IBM Developer
SKILLS NETWORK

Applied Data Science CapstoneCapstone Project

Final report

Clustering restaurants in Toronto with help of Foursquare location data and kMeans Algorithm

DRAFT

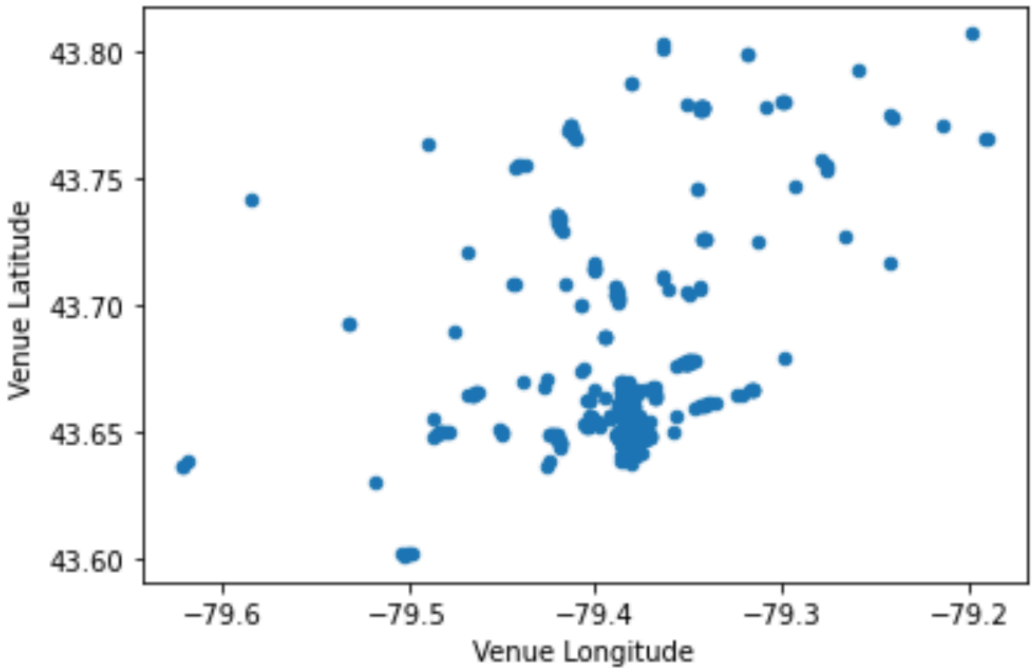
Introduction

In Berlin there are some touristic hotspots around certain streets or places. Those hotspots are mainly characterised by a high density of bars and restaurants. You can spend a whole night touring from restaurant to restaurant trying different dishes and drinks.
In your hometown you get to know those places by time. But what is, if you just moved to a new city or plan to do a restaurant trip during your holidays. You do not want to spend all your time in finding out, which restaurant you should try and which not.

This project aims to solve this problem for Toronto by finding answers to the following question: 1.) Are there some places / streets where the density of restaurants is higher than usual? 2.) Are there areas, where you can find a lot of restaurants with good ratings or are those restaurants speed all over the city?

The results of this project may help you to plan your next trip to Toronto, if you are a tourist, or just give you an overview of where high rated restaurants are located in Toronto.

As a motivation take a look at the picture below, which shows the distribution of 479 restaurants



Restaurants in Toronto; data retrieved with Foursquare location API

in Toronto. Where are high rated ones located/grouped?

Data

The goal is, to identify group restaurants in in Toronto based on the distance to each other and their rating.

To achieve this goal, the following information needs to be available:

a) Positional information for each neighbourhood in Toronto
b) Venue category
c) Positional information for each restaurant in form of longitudinal and latitudinal coordinates
d) Venue Id
e) Rating of the restaurant

Positional information for neighbourhood and restaurants

The position information of each neighbourhood is needed, because it will be used to find all venues in the corresponding neighbourhood. Because the restaurants will be grouped based on their Euclidean distance, their positional information is needed.
Both positional information is given in degrees of longitude and latitude.

The below picture shows an example result of querying the Foursqure API to find all venues in Toronto's neighbourhoods. It can bee seen, that the query returns all possible types of venues like Bars, Stores or Restaurants and their location based on latitude and longitude.

```
[16]: print(toronto_venues.shape)
      toronto_venues.head()
```

```
(2106, 7)
```

[16]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Malvern, Rouge	43.806686	-79.194353	Wendy's	43.807448	-79.199056	Fast Food Restaurant
1	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497	Royal Canadian Legion	43.782533	-79.163085	Bar
2	Guildwood, Morningside, West Hill	43.763573	-79.188711	RBC Royal Bank	43.766790	-79.191151	Bank
3	Guildwood, Morningside, West Hill	43.763573	-79.188711	G & G Electronics	43.765309	-79.191537	Electronics Store
4	Guildwood, Morningside, West Hill	43.763573	-79.188711	Sail Sushi	43.765951	-79.191275	Restaurant

Search query for all venues in Toronto

Venue category

The query for venues in Toronto results in a total of 2106 venues. Going straight forward could be to pick only those venues, which category equals “Restaurant”. But this would lead to a loss of much data since not all restaurants are categorised as such (see picture below of different restaurant types).

```
[21]: restaurant_cat=cat[cat.str.contains('Restaurant')]
      restaurant_cat.values

[21]: array(['American Restaurant', 'Asian Restaurant', 'Belgian Restaurant',
             'Brazilian Restaurant', 'Cajun / Creole Restaurant',
             'Caribbean Restaurant', 'Chinese Restaurant',
             'Colombian Restaurant', 'Comfort Food Restaurant',
             'Cuban Restaurant', 'Dim Sum Restaurant', 'Doner Restaurant',
             'Dumpling Restaurant', 'Eastern European Restaurant',
             'Ethiopian Restaurant', 'Falafel Restaurant',
             'Fast Food Restaurant', 'Filipino Restaurant', 'French Restaurant',
             'German Restaurant', 'Gluten-free Restaurant', 'Greek Restaurant',
             'Hakka Restaurant', 'Indian Restaurant', 'Italian Restaurant',
             'Japanese Restaurant', 'Korean BBQ Restaurant',
             'Korean Restaurant', 'Latin American Restaurant',
             'Mediterranean Restaurant', 'Mexican Restaurant',
             'Middle Eastern Restaurant', 'Modern European Restaurant',
             'Molecular Gastronomy Restaurant', 'Moroccan Restaurant',
             'New American Restaurant', 'Portuguese Restaurant',
             'Ramen Restaurant', 'Restaurant', 'Seafood Restaurant',
             'Sri Lankan Restaurant', 'Sushi Restaurant',
             'Taiwanese Restaurant', 'Thai Restaurant', 'Theme Restaurant',
             'Tibetan Restaurant', 'Turkish Restaurant',
             'Vegetarian / Vegan Restaurant', 'Vietnamese Restaurant'],
            dtype=object)

[22]: restaurants=toronto_venues['Venue Category'].isin(restaurant_cat.values)
      df_restaurants=toronto_venues[restaurants]
      df_restaurants.shape

[22]: (479, 7)
```

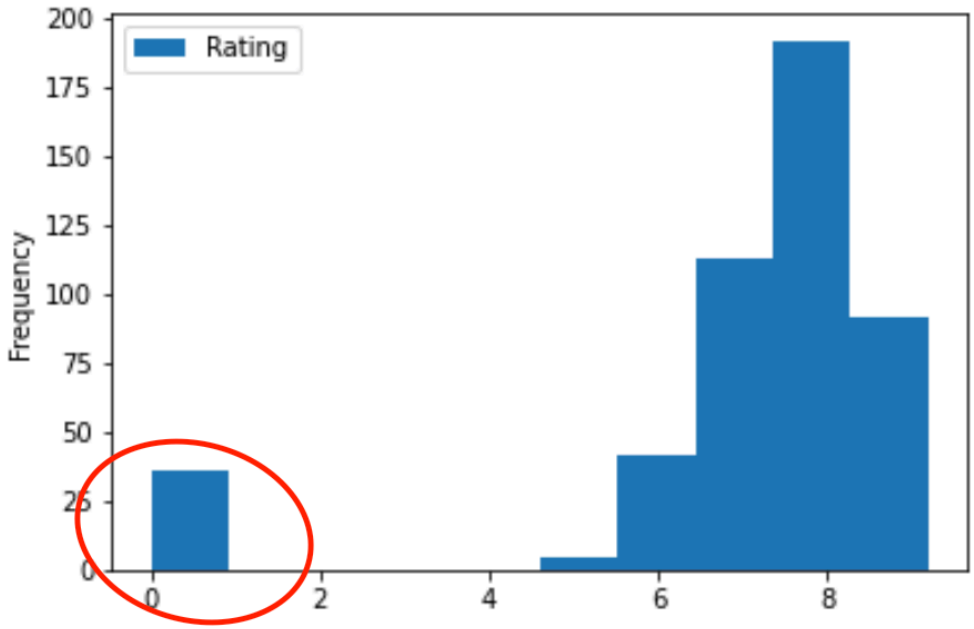
Restaurant types

As a result, the total 2106 venues are filtered for those venues, which category contains the string ‘Restaurant’ resulting in 479 restaurants of all all types (Asian, American, French etc.)

Venue Id and Rating

To be able to get the rating for each restaurant, its venue Id is needed, since each venue needs to be queried separately for its rating.

The below histogram shows the result of querying the rating for the 479 restaurants. There are 36 Restaurants which have not been rated yet. These will be excluded from the data set.



Result of rating query

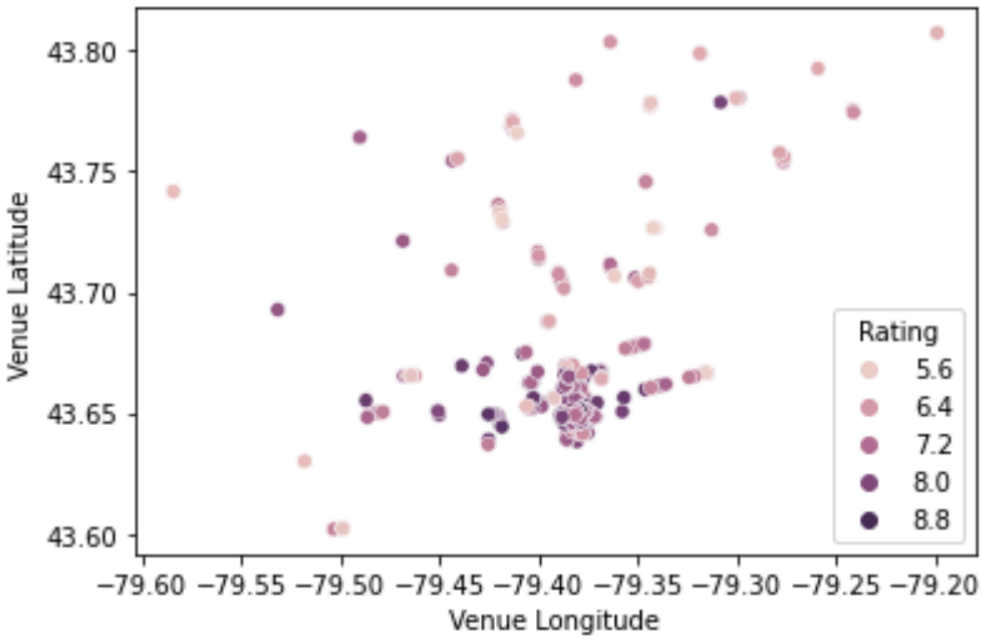
Final Dataset

The structure of the final data set is depicted in the picture below. It contains the following information for each of the 479 restaurants: its Foursquare ID, positional Information (longitude, latitude) and rating.

[102]:

	Venue id	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Rating
0	4bb6b9446edc76b0d771311c	Malvern, Rouge	43.806686	-79.194353	Wendy's	43.807448	-79.199056	Fast Food Restaurant	6.0
4	4b1711a6f964a520cbc123e3	Cedarbrae	43.773136	-79.239476	Federick Restaurant	43.774697	-79.241142	Hakka Restaurant	7.1
5	4e261f261f6eb1ae13930699	Cedarbrae	43.773136	-79.239476	Drupati's Roti & Doubles	43.775222	-79.241678	Caribbean Restaurant	7.0
6	4c27da423492a593158cb628	Cedarbrae	43.773136	-79.239476	Thai One On	43.774468	-79.241268	Thai Restaurant	6.6
9	4b6475aef964a520eab42ae3	Dorset Park, Wexford Heights, Scarborough Town Centre	43.757410	-79.273304	Kim Kim restaurant	43.753833	-79.276611	Chinese Restaurant	7.1

With this information a scatter plot could be generated which gives an approximate idea of the distribution of restaurants.



Rated restaurants in Toronto