Data-driven business site research

COURSERA CAPSTONE PROJECT

RAFFAELE NOLLI

The concept

- Data-driven tool to characterise neighbourhoods of a city, able to provide information relevant to business;
- Characterisation of neighbourhood based on venue composition (restaurants, services, shops…);
- Characterisation of neighbourhoods based on food venues average price tier and ratings;

- provide information on neighbourhoods (residential, commercial, high-end, working class…)
- Identify areas suited for business investment and expansion based on similarity between neighbourhoods;

The data

- General information about Milan's neighbourhoods, and table of names:

https://en.wikipedia.org/wiki/Zones_of_Milan

- Institutional database with all addresses in Milan and their spatial coordinates:

ds634_civici_coordinategeografiche_20190902_csv.zip

- Information on neighbourhood's venue composition gathered through **Foursquare API;**
- Information on neighbourhood's average venue price tier and rating gathered through **Yelp API;**

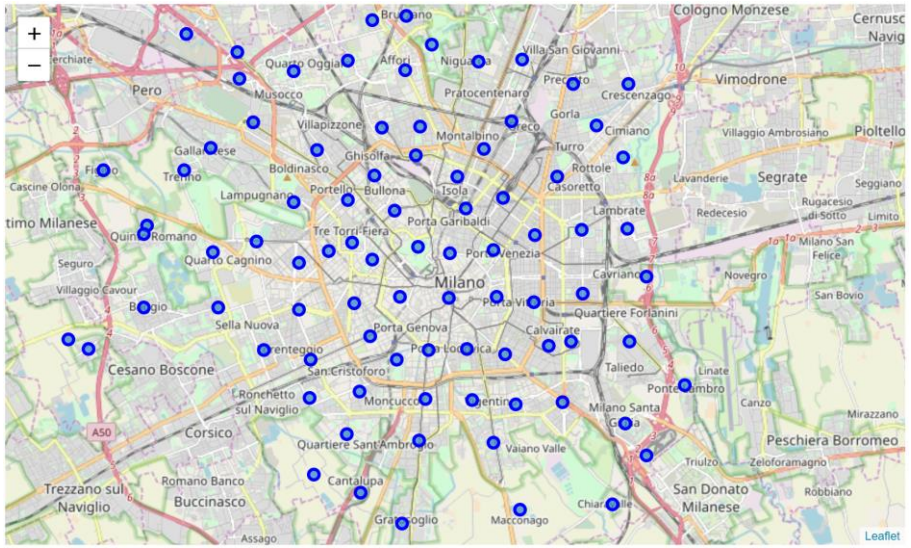
Data and features preparation

Data cleaning and extraction from institutional database.

Averaging of neighbourhood addresses coordinates to create a list of neighbourhoods.

Use of *folium* package to create an annotated map.

	Borough	Latitude	Longitude
0	DUOMO	45.463216	9.187042
1	BRERA	45.473397	9.187424
2	GUASTALLA	45.463486	9.202288
3	Giardini Pta Venezia	45.474080	9.201534
4	VIGENTINA	45.451611	9.192733



Data and features preparation: venue composition

Use of Foursquare API to create a database of popular venues of each neighbourhood.

Formatted in “one-hot encoding”, optimised for clustering.

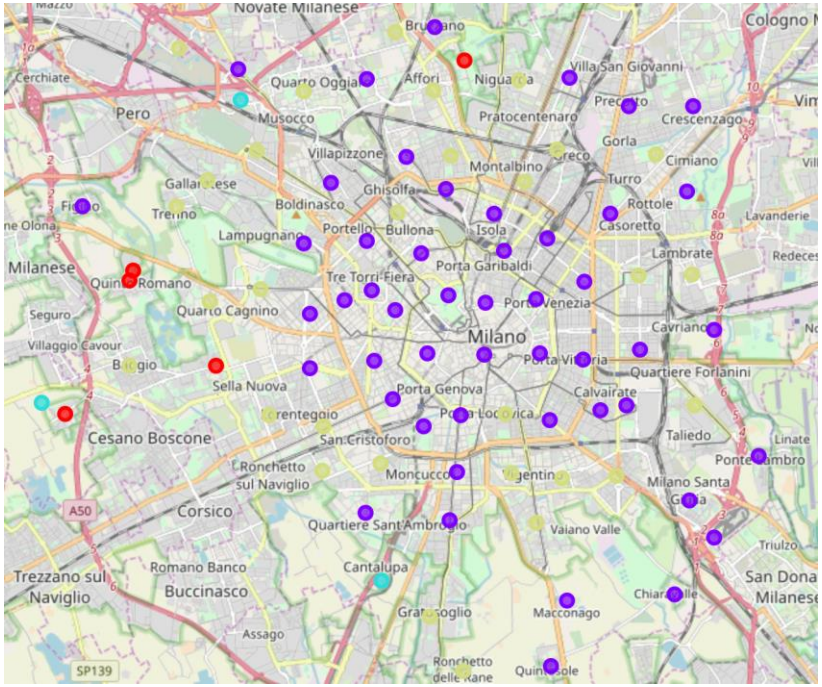
Formatted to express common venues in each neighbourhood, to make it readable and to highlight similarities.

	Borough	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	DUOMO	45.463216	9.187042	Piz	45.462163	9.185767	Pizza Place
1	DUOMO	45.463216	9.187042	Ciacco. Gelato senz'altro	45.463704	9.186796	Ice Cream Shop
2	DUOMO	45.463216	9.187042	Starbucks Reserve Roastery	45.464920	9.186153	Coffee Shop
3	DUOMO	45.463216	9.187042	Piazza del Duomo	45.464190	9.189527	Plaza
4	DUOMO	45.463216	9.187042	Bialetti Store	45.464775	9.188343	Kitchen Supply Store

	Borough	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
0	ADRIANO	Italian Restaurant	Supermarket	Café	Bakery	Arts & Entertainment	Soccer Field	Bistro	Pharmacy
1	AFFORI	Supermarket	Café	Italian Restaurant	Park	Pizza Place	Cocktail Bar	Fried Chicken Joint	Pool Hall
2	BAGGIO	Pizza Place	Italian Restaurant	Gastropub	Convenience Store	Café	Bar	Supermarket	Japanese Restaurant
3	BANDE NERE	Café	Ice Cream Shop	Hotel	Restaurant	Hobby Shop	Diner	Pub	Plaza
4	BARONA	Soccer Field	Bakery	Japanese Restaurant	Tennis Stadium	Athletics & Sports	Trattoria/Osteria	Theater	Café

Clustering by venue composition

- **Cluster 0**: suburban or semi-rural location; presence of sport infrastructure.
- **Cluster 1**: city centre and locations close to important transport ;
- **Cluster 2**: outer locations, close to the outer ring motorway;
- **Cluster 3**: ring of working class areas around town centre.



Data and features preparation: price tier and ratings

Use of Yelp API to collect information of price tier and ratings for the most relevant food venues in each neighbourhood.

	0	1	2	3	4	5	6
0	DUOMO	45.463216	9.187042	Risoelatte	GyilsXEoEw6V9tmV70ZvoA	€€	4.5
1	DUOMO	45.463216	9.187042	Luini	acm_-PPIeqdo7ZLRn5fXJw	€	4.0
2	DUOMO	45.463216	9.187042	Princi	YAvqWdS39-Lb2KyhmhYSmg	€	4.0
3	DUOMO	45.463216	9.187042	Piz	DxgcES-gFf3jFQ9OxYRI4A	€€	4.5
4	DUOMO	45.463216	9.187042	Trattoria Milanese	mmsfdsAJdKkZEr3ibl6SEA	€€€	4.5

Venue name

Price tier

Rating

Averaged per neighbourhood, prepared for clustering.

Creation of the *heat* indicator (simple sum or linear combination of price and rating)

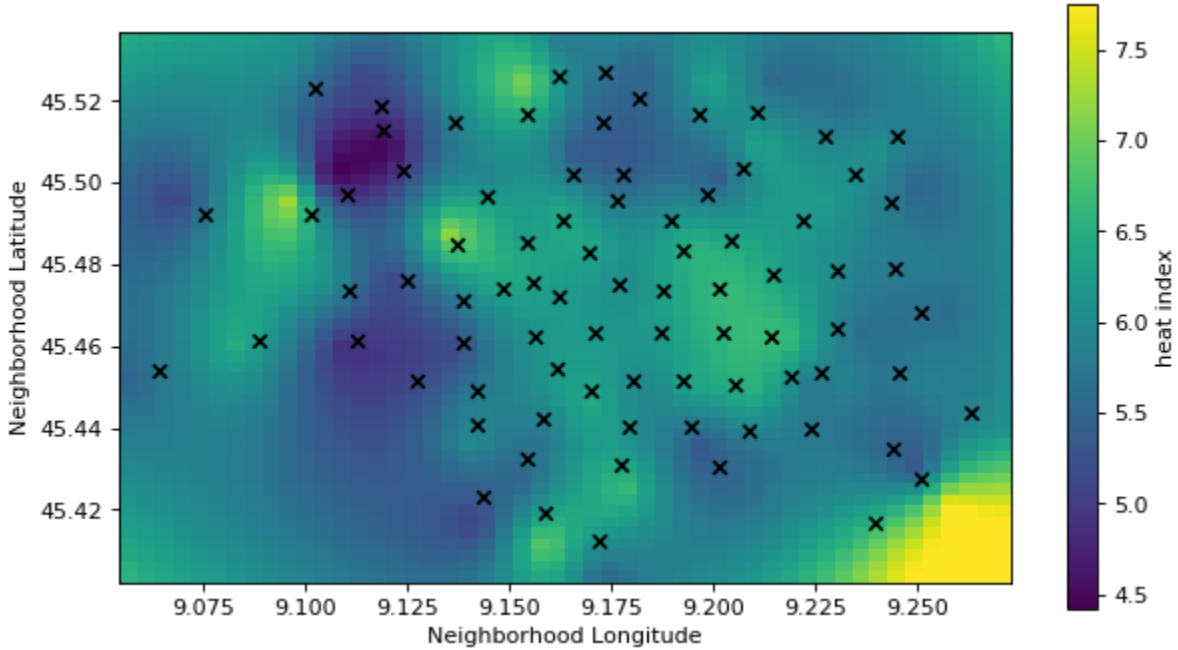
	Borough	Neighborhood Latitude	Neighborhood Longitude	Price	Rating	heat
0	ADRIANO	45.511551	9.244950	2.0	3.850000	5.850000
1	AFFORI	45.514835	9.172873	1.7	3.866667	5.566667
2	BAGGIO	45.461161	9.088360	2.6	3.888889	6.488889
3	BANDE NERE	45.460724	9.138703	2.0	3.166667	5.166667
4	BARONA	45.432509	9.154128	2.0	3.500000	5.500000

Data and features preparation: a *heat* map

Neighbourhoods location marked with crosses;

Map created by interpolation, extrapolated values on borders not reliable;

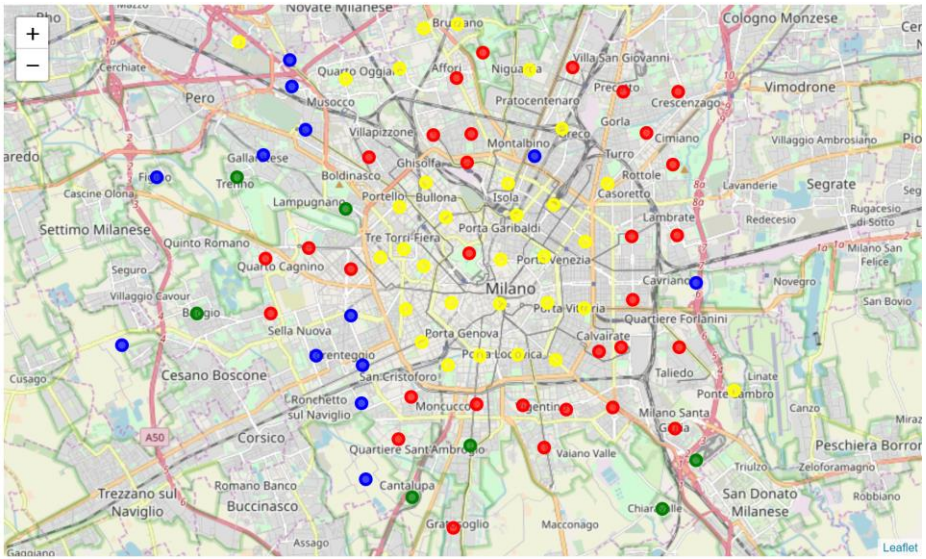
Central *Hot* area, surrounded by cooler ring.



Clustering by price and ratings

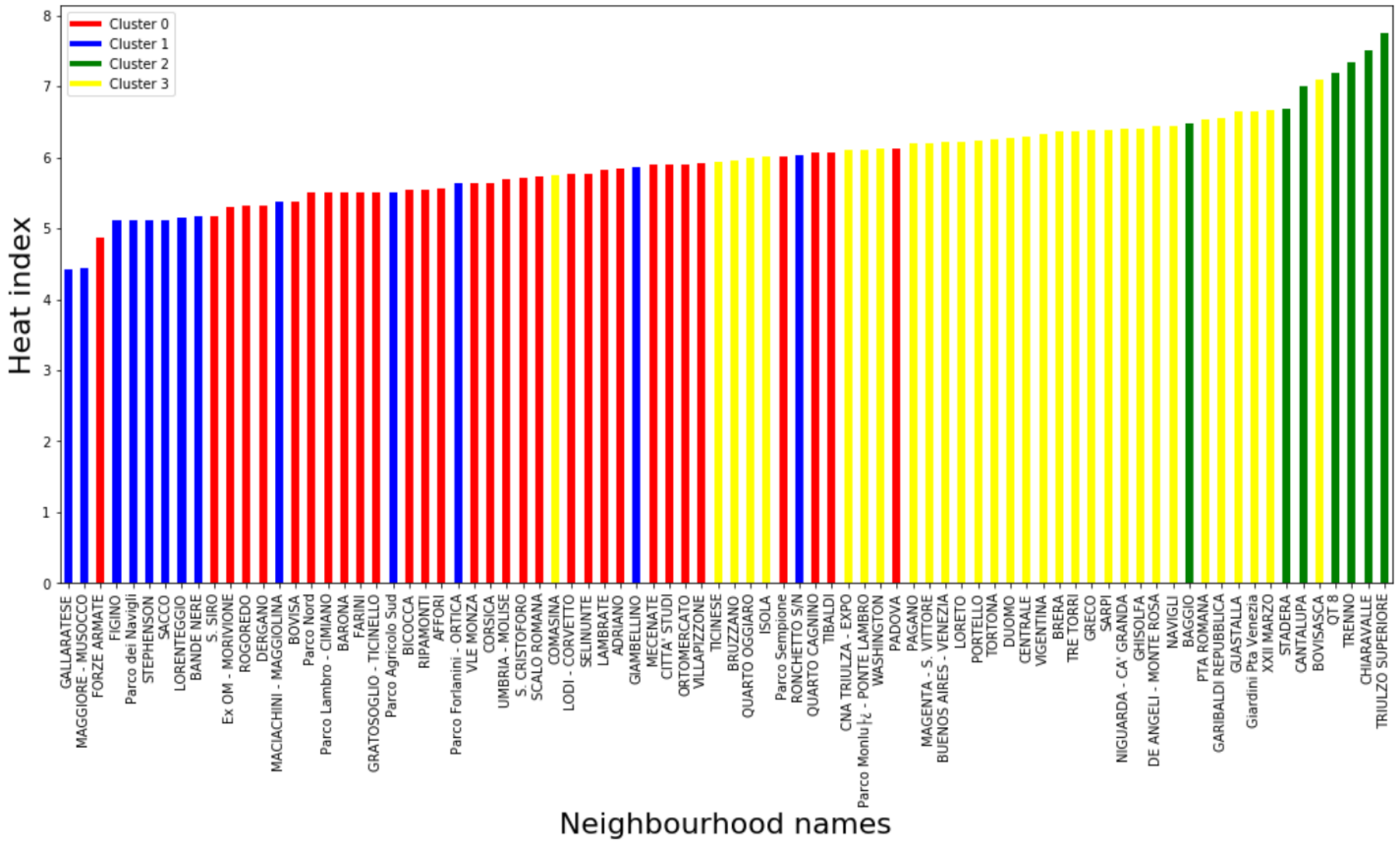
- **Cluster 0:** mostly working class areas, ust outside the inner ring road;
- **Cluster 1:** peripheral areas, close to transport infrastructures;
- **Cluster 2:** on the outskirts, presence of attractions, shopping malls, event venues;
- **Cluster 3:** central neighbourhoods, or newly redeveloped areas, up and coming areas.

	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Price	Rating	heat
0	0	45.471561	9.192746	1.933665	3.713331	5.646996
1	1	45.474506	9.134927	2.133075	3.097329	5.230404
2	2	45.447544	9.164748	2.907143	4.230486	7.137628
3	3	45.480652	9.180711	2.084517	4.224117	6.308634



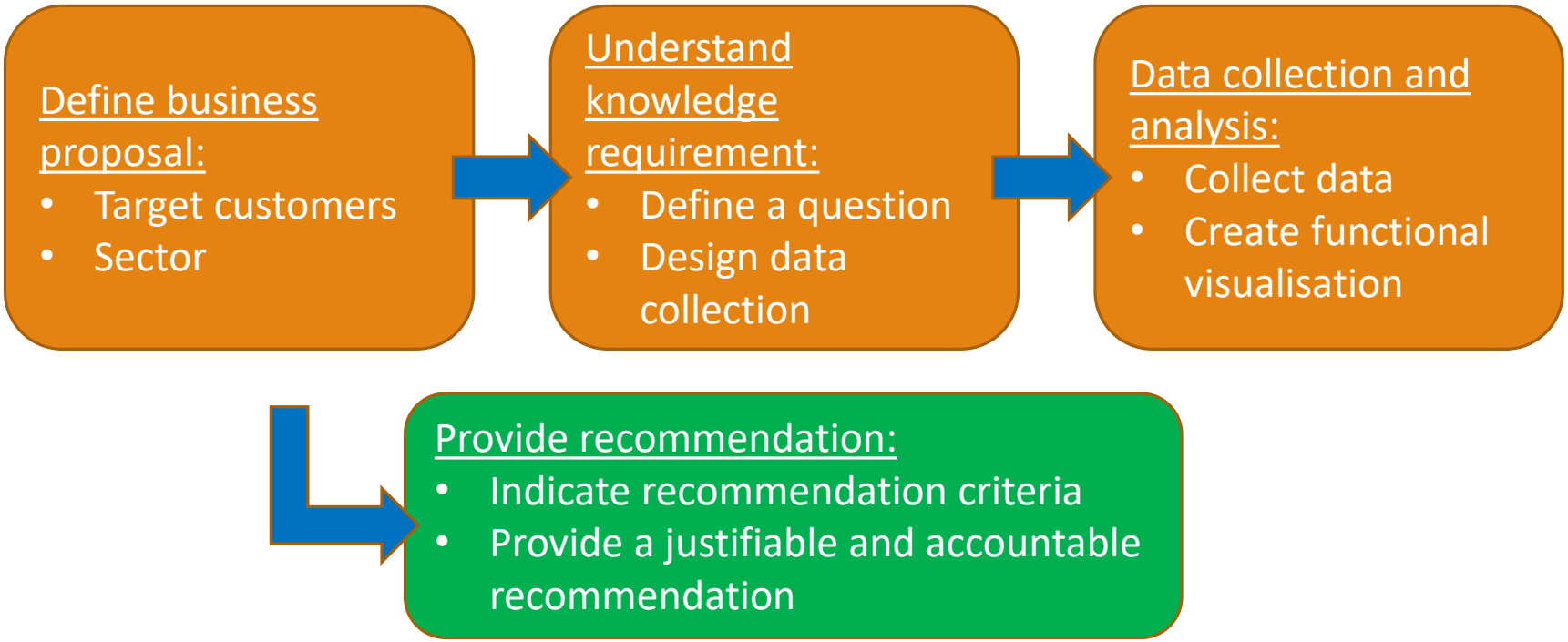
Clustering by price and ratings VS *heat*

Correlation between *heat* and clustering



Use of data: scenario creation

Demonstration of the use of the generated information to make data-informed business decisions.



Scenario 1: where to open a Korean restaurant

Problem definition:

The investor is looking for a location to open a high end Korean restaurant (price and rating indicators > 3).

Hypotheses on which the research is based:

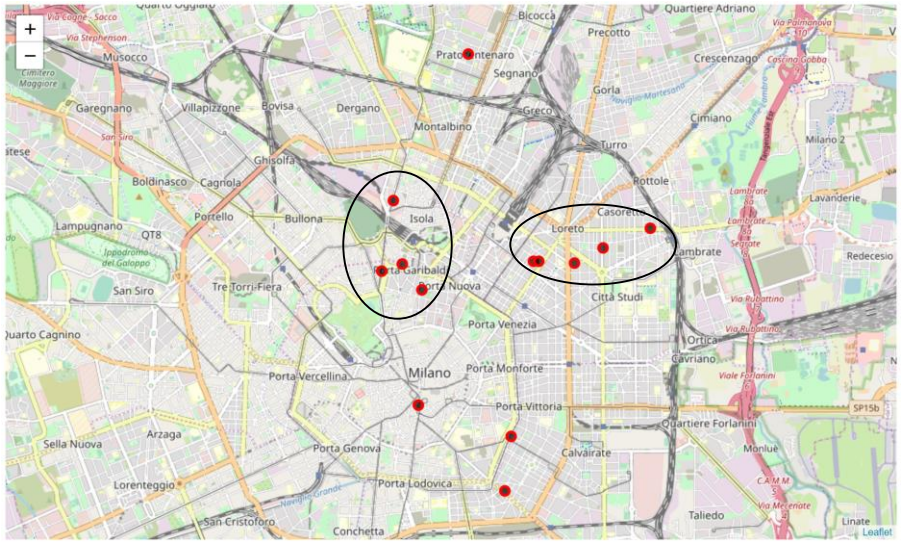
- Concentration of competition, i.e. similar kind of businesses, avoiding saturation or isolation;
- Kind of neighbourhood, i.e. central, high-end, etc.;
- Target price tier and rating of the area.

Scenario 1: where to open a Korean restaurant

Where can I find other Korean restaurants? Yelp API knows.

Two neighbourhoods have a concentration of them:

- Loreto and Cittá Studi, heavily populated with students
- Garibaldi, up-and-coming area, recent re-development



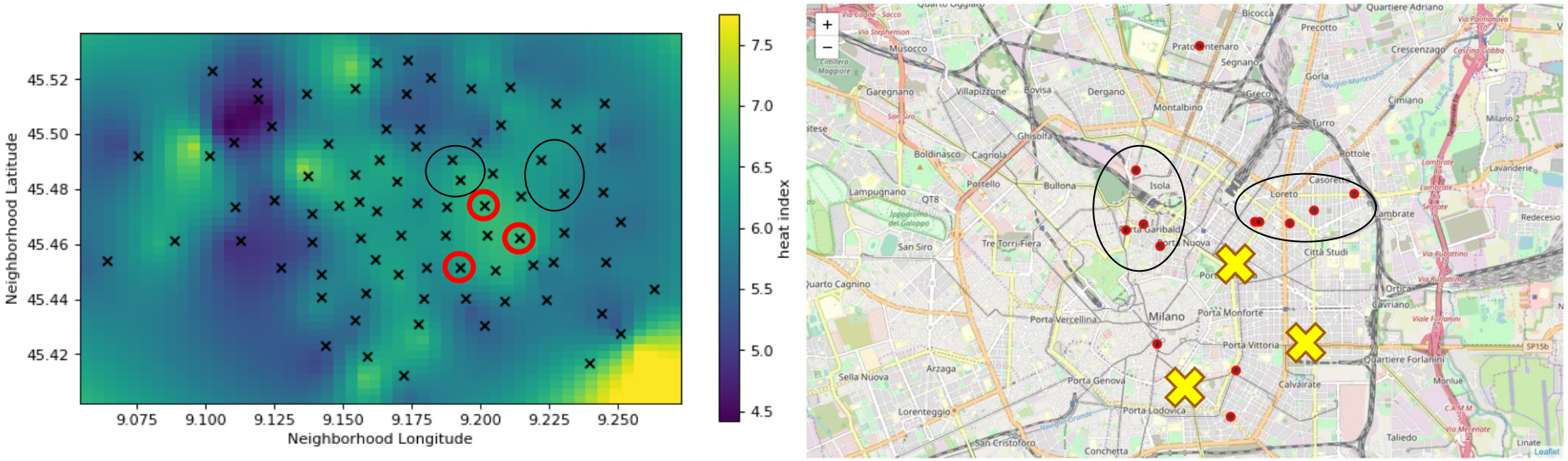
Main properties of the neighbourhoods, from the data:

Neighbourhood	Price tier	Rating	Heat	Price/rating cluster	Venue composition cluster
Garibaldi	2.29	4.27	6.55	3	1
Loreto	2.06	4.17	6.22	3	1

Scenario 1: where to open a Korean restaurant

We can filter the neighbourhood database to find similar areas:

Neighbourhood	Price tier	Rating	Heat	Price/rating cluster	Venue composition cluster
XXII Marzo	2.25	4.42	6.67	3	1
Pta Venezia	2.30	4.35	6.65	3	1
Vigentina	2.17	4.15	6.32	3	3



○ Recommended neighbourhoods ✖

Scenario 2: Identifying new areas for a luxury shop

Problem definition:

The investor is looking for an area to open a luxury business in, away from the traditional luxury districts, such as an up-and-coming neighbourhood.

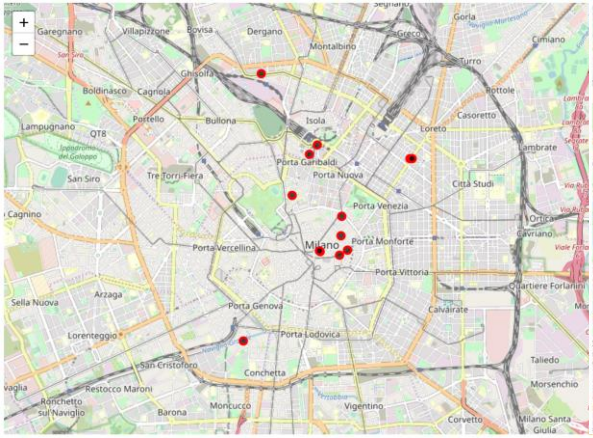
Hypotheses on which the research is based:

- Choice based on similarity to high-end, luxury areas, in terms of clustering, price, rating a heat indicators;
- Geographical proximity to current luxury areas, to avoid isolated or inconveniently located neighbourhoods;
- Low but not null presence of luxury venues, to remove risks involved with pioneering, or to avoid areas with concentrations of other kinds of business.

Scenario 2: Identifying new areas for a luxury shop

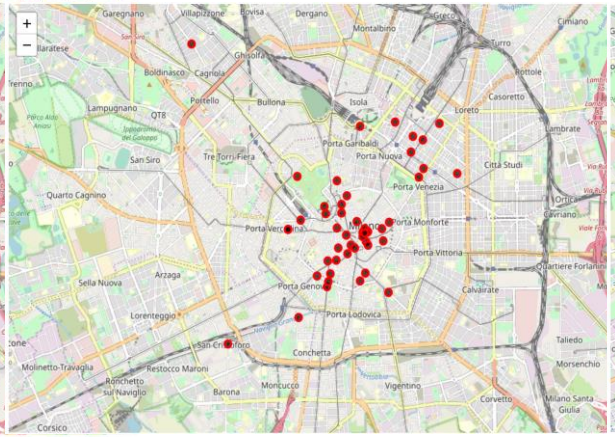
Use of Yelp API to acquire information on Milan's luxury businesses.

1st search: luxury retail



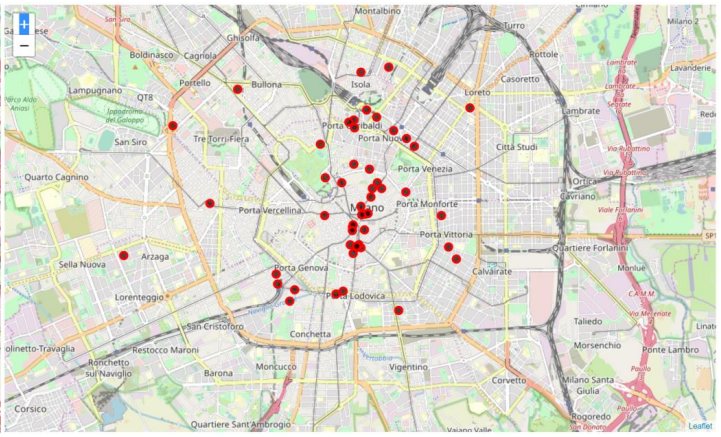
Too little information returned. Search needs to be broadened.

2nd search: all luxury tagged venues



Spurious information returned, such as ice cream shops, discount shops.

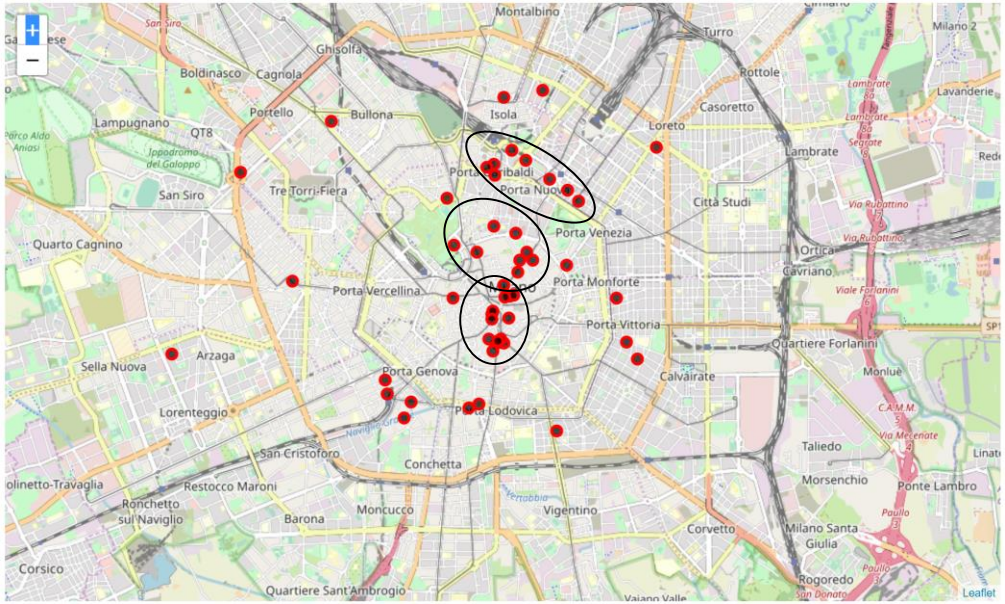
3rd search: venues on maximum price tier



Mix of retail and restaurants, clear geographical picture

Scenario 2: Identifying new areas for a luxury shop

Neighbourhoods with high concentration of luxury venues: identification and description.

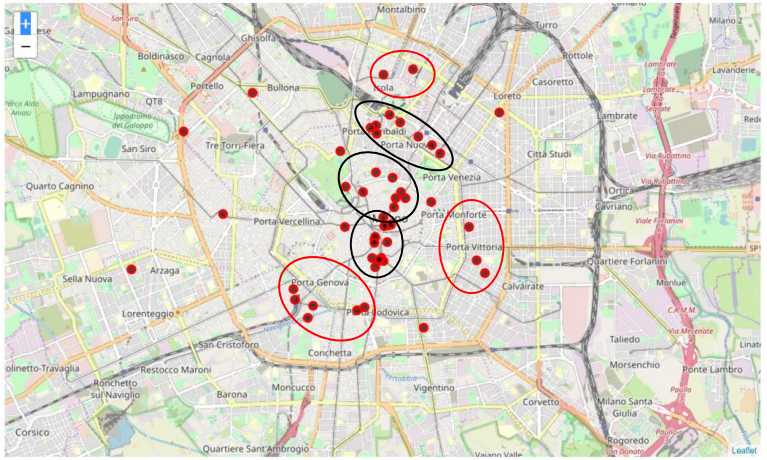


Neighbourhood	Price tier	Rating	Heat	Price/rating cluster	Venue composition cluster
Duomo	2.04	4.23	6.27	3	1
Garibaldi-Repubblica	2.29	4.27	6.55	3	1
Brera	2.08	4.28	6.37	3	1

Scenario 2: Identifying new areas for a luxury shop

First we look at adjacent neighbourhoods with lower concentration of high price venues.

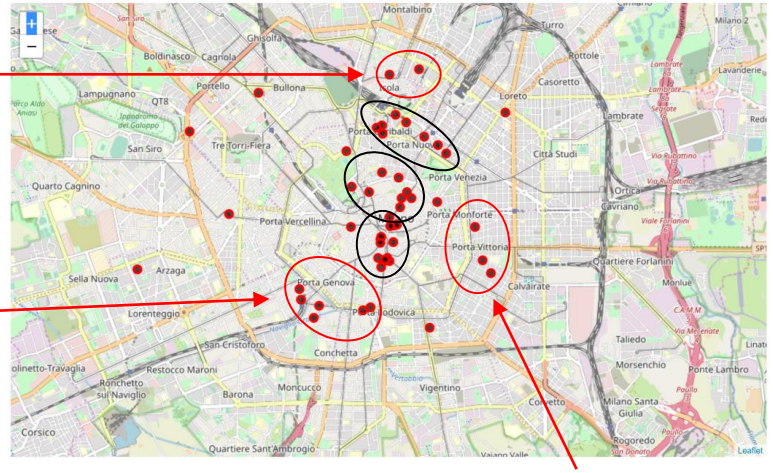
We exclude some on the basis of available information (low-mid end retail oriented areas, restaurants/entertainment areas…)



Neighbourhood	Price tier	Rating	Heat	Price/rating cluster	Venue composition cluster
Tortona	2.14	4.12	6.26	3	1
Navigli	2.17	4.27	6.44	3	1
Guastalla	2.23	4.42	6.64	3	1
Ticinese	1.71	4.22	5.93	3	1
Isola	1.79	4.22	6.01	3	1

Scenario 2: Identifying new areas for a luxury shop

The three proposed areas (one composed of three neighbourhoods) meet the selection criteria and represent our recommendation.



Neighbourhood	Price tier	Rating	Heat	Price/rating cluster	Venue composition cluster
Tortona	2.14	4.12	6.26	3	1
Navigli	2.17	4.27	6.44	3	1
Guastalla	2.23	4.42	6.64	3	1
Ticinese	1.71	4.22	5.93	3	1
Isola	1.79	4.22	6.01	3	1

Conclusions and future directions

- The presented projects gives an example of use of data to influence decision-making in business.
- Data and their visualisations provide information on the city structure on many layers.
- The collected data can easily be expanded and integrated for any purpose with the tools presented, e.g.:
  - data on available properties, to evaluate how feasible is investing in a given neighbourhood;
  - integration of data on residential and business property prices;
  - evolution of indicators in time and correlation with re-developments, construction of new transport links, etc.