Opening a New Restaurant in Houston

Greg Giem

June 27, 2019

Executive Summary (Week 2)

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 regression analysis associating restaurant categories with zip codes over time 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
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)

4. Results (Week 2)

5. Discussion (Week 2)

6. Conclusion (Week 2)

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)

9. Appendix A (Week 2)

10. Appendix B (Week 2)