Determining the Class of Districts in Toronto using Common Business Venues

September 4, 2021

1. Introduction

Districts in prominent cities are always seeking to attract investors to their neighborhoods. These investors often need to know if their proposed business type is suitable or associated with the class of the district of interest. Certain district class may have common factors that may distinguish their class to attract more of a particular type of service over other types. Hence it becomes important to study the relationship between the class categories and the common types of business venues found in a city's districts. Districts are mostly categorized into high, medium and low classes depending on relative cost of property in the districts. This study seeks to determine the extent to which the common business categories found in a district, defines the class type of the district and the accuracy of machine learning tools in making the predictions. It also aims to highlight the most common business categories associated with high, medium and low class areas.

Toronto city, a prominent city with available and reliable data sources of its neighborhoods, was used as a sample case study. Toronto city has 35 neighborhood groups with distinct district numbers. The cost of housing index in each district was used to establish the class categories. From location data set available, the distinct business venues and the frequency of visits was extracted for each district.

1.1 Business Problem

The problem is, a would-be investor needs to determine which types of business groups are mostly found in high, middle, and low class districts. This knowledge is critical to the potential investor, planning in consideration of his/her budget.

The broad objective of this assignment is to establish if identified groups of most common business categories are able to define the class of the districts.

The specific objectives includes identifying which types of businesses are most found in high, middle and low class districts.

2. Source and Description of Data.

The cost of housing index of Toronto city was extracted from Toronto Regional Real Estate Board as culled in Storey's real estate site https://storeys.com/median-home-price-35-toronto-neighbourhoods/. The specific variable used to represent Housing cost index was the Median Detached House Prices of Districts given in Canadian dollars

The coordinates of the districts was extracted using geolocator on the geopy python library. The Longitude and Latitudes were obtained for each Districts

The coordinates was applied for VENUES data from Foursquare data set API at 1000m radius - being the approximate radius of the smallest district in Toronto City, C10, to avoid overlap of venues returned (Toronto City - Wikipedia).

The various districts in Toronto city categorized by their district numbers was sourced from Toronto city Wikipedia site. https://en.wikipedia.org/wiki/Toronto

3. Methodology

In determining the relationships between the unique business venue categories and the class categories of the district, descriptive statistical method was applied. The number of the various unique venue categories found in each district was used as the independent variable. The dependent variable is the class category of the district. This is as shown in table 3.0

Table 3.0: The Dependent and Independent Variables and their Proxies

	Variable functions	Variables	Proxies
0	Dependent	Class of District	Median Detached House Price Index
1	Independent	Common Business Categories	Count of Unique Venue Categories

Source: Author's conception.

Inferential statistics, of histogram, box plot and the describe function in pandas data frame was used on the dependent variable in order to enable the categorization of the data into high, medium, and low classes.

To analyze the variables Python libraries were imported accordingly: Pandas, numpy, sklearn libraries to enable data wrangling and result testing; requests to enable data requests; geopy to convert addresses into latitude and longitude values; json to handle json files; matplotlib and folium to enable visualization.

The venue categories of each district was obtained from foursquare API dataset requests. These categories was summed up and the average was obtained for each unique business category. Using Onehot encoding, the venue categories were organized into a matrix.

Two machine learning algorithms were used to train the matrix.

1. K-means cluster algorithm

Reason - This was used in order to determine if groups which have not been explicitly labeled in the venue category data can define the district class category data

2. Support Vector Machine

Reason - This was used in order to determine if supervised learning will give better definition of the dependent variables especially considering the high dimension of the independent variable.

K-means cluster algorithm was used to group the districts into cluster groups and the best fit group for the district classes was used as a prediction for each class.

Support vector Machine was used by splitting the business categories and House price district classes into training and testing sets.

f1-score model was used to test the accuracy of the two predictions because it conveys the balance between the precision and the recall

3.1 Data Presentation

The master data has the main components of District's Number, Neighborhoods, Median House Price, Latitude and Longitude information of Toronto city. The District's Numbers and their prominent Neighborhoods were extracted form, Toronto Regional Real Estate Board site. The Latitude and Longitude co-ordinate of the neighborhoods in each district were gotten from iterative runs applying geolocator on the geopy python library.

The corresponding Median House Price index were obtained from Storey's real estate site. The variables were used to generate a CSV file stored in GitHub repository for the study. The first 10 of the 35 districts is in table 3.1

Table 3.1 : The First Ten Districts in Toronto City with Housing Price Index and Co-ordinates

	DistrictNo	Neighbourhoods	Med_HousePrice	Latitude	Longitude
0	C01	Downtown Harbourfront	1990000	43.640080	-79.380150
1	C02	The Annex Yorkville South Hill	2262000	43.674682	-79.399256
2	C03	Forest Hill South Oakwood Vaughan Humewood Ced...	1400000	43.682726	-79.438055
3	C04	Bedford Park Lawrence Manor North Toronto	2243000	43.729199	-79.403253
4	C06	North York Clanton Park Bathurst Manor	1252509	43.754326	-79.449117
5	C07	Willowdale West Newtonbrook West	1690000	43.789576	-79.417588
6	C08	Cabbagetown St. Lawrence Market Toronto waterf...	1650000	43.664473	-79.366986
7	C09	Moore Park Rosedale	3352500	43.690388	-79.383297
8	C10	Davisville Village Midtown Toronto Mount Pleasant	1860000	43.697936	-79.397291
9	C11	Leaside Thorncliffe Park Flemingdon Park	2100000	43.709215	-79.341791
10	C12	York Mills St.Andrew-Windfields Bridle Path	4165500	43.744039	-79.406657

Source: Author's Github Repository

To show distribution of the different districts of Toronto City, the geographical coordinates of the city was obtained and plotted on Folium as map_Toronto. Figure 3.1 shows the map of Toronto City and center point of each district. The blue dots represents the 35 districts.

Figure 3.1: The Center Point of the 35 districts in Toronto City

Source: Author's plotting on Folium

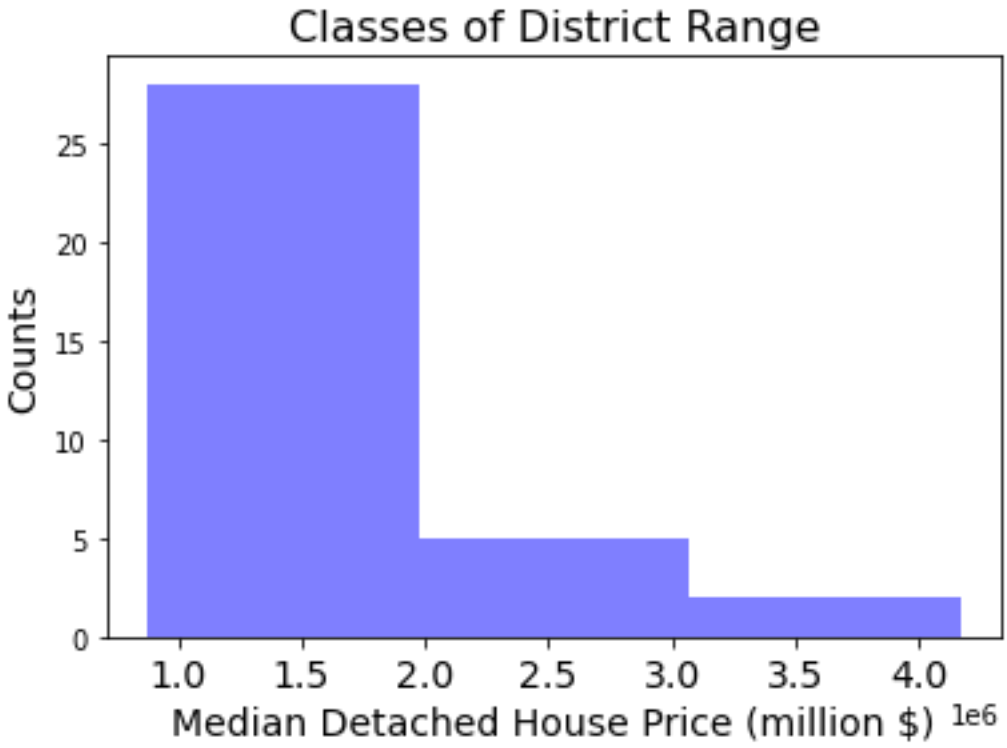
3.2 Data Analysis

Dependent Variable - House Price Analysis

The analysis to determine the proper levels to group the districts into high, middle and lower class area given their house price index was performed by: Sorting the data frame in ascending order by the House price index

The House price index versus the counts of the districts was then displayed in a histogram chart in 3 bins to represent the 3 classes required. Figure 3.2 displays the distribution.

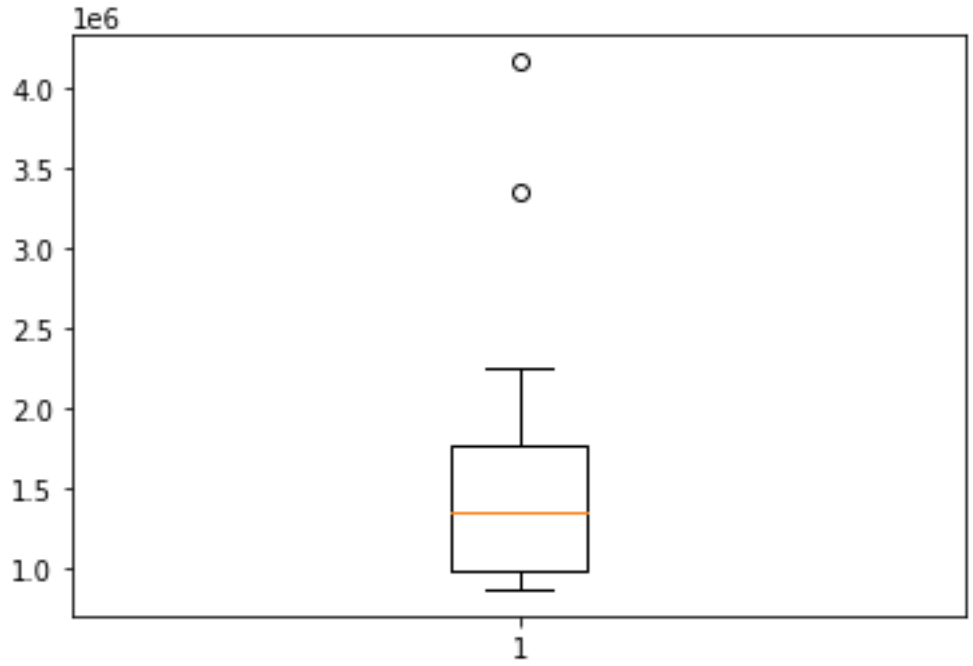
Figure 3.2 Histogram Chart of Median House Prices of districts in Toronto City



Source: Author's analysis on python

The histogram showed that the group counts is skewed towards the lower price class when grouped evenly. Hence there is need to visualize the distribution along the prices and to spot if there are outliers. Hence, a box plot was applied to show the distribution. Figure 3.3 shows the boxplot analysis of the housing price index.

Figure 3.3 Box plot distribution of Median Housing Prices



Source: Author's analysis

The box plot showed that distribution is centered mostly above 1 million and below 2 million and that two outliers are at the high end of the House price index. To get the exact values of the inferential statistics, the describe function was applied as follows

The 25%, 50% and 75% levels were obtained as approximately 0.998, 1.35 and 1.77 million dollar respectively. Hence the ranges were grouped and applied as follows

Lower Class Area = from 875,000 to 998,305.50

Mid Class Area = from 998,305.50 to 1,775,000

High Class Area = from 1,775,000 to 4,165,500.

The resulting classification is appended as a column to the House price data frame and first 5 rows presented in table 3.2

Table 3.2: First five rows of Districts in Toronto City with their Class Categorization

	DistrictNo	Neighbourhoods	Med_HousePrice	Latitude	Longitude	Area Class
0	W03	Keelesdale Eglinton West Rockcliffe Smythe Wes...	875000	43.690158	-79.474998	Lower Class Area
1	E09	Scarborough City Centre Woburn Morningside Ben...	890000	43.782601	-79.204958	Lower Class Area
2	W10	Rexdale Clairville Thistletown - Beaumond Heig...	891500	43.721823	-79.572268	Lower Class Area
3	E04	The Golden Mile Dorset Park Wexford	918000	43.750979	-79.276099	Lower Class Area
4	E10	Rouge (South) Port Union (Centennial Scarborou...	967000	43.768914	-79.187291	Lower Class Area

Source: Author's analysis

Independent Variables- Common Business Venues Analysis

To extract the business venues found in each district, the author's credentials were used to make the request to Foursquare API, and defined as nearby_venues

The latitude, longitude and venue category of the venues within a 1000 meters radius of the application co-ordinate was requested for from the Foursquare API.

The request is applied for each neighborhood in each district in Toronto city to create a data set defined Toronto_venues listing the various venues found in each neighborhood and the venue business type

In total the request returned 1887 venues for the 35 districts in Toronto city. The total number of venue categories found in each District Neighborhoods in alphabetical order was returned as follows. A view of first 5 venues is shown in table 3.3

Table 3.3: First Five Rows of the Total Number of Unique Business Categories in Toronto

	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Neighborhood						
Agincourt Malvern West Milliken	67	67	67	67	67	67
Bedford Park Lawrence Manor North Toronto	58	58	58	58	58	58
Birch Cliff Oakridge Hunt Club Cliffside	11	11	11	11	11	11
Bloor West Village Baby Point The Junction (Junction Area) High Park North	87	87	87	87	87	87
Cabbagetown St. Lawrence Market Toronto waterfront	100	100	100	100	100	100

Source: Author's request from Foursquare

The total number of unique venue category was determined as 246 unique categories.

A new data set was created, defined as Toronto_onehot, to show the one hot encoding neighborhood against the venue categories.

There are 1887 neighborhood venues and 246 distinct business venue categories. In order to group the business venues by districts in line with the objective of the study, the neighborhood venues were grouped and the average of each business category obtained.

The number of top venue was set at 10, to get the 10 most common venue for each district, with the first 5 rows shown in table 3.4

Table 3.4: First Five Rows of the 1st Most Common Venues in Toronto City Districts

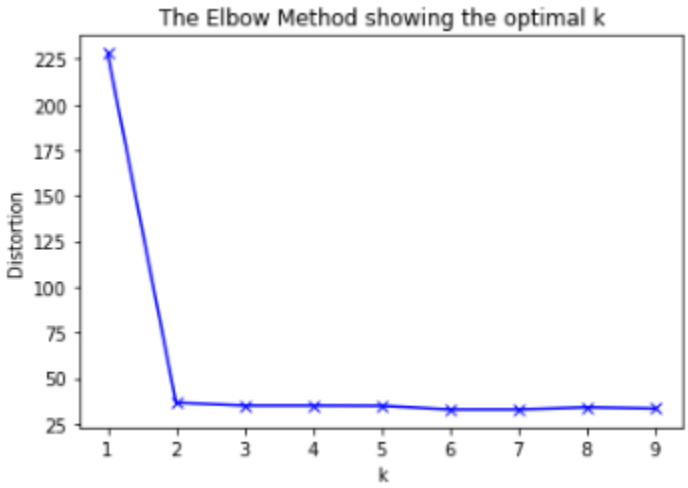
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Agincourt Malvern West Milliken	Clothing Store	Restaurant	Coffee Shop	Gym / Fitness Center	Department Store	Electronics Store	Sandwich Place	Gym	Bakery	Bank
1	Bedford Park Lawrence Manor North Toronto	Sushi Restaurant	Coffee Shop	Italian Restaurant	Pizza Place	Bakery	Sandwich Place	Pub	Fast Food Restaurant	Bank	Asian Restaurant
2	Birch Cliff Oakridge Hunt Club Cliffside	Park	Restaurant	Convenience Store	Thai Restaurant	College Stadium	General Entertainment	Gym	Skating Rink	Diner	Café
3	Bloor West Village Baby Point The Junction (Ju...	Café	Grocery Store	Coffee Shop	Restaurant	Italian Restaurant	Mexican Restaurant	Thai Restaurant	Bakery	Bar	Arts & Crafts Store
4	Cabbagetown St. Lawrence Market Toronto waterf...	Coffee Shop	Restaurant	Grocery Store	Café	Diner	Park	Pharmacy	Pizza Place	Thai Restaurant	Hotel

Source: Author's analysis of Foursquare request.

In order to determine groups of districts with similar characteristics of most common venues, the districts were clustered using K-means cluster algorithms. To achieve this the neighborhood column was dropped from the data frame in order to leave only numerical units for analysis.

The K-means algorithm requires establishing the optimal value of k for the unsupervised learning iterations. This value of k determines the number of clusters for optimal grouping and it was determined using the elbow method.

Figure 3.4: Plot of the Distortions against the Values of k ranging from 1 to 10



Source: Author's analysis on python

The optimal value of k is 2 using the elbow graph method for various runs with different radius values. This implies that one group is the optimal cluster for the data frame. However, for the purpose of the study we applied k to be 3 in line with the class category of the Distract to investigate any inherent relationship. This produced the following array of values from 0 to 2 for the 35 districts.

The house price data frame was merged to the sorted district venues in order to attach the k-means obtained cluster labels accordingly. A new data frame given as Toronto merged emerged

3.3 Analysis Testing

K-means cluster algorithm

In order to test if the K-means obtained cluster grouping has any significant relationship with the District classes, the best fit of the cluster array is required to be determined before application in the testing tool. Comparing the Cluster labels and the Area Class columns the best fit is

2 = Lower Class Area 1 = Mid Class Area 0 = High Class Area

Hence the prediction of the k-means cluster classification of districts in Toronto City was obtained accordingly.

Support Vector Machine

In order to apply the support vector machine the algorithm was supervised by splitting the Toronto_grouped data frame, that is the grouped average of each business category (independent variable), into a training set and a test set. The actual district's area classification (dependent variable was also split into training and test set. Shown as

```
Train set: (28, 245) (28,)
Test set: (7, 245) (7,)
```

The support vector machine was applied to train the train set and its prediction on the test set is given as follows

```
array([1, 1, 1, 1, 1, 1, 1])
```

4. Results

f1-score machine learning tool was used to test the accuracy of both the K-means cluster best fit classification and the support vector machine classification of the District's classes, because it conveys the balance between the precision and the recall values. The results were given as follows

The accuracy of the K-means cluster best fit classification was obtained as

```
0.45714285714285713
```

The accuracy of the support vector machine classification was obtained as

```
0.8333333333333333
```

The k-means cluster classification of the Districts in Toronto city was visualized on the city's map to show the distribution of its classification, using

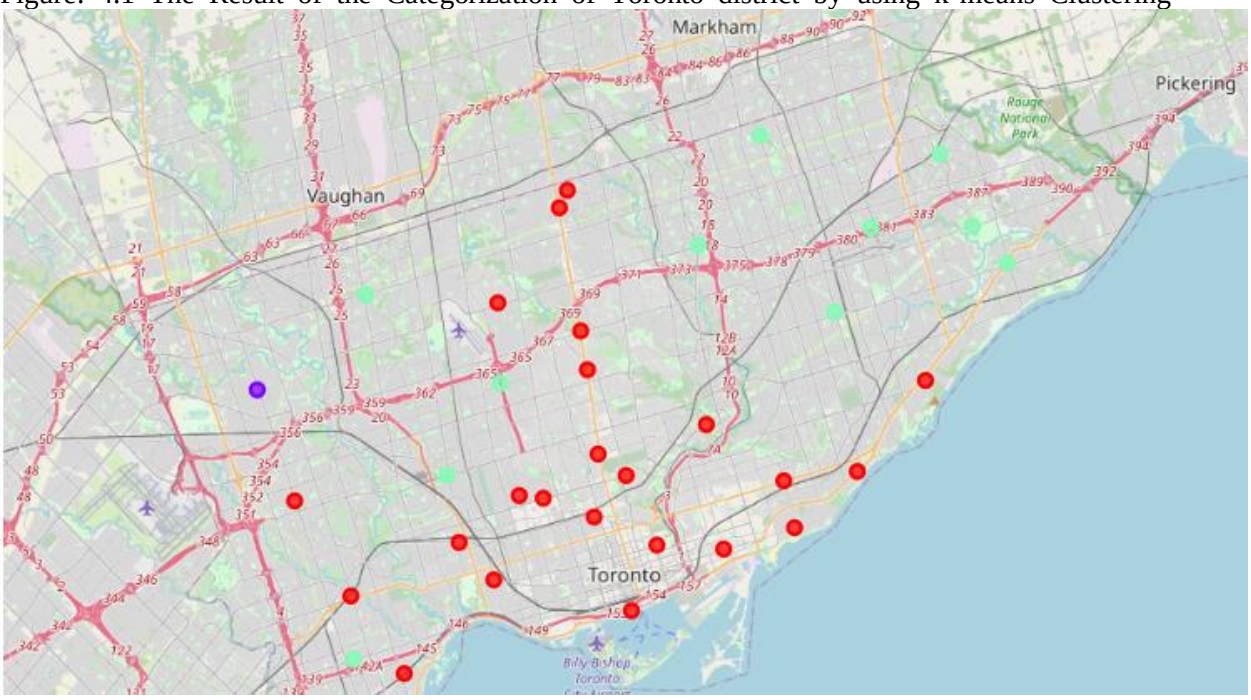
Green Circle markers - for Lower class districts

Purple Circle markers - for Mid class districts

Red Circle markers - High class districts.

The result of the Categorization is displayed in figure 4.1

Figure: 4.1 The Result of the Categorization of Toronto district by using k-means Clustering



Source: Author's analysis displayed on Folium

In order to obtain the most common business categories for each district the counts of the top ten venues were collated for every neighborhood in the districts and the first 5 rows is presented as

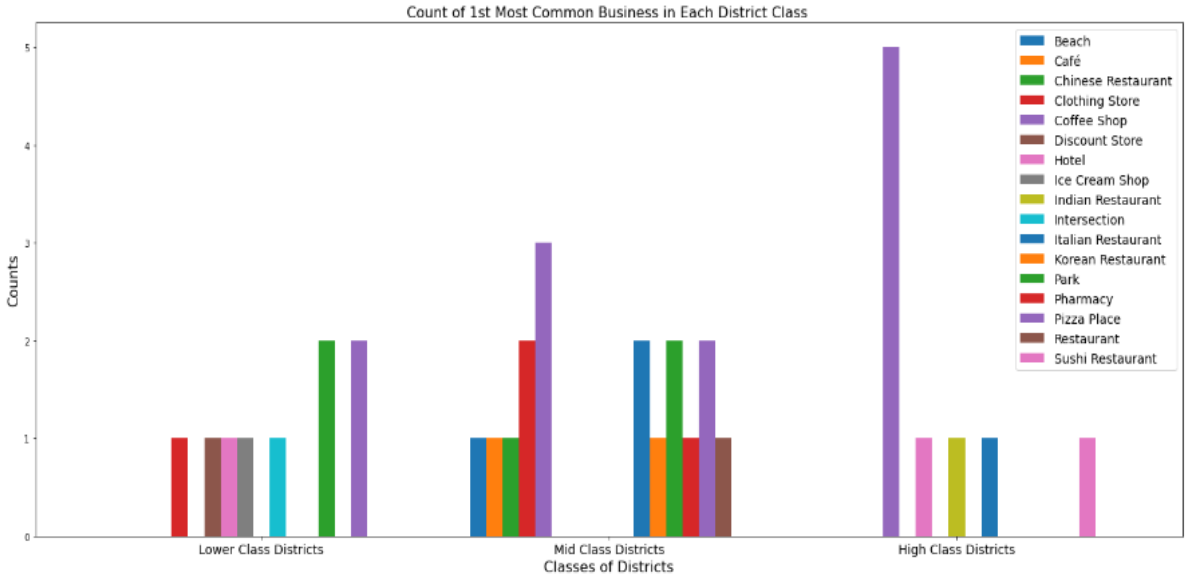
Table 4.1: First Five Rows of the top ten businesses of every neighborhood aggregated by districts

	Neighborhood	Join
0	Agincourt Malvern West Milliken	8 Clothing Store, 5 Restaurant, 4 Coffee Shop,...
1	Bedford Park Lawrence Manor North Toronto	4 Sushi Restaurant, 3 Bakery, 3 Coffee Shop, 3...
2	Birch Cliff Oakridge Hunt Club Cliffside	2 Park, 1 Café, 1 College Stadium, 1 Convenien...
3	Bloor West Village Baby Point The Junction (Ju...	6 Café, 5 Grocery Store, 4 Coffee Shop, 4 Rest...
4	Cabbagetown St. Lawrence Market Toronto waterf...	7 Coffee Shop, 6 Restaurant, 5 Café, 5 Grocery...

Source: Author's analysis

In order to determine the top business categories in the three classes of the district, that is lower, Mid and High classes, the 1st Most common venues was determined of all the three classes and presented in a bar chart as follows

Figure 4.2: Bar Chat distribution of the 1st Most Common Businesses in the 3 District Classes



Source: Author's analysis

To obtain a comprehensive table of the result containing: the districts of Toronto City by District Numbers; the district's co-ordinates; 1st to 10th most common venues; the median house price index; the districts class categorization; and the count of the top business joined for each neighborhoods in the district, the Toronto_merged data frame was merged with the top 10 data frame.

5. Discussion

The results showed that the accuracy of the K-mean cluster algorithm is weak 45.7% and hence not significant is describing the relationship between venues found in a district and the class

category of the district. From the plotting of the clusters on the map, it is obvious that the cluster groups represents other descriptions perhaps based on geographic location rather than how expensive the neighborhoods of the districts are.

However, modeling the business venue categories by supervision, that is, training the categories in line with the determined class categories produced a significant accuracy 83.3%, where the support vector machine was used.

From the over 240 unique business categories only 17 represented the top common business across the 35 districts. There are various distinctive characteristics found in the distribution of most common businesses amongst the classes of the districts. For instance European restaurants, especially Italian, are mostly found more in expense districts and less as the district becomes less expensive.

Hence, if an investor is interested in opening an Italian restaurant in Toronto City, this report would recommend siting it in a Lower class district if the investor is averse to competition. This is because there are much fewer Italian restaurants in the Lower class. However if the investor has a business edge and not averse to competition or intends to be a service provider to Italian restaurants, then most certainly, the higher class districts should be more favorable.

The 1st most common business venue among High and Mid class districts are Coffee Shops and the 1st most common business venue among Lower and Mid class districts are Clothing Shops.

6. Conclusions

The classes of the districts in Toronto City can be defined by the most common business categories. This evidenced by the finding that given the most common business venue categories found in a district, a trained set can predict the class of the district with accuracy of up to 83.3%. Also there is no 1st most common business venue category that is uniquely prominent in all 3 classes of the districts studied. The 1st most common business venue category that is prominent in more than one class, is mostly among High and Mid class or Lower and mid classes.

Coffee Shops are the by far the top business category found in high class districts. Coffee shops are also the most vastly found business in mid class district but as copious as in high class districts. Pizza places and parks are the most prominent businesses in Lower class districts but marginally.

References

1. Storey's Real Estate News (2021). The Median Home Price in 35 Toronto Neighborhoods Right Now, retrieved from https://storeys.com/median-home-price-35-toronto-neighbourhoods/
2. Wikipedia (2021). Toronto, retrieved from https://en.wikipedia.org/wiki/Toronto

Published by

Nonso Okeke