Best location for an Italian restaurant among top US cities

Final Project Assignment for IBM Data Science Course

by Ignacio de Juan
January 2021

Table of Contents

1. Introduction / Business Problem 3
2. Data 3
3. Methodology 4
4. Results 9
5. Discussion 9
6. Conclusions 10

1. Introduction / Business Problem

In this project we are going to undertake the mission of facilitating the decision making process of where is the best location to open a new restaurant of a certain cuisine type in a top city of the USA. In this case it will be Italian cuisine.

The decision of opening a new restaurant should take into consideration the competition, demand and likelihood of success, taking into account as much information available.

Through this exercise we will address the following questions:

- Which are the cities with a highest potential based on the level of concentration of these type of restaurants?
- Is the size of the local Italian descent community a variable that affects the number of Italian restaurants?
- What other factors should we take into account for such a decision?

The result of our data science work will be a framework to help narrow down the initial 100 cities to a target of 10-15 cities where we see the highest potential of an Italian restaurant to succeed.

With the aid of Wikipedia tables and the Foursquare API we will navigate through the data, looking for insights and arguments to help our final decision.

Lastly, we will point out the caveats and potential further developments of this exercise.

Target Audience

This analysis can be of interest to entrepreneurs, investors and corporates that are considering the set-up of a new business or the expansion of their current one. The provided framework can be useful to determine where to open the restaurant and what strategy should be followed.

2. Data

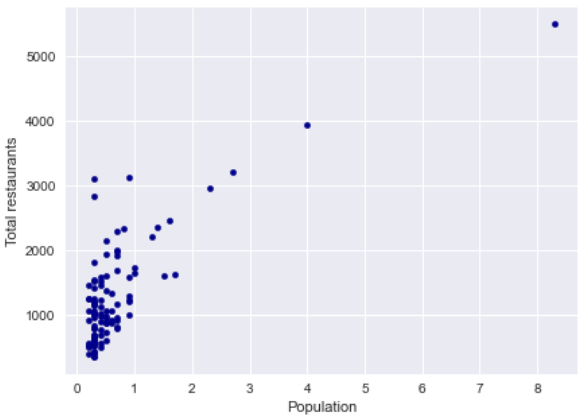
We will take data from top US cities, ranked by population. The reason for taking the largest cities is to achieve a sufficiently large dataset to draw valid conclusions. We will be using Wikipedia to identify the target cities of our study.

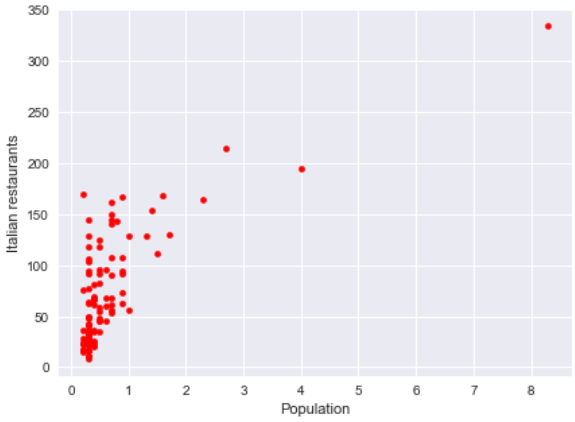
Foursquare will serve as our database for the restaurant data, retrieving the number of Italian restaurants and total restaurants, as well as the ratings for the Italian restaurants.

Additionally we are going to fetch from Wikipedia the statistics of how much of the population in each city comes from Italian descent. Since the information is not available at city level but only state level, we will take it as good enough for this exercise.

3. Methodology

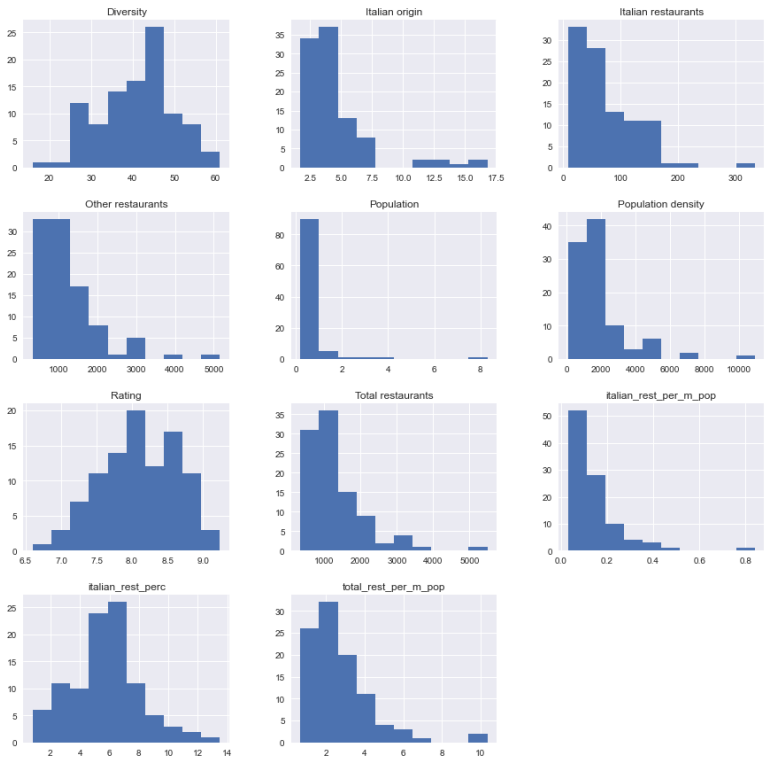
We will start by downloading the data, exploring it and plotting it to have a first approach of the problem. Then we will decide which models are more likely to help us to reach any conclusion and we will apply and test them.
With the data found in Foursquare we are able to calculate the number of Italian restaurants in each city. We can also see what the punctuation is given by the customers to these venues.





Another piece of data that we want to test for its relevance of our analysis is the amount of population from Italian descent. This information can be found in the Wikipedia at a state level, and we will use it despite not being as granular as the city level that we are using for our analysis in general.

The data exploration has been useful to detect outliers and clean the data, since New York and Los Angeles were too big to be compared with the rest, and there was the strange case of Jersey City, with a very high ratio of restaurants per population, and therefore we preferred to take it out of the study.



As part of this phase, we will use Python's powerful tools to get a quick overview of the correlation levels between the variables. We want to identify what are the variables that have a higher effect on the number of restaurants, and they are the following:

Irrelevant correlations (r < 0.3)

- Italian origin is correlated at 0.28 both with Italian restaurants as well as with Other restaurants. So this is clearly not a driver of Italian restaurants.

Weak correlations (0.3 < r < 0.5):

- Population density (0.36)
- Rating (0.36)

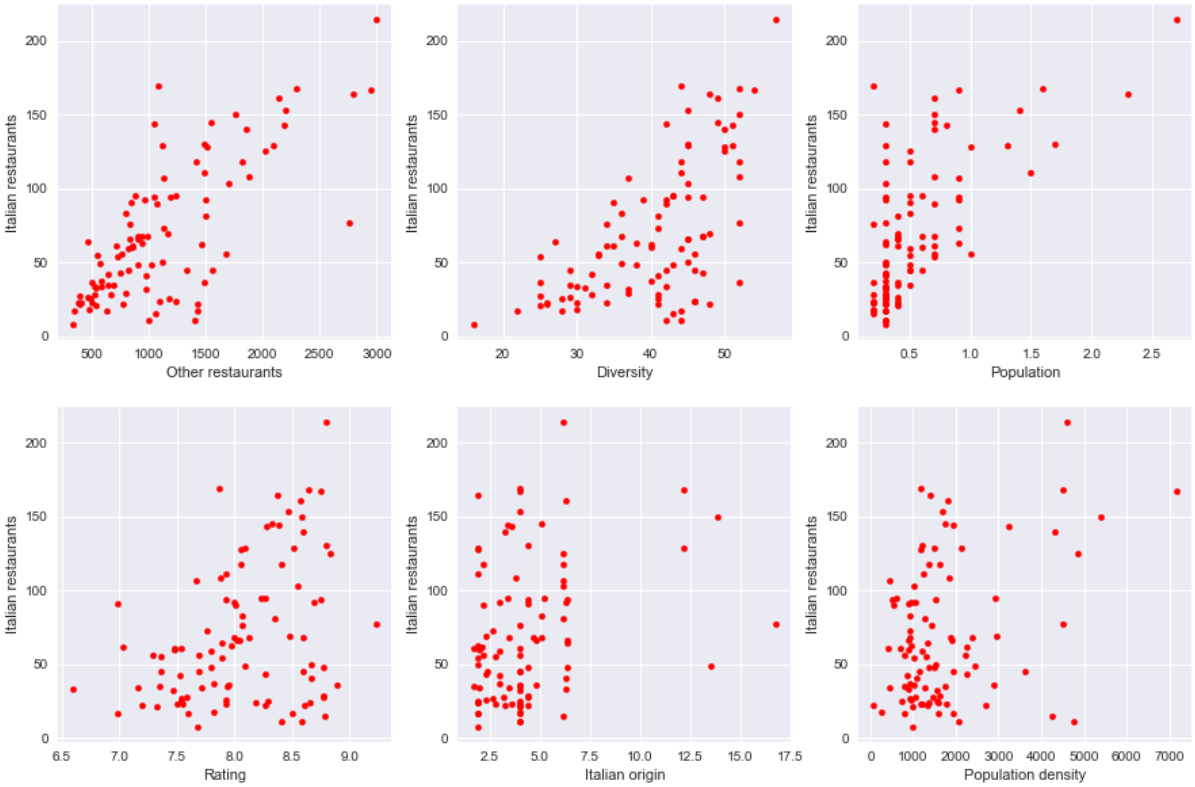
Moderate correlation (0.5 < r < 0.7):

- Diversity (0.64)
- Population (0.66)

Strong correlation (r > 0.7):

- Other restaurants (0.74)

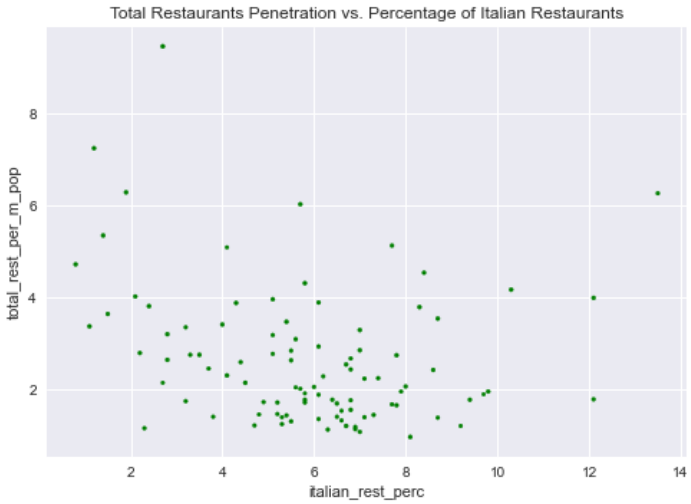
The chart below shows these relationships graphically:



From the scattered plots we deduct that Population and Other restaurants are the two features that have a higher correlation that seems to be linear. Diversity also has quite consistency and Rating and Italian origin do have less of a linear correlation.
The histograms show that in general the distributions are highly concentrated among the left side of the charts, which probably derives from the fact that most of the cities are concentrated below 1M inhabitants. The closest we see to a normal distribution can be found in the rating histogram as well as the italian_rest_perc histogram.

Level of market saturation

One useful calculation that we are going to make is the level of market saturation vs the percentage of Italian restaurants. This will give us an indication of the markets with potential growth.



The above chart shows on the Y axis the amount of restaurants per thousand inhabitants, and the X axis the percentage of Italian restaurants. The attractive market is where the Y axis is low, with low level of saturation of restaurants, and the X axis low as well, so there are less Italian restaurants and this means there is room for more.
Besides, we are going to rank the target space by ratings, so we can identify the cities where the ratings are lower, so it is going to be easier to face competition when we land with our new Italian restaurants.
We will start with the cities with less than 4 restaurants per 1000 inhabitants, and less or equal than 4% of Italian restaurants among those restaurants.

	City	State	Italian restaurants	total_rest_per_m_pop	italian_rest_perc	Rating
16	Stockton	California	17	2.13	2.7	6.98
2	Bakersfield	California	21	1.39	3.8	7.33
10	Riverside	California	32	3.34	3.2	7.47
27	Laredo	Texas	8	1.14	2.3	7.68
11	Sacramento	California	45	2.74	3.3	7.69
14	San Jose	California	56	1.73	3.2	7.69
53	Chesapeake	Virginia	18	2.44	3.7	7.82
18	Arlington	Texas	24	2.78	2.2	8.18
1	Anaheim	California	22	3.63	1.5	8.27
3	Chula Vista	California	11	3.36	1.1	8.58
7	Long Beach	California	45	3.19	2.8	8.60
43	Gilbert	Arizona	22	2.63	2.8	8.61
65	Henderson	Nevada	41	3.40	4.0	8.66
44	Glendale	Arizona	29	2.74	3.5	8.77
9	Oakland	California	36	3.80	2.4	8.89

We have a first list of 15 locations that seem attractive because there is still a low penetration of restaurants, and the percentage of Italian restaurants is also low. So we are looking for a good market opportunity in terms of size as well as differentiation opportunities versus the competition.

Next we are going to go one step further and model the data. We would like to make a prediction of how many restaurants should there be in those locations if they were following the "general rule" of the the

other cities. This will help us to confirm if there is a market opportunity, and rank the target cities by the largest gap of existing Italian restaurants vs. potential predicted Italian restaurants.

Modeling

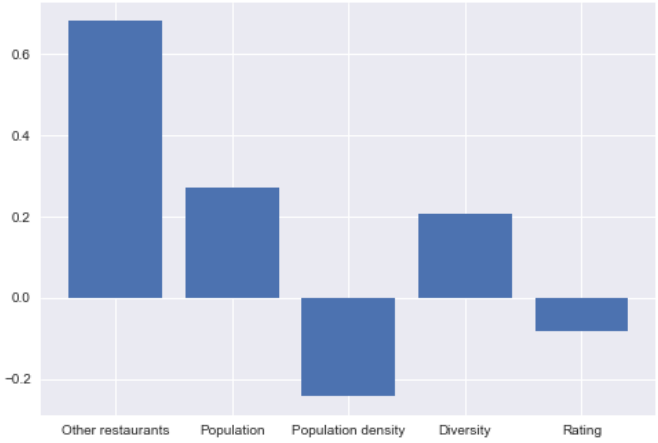
We have gone through our first approach to find the cities where the potential for a new Italian restaurant may be higher. Now, we are going to use a regression model to quantify what is the potential for each of these cities.

The results we have obtained are:

```
R-square : 0.32
```

And the model can be expressed with this equation:

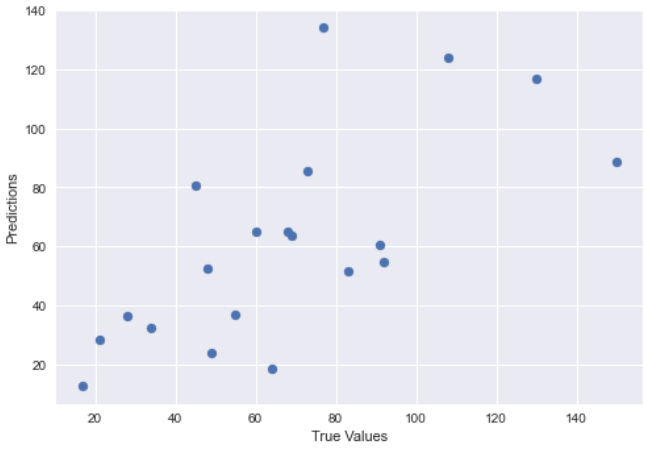
```
Italian restaurants = 0.053 * Number of other restaurants + 22.245 * Population (M) – 0.007
* Population density + 1.042 * Diversity – 6.596 * Rating
```




```
Feature: 0, Score: 0.67999 Other restaurants
Feature: 1, Score: 0.26997 Population
Feature: 2, Score: -0.24316 Population density
Feature: 3, Score: 0.20739 Diversity
Feature: 4, Score: -0.08421 Rating
```

In the above table we can see the importance of each of the variables (features) of the model. We can see that the amount of Other restaurants is the main driver to the number of Italian restaurants, followed by Population, Diversity and Population density. Rating is the feature of least importance, and it is negative, which means that the more Italian restaurants, the lower the ratings for them.

The below picture depicts the model predictions vs. the actual values:



4. Results

We have been able to gather all the information we needed from Foursquare and Wikipedia, and the data has shown to be quite consistent.
Looking at the ratios of amount of restaurants relative to population size, compared with the percentage of Italian restaurants, it looks as if there are some interesting cities to invest at first sight. Moreover, if we look at the ratings for Italian restaurants, we can have a hint that the competition may be weaker (it could also be that the population in different cities have different expectations).
Adding to the previous analysis, we have prepared a model to be capable of ranking the opportunities based on the gap between the "normal" amount of restaurants expected and the number in reality. Although the model has not shown to have a large R-square, it is nevertheless useful to rank opportunities.

5. Discussion

With the aid of data engineering, statistical analysis and machine learning we have created a framework that could serve as a basis of discussion with our investor. We have seen that despite the similarities in size, American cities show large variations of the amount of restaurants relative to its population. Most of the cities that have fulfilled our conditions to be targeted are in California (60% belong to this state, being 15% in the intial list), which could be a good thing so the investments can all be done on one area, but it could also show that California has other factors that make it not attractive. This is something that goes beyond the scope of this analysis.

6. Conclusions

Further research and caveats of this analysis

Caveats

- Foursquare data may be uneven in different cities even if we have chosen all cities in US
- Ethnic origin of population is at state level, not city level
- The population in the Wikipedia table may refer to a different geography than the geography applied by Foursquare in the queries

Further research

- Look at the evolution in time
- Analyse the preferences of the consumers also in terms of assessment of the existing restaurants
- Differentiate different level of budget in the restaurants
- Look at composition of the population and income in the city
- Zoom into the cities that have been selected to better understand the competition, both in geographical terms as well as other types of cuisines. We should make an analysis of the neighbourhoods to identify the best locations for our restaurants.
- Polinomial regression analysis: number or IT rest (y), number of rest, pop size, italian origin, average punctuation feedback

7. Source code

https://github.com/idejuan/data_science/blob/main/Final_Project_DSPC_w02.ipynb