Opening a New Restaurant in Houston

Greg Giem

July 1, 2019

Executive Summary (Week 2)

This report details the methodology used to determine 38 opportunities to open new restaurants in 16 different high-income zip codes in Houston. The analysis clustered all the zip codes in the greater Houston area and retained only the highest-income clusters, then compared those clusters to restaurant category frequency gathered from Foursquare. The restaurant categories that were dramatically underrepresented in each cluster provide possible opportunities for new restaurants in the Houston area.

Table of Contents

1. Introduction (Week 1)

Houston is the fourth largest city in the United States and is noted for its diversity. Not only is it culturally and ethnically diverse as a whole, it's someone unusual among large cities in its diversity within communities. Houston is also known as a city where dining out at restaurants is part of the evening entertainment. But perhaps most of all, Houston is known as an oilfield city, the chief technical center for a global oil industry (Gilmer). This means it has a large number of highly-paid executives and engineers that can drive a high-end restaurant industry.

If you wanted to open a high-end restaurant in Houston, what kind would you choose and where would you place it? To answer this question, we can look at the income distribution across Houston to find the zip codes where the wealthy have chosen to reside, and we can look at the types of restaurants already within easy reach. By looking at what's been historically popular in wealthy neighborhoods, we can find underserved wealthy areas to target.

2. Data (Week 1)

The main sources of data we'll look at to determine where to open a high-end restaurant are *income* and *existing offerings*. The income distributions can point us toward the neighborhoods we should focus our efforts, and an analysis of existing offerings can show us what types are over- or under-serving those neighborhoods.

We will focus mainly on the area inside Beltway 8, or the Sam Houston Tollway, to define rough city limits of Houston. We will limit the scope of the search to this area plus some of the oilfield-heavy areas on the West and Southwest. The Houston metropolitan area sprawls through neighboring cities in all directions, including the Woodlands to the north, Tomball to the southwest, Katy to the west, Sugar Land to the southwest, and Pearland to the south. We will use geographic data to narrow this search.

The existing offerings are limited in scope to restaurant types already found within Houston. These restaurants were included based on proximity to zip codes, with no use of ratings, price, popularity, or any other metrics. The purpose of the venue analysis was to gauge relative interest of restaurant categories rather than specific restaurants.

After income and geographic data are combined, a clustering algorithm can be used to narrow the zip codes of interest to those associated with high earners. Then a clustering analysis associating restaurant categories with zip codes can be used to determine outliers from the trend – representing possible opportunities.

2.1. Income Data

The income data used in this analysis came from income tax return summaries compiled by the United States Internal Revenue Service (SOI Tax Stats…). Data was available through 2016 tax year, so we will focus on the five-year period stretching from 2012 to 2016.

The data includes many different categories associated with federal income tax returns, but we will focus on several that were available through all of the years in question that seemed relevant to the task at hand. The descriptions for each of the many columns was available from the documentation on the IRS website referenced in Section 7. We renamed these columns to a more user-friendly description.

Renamed Identifier	Original Identifier	Description	Reference	Type
zip code	zipcode	5-digit Zip code		Char
income bracket	AGI_STUB	Size of adjusted gross income	1 = $1 under $25,000 2 = $25,000 under $50,000 3 = $50,000 under $75,000 4 = $75,000 under $100,000 5 = $100,000 under $200,000 6 = $200,000 or more	Num
returns	N1	Number of returns		Num
dependents	NUMDEP	Number of dependents	1040:6c	
total AGI	A00100	Adjust gross income (AGI) Does not include returns with adjusted gross deficit.	1040:37 / 1040A:21 / 1040EZ:4	Num

Table 1: IRS Data Description

This data can then be manipulated and combined with geographic data to get relative wealth scores of different types including average household income, income density, etc. These parameters can later be used to identify target markets.

2.2. Geographic Data

The geographic data used in this analysis consisted of the zip code boundaries provided by the 2010 US Census. The data was gathered from a GitHub repository (OpenDataDE) that maintained GeoJSON files for each state that were converted from the US Census Shapefiles (per the readme). These files contain boundaries, latitudes and longitudes, and areas for each zip code.

The geographic data was originally only consisting of the zip codes contained within the Sam Houston Tollway, but that left out some major areas of oilfield activity including the "Energy Corridor" around I10/Hwy 6 and Sugar Land, both of which revealed a notably high percentage of high earners.

When combined with the income data, the geographic data revealed that some very small, wealthy zip codes skewed the relative numbers despite having only a small number of returns. These zip codes (specifically 77010 and 77046) were found almost entirely within the bounds of other zip codes, so the income data for these zip codes was combined with the larger zip codes and the small zip codes were dropped from the analysis. This should not negatively affect the analysis.

In the end, 92 zip codes were chosen for inclusion in the analysis.

2.3. Venue Data

The venue data is gathered from Foursquare similar to lab exercises. For this exercise, only restaurant-type venues were desired in order to evaluate the types and quantities of restaurants in the vicinity of the chosen zip codes. The Foursquare API appeared to limit all queries to 100 results, independent of a higher value entered in the query, so it was necessary to search using restaurant subtypes that were available in the Foursquare documentation (Venue Categories). This resulted in many more queries, but each of those queries returned fewer than 100 results, meaning that venues weren't dropped because of density.

A somewhat arbitrary two-mile radius from the center of the zip code was used to consider restaurant venues accessible to that zip code. This assumption will have a large impact on the number of venues found and is somewhat difficult to apply to both the smaller downtown zip codes and the larger suburban zip codes, but it was used as a reasonable approximation for the purposes of this exercise.

After gathering the restaurant data as a function of zip code, the data was manually examined and a list of subcategories were removed based on inapplicability to the exercise (corporate or school cafeterias, grocery stores, and other venues that were returned as results without being what most would consider traditional restaurants).

In the end, 206 distinct restaurant categories were used within the included zip codes.

3. Methodology (Week 2)

The final purpose of this exercise is to link restaurant concentration to income concentration in a way that will indicate potential gaps in the market. In order to determine this relationship, the income, geographic, and venue data discussed in the previous section must be analyzed and correlated.

The methodology used for this project follows the sequence below, with each element being described in more detail further on in this section.

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

This is not the most sophisticated method available, and it will have opportunities for improvement, but it provides a good starting point for identification of outliers.

It should be noted that for the convenience of time, each of these steps was conducted separately with data then stored in CSV and JSON files for later access. This saved lots of web queries and processing time, allowing analysis of very large quantities of data over many different sessions.

3.1. Gather Income Information

The United States Internal Revenue Service publishes a wealth of statistical information related to tax filings that are organized by zip code and further broken down by income bracket (SOI Tax Stats…). This information is currently only available through 2016 filings, so it will not be totally representative of the current income levels of the city of Houston, but it should be close. In order to later be able to evaluate data relative to income trends, the income data over a 5-year period was gathered and organized into a single Pandas DataFile.

Documentation accompanying each year's income data describes each of the columns found in the data files. These are not completely consistent from year to year, with some titles changed and other columns added or dropped, so only columns of interest were kept that were present in all years' data

files. These columns were all renamed from their original IRS identifiers to something that would be much more user-friendly. Then some additional columns were created to create new columns that would be relevant to later analysis.

Calculated Column	Brief Description
Normalized Income Score	Weighted income score
Total High Earners	Households earning > $100k
Percent High Earners	Percentage of total households earning > $100k
Average Household Income	Total AGI divided by number of households

Table 2: Calculated Income Columns Added to Source Data

The "normalized income score" was created to apply relative weights to each of the income bracket populations within a zip code. The purpose of this was to capture a possible spending profile of a zip code, with the theory being that something like an average income score might skew the appearance of the high-end market in zip codes that consisted of a mix of extremely high earners with extremely low earners (a real scenario in some of the downtown zip codes). This normalized income score is higher when there are a large number of people in the upper income brackets that are expected to have disposable income for high-end restaurants but aren't necessarily the extremely high earners. The result was that the normalized income score increased the relative importance of high-earning suburbs.

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

[8]:

	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

Figure 1: Example Income Data

The original income data included the 6 brackets detailed in Table 1, but an additional income bracket "0" was added to sum the number of returns, dependents, and AGI to be able to analyze the zip code as a whole.

3.2. Determine Geographic Boundaries

The next step in the process is to determine the center and geographic boundaries of each zip code. This is extremely helpful for visualization and essential for relating income data to restaurants by latitude and longitude.

The data used for this exercise was sourced from the OpenDataDE github repository (OpenDataDE…), which is a processed JSON version of the shapefiles provided by the US Census Bureau (2010 TIGER…).

The JSON file was loaded into python as a dictionary where it could be easily converted into a Pandas DataFrame. This DataFrame gathered the zip code longitude, latitude, and land areas. The dictionary was also later useful for longitude/latitude lookups when querying restaurant data. The JSON file was retained for use by the Folium plotting functions, so the boundary information was not stored in the DataFrame.

```
[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

Figure 2: Example Geographic Data

3.3. Combine Income and Geographic Information

Once both income and geographic data are available, a new DataFrame was created that combined the two datasets and added new metrics unavailable without the combination. The metrics were essentially density functions that combined the density of households, income, and high earners in the area of the zip code.

```
[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

Figure 3: Example Combined Income and Geographic Data

The value of these metrics is somewhat questionable in real terms because a sparser population might make travel easier, extending the relative accessible range of restaurants. But in any case, it can later be used in the clustering algorithm to differentiate zip codes that may have a similar income makeup but reflect different population types.

When combining geographic areas with income information, it became clear that there were two extremely small zip codes within Houston that had very small geographic areas and very small populations. When looking at percentage metrics, these skewed the data and didn't provide much value because of there small populations, so their income data was combined with that of the zip codes surrounding them.

3.4. Narrow Zip Codes by Geography

The original source data used all zip codes in the greater Houston area. In addition to adding to the work load, most of the relevant zip codes were found inside the Sam Houston Tollway. Initially the search was going to be contained within that area, but a preliminary analysis of the combined geographic and income data revealed that several zip codes in suburbs just outside the Sam Houston Tollway had relatively high average incomes and even higher relative income scores. These areas coincided with headquarters of energy-related companies, explaining the presence of the additional high earners. In order to avoid missing out on this potentially lucrative market, these zip codes were added back into the larger analysis.

Chloropleth maps in folium were used to conduct this preliminary analysis. They allowed a quick visual indication of a particular metric over the entire Houston area. This led to an iterative process where more zip codes were added or removed to the datasets in the process above until the final choice was made to limit the geographic search area.

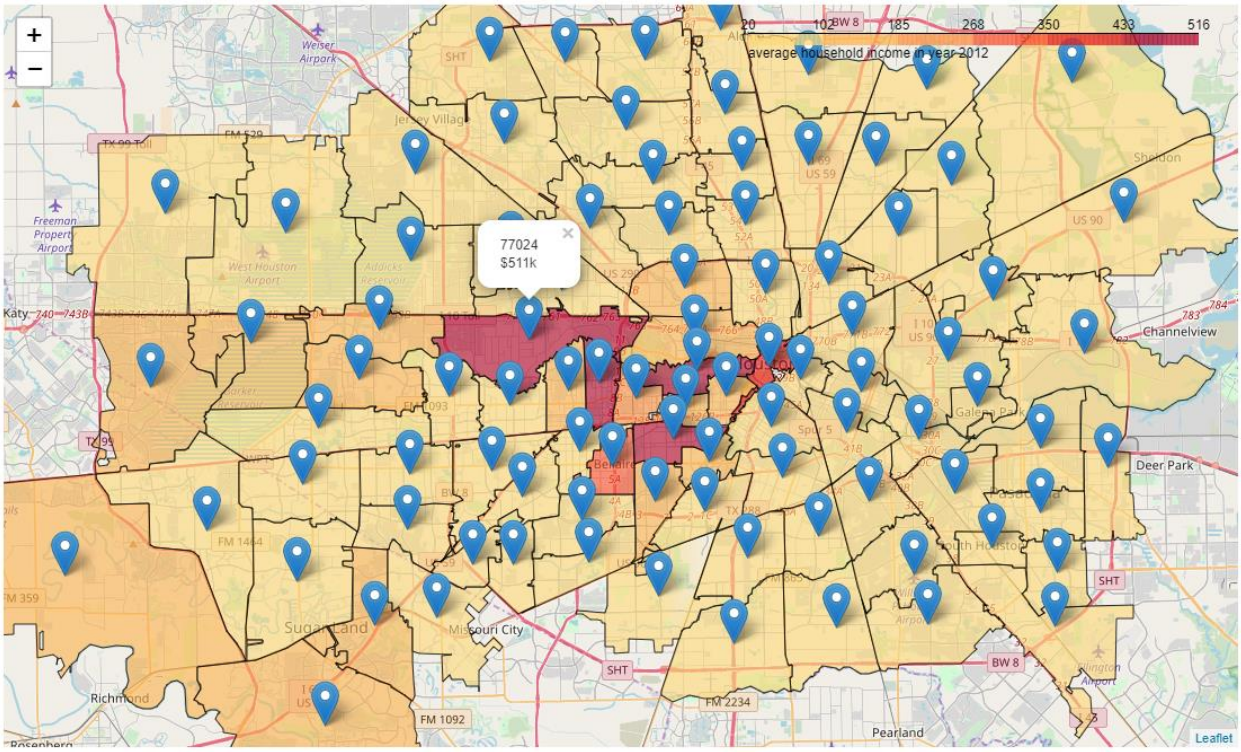

Figure 4: Chloropleth Map of Houston Average Household Income for 2012

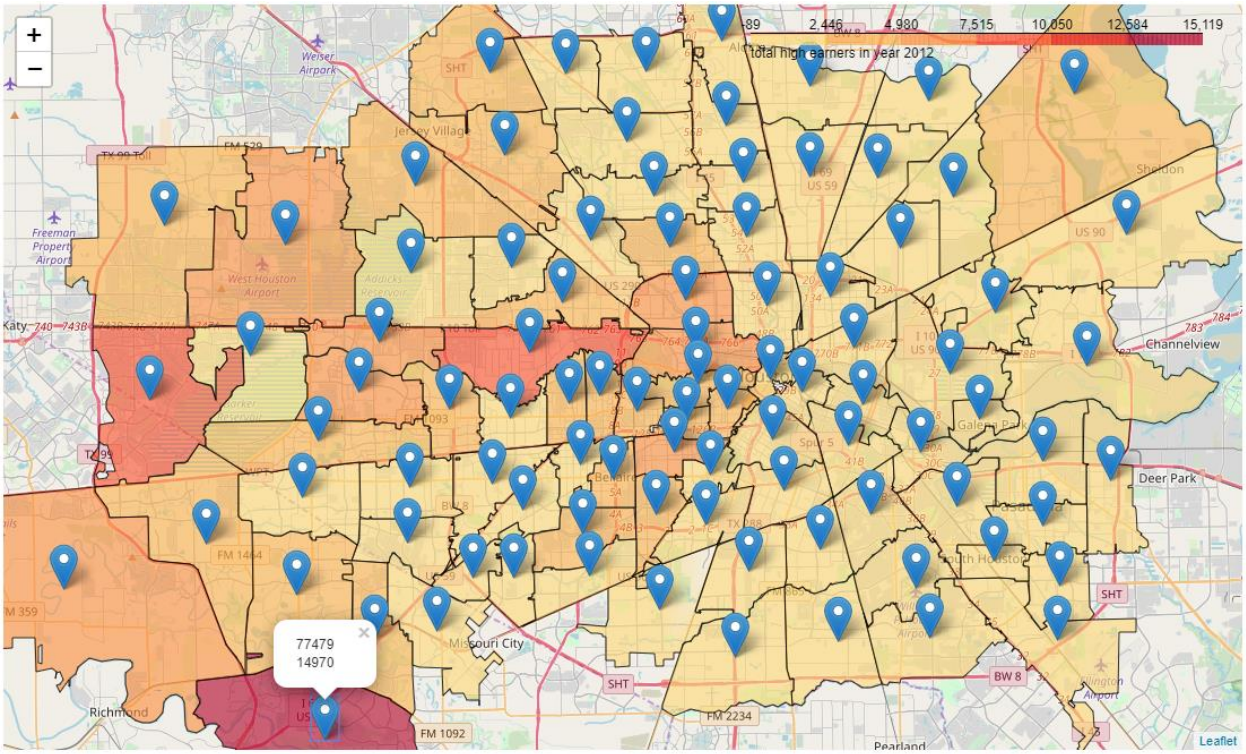

Figure 5: Chloropleth Map of Houston High Earners for 2012

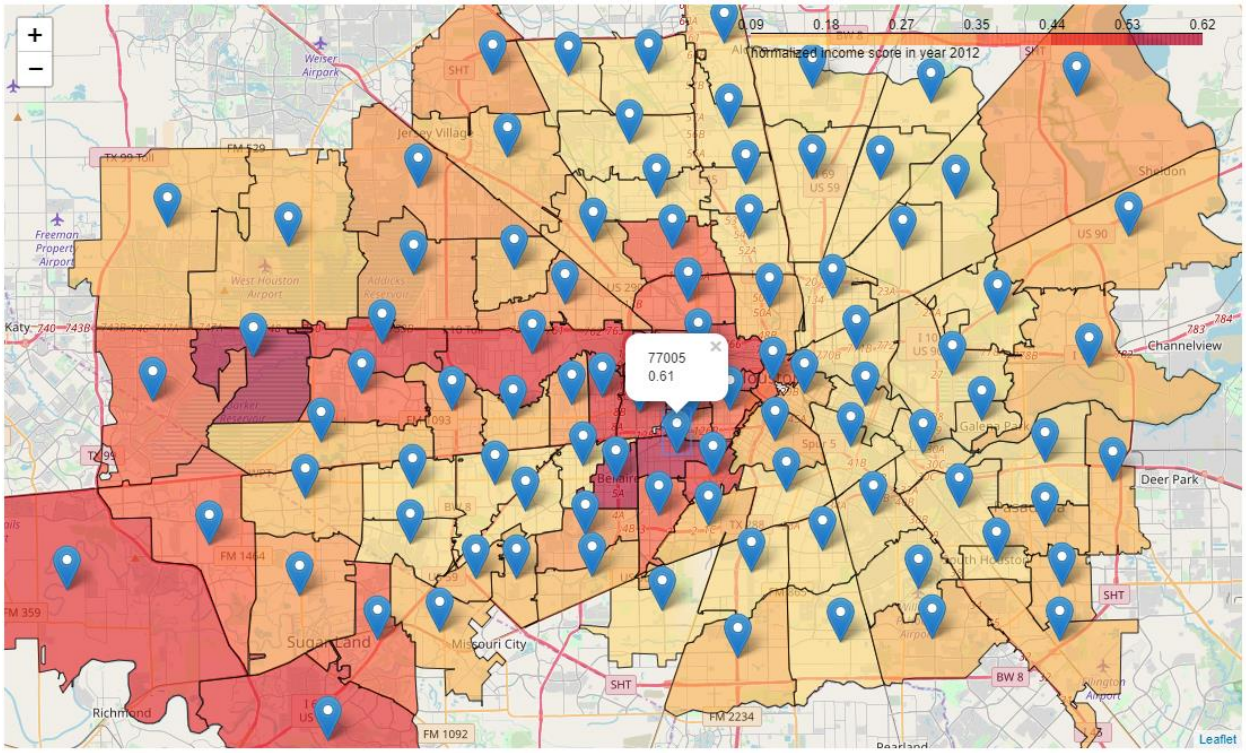

Figure 6: Chloropleth Map of Houston Normalized Income Score for 2012

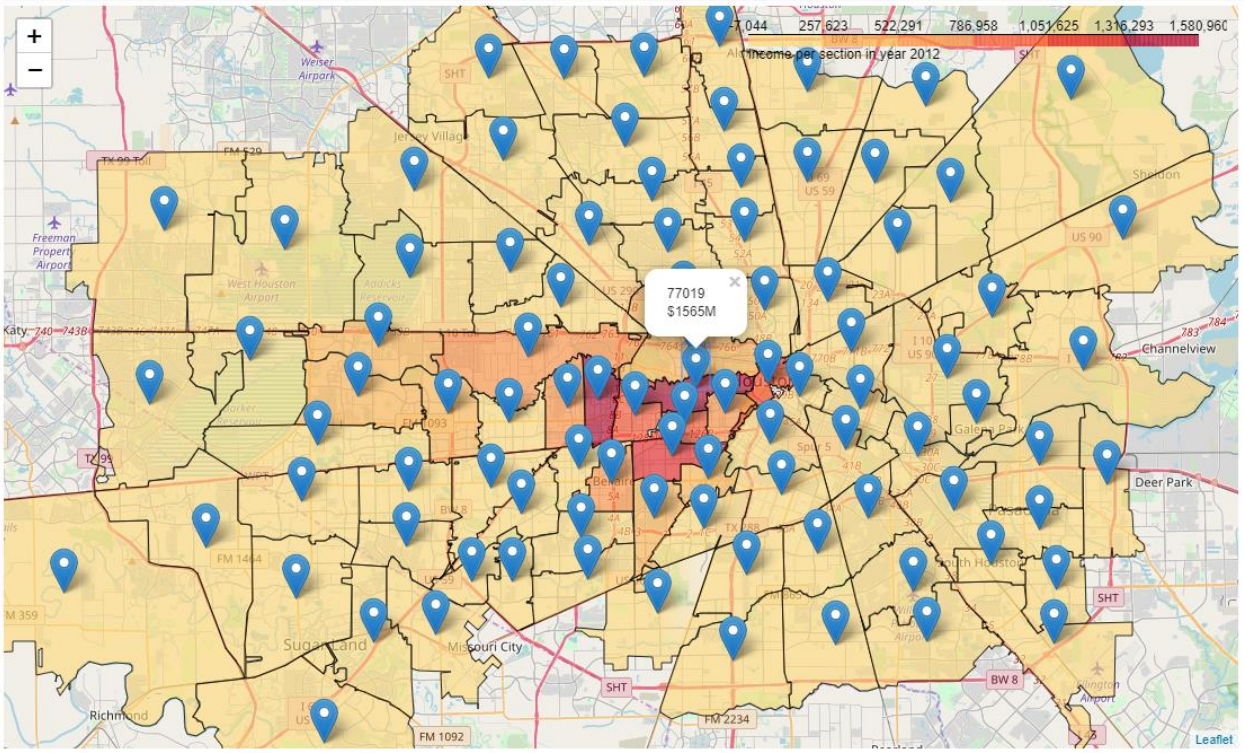

Figure 7: Chloropleth Map of Houston Income per Square Mile for 2012

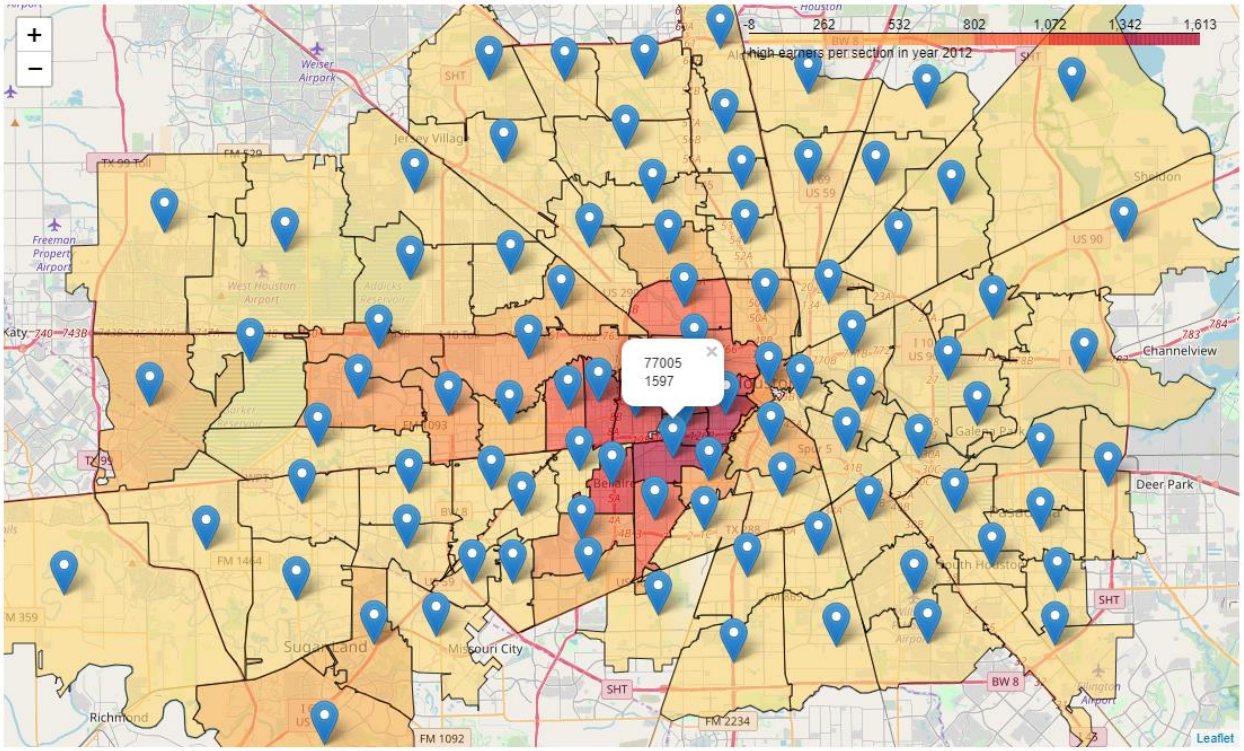

Figure 8: Chloropleth Map of Houston High Earners per Square Mile for 2012

The maps in the previous figures illustrate some of the challenges in determining the best location to target the high-income population. It's clear that the central part of Houston, just west of the

downtown area is where the highest-earners are concentrated, based on the average incomes and income density metrics. But it's also clear that the western and southwestern suburbs have large number of earners in the higher income brackets. When the metrics are considered together, it can be concluded that there are many high earners, but they are probably not earning incomes in the same order of magnitude as some of the central Houston residents. The normalized income score also implies that there are fewer low earners in some of the suburban zip codes.

Note that folium markers were added to each zip code with the zip code itself and the income metric. This made it easy to add or remove zip codes to the analyzed data by interacting with the map.

3.5. Gather Restaurant Data

With the zip codes narrowed, it is now possible to gather restaurant data relative to each zip code in the data set. The Foursquare API provides a radius and a maximum number of venue hits for the search query, bit experimentation showed the maximum number of venues returned was limited to 100, even when the argument was set to a higher number. To avoid artificially reducing the data, the search was conducted over a list of specific restaurant categories provided by the foursquare documentation.

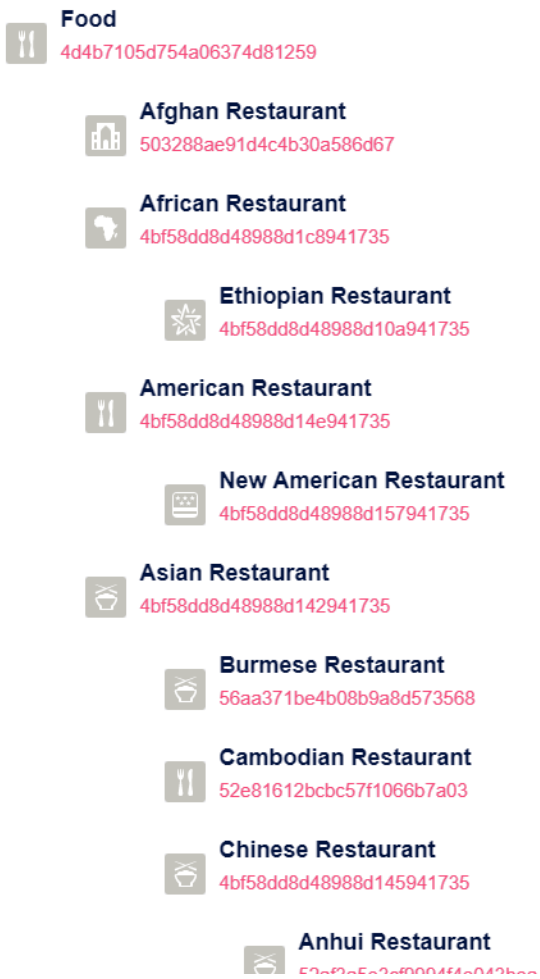

Figure 9: Example Restaurant Categories

This required nested loops where each of more than 100 restaurant categories were searched for each of 92 zip codes, with the results added to a Pandas DataFrame. Experimentation showed that the

Foursquare APi would occasionally stall, forcing the loop to be manually interrupted, so a set of timers and related exception handlers were added to automatically break and retry a query that took more than 5 seconds to return a result.

A somewhat arbitrary radius of 2 miles was used from the center of each zip code for the search results. This has some obvious issues, especially considering the wide range of zip code areas considered. In heavily populated areas where driving could be considered prohibitively inconvenient, this distance could be too far to walk. In less densely populated areas where driving is the norm and traffic is not an issue, 5 or 10 miles might be an easy distance to travel for a meal. A more sophisticated model has many other options to consider that were not explored as part of this exercise.

After the DataFrame was populated, it had to be manually analyzed to determine the contents. Several issues were discovered that forced the data to be manually trimmed to end up with useful information.

1. There were duplicate results within a zip code. It was no apparent why, but some results would be found more than once for a given zip code. It's expected (and desired) that a certain restaurant would be found for more than one zip code, but the results for a particular zip code needed to be unique.
2. There were results that were not part of any of the searched categories or their subcategories. Again it was not clear why Foursquare returned things like parks or schools that are not found under the "food" category.

These erroneous results were removed by filtering the data to unique rows and removing any hits from undesired categories. The resulting DataFrame included a list of restaurants with location and category information for each zip code.

```
venues_df.head()
```

	zip code	restaurant name	restaurant latitude	restaurant longitude	restaurant category	restaurant category ID
0	77094	Murphy's Deli	29.789579	-95.670630	Deli / Bodega	4bf58dd8d48988d146941735
1	77094	Park 10 Deli	29.786183	-95.663970	Deli / Bodega	4bf58dd8d48988d146941735
2	77094	High Tower Deli & Cafe #3	29.788591	-95.657420	Deli / Bodega	4bf58dd8d48988d146941735
3	77094	Smashburger	29.784339	-95.705163	Burger Joint	4bf58dd8d48988d16c941735
4	77094	Burger Tex Grill	29.785979	-95.687331	Burger Joint	4bf58dd8d48988d16c941735

Figure 10: Example Restaurants

The specific restaurants found are not relevant to the exercise, so a new DataFrame was created to aggregate restaurant category counts for each zip code, including a column that summed the total number of restaurants accessible within the given radius from the center of each zip code.

```
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

Figure 11: Example Restaurant Category Counts

Finally, the restaurant count information was merged into the main income/geography DataFrame previously discussed, adding two new columns associated with total restaurant count and restaurant density (restaurants within the given radius per household).

```
houston_df.head()
```

	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	restaurant count	restaurants per household
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	2062	0.422541
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	2062	1.472857
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	2062	2.454762
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	2062	3.172308
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	2062	4.909524

Figure 12: Example Combined Income, Geography, and Restaurant Data

3.6. Cluster Zip Codes by Income

A KMeans clustering algorithm was run on the income data in attempt to gather economically similar zip codes, using the columns for average income normalized income score, and number of high earners. The data was first normalized using sklearn after initial attempts revealed the much larger magnitude of the average household income column was skewing the results.

```
cluster_df.head()
```

	normalized income score	percent high earners	average household income
0	0.433197	0.321721	423.734426
1	0.246939	0.126531	49.377347
2	0.261221	0.158779	70.431374
3	0.611542	0.553652	449.370875
4	0.404315	0.277178	132.526888

```
cluster_scaled_df.head()
```

	normalized income score	percent high earners	average household income
0	0.654365	0.575022	0.821041
1	0.293395	0.217364	0.051178
2	0.321075	0.276453	0.094475
3	1.000000	1.000000	0.873763
4	0.598393	0.493403	0.222174

Figure 13: Example Income Data Before and After Scaling

The zip code data was clustered into 6 groups, then each of the groups was analyzed manually to make sure they made sense. This was first plotted using folium to see if the groupings seemed similar to those seen before when visualizing income data, then the data was examined as part of the DataFrame.

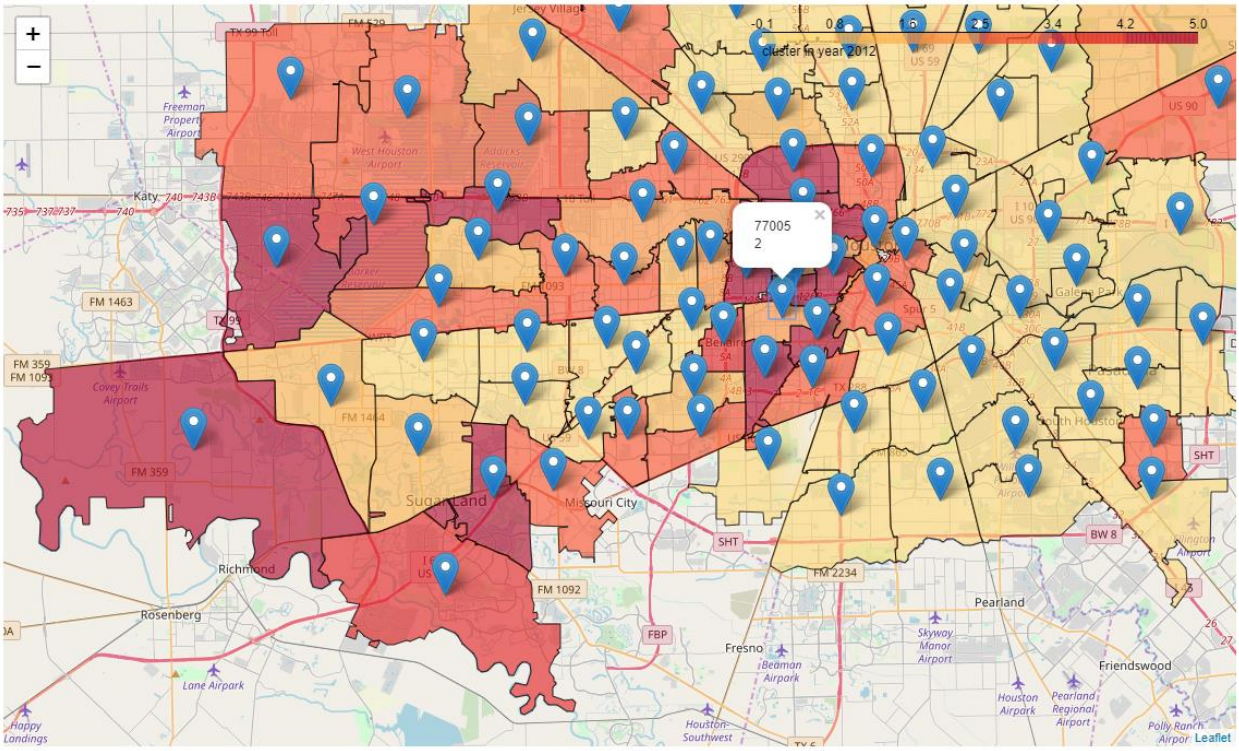

Figure 14: Clustered Zip Codes

```
folium_df[folium_df['cluster'] == 2].head()
```

	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	restaurant count	restaurants per household	cluster
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	9.999426e+05	759.208655	2062	0.422541	2
3	77005	0	11090.0	4983523.0	0.611542	6140.0	0.553652	2012	449.370875	9958464	720	29.718435	-95.423555	9959184	2884.068417	1.296016e+06	1596.770070	1725	0.155546	2
15	77019	0	11840.0	5572872.0	0.463007	4190.0	0.353885	2012	470.681757	9115666	104823	29.754150	-95.409498	9220489	3325.795313	1.565391e+06	1176.949524	1964	0.165878	2
20	77024	0	17570.0	8973976.0	0.526010	7830.0	0.445646	2012	510.755606	32905019	105158	29.771225	-95.511751	33010177	1378.547316	7.041008e+05	614.344080	737	0.041946	2
48	77056	0	11100.0	4857996.0	0.512613	4550.0	0.409910	2012	437.657297	8960556	34523	29.748202	-95.468948	8995079	3196.066195	1.398782e+06	1310.099206	1566	0.141081	2

```
folium_df[folium_df['cluster'] == 4].head()
```

	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	restaurant count	restaurants per household	cluster
73	77094	0	4590.0	843334.0	0.552941	2290.0	0.498911	2012	183.732898	29196393	810402	29.769285	-95.681292	30006795	396.178445	72791.013646	197.657656	174	0.037908	4
77	77401	0	7990.0	2316414.0	0.581477	4190.0	0.524406	2012	289.914143	9715941	1966	29.704019	-95.460905	9717907	2129.471389	617363.872146	1116.706523	807	0.101001	4
84	77479	0	33720.0	4987378.0	0.511447	14970.0	0.443950	2012	147.905635	85627045	4020265	29.566996	-95.636016	89647310	974.199881	144089.651661	432.496211	242	0.007177	4

Figure 15: Zip Codes from Clusters 2 and 4

The manual examination proved the zip codes were well clustered. For example cluster 2 had all of the zip codes with incomes over $400k, representing most of the hot spots from Figure 4. Cluster 4 contained two of the zip codes with relatively high incomes and normalized income scores that were more prominent in Figure 6. This is even clearer when scatterplotted with normalized income score against average household income

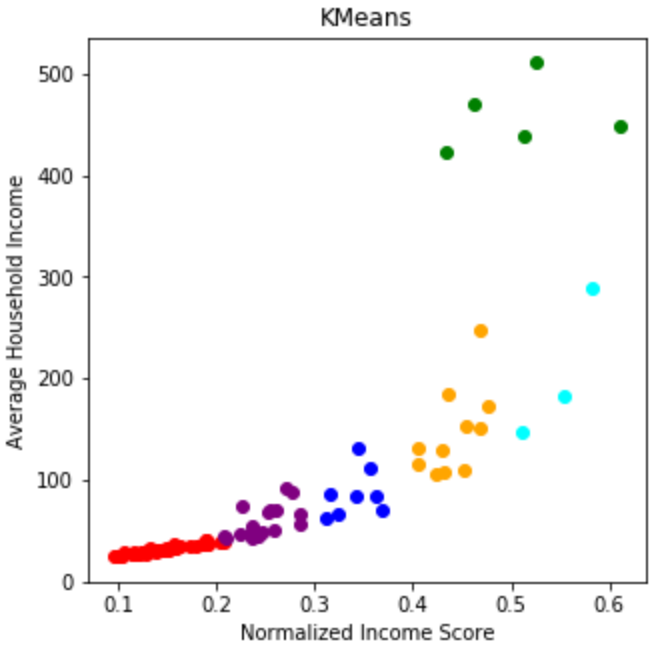

Figure 16: Scatterplot of Original Clusters

After reviewing the clusters, it appears that clusters 0, 1, and 3 can be eliminated, leaving only the higher-income zip codes.

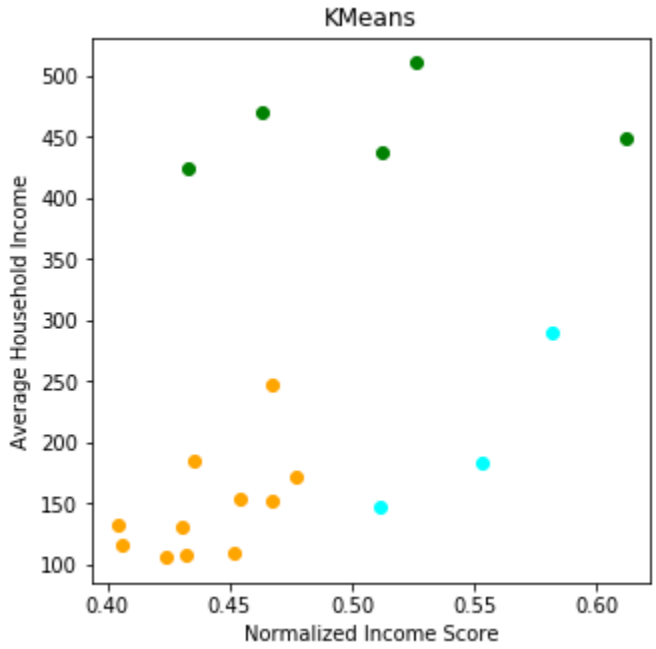

Figure 17: Scatterplot of High-Income Clusters

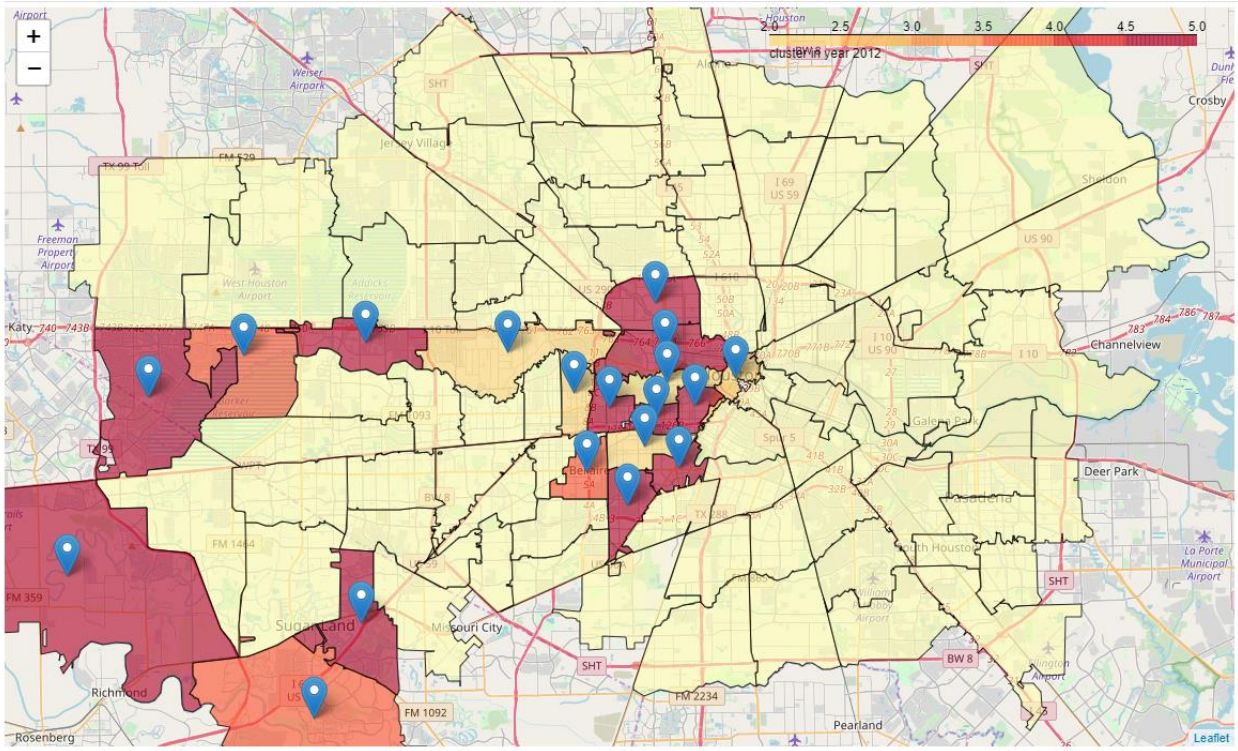

Figure 18: High-Income Zip Codes in Houston

After eliminating the lower-income zip codes, the remaining zip codes are those centered around the center of Houston and the suburbs with heavy energy-industry activity in the southwest and western areas of the city. This has reduced the analysis from 92 zip codes to 19.

3.7. Correlate Clusters to Restaurant Concentration

Now that the number of zip codes has been reduced to a manageable number, it's possible to examine most relevant data together. A quick examination of income and restaurant data for each zip code revealed that one of the high-income zip codes only has 18 restaurants.

```
report_columns = ['zip code',
                  'returns',
                  'total AGI',
                  'normalized income score',
                  'total high earners',
                  'percent high earners',
                  'average household income',
                  'restaurant count',
                  'restaurants per household']
houston_HI_df[(houston_HI_df['year'] == '2012') & (houston_HI_df['income bracket'] == 0)][report_columns]
```

	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

Figure 19: High-Income Zip Code Data from 2012

One of the reasons for this very low number of restaurants (0.001 per household) in zip code 77406 is because it's a large zip code in the western suburbs of the city. When the 2-mile radius was drawn around the center of the zip code, it left a lot of the zip code unqueried, covering an area that's almost completely made up of housing developments. By contrast, it's clear from Figure 20 that the 2-mile radius in the central zip codes overs not only most of the zip code area, but also much of nearby zip codes. The map indicates that zip code 77479 also has a low coverage and low number of restaurants per household, but it's a relatively wealthy zip code with enough restaurants to work with, so it was left in the dataset while 77406 was subsequently removed.

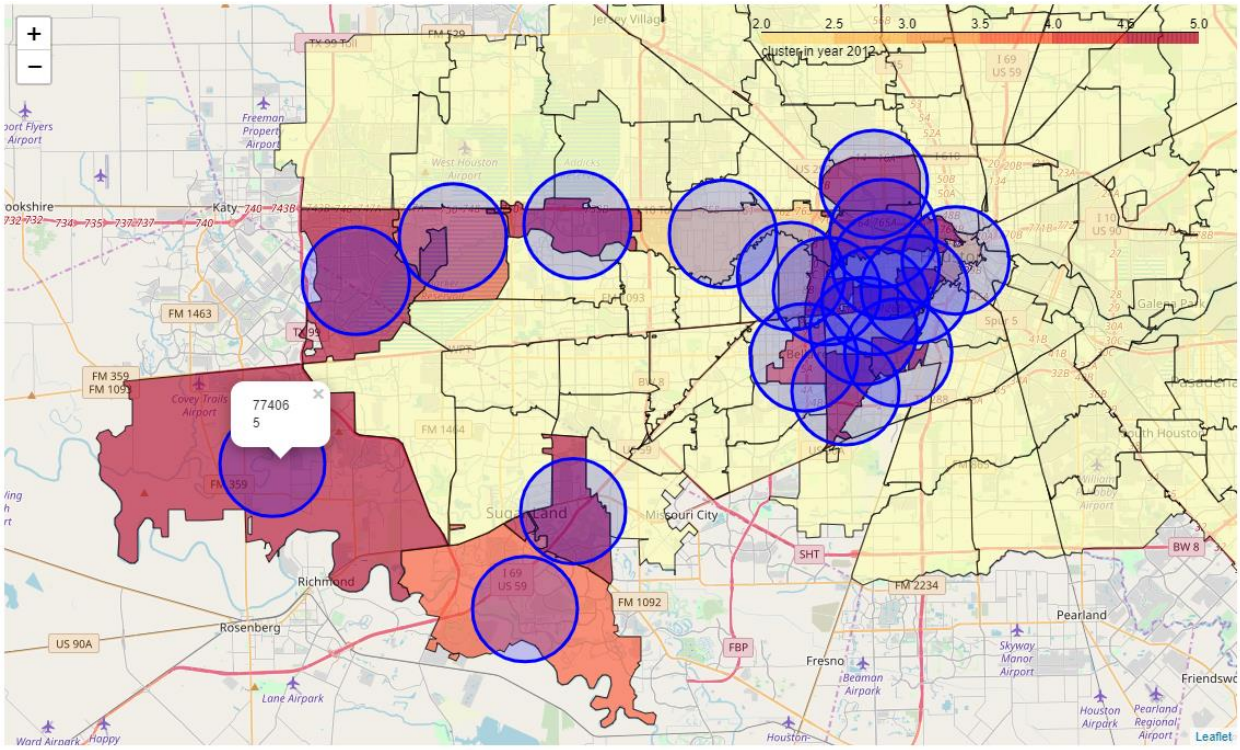

Figure 20: Restaurant Search Radius for Each Zip Code

3.8. Identify Outliers

With the clusters created and the restaurant data associated with them, the next step taken was to identify outliers within the cluster as possible opportunities. This can be done analytically and/or visually, but plotting the data visually was the easiest place to start. It's clear from the plots that the 139 remaining different restaurant categories are not all still valuable. Several issues are immediately apparent:

1. Within a cluster, not all restaurant categories are present
2. There are many restaurant categories that have very low representation in all zip codes, making relative outliers almost meaningless
3. There are some zip codes that have very high outliers for a specific reason, like proximity to a heavy ethnic concentration

The first two issues were easily resolved with data manipulation while the third can be interpreted on a case-by-case basis.

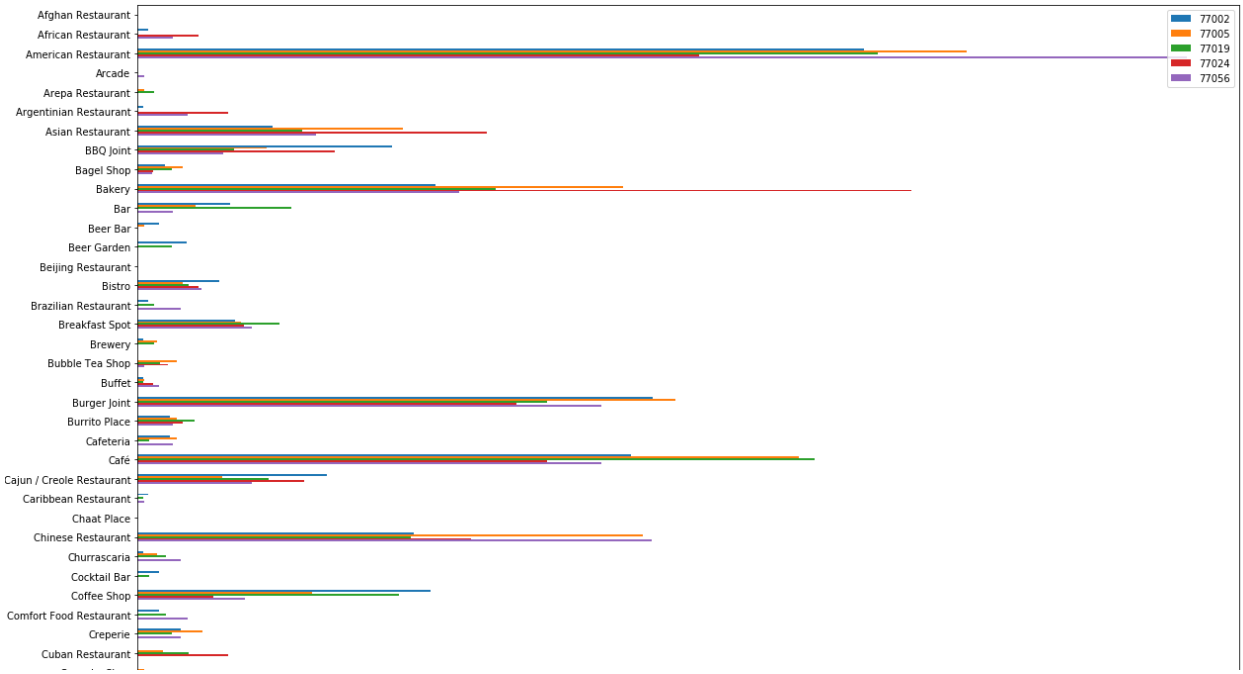

Figure 21: Bar Plot of Restaurant Category Frequencies for Cluster 2

The clusters were filtered to restaurant categories with a mean frequency exceeding 1% and then re-examined. This greatly reduced the number of categories to consider. The intersection of these results also showed the categories that were common to all high-income clusters.

```
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
```

Figure 22: Filtered Cluster Categories

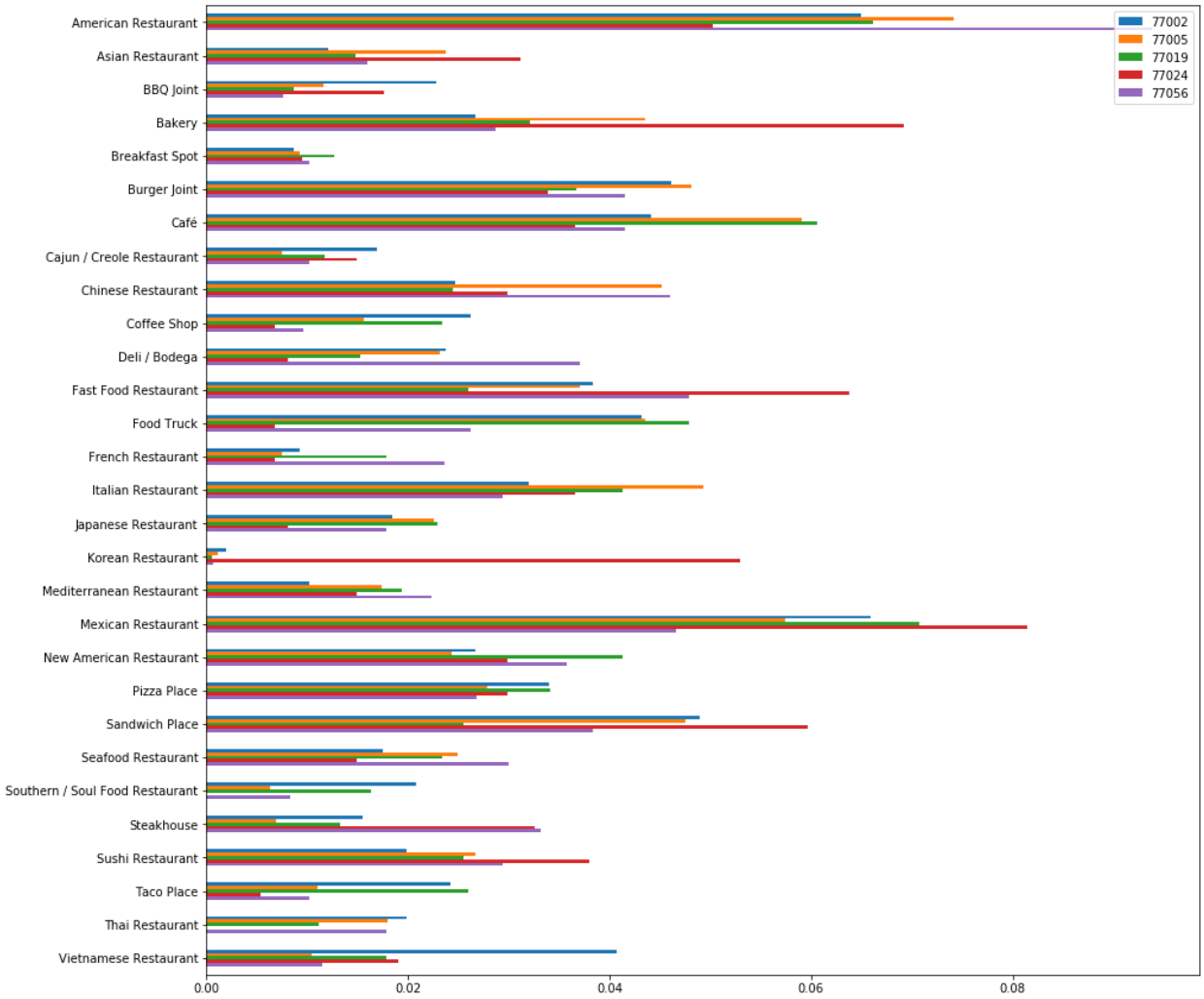

Figure 23: Restaurant Distribution for Cluster 2

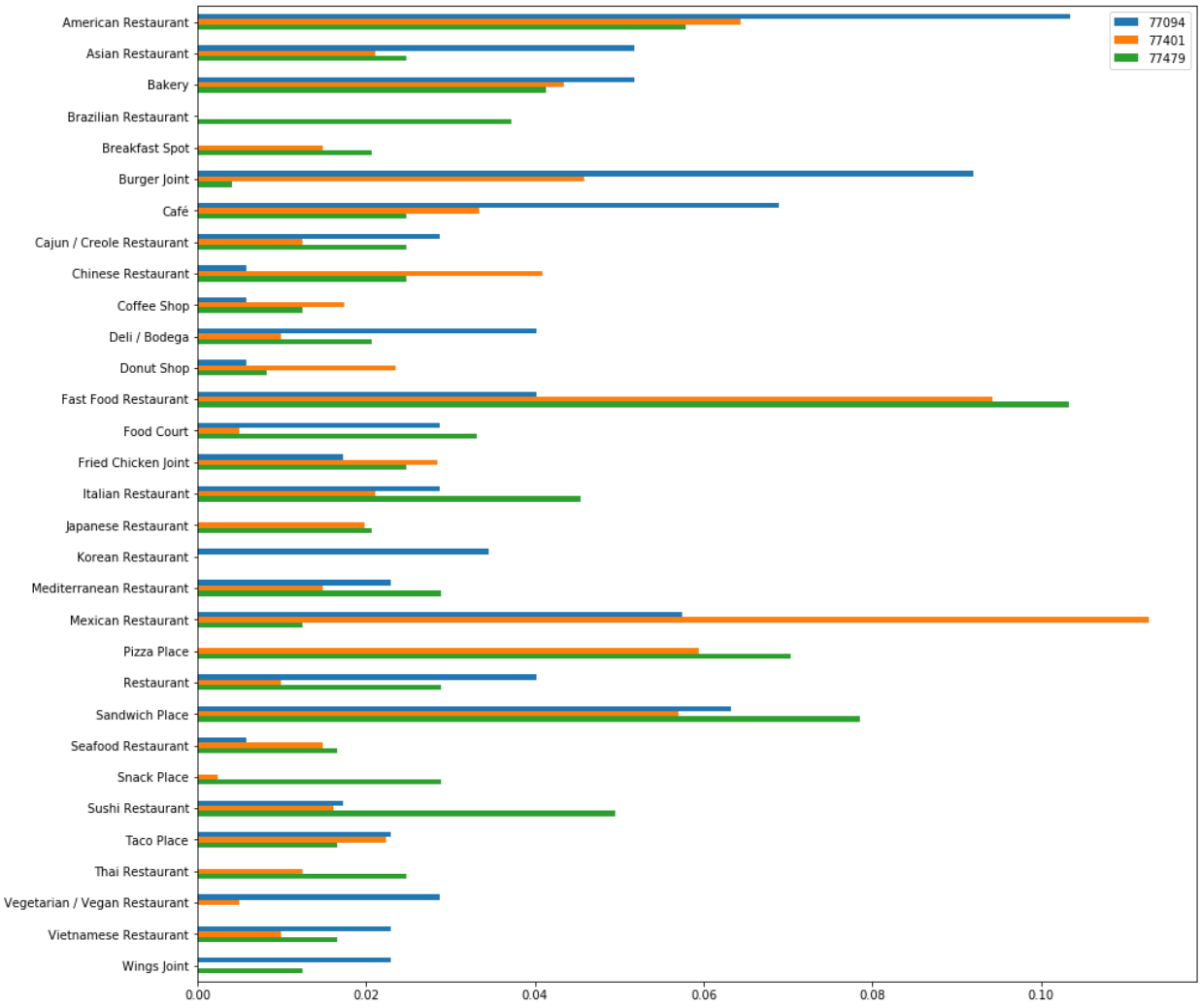

Figure 24: Restaurant Distribution for Cluster 4

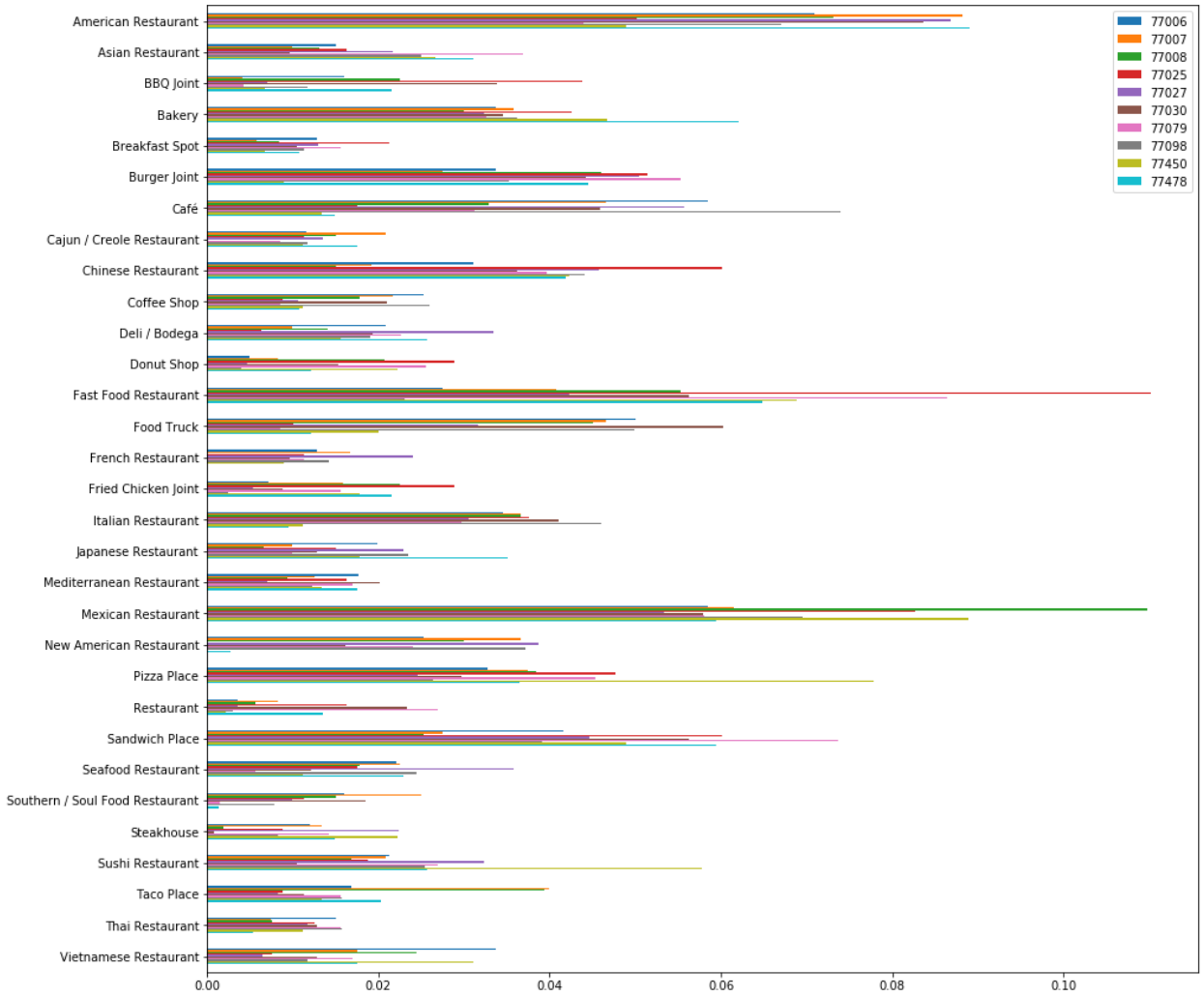

Figure 25: Restaurant Distribution for Cluster 5

Once the data sets are more manageable in size, it's easier to visually identify outliers. Examining Figure 24 reveals that zip codes 77094 and 77401 have no Brazilian restaurants, 77094 has no breakfast spots, 77479 has a relatively low number of burger joints, and so on.

This visual examination is essentially comparing the frequency of a restaurant category for a particular zip code to the general trend of the other zip codes in the cluster, so this can be accomplished programmatically by comparing the frequencies to the mean for that particular category. If the value for a zip code is well below the mean, perhaps less than one third, it may be worth looking at more closely.

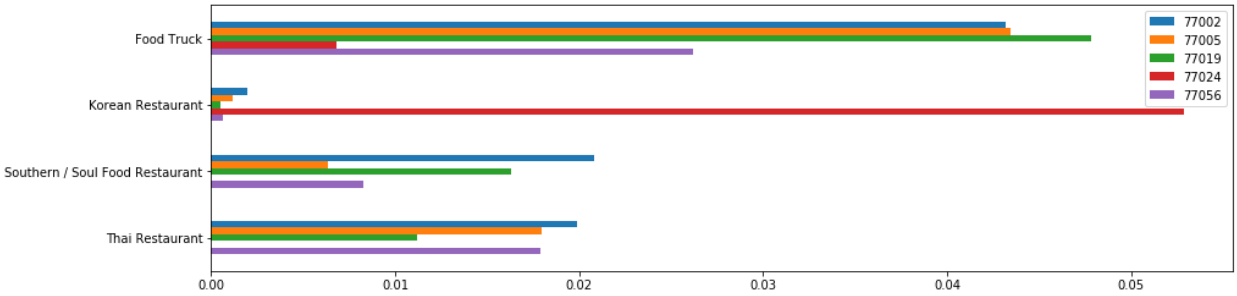

Figure 26: Restaurant Distribution for Cluster 2 Outliers

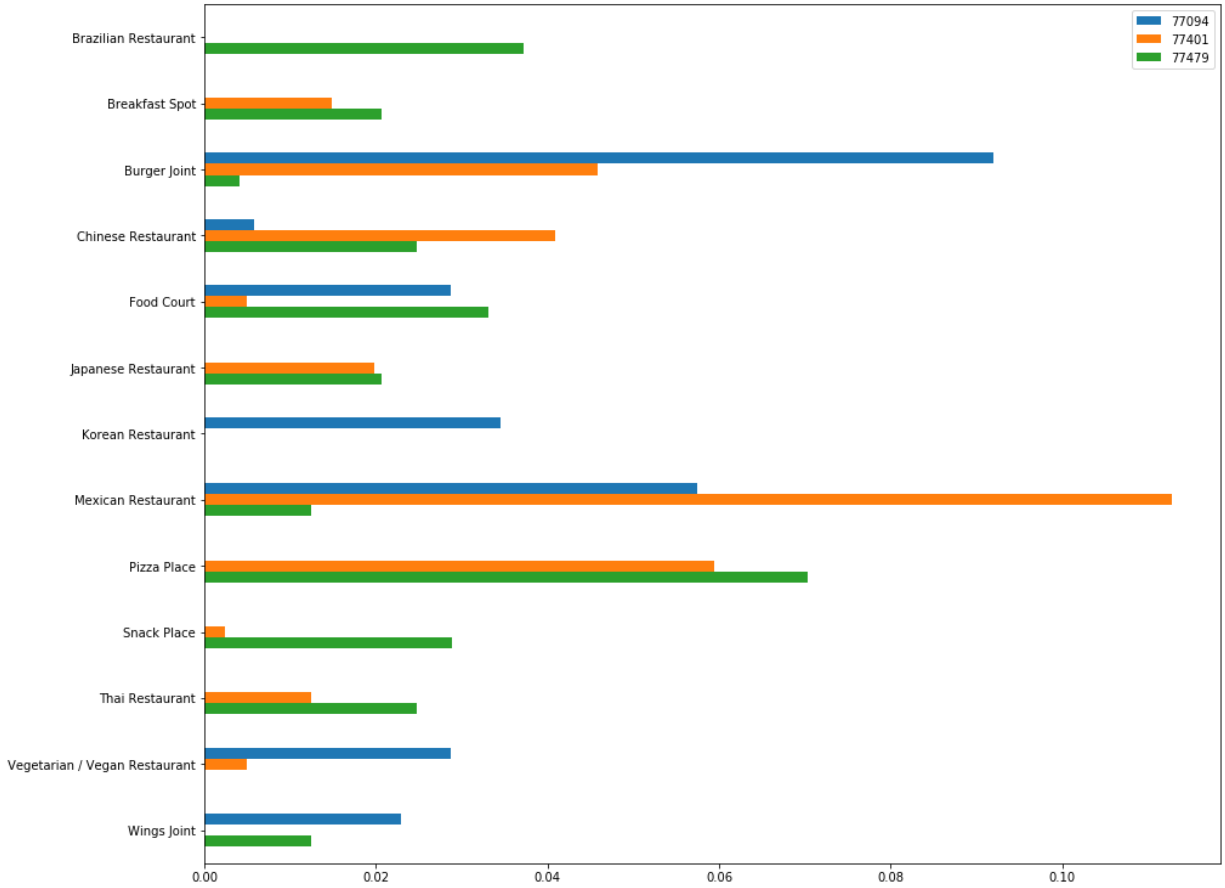

Figure 27: Restaurant Distribution for Cluster 4 Outliers

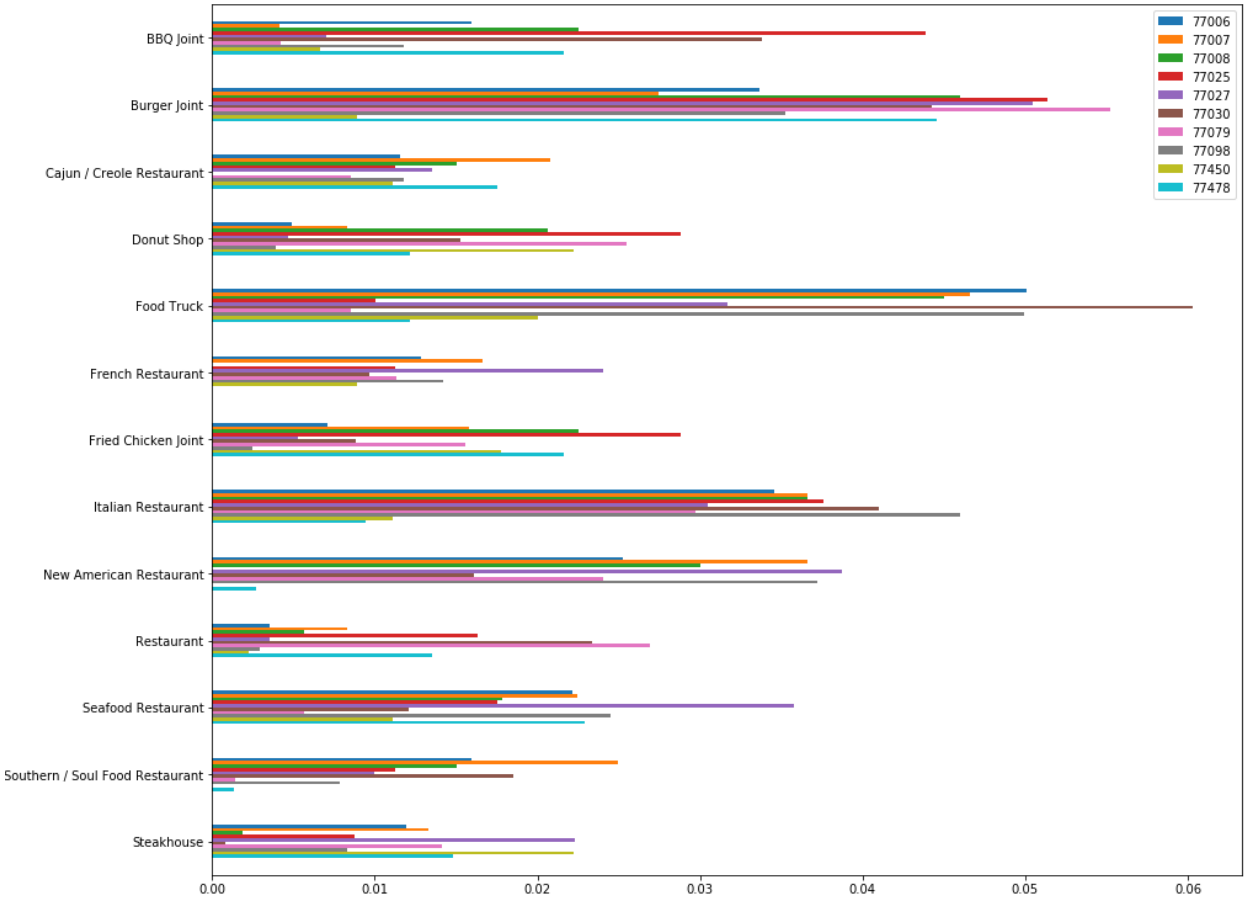

Figure 28: Restaurant Distribution for Cluster 5 Outliers

Another set of horizontal bar plots allows another visual evaluation of distributions that can be summarized in a table of opportunities.

Cluster	Zip Code	Restaurant Category Opportunities	Notes
2	77002	None	Proximity of 77024 to Korean neighborhood skewed results
	77005	Possibly Southern / Soul Food	
	77019	None	
	77024	Food Truck Southern / Soul Food Thai	
	77056	Possibly Southern / Soul Food	
4	77094	Breakfast Spot Chinese Thai	Anything present or largely present in only one of the 3 zip codes was ignored. Would have worked better with a larger set
	77401	Food Court Wings	
	77479	Burger Joint Mexican	
5	77006	Donut Shop	10 zip codes made for clearer distributions and opportunities
	77007	BBQ Joint	
	77008	French Steakhouse	
	77025	Food Truck New American	
	77027	BBQ Joint Fried Chicken Joint	
	77030	Cajun / Creole Steakhouse	
	77079	BBQ Joint Food Truck Seafood Southern / Soul Food	
	77098	Donut Shop Fried Chicken	
	77450	BBQ Joint Burger Joint Italian New American Southern / Soul Food	
	77478	Possibly Food Truck French Italian New American Southern / Soul Food	

Table 3: Restaurant Opportunities from Outliers

There were several occasions where a very strong positive outlier made many others appear to be negative outliers / opportunities (e.g. Korean in Cluster 2); these categories were ignored for the other zip codes. The lower number of zip codes in Cluster 4 also increased relative variation, making the conclusions more suspect. It's also interesting to note that although the high-frequency categories were

fairly common between the three clusters, there were no restaurant categories in common to the categories with outliers.

4. Results (Week 2)

The analysis conducted in this exercise suggests that there are opportunities for new restaurants in 16 of the 18 identified high-income neighborhoods in Houston. This analysis is based on zip code groups that were clustered by income and the evaluation of relative accessibility of restaurants to each zip code in the clusters.

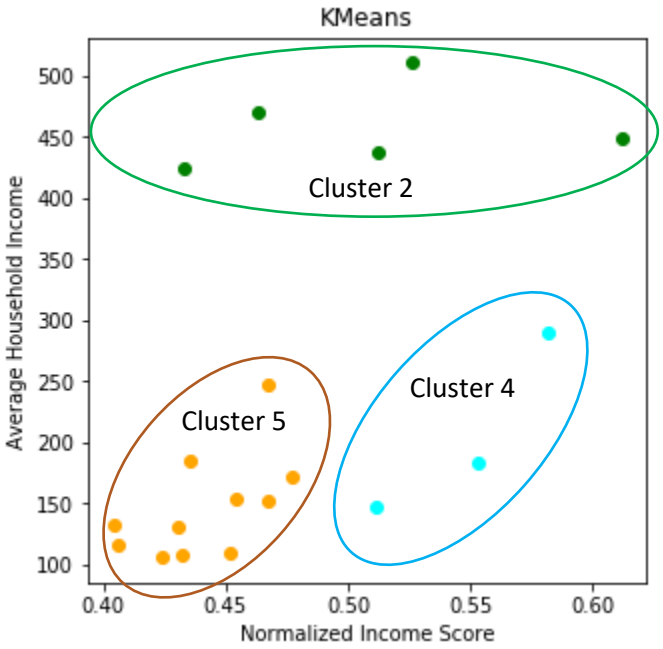

Figure 29: Analyzed High-Income Clusters, from Figure 17

The clusters chosen for analysis represent the high-income zip codes in Houston, with all having an average household income greater than $100k. The restaurant categories are compared in context with the other zip codes in the cluster to identify outliers that may represent an underserved market for that restaurant type.

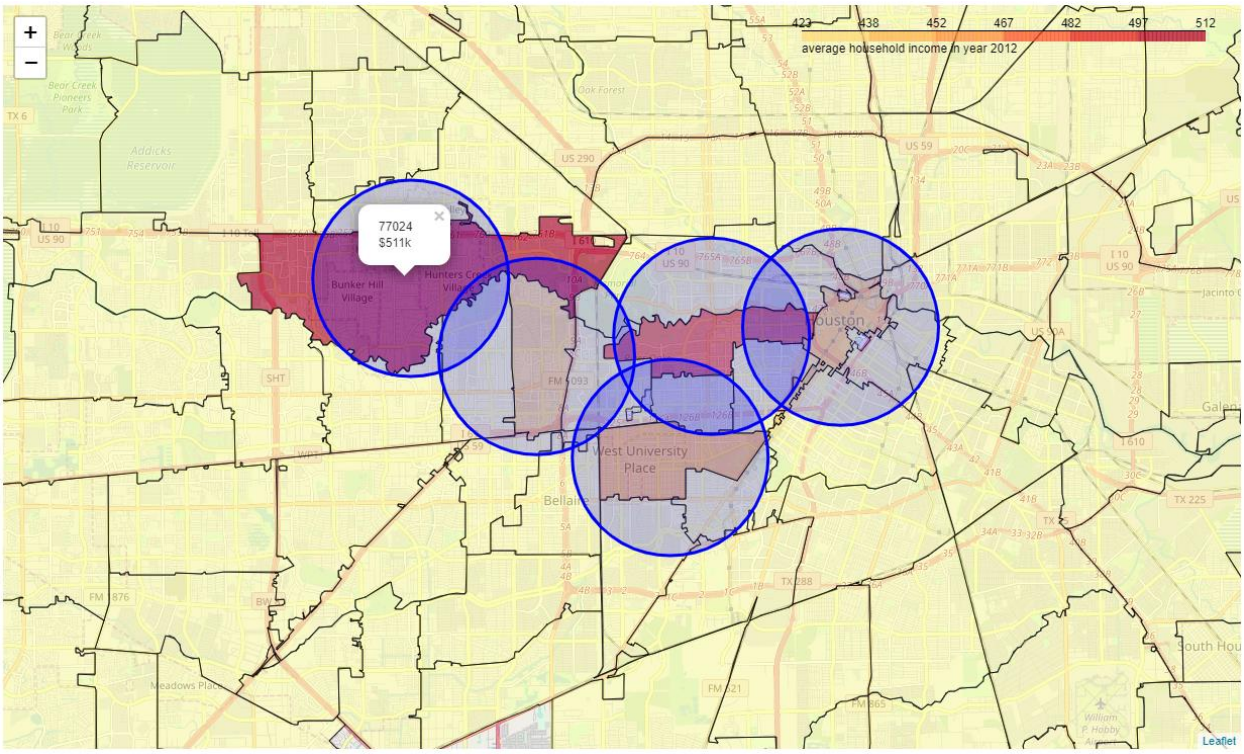

Figure 30: Zip Codes and Restaurant Circles for Cluster 2

Cluster 2, shown in Figure 30, represents the highest income zip codes in Houston and is found closer to the western center of the city. Outlier analysis indicated there were fewer opportunities in this cluster than in the others, which may partly have been skewed by the proximity of zip code 77024 to Korea town just north of I10 near Gessner. It does appear that there may be opportunities for new Southern / Soul food restaurants in zip codes 77005 around West University and 77056 around the Galleria area.

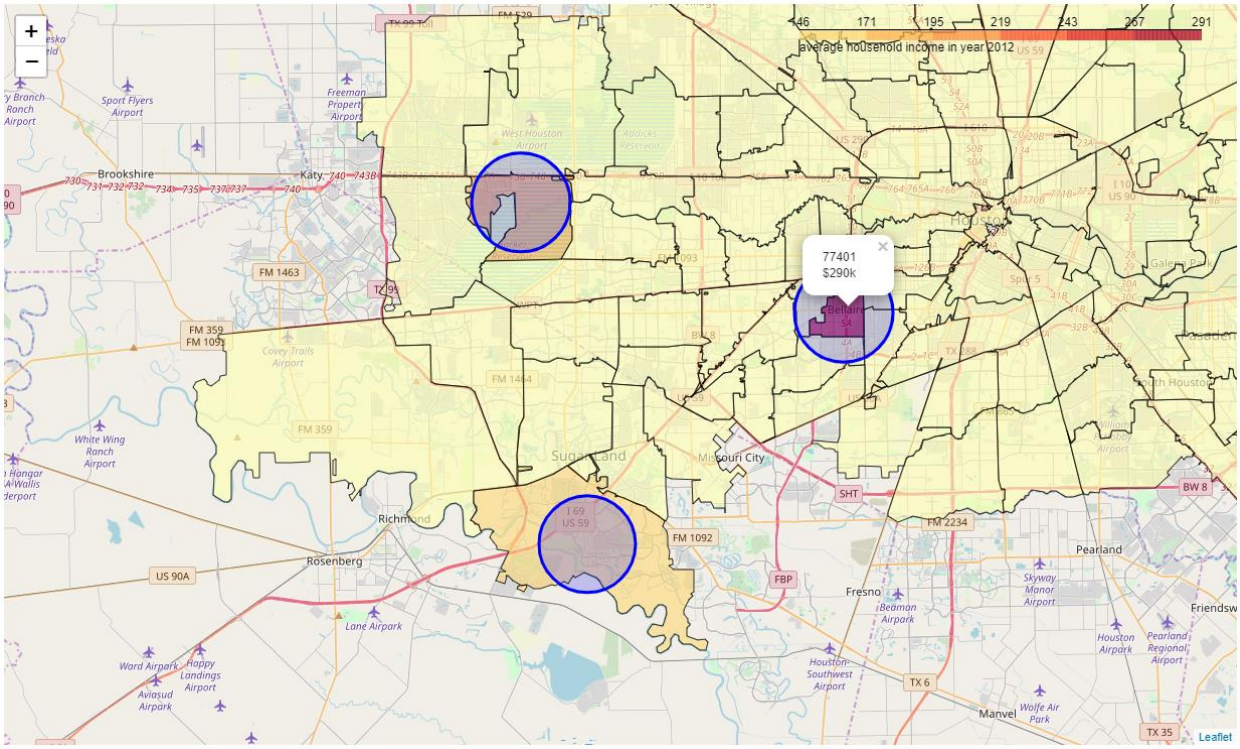

Figure 31: Zip Codes and Restaurant Circles for Cluster 4

Cluster 4, shown in Figure 31, consists of 3 zip codes spread around the western and southwestern portions of the city. Their economic similarity comes from a high average income combined with a high percentage of households falling in the higher income brackets. Their geographic similarity is not immediately apparent until the industry of the city of Houston is considered. Zip code 77094 on the west is home to the "Energy Corridor" where many energy industry companies are headquartered, and 77478 in the Southwest is a more affluent suburb home to some energy companies and many neighborhoods that feed the downtown energy corporations. Zip code 77041 is a high-income residential neighborhood immediately adjacent to the Galleria area, another major hub for energy activity.

The opportunities in Cluster 4 were somewhat less clear because of the small number of zip codes considered, but those opportunities should be investigated per the outliers shown in Figure 27 and Table 3. Burger Joint and Mexican Restaurant for 77479 and Pizza Place for 77094 seemed to be the most significant opportunities.

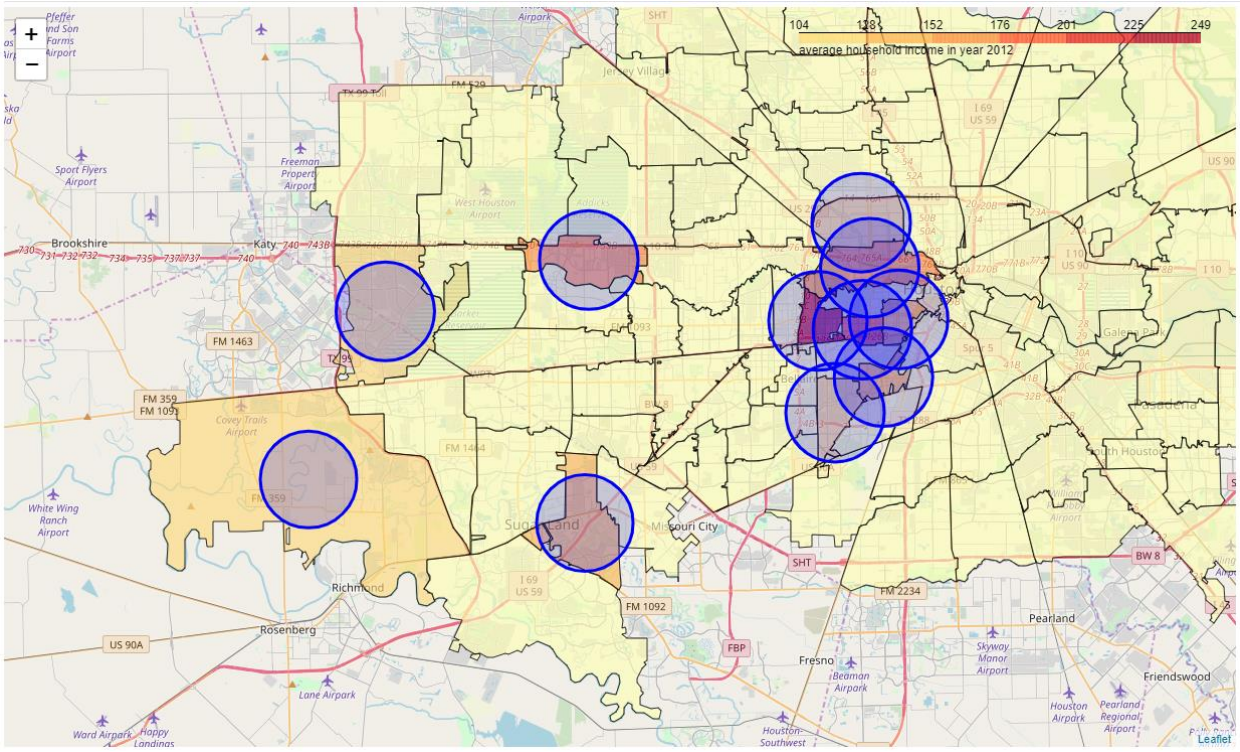

Figure 32: Zip Codes and Restaurant Circles for Cluster 5

Cluster 5, shown in Figure 32, consists of the remaining high-income zip codes in central Houston plus some additional high-income zip codes from the suburbs featured in Cluster 4. Geographically, it looks like this cluster could have been split between the other two clusters, but economically it's significantly lower average income than Cluster 2 and lower top-end income than Cluster 4.

Figure 27 and Table 3 in the previous section list some of the many possible restaurant opportunities found in these zip codes. There are multiple interesting examples where most of the zip codes have a significant amount of a certain category while one or more similar zip codes have none whatsoever (e.g. 77030 with no Cajun / Creole, 77008 and 77478 with no French, or 77450 with no Southern / Soul Food). A more in-depth analysis may reveal some traps with this data. 77030 has no Cajun / Creole, but it has the second highest frequency of Southern / Soul Food. These don't necessarily overlap, but if the categories didn't have as many distinctions as they do, they could easily both be considered "Southern Food" or something similar, and the outliers would no longer exist.

5. Discussion (Week 2)

The data processing, clustering, and further data processing used for this exercise convincingly suggest that reasonable zip code targets have been found and that restaurant opportunities exist, but there are clear improvements that could be made to the process.

1. The only demographic information accounted for is income
    a. The data assumes that outliers represent an abundance or shortage of a restaurant type in a zip code by comparing its frequency to others in the same cluster

    b. Houston is a diverse city with diverse neighborhoods, but there are still clear pockets of certain ethnic populations.
    c. The popularity of a restaurant category may be related to the ethnic population (e.g. Korean Food in 77024, which is adjacent to Koreatown) rather than an overserved restaurant category
    d. The use of restaurant frequency in these situations may conceal actual underserved categories by skewing the results
2. Household demographics were not considered in the results
    a. Dependent information was available in the IRS data, but was not used
    b. There are some assumptions that could be made relating household size to restaurant type, or to zip code residence, but these were not attempted
    c. It would make sense to include family size in further analysis to avoid opening a family-oriented restaurant in a neighborhood of single people or an upscale restaurant in a suburban residential neighborhood
3. Income trends over time were not considered
    a. The data was originally gathered intending to compare trends in income to trends in restaurant frequency
    b. This was based on the mistaken understanding that use of a date in the Foursquare API would pull information from that date, which is not the case
    c. When it was seen that only current restaurant data was available, the methodology was changed to clustering with outlier analysis
4. The restaurant categories may be overly distinct
    a. 157 different restaurant categories were gathered by the initial venue search, with 141 remaining after limiting the zip codes to the high-income clusters
    b. Many of these categories were considered subsets of other categories within Foursquare, but it's not a clear to decision whether to keep them distinct or to lump them together
    c. A clear example can be seen in Figure 22, where out of the 22 categories with frequencies exceeding 1% common to all clusters, 6 could easily have been considered to fall under a more general "Asian Restaurants" category (Asian, Chinese, Japanese, Sushi, Thai, Vietnamese)
    d. Further grouping restaurant categories could have simplified analysis, but would have then been more general/vague for what type of restaurant was missing from the area.
5. Proximity does not reflect travel time
    a. A single radius of 2 miles was used for all zip codes to gather nearby restaurants
    b. This could represent a dramatically different travel time depending on the population density and road accessibility of the zip code
    c. A better metric would have been based on travel time rather than proximity, but this was out of the scope of the exercise
    d. A related issue was that for some zip codes, the radius used did not even encompass the entire zip code. It's hard to believe the needs of that zip code population were really met by the metric, but it should at least still provide a good starting position for a more in sophisticated analysis

6. Conclusion (Week 2)

An analysis was conducted on the income data, geography, and restaurant category concentration for the city of Houston to determine likely restaurant opportunities. The high-income neighborhoods were considered after a clustering algorithm was applied to all the zip codes in the larger Houston area, and a list of different categories of restaurants that were underrepresented in 16 high-income zip codes was compiled. A restauranteur targeting the high-income population of Houston could start with these zip codes and associated restaurant categories to determine the best option for a new restaurant.

	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

Table 4: Restaurant Opportunities (from Table 3)

7. References (Week 2)

"2010 TIGER/Line® Shapefiles." United States Census Bureau, https://www.census.gov/cgi-bin/geo/shapefiles2010/main (accessed April 21, 2019)

Gilmer, Bill. "Proximity Counts: How Houston Dominates the Oil Industry." *Forbes*, Forbes Magazine, 23 Aug. 2018, www.forbes.com/sites/uhenergy/2018/08/22/proximity-counts-how-houston-dominates-the-oil-industry/ (accessed April 21, 2019)

"OpenDataDE/State-zip-code-GeoJSON." GitHub Repository, https://github.com/OpenDataDE/State-zip-code-GeoJSON (accessed April 21, 2019)

"SOI Tax Stats - Individual Income Tax Statistics - ZIP Code Data (SOI)." United States Internal Revenue Service, https://www.irs.gov/statistics/soi-tax-stats-individual-income-tax-statistics-zip-code-data-soi (accessed April 21, 2019)

"Venue Categories." Foursquare Developers, https://developer.foursquare.com/docs/resources/categories (accessed June 27, 2019)

8. Acknowledgements (Week 2)

Many functions and algorithms used in this project were adopted from previous courses in the Data Science Professional Specialization on Coursera

9. Appendix A (Week 2)

This report, the associated presentation, and the Jupyter notebook used to perform the analysis can all be found in the GitHub repository https://github.com/ggiem/Coursera_Capstone.