88dc0107/download/covidars.csv is a dataset daily updated that includes information about lethality, number of infected people, city district etc.
- Popolazione_indicatori_2019.xls contains old age index by municipality updated to 31.12.2019. The old age index was calculated as: (over-65 inhabitants/0-14 inhabitants)*100.
- An additional dataset was used to obtain data about the number of male inhabitants by municipality and by district. The sex is regarded as an interesting feature because it seems that male people are more prone to be infected with the virus.
- Finally, at https://it.wikipedia.org/wiki/Comuni_della_Toscana I got access to extension of the area and number of people for each municipality, thus I could calculate population density by municipality and by district.

2.2 Datasets feature choice

Age and Sex have been already identified as features strictly linked to likelihood of infection. So, I used multiple datasets to access all information that could be useful for further investigation.
Population Density Since geographical coordinates are not very precise through *geolocator* tool, I have retrieved them from a website reporting latitudes and longitudes for every Italian town and city (https://www.dossier.net/utilities/coordinate-geografiche/).

I will use Foursquare API to select the total number of cafés and schools for every district and use them as additional features since they are the most relevant venues for people's social lives and meetings.

2.3 Datasets cleaning

Now, I would single out the features essential for the analysis and I would merge all useful data into a single dataframe, so that working with data is easier.

I decided to retain the following as features for each Tuscany district: old age index, population density, male fraction of population, fraction of infected people.
Population density was calculated dividing the Population by the Area extension.
The male and infected people fractions were calculated dividing their total value by the total population of the referring district or municipality.
Other pre-existing features were excluded because were unrelated to the topic or redundant.
In fact, the purpose of this project is clustering Tuscany districts on the basis of risk, but we have only 10 cities, so selecting too many features would be a mistake for the analysis.