# **Where can we start an Indian restaurant in Austin?**

**Raghuram Moturi**

**December 30, 2020**

****

## Introduction

## Background

## Austin is the capital of Texas in the United States and is one of the fastest growing cities in America. It was recently voted the No. 1 place to live in America for the third year in a row (U.S. News & World Report) and was ranked No.4 of the best large cities to start a business (WalletHub). According to Austin City Government, the City of Austin has crossed the threshold of becoming a Majority-Minority city, meaning that no demographic group exists as a majority of the City's population. One notable trend is the growing number of Latino and Asian households.

## Being a fast-growing city with diverse ethnicities, Austin is a great place for entrepreneurs to start and grow their business. The city is also well known for its outstanding food and great live music venues.

## Business problem

## Having visited Austin recently, I wondered why there aren't many authentic Indian restaurants in the Austin area and would love to see more. The objective of this project is to segment and cluster the neighborhoods of Austin using different data sources including foursquare location data to find the 'best' neighborhood to open an Indian restaurant based on the venues in the area.

## Who would be interested in this project

## The target audience of this project would be anyone that is interested in opening or growing a Indian restaurant in Austin, TX. The cluster analysis of Austin neighborhoods and demographic data will help entrepreneurs make an informed decision about which neighborhoods to aim for.

## Data Acquisition and cleaning

## Data Source

## To begin with, I gathered data on the reporting neighborhoods in the city of Austin. The neighborhood data I found is from the Housing and Planning Department of the Austin City Government: <https://data.austintexas.gov/Building-and-Development/Neighborhoods/a7ap-j2yt;>this dataset includes the names and geographic information of the different neighborhoods, and their sizes and shapes.

## Use Foursquare API to find the top 100 venues within a radius of 1500 meters of the center of the neighborhoods

## There are other things to consider when trying to open any restaurant. It is a big decision to make! Another factor I would imagine a business owner taking into account is the demographic data based ethnicity. I found a related dataset on <https://www.austintexas.gov/page/demographic-data>, it shows the information on population, race and ethnicity, and housing and density, grouped by neighborhood reporting areas in Austin (based on the 2010 Census data). These neighborhoods are the same as the one we did the cluster analysis on. Thus, we can merge the data frames together to see the neighborhoods with different Asian population densities in their clusters.

## Data Cleaning

## In order to gets the correct latitude and longitude, I have added the city name, state and country to the neighborhood names (Austin, TX, USA)

## 

## Removed the duplicate entries from the data set

## 

## Exploratory Data Analysis

## Exploring the Neighborhoods

## Created the map of Austin with the neighborhoods using the latitude and longitude values

## 

## To find the venues around the center of these neighborhoods. I have requested the data from Foursquare API and collect information for the top 100 venues in the neighborhoods within a radius of 1500 meters

## 

## Now I have grouped the venues by the neighborhood and we can see that some neighborhoods have more venues closer together while other neighborhoods' venues are more spaced out.

## 

## Exploring Indian Restaurants in Austin

## Before we get into clustering, I'm creating a new data frame with all the Indian restaurant data that was returned by Foursquare API. Since some of these venues were double counted, I will drop them in order to make a map of these restaurants.

## 

## Rendered the map of existing Indian Restaurants

## 

## One Hot Encoding

## Previously, we collected data on venues in Austin with their names and coordinates. However, to run machine learning algorithms on the data, we need numerical data about the existence of these venues. One hot encoding helps us do that by creating new (binary) columns to indicate the presence of each possible value from the original data. This means that each venue in each neighborhood will be labeled as 1 in their correct category. After this, we group the dataframe by the neighborhoods to get the mean of the frequency of occurence of each venue category.

## 

## Machine Learning-Clustering

## For this project, I used k-means clustering. To begin with, a cluster is a collection of data points aggregated together based on their similarities? Using machine learning algorithms, we can cluster the neighborhoods based on their similarities with each other. K-means algorithm, in particular, first identifies k number of centroids, and then allocates every data point to the cluster, in a way that the data point is closer to that cluster's centroid than any other centroid. K-means algorithm runs this in a repetitive fashion before the centroids are stabilized and the clusters are formed. I am using this method because it is an unsupervised learning method meaning that the algorithm will find the similarities between the data points for us given we don't know them to begin with.

## Finding the best K

## One limitation of k-means clustering is that the algorithm does not decide how many clusters to form on its own and we need to find the best K to make clustering more accurate. The Elbow Method is one of the most popular methods to determine this optimal value of k. We iterate the values of k from 1 to 10 and calculate the distortion and inertia values for each value of k in the given range. Distortion is the average of the squared distances from the cluster centers of the respective clusters while inertia is the sum of squared distances of samples to their closest cluster center.

## 

## 

## To determine the optimal number of K, we select the value of k at the “elbow” of the plots, the point after which the distortion/inertia starts decreasing in a linear fashion. Given these plots, we conclude that 4-5 clusters would work best for our data. Ultimately, I decided to go with 5

## Ran the k-means to cluster the neighborhoods into clusters

## 

## Examining the Clusters

## We created 5 clusters (cluster 0 - 4) using k-means. Now let's look at each cluster more closely.

## 

## 

## Joined the cluster labels to the Indian Restaurant dataset

## 

## 

## Data Visualization

## Created a map to examine these clusters further.

## 

## 

## 

## Results

## As we see in the map, most Indian restaurants are located in Clusters 0,2 and 3, which are include neighborhoods in the northern and southern parts of Austin. Some of the neighborhoods in Cluster 3 (blue) are in South Metropolitan area but Cluster 3 has the second lowest frequency of Indian restaurants. Overall, Indian restaurants are concentrated in north and south Austin. This is interesting information so far!

## Heat Map by Asian Population Density in Austin

## 

## Discussion

During our cluster analysis, we found that Cluster 0 has the lowest average frequency of Indian restaurants while Cluster 2 has the highest. However, Cluster 1 has the highest number of Indian restaurants, with a lower average frequency, which might be because Cluster 1 has a high number of neighborhoods compared to other clusters, and that there are other common venues in the neighborhood which makes the frequency of Indian restaurants lower. But all in all, compared to other venues, Cluster 1 doesn't have a high frequency of Indian restaurants. Based on the demographic data I found, the top 9 neighborhoods with the highest Asian population density don't have a high frequency of Indian restaurants. These neighborhoods include UT, Lake Line, Gateway, West University, and Anderson Mill and so on. It makes sense since some of these neighborhoods have a very small total population. However, highly populated college neighborhoods like UT, West University with many Indian students don't have a high frequency of Indian restaurants. I might be biased but this should be changed (because Indian food is awesome!). Other highly populated neighborhoods like Anderson Mill (with a total population of 28473!) also don't have an average high frequency of Indian restaurants either. Does this mean that someone should open an Indian restaurant there?

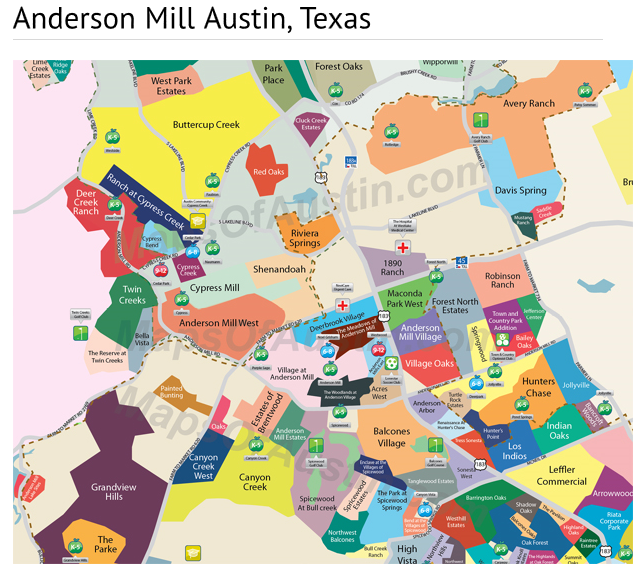
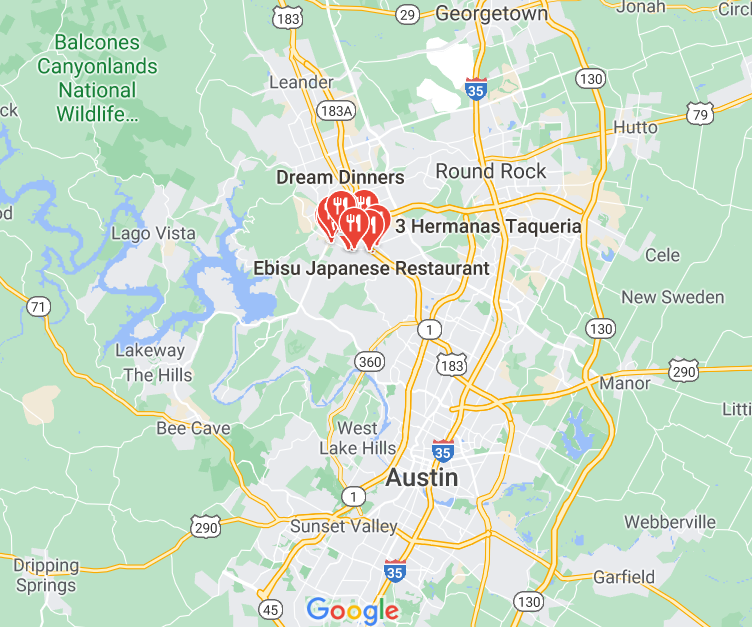
After discussing (possibly) the best neighborhood to open an Indian restaurant, we should also note some limitations to this analysis. To start with, the coordinates of neighborhoods are not 100% accurate. They were taken from Google Geocoding API based on the names. Thus, we could only approximate the location of each neighborhood. Furthermore, I set a limit to the Foursquare API to return only the top 100 venues within the radius of 1500 meters. But the neighborhoods are very different in shapes and sizes. Some neighborhoods are much larger but less populated while others are more densely populated with a smaller area. Thus, the Foursquare API might not have been able to capture all the Indian restaurants in each neighborhood. However, we calculated the frequency of Indian restaurants within the 1500 meter radius, which could still reflect the average frequency of Indian restaurants within that neighborhood. Lastly, because we don't have specific population data on the Indian ethnicity, it is hard to tell how much of the Asian population is Indian .Thus, before opening a restaurant, it might be better to do some research on that specific neighborhood, for example, on its commercial pricing, consumers, competitors, and take other factors into consideration.

## Conclusion

At the start of this project, we defined a business problem: **Where to open an Indian restaurant in Austin, TX?**

Then we collected the neighborhood zoning information from Austin City government and used Google Geocoding API to find the approximate coordinates for those neighborhoods. We then used Foursquare API to discover the 100 venues within the radius of 1500 meters in each neighborhood and took the average frequency of Indian restaurants in comparison to other venues. Using clustering algorithm, we grouped the neighborhoods in a total of 5 clusters, with Cluster 0 having the lowest average frequency of Indian restaurants. Finally, we compared the clusters to the neighborhood demographic data provided by the government and found that the neighborhoods with the highest percentages of Asian population don't have a very high frequency of Indian restaurants.

**Then I concluded that Anderson Mill is the best neighborhood to open an Indian restaurant based on our cluster analysis and the demographic data.**

**** ****

**Anderson Mill is a wonderful Austin community with numerous local conveniences. The area is located in the northwest region of the city near Highways 620 and 183, and Toll Roads 183A and 45. Anderson Mill’s phenomenal location makes it easy for residents to travel to just about any other Austin region or surrounding area. Aside from being within extremely close proximity to a number of Austin greenbelts, recreational facilities, shops and restaurants, Anderson Mill also offers beautiful, affordable homes in quiet neighborhoods. The majority of homes in the area were built in the 1970s and 1980s, but there are still some being built today. Anderson Mill truly has something for every taste and every budget.**

Overall, this project was a great practice utilizing data science concepts and machine learning algorithms. I also believe that this project would give our target audience a starting point to consider the possibility and value of a new Indian restaurant in a given neighborhood in Austin, TX.