Tuscany COVID-19 Analysis

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 (Tavole_maschi_per_età_e_classi_età_31_12_2019.xls) 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 territories and number of people for each municipality, thus I could calculate population density by municipality and by Province.

2.2 Datasets feature choice

Age and sex have been already identified as features strictly linked to likelihood of infection. In addition, I selected population density and male percentage of population to assess the risk, which is described by lethality and total infected people variables.
So, I used multiple datasets to access all information that could be useful for further investigation.

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

2.3 Datasets cleaning

At this point, I have singled out the features essential for the analysis and I have merged all useful information into a single DataFrame, so that working with data was easier.
I have calculated the following features from the available ones:
- Population density (number of people by square kilometer) was obtained dividing the population by the area extension.
- The male and infected people's percentages were obtained dividing their total value by the total population of the referring district or municipality and multiplied by 100.

Other datasets preexisting features were excluded because were unrelated to the topic or redundant.
In fact, the purpose of this project is clustering Tuscany Provinces on the basis of risk, but we have only 10 cities, so selecting too many features would lead to erroneous conclusions.

Province	Lethality	Old Age Index	Population Density	Male Fraction	Positives Fraction
Arezzo	3.917334	211.187	107.144248	48.0814	4.673285
Firenze	7.176198	205.593	288.072756	47.6895	4.700139
Grosseto	3.490904	251.105	50.736146	47.5342	2.745171
Livorno	6.981786	233.98	281.740384	47.2586	3.910776
Lucca	6.343814	218.922	225.385316	47.7157	4.635390
Massa-Carrara	9.808577	248.68	173.588452	47.0306	5.355890
Pisa	5.726846	192.581	173.235272	48.7468	5.321832
Pistoia	7.018945	205.564	304.646505	48.4506	5.761835
Prato	5.356232	161.049	701.010295	49.2808	5.936720
Siena	4.419087	213.61	70.628832	47.3745	3.542182

Table 1. Demographic features cleaned DataFrame, ready for the application of the first clustering approach.

3. Data exploratory analysis

3.1 Trend of COVID-19 infection

First of all, I have decided to focus my attention on the analysis of the trend of COVID-19 infection over time. For this purpose, I have selected three cities: Siena, because of simple personal interest; Prato, because it is where the percentage of infected people compared to the total population has been the highest in Tuscany; Massa-Carrara Province, where the lethality index is the most relevant among the other cities.

I have generated a graph to visualize this trend, where the total number of infected people over time is shown. To understand even better the current and past situation, I have calculated the everyday new cases (positives to the infection) and visualized the outcome in another bar plot. Finally, I have analyzed the lethality index fluctuations over time.

3.1.1 Siena

The steepest portions of the first bar plot (the one that shows the total number of infected people over time) are located at the beginning of November 2020 and during the last fifteen days of February 2021. In fact, those two periods corresponded to the governor decision of adopting local lockdown regimens for at least two weeks.
The second bar graph, representing the everyday new positive cases, shows peaks around the first week of November 2020 (where the value reaches its maximum of almost 160 new infected people), then in the first week of January 2021. From the last days of February 2021 until the first two weeks of March there was another rise in the number of new infected people, which remained constant. This number has been descending since then.
The last graph, showing the lethality index trend, indicates that it was alarming (between a value of 6 and 8) only during the summer, probably because who got infected in the summer was more fragile and got much worse consequences than an average healthy adult getting a flu. In fact, the amount of infected people in June, July and August 2020 was very low - always below 20 cases per day. In the end, there has been a dramatic reduction in the lethality index since the beginning of November 2020, maybe due to the restricting measures that were adopted to protect the older and defied.

3.1.2 Massa-Carrara

Massa-Carrara's whole Province showed an increasing trend of the amount of total infected people between the end of October 2020 and the beginning of November 2020. After that, the increase was slower and progressive.
The maximum number of daily new positive cases was reached around the first half of November 2020, with a value of over 350. Recently, this number has been pretty stable, and varying between 100 and 50 cases per day.
Moreover, Massa-Carrara's lethality index reached its highest values - about 16 - during the Summer, resembling Siena’s trend, so that we can infer the same conclusions. Since the first half of November 2020, lethality has remained constant, around a value of 4-5. As stated before, the only recent events that took place have been the lockdown measures and later re-openings. Anyway, to justify the latter, Tuscany has already begun its vaccination campaign since the latest re-openings.

3.1.3 Prato

The first bar graph of Prato, that shows the total infected people with COVID-19, is almost identical to the corresponding graph for Massa-Carrara.
On the contrary, the bar plot investigating the trend of the infection with a focus on the daily new positives indicates a longer period, from the end of October 2020 to the end of November 2020, when the number of new infected people was large (even over 250, but generally around 200). Recently, this amount has been rising steeply and has been reaching again the value of 200. This trend is different from Siena and Massa-Carrara, where the number of new cases is currently almost constant or slowly descending.
The lethality index graph is overall similar to Siena and Massa-Carrara, with its maximum values varying from 8 to 10, that were reached last Summer. After the restricting measures, the lethality consistently decreased to flatten around a value of 2.

3.1.4 Latest developments

I have analyzed the latest trends for Siena, Massa-Carrara and Prato Provinces to get a better understanding of the data for the three areas, because, although the graphs are generally similar, the amounts can be very different (see Figures 1-2).

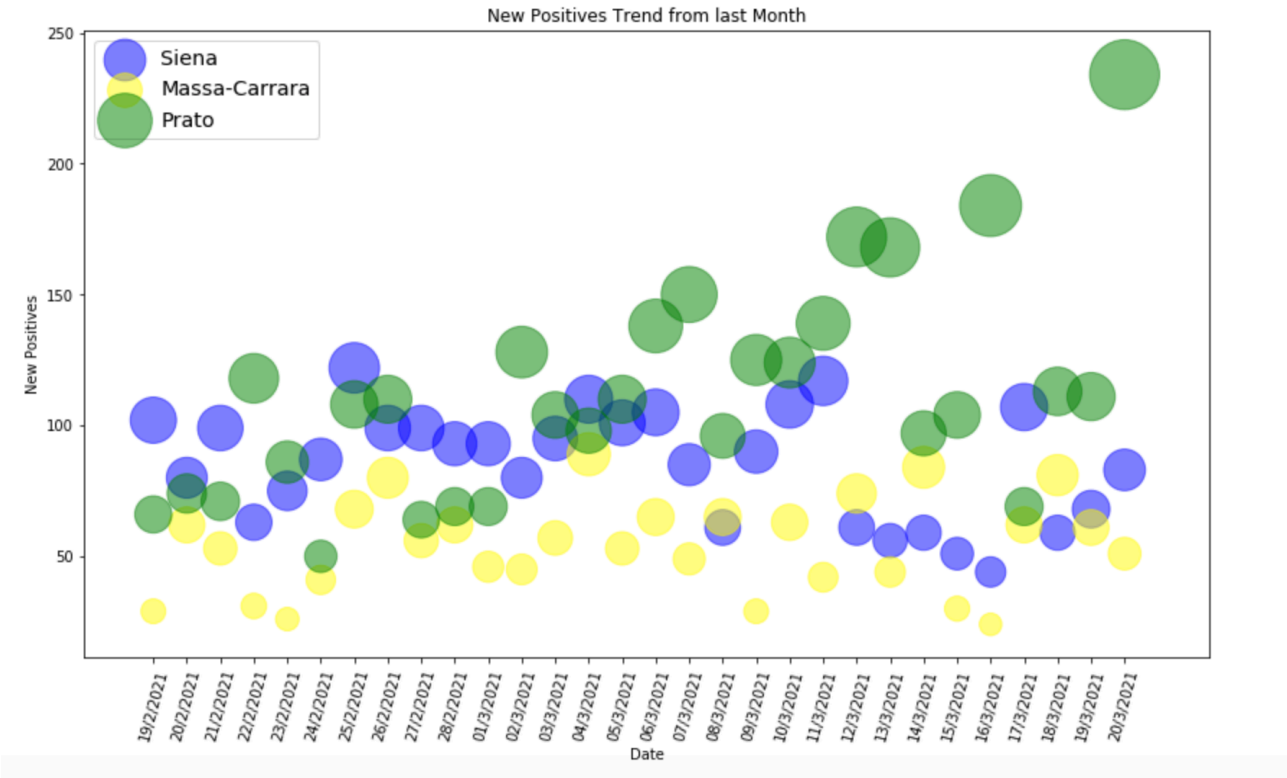

Figure 1. Bubble Plot to show the overall recent trend of the infection. As we can see, Prato's situation is worsening while it's quite stable for Massa-Carrara and Siena's Provinces.

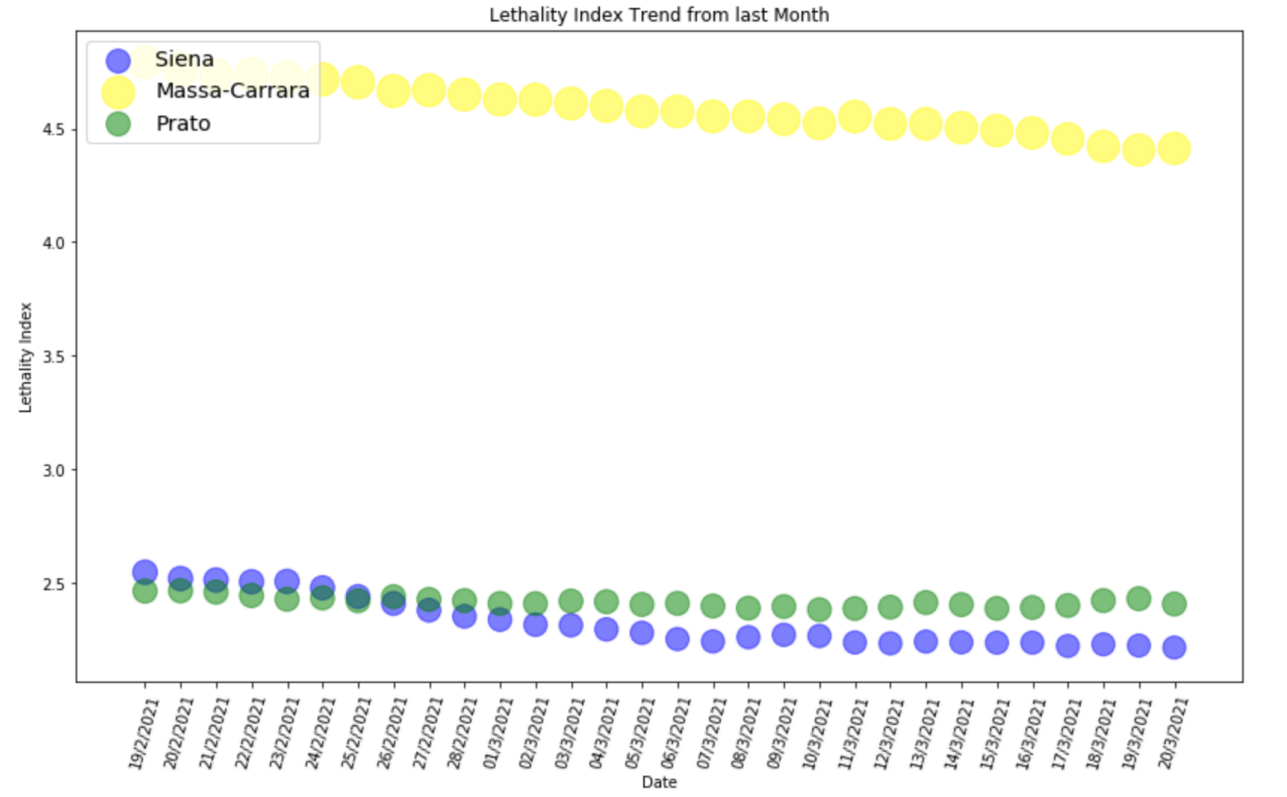

Figure 2. Even if Massa-Carrara's amount of new infected people was the lowest from the previous graph, it still presents the highest lethality index. Instead, Prato has a low index, similar to Siena, even though there has been a recent rising trend in the new cases. This difference could be associated with a diversity of population, such as one younger for Prato and one older for Massa-Carrara. This could mean that, in general, older or defied people with SARS-CoV-2 virus are more prone to get to the worse consequences.

3.2 Features Correlation with COVID-19

The features that I have selected for the analysis of the problem and model building are correlated with the positives' percentage of the population:

- Old age index has a correlation coefficient of -0.636132. This means that as the percentage of infected people rises, its value gets smaller. This might be due to the fact that younger people go to school and have a much more relevant social life than the older. Therefore, young people are the category of the population that is more likely to unwillingly spread the virus.
- Lethality index has a correlation coefficient of 0.490763. This behavior is simple to understand: as more people get infected, also more people might die from it.
- Male fraction of the population has a correlation coefficient of 0.637767. Some studies reported that it was more likely for men to get infected, so I have chosen this category as a feature to check the trend over the population. Another reason is that male percentage is a demographic descriptor that could prove useful. From the correlation graph, it seems that a higher male percentage of the population is associated with a higher number of infected people.
- Population density has a correlation coefficient of 0.651351. This is the feature with the highest correlation coefficient with the percentage of infected people. In fact, it should be an obvious consequence that people that live in the countryside or in a less-dense environment are less likely to get infected.

In addition, I have investigated the correlation coefficients of the same selected features with lethality index as the target variable. In this case, the only feature that requires commenting is the positives' percentage: as the number of infected people among the population rises, also the lethality index gets larger (correlation coefficient of 0.490763). This can be explained easily: as the number of infected people grows larger, the most likely it gets that older or sick people get COVID-19 disease and eventually die.

4. The Model

4.1 First Clustering Approach

I have decided to cluster Tuscany's Provinces so that they could be grouped on the basis of the risk that they are experiencing and have been facing since COVID-19 pandemic spread.
The first clustering approach is based on the demographic characteristics of the Provinces and their data regarding the virus. For this reason, I normalized data after cleaning the DataFrame and singled out the best number of clusters to use. In fact, I decided to apply KMeans algorithm, for which it was essential to provide an initial number of clusters, then the algorithm would converge to its optimum.
I reckoned that the Normalizer from sci-kit learn was the best option to bring all data to the same scale. To identify the best number of clusters to use, I have applied the elbow graph method and I have obtained an optimal value of 4. So, I have built the model again, I have assigned the resulting clusters' labels to each Province and, finally, produced a map.

4.2 Second Clustering Approach

The second KMeans algorithm for clustering was applied to another DataFrame, where only the number of cafés and schools was present. In fact, I decided to build a model based only on environmental factors to distinguish between Provinces. As in the previous clustering approach, I have estimated the best number of clusters to use with the elbow method graph, which provided the same result.
So, I have applied this model using 4 clusters and I have obtained 4 classes based on the incidence of cafés and schools on each area. The data were obtained with the same method and given in percentage, so this time I did not apply the normalization process. At the end, I have produced a map showing the results.
I have chosen the amount of schools and cafés situated in the Provinces' areas as features because, as we know, the virus spread is closely associated with people's social lives. Especially for young people, these are the places where we meet the closest friends, make new friendships and lower our guards. Of course, schools are places where to get educated and are essential for

Province	Bars Incidence	Schools Incidence
Arezzo	3.616950	0.556454
Firenze	6.934278	1.341198
Grosseto	3.771633	0.361354
Livorno	12.675111	0.828439
Lucca	7.903916	0.630022
Massa-Carrara	9.049223	1.131153
Pisa	4.270661	0.903409
Pistoia	5.529069	1.147543
Prato	5.216198	1.372684
Siena	4.047378	0.339457

Table 2. The DataFrame used for the application of the second clustering approach.

children's and teen's growing up, so they can't stay closed for long. We shall not forget that these analysis does not have the intention of stating anything, but just to point out the difficulty in pandemic handling and factors linked to it.

4.3 Final Clustering Approach

After assessing the correlation between the bars' (or cafés') incidence and schools' incidence with the percentage of infected people of the population and the lethality index, I have used those environmental factors as additional features to build a final clustering model and get the ultimate groupings.
I have chosen the cafés' and schools' incidence in place of their simple amount by Province because they were correlated, sometimes even strongly, with the total infected people percentage and lethality index. I can't state the same for data reporting only their amount, which showed a correlation coefficient about 0, and therefore couldn't be used for assessing the risk.

After that, I have normalized data, I have chosen the best possible number of clusters (which was 3 this time), I have built the model and assigned labels to each Province. At the end, I have produced the final map with the resulting clusters.

5. Results

5.1 First Clustering Approach

5.1.1 Cluster 0 - Low Risk

This cluster includes Siena, Grosseto and Arezzo Provinces. It defines low-risk areas because its population is quite old and not very dense. In fact, the old age index is negatively correlated with the number of infected people. Also, the percentage of infected people is quite low (the average is near the value of 3%). Lethality index is not alarming (the value ranges from 3 to 4.5) and the male fraction is almost constant.

5.1.2 Cluster 1 - High Risk

Firenze, Livorno and Pistoia belong to this group, which is defined as an high-risk area. The reasons are various: the population is less old than Cluster 0, the Provinces' areas are very

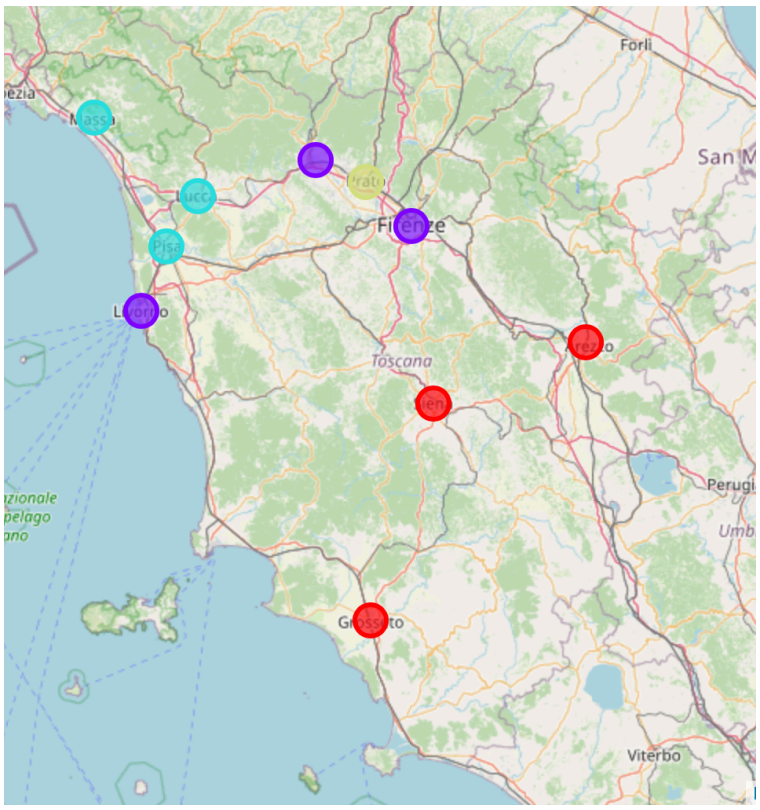

Figure 3. First clustering approach resulting map. Cluster 0 is red, Cluster 1 is purple, Cluster 2 is light blue and Cluster 3 is pale green.

densely populated and the lethality index has a stable value of 7. The positives' percentage ranges from a value of 4 to almost 6%, while the male fraction of the population resembles the one identified in Cluster 0.

5.1.3 Cluster 2 - Medium/Variable Risk

Lucca, Massa-Carrara and Pisa Provinces fall into this cluster. Although they present an overall similar male percentage of the population, their population density is higher than Cluster 0 but lower than Cluster 1. The risk might be variable because they have a variable old age index and a variable lethality index. For the latter, the most at risk Province is Massa-Carrara, where the index reaches its maximum value, about 9. The positives percentage ranges from 4.7 to 5.4%, so its value is lower than the one from Cluster 1.

5.1.4 Cluster 3 - Potential Risk

This group contains only Prato's Province, that seems different from all the other Provinces because of demographic factors. First of all, it presents the highest male percentage of population (of approximately 49%), the lowest old age index (approximately 161) and the highest population density (more than 700 people by square kilometer). Lethality index is average (approximately 5%), while the total infected people percentage is the highest (above 6%).

5.2 Second Clustering Approach

5.2.1 Cluster 0 - High School Density & Medium Café Density

This cluster includes Firenze, Pistoia and Prato. They have a school incidence of 1 school by square kilometer or above and a bars incidence that ranges from 5 to almost 7 cafés by square kilometer. This means that these places have many schools and therefore many young people.

This latter group of people might spend much time in the cafés that are distributed on the whole land.

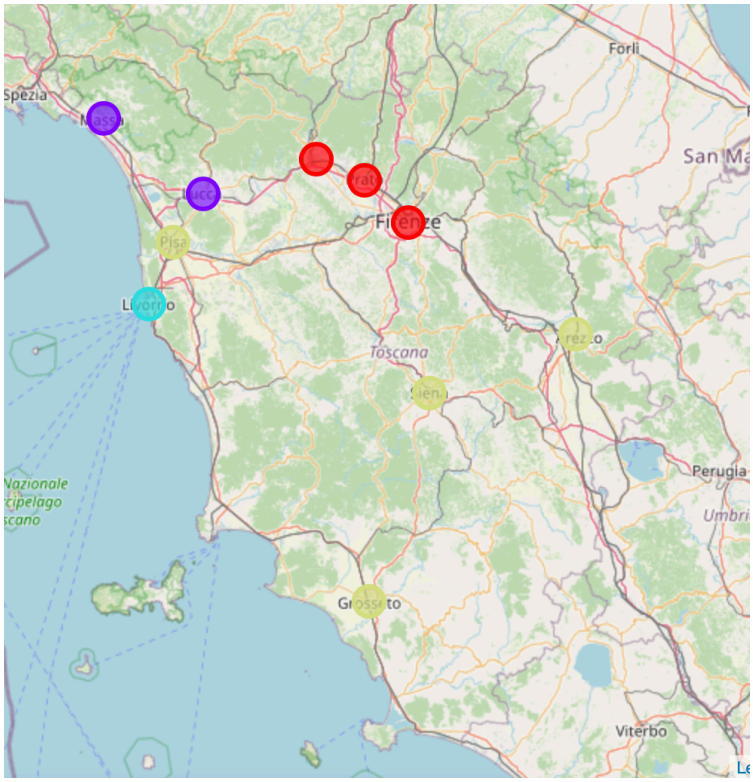

Figure 4. Second clustering approach resulting map. Cluster 0 is red, Cluster 1 is purple, Cluster 2 is light blue, Cluster 3 is pale green.

5.2.2 Cluster 1 - Medium School Density & High Café Density

Lucca and Massa-Carrara's Provinces belong to this group. They have less schools on their area than Cluster 0, but more cafés. In this case, bars are probably attended by young people, but not only.

5.2.3 Cluster 2 - Medium School Density & Very High Café Density

This cluster is defined by a single Province: Livorno. This time, Livorno is the most different Province among Tuscany cities instead of Prato. Although its school density is almost similar to the values identified for Cluster 1, its cafés incidence is much more influential (it doubles the value of Cluster 0). This means that people's social lives are very often spent in cafés and this may be a common behavior for the whole population.

5.2.4 Cluster 3 - Low School Density and Low Café Density

This cluster comprehends the left-out Provinces of Siena, Grosseto, Pisa and Arezzo. The schools' incidence is relatively low and bars' incidence is lower than the one shown by Cluster 0 (maximum value of 4 cafés by square kilometer). Although the values are lower by area extension, their value could be proportioned to the inhabitants. Their presence could be frequent in certain areas, while the remaining land is mostly countryside. Therefore, we cannot infer much from this cluster directly, because we need confronting with demographic data.

5.3 Final Clustering Approach

5.3.1 Cluster 0 - High Risk

Firenze, Livorno, Lucca, Massa-Carrara, Pisa and Pistoia are included in this cluster. They are characterized by a high population density along with a large lethality index (from almost 6 to the maximum of 9.8 of Massa-Carrara's Province). The population is relatively old and the positives' percentage ranges from 4.0 to 5.9%. Although 4.0% is a small percentage of infected people, this value is attributed to Livorno, that has a very high café density and a medium schools' incidence as underlined before. The café density could be associated with a higher lethality index. The schools' incidence is overall medium or high for all these Provinces. Because of all these factors, these cities sustain a major potential risk of infection.

5.3.2 Cluster 1 - Potential Very High Risk

Prato alone belongs to this class and it is the Province that again differs from all the others as it was identified with the first clustering approach. In fact, it has a very high schools' incidence, which could be predictable given the high population density. The risk is therefore very high and supported both by demographic and environmental factors. The conditions could get less preoccupying in the future because the lethality index is quite low (about 5), but, if the situation is not kept under control, it could precipitate pretty easily.

5.3.3 Cluster 2 - Low Risk

From this analysis, Siena, Arezzo and Grosseto result as the safest cities in Tuscany. Their lethality index is pretty low, as well as their total infected people. Population is well distributed on the territory providing easy isolation of the areas where the virus is spreading. This is supported by environmental factors too, since the schools are of a manageable amount, as well as bars. In addition, the population is generally old, so the positives' percentage of population is not believed to exponentially grow.

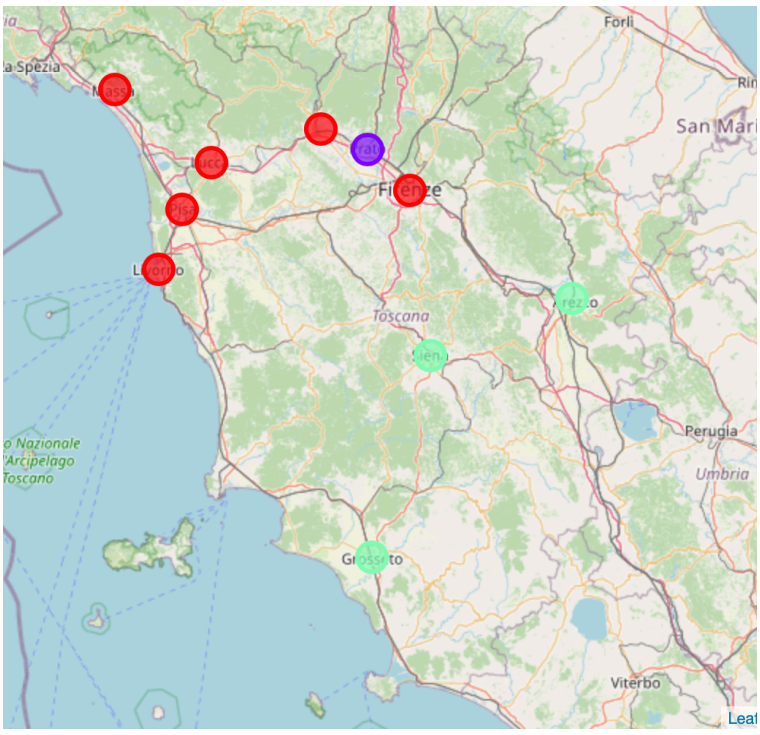

Figure 5. Final clustering approach resulting map. Cluster 0 is red; Cluster 1 is purple; Cluster 2 is green.

6. Discussion of Results

The areas of risk in the maps are well defined by lots of features. From this analysis, it seems that the spread of the virus is mainly due to young people and children going to school. On the contrary, old people could be more scared and could tend not to leave their homes. This may be an explanation for the negative correlation between the old age index and the total infected people. It is also true that it is impossible to deprive the young of their social and educational life for a long time, as it would have serious psychological consequences on them. This is why vaccination campaigns and citizenship are the most important values to be transmitted by the government to the population.

The attendance of schools and cafés is also of primarily importance for economic factors and development of the future generations, activities that cannot be suspended for long, especially in Tuscany and the whole country as tourism is one of the main sources of income. In addition, cafés' incidence does not seem to be correlated with the positives' fraction of population. Although it is correlated with the lethality index, that could be an intrinsic tendency not explained otherwise. For this reason, I think that re-openings of cafés could be relevant for people that own them, to allow them to be financially stable, given the low risk of getting infected if distancing measures are respected.

Population density is one of the features which is more steeply correlated with the infected people's percentage. This is very well resembled in the maps, where the biggest and closest cities are also the ones that are identified as high-risk areas, while the Provinces rich in countryside are also those that have been less touched. Prato is known to be different from all the other cities. This is because it has a large population made of young people, generally known to work in textile factories and in a multicultural environment since they have different nationalities. The Italian population is old, while immigrants are generally younger and with larger families. That's why Prato's population is much more dense and much younger than average. COVID-19 seems to provoke the worse consequences in older people, so even though young people might get infected more easily, they should be able to overcome the disease in a short period of time. Still, they could transmit the infection to many others, and the situation in hospitals could escalate easily. This is why Prato's situation is being closely monitored and it has been experiencing the most restricting measures in Tuscany.

The male percentage of population is predicted to be associated with the total infected people's percentage, but presents a slightly negative correlation coefficient with the lethality index. This does not necessarily confirm the hypothesis that male people are more keen to get infected. In fact, even if they would be more prone to get COVID-19, they are not likely to be the most-at-risk category, because the deaths do not only depend on the sex, but probably on the general condition of the patient and his/her chronic or pre-existing illnesses.

7. Conclusions

This analysis provides new insights into the current and past Tuscany situation regarding COVID-19 infection. The present trends could evolve in the future and this model could be re-applied simply by using the daily updated dataset supplied by "Open Data Toscana". Therefore, it is a very flexible and customizable model that can be used for any type of data analysis. I applied clustering but it is also possible to build a regression model to predict the total infected people's percentage or lethality index, based on both demographic and environmental factors.