*#This document combines the ‘Introduction/Business Problem’ and ‘Data/Analytic Approach’ section for Week 4 assignment of IBM Applied Data Science Capstone Report. See attached Annexes for Data*

**INTRODUCTION/BUSINESS PROBLEM**

1. Business. Nanyang Coffee Company (NYC) is a major supplier of the processed coffee beans used in the *Nanyang* style of coffee popularly drank by the people of Southeast Asia. [*Note: Southeast Asian in this context refers to the ten nations that make up the Association of South East Asian Nations i.e. Brunei, Cambodia, Indonesia, Laos, Malaysia, Myanmar, Philippines, Singapore, Thailand and Vietnam.*] NYC is a dominant player in Southeast Asia, with a market share of the 20% of the roasted coffee beans market for *Nanyang* coffee in the region. Its key products are roasted coffee beans and coffee powders. Its business lines are (1) the wholesale of *Nanyang* style roasted coffee beans and processed products to coffee shops, restaurants, and coffee product retailers; and (2) its own coffee specialty retail outlets that sells its coffee products and take-away coffee beverages.

1. Product Introduction. Unlike the coffee preferred by consumers in America or Europe which uses primarily Arabica beans, the *Nanyang* coffee is made from a mixture of Arabica, Robusta and Liberica coffee beans. In particular, the significant proportion of the Robusta beans in *Nanyang* coffee confers a stronger and thicker flavor on the *Nanyang* coffee, as compared to the smoother and thinner flavor of the coffee drunk by European and American consumers e.g. expressos. The Robusta beans are commonly grown and sourced from Vietnam, Indonesia and Malaysia. Beside the composition of the beans, the roasting of the beans for *Nanyang* coffee also differs substantially, with a longer roasting period at lower heat necessary to mellow the stronger flavor of the Robusta beans, as compared to the shorter period and higher temperature involved in the roasting of the predominantly Arabica coffee beans. In addition, butter, margarine, sugar and/or palm oil are also added to the beans in the roasting process of the Robusta beans to impart a unique and complex flavor to the coffee. Finally, the *Nanyang* coffee is served either black, or with condensed and/or evaporated milk to add a creamy and thick texture to the coffee.
2. Business Problem. NYC is keen to expand beyond its existing market in Southeast Asia to North America. It has identified Canada as a promising country for its maiden entry into the North American market and earmarked the city of Toronto as its first city in Canada to expand it. The latest 2016 Census of Population prepared by Statistic Canada indicated that Toronto had a population of about 2.7 million and is the largest city in Canada. Within this population, Southeast Asians comprise about 7.0% or about 195,000[[1]](#footnote-1). This represents the largest concentration of Southeast Asian in any Canadian city, significantly more than the next largest such concentration of about 50,000 in Vancouver as similarly indicated in the 2016 Census of Population[[2]](#footnote-2). NYC views that the large Southeast Asian population in Toronto will constitute a ready market for its product.
3. NYC’s expansion plan calls for the setting up of a specialty coffee retail outlet offering a range of roasted coffee products and take-away drinks. Its coffee product will be positioned as a mid-range coffee product that appeals to the middle-income coffee lovers who currently patronize the cafes and coffee shops in Toronto. Within this segment, it seeks to establish its bridgehead customers among the Southeast Asians who are familiar with its products from their home countries, and through their customs and complemented by its marketing campaign, to build awareness amongst the wider coffee lover population in Toronto. Should the above succeed, NYC will consider setting up my outlets in other parts of Toronto, as well as expansion in other Canadian cities. NYC has engaged our consultancy company to advise on the expansion, with a key component being to decide on the location of the first NYC outlet.
4. Goals and Objectives, **Given the above, NYC has determined that the Goal of the expansion into Toronto is to maximize the awareness and penetration of the *Nanyang* coffee among the coffee drinking consumers, measures in terms of sales of this first outlet.** **The objective of the study is therefore to identify the neighborhood in Toronto to locate this first NYC outlet that best maximize the awareness and sales of the Nanyang coffee among the targeted segment of the coffee drinking consumers in Toronto, with the bridgehead customers being the ethnic Southeast Asians who enjoy drinking coffee.**

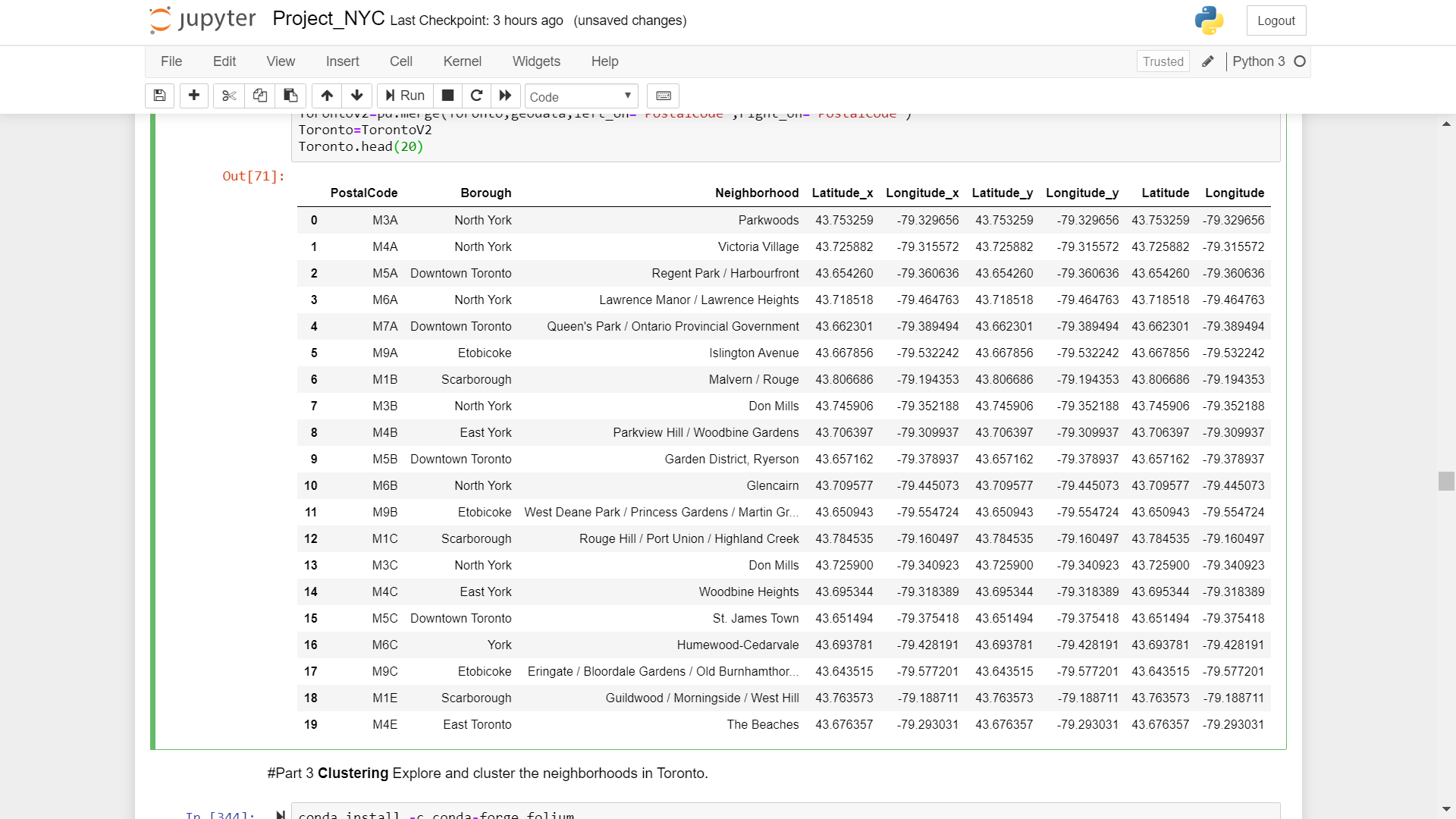
**ANALYTICS APPROACH/DATA**

1. **The goal of the study is to identify the neighborhood to locate the first NYC store in Toronto.** The hypothesis is that the location should be in the neighborhood with a relatively high concentration of (1) coffee shops and cafes, as well as (2) restaurants serving ethnic Southeast Asian cuisines. This assumes that a high concentration of coffee shops and cafes, along with a high parallel concentration of ethnic Southeast Asian restaurants would be indicative of a population of ethnic Southeast Asian customers who drinks coffee. This represents the most likely early customers for the *Nanyang* coffee products sold by NYC. This hypothesis is also based on surveys undertaken by NYC in its home market in Southeast Asia that Southeast Asians often patronize coffee outlets to drink coffee and socialized as a post dinner activity. Situating the NYC outlet in this neighborhood is therefore assessed as giving the highest likelihood of building awareness of the *Nanyang* coffee and therefore boosting sales.
2. This study has two components. For the first component of the study which seeks to identify the most optimal neighborhood to site the NYC outlet, the analytics approach will be a descriptive one that seeks to identify and rank neighborhood with high concentration of coffee shops and cafes, and ethnic Southeast Asian restaurants. A greater weightage will be accorded to the concentration of ethnic Southeast Asian restaurants as proxy for the Southeast Asians consumers in the selected neighborhood, given the bridgehead customers being ethnic Southeast Asian who like drinking coffee. For the second component of the study, this will seek to understand the strength of the relationship between the concentration of ethnic Southeast Asian and the concentration of coffee shops/cafes. This seeks to provide data on the consumption behavior of the ethnic Southeast Asians in Canada i.e. whether they enjoy going to coffee shops/cafes as a post-meal activity like the Southeast Asian consumers in NYC’s home region. This assumes that there is a proportional relationship between the two. The analytics approach will be to both use statistical correlation, as well as build a regression model to understand the strength of the relationship which could serve as model for other large Canadian cities.

**DATA**

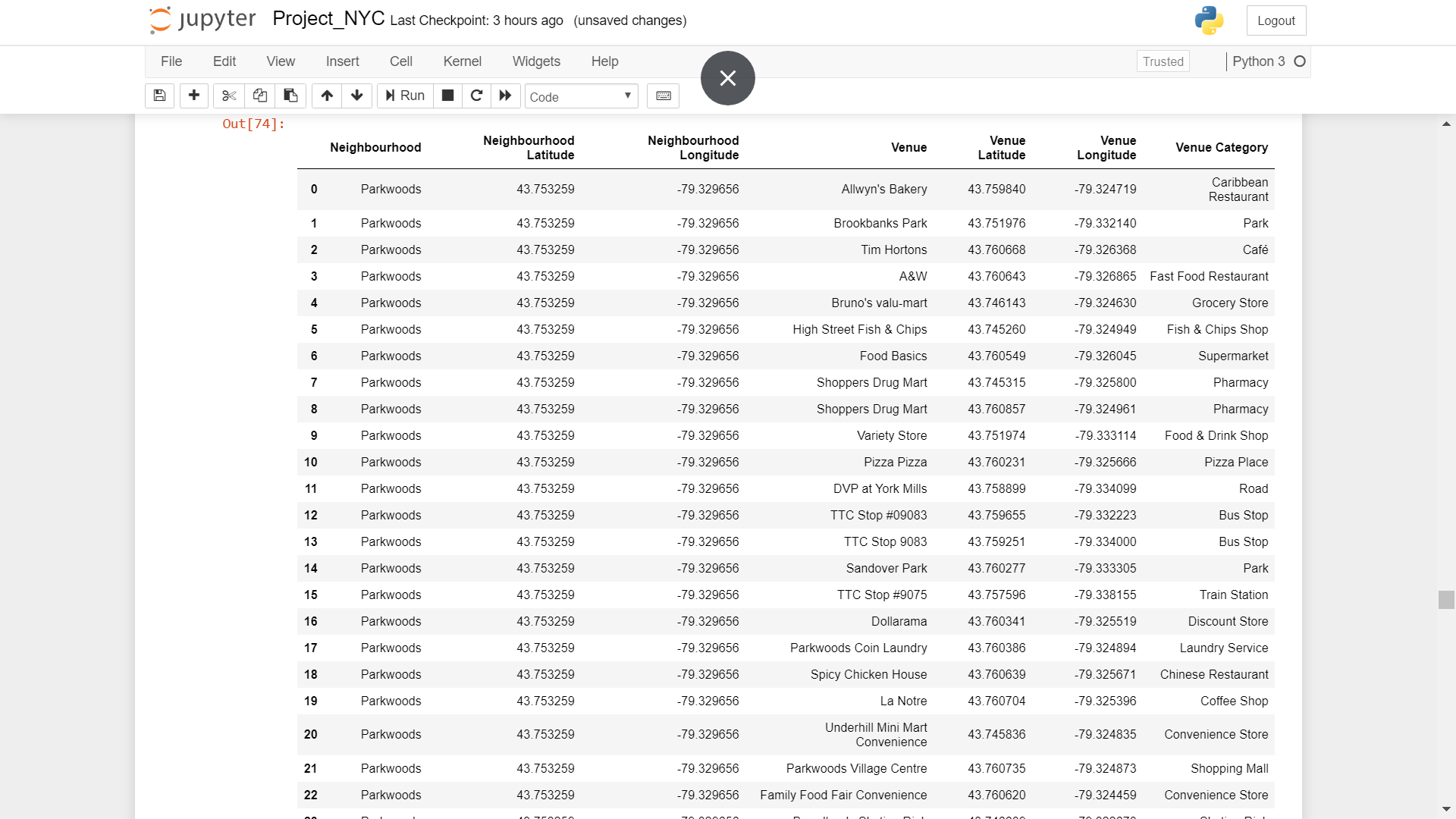
1. Geographical Data. This comprises (1) postal codes of boroughs in Toronto, (2) boroughs and associated neighborhoods in Toronto and (3) latitudes and longitudes of the neighborhood. For (1) and (2), these are scrapped using the BeautifulSoup software off the Wikipedia website of the Toronto Postal Codes <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>. For the geographical coordinates of each neighborhood in (3), these are obtained from IBM at <http://cocl.us/Geospatial_data>. The geographical data are then combined into a data frame, to be used as the input via the Foursquare API to extract the venue data of the coffee shops/cafes and Southeast Asian restaurants in each neighborhood for further analysis to address the goals of the study. A sample of the geographical data frame is shown in Figure 1 and the geographical data set (csv format) is attached as Annex A to this document (see github link).

Figure 1: Geographical Data Sample



1. Venue Data. The coffee shops/cafes, as well as the various ethnic Southeast Asian restaurants in each Toronto neighborhood will be extracted from Foursquare using the Foursquare API. This will use the geographical data obtained and formatted as outlined in Para 8 as the starting input. The sample of the master data showing all venues for each neighborhood is as shown in Figure 2 and the venue data set (csv format) is attached as Annex B to this document (see github link). The coffee shops/cafes and the various ethnic Southeast Asian restaurants will be extracted from this master data set subsequently.

Figure 2: Neighborhood Venues Data Sample



1. TBC

1. <https://en.wikipedia.org/wiki/Demographics_of_Toronto> [↑](#footnote-ref-1)
2. <https://en.wikipedia.org/wiki/Vancouver#Demographics> [↑](#footnote-ref-2)