Coursera Capstone Project

Author: James Crandall

**Introduction**

Within the last 20 years, Houston, Texas has been an exponentially growing city for housing, population, and the energy industry. With a flood of people flocking to the city for jobs, the diversity of food and beverage selections has grown with it. Though Houston has always been known for being a multicultural city, it has greatly increased the Millennial population. Millennial’s, fueled by social media, social hangouts, social everything, are looking for institutions that allow for gatherings. With the club night life still breathing heavily, the daytime festivities are increasing with social drinking, primarily at breweries.

Microbreweries have erupted in popularity to offer as an alternative to bars and other day time hangouts. The selection of breweries come and go as fast as the fads last on social media.

**Problem Statement**

What is needed is a brewery that is not surrounded by other breweries, that can tend to the local/immediate population, has good beer, and can stay busy. My client is looking for the ideal place for her brewery to call home. It will serve local clienteles and be able to ship beer to eateries and groceries. Having not only an indoor/outdoor entertainment space is critical, but there also needs to be space for brewing, packing, shipping, trucking, and of course drinking (sampling).

**Data**

Foursquare offers great access to business geographic data, along with reviews, ratings, specific business types, etc…

The goal is to leverage this data, isolate establishments that relate to drinking, brewing, alcohol as a few of the keywords. Preferably, areas that have little to no data of surrounding establishments would be a plus, as there might be a market for my client to serve as well as areas that have data, many establishments, but the ratings are less than favorable for those. This could allow for someone new to come in and woo the existing customer base.

Houston is one of the largest cities by area in the United States, isolating areas may be best. Clustering neighborhoods first will be ideal to narrow down options, possibly to target certain clients like Millennials.

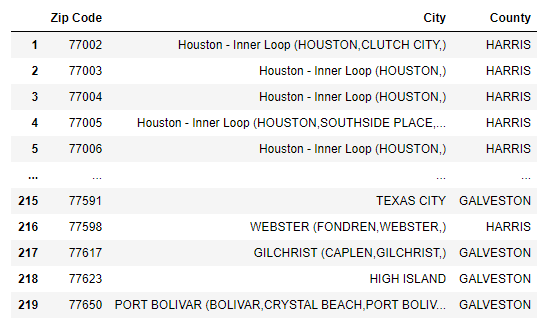
I plan to join existing geographic data of Houston with Foursquare to appropriately cluster areas based on existing infrastructure. Then look at ratings, and existing businesses to determine potential competition and find a location suitable for my client.

**HAR**

The Houston Area Realtors (HAR) website was used to collect initial city data within Houston.

This is to determine the base line data for the following categories

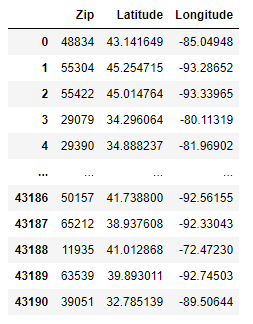
* Zip Codes
* Cities
* Counties



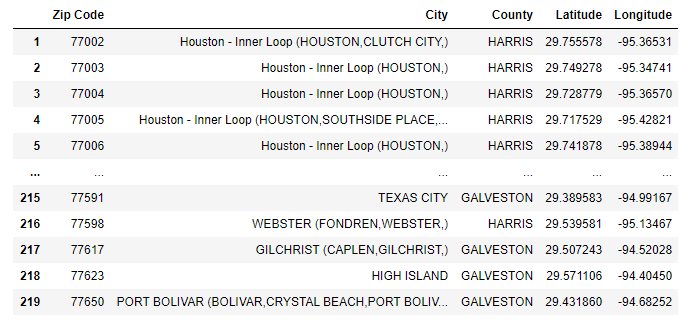
**USA Gov Data**

The United States government data was then utilized to give an initial list of the following categories

* Zip Codes
* Latitude Coordinates
* Longitude Coordinates



The HAR and USA Gov data were joined to consolidate Zip Codes, Cities, Latitude and Longitude coordinates



**FOURSQUARE**

Foursquare uses geolocational data related to businesses and establishments, along with other attributable factors such as reviews, timestamps, and ratings.

Foursquare was leveraged to include Venue data based on the desired location. Our selection was Houston, Texas

The initial set up to access Foursquare data required the use of creating an account as a developer. This allowed for the initiation of an access token, client ID, and client secret used for retrieval.

Once the account had been established, access is allowed using python script.

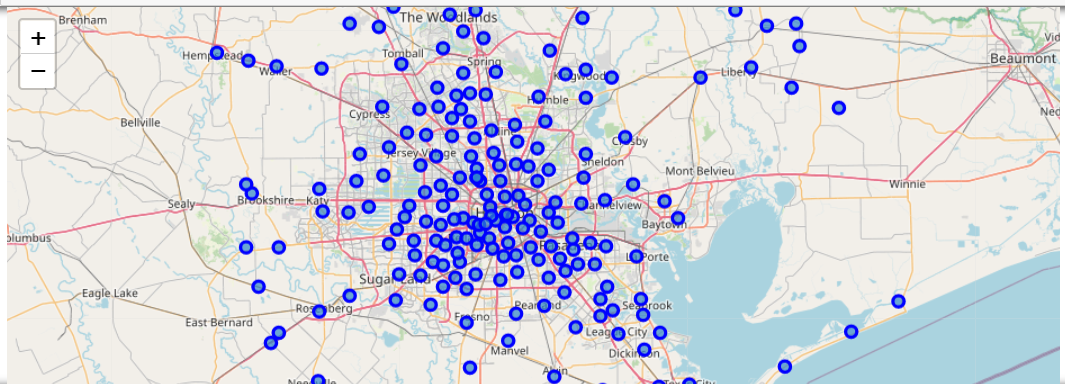
**Methodology**

Prior to retrieving Foursquare data through the use of an API call, we must establish the location of interest, in this case, Houston, Texas.

Nominatim and Folium packages are used within python to establish our address and map our Zip Codes



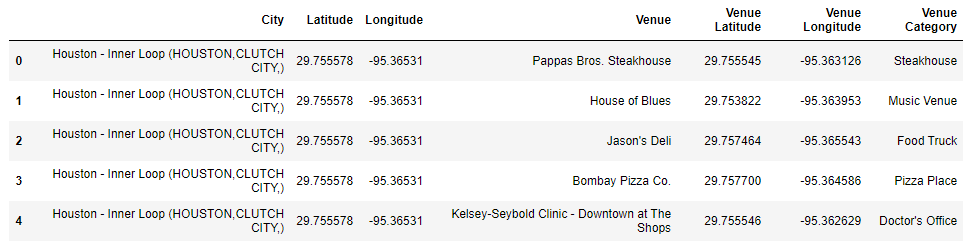




Now that the location of interest has been established the use of the Foursquare platform can be fully leveraged through python

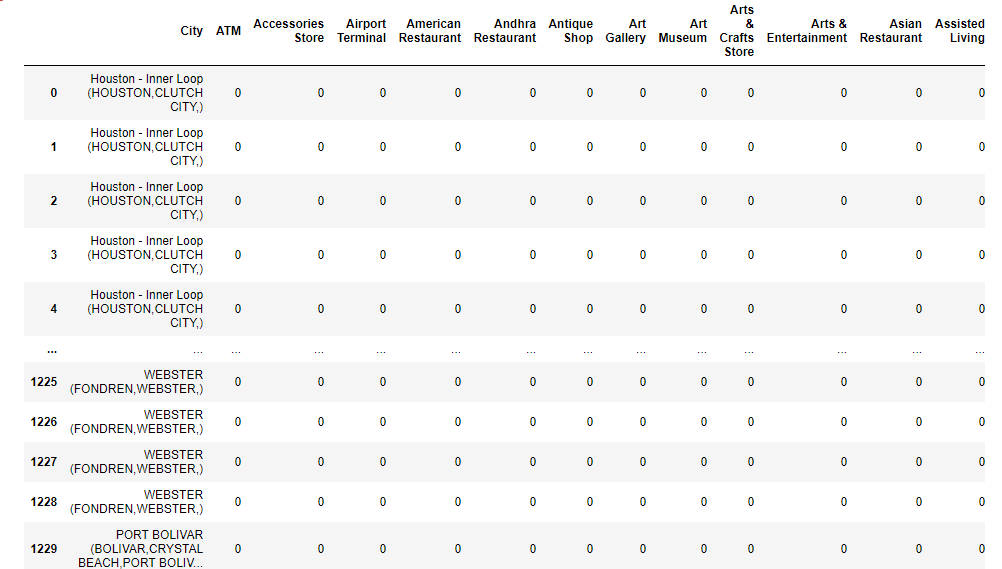


1 Foursquare Script for Inputs



2 Foursquare Script for Output

Next is the analysis of the Foursquare data. Grouping our venue categories by City

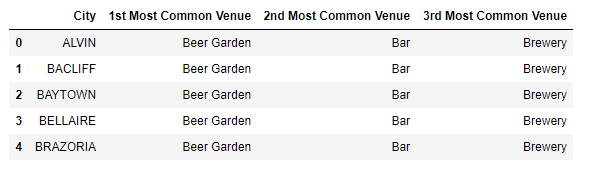


Next is the isolation of venues labeled “Bar”, “Beer Garden”, “Brewery” to isolate alike venues



Followed with establishing the top 3 venues in order of frequency within each city





3 Final Output for Data frame

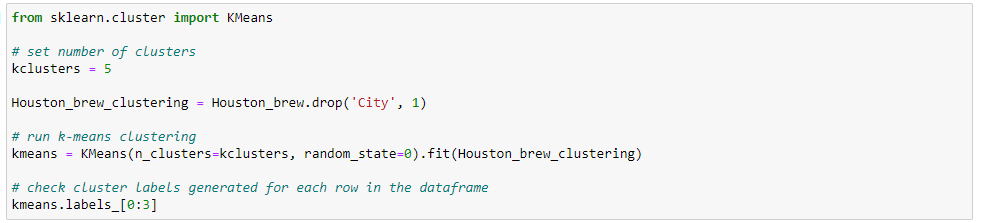
**Analysis**

The goal is to determine the best location to establish our brewery, ideally we want to avoid as much established competition as possible in the search.

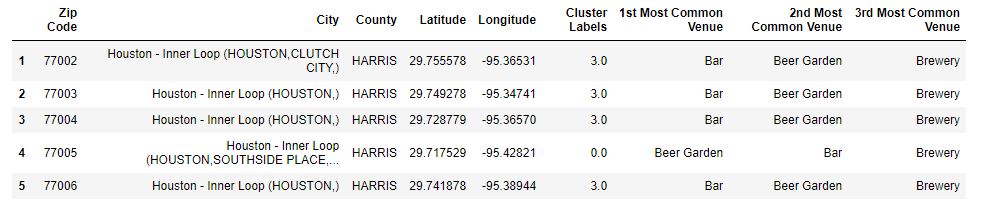
We use an unsupervised learning technique called k-Means.

According to the Wikipedia article “***k*-means clustering** is a method of [vector quantization](https://en.wikipedia.org/wiki/Vector_quantization), originally from [signal processing](https://en.wikipedia.org/wiki/Signal_processing), that aims to [partition](https://en.wikipedia.org/wiki/Partition_of_a_set) *n* observations into *k* clusters in which each observation belongs to the [cluster](https://en.wikipedia.org/wiki/Cluster_(statistics)) with the nearest [mean](https://en.wikipedia.org/wiki/Mean) (cluster centers or cluster [centroid](https://en.wikipedia.org/wiki/Centroid)), serving as a prototype of the cluster”. We are using k-Means to determine, based on commonality of venue rankings, which Cities are alike with each other.

We need to define each city to a cluster based on the nearest neighbors of alike qualities in venue commonality.



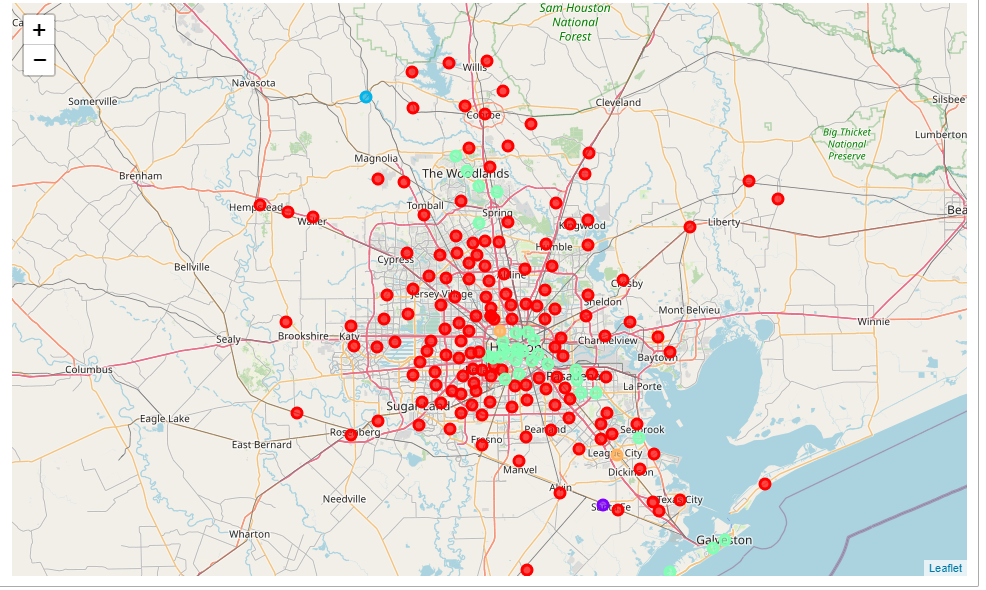
4 k-Means implementation



5 Cluster Labeling to Cities

**Results**

As the final application of Cluster Labeling defines which city belongs to which cluster, it is helpful to circle back around to graphically show how this is interpreted on a map of Houston with the corresponding Zip Codes



6 Map of Houston with Clusters

Clustering Outcomes

* Cluster 1 – Red
  + Largest presence of Beer Gardens and Bars per Zip Code
* Cluster 2 – Purple
  + Bar is Dominate
* Cluster 3 – Light Blue
  + Beer Garden and Bars are near equal weight
* Cluster 4 – Seafoam Green (in table)
  + Largest Concentration of Bars and Beer Gardens
* Cluster 5 – Orange
  + Beer Gardens, Bars, and Breweries all equal weight

**Discussion**

The map shows the clustering representation of clustering based on venue commonality per zip code.

Biggest Notes

Within the 610 Loop is highly dominate of Bars, Beer Gardens, and Breweries (Cluster 4)

* This makes sense as a large population of millennial’s live in this area. Those that encourage social engagements over many other things. Bars and Breweries encourage this behavior.

Outside of the 610 Loop is available for a new establishment as the concentration of competing venues is negligible (Cluster 1, 2, 3, & 5). Reasons for this include the following

* Suburban areas have fewer bars, beer gardens, and breweries per capita than inside downtown areas
* Fewer establishments in the clusters are found through Foursquare
  + As Foursquare uses input data from other applications it is up to the establishments to determine if they want their data available through Foursquare
  + Particular clientele could not be users of applications that leverage foursquare, resulting in fewer venues available to pull from per Zip Code

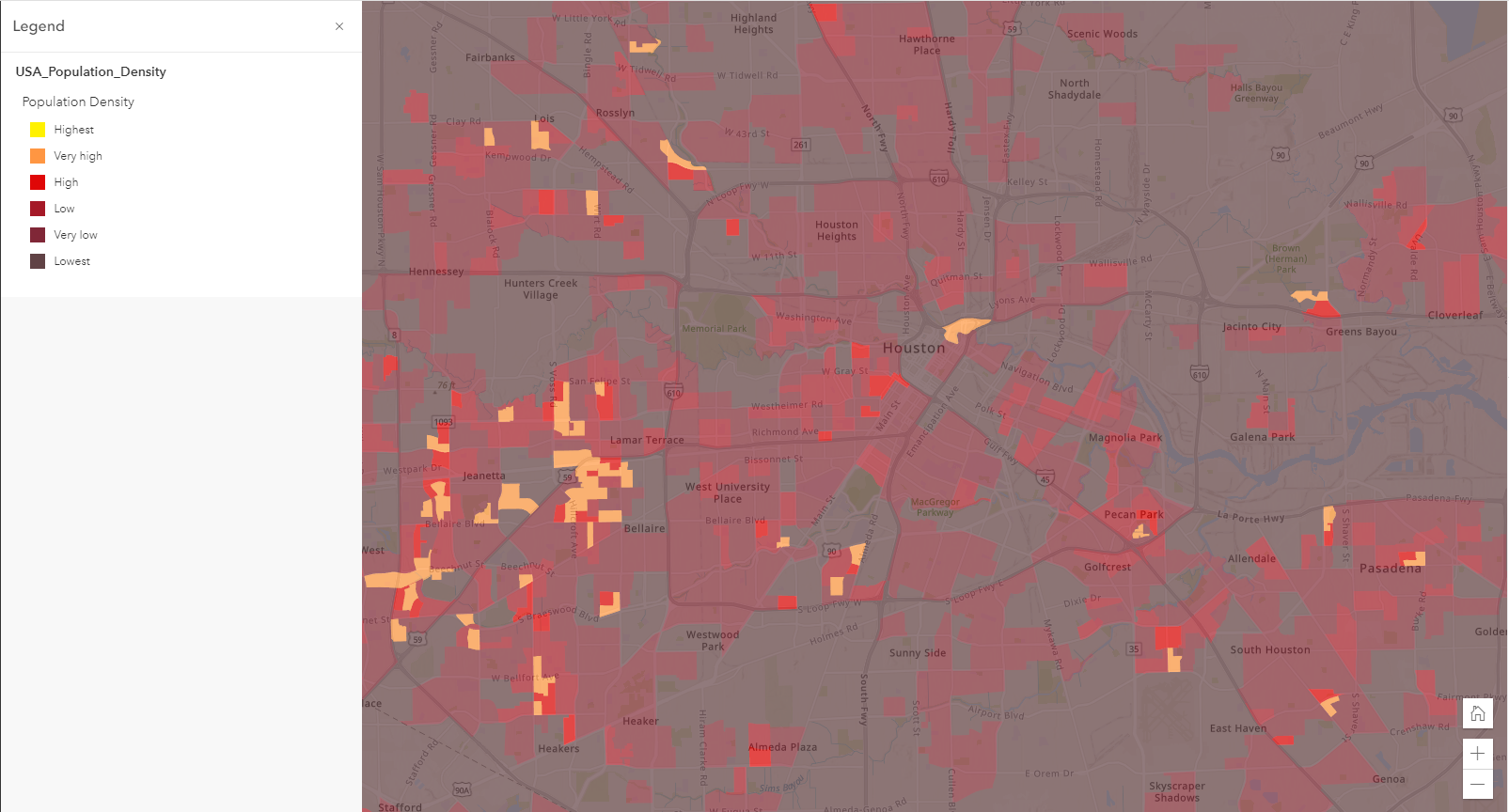
**Conclusion**

It is to my best knowledge that the client should seek anywhere outside of the 610 Loop where the density of Brewery like establishments is the highest.

The further away from the center of the city, the more spread out becomes of the population. It may be wise to still remain close to the 610 Loop to attract customers but not be within competition of the existing establishments.

My best suggestion is to be near the highest density of population but still within Cluster 1 (red)

Below is a map of population density matched with areas of Houston



7 Population Density by Zip Code (ESRI)

Given that the population density is the highest to the West – South West of 610 Loop (yellow). I would first consider building my brewery there. Still close to the 610 Loop but far enough to avoid competition.