Finding a Suitable Location for a New Indian/Asian Restaurant within Houston or Dallas

Kheya Banerjee

December 22, 2020

# Introduction

## Background

According to reports [[1](#wilipediaindian)], Dallas, TX and Houston, TX have the 7th and 8th largest Indian immigrant population in US. The other 6 cities are:

1. New York, NY
2. Chicago, IL
3. Washington, DC
4. Los Angeles, CA
5. San Francisco, CA
6. San Jose, CA

## Problem

Although the Indian immigration population in Dallas and Houston is large, the number of Indian restaurants is few. Hence, there is always demand for new authentic Indian dining restaurants. Due to cultural and culinary proximity, we have used Asian restaurants along with Indian.

## Interest

It is extremely important to choose a suitable location for a new restaurant as this will help the business to thrive.

**The goal of this project is to determine a location for a new Indian/Asian restaurant within Dallas or Houston area.**

# Data Acquisition and Cleaning

## Data Sources

Various data sources are used in this project. They are:

1. The center latitude and longitude of Dallas, TX and Houston, TX are obtained by the help of Nominatim package which is a child package of geopy.geocoders. Although these are approximate address, it is quite accurate for our purpose.
2. To determine the boundaries of Dallas and Houston city, we have downloaded the list of all zip codes with the respective latitude and longitude values from the website [[2](#zipdownload)].
3. The restaurant location and categories are downloaded from foursquare website [[3](#foursquare)] as .json file. The total number of venues in food category are:

|  |  |
| --- | --- |
| **Dallas** | 457 |
| **Houston** | 572 |

## Data Cleaning

Some preprocessing and cleaning of the zip code data is needed as the csv file had only 1 column with all the values separated by ';'. The processing is done as follows:

1. The dataframe is converted to a list.
2. The column labels are also converted to list and split to make column labels for new dataframe
3. Every element(row) is split and separated into columns. Values are written into a new dataframe.

The restaurant information from Foursquare [[3](#foursquare)] also needed some selection as only the required fields are downloaded and stored into a dataframe. The fields saved in dataframe are:

1. Zip code
2. Zip code Latitude
3. Zip code Longitude
4. Venue name
5. Venue Latitude
6. Venue Longitude
7. Venue Category
8. Venue Category ID
9. For clustering, some preprocessing of data is needed as k-means algorithm does not accept categorical variables. The dataframe needed to be converted to numpy array.