Business Problem

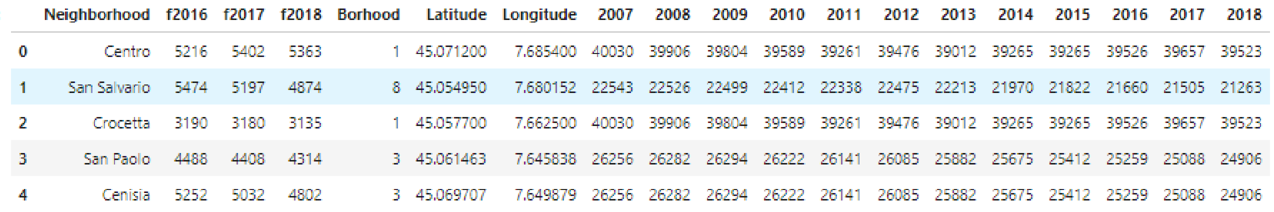
The aim of this work is to find the best place in Turin to open a restaurant. This can be particular useful for new business owner that wants to find the perfect area of the city to start their new activity.
In detail this work shows what would be the best place for a new restaurant and what kind of restaurant will be more profitable in that particular area.
The choice of the place where locate a venue for a new business acitivity is very important and can change the future gain.

Data Source

The data sources that I am going to use are:
- Turin Municipality statistics data (http://www.comune.torino.it/statistica/dati/): this are data provided by Turin municipality and related with the demography of the different Neighborhoods. This data included the overall people that live in a particulare neighborhood as well as the percentage of the foreign residents.
- Foursquare (https://it.foursquare.com/): the data retrived from forusquare website are related with the different business venues located in turin. We will use these data to understand where are located the current restaurants and what kind of restaurant (category) we are talking about.

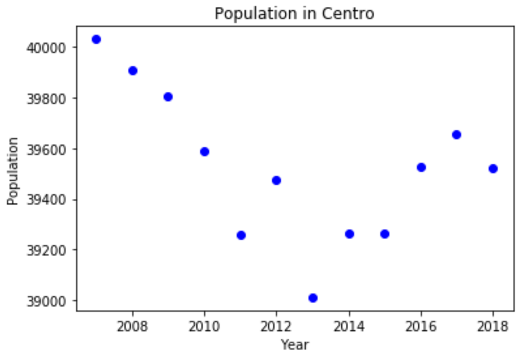
Methodology

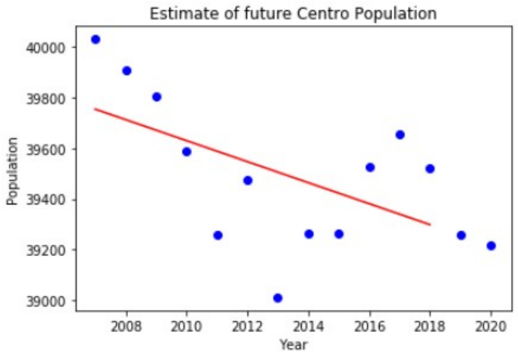
The first part of this work is data cleaning and wrangling in particular regarding the data provided by Turin municipality. The raw data provided are indeed fragmented, moreover some data are divided by Neighborhood, other data are organized in more greater Dinstrict (in italian *Circoscrizioni*) therefore it is necessary to put everything in a single dataset organized by Neighborhood.
Where were present only the District data and not the Neighborhood one, it was splitted as approximation the population equally between its Neighborhood.
After the data cleaning the population dataset have the following structure, where under the year we have the population of that specific Neighborhood at that year.



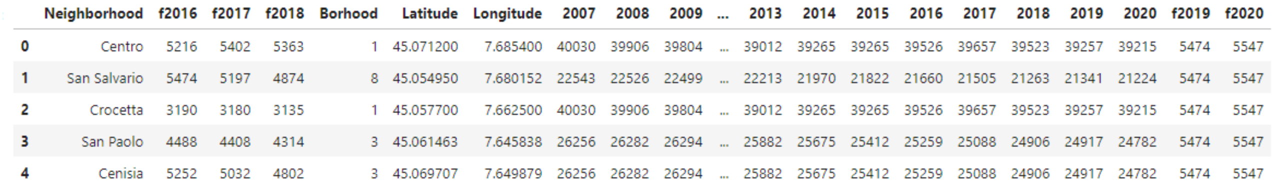
	Neighborhood	f2016	f2017	f2018	Borhood	Latitude	Longitude	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
0	Centro	5216	5402	5363	1	45.071200	7.685400	40030	39906	39804	39589	39261	39476	39012	39265	39265	39526	39657	39523
1	San Salvario	5474	5197	4874	8	45.054950	7.680152	22543	22526	22499	22412	22338	22475	22213	21970	21822	21660	21505	21263
2	Crocetta	3190	3180	3135	1	45.057700	7.662500	40030	39906	39804	39589	39261	39476	39012	39265	39265	39526	39657	39523
3	San Paolo	4488	4408	4314	3	45.061463	7.645838	26256	26282	26294	26222	26141	26085	25882	25675	25412	25259	25088	24906
4	Cenisia	5252	5032	4802	3	45.069707	7.649879	26256	26282	26294	26222	26141	26085	25882	25675	25412	25259	25088	24906

After this first data wrangling activity, it was used liner regression in order to extimate the population in 2019 and 2020, both for the overall one and for the foreign one. Below it is shown an example regarding the Neighborhood "Centro". On the left side it is possible to see the data collected between 2007 and 2018,



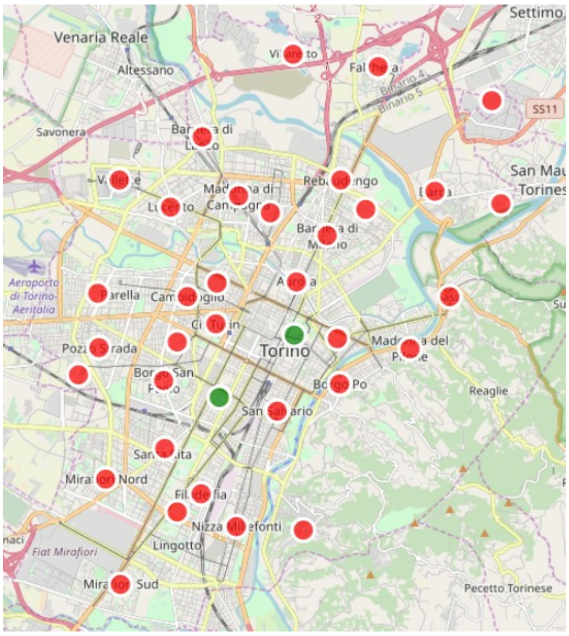


The same linear regression was applied to every Neighborhood. Below it is possible to see the result dataframe, where "f" before the year stays for foreigners population.

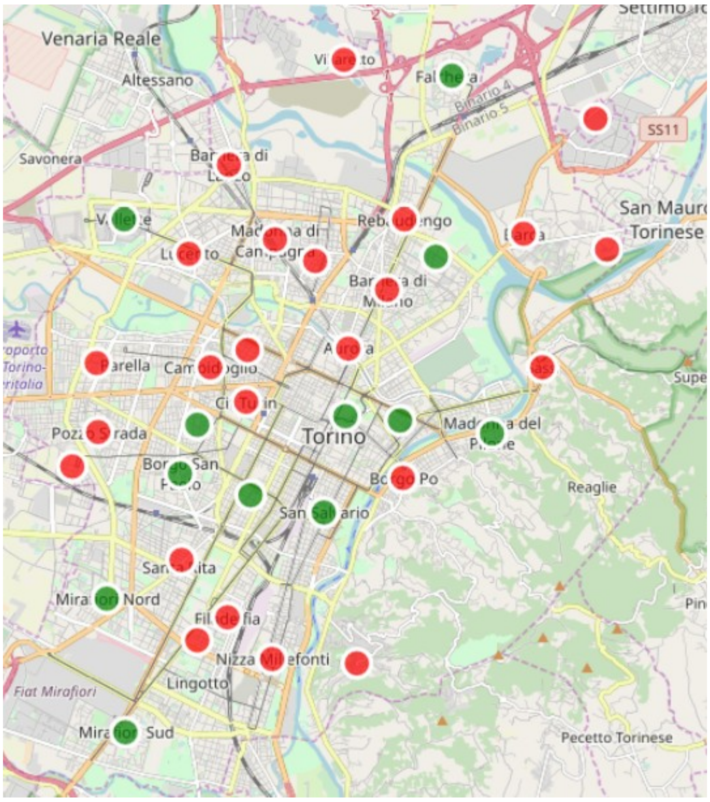


	Neighborhood	f2016	f2017	f2018	Borhood	Latitude	Longitude	2007	2008	2009	...	2013	2014	2015	2016	2017	2018	2019	2020	f2019	f2020
0	Centro	5216	5402	5363	1	45.071200	7.685400	40030	39906	39804	...	39012	39265	39265	39526	39657	39523	39257	39215	5474	5547
1	San Salvario	5474	5197	4874	8	45.054950	7.680152	22543	22526	22499	...	22213	21970	21822	21660	21505	21263	21341	21224	5474	5547
2	Crocetta	3190	3180	3135	1	45.057700	7.662500	40030	39906	39804	...	39012	39265	39265	39526	39657	39523	39257	39215	5474	5547
3	San Paolo	4488	4408	4314	3	45.061463	7.645838	26256	26282	26294	...	25882	25675	25412	25259	25088	24906	24917	24782	5474	5547
4	Cenisia	5252	5032	4802	3	45.069707	7.649879	26256	26282	26294	...	25882	25675	25412	25259	25088	24906	24917	24782	5474	5547

According with this data it is possible to see below the Neighborhood that will have in 2020 more population than 2018 (in green).



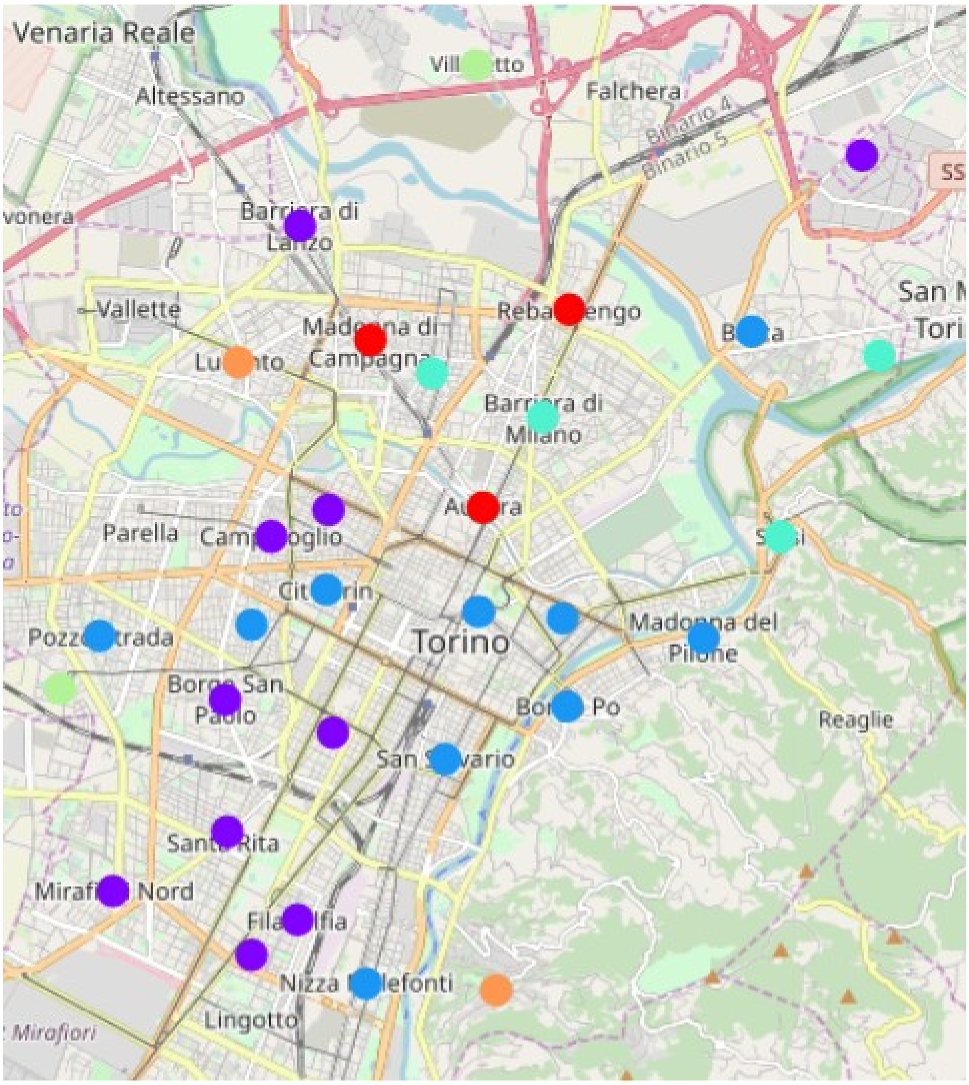
Below it is possible to see which Neighborhood will have in 2020 an increase in foreigners compared with what we have today. Again the green means increase in population.



The second data used are the one provided by foursquare API. With this data it was retrived for each Neighborhood all the venues in that area (with a radius of 500m compared with the center of Neighborhood). The dataset was then cleaned up dropping all the value that were not concerning restaurants.

Using this reastaurant dataset it was clustered the area of turin with the different Negiborhood. This was used with k-mean clustering. It was used this tecnique because we want to identify some area that have similarities in therm of restaurants.
The results are shown below:

- Cluster 0 (Red): 1st Chinese/Pizza/Indian Restaurant
- Cluster 1 (Purple): 1st Pizza/Italian/Mediterranean Restaurant
- Cluster 2 (Blue): 1st Italian/Pizza Restaurant
- Cluster 3 (Aqua-Blue): Pizza Place/Mexican/Indian Restaurant
- Cluster 4 (Light Green): Mainly foreign restaurant, this area was difficult to identify due to the fact that Foursquare has several restaurant that hase onlu the category "Restaurant" and no additional information about the cousin type. For this reason these area (Villaretto and Borgata Lesna) should be excluded from the study
- Cluster 5 (Orange): Italian Restaurant



Results

Considering the clustering results obtained we can consider that:

- The best place to open a new italian resturant (based on the fact that the concentration of italian restaurat today is low) is cluster 0 and cluster 3
- The best place to open a new restaurant with non-italian cusine (based on the fact that the concentration of italian restaurat today is high) is cluster 5, cluster 1 and cluster 2

If integrate this data with the one obtained with the demographic studies, it looks like that if we want to open a new restaurant it is better to open it in the center of the city. Unfortunatelly it looks like that these were the area were it was not convenient to open a new italian restaurant, for this reason the population study does not help us to understand the conveniency to open a new restaurant.
If we consider instead the a non italian kitchen, and imagine that this can be preferred by foreigners we have several places that can be good to open a new restaurant. In particular all the area around the downtown.

Observation and recommendations

The work that was done has some limits, in particular:

- The data used for the population study were very fragmented, some hypotesis were done in order to work with these data. However haveing more complete data could help to have a better statistics.
- The linear regression used to predict the behaviour of the next year is very depending on the number of year that we used to train the model. In 2014 Turin stopped its population growing and it slowly started to decrease. If it is used the data from 2014 it gives at results that the population is going down, if they are used also the data before (2007-2014), this gives as results a growth in the population. It is more reliable the first case, this is the one used in this work.
- More optimization can be done in clustering the venues, as well as retriving the venue from foursquare. An interesting case could be retriving also the rating of each structure in order to check also the quality of the service and not only if a restaurant is present or not. It was not possible to do this since it requires a foursquare premium license.

Conclusion

In this work some specific area to open a new resturant in the city of Turin were identified.