Where to Retire in San Diego

IBM Data Science Capstone Project

Dapeng Wang

2020/02/08

Table of Contents

Abstract

San Diego is well known for its good weather and beautiful city, which make it a top choice for retirement. There are more than 100 neighborhoods in San Diego alone, not including the surrounding cities. Which neighborhood to choose if one decides to retire in this city? It all depends on what each neighborhood can offer and each individual's preference.

In this study, data were scraped and analyzed on each neighborhood, they were clustered into different groups by what venues is available from each neighborhood. A recommendation model can give top 5 neighborhoods based on each individual's preference.

Introduction

I have lived in San Diego for many years and really enjoy the life here. Good weather, warm winter, and cool summer, beautiful coastal line, and variety of food, just to name a few. But there are over 100 neighborhoods in San Diego, I do not have the luxury to experience all of them.

Many friends, like John, often asked me:"If I want to retire in San Diego, which area should I live?" That is really a big question to answer without diving in to see what each area can offer.

Lucky, after a few hours of research, I found a complete list of neighborhoods in San Diego. Armed with newly acquired skills on web scraping, Foursquare API, and machine learning, I can quickly make some comparison on the neighborhoods and give a recommendation based on John's preference, whether he likes to eat restaurant or fast food, whether he likes to enjoy museum or parks.

This is only a preliminary study on limited data. Many other factors, like the average house price or crime rate in the area, are not considered in the study. These can be included in the future for a improved model.

Methodology

Business Problem

The objective of this study is to analyze and select the best locations in San Diego for a person to retire. Using data scraping and machine learning techniques like clustering and recommendation, this project will answer the question: If a person wants to retire in San Diego, which neighborhood can you recommend?

Data

The following data is collected in this study:

1. A complete list of neighborhoods in San Diego from Wikipedia. The link is: https://en.wikipedia.org/wiki/List_of_communities_and_neighborhoods_of_San_Diego
2. Latitude and Longitude data on each neighborhood.
3. Venue data in each neighborhood from Foursquare API

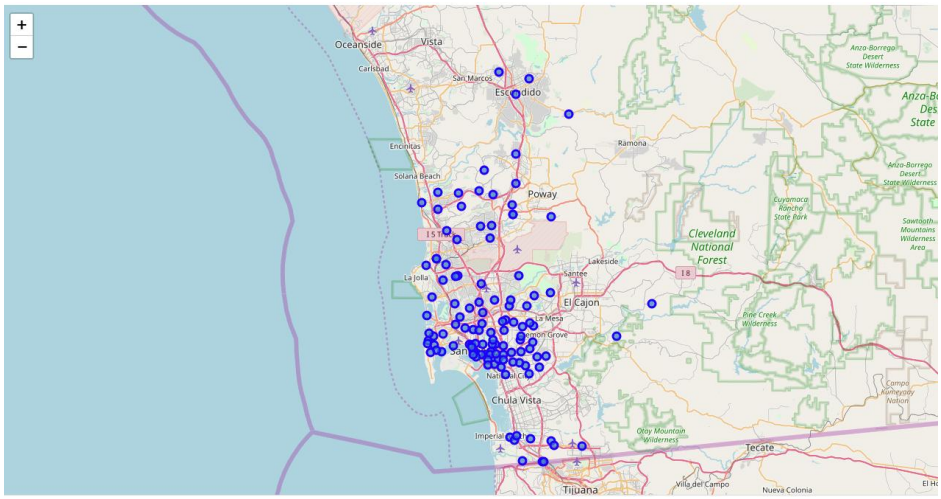
Analysis Techniques

1. I used Python requests and beautifulsoup packages to scrape the San Diego neighborhood list from Wikipedia page. The data is the name of the neighborhoods saved in a list.
2. The latitude and longitude of each neighborhood is retrieved with the geocoder package, data is saved along with the neighborhood names. All is saved in a dataframe.
3. Use Folium package to display the neighborhood overlaid on map of San Diego.
4. Use Foursquare API to retrieve up to 100 venue data for each neighborhood within 500 meters. Data is in JSON format. Venue name, latitude, longitude and category is transformed in to DataFrame.

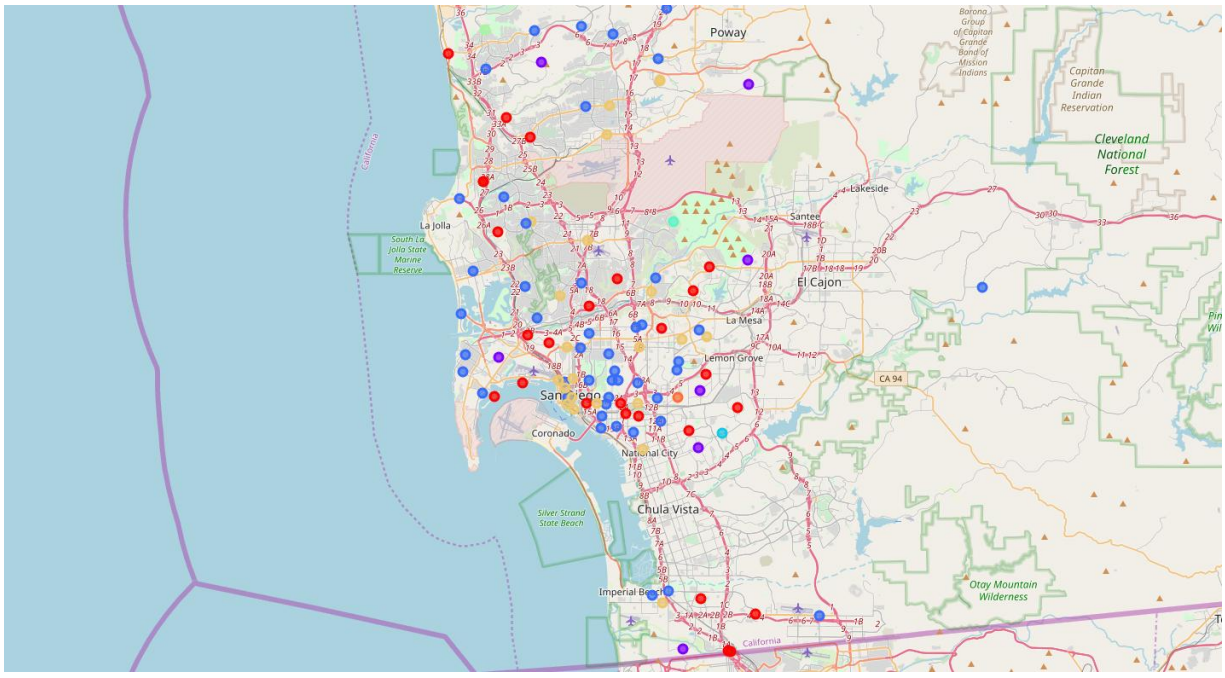
5. Kmeans clustering is used to divide the neighborhoods in 8 clusters. The data is overlaid on map with different colors
6. Based on the user preference on Restaurant, Fast Food, Transit, ParkRec, Gym, etc. the neighborhood matching score is calculated and is sorted. Top 5 neighborhoods is overlaid on the map of San Diego

Results

The neighborhoods is overlaid on map of San Diego



Kmeans clustering divides the neighborhoods in 8 clusters. Each cluster is overlaid on map of San Diego with different color.



The clusters are as following:

Cluster1: Park, Store, Food, Pet and Others

Cluster 2: ParkRec and Yoga Studio

Cluster 3: Food, Park, Doctor and others

Cluster 4: Sport

Cluster 5: IT Services

Cluster 6: Stables

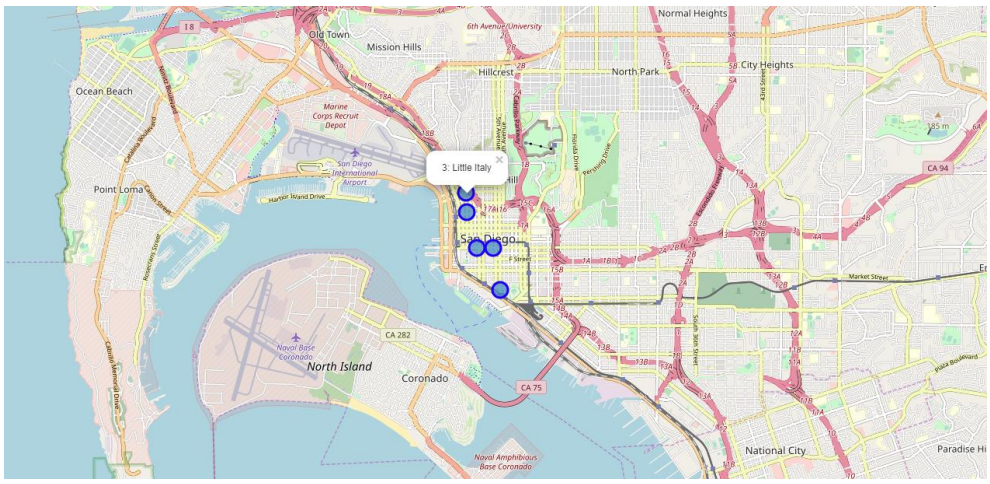
Cluster 7: Food Yoga and Landscaping

Cluster 8: Recording Studio

The model can also make recommendations. John really likes ParkRec and Theater, then likes Fast Food and Public Transit. He does not like Gas Station, and want to stay away from it. Based on His preference, I constructed the model to rank the neighborhoods and recommended these 5 neighborhoods:

1. Gaslamp Quarter
2. Midtown
3. Little Italy
4. Columbia
5. Horton Plaza

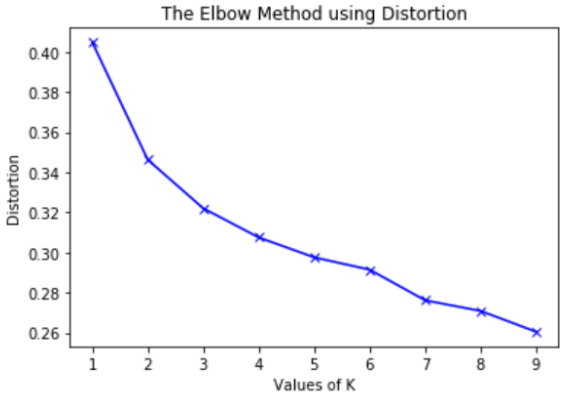
These 5 areas are overlaid on map of San Diego. They are very close to each other in the San Diego Downtown area, where there are parks, restaurant, public transit and recreation area, and very few or no Gas Station.



Discussion

A few interesting observation that needs further discussion here:

1. The combined dataset with venues has 158 features after the one-hot encoding. The distortion on the number of clusters after the Kmeans clustering does not show a very clear elbow point. The Distortion reduces gradually. The venue categories need to be examined further to combine the similar categories together so we can get more meaningful clusters.



2. The recommendation ranking method is borrows from the content based recommendation. It is a useful tool to make good recommendations based on user profile.
3. There are a lot more important factors to consider when to choose an area to live, like the house price, crime rate, etc. These are not included in the study but can be included in the future.

Conclusion

In this project, I have demonstrated defining a business problem, specifying data needed, extracting data, preprocessing data, performing machine learning model on the data and making recommendation based on the model. At last, the model recommended 5 areas for based on the user's preference. This will be a good start for John to select his retirement place.