I write introduction and Data section in one file. Please scroll down to see all.

Analysis of opening a new shopping centre in Sydney

Introduction

A shopping centre is a multipurpose place that contributes significantly in liveability of a region. A variety of stores, restaurants, entertaining places, cinemas and sport complex are examples of places in a shopping centre that motivates people go shopping. In addition to shopping, entertaining activities are of prime importance in these places. Therefore, building a new shopping mall in a region could have a significant impact on people lives in the region.

Probably the main question for market study would be the place that shopping mall is aimed to be built. One of the major factors is the number of neighbouring shopping centres in the area and also in the city.

Current study reveals specific areas in Sydney which are potentially suitable for opening a new shopping centre. The study uses data analysis techniques together with clustering machine learning to suggest new places for this purpose.

Data Source

Data of Sydney suburbs is web scraped from Wikipedia. The link to data of Sydney suburbs is https://en.wikipedia.org/wiki/Category:Suburbs_of_Sydney. The Geographical data for location of each suburb is obtained from Geopy and Geocoder Python libraries. Data for venue of shopping centres in Sydney is obtained from Fouresquare.

Methodology

Data acquisition and cleaning

List of neighbourhoods

Data downloaded or scraped from multiple sources and combined in a table for further processing. First, data for Sydney suburbs were scraped from Wikipedia

web page (https://en.wikipedia.org/wiki/Category:Suburbs_of_Sydney). There were 200 neighbourhood in this web page. However, there was not a table for data in this web page. So, initially list of neighbourhoods were scraped from this page by using Beautifulsoup package. After scraping the list of suburbs three steps were followed to create data frame. I) Creating a list to store neighbourhood data, II) appending the data into the list and III) create a new data frame from the list.

Geographical data for neighbours

The Wikipedia web page does not have geographical data for locations in neighbourhoods. Therefore, the next step was acquiring longitude and latitude of each location using Geocoder Python package. This includes following steps: defining a function to get coordinates, calling the function to get the coordinates, store in a new list using list comprehension, creating temporary data frame to populate the coordinates into Latitude and Longitude and finally merging the coordinates into the original dataframe.

A map from dataframe was created using Folium package and is shown in Figure 1.

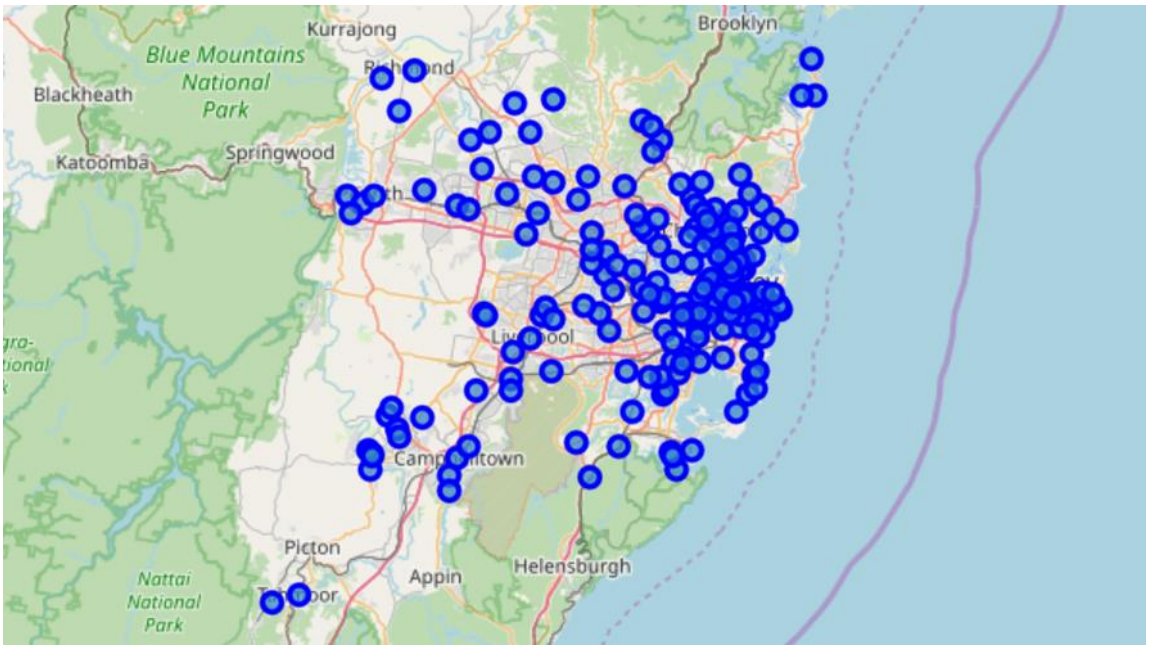

Figure 1. A map of 200 neighbourhood locations in Sydney created by Folium package.

Venue data for neighbours

Foursquare is a popular web service that provides comprehensive data for more than 150 million locations around the globe. The source for venue data was chosen to be Foursquare. A developer app was created in Foursquare web site and through its API, venue data for each neighbourhood was stored in a new data frame which was later merged to original data frame. The first five rows of the resulting data frame is shown in Table 1.

Table 1. List of neighbourhoods in Sydney with geographical and venue data. First column was scraped from Wikipedia, 2nd and 3rd columns were got from Geocoder and last four columns were obtained from Foursquare.

	Neighborhood	Latitude	Longitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
0	Agnes Banks, New South Wales	-33.61445	150.71083	Wog Mobile	-33.619594	150.706412	Rental Car Location
1	Agnes Banks, New South Wales	-33.61445	150.71083	Yarramundi Reserve	-33.613377	150.698378	Nature Preserve
2	Agnes Banks, New South Wales	-33.61445	150.71083	D & V Turf Supplies Pty Ltd	-33.623196	150.702574	Other Repair Shop
3	Agnes Banks, New South Wales	-33.61445	150.71083	Navua Reserve	-33.608786	150.696020	Park
4	Agnes Banks, New South Wales	-33.61445	150.71083	Trees Adventure	-33.612809	150.692359	Rock Climbing Spot