Analyzing better choices of opening stores in Bay area based on companies ‘locations

Cheng Han

May 31,2020

1. Introduction

1.1 Background

The **Bay Area** (more fully, the **San Francisco Bay Area**), ringing the San Francisco Bay in northern California, is a geographically diverse and extensive metropolitan region that is home to over 7 million inhabitants in cities such as San Francisco, Oakland, and San Jose.

With the influence of COVID-19, American stores have been severely hit, and some stores have not been open for a long time. When the epidemic situation gets better, the store will reopen and the business situation will improve. Therefore, it is a necessary question to determine which type of store is suitable before this.

The Bay Area is also known as its technology companies, with large numbers companies and staff, so paying attention to the stores around the companies can help us make a reasonable guess about the address and type of the store.

Napa County
Sonoma County
Napa
Solano County
North Bay
Fairfield
Novato
Vallejo
Marin County
San Rafael
Concord
Richmond
Contra Costa County
Berkeley
East Bay
Oakland
San Francisco
San Francisco County
Oakland International Airport
San Francisco International Airport
Hayward
Alameda County
Fremont
San Mateo
Palo Alto
San Jose International Airport
San Jose
San Mateo County
Peninsula
South Bay
Santa Clara County
Gilroy
10 Kilometers
10 Miles
N
101
80
680
580
280
880

1.2 Problem

The main content of the problem is to find the best choice by analyzing the stores near the Bay Area company. The factors that influence the choice mainly include the number and type of stores near the company. If there are too many stores of the same type, it will be an unwise decision to open a new store of the same type in this area. Therefore, we can use machine learning methods such as clustering to solve this problem.

1.3 Interest

This project is useful for everyone who wants to open a store in the Bay Area. In this project we considered almost all kinds of stores, including but not limited to : electronics store, Arts & crafts store, Furniture store etc. And all the recommendations are equally applicable to those who want to transform the store later.

2. Data acquisition and cleaning

2.1 Data requirements

---List of tech companies (name, category...)

#This is the name and type of the company we want to investigate; it helps us to initially define the scope of the research.

---The coordinates of the companies

#This is required in order to plot the map.

---List of all stores (from Foursquare API)

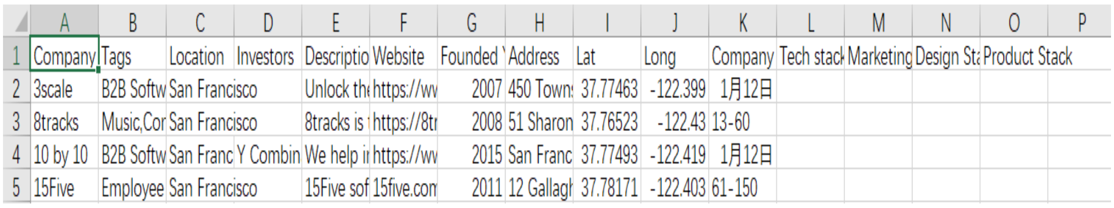
#We will use this data to perform clustering and analyzing.

2.2 Data sources

The basic source is the Bay Area companies list which is an open source by Mr.Connor Leech(https://github.com/connor11528/tech-companies-bay-area).

The csv file has been uploaded on IBM cloud and will be used to analyze.

The head of the list is shown below:

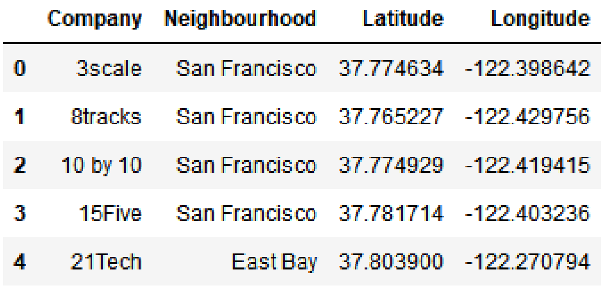


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Company	Tags	Location	Investors	Descriptio	Website	Founded `	Address	Lat	Long	Company	Tech stack	Marketing	Design Sta	Product Stack	
2	3scale	B2B Softw	San Francisco		Unlock th	https://wv	2007	450 Town:	37.77463	-122.399	1月12日					
3	8tracks	Music,Cor	San Francisco		8tracks is	https://8tı	2008	51 Sharon	37.76523	-122.43	13-60					
4	10 by 10	B2B Softw	San Franc	Y Combin	We help ir	https://wv	2015	San Franc	37.77493	-122.419	1月12日					
5	15Five	Employee	San Francisco		15Five sof	15five.con	2011	12 Gallagh	37.78171	-122.403	61-150					

2.3 Data cleaning and feature selection

Firstly, we clean the list of tech companies to get part of the data. There are some redundancies so we only keep the columns that we need (Company name, Location, Lat, Long) Then we use them to explore the venues near the offices within radius of 1000 through Foursquare API. Next extract all stores from the venues and finally perform KMeans clustering.

The head of data after data cleaning is shown below:



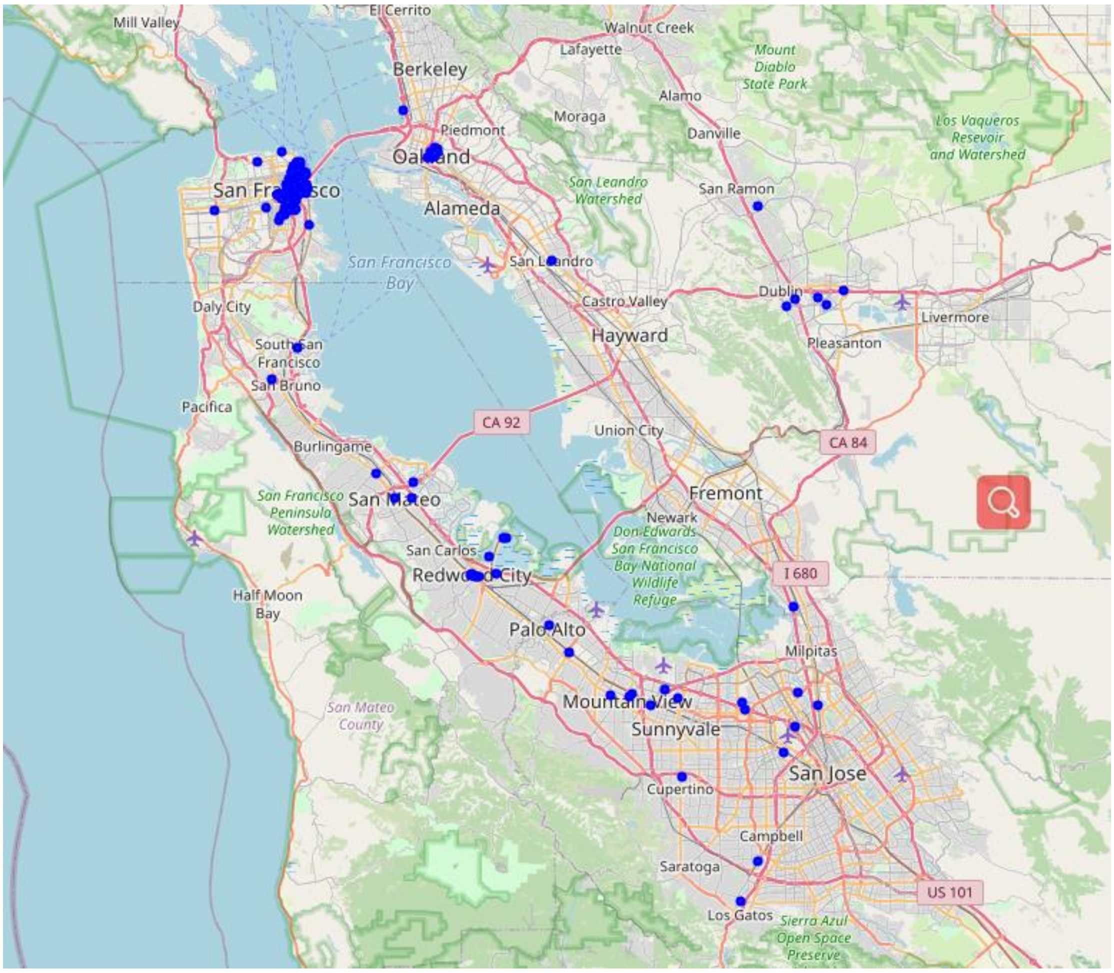
	Company	Neighbourhood	Latitude	Longitude
0	3scale	San Francisco	37.774634	-122.398642
1	8tracks	San Francisco	37.765227	-122.429756
2	10 by 10	San Francisco	37.774929	-122.419415
3	15Five	San Francisco	37.781714	-122.403236
4	21Tech	East Bay	37.803900	-122.270794

3.Methdology

I： Firstly，we need to get the list of Bay Area companies , as we have mentioned. It can de found by Mr.Connor Leech(https://github.com/connor11528/tech-companies-bay-area). We do not need to scraping any data since the list has contained all information we want. In order to use the data, we import it to my IBM CLOUD.

II: After gathering the data, we transform the data into a Dataframe using pandas. The columns' names of the dataframe is not standard, so we rename them and find there are 756 rows. For our project, the number of companies is too large (756), so we only consider the top 200 companies. It will hardly affect our conclusion since 200 companies are enough for us to draw some obvious conclusions.

III: Plot the map for all the tech companies in the Bay Area:

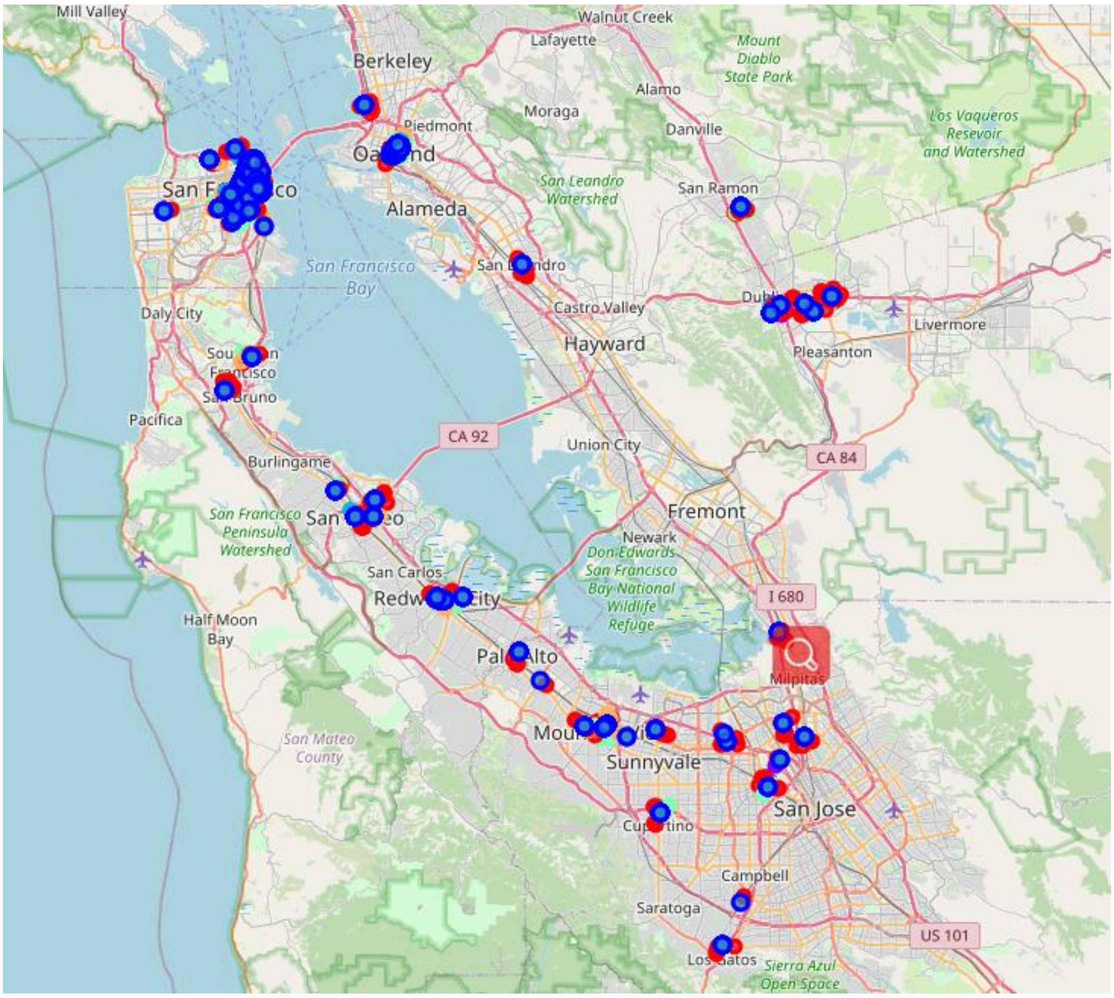


Iv: To do the analysis, we must use Foursquare API. I already have an account and just input the client ID and client Secret. Then use Foursquare API to get the top 200 venues that are within a radius of 1000 meters. Some details are making API calls to Foursquare passing in the geographical coordinates of the offices in a Python loop. Finally, the procedure will return the venue data in JSON format and we will extract Neighborhood, Company, Latitude, Longitude, Venue name, VenueLatitude, VenueLongitude, Venuecategory. Next, we calculate that there are 18459 venues and 365 unique categories.

V: We perform onehot coding and generating a copy of dataframe. Merge the dummy variables with the original dataframe and drop the useless column.

VI: Using KMeans Machine Learning Algorithm for the clustering of the stores near the bay area companies. *k*-means clustering is a method of vector quantization, originally from signal processing, that aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells.We will cluster the stores into 5 clusters based on their frequency of occurrence.

VII: The plot of all the classified stores and the tech companies is shown below:

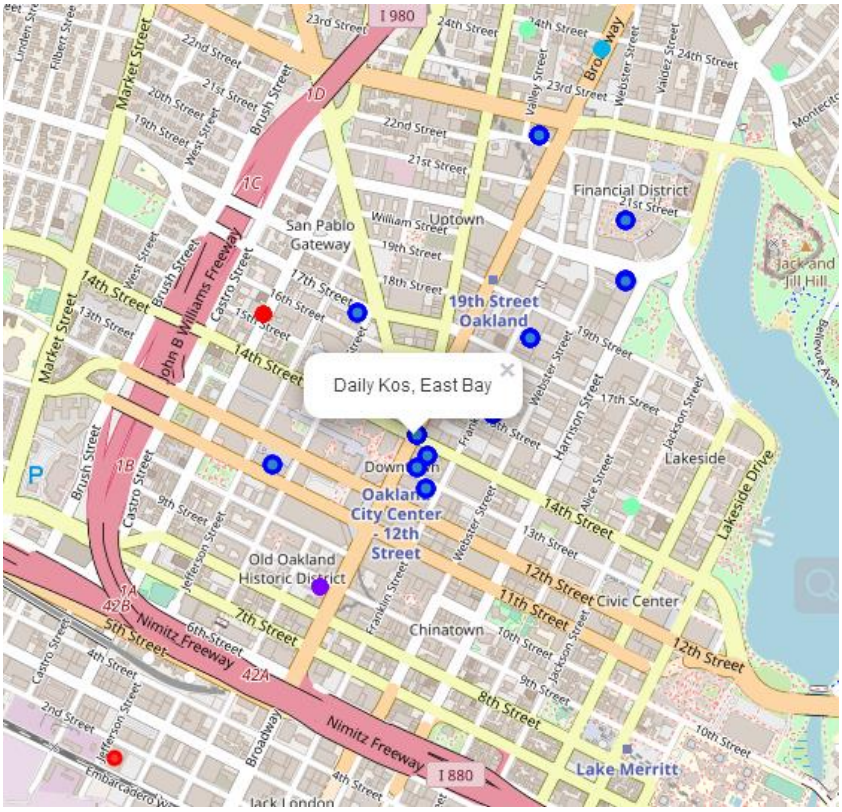


4.Results

Clusters	Colors	Stores types
0	Red	Electronics Store ,Arts & Crafts Store ,Furniture / Home Store
1	Purple	Clothing Store
2	Light Blue	Men's Store
3	Green	Grocery Store
4	Orange	Liquor Store

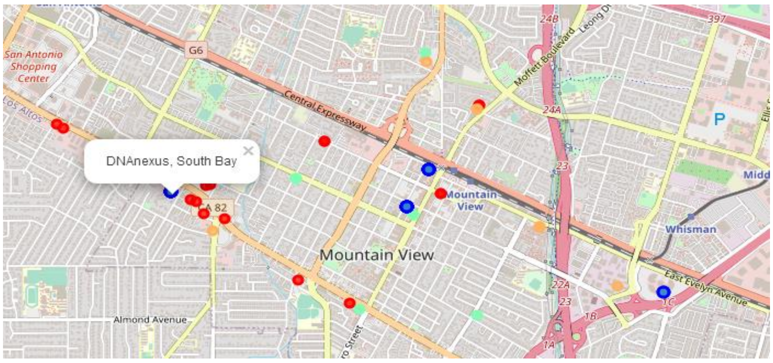
5.Discussion

I:



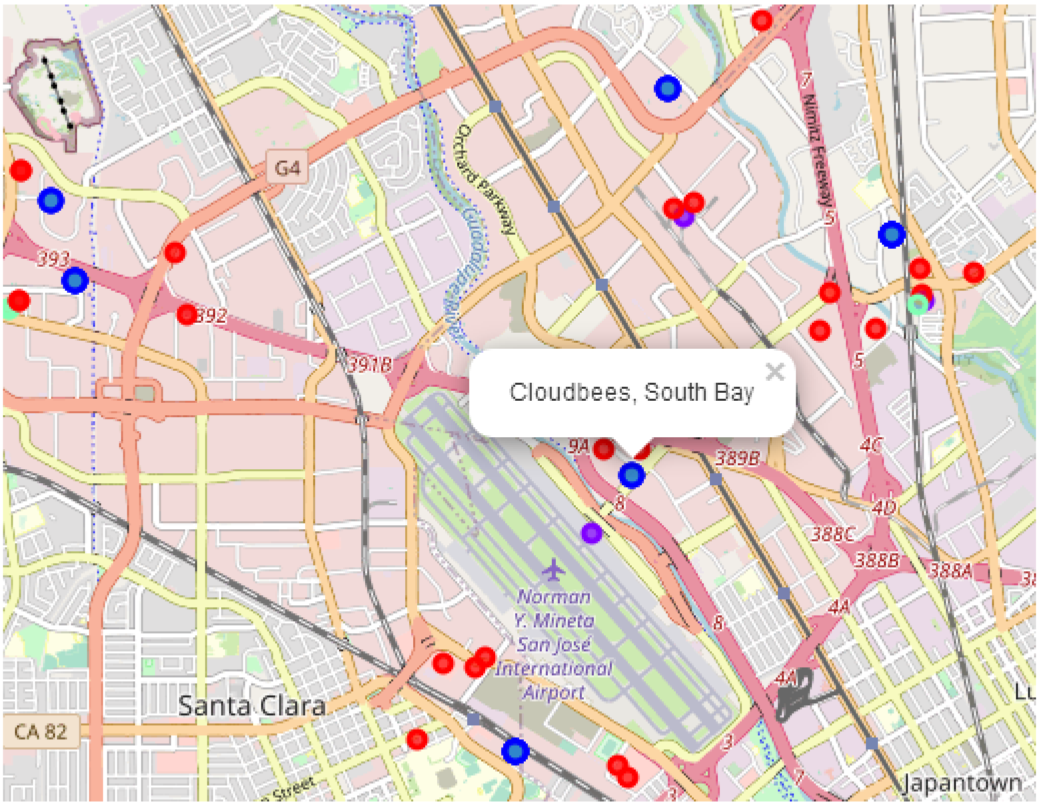
In general, the red represents Electronics Store, Arts & Crafts Store, Furniture / Home Store etc. is the main part, which is also because the base of these stores is relatively large, but surprisingly there are only 2 such stores near Daily Kos, East Bay. So, it would be a wise decision to open an Electronics Store here.

II:



Similarly, the numbers of purple point and light blue points represents clothing store and men's store are **0** near DNAnexus, South Bay. Therefore, if you start a clothing store or men's store then nearly all the customers working near DNAnexus, South Bay will come here.

III:



From the graph, we can conclude that a better choice to open stores near Cloudbees, South Bay is Grocery Store.

IV: We can find the orange points (Liquor Store) mostly located near San Francisco and Peninsula. So, the other places are all good choices.

6.Conclusion

In this study, I analyzed the best choice to open the stores near the Bay Area companies. Main judgment indicator is frequency of occurrence. I have solved the problem with some steps: collect data, clean data, KMeans clustering, data visualization and observation. Finally, I gave recommendations about the suitable kinds of stores for specific regions. The results of this project may help the relevant people to start new stores in Bay Area.