K-Means Cluster Comparison of the number of Sports Teams in two Cities of the United States and Canada.

Aaron Smith. Coursera Student.

Applied Data Science Capstone, Capstone Project - The Battle of Neighbourhoods.
Analysis of two cities with high number of sporting venues.

1. Introduction

I will compare the number of sport teams within the US and Canada to see which two cities (in each country) has the most sports teams.

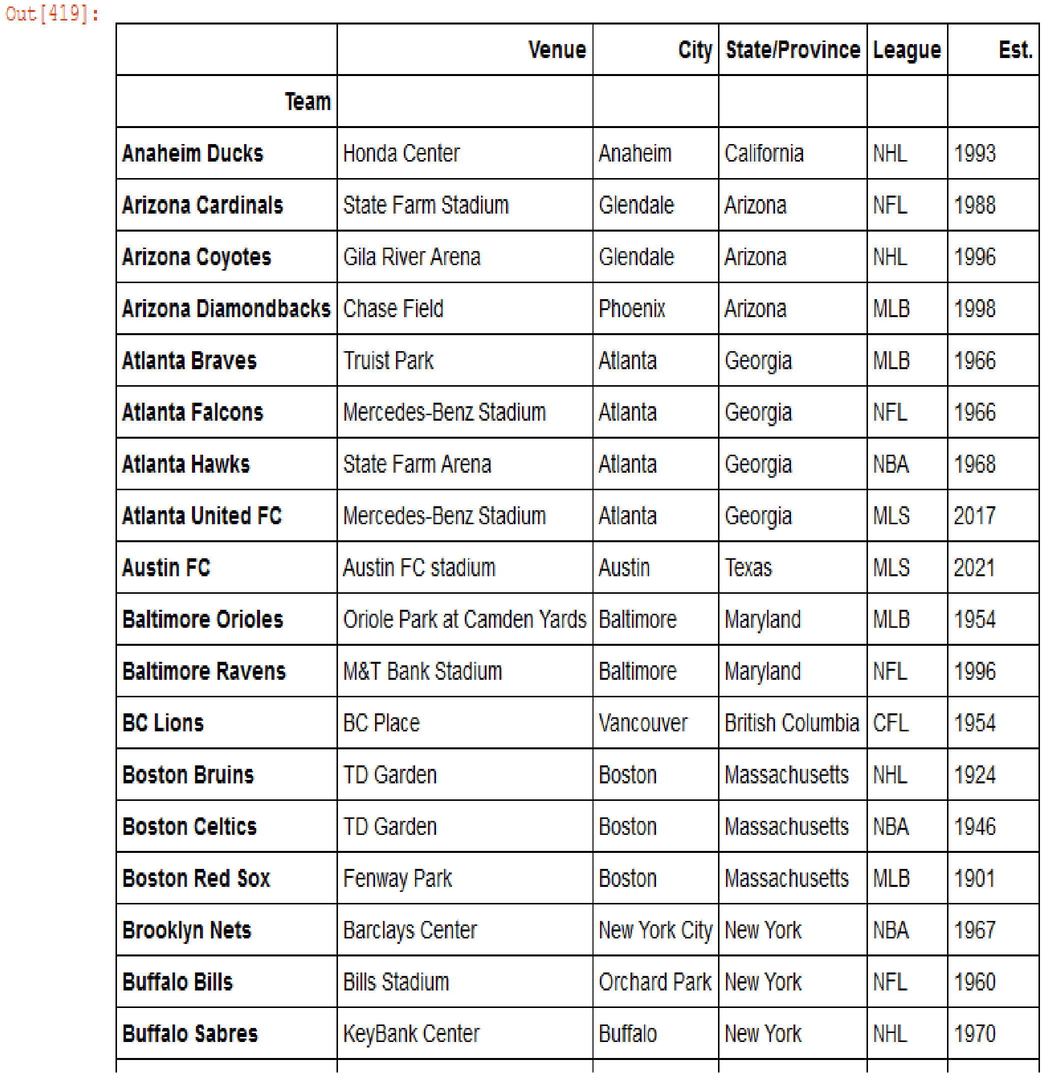
The second part of this assignment is to discuss a business problem regarding someone with wishes to open a restaurant. Sporting venues are popular with fans and tourists both during match day and non-playing days; I would therefore assume that a restaurant operating within the close proximity of the stadium would be beneficial for the owners.

I will use the K-Means algorithm Clustering attribute to cluster the objects (venues) into different groups (sporting and food/drink). Moreover, this procedure will help identify areas where upon a restaurant venue will be most profitable.

Contents.

2. Data Description

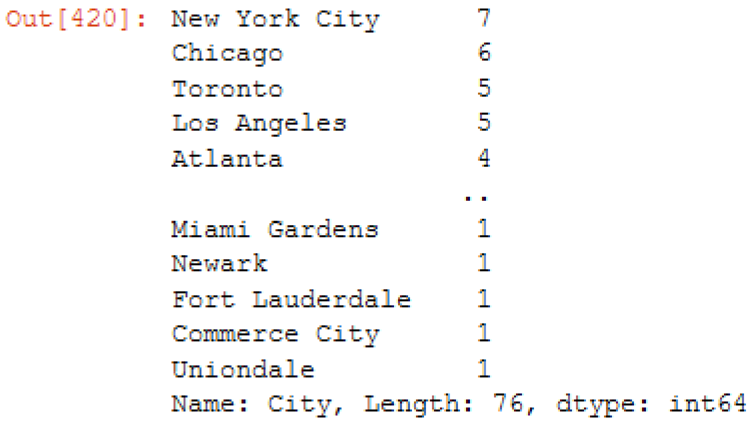
I have used a page from Wiki data to upload a dataset of major professional sports teams of the United States and Canada[1]. Moreover, the Wiki dataset gave me information on name of Team, Venue, City and State. Below is the example of the dataset which I used to upload the data.



Out[419]:

	Venue	City	State/Province	League	Est.
Team					
Anaheim Ducks	Honda Center	Anaheim	California	NHL	1993
Arizona Cardinals	State Farm Stadium	Glendale	Arizona	NFL	1988
Arizona Coyotes	Gila River Arena	Glendale	Arizona	NHL	1996
Arizona Diamondbacks	Chase Field	Phoenix	Arizona	MLB	1998
Atlanta Braves	Truist Park	Atlanta	Georgia	MLB	1966
Atlanta Falcons	Mercedes-Benz Stadium	Atlanta	Georgia	NFL	1966
Atlanta Hawks	State Farm Arena	Atlanta	Georgia	NBA	1968
Atlanta United FC	Mercedes-Benz Stadium	Atlanta	Georgia	MLS	2017
Austin FC	Austin FC stadium	Austin	Texas	MLS	2021
Baltimore Orioles	Oriole Park at Camden Yards	Baltimore	Maryland	MLB	1954
Baltimore Ravens	M&T Bank Stadium	Baltimore	Maryland	NFL	1996
BC Lions	BC Place	Vancouver	British Columbia	CFL	1954
Boston Bruins	TD Garden	Boston	Massachusetts	NHL	1924
Boston Celtics	TD Garden	Boston	Massachusetts	NBA	1946
Boston Red Sox	Fenway Park	Boston	Massachusetts	MLB	1901
Brooklyn Nets	Barclays Center	New York City	New York	NBA	1967
Buffalo Bills	Bills Stadium	Orchard Park	New York	NFL	1960
Buffalo Sabres	KeyBank Center	Buffalo	New York	NHL	1970

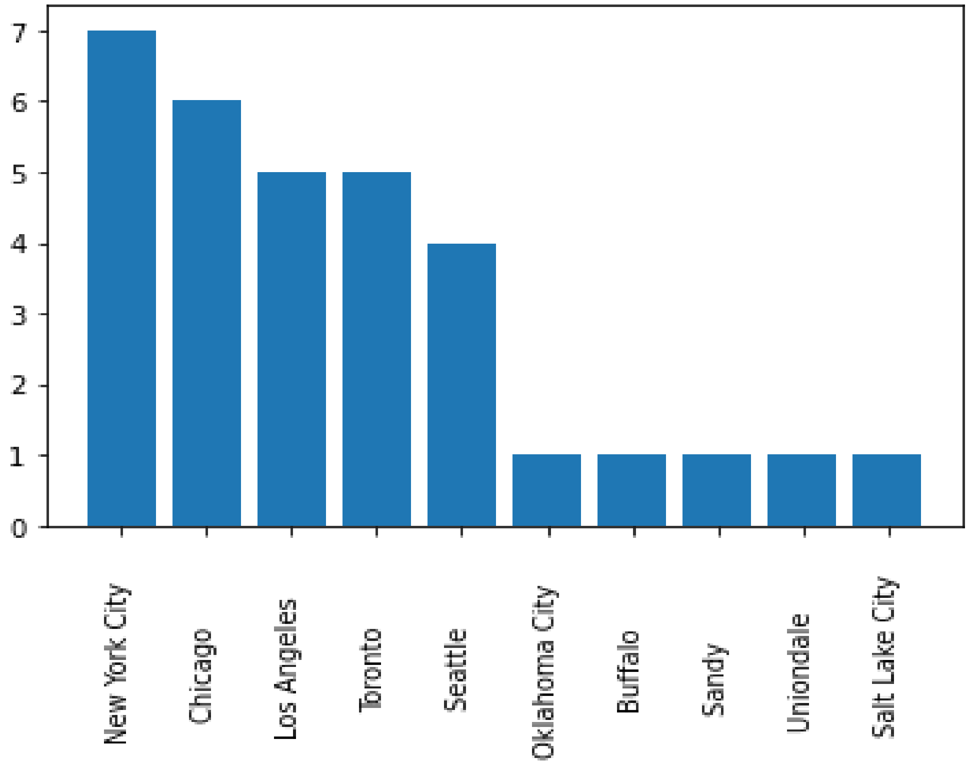
I proceeded to extract the number of cities (of both the USA and Canada) with the highest number of sports teams. I found that the two cities with the most sports teams are New York City (USA) and Toronto (Canada).



```
Out[420]: New York City       7
          Chicago             6
          Toronto             5
          Los Angeles         5
          Atlanta             4
                             ..
          Miami Gardens       1
          Newark              1
          Fort Lauderdale     1
          Commerce City       1
          Uniondale           1
          Name: City, Length: 76, dtype: int64
```

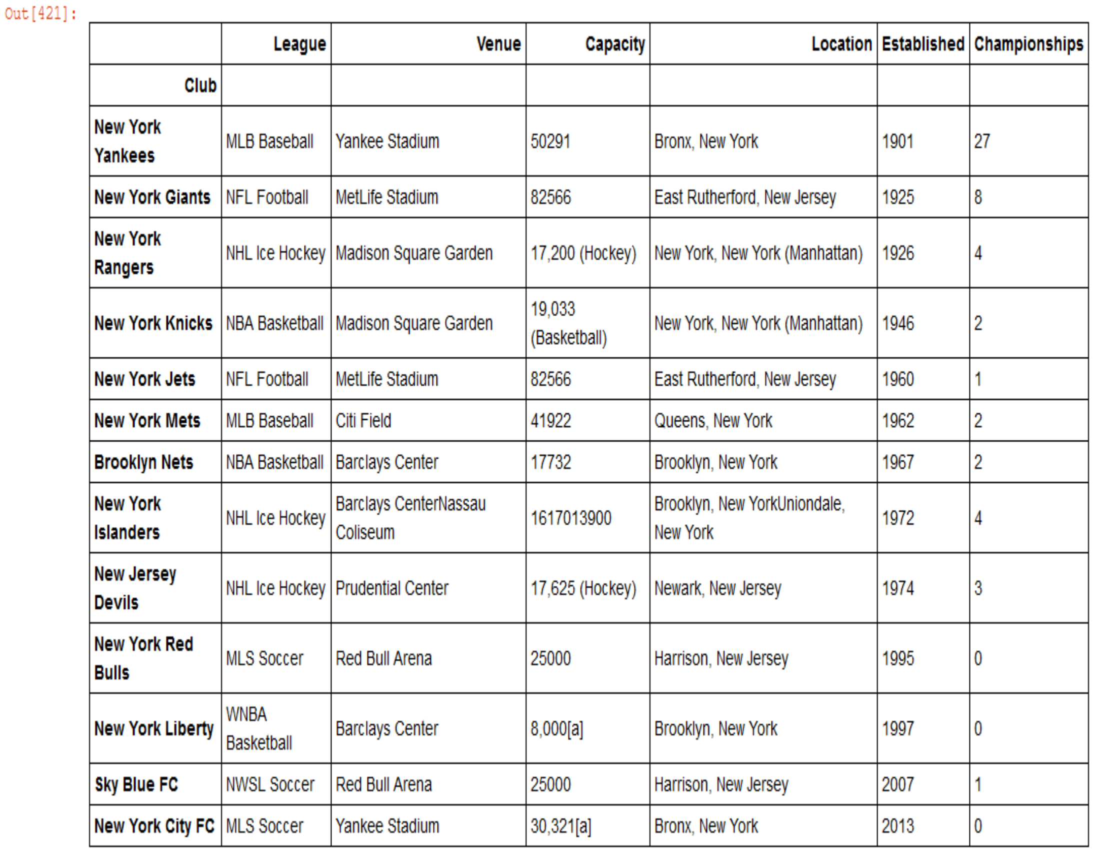
Below is a graph of the dataset to highlight which cities show the highest number of cities with a sports team.

- Bar Chart Example.



Using the data above *Bar Chart Example I can attain that the cities (within different counties) with the most sports teams are New York City (7 sports teams) and Toronto (5 sports teams). I again used Wiki datasets to scrape data from New York City area sports teams[2] and Toronto area sports teams[3] to confirm and detail information (venues and clubs) of the two cities.

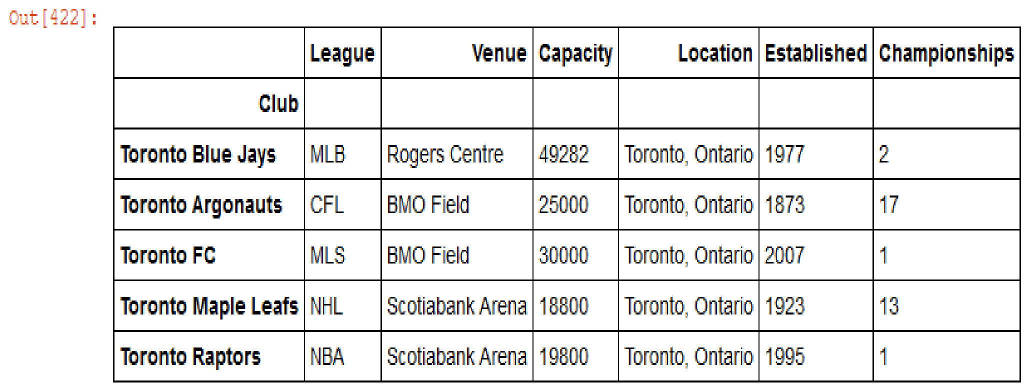
New York City



Out[421]:

	League	Venue	Capacity	Location	Established	Championships
Club						
New York Yankees	MLB Baseball	Yankee Stadium	50291	Bronx, New York	1901	27
New York Giants	NFL Football	MetLife Stadium	82566	East Rutherford, New Jersey	1925	8
New York Rangers	NHL Ice Hockey	Madison Square Garden	17,200 (Hockey)	New York, New York (Manhattan)	1926	4
New York Knicks	NBA Basketball	Madison Square Garden	19,033 (Basketball)	New York, New York (Manhattan)	1946	2
New York Jets	NFL Football	MetLife Stadium	82566	East Rutherford, New Jersey	1960	1
New York Mets	MLB Baseball	Citi Field	41922	Queens, New York	1962	2
Brooklyn Nets	NBA Basketball	Barclays Center	17732	Brooklyn, New York	1967	2
New York Islanders	NHL Ice Hockey	Barclays CenterNassau Coliseum	1617013900	Brooklyn, New YorkUniondale, New York	1972	4
New Jersey Devils	NHL Ice Hockey	Prudential Center	17,625 (Hockey)	Newark, New Jersey	1974	3
New York Red Bulls	MLS Soccer	Red Bull Arena	25000	Harrison, New Jersey	1995	0
New York Liberty	WNBA Basketball	Barclays Center	8,000[a]	Brooklyn, New York	1997	0
Sky Blue FC	NWSL Soccer	Red Bull Arena	25000	Harrison, New Jersey	2007	1
New York City FC	MLS Soccer	Yankee Stadium	30,321[a]	Bronx, New York	2013	0

Toronto



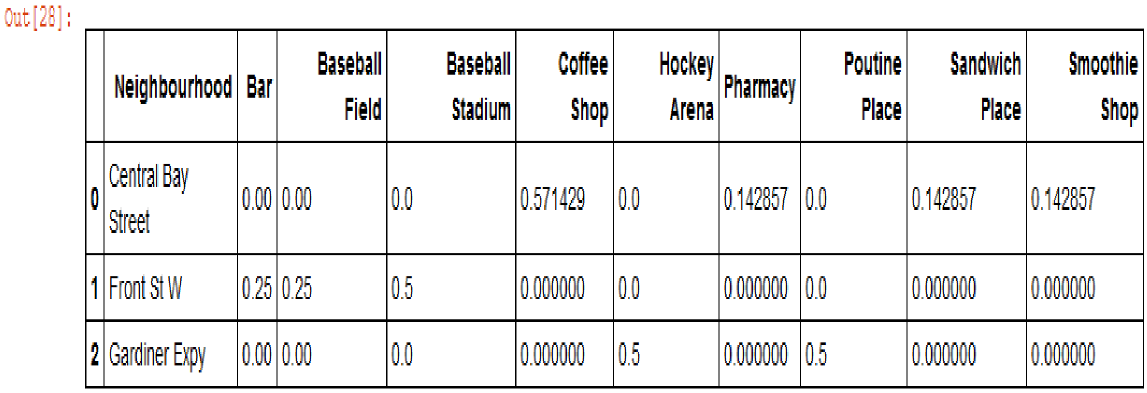
Out[422]:

	League	Venue	Capacity	Location	Established	Championships
Club						
Toronto Blue Jays	MLB	Rogers Centre	49282	Toronto, Ontario	1977	2
Toronto Argonauts	CFL	BMO Field	25000	Toronto, Ontario	1873	17
Toronto FC	MLS	BMO Field	30000	Toronto, Ontario	2007	1
Toronto Maple Leafs	NHL	Scotiabank Arena	18800	Toronto, Ontario	1923	13
Toronto Raptors	NBA	Scotiabank Arena	19800	Toronto, Ontario	1995	1

3. Methodology Section.

I will now use the data to analyse the results to produce a visual map of the locations (both sporting venues and food/drink service industry venues). In order to view location of teams within their chosen city I had to manually update a csv file to include Latitude and Longitude information this was based on venue data. Moreover, the Latitude and Longitude information was found on google maps whereupon I entered the venue name and sourced the lat and long details.

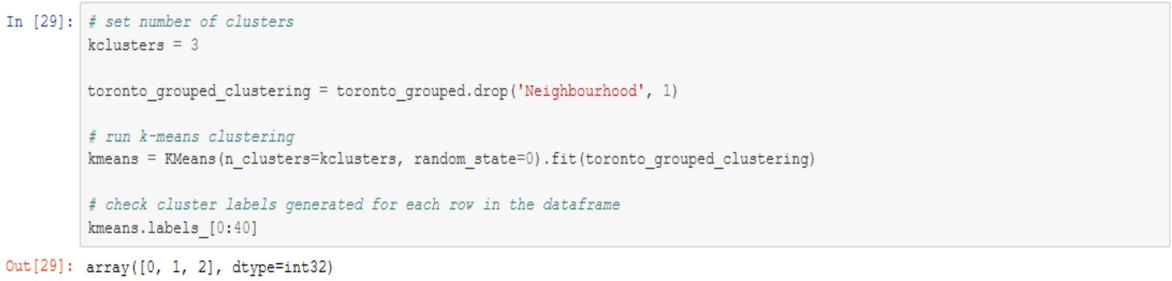
For the venues around the stadium/sporting venues I used the foursquare API to gather data, the radius I entered was 150 meters with a limit of 100 venues. The data results gave me twelve venues within the area. Once the result data was finalised, I had to process the code using 'One Hot Encoding' [ref]. I wrote the K-Means cluster algorithm with three clusters as this was the same number of neighbourhoods I was working with. I have identified three areas within the city of Toronto that I feel would make an excellent business location for the restaurant, all three locations are within proximity to the stadiums and are also close to public transport.



Out[28]:

	Neighbourhood	Bar	Baseball Field	Baseball Stadium	Coffee Shop	Hockey Arena	Pharmacy	Poutine Place	Sandwich Place	Smoothie Shop
0	Central Bay Street	0.00	0.00	0.0	0.571429	0.0	0.142857	0.0	0.142857	0.142857
1	Front St W	0.25	0.25	0.5	0.000000	0.0	0.000000	0.0	0.000000	0.000000
2	Gardiner Expy	0.00	0.00	0.0	0.000000	0.5	0.000000	0.5	0.000000	0.000000

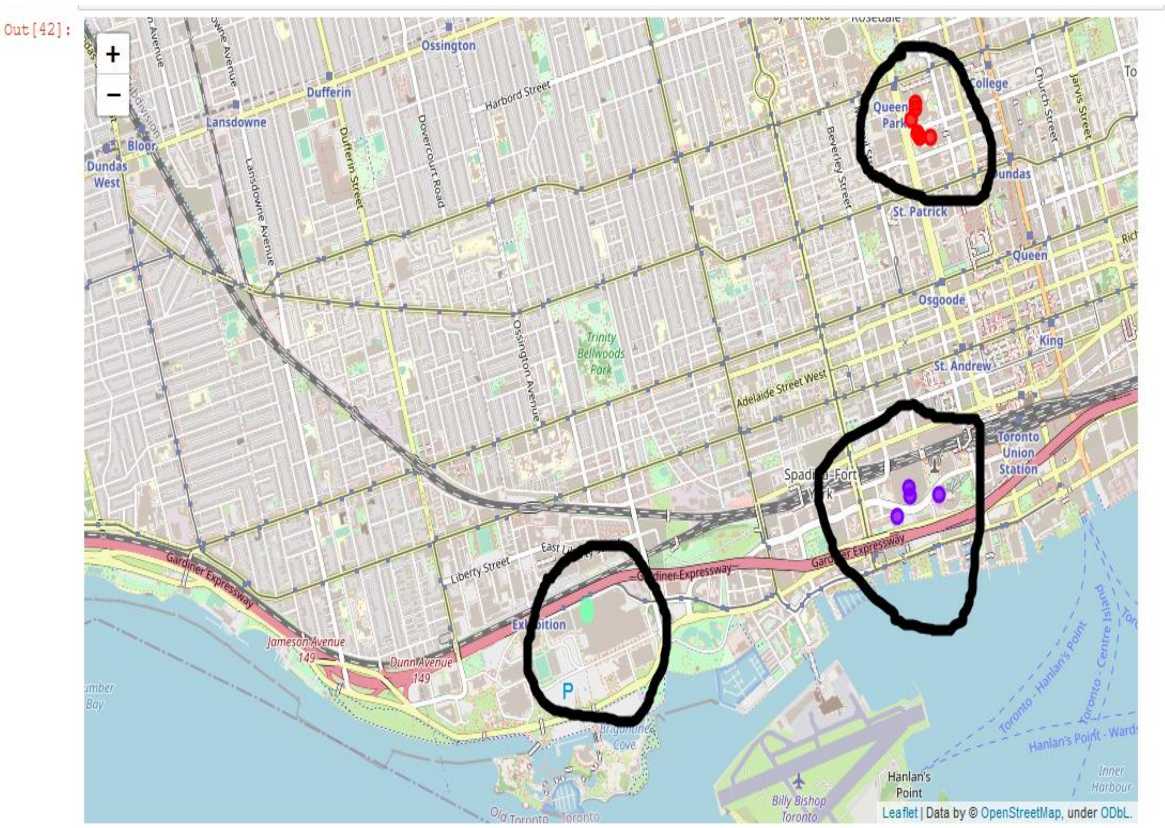
The resulting data frame [above] is an example of the One Hot Encoding process which includes the venues that are situated in the neighbourhoods. The data is now stored for use with the cluster modelling algorithm.



```
In [29]: # set number of clusters
         kclusters = 3

         toronto_grouped_clustering = toronto_grouped.drop('Neighbourhood', 1)

         # run k-means clustering
         kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(toronto_grouped_clustering)

         # check cluster labels generated for each row in the dataframe
         kmeans.labels_[0:40]

Out[29]: array([0, 1, 2], dtype=int32)
```

4. Results Section

In the results section I want to print the results of the K-Means algorithm to show if my hypothesis was correct. The result of the final map showed only a small fraction of food/drink venues within close range to the sporting venues therefore any new restaurant opening within this area would be a success.



Out[42]:

Viewing the map after the results I could see that the 'Gardiner Expy' neighbourhood would be perfect for the restaurant with only two food/drink venues in this area.

5. Discussion Section

Hitherto, there does not seem to be a restaurant situated in the area, only coffee and snack shops. Although Toronto is small geographically the city has a very high number of tourists who visit the area and with the stadiums so close this will only help business in the long-term.

After exploring the area using the maps, I feel it will be a good opportunity for any restaurant owners to open a business in this area. Competition from the stadium(s) would have little or no effect on business during non-playing days, I would also add the most popular places are the Coffee shops and public transport which would be beneficial.

6. Conclusion Section

My conclusion is that Toronto seems to be a very popular place among tourists, the only doubt I have is I could have sought data on the ages of the people visiting Toronto. Moreover, if the restaurant is aimed towards the elder generation maybe having the business situated near a sports area could damage business?

References.

[1]

url =
https://en.wikipedia.org/wiki/Major_professional_sports_teams_of_the_United_States_and_Canada

(Accessed: 21.01.2021)

[2]

url =
https://en.wikipedia.org/wiki/List_of_New_York_City_metropolitan_area_sports_teams
(Accessed: 21.01.2021)

[3]

url =
https://github.com/aaron1986/Coursera_Capstone/blob/master/Canada.csv
(Accessed: 21.01.2021)

[ref]

‘One Hot Encoding’. (https://machinelearningmastery.com/) (Accessed: 21.01.2021) For categorical variables where no such ordinal relationship exists, the integer encoding is not enough.

In fact, using this encoding and allowing the model to assume a natural ordering between categories may result in poor performance or unexpected results (predictions halfway between categories).