

Opening a New Restaurant In Houston

Greg Giem

July 1, 2019

Objective

Determine location and category opportunities for opening a new restaurant targeting Houston's high-income population using IRS and Foursquare data

- ☐ Identify and cluster high-income zip codes
- ☐ Associate zip codes with nearby restaurant venues
- ☐ Analyze restaurant venues to find underrepresented restaurant categories

Data – IRS Income Data

IRS.gov Tax Return Information

- Income Brackets
- Returns
- Total Adjusted Gross Income

Income Metrics

- Average Household Income
- Normalized Income

```
[8]: irs_data.head(10)
```

	zip code	income bracket	returns	dependents	total AGI	normalized income score	total high earners	percent high earners	year	average household income
0	77002	0	4880.0	1100.0	2067824.0	0.433197	1570.0	0.321721	2012	423.734426
1	77002	1	1400.0	270.0	16191.0	0.433197	1570.0	0.321721	2012	11.565000
2	77002	2	840.0	120.0	31306.0	0.433197	1570.0	0.321721	2012	37.269048
3	77002	3	650.0	70.0	39887.0	0.433197	1570.0	0.321721	2012	61.364615
4	77002	4	420.0	60.0	36213.0	0.433197	1570.0	0.321721	2012	86.221429
5	77002	5	680.0	150.0	94786.0	0.433197	1570.0	0.321721	2012	139.391176
6	77002	6	890.0	430.0	1849441.0	0.433197	1570.0	0.321721	2012	2078.023596
7	77003	0	4900.0	2970.0	241949.0	0.246939	620.0	0.126531	2012	49.377347
8	77003	1	2160.0	1620.0	28584.0	0.246939	620.0	0.126531	2012	13.233333
9	77003	2	1160.0	870.0	41995.0	0.246939	620.0	0.126531	2012	36.202586

<https://www.irs.gov/statistics/soi-tax-stats-individual-income-tax-statistics-zip-code-data-soi>

Data – Geographic Data

Census.gov GEOJSON Data

□ Boundaries

□ Areas

□ Latitudes and Longitudes

Combination with Income

□ Income Density

```
[9]: houston_zips_df.head()
```

```
[9]:
```

	land area	water area	latitude	longitude	zip code
0	16112274	11938	29.670870	-95.585990	77099
1	17881915	318808	29.773179	-95.314327	77020
2	22650287	213335	29.791808	-95.228991	77013
3	6567398	67910	29.749808	-95.345901	77003
4	16006249	20223	29.795344	-95.367590	77009

```
[16]: houston_df.head(10)
```

```
[16]:
```

	zip code	income bracket	returns	total AGI	normalized income score	total high earners	percent high earners	year	average household income	land area	water area	latitude	longitude	total area	households per section	income per section	high earners per section
0	77002	0	4880.0	2067824.0	0.433197	1570.0	0.321721	2012	423.734426	5227914	128033	29.756845	-95.365652	5355947	2359.833271	999942.597060	759.208655
1	77002	1	1400.0	16191.0	0.433197	1570.0	0.321721	2012	11.565000	5227914	128033	29.756845	-95.365652	5355947	677.001348	7829.520592	759.208655
2	77002	2	840.0	31306.0	0.433197	1570.0	0.321721	2012	37.269048	5227914	128033	29.756845	-95.365652	5355947	406.200809	15138.717291	759.208655
3	77002	3	650.0	39887.0	0.433197	1570.0	0.321721	2012	61.364615	5227914	128033	29.756845	-95.365652	5355947	314.322055	19288.251983	759.208655
4	77002	4	420.0	36213.0	0.433197	1570.0	0.321721	2012	86.221429	5227914	128033	29.756845	-95.365652	5355947	203.100404	17511.607017	759.208655
5	77002	5	680.0	94786.0	0.433197	1570.0	0.321721	2012	139.391176	5227914	128033	29.756845	-95.365652	5355947	328.829226	45835.892709	759.208655
6	77002	6	890.0	1849441.0	0.433197	1570.0	0.321721	2012	2078.023596	5227914	128033	29.756845	-95.365652	5355947	430.379429	894338.607468	759.208655
7	77002	0	4530.0	1881196.0	0.473731	1640.0	0.362031	2013	415.275055	5227914	128033	29.756845	-95.365652	5355947	2190.582934	909694.448763	793.058722
8	77002	1	1130.0	13056.0	0.473731	1640.0	0.362031	2013	11.553982	5227914	128033	29.756845	-95.365652	5355947	546.436802	6313.521145	793.058722
9	77002	2	700.0	26353.0	0.473731	1640.0	0.362031	2013	37.647143	5227914	128033	29.756845	-95.365652	5355947	338.500674	12743.583235	793.058722

<https://www.census.gov/cgi-bin/geo/shapefiles2010/main>

<https://github.com/OpenDataDE/State-zip-code-GeoJSON>

Data – Restaurant Data

Foursquare Data

- ☐ Restaurant Categories
- ☐ Restaurant Counts

Combination with Zip Codes


- ☐ Restaurants per Household


```
houston_venue_count_df.head()
```


	zip	Afghan Restaurant	African Restaurant	American Restaurant	Arcade	Arepa Restaurant	Argentinian Restaurant	Asian Restaurant	BBQ Joint	Bagel Shop	...	Turkish Restaurant	Udon Restaurant	Vegetarian / Vegan Restaurant	Venezuelan Restaurant	Vietnamese Restaurant	Whisky Bar	Wine Bar	Wings Joint	Xinjiang Restaurant	restaurant count
0	77002	0	2	134	0	0	1	25	47	5	...	2	0	15	0	84	0	8	3	0	2062
1	77003	0	2	94	0	0	1	19	50	3	...	2	0	6	0	76	0	5	4	0	1625
2	77004	0	0	70	0	0	0	19	31	7	...	2	0	10	0	75	0	4	6	0	1199
3	77005	0	0	128	0	1	0	41	20	7	...	5	0	18	0	18	0	6	7	0	1725
4	77006	0	2	160	0	1	1	34	36	8	...	0	0	28	0	76	1	11	5	0	2257


5 rows × 157 columns


<https://foursquare.com/>


**Food**
4d4b7105d754a06374d81259


**Afghan Restaurant**
503288ae91d4c4b30a586d67


**African Restaurant**
4bf58dd8d48988d1c8941735


**Ethiopian Restaurant**
4bf58dd8d48988d10a941735


**American Restaurant**
4bf58dd8d48988d14e941735


**New American Restaurant**
4bf58dd8d48988d157941735

**Asian Restaurant**
4bf58dd8d48988d142941735

**Burmese Restaurant**
56aa371be4b08b9a8d573568

**Cambodian Restaurant**
52e81612bcb57f1066b7a03

**Chinese Restaurant**
4bf58dd8d48988d145941735

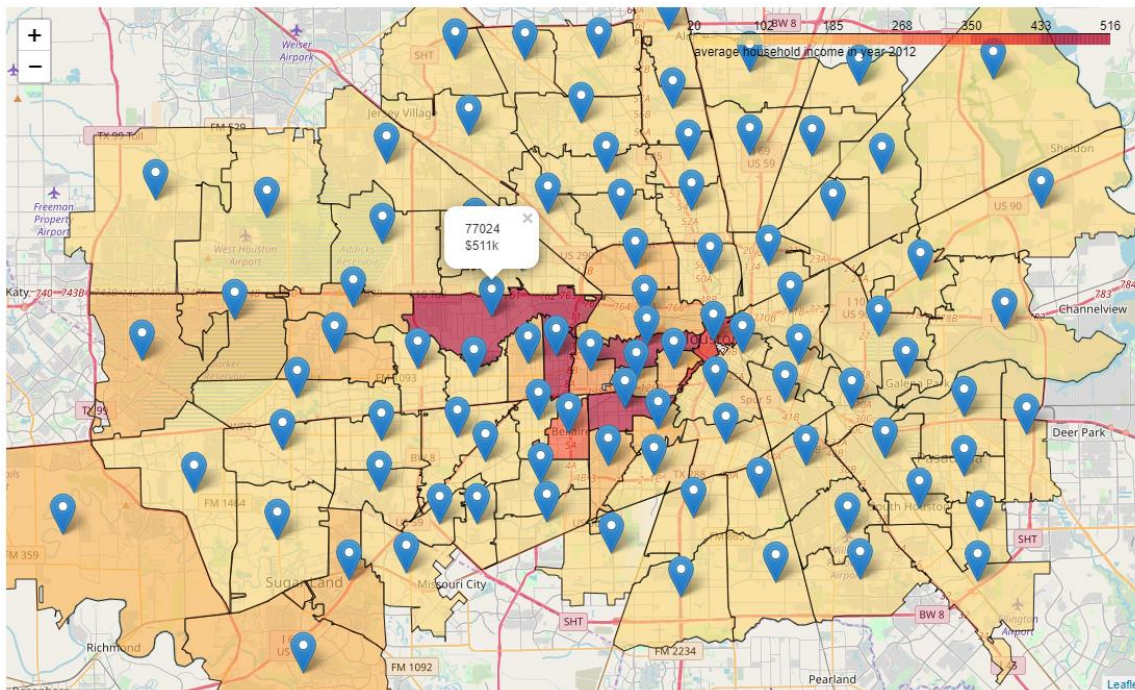
**Anhui Restaurant**
52e81612bcb57f1066b7a03

Overall Methodology

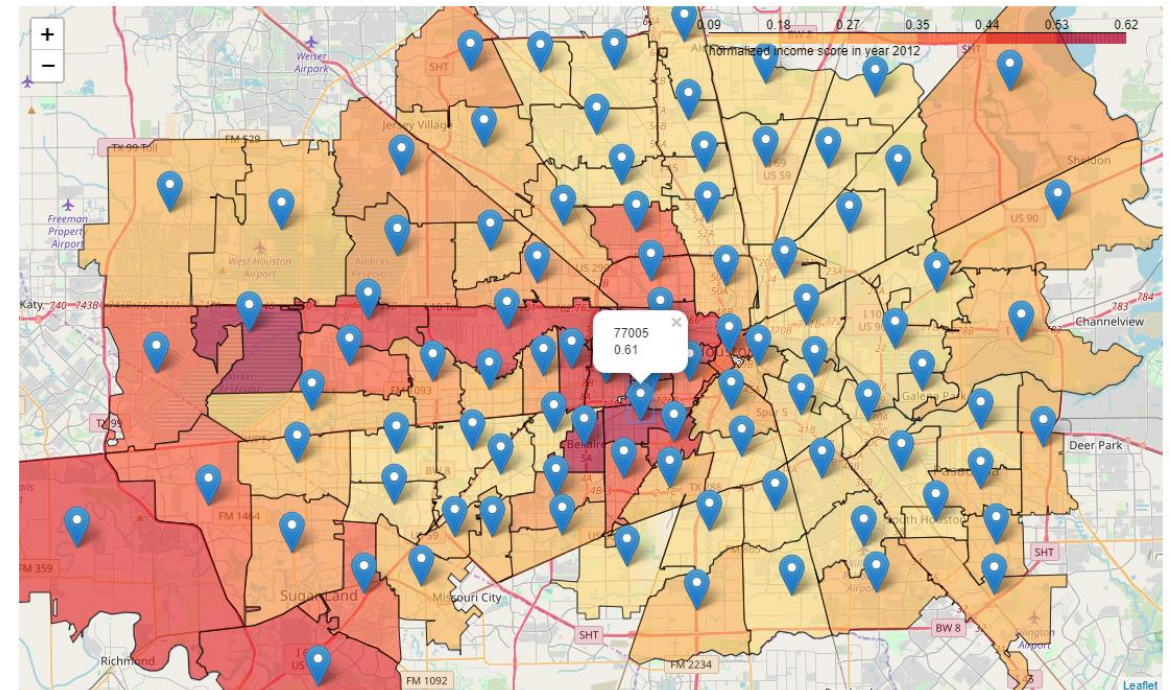
1. Gather income information related to the Houston population, by year and zip code
2. Determine geographic boundaries of each of the Houston zip codes
3. Combine the geographic information with the income information
4. Narrow the included zip codes to those within a certain geographic area
5. Gather restaurant data associated with each zip code
6. Cluster zip codes using income information
7. Correlate income information to venue type information
8. Identify zip codes with notable differences in venue concentrations from others in the same cluster

Income and Geographic Analysis

1. Gather income information related to the Houston population, by year and zip code
2. Determine geographic boundaries of each of the Houston zip codes
3. Combine the geographic information with the income information
4. Narrow the included zip codes to those within a certain geographic area



Average Household Income



Normalized Income Score (Weighted by Returns in Income Bracket)

Gather and Filter Restaurant Data

5. *Gather restaurant data associated with each zip code*

❑ All restaurants by Foursquare category within 2 miles of each zip code

❑ Filter out bad results

❑ Count and frequency of category types

houston_venue_count_df.head()

	zip code	Afghan Restaurant	African Restaurant	American Restaurant	Arcade	Arepa Restaurant	Argentinian Restaurant	Asian Restaurant	BBQ Joint	Bagel Shop	...	Turkish Restaurant	Udon Restaurant	Vegetarian / Vegan Restaurant	Venezuelan Restaurant	Vietnamese Restaurant	Whisky Bar	Wine Bar	Wings Joint	Xinjiang Restaurant	restaurant count
0	77002	0	2	134	0	0	1	25	47	5	...	2	0	15	0	84	0	8	3	0	2062
1	77003	0	2	94	0	0	1	19	50	3	...	2	0	6	0	76	0	5	4	0	1625
2	77004	0	0	70	0	0	0	19	31	7	...	2	0	10	0	75	0	4	6	0	1199
3	77005	0	0	128	0	1	0	41	20	7	...	5	0	18	0	18	0	6	7	0	1725
4	77006	0	2	160	0	1	1	34	36	8	...	0	0	28	0	76	1	11	5	0	2257

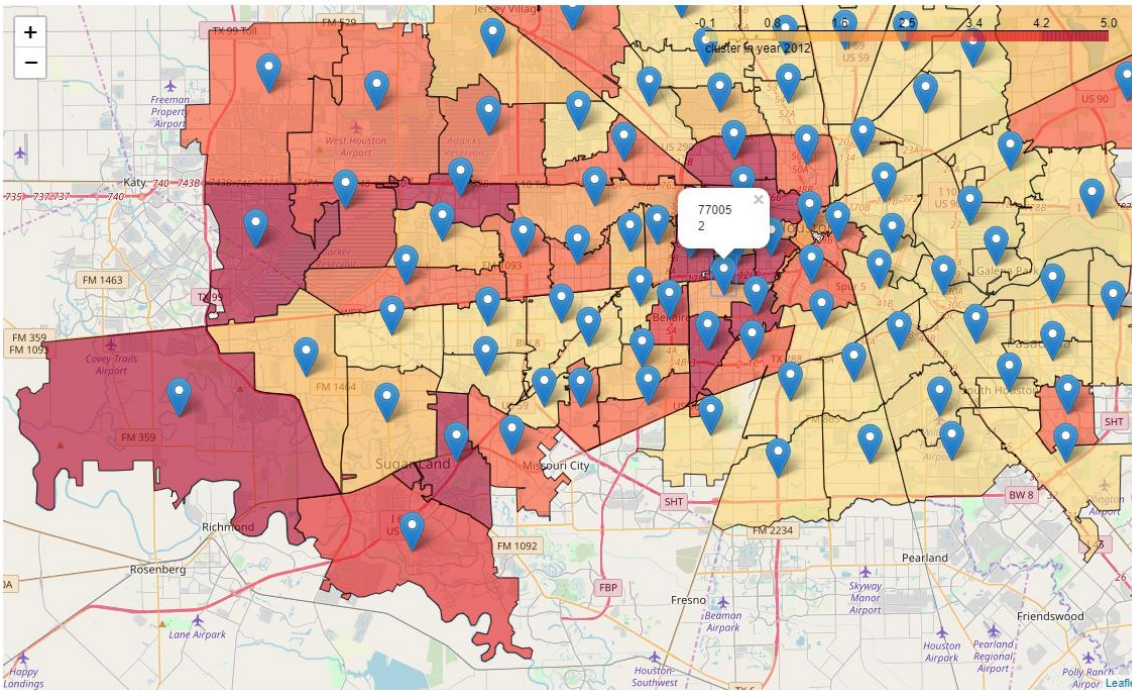
5 rows × 157 columns

	zip code	cluster	Afghan Restaurant	African Restaurant	American Restaurant	Arcade	Arepa Restaurant	Argentinian Restaurant	Asian Restaurant	BBQ Joint	...	Theme Restaurant	Tiki Bar	Turkish Restaurant	Vegetarian / Vegan Restaurant	Venezuelan Restaurant	Vietnamese Restaurant	Whisky Bar	Wine Bar	Wings Joint	restaurant count
0	77002	2	0.0	0.000970	0.064985	0.0	0.000000	0.000485	0.012124	0.022793	...	0.002910	0.000000	0.000970	0.007274	0.0	0.040737	0.000000	0.003880	0.001455	2062
1	77005	2	0.0	0.000000	0.074203	0.0	0.000580	0.000000	0.023768	0.011594	...	0.000000	0.000000	0.002899	0.010435	0.0	0.010435	0.000000	0.003478	0.004058	1725
2	77006	5	0.0	0.000886	0.070891	0.0	0.000443	0.000443	0.015064	0.015950	...	0.000886	0.000000	0.000000	0.012406	0.0	0.033673	0.000443	0.004874	0.002215	2257
3	77007	5	0.0	0.000000	0.088186	0.0	0.001664	0.000000	0.009983	0.004160	...	0.000000	0.000000	0.000000	0.011647	0.0	0.017471	0.000000	0.000000	0.007488	1202
4	77008	5	0.0	0.000938	0.073171	0.0	0.000000	0.000000	0.013133	0.022514	...	0.000000	0.000938	0.000000	0.015009	0.0	0.024390	0.000000	0.000000	0.005629	1066

5 rows × 142 columns

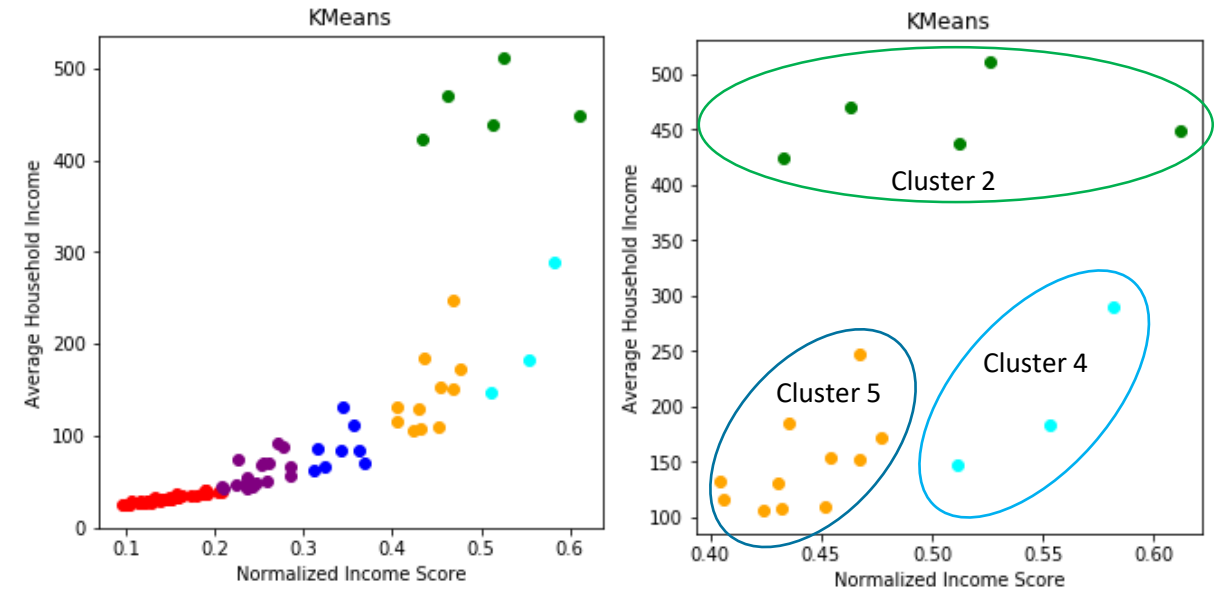
Cluster by Income

6. Cluster zip codes using income information



6 Clusters using Income Metrics

1. Average Household Income
2. Number of High Earners
3. Normalized Income Score



Retain only High-Income Clusters

Cluster 2: 5 zip codes

Cluster 4: 3 zip codes

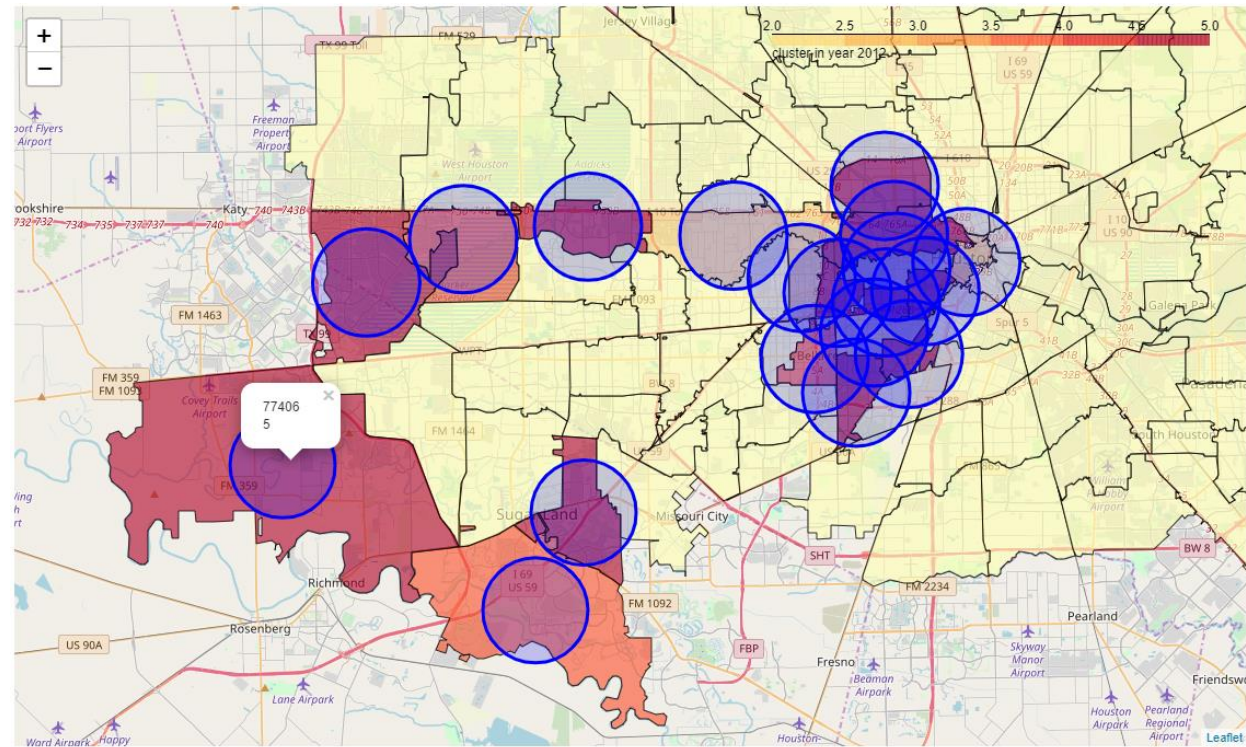
Cluster 5: 11 zip codes

Correlate Clusters to Restaurant Concentration

7. Correlate income information to venue type information

	zip code	returns	total AGI	normalized income score	total high earners	percent high earners	average household income	restaurant count	restaurants per household	
	0	77002	4880.0	2067824.0	0.433197	1570.0	0.321721	423.734426	2062	0.422541
	35	77005	11090.0	4983523.0	0.611542	6140.0	0.553652	449.370875	1725	0.155546
	70	77006	12050.0	1596949.0	0.404315	3340.0	0.277178	132.526888	2257	0.187303
	105	77007	19360.0	3333203.0	0.476756	6990.0	0.361054	172.169576	1202	0.062087
	140	77008	17300.0	1858315.0	0.431792	5500.0	0.317919	107.417052	1066	0.061618
	175	77019	11840.0	5572872.0	0.463007	4190.0	0.353885	470.681757	1964	0.165878
	210	77024	17570.0	8973976.0	0.526010	7830.0	0.445646	510.755606	737	0.041946
	245	77025	12120.0	1401085.0	0.405776	3590.0	0.296205	115.601073	798	0.065842
	280	77027	9710.0	2401290.0	0.467559	3260.0	0.335736	247.300721	1704	0.175489
	315	77030	5100.0	774234.0	0.467451	1780.0	0.349020	151.810588	1243	0.243725
	350	77056	11100.0	4857996.0	0.512613	4550.0	0.409910	437.657297	1566	0.141081
	385	77079	15570.0	2384056.0	0.454207	5760.0	0.369942	153.118561	706	0.045344
	420	77094	4590.0	843334.0	0.552941	2290.0	0.498911	183.732898	174	0.037908
	455	77098	8110.0	1497937.0	0.435758	2550.0	0.314427	184.702466	2043	0.251911
	490	77401	7990.0	2316414.0	0.581477	4190.0	0.524406	289.914143	807	0.101001
	525	77406	16400.0	1799280.0	0.451707	5770.0	0.351829	109.712195	18	0.001098
	560	77450	30590.0	3231062.0	0.423799	9980.0	0.326250	105.624779	450	0.014711
	595	77478	12440.0	1621592.0	0.430386	4050.0	0.325563	130.353055	741	0.059566
	630	77479	33720.0	4987378.0	0.511447	14970.0	0.443950	147.905635	242	0.007177

Restaurants vs. Remaining Zip Codes



Restaurant Search Areas

Outlier Identification - Filtering

8. *Identify zip codes with notable differences in venue concentrations from others in the same cluster*

For Each Cluster

- ☐ Filter to categories with mean > 1% of total restaurants
- ☐ Filter to categories with at least one value less than 1/3 of mean
- ☐ Examine outliers on bar chart

```
Cluster 2 has 29 restaurant categories with an average frequency exceeding 0.01
Cluster 4 has 31 restaurant categories with an average frequency exceeding 0.01
Cluster 5 has 31 restaurant categories with an average frequency exceeding 0.01
```

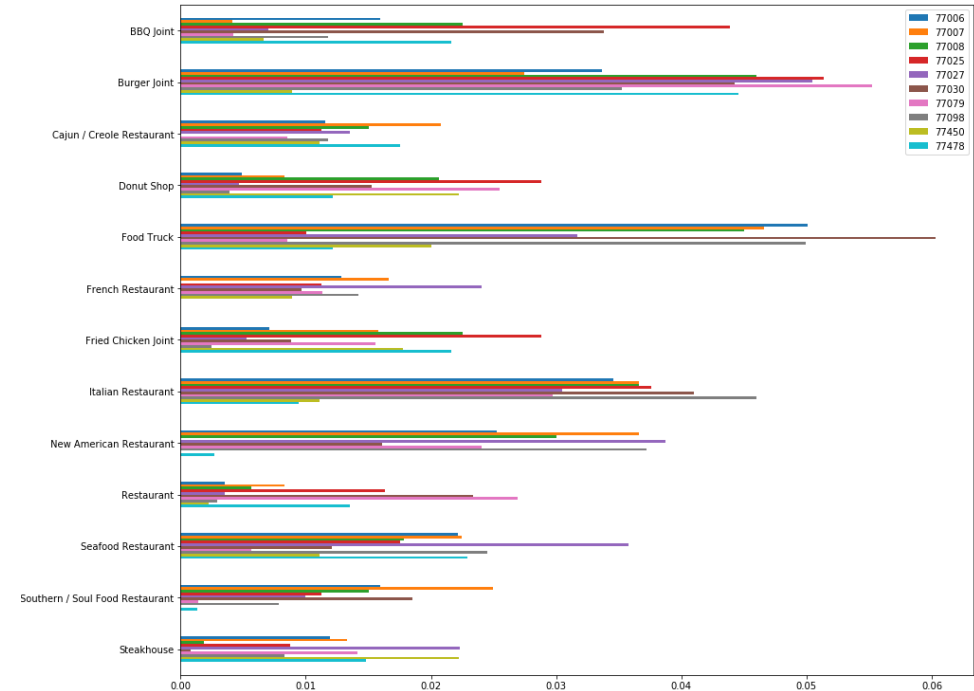
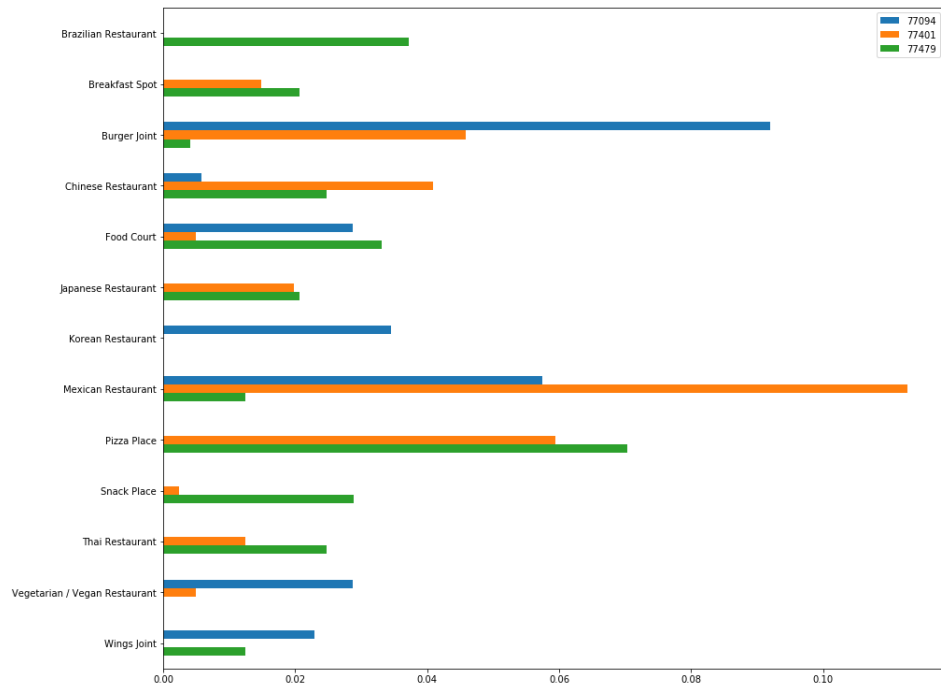
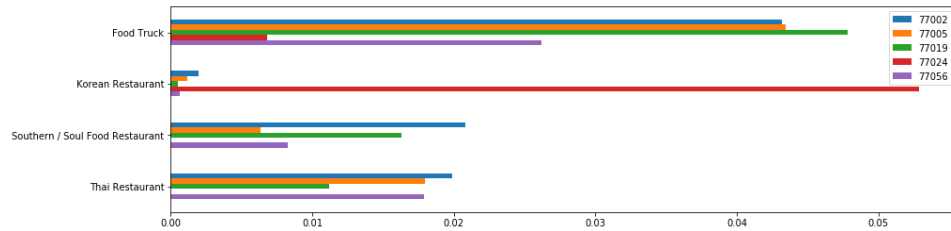
There are 22 restaurant categories common to all selected clusters:

```
American Restaurant
Asian Restaurant
Bakery
Breakfast Spot
Burger Joint
Café
Cajun / Creole Restaurant
Chinese Restaurant
Coffee Shop
Deli / Bodega
Fast Food Restaurant
Italian Restaurant
Japanese Restaurant
Mediterranean Restaurant
Mexican Restaurant
Pizza Place
Sandwich Place
Seafood Restaurant
Sushi Restaurant
Taco Place
Thai Restaurant
Vietnamese Restaurant
```

```
4 outlier categories found for cluster 2
13 outlier categories found for cluster 4
13 outlier categories found for cluster 5
```


Outlier Identification - Visualizing

8. *Identify zip codes with notable differences in venue concentrations from others in the same cluster*



Outlier Identification - Results

8. *Identify zip codes with notable differences in venue concentrations from others in the same cluster*

- ❑ Identify all $< 1/3$ of mean
- ❑ Ignore cases with one positive outlier skewing data

Note: Some zip codes had no opportunities, and some were more significant outliers than others

	Food Truck	Southern / Soul Food	Thai	Breakfast Spot	Chinese	Food Court	Wings	Burger Joint	Mexican	Donut Shop	BBQ Joint	French	Steakhouse	New American	Fried Chicken Joint	Cajun / Creole	Seafood	Italian
77002																		
77005		x																
77019																		
77024	x	x	x															
77056		x																
77094			x	x	x													
77401						x	x											
77479								x	x									
77006										x								
77007											x							
77008												x	x					
77025	x													x				
77027											x				x			
77030													x			x		
77079	x	x									x						x	
77098									x						x			
77450		x						x			x			x				x
77478	x	x											x	x				x

Conclusions and Possible Improvements

- ❑ Income clusters were very clear
- ❑ Clusters had fairly clear similarities in restaurant frequency
- ❑ Many Opportunities were identified
- ❑ Opportunities should be investigated in more detail
- ❑ Population density and travel time should be accounted for (urban vs. suburban)
- ❑ Ethnic and family demographics should be accounted for
- ❑ Categories may need to be grouped to get higher frequencies