**Retail Rental Prices & Venues Data Analysis of San Diego, CA**

Qing Lin

Jul 13, 2020

# Introduction

## **Background**

San Diego is the second largest California city and there are more than 100 neighborhoods. By 2020, the city’s population is around 1.54 million. In such a big city, the availability and price of retail space lease varies among different neighborhoods.

## Business Problem

Milk tea is enjoyed throughout the world as both a hot and cold beverage. Now an investor plans to look for a storefront used for a milk tea shop, and need to consider the surroundings, potential customers, and rental cost. So, this project will investigate the neighborhoods in San Diego city and try to find some areas that fulfill the requirements of the investor.

# Data Preparation

## Data Source

* Average Rent in San Diego, CA by Neighborhood from <https://www.rentcafe.com/average-rent-market-trends/us/ca/san-diego/>, the data was updated February 2020
* I use Nominatim Geocoding service to get the coordinates of the neighborhoods
* I use Foursquare API to get the venues categories of the given neighborhoods.

## Data Pre-processing

**Html-scraping:** Rentcafe.com website provides a table listing the average rent of each neighborhood. I make a DataFrame to store the original data.

**Data cleaning:** By using the Nominatim Geocoding service, I get the geographical coordinates with the neighborhoods’ name, but part of the names cannot get the valid return, so I remove those data. Finally, there are 97 neighborhoods for clustering.

**Venues exploring:** Foursquare API can provide the categories applied to venues with the coordinates, radius input. I take the level 0 of the category hierarchy.

# Methodology

* A screenshot of a cell phone

  Description automatically generatedData Structure: Pandas DataFrame, the two-dimensional data structure, is used for data storage. Columns can be neighborhood name, average rent, latitude, longitude. Each neighborhood is a row.
* A close up of a map

  Description automatically generatedPython Folium library is used for geospatial visualization. I create a map of San Diego with neighborhoods superimposed on top.
* Get Venue Categories. Foursquare Venue Category Hierarchy <https://developer.foursquare.com/docs/build-with-foursquare/categories/>. I only take the top level as following:
* Arts & Entertainment (4d4b7104d754a06370d81259)
* College & University (4d4b7105d754a06372d81259)
* Event (4d4b7105d754a06373d81259)
* Food (4d4b7105d754a06374d81259)
* Nightlife Spot (4d4b7105d754a06376d81259)
* Outdoors & Recreation (4d4b7105d754a06377d81259)
* Professional & Other Places (4d4b7105d754a06375d81259)
* Residence (4e67e38e036454776db1fb3a)
* Shop & Service (4d4b7105d754a06378d81259)
* Travel & Transport (4d4b7105d754a06379d81259)
* K-means clustering is an unsupervised machine learning algorithm and sklearn.cluster.KMeans is used in this project. I run k-means to cluster neighborhoods into 4 clusters.

# Result

1. A close up of text on a white background

   Description automatically generatedA screenshot of a computer

   Description automatically generatedCategorical data are converted into numeric data using sklearn.preprocessing.MinMaxScaler and Boxplot is used to show the frequency of each category of all neighborhoods
2. K-Means clustering and partition neighborhoods
   1. Visualize the frequencies of categories for each clusterA picture containing pencil

      Description automatically generated

* Cluster0 was found to have the lowest frequency for ‘Food’ category
* Cluster1 was found to have the highest frequency for ‘Food’ category
* Cluster2 was found to have the medium frequency for ‘Food’ category
* Cluster3 was found to have the low to medium frequency for ‘Food’ category
  1. A close up of a map

     Description automatically generatedVisualize the clusters on the map
  2. A screenshot of a cell phone

     Description automatically generatedVisualize the average rent of each cluster with bar chart

# Discussion and Conclusion

* Discussion
* Based on the boxplot of frequencies of categories, we can tell that Cluster1 has more venues for food already, while Cluster 2 is second cluster has high frequency of food category.
* Cluster1 also has the highest rental cost, and then Cluster2
* Conclusion

Considering the prospective customers, neighborhoods in Cluster1 are the best choice for opening a milk tea shop. But the investor also needs to evaluate the rental cost. If the cost is over his/her budget, Cluster2 is still a good option.