Coursera Capstone Project: Battle of the Neighborhoods

Project Title: Investment opportunities in the Healthy Food and Wellness Restaurants sector in the city of Toronto, Canada

Introduction/Business problem

The health and wellness food market is projected to grow about USD 235 billion during 2020 to 2024 with an average grow of 6% per year. In this period, 37% of this grow is coming from the Americas. Also, this market is fragmented with a few players occupying the market share, the key to this market is the increasing adoption of healthy habits in the society.

Market and geographical data are key to localize the best places to build, install or open a new restaurants or coffee shops within this new and profitable market. In this project we propose to use foursquare and data analysis to propose the best places for investors to open restaurants and or coffee shops in the healthy food market in the city of Toronto

Fig 1 Global Health and Wellness Food Market 2018-2020

Problem Which Tried to Solve:

The major purpose of this project, is to suggest a better neighborhood investment opportunities in the health and wellness food sector in Toronto, Canada

The Location:

Greater Toronto Area

Data Description

We will use the Toronto postal codes Dataset consisting of latitude and longitude, zip codes and available here: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Foursquare API Data:

For this project we will use data about different venues in different neighborhoods of that specific borough. In order to gain that information we will use "Foursquare" locational information that includes information as venue names, locations, menus and photos. After finding the list of neighborhoods, we then connect to the Foursquare API to gather information about venues in every neighborhood. For each neighborhood, we have chosen the radius to be 100 meters. Specifically, for each venue, the information obtained is:

1. Neighborhood
2. Neighborhood Latitude
3. Neighborhood Longitude
4. Venue
5. Name of the venue e.g. the name of a store or restaurant
6. Venue Latitude
7. Venue Longitude
8. Venue Category

Libraries Which are Used to Develop the Project:

- Pandas
- Folium
- Scikit Learn.
- JSON
- XML
- Geocoder
- Matplotlib

Methodology

The methodology applied for this study is the CRISP-DM. The follow scheme shows the steps followed in our specific study case.

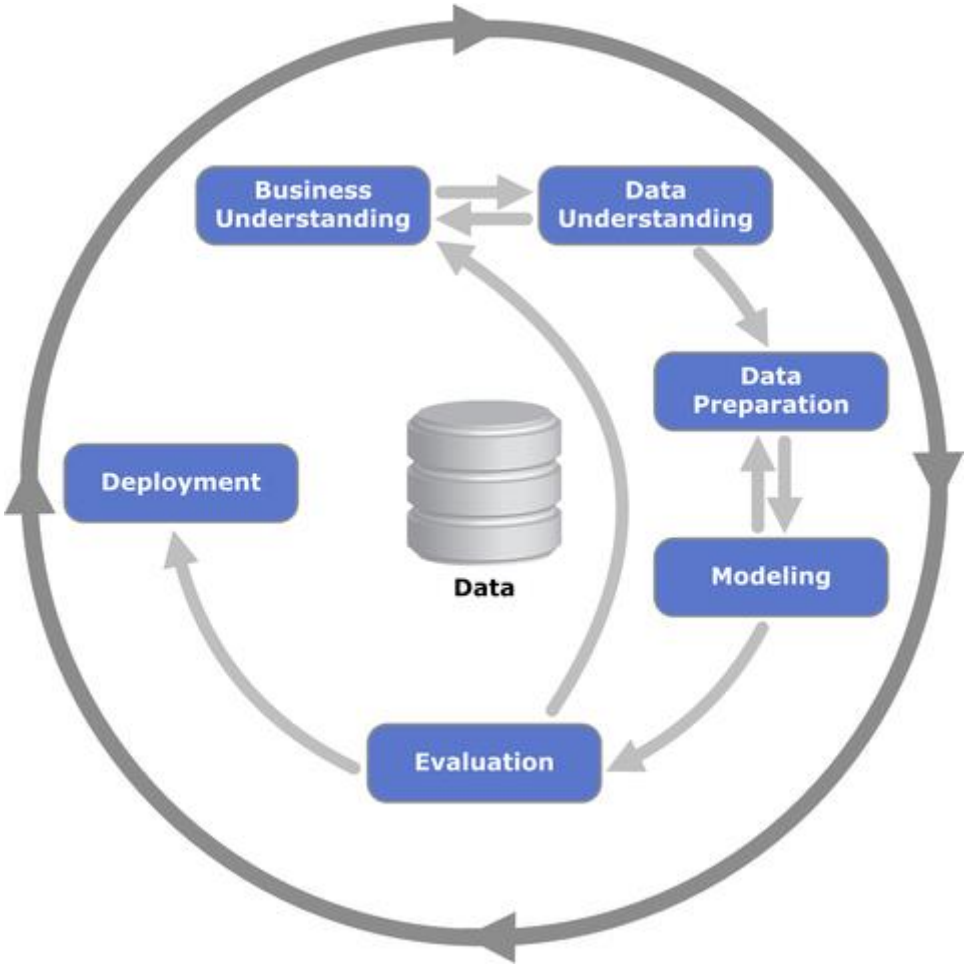

Fig. 2 CRISP-DM Methodology

Each point is explained in the next lines:

Business understanding: Analysis of the problem to be solved. In our case, to get information about the best places to open new Healthy Food and wellness restaurants in Toronto.

Data Understanding: The data needed for performing the study. In our case the data is the demographic and geographic data of the boroughs and neighborhoods in Toronto, the postal code and the respective geospatial coordinates for each neighborhood and the list of Venues for each location.

Data Preparation: Prepare pandas data frame with the neighborhoods, boroughs and postal code (pandas data frame with the geospatial coordinates for each neighborhood, and the list of venues using the FoursquareAPI)

Modeling: Create a map of Toronto's neighborhoods, a machine learning algorithm based on a Kmean approach to segment and cluster the list of venues in each neighborhood. Also we need to evaluate the model creating a map of the clusters and evaluate the accuracy of the kmean algorithm during the clustering process.

Deployment: According to the frequency of the Healthy Food Stores and other venues like gyms or parks in each neighborhood, make a conclusion about which one would be the best to open a new store

Results

In our first step we obtained the all the data needed for the study. As we want to study the city of Toronto we take the Toronto postal code and neighborhoods list from the link https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M . Then, we created a data frame with all the information obtained from the link as we can see in the next image:

	Postal Code	Borough	Neighborhood
0	M1S	Scarborough	Agincourt
1	M8W	Etobicoke	Alderwood, Long Branch
2	M3H	North York	Bathurst Manor, Wilson Heights, Downsview North
3	M2K	North York	Bayview Village
4	M5M	North York	Bedford Park, Lawrence Manor East
...	...	...	...
98	M2R	North York	Willowdale, Willowdale West
99	M1G	Scarborough	Woburn
100	M4C	East York	Woodbine Heights
101	M2P	North York	York Mills West
102	M2L	North York	York Mills, Silver Hills

Fig. 3 Neighborhood and Postal Codes data frame

Later, using the coordinates obtained from geocoder we can merge it with the anterior data frame to obtain a data frame where we have the following column names and information: Postal Code, Borough, Neighborhood, Latitude and Longitude.

	Postal Code	Borough	Neighborhood	Latitude	Longitude
0	M5E	Downtown Toronto	Berczy Park	43.64531	-79.37368
1	M6K	West Toronto	Brockton, Parkdale Village, Exhibition Place	43.63941	-79.42676
2	M7Y	East Toronto	Business reply mail Processing Centre, South C...	43.64869	-79.38544
3	M5G	Downtown Toronto	Central Bay Street	43.65609	-79.38493
4	M6G	Downtown Toronto	Christie	43.66878	-79.42071
5	M4Y	Downtown Toronto	Church and Wellesley	43.66659	-79.38130
6	M5V	Downtown Toronto	CN Tower, King and Spadina, Railway Lands, Har...	43.64082	-79.39956
7	M5L	Downtown Toronto	Commerce Court, Victoria Hotel	43.64823	-79.37890
8	M4S	Central Toronto	Davisville	43.70340	-79.38596
9	M4P	Central Toronto	Davisville North	43.71276	-79.38851
10	M6H	West Toronto	Dufferin, Dovercourt Village	43.66509	-79.43871
11	M5X	Downtown Toronto	First Canadian Place, Underground city	43.64828	-79.38146
12	M5P	Central Toronto	Forest Hill North & West, Forest Hill Road Park	43.69479	-79.41440
13	M5B	Downtown Toronto	Garden District, Ryerson	43.65736	-79.37818

Fig. 4 Neighborhood and Postal Codes data frame merged with longitude and latitude data frame for each location.

After obtaining this information we have all the information needed to the next step: Creating a map of Toronto Neighborhoods.

Creating a map of Toronto's neighborhoods:

After we had available all the information into a data frame, we created a map of the neighborhoods of Toronto, to check the accuracy of the data. First, we needed to get the location of the city of Toronto using the geopy library and finally, using this information, the data frame 4 and the Folium package we got the map

```
address = 'Toronto, ON' #Use geopy library to get the latitude and longitude values of Toronto

geolocator = Nominatim(user_agent="toronto_explorer")
location = geolocator.geocode(address)
latitude = location.latitude
longitude = location.longitude
print('The geograpical coordinate of Toronto are {}, {}.'.format(latitude, longitude))
```

```
The geograpical coordinate of Toronto are 43.6534817, -79.3839347.
```

```
map_toronto = folium.Map(location=[latitude, longitude], zoom_start=8)

# add markers to map
for lat, lng, borough, neighborhood in zip(df_toronto_coor['Latitude'], df_toronto_coor['Longitude'], df_toronto_coor['Borough'], df_toronto_coor['Neighborhood']):
    label = '{}, {}'.format(neighborhood, borough)
    label = folium.Popup(label, parse_html=True)
    folium.CircleMarker(
        [lat, lng],
        radius=5,
        popup=label,
        color='purple',
        fill=True,
        fill_color='#c931cc',
        fill_opacity=0.7,
        parse_html=False).add_to(map_toronto)

map_toronto
```

Fig. 5 Code needed to generate the Toronto City Map

The map obtained shows the differeten neighborhood in the city of Toronto accrding to our data

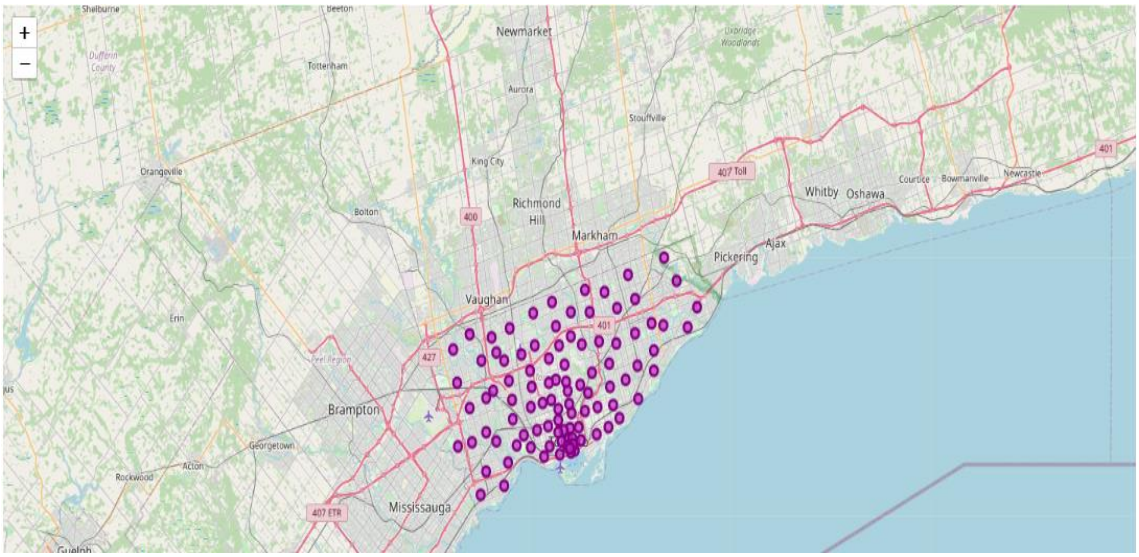

Fig. 6 Map of Toronto: each purple point is a neighborhood.

As we can see the map created shows in purple circle markers the neighborhoods in Toronto. A total of 102 markers are spread around the downtown, representing the 102 neighborhoods of the city

Using the Explore of Foursquare query to create a list of venues

After the map is displayed, we used the Foursquare API to find the list of all the venues in Toronto, having as return a data frame with the venues of every neighborhood like is showed in the Fig 7.

```
# Get the venues for each neighborhood
toronto_venues = getNearbyVenues(
                                  postalCodes=df_toronto_ngh['Postal Code'],
                                  boroughs=df_toronto_ngh['Borough'],
                                  neighbourhoods=df_toronto_ngh['Neighborhood'],
                                  latitudes=df_toronto_ngh['Latitude'],
                                  longitudes=df_toronto_ngh['Longitude']
                                 )

toronto_venues.head()
```

	Postal Code	Borough	Neighborhood	Latitude	Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	M5E	Downtown Toronto	Berczy Park	43.64531	-79.37368	The Keg Steakhouse + Bar - Esplanade	43.646712	-79.374768	Restaurant
1	M5E	Downtown Toronto	Berczy Park	43.64531	-79.37368	Fresh On Front	43.647815	-79.374453	Vegetarian / Vegan Restaurant
2	M5E	Downtown Toronto	Berczy Park	43.64531	-79.37368	Meridian Hall	43.646292	-79.376022	Concert Hall
3	M5E	Downtown Toronto	Berczy Park	43.64531	-79.37368	Goose Island Brewhouse	43.647329	-79.373541	Beer Bar
4	M5E	Downtown Toronto	Berczy Park	43.64531	-79.37368	LCBO	43.642944	-79.372440	Liquor Store

Fig. 7 Toronto Foursquare list of venues

Machine learning approach to cluster the neighborhoods

Machine learning approach to cluster the neighborhoods the clustering of the neighborhoods according to the venues, to evaluate the best candidates to open a new Healthy Food store was made using the K-means algorithm. This algorithm is unsupervised, fast and easy to apply, so it is largely used in application where clustering and segmentation of people, locations and data in general is required. The study was performed with K = 5, which means the data was segmented in 5 different clusters

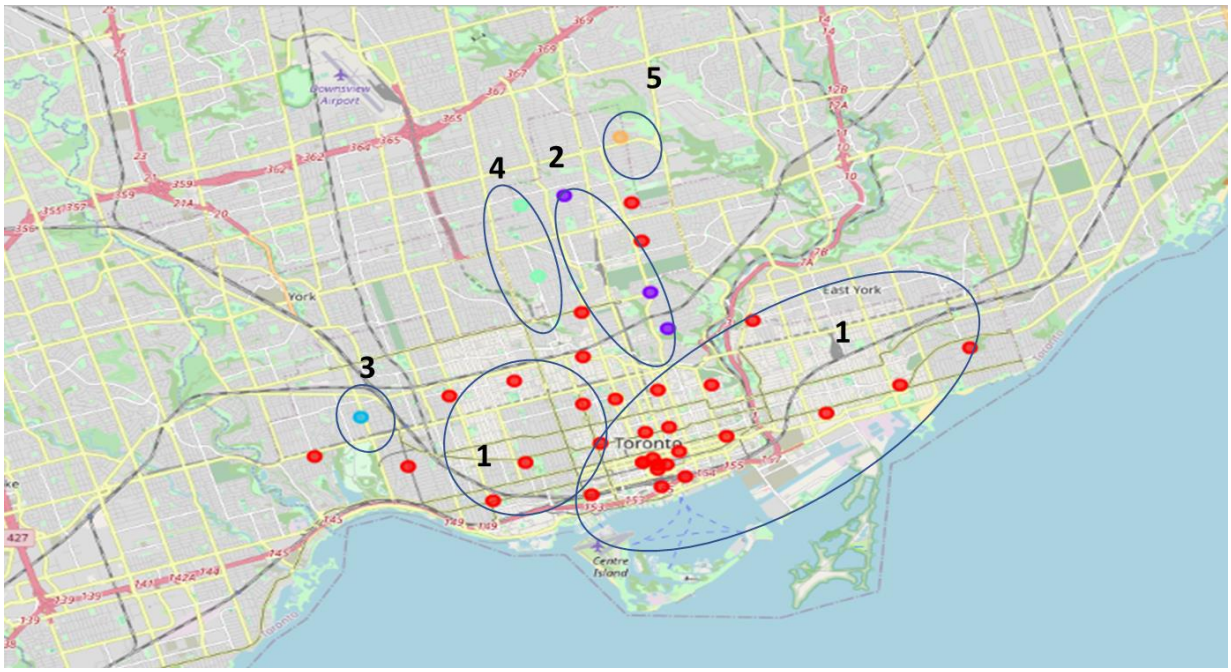

Fig. 8 Map of Toronto showing the segmented neighborhoods by venue type. Each circle is one cluster. The red points are the same cluster number 1

The visualization of the clusters in the map was made by the Folium package and it is possible to observe the 5 different color markers, although one of them, the red marker is largely representative.

Discussion

In this section we are going to discuss every cluster obtained from the previous segmentation of the data:

Cluster 1 (cluster label = 0) (Red color)

The Cluster 1 is formed by neighborhood with a high density of restaurants of a diverse range of options. It includes food from different places of the world, being the Asian food restaurants extremely popular. This cluster is probably has a lot of competitors but also a big population density

```
#examining the cluster 1
toronto_merged.loc[toronto_merged['Cluster Labels'] == 0]
```

	Postal Code	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	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
0	M5E	Downtown Toronto	Berczy Park	43.64531	-79.37368	0	Coffee Shop	Restaurant	Hotel	Breakfast Spot	Beer Bar	Italian Restaurant	Seafood Restaurant
1	M6K	West Toronto	Brockton, Parkdale Village, Exhibition Place	43.63941	-79.42676	0	Café	Coffee Shop	Bar	Restaurant	Nightclub	Bakery	Sandwich Place
2	M7Y	East Toronto	Business reply mail Processing Centre, South C...	43.64869	-79.38544	0	Coffee Shop	Hotel	Asian Restaurant	Restaurant	Italian Restaurant	Café	American Restaurant
3	M5G	Downtown Toronto	Central Bay Street	43.65609	-79.38493	0	Coffee Shop	Clothing Store	Plaza	Hotel	Middle Eastern Restaurant	Electronics Store	Italian Restaurant
4	M6G	Downtown Toronto	Christie	43.66878	-79.42071	0	Café	Grocery Store	Italian Restaurant	Baby Store	Coffee Shop	Athletics & Sports	Candy Store
5	M4Y	Downtown Toronto	Church and Wellesley	43.66659	-79.38130	0	Coffee Shop	Japanese Restaurant	Restaurant	Gay Bar	Sushi Restaurant	Men's Store	Bubble Tea Shop
6	M5V	Downtown Toronto	CN Tower, King and Spadina, Railway Lands, Har...	43.64082	-79.39956	0	Coffee Shop	Italian Restaurant	Café	Bar	Park	Gym / Fitness Center	Speakeasy

Fig. 9 Part of the neighborhoods and venues of Cluster 1

33	M4V	Central Toronto	Summerhill West, Rathnelly, South Hill, Forest...	43.68569	-79.40232	0	Coffee Shop	Light Rail Station	Liquor Store	Supermarket	Women's Store	Dumpling Restaurant	Fast Food Restaurant
34	M5R	Central Toronto	The Annex, North Midtown, Yorkville	43.67484	-79.40185	0	Italian Restaurant	Sandwich Place	Coffee Shop	Café	Vegetarian / Vegan Restaurant	Mexican Restaurant	Burger Joint
35	M4E	East Toronto	The Beaches	43.67703	-79.29542	0	Health Food Store	Pub	Trail	Women's Store	Distribution Center	Farmers Market	Farm
36	M4K	East Toronto	The Danforth West, Riverdale	43.68375	-79.35528	0	Discount Store	Intersection	Grocery Store	Bus Line	Park	Coffee Shop	Ice Cream Shop
37	M5K	Downtown Toronto	Toronto Dominion Centre, Design Exchange	43.64710	-79.38153	0	Coffee Shop	Hotel	Café	Restaurant	Japanese Restaurant	Seafood Restaurant	American Restaurant
38	M5S	Downtown Toronto	University of Toronto, Harbord	43.66311	-79.40180	0	Café	Restaurant	Bakery	Coffee Shop	Bookstore	Gym	Bar

Fig. 10 Part of the neighborhoods and venues of Cluster 1

Cluster 2 (cluster label = 1) (Purple color)

In this cluster we can see the venues of three neighborhoods where there are not so much restaurants but there are a lot of parks, playgrounds, gyms and sport venues that can be a good sign to invest in the healthy food industry.

```
#cluster 2
toronto_merged.loc[toronto_merged['Cluster Labels'] == 1]
```

	Postal Code	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	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
20	M4T	Central Toronto	Moore Park, Summerhill East	43.69066	-79.38356	1	Playground	Gym	Park	Tennis Court	Women's Store	Distribution Center	Farm	Falafel Restaurant
21	M4R	Central Toronto	North Toronto West, Lawrence Park	43.71452	-79.40696	1	Gym Pool	Playground	Park	Garden	Women's Store	Distribution Center	Farm	Falafel Restaurant
26	M4W	Downtown Toronto	Rosedale	43.68190	-79.37850	1	Japanese Restaurant	Playground	Park	Bike Trail	Dog Run	Farmers Market	Farm	Falafel Restaurant

Fig. 12 Neighborhoods and venues of Cluster 2

Cluster 3 (cluster label =2) (Blue color)

The Cluster 3 is formed by one neighborhood in which there are some restaurants and some options of leisure. Maybe is not best place to invest because low density population and a lot of competitors

```
#cluster 3
toronto_merged.loc[toronto_merged['Cluster Labels'] == 2]
```

	Postal Code	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	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
15	M6P	West Toronto	High Park, The Junction South	43.65994	-79.46302	2	Sandwich Place	Park	Residential Building (Apartment / Condo)	Women's Store	Farm	Falafel Restaurant	Event Space	Ethiopian Restaurant

Fig. 13 Neighborhoods and venues of Cluster 3

Cluster 4 (cluster label = 3) (Green color)

The Cluster 4 is formed by 2 neighborhood and it is characterized by having some places to eat and stores at the same proportion. There are not so much food options so this neighborhood could be a great investment opportunity

```
#cluster 4
toronto_merged.loc[toronto_merged['Cluster Labels'] == 3]
```

	Postal Code	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	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
12	M5P	Central Toronto	Forest Hill North & West, Forest Hill Road Park	43.69479	-79.41440	3	Home Service	Women's Store	Dog Run	Fast Food Restaurant	Farmers Market	Farm	Falafel Restaurant	Event Space
27	M5N	Central Toronto	Roselawn	43.71194	-79.41912	3	Home Service	Clothing Store	Women's Store	Dog Run	Fast Food Restaurant	Farmers Market	Farm	Falafel Restaurant

Fig. 14 Neighborhoods and venues of Cluster 4

Cluster 5 (cluster label = 4) (Orange color)

The Cluster 5 is also formed by one neighborhood in which there are few restaurants and few options of leisure. Probably there is a good place to open a healthy food restaurant but near to the park or dog run.

```
#cluster 5
toronto_merged.loc[toronto_merged['Cluster Labels'] == 4]
```

	Postal Code	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	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
18	M4N	Central Toronto	Lawrence Park	43.72898	-79.39173	4	Park	Electronics Store	Women's Store	Dog Run	Fast Food Restaurant	Farmers Market	Farm	Falafel Restaurant

Fig. 15 Neighborhoods and venues of Cluster 5

Conclusions

After preparing, modelling and analyzing the data, it can be concluded that the k-mean algorithm of the machine learning approach is very useful to find segmentation in real life application that involves studying the market, competitors, target, segments, etc. We found there are 2 neighborhoods(cluster 4 and 2) with great potential to open a healthy food and wellness restaurant or facilities and two more neighborhoods in which, even with a lot of competitors, the target customers can be found because of the great population density (cluster 1 and 5).

References:

[1] https://www.businesswire.com/news/home/20200317005775/en/Health-Wellness-Food-Market-2020-2024Increasing-Adoption-Healthy