Deciding where open a Baseball shop in Toronto

Marcus Lopes

March 30, 2019

Table of Contents

1. Introduction

1.1. Background

The neighborhoods in a city are very diversificated, but we can cluster them in a way that are possible to identify patterns in what venues are most common in each group of similar neighborhoods.

1.2. Problem

This will be especially important to us, because we are contracted to determine where should be open a baseball's stuff shop in the city of Toronto. So we will identify where are the neighborhoods that will have a stablished potencial consumers based in the atual venues nearest.

1.3. Interest

This will be important to give the entrepreneur the best chances to have succed with him shop and have a better return for him investment.

2. Data

The data we will use to solve the problem presented above will be the table contained in the Wikipedia page https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M in wich we will extract the postal codes, boroughs and neighborhoods of Toronto, Canada. And with geospatial data provided in the csv http://cocl.us/Geospatial_data from the Capstone Project course of Coursera and IBM we will append the latitude and longitude from postal codes to our table previously extracted. Finally we will use the location data from Foursquare to identify wich categories of venues are more contained in each neighborhood to then clustering this data and so identify the principals caregories of venues contained in each neighborhood's cluster. This will permit make a decision about where open the baseball's stuff store.

3. Methodology

Segmenting and Clustering Neighborhoods in Toronto.

3.1. Catching neighborhoods data from source

Here we extract table from html...

We will use just the table containing the postal codes, boroughs and neighborhoods data from wikipedia page.

... and prepare data.

We need to transform html data in a data frame with organized informations, so we will can work better with the data.

	PostalCode	Borough	Neighborhood
0	M1B	Scarborough	Rouge, Malvern
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union
2	M1E	Scarborough	Guildwood, Morningside, West Hill
3	M1G	Scarborough	Woburn
4	M1H	Scarborough	Cedarbrae
5	M1J	Scarborough	Scarborough Village
6	M1K	Scarborough	East Birchmount Park, Ionview, Kennedy Park
7	M1L	Scarborough	Clairlea, Golden Mile, Oakridge
8	M1M	Scarborough	Cliffcrest, Cliffside, Scarborough Village West
9	M1N	Scarborough	Birch Cliff, Cliffside West
10	M1P	Scarborough	Dorset Park, Scarborough Town Centre, Wexford ...
11	M1R	Scarborough	Maryvale, Wexford
12	M1S	Scarborough	Agincourt
13	M1T	Scarborough	Clarks Corners, Sullivan, Tam O'Shanter
14	M1V	Scarborough	Agincourt North, L'Amoreaux East, Milliken, St...
15	M1W	Scarborough	L'Amoreaux West

. . .

Table 1. Postal codes, boroughs and neighborhoods of Toronto (size 103 rows × 3 columns).

3.2. Appending Latitude/Longitude to neighborhoods data from Toronto

Here we get the latitude and longitude data for the postal codes from our data source (http://cocl.us/Geospatial_data).

	Postal Code	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917
4	M1H	43.773136	-79.239476
5	M1J	43.744734	-79.239476
6	M1K	43.727929	-79.262029
7	M1L	43.711112	-79.284577
8	M1M	43.716316	-79.239476
9	M1N	43.692657	-79.264848
10	M1P	43.757410	-79.273304
11	M1R	43.750072	-79.295849
12	M1S	43.794200	-79.262029
13	M1T	43.781638	-79.304302
14	M1V	43.815252	-79.284577
15	M1W	43.799525	-79.318389

. . .

Table 2. Postal codes with respective latitudes and longitudes (size 103 rows × 3 columns).

And here we append the latitude and longitude data to corresponding postal codes.

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Rouge, Malvern	43.806686	-79.194353
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476
5	M1J	Scarborough	Scarborough Village	43.744734	-79.239476
6	M1K	Scarborough	East Birchmount Park, Ionview, Kennedy Park	43.727929	-79.262029
7	M1L	Scarborough	Clairlea, Golden Mile, Oakridge	43.711112	-79.284577
8	M1M	Scarborough	Cliffcrest, Cliffside, Scarborough Village West	43.716316	-79.239476
9	M1N	Scarborough	Birch Cliff, Cliffside West	43.692657	-79.264848
10	M1P	Scarborough	Dorset Park, Scarborough Town Centre, Wexford ...	43.757410	-79.273304
11	M1R	Scarborough	Maryvale, Wexford	43.750072	-79.295849
12	M1S	Scarborough	Agincourt	43.794200	-79.262029
13	M1T	Scarborough	Clarks Corners, Sullivan, Tam O'Shanter	43.781638	-79.304302
14	M1V	Scarborough	Agincourt North, L'Amoreaux East, Milliken, St...	43.815252	-79.284577
15	M1W	Scarborough	L'Amoreaux West	43.799525	-79.318389

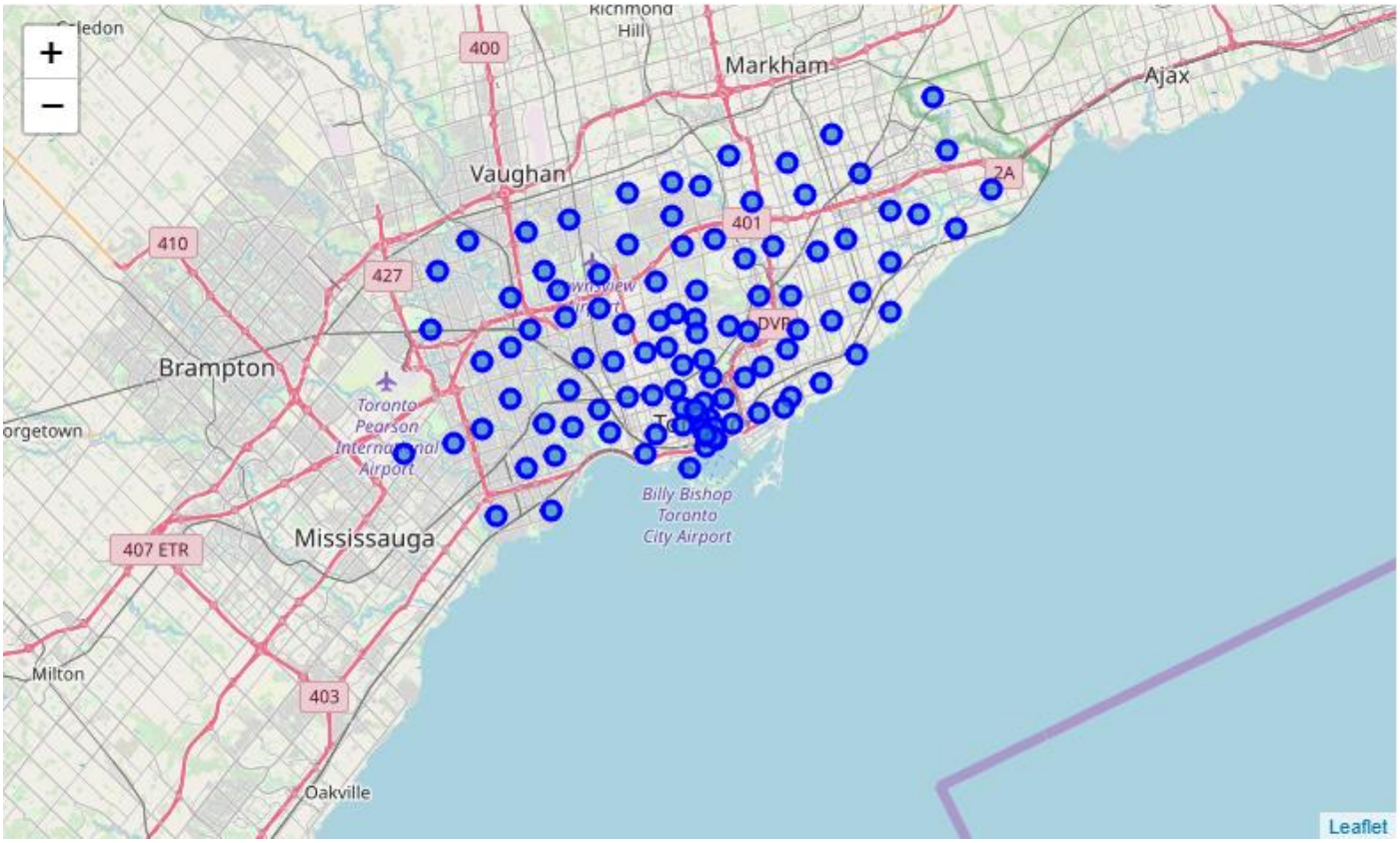
. . .

Table 3. The complete list with postal codes, boroughs, neighborhoods, latitudes and longitudes (size 103 rows × 5 columns).

3.3. Exploring the neighborhoods in Toronto

We will visualize the neighborhoods in a map of Toronto. For this, first we use geopy library to get the latitude and longitude values of Toronto. The geograpical coordinate of Toronto are 43.653963, -79.387207.

And then we create a map of Toronto with neighborhoods superimposed on top.



Map 1. The neighborhoods of Toronto.

3.4. Now we will use Foursquare to request a exploratory search of venues in the neighborhoods.

First we define Foursquare Credentials and Version and then we explore the first neighborhood in our dataframe.

Here we get the first neighborhood's name, that is “Rouge, Malvern”. So we get the neighborhood's latitude and longitude values. Latitude and longitude values of Rouge, Malvern are 43.806686299999996, -79.19435340000001. And now, we will get the top 100 venues that are in “Rouge, Malvern” within a radius of 500 meters.

First, we create the GET request URL. And then send the GET request and examine the resutls.
Now we are ready to clean the json result and structure it into the *pandas* dataframe below.

	name	categories	lat	lng
0	Wendy's	Fast Food Restaurant	43.807448	-79.199056

Table 4. The venues in “Rouge, Malvern” neighborhood.

And how many venues were returned by Foursquare? Just 1 venue were returned by Foursquare.

3.5. Explore all neighborhoods in Toronto

Here we create a function to repeat the same process to all the neighborhoods in Toronto, resulting in a dataframe with 2235 rows and 7 columns with each row representing one venue with respective neighborhood, latitude, longitude and category, like the five first rows below:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Rouge, Malvern	43.806686	-79.194353	Wendy's	43.807448	-79.199056	Fast Food Restaurant
1	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497	Royal Canadian Legion	43.782533	-79.163085	Bar
2	Guildwood, Morningside, West Hill	43.763573	-79.188711	Swiss Chalet Rotisserie & Grill	43.767697	-79.189914	Pizza Place
3	Guildwood, Morningside, West Hill	43.763573	-79.188711	G & G Electronics	43.765309	-79.191537	Electronics Store
4	Guildwood, Morningside, West Hill	43.763573	-79.188711	Marina Spa	43.766000	-79.191000	Spa

Table 5. List of venues in the neighborhoods of Toronto

Then we check how many venues were returned for each neighborhood:

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Adelaide, King, Richmond	100	100	100	100	100	100
Agincourt	4	4	4	4	4	4
Agincourt North, L'Amoreaux East, Milliken, Steeles East	2	2	2	2	2	2
Albion Gardens, Beaumond Heights, Humbergate, Jamestown, Mount Olive, Silverstone, South Steeles, Thistletown	10	10	10	10	10	10
Alderwood, Long Branch	10	10	10	10	10	10
Bathurst Manor, Downsview North, Wilson Heights	18	18	18	18	18	18
Bayview Village	4	4	4	4	4	4
Bedford Park, Lawrence Manor East	24	24	24	24	24	24
Berczy Park	55	55	55	55	55	55
Birch Cliff, Cliffside West	4	4	4	4	4	4

. . .

Table 6. Count of venues in each neighborhood.

In the “Venue” colum of table above we observe that some neighborhood reach to limit of 100 venues in exploratory seach requested in Foursquare, but in other hand some neighborhoods have little quantity of venues.

Now we will find out how many unique categories can be curated from all the returned venues. There are 273 uniques categories.

3.6. Analyze Each Neighborhood

For each neighborhood we use one hot encoding for determin wich categories of venues are most common. We have 2235 venues in the neighborhoods and 272 categories of venues. So we group data by neighborhood and by taking the mean of the frequency of occurrence of each category. After group by neighborhood we confirm

the new size of data frame is 98 neighborhoods with venues and 272 unique categories. Below is each neighborhood along with the top 5 most common venues:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Adelaide, King, Richmond	Coffee Shop	Steakhouse	Thai Restaurant	Bar	Café
1	Agincourt	Lounge	Sandwich Place	Breakfast Spot	Skating Rink	Drugstore
2	Agincourt North, L'Amoreaux East, Milliken, St...	Playground	Park	Yoga Studio	Drugstore	Dim Sum Restaurant
3	Albion Gardens, Beaumond Heights, Humbergate, ...	Grocery Store	Pizza Place	Fried Chicken Joint	Coffee Shop	Sandwich Place
4	Alderwood, Long Branch	Pizza Place	Skating Rink	Dance Studio	Pharmacy	Coffee Shop

...

Table 7. The top 5 venues for each neighborhood.

3.7. Clustering neighborhoods using the k-means methodology

We will use a machine learning algorithm to cluster the neighborhoods based in the categories of venues contained in each of them. This way we expect to better understand how similar are the neighborhoods and agroup them to form five clusters, so we can take a look and decide latter where a Baseball shop should be open to have the most likely public based on categories of venues more frequently in a cluster.

So we run *k*-means to cluster the neighborhood into 5 clusters and create a new dataframe that includes the cluster as well as the top 10 venues for each neighborhood. Below a example with just the tope 3 venues:

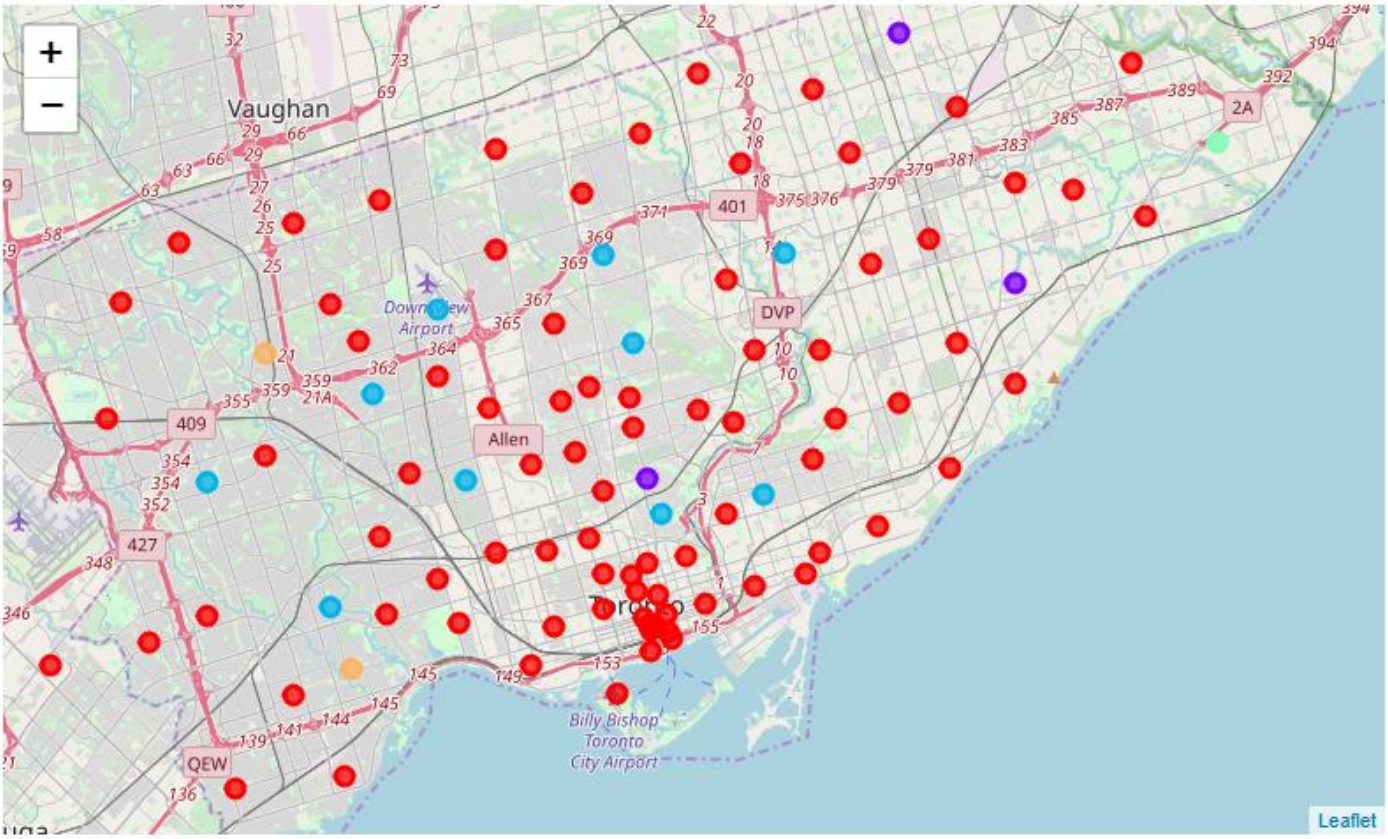
	PostalCode	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
0	M1B	Scarborough	Rouge, Malvern	43.806686	-79.194353	0	Fast Food Restaurant	Dumpling Restaurant	Diner
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497	3	Bar	Yoga Studio	Discount Store
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711	0	Pizza Place	Spa	Electronics Store
3	M1G	Scarborough	Woburn	43.770992	-79.216917	0	Coffee Shop	Korean Restaurant	Yoga Studio
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476	0	Hakka Restaurant	Caribbean Restaurant	Bank

...

Table 8. Our data now with the respective cluster labels.

4. Results

Finally, we will visualize the resulting clusters in Toronto map. This way we can see clearly where are the similar neighborhoods distributed along the city.



Map 2. The neighborhoods of Toronto with the color of respective cluster label.

In the resulting map we can see that are a very large cluster, one medium size and other three are small clusters. In next section we will search for the most common venues categories presented in each cluster to understand the patterns that resulted in this unequal distribution.

5. Discussion

After show the cluster in a map, we need to examine clusters, understand what are the similarities between each group. Now, you can see each cluster and the most common venue categories contained in the neighborhoods that distinguish each cluster.

Cluster 1

	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Rouge, Malvern	0	Fast Food Restaurant	Dumpling Restaurant	Diner	Discount Store	Dog Run
2	Guildwood, Morningside, West Hill	0	Pizza Place	Spa	Electronics Store	Breakfast Spot	Rental Car Location
3	Woburn	0	Coffee Shop	Korean Restaurant	Yoga Studio	Dumpling Restaurant	Discount Store
4	Cedarbrae	0	Hakka Restaurant	Caribbean Restaurant	Bank	Bakery	Fried Chicken Joint
6	East Birchmount Park, Ionview, Kennedy Park	0	Discount Store	Chinese Restaurant	Bus Station	Department Store	Coffee Shop

...

Table 9. Example of five neighborhoods of Cluster 1 with respective top 5 common venues.

Cluster 2

	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
5	Scarborough Village	1	Playground	Drugstore	Dim Sum Restaurant	Diner	Discount Store
14	Agincourt North, L'Amoreaux East, Milliken, St...	1	Playground	Park	Yoga Studio	Drugstore	Dim Sum Restaurant
48	Moore Park, Summerhill East	1	Playground	Gym	Drugstore	Dim Sum Restaurant	Diner

...

Table 10. The 3 neighborhoods of Cluster 2 with respective top 5 common venues.

Cluster 3

	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
23	York Mills West	2	Park	Bank	Yoga Studio	Dumpling Restaurant	Discount Store
25	Parkwoods	2	Fast Food Restaurant	Food & Drink Shop	Park	Diner	Discount Store
30	CFB Toronto, Downsview East	2	Bus Stop	Airport	Park	Yoga Studio	Dumpling Restaurant
40	East Toronto	2	Park	Coffee Shop	Convenience Store	Yoga Studio	Dumpling Restaurant
44	Lawrence Park	2	Gym / Fitness Center	Park	Bus Line	Swim School	Yoga Studio

...

Table 11. Example of five neighborhoods of Cluster 3 with respective top 5 common venues.

Cluster 4

	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
1	Highland Creek, Rouge Hill, Port Union	3	Bar	Yoga Studio	Discount Store	Dog Run	Doner Restaurant

...

Table 12. The unique neighborhood of Cluster 4 with respective top 5 common venues.

Cluster 5

	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
91	Humber Bay, King's Mill Park, Kingsway Park So...	4	Baseball Field	Pool	Yoga Studio	Diner	Discount Store
97	Emery, Humberlea	4	Baseball Field	Paper / Office Supplies Store	Yoga Studio	Discount Store	Dog Run

...

Table 13. The 2 neighborhoods of Cluster 5 with respective top 5 common venues.

Looking at results we observe that the cluster number 1 is the largest of them, so is hard to find particular venues that determines the link to this group. In the others clusters we have less neighborhoods and is possible see more clearly what venues is more common between the group. The cluster number 2 is distinguished by playgrounds and drugstore. In the cluster number 3 is more common to a neighborhood have park, yoga studio, discount store and dog run. Having just one neighborhood in the cluster number 4, the most common venue is bars. So, that neighborhood have too some similarities with the pattern of cluster 3, but strong distinguished by the bars, justifying be alone in a different group. And least cluster, number 5, we have the baseball field as most common venue of the group and second most common is yoga studio.

By this results and based in our need to recommend where open a Baseball shop, we can properly say that the better area in the city of Toronto to open this shop is in any neighborhood of cluster number 5.

Conclusion

In this report we can see that clustering neighborhoods of a city considering the most common venues categories is a very powerful tool to understanding how similar or dissimilar are the areas along a location of interest. For this we need to have the data about neighborhoods of a city and them latitude and longitute. With this we can call the Foursquare API to make a exploratory seach for venues categories more contained in each neighborhood. In a organized data we then use the k-means to grouping the neighborhoods based in their similarities. Based on our observations is possible see the city in a new way and so we have a easy view of

where is the locations better to open a Baseball shop. As we can see the answer is cluster number 5 because there are baseball fieds as most common venues and this is a attraction for public that are potential clients for the new Baseball shop.

So we recommend the Baseball shop open in the area of cluster number 5 to have better financial expectatives, resulting in more salles and consequently revenues.

Our job permit us say that for a more accurate result is possible search for data of other Baseball shops and make linear regressions with the revenues of this shops and the categories of venues more contained in respective neighborhoods of this shops. This way would tell what venues categories contribute more on less in succes of a Baseball shop.