Culinary Constellations

Data Science Capstone Project

Introduction

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 haven’t 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 perspectives, you may wonder questions like:

- How are the different venues integrate 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 different cuisines composition?

But 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.

Introduction

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 looks like?

Business Problem

The Question

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

We can answer this questions by using analytic techniques to recognize food venue clusters and classifying then to focus in 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.

Data

Dataset choice

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 criteria which is generally valid across America, so we will be collecting data from the most cosmopolitan cities in the country. The following URL will be used to fulfill this criteria.

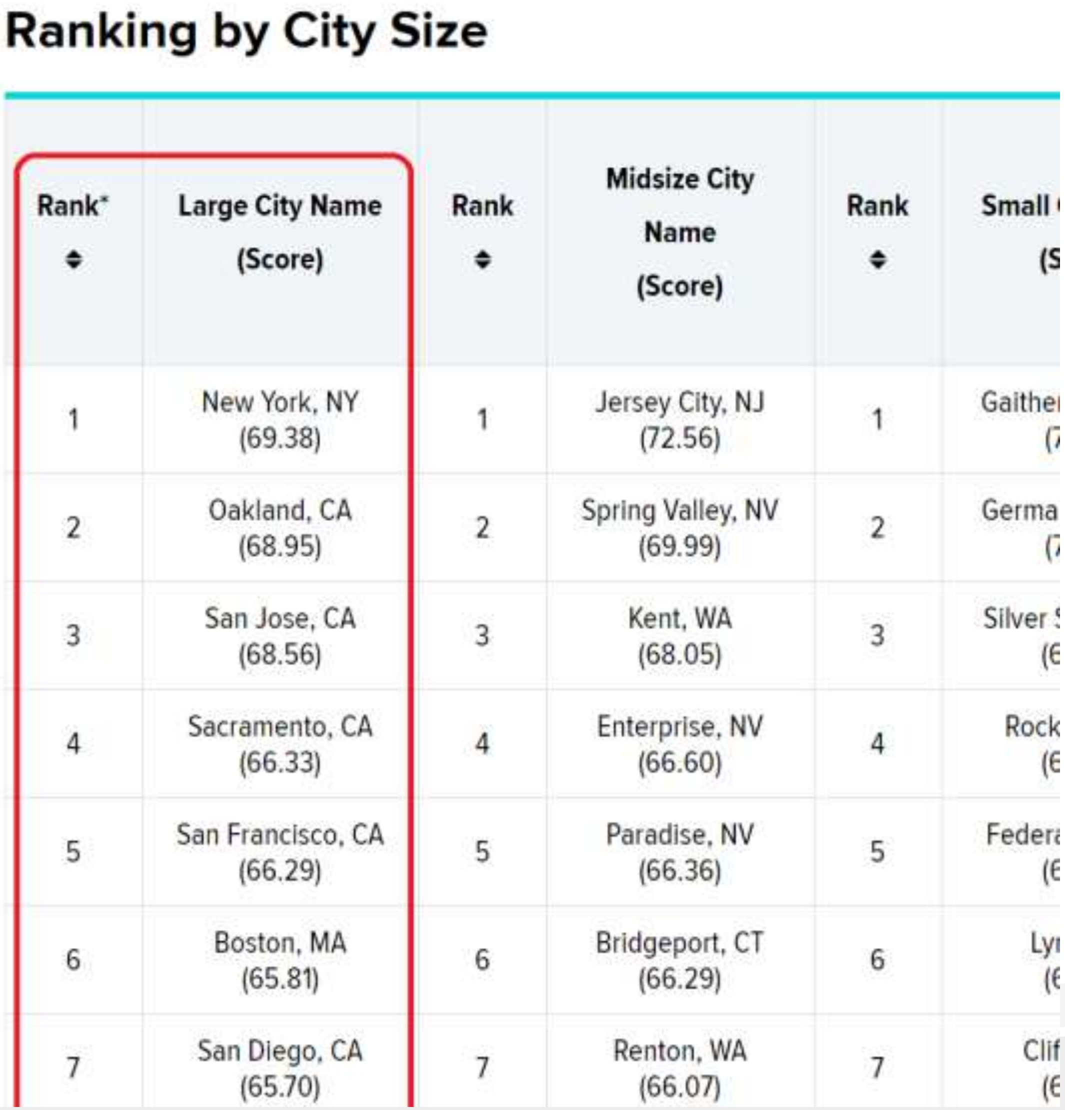
https://wallethub.com/edu/cities-with-the-most-and-least-ethno-racial-and-linguistic-diversity/10264

Data

Data considerations

- We only be considering large cities were culinary constellations are most likely to be found, so we can have access to enough data.

- We will be only be using data with diversity score greater than 50. See the after mentioned URL for diversity score clarification. Data from places with most diversity will help to archive region neutrality.



Ranking by City Size

Rank* ⬍	Large City Name (Score)	Rank ⬍	Midsize City Name (Score)	Rank ⬍	Small (S
1	New York, NY (69.38)	1	Jersey City, NJ (72.56)	1	Gaithe (7
2	Oakland, CA (68.95)	2	Spring Valley, NV (69.99)	2	Germa (7
3	San Jose, CA (68.56)	3	Kent, WA (68.05)	3	Silver S (6
4	Sacramento, CA (66.33)	4	Enterprise, NV (66.60)	4	Rock (6
5	San Francisco, CA (66.29)	5	Paradise, NV (66.36)	5	Federa (6
6	Boston, MA (65.81)	6	Bridgeport, CT (66.29)	6	Ly (6
7	San Diego, CA (65.70)	7	Renton, WA (66.07)	7	Clif (6

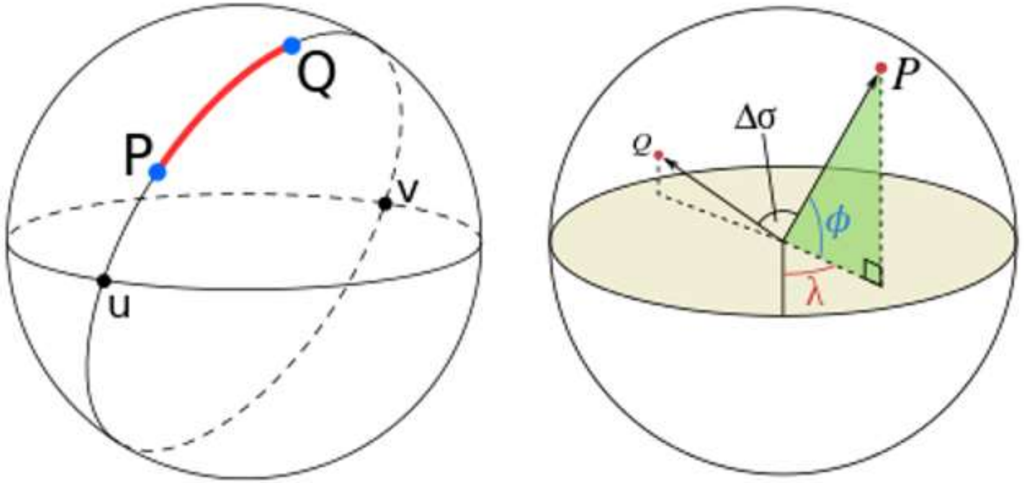
Methodology

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.

Methodology

Distance function for DBSCAN



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

  mean earth radius = 6,371.009 km per radian

- The maximum distance between two samples for one to be considered as in the neighborhood of the other:

  Epsilon = 200 feet = 61 meters

Methodology

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.

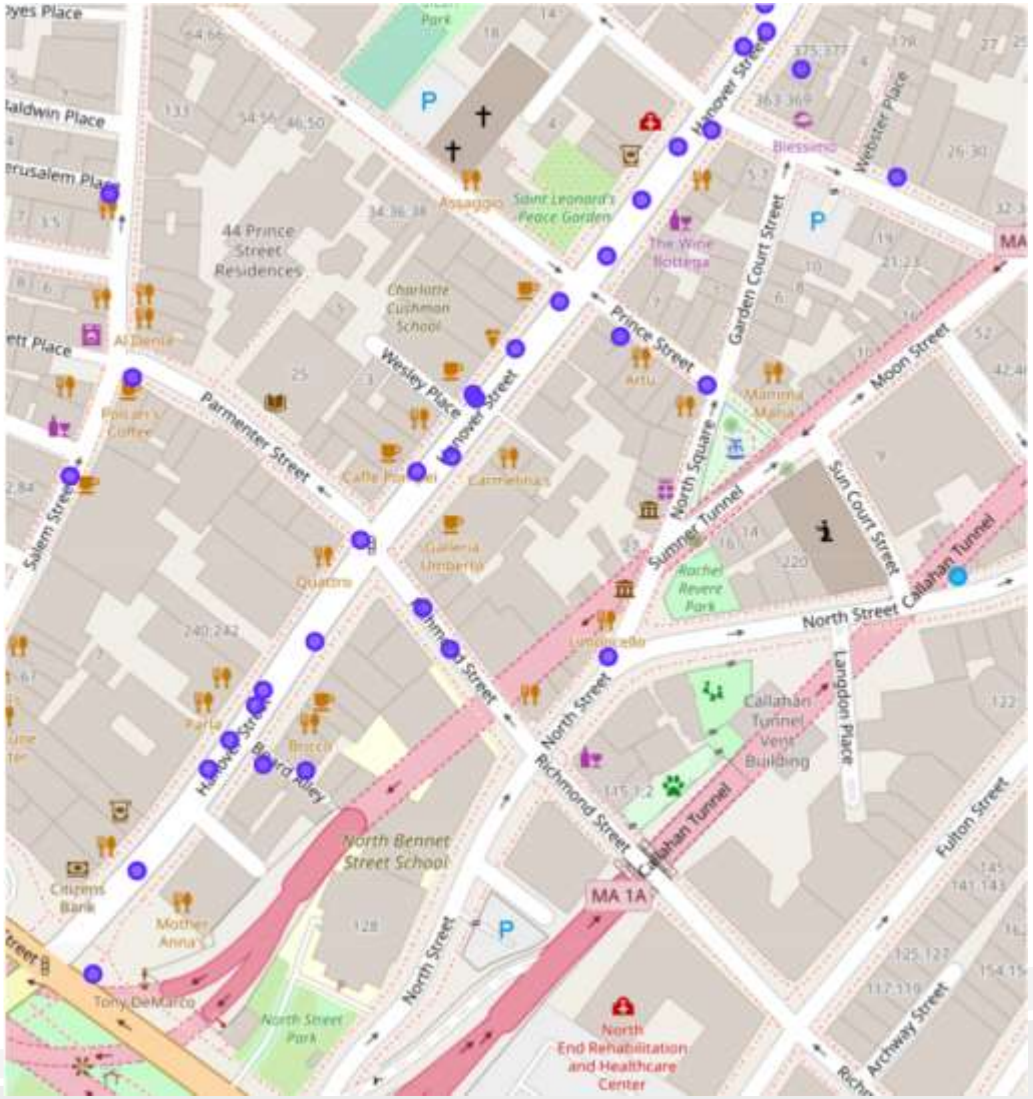
- $$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m = \beta_0 + \sum_{i=1}^{m} \beta_i x_i$$

- 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.

Methodology

Data visual exploration

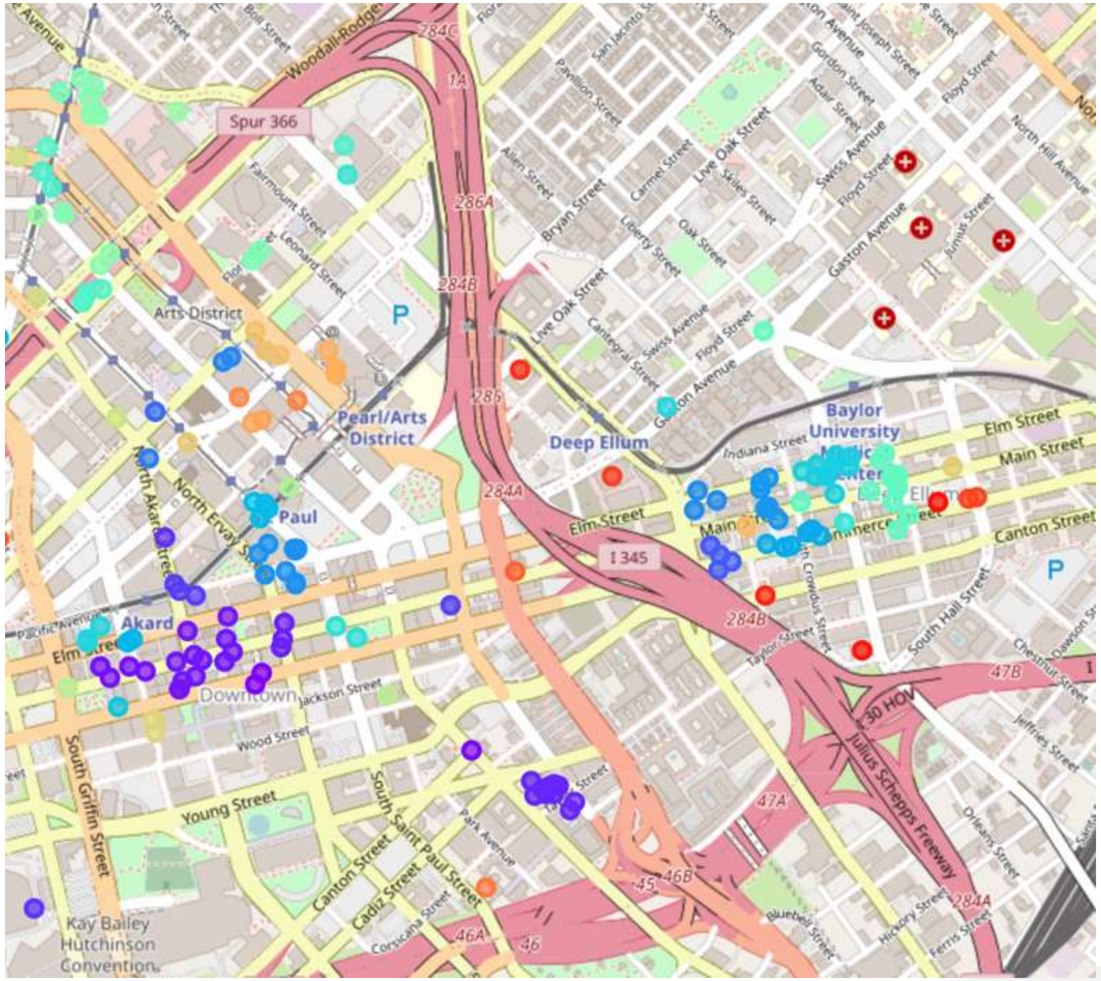
- The largest food venue cluster found, located in Boston, MA.
- This visualization shows the largest food venue cluster in our dataset, a food venue cluster over Hanover Street in Boston Massachusetts. We have been able to cluster arbitrary cluster shapes because we are using the DBSCAN algorithm, where 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.



Methodology

Data visual exploration

- Downtown Dallas, TX.
- To the left, multiple clusters clumped into a big cluster,
- To the right, multiple cluster, very close together, but not becoming one,
- 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.



Methodology

Training data selection criteria

- **City diversity score >=** 50
- **Cluster size >=** 5

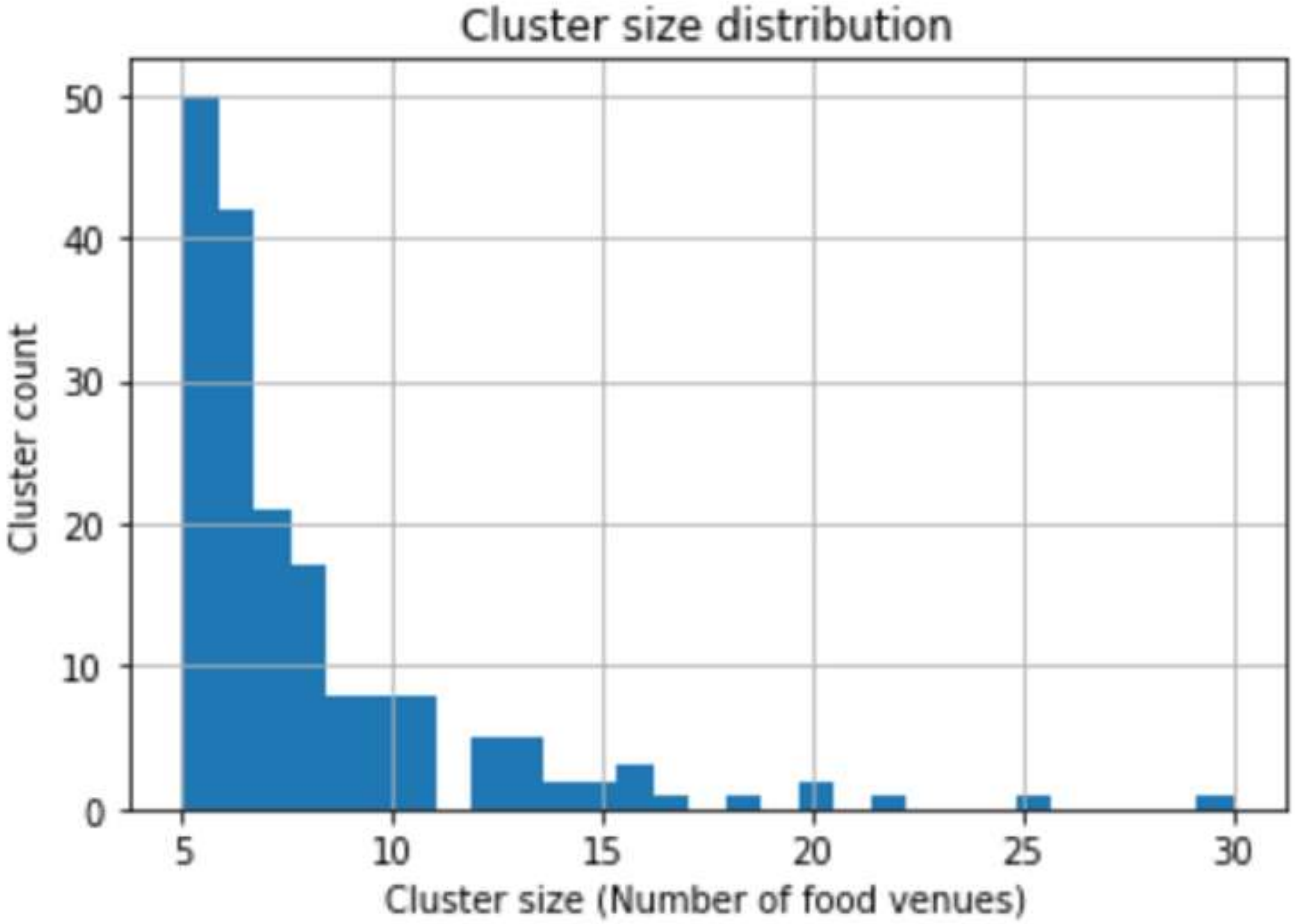
Data exploration facts

- Total venues in dataset: 1408
- Total clusters in dataset: 178
- Cluster min size: 5
- Cluster max size: 30
- Cluster mean size: 7.91

Methodology

Cluster size distribution

- 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

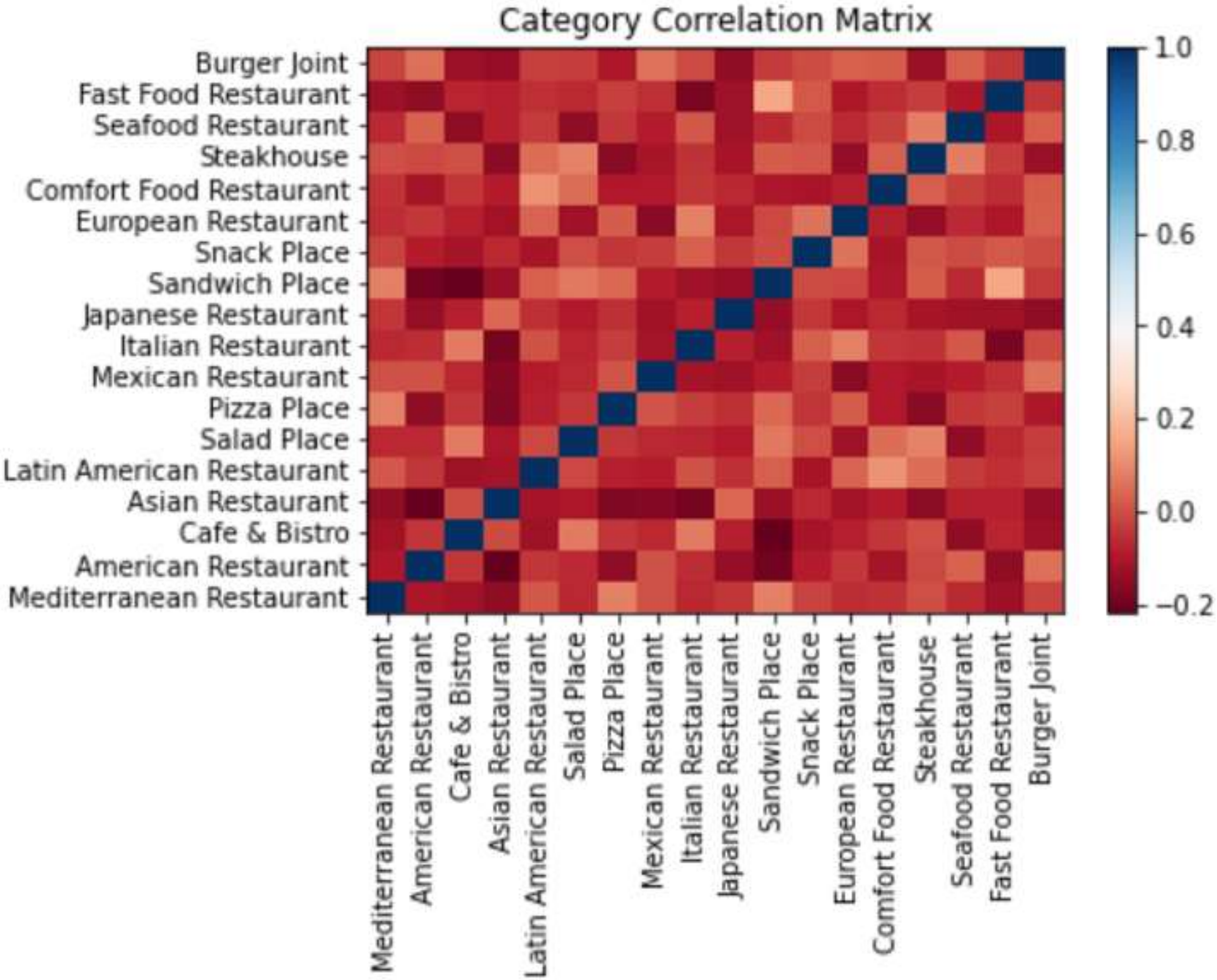


Methodology

Correlation Heat Map

- 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.



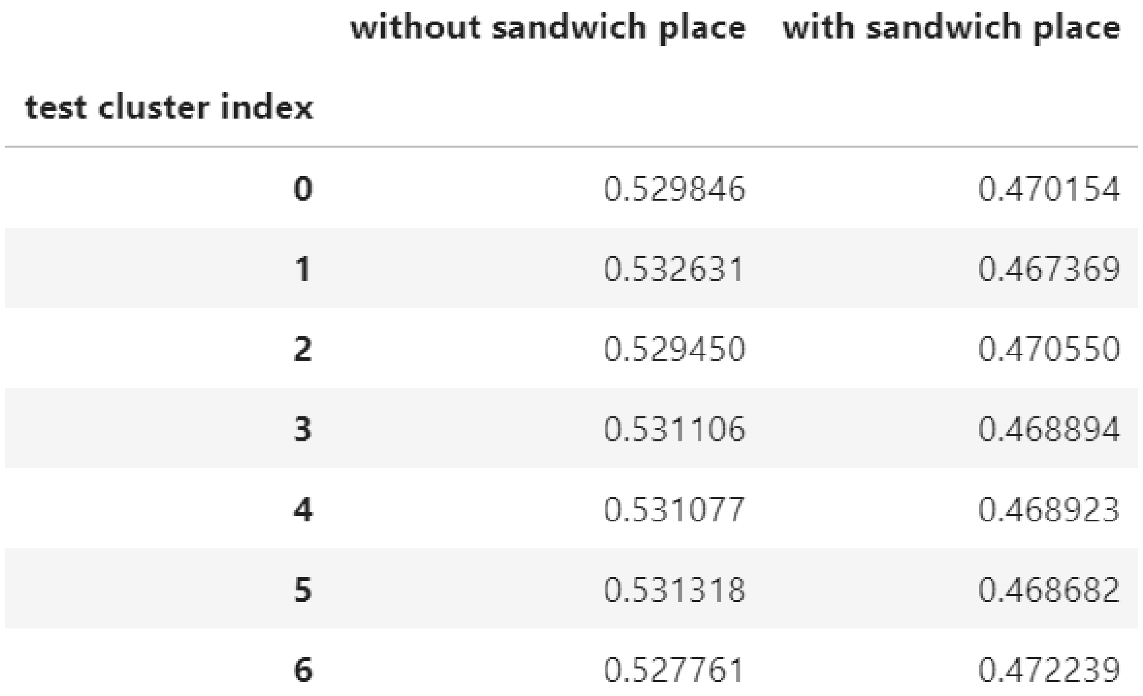
Results

Sandwich Place & Asian Resturant Case Studies

Sandwich Place Model

Test dataset results

- 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..

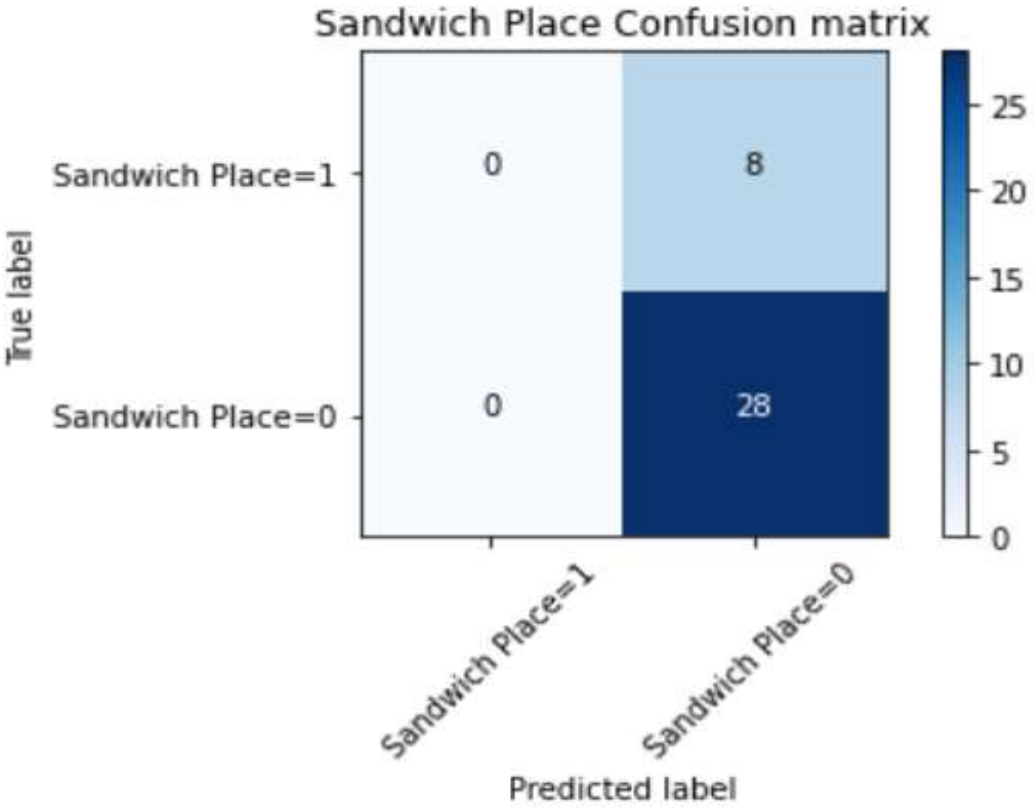


	without sandwich place	with sandwich place
test cluster index		
0	0.529846	0.470154
1	0.532631	0.467369
2	0.529450	0.470550
3	0.531106	0.468894
4	0.531077	0.468923
5	0.531318	0.468682
6	0.527761	0.472239

Sandwich Place Model

Model Evaluation

- 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%.
- **Jaccard similarity score >=** 0.7778
- This is surprisingly high score give so little probability difference in the model results.



Sandwich Place Model

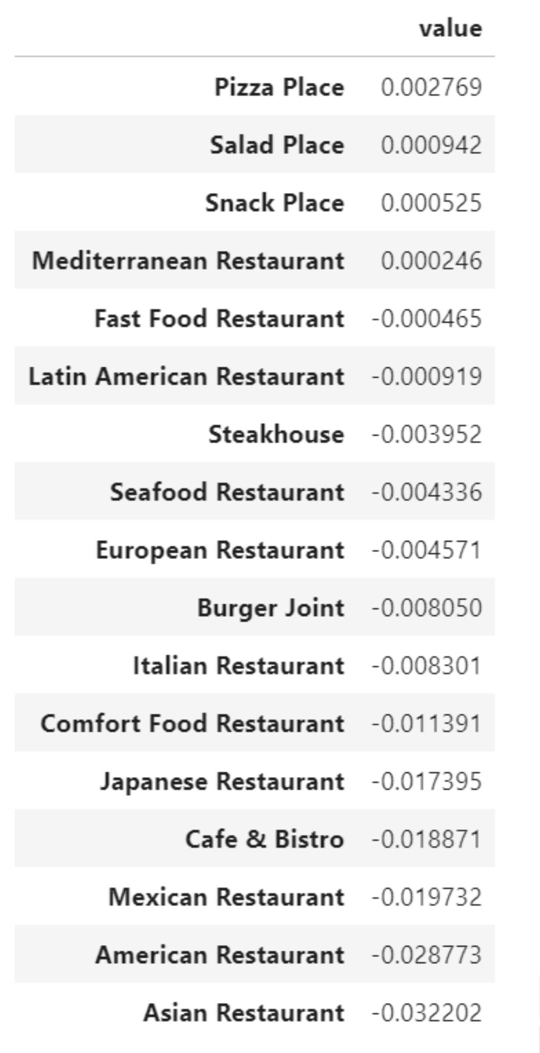
Model Coefficients

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

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

- The coefficients are listed from the greater to the lesser value:

- 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.

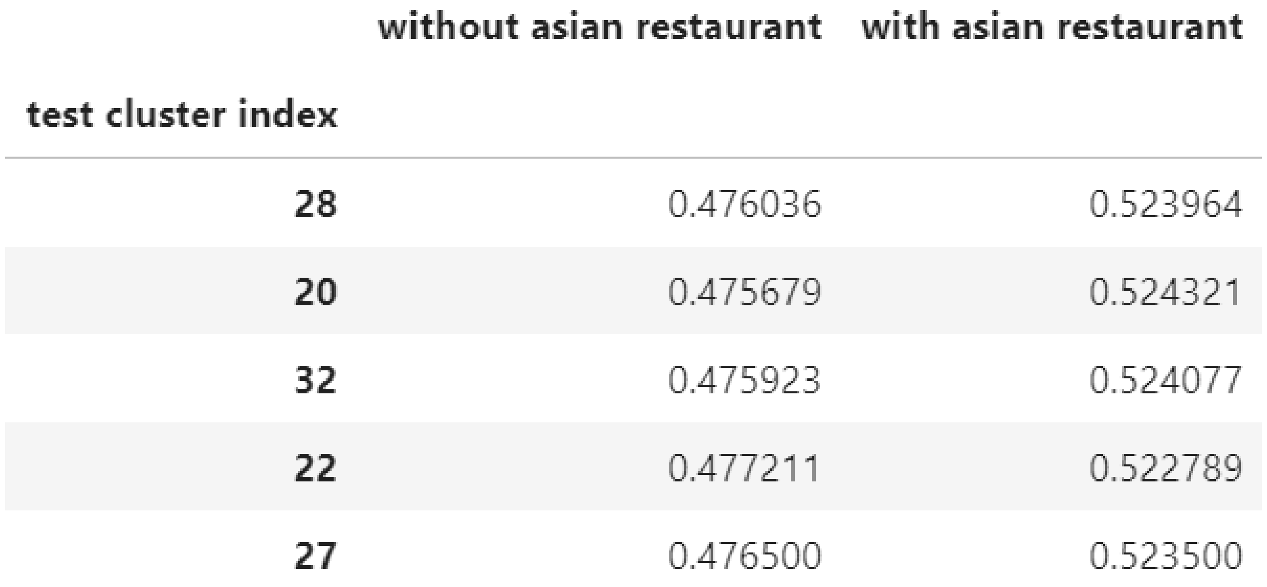


	value
Pizza Place	0.002769
Salad Place	0.000942
Snack Place	0.000525
Mediterranean Restaurant	0.000246
Fast Food Restaurant	-0.000465
Latin American Restaurant	-0.000919
Steakhouse	-0.003952
Seafood Restaurant	-0.004336
European Restaurant	-0.004571
Burger Joint	-0.008050
Italian Restaurant	-0.008301
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

Asian Restaurant Model

Test dataset results

- 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.

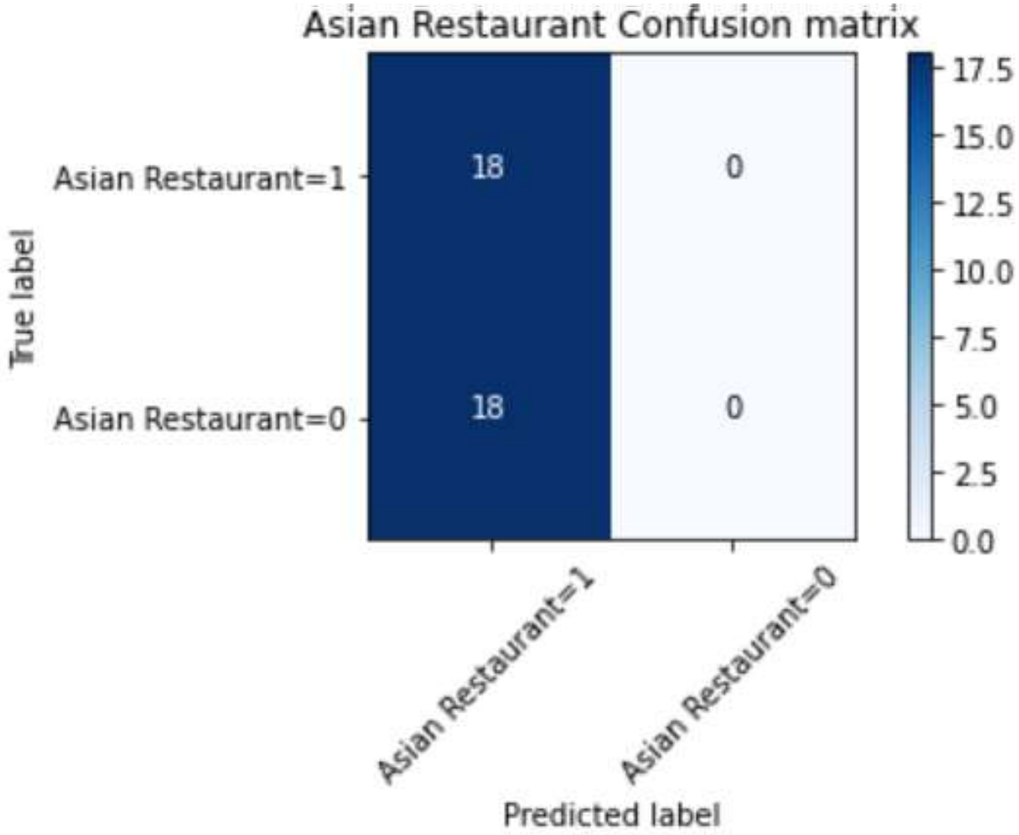


	without asian restaurant	with asian restaurant
test cluster index		
28	0.476036	0.523964
20	0.475679	0.524321
32	0.475923	0.524077
22	0.477211	0.522789
27	0.476500	0.523500

Sandwich Place Model

Model Evaluation

- 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)
- **Jaccard similarity score =** 0.5
- The cluster composition give no information about the probability of an Asian Restaurant to be present.



Conclusions

- 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.

Thank you