CAPSTONE PROJECT

IBM Data Science Professional Certificate

Dmitriy Zubenko
2020

New Bakery Business Inspection in Toronto, Ontario, Canada

Description

Problem

Toronto is the capital city of the Canadian province of Ontario. With a recorded population of approximately 2.7 million in 2016, it is the most populous city in Canada and the fourth most populous city in North America. The diverse population of Toronto reflects its current and historical role as an important destination for immigrants to Canada. More than 50 percent of residents belong to a visible minority population group, and over 200 distinct ethnic origins are represented among its inhabitants. Toronto is an international centre of business, finance, arts, and culture, and is recognized as one of the most multicultural and cosmopolitan cities in the world. Toronto covers an area of 630 square kilometres (243 sq mi), with a maximum north–south distance of 21 km (13 mi). It has a maximum east–west distance of 43 km (27 mi) and it has a 46-kilometre (29 mi) long waterfront shoreline, on the northwestern shore of Lake Ontario.

Data

To proceed with research we will use such **data**:

- postal codes, boroughs, neighborhoods info on Toronto, Canada
- postal codes latitude and longitude coordinates
- venue data of bakeries and pastries

Data sources to help:

- postal codes, boroughs, neighborhoods info on Toronto, Canada by Wikipedia
- postal codes latitude and longitude coordinates by Google Maps API
- venue data of bakeries and pastries by Foursquare API

The **objective**

of this problem is to analyze and select the best locations in the city of Toronto, Canada to open new bakery.

Utilizing data science methodology and instruments such data analysis and data visualization project aims to provide new insights for declared business problem.

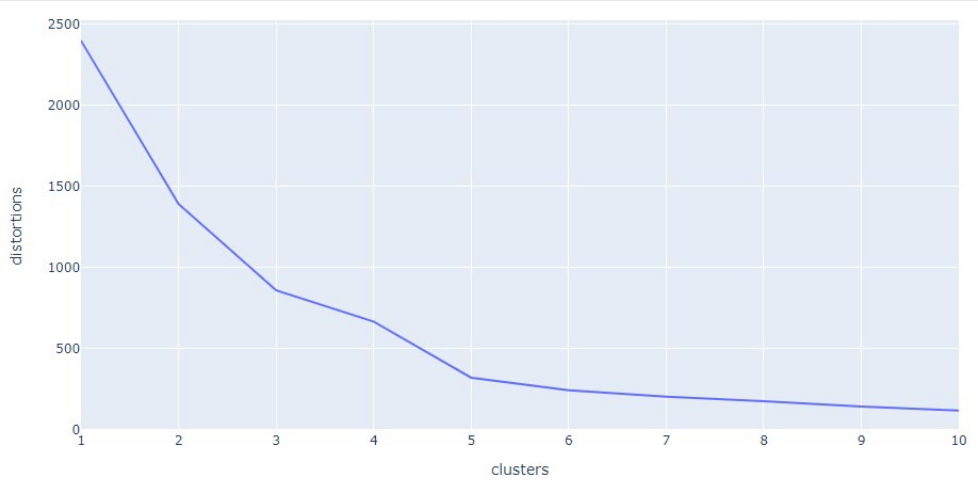
Methodology

- Filter out competitors in all the venues categories
- Calculate competitors' number and their percentage from all vanues
- Wrangle data to prepare for K-Mean cluster analysis
- Clustering
- Clusters exploration and analysis
- Conclusions

Results

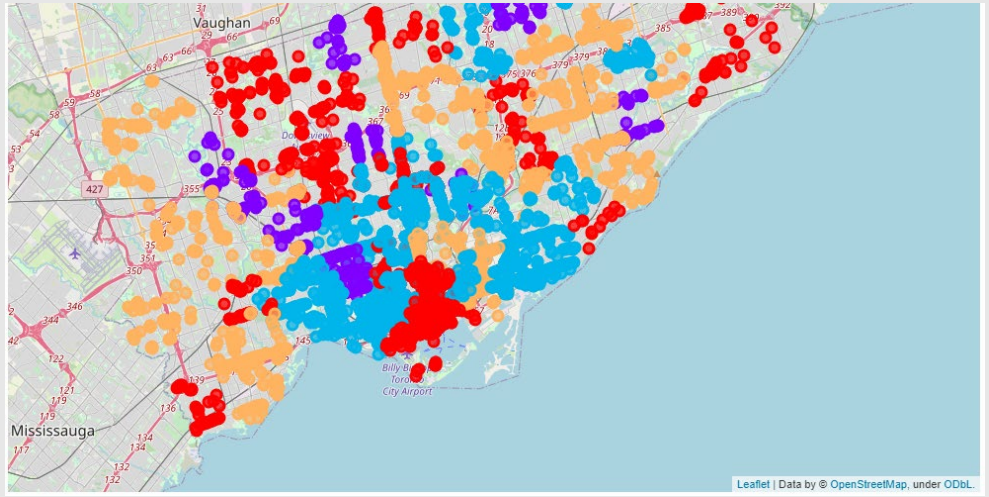
‘Elbow’ method

Helped to calculate number five clusters



Mapping Clusters

Helped to segment venues and see their locations over Toronto



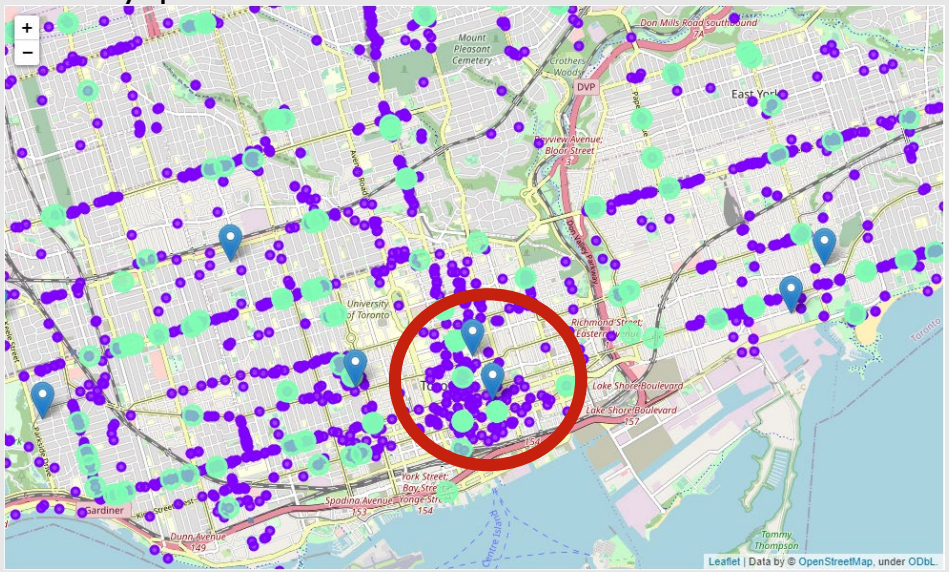
Discussion

	Cluster	Num of Competitors	Venues Total	Percent of Competitors
0	0	46	2201	1.803333
1	4	58	474	12.645556
2	2	169	2658	6.551613
3	1	1	3	33.330000
4	3	78	1558	5.363793

As we examine total of all clusters, we can state that Cluster 2 has highest number of venues and competitors.

We applied filters to Cluster 2 to exclude clusters with low number of venues and high number of competitors.

The last step was to add three layers to the maps: venues, competitors and preferred postal codes, so we can see easily postal codes we need.



Conclusion

As we can see from the map tho postal codes *M5B* and *M5C* situated in area with high amount of venues (so there is nice traffic) and have not so much competitors around.

Let's discover what boroughs these are!

```
# import JSON file with Toronto boroughs from previous task
toronto_boroughs = pd.read_json(r'toronto_boroughs.json')

# looking for the most perspective boroughs by posta code
toronto_boroughs.loc[(toronto_boroughs['Postal Code'] == 'M5B') | (toronto_boroughs['Postal Code'] == 'M5C')]
```

	Postal Code	Borough	Neighborhood	Latitude	Longitude
9	M5B	Downtown Toronto	Garden District, Ryerson	43.657162	-79.378937
15	M5C	Downtown Toronto	St. James Town	43.651494	-79.375418

Thank you!