The Battle of Neighborhoods

Introduction and Business Problem

Around the world urbanization is a megatrend which can observe in all countries around the world, The result are cities which have a very big land area and a high population density. Through this development such cities become more and more confusing and within in a city there could are big differences between each neighborhood.
In particular for founder of businesses like restaurants, gyms or stores it could be a nearly impossible task to find the right location for their new business.

For such founder here is a solution to simplify this problem. There are a lot of different criterias which influences whether the location of the business is promising. The method described here uses the following criterias to evaluate a location:

1. Crimes within a neighborhood
2. Population density of a neighborhood
3. Number of competitors within the neighborhood

The founder gets a list of neighborhoods within it is worth to start a new business. Moreover if the founder has already a location he or she can check if it was good choice.

The method was used for the first time to evaluate the City of Toronto but it can be used for each other city.

1. Description of the Data

In this section each data source is briefly described. For the evaluation of the City of Toronto there are the following data sources used.

1.1 Dataset about Neighborhood Boundaries

For a good estimation of the data it is helpful to visualize the data and the results with a choropleth map. For this map type a geometry file is necessary.

The City of Toronto operates a data portal which provides such a file:
https://open.toronto.ca/dataset/neighbourhoods

The second important use is to compute the land area from the boundaries data. The data is given in the World Geodetic System 1984 (WGS84).

For a direct access the data was downloaded and was placed on github:
https://github.com/DirkFritz/Coursera_Capstone/blob/master/Neighbourhoods.geojson

IMPORTANT: For the use of the file with jubyter and folium the browser Firefox is necessary. If you uses chrome the map wouldn't plot. The file size is seemingly to big.

1.2 Dataset for Crime Statistics

The Police of Toronto has public safety data portal. The service provides a detailed listing for all crime types for each neighborhoods which happens in the last years.

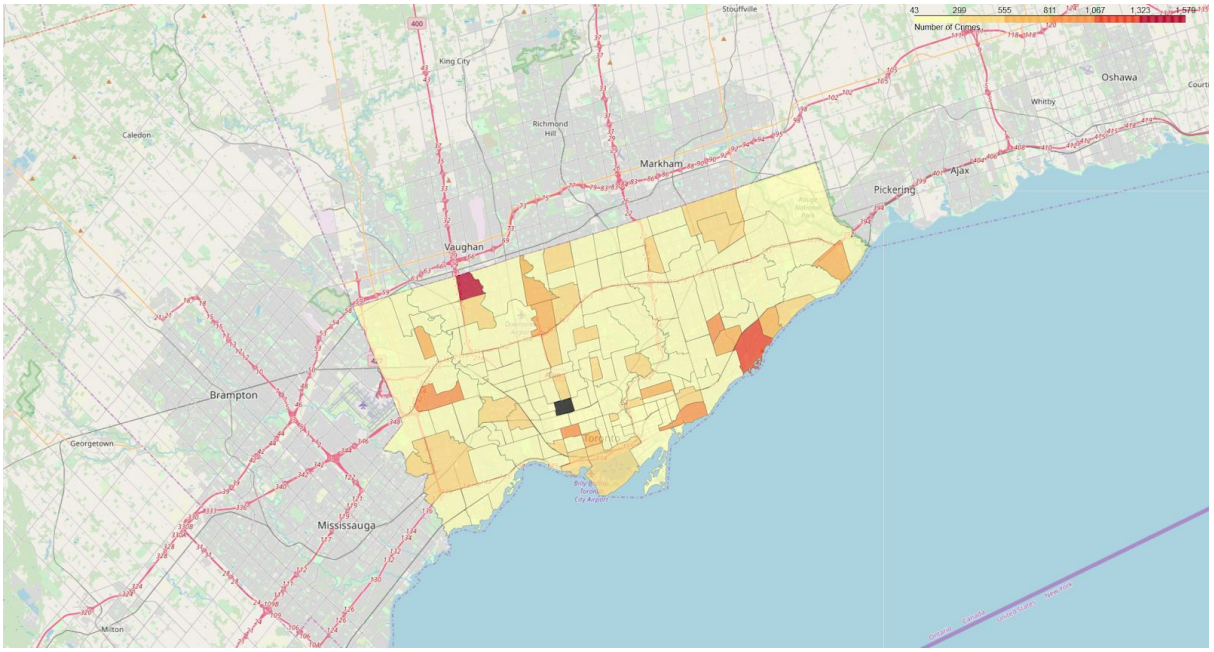
The link to the data portal is:
https://data.torontopolice.on.ca/datasets/neighbourhood-crime-rates-boundary-file-/data?geometry=-80.334%2C43.526%2C-78.479%2C43.874

The provided dataset has the following crime types:

1. Assaults
2. Auto Thefts
3. Breaks and Enters
4. Robberies

5. Theft-Overs
6. Homicide

For the evaluation the sum of all crime types is used. The download of the dataset can accessed over the following link:
https://github.com/DirkFritz/Coursera_Capstone/blob/master/Neighbourhood_Crime_Rates_Boundary_File.csv

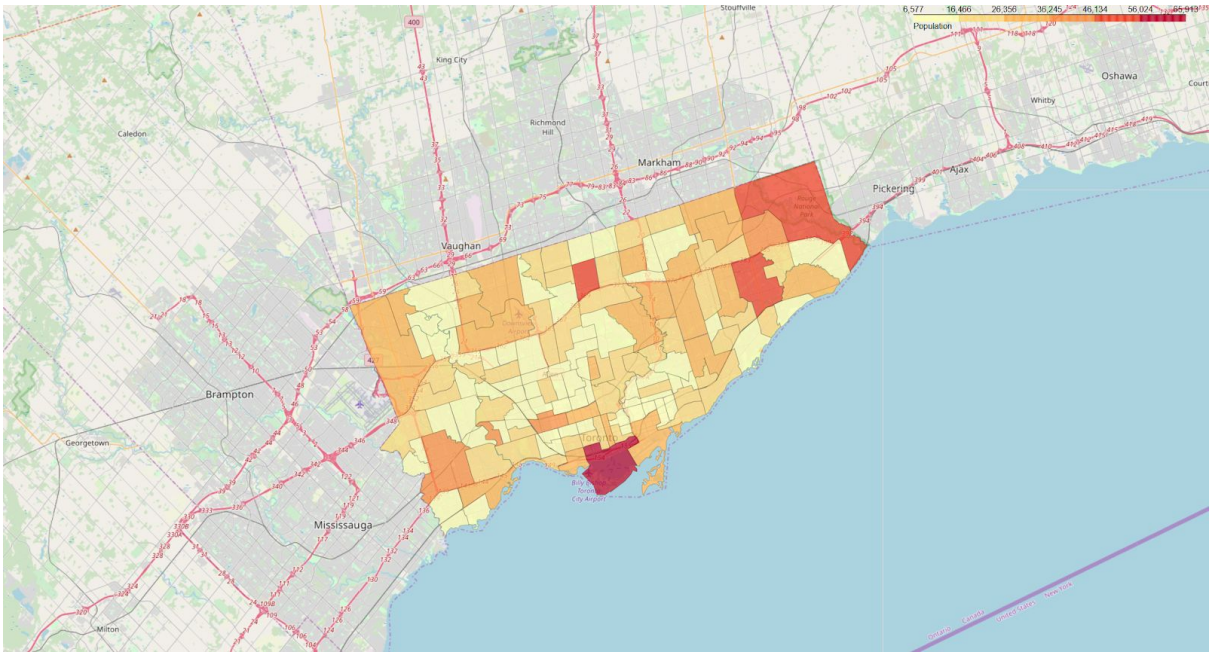


Map with the overall number of crimes

1.3 Dataset about the Neighborhood Profile

For the criteria population there is also data provided by The City of Toronto's Open Data Portal.
https://open.toronto.ca/dataset/neighbourhood-profiles/

In the dataset are a lot further information of each neighborhood.



Map of the population for each neighborhood

The complete file can be directly accessed here:
https://raw.github.com/DirkFritz/Coursera_Capstone//master/neighbourhood-profiles-2016-csv.csv

Dataset about the Number of Competitors

The number of competitors is important criteria if a further business at a specific location shall be economically. For getting this data the Foursquare API is used:
https://developer.foursquare.com/places

Foursquare is a location-based recommendation service for restaurants and other places.

For the analysis the API function explore is used. In the following evaluation the restaurant type “pizza place” is chosen as an example. With the Foursquare API places can be clearly identified by an id.
The category “Food” the correspond to 4d4b7105d754a06374d81259 and a pizza place is 4bf58dd8d48988d1ca941735.

Venue Category	Venue Id
Chinese Restaurant	4bf58dd8d48988d145941735
Sandwich Place	4bf58dd8d48988d1c5941735
Fast Food Restaurant	4bf58dd8d48988d16e941735
Pizza Place	4bf58dd8d48988d1ca941735
Fried Chicken Joint	4d4ae6fc7a7b7dea34424761
Wings Joint	4bf58dd8d48988d14c941735

Result of a Foursquare request for category food

2. Methodology

In this section is described how the evaluation of a business location for each neighborhood works. There are two step to get the result:

1. Collecting all relevant data to one dataset
2. Cluster the data with K-means algorithm

For the further analysis the following data was acquired:

1. Name of the neighborhood
2. Geographic position
3. Number of crimes
4. Land area
5. Population
6. Population density
7. Number of pizza places for each neighborhood

2.1 Collecting the data

Name and geographic position

The first step is to get the name of each neighborhood and it geographic position, so that the exploration of the data can be started.

The geographic position can be directly read into a dataframe from the file Neighbourhoods.geojson**.** The geographic data is in WGS84 format.

AREA_SHORT_CODE	AREA_NAME	LONGITUDE	LATITUDE
94	Wychwood (94)	-79.425515	43.676919
100	Yonge-Eglinton (100)	-79.403590	43.704689
97	Yonge-St.Clair (97)	-79.397871	43.687859
27	York University Heights (27)	-79.488883	43.765736
31	Yorkdale-Glen Park (31)	-79.457108	43.714672

Example from Neighborhoods.geojson file

Number of Crimes

From the file Neighbourhood_Crime_Rates_Boundary_File.csv the number of crimes is read out. In the file are data for the year 2016, 2017 and 2018. For the analysis data from year 2018 is used. The evaluated number of crimes is the sum of all listed crime types.

AREA_SHORT_CODE	AREA_NAME	LONGITUDE	LATITUDE	Number of Crimes
129	Agincourt North (129)	-79.266712	43.805441	513.0
128	Agincourt South-Malvern West (128)	-79.265612	43.788658	154.0
20	Alderwood (20)	-79.541611	43.604937	223.0
95	Annex (95)	-79.404001	43.671585	154.0
42	Banbury-Don Mills (42)	-79.349718	43.737657	81.0

Example from Neighborhood_Crime_Rates_Boundary_File.csv

Land Area, Population and Population Density

For determining the land area also the Geojson file Neighbourhoods.geojson is used. The boundaries in the file are polygons in WSG84 format. From this polygons the land area in square meter is calculated. The population of each neighborhood can be read from the file **neighbourhood-profiles-2016-csv.csv**.

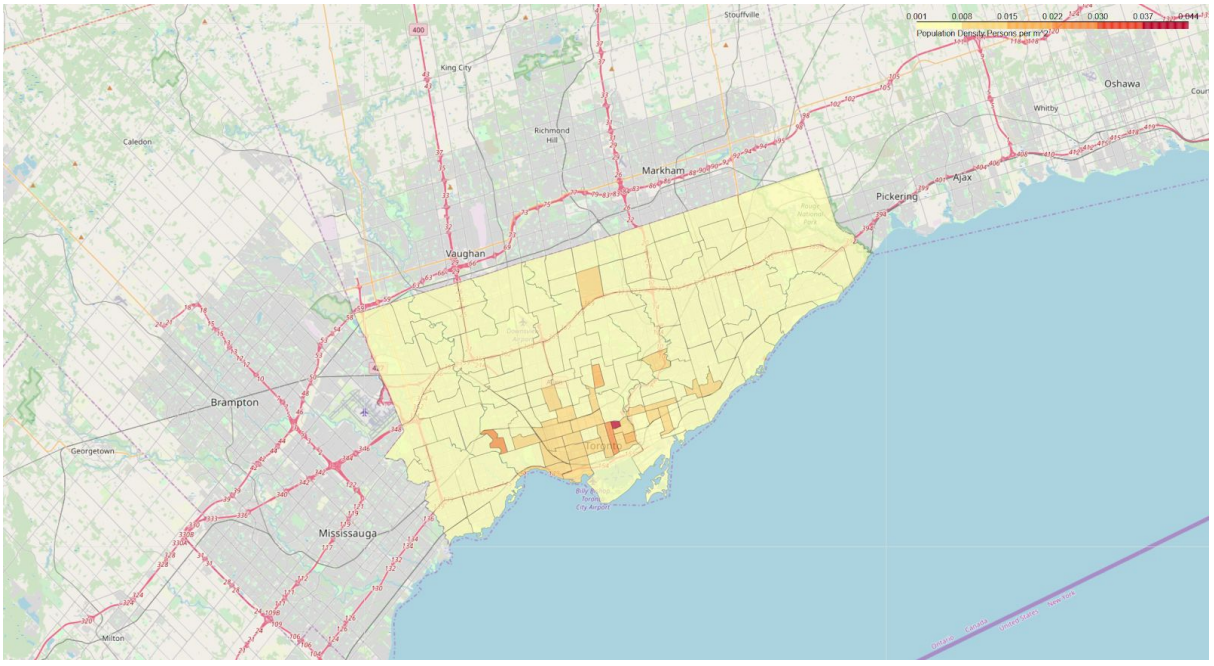
For the later analysis of a place like a restaurant the population density is a more important criteria as the land area or population separately.

The population density is calculated by

$$\text{population density} = \text{land area} / \text{population}$$

AREA_SHORT_CODE	AREA_NAME	LONGITUDE	LATITUDE	Number of Crimes	Area m^2	Population	Person Per m^2	Probability Venue	Sum Venue	Labels	Persons per m^2
129	Agincourt North (129)	-79.266712	43.805441	513.0	7264393.0	29113	0.004008	1.000000	1	3	0.004008
128	Agincourt South-Malvern West (128)	-79.265612	43.788658	154.0	7875876.0	23757	0.003016	0.500000	1	0	0.003016
20	Alderwood (20)	-79.541611	43.604937	223.0	4980675.0	12054	0.002420	1.000000	2	0	0.002420
95	Annex (95)	-79.404001	43.671585	154.0	2791395.0	30526	0.010936	0.500000	1	2	0.010936
42	Banbury-Don Mills (42)	-79.349718	43.737657	81.0	10045354.0	27695	0.002757	0.666667	2	0	0.002757

Example of the calculated population density



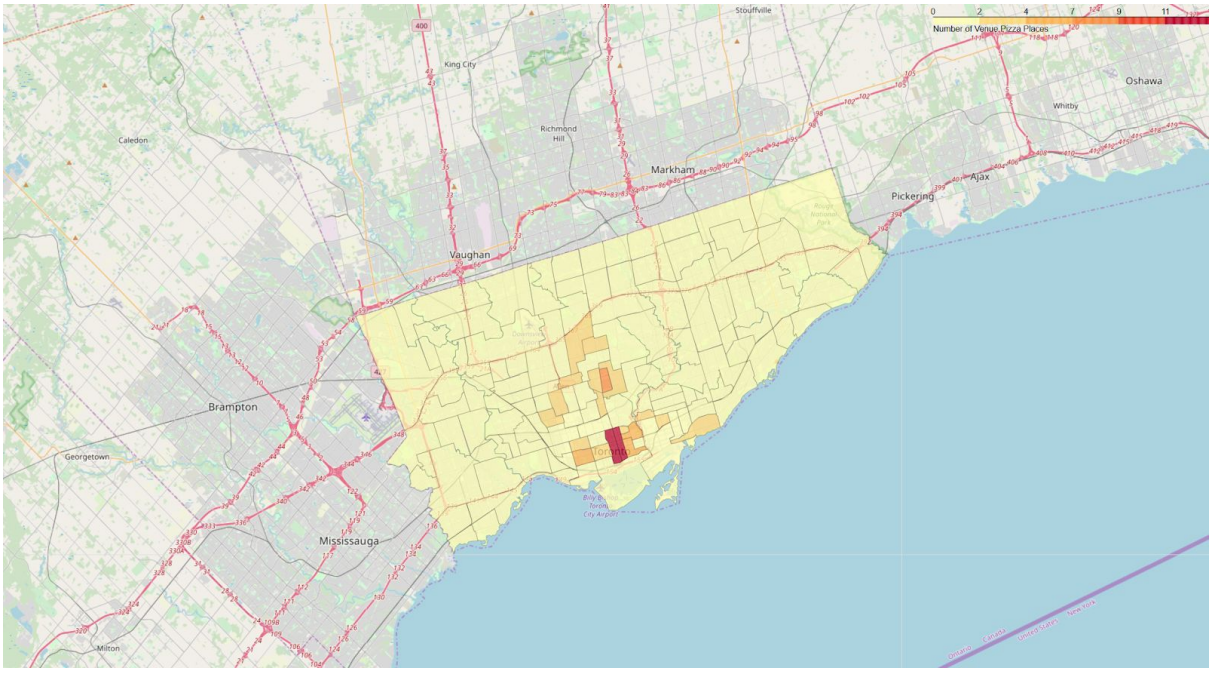
Population Density City of Toronto

Number of Pizza Places for each Neighborhood

The Foursquare API offers the function explore. This function allows to find all venues within a specific radius around a position. For this evaluation the venue “pizza place” is chosen.
With the geographic position data from Neighbourhoods.geojson and with a defined radius of 500m we get all necessary information for following evaluation.

AREA_SHORT_CODE	AREA_NAME	LONGITUDE	LATITUDE	Number of Crimes	Area m^2	Population	Persons per m^2	Sum Venue
129	Agincourt North (129)	-79.266712	43.805441	513.0	7264393.0	29113	0.004008	1
128	Agincourt South-Malvern West (128)	-79.265612	43.788658	154.0	7875876.0	23757	0.003016	1
20	Alderwood (20)	-79.541611	43.604937	223.0	4980675.0	12054	0.002420	2
95	Annex (95)	-79.404001	43.671585	154.0	2791395.0	30526	0.010936	1
42	Banbury-Don Mills (42)	-79.349718	43.737657	81.0	10045354.0	27695	0.002757	2

Example for the number of venue pizza place

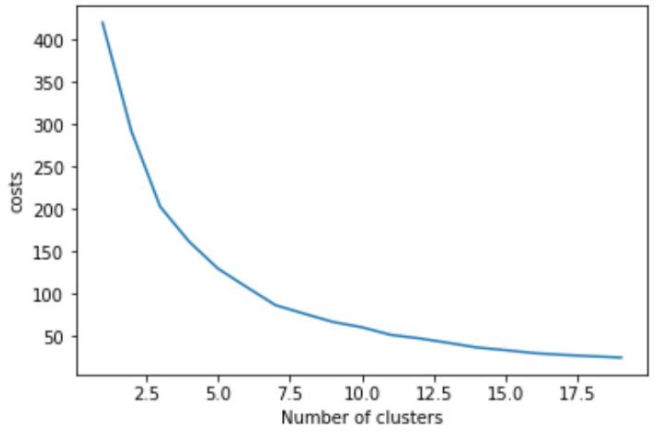


Number of pizza places City of Toronto

2.2 Cluster the data with K-means algorithm

After the data collection is accomplished the evaluation can start. Because the data is unstructured, an unsupervised learning algorithm should be the right tool to analyse the data. In the end, the neighborhoods should be segmented in different clusters. This clusters should us allow to estimate whether a cluster is more suitable for a pizza place as the other. For this task the algorithm K-means should give good results.

The K-means algorithm needs the number of clusters as parameter and the data should be scaled before. The elbow method showed that a good choice for the number of clusters should be between 3 and 6



Evaluation of K-mean with different number of clusters

After by hand evaluation of this range the number of cluster is defined to five, because the cluster have had the most clear differences.

AREA_SHORT_CODE	AREA_NAME	LONGITUDE	LATITUDE	Number of Crimes	Area m^2	Population	Persons per m^2	Sum Venue	Labels
129	Agincourt North (129)	-79.266712	43.805441	513.0	7264393.0	29113	0.004008	1	5
128	Agincourt South-Malvern West (128)	-79.265612	43.788658	154.0	7875876.0	23757	0.003016	1	0
20	Alderwood (20)	-79.541611	43.604937	223.0	4980675.0	12054	0.002420	2	0
95	Annex (95)	-79.404001	43.671585	154.0	2791395.0	30526	0.010936	1	1
42	Banbury-Don Mills (42)	-79.349718	43.737657	81.0	10045354.0	27695	0.002757	2	0
34	Bathurst Manor (34)	-79.456055	43.764813	674.0	4762402.0	15873	0.003333	0	5
76	Bay Street Corridor (76)	-79.385721	43.657511	59.0	1809967.0	25797	0.014253	13	3
52	Bayview Village (52)	-79.377117	43.776361	139.0	5161147.0	21396	0.004146	0	0
49	Bayview Woods-Steeles (49)	-79.382118	43.796802	96.0	4090512.0	13154	0.003216	0	0
39	Bedford Park-Nortown (39)	-79.420227	43.731486	135.0	5520505.0	23236	0.004209	4	1

K-means assigned each neighborhood a cluster label

After the execution of the K-means algorithm the neighborhoods are grouped by the cluster labels and sum upped with the mean function.

Labels	LONGITUDE	LATITUDE	Number of Crimes	Area m^2	Population	Persons per m^2	Sum Venue
0	-79.356925	43.729088	810.142857	4.016187e+06	17517.000000	0.005073	0.857143
1	-79.410995	43.718262	187.433735	5.914220e+06	20042.506024	0.004284	0.361446
2	-79.381349	43.669528	148.000000	1.483088e+06	26825.250000	0.018447	9.500000
3	-79.392426	43.686745	220.675676	2.365391e+06	17645.837838	0.008498	2.108108
4	-79.435646	43.663522	303.500000	1.103498e+06	31304.000000	0.034273	1.500000

Result K-means grouped by cluster labels and the mean function

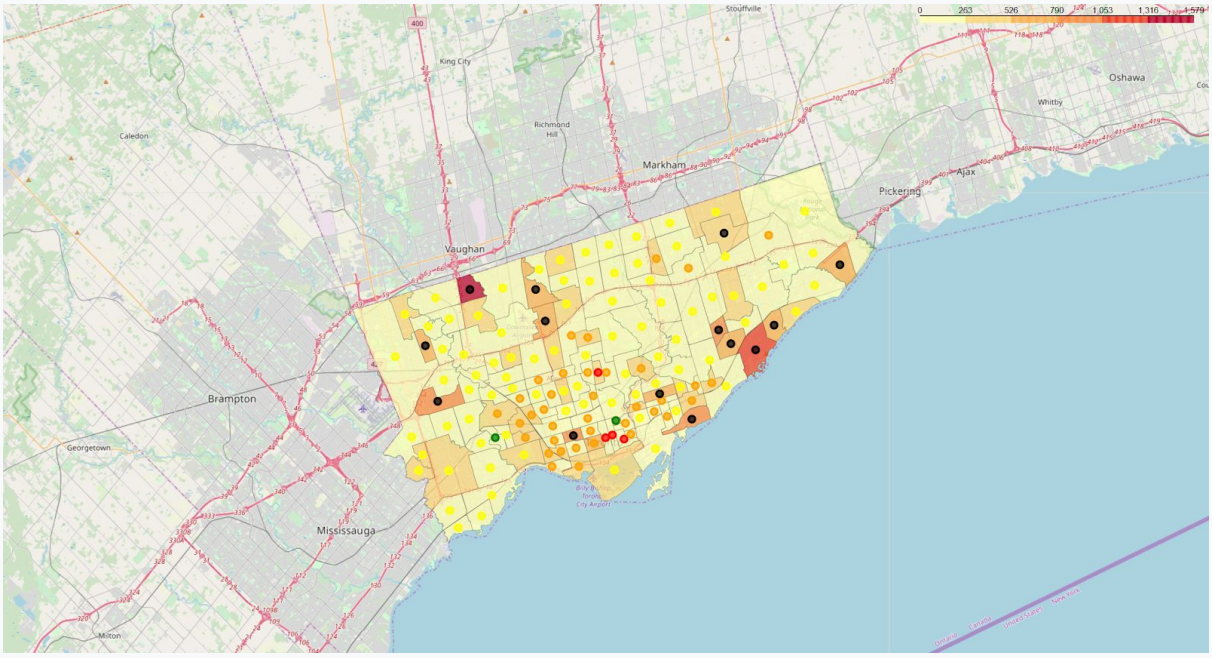
3. Results

The result of K-means algorithm can be interpreted as follows.

Cluster 0: High crime, low population density, a few competitors
Cluster 1: Low crime, low population density, a few competitors
Cluster 2: Low crime, medium population density, a lot of competitors
Cluster 3: Low crime, low population density, some competitors
Cluster 4: Low/medium crime, high population density, some competitors

The neighborhoods with label 0 aren't suitable because the crime is very high. The cluster with label 1 and 3 are also not suitable. In this neighborhoods the number of crimes are low but the population density is also very low. Cluster with label 2 has low number of crimes and a medium population density but the number of competitors are extremely high. Therefore, this neighborhood can't be recommended as a good location for a new pizza place.

Cluster with label 4 has low to medium crime but a very high population density and not so much competitors. Therefore, this neighborhoods can be recommended for new pizza places.



Result K-means algorithm: Label 0: Black, Label: 1 Yellow, Label 2: Red, Label 3: Orange, Label 4: Green (recommended)

4. Discussion and Conclusion

We see in the map above that we have a concentration of pizza places by only a few neighborhoods, even though, the population density isn't so high like in other neighborhoods. The question is now, what are the reason for this? Are there rational reasons or have the owner chosen this place because their think if there is already a other restaurant then it should be a good place for an new restaurant.

On other hand it seems to be that neighborhoods with high crime aren't a good environment for such businesses. Just like that the neighborhoods with a low population density seems to be also not suitable places for restaurants like pizza places.

The really interesting point is that the K-means algorithm could identified two neighborhoods which seem to be a very suitable location for a new pizza place.

The conclusion of the evaluation is, that data science can be a valuable contribution to find good location in big cities. It is clear that a suitable location depends of many further criterias. This criterias could be average income, are there sights nearby or which ethnic groups live there and so on. Therefore for safe estimation of location more research is necessary - but the results of this first approach are promising.