Capstone project report

INTRODUCTION

New York City is the US's top city for small business for the second year in a row, according to Biz2Credit's annual study of the Top Small Business Cities in America, which analyzed the financial performance of 27,000 small businesses and their local market economic conditions.

A Canadian restaurant owner who has multiple restaurants in Toronto is planning to expand to the US. Seeing the potential of NYC, he decides to open a new restaurant in this city. But there is a question: Where should the new restaurant be located?

The main purpose of this report is showing a Data Science approach to solve that question.

DATA

Based on the purpose of the project, New York City neighborhoods and client's restaurants in Toronto were chosen as the observation target. We collect the data using FourSquare API which provides the surrounding venues of a given coordinate.

The process of collecting and clean data:

- Find the geographic data of the NYC neighborhoods and client's restaurants in Toronto.
- For each neighborhood or restaurant, pass the obtained coordinates to FourSquare API. The "explore" endpoint will return a list of surrounding venues in a pre-defined radius.
- Count the occurrence of each venue type, then apply one hot encoding to turn each venue type into a column with their occurrence as the value.
- Then merge NYC neighborhoods dataset with client's restaurants dataset. The resulting dataset is a 2 dimensions dataframe:

```
(301, 432)
```

	Neighborhood	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport Terminal	Airport Tram	American Restaurant	Animal Shelter	Antique Shop	...	Watch Shop	Waterfront	Weight Loss Center	Whisky Bar	Wine Ba
0	Allerton	0.0	0.0	0.0	0.0	0.0	0.0	0.03125	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
1	Annadale	0.0	0.0	0.0	0.0	0.0	0.0	0.00000	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
2	Arden Heights	0.0	0.0	0.0	0.0	0.0	0.0	0.00000	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
3	Arlington	0.0	0.0	0.0	0.0	0.0	0.0	0.12500	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
4	Arrochar	0.0	0.0	0.0	0.0	0.0	0.0	0.00000	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0

5 rows × 432 columns

METHODOLOGY

The assumption is that a neighborhood in NYC which is similar to the one that contains the client's restaurant in Toronto is a good place to open a new restaurant. Thus, the clustering technique will be

used to analyze the dataset. In the end, we will find good neighborhoods in NYC which suits the client's purpose.

Python data science tools will be used to help analyze the data.

In order to have the first insight into our data, we calculate the top 10 venues in each NYC neighborhood and each client's restaurant.

Client's restaurant:

```
(7, 12)
```

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	Competitors
0	Moore Park,Summerhill East	Playground	Park	Tennis Court	Gym	Doner Restaurant	Donut Shop	Dumpling Restaurant	Eastern European Restaurant	Electronics Store	Ethiopian Restaurant	0.0
1	Forest Hill North,Forest Hill West	Trail	Sushi Restaurant	Bus Line	Jewelry Store	Yoga Studio	Donut Shop	Fish & Chips Shop	Filipino Restaurant	Fast Food Restaurant	Farmers Market	1.0
2	Harbourfront,Regent Park	Coffee Shop	Bakery	Pub	Park	Café	Breakfast Spot	Restaurant	Mexican Restaurant	Theater	Bank	7.0
3	Rosedale	Park	Playground	Trail	Dog Run	Fish & Chips Shop	Filipino Restaurant	Fast Food Restaurant	Farmers Market	Falafel Restaurant	Event Space	0.0
4	CN Tower,Bathurst Quay,Island airport,Harbourf...	Airport Lounge	Airport Service	Airport Terminal	Boat or Ferry	Sculpture Garden	Plane	Airport	Airport Food Court	Airport Gate	Harbor / Marina	0.0
5	Chinatown,Grange Park,Kensington Market	Bar	Café	Vegetarian / Vegan Restaurant	Bakery	Vietnamese Restaurant	Coffee Shop	Dumpling Restaurant	Chinese Restaurant	Mexican Restaurant	Dim Sum Restaurant	33.0
6	Central Bay Street	Coffee Shop	Café	Italian Restaurant	Bar	Burger Joint	Thai Restaurant	Sandwich Place	Salad Place	Indian Restaurant	Ice Cream Shop	25.0

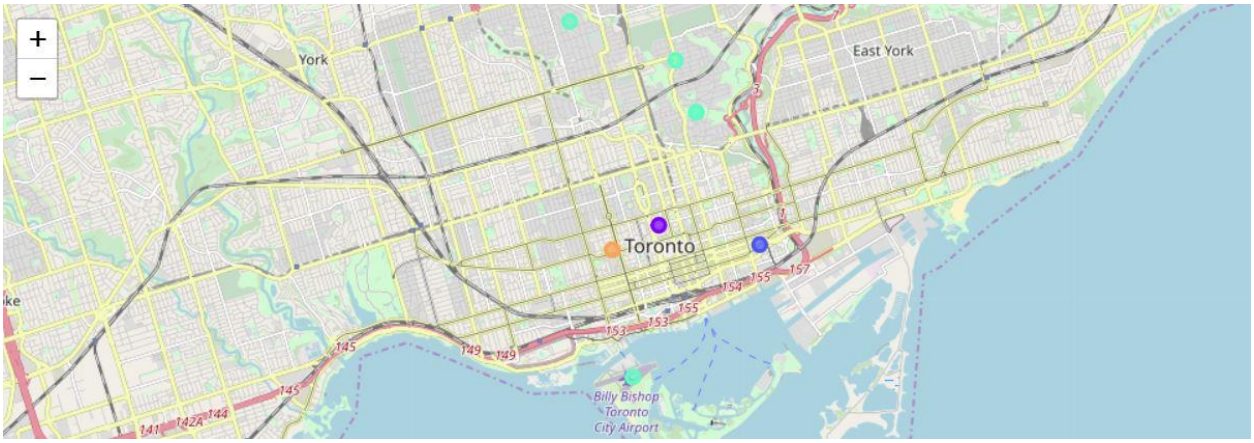
NYC neighborhoods:

```
(301, 12)
```

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	Competitors
0	Allerton	Pizza Place	Spa	Supermarket	Chinese Restaurant	Deli / Bodega	Food	Fast Food Restaurant	Bakery	Electronics Store	Pharmacy	5.0
1	Annadale	Pub	Cosmetics Shop	Diner	Train Station	Liquor Store	Sports Bar	Pizza Place	Restaurant	Pet Store	Exhibit	1.0
2	Arden Heights	Pharmacy	Deli / Bodega	Pizza Place	Coffee Shop	Home Service	Filipino Restaurant	Event Space	Exhibit	Eye Doctor	Factory	0.0
3	Arlington	Bus Stop	Deli / Bodega	Intersection	American Restaurant	Food Service	Boat or Ferry	Coffee Shop	Yoga Studio	Fish & Chips Shop	Filipino Restaurant	1.0
4	Arrochar	Deli / Bodega	Bus Stop	Italian Restaurant	Liquor Store	Middle Eastern Restaurant	Taco Place	Sandwich Place	Food Truck	Pizza Place	Cosmetics Shop	4.0

Then we use k-means clustering technique to find similar groups of neighborhoods. And plot the groups that contain the client's restaurant.

Toronto:



NYC:

RESULTS

Based on 2 plots about the similar groups of client's restaurant and NYC neighborhoods, we have a conclusion that if neighborhoods A in NYC is in the same cluster of client's restaurant B in Toronto, then the client should open a new restaurant in A with the same business strategy that he applied in B.

CONCLUSION

It's unfortunate that the analysis couldn't produce a precise model or showing any strong coefficient correlation for any venue type. But we can still get some meaningful and logical insights from the result.

Doing this project helps to practice every topic in the specialization, and thus, equipping learners with Data Science methodology and tools using Python libraries. Also doing a real project certainly helps one learns so much more outside the curriculum, as well as realizes what more to research into after completing the program. And as this report shows, there are surely a lot of things to dig into.

Toward the person that went through this project, many thanks for the time and patient.