
CULINARY CONSTELLATIONS

Daniel Calderon Castellanos

November 12, 2020

1. Introduction

1.1 Background

Don't you like to go to places where you can choose among a great variety of cuisine choices? Aren't they helpful when you just have not made up your mind on where to eat?

Places having many food venue choices in the same place target this kind of market. We can see this kind of pattern in food courts, food boulevards and restaurant areas. We will call this phenomena "food venue clustering" from now on,

When looking at culinary constellations from a business perspective, you may wonder questions like:

- How the different venues are integrated one with another?
- Do they complement each other by offering a different kind of cuisine? Or same cuisines are crumped together?
- How important is the cuisine cluster composition?

Why should we should bother to analyze this phenomena?, well, by understanding how a food venue cluster is composed, you can use a pattern of existing clusters to predict a successful cluster composition.

1.2 Target Audience

If you are a food industry investor, or are interested in deciding on where to place your food venue in America, this analysis can give you valuable insights.

For example:

- Is there a correlation between some kinds of cuisine being close to your particular kind of cuisine?
- Close to which kind of venues should you locate your venue?
- How does a food venue cluster look like?

1.3 Business Problem

“Is it possible to predict a food venue’s success or failure factor due to its location by looking at other food venues which are nearby?”

We can answer this question by using analytic techniques to recognize food venue clusters and classifying them to focus on their similarity and feature correlation, we can answer this question with as little as geographic location and category provided by the Foursquare basic (developer’s) API.

2. Data

2.1 Dataset selection

We already established the dataset will be generated using the developer’s **Foursquare API** and we want to invest in a restaurant somewhere in America

<https://foursquare.com/>

We need a choice criterion, which is generally valid across America, so we will be collecting data from the most cosmopolitan cities in the country. Following article in **WalletHub** will be used to fulfill this criterion.

<https://wallethub.com/edu/cities-with-the-most-and-least-ethno-racial-and-linguistic-diversity/10264>

We will need a Geolocation tool; its use will be very extensive since geographic distance is used for selecting and clustering the different food venues.

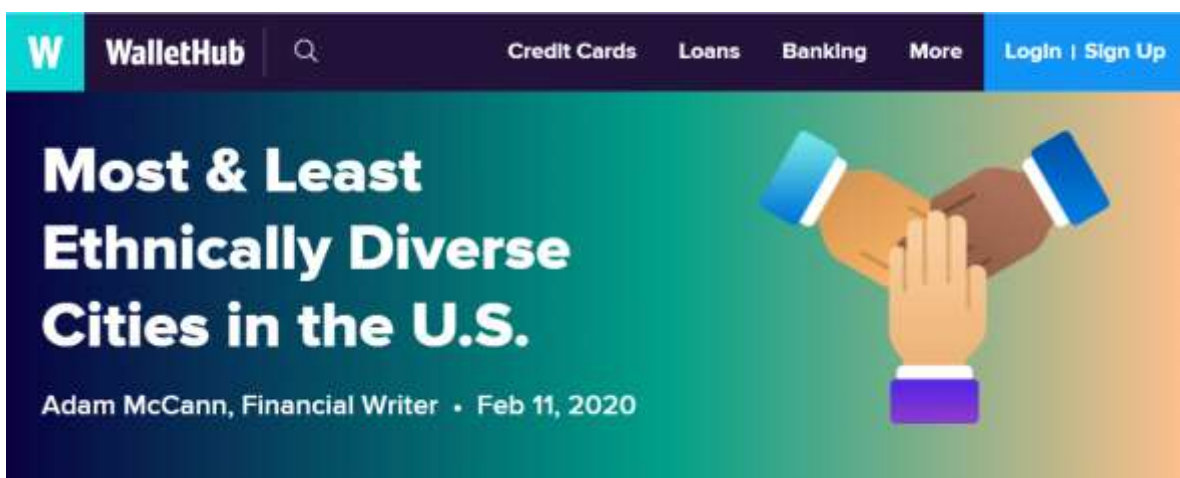
The **GeoPy** API was chose for this task, it's free and is stable enough. Here's the URL:

<https://geopy.readthedocs.io/en/stable/>

2.1.1 WalletHub data source

It is considered important that the model can be applied in general across all regions in US. Is also important that the model should consider as many cuisines as possible in order to explore all possible cuisine categories.

In order to archive this goal, we will be targeting a culturally neutral dataset as much as possible. The article WalletHub article on "Most as Least Ethnically Diverse Cities in the U.S." will help us to select cities with high diversity as much as possible. This is how the web page looks like:



<https://wallethub.com/edu/cities-with-the-most-and-least-ethno-racial-and-linguistic-diversity/10264>

The website is loaded with plenty of diversity statistics and heat maps.

The article contains an explanation on how the diversity score is calculated, the details are in the article but they boil down to the following:

- Ethnoracial Diversity: Total Points – 50%
- Linguistic Diversity: Total Points – 33%
- Birthplace Diversity: Total Points – 17%

Among the many statistic tables provided by the web site we are interested in one in particular, it looks like this:

Ranking by City Size

Rank* ↕	Large City Name (Score)	Rank ↕	Midsize City Name (Score)	Rank ↕	Small City Name (Score)
1	New York, NY (69.38)	1	Jersey City, NJ (72.56)	1	Gaithersburg, MD (72.01)
2	Oakland, CA (68.95)	2	Spring Valley, NV (69.99)	2	Germantown, MD (71.55)
3	San Jose, CA (68.56)	3	Kent, WA (68.05)	3	Silver Spring, MD (69.60)
4	Sacramento, CA (66.33)	4	Enterprise, NV (66.60)	4	Rockville, MD (69.09)
5	San Francisco, CA (66.29)	5	Paradise, NV (66.36)	5	Federal Way, WA (64.94)
6	Boston, MA (65.81)	6	Bridgeport, CT (66.29)	6	Lynn, MA (64.74)

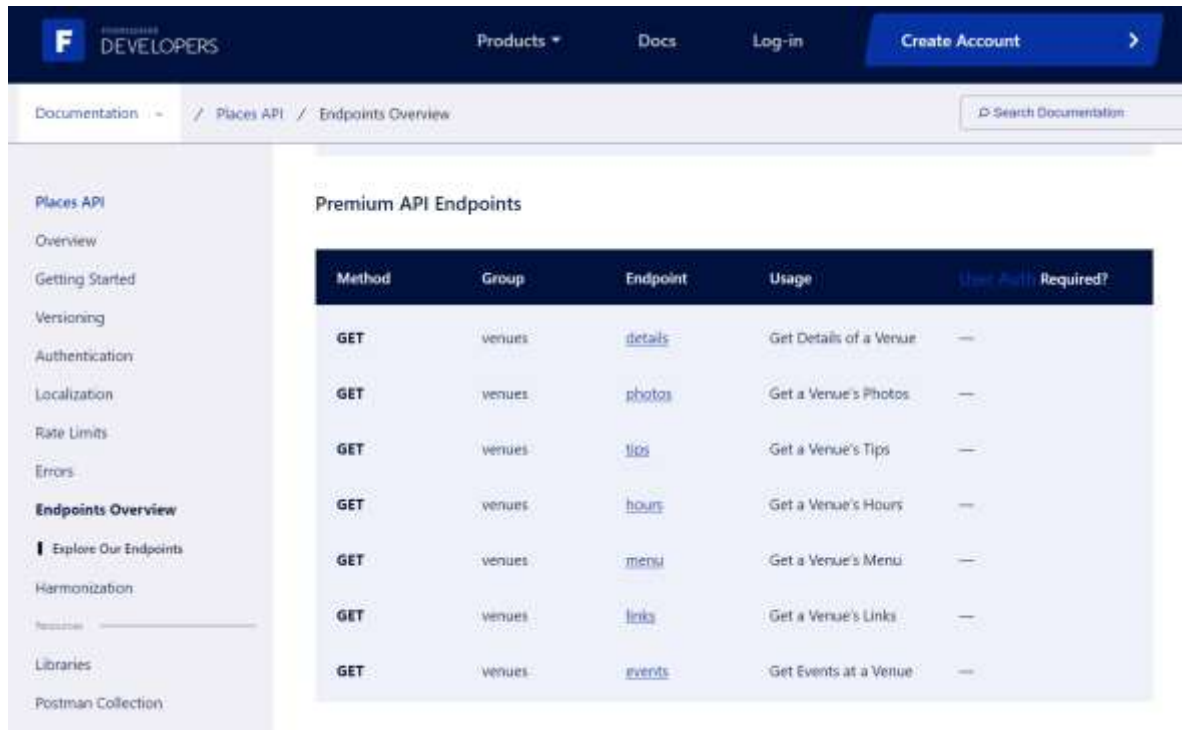
We are particularly in large cities because (as we are going to explain in the next section) the Foursquare API explore endpoint only return recommended venues. So without a complete venue listing, is difficult to find a food venue in proximity with each other, unless both are recommended venues, and that only happens in areas with great concentration of food venues, like the ones we find in the large cities.

Below is a sample of city_info.csv. Some city names needed to be rewritten in order the geolocation API to correctly identify them:

Rank	City	Diversity Score
1	New York, NY	69.38
2	Oakland, California	68.95
3	San Jose, CA	68.56
4	Sacramento, CA	66.33
5	San Francisco, California	66.29

2.1.2 Foursquare data source

We will be using Foursquare API, free regular web services end points. These endpoints make it right for a social application, but true analytical data is found in the premium endpoints, like menu details, price ranks, venues hours. See the Foursquare page below:



The screenshot shows the Foursquare Developers API documentation page. The top navigation bar includes the Foursquare logo, 'DEVELOPERS', and links for 'Products', 'Docs', 'Log-in', and 'Create Account'. The breadcrumb trail indicates the current location: 'Documentation / Places API / Endpoints Overview'. A search bar is also present. The left sidebar lists various API sections, with 'Endpoints Overview' selected. The main content area is titled 'Premium API Endpoints' and contains a table with the following data:

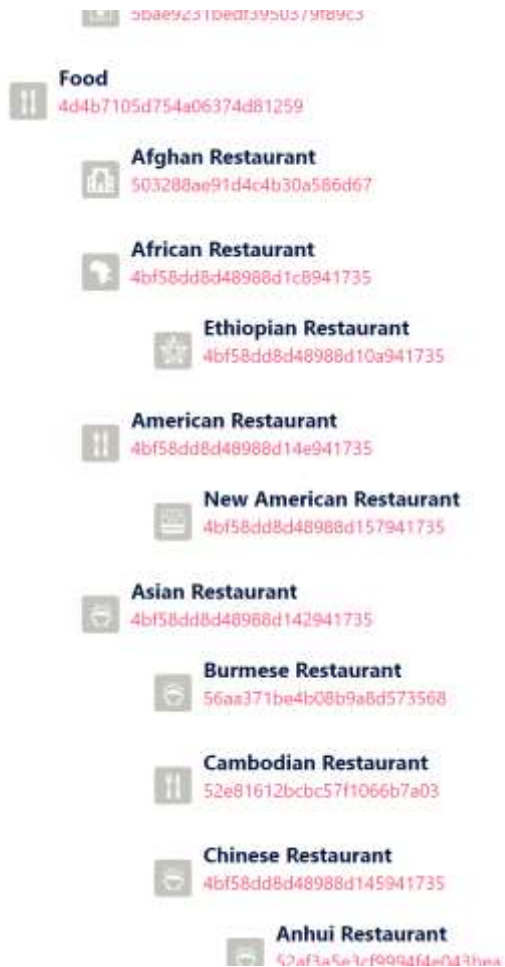
Method	Group	Endpoint	Usage	Use Auth?	Required?
GET	venues	details	Get Details of a Venue	—	—
GET	venues	photos	Get a Venue's Photos	—	—
GET	venues	tips	Get a Venue's Tips	—	—
GET	venues	hours	Get a Venue's Hours	—	—
GET	venues	menu	Get a Venue's Menu	—	—
GET	venues	links	Get a Venue's Links	—	—
GET	venues	events	Get Events at a Venue	—	—

It is important to point out that our cluster features will consist only of the food category composition of the venue cluster. That is the percentage that each cuisine represents from all cuisines in a set of food venues that are close to each other. Additional features are desirable for a more robust cluster model, but we are limited by the paid premium end point.

It is also important to point out that the explore end point of Foursquare regular set only returns recommended venues, so is difficult to locate small food venues concentrations like food courts or plazas.

Only when recommended venues are next to each other, say in the same block or across the street, we will be able to identify a cluster of venues, a very large cluster of venues like food boulevards or restaurant areas. This is the kind of food venue clusters we will be exploring.

So we are focusing in venues in the food category. Foursquare has a hieratical system to classify this kind of venues:



This hieratical tree will be simplified into dozen flat food categories, where some of the bottom categories can be raised at main categories levels given some categories importance. More details in the “Data Preparation” section.

<https://developer.foursquare.com/docs/build-with-foursquare/categories/>

2.1.3 GeoPy data source



Geolocalization will be provided by the GeoPy data source API

GeoPy will be used to determine the geospatial coordinates of the cities of interest. Those will be the starting point of Foursquare explore end point.

Since the geospatial coordinates are included in the json structure of the Foursquare explore endpoint, there will be no need to call the GeoPy to complete the food venue data set.

3. Methodology

We will aboard the subject of choosing machine learning algorithms in this section because we need to run 2 kinds of algorithms: the one that clusters the food venues from the dataset and the one that is used to implement the predictive model.

3.1 Machine learning algorithms used in the analysis

This section will cover the rationale and the goals meant to be accomplished in decision making process of choosing the right tools for this analysis:

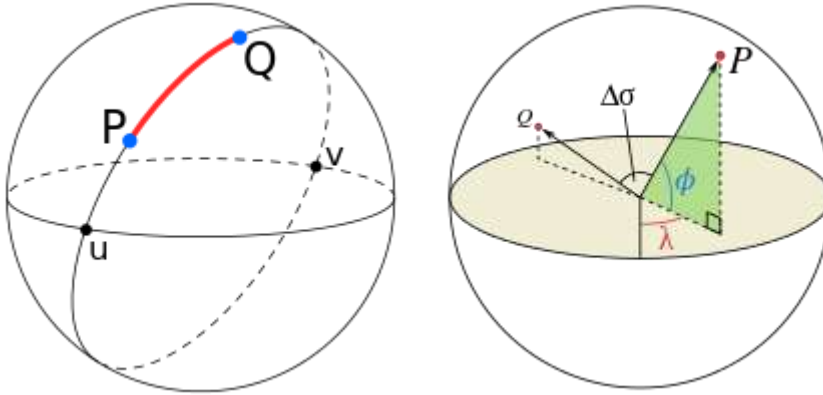
3.1.1 Clustering using DBSCAN

When it comes to choose and algorithm to cluster geospatial coordinates the decision is obvious: DBSCAN algorithm

DBSCAN is especially very good for tasks like class identification on a spatial context. The wonderful attribute of DBSCAN algorithm is that it can find out any arbitrary shape clusters without getting affected by noise.

Distances between different venues will be derived from geospatial coordinates using *haversine coordinates*. Scikit Learn library implementation implements haversine geometry natively.

The haversine formula determines the great-circle distance between two points on a sphere given their longitudes and latitudes as illustrated in the next figures:



The mean earth radius that will be used for the harvestine formula is:

6,371.009 km per radian

The explanation of this constant as well as a more complete explanation of the harvestine formula can be found at the following link:

https://en.wikipedia.org/wiki/Haversine_formula

3.1.2 Prediction Model using Logistic Regression

The choice of a model is also easy here: Logistic Regression. This algorithm not only is able to provide the probability of the prediction, which is desirable, it also helps to understand the relationship of the outcome with the underling features.

Logistic Regression can provide the weights factor or confidence of the equation, which allow us to predict the impact of a change in a feature.

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m = \beta_0 + \sum_{i=1}^m \beta_i x_i$$

Features of a cluster will be exclusively composed by the food categories included in that cluster, or more specifically, the percent of categories that compose a certain cluster.

As an example, three clusters are shown below showing part of their respective features, which are the share proportion of each category that conforms that cluster (those values have to be multiplied by 100 in order to express them as a percentage of the whole clusters:

	Mediterranean Restaurant	American Restaurant	Cafe & Bistro	Asian Restaurant	Latin American Restaurant	Salad Place	Pizza Place	Mexican Restaurant	Italian Restaurant
global label									
48	0.200000	0.200000	0.200000	0.200000	0.200000	0.000000	0.000000	0.000000	0.000000
50	0.285714	0.000000	0.000000	0.000000	0.000000	0.142857	0.285714	0.142857	0.142857
62	0.166667	0.000000	0.166667	0.333333	0.000000	0.000000	0.000000	0.000000	0.333333

3.2 Explanatory Data Analysis

3.2.1 Initial Dataset

Our initial data was retrieved by the Foursquare API based on the city information provided by WalletHub. All venues retrieved belong to the 'food' category. (*id=4d4b7105d754a06374d81259*)

Here are some facts:

Total cities explored for food venues: 60

Total food venues found: 8881

Food venue count per city average: 148.

Top 5 cities by food venue count:

food venues found	
city name	
Baltimore, MD	249
Honolulu, HI	249
New York, NY	247
Boston, MA	247
Seattle, WA	247

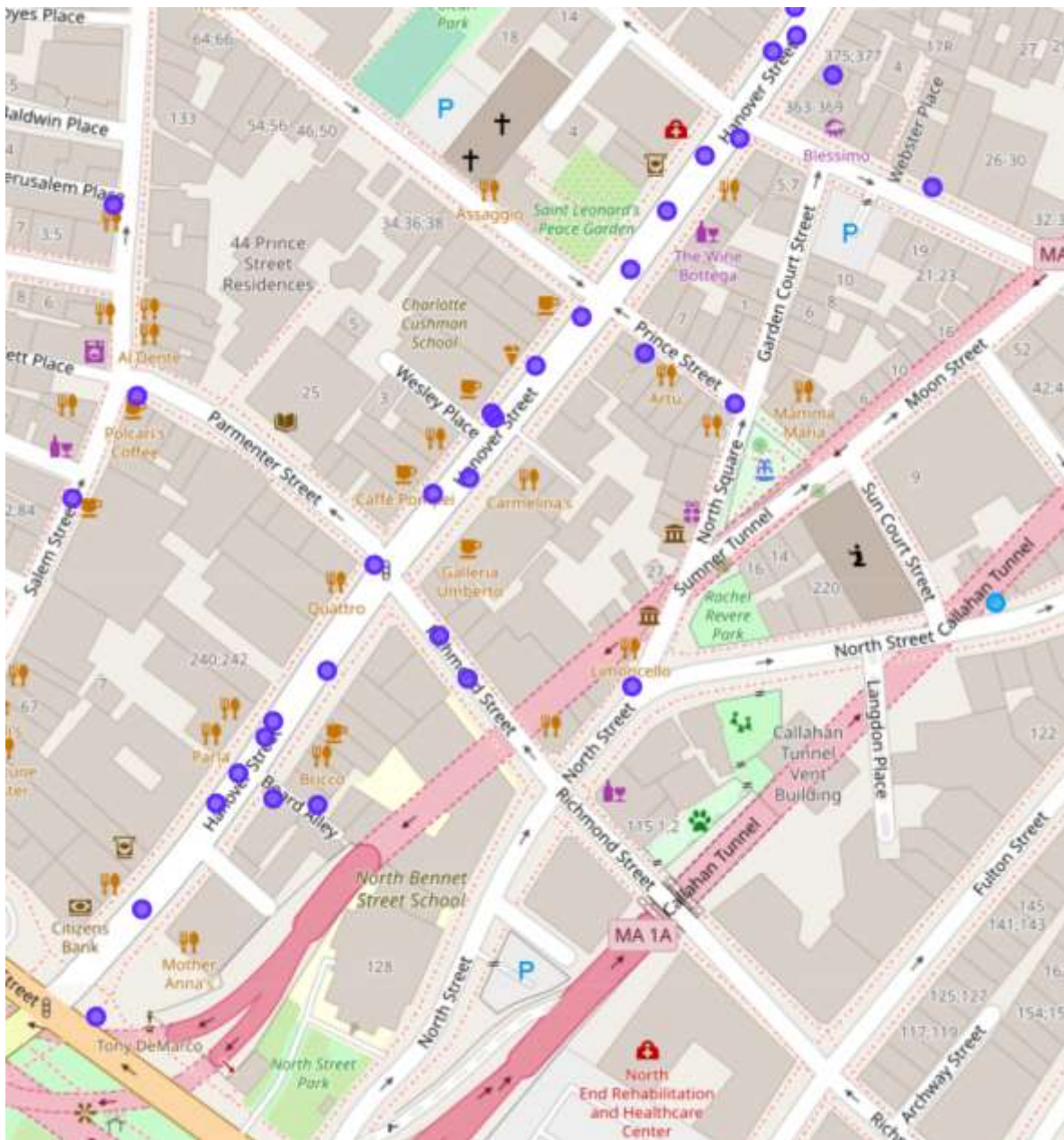
3.2.2 Clustering nearby food venues

We test different epsilons for cluster creations from our initial dataset. Epsilon is the maximum distance between two samples for one to be considered as in the neighborhood of the other.

There is no better way to explore geolocation dataset other than visualize them over a map.

Several map files were created. Here are two of the most interesting:

The largest food venue cluster found, located in Boston, MA:

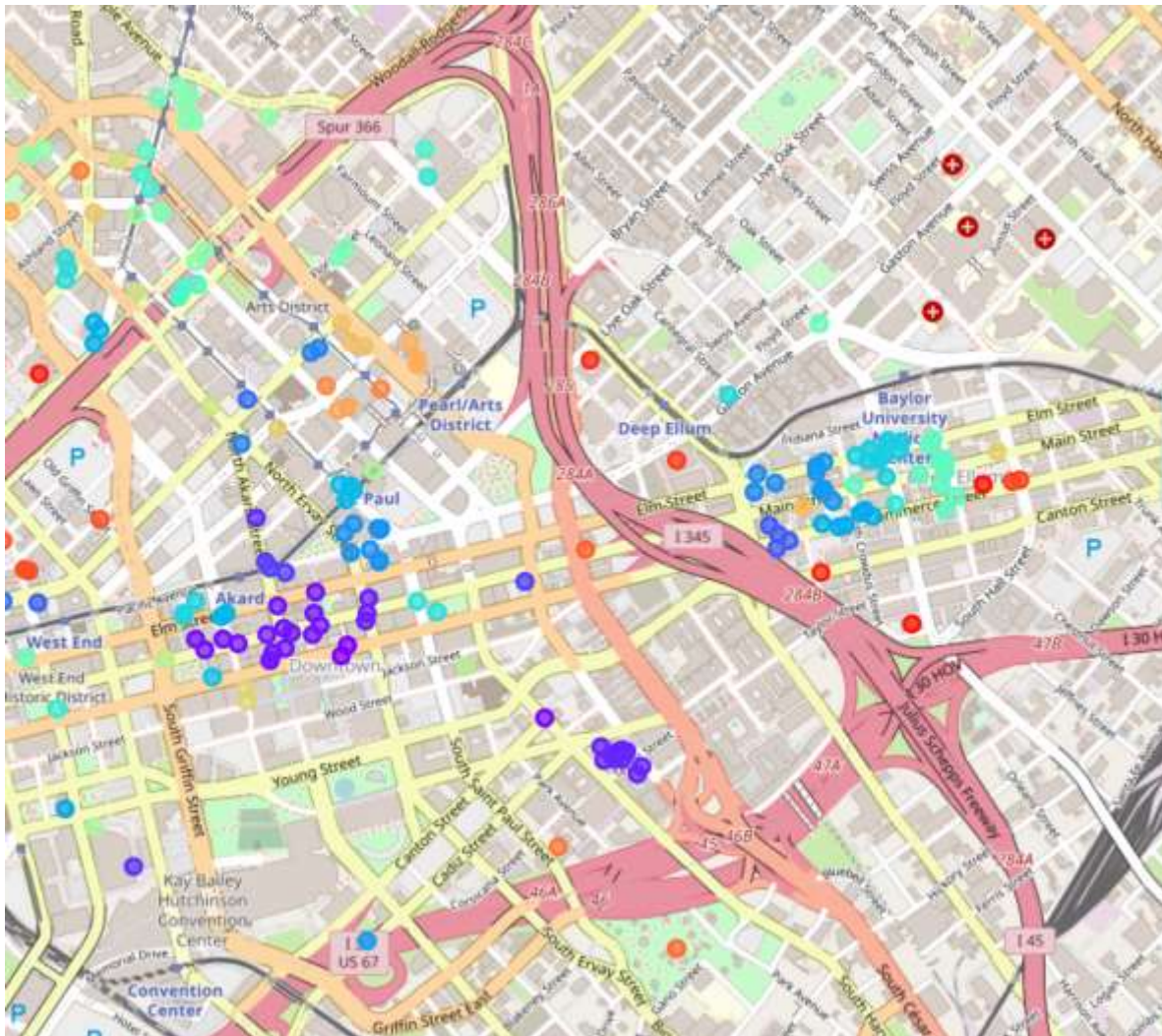


This visualization shows the largest food venue cluster in our dataset, a food venue cluster over Hanover Street in Boston Massachusetts. We have been able to cluster arbitrary cluster shapes because we are using the DBSCAN algorithm, where we decide on the distance between each sample, not on the size of the whole cluster. In this case, we are detecting a large food boulevard.

To the left, multiple clusters clumped into a big cluster,

To the right, multiple cluster, very close together, but not becoming one,

Downtown Dallas, TX:



This example is a great way to visualize the impact of the epsilon parameter in the DBSCAN. Epsilon is the maximum distance between two samples for one to be considered as in the neighborhood of the other. This is the most important DBSCAN parameter to choose appropriately for our data set and distance function. Epsilon was set to the following value:

Eps = 200 feet = 61 meters

This will cluster venues in the same regular blocks, around the corner, or across the street.

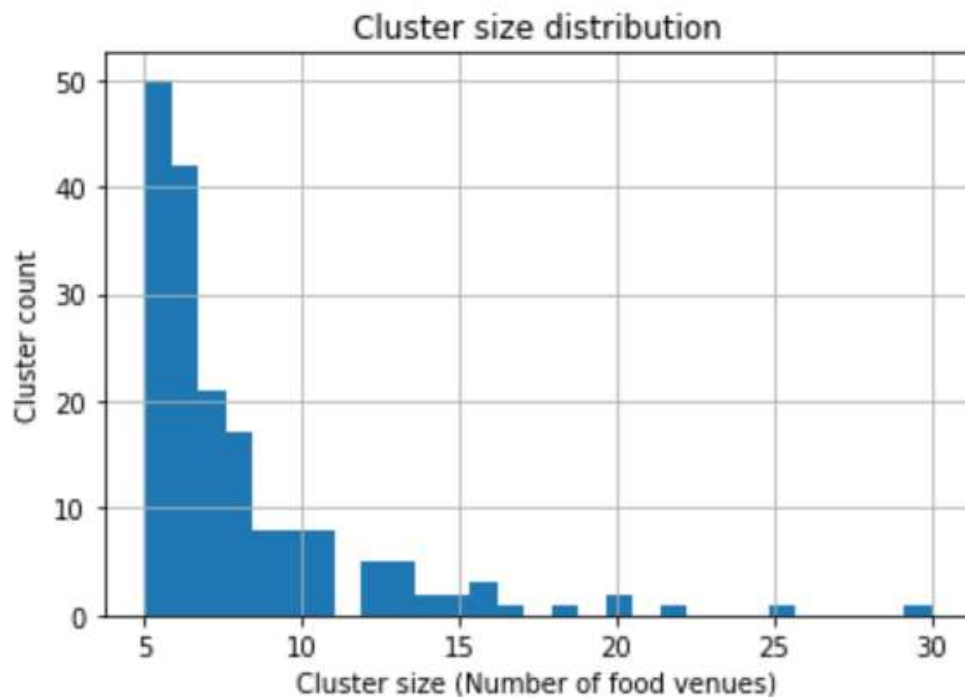
3.2.3 Criteria Selection for Final Dataset

Not only we want our model to be applied in general across the country, but also is desirable to be able to apply the model no matter the size of the cluster.

By generating a frequency distribution graph it can be shown that the occurrence of clusters with higher number of venues decays exponentially as the size of the cluster increases. So the frequency of each of the smaller size varies greatly among them.

The clusters of size 1 (if they can be called “clusters” because a single entity does not conform a cluster) conforms the largest part of the dataset.

Clusters of lesser size are not shown for clarity, the frequency graph below shows that if we consider 5 food venues placed next to each other as cluster, the majority of the clusters fall into the 5 to 11 size range:



We will assume that the food category composition of large clusters is the same as the smaller ones, so we will focus our analysis on food category proportion as the only features, not cluster size or the particular culture of a region. We will train the model with data that meet the following criteria:

City diversity score ≥ 50

Cluster size ≥ 5

Here are some facts about our final dataset:

total venues in dataset: 1408 (actually belong to a cluster)

total clusters in dataset: 178

cluster min size: 5

cluster max size: 30

cluster mean size: 7.91

Clusters with the largest venue count:

	city	rank	score	venue_count
global label				
567	Boston, MA	6	65.81	30
3105	Seattle, WA	36	52.50	25
708	San Diego, CA	7	65.70	22
342	Sacramento, CA	4	66.33	20
1927	Honolulu, HI	22	59.63	20

3.2.2 Data Preprocessing

3.2.2.1 Normalization

Since all dimensions are in the same domain, same units, no other transformation is needed to normalize the data except converting the total venues in a category to ratios against all venues in a given cluster. It has to be pointed out the non-symmetrical distribution: the values range will be from zero to one, and there will be no zero averages.

3.2.2.2 Feature set

The number of food categories were reduced to about a dozen, in order to make the feature set more understandable.

The categories were grouped together in a hierarchical fashion, only if a category in the most specific branches of the hierarchical tree were very relevant (found very frequently in the selected venues) they were kept in their own category:

Sample of food_grouped_categories.csv

main_category	food_grouped_category	Count
Mexican Restaurant	Mexican Restaurant	101
Pizza Place	Pizza Place	87
Italian Restaurant	Italian Restaurant	84
American Restaurant	American Restaurant	81
Sandwich Place	Sandwich Place	63
Seafood Restaurant	Seafood Restaurant	53
Sushi Restaurant	Japanese Restaurant	53
Japanese Restaurant	Japanese Restaurant	48
Bakery	Cafe & Bistro	45
Café	Cafe & Bistro	42
Burger Joint	Burger Joint	41
New American Restaurant	American Restaurant	36
Asian Restaurant	Asian Restaurant	35
Mediterranean Restaurant	Mediterranean Restaurant	35
Restaurant	Comfort Food Restaurant	33
Chinese Restaurant	Asian Restaurant	27
BBQ Joint	Steakhouse	27
Food Truck	Fast Food Restaurant	27
Vietnamese Restaurant	Asian Restaurant	26
Steakhouse	Steakhouse	22
Gastropub	European Restaurant	21
French Restaurant	European Restaurant	20

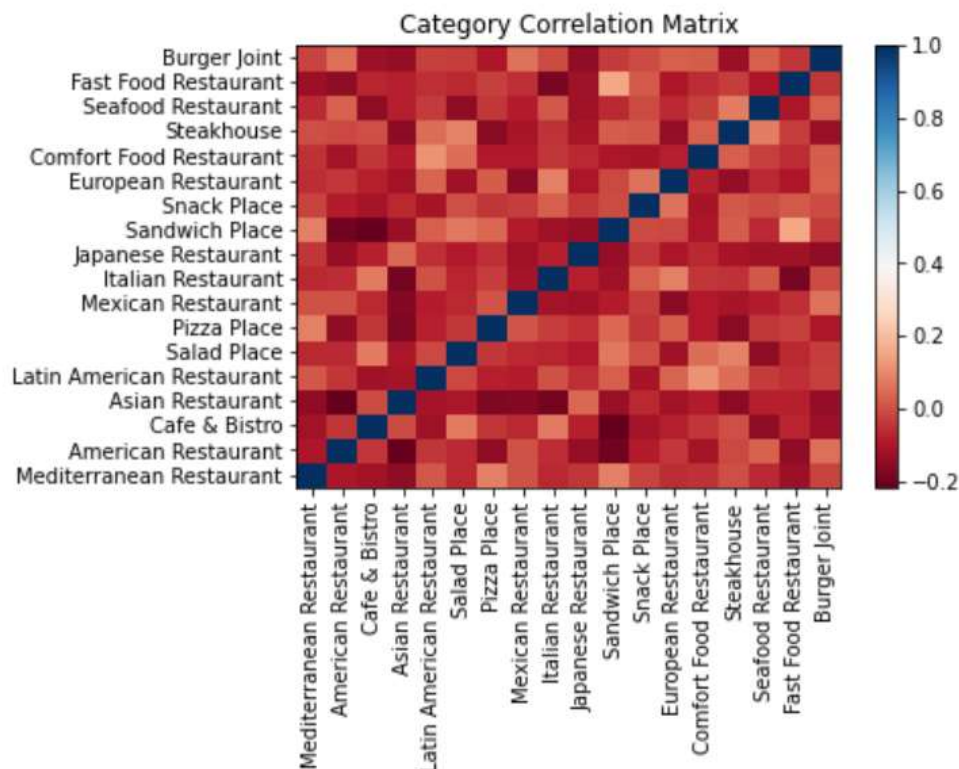
3.3 Inferential statistical testing

In order to check for any correlation, the feature correlation matrix was generated, below is a fragment of it:

	Mediterranean Restaurant	American Restaurant	Cafe & Bistro	Asian Restaurant	Latin American Restaurant	Salad Place	Pizza Place	Mexican Restaurant	Italian Restaurant	Japanese Restaurant	Sandwich Place
Mediterranean Restaurant	1.000000	-0.100630	-0.117413	-0.144372	0.014968	-0.064653	-0.088443	0.006334	-0.064218	-0.041928	0.086915
American Restaurant	-0.100630	1.000000	-0.040371	-0.215129	-0.042705	-0.059624	-0.145048	0.006663	-0.054627	-0.134604	-0.192205
Cafe & Bistro	-0.117413	-0.040371	1.000000	-0.005590	-0.122819	0.074759	-0.042023	-0.065382	0.077601	-0.081365	-0.213537
Asian Restaurant	-0.144372	-0.215129	-0.005590	1.000000	-0.113150	-0.105678	-0.175940	-0.163667	-0.190972	0.044247	-0.130893
Latin American Restaurant	0.014968	-0.042705	-0.122819	-0.113150	1.000000	-0.008754	-0.077534	-0.085165	0.004283	-0.052175	0.027879
Salad Place	-0.064653	-0.059624	0.074759	-0.105678	-0.008754	1.000000	-0.036593	-0.064691	-0.072066	-0.089831	0.070223
Pizza Place	-0.088443	-0.145048	-0.042023	-0.175940	-0.077534	-0.036593	1.000000	0.010521	-0.026857	-0.052512	0.041685
Mexican Restaurant	0.006334	0.006663	-0.065382	-0.163667	-0.085165	-0.064691	0.010521	1.000000	-0.113217	-0.123102	-0.082758
Italian Restaurant	-0.064218	-0.054627	0.077601	-0.190972	0.004283	-0.072066	-0.026857	-0.113217	1.000000	-0.079724	-0.122829

Most of the correlation values are close to zero, meaning no correlation among the food categories in a cluster. The analysis is easier using a heatmap:

Category Correlation Heat Map



Most of the correlation matrix are between -0.2 and 0.2, thus the heatmap is painted in red, so there is very little correlation among the features as stated before. There are some remarks that can be noted after a visual inspection:

- There is some positive correlation (light red) between a Sandwich Place and a Fast Food Restaurant. Or at least is a better correlation than other relationships in the matrix.
- Asian Restaurant is shifted to heavier red, meaning the most negative values. This means Asian Restaurant is found slightly more often when other categories are absent. This is mainly because the omnipresence of an Asian Restaurant in most of the clusters.

Some additional statistical testing will be provided using a Jaccard Index after the results of the Logistic Regression are presented.

4. Results

Let's model the two after mentioned categories, the Sandwich Place and the Asian Restaurant. The same feature set is going to be used for both, only that for training the Sandwich model, the Sandwich Place category will be the dependent variable and all others the independent variables. Same with the Asian Restaurant but the independent variable would be the Asian Restaurant category.

4.1 The Sandwich Place Model

Below are presented a fragment of the results of the Sandwich Place model when using the testing dataset (20% of the samples)

	without sandwich place	with sandwich place
test cluster index		
0	0.529846	0.470154
1	0.532631	0.467369
2	0.529450	0.470550
3	0.531106	0.468894
4	0.531077	0.468923
5	0.531318	0.468682
6	0.527761	0.472239

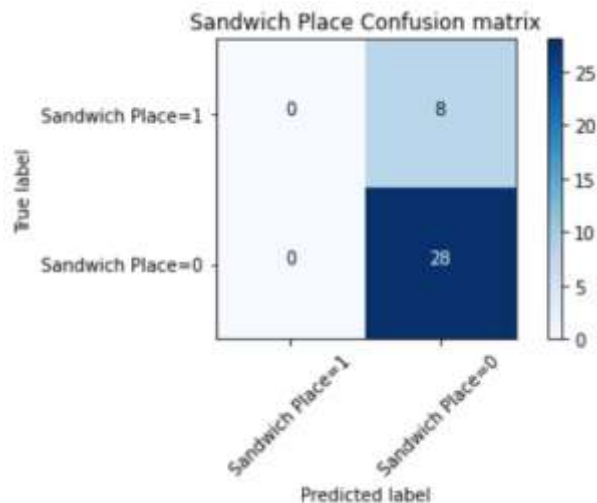
The probability of finding a Sandwich Place in a food venue cluster fluctuates between %46 and 47% from cluster to cluster. This small 1% variation means very little dependency on the cluster kind.

Now let's test our model results against the real data:

Jaccard similarity score ≥ 0.7778

This is surprisingly high score given so little probability difference in the model results.

Let's look at the Sandwich Place Confusion Matrix:



The model got it right 78% of the time by always guessing there is no Sandwich venue in a cluster composition. This sounds right since the probability is always between 46% and 47%.

Let's look at the Sandwich Place model's coefficients, which are related to how much each independent variable contributes to the outcome. This is why we chose the Logistic Regression in first place.

The coefficients are arranged from the greater to the lesser value:

	value
Pizza Place	0.002769
Salad Place	0.000942
Snack Place	0.000525
Mediterranean Restaurant	0.000246
Fast Food Restaurant	-0.000465
Latin American Restaurant	-0.000919
Steakhouse	-0.003952
Seafood Restaurant	-0.004336
European Restaurant	-0.004571
Burger Joint	-0.008050
Italian Restaurant	-0.008301
Comfort Food Restaurant	-0.011391
Japanese Restaurant	-0.017395
Cafe & Bistro	-0.018871
Mexican Restaurant	-0.019732
American Restaurant	-0.028773
Asian Restaurant	-0.032202

It seems that Sandwich Places prefer to live next to a Pizza Place, a Salad Place or a Snack Place, than next to an Asian Restaurant, an American Restaurant or a Mexican Restaurant.

4.2 The Asian Restaurant Model

We will train, test, and analyze our Asian Restaurant model same as we did for the Sandwich Place, but using the Asian Restaurant as the dependent variable this time.

Here is a sample of the output of the Asian Restaurant Model:

	without asian restaurant	with asian restaurant
test cluster index		
28	0.476036	0.523964
20	0.475679	0.524321
32	0.475923	0.524077
22	0.477211	0.522789
27	0.476500	0.523500

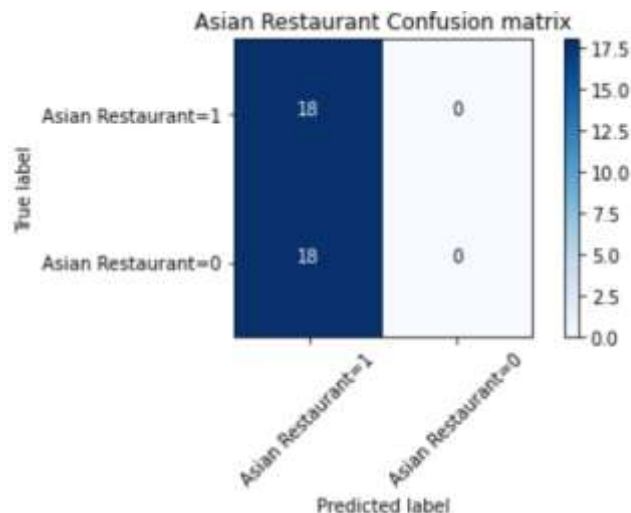
The results give a solid 52% probability of chance of an Asian Restaurant being present on a food venue cluster regardless of the cluster composition.

This is why is no surprise when we compute the following Jaccard similarity score:

Jaccard similarity score ≥ 0.5

The cluster composition give no information about the probability of an Asian Restaurant to be present.

This is confirmed when we take a look at the confusion matrix.



The model has a 50 – 50 chance to correctly predict if an Asian Restaurant is present in a cluster by always guessing there is one (It's always a 52% chance that there is an Asian Restaurant in any food venue cluster)

5. Discussion

The analysis clearly shows there is very little correlation between a food venue cluster composition and the kind of food venues that are present on that cluster.

It is true we found an example of a venue with a food category that had some correlation with the category of the fellow venues in the same cluster. But we found it only by looking for an extreme example in the correlation heat map.

It can be argued that the Sandwich category is correlated to other categories simply because of the fact that the other related categories (pizza, salad, snack) belong to the same category umbrella, fast food, and they are usually congregated together in a food court.

But remember, we are not analyzing food courts, we can't given the available dataset. We are analyzing large concentrations of venues like food boulevards or restaurant areas due to incompleteness in the information. The missing venues are evident when looking at the plots in the maps, the maps show more venues than listed in the recommended venues of Foursquare.

If we could repeat the analysis at a food court level, we can then still expect cluster compositions to be diverse in their categories, but that does not ensure that we will find a pattern, which is the desirable outcome, a unbiased analysis has to be done.

5. Conclusion

After all the work devoted to this analysis, and after looking at the very small correlation found in the results, we arrive into a conclusion that we already knew intuitively: People like variety, but this time, the statement has been proved by facts instead of perception.

There is another conclusion we can arrive. Given the reasons explained above of why our data exhibit such a little correlation, we can think that maybe we aimed for such a universal model of a food cluster venue, that in the end, we found very loose connections between the kinds of food venues in a cluster.

Maybe if we aim for a particular kind of food venue cluster, like ones found in regions with cultural predominance, a more precise location like food courts or plazas, add features like price range, size of the venues, open hours, then we could find a more deterministic pattern in the results, and by extension more helpful.

Unfortunately, the level of detail of a dataset is correlated to it's price, so even when this analysis is not enough for use in a real life scenario, it can be used as a proof of concept. If we were able to arrive to some basic conclusions with so little detail, a promising outcome is expected from a more comprehensive analysis.

In the meanwhile, if you are planning to place a restaurant in a large food venue area, remember that an Asian restaurant is one of your safest bet. And if you are betting for a sandwich place, make sure is in the same area as a pizza place, a salad place, or a snack place.

Thanks

5. References

- <https://foursquare.com/>
- <https://wallethub.com/edu/cities-with-the-most-and-least-ethno-racial-and-linguistic-diversity/10264>
- <https://geopy.readthedocs.io/en/stable/>
- <https://wallethub.com/edu/cities-with-the-most-and-least-ethno-racial-and-linguistic-diversity/10264>
- <https://developer.foursquare.com/docs/build-with-foursquare/categories/>
- <https://geoffboeing.com/2014/08/clustering-to-reduce-spatial-data-set-size/>
- https://en.wikipedia.org/wiki/Haversine_formula
- <https://python-visualization.github.io/folium/>