Capstone Research Report

"An exploration of the practicability of opening new shopping mall in San Diego, California."

Author: Ruikun Li
11 / 4 / 2019

Introduction:

San Diego, as the second largest population city in California, has long history and great potential for future development in bio-medical and chips industry. Nowadays, with new immigrations and domestic people move in, old downtown, established in 1860s and developed in 50s, has lack of strength of capability to provide civilians convenient facility, particularly for shopping mall. However, there are lot of shopping mall are built surround downtown area. For great population at San Diego downtown, visiting shopping mall is great way to relax and enjoy themselves during weekends and holidays. They can do grocery shopping, dine at restaurants, shop at the various fashion outlets, watch movies and perform many more activities. Shopping malls are like a one-stop destination for all types of shoppers. For retailers, the central location and the large crowd at the shopping malls provides a great distribution channel to market their products and services. Property developers are also taking advantage of this trend to build more shopping malls to cater to the demand. As a result, whether build a new shopping mall at San Diego downtown area become new problem. Opening shopping malls allows property providing service for local civilians and generating more tax income for local government, but there are lot of giant shopping mall surround downtown area. Of course, as with any business decision, opening a new shopping mall requires serious consideration and is a lot more complicated than it seems. Particularly, the location of the shopping mall is one of the most important decisions that will determine whether the mall will be a success or a failure.

Business Problem:

My primary objective of this capstone is to analyse and select the best location in the city of San Diego for building new giant shopping mall. Applying data science and machine learning technologies to make decision of problem: "Whether a new shopping property developer should into San Diego Downtown Area."

Target Audience of this Project:

This project is primarily niche to property developers and investors looking to open or invest in new shopping malls in the San Diego, California. This project is timely as the shopping service area far from enough for well serving all cities civilians of San Diego.

Data Collection:

To solve this target problem, we need to collect those data shown above:

- List of cities in San Diego. This defines the scope of this project, which is confined to the Great San Diego, CA.
- Latitude and longitude coordinates of those cities. This is required in order to plot the map and also to get the venue data.
- Venue data, particularly data related to shopping malls. We will use this data to perform clustering on the cities.

Data Sources:

For dictionary purpose, we going to focus on the Wikipedia category page (https://en.wikipedia.org/wiki/Category:Cities_in_San_Diego_County,_California) to search for all cities contained by San Diego and try to extract all cities list from that page. To achieve this step, we are going to deploying web scraping tool to make extracting successfully, by helping with Python requests package and BeautifulSoup library. Then we will get geographical information such as coordinates of San Diego’ s cities by Python Geocode library tools, it will provide latitude and longitude coordinates information of the cities.

Meanwhile, we are going to apply Foursquare API to get the venues data for those cities. Foursquare has one of the largest databases of 105+ million places and is used by over 125,000 developers. Foursquare API will provide many categories of the venue data, we are particularly interested in the Shopping Mall category in order to help us to solve the business problem put forward. This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium). In the next section, we will present the methodology and the data analysis research sections.

Methodology:

Firstly, we need to get the list of cities in the county of San Diego. Fortunately, we can easily browse those names from Wikipedia page of Cities of San Diego. In order to get names of those cities, we applied web scraping tools which are Python requests and BeautifulSoup libraries to extract the list of cities names data. For next step, request geographical coordinates in such form of latitude and longitude to be able to next use for Foursquare API inquiry. For that purpose, there is a coordinator inquiring tool, the Geocoder package, that will allow us to convert address into geographical coordinates in the form of latitude and longitude, and then we populated the data into a pandas DataFrame and then located all cities address in a map of San Diego by Folium package. This allows us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the great San Diego County. Next, we deployed

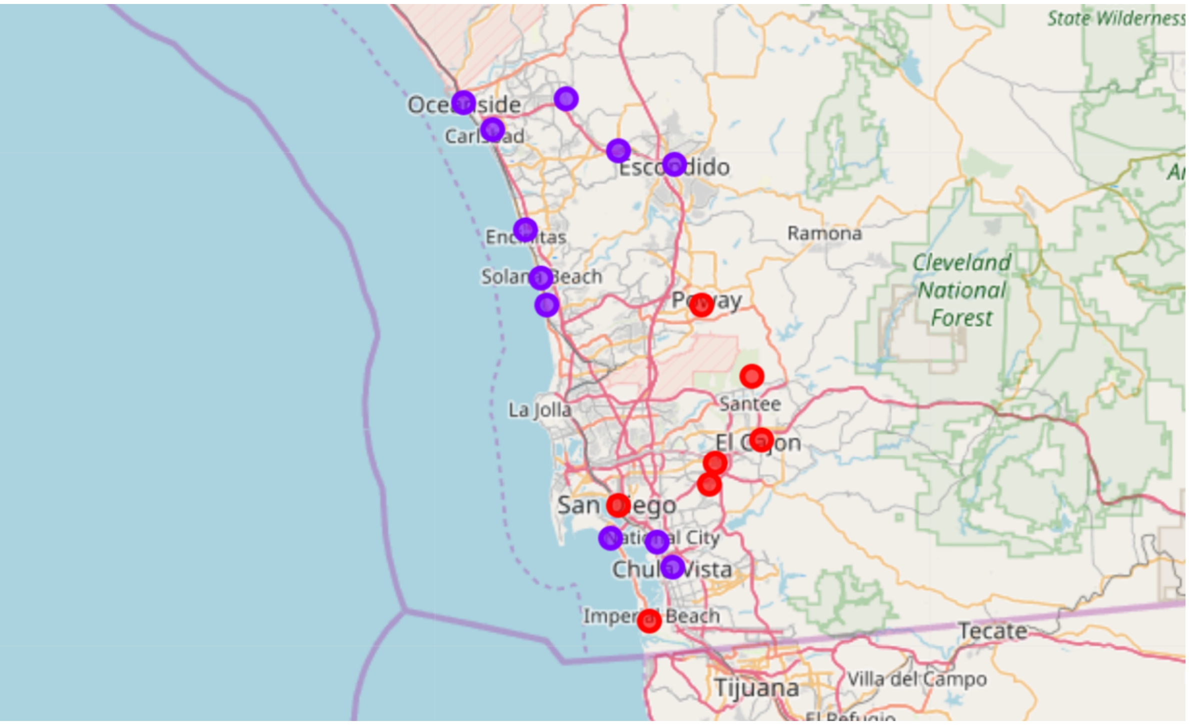
Foursquare API to get the top 100 venues within a radius of 20 kilometers of each city. We then made API calls for Foursquare passing us the geographical coordinates of the cities in a Python loop. The returned data of FourSqaure API calls were in JSON format, and we then extracted useful information under format by the venue name, venue category, venue latitude and longitude. To deal with data, we applied count function to check out how many venues were returned for each city and how many unique categories, and we analyzed each city by grouped the rows by cities and took the mean of the frequency of each venue category. For next step, we prepared the data for clustering. Since we researched the “Shopping Mall” data, we filtered out the “Shopping Mall” as venue category for each city. Lastly and performed clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. Then clustered the cites into 2 clusters based on their frequency of occurrence for “Shopping Mall”. Based on the occurrence of shopping malls in various cities, we could conclude those cities are over or under service of shopping mall.

Results:

After clustering process, the cluster indicated us that those cities have been classified as two types, based on those cities’ numbers of shopping malls. Its easily to find those only seven cities have shopping mall and can be diversify as two types:

- Cluster 0 indicate that city has one or more shopping mall.
- Cluster 1 indicate that city has none shopping mall.

To visualize this result, I plot the map above.



Discussion:

It’s clear to see that from map, there are total seven shopping mall that are searched by this research. Because of this reason, it is easy to classify types of cities of San Diego county and by whether there are shopping in those cities. For this reason, the cluster 0 (indicating city that has shopping mall) and verse visa. Investigating by map, we found the most of shopping mall are located in south of San Diego, and there is no shopping mall located at north. It clearly has an opportunity to open a new shopping mall with high success probability.

Limitations and Suggestions for Future Research:

As state in discussion section, because of reason of low numbers of shopping mall in San Diego County, it is easy to classify those cities in two types by whether there are or not shopping mall, and normally there is no city has two or more shopping mall. So, there will lead a misleading for developers, for lacking shopping mall for some cities and there is opportunity, but there are still some shopping plaza or outlet mall may play role alternate with shopping mall for those cities. Purely consider may lead a misunderstanding for business decision of developers. Meanwhile, for some reason, we only consider the San Diego County and there is nothing more about neighbor counties, such as Santa Ana or San Clement. As we research, the civilians of north part of San Diego County become more richer than south part, they may choose to drive to other county for example, Santa Ana, for more Luxury option such as South Coast Shopping Center. To simply conclude if there is opportunity of shopping mall opening will become a misleading.

Conclusion:

In this project, we gone through the process of identifying the business problem, collecting and transform data for well using, applying machine learning by clustering the data into two clusters base on those shopping malls they have, and conclude and show some limitation of this research. Based on our research, we found out there are lack of shopping mall serving power in north county of San Diego, but we need to consider more factor or elements to figure out whether there should build a new shopping mall.