Capstone Project

IBM Applied Data Science Capstone

Opening a New Gym in the city of Hayward, California

By: Ashneel Kumar

March 2020

Intro:

As much as we'd like to believe that chocolate cake and ice cream are the elixir for a long and healthy life, we know that's not the case. We also know that a regular exercise routine is one of the surest paths to a longer and healthier life.

The Centers for Disease Control recommends that adults get at least 150 minutes of moderately intense aerobic exercise every week – but less than half of all Americans do so. And for the millions of people who would like to do better, this is often the time of year to consider joining a gym. We can all find lots of excuses for not getting our bodies in motion, but there's an equally compelling list of reasons why we should do so.

Whether you want to lose weight, gain muscle, tone up or improve your overall health, you'll have to start somewhere and that's the No. 1 reason people of all fitness levels pony up the bucks and decide to join a conveniently located gym.

Business Problem:

The focus of this project is to analyze and pick a prime location in the city of Hayward, California to open a brand-new gym. Using data science methodology as the steppingstone with machine learning method and techniques like clustering to answer this business question:

Where would you recommend an investor/developer in the city of Hayward to open a brand-new Gym?

Target Audience:

This project is useful for investors, developers or entrepreneurs.

Data to solve this business problem:

- List of all the gyms in the city of Hayward CA

- Latitude and longitude coordinates of those gyms, which is a requirement in order to plot the maps and get venue data etc.

- Data involving all the gyms in the city of Hayward. The data will be used to preform clustering on the neighborhood.

Data Source, Web-scraping & Cleaning:

Data Source:

The data source will be this Wikipedia page:

https://en.wikipedia.org/wiki/Category:Neighborhoods_in_Hayward,_California

Data includes all the neighborhoods of Hayward, California.

Web-Scraping & Foursquare:

Using web scraping techniques to extract data from this Wikipedia page, in Jupyter Notebook using the python language and beautifulsoup packages as well as Python Geocoder package which will give us the latitude and longitude coordinate of the neighborhood in the city of Hayward California.

Also using Foursquare API to get all the data necessary regarding “Gyms” for those neighborhoods to solve the business problem.

At display will be a lot off data science skills such as, web scraping (Wikipedia), Foursquare (API), data cleaning, data wrangling, machine learning technique (K-Means Clustering) and visualization using (Folium).

Data Cleaning/Selection Process:

Data cleaning involved repeated cycles of screening, diagnosing, treatment and documentation of this process. As patterns of errors are identified, data collection and entry procedures were adapted to correct those patterns and reduce future errors. Data scraped using web-scraping were transferred and combined into one table. Missing data were answered by dropping that rows from the table. Duplicated data were removed from the analysis.

After finishing the data cleaning process there were only 5 neighborhoods left with 3 features which included the neighborhood name, latitude & longitude. That’s all the information I needed to come to my conclusion regarding which area to open a brand-new gym in the city of hayward.

Methodology:

1.1 - The first process if getting the list of neighborhoods in the city of Hayward California. After researching I found the list available in a Wikipedia page (https://en.wikipedia.org/wiki/Category:Neighborhoods_in_Hayward,_California). The next step was web scraping using Python request and beautiful soup packages to extract the neighborhoods in the city of hayward.

	Neighborhood
0	Downtown Hayward
1	Eden Landing, California
2	Hayward Heath, California
3	Mount Eden, California
4	Schafer Park, California

Figure 1.1

1.2 - Also needed was geographical coordinates in the form of latitude and longitude in order to able to use Foursquare API. In that case I used a public data source "Opendatasoft.com", to search for the coordinates in the city of hayward. Here is the link for that data source: (https://public.opendatasoft.com/explore/embed/dataset/us-zip-code-latitude-and-longitude/table/?refine.state=CA&q=hayward). Then I had to transfer those coordinates into a csv file and upload to my notebook for data analysis. After gathering the data, the next solution was to transfer the data into a pandas Data Frame.

	Zip	City	State	latitude	longitude
0	94544	Hayward	CA	37.633732	-122.061010
1	94557	Hayward	CA	37.680181	-121.921498
2	94543	Hayward	CA	37.680181	-121.921498
3	94542	Hayward	CA	37.657381	-122.050760
4	94545	Hayward	CA	37.635582	-122.104180

Figure 1.2

1.3 - Now I had to populate the coordinates into latitude and longitude.

	latitude	longitude
0	37.633732	-122.061010
1	37.680181	-121.921498
2	37.680181	-121.921498
3	37.657381	-122.050760
4	37.635582	-122.104180
5	37.680181	-121.921498
6	37.674431	-122.088830

Figure 1.3

1.4 – Next step was to merge the coordinates into an original data frame.

	Neighborhood	latitude	longitude
0	Downtown Hayward	37.633732	-122.061010
1	Eden Landing, California	37.680181	-121.921498
2	Hayward Heath, California	37.680181	-121.921498
3	Mount Eden, California	37.657381	-122.050760
4	Schafer Park, California	37.635582	-122.104180

Figure 1.4

1.5 - Next was to visualize the neighborhoods using the Folium package. This allows us to preform a data visualization to make sure that the geographical coordinates data are correctly plotted in the city of Hayward, California.

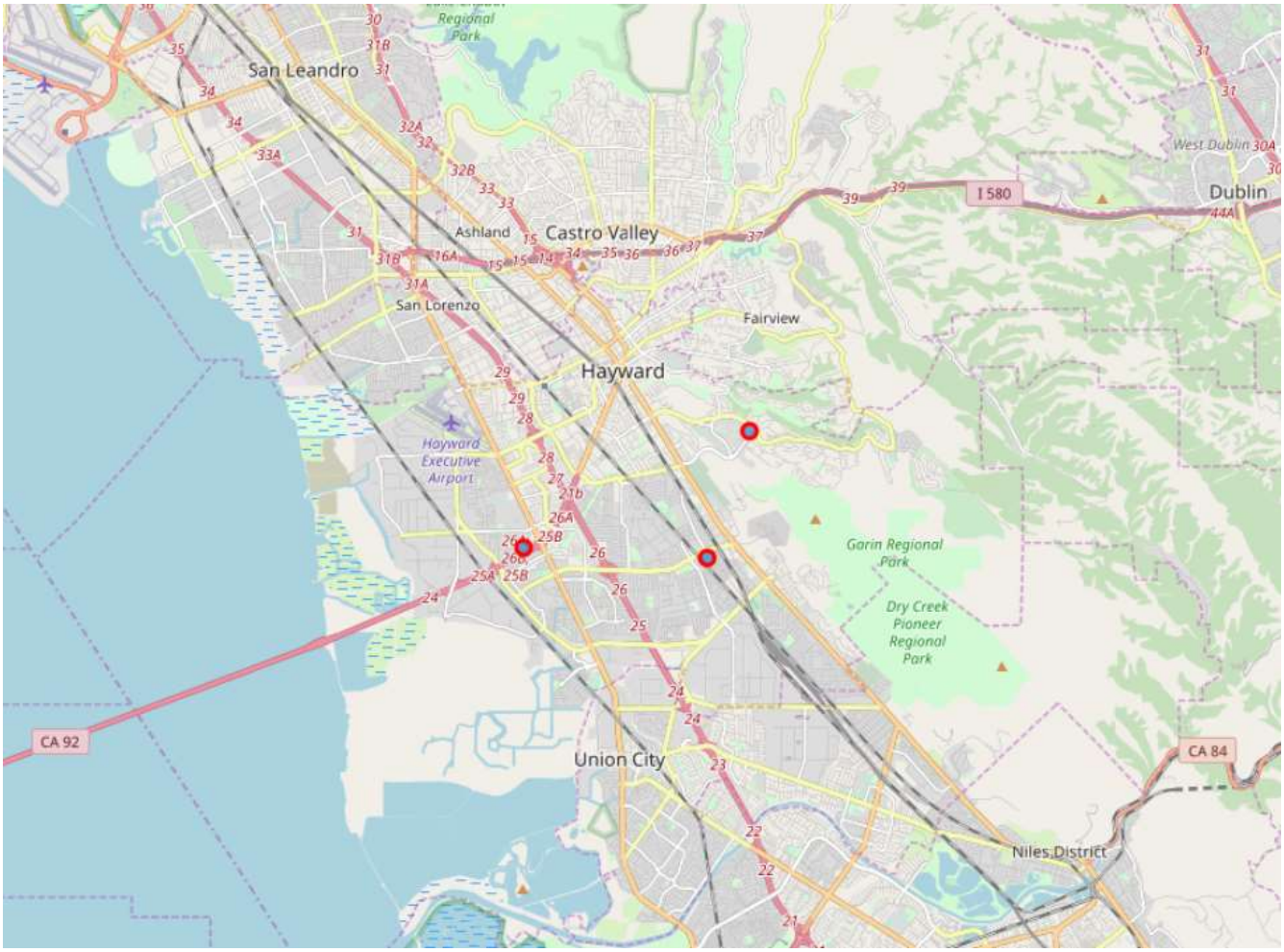

Figure 1.5

1.6 – Now we utilized the Foursquare API to get the top 100 venues that area within a radius of 9000 meters (5 miles). Using your own unique foursquare Id and Secret key, we make an API call to foursquare passing in the geographical coordinates of the neighborhood in a Python loop. Foursquare returns the venue data in JSON format and we will extract all the venue name, venue category, venue latitude and longitude.

	Neighborhood	Latitude	Longitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
0	Downtown Hayward	37.633732	-122.06101	Pupuseria Las Cabanas	37.627032	-122.043346	Latin American Restaurant
1	Downtown Hayward	37.633732	-122.06101	Bob's Hoagy Steaks	37.629506	-122.048341	Sandwich Place
2	Downtown Hayward	37.633732	-122.06101	Tacos Uruapan	37.622047	-122.056986	Mexican Restaurant
3	Downtown Hayward	37.633732	-122.06101	Bronco Billy's Pizza Palace	37.655525	-122.048817	Pizza Place
4	Downtown Hayward	37.633732	-122.06101	Garin/Dry Creek Pioneer Regional Parks	37.628355	-122.029091	Park

Figure 1.6

1.7 – Now with the data, we can check how many venues were returned for each neighborhood and examine how man With the data, we can check how many venues were returned for each neighborhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyze each neighborhood by grouping the rows by neighborhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Next, we filter the data as “Gym” as a venue category.

	Neighborhoods	Afghan Restaurant	Airport	American Restaurant	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Auto Dealership	BBQ Joint	Bakery	Bank	Bar	Beer Garden	Beer Store	Breakfast Spot	Bubble Tea Shop	Burger Joint	Burmese Restaurant	Café	Chinese Restaurant	City Hall	Coffee Shop	Comedy Club	Comic Shop	Convenience Store
0	Downtown Hayward	0.01	0.00	0.02	0.01	0.01	0.00	0.00	0.02	0.03	0.00	0.01	0.00	0.00	0.03	0.00	0.05	0.00	0.03	0.04	0.00	0.08	0.00	0.00	0.00
1	Eden Landing, California	0.01	0.00	0.02	0.00	0.00	0.01	0.01	0.00	0.03	0.00	0.00	0.01	0.01	0.01	0.00	0.04	0.01	0.01	0.00	0.00	0.05	0.01	0.01	0.01
2	Hayward Heath, California	0.01	0.00	0.02	0.00	0.00	0.01	0.01	0.00	0.03	0.00	0.00	0.01	0.01	0.01	0.00	0.04	0.01	0.01	0.00	0.00	0.05	0.01	0.01	0.01
3	Mount Eden, California	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.02	0.02	0.00	0.03	0.00	0.00	0.03	0.01	0.05	0.00	0.04	0.03	0.01	0.08	0.00	0.00	0.00
4	Schafer Park, California	0.01	0.01	0.01	0.00	0.02	0.00	0.00	0.03	0.01	0.01	0.03	0.00	0.00	0.02	0.00	0.07	0.00	0.03	0.02	0.01	0.06	0.00	0.00	0.00

Figure 1.7

	Neighborhoods	Gym
0	Downtown Hayward	0.02
1	Eden Landing, California	0.01
2	Hayward Heath, California	0.01
3	Mount Eden, California	0.01
4	Schafer Park, California	0.01

Figure 1.7 (2)

1.8 – The final steps were to preform clustering on the data by using k-means clustering. The K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. K-means algorithm identifies K number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. This algorithm is particularly suited to solve the problem for this project.

Now we will cluster the neighborhoods into 2 clusters (0 & 1) based on their frequency of occurrence for “Gym”. The results will allow us to identify which neighborhoods is over-saturated with gyms and which neighborhoods are not. Based on the occurrence of Gyms in different neighborhoods, it will help us to answer the question as to where would you recommend an investor/developer in the city of Hayward to open a brand-new Gym?

	Neighborhood	Gym	Cluster Labels	latitude	longitude
1	Eden Landing, California	0.01	0	37.680181	-121.921498
2	Hayward Heath, California	0.01	0	37.680181	-121.921498
3	Mount Eden, California	0.01	0	37.657381	-122.050760
4	Schafer Park, California	0.01	0	37.635582	-122.104180

Figure 1.8 (Cluster 0)

	Neighborhood	Gym	Cluster Labels	latitude	longitude
0	Downtown Hayward	0.02	1	37.633732	-122.06101

Figure 1.8 (Cluster 1)

Observation:

After investigating the data regarding all the gyms in the city of Hayward, California. The highest number of gyms are in cluster 0 and smaller number of gyms are in cluster 1. Cluster 1 checks all the boxes which includes: traffic area (more traffic to your business), downtown area (more appeal to your business) & less competition (Gyms). Property investors/developers with unique selling propositions to stand out from the competition can also open a brand-new Gym in neighborhoods in cluster 1 with less competition. Lastly, property developers are advised to avoid neighborhoods in cluster 0 which already have high concentration of gyms and suffering from intense competition. So, to answer the business question I would recommend an investor/developer to open a brand-new Gym in the city of Hayward in Cluster 1 (Downtown Hayward).

Future Research:

This project consisted of two factors including location and frequency of occurrence of Gym's in different neighborhood in the city of Hayward, California. Other factors such as income of residents, population, crime rate could influence the location decision of a new Gym. However, the business question was just to find a location for a brand-new gym. As stated above, those factors could be utilized to devise a different methodology to estimate such data to be used in a clustering algorithm to determine a preferred location to open a gym as well.

Conclusion:

This project consisted of all the necessary steps that a regular data scientist use.

1. *Business Understanding:* Ask relevant questions and define objectives for the problem that needs to be tackled
2. *Data Mining:* Gather and scrape the data necessary for the project.
3. *Data Cleaning:* Fix the inconsistencies within the data and handle the missing values.
4. *Data Exploration:* Form hypotheses about your defined problem by visually analyzing the data.
5. *Model:* Construct a model to predict and forecast
6. *Interpret:* Put the results into good use.

Lastly providing recommendations to the relevant stakeholders i.e. property developers and investors regarding the best locations to open a new gym. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighborhoods in cluster 1 are the most preferred locations to open a new gym. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding competition areas in their decision making to open a new gym.