Capstone Project

Analysis of Gas Stations in Bay Area

By Vivek Advani

Introduction

- This project is completed as part of IBM's Data Science Professional Certificate course offered by Coursera.org.
- This project is about analyzing 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.

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.
- Are there enough gas stations in our neighbourhood?
- Should population density be correlated to the number of gas stations?

Data Source

"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 in Northern California.

"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.

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. Also, using Foursquare.

Methodology (continued)

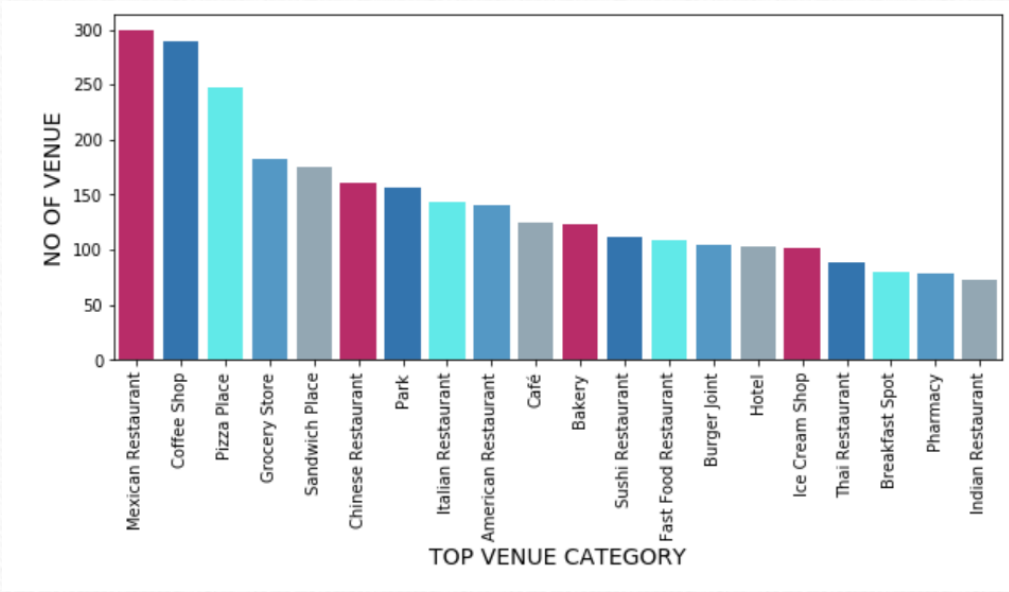
- Data Understanding and Preparation The data understanding, and preparation part would be the most difficult as the collected data would not be clean. Goal here is to clean the data.
- Modelling and Evaluation We would use K-Means algorithm to create K clusters. There is no such accuracy metric for Unsupervised Learning algorithm, we would use elbow graph to select the best K.

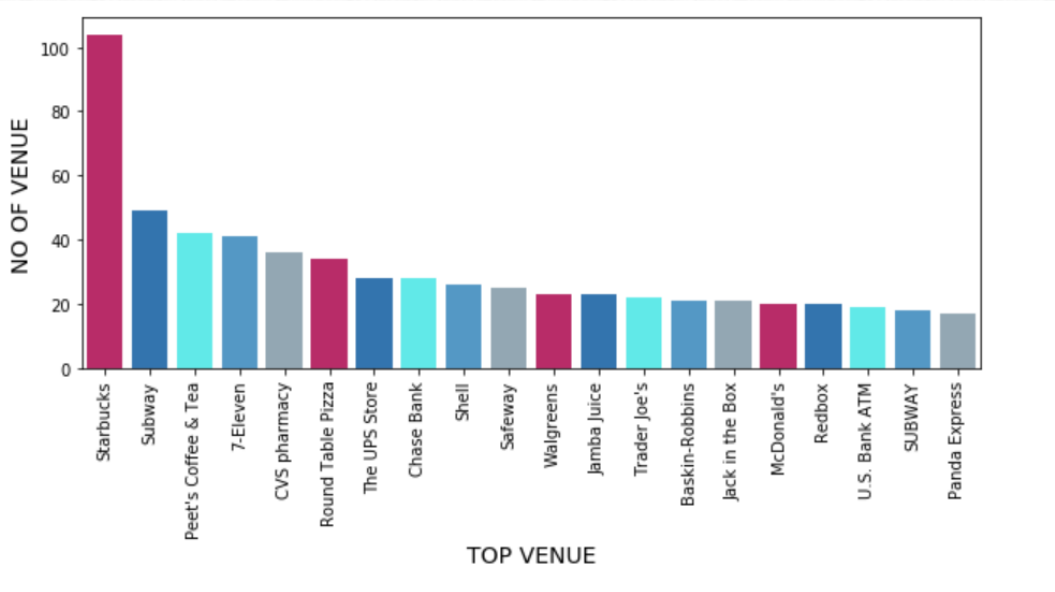
Data Gathering

- Scraped the data using Site Map browser extension.
- Added Latitude and Longitude using geopy.geocoders import Nominatim library.
- Below are 5 records from dataset.

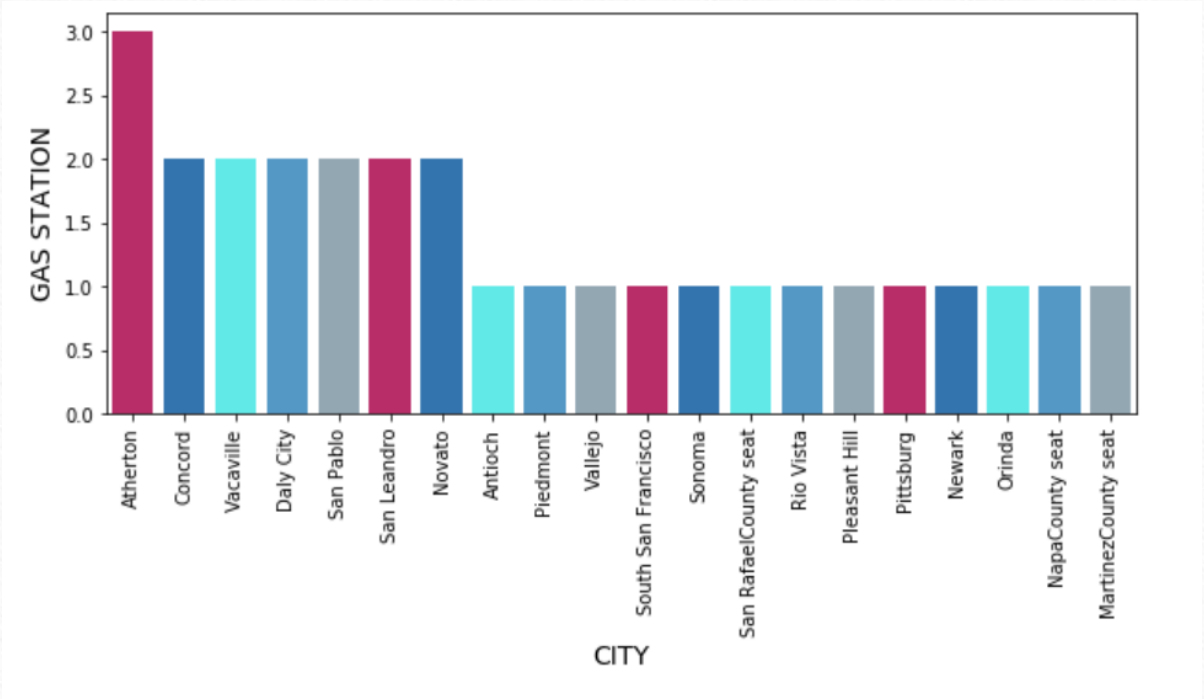
	County	City	Population	Land area (sq mi)	Latitude	Longitude
0	Alameda	Alameda	73,812	10.61	37.76707	-122.24584
1	Alameda	Fremont	2,14,089	77.46	37.55135	-121.98305
2	Alameda	Hayward	1,44,186	45.32	37.67134	-122.08556
3	Alameda	Livermore	80,968	25.17	37.67570	-121.75856
4	Alameda	Newark	42,573	13.87	37.53682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

Data Visualization





Gas Station in Bay Area(within a mile)

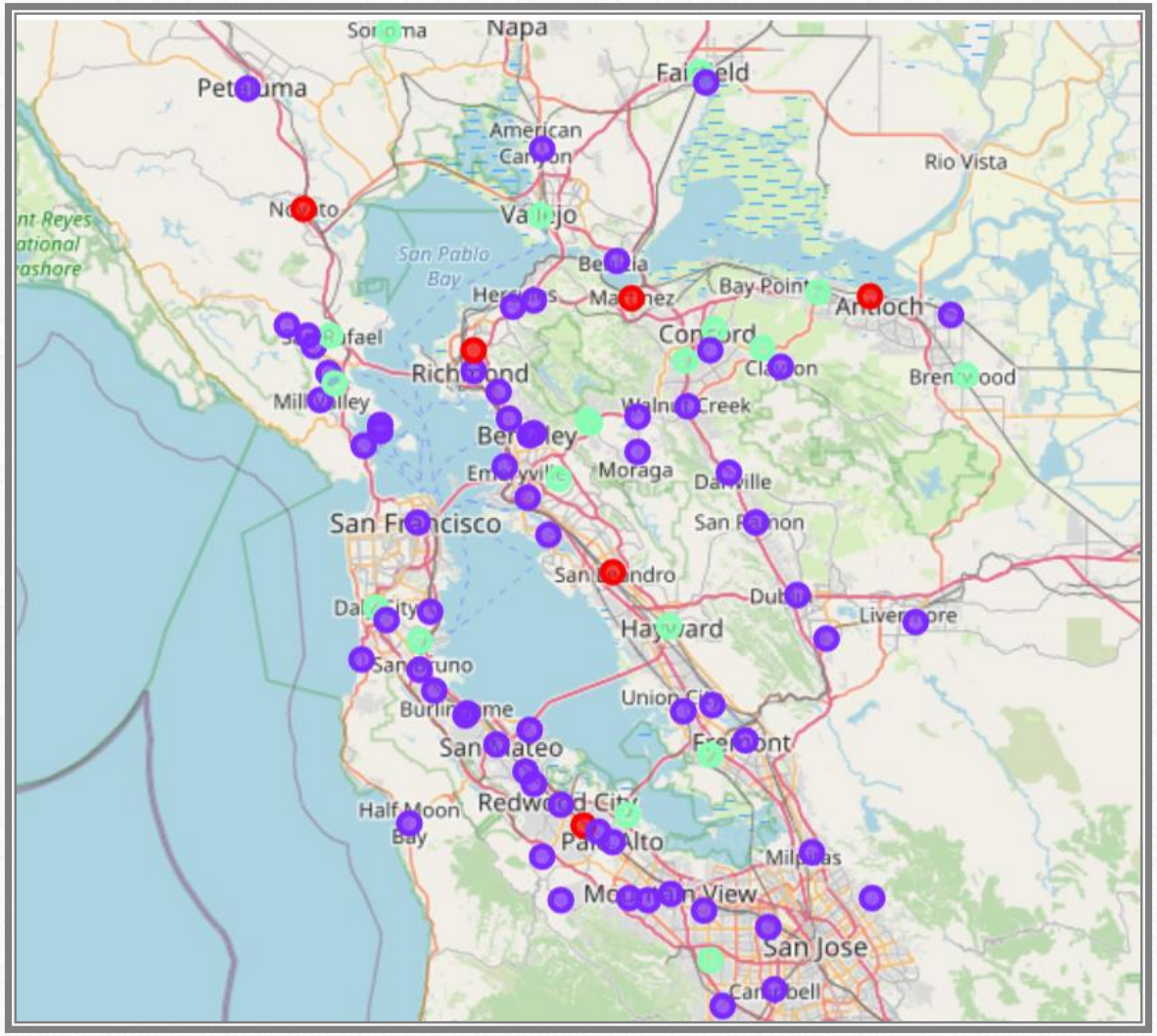


List of Top 10 common venue

	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

- Used K-Means clustering to form clusters of cities.
- The value of K chosen is 3.
- 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.



Recommendation

- As per analysis done in this project, it is recommended to open a gas station within the cities which are included in cluster 1.
- Cluster 1 includes cities such as Fremont, Oakland, they are the one's 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.

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 problem.
- Basis on the Data visualization we can get insights that there are lots of variety of Restaurants and Coffee shops in the Bay.

Challenges Faced

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.