Ryszard Milewicz

Report

Toronto - Where to Open a Restaurant

Issued: 31 August, 2020

Table of contents

1. Introduction

Toronto is the biggest city in Canada. It creates a big number of business opportunities. Problem where to open a restaurant in Toronto is important to several businesspeople who are involved in restaurant business or plan to enter this business. In order to limit business risk, it is important to know where there is the best location for this business in Toronto based on available data about Toronto population and neighborhoods.

This analysis based on data science methods should give substantial support to take right decision about restaurant location. According to Dalhousie University report:

https://www.dal.ca/news/2017/12/13/canadians-will-spend-more-in-restaurants-in-2018--canada-s-food-.html

almost 30 per cent of consumers' food budget will be spent on eating out. So, information about consumer's income in neighborhoods and potential spending per restaurant should help finding right place for opening a restaurant.

Interested parties:

- entrepreneurs from other parts of country who are looking to open a restaurant in Toronto,

- business people from Toronto who are looking to change business area,

- restaurants owners who are looking to expand their businesses.

2. Data

In order to solve above mentioned problem there is needed data about restaurants in neighborhoods and demographic data (population, income). Foursquare location data will be used to find restaurants. In order to find venues using Foursquare API there is needed information about location of every neighborhood.

Location data can be found in Internet in several locations. In Wikipedia it is possible to get postal codes for each Toronto neighborhood - "List of postal codes of Canada: M":

https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Then use Geocoder Python package:

https://geocoder.readthedocs.io/index.html.

to get the latitude and longitude coordinates of a given postal code. However, sometimes you will get "None" instead of coordinates, so multiple queries are needed.

More reliable data source should be available at government sites like Natural Resources Canada or local government's site. Toronto Open Data database covers wide amount of data about Toronto under following link:

https://www.toronto.ca/city-government/data-research-maps/open-data/

and will be used for this project as it is up to date and free for commercial use. Data about neighborhood coordinates can be found under following link:

https://ckan0.cf.opendata.inter.prod-toronto.ca/download_resource/a083c865-6d60-4d1d-b6c6-b0c8a85f9c15?format=csv&projection=4326

The same database will be used to get information about population of each neighborhood and average income:

https://ckan0.cf.opendata.inter.prod-toronto.ca/download_resource/ef0239b1-832b-4d0b-a1f3-4153e53b189e?format=csv

By getting data from one source we expect less problems with data compatibility.

After merging location data with population and income, neighborhood venues will be added using Foursquare API. Then venues will be filtered after restaurants and next K-mean clustering will be performed to get classification of neighborhoods versus number of restaurants, population and income. Finally, best cluster will be analyzed and recommendation given.

3. Methodology

Firstly population data was imported from Toronto Open Data database using above mentioned link in convenient csv file form. Resulted dataset was huge (2383 columns), so it was filtered in order to cover data important for this analysis:

Characteristic	Neighborhood	Neighbourhood Number	Population, 2016	Total income: Population with an amount	Total income: Average amount ($)
0	City of Toronto	NaN	2,731,571	2,187,220	52,268
1	Agincourt North	129	29,113	23,505	30,414
2	Agincourt South-Malvern West	128	23,757	19,370	31,825
3	Alderwood	20	12,054	9,915	47,709
4	Annex	95	30,526	25,615	112,766
...	...	...	...	...	...
136	Wychwood	94	14,349	11,030	54,460
137	Yonge-Eglinton	100	11,817	9,555	89,330
138	Yonge-St.Clair	97	12,528	10,805	114,174
139	York University Heights	27	27,593	22,230	29,958
140	Yorkdale-Glen Park	31	14,804	11,490	38,527

As we can see in this dataset, there are 140 neighborhoods in Toronto, so wide variety of places, where someone can open a restaurant. Advice, where is the best place to do so will be important. We are interested in neighborhoods data, so row with whole Toronto data should be removed. Also, long column names can be shorter. Number format should be modified to remove unnecessary comma marks as for further analysis we will need some operations like multiplication which will be not possible on numbers with commas. After modifications dataset looks clearly:

Characteristic	Neighborhood	Neighbourhood Number	Population	Average income	Neighborhood Income
1	Agincourt North	129	23505	30414	714881070
2	Agincourt South-Malvern West	128	19370	31825	616450250
3	Alderwood	20	9915	47709	473034735
4	Annex	95	25615	112766	2888501090
5	Banbury-Don Mills	42	22335	67757	1513352595
...	...	...	...	...	...
136	Wychwood	94	11030	54460	600693800
137	Yonge-Eglinton	100	9555	89330	853548150
138	Yonge-St.Clair	97	10805	114174	1233650070
139	York University Heights	27	22230	29958	665966340
140	Yorkdale-Glen Park	31	11490	38527	442675230

In order to find restaurants near each neighborhood, we need coordinates (latitude, longitude). This data is also taken from Toronto Open Data database in csv form. Raw dataset contains information useful and not useful for this analysis:

	_id	AREA_ID	AREA_ATTR_ID	PARENT_AREA_ID	AREA_SHORT_CODE	AREA_LONG_CODE	AREA_NAME	AREA_DESC	X	Y	LONGITUDE	LATITUDE	OBJECTID	Shape__Area	Shape__Length	geometry
0	7141	25886861	25926662	49885	94	94	Wychwood (94)	Wychwood (94)	NaN	NaN	-79.425515	43.676919	16491505	3.217960e+06	7515.779658	{u'type': u'Polygon', u'coordinates': (((-79.4...
1	7142	25886820	25926663	49885	100	100	Yonge-Eglinton (100)	Yonge-Eglinton (100)	NaN	NaN	-79.403590	43.704689	16491521	3.160334e+06	7872.021074	{u'type': u'Polygon', u'coordinates': (((-79.4...
2	7143	25886834	25926664	49885	97	97	Yonge-St.Clair (97)	Yonge-St.Clair (97)	NaN	NaN	-79.397871	43.687859	16491537	2.222464e+06	8130.411276	{u'type': u'Polygon', u'coordinates': (((-79.3...
3	7144	25886593	25926665	49885	27	27	York University Heights (27)	York University Heights (27)	NaN	NaN	-79.488883	43.765736	16491553	2.541821e+07	25632.335242	{u'type': u'Polygon', u'coordinates': (((-79.5...
4	7145	25886688	25926666	49885	31	31	Yorkdale-Glen Park (31)	Yorkdale-Glen Park (31)	NaN	NaN	-79.457108	43.714672	16491569	1.156669e+07	13953.408098	{u'type': u'Polygon', u'coordinates': (((-79.4...

Data is filtered to contain only required information: neighborhood number, longitude and latitude:

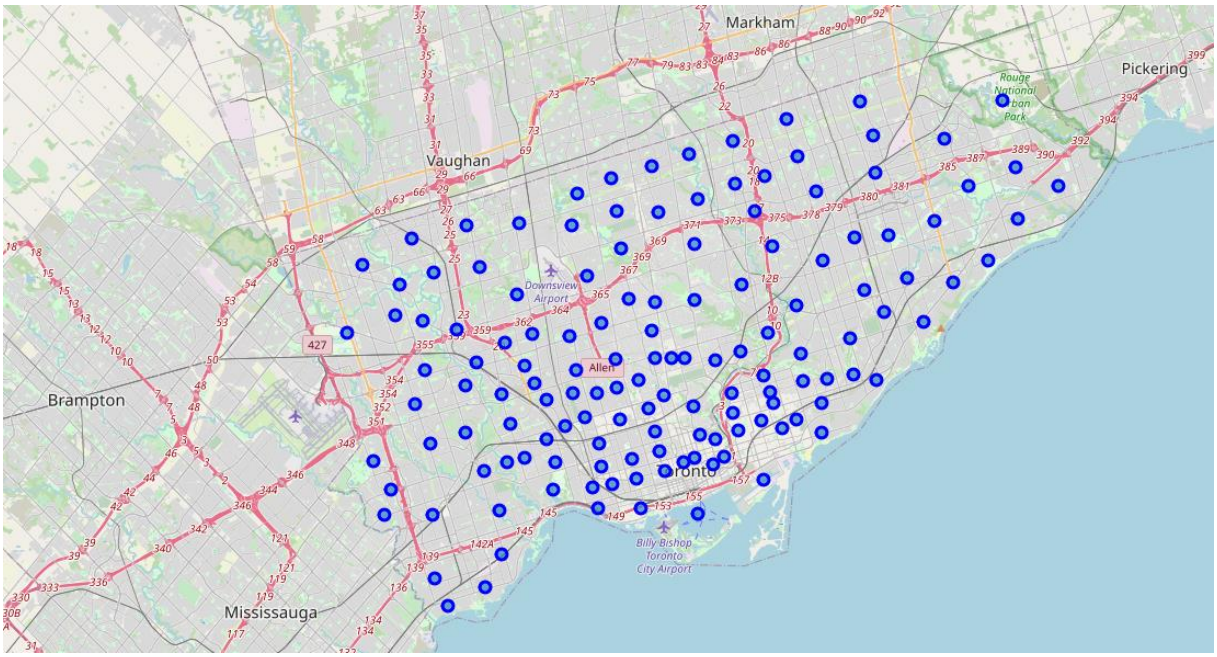
	Neighbourhood Number	Longitude	Latitude
0	94	-79.425515	43.676919
1	100	-79.403590	43.704689
2	97	-79.397871	43.687859
3	27	-79.488883	43.765736
4	31	-79.457108	43.714672
...	...	...	...
135	124	-79.260382	43.725556
136	78	-79.397240	43.653554
137	6	-79.547863	43.698993
138	15	-79.510577	43.653520
139	117	-79.314084	43.795716

Now it is possible to add coordinates to population data and create combined dataset:

	Neighborhood	Neighbourhood Number	Population	Average income	Neighborhood Income	Longitude	Latitude
0	Agincourt North	129	23505	30414	714881070	-79.266712	43.805441
1	Agincourt South-Malvern West	128	19370	31825	616450250	-79.265612	43.788658
2	Alderwood	20	9915	47709	473034735	-79.541611	43.604937
3	Annex	95	25615	112766	2888501090	-79.404001	43.671585
4	Banbury-Don Mills	42	22335	67757	1513352595	-79.349718	43.737657
...	...	...	...	...	...	...	...
135	Wychwood	94	11030	54460	600693800	-79.425515	43.676919
136	Yonge-Eglinton	100	9555	89330	853548150	-79.403590	43.704689
137	Yonge-St.Clair	97	10805	114174	1233650070	-79.397871	43.687859
138	York University Heights	27	22230	29958	665966340	-79.488883	43.765736
139	Yorkdale-Glen Park	31	11490	38527	442675230	-79.457108	43.714672

So, having above dataset it is possible to use Foursquare API to find venues for every neighborhood.

Firstly, lets look where neighborhoods are located on Toronto map:



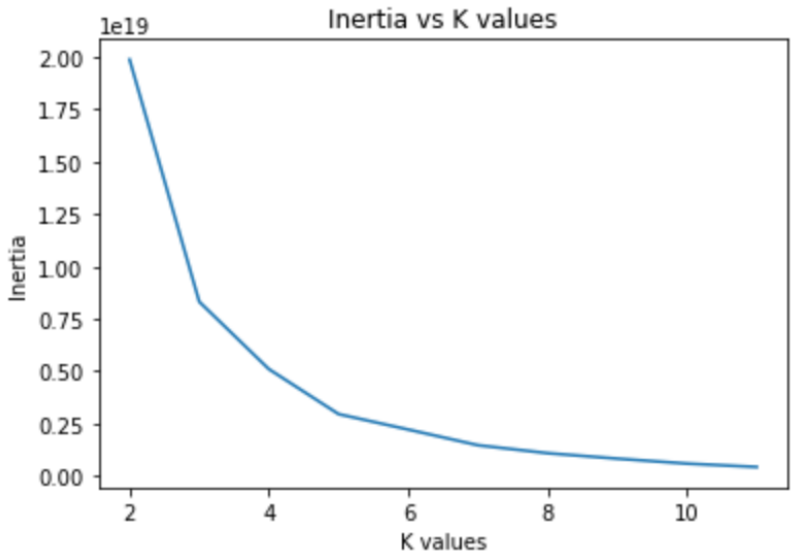
Let's assume that neighborhood radius is 700 meters. We will look for top venues (limited to 100) in each neighborhood using Foursquare API. Resulting dataset is quite huge, first 20 rows looks as follows:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Agincourt North	43.805441	-79.266712	Menchie's	43.808338	-79.268288	Frozen Yogurt Shop
1	Agincourt North	43.805441	-79.266712	Saravanaa Bhavan South Indian Restaurant	43.810117	-79.269275	Indian Restaurant
2	Agincourt North	43.805441	-79.266712	Shoppers Drug Mart	43.808894	-79.269854	Pharmacy
3	Agincourt North	43.805441	-79.266712	Booster Juice	43.809915	-79.269382	Juice Bar
4	Agincourt North	43.805441	-79.266712	Dollarama	43.808894	-79.269854	Discount Store
5	Agincourt North	43.805441	-79.266712	Congee Town 太皇名粥	43.809035	-79.267634	Chinese Restaurant
6	Agincourt North	43.805441	-79.266712	Pizza Pizza	43.808318	-79.268537	Pizza Place
7	Agincourt North	43.805441	-79.266712	Subway	43.809372	-79.269474	Sandwich Place
8	Agincourt North	43.805441	-79.266712	Popeyes Louisiana Kitchen	43.808652	-79.267929	Fried Chicken Joint
9	Agincourt North	43.805441	-79.266712	TD Canada Trust	43.809828	-79.268764	Bank
10	Agincourt North	43.805441	-79.266712	RBC Royal Bank	43.808757	-79.269280	Bank
11	Agincourt North	43.805441	-79.266712	Tim Hortons	43.809993	-79.269032	Coffee Shop
12	Agincourt North	43.805441	-79.266712	The Beer Store	43.809286	-79.263676	Beer Store
13	Agincourt North	43.805441	-79.266712	Wild Wing	43.808799	-79.267808	Wings Joint
14	Agincourt North	43.805441	-79.266712	Xe Lua Vietnamese Cuisine 火車頭	43.809224	-79.269547	Vietnamese Restaurant
15	Agincourt North	43.805441	-79.266712	Woodside Cinemas	43.809900	-79.269521	Movie Theater
16	Agincourt North	43.805441	-79.266712	Kin Kin Bubble Tea Co	43.807852	-79.270296	Chinese Restaurant
17	Agincourt North	43.805441	-79.266712	LCBO	43.808126	-79.270046	Liquor Store
18	Agincourt North	43.805441	-79.266712	Smokers Corner Newstand	43.808624	-79.269437	Convenience Store
19	Agincourt North	43.805441	-79.266712	Bento Box	43.809248	-79.269029	Japanese Restaurant

Restaurant venues are filtered and counted for each neighborhood. It was found that there are from 0 to 37 restaurants per neighborhood. Surprisingly, there are neighborhoods without any restaurant. Results are added to dataset with population data:

	Neighborhood	Neighbourhood Number	Population	Average income	Neighborhood Income	Longitude	Latitude	Number_of_restaurants
0	Agincourt North	129	23505	30414	714881070	-79.266712	43.805441	7
1	Agincourt South-Malvern West	128	19370	31825	616450250	-79.265612	43.788658	14
2	Alderwood	20	9915	47709	473034735	-79.541611	43.604937	0
3	Annex	95	25615	112766	2888501090	-79.404001	43.671585	25
4	Banbury-Don Mills	42	22335	67757	1513352595	-79.349718	43.737657	12
...	...	...	...	...	...	...	...	...
135	Wychwood	94	11030	54460	600693800	-79.425515	43.676919	20
136	Yonge-Eglinton	100	9555	89330	853548150	-79.403590	43.704689	29
137	Yonge-St.Clair	97	10805	114174	1233650070	-79.397871	43.687859	19
138	York University Heights	27	22230	29958	665966340	-79.488883	43.765736	4
139	Yorkdale-Glen Park	31	11490	38527	442675230	-79.457108	43.714672	12

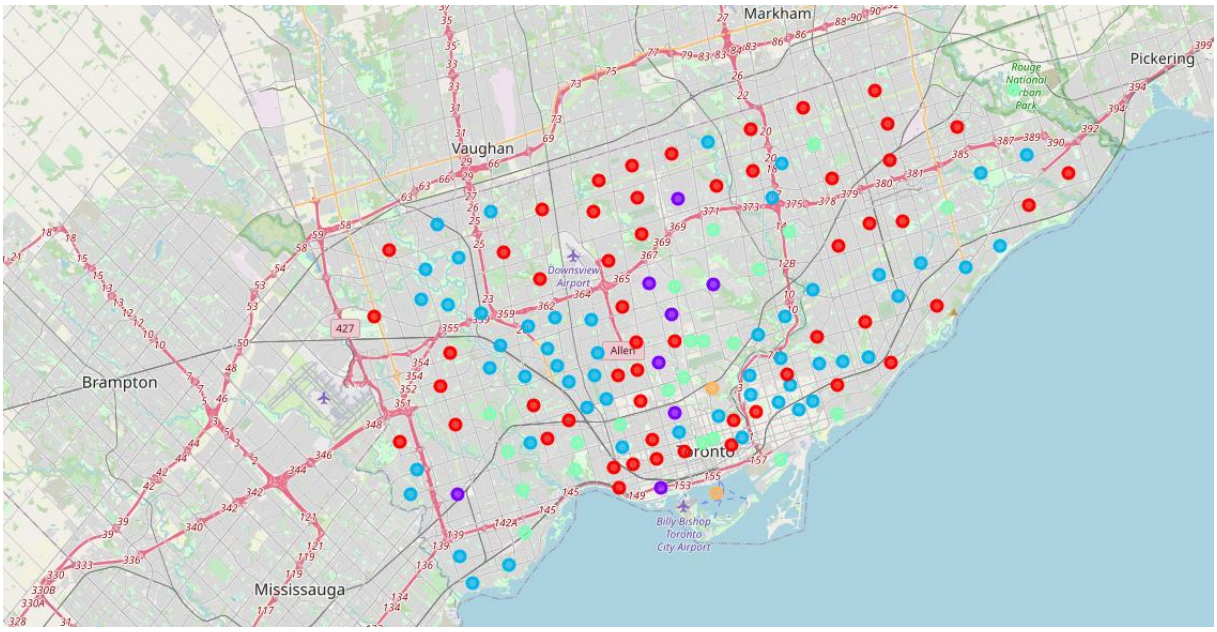
On above dataset was performed K-means clustering using different number of clusters: 2 to 12. For each number of clusters inertia was calculated:



Optimal cluster number (based on change of inertia) is 5. Cluster numbers were added to dataset:

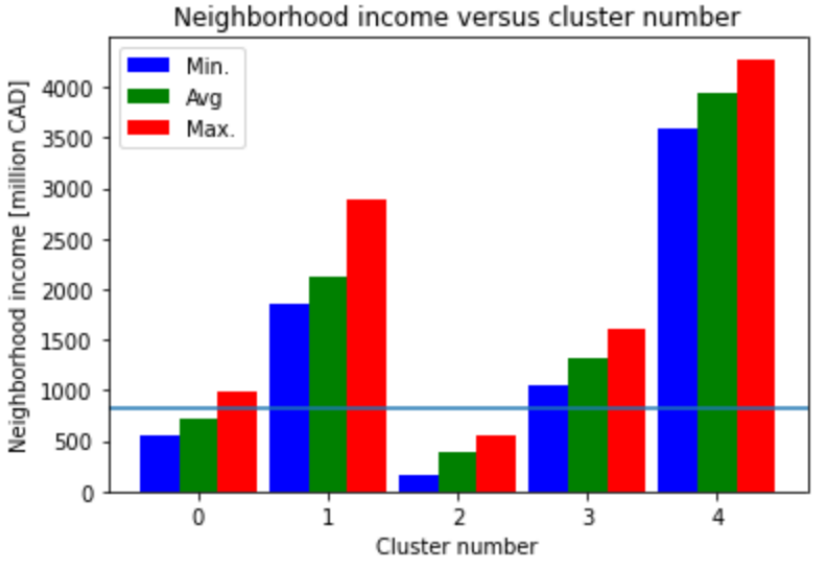
	Cluster Labels	Neighborhood	Neighbourhood Number	Population	Average income	Neighborhood Income	Longitude	Latitude	Number_of_restaurants
0	0	Agincourt North	129	23505	30414	714881070	-79.266712	43.805441	7
1	0	Agincourt South-Malvern West	128	19370	31825	616450250	-79.265612	43.788658	14
2	2	Alderwood	20	9915	47709	473034735	-79.541611	43.604937	0
3	1	Annex	95	25615	112766	2888501090	-79.404001	43.671585	25
4	3	Banbury-Don Mills	42	22335	67757	1513352595	-79.349718	43.737657	12

Clusters are marked by different colors on the map:



Cluster number	Color
0	Red
1	Violet
2	Light blue
3	Light green
4	Orange

Let's analyze clusters. Neighborhood income versus cluster:

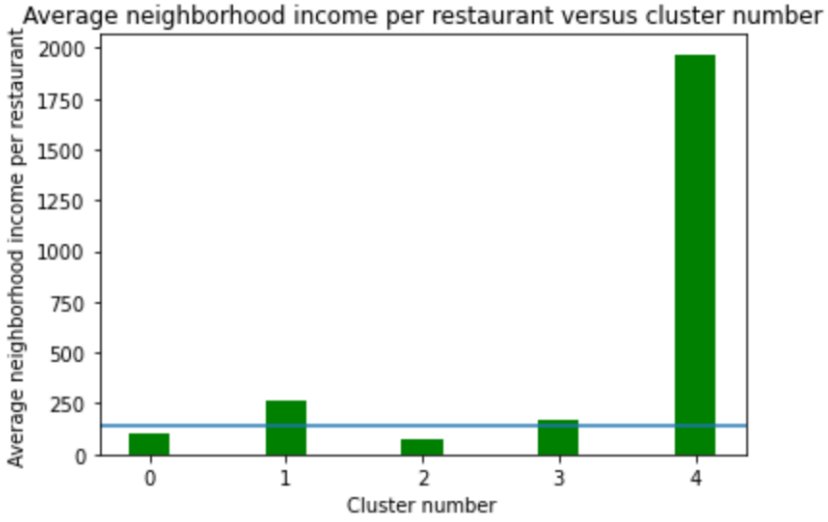


Average neighborhood income is 816.5 million $ (marked by horizontal line on above chart). We can see that average neighborhood income for clusters 0 and 2 is below average, so we will concentrate on other clusters as potential customers have more free resources to spend in restaurants.

Let’s analyze average number of restaurants in neighborhood versus cluster:

Average number of restaurants in neighborhood in Toronto is 6. Average number of restaurants in neighborhood for clusters 0 to 3 is not far from the average, for cluster 4 is significantly lower – this means that there is less competition.

Let's check average neighborhood income per restaurant:



Clearly cluster 4 is far above the average meaning that potential income of the new restaurant will be also above the average. Let’s examine cluster 4:

	Cluster Labels	Neighborhood	Neighbourhood Number	Population	Average income	Neighborhood Income	Longitude	Latitude	Number_of_restaurants
104	4	Rosedale-Moore Park	98	17285	207903	3593603355	-79.379669	43.68282	0
122	4	Waterfront Communities-The Island	77	60620	70600	4279772000	-79.377202	43.63388	5

There are only two neighborhoods in cluster 4: Rosedale-Moore Park and Waterfront Communities-The Island.

4. Results

As we see from above analysis, neighborhoods joined as cluster 4 are the potential best places to open an restaurant. Cluster 4 contains only two neighborhoods out of 140 in Toronto.

One of them, Rosedale-Moore Park has no restaurants at all, so it is the first choice (no competition, so highest neighborhood income per restaurant if there will be one).

Looking at average income, it can be quite expensive place, so as the second choice it is recommended to check possibilities in Waterfront Communities-The Island.

5. Discussion

Joined data from Toronto Open Data database about Toronto population with venues data from Foursquare API created a dataset which with help of K-mean clustering gave clear recommendation to choose location of an restaurant. However, as last census data comes from year 2016, it is not up to date and there can be differences in situation in year 2020. However, newer data is not available.

For final selection out of two neighborhoods it would be recommended to check other data, e.g. perform on-site reviews about spending on eating-out.

It would be also interesting to check why in some neighborhoods there are no restaurants.

It must be also taken into account that venue data is changing fast, four days before finishing this report there was max. 36 restaurants per neighborhood, this value changed to 37 at the date of finishing this report. Newest possible data should be taken as source for analysis.

6. Conclusion

Publicly available data sources can give enough data to build solid dataset as the ground for data science methods. Local government databases gives rich set of information about the population of Toronto. There must be performed much preparation, filtering, cleaning but finally it is possible to get right dataset. Foursquare API is helpful to get information about venues, however it is dynamically changed and updated. Filtered and joined data is analyzed using data science methods to obtain answer to real life question like “where to open a restaurant”.