

Capstone Project

Analysis of Gas Stations in Bay Area



By: Vivek Advani

Date: 04/10/202

Introduction

This project is completed as part of IBM's Data Science Professional Certificate course offered by Coursera.org. This project is about analysing the Gas stations in the Bay Area. Given the number of vehicles in the Bay Area makes it a great location for a gas station business. The Bay has n number of cities and our goal is to extract hundreds of venues along within a mile from each city. This project will look for cities where there is little or no presence of gas stations to recommend a need for gas station in those cities. Based on the analysis we would be able you recommend a city with the need to have more gas stations.

Business Problem

The object of this project is to analyse which neighbourhood in Bay Area requires a gas station. With the help of Data Science Methodology and machine learning techniques, we can certainly build an analysis to provide solution for following questions.

Methodology

Analytic Approach The analytic approach for this problem is to perform unsupervised learning technique such as K-means Clustering. This will help to identify various patterns based on neighbourhoods in Bay Area.

Data Requirements We would require data such as list of Boroughs and Neighbourhood of San Francisco Bay Area, also the census data for Bay Area)

Data Collection Once we have noted the data requirements, the next step is data collection. We need to scrape data from the online websites using libraries such as Beautiful Soup, site map. Also, using Foursquare.

Data Understanding and Preparation The data understanding, and preparation part would be the most difficult as the collected data would not be clean. Goal here is to clean the data.

Modelling and Evaluation We would use K-Means algorithm to create K clusters. There is no such accuracy metric for Unsupervised Learning algorithm, we would use elbow graph to select the best K.

Data Source

["https://en.wikipedia.org/wiki/List_of_cities_and_towns_in_the_San_Francisco_Bay_Area"](https://en.wikipedia.org/wiki/List_of_cities_and_towns_in_the_San_Francisco_Bay_Area)

The San Francisco Bay Area, commonly known as the Bay Area, is a metropolitan region surrounding the San Francisco Bay estuaries in Northern California. According to the 2010 United States Census, the region has over 7.1 million inhabitants and approximately 6,900 square miles (18,000 km²) of land. The region is home to two major cities: San Francisco and Oakland and, the largest city, San Jose.

["https://foursquare.com/city-guide"](https://foursquare.com/city-guide)

Using Foursquare to get 100 venues for each neighbourhood in Bay Area and then select only Gas Stations to build model.

Data set

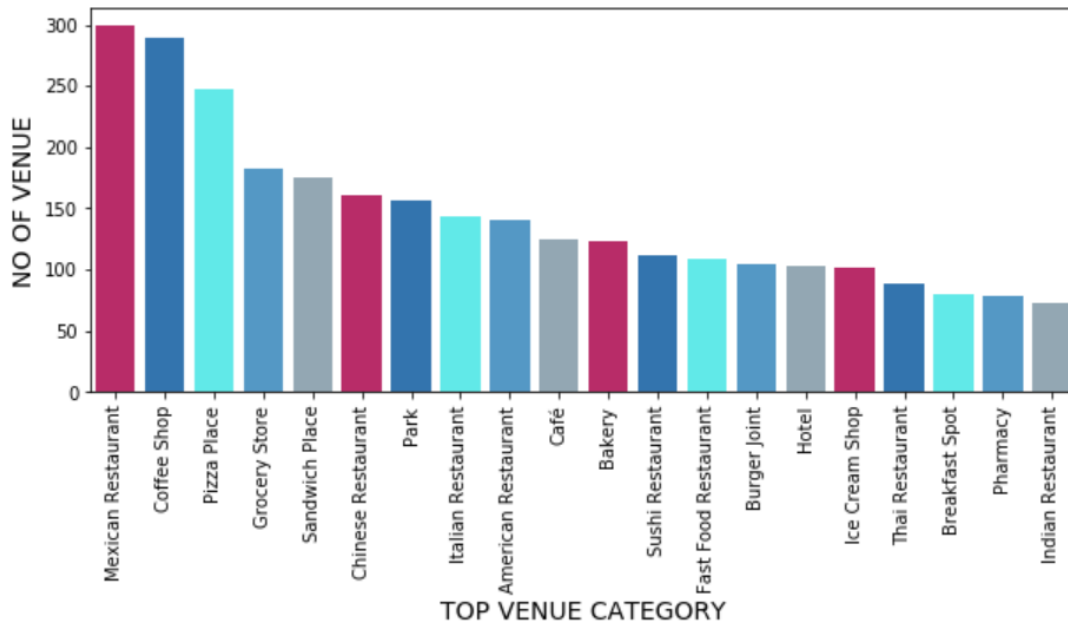
Scraped the data using Site Map browser extension. Added Latitude and Longitude using geopy.geocoders import Nominatim library.

Below are 5 records from dataset.

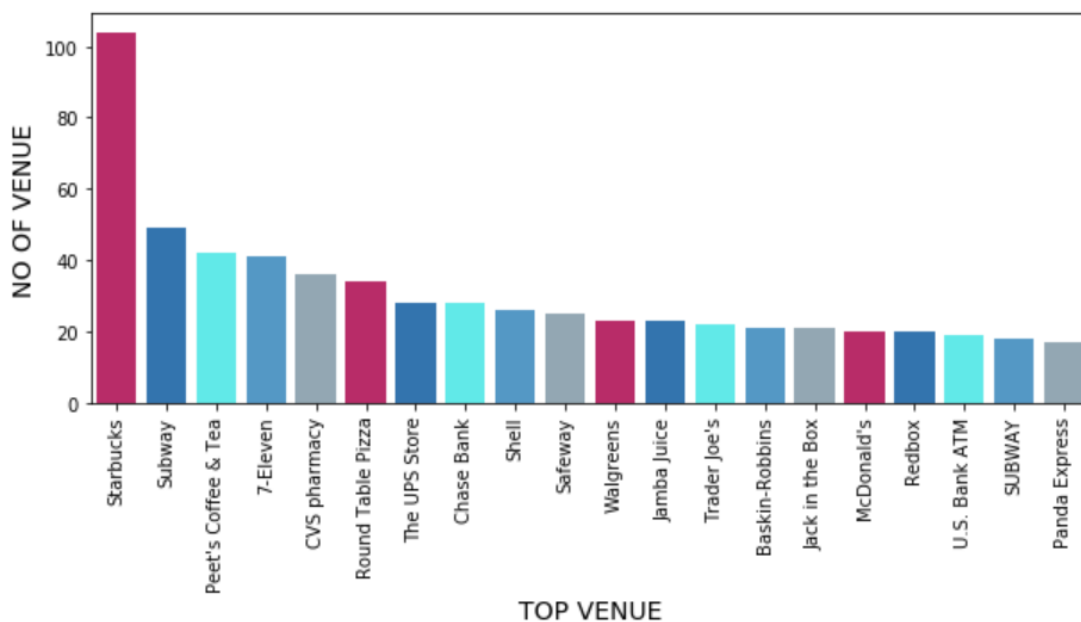
	County	City	Population	Land area (sq mi)	Latitude	Longitude
0	Alameda	Alameda	73,812	10.61	37.76707	-122.24584
1	Alameda	Fremont	2,14,089	77.46	37.55135	-121.98305
2	Alameda	Hayward	1,44,186	45.32	37.67134	-122.08556
3	Alameda	Livermore	80,968	25.17	37.67570	-121.75856
4	Alameda	Newark	42,573	13.87	37.53682	-122.03032

Data Visualisation

Scraped hundreds of venues within a radius of 1 mile for each city using Foursquare api. Bar graph below shows the Top Venue Category in the Bay Area, from which we can depict that Coffee shops and restaurants are the major venues. While Parks, Hotel and Pharmacy are few other top venues across the Bay Area.



Bar graph below shows the Top Venue in the Bay Area, from which we can depict that there are over hundred of Starbucks in Bay Area within a mile radius. This shows the love for Starbucks we have.



Bar graph below shows the spread of Gas stations across the Bay Area within a mile of their location, from which we can depict only few cities have more than 1 gas station within a radius of 1 mile. While some of them have 1 and the rest don't have gas station within a mile radius.

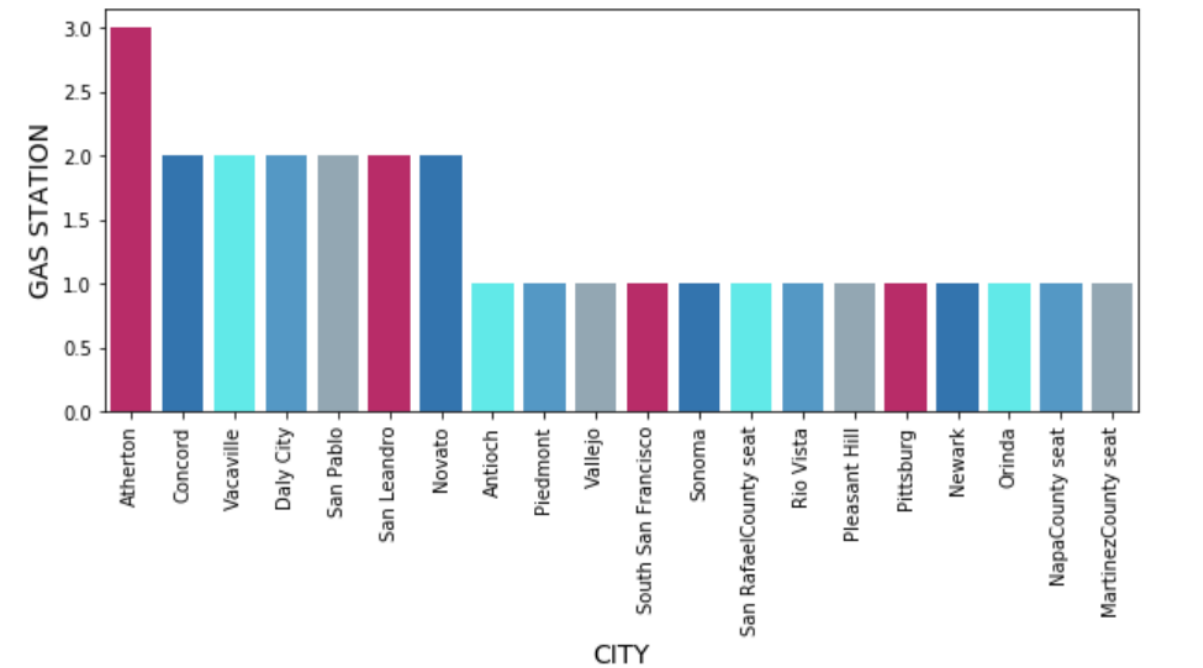


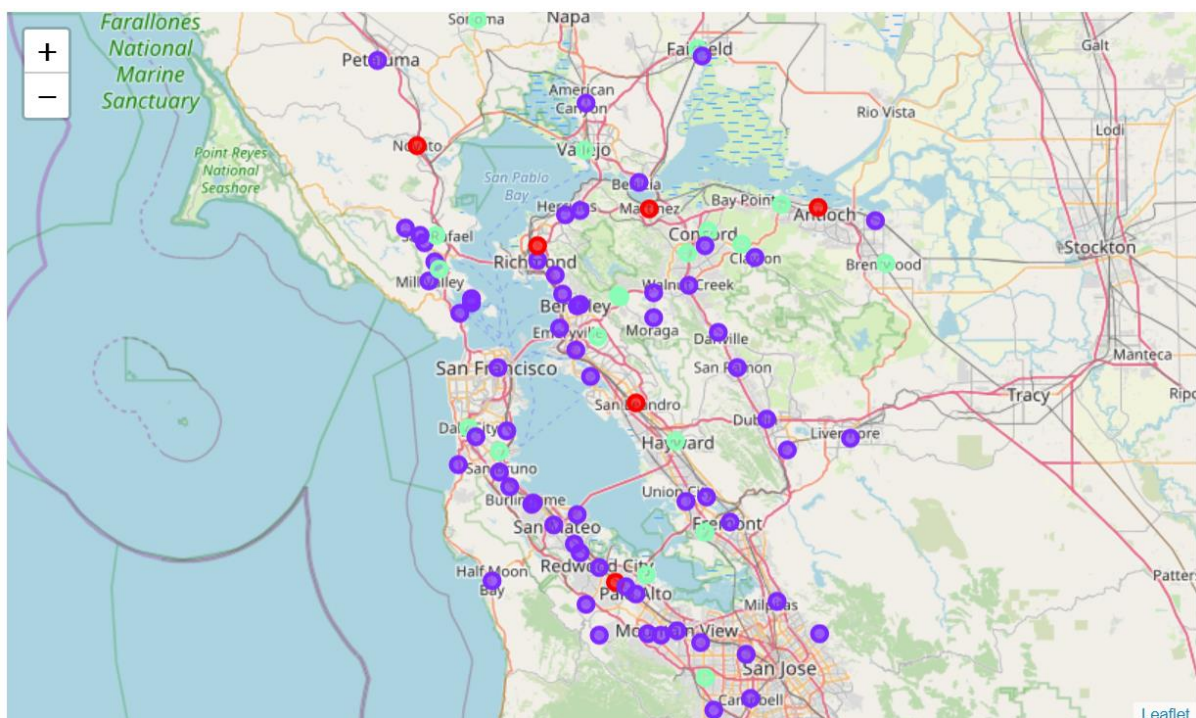
Table below shows 10 most common venue across the cities in Bay Area within a mile of their location, from which we can depict only few cities have more than 1 gas station within a radius of 1 mile. While some of them have 1 and the rest don't have gas station within a mile radius.




	City	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Alameda	Café	Mexican Restaurant	Coffee Shop	Grocery Store	Sandwich Place	Pizza Place	Asian Restaurant	Chinese Restaurant	Middle Eastern Restaurant	Dessert Shop
1	Albany	Coffee Shop	Grocery Store	Brewery	Burger Joint	Trail	Pizza Place	Japanese Restaurant	Thai Restaurant	Mexican Restaurant	Breakfast Spot
2	American Canyon	Mexican Restaurant	Convenience Store	Fast Food Restaurant	Pizza Place	Chinese Restaurant	Sandwich Place	Hotel	Pharmacy	Park	Ice Cream Shop
3	Antioch	Rental Car Location	Mexican Restaurant	Gym	Restaurant	Construction & Landscaping	Pizza Place	Burger Joint	Food Court	Bookstore	Gas Station
4	Atherton	Hotel	Mexican Restaurant	Pizza Place	Gas Station	Athletics & Sports	Chinese Restaurant	Asian Restaurant	Grocery Store	Coffee Shop	Food & Drink Shop

Clustering

Clustering technique is to find subgroups of observations within a data set. When we cluster the data, we want our similar data points or observations to be in the same group and dissimilar observations in different groups. As there is no target variable, this is an unsupervised algorithm, which implies that it seeks to find relationships between observations without being trained on a target variable. With the help of this technique we find identify the data points that are alike and categorize them in a cluster. K-means clustering was used in order to perform analysis on gas station, it is the most used clustering method to split the dataset into a set of k groups.

The value of K chosen is 3, Map below shows 3 clusters formed using K-Means clustering.



-  Cluster 0: Cities with high frequency of gas station within a mile range.
-  Cluster 1: Cities with low frequency of gas station within a mile range.
-  Cluster 2: Cities with moderate frequency of gas station within a mile range.

Cluster 0

```
df_merged.loc[df_merged['Cluster Label'] == 0].head()
```

	County	City	Population	Land area (sq mi)	Latitude	Longitude	Cluster Label	Gas Station
10	Alameda	San Leandro	84,950	13.34	37.72832	-122.15856	0	0.027027
19	Contra Costa	San Pablo	29,139	2.63	37.96033	-122.34239	0	0.028571
21	Contra Costa	MartinezCounty seat	35,824	12.13	38.01393	-122.13494	0	0.032258
29	Contra Costa	Antioch	1,02,372	28.35	38.01583	-121.81974	0	0.038462
38	Marin	Novato	51,904	27.44	38.10609	-122.56790	0	0.028169

Above chart shows that cities such as Antioch, San Leandro, etc has the highest frequency of gas station within a mile range.

Cluster 1

```
df_merged.loc[df_merged['Cluster Label'] == 1].head()
```

	County	City	Population	Land area (sq mi)	Latitude	Longitude	Cluster Label	Gas Station
0	Alameda	Alameda	73,812	10.61	37.76707	-122.24584	1	0.0
1	Alameda	Fremont	2,14,089	77.46	37.55135	-121.98305	1	0.0
3	Alameda	Livermore	80,968	25.17	37.67570	-121.75856	1	0.0
5	Alameda	OaklandCounty seat	3,90,724	55.79	37.80508	-122.27307	1	0.0
7	Alameda	Dublin	46,036	14.91	37.70423	-121.91635	1	0.0

Above chart shows that cities such as Fremont, Oakland, etc has the least frequency of gas station within a mile range. The cities in this cluster are densely populated and would require high frequency of gas stations.

Cluster 2

```
df_merged.loc[df_merged['Cluster Label'] == 2].head()
```

	County	City	Population	Land area (sq mi)	Latitude	Longitude	Cluster Label	Gas Station
2	Alameda	Hayward	1,44,186	45.32	37.67134	-122.08556	2	0.010000
4	Alameda	Newark	42,573	13.87	37.53682	-122.03032	2	0.022222
6	Alameda	Piedmont	10,667	1.68	37.82475	-122.23235	2	0.023810
14	Contra Costa	Pleasant Hill	33,152	7.07	37.94799	-122.06271	2	0.010638
15	Contra Costa	Pittsburg	63,264	17.22	38.01946	-121.88851	2	0.018519

Above chart shows that cities such as Hayward, Pittsburgh, etc has the moderate frequency or a min of 1 gas station within a mile range.

Recommendation

As per analysis done in this project, it is recommended to open a gas station within the cities such as Fremont, Oakland, etc which are included in cluster 1. Cluster 1 includes cities such as Fremont, Oakland, they are the ones with high population density and low on gas station within a mile range. However, this might depend on some other factors like distance to highway which is out of scope for this capstone project.

Also, we can conclude that population should be correlated with the frequency of gas station. This project only considers the frequency of gas stations within the cities in the Bay Area. However, several other factors might influence the recommendation such as , the purchasing power of consumers, distance from the community, number and type of vehicle owned by the people in those cities, etc. are not considered for recommendation.

Conclusion

With the help of Foursquare API and various machine learning techniques we can perform analysis on various other venues and can answer many business problems. Basis on the Data visualization we can get insights that there are lots of variety of Restaurants and Coffee shops in the Bay. So, if entrepreneurs are interested in opening a coffee place or a restaurant, they can use this model to identify the best location to start their business.

We can also use this model to identify the best cities to stay in the Bay Area considering the schools, parks and safety factor of the city.