A. Introduction:

a. Background

Food business is one of the most common business by its number of type and its store quantity. There are stores starting and stores closing every day. One of the most important parameters of their successes is location. However, the same location means different to different types of food business. Near a shopping center may be good for a fast food restaurant such as McDonald but may be not suitable for a high-end French restaurant. Finding a good location is more of finding a good location for a class or a type of food business rather than finding a location for all food businesses. Businesses of the same class do compete to each other. But it is riskier to open a business just by your own. That is why in many cases, same types of restaurants cluster together.

b. Problem

The problem we want to solve in my project is: if we are going to open a food business of one kind, where should we open inside a city? Our solution to the problem is finding clusters of locations for different types of food business. New York, Toronto and London are cities we are going to perform analyses. For simplicity, we are only going to distinguish food businesses into two classes: fast food business and non-fast food business. To better answer the above question, we are also going to find why this cluster of locations is good for one class and is there any difference among three cities? Both maps and tables will be used to analyze information collected and interpret findings and results.

c. Audience and interest

Our analysis will be valuable for food business owners, food business investors, and may be interested by investors who follow publicly listed food companies those have franchises in those three cities.

B. Data description

a. Data Sources

For each city, we will divide the data gathering and processing procedure into three steps by different data sources. Step 1: Get Borough, Neighborhood and Postal code of each city. For NYC, we get those data from[2] and for Toronto and London, we get them from Wikipedia; Step 2: Get geographical coordinates of neighborhoods. For all three cities, we get the data from pgeocode package from PyPL[1]; Step 3: Get nearby venue information for each geographical coordinate. We use Foursquare to get those data.

b. Data cleaning

As mentioned in Data Sources section, there are generally three steps to get our data. The Step 2 and 3 are quite similar for all three cities, however, the step 1 is a little different among cities. Here, we will explain this step in-detail for each city.

Step 1: Get Borough, Neighborhood and Postal code information for each city.

New York City

We find borough, neighborhood and postal code information from website[2] by using the pandas read-html function. Those data are quite clean. The only thing we need to adjust is the postal codes column since there are more than one postal code in the same neighborhood, but we only need one. Since those postal codes are separated by comma, we can split them and only take the first one. We then create a new column for postal code with only one postal code and we drop the old column. Now we have our table ready for the next step. We aim to create similar table for all three cities before going to the next step.

	Borough	Neighborhood	Postcode
0	Bronx	Central Bronx	10453
1	Bronx	Bronx Park and Fordham	10458
2	Bronx	High Bridge and Morrisania	10451
3	Bronx	Hunts Point and Mott Haven	10454
4	Bronx	Kingsbridge and Riverdale	10463

Toronto

We find borough, neighborhood and postal code information from wikipedia[3] by using the pandas read-html function. The initial table we get is in very good shape, however, there are several adjustments we need to make before entering to the next step.

Firstly, there are 'not assigned' cells in column 'Borough'. We need to drop those rows since they can not provide us enough information. Secondly, there are some 'not assigned' cells in column 'Neighborhood' with a valuable cell under the column 'Borough'. We replace those 'not assigned' with the value of 'Borough'. Thirdly, there are different values of Neighborhood with the same Postcode. Those Neighborhood value should be merged together since we need unique postal code to get local information to avoid duplicate results.

	Postcode	Borough	Neighborhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Harbourfront
3	M6A	North York	Lawrence Heights, Lawrence Manor
4	M7A	Downtown Toronto	Queen's Park

London

We find borough, neighborhood and postal code information from wikipedia[4] by using the pandas read-html function. The initial table looks messy compare to initial tables of other two cities. London is a very big city. In order to simply our analysis, we only focus on the city of London rather than the entire London area. Then, we remove several columns since they provide irrelevant information. There are some '[#]' characters behind some values of column 'Borough'. Since this could mistakenly differentiate the same borough with those special characters, we need to remove them. Finally, some rows have more than one postal code. We will only use the first code if there are many.

	Postcode	Borough	Neighborhood
0	SE2	Bexley, Greenwich	Abbey Wood
1	W3	Ealing, Hammersmith and Fulham	Acton
2	EC3	City	Aldgate
3	WC2	Westminster	Aldwych
4	SE20	Bromley	Anerley

Step 2: Get geographical coordinates of neighborhoods. We will use the postal code from Step 1 to get coordinates. Each postal code has only one set of coordinates. Below is the result for City of London.

	Postcode	Borough	Neighborhood	Latitude	Longitude
0	SE2	Bexley, Greenwich	Abbey Wood	51.4869	0.107500
1	W3	Ealing, Hammersmith and Fulham	Acton	51.5114	-0.265717
2	EC3	City	Aldgate	51.5085	-0.125700
3	WC2	Westminster	Aldwych	51.5142	-0.123382
4	SE20	Bromley	Anerley	51.4154	-0.056950

Sometimes there may be no coordinates for some boroughs. This could result 'NaN' in the value of coordinates, which could make some errors in Step 3. We will remove those rows if there is any.

Step 3: Get nearby venue information for each geographical coordinate. We set a limit of maximum 500 venues for each set of coordinates and a radius of 500 meters. Once the data of venue is ready, we group them by the value of neighborhood and get weight by each venue category. Below is the result for City of London. (The sum of weight of neighborhood by each venue category is 1)

	NEIGHBORHOOD	African Restaurant	Airport	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports
0	Abbey Wood	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.00	0.00	0.00	0.0
1	Acton	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.00	0.00	0.00	0.0
2	Aldgate	0.01	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.01	0.02	0.02	0.02	0.0
3	Aldwych	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.00	0.01	0.00	0.0
4	Anerley	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.00	0.00	0.00	0.0

Then we get the 10 most common venue category for each neighborhood.

	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
0	Abbey Wood	Chinese Restaurant	Grocery Store	Indian Restaurant	Men's Store	Fish Market	Falafel Restaurant	Farmers Market	Fast Food Restaurant	Film Studio	Fish & Chips Shop
1	Acton	Pub	Park	Convenience Store	Mini Golf	Gas Station	Grocery Store	Bed & Breakfast	Train Station	Bakery	Japanese Restaurant
2	Aldgate	Theater	Hotel	French Restaurant	Pub	Plaza	Pizza Place	Wine Bar	Steakhouse	Café	Ice Cream Shop
3	Aldwych	Theater	Clothing Store	Coffee Shop	Bakery	Ice Cream Shop	Indian Restaurant	Dessert Shop	Cosmetics Shop	Museum	Wine Bar
4	Anerley	Fast Food Restaurant	Pub	Train Station	Coffee Shop	Pizza Place	Supermarket	Café	Hardware Store	Tunnel	Garden Center

Now we have our data ready to put into the model for analyzing.

C. Methodology

D. Results

E. Discussion

F. Conclusion

G. References

1. Pgeocode package: https://pypi.org/project/pgeocode/
2. New York City Borough, Neighborhood and Postal Code:

https://www.health.ny.gov/statistics/cancer/registry/appendix/neighborhoods.htm

3. Toronto Borough, Neighborhood and Postal Code:

https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

4. London Borough, Neighborhood and Postal Code:

https://en.wikipedia.org/wiki/List_of_areas_of_London