

Data Science Capstone Project

Exploring the Cities in Texas

Introduction

Texas is one of the most diverse states in the US in term of both ethnic and economic diversity. This makes it a prospective location for restaurant business especially Thai cuisine that has captivated people all over the world with its aromatic ingredients, spicy taste and unique flavor.

The following data analysis is conducted to explore the potential of opening a Thai restaurant in Texas. The analysis will be focused on the five most populous cities in Texas which are Houston, San Antonio, Dallas, Austin, and Fort Worth. Several factors will be taken into consideration such as demographic and economic information, people's interests, current competition and other common venues in the cities.

Target audiences are potential investors who are looking to open or expand a restaurant in Texas, as well as current restaurant customers who might want to explore the cities further for other restaurants. The analysis should also be beneficial for current and future data scientists who want to use exploratory data analysis, data visualization, and machine learning to find insights from the data and further answer other questions.

Data

1. Geographical Coordinates of Each City: Geopy

The latitude and longitude of each city were obtained using Geopy and stored in a dataframe.

2. Demographic and Economic Information: Web Scraping

Statistical information such as population, race, income and housing for each city was obtained from the United States Census Bureau website

(<https://www.census.gov/quickfacts/fact/table/US/PST045219>). The data tables were read from

the website into a dataframe using pandas. The parameters of interest are %Asian, Median household income, and Median gross rent.

3. Restaurants Search and Venues Exploring: Foursquare location data

Foursquare API was used to search for restaurants to study the competition as well as explore common venues in the city to study the interests of people. The calls returned JSON files from which relevant data could be extracted. The limit for search is 50 places while the limit for explore is 100 venues. The radius of interest is within 10 km.