Description of the data and how it will be used to solve the problem

The data that are relevant for answering the business question is the location data of popular restaurants as well as the category of the restaurants. Before we use that data, however, we must determine what neighbourhoods or areas in Toronto that is suitable for opening up restaurants. It can be assumed since Toronto is the capital city of Canada; there are many restaurants in the whole city. However, the distribution of restaurants will not be uniform; there will be areas with more restaurants and areas that have fewer restaurants, even none. It is a simple assumption that the number of restaurants is proportional to the population density. The area with the highest population density then is the best location for opening up a new restaurant.

The city of Toronto is divided into several levels of areas: community council areas, wards, and neighbourhoods. The community council areas are the first level of area division. It consists of several wards, which in turn consists of several neighbourhoods. There are four community council areas, North York, Etobicoke York, Toronto and East York, and Scarborough with different population densities. To find the population density of each community council areas, the 2016 Toronto Census data will be used. The census dataset consists of several categories such as total population per community council areas and the number of population in a certain age range. There is also a separate but related dataset that has the area size of the community council areas. The population density is then calculated by dividing the total population of a community council area by the area size. For example, Etobicoke York has 642415 people and an area 187.5 km2 or 18750 hectares. The population density then will be 34 persons per hectare (rounded to nearest integer). Applying this method to all the community council areas, it is found that Toronto and East York has the highest population density: 80 persons per hectare. Since more population density means more potential customers this community council area is selected to be the area for the new restaurant.

Toronto and East York are divided into several postal code areas and boroughs. Wikipedia provided the postal code data in Toronto, as well the corresponding boroughs and neighbourhoods in a table form. For example, in the column Postal Code, there is a code M1B. The Borough and Neighborhood columns have Scarborough and Rouge respectively. A postal code can correspond to more than one neighbourhood in a borough. The code M1B also correspond to the Malvern neighbourhood. This data will be used to divide the community council areas further. Before use, the data needs to be preprocessed so that neighbourhoods that have the same postal code will be joined together in one row. This postal code data will be combined with the latitude and longitude of each postal code area as provided by <http://cocl.us/Geospatial_data>. The geospatial coordinates provided by the link consist of postal code column as well as latitude and longitude column. For example, the postal code M1B has the values 43.80669 and -79.1943534 in the corresponding Latitude and Longitude columns respectively. The combined data will provide a centre from which to search for popular restaurants. This data is finally going to be fed into the Foursquare search API to search for popular restaurants around the coordinates.

The result from the Foursquare search API is a JSON file consisting of restaurant name such as “Subway”, restaurant category such as “sandwich place”, restaurant address, restaurant website address if it has one as well as the restaurant latitude and longitude and some more information. The features that will be extracted from the JSON file will be the restaurant’s name, the restaurant’s category, and the restaurant’s latitude and longitude. These features will be combined with the dataset containing neighbourhood names, and the neighbourhoods’ latitude and longitude. The resulting data frame will consist of neighbourhood name, latitude and longitude, restaurant name, latitude and longitude, and the restaurant category. Then the number of restaurants in each postal code area or combined neighbourhoods will be counted so that we will know which neighbourhoods or postal code areas that has many restaurants. On the assumption that if an area has many restaurants then that location is a good location to open up new restaurants, the location with the highest or high number of operating restaurants will be chosen as a good location for the client’s new restaurant site. This answers the first part of the business question.

The categories of restaurants in the data frame will also be counted and sorted according to the order of frequency and several of the high-frequency categories will then be chosen as the popular types of restaurants. This assumes that the number of certain restaurant category in an area is proportional to its popularity. For example, if a certain area has 3 restaurants in the “Chinese Restaurant” category and 7 restaurants in the “Italian Restaurant” category, then this data will be interpreted as Italian restaurants are more popular than Chinese restaurants in that area. This answers the second part of the business question.