Italy a place of great food

SEARCHING FOR AUTHENTIC TASTES

Minu M J | IBM Capstone Project

Introduction

No matter which country you are from, there is an Italian restaurant near you. The popularity of Italian cuisine can be attributed to the fact that they are very approachable and delicious. There is a large variety of dishes to choose from and what makes it so special is that these are made with very few ingredients making it a common mans food. The popularity of Italian cuisine has had many tourists visit Italy just to get a taste of the local food. In this project we explore the various neighborhoods in Italy and the best places to open a new Italian restaurant.

BUSINESS PROBLEM

In a place famous for its cuisine, opening a new restaurant and attracting customers can be tricky.

- One need to make sure that the place they choose to open the restaurant is not over crowded with similar eateries making competition intense.

TARGET AUDIENCE

This is intended for anyone who wishes to open authentic Italian restaurant. It can be an investor looking to invest in Italy or it can be multinational hotel chain looking to increase their customer base.

Data

To find a solution using data science methodology we need the following information:

- A list of all the regions/neighborhoods in Italy.
- Latitude and longitude values of each region.
- Data on different venues in each region.

DATA SOURCES:

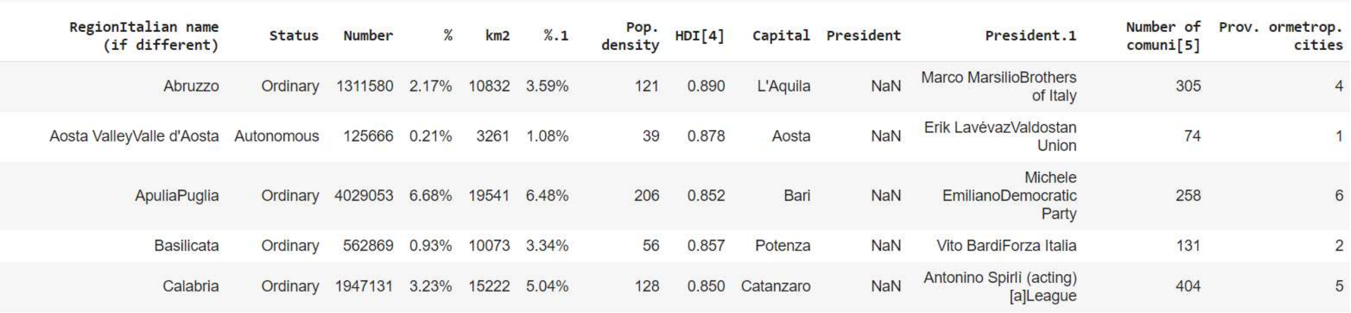
A list of all regions in Italy can be found from this Wikipedia page (https://en.wikipedia.org/wiki/Regions_of_Italy). We use web scraping and python packages to extract data from the page. It gives us a list of all the 21 regions in Italy. Python geocoder package is used to extract the latitude and longitude details of each region. We then use FourSquare AI to get a list of all the venues within a predefined radius of each region. This data is then used to get a list of all the Italian Restaurants in a region with which we shall apply further machine learning techniques to cluster neighborhoods.

Methodology

Our aim is to get a list of all the Italian restaurants in each region of Italy. In order to obtain the data we first get a list of all the regions from the Wikipedia page. The data is extracted using web scraping with python pandas package. Once we get a list of regions we need to obtain the latitude and longitude values to allow us to plot each region on a map. This is where we use pythons geocoder package to get the latitude and longitude values for each region. Once we get this information we can now create a pandas dataframe for further analysis. In order to visualize any regional data, the most appropriate method is to use a map and for that we use Folium package from python. Next we need to list all the venues in different regions. This is accomplished by calling the Foursquare API. The final step is to cluster neighborhoods based on the number of Italian restaurants in the region. This will give us a better idea on the spread of competition.

Result

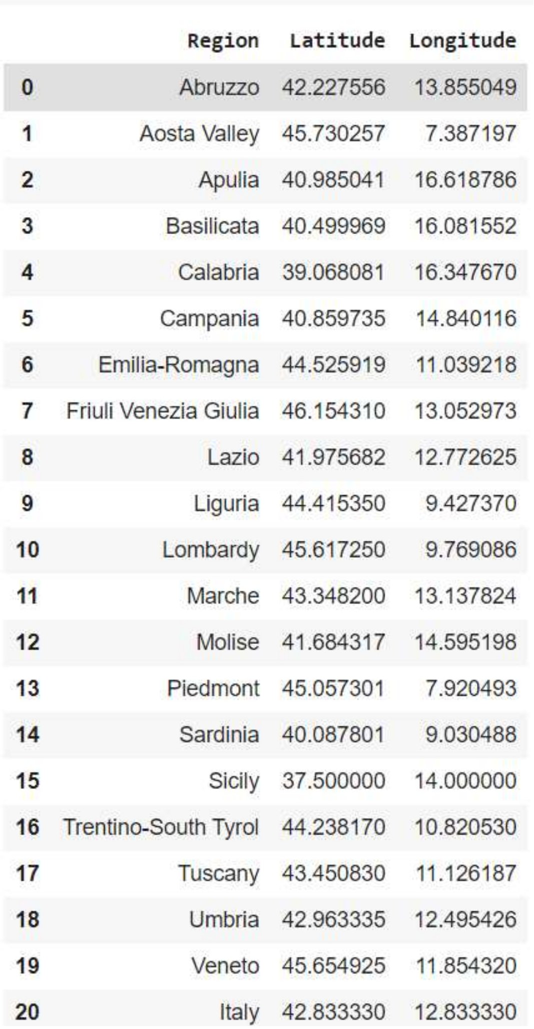
After web scraping the data we obtained is:



RegionItalian name (if different)	Status	Number	%	km2	%.1	Pop. density	HDI[4]	Capital	President	President.1	Number of comuni[5]	Prov. ormetrop. cities
Abruzzo	Ordinary	1311580	2.17%	10832	3.59%	121	0.890	L'Aquila	NaN	Marco MarsilioBrothers of Italy	305	4
Aosta ValleyValle d'Aosta	Autonomous	125666	0.21%	3261	1.08%	39	0.878	Aosta	NaN	Erik LavévazValdostan Union	74	1
ApuliaPuglia	Ordinary	4029053	6.68%	19541	6.48%	206	0.852	Bari	NaN	Michele EmilianoDemocratic Party	258	6
Basilicata	Ordinary	562869	0.93%	10073	3.34%	56	0.857	Potenza	NaN	Vito BardiForza Italia	131	2
Calabria	Ordinary	1947131	3.23%	15222	5.04%	128	0.850	Catanzaro	NaN	Antonino Spirlì (acting) [a]League	404	5

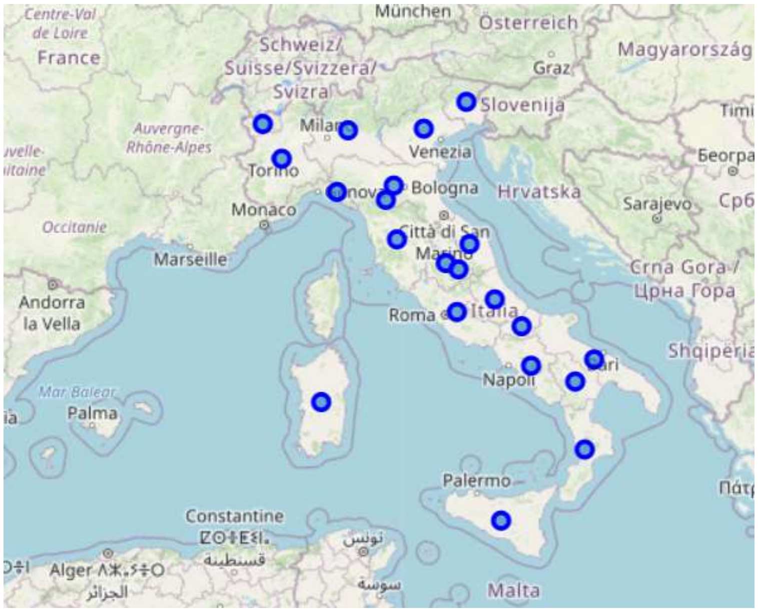
It can be observed that the region column contains region names in both English and Italian, also a lot of the data is not useful for our project. So we cleanup the data and then invoke geocoder function to find latitude and longitude for each region. This is then stored in a dataframe. It can be seen that there are 21 regions in Italy.

The cleaned data is as follows:

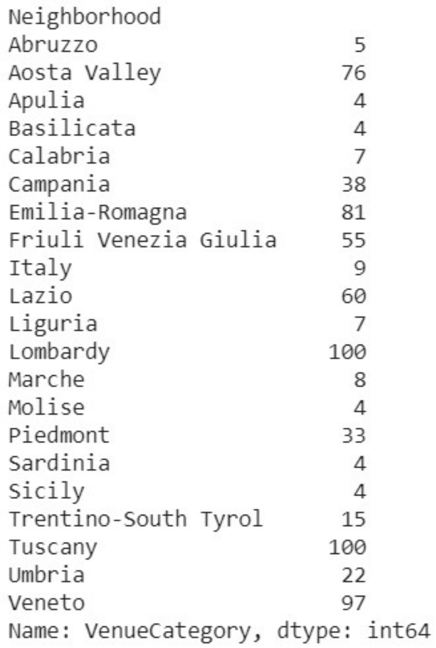


	Region	Latitude	Longitude
0	Abruzzo	42.227556	13.855049
1	Aosta Valley	45.730257	7.387197
2	Apulia	40.985041	16.618786
3	Basilicata	40.499969	16.081552
4	Calabria	39.068081	16.347670
5	Campania	40.859735	14.840116
6	Emilia-Romagna	44.525919	11.039218
7	Friuli Venezia Giulia	46.154310	13.052973
8	Lazio	41.975682	12.772625
9	Liguria	44.415350	9.427370
10	Lombardy	45.617250	9.769086
11	Marche	43.348200	13.137824
12	Molise	41.684317	14.595198
13	Piedmont	45.057301	7.920493
14	Sardinia	40.087801	9.030488
15	Sicily	37.500000	14.000000
16	Trentino-South Tyrol	44.238170	10.820530
17	Tuscany	43.450830	11.126187
18	Umbria	42.963335	12.495426
19	Veneto	45.654925	11.854320
20	Italy	42.833330	12.833330

Let us now plot these on the map.

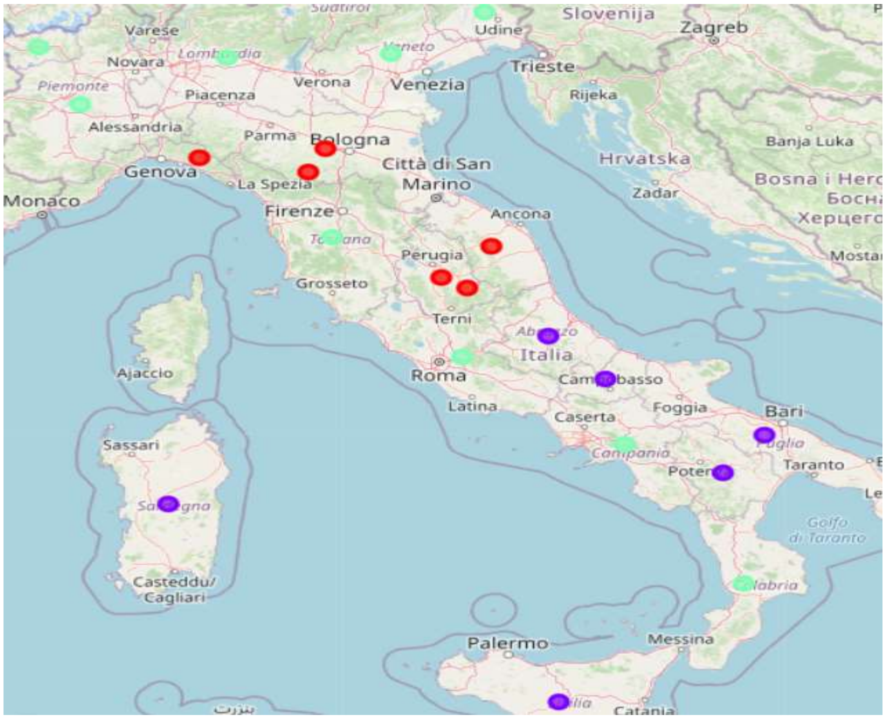


We now use Foursquare API to get a list of venues in each region. Grouping the 733 unique venues by region we get:

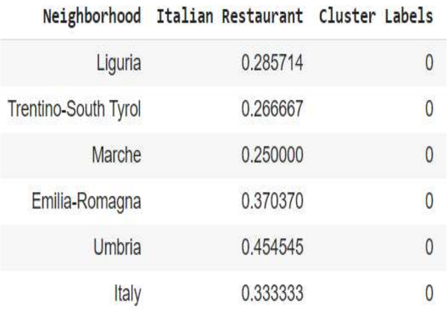


```
Neighborhood
Abruzzo                   5
Aosta Valley             76
Apulia                    4
Basilicata                4
Calabria                  7
Campania                 38
Emilia-Romagna           81
Friuli Venezia Giulia    55
Italy                     9
Lazio                    60
Liguria                   7
Lombardy                100
Marche                    8
Molise                    4
Piedmont                 33
Sardinia                  4
Sicily                    4
Trentino-South Tyrol     15
Tuscany                 100
Umbria                   22
Veneto                   97
Name: VenueCategory, dtype: int64
```

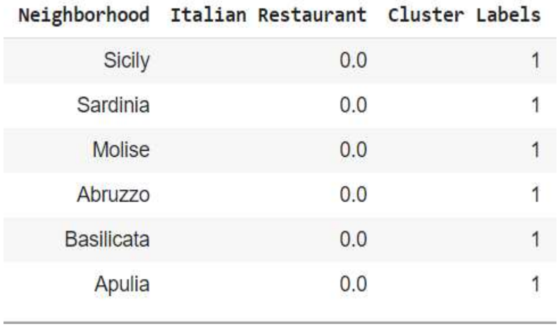
We now apply k means clustering to the data and divide the regions into three clusters. The result is:



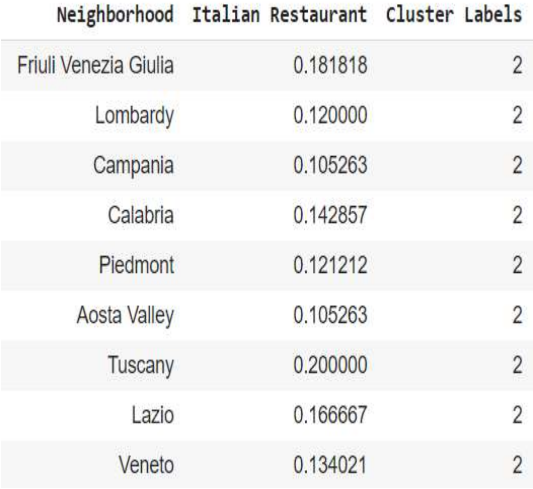
The regions in each of the clusters are:



Neighborhood	Italian Restaurant	Cluster Labels
Liguria	0.285714	0
Trentino-South Tyrol	0.266667	0
Marche	0.250000	0
Emilia-Romagna	0.370370	0
Umbria	0.454545	0
Italy	0.333333	0



Neighborhood	Italian Restaurant	Cluster Labels
Sicily	0.0	1
Sardinia	0.0	1
Molise	0.0	1
Abruzzo	0.0	1
Basilicata	0.0	1
Apulia	0.0	1



Neighborhood	Italian Restaurant	Cluster Labels
Friuli Venezia Giulia	0.181818	2
Lombardy	0.120000	2
Campania	0.105263	2
Calabria	0.142857	2
Piedmont	0.121212	2
Aosta Valley	0.105263	2
Tuscany	0.200000	2
Lazio	0.166667	2
Veneto	0.134021	2

Discussion

It can be observed from the above that Cluster 0 has highest number of Italian Restaurants, followed by cluster 2. Cluster 3 represents regions with almost no Italian restaurants. Hence Cluster 3 would be ideal location to begin Italian restaurant as the competition would be very less in those regions.

Acknowledgments:

Cover picture: https://unsplash.com/@inasestrjevitovschii