Applied Data Science Capstone Project
Restaurant Recommender System

Rodrigo Bagdadi Benoliel

07/15/2020

1. Introduction

1.1 Background

More than ever, gastronomy is appreciated as one of main kinds of leisure available, especially in the big cities. Not only the quantity of restaurants, coffee shops, and other venues has increased, but as the global cultural interchange develops, its variety has become massive as well. A few generations ago, the idea of having african food on New York would be inconceivable, but nowadays, that had become something usual.

1.2 Problem

Although, the amount of offer has become such that people can even get disorientated by it. In this scenario, it would be very useful if costumers had a compass to point where to find venues of their taste, wherever they are. Given the situation, a recommendation system can be developed to help people find restaurants nearby that attend to their preferences, and be available as a mobile app.

1.3 Interest

Such a project would not only be interesting for the consumers, but for the venues as well, who will get publicity targeting costumers that will potentially enjoy their products. Having this interest in mind, being included in the system can be offered as a service to businesses on the gastronomy sector. Given its versatility and potentially low cost, such a service would be very attractive.

2. Data acquisition and cleaning

2.1 Data source

As for the system data, the Foursquare API will be used to locate and extract information on restaurants near the user, which will be assumed to be acquired by the device GPS. For this project, the user experience was manually acquired from a volunteer, but it's important to consider that, in the case of an actual business, the same data could be automatically gathered by Foursquare, under a premium account. The user manually collected data was acquired from the Foursquare social media, Google Maps, and Foursquare documentation.

2.2 Feature Selection and Modelling

The user volunteer lives in Barra da Tijuca, a neighborhood in the city of Rio de Janeiro, Brazil. Her coordinates were set as (-23.013493, -43.309671), as being the location where her house is at. That was chosen for a matter of simplicity, although for the system, any coordinates could have been given.

The user experience was a set of restaurants, with its respective rate, category, and coordinates. More specifically, it contained the: venue id, venue name, rate, category id, category name, latitude and longitude.

At first, there was no intention to gather the venue names, as it could be redundant considering the id was already known. But surprisingly, the venue ids available on the Foursquare social media were not always the same one returned by the API, so it turned out to be more accurate to use the venue name as key identifier instead. Both features were modelled as strings.

The venue rate was obviously important as well. To fit the "5-star" rating model, the user rate was set as an integer between 0 and 5.

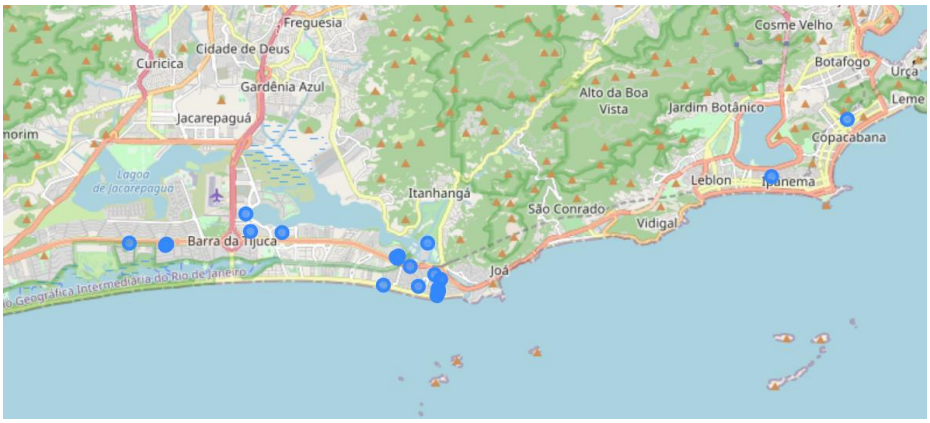
The category id, although not being as intuitive and comprehensible as the category name, was very useful as a filter while using the explore query. The explore query showed itself to return 100 venues on maximum, which is very few, especially considering most of those aren't even restaurants. As a countermeasure, for each category present in the user experience, the category id was passed as an argument to the API URL, filtering unwanted results, and therefore maximizing its relevancy. Still, the category name was kept for its advantages mentioned before. Both category names and ids are available at the Foursquare documentation, and were modelled as strings.

At last, the venue latitude and longitude were acquired on Google Maps, directed from the Foursquare social media, and set as float numerical values.

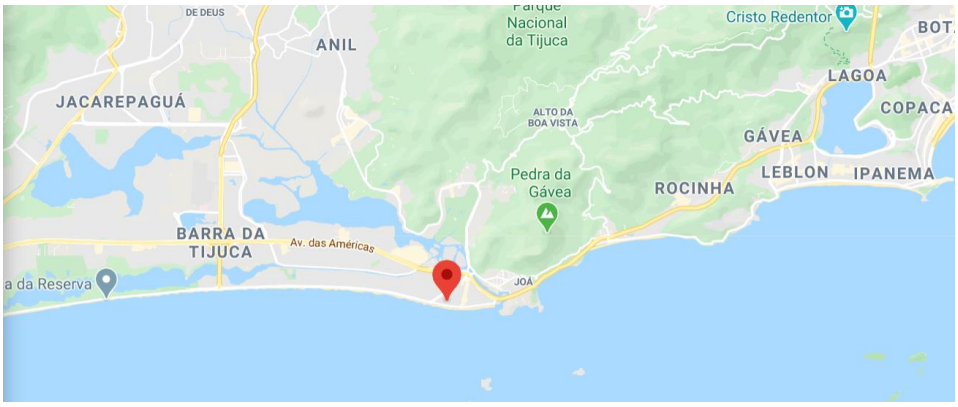
The venues' features returned by the explore query were filtered and modelled the same way as the user experience's.

3. Exploratory Data Analysis

Starting by examining the user experience, a few patterns can be identified. By looking at the coordinates of the restaurants known, a higher density area can be seen at the map center.

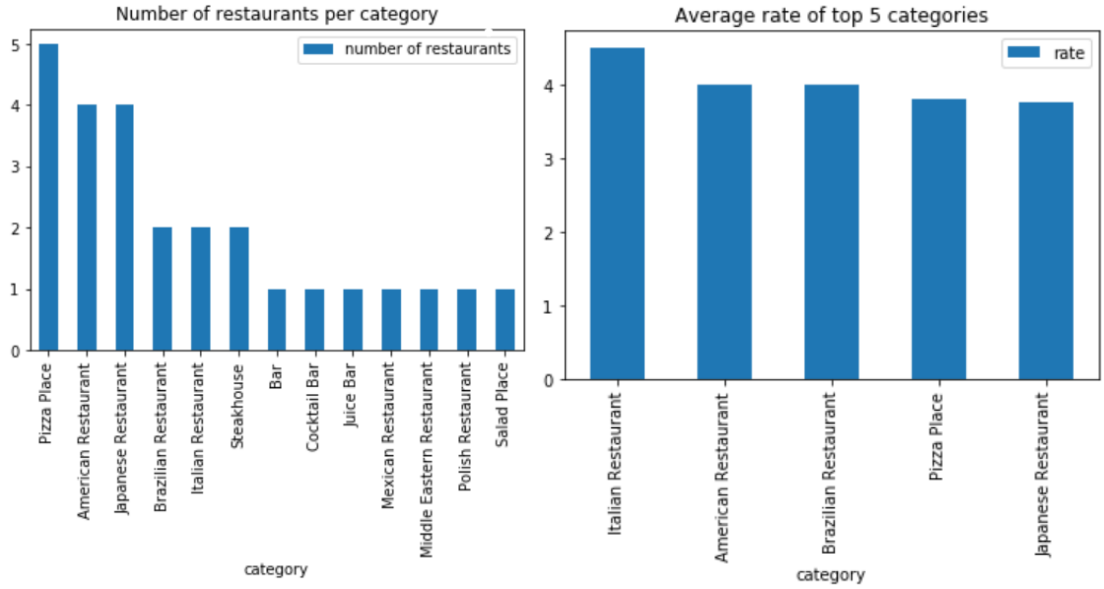


That higher density is no coincidence, by searching for the user coordinates on Google Maps, we see that that's the area where the user lives in.



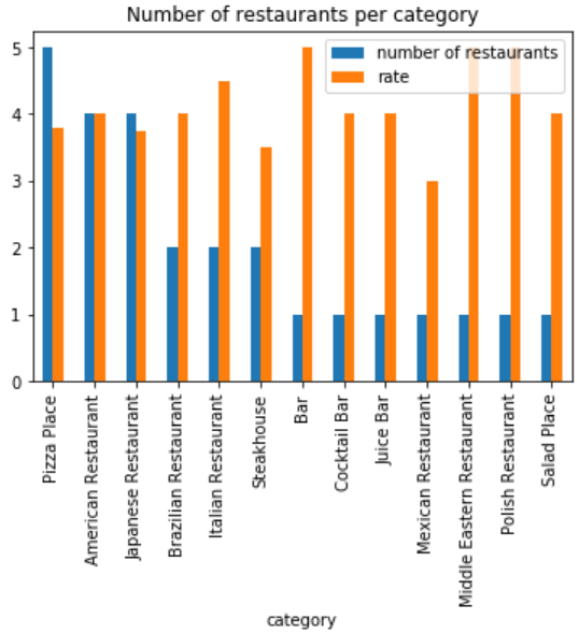
That indicates that, as expected, the user usually rather eat at restaurants closer to her home. Still, if she finds herself anywhere else, the same data can be used to find restaurants of her taste nearby.

Now, by analyzing the restaurant types, we can observe which restaurants the user frequents the most. By selecting the top five most frequent categories, and calculating its average rate, we can confirm that, generally, the user enjoys those categories.



Considering the highest rates and frequencies shown above, we can expect that the recommended restaurants should be from any of the top five categories. That is, if there are any venues of that kinds available nearby, of course.

Although, an interesting phenomenon can be seen if compare all categories and means are compared, as done bellow.



We can see that the rate deviation is higher between the less frequent categories. But it is no surprise that, with the lack of samples, the data turns biased. If the experience had closer number of restaurants for each category, we would expect the average rate to be more representative about the user's preference. On the other hand, the deviation on frequency per category is also a valuable information about her taste.

Besides, we cannot expect that the number of restaurants known by category depends only on the user's choice, because it also relies on their existing offer. Like so, there would be more pizzerias and american restaurants available than salad places or polish.

4. Recommendation System Modelling

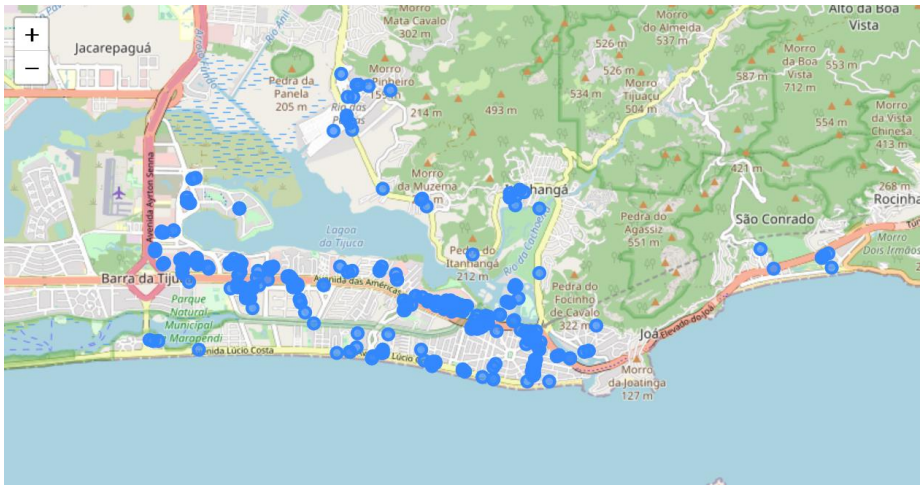
4.1 New Venues Gathering

The gathering of potential recommendations was done with the Foursquare explore query. For each query, a few parameters had to be specified on the API URL, which are explained bellow. Reminding that a new query were made for each venue category.

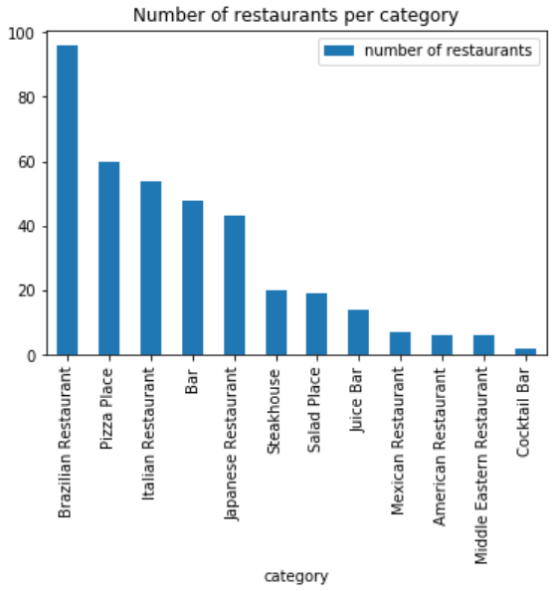
Client id	Client Secret	Version	Radius	Limit	Category	Latitude	Longitude
-	-	20180605	6000	100	Looped over each	-23.013457	-43,309648

After searching for each category, a new data frame was set, that included the venue id, name, category, latitude and longitude. Also, a new feature were added, calculated based on the venue coordinates, which is the distance. That measure wasn't in fact the distance, but is a representation of it, as being calculated as the absolute difference between the user and venue coordinates.

The location of acquired venues are displayed bellow.



It can be seen that there is a significant amount of restaurants at the user's neighborhood. The count of restaurants by categories is displayed as:



4.2 Data Preparation

For the recommendation system to be applied, the distance and category features were considered. By nature, the distance is numerical already, but the same can't be said about category, which had to be turned into it by one-hot-encoding. The restaurant name and id were dropped, the record of which row corresponded to each venue were made by the its index. The result is exampled bellow:

index	American Rest.	Pizza Place	Japanese Rest.	…	Bar	Juice Bar	Mexican Rest.	Italian Rest.	Distance
1	0	0	1	…	0	0	0	0	0.00828
2	0	0	0	…	0	0	0	1	0.01223
3	0	0	0	…	1	0	0	0	0.01194

The user experience was prepared similarly, but each value from the category columns were multiplied by the restaurant rate.

index	American Rest.	Pizza Place	Japanese Rest.	…	Bar	Juice Bar	Mexican Rest.	Italian Rest.
1	0	0	5	…	0	0	0	0
2	0	0	0	…	0	0	0	4
3	0	0	0	…	4	0	0	0

Then, the mean value for each category column was taken and added to an array, which was then normalized. Also a bias value was included as a weight for the distance feature. The bias value was chosen manually, as a way to balance the distance and category variables. That array is the user profile.

American Rest.	Pizza Place	Japanese Rest.	…	Bar	Juice Bar	Mexican Rest.	Italian Rest.	Distance Bias
0,8421	0,6573	0,5864		0,2685	0,2648	0,2567	0,3489	-120

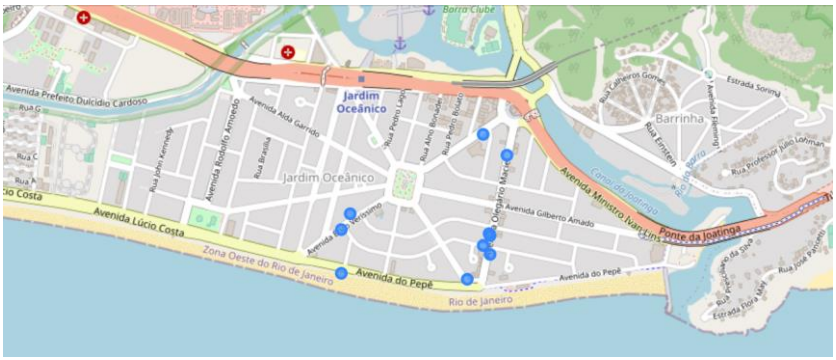
The values on the tables above are a mere illustration, the actual values can be found on the Jupyter notebook tables, available at the project repository.

4.3 Recommendation Yielding

To calculate how much each venue is recommendable to the user, the venues numerical table was multiplied by the user profile as matrixes. The product is the recommendation array, which was added to the original venues table. The top 10 recommended restaurants are displayed, followed by its locations:

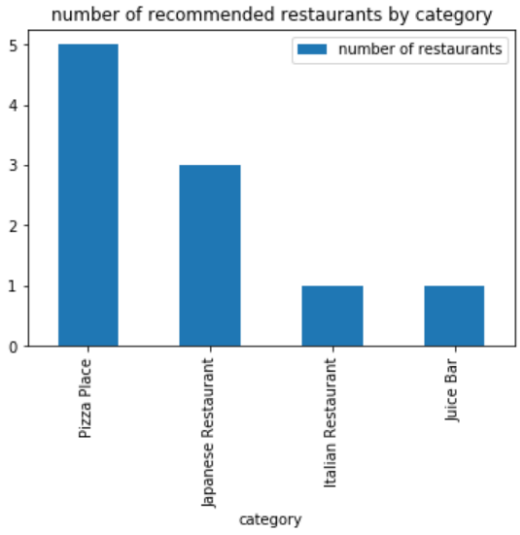
id	name	category	lat	lng	distance	Recommendation
4c81acfed4e2370464956088	Pizzaria Zona Sul	Pizza Place	-23.012214	-43.311411	0.002160	0.568043
587c01a8c7ec6b7a1e21b1ad	Skipper 1992 Pizza Bar	Pizza Place	-23.013088	-43.305379	0.004311	0.137705
4c0bd1fda1b32d7f2c899bf0	Capriccios a	Pizza Place	-23.013811	-43.305366	0.004317	0.136652
4bcbaaa368f976b0af9d6183	Pe'ahi	Japanese Restaurant	-23.014807	-43.306366	0.003556	0.078222
4e9d880010814f7188b647d2	Gero	Italian Restaurant	-23.012820	-43.311783	0.002217	0.030332
50b9ff41e4b05d6264f5dd43	Japa Jato	Japanese Restaurant	-23.013463	-43.305665	0.004006	-0.011753

4c1ba65ab4e62d7f6bb7d993	Koni Store	Japanese Restaurant	-23.012976	-43.305392	0.004311	-0.072640
4d702dbc516b8cfa1af25d10	Mr Lenha - Pizzaria e Restaurante	Pizza Place	-23.009028	-43.305679	0.005989	-0.197892
4eb5abee469073bbc64345de	Domino's Pizza	Pizza Place	-23.009882	-43.304617	0.006211	-0.242207
4b8aa188f964a520487632e3	Barraca do Pepê	Juice Bar	-23.014555	-43.311790	0.002370	-0.263548

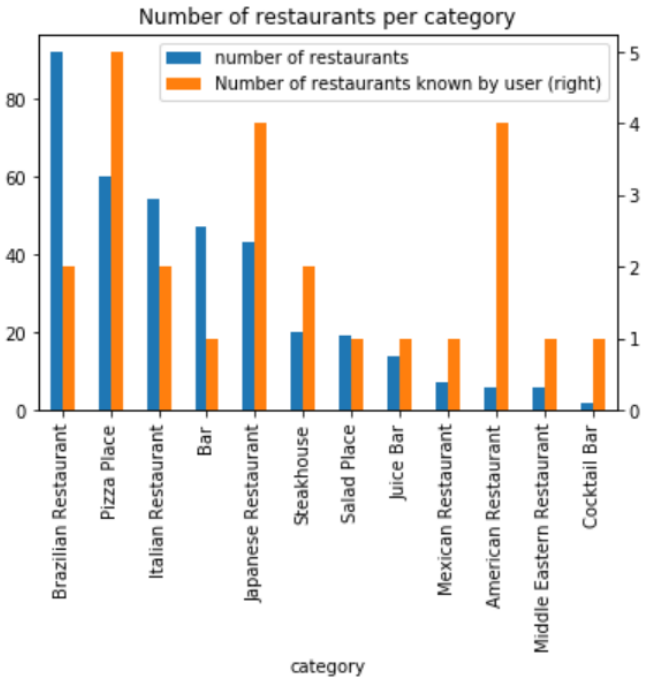


5. Result Analysis

Judging by the top 10 recommended venues, it can be seen on the map that the algorithm successfully yielded venues close to the user, the farthest one being mere 8 minutes away by feet, as estimated by Google Maps. Besides the distance component, most of the venues also belong to the categories the user frequents the most and rates highly, as it should be.



A comparison can also be done between the user's experience and the local offer:



The user tends to known less restaurants from categories that have low availability, which makes sense. But there also are exceptions to that trend, which is a manifestation of the user's personal taste. That can clearly be seen at the "american restaurant" category.

As expected, the recommended system yielded venues: close to the user; from categories with high availability, which makes them easier to be found nearby; from categories the user rates highly. Although, if a restaurant belongs to a category that the user has experience and high satisfaction with, but low availability, it still might be recommended as well, as long as the user finds herself close to it.

6. Conclusion and Future Directions

The model implemented showed itself successful by yielding the expected results given the user profile. On the other hand, with higher effort and investment, there are improvements that can be done. For example, by subscribing to the Foursquare premium account, more information could be gathered about each venue, adding a new layer to the recommendation analysis, which could make its results even more fitting. Besides, if there were a record of not only of which venues the user knowns, but also how often she goes to each of them, her profile could be more accurate as well.