

APPLIED DATA SCIENCE CAPSTONE PROJECT

Opening a New Japanese Restaurant in Milan, Italy



Table of Contents

- Introduction: The Business Problem
- The Dataset
- Methodology
- Analysis
- Results and Discussion
- Conclusion

Introduction: The Business Problem

Writers, artists, travelers, and explorers have extolled the beauty of Italy and its alluring lifestyle for centuries. Located in Northern Italy, the **city of Milan**, in particular, is for sure one of the world capitals of fashion, design, and stunning food and beverage experiences. Sipping a lively cup of espresso in a caffetteria looking out on Piazza Duomo or enjoying a reviving plate of risotto alla milanese with a scented glass of Chianti, one of the easiest ways to come to terms with this charming city is to spend some time in a local restaurant or café.

FoodExperiences Inc., a fictitious American corporation that owns and manages a number of restaurants around the world, knows this very well and it is in fact planning to open one of its restaurants in Milan. Thanks to extensive Market Research, it has already established that an innovative, high-quality and premium-price Japanese restaurant concept would gain a lot of traction in the city and be, therefore, a great business opportunity. As FoodExperiences Inc. makes rapid progress with the development of its business plan, it's now time to start identifying potentially good locations for the restaurant. Picking a location in the city center, a guarantee of huge foot traffic, seems to be a no-brainer, but at the same time, of course, the management team does not want to choose one where other Japanese and/or Asian restaurants are within reach.

In order to address this key business problem, FoodExperiences Inc.'s mandates **DataScienceWizards srl**, a fictitious Italian data science boutique that specializes in addressing and solving key business questions through data and location data, to:

 conduct a quick, preliminary analysis of the current food and restaurant offering in Milan; and 2. provide FoodExperiences Inc. with a first recommendation of good restaurant location neighborhoods in the city center.

The Dataset

As we said, DataScienceWizards srl offers analytics services that exactly cater to the needs of small businesses, or larger companies like FoodExperiences Inc., who are faced with business questions that can be answered, at least partly, by exploring and analyzing location data and performing geospatial intelligence. The **Foursquare API** is a gold standard for this task. Querying the Foursquare API allows Data Scientists to get a rich description of locations and venues that include: the name of the venue, the geospatial coordinates of the venue, the full physical address of the venue, the venue's distance for the set point, etc.

The Team at DataScienceWizards srl immediately gets down to work. It gets the most recent copy of a <a href="https://high.com/hig

Finally, the neighborhoods are clustered by using unsupervised Machine Learning (K-Means) in order to identify the neighborhoods that offer similar food experiences. Clusters are analyzed to determine their differentiating features and a narrowed-down list of Milan center neighborhoods, that best match the business requirements, are recommended to FoodExperiences Inc.

THE TRENDIEST NEIGHBORHOODS IN THE CITY OF MILAN, ITALY

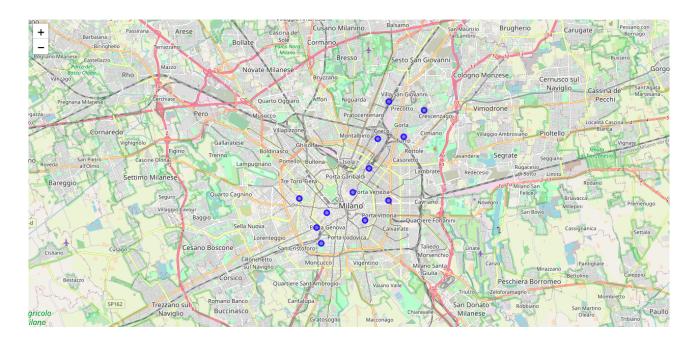
We kick the exercise off by importing the necessary Python libraries into a new Jupyter Notebook. These libraries are very common in the standard Data Scientist's tool set.

```
import pandas as pd
import requests
import json
lconda install -c conda-forge geopy --yes
from geopy.geocoders import Nominatim
from pandas.io.json import json_normalize
import matplotlib.cm as cm
import matplotlib.colors as colors
from sklearn.cluster import KMeans
lconda install -c conda-forge folium=0.5.0 --yes
import folium
%matplotlib inline
```

We then capture the trendiest neighborhoods in the center of Milan (based on information provided in lifestyle reports like <u>this one</u>) into a Pandas dataframe.

	Borough	Neighborhood	latitude	longitude
0	Milan City Center	20121	45.473098	9.191635
1	Milan City Center	20122	45.459848	9.199988
2	Milan City Center	20123	45.463481	9.174259
3	Milan City Center	20124	45.484534	9.202812
4	Milan City Center	20125	45.498444	9.208493
5	Milan City Center	20126	45.515986	9.216185
6	Milan City Center	20127	45.499361	9.226081
7	Milan City Center	20128	45.511964	9.239731
8	Milan City Center	20129	45.469305	9.215967
9	Milan City Center	20145	45.470361	9.155620
10	Milan City Center	20144	45.456529	9.167259
11	Milan City Center	20143	45.449025	9.170601

With **Folium**, we can easily visualize Milan center and the neighborhoods that we have in our dataframe. For the center of Milan, we use the geospatial coordinates of "Piazza Duomo", one of the key symbols of the city.



Next, we set up to use the **Foursquare API** to start exploring the restaurants that we have in these neighborhoods. The Foursquare API is a gold standard for a task like this.

Querying the Foursquare API allows Data Scientists to get a rich description of locations and venues that include: the name of the venue, the geospatial

coordinates of the venue, the full physical address of the venue, the venue's distance for the set point, etc. To use the Foursquare API, we need to register ourselves to the web service and get the necessary Client ID, Client Secret, and API Version. These will be passed to the web service every time we make an API call.

To get a sense of what we can get, we explore the first neighborhood in Milan. We query the Foursquare API to get the nearby "food" venues (using Foursquare's categoryID '4d4b7105d754a06374d81259' according to Foursquares API documentation), get a JSON response, and structure it into a Pandas dataframe:

	name	categories	lat	Ing
0	Nobu	Japanese Restaurant	45.470974	9.192958
1	Ristorante Yazawa	Shabu-Shabu Restaurant	45.475806	9.189566
2	Bice	Italian Restaurant	45.470152	9.194140
3	22	Bistro	45.474928	9.193852
4	Antica Osteria Stendhal	Italian Restaurant	45.473978	9.187678

But how many real restaurants do we have in this neighborhood? Coffee shops, pizza places, bakeries, etc. are not direct competitors, so we don't care about those. Let's then pick the venues that actually have the string 'Restaurant' in the 'categories' column.



We can also easily get the total number of restaurants in the neighborhood with a single line of code:

```
print('This neighborhood has {} restaurants.'.format(nearby_restaurants.shape[0]))
This neighborhood has 32 restaurants.
```

We write a function in order to be able to get the restaurants in all the neighboroods and we leverage it to capture all the food venues in the selected neighborhoods into the Pandas dataframe 'milan_restaurants'.

With the following line of code, we capture all the food venues in the selected neighborhoods into the Pandas dataframe 'milan_restaurants'. milan_restaurants = getNearbyRestaurants(names=df_milan['Neighborhood'],latitudes=df_milan['latitude'],longitudes=df_milan['longitude']) print(milan_restaurants.shape) milan_restaurants.head() (288, 7)Neighborhood Neighborhood Latitude Neighborhood Longitude Venue Venue Latitude Venue Longitude **Venue Category** 0 20121 45.473098 9.191635 Nobu 45.470974 9.192958 Japanese Restaurant 9.191635 1 20121 45.473098 Ristorante Yazawa 45.475806 9.189566 Shabu-Shabu Restaurant 20121 45.473098 9.191635 45.470152 9.194140 20121 45.473098 9.191635 45.474928 9.193852 20121 45.473098 9.191635 Antica Osteria Stendhal 45.473978 9.187678 Italian Restaurant

Once again, though, we are interested in the venues that are real restaurants, excluding all the others.



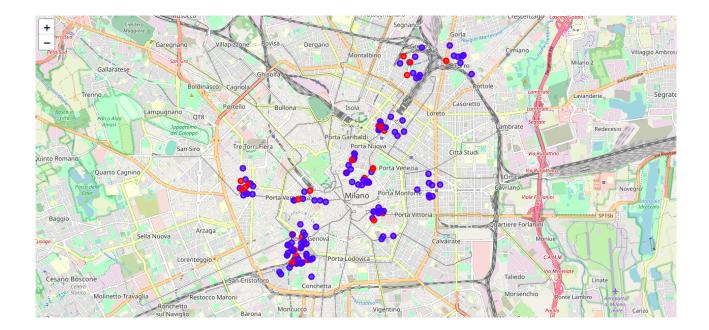
We want to get an idea of how many restaurants we have, by restaurant category. All we have to do is to call the 'value_counts()' method on the 'Venue Category' feature in the our dataframe.

```
milan_restaurants['Venue Category'].value_counts()
Italian Restaurant
                                  20
Restaurant
Seafood Restaurant
                                  12
Japanese Restaurant
                                 11
Chinese Restaurant
                                  6
Sushi Restaurant
Spanish Restaurant
Vegetarian / Vegan Restaurant
Asian Restaurant
Campanian Restaurant
                                  2
Mediterranean Restaurant
                                  2
Empanada Restaurant
Dim Sum Restaurant
Kebab Restaurant
Sardinian Restaurant
                                  1
Southern / Soul Food Restaurant
Shabu-Shabu Restaurant
Filipino Restaurant
Tuscan Restaurant
                                  1
Thai Restaurant
French Restaurant
Greek Restaurant
Abruzzo Restaurant
Fast Food Restaurant
Indian Restaurant
American Restaurant
Name: Venue Category, dtype: int64
print('Our focus area in Milan center has {} restaurants in total.'.format(milan_restaurants.shape[0]))
Our focus area in Milan center has 153 restaurants in total.
```

The first observation that we can make is that, not surprisingly, **the Italian food represents the dominant food experience** in our focus area.

However, there are some restaurants offering Japanese or Asian food in general. We process the data a bit further and determine that **our focus area** in Milan has 30 Oriental restaurants, out of the 153 total restaurants (about 20%).

One good way to get a visual clue of the situation is to display on a map all the collected restaurants in our area of interest. We also show the Oriental restaurants in different color (red) to better appreciate their density.



This all looks good. To recap, we now have all the restaurants in the best neighborhoods in Milan center and we have a sense of how many Oriental food ones are there, and where exactly they are located across the neighborhoods.

This concludes the data gathering and preliminary exploration phase. We're now ready to use this data to perform further analysis in order to come up with ideal locations for the FoodExperiences Inc.'s Japanese restaurant in Milan.

Methodology

In this project we are directing our efforts on understanding the food offering in the trendiest neighborhoods in Milan center, particularly the offering of Oriental food as opposed to non-Oriental food experiences.

In the first part of the exercise, we have collected the required data, location and type (category) of every restaurant of the neighborhoods in Milan center ("Piazza Duomo"). We have also specifically identified the Oriental food restaurants as these might be seen as competitors of the restaurant that FoodExperiences Inc. wants to open in Milan.

In the second and last part, we will use **Machine Learning (K-Means Clustering)** in order to identify neighborhoods / addresses which should be a starting point for a final 'street-level' exploration and search for the optimal venue location by the management team of our Client, FoodExperiences Inc.

Analysis

We start our in-depth analysis by checking how many venues (restaurants) were returned for each neighborhood.

milan_restaurants.groupby('Neighborhood').count()								
	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category		
Neighborhood								
20121	23	23	23	23	23	23		
20122	14	14	14	14	14	14		
20123	11	11	11	11	11	11		
20124	17	17	17	17	17	17		
20125	11	11	11	11	11	11		
20126	5	5	5	5	5	5		
20127	8	8	8	8	8	8		
20128	1	1	1	1	1	1		
20129	8	8	8	8	8	8		
20143	21	21	21	21	21	21		
20144	19	19	19	19	19	19		
20145	15	15	15	15	15	15		

What this quick picture tells us is that there are neighborhoods with fewer restaurants in general, like '20125', '20126', '20127', or '20128'.

For each of the eleven neighborhoods, we want to establish the relative frequency of each restaurant type, so we first produce a dataframe like this.

	Neighborhood	Abruzzo Restaurant	American Restaurant	Asian Restaurant	Campanian Restaurant	Chinese Restaurant	Dim Sum Restaurant	Empanada Restaurant	Fast Food Restaurant	Filipino Restaurant	 Restaurant	Sardinian Restaurant	Seafood Restaurant	Shabu- Shabu Restaurant	Southern / Soul Food Restaurant	Re
0	20121	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000	0.000000	 0.217391	0.000000	0.043478	0.043478	0.000000	(
1	20122	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000	0.000000	 0.285714	0.071429	0.000000	0.000000	0.000000	
2	20123	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000	0.000000	 0.181818	0.000000	0.000000	0.000000	0.000000	(
3	20124	0.000000	0.0	0.058824	0.000000	0.117647	0.058824	0.000000	0.000	0.000000	 0.058824	0.000000	0.000000	0.000000	0.000000	(
4	20125	0.000000	0.0	0.000000	0.000000	0.181818	0.000000	0.000000	0.000	0.000000	 0.272727	0.000000	0.181818	0.000000	0.000000	(
5	20126	0.000000	0.2	0.000000	0.000000	0.000000	0.000000	0.000000	0.000	0.000000	 0.200000	0.000000	0.000000	0.000000	0.000000	(
6	20127	0.000000	0.0	0.000000	0.000000	0.125000	0.000000	0.000000	0.125	0.000000	 0.000000	0.000000	0.000000	0.000000	0.000000	(
7	20128	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000	0.000000	 0.000000	0.000000	0.000000	0.000000	0.000000	(
8	20129	0.000000	0.0	0.000000	0.125000	0.000000	0.000000	0.000000	0.000	0.000000	 0.375000	0.000000	0.125000	0.000000	0.000000	(
9	20143	0.047619	0.0	0.000000	0.000000	0.047619	0.000000	0.047619	0.000	0.000000	 0.000000	0.000000	0.047619	0.000000	0.047619	(
10	20144	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000	0.052632	 0.000000	0.000000	0.368421	0.000000	0.000000	(
11	20145	0.000000	0.0	0.133333	0.066667	0.066667	0.000000	0.000000	0.000	0.000000	 0.066667	0.000000	0.000000	0.000000	0.000000	(
12 ro	ws × 27 columr	ns														

Then we print each neighborhood along with the top 3 most common restaurant types.

```
----20121----
              venue freq
0 Italian Restaurant 0.43
1
    Restaurant 0.22
2 Japanese Restaurant 0.17
----20122----
               venue frea
0 Italian Restaurant 0.36
   Restaurant 0.29
2 Japanese Restaurant 0.14
----20123----
              venue frea
0 Italian Restaurant 0.55
1 Sushi Restaurant 0.18
     Restaurant 0.18
----20124----
              venue freq
0 Italian Restaurant 0.59
  Chinese Restaurant 0.12
2 Japanese Restaurant 0.06
----20125----
             venue freq
venue freq
0 Restaurant 0.27
1 Chinese Restaurant 0.18
2 Seafood Restaurant 0.18
```

Now let's create a new dataframe and display the top 5 restaurants (i.e. most common) for each neighborhood:

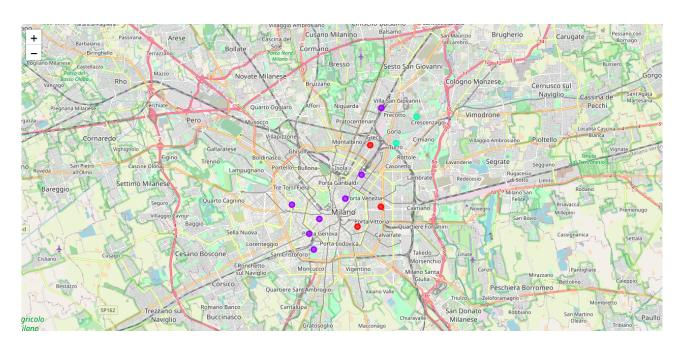


We run K-Means to cluster the neighborhoods into 3 clusters and generate a "merged dataframe" that, along with the details of the selected 12 neighborhoods, shows the cluster that each neighborhood belongs in as well.

K-Means (MacQueen, 1967) is one of the simplest unsupervised learning algorithms that solve clustering problems. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed *a priori*.

	Borough	Neighborhood	latitude	longitude	Cluster Labels	1st Most Common Restaurant	2nd Most Common Restaurant	3rd Most Common Restaurant	4th Most Common Restaurant	5th Most Common Restaurant
0	Milan City Center	20121	45.473098	9.191635	1	Italian Restaurant	Restaurant	Japanese Restaurant	French Restaurant	Sushi Restaurant
1	Milan City Center	20122	45.459848	9.199988	0	Italian Restaurant	Restaurant	Japanese Restaurant	Sardinian Restaurant	Vegetarian / Vegan Restaurant
2	Milan City Center	20123	45.463481	9.174259	1	Italian Restaurant	Sushi Restaurant	Restaurant	Spanish Restaurant	Greek Restaurant
3	Milan City Center	20124	45.484534	9.202812	1	Italian Restaurant	Chinese Restaurant	Asian Restaurant	Spanish Restaurant	Dim Sum Restaurant
4	Milan City Center	20125	45.498444	9.208493	0	Restaurant	Italian Restaurant	Chinese Restaurant	Seafood Restaurant	Greek Restaurant
5	Milan City Center	20126	45.515986	9.216185	1	Italian Restaurant	American Restaurant	Restaurant	Kebab Restaurant	Indian Restaurant
6	Milan City Center	20127	45.499361	9.226081	2	Italian Restaurant	Chinese Restaurant	Fast Food Restaurant	Tuscan Restaurant	American Restaurant
7	Milan City Center	20128	45.511964	9.239731	2	Italian Restaurant	Tuscan Restaurant	American Restaurant	Asian Restaurant	Campanian Restaurant
8	Milan City Center	20129	45.469305	9.215967	0	Italian Restaurant	Restaurant	Campanian Restaurant	Seafood Restaurant	Indian Restaurant
9	Milan City Center	20145	45.470361	9.155620	1	Italian Restaurant	Asian Restaurant	Restaurant	Campanian Restaurant	Chinese Restaurant
10	Milan City Center	20144	45.456529	9.167259	1	Italian Restaurant	Seafood Restaurant	Japanese Restaurant	Sushi Restaurant	Filipino Restaurant
11	Milan City Center	20143	45.449025	9.170601	1	Italian Restaurant	Mediterranean Restaurant	Chinese Restaurant	Empanada Restaurant	Indian Restaurant

The final step is to display the resulting clusters on a map, which will allow us to derive some insights.



We have already maintained that the dominant food experience in this part of Milan is the Italian one, which of course is not a surprise.

However, we can now examine each cluster of restaurants and label it based on the predominant "food experience" after the Italian one.



These are the Cluster Labels that we can associated to each cluster:

- Cluster 0: "Blending Italian with other food experiences and suggestions"
- Cluster 1: "Young and easy food experiences"
- Cluster 2: "Dominant Italian and Oriental food experiences"

Results and Discussion

Our work has shown that **Milan center** has a lot to offer to those who are planning to dine out, with Italian restaurants being the mainstream offering (with 70 venues out of a total of 155 - a 45% incidence), but where Oriental restaurants in general represent close to 20% of the offering. Venues are not distributed evenly across the neighborhoods and there are some where, as a whole, the restaurant density is lower, like in '20125', '20126', '20127', 20128', or '20129' as already noted above.

We also processed the data in such a way to reveal, for each targeted neighborhood, the top 5 most present restaurant categories, thus spotting the ones where Oriental food is a significant part of the offering. There are in fact neighborhoods where this is quite apparent and, as such, all other factors

being equal, these would not be ideal candidates for starting up a new Japanese restaurant.

Next, we used k-means to cluster the neighborhoods into 3 clusters in order to establish their predominant "food color(s)". This allowed us to identify the neighborhoods for "Dominant Italian and Oriental food experiences" (Cluster 2), therefore not a good pool of location candidates, as well as areas where the predominant food experience is non-Oriental (Cluster 0 and Cluster 1).

Based on the combined evidences of the analysis, DataScienceWizards srl would recommend neighborhood '20129' as the ideal location candidate, followed by '20126' and '20127'. The differential advantage of '20129' is that it is closer to Milan center ("Piazza Duomo") and at a walking distance form it.

This, of course, does not automatically imply that the neighborhoods '2019', '20126', and '20127' are indeed optimal locations for the new Japanese restaurant that FoodExperiences Inc. wants to establish. It is entirely possible that good reasons are there for fewer restaurants in these areas, and fewer Oriental restaurants in particular, reasons which would make them unsuitable for a new restaurant regardless of the apparently reduced competition in the area.

These recommendations are only a starting point and will require more indepth analysis and evaluations.

Conclusion

The goal of this project was to analyze the **trendiest neighborhoods**, **for lifestyle and night life**, **in the city center of Milan** in order to better understand the food experiences offered there (particularly Oriental as opposed to Italian) and thus help the management team of FoodExperiences Inc. narrow down the search for an optimal location for a new Japanese restaurant in Milan.

By performing an extensive location data analysis, made possible by the Foursquare API, DataScienceWizards srl was able to come up with a recommended neighborhood, '20129', which satisfies a number of basic requirements for the location selection. Neighborhoods '20126' and '20127' were also identified as 'second-' and 'third-best' recommendations.

The final decision of the optimal restaurant location will be made by FoodExperiences Inc.'s management team based on other features of the locations in the three recommended neighborhoods, taking into consideration additional elements like attractiveness of each location, level of noise,

proximity to major roads, real estate availability, prices, social and economic factors, and the like.