Creator: diegograndi | Credit: Getty Images/iStockphoto Copyright: diegograndi

Capstone Project - Toronto Restaurant Map

Published on August 4, 2020

Mingyu Cui
Senior Consultant & Data Science Specialist @ IBM GBS

INTRODUCTION / BUSINESS PROBLEM

The continent of North America has a long history of immigration dated back to several centuries. As the general public describes, it is truly a "melting pot" both culture-wise and dinning-wise. As one of the biggest cities of Canada, Toronto has a considerable diverse population in terms of ethnics settling in its lively boroughs and neighborhoods. Meanwhile, attracted by the aesthetic and gourmet characteristics of the city, as well as the free, open and welcoming culture it had, Toronto accommodates millions of tourists from all over the world.

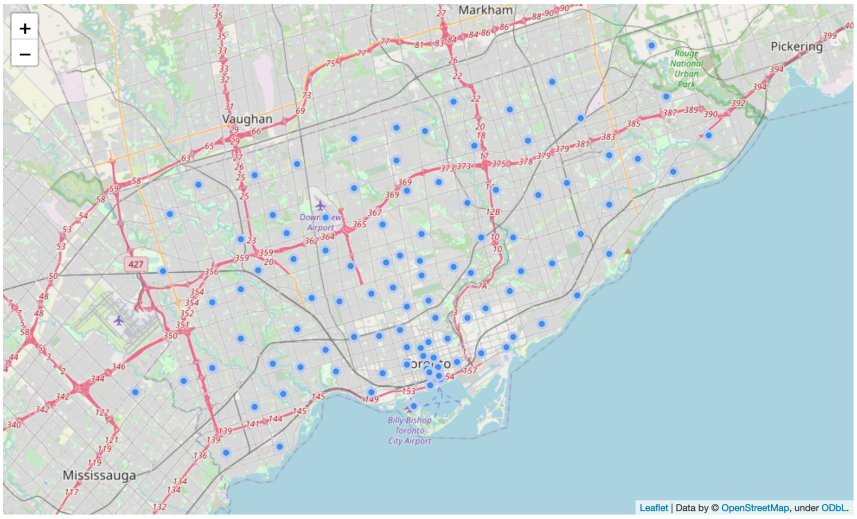
Hypothetical Scenario

Commissioned by a US-based investor who's investing and culturing a few start-up restaurants in Toronto, our objective is to assist the restaurants' operation staff in getting acquainted with the city and setting up standards for assessing and evaluating the restaurants business conditions.

Playback

During the initiating stage we have scrapped Toronto's neighbourhood and postal

from Wikipedia, searched and matched the geographical coordinates against each Postal Code, utilized Foursquares APIs to examine venues around, and eventually segmentedand clustered these neighbourhoods by their unique venue composition in the number of the investor's intended number of restaurants to be opened (*reference to "Segmenting and Clustering Neighbourhoods in Toronto 3: Map clustering" notebook*). Based upon the information we provided, the stakeholders now have a clearer expectation on how should the intended restaurants should be positioned in each of the fitting clusters.



Objective

Now that since all the stakeholders have a concept of how the 5 restaurants shall be distributed and their respective surrounding municipal context, the next step is to further explore venues around them to grasp information on:

1. *Major cuisines in dining themed clusters as suggestions to our restaurants' cuisine selection;*
2. *Ratings of the restaurants in dining themed clusters and how they are distributed. Thus we could set a rating target for our restaurants sitting in the context;*
3. *Naming suggestions for the restaurants in the clusters with theme other than dining, referencing to their venue names.*

Based upon these information our restaurants could then set up and get running.

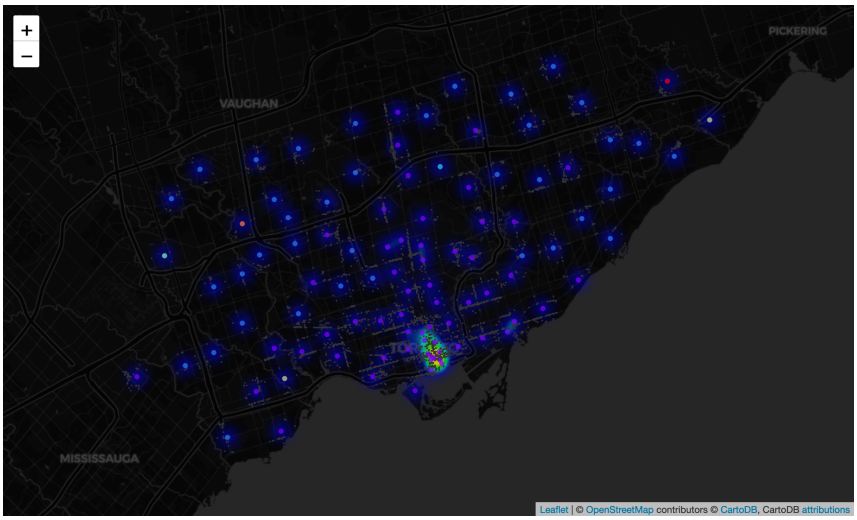
DATA

Base Data

The data to be used as foundation for this research will be the output from last stage, "Segmenting and Clustering Neighbouthoods in Toronto", which is a CSV file containing neighbourhood Postal Codes, boroughs, geographical coordinates, venues of interest and cluster IDs.

The source data is scrapped from Wikipedia: List of postal codes of Canada: M, of which the Postal Codes are used a the primary key of the table. During the process we convert the HTML table to Pandas dataframe, dropped all unassigned entries, and utilized Geopy API to acquire their respective geographical coordinates.

Venue information is then acquired through Foursquares APIs. The numerical limit was set to 100 and search radius is within 1000. The output data includes each venues' latitude, longitude as well as venue category. After normalizing the venue category count in each neighbourhood, with neighbourhoods with zero venues being dropped, the remaining neighourhoods are then clustered into 7 realms by K-MEANS, each has a distinct theme. Our work then start from here.



Additional Data

Before we dive into further exploring and analyzing, additional information will be required. We will utilize Foursquare APIs again to allocate ratings on all "restaurant" category venues in cuisine theme clusters. With all raw ingredients we need scrapped and processed, it's time to cook our insights!

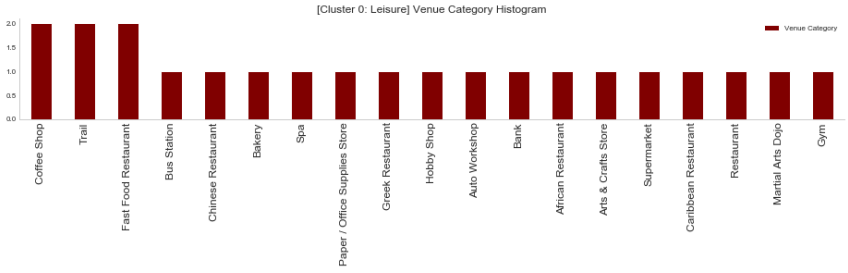
Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	ID	Venue	Venue Latitude	Venue Longitude	Venue Category	Postal Code	Borough	Cluster_Labels
Parkwoods	43.753259	-79.329656	4b8991cbf964a520814232e3	Allwyn's Bakery	43.759840	-79.324719	Caribbean Restaurant	M3A	North York	2
Parkwoods	43.753259	-79.329656	58a8dcaa6119f47b9a94dc05	A&W	43.760643	-79.326865	Fast Food Restaurant	M3A	North York	2
Parkwoods	43.753259	-79.329656	4c0150f4716bc9b65b9dbb55	Spicy Chicken House	43.760639	-79.325671	Chinese Restaurant	M3A	North York	2
Victoria Village	43.725882	-79.315572	4f3ecce6e4b0587016b6f30d	Portugril	43.725819	-79.312785	Portuguese Restaurant	M4A	North York	1
Victoria Village	43.725882	-79.315572	4d689350b6f46dcb77ee15b2	The Frig	43.727051	-79.317418	French Restaurant	M4A	North York	1

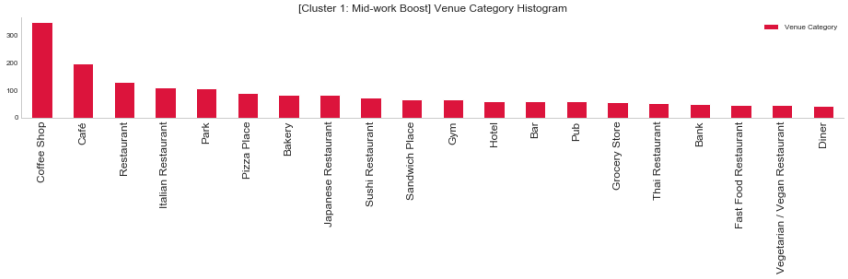
METHODOLOGY

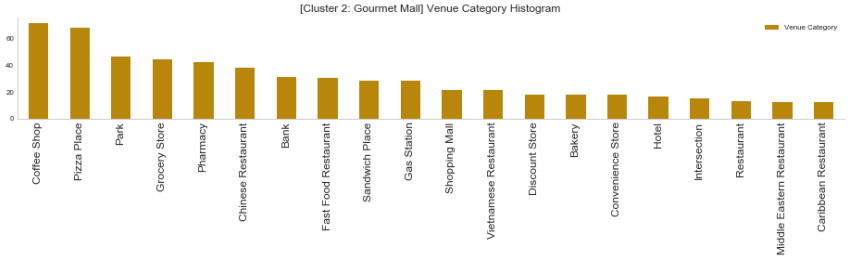
Feature description and Histogram Chart

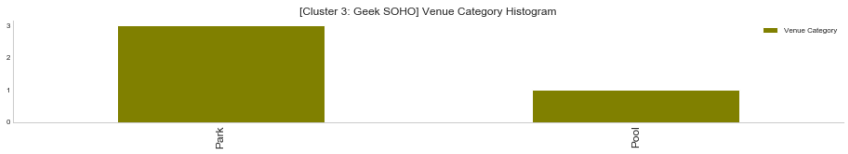
By observing the output data from neighbourhood clustering, we will name each clusters by

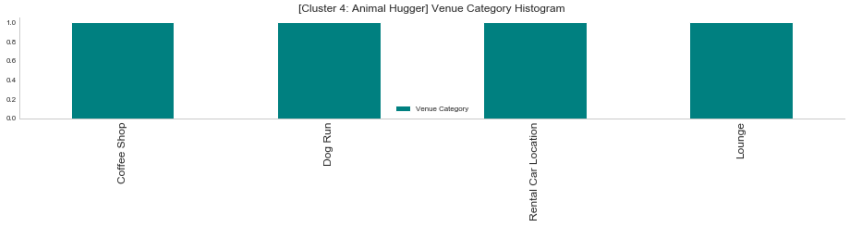
their unique themes. Then we will be able to decide which clusters could accommodate 23 restaurants of what type. After that, we will utilize histogram charts to visualize the capacity of venue categories in each clusters. Thus we will be able to **give suggestions on cluster selection by their dining needs capacity**.

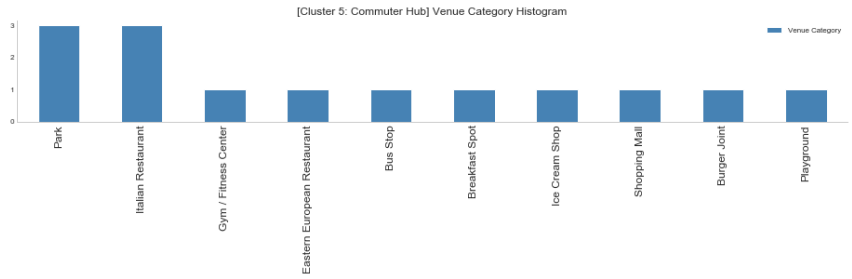


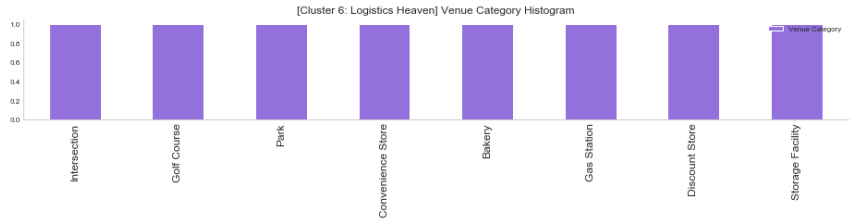








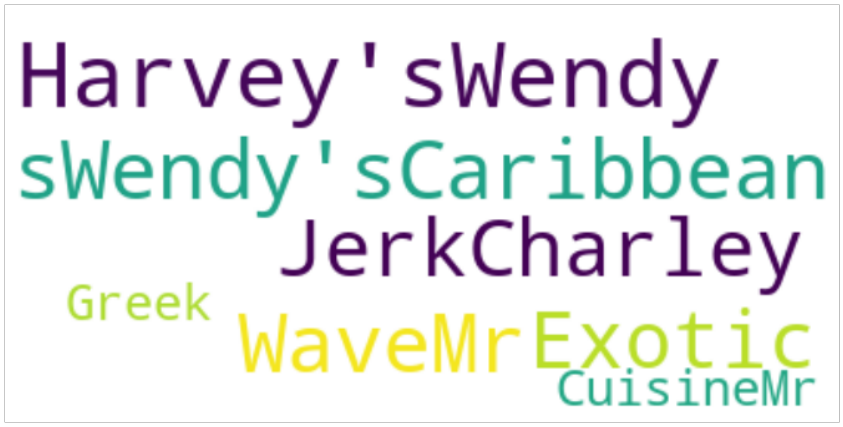




Wordcloud

In cuisine themed clusters, we are going to use word cloud to identify the top cuisines in terms of total number count. Therefore **obtaining the suggeted cuisines which is most likely to acquire customers**. By applying the same methodology, we can also get inspirations for naming suggestions in each selected neighbourhood.

Cluster 0: Leisure



Cluster 1: Mid-work Boost

Cluster 2: Gourmet Mall

Cluster 5: Commuter Hub

Foursquares API

We will scrap ratings of each restaurants in cuisine themed clusters by calling Foursquares API. Then once again, using Histogram to visualize the volume distribution of the rating scores. By doing so, we can give **suggestions on restaurant rating targets** based on these numbers.

RESULTS

Location-based Tactics

The output from K-means analysis allows us to define the theme of the 7 clusters. Amongst them, cluster 2 (Gourmet Mall) is densely populated by restaurants, followed by cluster 1 (Mid-work Boost) and cluster 0 (Leisure) where coffee shops and pizza places take a larger chunk of the venue types. Cluster 5 (Commuter Hub) has the least number of restaurants. Cluster 3 (Geek SOHO), 4 (Animal Hugger) and 6 (Logistics Heaven) has no restaurants at all. Restaurants opened in these clusters (0: Leisure, 1: Mid-work Boost and 2: Gourmet Mall) will face brutal competition.

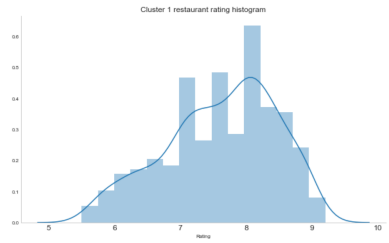
After analyze the name of the restaurants in cluster 0 (Leisure), 1 (Mid-work Boost), 2 (Gourmet Mall) and 5 (Commuter Hub), if we were to open a restaurant in these regions, it will be the best to serve **Caribbean and Mediterranean** cuisines in **Cluster 0: Gourmet Mall**, **Japanese, Italian and Thai** cuisines in **Cluster 1: Mid-work Boost**, **Chinese and Vietnamese** cuisines in **Cluster 2: Leisure**, and possibly **European** cuisines in **Cluster 5: Commuter Hub**(asserted by the name convention in that cluster).

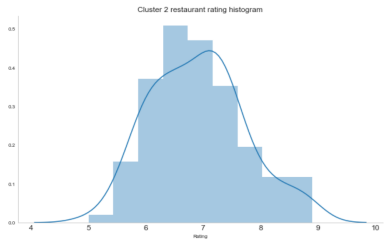
	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	ID	Venue	Venue Latitude	Venue Longitude	Venue Category	Postal Code	Borough	Theme	Rating
0	Parkwoods	43.753259	-79.329656	4b8991cbf964a520814232e3	Allwyn's Bakery	43.759840	-79.324719	Caribbean Restaurant	M3A	North York	Gourmet Mall	6.3
1	Parkwoods	43.753259	-79.329656	58a8dcaa6119f47b9a94dc05	A&W	43.760643	-79.326865	Fast Food Restaurant	M3A	North York	Gourmet Mall	6.3
2	Parkwoods	43.753259	-79.329656	4c0150f4716bc9b65b9dbb55	Spicy Chicken House	43.760639	-79.325671	Chinese Restaurant	M3A	North York	Gourmet Mall	6.3
3	Victoria Village	43.725882	-79.315572	4f3ecce6e4b0587016b6f30d	Portugril	43.725819	-79.312785	Portuguese Restaurant	M4A	North York	Mid-work Boost	6.3
4	Victoria Village	43.725882	-79.315572	4d689350b6f46dcb77ee15b2	The Frig	43.727051	-79.317418	French Restaurant	M4A	North York	Mid-work Boost	6.3

Restaurant Ratings

At the final stage of our data analysis, we used Foursquares API to get the ratings from all listed restaurants in the 4 clusters using venue ID as the search key, and store the returned ratings in the restaurant dataframe we created. The final outputted Histograms reveals that:

- **Cluster 1: Mid-work Boost** has the highest average restaurant rating, while the top rated restaurant is from **Cluster 1: Mid-work Boost** as well;
- Restaurant ratings in **Cluster 0: Leisure** has a **relatively concentrated** distribution with the top 10% lies in the range of **6.0** to **6.5**, with an extremely small sample;
- Restaurant ratings in **Cluster 1: Mid-work Boost** has the **most-widely spreaded** distribution with the top 10% lies in the range of **8.5** to **9.2**;
- Restaurant ratings in **Cluster 2: Gourmet Mall** has a **relatively spreaded** distribution with the top 10% lies in the range of **7.5** to **8.8**;
- Restaurant ratings in **Cluster 5: Commuter Hub** has a mean rating of **7.7** with an extremely small sample;





RESULTS

Location-based Tactics

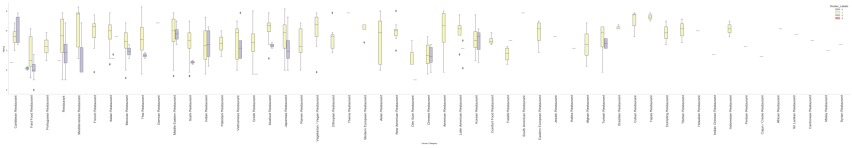
The output from K-means analysis allows us to define the theme of the 7 clusters. Amongst them, cluster 2 (Gourmet Mall) is densely populated by restaurants, followed by cluster 1 (Mid-work Boost) and cluster 0 (Leisure) where coffee shops and pizza places take a larger chunk of the venue types. Cluster 5 (Commuter Hub) has the least number of restaurants. Cluster 3 (Geek SOHO), 4 (Animal Hugger) and 6 (Logistics Heaven) has no restaurants at all. Restaurants opened in these clusters (0: Leisure, 1: Mid-work Boost and 2: Gourmet Mall) will face brutal competition.

After analyze the name of the restaurants in cluster 0 (Leisure), 1 (Mid-work Boost), 2 (Gourmet Mall) and 5 (Commuter Hub), if we were to open a restaurant in these regions, it will be the best to serve **Caribbean and Mediterranean** cuisines in **Cluster 0: Gourmet Mall**, **Japanese, Italian and Thai** cuisines in **Cluster 1: Mid-work Boost**, **Chinese and Vietnamese** cuisines in **Cluster 2: Leisure**, and possibly **European** cuisines in **Cluster 5: Commuter Hub**(asserted by the name convention in that cluster).

Restaurant Ratings

At the final stage of our data analysis, we used Foursquares API to get the ratings from all listed restaurants in the 4 clusters using venue ID as the search key, and store the returned ratings in the restaurant dataframe we created. The final outputted Histograms reveals that:

- **Cluster 1: Mid-work Boost** has the highest average restaurant rating, while the top rated restaurant is from **Cluster 1: Mid-work Boost** as well;
- Restaurant ratings in **Cluster 0: Leisure** has a **relatively concentrated** distribution with the top 10% lies in the range of **6.0** to **6.5**, with an extremely small sample;
- Restaurant ratings in **Cluster 1: Mid-work Boost** has the **most-widely spreaded** distribution with the top 10% lies in the range of **8.5** to **9.2**;
- Restaurant ratings in **Cluster 2: Gourmet Mall** has a **relatively spreaded** distribution with the top 10% lies in the range of **7.5** to **8.8**;
- Restaurant ratings in **Cluster 5: Commuter Hub** has a mean rating of **7.7** with an extremely small sample;



DISCUSSION

Location Selection

By extracting all "Restaurant" tags from Venue Category, we have limited our scope down to venues falling in restaurant category in each clusters. Filtered with the selected tags, it reveals that Cluster 3 (Geek SOHO), 4 (Animal Hugger) and 6 (Logistics Heaven) has ZERO restaurant. Further examination on specific venue category composition of the 3 clusters shows that these neighbourhood are suburban regions with land uses mostly dedicated for public storage, trails, parks, etc. which accommodate very small residential population. A proposed restaurant in these clusters will expect abhorring average count of customers per day and could highly possibly result in a failure in terms of investment. Therefore, suggested locations are within Cluster 0 (Leisure), 1 (Mid-work Boost), 2 (Gourmet Mall) and 5 (Commuter Hub). among which Cluster 1 and 2 has the largest numbers of restaurants with difference in their cuisines. Cluster 0 and 5 are second best choises if taking other factors (utility infrastructures, rent prices, consumption level of the approximate areas, etc.) into consideration.

Speaking of proposed cuisines, Fast Food is the dominion in most (0, 1, 2) clusters and has the largest total number in the entire city. If we are purely seeking revenues, Fast Food restaurant is definitely the best choice of all. If otherwise we are seeking a long term market brand, it's inevitable to consider a cuisine's cultural background.

Caribbean cuisine, Japanese cuisine (especially Sushi), Chinese cuisine and Eastern European cuisin are the first choices for restaurants in Cluster 0, 1, 2 and 5 respectively. For Cluster 0, the second best choice is Mediterranean cuisines. Meanwhile Cluster 1 has the best diversity, with its second and third best choices being Italian and Thai cuisines. At the same time, we should notice that Vegiterian cuisines has a strong presence in this Cluster. Solely from the data at hand we can tell that there's a good chance that Chinatown is located in Cluster 2 neighbourhoods, but Vietnamese cuisines also occupies a considerable chunck of the pie. Cluster 5 has very strong characteristics from the opposite shore of the Atlantic Ocean by naming its restaurant in a European fashion. Last but not the least, for restaurant naming, it's a good tactic to follow the naming conventions of the servicing cuisines' origin country or area.

Operation Expectation

The investors expected all proposed restaurants could rank among top 10% of all restaurants in their clusters respectively within 3 years since opening. In the 4 targeted clusters, the proposed restaurants' ratings have to reach a certain score to meet our expectations. By examining the histograms of restaurant rating distribution in each of the four restaurant-accommodating clusters, we could confidently decide that:

- Restaurant sitting in **Cluster 0: Leisure will** be expected to achieve a rating higher than

- Restaurant sitting in **Cluster 1: Mid-work Boost** will be expected to achieve a rating higher than **8.5** to rank among the top 10% of all restaurants in this cluster;
- Restaurant sitting in **Cluster 2: Gourmet Mall** will be expected to achieve a rating higher than **7.0** to rank among the top 10% of all restaurants in this cluster;
- Restaurant sitting in **Cluster 5: Commuter Hub** will be expected to achieve a rating higher than **7.7** to rank among the top 10% of all restaurants in this cluster;

The rating data used in this research is acquired from the single source of Foursquares and therefore the data is to some extent biased since Foursquare is a professional geo-information provider rather than a specialized restaurant rating platform. Even among location service providers, such information could vary slightly in different source and statistical scopes (in perspective of customer groups). Not even mention the data from providers such as Yelp of which consumer active rate is much higher, rating participant population is larger, and customer spectrum is more focused, that might be more accurate. But, in the business scenario set for this research, the accuracy is good enough for an early stage targeting on restaurant operation expectation. The diviant contributes little in defecting the findings' business value and therefore can be omitted.

One last thing is that if we observe the acquired rating data closely, there are duplicated venues. This is because the way we getting the venues is by defining the distance radius (1000) from the centroid geographical coordinates of each borough / Postal Code. This methodology forced us to choose between a better coverage (with duplicated venue data) and a less confusing data (with defect on venue sampling in each borough). Surely we can drop the duplicants when scrapping the ratings, but the reason not doing so is because firstly, the duplication amount is small enough to be accepted; secondly, since no borough is shaped in a perfect circle which perfectly fits the venue coverage scope created by our venue searching methodology, if not examine the address of each venue, we cannot completely eliminate the possibility of incorrectly categorizing venue from one borough into another. Not mentioning the total area of the boroughs are not equal and the variance is considerably great, which can be revealed by observing our Folium visualizations and verified by the fact that the closer a borough is to downtown, the more duplicated venues we get. As a conclusion on this topic, it is an error born with the methodology we employed for the reasion of simplicity and therefore cannot be alleviated easily without employing a more sophisticated (to be accurate, coordinate-sensitive) venue searching methodology.

CONCLUSION

Business Values of Applied Data Science

In this project we utilized multiple libraries tackling data analyzing as well as Foursquare APIs with python to gain business insights on geogrsphical and social context. These emerging tools help us gain a complete market view without employing traditional ways such as sending investigators or speading questionnaires, which cost dearly in terms of time and expenses. Surely there is improvement space for the process above, as mentioned in DISCUSSION section. The model we use can be improved and certain errors coming with the methodology can be mitigated to some extent. But as we have discussed and concluded

the effect of these errors on our results can be omitted considering the business stage we're at and the scope we are dealing with. Hence we can wrap everything up here and,

Until next time!

Mingyu Cui

Senior Consultant & Data Science Specialist @ IBM GBS
