A beautiful city for a good investment

Claudio Calamita – Physicist – Data Scientist

Contents

Introduction **1**

Data Requirements **2**

Data Collection and Understanding **4**

Analyze economy of the city 4

Build dataframe for Municipalities and Neighborhoods 5

Extract climate data 7

Descriptive analysis 8

Foursquare API data collection 10

Model Clustering Neighborhoods **12**

Determining the best value for k 12

Evaluating the model 13

Results **14**

Conclusion **17**

Introduction

Naples is the city where I was born. It is in Italy and is a wonderful city, near the sea and with a mild climate.

In this city you can eat very good food everywhere, from pizza to pasta and thousand of different cakes.

It is a city of art, full of museums and churches where you can find different art styles as Baroque, Neoclassicism and Romantic.

It is historic since the city was conquered by different population in the past as Angevin and Aragonese.

The people are very kind with everyone and it is an alive city also for night life.

It is a city of culture, indeed, there are different universities which are very important as University Federico II that is one of the oldest in the world.

The site Teleport ([1]) asserts:

> *Naples, Italy, is characterized by reasonably priced housing. Our data reflects that this city has a good ranking in health-care and tolerance.*

The site also considers Naples as 17th from a total of 163 countries for what each country on earth contributes to the common good of humanity, and what it takes away, relative to its size.

The mayor is trying to give an impulse to the city. He is facing the criminality and dealing with public debt, aiming to increment tourism. In the 2014 and 2013 he could get the charge to host respectively the **Copa Davis** and **Copa America**.

Naples is a city with a very high density of population so it could be a good investment for a local as a restaurant, or hotels, or pizza shop, coffee shop.

I want to use data to show what is the best area for an investment by stakeholders in these city.

Data Requirements

Naples is structured in Municipalities and Neighborhoods. There are 10 municipalities and 31 neighborhoods. I want to find the best area for an investment in commercial area. So I want to get the venues for every municipality, do a clustering of municipalities according their venues and then find the area. Mainly I need for this geospatial data. In particular, the data I will need for my notebook are:

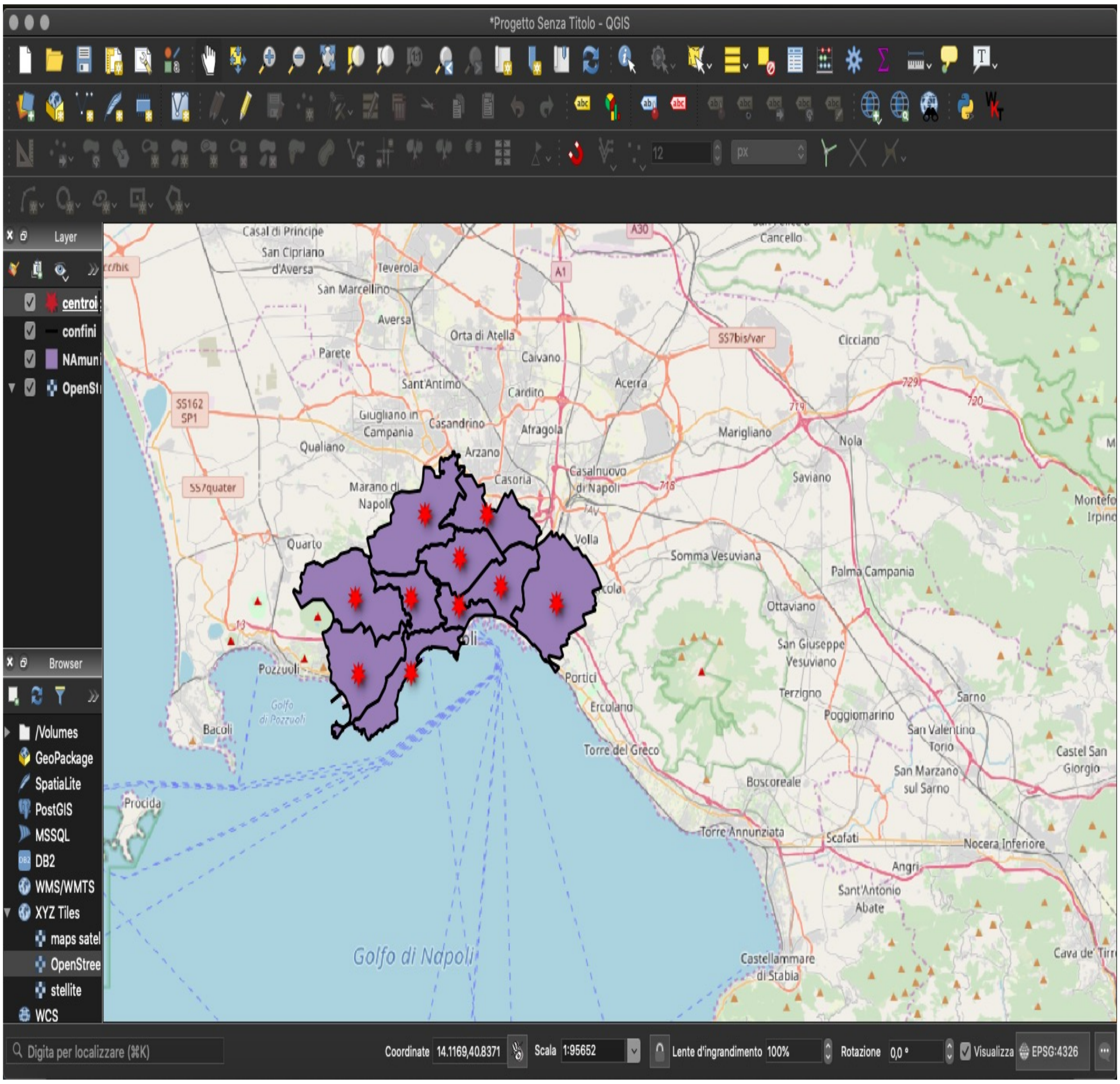
- Data for economy of the city. I will use BeautifulSoap[1] to scrape these data from Wikipedia [2];
- Data for municipality and neighborhoods. I will use BeautifulSoap to scrape these data from Wikipedia[3];
- Data for boundaris of every municipality. I will download the data from open data of the website of the city[4]. These data are in the shapefile format. This format is a GIS (Geospatial Information System) standard for geospatial data. Every data is described in the standard WKT (Well Known Text) that describes an element of a map with Point, Linestring, Polygon. In this case the data are polygons that are difficult to manipulate. Yet, from polygons is possible to extract the boundaries as Linestring and the centroids as Point. I did this with an open-source tool QGIS. With shapefiles of boundaries and centroids it is easier to visualize Municipalities on Folium[2] map. In figure 1 you can see an example of what I mean.;
- Data for climate of the city. Being a city of sun, with a good climate, it is full of tourists the whole year. I will scrape them from a site[5] with BeautifulSoap;
- Data from Foursquare API[3] to extract venues for every municipality. The referring point for every municipality is the centroid of the municipality extracted as seen before;

As I mentioned the idea is to cluster municipality by venues, analyze the features of clusters. Then do a heat-map for the area of investment described previously and for every area find what is the best place to invest using density map or contour map.

[1] BeutifulSoap is a powerful Python library for scraping website

[2] Folium is a Python library to visualize geospatial data on a map

[3] he Foursquare Places API provides location based experiences with diverse information about venues, users, photos, and check-ins



*Progetto Senza Titolo - QGIS

Coordinate 14.1169,40.8371 Scala 1:95652 Lente d'ingrandimento 100% Rotazione 0,0 ° ☑ Visualizza EPSG:4326

Figure 1: Example of shapefile imported in QGIS. In violet the polygons representing the municipalities. In black the boundaries. In red star the centroids.

Data collection and Understanding

In this chapter we will see the collection of the data and the analysis of them. It's a fundamental step for preparing to modeling the problem.

The graphs in this chapter and next will be done using matplotlib, a data visualization library in Python.

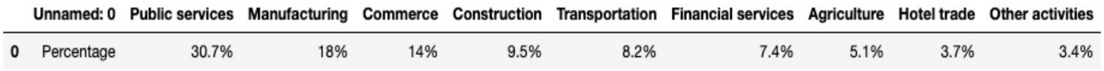
Analyze economy of the city

Naples is Italy's fourth-largest economy after Milan, Rome and Turin, and is the world's 103rd-largest urban economy by purchasing power, with an estimated 2011 GDP of US dollar 83.6 billion, equivalent to $ 28749 per-capita.

Naples is a major cargo terminal, and the port of Naples is one of the Mediterranean's largest and busiest. The city has experienced significant economic growth since World War II.

Naples is a major national and international tourist destination, being one of Italy and Europe's top tourist cities. Tourists began visiting Naples in the 18th century, during the Grand Tour. In terms of international arrivals, Naples was the 166th-most-visited city in the world in 2008, with 381000 visitors (a 1.6 per cent decrease from the previous year), coming after Lille, but overtaking York, Stuttgart, Belgrade and Dallas [2].

Figure 2 shows how is distributed the economy of the city. Investment in hotel is just 3.7 %, commerce 14 % so it could be a good investment in this area since the city is not filled.

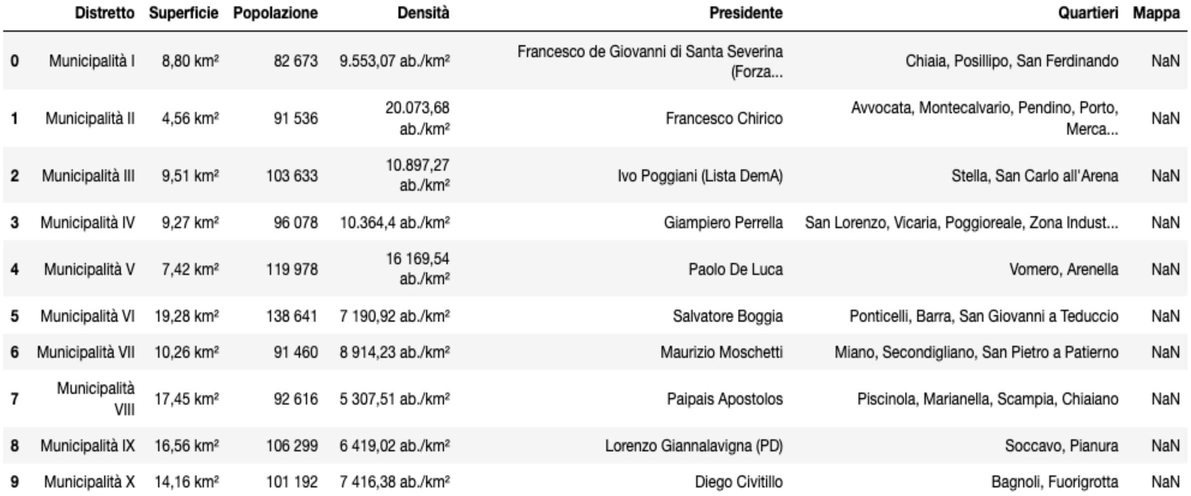


	Unnamed: 0	Public services	Manufacturing	Commerce	Construction	Transportation	Financial services	Agriculture	Hotel trade	Other activities
0	Percentage	30.7%	18%	14%	9.5%	8.2%	7.4%	5.1%	3.7%	3.4%

Figure 2: Dataframe describing economy of the city.

Build dataframe for Municipalities and Neighborhoods

The dataset for municipalities and neighborhoods is scraped from Wikipedia[3] using the Python library BeautifulSoap and imported in the notebook as a Pandas dataframe[4]. In figure 3 you can see the raw data imported as a dataframe. This dataset should be cleaned ad adjusted to be usable. Let's drop the columns "Presidente" indicating the president for municipality (of no use in this case) and "Mappa" which contained the maps of each municipality in Wikipedia as images (clearly the images are not scraped). Then we should add a referring for latitude and longitude of every municipality. For that I download the open-data [4] and I extracted the centroids of the polygons of municipalities to get the referring coordinates. As seen in the chapter "Data Requirements" I did it with the open source software QGIS and used the library shapefile of Python to read the centroids saved locally and then updated in my github[6].



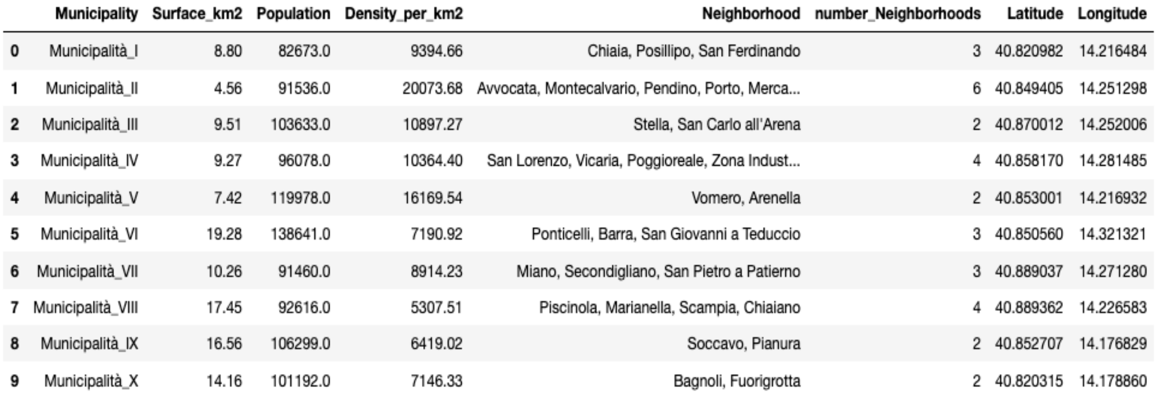
	Distretto	Superficie	Popolazione	Densità	Presidente	Quartieri	Mappa
0	Municipalità I	8,80 km^2	82 673	9.553,07 ab./km^2	Francesco de Giovanni di Santa Severina (Forza...	Chiaia, Posillipo, San Ferdinando	NaN
1	Municipalità II	4,56 km^2	91 536	20.073,68 ab./km^2	Francesco Chirico	Avvocata, Montecalvario, Pendino, Porto, Merca...	NaN
2	Municipalità III	9,51 km^2	103 633	10.897,27 ab./km^2	Ivo Poggiani (Lista DemA)	Stella, San Carlo all'Arena	NaN
3	Municipalità IV	9,27 km^2	96 078	10.364,4 ab./km^2	Giampiero Perrella	San Lorenzo, Vicaria, Poggioreale, Zona Indust...	NaN
4	Municipalità V	7,42 km^2	119 978	16 169,54 ab./km^2	Paolo De Luca	Vomero, Arenella	NaN
5	Municipalità VI	19,28 km^2	138 641	7 190,92 ab./km^2	Salvatore Boggia	Ponticelli, Barra, San Giovanni a Teduccio	NaN
6	Municipalità VII	10,26 km^2	91 460	8 914,23 ab./km^2	Maurizio Moschetti	Miano, Secondigliano, San Pietro a Patierno	NaN
7	Municipalità VIII	17,45 km^2	92 616	5 307,51 ab./km^2	Paipais Apostolos	Piscinola, Marianella, Scampia, Chiaiano	NaN
8	Municipalità IX	16,56 km^2	106 299	6 419,02 ab./km^2	Lorenzo Giannalavigna (PD)	Soccavo, Pianura	NaN
9	Municipalità X	14,16 km^2	101 192	7 416,38 ab./km^2	Diego Civitillo	Bagnoli, Fuorigrotta	NaN

Figure 3: Dataframe describing raw data for municipalities and neighborhoods.

After cleaned the dataframe and added latitude, longitude and number neighborhoods for every municipality, it appears as in figure 4

Let's import also the boundaries in the same manner as centroids (the dataset is always in my github [6]) and plot on the map the data for centroids and boundaries. See figure 5 showing a Folium map for boundaries and centroids.

[4] Pandas is a fundamental Python library for data analysis. A dataframe is data structure that can be imagined as a table with indices for rows and columns.



	Municipality	Surface_km2	Population	Density_per_km2	Neighborhood	number_Neighborhoods	Latitude	Longitude
0	Municipalità_I	8.80	82673.0	9394.66	Chiaia, Posillipo, San Ferdinando	3	40.820982	14.216484
1	Municipalità_II	4.56	91536.0	20073.68	Avvocata, Montecalvario, Pendino, Porto, Merca...	6	40.849405	14.251298
2	Municipalità_III	9.51	103633.0	10897.27	Stella, San Carlo all'Arena	2	40.870012	14.252006
3	Municipalità_IV	9.27	96078.0	10364.40	San Lorenzo, Vicaria, Poggioreale, Zona Indust...	4	40.858170	14.281485
4	Municipalità_V	7.42	119978.0	16169.54	Vomero, Arenella	2	40.853001	14.216932
5	Municipalità_VI	19.28	138641.0	7190.92	Ponticelli, Barra, San Giovanni a Teduccio	3	40.850560	14.321321
6	Municipalità_VII	10.26	91460.0	8914.23	Miano, Secondigliano, San Pietro a Patierno	3	40.889037	14.271280
7	Municipalità_VIII	17.45	92616.0	5307.51	Piscinola, Marianella, Scampia, Chiaiano	4	40.889362	14.226583
8	Municipalità_IX	16.56	106299.0	6419.02	Soccavo, Pianura	2	40.852707	14.176829
9	Municipalità_X	14.16	101192.0	7146.33	Bagnoli, Fuorigrotta	2	40.820315	14.178860

Figure 4: Dataframe describing cleaned data for municipalities and neighborhoods.

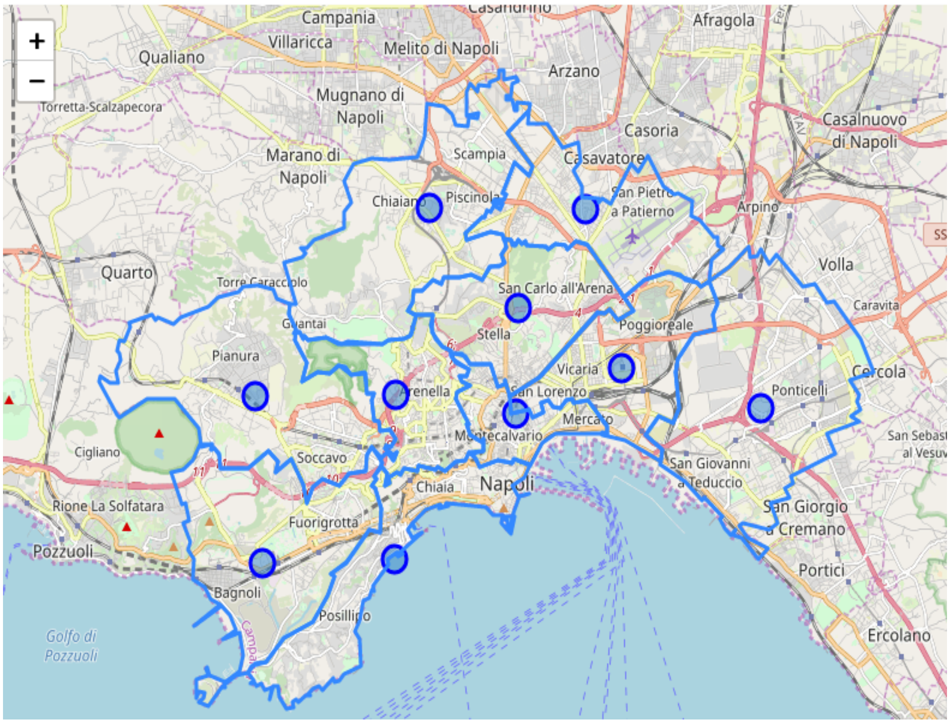
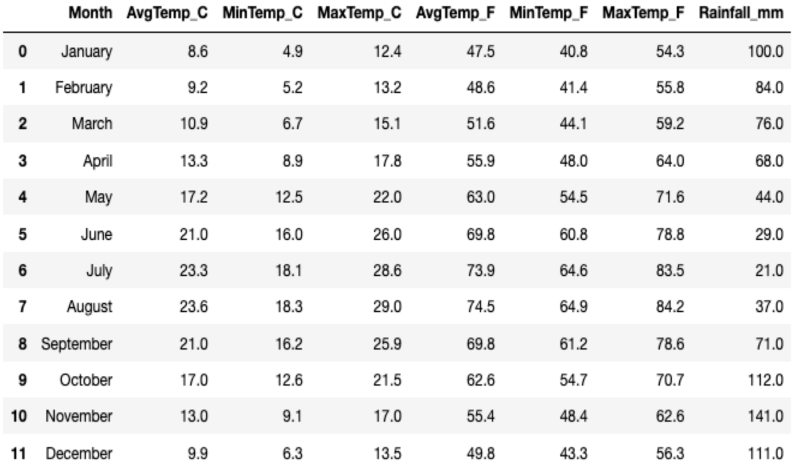


Figure 5: A Folium map for municipalities and neighborhoods.

Extract climate data

As I mentioned previously, Naples is a city with a very mild climate. Let's see it. I scraped the data from a website [5] and in figure 6 you can see a dataframe for climate data.



	Month	AvgTemp_C	MinTemp_C	MaxTemp_C	AvgTemp_F	MinTemp_F	MaxTemp_F	Rainfall_mm
0	January	8.6	4.9	12.4	47.5	40.8	54.3	100.0
1	February	9.2	5.2	13.2	48.6	41.4	55.8	84.0
2	March	10.9	6.7	15.1	51.6	44.1	59.2	76.0
3	April	13.3	8.9	17.8	55.9	48.0	64.0	68.0
4	May	17.2	12.5	22.0	63.0	54.5	71.6	44.0
5	June	21.0	16.0	26.0	69.8	60.8	78.8	29.0
6	July	23.3	18.1	28.6	73.9	64.6	83.5	21.0
7	August	23.6	18.3	29.0	74.5	64.9	84.2	37.0
8	September	21.0	16.2	25.9	69.8	61.2	78.6	71.0
9	October	17.0	12.6	21.5	62.6	54.7	70.7	112.0
10	November	13.0	9.1	17.0	55.4	48.4	62.6	141.0
11	December	9.9	6.3	13.5	49.8	43.3	56.3	111.0

Figure 6: Climate dataframe for Naples city.

Let's analyze and visualize these data. We can see the trend of temperature in time and by this way also the rainfall in time. In figure 7 I reported the trend for averae temperature in degrees centigrade with minimum and maximum variation (left) and a bar plot of the trend for rainfall in mm.

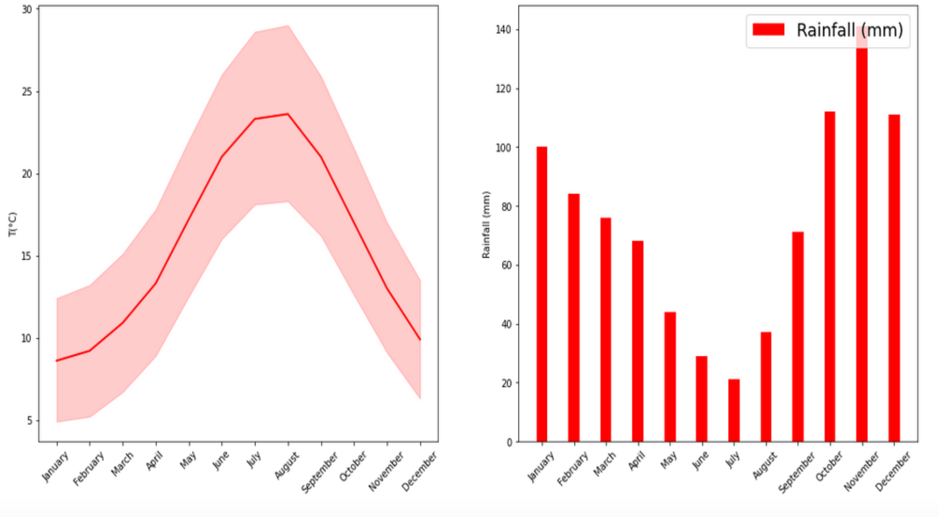


Figure 7: Trend for average temperature in degrees centigrade (left) and barplot for rainfall in mm (right).

In order to have a better idea of these data let's do a comparison with another city as New York. In figure 8 you can see this comparison red for Naples and blu for New York.

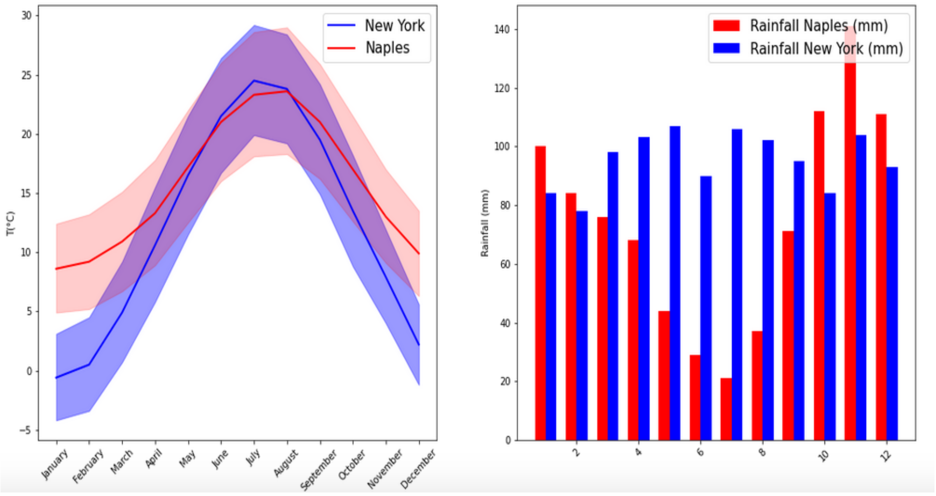
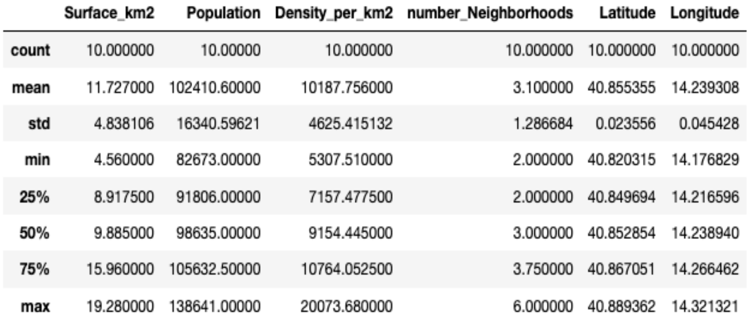


Figure 8: Trend for average temperature in degrees centigrade (left) and barplot for rainfall in mm (right).

We can see that Naples has generally an higher temperature over 10 °C and a lower rainfall than New York.

Descriptive analysis

Let's consider again the dataframe extracted for municipalities and let's do a describe of this dataframe (figure 9)



	Surface_km2	Population	Density_per_km2	number_Neighborhoods	Latitude	Longitude
count	10.000000	10.00000	10.000000	10.000000	10.000000	10.000000
mean	11.727000	102410.60000	10187.756000	3.100000	40.855355	14.239308
std	4.838106	16340.59621	4625.415132	1.286684	0.023556	0.045428
min	4.560000	82673.00000	5307.510000	2.000000	40.820315	14.176829
25%	8.917500	91806.00000	7157.477500	2.000000	40.849694	14.216596
50%	9.885000	98635.00000	9154.445000	3.000000	40.852854	14.238940
75%	15.960000	105632.50000	10764.052500	3.750000	40.867051	14.266462
max	19.280000	138641.00000	20073.680000	6.000000	40.889362	14.321321

Figure 9: A statistical description of the dataframe for municipalities.

These data show a mean density of population of ~ 10000 people per km^2, a very high number for density.

Let's do now some bar plot to deepen these data.

From the figure 10 we can note that population is concentrated primarily in municipality with lower surface (Black and red graphs up to municipality 5. The black graph grows from municipality 5 while red graph is almost constant). So in this area there is a greater density of population (green graph). There is a greater number of neighborhoods for second municipality that shows 6 neighborhoods and

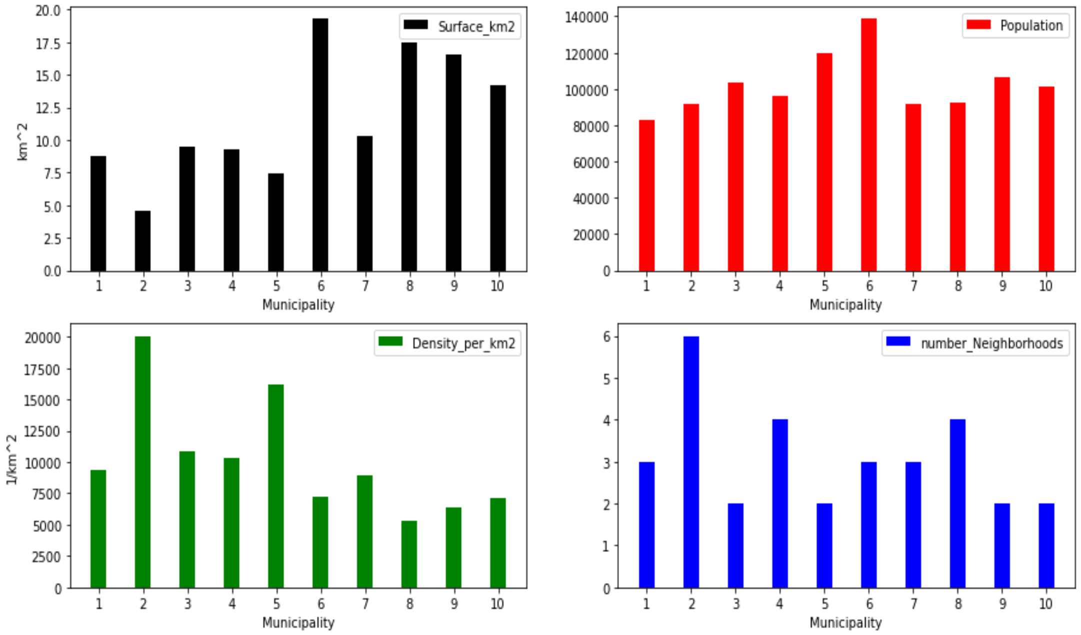


Figure 10: Bar plots for several variables of the dataframe of municipalities.

~ 20000 people per km^2. Let's do a scatter plot to better highlight these features. In figure 11 there is a net separation at 12 km^2.

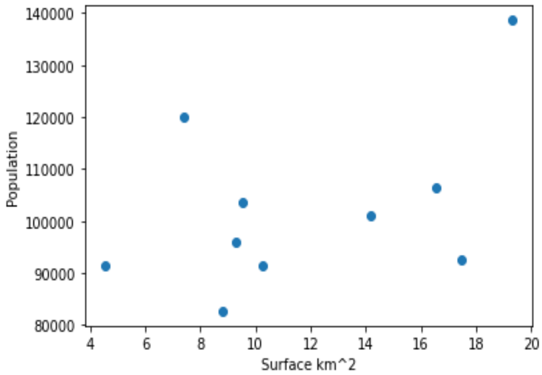


Figure 11: Scatter plot of Population Versus surface of municipalities.

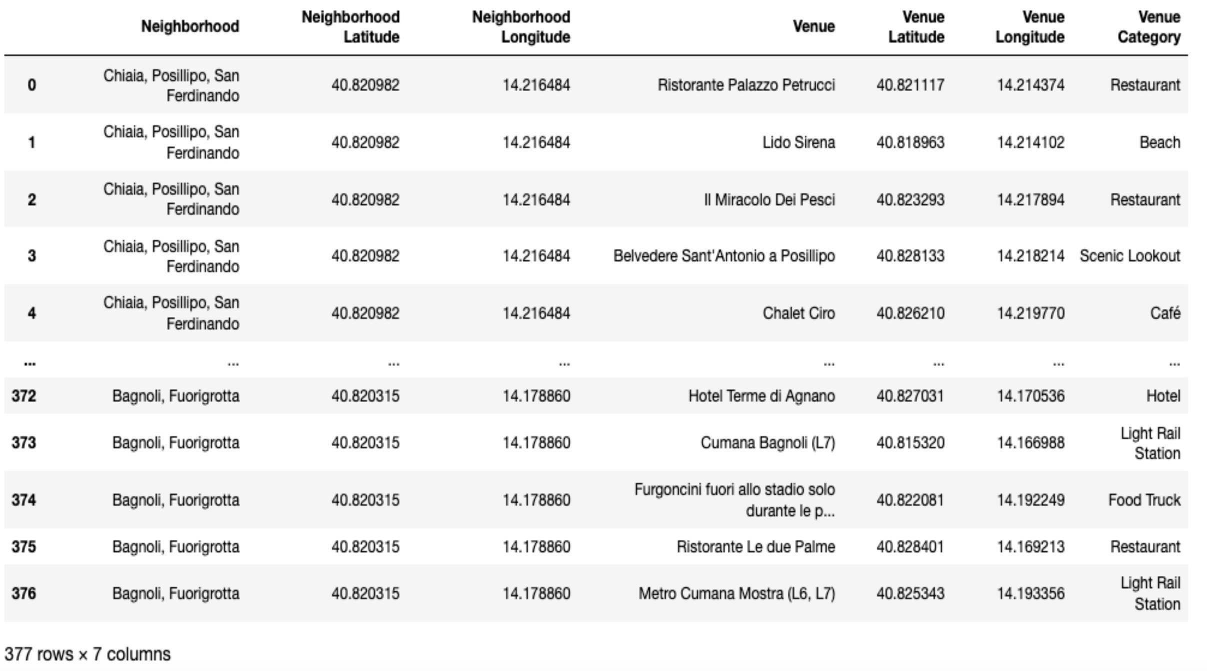
If we count the overall population at this cut value we find:

- Population in municipality with Surface lower than 12 km^2: 585358.
- Population in municipality with Surface greater than 12 km^2: 438748.

Since the Municipality II has a surface of ~ 4 km^2 and a density in population of $\sim 20 \cdot 10^3$ km^2 it could be a good candidate for an eventual investment in commerce area.

Foursquare API data collection

Using the Foursquare API I can do some requests to get some information starting from a coordinate. I can do an "explore" query to get all the venues in a certain range starting from a coordinate. Considering every coordinate extracted previously from shapefiles and doing an explore request on the API for each of it we can get all the venues for Naples city according Foursquare. In figure 12 there is the dataframe with all the venues extracted, the category and coordinates. There are 377 venues.



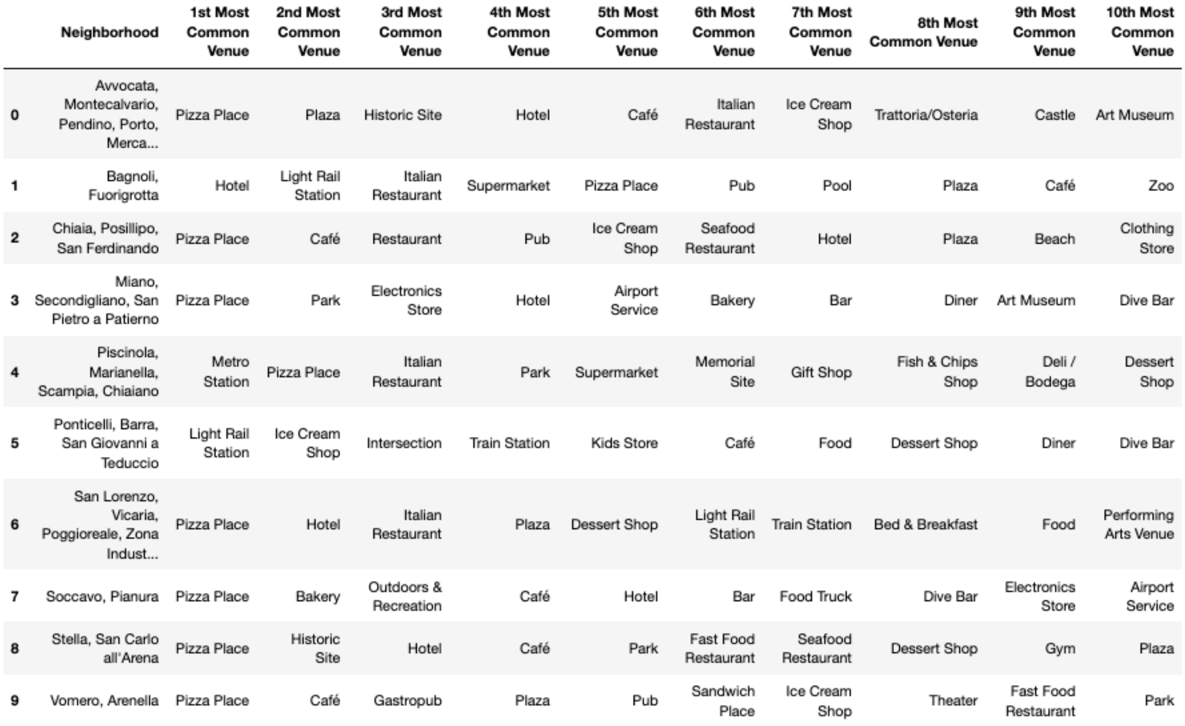
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Chiaia, Posillipo, San Ferdinando	40.820982	14.216484	Ristorante Palazzo Petrucci	40.821117	14.214374	Restaurant
1	Chiaia, Posillipo, San Ferdinando	40.820982	14.216484	Lido Sirena	40.818963	14.214102	Beach
2	Chiaia, Posillipo, San Ferdinando	40.820982	14.216484	Il Miracolo Dei Pesci	40.823293	14.217894	Restaurant
3	Chiaia, Posillipo, San Ferdinando	40.820982	14.216484	Belvedere Sant'Antonio a Posillipo	40.828133	14.218214	Scenic Lookout
4	Chiaia, Posillipo, San Ferdinando	40.820982	14.216484	Chalet Ciro	40.826210	14.219770	Café
...	...	...	...	...	...	...	...
372	Bagnoli, Fuorigrotta	40.820315	14.178860	Hotel Terme di Agnano	40.827031	14.170536	Hotel
373	Bagnoli, Fuorigrotta	40.820315	14.178860	Cumana Bagnoli (L7)	40.815320	14.166988	Light Rail Station
374	Bagnoli, Fuorigrotta	40.820315	14.178860	Furgoncini fuori allo stadio solo durante le p...	40.822081	14.192249	Food Truck
375	Bagnoli, Fuorigrotta	40.820315	14.178860	Ristorante Le due Palme	40.828401	14.169213	Restaurant
376	Bagnoli, Fuorigrotta	40.820315	14.178860	Metro Cumana Mostra (L6, L7)	40.825343	14.193356	Light Rail Station

377 rows × 7 columns

Figure 12: Dataframe for venues in Naples city according Foursquare API.

Then let's group rows by neighborhood and by taking the mean of the frequency of occurrence of each category. At the end we can display the top 10 venues for each municipality. In figure 13 is visible a dataframe with venues for all municipalities.

This dataframe concludes the data preparation and now is possible the creation of the model.



	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Avvocata, Montecalvario, Pendino, Porto, Merca...	Pizza Place	Plaza	Historic Site	Hotel	Café	Italian Restaurant	Ice Cream Shop	Trattoria/Osteria	Castle	Art Museum
1	Bagnoli, Fuorigrotta	Hotel	Light Rail Station	Italian Restaurant	Supermarket	Pizza Place	Pub	Pool	Plaza	Café	Zoo
2	Chiaia, Posillipo, San Ferdinando	Pizza Place	Café	Restaurant	Pub	Ice Cream Shop	Seafood Restaurant	Hotel	Plaza	Beach	Clothing Store
3	Miano, Secondigliano, San Pietro a Patierno	Pizza Place	Park	Electronics Store	Hotel	Airport Service	Bakery	Bar	Diner	Art Museum	Dive Bar
4	Piscinola, Marianella, Scampia, Chiaiano	Metro Station	Pizza Place	Italian Restaurant	Park	Supermarket	Memorial Site	Gift Shop	Fish & Chips Shop	Deli / Bodega	Dessert Shop
5	Ponticelli, Barra, San Giovanni a Teduccio	Light Rail Station	Ice Cream Shop	Intersection	Train Station	Kids Store	Café	Food	Dessert Shop	Diner	Dive Bar
6	San Lorenzo, Vicaria, Poggioreale, Zona Indust...	Pizza Place	Hotel	Italian Restaurant	Plaza	Dessert Shop	Light Rail Station	Train Station	Bed & Breakfast	Food	Performing Arts Venue
7	Soccavo, Pianura	Pizza Place	Bakery	Outdoors & Recreation	Café	Hotel	Bar	Food Truck	Dive Bar	Electronics Store	Airport Service
8	Stella, San Carlo all'Arena	Pizza Place	Historic Site	Hotel	Café	Park	Fast Food Restaurant	Seafood Restaurant	Dessert Shop	Gym	Plaza
9	Vomero, Arenella	Pizza Place	Café	Gastropub	Plaza	Pub	Sandwich Place	Ice Cream Shop	Theater	Fast Food Restaurant	Park

Figure 13: Dataframe for top 10 venues for every municipality according Foursquare API.

Model Clustering Neighborhoods

The data are modeled using a clustering algorithm. I used a clustering K-means alghoritm. This is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster[7].

The centroids are initialized with casual values and updated in an iterative way every time the cluster are determined up to a tolerance value. The problem with this algorithm is it is dependent from the starting points as centroids so it could give different answers every time it is applied.

I used the implementation of this algorithm in the Python library Sklearn.

Determining the best value for k

In order to find the best k, is applied the Elbow method [8] that considers the sum of distances of points from clusters varying the value of the number of clusters k.

I did it and the best value for k is 4 as you can see in figure 14

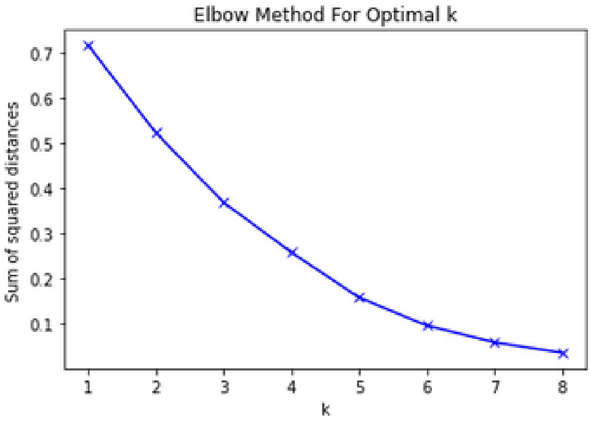


Figure 14: Best value of K for cluster K-means using Elbow method.

Evaluating the model

After having done clustering of the data we can see the clusters on the map and the features they have. Figure, 15 shows how municipalities are clusterized.

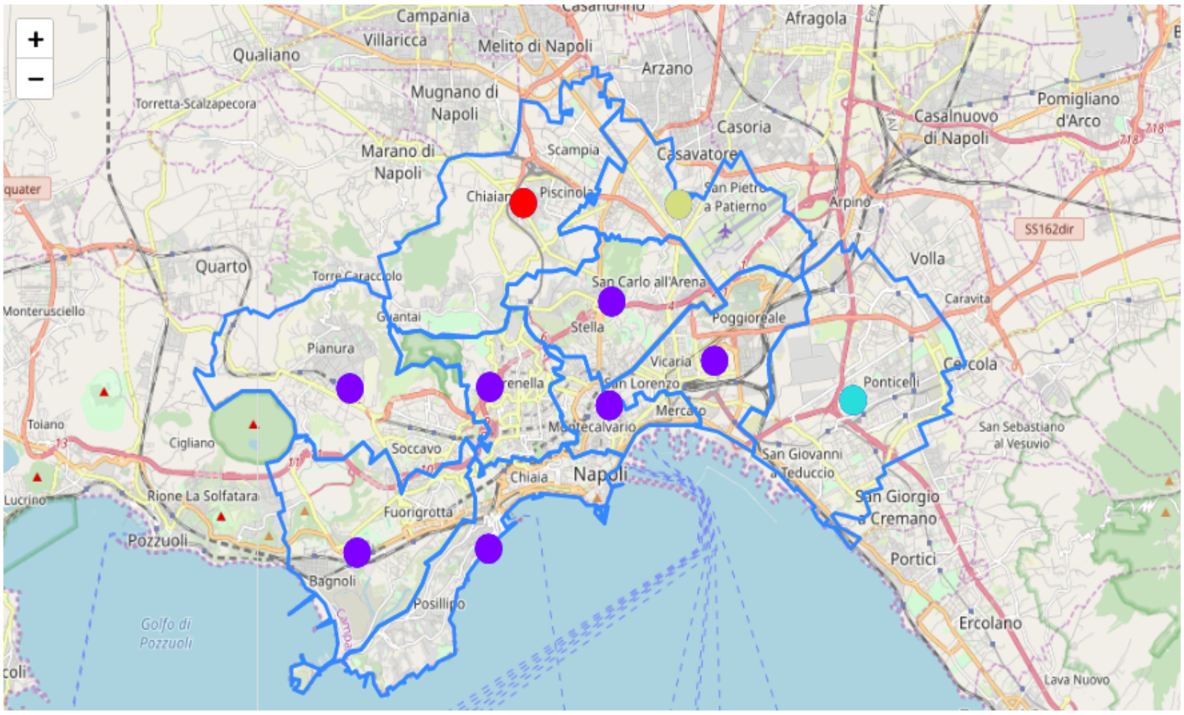


Figure 15: Map showing clusters of municipalities.

The figure shows 3 clusters on the periphery of the city characterized by train stations, parks, supermarkets, bakery.

Then there is another big cluster characterized by venues that are hotels, pizza place, coffe and cocktails shop. It is the storic center of the city with museums. It is the part of the city more alive. So it could be a good candidate for an investment. But let's analyze it with more precision in the next chapter.

Results

Let's analyze the model in this section. So I employ heatmaps and histo2d to find the "hottest" area for a good investment for an hotel, for a restaurant, for a pizza place or coffee and cocktails shop.

For each one of these activities I do a query on Foursquari API for category as "search" to get the venues for every activity.

I extract the coordinates of each venue and I do a heatmap with Folium to see where is the hottest zone for a specific category. Then I do 2d histogram, a density plot and a contour plot to get for every activity the best position for the investment.

I map the latitude and longitude in a plane (an approximation for little area). Then I get the max for the distributions of latitude and longitude and that could be a reasonable estimation for the most likely position, where the same activities are concentrated.

The figure 16 shows the scatter, histo2d, kde density, contour plots for latitude and longitude, for the category hotel.

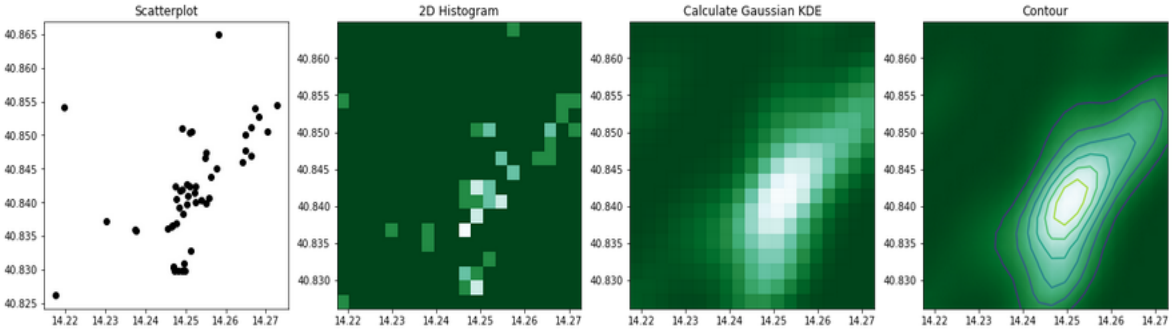


Figure 16: Scatter, histo2d, kde density, contour plots for latitude and longitude, for the category hotel.

The figure 17 shows clusters, the heat-map map and the best location (in black) for hotel category.

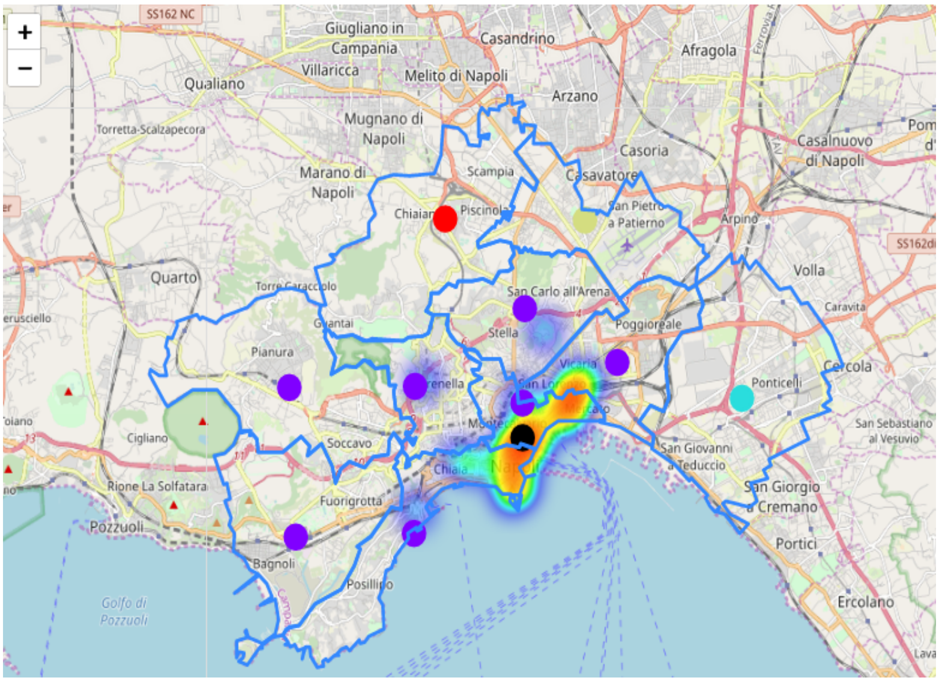


Figure 17: Clusters, the heat-map and the best location for hotel category (in black).

We can do the same analysis for other categories. I did it for the categories: hotel, pizza, restaurant, coffee, cocktails.

In figure 18 I show the best locations.

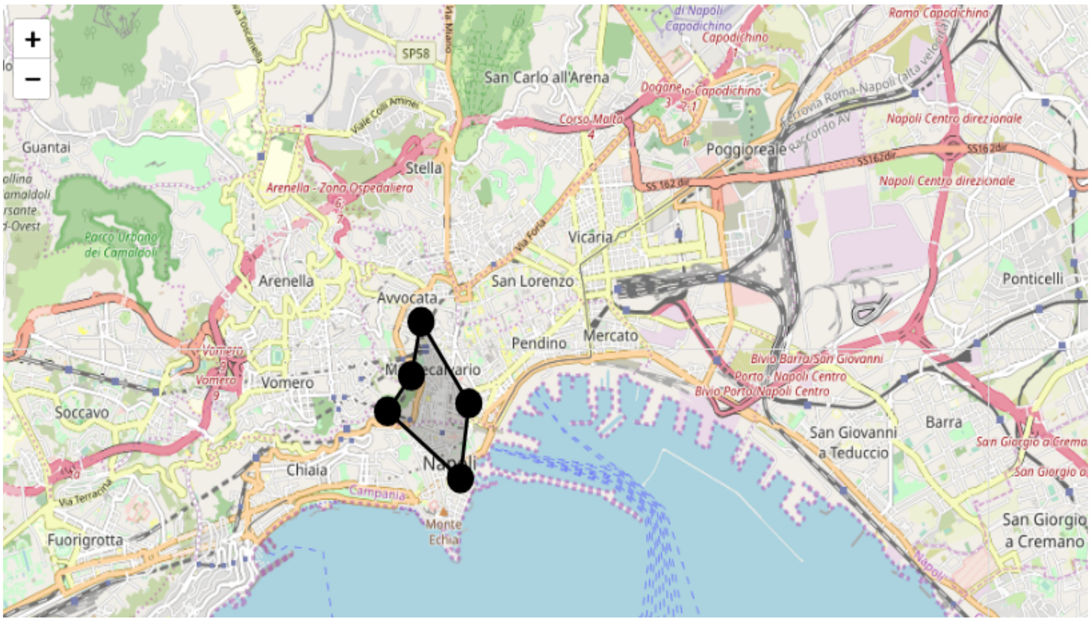


Figure 18: Best locations for the categories: hotel, pizza, restaurant, coffee, cocktails.

The points on the map could constitute a polygon for a very reasonable area for a good investment. It is located in municipality 2 which as we considered at the beginning was a good candidate for an investment. Indeed, it is the historic part of the city, full of museums, it is a short area $\sim 4\ km^2$ but with an high density of population 20000 people per km^2. It is near the sea and is an attraction for tourists.

Conclusion

In this project I analyzed the city where I was born. A beautiful city that needs to be relaunched with tourism, with investments. This is a city with a very good climate, mild. A city of sea, with very good food everywhere you see. It is a city with a very high density of population so everywhere you invest in a local it is always full of people.

The city is structured in municipalities and neighborhoods. I used open data to get information about municipality, for boundaries and centroids.

I used Foursquare API to get data about venues. I clustered municipalities with the clustering algorithm K-Means. I found 4 clusters. 3 are in the residential zone of the city, made up of train station, parks, bakery, supermarket. 1 of that is full of restaurants, hotels, pizza place, coffee and cocktails shop.

I plotted on heat-map these activities and found the best location, more reasonable, using histo 2d plot and contour plot.

I plotted all the best places on a map. These constitute a polygon. That could be a very reasonable area fora good investment.

This area is located in Municipality 2, the shortest with 4 squared kilometers but with a very very high density of population, 20000 people per squared kilometer.

The analysis can be further extended going to consider other categories as food, gym or deepen those already considered as doing a research for a better location of Italian restaurant, Chinese restaurant and so on.

Bibliography

[1] https://teleport.org/cities/naples/

[2] https://en.wikipedia.org/wiki/Naples

[3] https://it.wikipedia.org/wiki/Municipalit%C3%A0_di_Napoli

[4] http://www.comune.napoli.it/flex/cm/pages/ServeBLOB.php/L/IT/IDPagina/26531

[5] https://en.climate-data.org/europe/italy/campania/naples-4561/

[6] https://github.com/claudio−calamita/Coursera_Capstone/tree/master/dataset

[7] https://en.wikipedia.org/wiki/K−means_clustering

[8] https://blog.cambridgespark.com/how-to-determine-the-optimal-number-of-clusters-for-k-means-clustering-14f27070048f