Living in the MotorValley

Which is the best place to live for those who start working in Ferrari?

This is the final project of the Coursera Applied Data Science Capstone course

By Sebastian D'Amico

Introduction

Ferrari headquarter is in Maranello, a small town at about 18 km from Modena, with a population of 17,504 (as of 2017). It is known worldwide as the home of Ferrari and Scuderia Ferrari Formula One racing team. Several other towns surround Maranello, and Modena, with 184.000-ish inhabitants, is the closest and biggest town with Shopping Centers, University, nightlife and many other services that push most of the people, joining Ferrari, to look for a house. Obviously is difficult to find all the services that are present in Modena in any other small town that surround Maranello, but which are the main differences between all the towns? The aim of this project is to classify the Maranello and surrounding towns in terms of available services and venues to help people joining Ferrari to judge, with real data, which is the place that better suits their own requirements.

The data

Different datasources will be used for this project. First of all we will take all the towns in the province of Modena (47 total municipalities) from the following website: https://zip-codes.nonsolocap.it/emilia-romagna/91-cap-province-of-modena/
We will then try to use the Geocoder Python to get coordinates from each postal code. In case of failing, we will manually extract Latitude and Longitude from Google Maps for each town. On top of that, the distance from Ferrari will be associated to each town so that people can judge also based on the time they will spend to go to the office. If we don't manage to get the distance using Google API, we will extract it manually.
Finally, Foursquare will be used to explore each town, extracting information of all the venues categories that will be used for having a better picture of what can be found in each town. Clustering algorithm will be used to automatically create clusters based on most common venues, giving attention also to the optimization of the number of clusters.
Plots and tables will help to better analyze the data and to drive the analysis also based on the results, still having as main target what already described in the introduction.

Methodology

Data collection

The raw data table is as follows:

	Location	Latitude	Longitude	DistanceToMaranello
0	Modena	44.650177	10.921732	22.96
1	Marzaglia	44.650963	10.803609	20.87
2	Cittanova	44.650268	10.850284	21.80
3	Cognento	44.636302	10.871811	16.96
4	Baggiovara	44.607764	10.867751	11.27

Data has been collected from the following websites:

- Modena Neighborhood details: https://www.comune.modena.it/decentramento/il-decentramento-a-modena/la-frazioni-centri-di-periferia
- Towns in the province of Modena: https://zip-codes.nonsolocap.it/emilia-romagna/91-cap-province-of-modena/
- Coordinates for each town: https://www.coordinate-gps.it/
- Driving distance between locations: https://www.mapdevelopers.com/distance_from_to.php

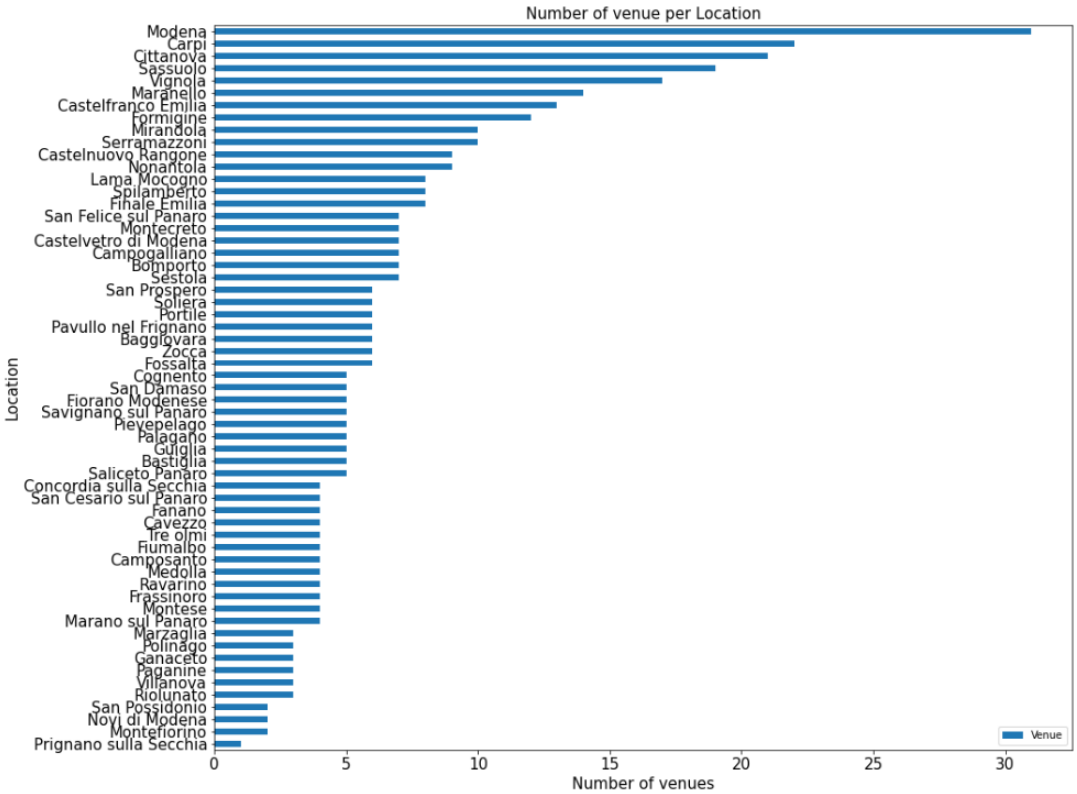
None of the sources above have clear or public APIs so the data has been gathered manually and the .csv file is available at the following link: https://github.com/sebas1986/TheMotorSportValley-WhereToLive-/blob/master/locations.csv
Having the raw table with locations and coordinates, the very next step has been to fetch, using Foursquare API, the venues for each location. Please refer to the public documentation for more details on Foursquare API: https://developer.foursquare.com/docs
Worth mentioning that when using the Foursquare API, we need to provide, on top of the coordinates, the radius of the area where to find venues and the maximum returned venues in that area. In this analysis a fixed radius of 1.0 km has been used as a good compromise considering all the different size of towns we have in our raw table. An improvement on this aspect would be to use a dynamic radius which adapts accordingly to the size of the town.
With the values fetches from Foursquare, we have created a table as follows:

	Location	Location Latitude	Location Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Modena	44.650177	10.921732	Piazza della Pomposa	44.649044	10.923808	Plaza
1	Modena	44.650177	10.921732	Osteria Ermes	44.649429	10.925370	Italian Restaurant
2	Modena	44.650177	10.921732	La Tenda	44.651706	10.919946	Event Space

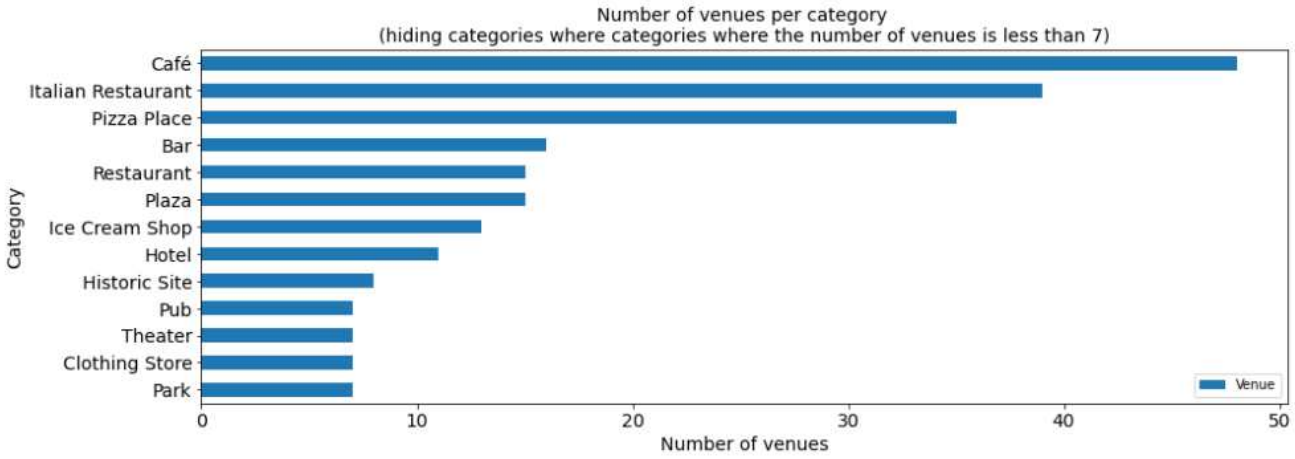
In the table above, each row is a venue with its coordinates, name and category.

Data overview
Before going further in the analysis, we wanted to have a clear view on the amount of venues, categories and how they distribute along the different towns.

The plot on the right shows the number of venues per location. On top we find Modena, which is not an unexpected result being the biggest town in the area (184.000-ish inhabitant). It is followed by Carpi and Cittanova. Please note that there are towns with very few venues.



Venues categories is also an important information to consider. The plot below shows the most important categories with the related number of venues.



Interesting to see that in the top 3 we can find:

- Cafè
- Italan Restaurant
- Pizza Place

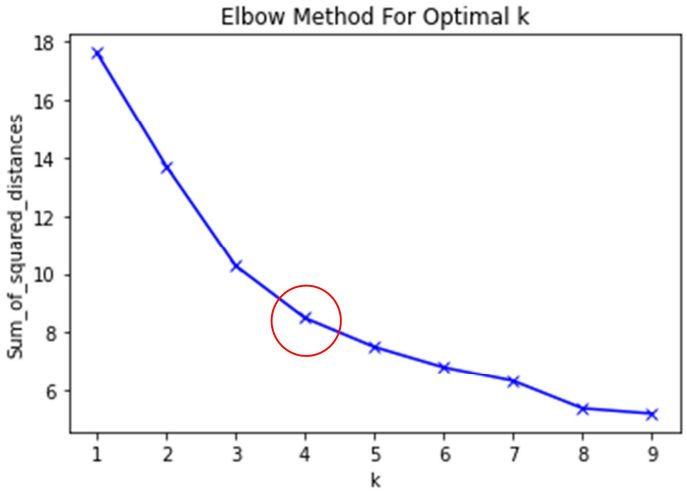
It is already clear that in this area the food is one of the main businesses, with high level restaurants (such as "La Francescana", one of the most famous restaurants from the chef Massimo Bottura), but in general very good quality in almost every place.

Clustering the data

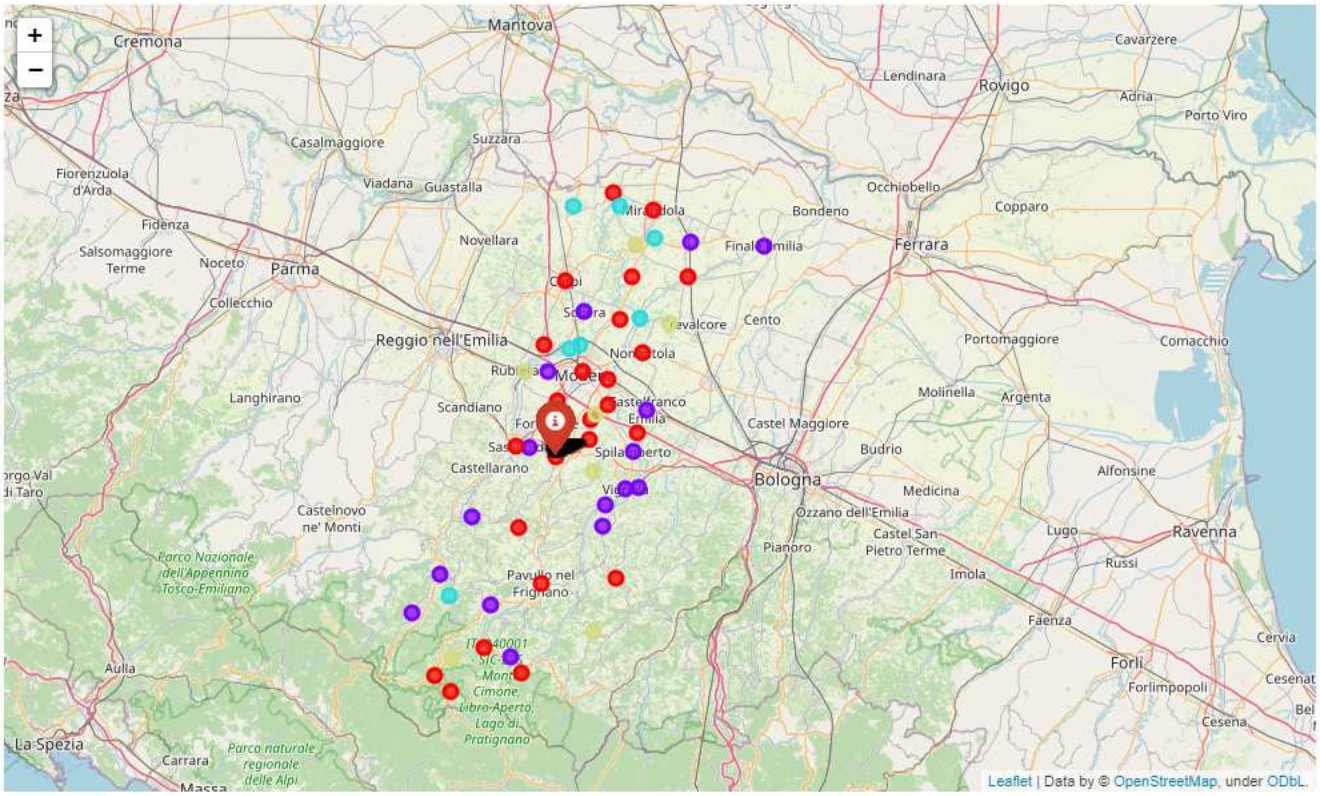
With the information collected so far, we decided to use the K-Means clustering algorithms to divide the different locations based on the venues categories. To avoid using data with very few samples, we decided to extract only categories with more than 7 venues and on top of that we renamed similar categories with the same name (e.g. Café and Bar renamed as Bar).
Then we grouped the data by location calculating the normalized mean for each location. The table below shows a sample:

	Location	Bar	Clothing Store	Historic Site	Hotel	Ice Cream Shop	Park	Pizza Place	Plaza	Pub	Restaurant	Theater
0	Baggiovara	0.500000	0.0	0.0	0.5	0.0	0.0	0.000000	0.0	0.000000	0.000000	0.0
1	Bastiglia	0.333333	0.0	0.0	0.0	0.0	0.0	0.333333	0.0	0.000000	0.333333	0.0
2	Bomporto	0.000000	0.0	0.0	0.0	0.0	0.0	0.666667	0.0	0.333333	0.000000	0.0

Table above was the input of the clustering algorithm. The optimal number of clusters has been calculated using the Elbow Method (see this link for further details: https://blog.cambridgespark.com/how-to-determine-the-optimal-number-of-clusters-for-k-means-clustering-14f27070048f) that suggested a number of 4 clusters (as shown in the following plot):



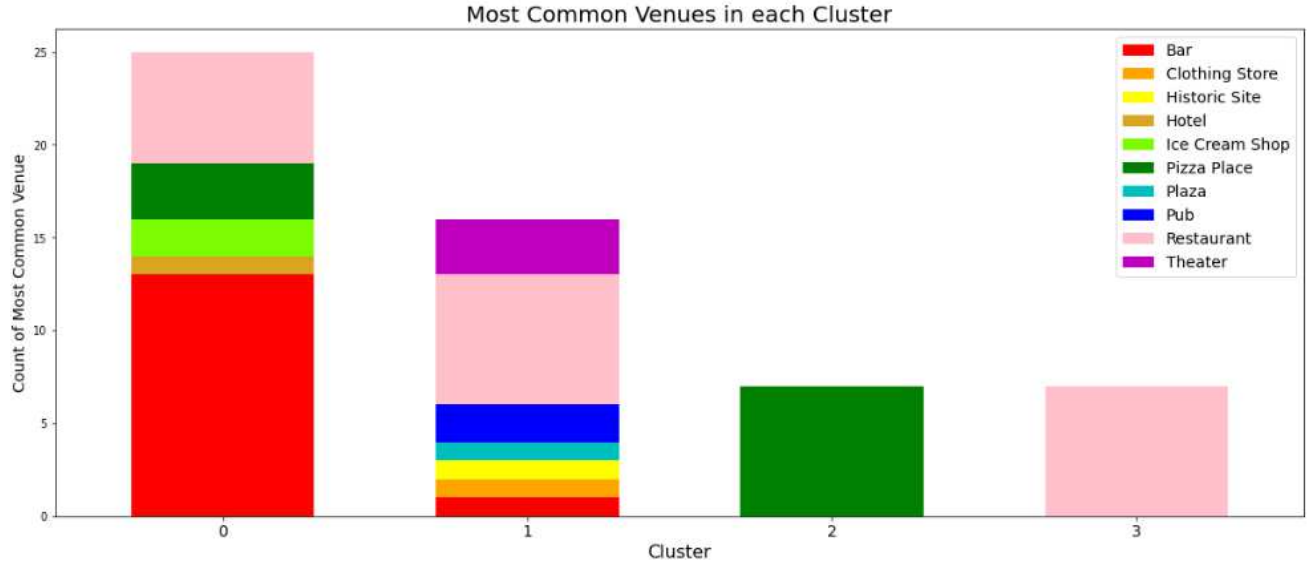
The four clusters are showed in the map below, where each color represent a different cluster and each marker a different location.



The marker with the "i" icon is where Ferrari Headquarter is located.
Maps has been produced using the "Folium" library. Please visit the following link for the full documentation: https://python-visualization.github.io/folium/

Results

The analysis of the clusters characteristics is one of the most interesting pars. Based on the 1st most common venue of each cluster, the plot below shows how the categories are distributed, in terms of number of venues, over the 4 clusters.



The information provided by the plot above can be used to label each cluster. The labels describe the characteristic of the cluster itself, which is the main goal of this analysis. Here we go:

- **Cluster 1**: several Bars and Restaurant/Pizza with Ice Cream shops and few Hotels;
- **Cluster 2**: several Restaurants but also social areas like theaters, pubs, Historic sites and clothing stores;
- **Cluster 3**: mainly pizza places;
- **Cluster 4**: mainly restaurants.

As already mentioned, the analysis could have been more precise by choosing a dynamic radius (instead of fixed) based on the location to properly get all the venues. For example, Modena is a big city but increasing the radius would affect the other small locations including outside venues. A compromise has been found.
Also, including the house pricing per square mt would be an important information for those who want to buy a house.
The distance to Ferrari (Maranello) is obviously an important information to consider while choosing the right place, together with the population density.
One of the main results from this analysis is, as already spotted from the beginning, that food is generally a common denominator for almost all the locations, and this can be confirmed by anyone that lives in the area.

Conclusions

People joining Ferrari have a lot of places where to find a house. Different services and different venues are available but the most common category here is "food". On top of that, based on the distance from the office, one can choose to stay in areas whit more green and less population density or places like Modena with more people.
Overall, the final decision to find an optimal location for a new resident will be made by the individuals themselves. As well as finding about how the different locations are characterized by the number of venues, they should also consider other factors such as transport, housing prices and access to different necessities.