Coursera IBM Data Science Capstone Project

Cloud Kitchen in City of Toronto

March 2020

Introduction where you discuss the business problem and who would be interested in this project.

A client is interested in setting up cloud kitchen in the city of Toronto. A cloud kitchen is a restaurant that has no physical space and no dine-in or takeaway facility as it does not exist as a restaurant of any sort. It relies entirely on third-party integrations or home delivery orders placed on call through a website or an app. This format gives the restaurant the flexibility to more than one brand using the same kitchen infrastructure. The restaurateur (client) currently owns an Iranian restaurant. He would also like to serve Lebanese and Egyptian cuisine to his customers under a different brand name. These cuisines can be grouped into "Middle Eastern" category since foursquare data has a similar category. A cloud kitchen will enable this business quite perfectly and the client is looking for a perfect location in Toronto to start his cloud kitchen. The target audience is the restaurant owner who is looking for a suitable location in the big city of Toronto. He is looking to leverage his existing business and expand through cloud kitchen. By providing recommendation for a suitable location, the client will be able to ideally build his cloud kitchen to expand his business.

Data where you describe the data that will be used to solve the problem and the source of the data.

- We will use the Notebook to build the code to scrape the following Wikipedia page, https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M, in order to obtain the data that is in the table of postal codes and to transform the data into a pandas dataframe.
- The dataframe will consist of three columns: Postal Code, Borough, and Neighborhood
- The data will then be divided into different boroughs which will be further divided into neighborhoods.
- Foursquare API will then be used to retrieve restaurant data of Toronto city.
- Using data analysis techniques that have been acquired over the last few months, three perfect locations to setup a cloud kitchen will be recommended to the client.
- The recommendation will be based on below criteria:
    - Number of competing restaurants in a particular neighborhood
    - Proximity to residential areas o Optimal location in terms of real estate cost (Additional data will be required for this criteria)
    - Conclude three best location for a cloud kitchen in Toronto serving Iranian, Lebanese and Egyptian cuisine

Methodology section which represents the main component of the report where you discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, if any, and what machine learnings were used and why.

First, I need to get the list of neighborhoods in Toronto, Canada. This is possible by extracting the list of neighborhoods from wikipedia page ("https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M") I did the web scraping by utilizing pandas html table scraping method as it is easier and more convenient to pull tabular data directly from a web page into dataframe. However, it is only a list of neighborhood names and postal codes. I will need to get their coordinates to utilize Foursquare to pull the list of venues near these

neighborhoods. To get the coordinates, I tried using Geocoder package but it was not working so I used the csv file provided by IBM team to match the coordinates of Toronto neighborhoods. After gathering all these coordinates, I visualized the map of Toronto using Folium package to verify whether these are correct coordinates. Next, I use Foursquare API to pull the list of top 100 venues within 500 meters radius. I have created a Foursquare developer account in order to obtain account ID and API key to pull the data. From Foursquare, I am able to pull the names, categories, latitude and longitude of the venues. With this data, I can also check how many unique categories that I can get from these venues. Then, I analyze each neighborhood by grouping the rows by neighborhood and taking the mean on the frequency of occurrence of each venue category. This is to prepare clustering to be done later. Here, I made a justification to specifically look for "Middle Eastern restaurants". Previously, when I ran the model, I was looking for "Iranian restaurants" but there are very few results (maybe due to Foursquare categorization) so I looked for the restaurants closest to Iranian cuisine taste (side note: Iranian food and Middle Eastern food are very similar in taste, so my justification is that if there are people who enjoyed Middle Eastern food, they likely are going to enjoy Iranian food too!) Lastly, I performed the clustering method by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and it is highly suited for this project as well. I have clustered the neighbourhoods in Toronto into 3 clusters based on their frequency of occurrence for "Middle Eastern food". Based on the results (the concentration of clusters), I will be able to recommend the ideal location to open the restaurant.

Results section where you discuss the results.
Clusters
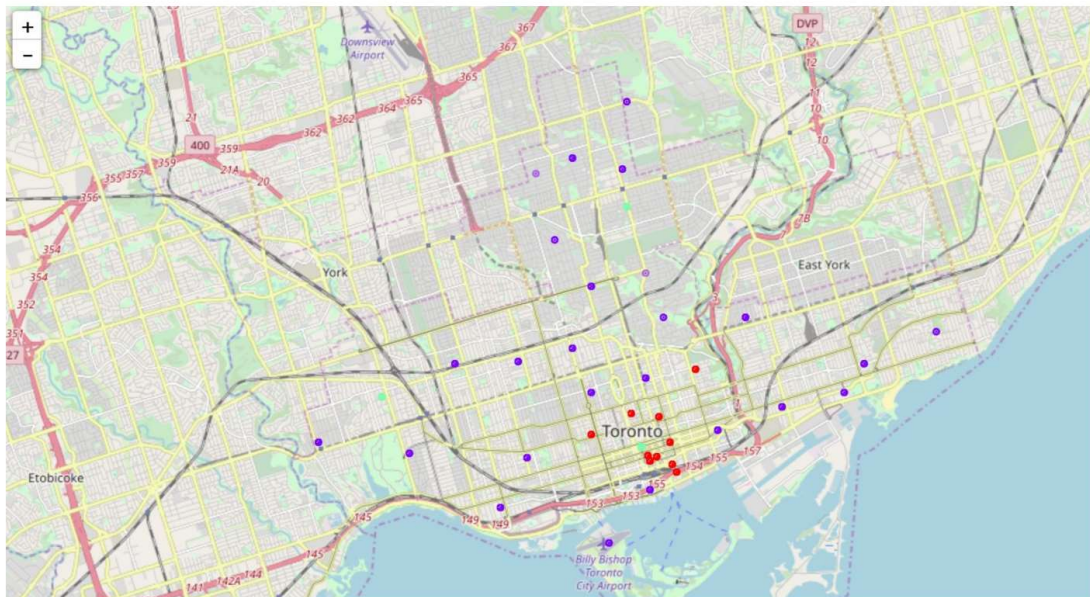
The results from k-means clustering show that we can categorize Toronto neighbourhoods into 3 clusters based on how many Middle Eastern restaurants are in each neighbourhood: ● Cluster 0: Neighborhoods with little or no Middle Eastern restaurants ● Cluster 1: Neighborhoods with some

Middle Eastern restaurants ● Cluster 2: Neighborhoods with high number of Middle Eastern restaurants The results are visualized in the above map with Cluster 0 in red color, Cluster 1 in purple color and Cluster 2 in light green color.

Discussion section where you discuss any observations you noted and any recommendations you can make based on the results.

In this project, I only take into consideration of one factor: the occurrence / existence of Middle Eastern restaurants in each neighbourhood. There are many factors that can be taken into consideration such as population density, income of residents, rent that could influence the decision to open a new restaurant. However, to put all these data into this project is not possible to do within a short time frame for this capstone project. Future research can take into consideration of these factors. In addition, I am relying on the existence of Middle Eastern restaurants only for this project but future research can take into consideration of other variables such as existence of Middle Eastern restaurants, Middle Eastern population level in each neighborhood etc.

Conclusion section where you conclude the report.

Most of Middle Eastern restaurants are in Cluster 1 and 2 which is around Studio District, Garden District, Central Bay Street areas. The lowest (close to zero) in Cluster 0 areas which are Adelaide, King, Richmond. Also, there are good opportunities to open near Harbourfront East, Toronto Islands, Union Station as the competition seems to be low. Looking at nearby venues, it seems Cluster 0 might be a good location as there are not a lot of Middle Eastern restaurants in these areas. Therefore, this project recommends the client to open an authentic Middle Eastern restaurant in these locations with little to no competition. Nonetheless, if the food is authentic, affordable and good taste, I am confident that it will have great following everywhere.

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing the machine learning by utilizing k-means clustering and providing recommendation to the stakeholder.

References

List of neighborhoods in Toronto:
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Foursquare Developer Documentation: https://developer.foursquare.com/docs