Evaluation of Cities for a Tutoring Business

Applied Data Science Capstone by IBM, Course 9

Sudha Kolathu Parambil

June 7, 2019

1.0 Introduction

1.1 Overview of the problem

Everyone wants to get a good grade in their school and get into an awesome University to eventually get their dream job. Some students prefer to have extra coaching out of the classroom to achieve their ideal GPA. There are several tutoring centers available close to schools to help such students to achieve their goals. As a Mathematics Professor, I am interested in analyzing the accessibility of a tutoring center between two different cities to help me to decide where I should start a Tutoring center to help those who are need extra coaching.

1.2 Discussion of the background

I first needed to decide if I wanted to be based on the East or West of the country. I narrowed down my choices to two states, California and Ohio. According to Forbes in 2016, Ohio was listed as one of the top states in the nation to raise a family. Therefore, I would imagine that many parents would want to provide extra tutoring for their children if given the opportunity. In California, there is an abundance of high-ranking universities. The academic expectations of these universities would mean that there is a chance of there being students interested in extra help outside of the classroom.

Personally, I made the move from Ohio to California a few years ago. Based on my experience with these two cities, I narrowed my study to focus on Roseville, California and Columbus, Ohio.

1.3 Who will be Interested

I believe that this data analysis will be useful for franchise tutoring businesses, ACT-SAT coaching centers, and after-school tutoring centers who would need help deciding where to open their businesses. In addition, schools may be interested in using this data analysis to help determine the need for a tutoring center located within school.

2.0 Description of the data and How to solve the problem

2.1 Source of Data and cleansing data

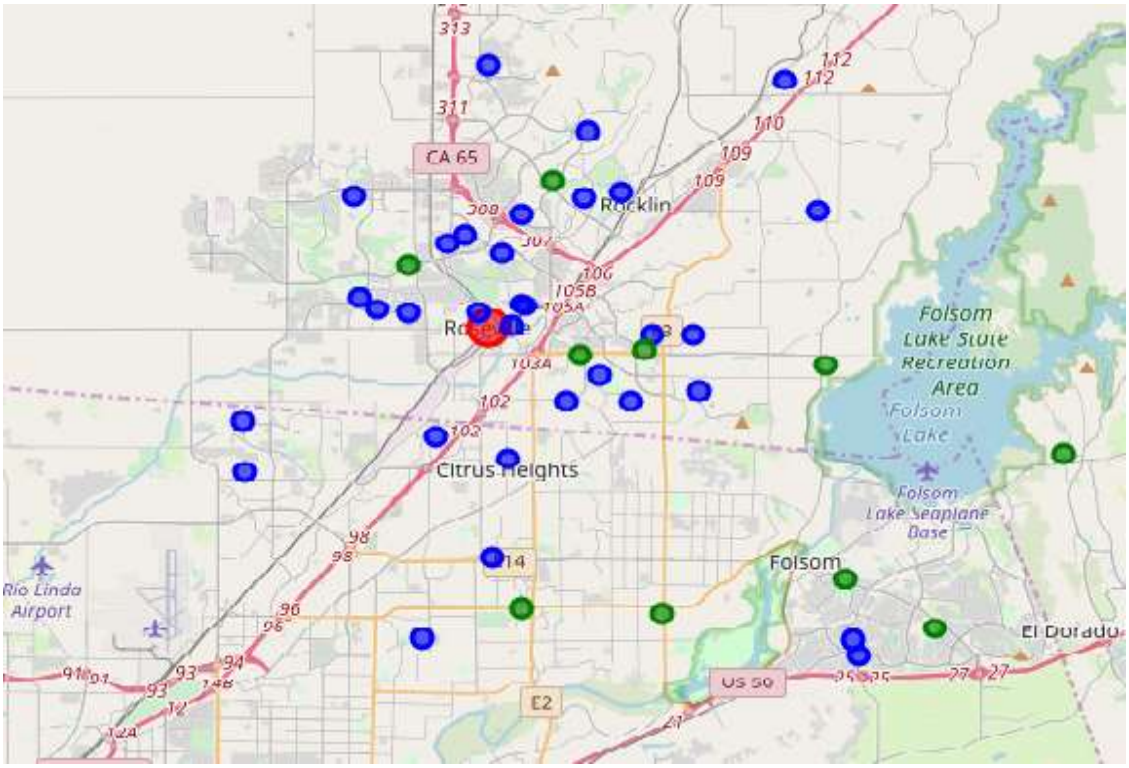
I constructed an URL to send a request to the **Foursquare API** to search for schools in the city of Roseville in California in a radius of 15000 meters around Roseville city I need to be analyzed using the longitude and latitude which I found using the geocoder. I converted the result of this in **JSON file** into a **Pandas dataframe**. While extracting data with the key word "school" it extracted even the play school , Music schools , Law schools etc. I am focusing on Elementary, middle and high schools only. So, Kept only those type of schools and drop the rest of the schools. Kept only the colums which I am intered in which are name of the school, category of the school (Elementary, middle and high schools), address, distance, longitude and latitude. I filtered the **columns** as name, categories, address, distance, lng and lat. I filtered and kept only the **rows** which are Elementary, Middle and High school. Then deleted the rows which has null cells and renamed the column heading 'name' to 'RosevilleSchools' and 'categories' to 'Type'.

	RosevilleSchools	Type	address	distance	lng	lat
0	Vencil Brown Elementary School	Elementary School	250 Trestle Rd	3328	-121.296026	38.781362
1	Oakmont High School	High School	1710 Cirby Way	3511	-121.260496	38.729000
2	Woodbridge Elementary School	Elementary School	515 Niles Ave	594	-121.291175	38.756861
3	Roseville High School	High School	1 Tiger Way	1296	-121.276113	38.759173
4	Heritage Oak Elementary School	Elementary School	2250-2254 Americana Dr	2442	-121.315551	38.756582

I did the above procedure to explore the tutoring centers in Roseville. While extracting the tutoring centers in it extracted the touring centers for yoga and colleges etc. So, I filtered only the tutoring centers related to the school. Also, I kept only the columns which I am interred in which are the name of the tutoring center, category (here is it is school), address, distance, longitude and latitude. Deleted the rows with null vales. I filtered the **columns** as name, categories, address, distance, lng and lat. I cleansed the data by deleting the rows which are not related to school. Also, deleted the rows which has null values. Then deleted the rows which has null cells and renamed the column heading 'name' to 'RosevileTutor' and 'categories' to 'Type'.

	RosevilleTutor	Type	address	distance	lng	lat
0	Kumon Math and Reading Center of Roseville	School	1271 Pleasant Grove Blvd #130	3238	-121.315506	38.771790
1	Kumon Math and Reading Center of Roseville - East	School	1850 Douglas Boulevard, Suite #908	2980	-121.255548	38.743404
2	Kumon Math and Reading Center of Rocklin	School	2331 Sunset Boulevard, Suite #230	5509	-121.265339	38.798358
3	Kingsfield Tutors	School	3017 Douglas Blvd Ste 300	4862	-121.232768	38.744904
4	Mathnasium of Fair Oaks	School	7840 Madison Ave	9896	-121.275734	38.663738

I used **Folium** to visualize the schools and tutoring centers in Roseville city in 15000 miles radius. I mapped both the schools and the tutoring center in Roseville together. I used the **red dots** color for the **city**, **blue dots** for the **schools** and the **green dots** for **tutoring centers**.



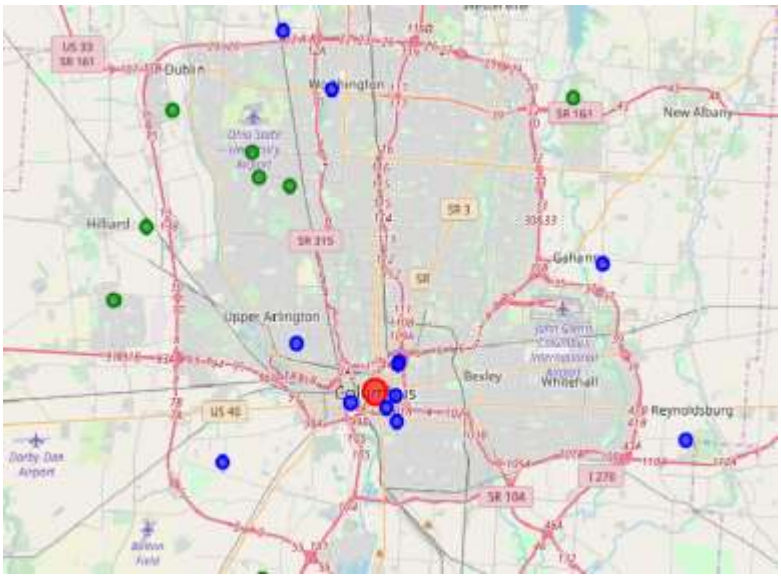
I did the same procedure for the Columbus city in Ohio. I created a dataframe for Schools in Columbus.

	ColumbusSchools	Type	address	distance	lng	lat
0	Columbus Downtown High School	High School	364 S 4th St	931	-82.994451	39.955401
1	Fort Hayes School	High School	546 Jack Gibbs Blvd	1840	-82.986692	39.974825
2	Grove City High School	High School	4665 Hoover Rd	11893	-83.072187	39.870556
3	Stewart Elementary School	Elementary School	670 Briggs St	1772	-82.988019	39.949653
4	Thomas Worthington High School	High School	300 W Dublin Granville Rd	14458	-83.026280	40.090659

I created a dataframe for Tutoring centers in Columbus.

	ColumbusTutor	Type	address	distance	lng	lat
0	Kumon Math and Reading Center of Upper Arlingt...	School	1214 Kenny Center Road	10643	-83.051741	40.049513
1	Mathnasium of Upper Arlington	School	4713 Reed Rd	11703	-83.069581	40.053202
2	Kumon Math and Reading Center of Grove City	School	2436 Stringtown Road	10543	-83.067743	39.882712
3	Mathnasium of Hilliard	School	2479 Hilliard Rome Rd	13954	-83.156059	40.001540
4	University Tutor - Columbus	School	5265 Captains Ct	12896	-83.073896	40.063646

Using Folium I mapped both the schools and the tutoring center in Columbus together. I used the **red dots** color for the **city**, **blue dots** for the **schools** and the **green dots** for **tutoring centers**.



2.2 Solving the problem

- First, I want to analyze the opportunity of a Tutoring Center business in each area by locating nearby schools in each city. Next, I would need to consider the likelihood of competing with other tutors in the area.
- Find the total number of elementary, middle, and high schools present and the number of Tutoring centers in each city.
- Then compare the total number of tutoring centers to the total number of schools in each city and derive a ratio that can be compared. The percentage obtained from the ratio can be analyzed as such that a smaller percentage ratio would mean it would be more beneficial to open my business in that city.

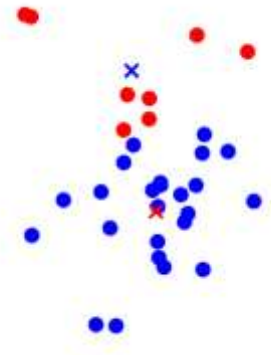
- This data can be analyzed using a pie bar graph and a bar graph.
- After making the decision on which city I am planning to open the Tutoring center. I need to find the approximate location of the prospective Tutoring center. I am going to locate that by finding a point on the map which is accessible from most of the schools.

- After making the decision I find out the schools near the new location using foursquare API. And converted the result of the JSON file to the datafarme. I cleansed the data and sorted the data by distance from the new location.

	name	categories	address	distance	lng	lat
0	Adelante High School	High School	350 Atlantic St	335	-121.279724	38.752476
1	Roseville High School	High School	1 Tiger Way	473	-121.276113	38.759173
2	Independence High School	High School	125 Berry St	517	-121.274504	38.758930
3	Woodbridge Elementary School	Elementary School	515 Niles Ave	1154	-121.291175	38.756861
4	Catheryn Gates Elementary School	Elementary School	1051 Trehowell Dr	2312	-121.282844	38.775627

3.0 Methodology

Used **K-mean** to find **two centroids** of the schools in the city to Identify two possible locations in the city. My aim is to find a new location for the Tutoring center where the schools are about average distance from the new location. So, I used K-mean because this will provide me two clusters of schools and two centroids where the distance from the school to the tutoring center has average distance. A cluster that has a smaller average distance is more compact than the cluster has a larger average distance. So, I mapped the schools and tutoring center and centroid to make decision based up on the cluster which is more compact. That means there will be more schools closer to that tutoring center so more students can access that facility.



Centroid 1: [38.73507468395327, -121.19612578361108]
Centroid 2: [38.75477366246383, -121.29437221487791