1 Problem description and background discussion

1.1 Problem description

Toronto as it the most populous city in Canada and the largest urban and metro area, with a population density of 4,149.5 people per square kilometer (10,750/sq mi) (1). The metro area of the city sprawls outward to a total surface area of 5,905.71 km2 (2,280.21 sq mi). Over the next 20 years, Toronto is expected to continue its moderate growth, surpassing 3 million by 2026, and reaching nearly 3.2 million in 2036. Foreign-born people account for nearly half of the population of Toronto. It is also reported that nearly 73% of the Canadians like to experience the ethnic food and flavors (2). There has been a recent increase the appreciation of the Afghan food(3). Keeping this statistics in mind company X is interested in opening a branch of their well-known Afghan restaurants in Toronto. They are looking for a suitable location to establish their new venture. They require some data-driven analysis of different neighborhoods to select an optimal location. The ideal location that they are interested in should be in a neighborhood popular for its eatery places and devoid of any existing Afghan restaurant. The company is also interested if they could be given more than one option, so that they can select an option by keeping in view of their economics. Let's see how data science can help in their selection process.

1.2 Background discussion

Data driven decision are becoming a norm in financial industry. Machine learning techniques can aid in making these decisions if enough and good quality data is available. In this study machine learning algorithm of *k*-mean clustering is applied for clustering of Toronto neighborhoods to identify a suitable location for a new Afghan restaurant. *k*-mean clustering is a type of unsupervised learning, mostly used for unlabeled data. The algorithm works iteratively to assign each data point to one of *k* groups based on the features that are provided. Data points are clustered based on feature similarity (4). Therefore, by using this algorithm we will be able to cluster neighborhoods with similar venues and may be able to identify more than one neighborhood suitable for the location of new restaurant. Initially the data of different boroughs is collected by scrapping a relevant webpage from Wikipedia. Then data wrangling, which involves data cleaning and transforming it to desired data frames is performed. In next step, the venue data for each of borough is collected from Foursquare and is transformed to data frames. Then top 20 venues data for each neighborhood are used in *k*-mean clustering algorithm. The resultant clusters are analyzed for good number of food venues but no Afghan restaurant. The results show that borough xyz is the most suitable option to establish a new Afghan restaurant.

2 References

1. Toronto Population 2019 [Available from: http://worldpopulationreview.com/world-cities/toronto-population/.
2. Canadians are craving ethnic foods 2019 [Available from: http://www.canadiangrocer.com/top-stories/canadians-increasingly-receptive-to-ethnic-foods-65179.
3. WHERE TO EAT THIS WEEKEND: OTTAWA 2019 [Available from: https://www.foodbloggersofcanada.com/where-to-eat-this-weekend-ottawa/.
4. Garbade MJ. Understanding K-means Clustering in Machine Learning 2019 [Available from: https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1.