EVALUATING THE BEST LOCATION FOR A NEW RESTAURANT IN MILAN

By Marco N.

1. Introduction

The restaurant business is among the most interesting ones at the present day, but it also involves big risks: newly set up restaurants require large capital investments and operational costs and, if a new outlet fails to break even in the first few years, the probability that the venue will close will largely increase, thereby incurring huge losses.

Among the many factors which may influence the success of the new venture, its location will undoubtedly play an essential role. Therefore, it is imperative for a newcomer to perform a thorough analysis of the future location before the investment is made.

The primary objective of this analysis is to help future investors make an informed and optimal decision about selecting a location for a new restaurant business in Milan, Italy, before actually opening the new outlet, helping them taking a calculated risk.

Since Milan's Expo in 2015, the city has enjoyed a remarkable gastronomic boom, with many new restaurant openings. So, the main idea to solve our business problem is to use the available information about already existing restaurants and neighborhood demographics to look for new opportunities in various locations within the city.

2. Data

I was able to obtain data about the restaurants currently existing in Milan by leveraging Foursquare's Global Database. Foursquare built an accurate dataset of location data by crowd-sourcing it, using the people's voluntary input to complete any missing information they had.

I leveraged the Explore endpoint from Foursquare's Places API to obtain data for all the food category venues in Milan, effectively filtering out all the places from other categories which are not in scope with this analysis (see the entire categories list here).

After cleaning up the API's output, I got a clean dataset containing the features I needed:

Restaurant Name	Restaurant Category ID	Longitude	Latitude

As per the neighborhood demographics data, luckily the government of the city of Milan has joined the Open Data movement and has made many datasets available to the general public. I was able to find data about the current neighborhood demographic here, and a projection of neighborhood population here.

The first dataset contained information about the population of the neighborhoods of Milan for the period 2011-2018, while the second one contained an estimate of the neighborhood population for each year up until 2038.

Finally, using the official geospatial data delimiting each neighborhood (also available from the Milan Open Data portal here), for each restaurant found from the Foursquare API I was able to get the corresponding neighborhood.

So, I had now all the data I needed to begin analyzing the relation between the restaurant locations and the density of the population of the restaurants' neighborhoods.

3. Methodology

Once I began the Exploratory Data Analysis phase, it was clear that the population datasets presented many problems.

The first one was that the first dataset divided the population of each neighborhood by age groups, while the second one was much more granular, providing the projected population per each age (age 0, age 1, age 2... until age 90). Since I was looking for restaurant profitability, I inferred that the people in minor or elder age are the ones with less spending capability: therefore, I selected just the data for people with age between 18 and 65.

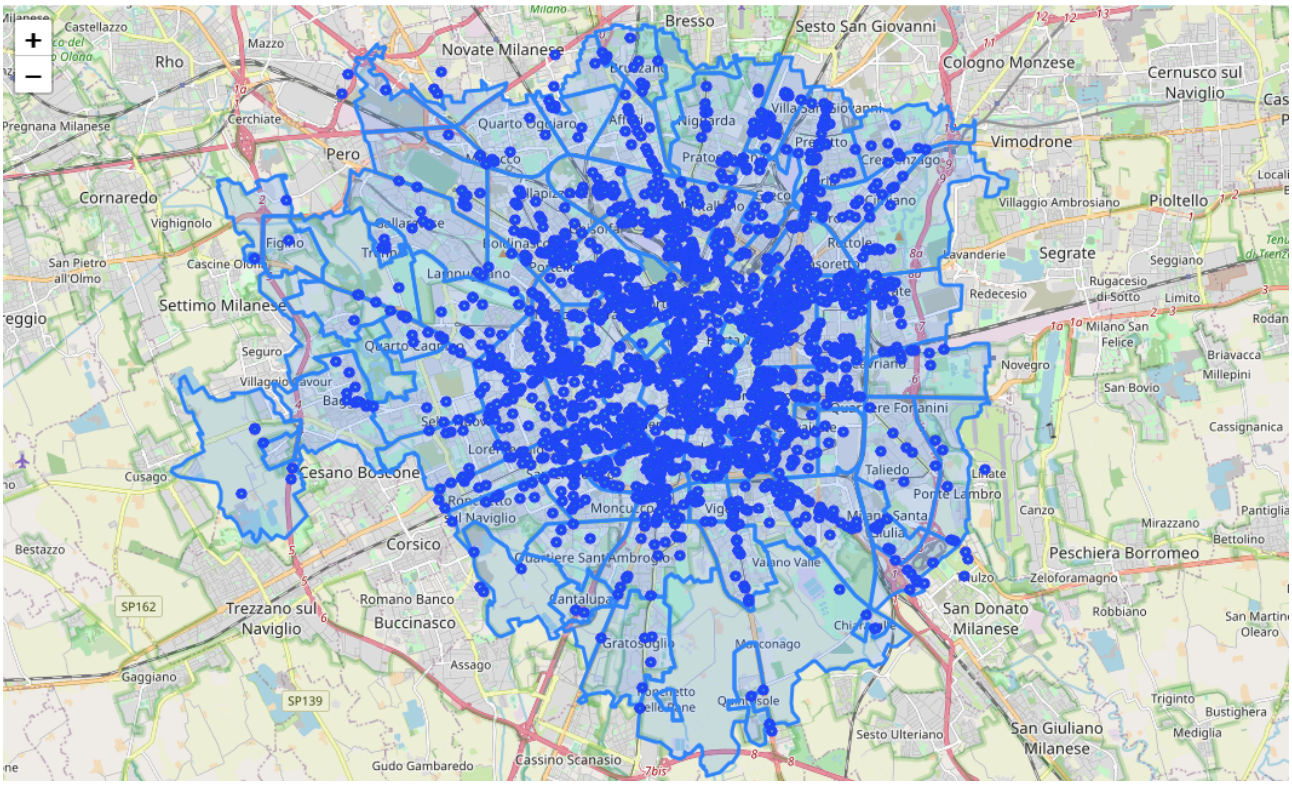
The second problem was about the discrepancies between the neighborhood names in the two datasets: I was able to resolve it by writing a script to check among the neighborhood names, select the ones which were different, and then change them manually.

Also, the datasets did not contain any information regarding the neighborhoods' location, so I had to gather the geo coordinates (latitude, longitude) per each neighborhood by using a geocoder service, which translated each neighborhood's address in GPS locations.

Lastly, the data range of the two datasets was too wide, so I selected just the latest official data for the current population (from 2018), and I decided to select the projected data for the year 2025, because a seven-year period gives a good estimate for a restaurant profitability.

As for the dataset coming from Foursquare, there has been some cleaning to be done, because the Foursquare output for venues in the "food" category was containing many venues which were out of scope, such as "bars" or "food trucks". Since I was only looking at restaurants, I had to manually drop all those unrelated categories.

I was able to obtain a GeoJSON file which contained all the boundary information for Milan's neighborhoods, so I was able to overlay it while visualizing all the restaurants dataset:



Due to how the Foursquare restaurant dataset has been constructed, many locations were assigned to the wrong neighborhood. I was able to resolve that by using the aforementioned GeoJSON file and check which of the neighborhoods' polygons contained each location.

Having now the correct neighborhood assigned to each venue, I was able to get the top 5 restaurant categories per each neighborhood by running a "One-Hot Encoding" on the restaurant dataset and grouping it by the mean of the categories' frequencies.

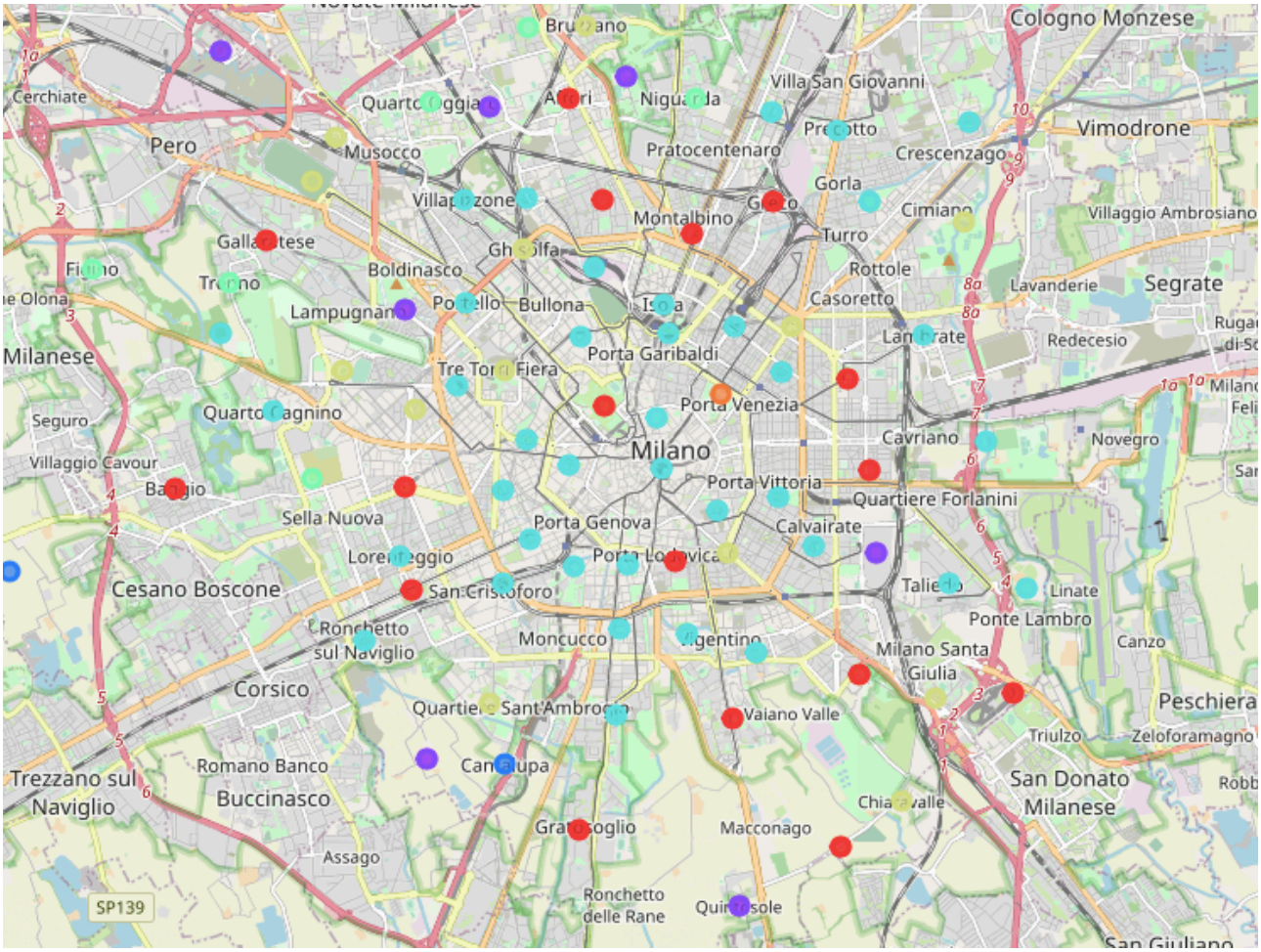
I was looking for similarities between the restaurants in each neighborhood to decide the best suited neighborhood to establish a particular restaurant type (for instance, a Pizzeria), so I decided to run a clustering algorithm. Among the various machine learning techniques, I chose the K-Means algorithm because it is very fast and because it can work with unlabeled data such as our restaurant categories.

I also looked for a relation between a restaurant's location and its neighborhood's population: I did that by calculating each neighborhood's population density and

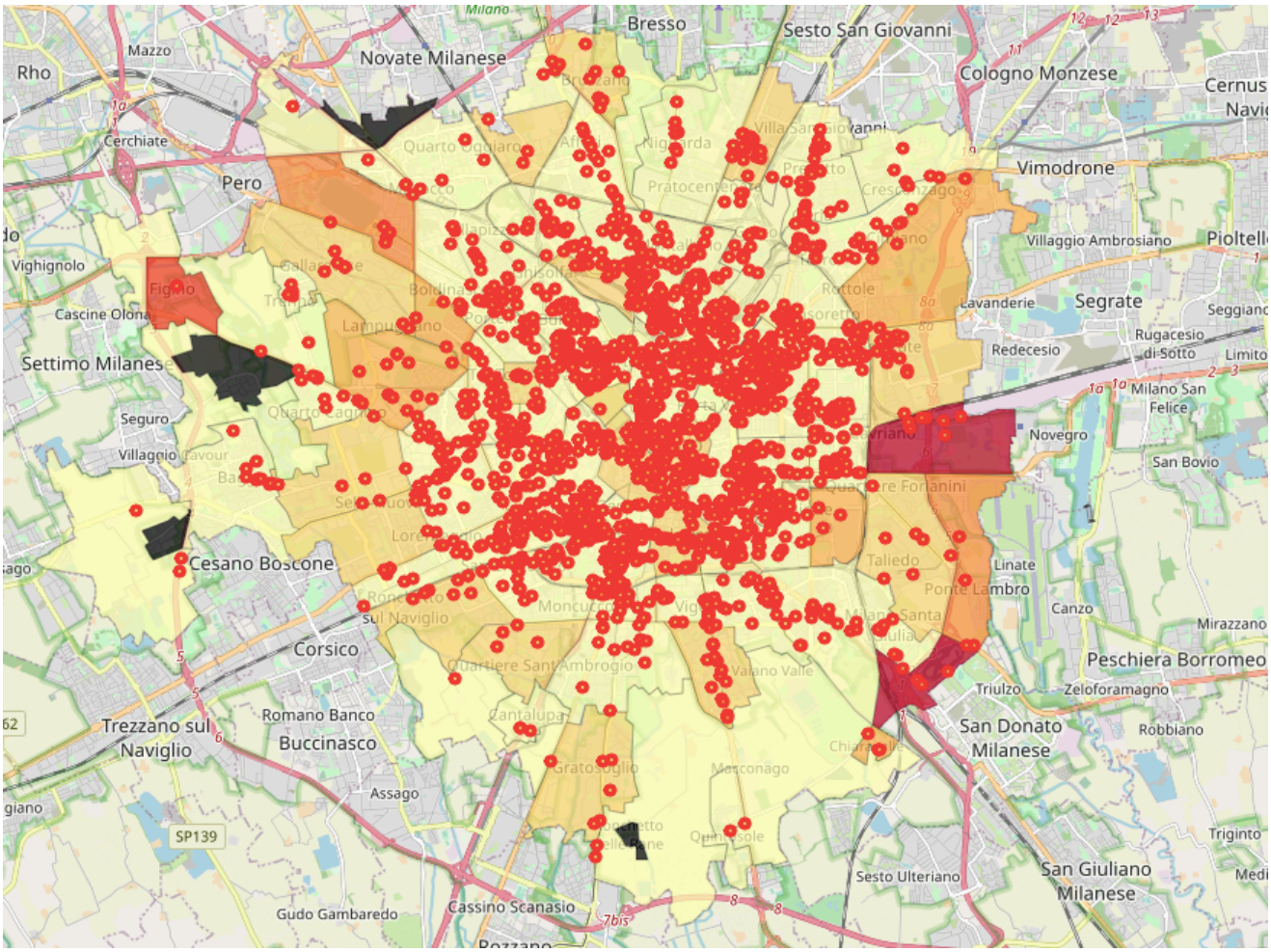
restaurant density (namely, the number of restaurants in the neighborhood divided by the neighborhood's area.)

4. Results

Using the K-Means algorithm, I was able to group the neighborhoods by the similarities between the restaurants in them. The image below represents the clustering by assigning the same color to the neighborhoods which share frequently the same restaurant categories:



As per the relation between each neighborhood population and restaurant density, I was able to make a choropleth map showing the restaurant locations and the neighborhood's population density at the same time:



The map above shows a more intense color for the neighborhoods with a greater working population density.

5. Discussion

The K-Means clustering by restaurant category clearly showed that the best neighborhoods for establishing an "Italian" restaurant are the ones belonging to Cluster n°3. Those are very central neighborhoods, where there is a high tourist demand for Italian restaurants.

	Quartiere	1	2	3	4	5
0	Duomo	Italian	Sandwiches	Pizza	Bistro	Restaurant
1	Brera	Italian	Japanese	Pizza	Bakery	Restaurant
3	Guastalla	Italian	Pizza	Bakery	Restaurant	Japanese
5	Ticinese	Italian	Japanese	Pizza	Restaurant	Vegetarian / Vegan
6	Magenta - S. Vittore	Italian	Pizza	Sushi	Sandwiches	Bakery

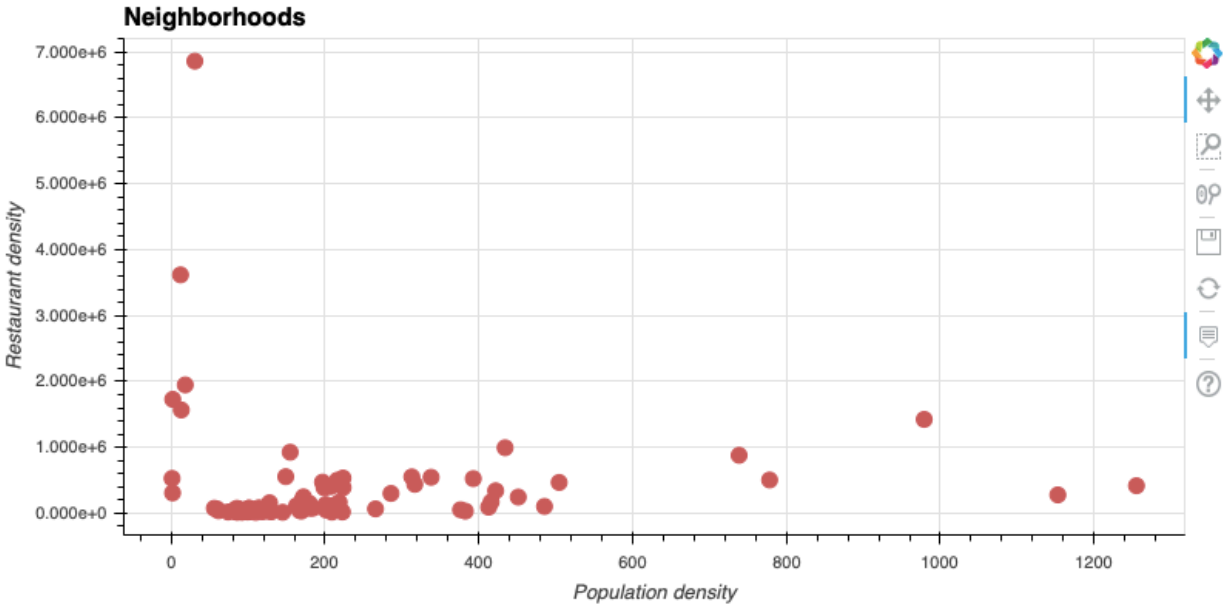
I'd suggest establishing a Pizzeria in one of the neighborhoods in Cluster n°0. Those are very "young" neighborhoods, suggesting there is much demand for low-cost food:

	Quartiere	1	2	3	4	5
4	Vigentina	Pizza	Restaurant	Italian	Sushi	Burgers
7	Parco Sempione	Italian	Restaurant	Pizza	Moroccan	Steakhouse
11	Maciachini - Maggiolina	Pizza	Seafood	Italian	Japanese	Bakery
12	Greco	Pizza	Italian	Restaurant	Greek	Gastropub
21	Città Studi	Pizza	Italian	Seafood	Bakery	Chinese

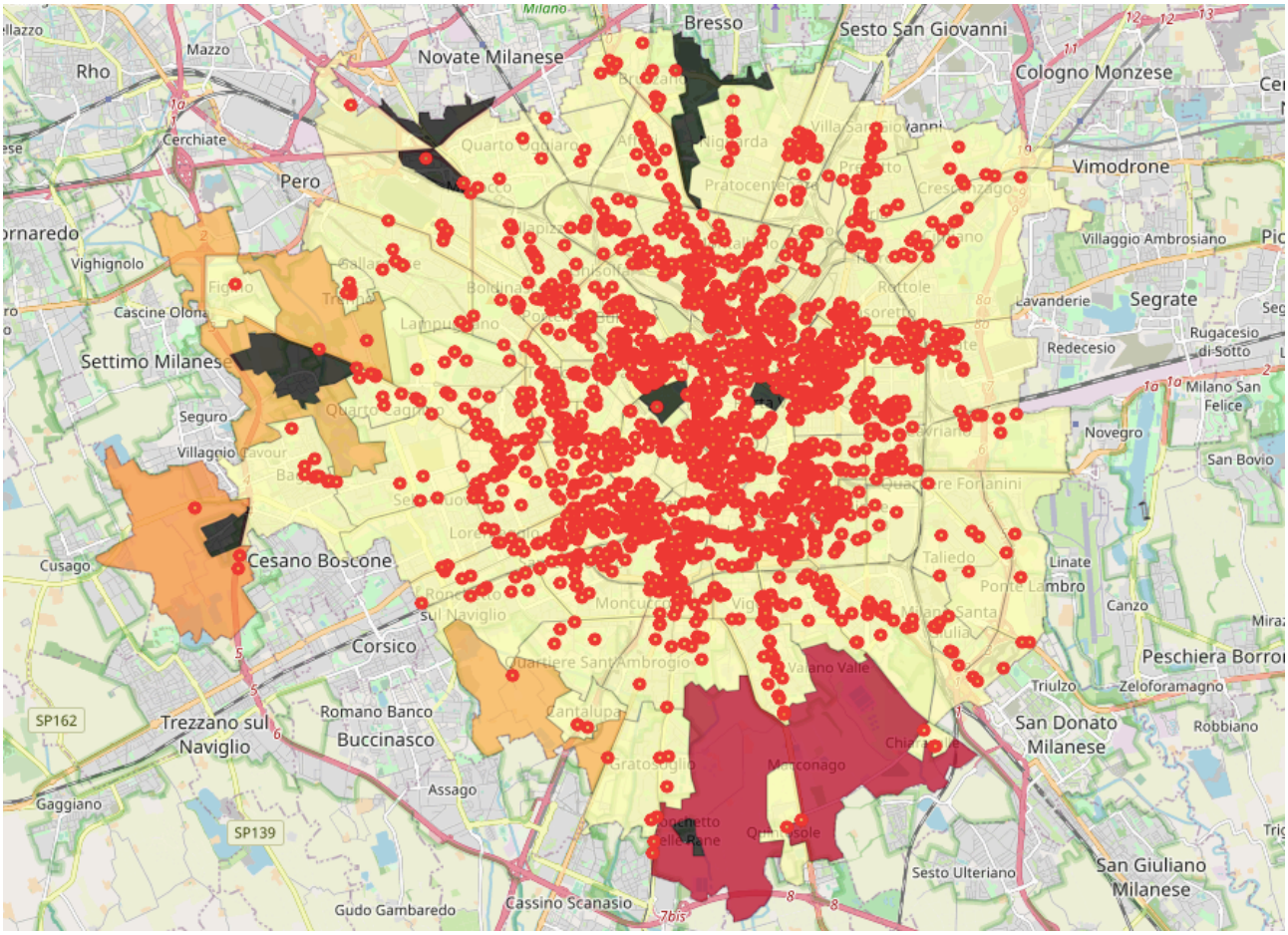
Lastly, I'd suggest establishing an "Ethnic" restaurant in one of the neighborhoods listed below (Cluster n°3). Those are suburban neighborhoods, where the prices for real estate are lower than the city center, thus making them the best place for foreign workers:

	Quartiere	1	2	3	4	5
28	Ortomercato	Italian	Pizza	Japanese	Empanada	Ethiopian
38	Quintosole	Italian	French	Empanada	Ethiopian	Falafel
62	QT 8	Italian	Restaurant	Japanese	Empanada	Ethiopian
69	Cascina Triulza - Expo	Italian	French	Empanada	Ethiopian	Falafel
76	Bovisasca	Italian	Restaurant	Japanese	Empanada	Ethiopian

As for the relation between the neighborhoods' population density and restaurant density, my study shows there is no clear correlation between them:



By looking at the projected neighborhood population (2025) though, the study shows there is a clear opportunity for investing in the suburban areas, which will show the most increase in high-spending population with the lowest restaurant density (meaning less competition for the new venue). The choropleth below shows the projected population density per each neighborhood in 2025 and the current restaurant locations:



Based upon the results of the clustering algorithm presented above, I would then recommend establishing an **Ethnic** restaurant in a **Suburban** neighborhood.

6. Conclusion

In this study, I tried to determine the best location for establishing a new restaurant venue in Milan. By leveraging the crowdsourced data from Foursquare and the data from the City of Milan's Open Data program, I was able to group the different neighborhoods by the most frequent category of restaurants in them. I also looked at the relation between a neighborhood's population and restaurant count, finding opportunities to fulfill a projected restaurant demand.