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

Qing Lin

Jul 07, 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 fullfill 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 most popular venues 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 93 neighborhoods for clustering.

**Venues exploring:** Foursquare API can provide the most popular venues with the coordinates, radius and LIMIT input. I keep the top 10 venues for each neighborhood.

# 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.
* A screenshot of a cell phone

  Description automatically generatedFoursquare API provides a range of tools for developers to incorporate the up-to-date location data. I set the limit as 100 venues and the radius 600 meters for each neighborhood from the given latitude and longitude information.

Totally, I got 280 unique categories.

* I use One-Hot Encoding to split the column ‘Venue Category’ to multiple columns and convert the categorical data to numerical data. The top 10 common venues are kept for each neighborhood.A screenshot of a computer

  Description automatically generated
* A close up of a device

  Description automatically generated K-means clustering is an unsupervised machine learning algorithm and sklearn.cluster.KMeans is used in this project. First, I try to find the optimal value of k with Elbow method.

A picture containing filled, group, full, computer

Description automatically generatedThen, I run k-means to cluster neighborhoods into 4 clusters.

# Result

## A close up of a map Description automatically generatedClusters on the map

## A screenshot of a cell phone Description automatically generatedAverage Rent of each cluster

A screenshot of a cell phone

Description automatically generatedThe bar chart of average rent:

## Cluster statistics

Pandas.DataFrame.describe method can generate descriptive statistics. For object data, the result’s index will include count, unique, top and freq. The top is the most common value. I found that even the ‘top’ value did not have remarkably high frequency. So I list the top 3 venues in ‘1st Most Common Venue’, ‘2nd Most Common Venue’ and ‘3rd Most Common Venue’ for each cluster.

The statistic output of each cluster:

* A screenshot of a computer

  Description automatically generatedCluster0:
* Cluster1:A screenshot of a social media post

  Description automatically generated
* Cluster2:

A screenshot of a social media post

Description automatically generated

* Cluster3:A screenshot of a computer

  Description automatically generated

The combined table of the top 3 venues for all clusters.A screenshot of a cell phone

Description automatically generated

# Discussion and Conclusion

* Discussion

The first three most common venues in Cluster 1 and Cluster 2 almost are restaurant, coffee shop or bar, which means that people like to go those neighborhoods for food and drink. But compared with other clusters, Cluster 1 has fewer common venues. So, Cluster 2 may have more visitors.

Cluster 1 has the lowest average rent, while Cluster 2 has the highest. So, Cluster 2 may be more prosperous in business.

* Conclusion
* Based on my current available data and result, I would like to suggest the investor to choose Cluster 2 as his target area.
* In my view, the clustering result does not partition the neighborhoods very well, because the most common venues show similarity to each other. I may try to add more data or use other algorithms to find a good suggestion in real life.