Taking a Look at Milanese Food Venues

By Francesco Rivano

1. Introduction

1.1 Background

Milan is the de facto economic capital of Italy, home to its stock exchange, famous for its role in the fashion industry and its important industrial past as part of the Italian industrial triangle (Genoa, Turin, Milan). It is obvious to expect a wide variety of cuisines and an explicit openness to foreign tastes (unlikely to be found in most areas of Italy, considering how many Italians are proud of their cuisine, both regional and national, and not that interested in recipes from abroad).

1.2 Business Problem

This report will show what kind of (food-related) venues are popular in a certain area and will, in a way, implement and suggest the opposite of "low density = good, high density = bad": we shall try to understand what are the most popular venues and their distribution among the relatively tiny Milanese neighbourhoods (Milan is a pretty compact-sized city). Lack of official datasets about crime means we will be unable to rely on that to avoid or suggest a certain area. "Domain knowledge" (that is, personal) or informations that could be found in guides online will give us a rule-of-thumb way to assess the quality of a certain area. The business problem to be solved is trying to find a trend in order to invest money in an area and/or finding the best venues to provide tourists with suggestions (maybe tailored to their purchasing power, tastes, and similar criteria). Also, are there differences in terms of foreign and Italian tastes (when they can be evaluated, as there could be too small a sample in terms of reviews to even compare them)? In a traditional European city, is the city centre home to the "best" restaurants, or do other neighbourhoods feature better choices?

1.3 Interesting for:

Investors, not only those interested in food venues per se, but also in related and unrelated venues and activities (museums, cinemas, hotels) that would benefit from proximity to renowned and healthy restaurants and similar businesses. Or else people that want to write a guide for tourists or just talk about food (journalists, perhaps, or bloggers).

2. Data

2.1 Data Sources

Foursquare and its API will provide most data concerning the venues. A small set of ad hoc Python functions will make retrieving information about each venue fairly easy. GeoJSON files can be found at http://dati.comune.milano.it/, and any kind of domain knowledge in terms of neighbourhood safety will be searched on Google, trying to rely on official sources whenever possible.

2.2 Data Cleaning

A few issues have arised: first and foremost, the GeoJSON file for Milanese neighbourhoods complies with the EPSG:32632 WGS 84 / UTM zone 32N standard instead of the more beginner-friendly latitude-longitude. We can expect more than a handful of NaN values, due to FourSquare's

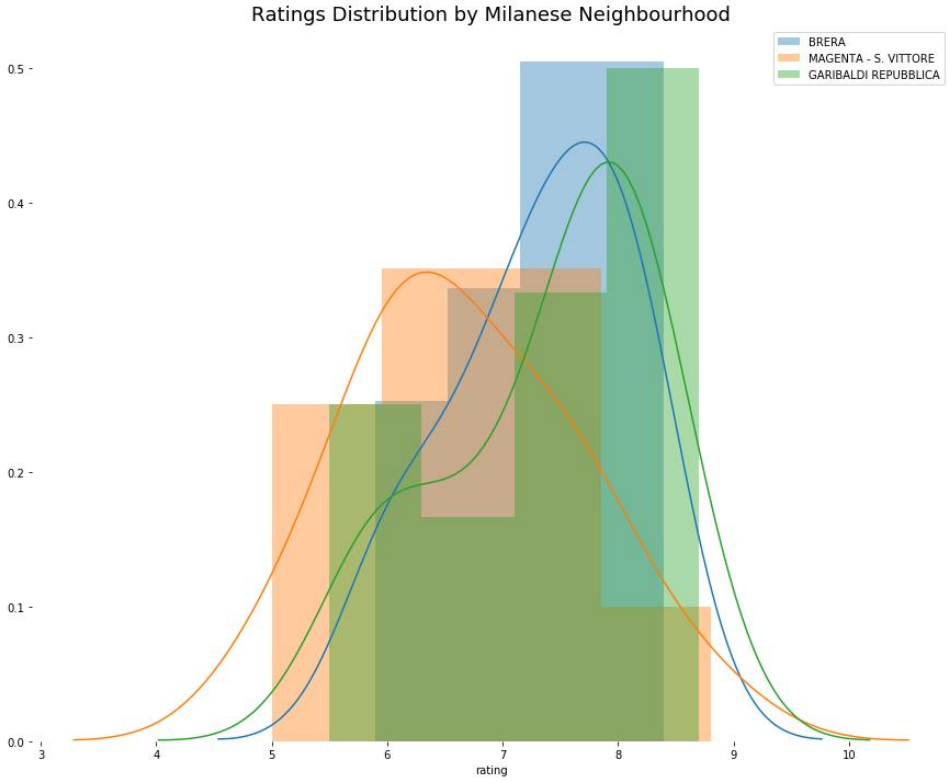
relatively small userbase in Italy/unpopular venues. Moreover, Foursquare does not provide the neighbourhood for most venues.

2.3 Feature Selection and Feature Engineering

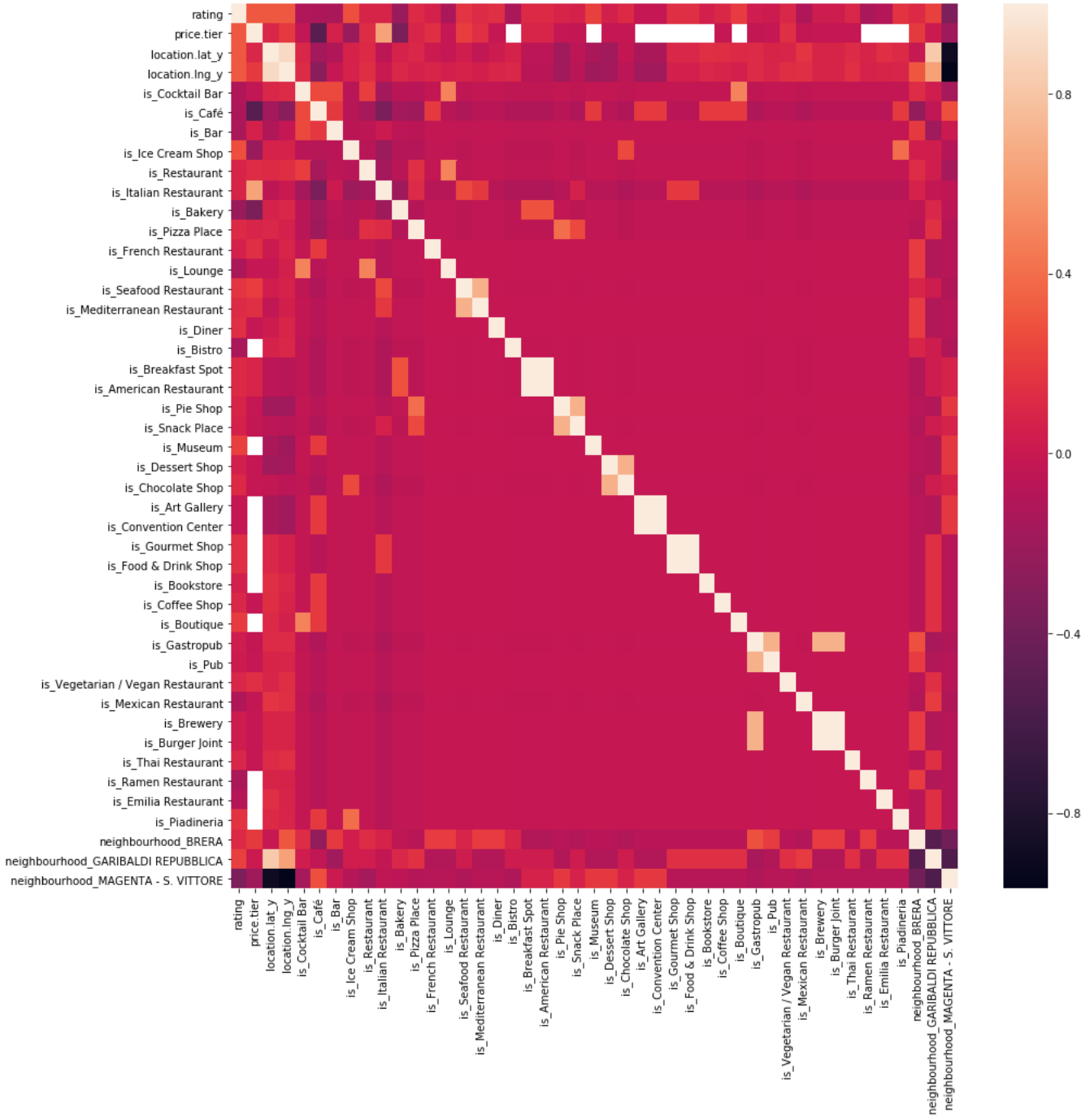
We will focus on things such as ratings, price tiers, and neighbourhoods. The first two are readily available when relying on the Foursquare API, while the latter did require a few extra steps (finding the centre of a neighbourhood, setting up a range, and including the results of the search into the dataset as part of said neighbourhood, while still considering that we may or may not get all the results to be inside our target neighbourhood).

3. Data Analysis

The FourSquare API put a considerable limit to the data acquisition part: that resulted in choosing only three popular and iconic neighbourhoods and hoping that they would feature interesting venues. Brera, Garibaldi – Repubblica, and Magenta – San Vittore were chosen as close to the city centre (or part of it) and popular, and therefore likely to have reviewed venues. Removing venues with no ratings from the cleaned "offline" dataset meant going from 116 venues to only 70. We can't just display the average rating by neighbourhood or the median in such cases, can we?

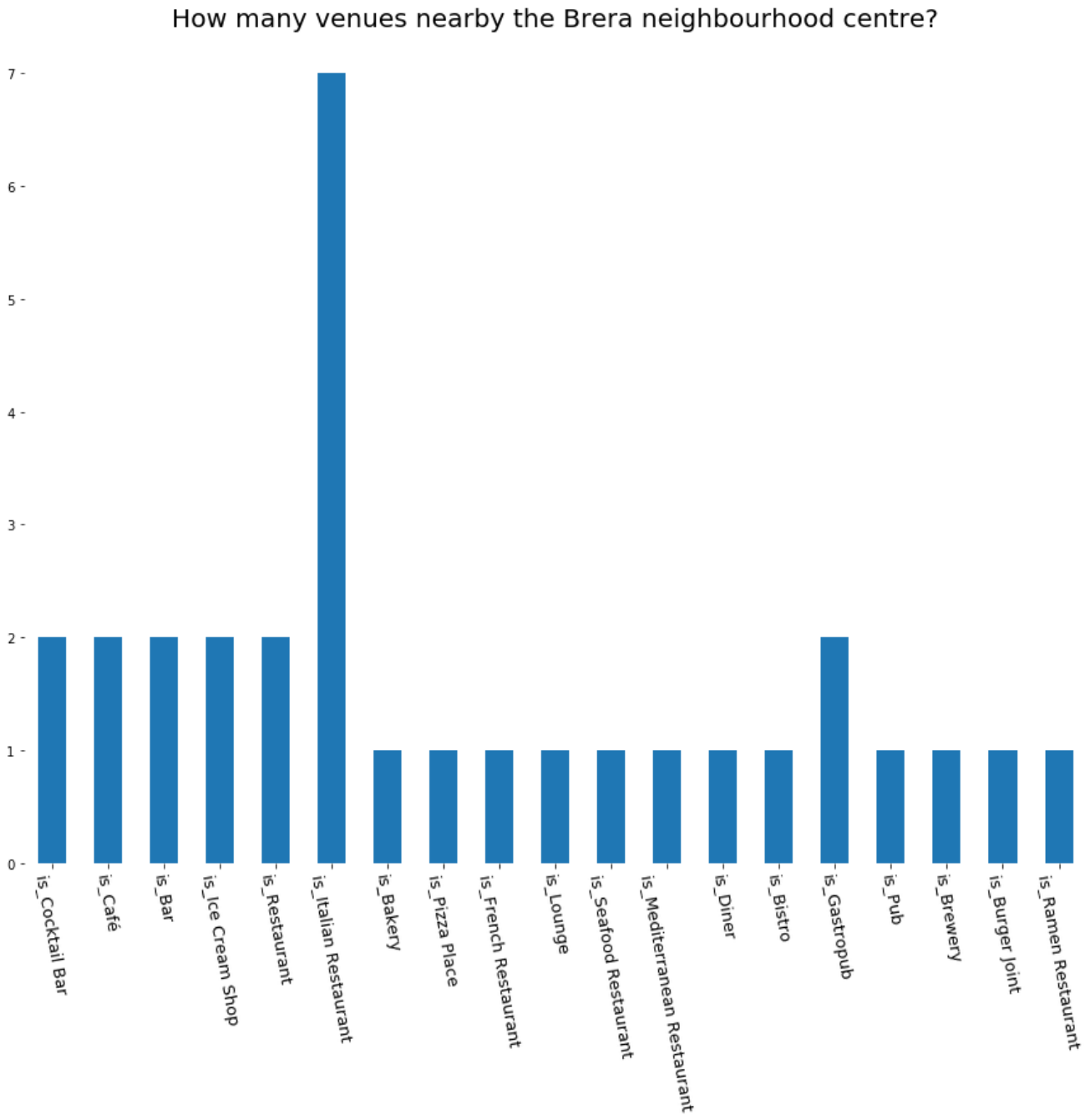


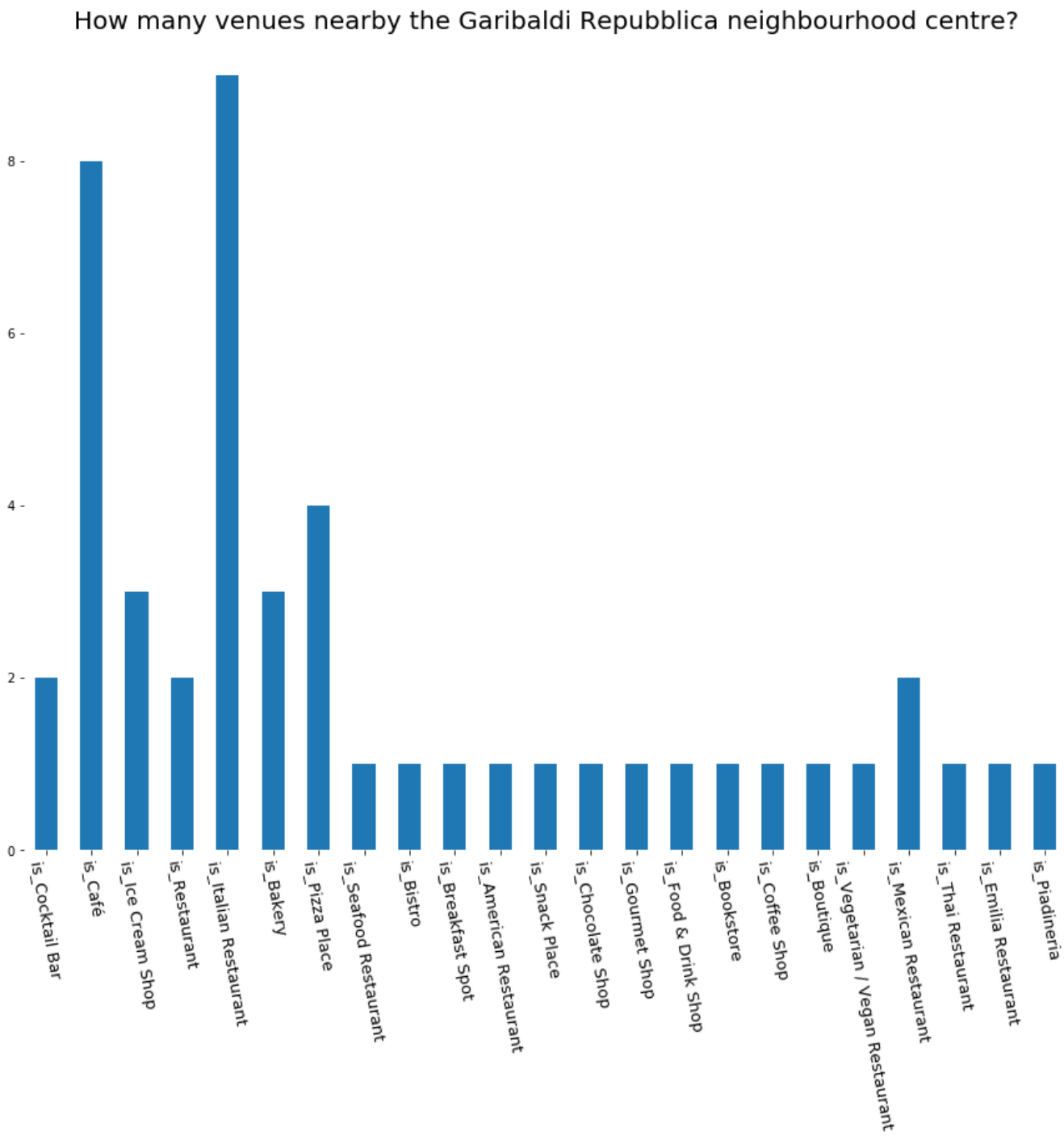
The ratings distribution for both Brera and Garibaldi – Repubblica, according to our tiny sample, looks left skewed, whereas Magenta – San Vittore isn't as skewed and the venues in the latter seem to be not so popular.



The correlation matrix is not a complete disappointment. While there are not so many strong correlations (except for the lone venues being the only representatives of 2 categories), we can notice how rating is weakly correlated with price tiers, being an ice-cream shop/boutique, being in Garibaldi - Repubblica, and weakly negatively correlated with being in Magenta - San Vittore or a bakery. Italian restaurants seem to be overall pricier than the alternatives (0.63 correlation with the price tier, going 1 to 3, with 3 being the most expensive tier), besides a way weaker correlation between price tiers and seafood restaurants. Bakeries tend to belong to the cheaper tier(s), it seems. Neighbourhood-wise, Brera venues seem to be slightly more likely to be more expensive, Magenta - San Vittore ones tend to be slightly less likely to be expensive.

Let’s take a look at the quantitative makeup in terms of venues types in each neighbourhood.


How many venues nearby the Brera neighbourhood centre?
7
6
5
4
3
2
1
0
is_Cocktail Bar
is_Café
is_Bar
is_Ice Cream Shop
is_Restaurant
is_Italian Restaurant
is_Bakery
is_Pizza Place
is_French Restaurant
is_Lounge
is_Seafood Restaurant
is_Mediterranean Restaurant
is_Diner
is_Bistro
is_Gastropub
is_Pub
is_Brewery
is_Burger Joint
is_Ramen Restaurant



4. Conclusions

Our brief Data Analysis, in spite of being heavily limited by FourSquare's limits, still gave us an insight about the kind of venues in each neighbourhood. We have noticed how bakeries seem to be more likely to criticised, how venues in Magenta – San Vittore are also not so popular, how Brera is more expensive than average, and that even in Milan Italian food is not on the cheap side (while still managing to dominate the food venues sector quantitatively, along with cafés). Had not there been so many limits to FourSquare APIs, a "carpet" analysis of the whole city would have become a viable option (or at we could have increased the sample size dramatically). Realistically speaking, the resulting dataset is too tiny to do any serious modeling. We can merely observe it and draw our own

conclusions. Brera is pricier, Garibaldi – Repubblica has popular venues, those in Magenta – San Vittore could be disappointing, according to reviewers.