Coursera Capstone Final Assignment Wee2 Final Report

Applied Data Science Capstone by IBM Coursera

Liu Xudong 2020-11-30

1. Background: Business Problem

"Is this place good enough to open a restaurant?" This may be the first question comes to your mind when you decide to open a restaurant in some place. So how do you know if you have chosen the right location?

There are many impacts can affect the operation of the restaurant later, such as number of existing restaurants nearby, number of populations in the location you choose, if there is a station or parking nearby, and so on. And sometimes it is hard to say which factor is decisive and which one is useless, we need a smarter way to take all these factors into consideration.

In this project, I will use data science method to help the stakeholders to make the decision by analyzing these factors.

2. Datasets

Based on description of our problem, we have identified following factors that will impact our decision are:

- number of existing restaurants around the location
- number of population or population change in past years in the neighborhood
- if there are bus stations/subway stations nearby
- how many parks, schools, hotels, markets, shopping malls, bars in the chosen location
- the total income for the population aged 15 years and over.

We use following data sources to create datasets which we will use in our project:

- The neighborhood name, population and total income information is extracted from Toronto Neighborhood Profile.
- The latitude and longitude of the neighborhood is given by google geocoding.
- Number of restaurants/stations/parks/etc and their location in every neighborhood will be obtained using Foursquare API.

Finally, we will have a data set including Latitude, Longitude, Population, Population Change, Total Income for the population aged 15 years and over, number of

Restaurant, Mall/Shops, Station/Subway, Bar/Drinks, Park, Hotel, School and Market. Each step of data cleaning will be described in detail in following cells.

2.1 Toronto Neighborhood Profile

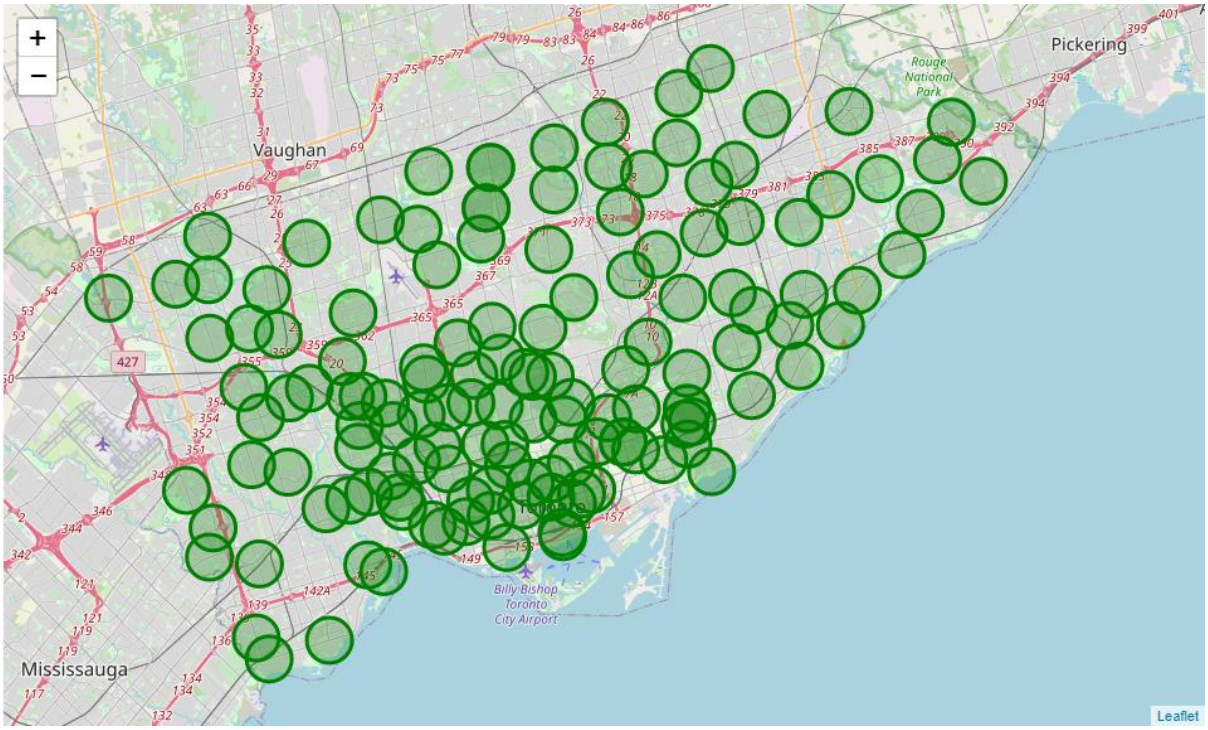
Read Toronto Neighborhood Profile file downloaded from Toronto government website (link) to data frame, transpose the neighborhood name and Characteristic, and pick out the data we need (Neighborhood names, Population(2016), Population Change(2011~2016), Total-Income statistics), call google geocoding API to get the coordinate data of each neighborhood, store all the information in data frame df_toronto_neighbors.

	Neighborhood	Population, 2016	Population Change 2011-2016	Total Income(>=15 years)	Latitude	Longitude
0	Agincourt North	29,113	-3.90%	25,005	43.808053	-79.266502
1	Agincourt South-Malvern West	23,757	8.00%	20,400	43.788009	-79.283882
2	Alderwood	12,054	1.30%	10,265	43.601710	-79.545238
3	Annex	30,526	4.60%	26,295	43.669833	-79.407585
4	Banbury-Don Mills	27,695	2.90%	23,410	43.744847	-79.340923
...	...	...	...	...	...	...
135	Wychwood	14,349	2.60%	11,345	43.677910	-79.420102
136	Yonge-Eglinton	11,817	11.70%	9,995	43.706431	-79.398642
137	Yonge-St.Clair	12,528	7.50%	11,170	43.688098	-79.394117
138	York University Heights	27,593	-0.40%	23,530	43.766449	-79.477446
139	Yorkdale-Glen Park	14,804	0.80%	12,065	43.708236	-79.453975

140 rows × 6 columns

2.2 Explore the neighborhoods in Toronto

Before using Foursquare API to explore the neighborhood, we marked each location on the map (latitude and longitude), and tried different radius to see if the radius is suitable to cover most of the area. Finally we use radius = 1000 meters, and from the map we can see almost 80% areas are covered.



By exploring the neighborhood, we got a list of venues within the areas, and marked them using some general categories, such as Restaurant, Bar/Drinks, Mall, Hotel and so on.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Restaurant	Fast Food	Mall/Shops	Station/Subway	Bar/Drinks
0	Agincourt North	43.808053	-79.266502	Menchie's	43.808338	-79.268288	Frozen Yogurt Shop	0	0	1	0	0
1	Agincourt North	43.808053	-79.266502	Saravanaa Bhavan South Indian Restaurant	43.810117	-79.269275	Indian Restaurant	1	0	0	0	0
2	Agincourt North	43.808053	-79.266502	Samosa King - Embassy Restaurant	43.810152	-79.257316	Indian Restaurant	1	0	0	0	0
3	Agincourt North	43.808053	-79.266502	Booster Juice	43.809915	-79.269382	Juice Bar	0	0	0	0	1
4	Agincourt North	43.808053	-79.266502	Congee Town 太皇名粥	43.809035	-79.267634	Chinese Restaurant	1	0	0	0	0

Finally, we merged the venue data frame and neighborhood profile data frame together with the coordinate data, and store in a new data frame, df_toronto_neighbors_merged.

	Neighborhood	Latitude	Longitude	Population, 2016	Population Change 2011-2016	Total Income(>=15 years)	Restaurant	Mall/Shops	Station/Subway	Bar/Drinks	Park	Hotel	School
0	Agincourt North	43.808053	-79.266502	29113.0	-0.039	25005.0	9	10	0	4	2	0	0
1	Agincourt South-Malvern West	43.788009	-79.283882	23757.0	0.080	20400.0	16	8	0	3	1	0	0
2	Alderwood	43.601710	-79.545238	12054.0	0.013	10265.0	3	9	0	2	2	0	0
3	Annex	43.669833	-79.407585	30526.0	0.046	26295.0	36	24	0	12	1	1	2
4	Banbury-Don Mills	43.744847	-79.340923	27695.0	0.029	23410.0	5	2	0	1	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
135	Wychwood	43.677910	-79.420102	14349.0	0.026	11345.0	34	25	0	13	2	0	1
136	Yonge-Eglinton	43.706431	-79.398642	11817.0	0.117	9995.0	38	26	0	16	1	0	0
137	Yonge-St.Clair	43.688098	-79.394117	12528.0	0.075	11170.0	17	10	1	4	2	1	0
138	York University Heights	43.766449	-79.477446	27593.0	-0.004	23530.0	8	11	0	6	0	0	0
139	Yorkdale-Glen Park	43.708236	-79.453975	14804.0	0.008	12065.0	14	10	1	3	0	0	0

3. Methodology and Analysis

Till now, we have collected all the data we needed, and stored the data in a data frame. But the data currently only contains the features data we needed, from these data, it is hard to say which location is better. In this chapter, we will discuss and find out a methodology to achieve our goal.

Our methodology will be:

1. Data Analysis

There are 11 features from above data frame, can all of them be used to cluster the dataset? We will analyze the data and select the most relevant features for clustering.

2. Label the current location

Because the dataset now only has the features we need, we will then use unsupervised learning algorithm to clustering the dataset, and use the result as the label of each location.

3. Define the cluster

We will understand the cluster, and based on the analysis result, we will assign a name for each cluster, for example very suitable, suitable, not suitable.

4. New dataset for classification

When we get the cluster labels, combine the cluster labels and data frame together as a new data frame, which will be used for training the classification model later.

5. Classification

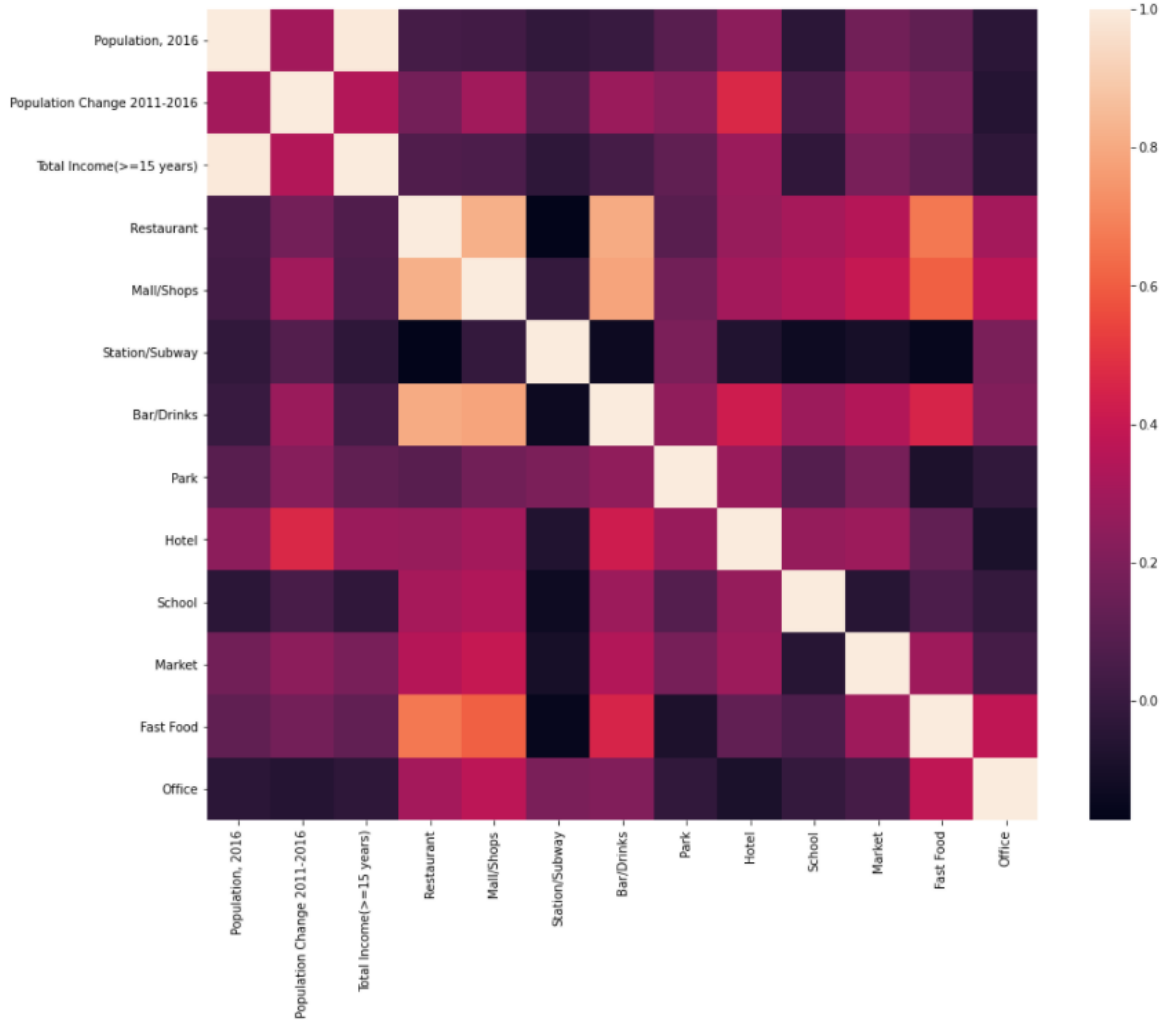
Use the new dataset to training the classification model, and the model will be used to predict if the new location is good enough to invest a restaurant or not.

3.1 Data Analysis

First of all, because the data values are not in the same magnitude, do the feature scaling to standardize them.

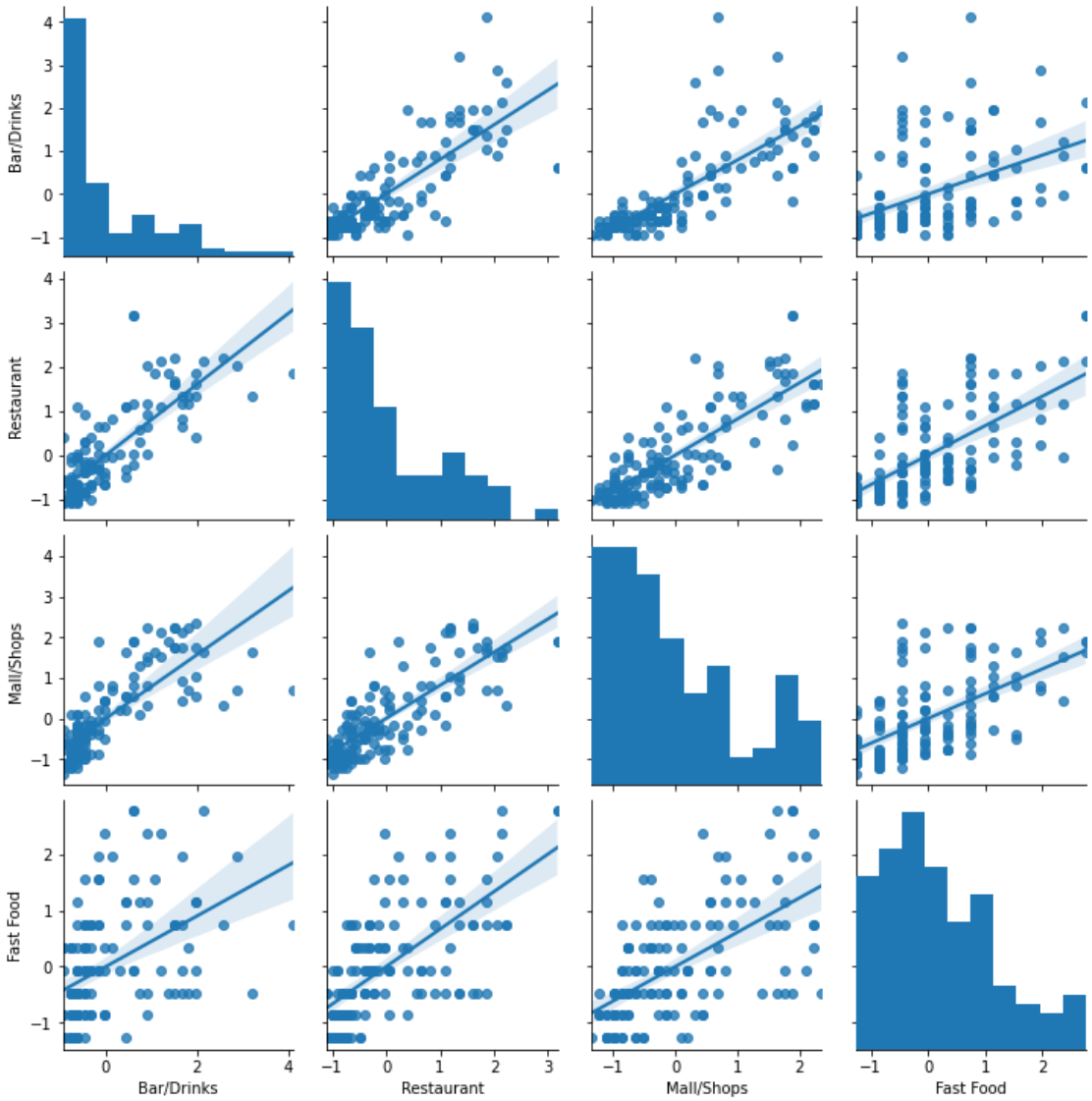
	Population, 2016	Population Change 2011-2016	Total Income(>=15 years)	Restaurant	Mall/Shops	Station/Subway	Bar/Drinks	Park	Hotel	School	Market	Fast Food	Off
0	0.960400	-0.895105	0.988683	-0.305871	-0.154866	-0.652051	-0.319429	0.239155	-0.342594	-0.262111	-0.738300	0.342584	1.0327
1	0.424676	0.446305	0.460132	0.303385	-0.394438	-0.652051	-0.472057	-0.444146	-0.342594	-0.262111	0.410167	0.342584	0.0216
2	-0.745894	-0.308943	-0.703141	-0.828091	-0.274652	-0.652051	-0.624685	0.239155	-0.342594	-0.262111	-0.738300	-0.866536	-0.9894
3	1.101732	0.063045	1.136747	2.044115	1.522138	-0.652051	0.901596	-0.444146	0.616670	4.630632	-0.738300	0.745624	0.0216
4	0.818567	-0.128585	0.805613	-0.654017	-1.113154	-0.652051	-0.777313	-1.127446	-0.342594	-0.262111	-0.738300	-1.269576	1.0327

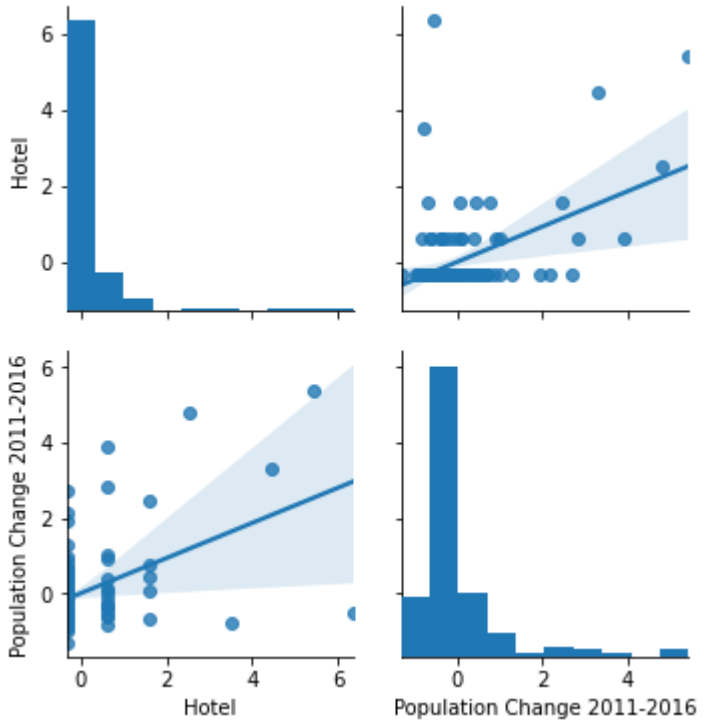
Then check the correlation between the features, and we got the heatmap plot as below:

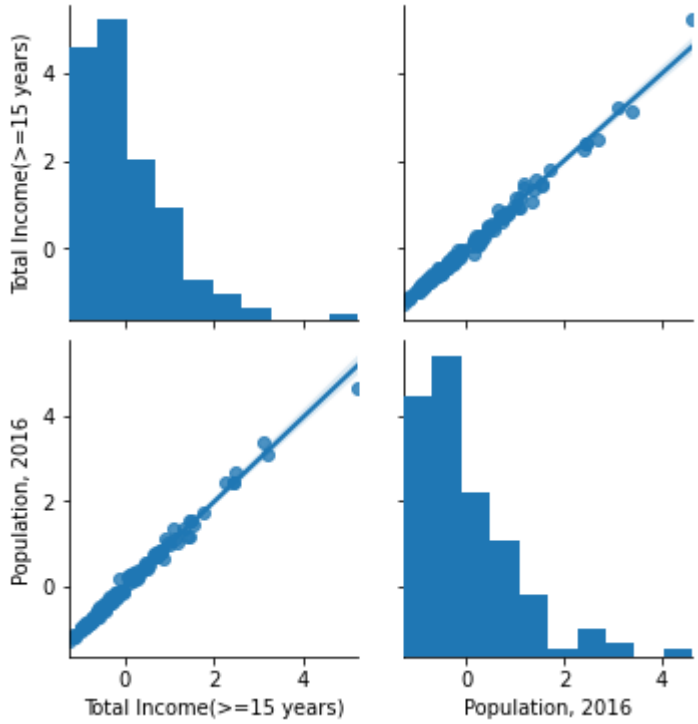


Above we can see the correlation plot of all the numerical features, there are some obvious bright spots, for instance

Bar/Drinks <-> Restaurant <-> Mall/Shops <-> Fast Food
Hotel <-> Population Change
Total income <-> Population
Look more into details about these features:


Bar/Drinks
Restaurant
Mall/Shops
Fast Food
Bar/Drinks
Restaurant
Mall/Shops
Fast Food

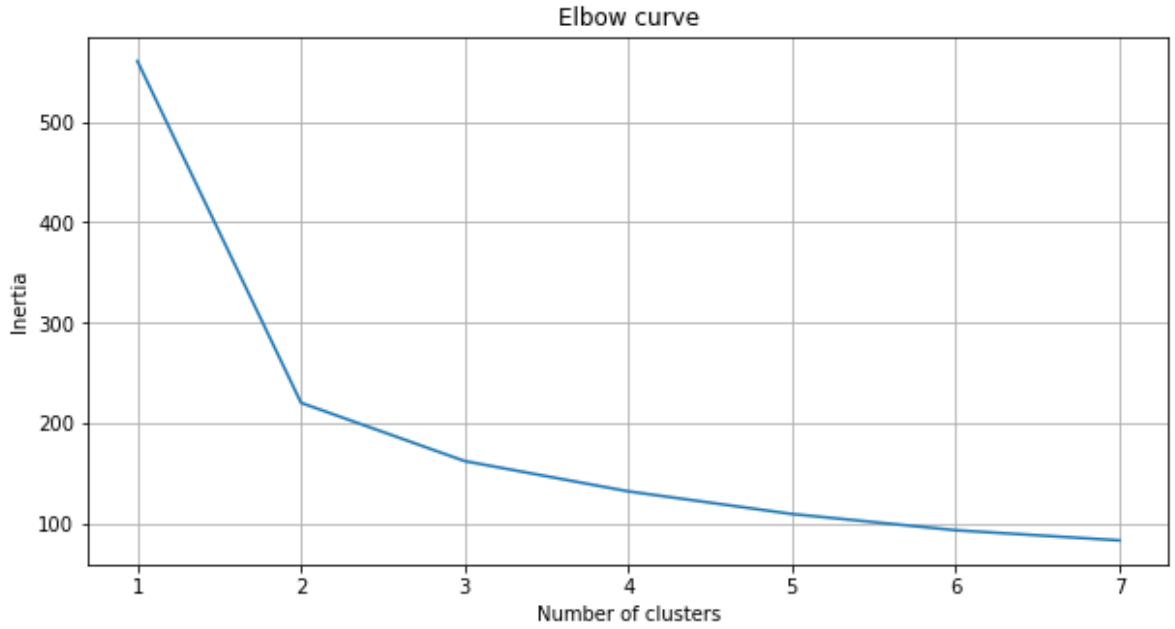




From the above plots, it is clear that linear relation among Restaurant, Bar/Drinks, Mall/Shops and Fast Food. And same for Total Income and Population, which is much stronger relationship than Restaurant. But it is hard to say clear relation between Hotel and Population Change. Based on this, in order to cluster, we will choose 'Restaurant', 'Mall/Shops', 'Bar/Drinks' and 'Fast Food' as selected features.

3.2 Label the current location

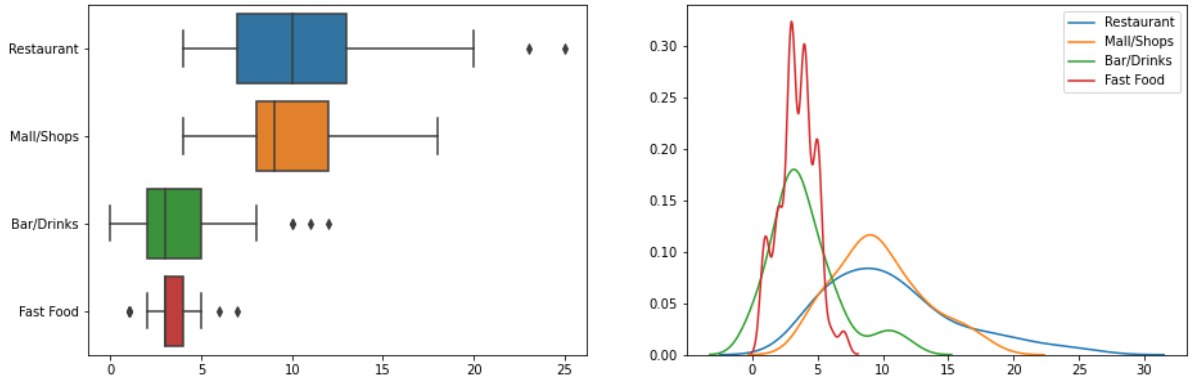
We use k-means algorithm to cluster the dataset and use the elbow method for finding the number of clusters. K-means clustering is a method of partitioning n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid). K-means clustering minimizes within-cluster variances (squared Euclidean distances).



Indeed, in the resulting plot, "the elbow" is located at k=5, which is evidence that k=5 is indeed a good choice of the number of clusters for this dataset.

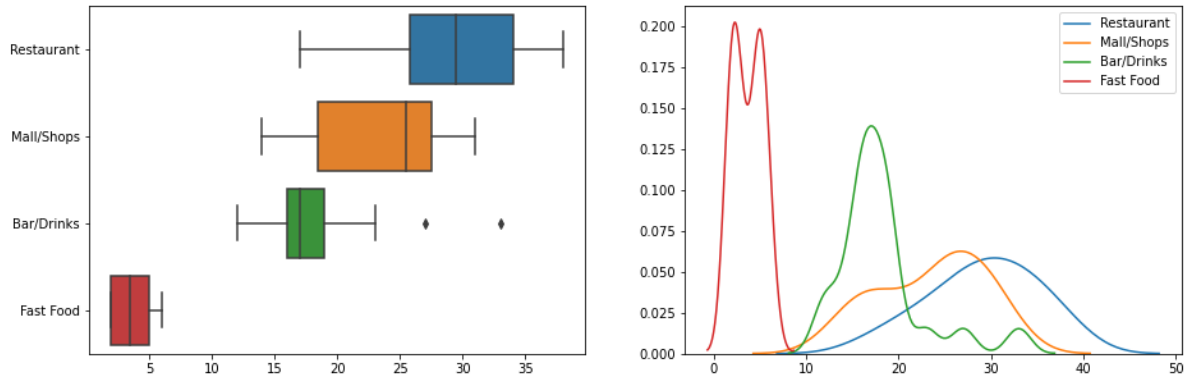
3.3 Define the cluster

After clustering the dataset, we looked into each cluster and try to define the meaning of each cluster:
Cluster_0:

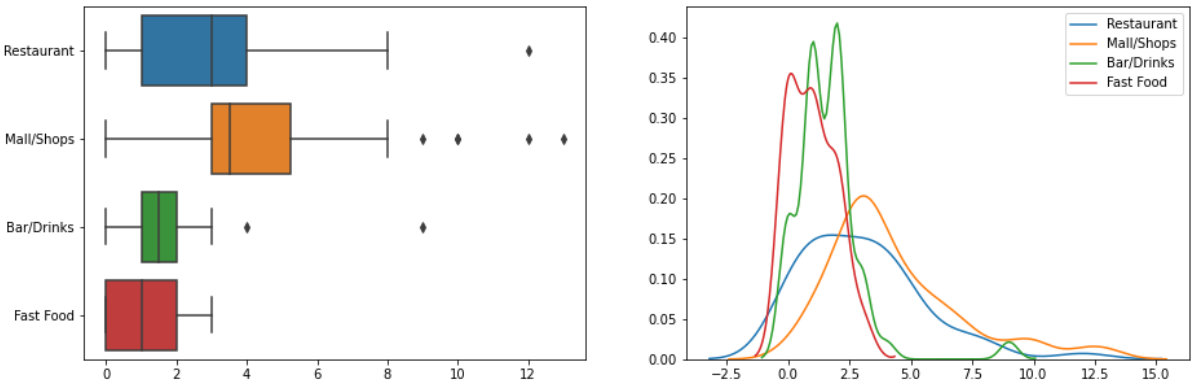


From the plot, for locations in cluster_0, the distribution of Restaurant is very consistent with the shopping Mall and the total number is not large. So considering to locate the

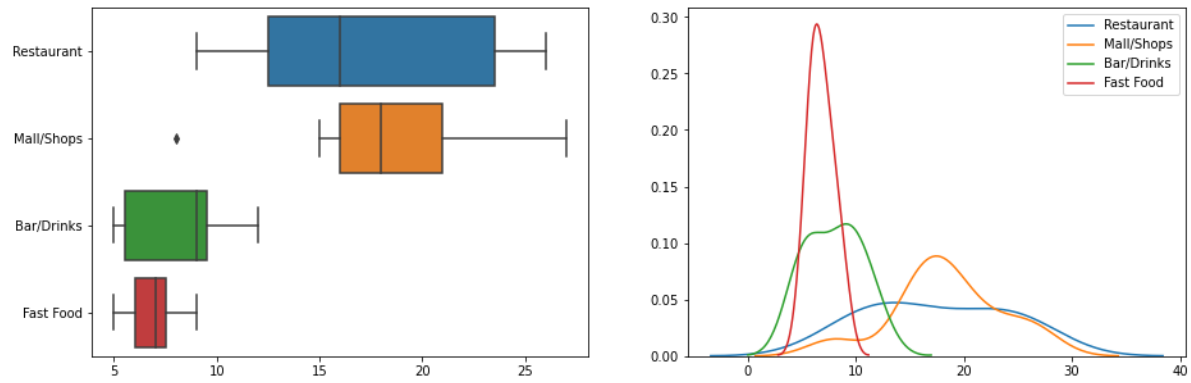
Restaurant near the shopping mall.
Cluster_1:



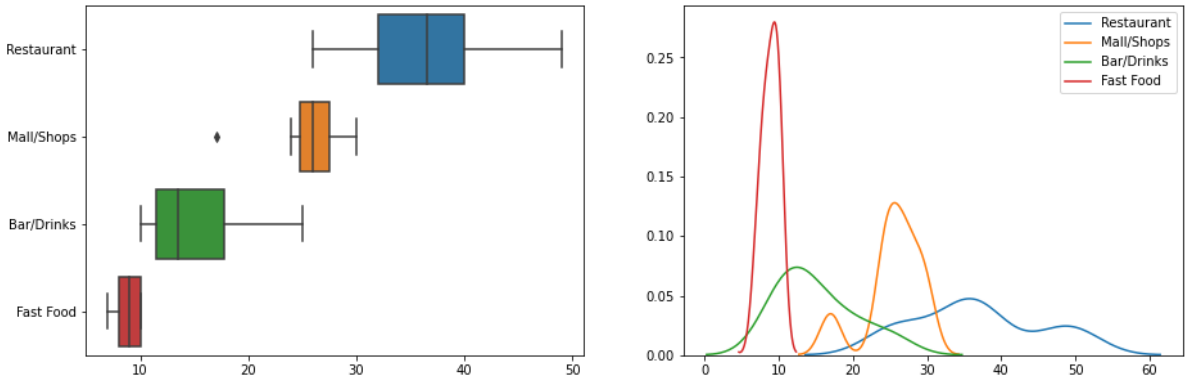
From the plot, for locations in cluster_1, the distribution of Restaurant is quite similarly with the shopping Mall (not as good as cluster_0), but where the Restaurant is concentrated the Bar/Drinks is few. So considering to invest Bar/Drinks near a restaurant.
Cluster_2:



From the plot, for locations in cluster_2, all of the venues are distributed similar, and number of each is not many. So this is a very good choice to invest Restaurant.
Cluster_3:



From the plot, for locations in cluster_3, the distribution range of Restaurant is large, and number is not small. But look at the number of Bar/Drinks, which is positive related with restaurant, there are not enough restaurant at the large number of Bar/Drinks, so a restaurant near a bar/drinks can be considered.
Cluster_4:



From the plot, for locations in cluster_4, the distribution of all the venues is very scattered, and number of Restaurant is large. So this is not a good choice to open a restaurant.

From the above analysis, we can define the clusters by following policy:
Cluster_0: Invest restaurant near shopping mall
Cluster_1: Invest bar/drinks near existing restaurant
Cluster_2: Good place to open a restaurant
Cluster_3: Invest restaurant near bar/drinks
Cluster_4: Not a good choice

3.4 New dataset for classification

We stored all the selected features and labels from clustering in a new dataset, called df_classify, and make dataset X contain all feature data, and dataset y contain label data. To prepare for classification, also split the data into training set and test set.

3.5 Classification

Because we have 5 clusters, which means 5 kind of values in dataset y, this is a multi-classes classification problem, we will use Multilayer Perceptron (MLP) provided by sklearn to build the model. By implementing the algorithm and training the model, we got the model performance report as below:

_____ Training Report _____

	precision	recall	f1-score	support
0	1.00	1.00	1.00	38
1	1.00	1.00	1.00	16
2	1.00	1.00	1.00	44
3	1.00	1.00	1.00	9
4	1.00	1.00	1.00	5
accuracy			1.00	112
macro avg	1.00	1.00	1.00	112
weighted avg	1.00	1.00	1.00	112

_____ Test Report _____

	precision	recall	f1-score	support
0	1.00	0.91	0.95	11
1	0.80	1.00	0.89	4
2	1.00	1.00	1.00	8
3	0.67	1.00	0.80	2
4	1.00	0.67	0.80	3
accuracy			0.93	28
macro avg	0.89	0.92	0.89	28
weighted avg	0.95	0.93	0.93	28

We can see that our neural networks classification (MLP) overall result (F1-scroe) is 93%, which shows a good performance of the model. But we can see the accuracy of cluster_3 and cluster_4 is not good as others, this may because the samples don't include as many samples as other clusters, so the model for this two cluster is not trained as good as other. This can be improved by collecting more data.

4. Conclusion

The purpose of this project is to help the stakeholders to make decision that if the location is good enough to make restaurant investment by analyzing the location nearby environment. In this project, we used a lot of methods to obtain the dataset we need, such as pandas, web request/API, google geocoding, used many methods to do the data cleaning, for example missing data handling, type conversion, correlation analysis (heatmap and regplot), used unsupervised learning algorithm (KMeans) and supervised learning algorithm (MLP) to build the model. With all these efforts, we can give a prediction (93% accuracy) of new location that which cluster it belongs to, and then give the suggestion on the restaurant investment in this location.

Future directions: During the project, I found that the neighborhood exploration using Foursquare API has impact on the dataset, you should carefully choose the exploration radius. I chose the radius 1000 meters which could cover 80% area of Toronto, and it was not uniform. In future the radius can be considered dynamically, it can be smaller at the dense areas and bigger at the sparse space. Another improvement area we can imagine is to collect more data, because we saw the prediction accuracy for some clusters is not as good as others.