Report
That Italian Place
From A to Z about venues in Milan, Italy

Fig. 1 Milan Cathedral from Piazza del Duomo

Iulian Corcoja

September 2019

Table of Contents

That Italian Place

Introduction

About Milan

Milan, the capital of Lombardy region in Italy, has a population of 1.3 million people. It is the biggest industrial city of Italy with many different industrial sectors. It is a magnetic point for designers, artists, photographers and models. Milan has an ancient city centre with high and interesting buildings and palazzos – that's why so many people from all over the world want to see the city of glamour.

Fig. 2 Castello Sforzesco in Milan (source – Castello Sforzesco)

Milan is famous for restaurants recommended by design-world insiders, chic bars where you can sample the classic Milanese *aperitivo*, top designer boutiques and typical Italian local shops that would fulfill the needs for any kind of tourist.

Even with such a great variety of shops and restaurants, no trip to this historic city would be complete without visiting at least a few of its magnificent landmarks, since there's a lot more to see than the iconic cathedral of Duomo.
With so many places to visit, it is often hard to decide what would be the next destination point or what venues to look for when searching for something specific.

Problem Description

Milan has a vast array of activities to offer that one can do, be it a tourist or a local who wants to explore new places or food:

- City sightseeing
- Outdoor activities
- Shopping and markets
- Milan nightlife and bars
- Taste Italian cuisine and fine-dinning
- Cafes and local barista specialists
- Visit museums and galleries
- Tours and trips
- Milan's architecture and design

Having so many options in possibilities and things to do, certainly is a good feature, but it might sometimes confuse people on what would be the best place or area for a particular activity.
The purpose of this project is to analyze the most common places in Milan and bundle them into distinct categories based on a list of properties. Every category would fit a single or a series of venue types that are similar between each other, offering the option to choose for a venue by the kind and location.

Target Audience

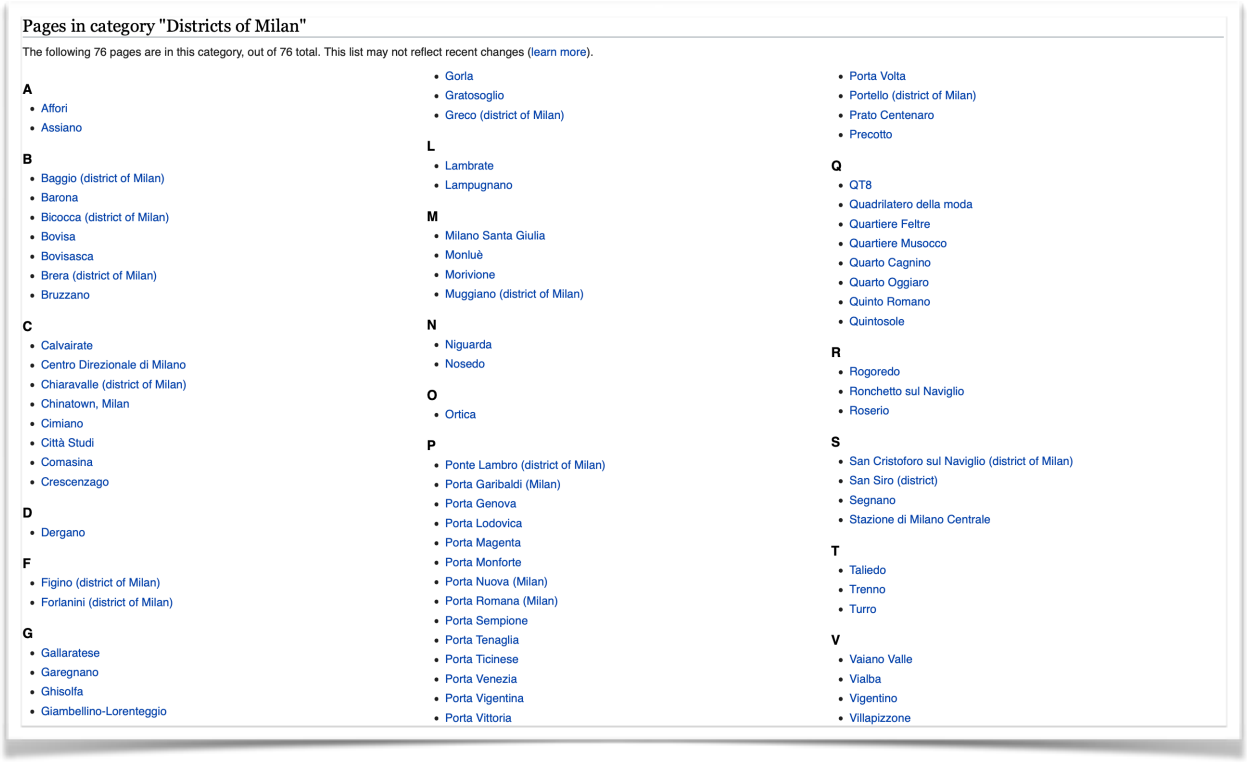
The following analysis would be very helpful for both tourists that are visiting Milan, whether it is for their first time or not, as well as for people that are living in the city but want to discover new places like: museums, parks, city specific attractions, cafes, restaurants, bar, etc. With this project, they can select the type of

activity they wish to do and choose the one of their liking that is the closest to them.

Data

Source of the Data

The analysis will be implemented for all the districts in the city of Milan, the source of this being the Wikipedia web-page that contains this information, grouped alphabetically:
https://en.wikipedia.org/wiki/Category:Districts_of_Milan



Pages in category "Districts of Milan"

The following 76 pages are in this category, out of 76 total. This list may not reflect recent changes (learn more).

A
- Affori
- Assiano

B
- Baggio (district of Milan)
- Barona
- Bicocca (district of Milan)
- Bovisa
- Bovisasca
- Brera (district of Milan)
- Bruzzano

C
- Calvairate
- Centro Direzionale di Milano
- Chiaravalle (district of Milan)
- Chinatown, Milan
- Cimiano
- Città Studi
- Comasina
- Crescenzago

D
- Dergano

F
- Figino (district of Milan)
- Forlanini (district of Milan)

G
- Gallaratese
- Garegnano
- Ghisolfa
- Giambellino-Lorenteggio
- Gorla
- Gratosoglio
- Greco (district of Milan)

L
- Lambrate
- Lampugnano

M
- Milano Santa Giulia
- Monluè
- Morivione
- Muggiano (district of Milan)

N
- Niguarda
- Nosedo

O
- Ortica

P
- Ponte Lambro (district of Milan)
- Porta Garibaldi (Milan)
- Porta Genova
- Porta Lodovica
- Porta Magenta
- Porta Monforte
- Porta Nuova (Milan)
- Porta Romana (Milan)
- Porta Sempione
- Porta Tenaglia
- Porta Ticinese
- Porta Venezia
- Porta Vigentina
- Porta Vittoria
- Porta Volta
- Portello (district of Milan)
- Prato Centenaro
- Precotto

Q
- QT8
- Quadrilatero della moda
- Quartiere Feltre
- Quartiere Musocco
- Quarto Cagnino
- Quarto Oggiaro
- Quinto Romano
- Quintosole

R
- Rogoredo
- Ronchetto sul Naviglio
- Roserio

S
- San Cristoforo sul Naviglio (district of Milan)
- San Siro (district)
- Segnano
- Stazione di Milano Centrale

T
- Taliedo
- Trenno
- Turro

V
- Vaiano Valle
- Vialba
- Vigentino
- Villapizzone

Fig. 3 Districts of Milan from Wikipedia page (source – Wikipedia)

There is a total of 76 districts in Milan, but for the project scope only 75 are used, because the geocoding capable framework wasn't returning acceptable results for some of the districts (more details in the "Nota Bene – Geocoding" section).

Additionally, three other frameworks are integrated into the project for the subsequent purposes:

- Python web scraping library.
  - Beautiful Soup: https://www.crummy.com/software/BeautifulSoup/bs4/doc/
- District geocoding (transforming district names to geo-coordinates).
  - HERE: https://developer.here.com
- Data platform for exploring venues in Milan.
  - Foursquare: https://foursquare.com

Web Scraping

Milan districts are scraped from the Wikipedia web-page with the help of Beautiful Soup python package. The districts are grouped alphabetically, so the scraping and processing of the data is split into several stages:

1. Scrape the alphabetical groups from the web-page.
2. From every group, extract the district names.
3. Union the groups with the district names into a single list with all the districts.

```
['Affori',
 'Assiano',
 'Baggio (district of Milan)',
 'Barona',
 'Bicocca (district of Milan)',
 'Bovisa',
 'Bovisasca',
 'Brera (district of Milan)',
 'Bruzzano',
 'Calvairate',
 'Centro Direzionale di Milano',
 'Chiaravalle (district of Milan)',
 'Chinatown, Milan',
 'Cimiano',
 'Città Studi',
 'Comasina',
 'Crescenzago',
 'Dergano',
 'Figino (district of Milan)',
 'Forlanini (district of Milan)',
 'Gallaratese',
 'Garegnano',
 'Ghisolfa',
 'Giambellino-Lorenteggio',
 'Gorla',
 'Gratosoglio',
 'Greco (district of Milan)',
 'Lambrate',
 'Lampugnano',
 'Milano Santa Giulia',
 'Monluè',
 'Morivione',
 'Muggiano (district of Milan)',
 'Niguarda',
 'Nosedo',
 'Ortica',
 'Ponte Lambro (district of Milan)',
 'Porta Garibaldi (Milan)',
 'Porta Genova',
 'Porta Lodovica',
 'Porta Magenta',
 'Porta Monforte',
 'Porta Nuova (Milan)',
 'Porta Romana (Milan)',
 'Porta Sempione',
 'Porta Tenaglia',
 'Porta Ticinese',
 'Porta Venezia',
 'Porta Vigentina',
 'Porta Vittoria',
 'Porta Volta',
 'Portello (district of Milan)',
 'Prato Centenaro',
 'Precotto',
 'QT8',
 'Quadrilatero della moda',
 'Quartiere Feltre',
 'Quartiere Musocco',
 'Quarto Cagnino',
 'Quarto Oggiaro',
 'Quinto Romano',
 'Quintosole',
 'Rogoredo',
 'Ronchetto sul Naviglio',
 'Roserio',
 'San Cristoforo sul Naviglio (district of Milan)',
 'San Siro (district)',
 'Segnano',
 'Stazione di Milano Centrale',
 'Taliedo',
 'Trenno',
 'Turro',
 'Vaiano Valle',
 'Vialba',
 'Vigentino',
 'Villapizzone']
```

Fig. 4 Districts of Milan after web scraping

Gecoding

Geocoding is one of the most important stages in the project (as well as the earliest one after the web scraping stage), since it is required to understand the exact location of every district in Milan. The reason HERE API was selected for this task is, that alongside other platforms like Google Maps or TomTom, HERE is one of the internet's preeminent online mapping platforms, known by a number of names through the years.

Compared to Google Maps geocoding APIs, which doesn't offer a free plan, with HERE one can garner up to 250,000 transactions each month at no cost. Going over the 250,000 calls limit, the additional cost is $1 for every 1,000 transactions, which is still fairly cheap compared to what other platforms currently offer.

Moreover, with HERE platform, the results were the most consistent in comparison to other freemium platforms.

The geocoding stage collects the below data, that further is used to explore and analyze the venues in Milan.

- Name of the district.
- Location latitude.
- Location longitude.
- District radius.

By design, HERE doesn't offer radius information inside the search query results, though it does offer the area's bounding box.

The radius is calculated from the bounding box as follows:

$$r = \frac{d(L_c, L_{tl}) + d(L_c, L_{br})}{2} / \sqrt{\pi}$$

where:

- $d(a,\ b)$ – geodesic distance between two geo-coordinates.
- L_c – location center (geo-coordinates).
- L_{tl} – location bounding box top-left geo-coordinates.
- L_{br} – location bounding box bottom-right geo-coordinates.

The final results of the Milan's district geocoding looks like this:

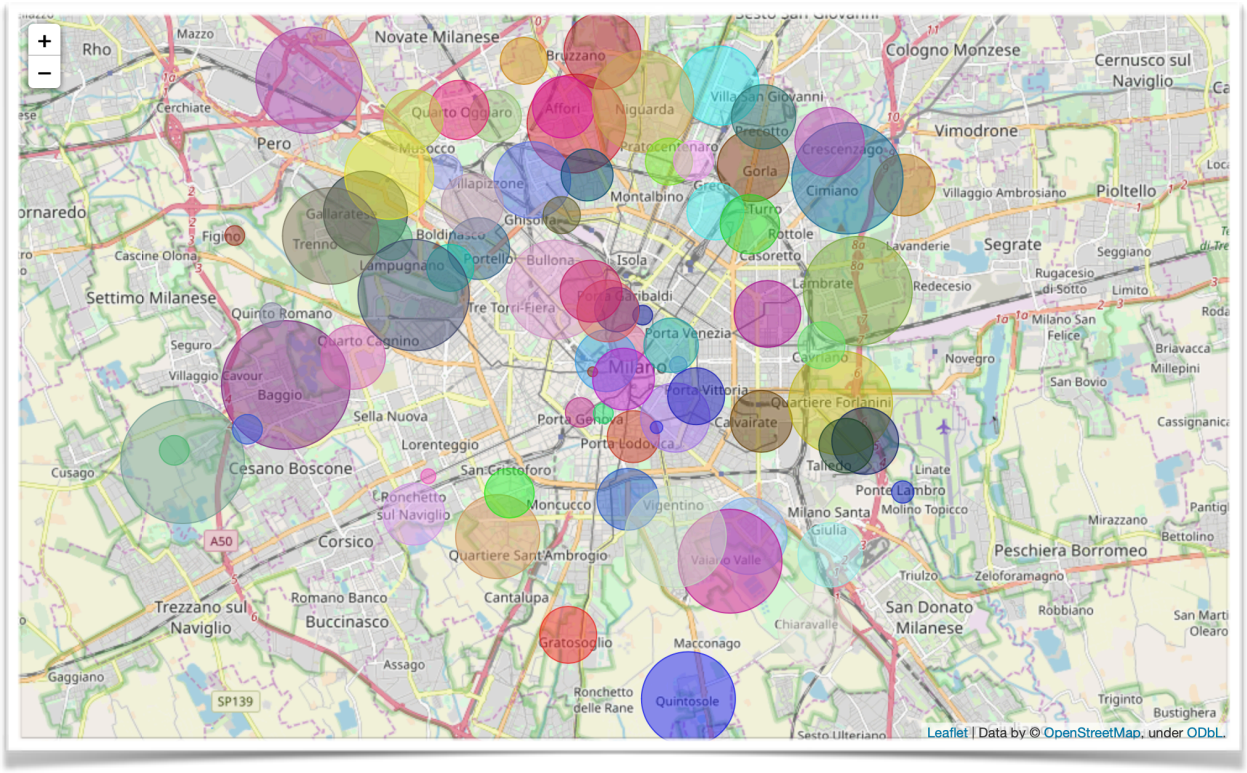


Fig. 5 Districts of Milan (geocoding – HERE framework)

	District Name	District Latitude	District Longitude	Venue Name	Venue Category	Venue Latitude	Venue Longitude
2020	Porta Vigentina, Milan, Italy	45.45500	9.19577	Viva	Salad Place	45.456250	9.195423
437	Chinatown, Milan, Italy	45.46796	9.18178	Six Inch Hair & Spa	Salon / Barbershop	45.470829	9.185033
1826	Porta Ticinese, Milan, Italy	45.45771	9.18096	Scout	Clothing Store	45.459396	9.180609
1766	Porta Tenaglia, Milan, Italy	45.47766	9.18224	Gam	Korean Restaurant	45.481481	9.184667
1885	Porta Ticinese, Milan, Italy	45.45771	9.18096	Tokuyoshi	Italian Restaurant	45.458383	9.176125
2751	Vigentino, Milan, Italy	45.43373	9.20105	Esselunga	Supermarket	45.443121	9.197132
1277	Porta Magenta, Milan, Italy	45.46582	9.17800	Gay Odin	Chocolate Shop	45.466187	9.180801
798	Greco, Milan, Italy	45.49702	9.21213	Alvin's Bar Pasticceria	Café	45.494939	9.207228
2489	San Siro, Milan, Italy	45.48075	9.12824	Tennis Lido	Tennis Court	45.480909	9.142764
614	Dergano, Milan, Italy	45.50412	9.17647	Ac Garibaldina 1932	Soccer Field	45.505813	9.172451
2767	Vigentino, Milan, Italy	45.43373	9.20105	Bar Chopin	Coffee Shop	45.425268	9.204760
2374	Rogoredo, Milan, Italy	45.43017	9.24401	Igiban	Japanese Restaurant	45.426923	9.250569
737	Gorla, Milan, Italy	45.50591	9.22265	Seven - La Casa dei Ciliegi	Italian Restaurant	45.503713	9.222976
427	Chinatown, Milan, Italy	45.46796	9.18178	L.O.V.E. Maurizio Cattelan	Monument / Landmark	45.464834	9.183314
2765	Vigentino, Milan, Italy	45.43373	9.20105	Barrio Alto	Gastropub	45.441364	9.201348
2100	Porta Vittoria, Milan, Italy	45.46106	9.20677	Piadiniamo	Piadineria	45.456505	9.205159
2638	Trenno, Milan, Italy	45.49225	9.10518	McCafé	Fast Food Restaurant	45.496345	9.113852
732	Giambellino-Lorenteggio, Milan, Italy	45.44548	9.13233	Fermata ATM Piazza Tirana / San Cristoforo [49]	Bus Stop	45.444730	9.130393
1615	Porta Romana, Milan, Italy	45.45691	9.20096	Osteria Delizie del Mare	Seafood Restaurant	45.452097	9.203307
1505	Porta Nuova, Milan, Italy	45.47685	9.19192	Panini Durini	Sandwich Place	45.475532	9.194549

Fig. 6 Table with geocoded districts of Milan (geocoding – HERE framework)

Venues Extraction

The next step after geocoding is the venue extraction for every district. Milan district's geographical coordinates data will be utilized as input for the Foursquare API, that will be leveraged to provision venues information for each district. Using the coordinates for every district, Foursquare API calls are made to get top venues in the radius of the district (computed as explained in the **geocoding** section). An example of all the venues extracted for a single district is shown below:

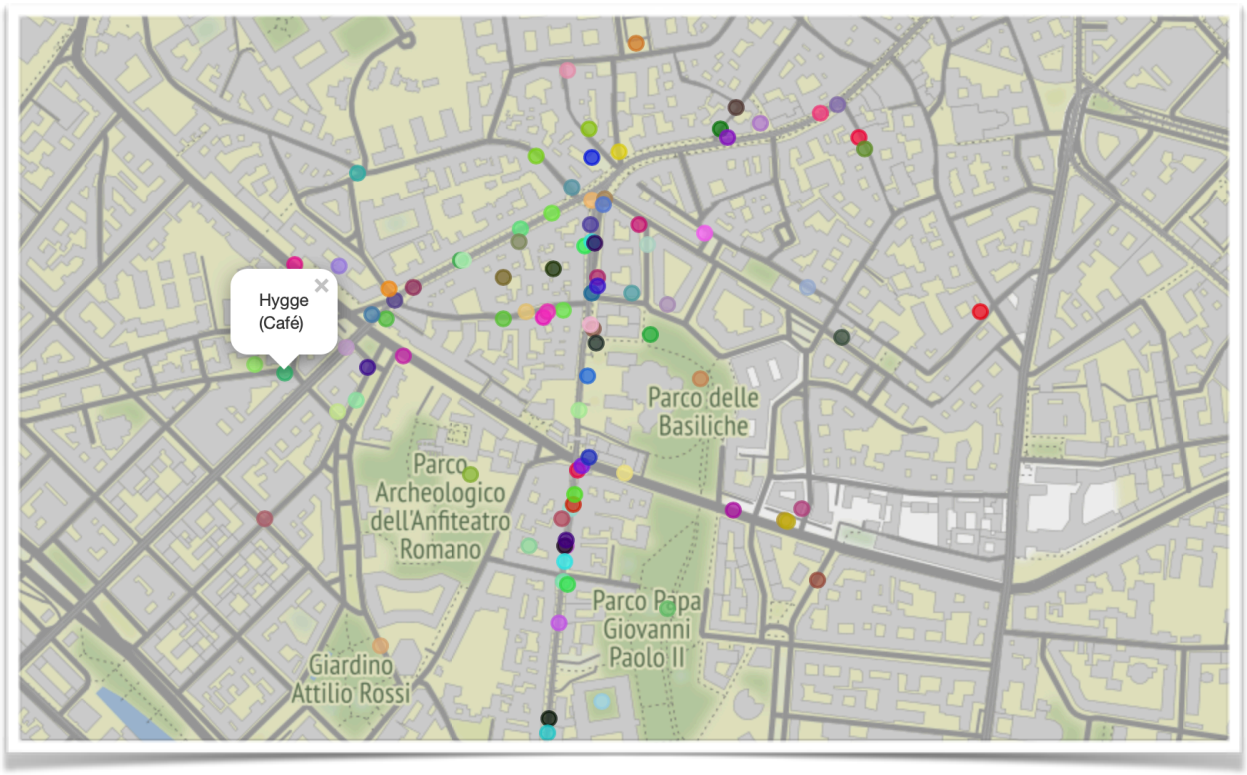


Fig. 7 Venues of Porta Ticinese district in Milan, Italy (retrieved via Foursquare API)

Nota Bene – Geocoding

In Milan, there is a region, named "Quadrilatero della moda", that HERE geocoding framework wasn't able to properly geocoded, causing issues with data processing. To prevent processing issues, this region is disregarded and removed from the Wikipedia's web-scraped list.

Methodology

Analytic Approach

The project scope implies clustering the main venues from Milan's districts using Machine Learning algorithms. Venue category type serves as the main feature when running the clustering algorithm. The custom clustering algorithm is composed of several stages, as explained in the exploratory data analysis and the algorithm sections.

Exploratory Data Analysis, Algorithm

1. Retrieve the data from the source.
2. Apply data processing and cleaning techniques on the retrieved data.
3. Extract venue categories and apply one-hot encoding.
4. Run the K-means algorithm for multiple K-values.
   - For this project, K-values where chosen to be the values from **1** to **10**.
5. Chose the best K-value using **Elbow** method.
6. Run the K-means algorithm with the K-value from the **Elbow** method.

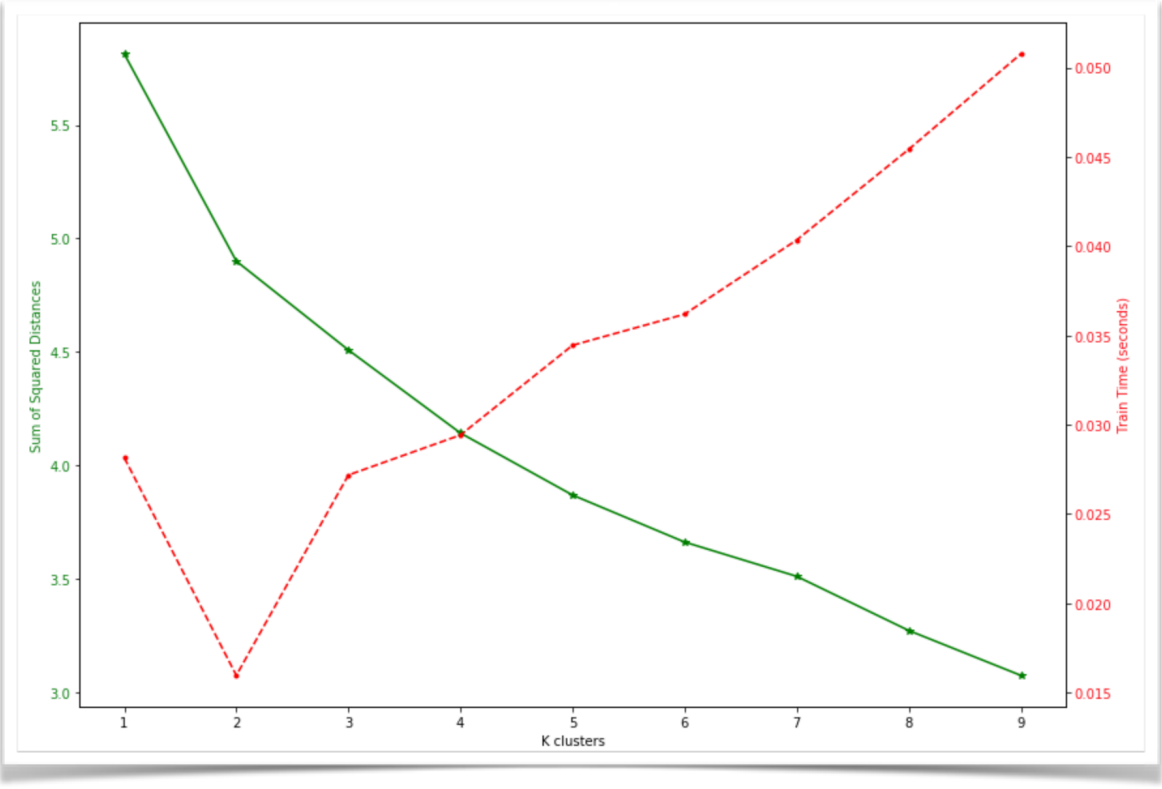


Fig. 8 Elbow method results when running the data from Milan's venues.

The clusters capture all the venue categories that are similar between each other, based on the type, frequency and quantity per district. The final results are presented in **fig. 9**.

Cluster Label	Accessories Store	Adult Education Center	African Restaurant	Airport	Airport Terminal	American Restaurant	Amphitheater	Argentinian Restaurant	Art Gallery	...	Video Game Store	Vietnamese Restaurant	Volleyball Court	Watch Shop
0	0.001311	0.002621	0.001311	0.001311	0.002621	0.000000	0.001311	0.000000	0.005242	...	0.001311	0.001311	0.003932	0.000000
1	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000
2	0.003405	0.000000	0.000973	0.000000	0.000000	0.002432	0.000000	0.002432	0.008268	...	0.000000	0.000000	0.000000	0.000486
3	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000
4	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000

Fig. 9 Clustering results after running K-means algorithm with a K-value chosen by Elbow method.

Results

Success Criteria

The success criteria of this project is to generate a good user recommendation of a district in such a manner, so their activities/places of choice are frequently met in the chosen area of Milan. Vice-versa, suggest to the user the most frequent activities/places for a specific district.

Milan Venues

Using Milan's districts and venues (mapped on every district), the venues were clustered using K-means algorithm into groups based on the mean occurrence of the venue category in each district. Furthermore, the clustered data has been utilized to create the most frequent venue types per each district, represented by a table (see **fig. 10**).

Another interesting aspect is to analyze all the venues throughout the entire city, so a general overview of the most frequent venue categories in Milan can be observed (see **fig. 11**).

	District Name	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	11th Most Common Venue
0	Affori, Milan, Italy	Hotel	Pizza Place	Park	Soccer Field	Italian Restaurant	Café	Skate Park	Shipping Store	Salon / Barbershop	Rock Club	Cocktail Bar
1	Assiano, Milan, Italy	Pizza Place	Yoga Studio	Filipino Restaurant	Event Space	Fabric Shop	Falafel Restaurant	Farm	Farmers Market	Fast Food Restaurant	Film Studio	Dumpling Restaurant
2	Baggio, Milan, Italy	Supermarket	Café	Plaza	Italian Restaurant	Bus Station	Campground	Park	Japanese Restaurant	Bar	Volleyball Court	Pizza Place
3	Barona, Milan, Italy	Soccer Field	Café	Japanese Restaurant	Theater	Bakery	Athletics & Sports	Trattoria/Osteria	Brewery	Tennis Stadium	Pub	Event Space
4	Bicocca, Milan, Italy	Café	Italian Restaurant	Sandwich Place	Steakhouse	Sushi Restaurant	Pizza Place	Hotel	Multiplex	Restaurant	Plaza	Supermarket
5	Bovisa, Milan, Italy	Italian Restaurant	Café	Pizza Place	Snack Place	Piadineria	Plaza	Ice Cream Shop	Smoke Shop	Sicilian Restaurant	Kebab Restaurant	Steakhouse
6	Bovisasca, Milan, Italy	Clothing Store	Shoe Store	Soccer Field	Restaurant	Café	Cosmetics Shop	Pizza Place	Supermarket	Park	Health & Beauty Service	Chinese Restaurant
7	Brera, Milan, Italy	Italian Restaurant	Hotel	Boutique	Ice Cream Shop	Wine Bar	Plaza	Restaurant	Café	Theater	Pizza Place	Arts & Crafts Store
8	Bruzzano, Milan, Italy	Italian Restaurant	Bakery	Pizza Place	Football Stadium	Gift Shop	Diner	Train Station	Gym / Fitness Center	Bus Station	Ice Cream Shop	Fish & Chips Shop
9	Calvairate, Milan, Italy	Italian Restaurant	Park	Pizza Place	Supermarket	Hotel	Ice Cream Shop	Bakery	Market	Other Nightlife	Tanning Salon	Bar
10	Centro Direzionale di Milano, Milan, Italy	Hotel	Italian Restaurant	Outdoors & Recreation	Lake	Rest Area	Café	Farmers Market	Fabric Shop	Falafel Restaurant	Farm	Fast Food Restaurant
11	Chiaravalle, Milan, Italy	Convenience Store	General Entertainment	Italian Restaurant	Restaurant	Yoga Studio	Farmers Market	Event Space	Fabric Shop	Falafel Restaurant	Farm	Filipino Restaurant

Fig. 10 Most frequent venue type per each district.

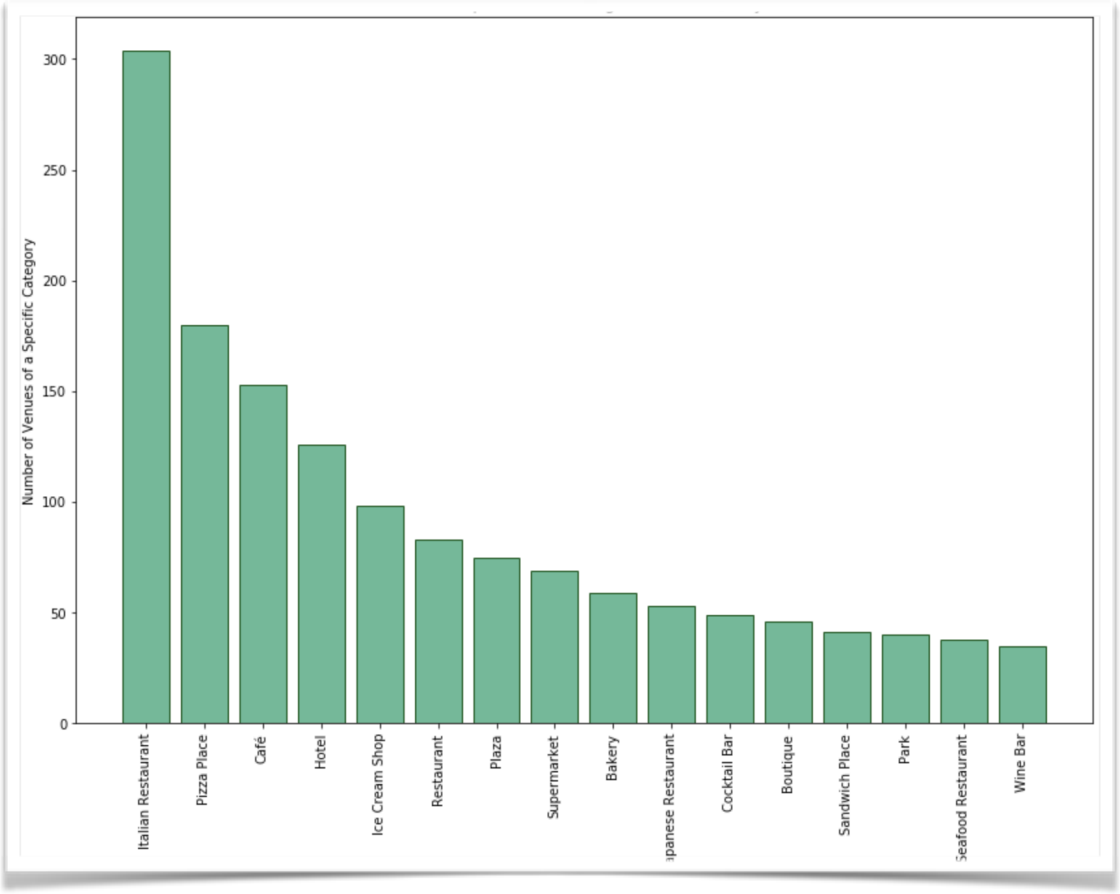


Fig. 11 Venue frequency in Milan based on the category type.

Bonus

Jupyter Notebook is a great tool that offers various GUI frameworks available to the developer. One example of such a framework is Jupyter Widgets, also named ipywidgets, that has of a collection of controls and UI elements that allows to create interactive components directly in the notebook.
Using a dropdown control, the notebook offers the user to select a single district in Milan. According to the selection, the 12th most frequent venue category types for that specific district will be presented to the user (see **fig. 12** and **fig. 13**).

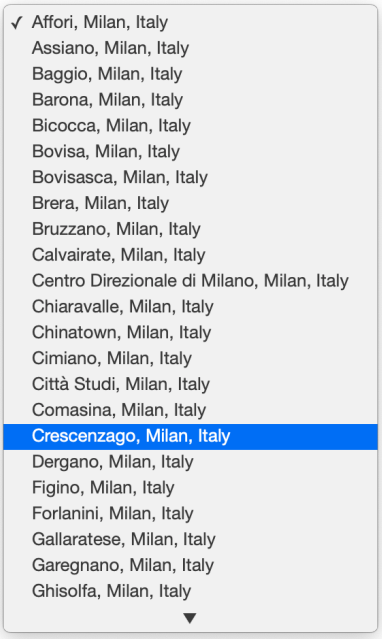


Fig. 12 A dropdown list with all the districts in Milan

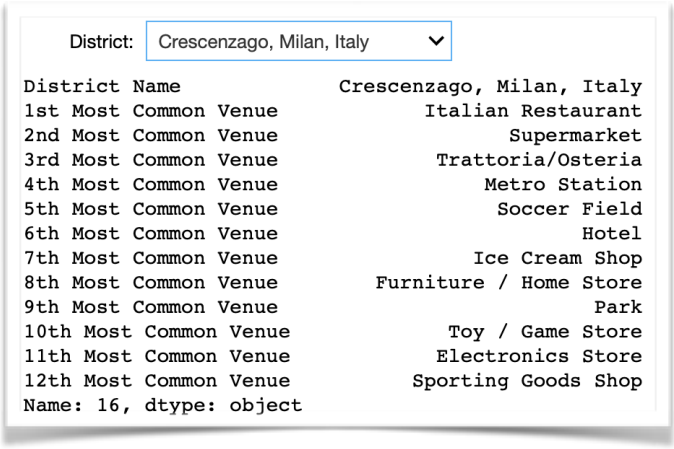


Fig. 13 Top most common venues for Crescenzago district in Milan, Italy