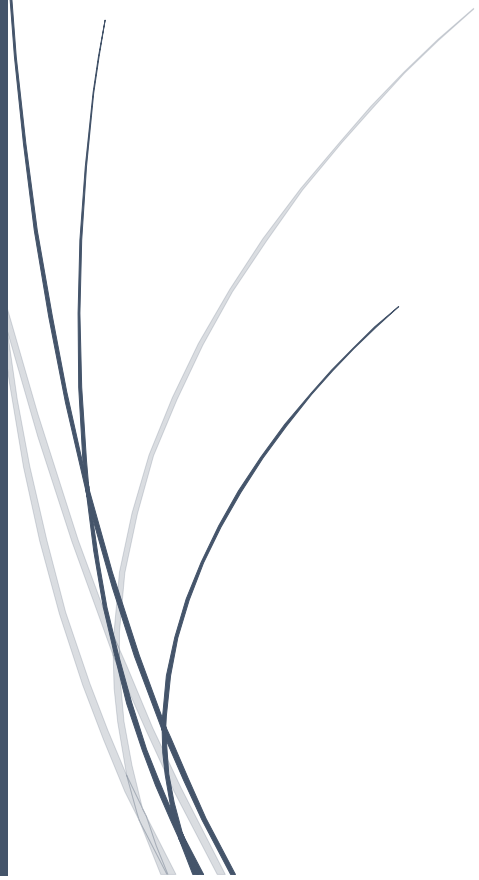
IBM Data Science Capstone

Opening a New Chinese Restaurant in Toronto

Leon Zhao



IBM Data Science Capstone

Opening a New Chinese Restaurant in Toronto

CONTENTS

INTRODUCTION/BUSINESS PROBLEM

Chinese restaurants are a indicator of Chinese cultural influence in many cities. Oftentimes, they are isolated to certain distinct districts or Chinatowns.

This project seeks to determine if there is a pattern in the distribution of Chinese restaurants in the city of Toronto and where would be an optimal location to open a new restaurant serving Chinese cuisine. The optimal location would consider factors that influence a food business such as proximity to a loyal customer base and competition from other restaurants.

DATA

For this problem, we used the following data:

- List of neighborhoods in Toronto
- Geospatial coordinates for neighborhoods
- Venue data of Chinese neighborhoods in the area so that we can do clustering

Data Sources:

- Wikipedia page containing list of neighborhoods of Toronto – Web scrapping technique to extract data from Wikipedia page, Pyhton requests and beautifulsoup packages were used.
- Geographical Coordinates of these neighborhoods – CSV file provided by Coursera (Geospatial_Coordinates.csv)
- Venues data for these neighborhoods – Foursquare API
- Map of Toronto – Python Folium package

METHODS

We get the list of neighborhoods in Toronto from Wikipedia page (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M). We perform web scrapping to extract the list of neighborhoods from the Wikipedia page. We take help of Python requests and beautifulsoup packages in performing this task.

Now, we need to get geographical coordinates for these locations. We read the csv file provided by Coursera (Geospatial_Coordinates.csv) for the coordinates. We could have used Python Geocoder package for this but we used already available data for simplicity. We merge the above two datasets.

Next, we need to get venues data for these neighborhoods. So, we take help of Foursquare API to get this data. We have to register for a developer account with Foursquare to use this API service. We can register for free account for this project. Post registration, we need to generate Foursquare ID and secret key (create an app and then generate on their website). We use these credentials to

call the API in our Python code. Foursquare returns the dataset in JSON file. We need to normalize the JSON file and extract the venue data from it into dataframe.

Then, we group rows of dataset by taking mean of frequency of each venue category. We prepare the dataset for clustering by keeping only the required data in the dataframe.

Lastly, we need to perform clustering on the dataset and analyze the thus obtained datasets all neighborhood.

We analyze the clusters and evaluate which is the best location to open an American restaurant in it on the basis of number of already running American restaurants in those areas.

RESULTS

The highest concentration of Chinese restaurants is in cluster 1, downtown Toronto.

There are locations Adelaide, King, Stn A PO Boxes, Church and Wellesley

CONCLUSION

We can open Chinese restaurants in clusters 1 and 2, in the following locations:

Cluster 1

Adelaide, King, Stn A PO Boxes, Church and Wellesley

Cluster 2

Studio District, Summerhill West, Rathnelly, South Hill, Forest