14.10.2020

IBM Data Science Capstone Project- The best place to open a new pub

PETR KANTEK

Contents

1. Introduction

Everyone loves beer, there is no question about that. But due to the current pandemic situation, many pubs are endangered by the governmental restrictions. Limitations of pubs include:

- **lowered capacity due to safety distance measures,**
- **reduced opening time,**
- **or in extreme cases closure.**

We can expect that many pubs will go bankrupt and will never reopen. This, on the other hand, provides an opportunity for brave entrepreneurs, who could open a new pub after the pandemic will have been vanished from the world, and thus fill the empty space in the pub and beer industry. We are here to help the potential pub-owners to choose the best location for their new pub! Since we already have some background knowledge of the Toronto area, we will focus on this location.

2. Data

We will use 3 data sources:

- **Toronto neighbourhoods dataset**
- **geocoordinates**
- **Foursquare Database**

We have already used the Toronto dataset in an earlier excercise. We will scrap it from a Wikipedia page using Pandas functionality. It comprises of 3 columns: Postal Code, Borough, and Neighbourhood. The geocoordinates dataset will be loaded from a geojson file and contains longitude and latitude values for each postal code. Lastly, Foursquare database is a huge database containing lots of data about various places on the planet. We will need only information about venues. Each venue contains information such as: Name, Location, and Category. We will access the database via REST API, particulary by GET requests on the search endpoint.

3. Analysis

We can print the dataframe of Toronto neighbourhoods. As I wrote in the Data section, the dataframe conists of 3 columns.

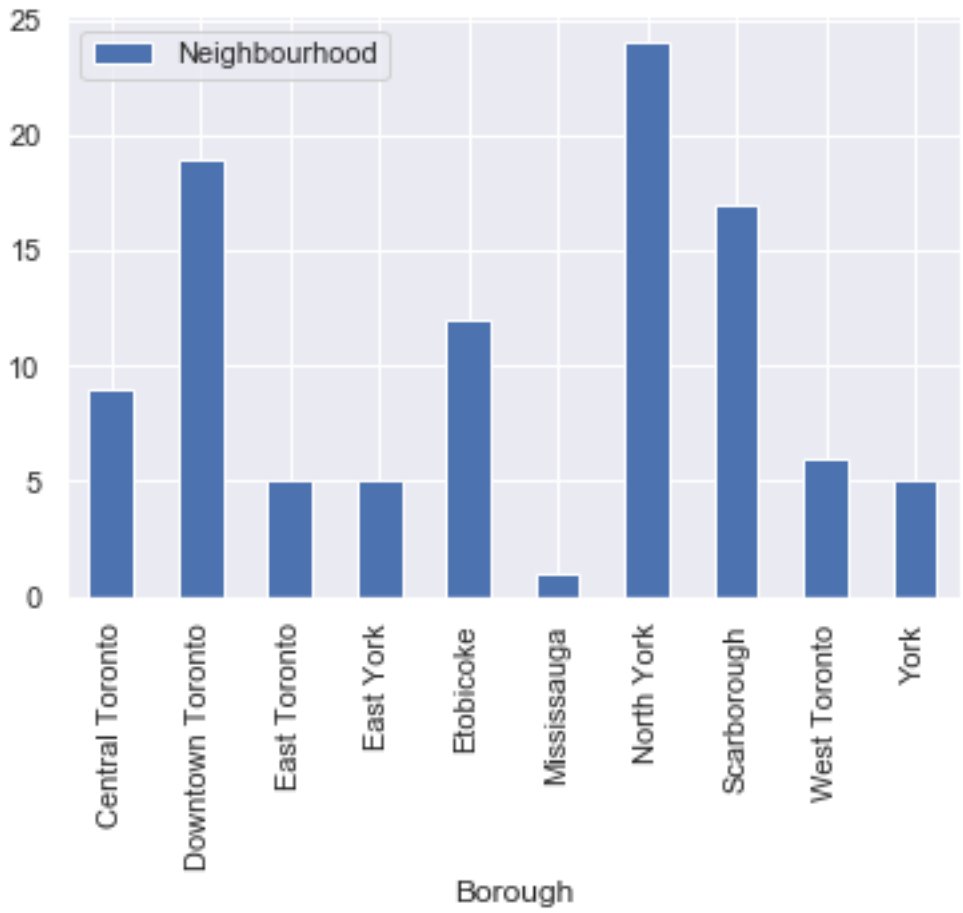
	Postal Code	Borough	Neighbourhood
0	M1A	Not assigned	Not assigned
1	M2A	Not assigned	Not assigned
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park, Harbourfront
...	...	...	...
175	M5Z	Not assigned	Not assigned
176	M6Z	Not assigned	Not assigned
177	M7Z	Not assigned	Not assigned
178	M8Z	Etobicoke	Mimico NW, The Queensway West, South of Bloor,...
179	M9Z	Not assigned	Not assigned

180 rows × 3 columns

As expected, the datatypes of the columns are objects.

```
Postal Code     object
Borough         object
Neighbourhood   object
dtype: object
```

In the preprocessing process, we will remove rows that contain ‘Not assigned’ values. Further, we can see the number of neighbourhoods of each borough.



Then, we can move to the geocoordinates. Unsurprisingly, they contain longitudes and latitudes.

	Postal Code	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917
4	M1H	43.773136	-79.239476
...	...	...	...
98	M9N	43.706876	-79.518188
99	M9P	43.696319	-79.532242
100	M9R	43.688905	-79.554724
101	M9V	43.739416	-79.588437
102	M9W	43.706748	-79.594054

103 rows × 3 columns

We will merge the geocoordinates with neighbourhood data. Now we can proceed to the venues data. Venues have over 272 categories, that we will one-hot encode, so we can create the clustering metric four our machine learning model.

	Neighbourhoods	Accessories Store	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge
0	Lawrence Park	0	0	0	0	0	0
1	Lawrence Park	0	0	0	0	0	0
2	Lawrence Park	0	0	0	0	0	0
3	Davisville North	0	0	0	0	0	0
4	Davisville North	0	0	0	0	0	0

After encoding, we can finally group the venues by neighbourhood, and compute the mean of pubs, which will be the clustering metric.

	Neighbourhoods	Pub
0	Agincourt	0.000000
1	Alderwood, Long Branch	0.142857
2	Bathurst Manor, Wilson Heights, Downsview North	0.000000
3	Bayview Village	0.000000
4	Bedford Park, Lawrence Manor East	0.045455

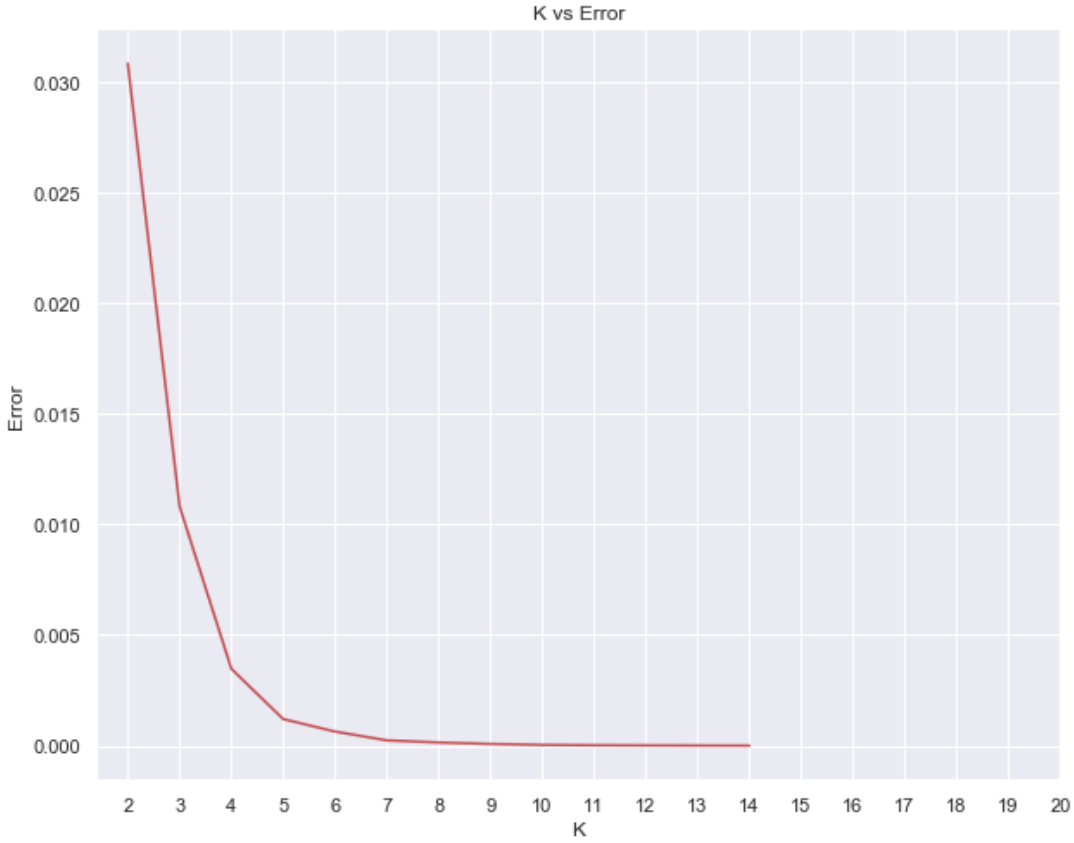
Finally, we have the data ready and we can proceed the to the training of K-means.

4. Modeling

Since this is a clustering and unsupervised task, I will use a K-means clustering algorithm. I have considered also other clustering algorithms, such as: Hierarchical clustering, DBSCAN, etc. but I decided to use K-means, as it is reliable in performing good results, its easy to understand, and code using the sk learn library.

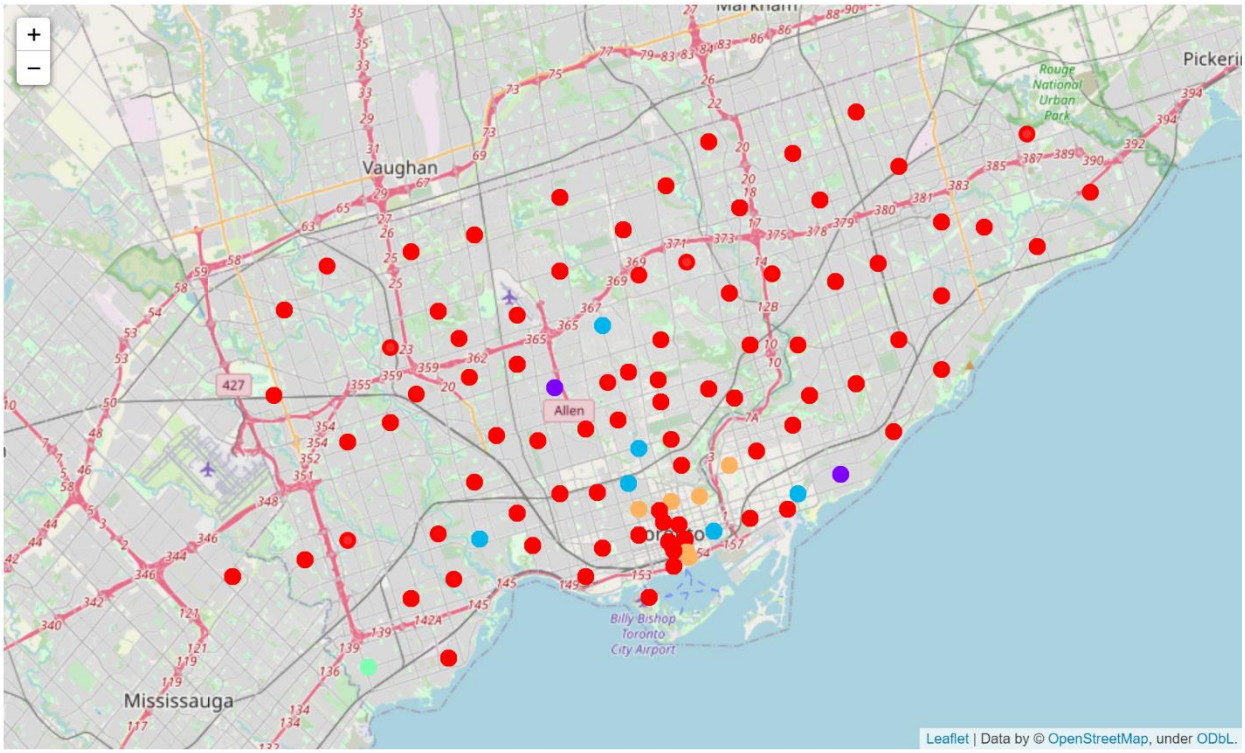
The decision to use K-means encompasses a question, how many clusters will there be? Because K-means requires the number of clusters in advance of the training process. There could be several methods to choose the most appropriate number, or at least some meaningful interval from which to choose it, that can be further refined by the specifications of the essence of machine learning task.

As with the decision of which clustering algorithm to use, there are also many options for evaluation of the best number of clusters. They include: Grid search, Probabilistic prediction, Elbow method, etc. I decided to use the Elbow method, as it is easy to implement using the inbuilt scikit-learn K-means algorithm feature called *inertia and it is also easy to evaluate it*.

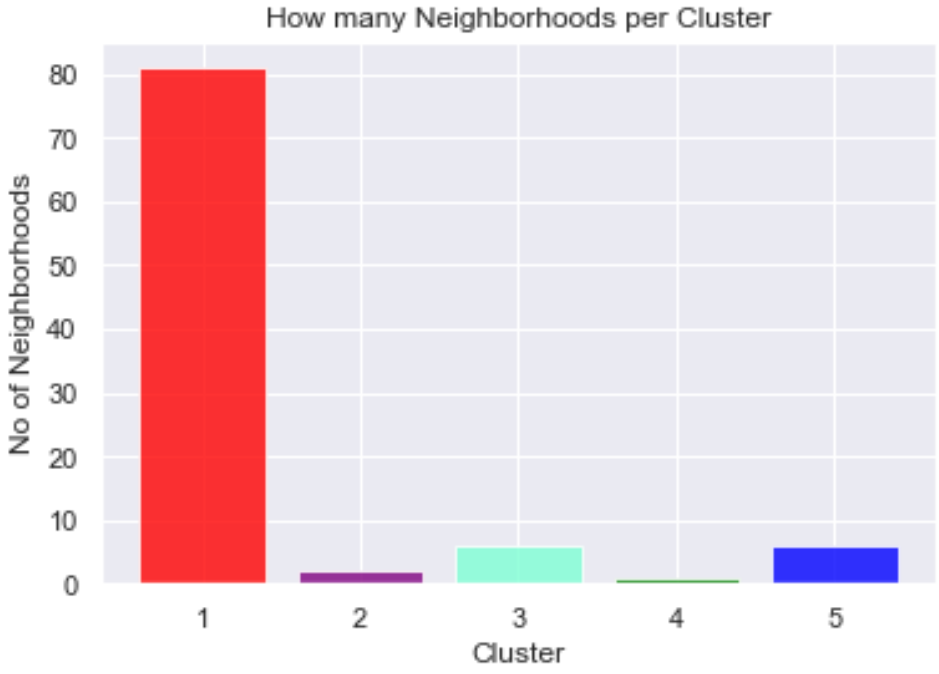


5. Results

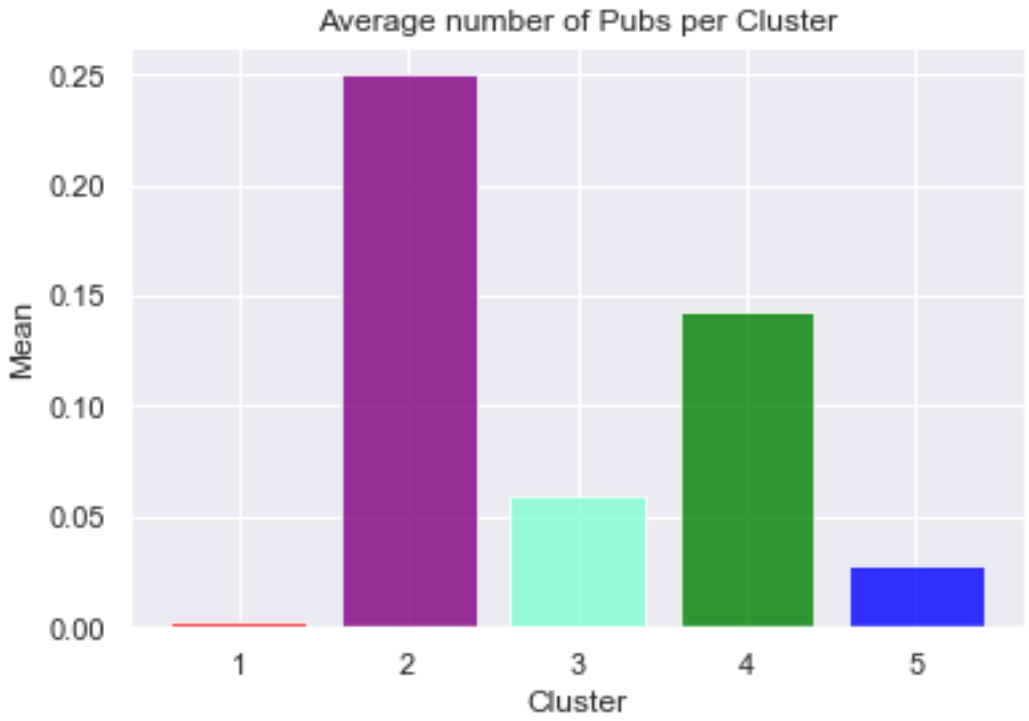
Let us look at the resulting clusters that the K-means algorithm created. On the following map we can see the cluster. It is apparent that neighbourhoods in the red cluster are the most frequent. On the other hand, neighbourhoods in the green clusters are the least frequent.



Let us visualize the number of neighbourhoods in clusters using barplot.



Furthermore, we can also visualize the average number of pubs per cluster.



6. Conclusion

We have gone through the whole machine learning (or data science) pipeline; from data loading and wrangling, to exploration, model training, optimization, and assessing the results. In the end there were 5 clusters, that k-means algorithm created. As per the initial task, to find the optimal place to set up a new pub, we can sketch a few answers. I would definitely not recommend to set up a pub in the any of the neighbourhoods in the 2. or 4. cluster, as the average number of pubs in them is too large and the number of neighbourhoods too small, which means, that the competition could be too large and setting a pub there would be inefficient. As for the other three clusters, those look more promising, with clusters 3 and 5 being close to each other both in term of the average number of pubs and the number of neighbourhoods. The most promising one seems to be the 1. cluster, as its average number of pubs is tiny and the number of neighbourshoods big. So, to sum up, for further consideration of other facts and conditions I would recommend looking at pub places in the neighbourhoods of the 1. or 5. clusters. :)

One of the drawbacks of this analysis could be the concern, that the data from the Foursquare database do not reliably represent the current situation in Toronto.