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

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

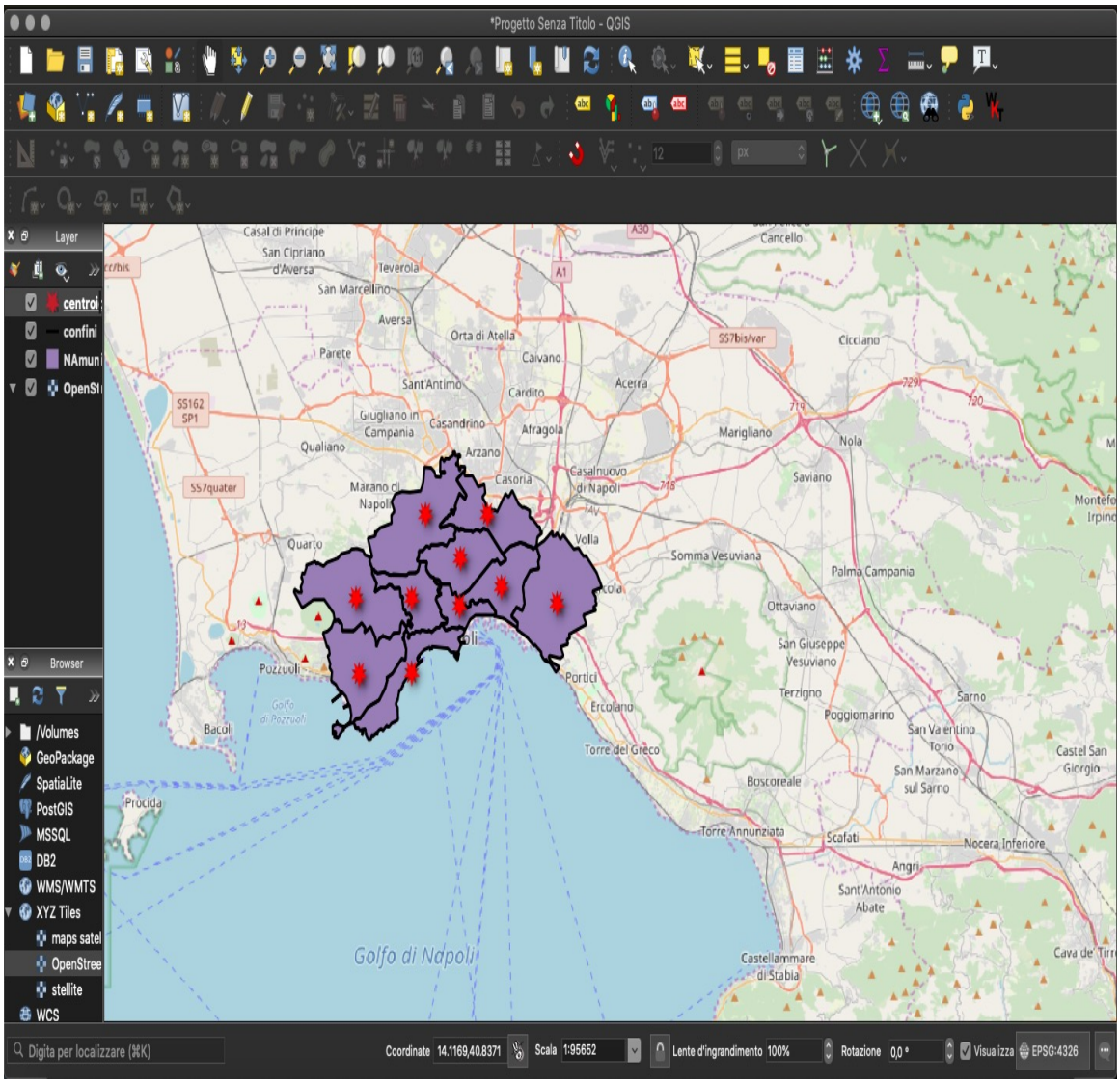

Figure 1: Example of shapefile imported in QGIS. In violet the polygons representing the municipalities. In black the boundaries. In red star the centroids.

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" indicatind 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 km²	82 673	9.553,07 ab./km²	Francesco de Giovanni di Santa Severina (Forza...	Chiaia, Posillipo, San Ferdinando	NaN
1	Municipalità II	4,56 km²	91 536	20.073,68 ab./km²	Francesco Chirico	Avvocata, Montecalvario, Pendino, Porto, Merca...	NaN
2	Municipalità III	9,51 km²	103 633	10.897,27 ab./km²	Ivo Poggiani (Lista DemA)	Stella, San Carlo all'Arena	NaN
3	Municipalità IV	9,27 km²	96 078	10.364,4 ab./km²	Giampiero Perrella	San Lorenzo, Vicaria, Poggioreale, Zona Indust...	NaN
4	Municipalità V	7,42 km²	119 978	16 169,54 ab./km²	Paolo De Luca	Vomero, Arenella	NaN
5	Municipalità VI	19,28 km²	138 641	7 190,92 ab./km²	Salvatore Boggia	Ponticelli, Barra, San Giovanni a Teduccio	NaN
6	Municipalità VII	10,26 km²	91 460	8 914,23 ab./km²	Maurizio Moschetti	Miano, Secondigliano, San Pietro a Patierno	NaN
7	Municipalità VIII	17,45 km²	92 616	5 307,51 ab./km²	Paipais Apostolos	Piscinola, Marianella, Scampia, Chiaiano	NaN
8	Municipalità IX	16,56 km²	106 299	6 419,02 ab./km²	Lorenzo Giannalavigna (PD)	Soccavo, Pianura	NaN
9	Municipalità X	14,16 km²	101 192	7 416,38 ab./km²	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.

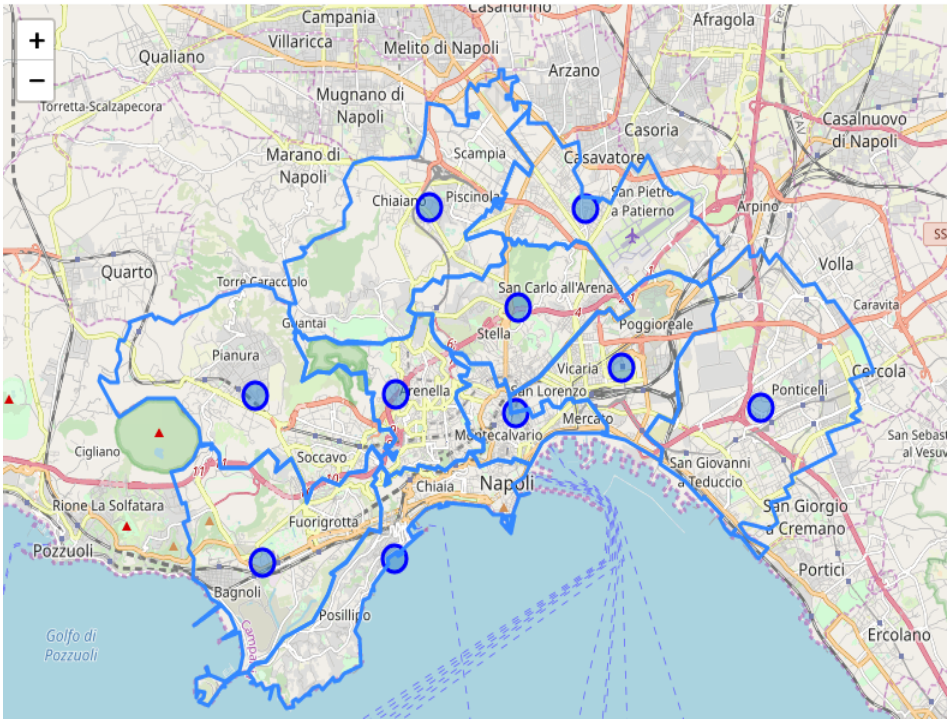

Figure 5: A Folium map for municipalities and neighborhoods.

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.

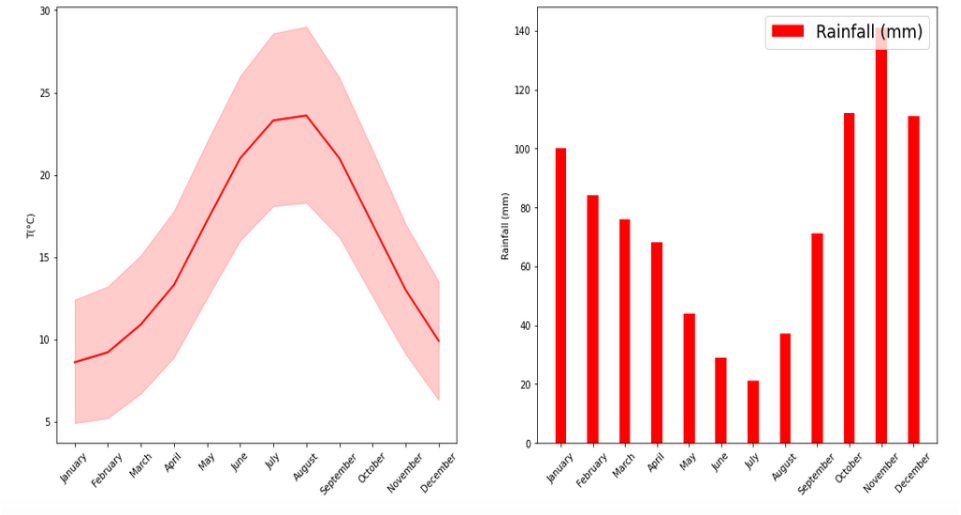

Figure 7: Trend for average temperature in degrees centigrade (left) and barplot for rainfall in mm (right).

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.

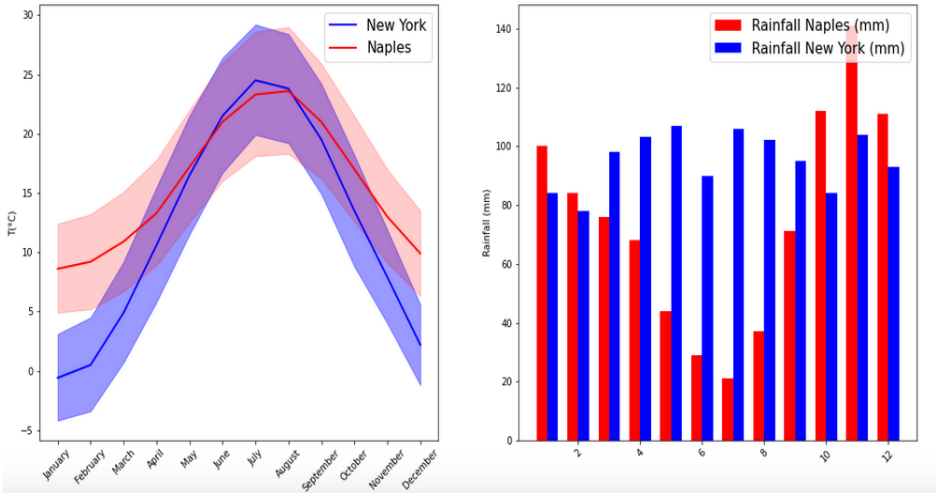

Figure 8: Trend for average temperature in degrees centigrade (left) and barplot for rainfall in mm (right).

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 $\sim$ 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

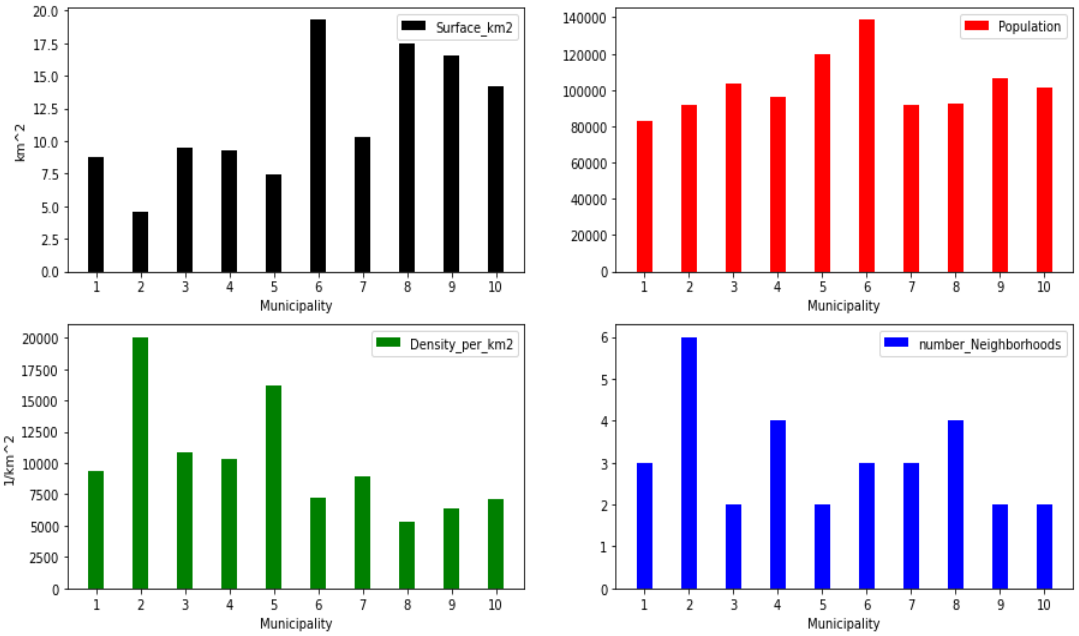

Figure 10: Bar plots for several variables of the dataframe of municipalities.

$\sim$ 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.

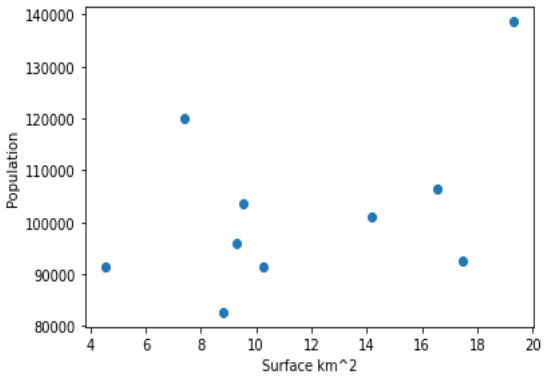

Figure 11: Scatter plot of Population Versus surface of municipalities.

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.

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