Predicting the best place to open a Beer Bar in Milan

Claudio Brutti

November/December, 2019

1. Introduction

1.1 Background

Food and drink venues are very popular in Italy, and in Milan, finding a good place where to open a venue, without being overwhelmed by a crowd of similar venues in the near, it's very hard.

1.2 Problem

The client has already opened successful beer bars in other towns and knows that the correct location is a key factor to have a good choice to ensure to have enough customers to run the business. He wants to analyze the Milan city neighborhoods and use this information in order to find the best place.

2. Data acquisition and cleaning

2.1 Data sources

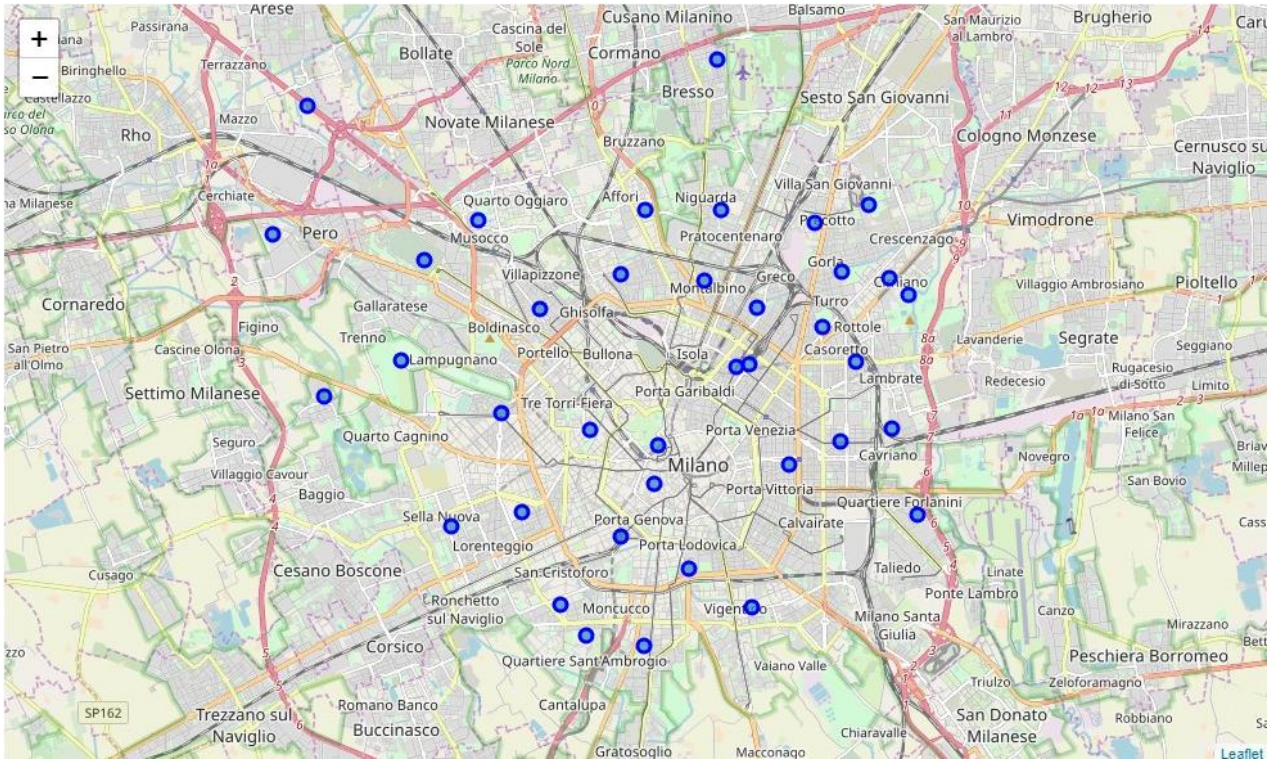
To group the city locations we decided to use the postal codes (CAP) that divides Milan. A CSV file with streets and their correspondent CAPs was previously produced form a webscraping of site http://www.omarpela.com/milano/cap-codice-di-avviamento-postale-vie-milano.html, to have a way to correctly place a CAP center in the city map. This CSV was uploaded to Github and published as https://raw.githubusercontent.com/clabru/Coursera_Capstone/master/capmi.csv.

A Foursquare data set was used to retrieve all the venues in the proximity of each CAP center.

2.2 Data cleaning

The webscraping produced some entries with street names in various format. A first cleaning of this dataset was needed in order to have street names in a standardized format.
A geocoder from geopy library helped to localize a center point for a group of streets belonging to the same CAP. Some of the streets could not be found, so we divided the streets I three different group with three different algorithms in order to have at least one street from one group to be

correctly located. If streets from more than one group were correctly located, the center of the corresponding CAP was assigned to the average coordinate.
The resulting centers set is represented in the following image:



For each of these CAP's coordinates we asked Foursquare to retrieve the venues in the nearby.

2.3 Feature selection

We were only interested in the quantity of venues grouped by category in the nearby of a CAP center, so no features other than coordinates and category name were extracted.

3. Exploratory Data Analysis

3.1 Calculation of data subset

The inherent nature of the problem pointed us to perform an analysis using a clusterization of the venues data to have a visual feedback of the results. The clusterization will group different areas with similar characteristics.

At first, we looked at all the categories in our analysis. Most of them were clearly not related to our problem. The presence of an art gallery or a toy store, were not considered influential for our scope. We decided then to consider only food related venues in our analysis. After some

exploratory analysis, we found that we could not trace a useful pattern using all the full food related categories. We decided, therefore, to filter our dataset furthermore, including only drink related venues.

4. Cluster Modeling

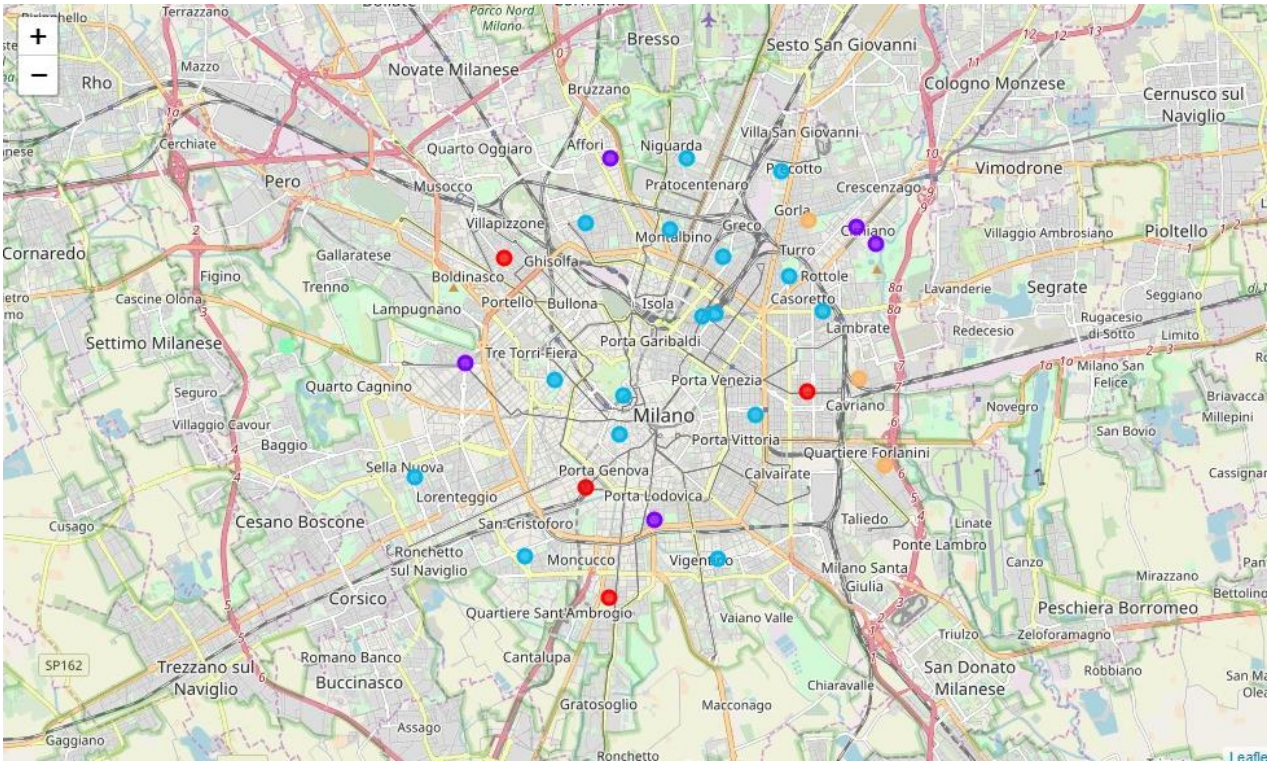
The clusterization method we are going to use will be the K-means, which is an unsupervised algorithm, and that will allow us to have a non-overlapping partitioning of the areas.

Analyzing these areas and their characteristics we will find possible patterns to use as a response to the original question.

The categories obtained by Foursquare, will be grouped for frequency in that area and then sorted, on order to have information about what types of drinking venues are more present in each area.

After a few tries, we found that a partition of 5 groups would result in a good segmentation of the data and a meaningful grouping.

The map obtained applying a K-mean clusterization to our filtered dataset is represented in the following map.



The group 5, represented in orange, has the interesting characteristic of having as the most common venue a Brewery/Beer Bar. In all other groups, Beer bar are at most in third place.

	cap	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
6	20127	Beer Bar	Brewery	Wine Bar	Food & Drink Shop	Cafeteria
12	20134	Beer Bar	Brewery	Wine Bar	Food & Drink Shop	Cafeteria
16	20138	Brewery	Wine Bar	Food & Drink Shop	Beer Bar	Cafeteria

We found, here, two interesting patterns:

- The outer part of the city is the one where Beer bars are most common, and therefore more likely to have a good “audience”
- The eastern part is already full of those venues, so, it would be wise to choose the western part of the town in order not to set our location in an already overcrowded area.

5. Conclusions

A good choice of location would be the outer western part of the city. All the CAPs of group 1 (in red) in that part of the city, would be a good choice since they represent areas where Beer venues are almost absent. For instance, the CAP 20156 area would be a good choice, since it’s a group 1 zone and also in the outer western part of the city.