The Hidden Taste Of US ZipCodes

Ariel Z

Coursera DataSience Course

Abstract

The goal of this research is to explore if we can utilize food venues and sort of cuisine to understand a pattern in the test people across zip codes. During the research, data was collected from three different sources and organized for analysis. The analysis has not found a way to segment the zip codes by food categories in a significant way.

A second analysis was done to check whether or not we can predict the price of a house or at least if the types of restaurants around have any correlation with it; The analysis found a small correlation and almost zero correlation between the number of food venues and the price of houses around.

Introduction

Food is an important part of our life, but can we try and find a hidden pattern in the preference of an area towards a specific kind of cuisine? for example, do people who live in different streets/neighborhoods, have different preferences towards restaurants. In this project, I'm going to utilize online data about different restaurants across the US to try and answer this question.

Food is a huge business segment, and this kind of data can help business owners to identify the best location where they should consider opening their next venues; This is a hard thing to answer because it'll require also to understand whether we prefer to open in an area where this type of cuisine is one of the most favorite or the opposite.

As part of the research I'm using housing price as a benchmark for economical data, the research will try to combine the restaurant's data with the values of the houses in the area. Using this combination to look for a correlation between the price and the sort of food venues, if such even exists. And if actually, such a correlation exist, should house owner prefers certain kind of restaurants venues around their block? Can this influence the price of their property?

Data Collection

Data Sources

For this research I'm going to use three data sources:

1. OpenDataSoft us zip code by geo locations.
2. Zillow data research for the house values data - this data includes measurements from 1998, I downloaded the file and removed all price points except the latest.
3. Forsquare - to get the different food venues we will be utilizing forsquare API, the venue search function.

Data Cleaning

Starting from the US zip codes data and the houses prices.

Looking at the size of both datasets there doesn't match, we have 43,191 zip codes versus 30,230 house prices. Also, we have some NA values in the hose dataset.

After aggregating both data sets and removing the null values we are left with a combined data set that includes 30,006 different zip codes with their associated geolocation, and house values.

Checking the number of states. The US has 50 states but the data has 53. With the help of Wikipedia,[1] I found out that the US has 5 inhabited territories which can also be counted as states, and DC which is considered to be a separate state.

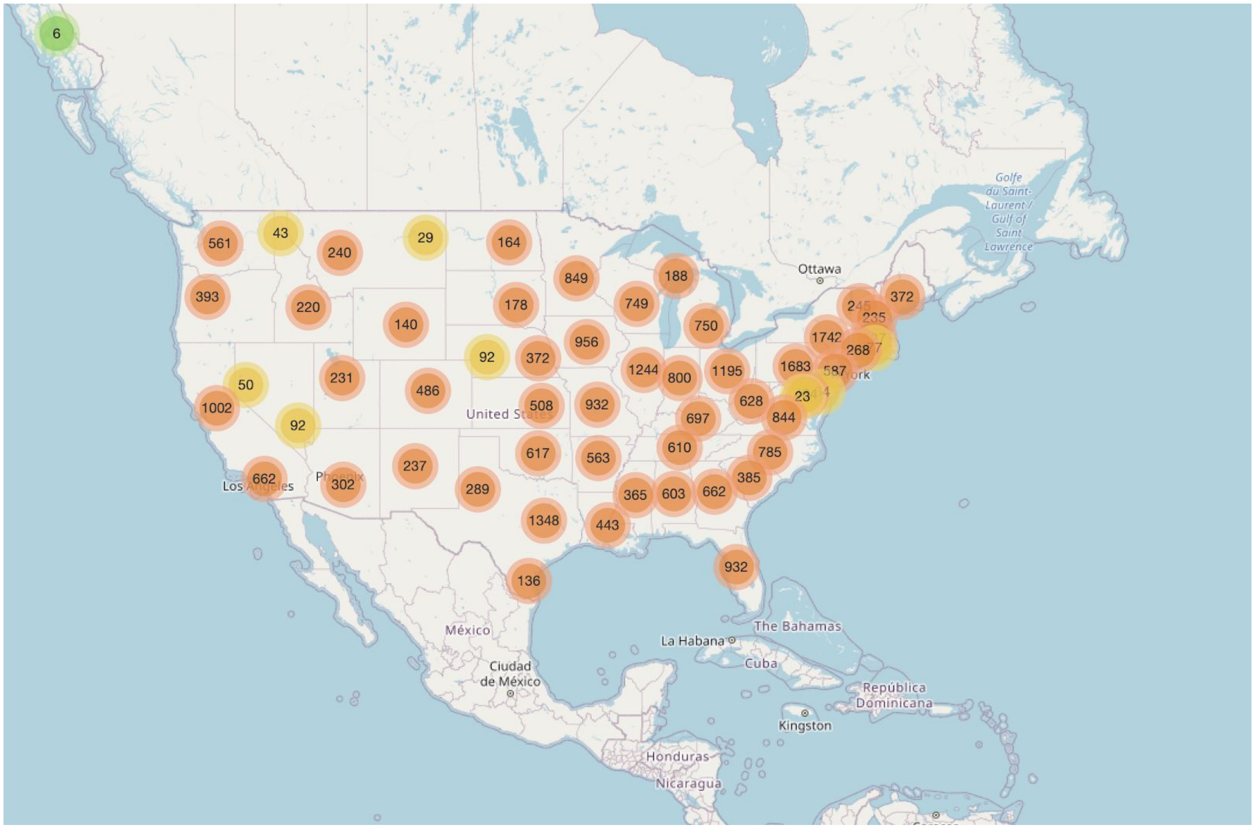
We had 2 inhabited territories in our data, PR, VI, and the state of DC.

I removed the data for PR and VI and left DC as part of the data.

[1] https://en.wikipedia.org/wiki/List_of_states_and_territories_of_the_United_States

Validating location data. Using a python package named folium we can plot all our zip code information on a map. It's hard to manually validate all the points but by plotting them on a map we should expect to have the whole US covered and all the markers on the map should be covering the US territory and not any other.

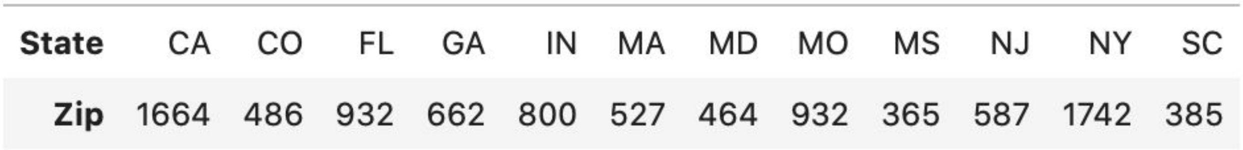
The plot worked as expected:



Selecting a state. Since the Foursquare API is limiting the number of requests and for each zip code we have to do a separate request I wanted to limit the number of states we are working with. I decided to use a box plot to check the different house prices across all states. Since I’m looking to find something that is unknown and compare different states I decided to choose those with the highest variation in prices. The process was done manually by looking at the plot.

I choose the following states: CA, FL, NY, MD, CO, MA, WA, IN, SC, GA, MS, MO.

The figure below display the number of zip codes in each state:



State	CA	CO	FL	GA	IN	MA	MD	MO	MS	NJ	NY	SC
Zip	1664	486	932	662	800	527	464	932	365	587	1742	385

Restaurant data. The retrieved data contained around 160,000 restaurants across 381 different categories. The most frequent restaurant in our dataset is Starbucks.

Although we have so many different categories, 196 of them contain less than 10 venues, so to remove the dimensionality of the data it decided to remove those categories; Finally, we ended up with a dataset containing 184 categories and 158,917 restaurants.

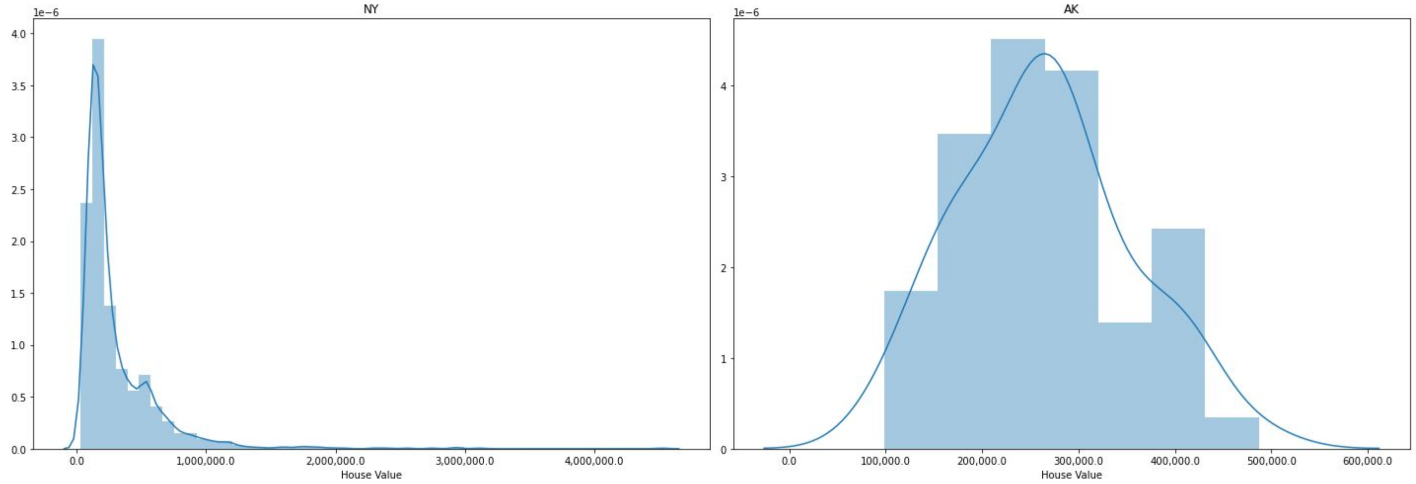
Missing restaurant categories. The restaurant data contained very general categories ‘Food’ and ‘Restaurant’ which does not give a lot of information. I found out that some of those exist in the data set in other locations and do have a more descriptive category (this makes sense considering Foursquare data is public and each user can choose what kind of data to enter). In total, the data contained 4200 items with the category restaurant and around 5000 with the category Food. This method managed to fill up around 500 food categories and 600 restaurants. Still missing 7,800 more descriptive categories I decided to remove them and prevent a bias in the data.

Data Exploration

Exploring The Data

After cleaning the data we have 151,000 different venues, with 184 different categories. On top of that, we also have our zip code location data state cities and house values. During this section, we will explore the data to better understand how it looks.

House prices. House prices in the same area are expected to be around a certain mean value or have what we call a bell curve. By plotting the prices in each state we can check if the data make sense. We can see for example that in the state of NY the price bell curve is around 0-1 M$ but they have a very long tail towards higher prices, whereas in the state of AK there are centered around 300K $ and the bell is pretty much centered.



Restaurant distribution by frequency. To better understand the data, sorting the different zip codes by the frequency of a restaurant category can help us better understand how it looks. By converting the data into dummy variables, each category becomes a feature, we can group the data by zip code and calculate the mean of data in each row; This method allows us to sort the data by the most frequent categories.

By this method, the most frequent type of restaurant is an American Restaurant which is probably not a big surprise since we are looking at the US data. But the second most frequent is a Pizza place - who doesn't live a hot slice of pizza.

On the second most frequent we can see Wings places leading the chars and then again a pizza place (the duplication with the first place is ok, since the second aggregate those zip codes where pizza wasn’t leading the first place).

In the third most common category, pizza is leading the chart and then a Mexican restaurant.

This kind of mixup between the places can be a problem when trying to cluster the data.

The full chars for this part available in Appendix A.

Leading category in each state. In an attempt to look at the data by different states, here we grouped the zip code data by states and sorted by the frequency of each category in each state. 15 categories in total for each state.

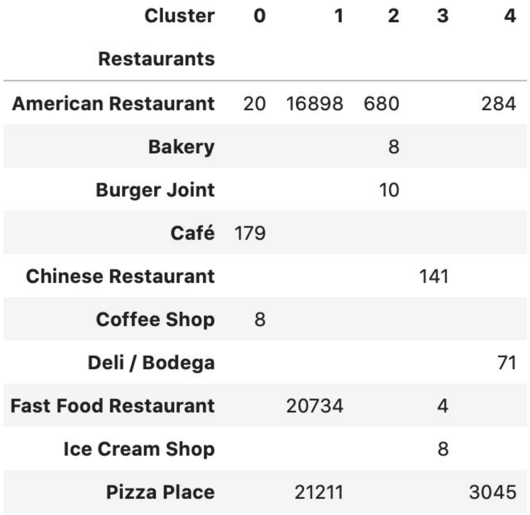
A list of the leading places in each state, the full plot is available under Appendix B:

- California - Mexican Restaurant
- Florida - A pizza place
- Colorado - Mexican Restaurant
- Georgia - American Restaurant
- Indiana - Pizza Place
- Massachusetts - Pizza Place
- Maryland - Pizza Place
- Missouri - Pizza Place
- Mississippi - Fast Food
- New Jersey - Pizza Place
- New York - Pizza Place
- South Carolina - American Restaurant

Clustering Manually. Our data has a lot of different categories which makes it impossible to explore by hand, so one approach can be to explore by manually clustering it.

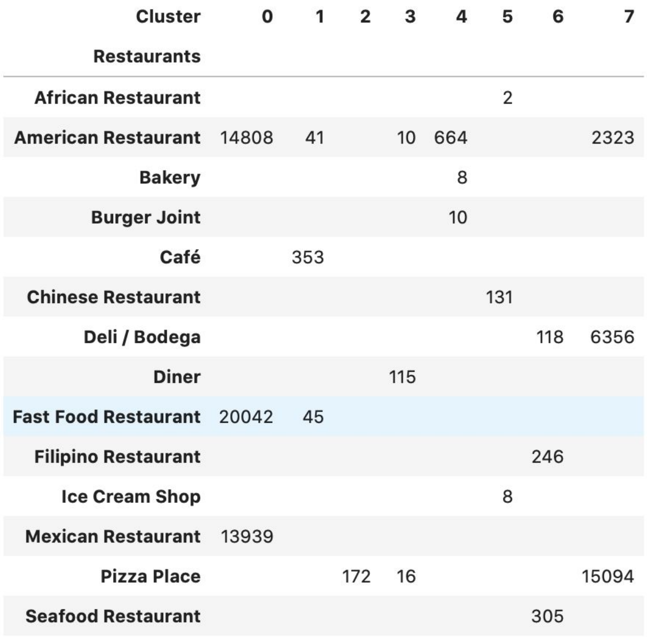
Using a basic K-Means algorithm a two clustering attempt were made one with K=5 and another with K=8

Both pivots displayed below with the top 3 for each cluster (leading zeros count):



Cluster	0	1	2	3	4
Restaurants					
American Restaurant	20	16898	680		284
Bakery			8		
Burger Joint			10		
Café	179				
Chinese Restaurant				141	
Coffee Shop	8				
Deli / Bodega					71
Fast Food Restaurant		20734		4	
Ice Cream Shop				8	
Pizza Place		21211			3045

8 cluster top categories



Cluster	0	1	2	3	4	5	6	7
Restaurants								
African Restaurant						2		
American Restaurant	14808	41		10	664			2323
Bakery					8			
Burger Joint					10			
Café		353						
Chinese Restaurant						131		
Deli / Bodega							118	6356
Diner				115				
Fast Food Restaurant	20042	45						
Filipino Restaurant							246	
Ice Cream Shop						8		
Mexican Restaurant	13939							
Pizza Place			172	16				15094
Seafood Restaurant							305	

5 cluster top categories

we can see that by increasing the number of clusters it's easier to differentiate the locations. The increased number of clusters helped to differentiate the data and brought more ethnic restaurants as criteria, But the result still has some duplications between the clusters for example pizza is the leading category in both clusters 2 and 7.

Model Improvement

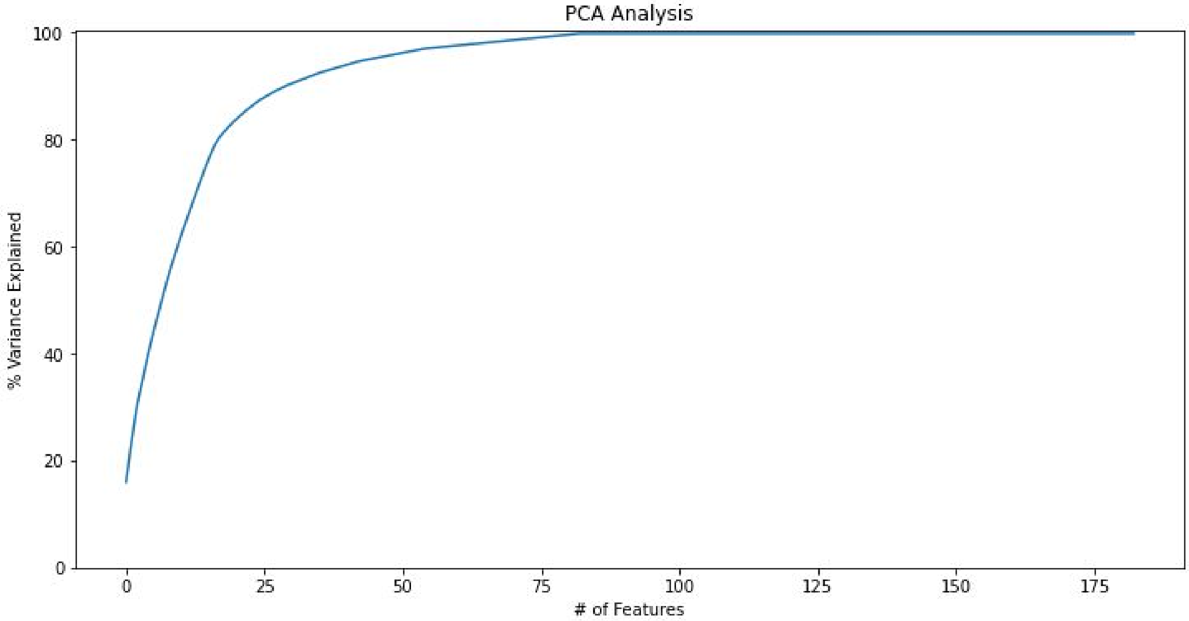
K-means vs DBSCAN. Kmeans is very simple although efficient clustering algorithm, his biggest drawback is that we have to define the number of clusters.

K Means segment the data into clusters based on their distance from each other.

DBScan is based on the density of the area, the idea is that a cluster formed if we have enough points around each center. We can specify the distance and the number of min points to form a cluster but we don’t specify the number of clusters in advance.

Feature Selection

PCA. It's a commonly used method to reduce the dimensionality of the data, the algorithm runs a test to identify the number of components we need to explain a certain percentage of our data set. In this data, we can explain 95% of the data using just 44 components.



The goal of this research is to find a pattern that will allow us to differentiate between the different zip codes, so we can run two separate tests, one that will use the top 44 features the PCA algorithm has found and a second test that will use only those not selected by PCA. By running those two tests we can try and check whether those small venues that can't explain a big sample of the data can do better in clustering it.

In the analysis section, the following tests will be performed, Kmeans, DBScan, and Gaussian Mixture. Utilizing three sets of features, a full list of them, PCA selected, and the opposite of the list selected by the PCA.

Data Analysis

Clustering

To run our test analysis for each one of the algorithms, we have to set a function to run each one of them with a certain type of variable. They all must be kept the same for all features set.

K Means configured with 40 random states and a range of clusters between 2-20

DBSCAN is configured with a min_sample of 9 and a step of 0.1 in the range 0.1-1

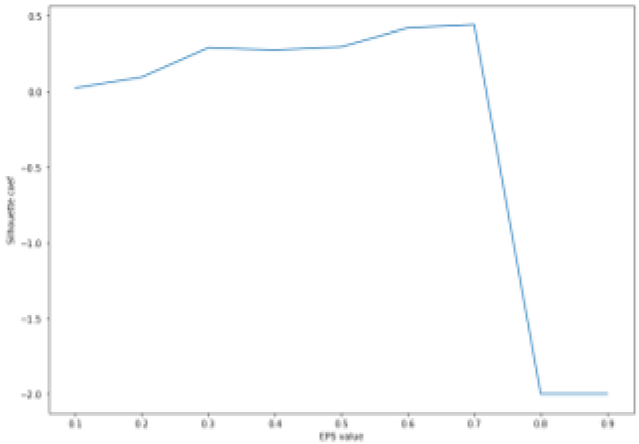
GaussianMixture is configured with the following test data, covariance_type='full', a number of components in the range of 1-20, and random state of 10. During all the tests this algorithm always returned 2 clusters and a negative silhouette coefficient score.

Comparing K-means and DBScan

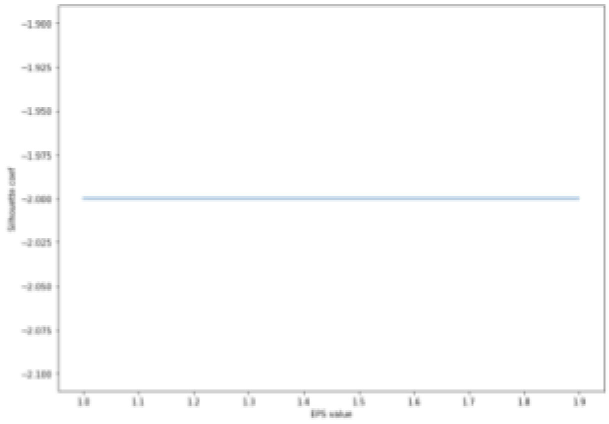
With a full list of features running the scoring test has resulted in a sil_c (=silhouette coefficient) of 0.44 and a eps of 0.7, as can be seen in the graph it couldn't find clusters above a value of 0.7 During the tasting, I have tried a value height then 1 to understand if this worth exploring but provided no clustering.

It’s important to note that the sil_c can score values between -1 and 1 so in places where the graph displays values of -2 it means no clusters found or 1 cluster found (if only one cluster found the sil_c can’t be tested)

<u>Here is one example of the result when testing dbscan with both ranges (1-2, 0.1-1):</u>



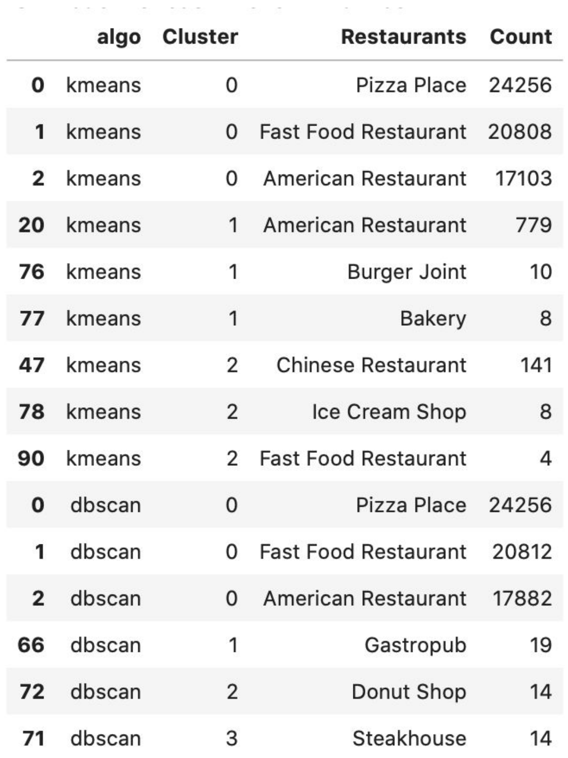
K-means test range 0.1-1



K-means test range 1-2

After clustering the data the differences between the categories selection can be observed. K-means seems to be able to partition the data in a better way: #1 is based on pizza and fast food, Cluster #2 American restaurants, Cluster #3 Chinese restaurants.

<u>A full table result:</u>



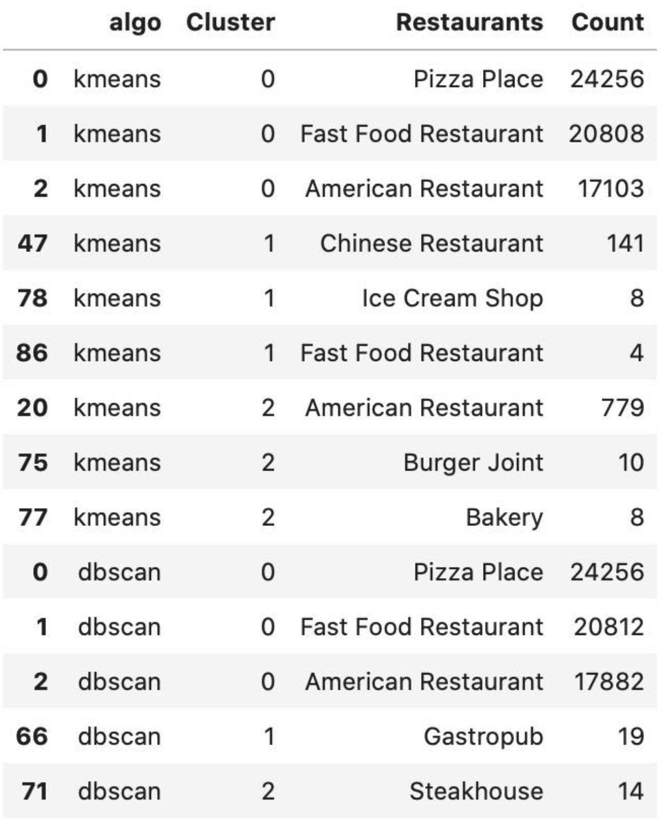
	algo	Cluster	Restaurants	Count
0	kmeans	0	Pizza Place	24256
1	kmeans	0	Fast Food Restaurant	20808
2	kmeans	0	American Restaurant	17103
20	kmeans	1	American Restaurant	779
76	kmeans	1	Burger Joint	10
77	kmeans	1	Bakery	8
47	kmeans	2	Chinese Restaurant	141
78	kmeans	2	Ice Cream Shop	8
90	kmeans	2	Fast Food Restaurant	4
0	dbscan	0	Pizza Place	24256
1	dbscan	0	Fast Food Restaurant	20812
2	dbscan	0	American Restaurant	17882
66	dbscan	1	Gastropub	19
72	dbscan	2	Donut Shop	14
71	dbscan	3	Steakhouse	14

PCA selection

The algorithm selected 44 features. We can look at them and check which features were selected, the table in Appendix C shows the features list sorted from most freq to the least.

The clustering, in this case, is the same as with the raw data, which is equivalent to the fact the PCA utilizing the most explanatory variables. But using the PCA data our s_ci from dbscan is improved from 0.44 to 0.46.

PCA clustering results by categories:



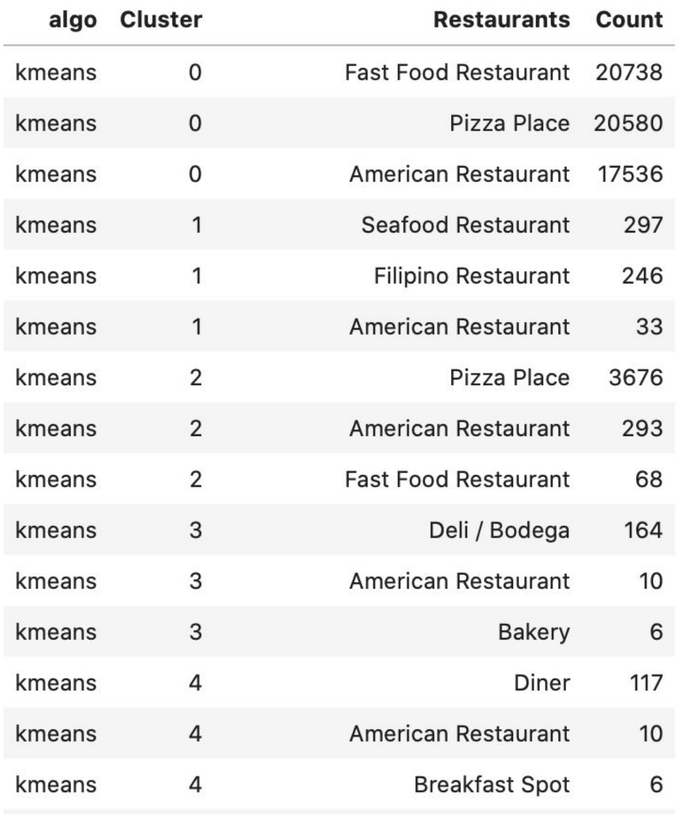
	algo	Cluster	Restaurants	Count
0	kmeans	0	Pizza Place	24256
1	kmeans	0	Fast Food Restaurant	20808
2	kmeans	0	American Restaurant	17103
47	kmeans	1	Chinese Restaurant	141
78	kmeans	1	Ice Cream Shop	8
86	kmeans	1	Fast Food Restaurant	4
20	kmeans	2	American Restaurant	779
75	kmeans	2	Burger Joint	10
77	kmeans	2	Bakery	8
0	dbscan	0	Pizza Place	24256
1	dbscan	0	Fast Food Restaurant	20812
2	dbscan	0	American Restaurant	17882
66	dbscan	1	Gastropub	19
71	dbscan	2	Steakhouse	14

Testing non-PCA

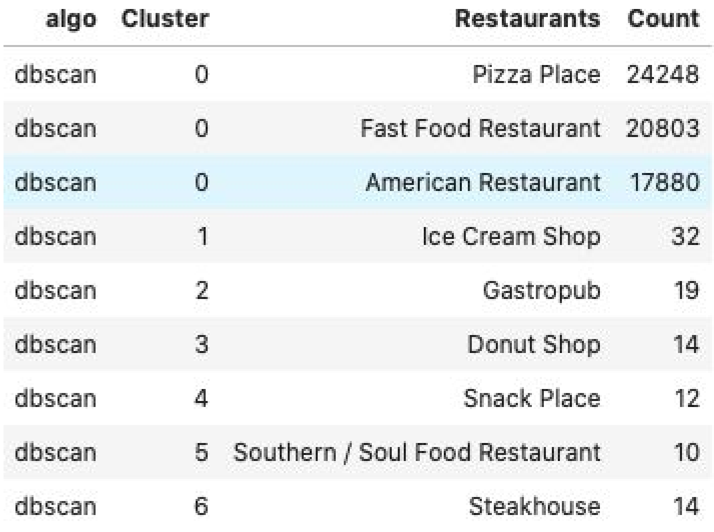
I decided to work with those categories PCA not choose, the reason as explained before, we are trying to find the best way to segment zip code, and just looking at the top venues may be preventing that.

The K-means best result is K=5, with a much lower coefficient, of 0.31, this time our leading categories changes a little bit: Clust 1-Fast food, Cluster 2 - SeaFood restaurant, Cluster 3 - Pizza place (although the second most common venue in cluster 1 is pizza).

The dbscan managed to create 7 clusters with this data set however the first one captures most of the data and the rest are very small, the coefficient has improved as well to 0.55



algo	Cluster	Restaurants	Count
kmeans	0	Fast Food Restaurant	20738
kmeans	0	Pizza Place	20580
kmeans	0	American Restaurant	17536
kmeans	1	Seafood Restaurant	297
kmeans	1	Filipino Restaurant	246
kmeans	1	American Restaurant	33
kmeans	2	Pizza Place	3676
kmeans	2	American Restaurant	293
kmeans	2	Fast Food Restaurant	68
kmeans	3	Deli / Bodega	164
kmeans	3	American Restaurant	10
kmeans	3	Bakery	6
kmeans	4	Diner	117
kmeans	4	American Restaurant	10
kmeans	4	Breakfast Spot	6



algo	Cluster	Restaurants	Count
dbscan	0	Pizza Place	24248
dbscan	0	Fast Food Restaurant	20803
dbscan	0	American Restaurant	17880
dbscan	1	Ice Cream Shop	32
dbscan	2	Gastropub	19
dbscan	3	Donut Shop	14
dbscan	4	Snack Place	12
dbscan	5	Southern / Soul Food Restaurant	10
dbscan	6	Steakhouse	14

Removing common places

We have seen that a few places rule most of the data and the clustering towards those categories. Checking how they interact if I remove the most common places from the features list, Fast Food, Pizza, American Restaurant. Since the clusters are always built around them, I assume that we can do better without this information.

The result of this experiment shows no improvement, the DB scan algorithm managed to create many more clusters but there weren't significant in a way that can tell us anything.

So removing variables doesn't do us any good.

It's important to note that although they have clustering was done without those categories they can still appear when looking at the cluster by leading places since we just removed them from the clustering. And there results in this case are low coefficents and most of the data in one or two clusters.

Test by Cities

Running those two clustering algorithms on the same data but grouped by state or city, the state and city data is based on the zip code dataset.

The result of the two algorithms was disappointing and didn't provide anything significant.

Summary Table - Clustering

The table summaries the best silhouette coefficient results:

coefficient value, number of clusters)

	All features	PCA selected	Non pca selected	Not common	Cities
K-means	0.39 / 3	**0.41 / 3**	0.31 / 5	NA / 2	0.44 / 3
DBScan	0.44 / 4	0.46 / 3	**0.55 / 7**	0.4 / 14	-0.2 / NA

The food venues data found to be not significant in clustering different zip codes.

Regressions

We haven’t been able to find any good clustering of the zip codes by the type of restaurants.

The next part of the research is to test if any kind of correlation between the type of a restaurant and the price of a house exists. Hoping to use the clustering as a part of the testing of the pricing wasn’t an option.

Another test that will be performed is regression analysis between the total number of restaurants and the price of a house._Since the data has a very low correlation in between I decided to use the following type of regression and compare them

1. Linear regression, just a simple benchmark to other regressions
2. Polynomial regression

3. Ridge regression,
4. Lasso regression, another type of regression that also performs some regularization, similar to ridge regression but with different penalty mechanisms.

Note: To make the number easier to read I divided the houses price by 1000

Linear regression

By running a linear regression trying to predict a house price based on the restaurants around the model R squared is just 0.25 which is not significant enough.

What I have found to be interesting is that some of the locations around have positive coefficients with the price. For example, a coffee shop in the same zip code has a coefficient of 60,000 and a juice bar a coeficient of 50,000.

At the end of this chapter, I'll provide a summary table of all the top coefficients found by all the different algorithms.

But since the R^2 is very low, a second approach is to select the best feature for linear regression using SKlearn package for feature selection; I'll combine both mutual information and f_regression, the first check how much mutual information two a param brings to the model and the later the f score of this param inside the model.

Using thos methods we were able to find 21 different restaurants categories that provide better information: Asian Restaurant, Bagel Shop, Bakery, Bubble Tea Shop, Burger Joint, Café, Chinese Restaurant, Coffee Shop, Deli / Bodega, Donut Shop, French Restaurant, Italian

Restaurant, Japanese Restaurant, Juice Bar, Mediterranean Restaurant, Mexican Restaurant, New American Restaurant, Pizza Place, Salad Place, Sushi Restaurant, Thai Restaurant.

After selecting this list of categories, our regression model scored an R^2 of 0.239 which is lower than with the full list of categories but we used just 11% of the categories and the values decreased by 4.4%

Since the data has a lot of features a polynomial regression was performed as well, leading to worse results than the linear one.

Ridge regression

A regression method that is mostly used when we have high collinearity in the data, but in this case I decided to use it as well since it can provide a recommended coefficient as well.

Tuning the algorithm using grid search from SKlearn, it recommended that the best alpha value will be 1500.

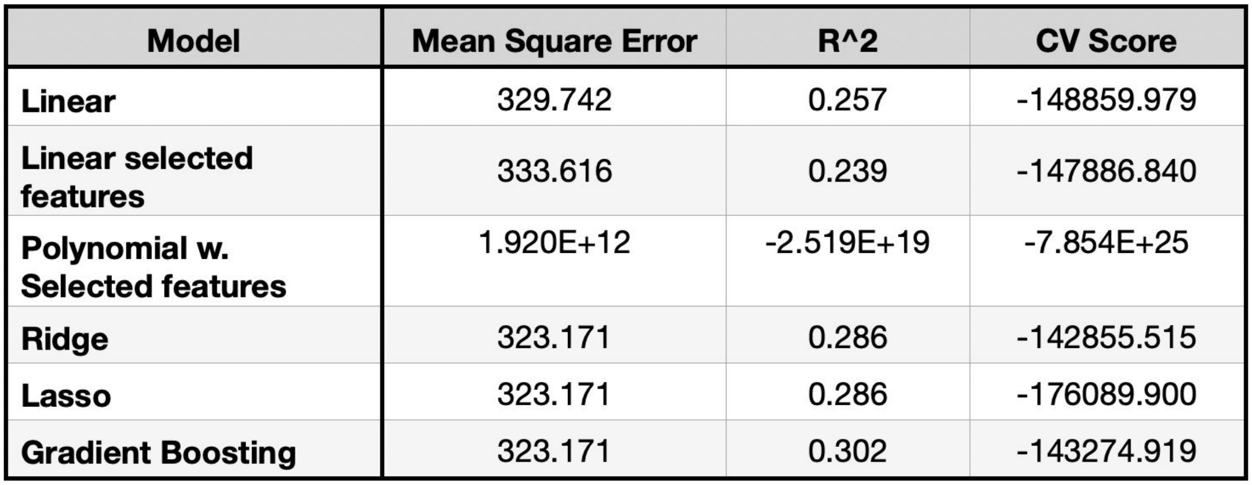
Lasso Regression

very similar to the ridge algo, but with a different coefficient. This algorithm can be useful for feature selection, since it tried to minimise those feature who found to be not important to zero. Tuning the algorithm using grid search from SKlearn, it recommended that the best alpha value will be 100.

Gradient Boosting regression

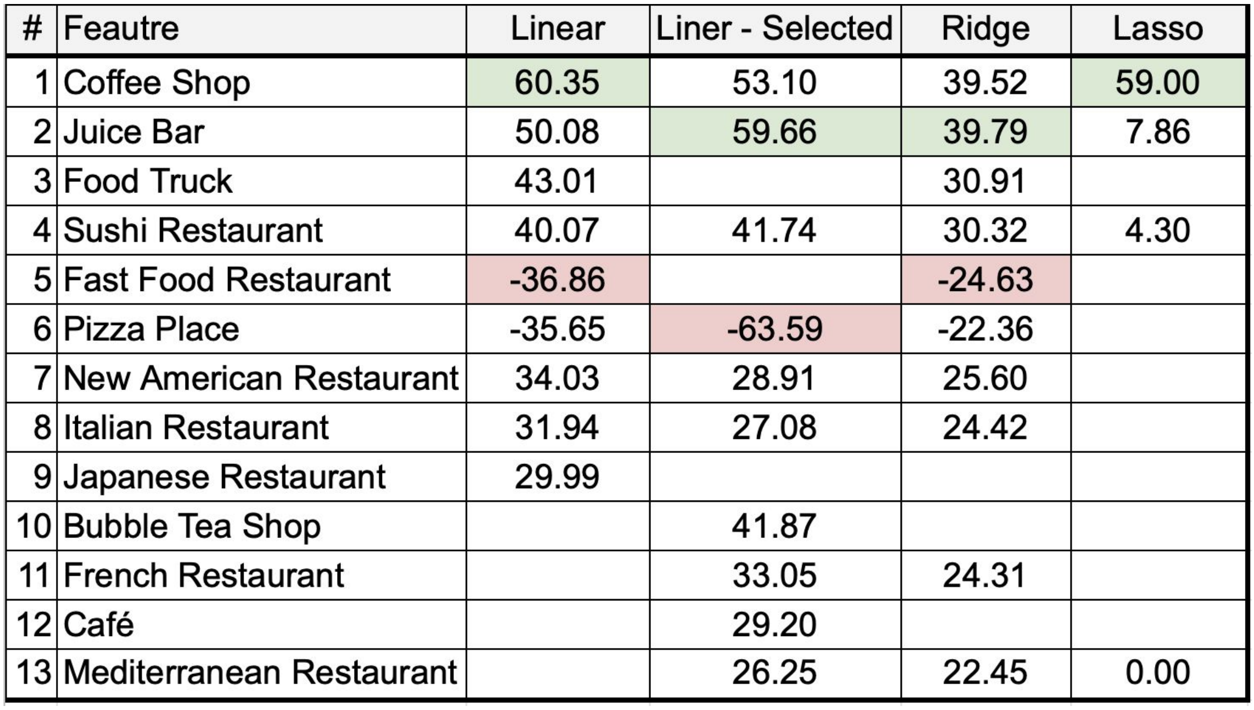
Boosting is a technique used in machine learning and can be used both in clustering and regression. The goal of this technique is to try and find the best features by minimizing a loss function. In this way even a params that seems weak can become significant if it can minimize the loss function. This algorithm depend on more parameters and tuning options, I used grid search to find the best learning rate and the optimal number of estimators, maximzing the feature using the sqrt algorithm. The values it found as best estimators, learning rate of 0.005 and number of estimators of 1001.

Summary Table - Regression



Model	Mean Square Error	R^2	CV Score
Linear	329.742	0.257	-148859.979
Linear selected features	333.616	0.239	-147886.840
Polynomial w. Selected features	1.920E+12	-2.519E+19	-7.854E+25
Ridge	323.171	0.286	-142855.515
Lasso	323.171	0.286	-176089.900
Gradient Boosting	323.171	0.302	-143274.919

Looking at the best coefficients our different regression algorithm has found (those with the heights values) can show that some of them may have an influence of the price, like a juice par and coffee witch have a positive coefficients and on the opposite it found places like fast foods and pizza places wih negative values. However since the score of all the different regression methods are at the lower end, I'll not give so params a segnifincant consideration.



#	Feautre	Linear	Liner - Selected	Ridge	Lasso
1	Coffee Shop	60.35	53.10	39.52	59.00
2	Juice Bar	50.08	59.66	39.79	7.86
3	Food Truck	43.01		30.91	
4	Sushi Restaurant	40.07	41.74	30.32	4.30
5	Fast Food Restaurant	-36.86		-24.63	
6	Pizza Place	-35.65	-63.59	-22.36	
7	New American Restaurant	34.03	28.91	25.60	
8	Italian Restaurant	31.94	27.08	24.42	
9	Japanese Restaurant	29.99			
10	Bubble Tea Shop		41.87		
11	French Restaurant		33.05	24.31	
12	Café		29.20		
13	Mediterranean Restaurant		26.25	22.45	0.00

House price to number of restaurants

Can the number of restaurants around a zip code be correlated to the price of a house? This was a question the data we have can answer. Summing up the number of restaurants in each zip code and running a simple linear regression with the price of a house it found an R^2 of 0.1; Which is not significant at all. A polynomial regression did slightly better improving the R^2 to 0.1.

Results

The research tried to answer a question whether or not we can identify a different food taste across zip codes in the US. and while doing so if we can find any correlation between the food venues categories and price of a house in the same zip code.

Clustering

The research couldn't identify any good way to partition the zip codes by different restaurant types. It is found out the American and Fast food are leading the top on almost all zip codes but this doesn’t help us to identify different preferences.

Regresion

Trying to understand if a price of house can be dependent on the type of a restaurant seems to be more promising. We have found out that there are 10 coefficients which represent the categories and have a positive impact with the price of the house.

Discussion

In a future work a combination of regular house features like, size, bedrooms conditions and etc can be combined with the top features found in this research. By combining them together we can try and answer the question if those actually can help predict a price when combined with the usual predictors. Only by doing so I believe that we can further understand doest it really significant factor or just a bias of this stand alone research.

As for the restaurant grouping a similar test can be done across different countries and in this way to remove a home bias toward national preferences that we can see in the same country.

References

[1] US Zip code dataset - OpenDataSoft

https://public.opendatasoft.com/explore/dataset/us-zip-code-latitude-and-longitude/export

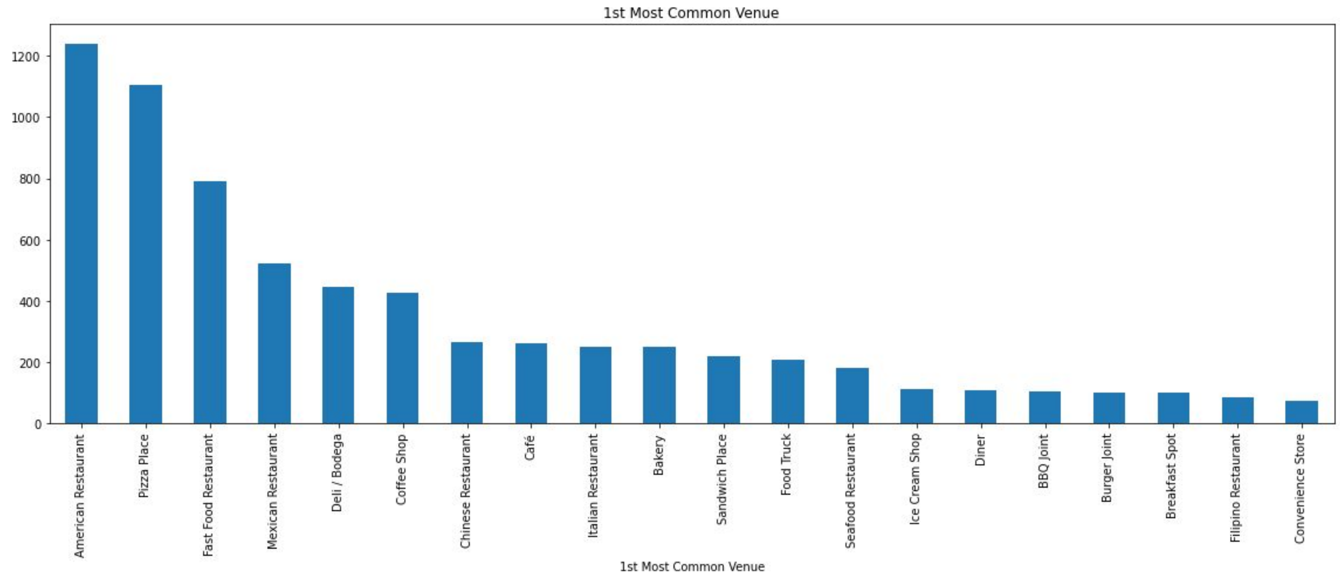
[2] House Values - Zillow research

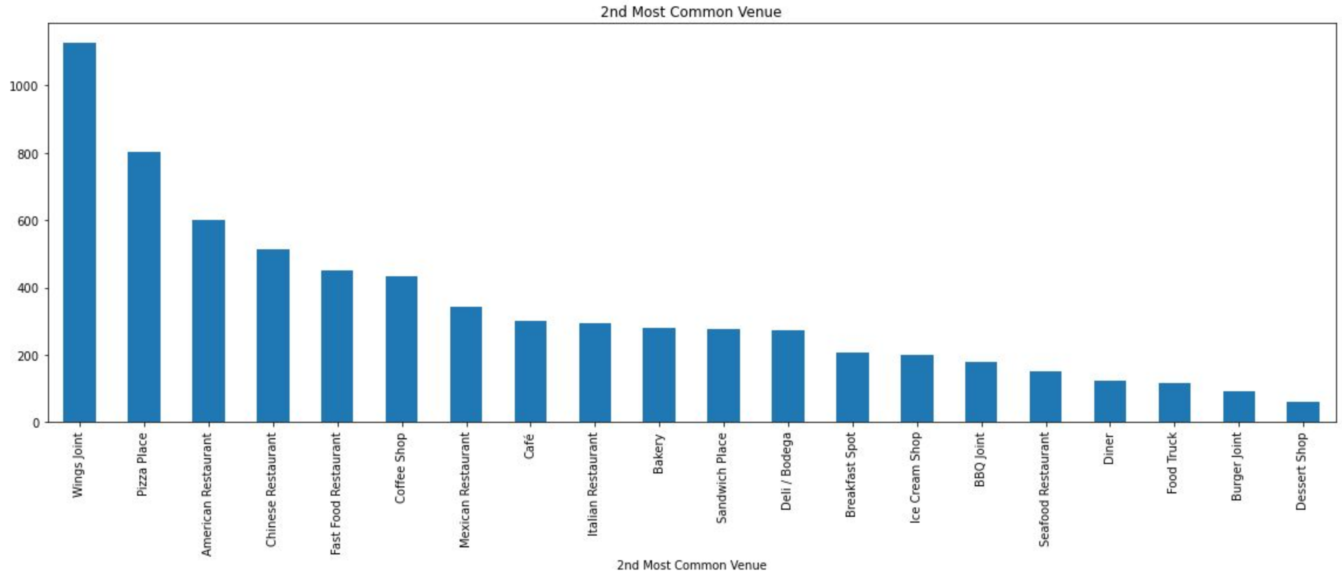
https://www.zillow.com/research/data/

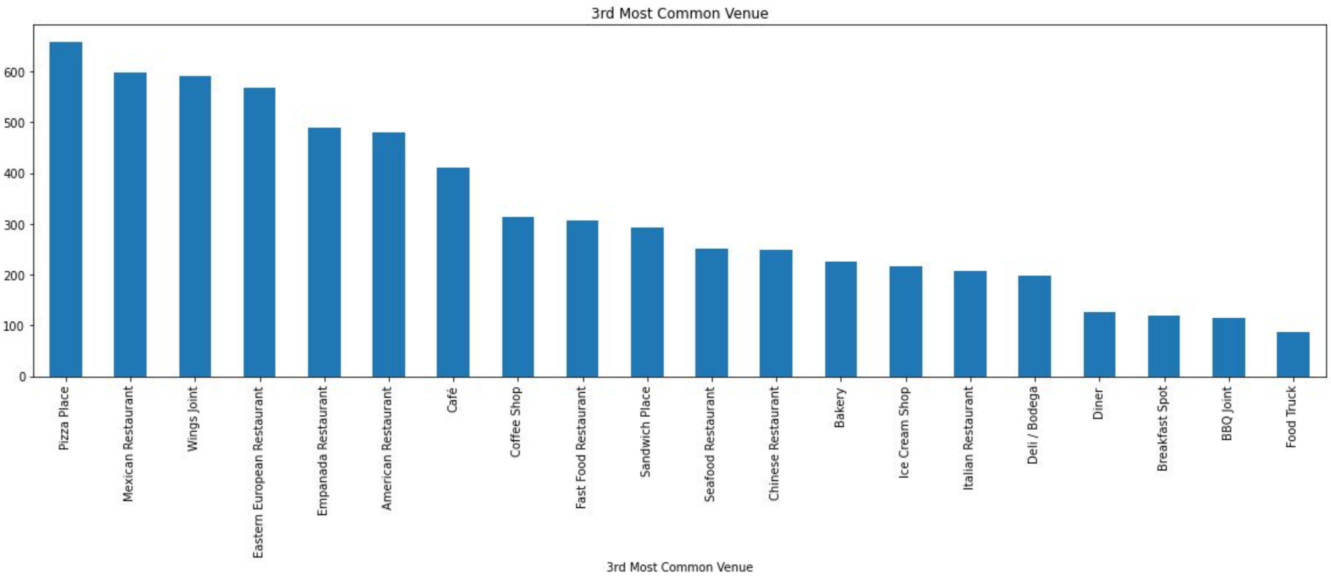
[3] Location data - Foursquare api

https://developer.foursquare.com/docs/api-reference/venues/search-enterprise/

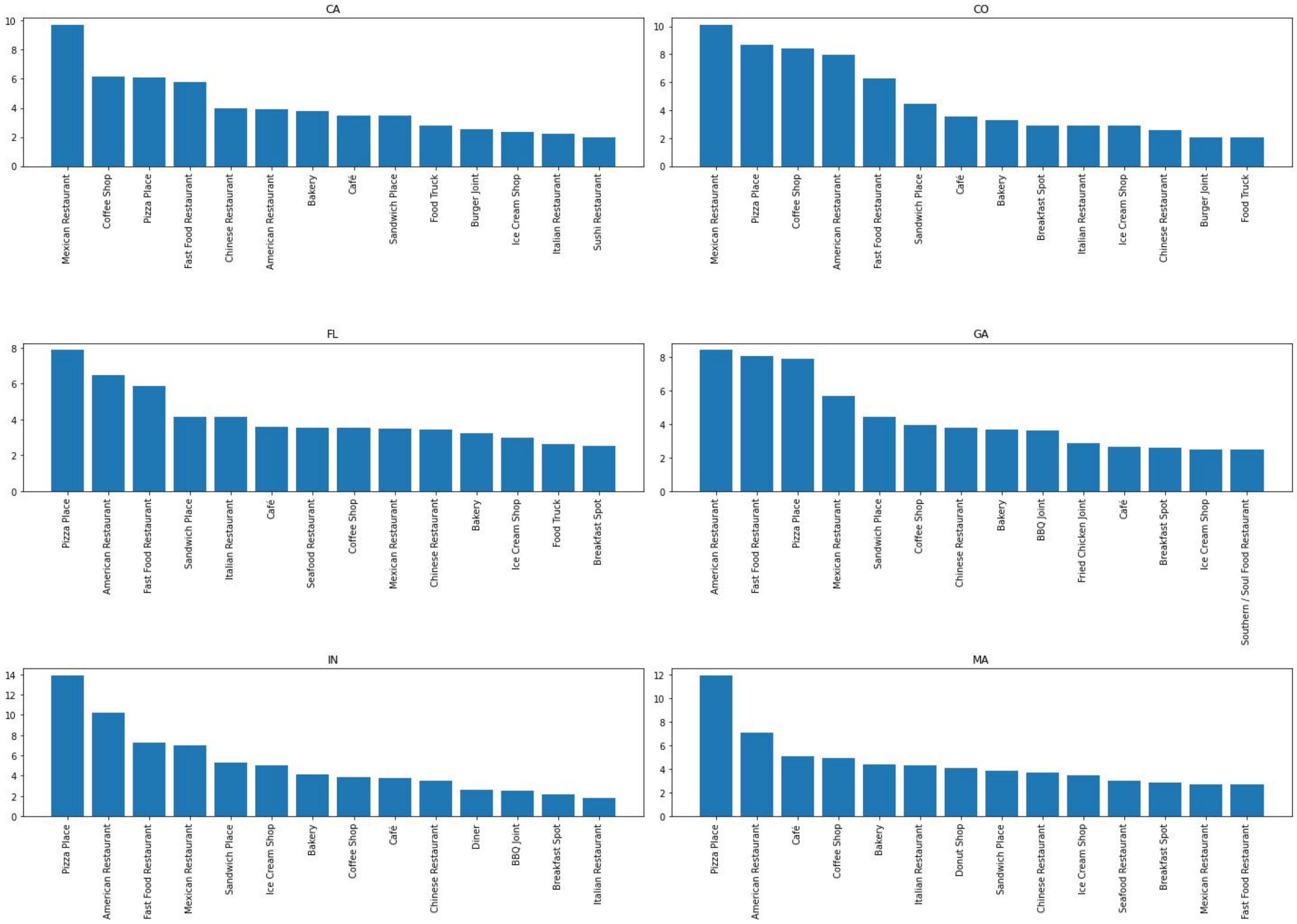
Appendix A

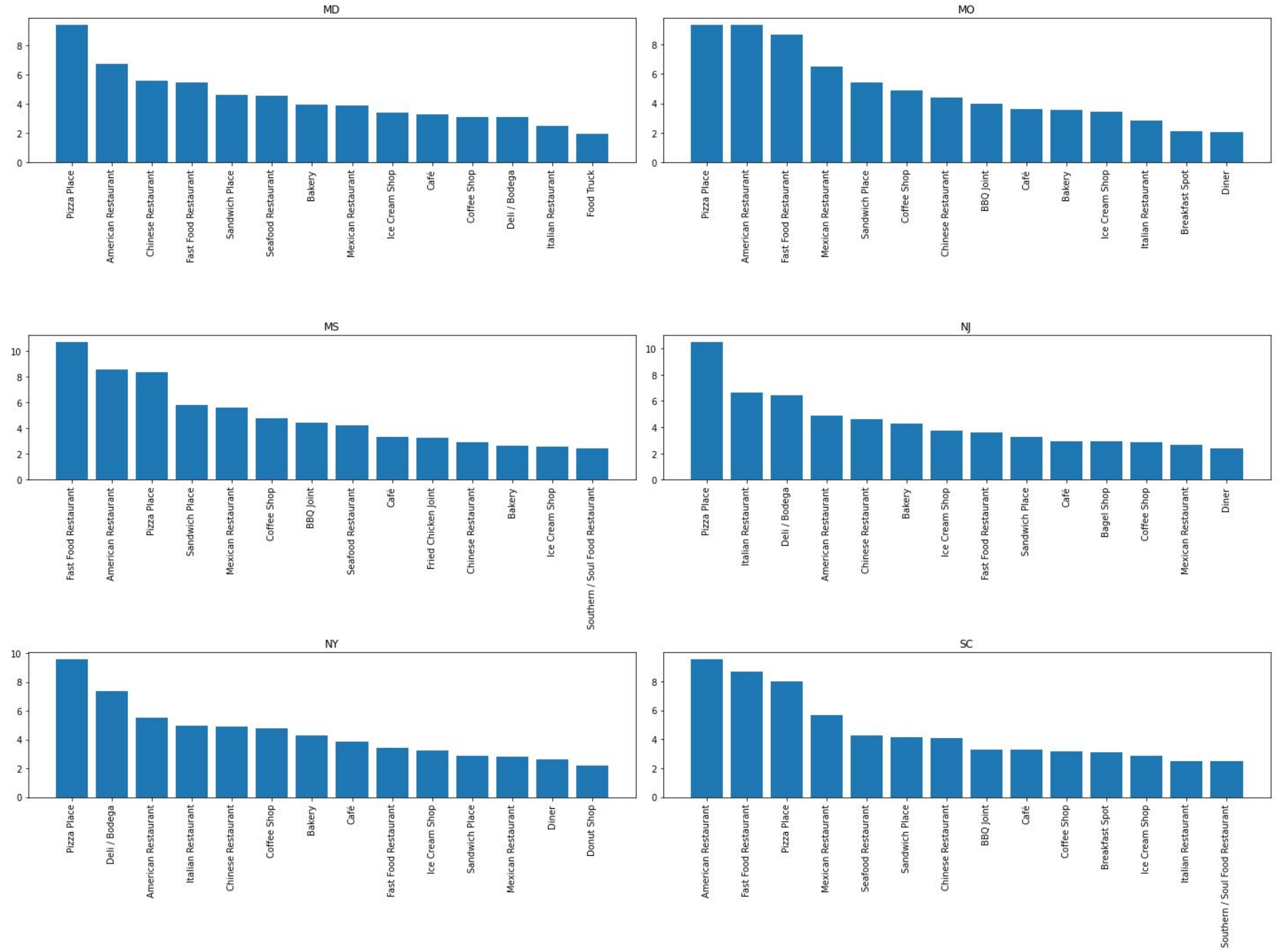

1st Most Common Venue
0
200
400
600
800
1000
1200
American Restaurant
Pizza Place
Fast Food Restaurant
Mexican Restaurant
Deli / Bodega
Coffee Shop
Chinese Restaurant
Café
Italian Restaurant
Bakery
Sandwich Place
Food Truck
Seafood Restaurant
Ice Cream Shop
Diner
BBQ Joint
Burger Joint
Breakfast Spot
Filipino Restaurant
Convenience Store
1st Most Common Venue


2nd Most Common Venue
0
200
400
600
800
1000
Wings Joint
Pizza Place
American Restaurant
Chinese Restaurant
Fast Food Restaurant
Coffee Shop
Mexican Restaurant
Café
Italian Restaurant
Bakery
Sandwich Place
Deli / Bodega
Breakfast Spot
Ice Cream Shop
BBQ Joint
Seafood Restaurant
Diner
Food Truck
Burger Joint
Dessert Shop
2nd Most Common Venue


3rd Most Common Venue
0
100
200
300
400
500
600
Pizza Place
Mexican Restaurant
Wings Joint
Eastern European Restaurant
Empanada Restaurant
American Restaurant
Café
Coffee Shop
Fast Food Restaurant
Sandwich Place
Seafood Restaurant
Chinese Restaurant
Bakery
Ice Cream Shop
Italian Restaurant
Deli / Bodega
Diner
Breakfast Spot
BBQ Joint
Food Truck
3rd Most Common Venue

Appendix B


CA
0
2
4
6
8
10
Mexican Restaurant
Coffee Shop
Pizza Place
Fast Food Restaurant
Chinese Restaurant
American Restaurant
Bakery
Café
Sandwich Place
Food Truck
Burger Joint
Ice Cream Shop
Italian Restaurant
Sushi Restaurant
CO
0
2
4
6
8
10
Mexican Restaurant
Pizza Place
Coffee Shop
American Restaurant
Fast Food Restaurant
Sandwich Place
Café
Bakery
Breakfast Spot
Italian Restaurant
Ice Cream Shop
Chinese Restaurant
Burger Joint
Food Truck
FL
0
2
4
6
8
Pizza Place
American Restaurant
Fast Food Restaurant
Sandwich Place
Italian Restaurant
Café
Seafood Restaurant
Coffee Shop
Mexican Restaurant
Chinese Restaurant
Bakery
Ice Cream Shop
Food Truck
Breakfast Spot
GA
0
2
4
6
8
American Restaurant
Fast Food Restaurant
Pizza Place
Mexican Restaurant
Sandwich Place
Coffee Shop
Chinese Restaurant
Bakery
BBQ Joint
Fried Chicken Joint
Café
Breakfast Spot
Ice Cream Shop
Southern / Soul Food Restaurant
IN
0
2
4
6
8
10
12
14
Pizza Place
American Restaurant
Fast Food Restaurant
Mexican Restaurant
Sandwich Place
Ice Cream Shop
Bakery
Coffee Shop
Café
Chinese Restaurant
Diner
BBQ Joint
Breakfast Spot
Italian Restaurant
MA
0
2
4
6
8
10
12
Pizza Place
American Restaurant
Café
Coffee Shop
Bakery
Italian Restaurant
Donut Shop
Sandwich Place
Chinese Restaurant
Ice Cream Shop
Seafood Restaurant
Breakfast Spot
Mexican Restaurant
Fast Food Restaurant


MD
0
2
4
6
8
Pizza Place
American Restaurant
Chinese Restaurant
Fast Food Restaurant
Sandwich Place
Seafood Restaurant
Bakery
Mexican Restaurant
Ice Cream Shop
Café
Coffee Shop
Deli / Bodega
Italian Restaurant
Food Truck
MO
0
2
4
6
8
Pizza Place
American Restaurant
Fast Food Restaurant
Mexican Restaurant
Sandwich Place
Coffee Shop
Chinese Restaurant
BBQ Joint
Café
Bakery
Ice Cream Shop
Italian Restaurant
Breakfast Spot
Diner
MS
0
2
4
6
8
10
Fast Food Restaurant
American Restaurant
Pizza Place
Sandwich Place
Mexican Restaurant
Coffee Shop
BBQ Joint
Seafood Restaurant
Café
Fried Chicken Joint
Chinese Restaurant
Bakery
Ice Cream Shop
Southern / Soul Food Restaurant
NJ
0
2
4
6
8
10
Pizza Place
Italian Restaurant
Deli / Bodega
American Restaurant
Chinese Restaurant
Bakery
Ice Cream Shop
Fast Food Restaurant
Sandwich Place
Café
Bagel Shop
Coffee Shop
Mexican Restaurant
Diner
NY
0
2
4
6
8
10
Pizza Place
Deli / Bodega
American Restaurant
Italian Restaurant
Chinese Restaurant
Coffee Shop
Bakery
Café
Fast Food Restaurant
Ice Cream Shop
Sandwich Place
Mexican Restaurant
Diner
Donut Shop
SC
0
2
4
6
8
American Restaurant
Fast Food Restaurant
Pizza Place
Mexican Restaurant
Seafood Restaurant
Sandwich Place
Chinese Restaurant
BBQ Joint
Café
Coffee Shop
Breakfast Spot
Ice Cream Shop
Italian Restaurant
Southern / Soul Food Restaurant

Appendix C

Category	Sum	Category	Sum
American Restaurant	9002	**Cocktail Bar**	107
Coffee Shop	7623	**Argentinian Restaurant**	97
Chinese Restaurant	6741	**College Cafeteria**	97
Bakery	6157	**Butcher**	89
Café	5326	**Beer Garden**	84
Breakfast Spot	2858	**Afghan Restaurant**	72
BBQ Joint	2786	**Bookstore**	69
Burger Joint	2451	**Cantonese Restaurant**	58
Asian Restaurant	2103	**Arepa Restaurant**	55
Bagel Shop	1729	**Australian Restaurant**	54
Caribbean Restaurant	1357	**Colombian Restaurant**	44
Bar	853	**Candy Store**	43
Bubble Tea Shop	681	**Arcade**	42
Cafeteria	522	**Cheese Shop**	42
Cajun / Creole Restaurant	379	**Chocolate Shop**	40
Burrito Place	318	**Churrascaria**	32
African Restaurant	242	**Beer Bar**	27
Bistro	239	**Burmese Restaurant**	23
Buffet	214	**Bowling Alley**	18
Comfort Food Restaurant	192	**Bed & Breakfast**	16
Brewery	164	**Andhra Restaurant**	14
Brazilian Restaurant	163	**Building**	11