Research Proposal: Capstone

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

Author: Ruikun Li

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.

Research Question:

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."

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.