SAN DIEGO HIGH SCHOOL

Capstone project

Abstract

This document is a report of applying data science techniques to define and solve a problem

Yadavalli, Sunil

February 2020

Table of contents

Introduction

In today’s world people keep switching between jobs for various reasons like career prospects, subject of interests, personal agenda etc. Thanks to the various channels of opportunities, we are able to find a good job that addresses our needs. However, this decision is governed by a lot of factors that takes priority and enable us to make a decision. One such factor, for people having school going kids, is to learn about the good schools in the neighborhood of the perspective job location. This is very important because we would want the kids get the best education and there is no compromise with their career.

Business Problem

Assume that someone got a job opportunity in the beautiful city of San Diego, California. The question is, where you would recommend the locality to stay so that there are enough high schools around with a minimal commute. Also, in the available options, which school would be the best option to opt for if there is a chance of admission? Can machine learning techniques like clustering, address this problem? Let’s see.

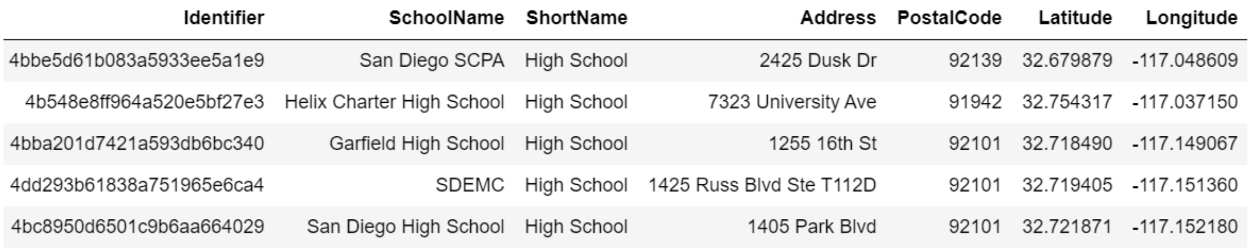
Data

To address the above problem, we would need the below:

- List of the neighborhoods in San Diego
- The latitude and longitude of those neighborhoods
- Venue data, particularly the high schools. This data will be used to perform clustering

We will leverage the Foursquare API to collect the data. We then apply the machine learning techniques on this data and also utilize map visualization (Folium with heat map)

Below is the sample data



Identifier	SchoolName	ShortName	Address	PostalCode	Latitude	Longitude
4bbe5d61b083a5933ee5a1e9	San Diego SCPA	High School	2425 Dusk Dr	92139	32.679879	-117.048609
4b548e8ff964a520e5bf27e3	Helix Charter High School	High School	7323 University Ave	91942	32.754317	-117.037150
4bba201d7421a593db6bc340	Garfield High School	High School	1255 16th St	92101	32.718490	-117.149067
4dd293b61838a751965e6ca4	SDEMC	High School	1425 Russ Blvd Ste T112D	92101	32.719405	-117.151360
4bc8950d6501c9b6aa664029	San Diego High School	High School	1405 Park Blvd	92101	32.721871	-117.152180

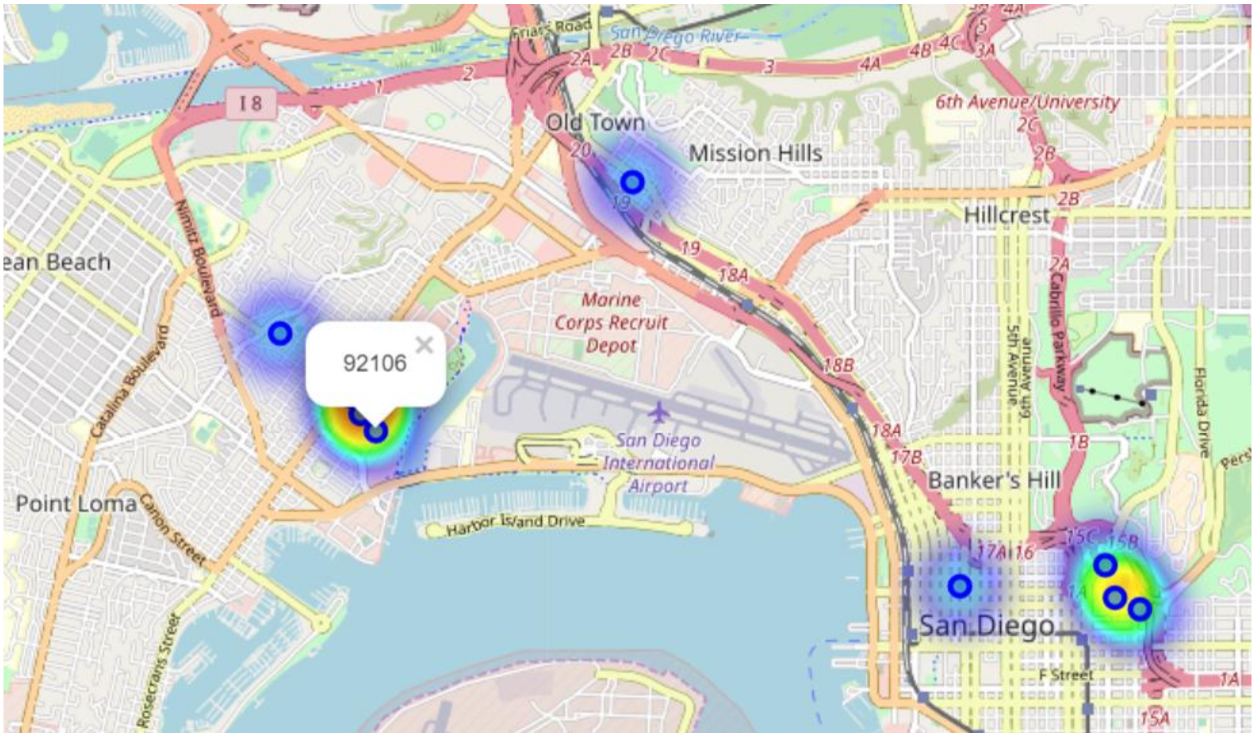
In the further sections we will talk about the methodology of using this data and apply data analysis.

Methodology

Firstly, we need to get the list of neighborhoods in the city of San Diego. Fortunately, the list can be populated using the Foursquare API. We use the explore API of the group Venue. We use the CategoryID of the High School (which can found on the Foursquare API documentation) and pass it as a parameter to the call to get the high school information. After gathering the data, we will populate the data into a pandas DataFrame and then visualize the neighborhoods in a map using Folium package.

Also, we will use Foursquare API to get the top 50 venues that are within a radius of 10000 meters. Essentially we are looking at a maximum radius of 6 miles. Foursquare will return the venue data in JSON format. With the data, we can check how many venues were returned for each neighborhood and examine how many high schools can be curated from all the returned venues.

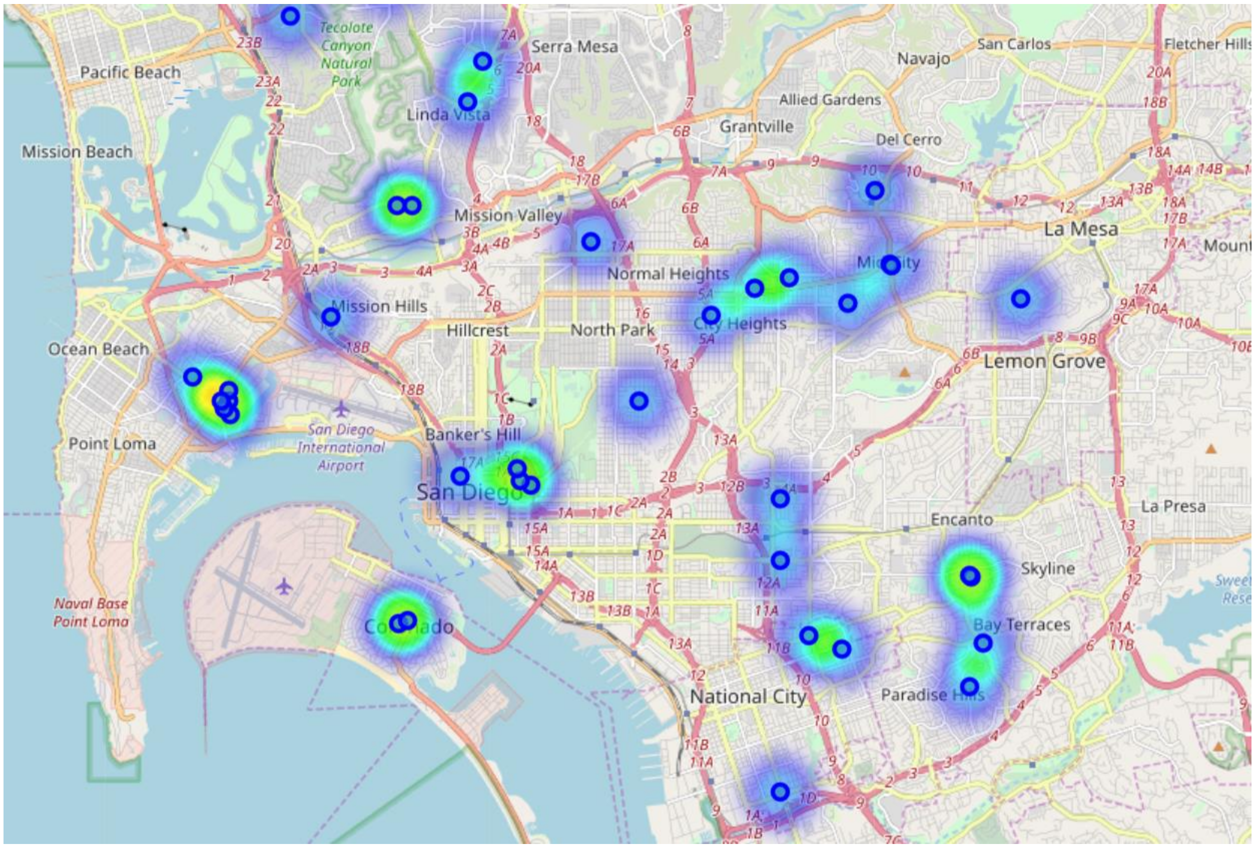
Lastly, Heatmap-based kernel density estimation was used. Heatmap is already implemented as plugin for Folium, which we use to visualize data to map. We label the points in the map with the Postal code so that it will be easier for us to know the majorly clustered neighborhood. From the below figure we can see that the neighborhood with the postal code 92106 has the most number of school in the locality.



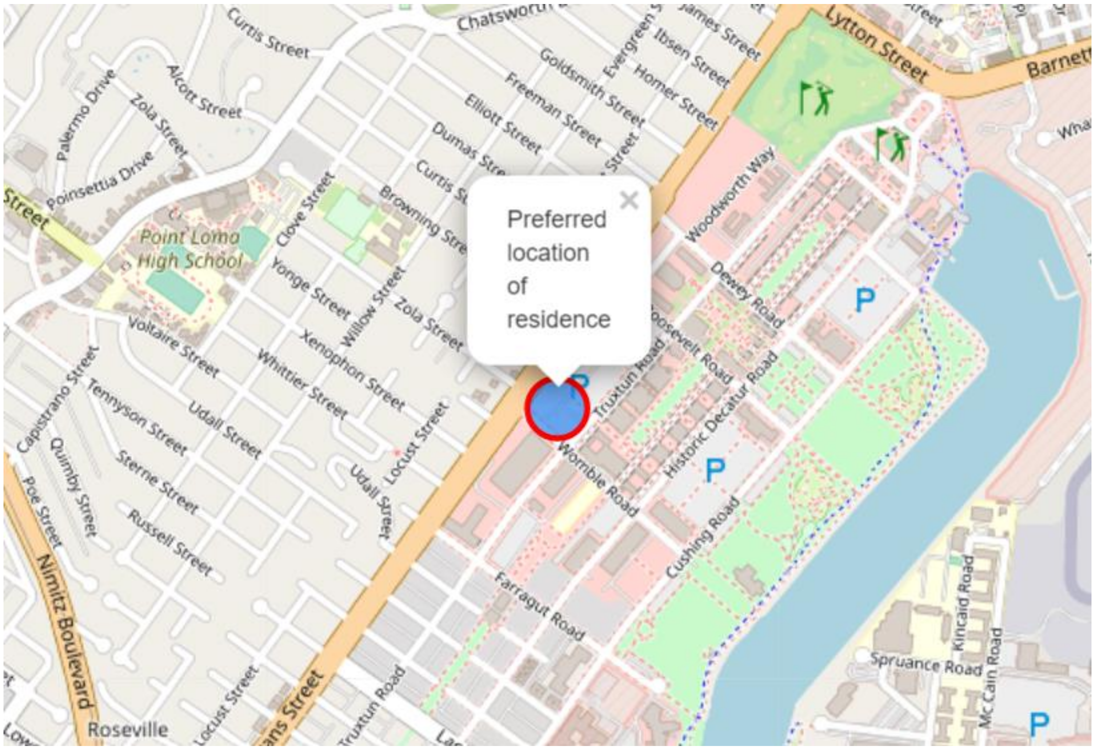
We filter our dataframe with this postal code and randomly pick one of the latitude/longitude to propose the area of residence we can look for so that we are in the vicinity of most of the schools.

Results

Heatmap with the high schools visualization is shown in the below figure.



Based on the randomly picked latitude and longitude the proposed area of residence to look for is around the Womble Road



Discussion

We could further update our dataframe based on the ratings for each of the high school. Also, additional tips on them would help us further make a better choice. However, due to the limitation of the regular API membership we could not do that. Ratings/ Tips would need a Premium membership. This could be the potential area of interest when we deal with a real-world problem.

Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning and lastly providing recommendations to the relevant stakeholders i.e. job seekers who would find a good locality to pick for residence that has high schools around them with minimal commute. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighborhoods having the postal code 92106 are the most preferred locations to find a residence. The findings of this project will help the relevant stakeholders to capitalize on the decision or choice to be made.