

# **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.



## 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 St	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://8tr	2008	51 Sharon	37.76523	-122.43	13-60					
4	10 by 10	B2B Softw	San Franc	Y Combin	We help ii	https://wv	2015	San Franc	37.77493	-122.419	1月12日					
5	15Five	Employee	San Francisco		15Five sof	15five.com	2011	12 Gallag	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