Data Scientist Relocation Identifier Program

by Hyun Gu Kang, PhD
Coursera/IBM Data Science Professional Certificate Capstone Project
December 4, 2019

Executive Summary

Newly minted data scientists likely need to or have the ability to relocate to new places that meet their professional and personal interests. Where are the likely places to which they would relocate? This question is important for both data scientists and employers alike.

Tax return data from the US Internal Revenue Service, commercial venue location data from FourSquare, geographical data from the US Census 2010 were combined with my professional and personal interests to (1) identify the types of desirable areas using K-means clustering, (2) predict areas of high desirability using regression methods. Data from Texas (U.S.) area were considered for this pilot project. Proximity to data science jobs, income, households with children, home ownership, organic food, and choir (singing) venues were considered.

Identifying the types of desirable neighborhoods lead to five types of areas: (1) Urban family settings with neighborhood amenities, (2) urban "yuppie"/single areas with amenities (3) suburban family settings with few amenities, (4) high-income "estates" areas with few amenities, and (5) rural areas with minimal amenities.

When combined with my professional and interests, the usual urban areas in Texas (Dallas, Houston, and Austin) were identified. In addition, surprisingly, few areas in west Texas (San Angelo and Rock Springs area were identified as possible desirable areas. These areas appear to be spurious findings at first glance, and more investigation is needed.

As a pilot, this project offers minimal features suited to my interests. A fully developed application would allow a user to list more specific list of interests and preferences, allow users to sort preferences, and be integrated with real estate and job board search engines. Aggregated data from this service would be used to inform employers on locating new offices.

1. Introduction

Data science is a fast-growing field. There are many data scientists being newly minted, who are looking for work. Since data science positions also pays well, a data scientist may be able to relocate to an area that best matches various preferences. Likewise, for employers, strategically locating offices near such desirable locations will help it attract talent.

The goals of this project is to identify neighborhoods in Texas – as defined zip codes – that offer different amenities or cater to different demographics which will help decision making for the individual. As a pilot project, it will feature hard coded features that align with my personal interest. Texas was chosen as the study area for this pilot project, because of its several useful features. Texas encompasses a very large geographical area with diverse population, diverse population density, and industry concentrations. This will make statistical modeling more robust, as there is much varying data available. Texas also features a growing tech sector and pro-business atmosphere that may cater to certain employer's needs.

2.1 Data:

Three main sources of data are as follows:

Neighborhood characteristics data was derived from the **U.S. Internal Revenue Service** aggregated tax return data. This dataset includes some information about incomes, home ownership, dependents, etc., sorted by zip codes in Texas. Data from other states are available and could be included in future work. The dataset includes a list of numbers of tax filings from each zip code in Texas, which can be used to infer the number of working and earning households in each area. For each zip code, the number of filings, amounts filed in different income categories (AGI: adjusted gross income) [<$25,000, [USD $25k, 50k), $50k-75k, $75-100k, $100-200k, and $200k+], as well as number of tax exemptions and the type of them, including dependents, home ownership, etc. For those readers not familiar with the U.S. tax system, Adjusted Gross Income (AGI) refers to wages and other contract work minus work-related expenses.

Of note, the IRS is not the best source for estimating ownership or owner-occupancy. Nor is this the best source for children or services for families nearby. However, it summarizes financials and demographic information succinctly.

Neighborhood venue data was derived from the Foursquare. FourSquare API was used to query the server for each geographic location. Because of its nature, FourSquare data does not query for all the venues in a given geographic boundary area, and each query is limited to 100 responses. Therefore, the venue information is a rough estimate based on a 1 mile (1.61 km) radius around the centroid location of each zip code, rather than within the boundaries.

Neighborhood Shape data was derived from the US Census Bureau. Instead of using zip codes (postal codes of the US), the data is based on ZCTA (ZIP code tabulation areas that excludes zip codes with no areas associated with them).

These data sources were supplemented by:

Zipcodes – from Python USZipcodes library 0.2.4, Presumably, this data is based on Census2010 data, but the documentation. This library provided the area (size) of each zip code, location of each zip code, population density, number of housing unit, etc.

2.2 Features to be extracted from each dataset

US IRS data – for each zip code:

- Estimated median income – this will be based on the number of filings in each category.
- Estimated mean income – this will be the arithmetic mean of AGI, calculated from the total income for the zip code divided by the number of filings.
- Number of households (population) – rather than demographics/census data, this report will use only tax filers (which ignores children or low-income people).
- Proportion of homeowners – this will be estimated from the Number of homeowner exemption claims divided by the "population." This will indicate owner occupancy of the area.
- Proportion of households with children – this will be estimated from the Number of "dependents" exemption claims divided by the "population". This feature indicates whether the households in this zip code have children

After examination of the data, mean income, and number of households were dropped from the dataset. Income data is much skewed, and therefore arithmetic mean is not a useful metric. USZipcodes provides population density, and therefore population value was dropped as it is no longer necessary.

Foursquare – for each zip code:

- **Venues (density)**: The number of venues as identified by the Foursquare API within a 1-mile radius. Due to the hard limit of 100 items per search, and due to large zip codes with high number of venues (>100), the search radius for the venues was limited to 1609 meters or 1 mile. This indicates venue density, or the presence of neighborhood features. Here, only the count, or the density of venues in the area are considered. This is a feature to indicate how commercial the area may be and likely has many neighborhood amenities, whatever they may be.

- Venue density of **data science jobs**. Job venue density was determined with a 100 km radius (100,000 m radius). Although FourSquare is not designed to map jobs, this is a proxy measure for relevant job opportunities. In future implementations, this could be replaced by job posting sites.
- Personally relevant features:
    - **Choral** venue density within a 10 km radius. Search word was 'choir' and does not include all the variation, as this keyword yielded useful distribution of hits. Yes, I would like to have access to live choral music for listening or for participation.
    - **Organic** venue density within a 10 km radius. Search word was 'organic.' This includes restaurants or grocery stores.

US Census Shapefile – for each zip code or ZCTA:
- Geographical boundary information was obtained.

<u>USZipcodes library – for each zip code</u>

- [Land] area of each zip code

- **Population density**. [the exact data source is unclear]

- Number of housing units

- Latitude and Longitude (used for mapping/query, not as a descriptive feature). Physical location could be used as a clustering feature, and would consider physical location as one of the clustering variable.

All of these features were hard coded, but eventually can be a part of interactive interface that can be customized to each person's preferences

2.3 Analysis Process

2.3.1 Data Cleaning

The IRS dataset is a standalone Excel file. Due to the large number of merged cells that does not cleanly import using Pandas library, this was manually cleaned in Excel before being imported

Each zip code occupies 7 lines, divided into different AGI income categories as well as the aggregate. This was cleaned and aggregated into individual features for each Zip code.

2.3.2 Data Merge

FourSquare data was obtained and merged with the cleaned neighborhood data using location information available from the USZipCodes library.

Here are the features and the correlation matrix between them:

	kids	income_median	home	pop density	area	Venue	data science	organic	choir
kids	1.000000	-0.174438	-0.132270	-0.138634	0.003333	-0.393811	-0.067740	-0.306344	-0.107228
income_median	-0.174438	1.000000	0.783091	-0.067679	-0.050867	0.095760	0.131913	0.172305	0.109721
home	-0.132270	0.783091	1.000000	-0.089031	-0.115141	0.026223	0.247146	0.081825	0.114860
pop density	-0.138634	-0.067679	-0.089031	1.000000	-0.405301	0.698830	0.193007	0.632719	0.499968
area	0.003333	-0.050867	-0.115141	-0.405301	1.000000	-0.321931	-0.267357	-0.260297	-0.311927
Venue	-0.393811	0.095760	0.026223	0.698830	-0.321931	1.000000	0.156257	0.612297	0.433876
data science	-0.067740	0.131913	0.247146	0.193007	-0.267357	0.156257	1.000000	0.231590	0.238479
organic	-0.306344	0.172305	0.081825	0.632719	-0.260297	0.612297	0.231590	1.000000	0.349659
choir	-0.107228	0.109721	0.114860	0.499968	-0.311927	0.433876	0.238479	0.349659	1.000000

2.3.3 Neighborhood classification

K-means nearest neighborhoods, an unsupervised learning algorithm, was used to classify the neighborhoods into five (5) categories. The choice of five is was chosen to make the interpretations easier, as well as manual ranking of the neighborhoods:

Clustering Results

- Cluster 0: lower income, lower home in the household, high density, high venue density, including data, organic, and choir: up and coming urban
- Cluster 1: lower income, low density. - I call this suburban with kids
- Cluster 2: higher income, lower kids, or home with high density - Yuppie neighborhood
- Cluster 3: high income, lower venues density - The estates
- Cluster 4: lower income, lower home, lower density or venues, and low data science - rural

Next, I gave these a desirability score by rank ordering them according to my preferences:

- Cluster 0 [up and coming) - score = 4 (affordable, has desirable venues)
- Cluster 1 (suburban) - score = 2 (lacks desirable venues)
- Cluster 2 (yuppie) - score = 3 (has desirable venues)

- Cluster 3 (estates) = score = 1 (too expensive, even with a data scientist income)
- Cluster 4 (rural) - score = 0 (lacks jobs)

In an interactive system, the user could be asked to rank these groupings to determine which best fit their preferences. Coming up with clever cluster descriptions would remain at human hands, however.

2.3.4 Regression

Once the neighborhoods were labeled with a desirability score, regression was used to identify most desirable neighborhoods.

Two different regression models were used: 1. Random Forest 2. Linear. KNN was not used as the desirability estimates were built using a clustering.

Random forest was used because it is a powerful classifier, and the exact nature of the model isn't very important. Random forest provided R^2= 0.9 on the partitioned testing data set.

Random forest identified Dallas, Houston, and Austin as top neighborhoods.

Linear regression was used because of its simplicity. Linear regression produced R^2= 0.75.

Linear regression identified only Austin and Houston as top neighborhoods.

Linear Regression Predictions

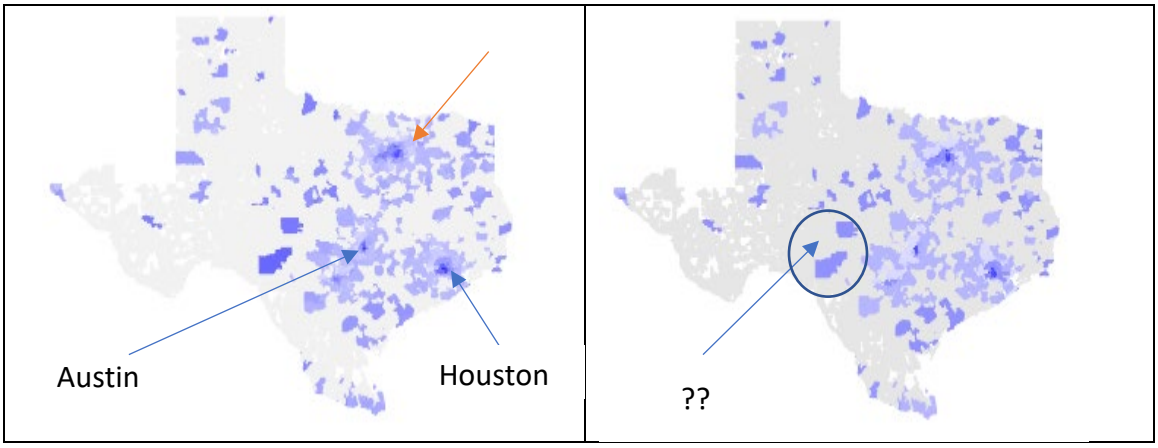
city	lr prediction
Austin	5.461262
Houston	4.970501
Austin	4.745056
Houston	4.709116
Houston	4.506680
Houston	4.506342
Houston	4.455803
Austin	4.397563
Houston	4.368335
Austin	4.327493
Austin	4.233221
Houston	4.178711
Houston	4.133718
Houston	4.129572
Austin	4.119329

Random Forest Regression Predictions

	Zip	city	rf prediction
108	75219	Dallas	4.0
99	75209	Dallas	4.0
441	77057	Houston	4.0
440	77056	Houston	4.0
431	77046	Houston	4.0
427	77042	Houston	4.0
807	78702	Austin	4.0
415	77030	Houston	4.0
412	77027	Houston	4.0
841	78751	Austin	4.0
404	77019	Houston	4.0
481	77098	Houston	4.0
806	78701	Austin	4.0
395	77010	Houston	4.0
96	75206	Dallas	4.0

2.4 Visualization.

Choropleth maps were created to show was created. Unfortunately, due to technical difficulties in creating a GeoJSON file that worked with Folium library, a static map was created by simply graphing the shapefile points.



Linear Regression	Random Forest Regression

3. Results Summary

The five general types of area identified areas that were "up and coming," "suburban," "yuppie," "estates" and "rural."

The major urban areas were identified as desirable locations (for my tastes) – Dallas, Houston, Austin, although the two regression methods differed on the order.

Surprisingly, a few areas in central-west Texas were identified as being desirable. These include San Angelo, TX and Rock Springs, TX. A casual inquiry into these neighborhoods did not yield any clues as to why these were ranked high.

4. Discussion

Based on my rankings, the major urban areas were selected as the most desirable neighborhoods for my tastes and preferences. This is not unexpected, as major urban areas tend to offer both venues as well as jobs.

The few areas in central-west Texas was a surprise. More investigation is needed to determine if these are spurious findings, or if there are highly desirable features in these neighborhoods that deserve a closer look.

The features selected for classification are correlated to each other, and therefore do not always yield insights into meaningful clusters. The correlation matrix indicates that all personal venue densities (overall, choral, organic) are strongly related, and therefore may not yield new insights. However, *data science venue density* was not strongly correlated with other variables, indicating that either [1] this adds new information that previous features do not or [2] FourSquare data is too noisy or unsuitable for estimating this feature. Connection to the job board site, such as SimplyHired, LinkedIn, etc. may improve this.

5. Limitations and Future work.

Currently, this code identifies the features that I would like to have. An interactive interface can be developed that allows any user to specify features that are particularly relevant. Even if the features are correlated, it would provide additional weight in a K-means clustering approach.

Use of FourSquare was limited to 100 venues per search. A way to go around the data limitations – 100 limit, radius based search needs to be identified.

Zip codes can be very large area tracts. Once zip codes are identified, then a similar process can be undertaken within smaller areas within the zip codes. Many real-estate websites such as Redfin and Zillow could be integrated to create a hyperlocal home search function that integrates venue information. For example, my personal ideal house would have very low population/venue density within 5 minutes of walking distance (300 meters), and then have very high density of desirable venues just beyond 5 minutes.

Rather than physical distance as used in FourSquare, traffic pattern data could be used to estimate driving time/commute time from work as particular feature of a particular house in a real estate market.

The model was built using my expectation for my interests and needs. Instead of being prescriptive using venue preferences, the model could be descriptive: training the model using home addresses of current data scientists – this would indicate the neighborhood characteristics of high desirability areas of data scientists for home locations. Regression on this data model would indicate how a given area would be attractive to a general data scientist. This model, or an aggregate of the data in the prescriptively developed model could be used to inform employers on job locations.

This small project does not replicate all the features of other location services, although this would be great. Real estate information available in Zillow, Redfin, or other MLS database companies, such as school districts, tax rates, neighborhood walkability, etc. is not incorporated. It will also not incorporate data from U.S. counties that offer property ownership, property value or tax status. "Good for kids" information from yelp.com are also not included.