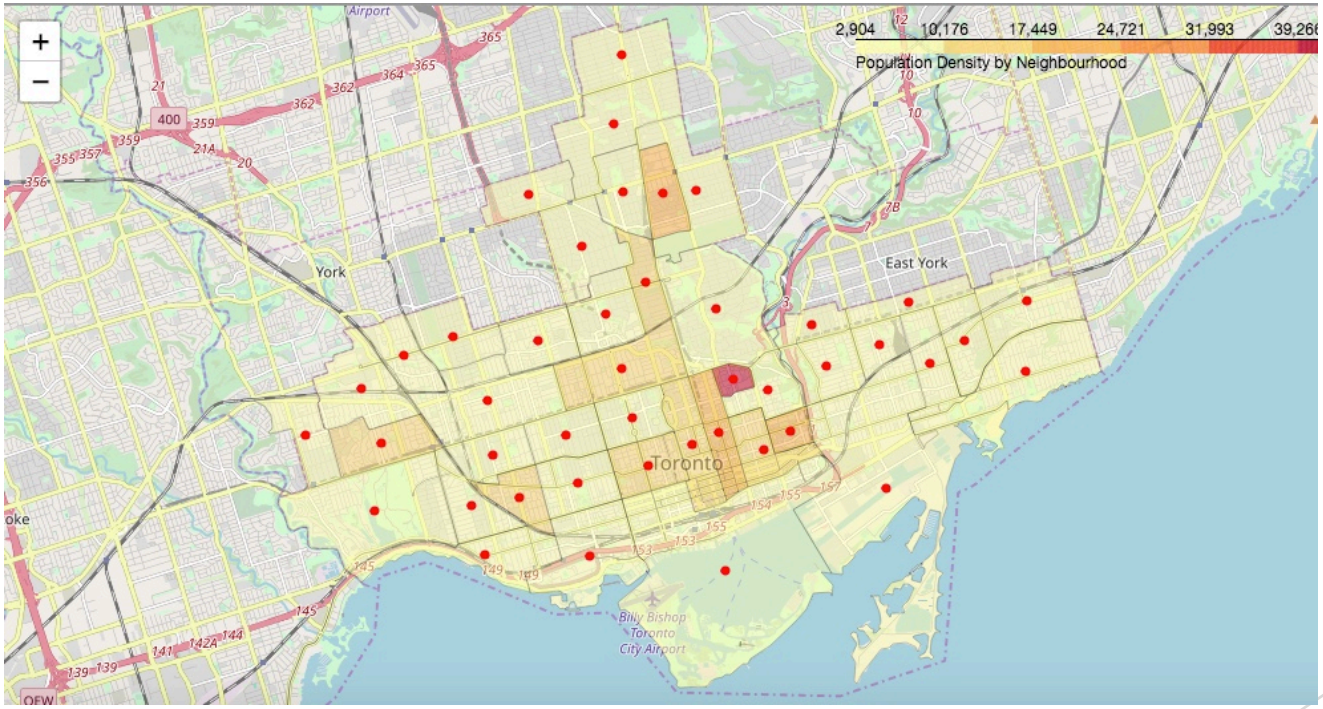
Toronto & Venue Location Choices

Marc van der Valk

February 24 2019

Toronto & It's Neighborhoods

Toronto is Canada's largest city with a population of 2,7 million (2016). The area researched is the former city of Toronto with 43 neighborhoods, as shown below with population densities.



Introduction & Business Problem

- Finding the right business location is one of the primary steps in preparing to set up a new business. It is not always an easy task.
- The aims of this project is to help current and future business owners in the process of selecting business locations in certain neighborhoods.
- By using data from the location based social network service Foursquare, as well as neighborhood area statistics it should be possible to recommend possible business locations.
- By using a machine learning algorithm called K-Means Clustering, an attempt is made to classify neighborhoods based on local venue categories and numbers. Statistics of average household income and population density are also included.
- The geo-spatial features (neighborhood boundaries) and relative location to the center of the city are also taken into account

What is a neighborhood?

- According to the website of the city of Toronto, the definition of a **neighborhood** is:
- An area that respects existing boundaries such as service boundaries of community agencies, natural boundaries (rivers), and man-made boundaries (streets, highways, etc.)
- Small enough for service organizations to combine them to fit within their service area.
- Represent municipal planning areas as well as areas for public service like public health.
- A neighborhood has a population roughly between 7,000 and 12,00 people.

Data used:

- Neighborhood geo-spatial boundaries files from the city of Toronto's data, research and maps site: https://www.toronto.ca/city-government/data-research-maps/
- Neighborhood statistics with population, area size in km2, average household income from the city of Toronto's neighborhood wellbeing site: http://map.toronto.ca/wellbeing/
- A list of boroughs and their neighborhoods of the city of Toronto from Wikipedia: https://en.wikipedia.org/wiki/List_of_city-designated_neighbourhoods_in_Toronto
- Venue categories from Foursquare's API
- Venues by geo-location coordinates from Foursquare's API
- Note: Foursquare is a location based social network where users visit (check-in) to venues , give tips etc. Venues are geo-located and categorized according to the venue type: for example restaurants are categorized under Food, stores and ships under Shops & Services etc.

Data Wrangling

... Cleaning, fixing and reformatting

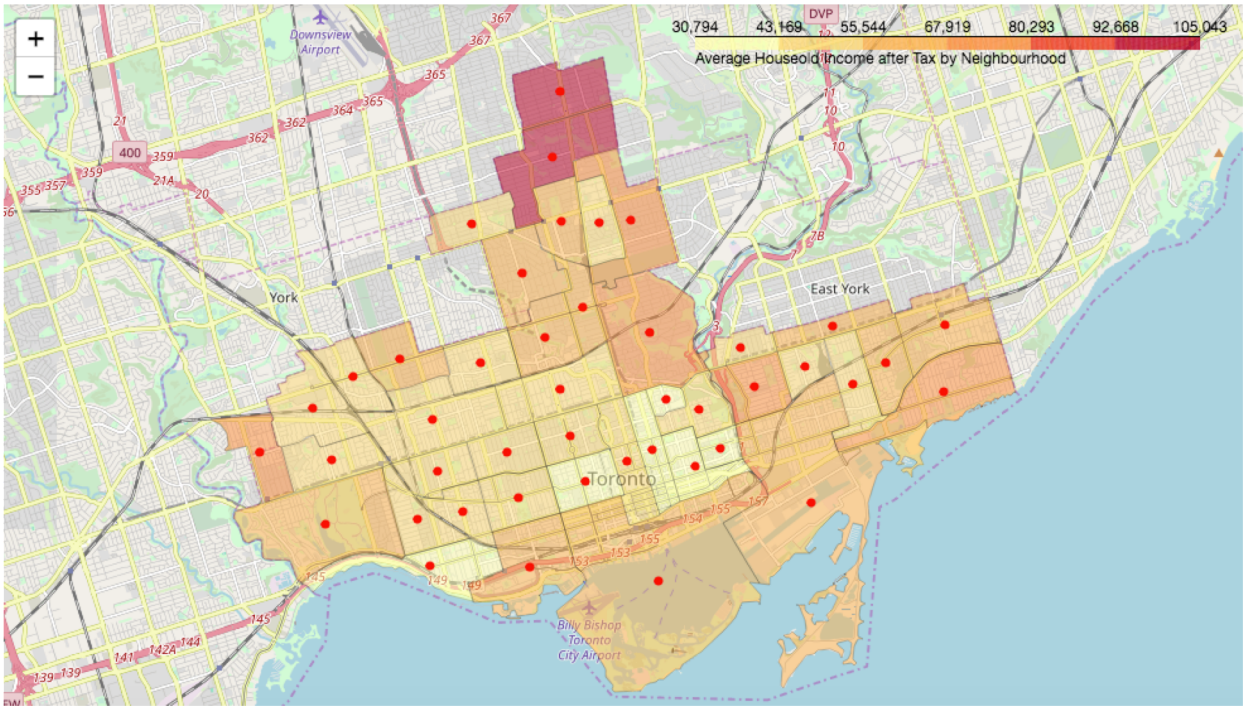
- In order to use the neighborhood's boundaries data for map visualizations it was necessary to convert it to a so-called GEO JSON format. The data could be imported into a tabular format. Only the column headers needed to be renamed to be consistent
- The neighborhood statistics file did not have any missing data but the column headers needed to be renamed to be consistent with the other data
- By using the Wikipedia list of neighborhoods, it was possible to filter both data tables on neighborhoods only within the boundaries of the former city of Toronto (from 140 down to to 44 neighborhoods)
- The Foursquare venue category data needed to be converted from JSON format to a tabular one. (JSON = JavaScript object notation format, a free format text based database based on key value pairs). 8 main categories with 455 subcategories found.
- The Foursquare venues data was more tricky. The geo-coordinates of the venue data needed to be located within a neighborhood to assign the correct neighborhood, the venue categories were incomplete and needed fixing
- In the last stage of data processing the tables needed to be aggregated to neighborhood level for further analysis
- Resulting in a table with 44 neighborhoods (rows) and 10 columns: 8 categories, population density and average household income

Example of the venue table

	CDN_Number	Neighbourhood	Venue	Latitude	Longitude	SubCategory	Category
0	095	Annex	Rose & Sons	43.675668	-79.403617	American Restaurant	Food
1	096	Casa Loma	Ezra's Pound	43.675153	-79.405858	Café	Food
2	095	Annex	Roti Cuisine of India	43.674618	-79.408249	Indian Restaurant	Food
3	095	Annex	Fresh on Bloor	43.666755	-79.403491	Vegetarian / Vegan Restaurant	Food
4	095	Annex	Playa Cabana	43.676112	-79.401279	Mexican Restaurant	Food

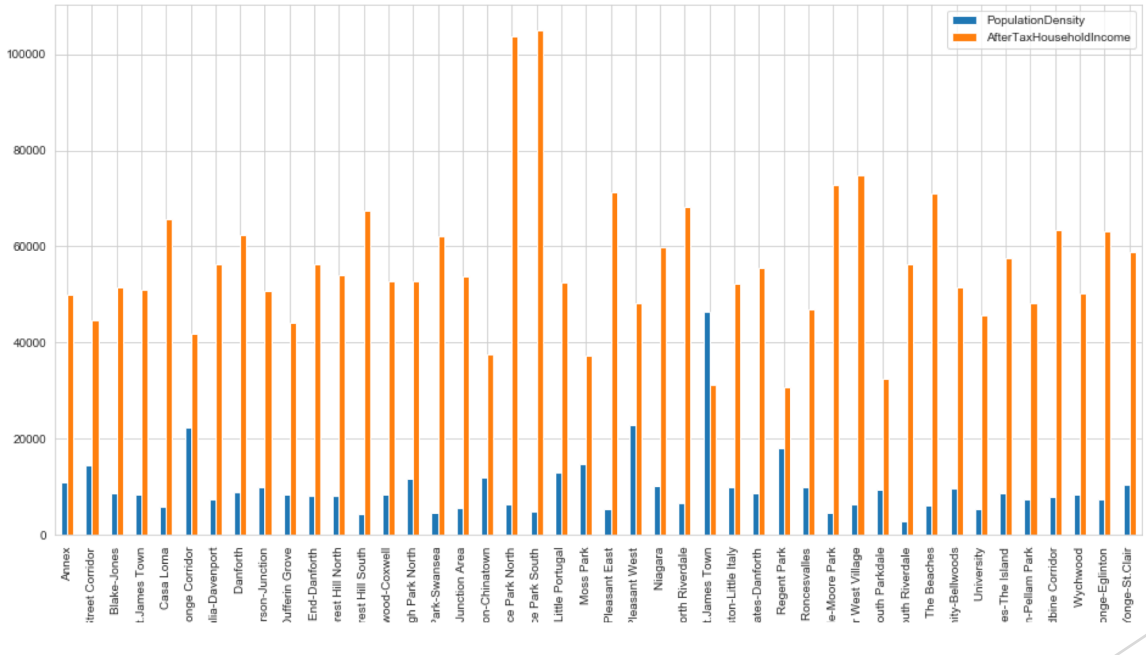
Neighborhood Average Household Income

Visualization of average household income by neighborhood.
The scale is from light yellow (lower) to darker red (highest average income)
There seems to be a trend of higher average income towards the outskirts of the research area with lower average income towards the center of the city.



Population density (km2) by Average Household Income by Neighborhood

Although the graphic is somewhat small, the orange lines depict the average income and the blue lines: population density. Visible in most cases is that lower population density usually corresponds with higher average income

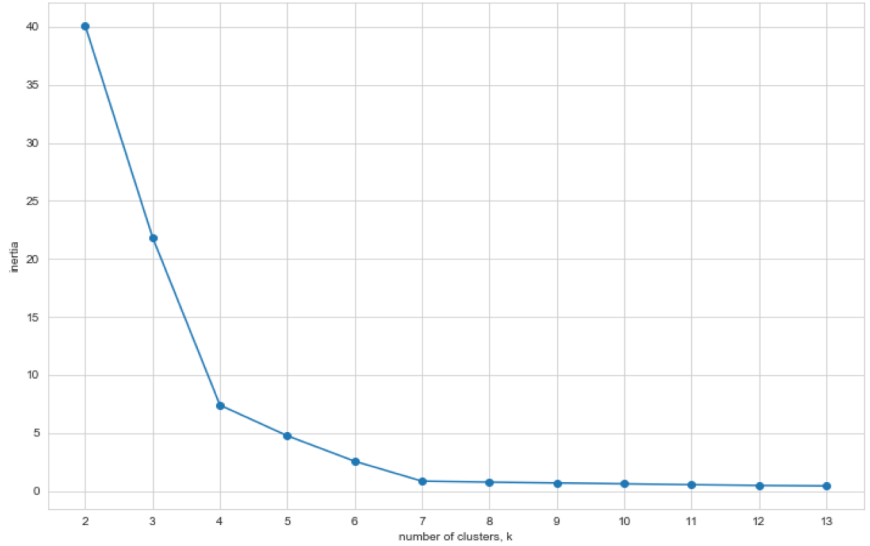


Classification of Neighborhoods using the K-Means clustering algorithm

- The K-Means algorithm is an unsupervised machine learning algorithm used to classify data into groups based on similar features (=attributes). The process is called clustering. The groups are assigned cluster numbers when done.
- Called unsupervised because the algorithm detects the similarities by itself and there is no known classification or outcome beforehand.
- K-Means can give you sights to data that you can not always detect beforehand. Especially if you have many features.
- It is up to the researcher to come up with a meaningful description for each cluster assigned by analyzing the outcome.
- The only thing you need to do before running the algorithm is to find the optimal number of clusters to use with the algorithm

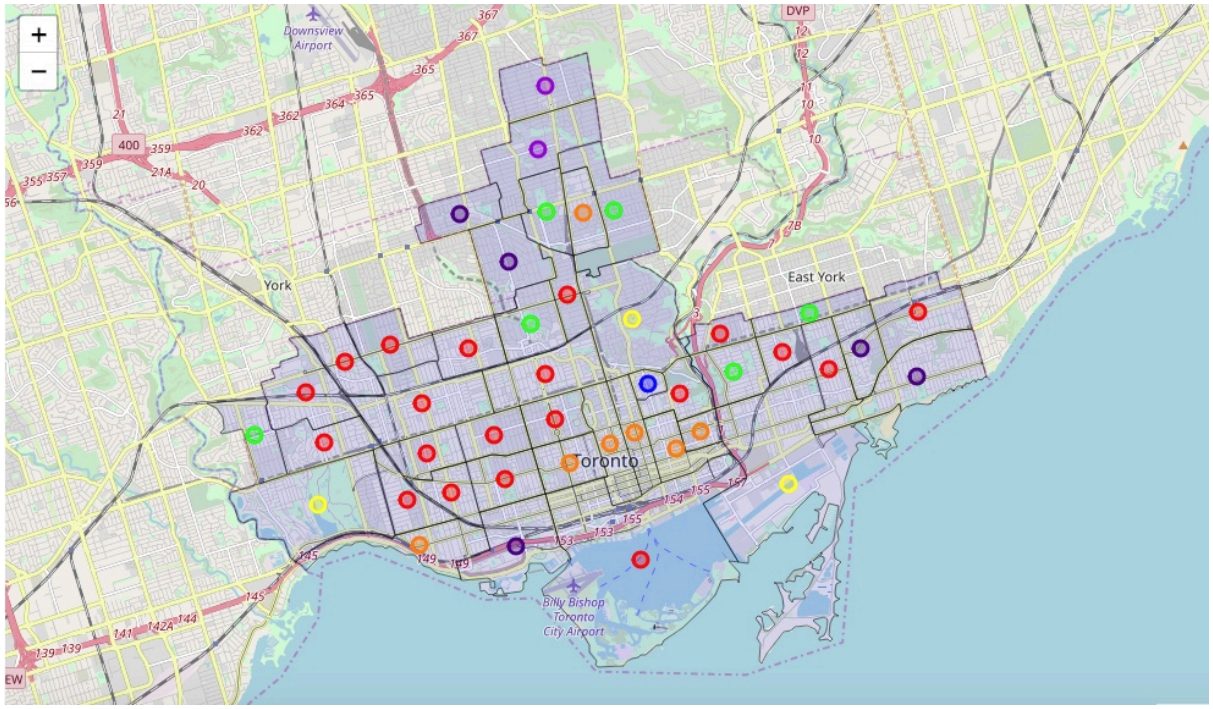
Finding the optimal number of clusters for the K-Means model

The process is to run the model (= algorithm) several times using a range of numbers. The model returns an inertia variable each time. As the number increases, the inertia value flattens out, shown below in the diagram. This would be the optimal number. In this case 7 clusters was found to be optimal



Visualizing the clustering on a map

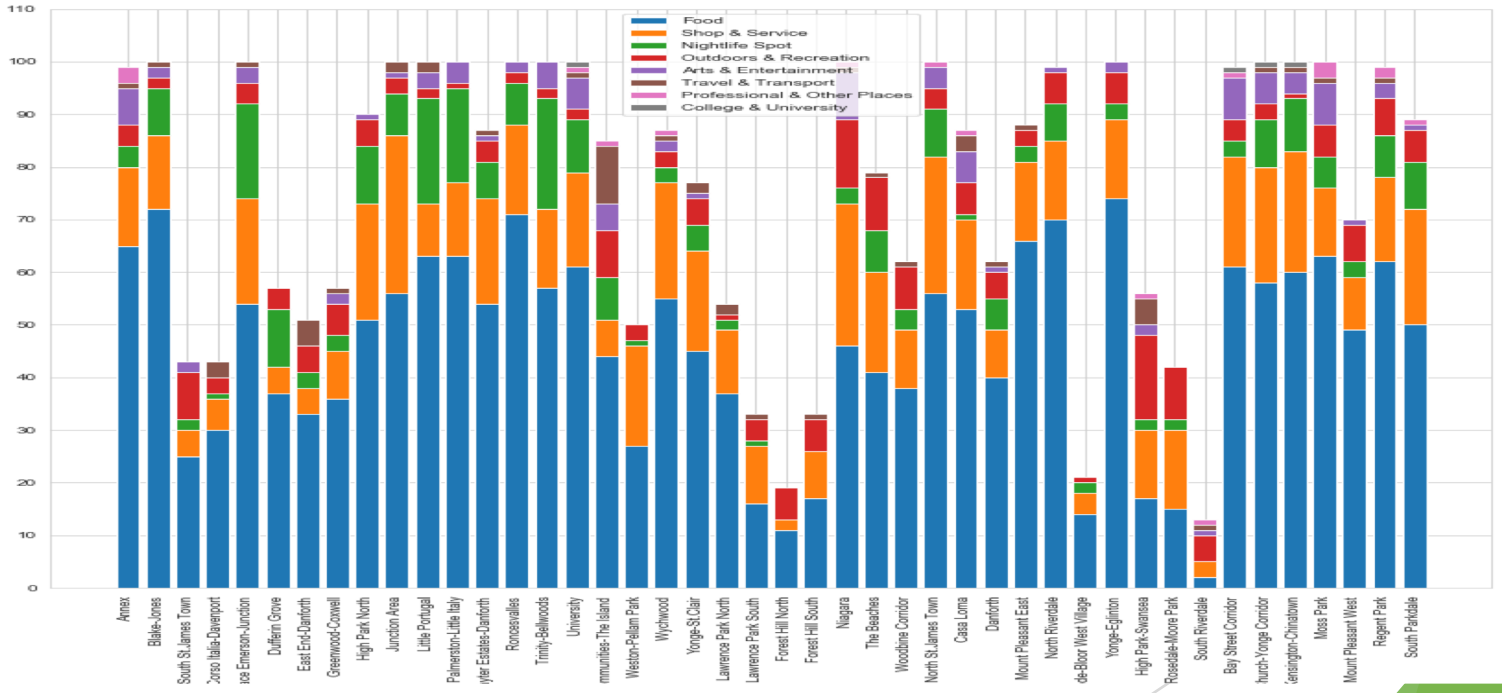
Each color dot represents a cluster. Already you can see some similar clusters on the map. Red is abundant, orange in downtown area. Yellow in neighborhoods with parks



- Red = Cluster 0
  - Most prominent, grouped around the downtown area
- Light Purple = Cluster 1
  - Most northern neighborhoods with a low population density and high avg. income
- Dark Purple = Cluster 2
  - Appear to be located on outer neighborhoods of the research area or close to a recreational area, needs further investigation
- Blue = Cluster 3
  - Only one neighborhood represented, needs some further investigation
- Green = Cluster 4
  - Also appear to be located on outer neighborhoods of the research area
- **Yellow** = Cluster 5
  - Seems to be neighborhoods with a larger park or recreational facilities
- Orange = Cluster 6
  - Mainly grouped in the downtown area

Analyzing the most prominent venue categories by neighborhood

By visualizing the venue categories by venue count by neighborhood more insight was given to the reasoning behind clustering, The Food category (blue) is the most prominent overall, then orange = Shops & Services, green = Nightlife Spots and red = Outdoors & Recreation. The ratio between the venue categories is also useful. For example the first 18 bars belong to cluster 0 which have relatively many nightlife venues in comparison.



Observations from the clustering map and graph

- **From Annex to Wychwood (red) - Cluster 0 -** neighborhoods are in the majority. They have a relatively large number of food related venues. And shops, services and nightlife venues are also prominent. The neighborhoods are located to the west/north-west of the downtown area, as well as to the east.
- **Lawrence Park North and South (light purple) - Cluster 1 -** Both neighborhoods have a high average income and low population density which would lead to a conclusion that these are mainly residential areas.
- **Forest Hill North to Woodbine Corridor (dark purple) - Cluster 2 -** There appears to be a relatively large number of shops and services in these neighborhoods. There are also relatively less nightlife spots (bars, clubs etc.) compared to the cluster 0 (red) neighborhoods. The outdoors and recreational venues are prominent.
- **North St. James Town (blue) - Cluster 3 -** This neighborhood has been separately clustered due to the fact that there it has the highest population density of all the researched neighborhoods. It has a relatively high number of shops and service venues.
- **From Casa Loma to Yonge-Eglinton (green)** - Cluster 4 - The neighborhoods are located in the north and eastern part of the research area with one exception located in the far west. Mainly on the outskirts. There are a relatively large number of shops and services as well as outdoor and recreational venues. Nightlife venues are prominent as well.
- **High Park-Swansea to South Riverdale (yellow) - Cluster 6 -** These neighborhoods are located in recreational areas with parks or close to the waterfront and have a high percentage of outdoor and recreational venues and venues of the travel and transportation category (hotels, transportation hubs: bus, metro or train stations)
- **From Bay Street Corridor to South Parkdale (orange) - Cluster 6 -** These are neighborhoods in the downtown area of Toronto with a high number of venues in the food category like restaurants. Shops and services are also prominent as well as venues in the nightlife spot category

Dataframe of the neighborhood features for further analysis

A dataframe (=table format) was created representing the order of most prominent venue categories by neighborhood. The cluster number, average income and population density are also included. After analyzing the clustering map and stacked bar graph with venue categories by count, the table proved to be very useful in detecting more similarities. For example comparing the average income with the overall median income helped in detecting these.

	Neighbourhood	Cluster	AvgIncome	PopDensity	1st Category	1st # Venues	2nd Category	2nd # Venues	3rd Category	V
0	Annex	0	49912	10902.0	Food	65	Shop & Service	15	Arts & Entertainment	
2	Blake-Jones	0	51381	8586.0	Food	72	Shop & Service	14	Nightlife Spot	
3	Cabbagetown-South St.James Town	0	50873	8335.0	Food	25	Outdoors & Recreation	9	Shop & Service	
6	Corso Italia-Davenport	0	56345	7438.0	Food	30	Shop & Service	6	Outdoors & Recreation	
8	Dovercourt-Wallace Emerson-Junction	0	50741	9899.0	Food	54	Shop & Service	20	Nightlife Spot	

Overall Statistics

	AvgIncome	PopDensity
count	44.000000	44.000000
mean	55981.727273	9972.568182
std	15130.504763	7034.026401
min	30794.000000	2904.000000
25%	48171.000000	6336.750000
50%	53315.500000	8461.000000
75%	62678.250000	10153.500000
max	105043.000000	46538.000000

Analyzing the results

- The clustering proved to be useful to giving a (future) business owner information to where to locate a certain business
- There was a clustering around certain areas even though the geo-coordinates were not included. I was surprised to find that the downtown area had such a low average household income and two neighborhoods in the north were outliers regarding high average income. One neighborhood was by itself due to a very high population density
- The clustering resulted in neighborhoods being grouped around most prominent venue types, as well a location based on the center of the city.
  If a neighborhood has a large number of venues of a certain type that could be a good reason to locate there as well as the area is already known for this reason and there will consumers visiting for the same reasons.
  Lack of certain venues of a category could mean that that it might not be a good idea to locate in that neighborhood
- The type of neighborhood might be a good reason to locate a specific type of business, for example in an area with more recreational venues and parks you could be looking at venue like a café or day restaurant where visitors can get some refreshment and a bite to eat

Conclusion and thoughts for the future

- The combination of business location, machine learning and geo-spatial tools are beneficial in providing information for possible business location decisions
- Having more data sources like business rental prices, transport and person movement statistics, parking facilities, building usage, this would contribute to make even better machine learning outcomes. This research has just scratched the surface as to what is possible.
- You could add specific financial business data to the analysis: how much income does a certain type of business need to make profit, what are the minimum space requirements, where is the labour force located etc.
- There is a whole field of spatial analysis and algorithms which could be incorporated in location analysis