1. Project Requirements and Criteria

Part1

1. A description of the problem and a discussion of the background. (15 marks)
2. A description of the data and how it will be used to solve the problem. (15 marks)

Part1

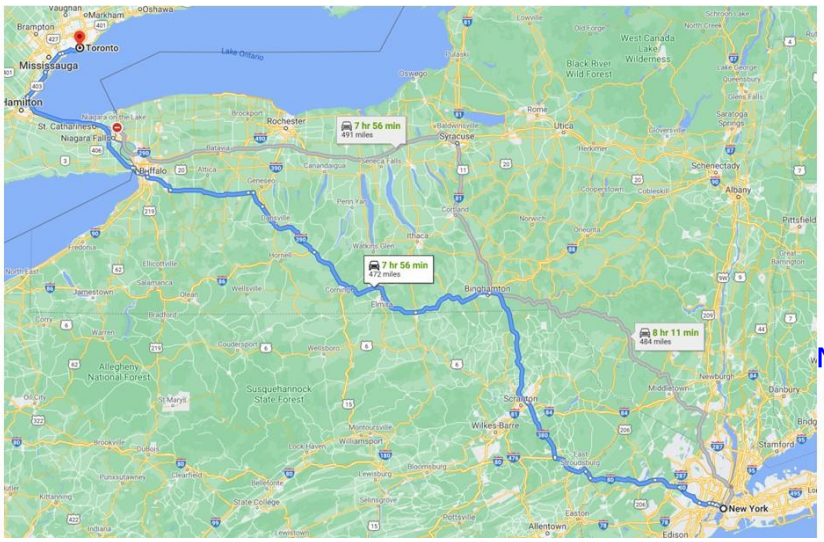
1. A link to your Notebook on your Github repository, showing your code. (15 marks)
2. A full report consisting of all of the following components (15 marks):
   - Introduction where you discuss the business problem and who would be interested in this project.
   - Data where you describe the data that will be used to solve the problem and the source of the data.
   - 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.
   - Results section where you discuss the results.
   - Discussion section where you discuss any observations you noted and any recommendations you can make based on the results.
   - Conclusion section where you conclude the report.
3. Your choice of a presentation or blogpost. (10 marks)

2. Background and Business Problem Definition

A construction company called ***ABC*** makes quite good business in New York city. This companies builds condos, semi-detached houses and houses according to the needs of New York customers and achieved quite good business profit.

New York is the biggest city and economic center of US, Toronto is the biggest city and economic center of Canada. Both cities are mage cities and have some similarity. But there might be also differences.

Toronto



New York

As US and Canada have strong economic tire and both countries are in the NAFTA (now USMCA) ， most of the successful companies will naturally plan to extend the business to the other side of the border.

New York is geographically close to Toronto, so ***ABC*** plans to extend the business in Toronto. Although there are many similarities of the 2 big cities, the companies still want to compare the 2 cities and know what kind of houses are most wanted in Toronto. So ***ABC*** company hires data scientist to use geographic data and other data source to help to understand 2 things:

(1) Are the Toronto neighborhoods similar as New York?

Even this is not directly linked to the house construction business, but the company wants to know what kind of venues are in Toronto neighborhood, what is the savor of the city.

The ***ABC*** company can reuse the design and modeling experience they have built in New York while they also want to know what they should do differently in Toronto.

This impacts the soft customer perceived values in the house design, decoration and so on. So, it is very important.

(2) What kind of houses the potential clients in Toronto most want?

Do they like Condos over town houses or houses? Real estate with how many rooms is most popular in each different neighborhood? What kind of family size and income the potential customers have?

Those are the key information needed to make business decisions.

The data scientist needs to dig out the necessary information to support ***ABC*** company for its business decision which market segments it should focus, which customers it should target. And accordingly, ***ABC*** will design, build and provide the dwelling solutions to meet Toronto customer's needs.

3. Data Source

3.1 Types of Data

There are some types of data needed:
- The geolocation data of Toronto areas. It should have the following info:
  - The neighborhood names
  - The standard and unique identity of each neighborhoods (e.g., Post Code)
  - The information about the venues (categories, location info)

- The list of neighborhood identities.
This is important as this can be used as index to filter the data and to scale the analysis (e.g., limit the analysis to Toronto downtown only; e.g., to expand the analysis to Greater Toronto Area like Markham, Richmond Hill if needed)

- The detail information of each neighborhood
  - The standard and unique identity of each neighborhoods (e.g., Post Code)
  - Statistic information about dwelling (size, number of rooms, type of dwelling)
  - Statistic information about income, age
  - Statistic information about education
  - Statistic information about family size

Not all the information will be used for the analysis. But with the important data collected, it is very easy to plug other analysis or to add dimensions of analysis.

3.2 Sources of Data

The geolocation data

- We can start with New York data used in the course from the following link:
  https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs/newyork_data.json

  This data has 306 “features”. The information will help to get the mapping of ['Borough', 'Neighborhood', 'Latitude', 'Longitude'] to start the next step


```
{'type': 'Feature',
 'id': 'nyu_2451_34572.1',
 'geometry': {'type': 'Point',
  'coordinates': [-73.84720052054902, 40.894705176
 'geometry_name': 'geom',
 'properties': {'name': 'Wakefield',
  'stacked': 1,
  'annoline1': 'Wakefield',
  'annoline2': None,
  'annoline3': None,
  'annoangle': 0.0,
  'borough': 'Bronx',
  'bbox': [-73.84720052054902,
   40.89470517661,
   -73.84720052054902,
   40.89470517661]}}
```


	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

- We can start with Toronto data from the following links:
  https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
  It is a list of mapping among ['Post Codes', 'Borough', 'Neighborhood'], but without the Latitude and Longitude information

Postal Code	Borough	Neighbourhood
M3A	North York	Parkwoods
M4A	North York	Victoria Village
M5A	Downtown Toronto	Regent Park, Harbourfront
M6A	North York	Lawrence Manor, Lawrence Heights
M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government

http://cocl.us/Geospatial_data
It is a list of mapping among ['Postal Code', 'Latitude', 'Longitude']

With the above, it should be easy to get the similar data like New York.

	Borough	Neighborhood	Latitude	Longitude
0	North York	Parkwoods	43.753259	-79.329656
1	North York	Victoria Village	43.725882	-79.315572
2	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
3	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494

- The FourSquare geolocation data

Once we have the list of neighborhoods and their latitudes and longitudes, we can pull out more information by using FourSquare queries

Remark: As there are >300 neighborhoods in New York and >100 neighborhoods in Toronto, but the FourSquare sandbox account can only have 950 queries/day, once we are confident with the data, we should save the pulled data into IBM storage. The future testing and debugging can be done on the saved data sets. This saves the queries.

- The Statistic Canada Census data:
https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/details/download-telecharger/comp/page_dl-tc.cfm?Lang=E
This data is the 2016 census data. For each Post Code, there are 2247 lines of data to cover different aspects:

0. General (8 lines)
1. Population age distribution (26 lines)
2. Dwelling structure (28 lines)
3. Family structure (41 lines)
4. Knowledge on languages (561 lines)
5. Income (211 lines)
6. Language (263 lines)
7. Citizenship and migration status (482 line)

8. Dwelling situation	(66 lines)
9. Education	(182 lines)
10. Career related information	(85 lines)
11. Working languages	(289 lines)
12. Mobility	(9 lines)

Not all the data will be used. But with the experiment of using very small part of the data, there is possibility to explore more analysis with downloaded data.

This is free data set from Canada government. And there are more updated data with more details available with subscription payment. Here we only demonstrate the usage of the data with data science methodologies.

4 How to solve the problems with the data

Here I only give out high level thinking.

(1) Question regarding the similarity of New York and Toronto

- Pull out the venue categories of New York and Toronto neighborhoods
- Use the K-Means clustering to break down the neighborhood into 5 clusters
    - Check the similarity of cluster distribution
    - Check the most popular venue categories of in the most popular cluster to see if there is any similarity and difference
- Use histogram to compare the top 10 venue categories of all neighborhoods between New York and Toronto

(2) Question regarding the dwelling types in Toronto

As there are many possible usages of the data, we only focus on “Average number of rooms per dwelling” and “Average household size” to give the indication how many rooms (including bed rooms) a house should in different neighborhood.

Of course, with the same logic, reference information about whether a condominium is more popular, how many rooms a dwelling should have and so on.

1619	8. Dwelling situation	Total - Private households by tenure - 25% sample data
1620		Owner
1621		Renter
1622		Band housing
1623		Total - Occupied private dwellings by condominium status - 25% sample data
1624		Condominium
1625		Not condominium
1626		Total - Occupied private dwellings by number of bedrooms - 25% sample data
1627		No bedrooms
1628		1 bedroom
1629		2 bedrooms
1630		3 bedrooms
1631		4 or more bedrooms
1632		Total - Occupied private dwellings by number of rooms - 25% sample data
1633		1 to 4 rooms
1634		5 rooms
1635		6 rooms
1636		7 rooms
1637		8 or more rooms
1638		Average number of rooms per dwelling
1639		Total - Private households by number of persons per room - 25% sample data
1640		One person or fewer per room
1641		More than 1 person per room
1642		Total - Private households by housing suitability - 25% sample data

But we only focus on the “Average number of rooms” and “Average dwelling value” as reference. And with the data available, a professional should be able to analysis other data presentation. And with the census data from difference years, additional information can be explored (e.g. whether the population is flowing in or out, whether the age in the neighborhood becomes older or younger, whether people’s income is increasing to afford more expensive real estates).

With the time limitation, only “Average number of rooms per dwelling” or “Average household size” will be presented in detail.

Other possibilities will be also provided to ***ABC*** company for the next step investigation.

The key to link different data sets is the post codes. Different data sets have different “index” like:

- the names of communities
- the names of neighborhoods
- the names of borough

And they are not standardized!!! E.g. there is no standard definition of “Yorkdale” area and “North and East York”.

But the post code is very standardized. We use the post code mapping to link different data sets.

There are some details to clean the data, e.g.

- M7A, M7R, M7Y are the Post Codes reserved for Ontario Government, Post Canada Forwarding and Processing Center, there is no residence, need to be removed
- And I found M5K, M5L, M5W, M5X are the Post Codes reserved for "Toronto Dominion Centre, Design Exchange", "Commerce Court, Victoria Hotel", "Stn A PO Boxes", "First Canadian Place, Underground city". There is no residence and “value for dwelling”, “number of rooms” in those post codes are 0. They need to be filtered out also.

5 Discovery, result and findings

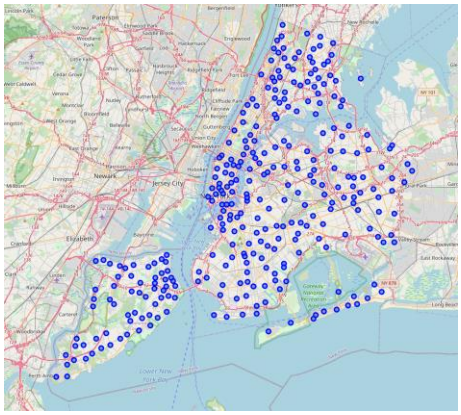
In this session, I will introduce the finds step by step following the coding.

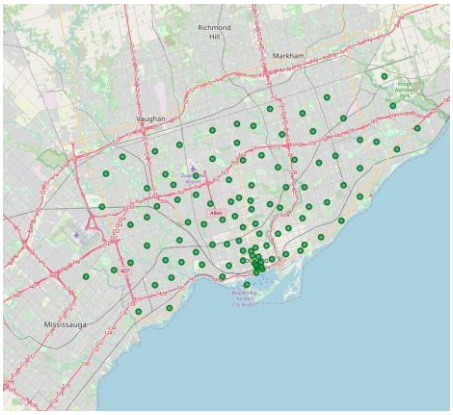
1. Get the data set of New York and Toronto with “Neighborhood”, “Latitude”, “Longitude”

New York data json file from previous course is used https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs/newyork_data.json

And for Toronto, the information is retrived from https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M and http://cocl.us/Geospatial_data. The previous one is a table about “Neighborhood – Post Code” and the second one is a table of “Post Code – Latitude - Longitude”

With the Lati/Logi information, we should be able to address neighborhoods of New York and Toronto on the maps





2. Get the venue information of New York and Toronto

Once we have the neighborhood data with Latitude and Longitude info, we should be able to pull out the venue info by querying Four Square geo data

FOURSQUARE

There are >300 neighborhoods in New York and >100 neighborhoods in Toronto, the four square sanbox account only allows 950 queries per day. We have to save the result to IBM Cloud Storage.

The result is a >10,000 New York venues and >2,000 Toronto venues in the following format:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Parkwoods	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park
1	Parkwoods	43.753259	-79.329656	Variety Store	43.751974	-79.333114	Food & Drink Shop
2	Victoria Village	43.725882	-79.315572	Victoria Village Arena	43.723481	-79.315635	Hockey Arena
3	Victoria Village	43.725882	-79.315572	Portugril	43.725819	-79.312785	Portuguese Restaurant
4	Victoria Village	43.725882	-79.315572	Tim Hortons	43.725517	-79.313103	Coffee Shop
...	...	...	...	...	...	...	...
2113	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999	Islington Florist & Nursery	43.630156	-79.518718	Flower Shop
2114	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999	Koala Tan Tanning Salon & Sunless Spa	43.631370	-79.519006	Tanning Salon
2115	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999	Once Upon A Child	43.631075	-79.518290	Kids Store
2116	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999	Kingsway Boxing Club	43.627254	-79.526684	Gym
2117	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999	Burrito Boyz	43.626657	-79.526349	Burrito Place

With the 10045 venues in New York, there are 440 unique categories.

With the 2112 venues in Toronto, there are 265 unique categories.

The venue categories are used to compare New York and Toronto

New York

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
1	Wakefield	40.894705	-73.847201	Rite Aid	40.896649	-73.844846	Pharmacy
2	Wakefield	40.894705	-73.847201	Walgreens	40.896528	-73.844700	Pharmacy
3	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop
4	Wakefield	40.894705	-73.847201	Dunkin'	40.890459	-73.849089	Donut Shop

(10045, 7)
There are 440 uniques categories.

Toronto

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Parkwoods	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park
1	Parkwoods	43.753259	-79.329656	Variety Store	43.751974	-79.333114	Food & Drink Shop
2	Victoria Village	43.725882	-79.315572	Victoria Village Arena	43.723481	-79.315635	Hockey Arena
3	Victoria Village	43.725882	-79.315572	Portugril	43.725819	-79.312785	Portuguese Restaurant
4	Victoria Village	43.725882	-79.315572	Tim Hortons	43.725517	-79.313103	Coffee Shop

(2112, 7)
There are 265 uniques categories.

3. Analysis

For each neighborhood, we should be able to get the information of venue category frequency (0-100%), just to check the similarity and differences between neighborhoods.

——Allerton——

	venue	freq
0	Pizza Place	0.12
1	Deli / Bodega	0.08
2	Supermarket	0.08
3	Chinese Restaurant	0.08
4	Department Store	0.04

——Annadale——

	venue	freq
0	Liquor Store	0.11
1	Diner	0.11
2	Train Station	0.11
3	Park	0.11
4	Pizza Place	0.11

——Arden Heights——

	venue	freq
0	Pharmacy	0.25
1	Coffee Shop	0.25
2	Bus Stop	0.25
3	Pizza Place	0.25
4	Outlet Store	0.00

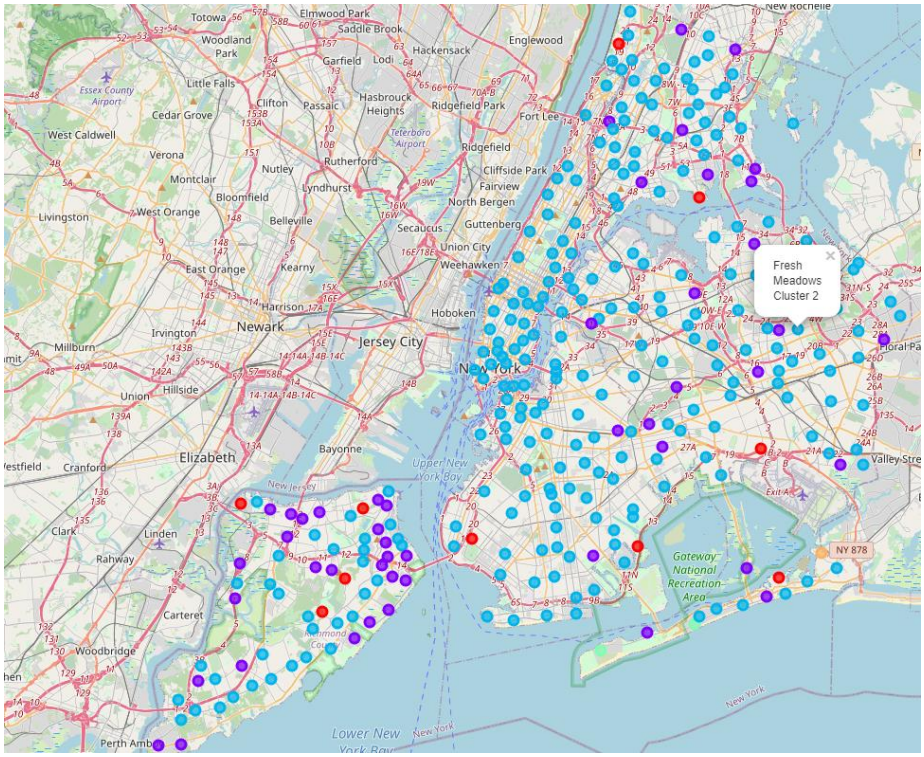
And for each neighborhood in New York, we listed the top venue categories

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Allerton	Pizza Place	Bus Station	Supermarket	Chinese Restaurant	Deli / Bodega
1	Annadale	Pizza Place	Pharmacy	Train Station	Liquor Store	Food
2	Arden Heights	Pizza Place	Deli / Bodega	Pharmacy	Coffee Shop	Bus Stop
3	Arlington	Boat or Ferry	Deli / Bodega	Intersection	American Restaurant	Coffee Shop
4	Arrochar	Bus Stop	Deli / Bodega	Pizza Place	Italian Restaurant	Liquor Store

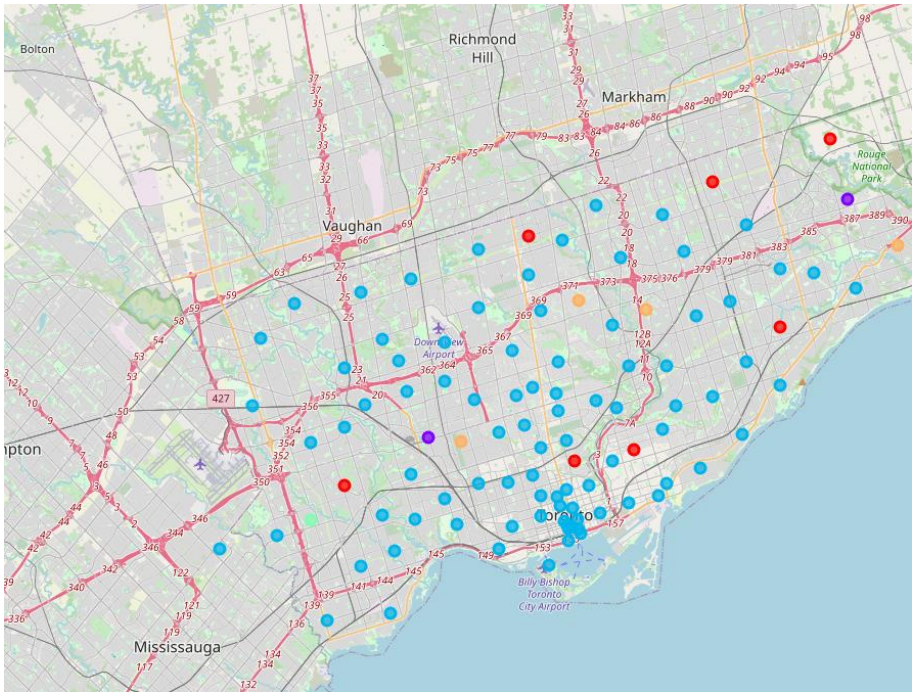
We can do the same for Toronto, for each neighborhood in Toronto, we listed the top venue categories.

The topx venue category for each neighborhood, can be used as “input features” for K-Means clustering.

In this experiment, we use 5 clusters



And the same can be done for Toronto



And if we can look into the top venue statistics of New York and Toronto:

```
newyork_merged['Cluster Labels'].value_counts()
```

```
2    245
1     48
0     10
3      2
4      1
Name: Cluster Labels, dtype: int64
```

```
toronto_merged['Cluster Labels'].value_counts()
```

```
1    87
0    12
2     2
4     1
3     1
Name: Cluster Labels, dtype: int64
```

We can see, both New York and Toronto have a “dominant category” (but as we did K-Mean separately for each city, so the categories are not the same)

And if we check the most popular categories in biggest cluster in New York, we can say, it i s about Pizza Place, Italian Restaurant, Coffee Shop. . .

```
df_cluster = newyork_merged[(newyork_merged['Cluster Labels'] ==2)]
df_cluster['1st Most Common Venue'].value_counts()
```

```
Pizza Place                32
Italian Restaurant         25
Coffee Shop                15
Chinese Restaurant         14
Pharmacy                   10
                           ..
Latin American Restaurant   1
Cosmetics Shop              1
Spanish Restaurant          1
Baseball Field              1
Other Repair Shop           1
Name: 1st Most Common Venue, Length: 67, dtype: int64
```

And if we check the most popular categories

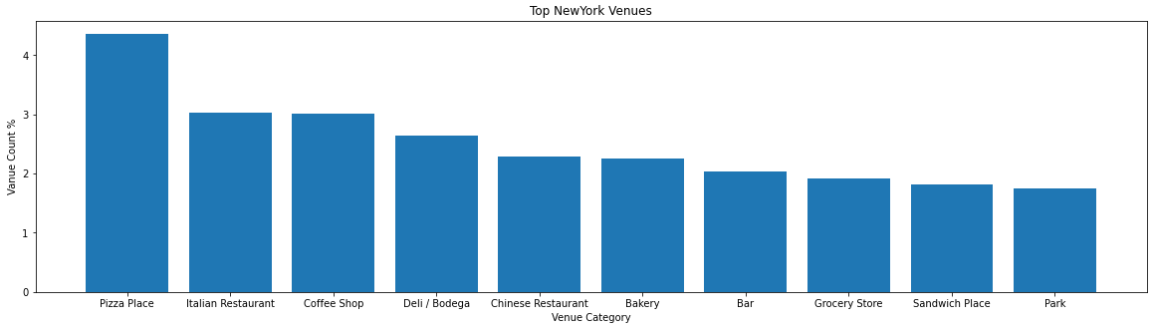
And if we check the most popular categories in biggest cluster in Toronto, we can say, it is about Coffee Shop, Grocery Store, Café. . .

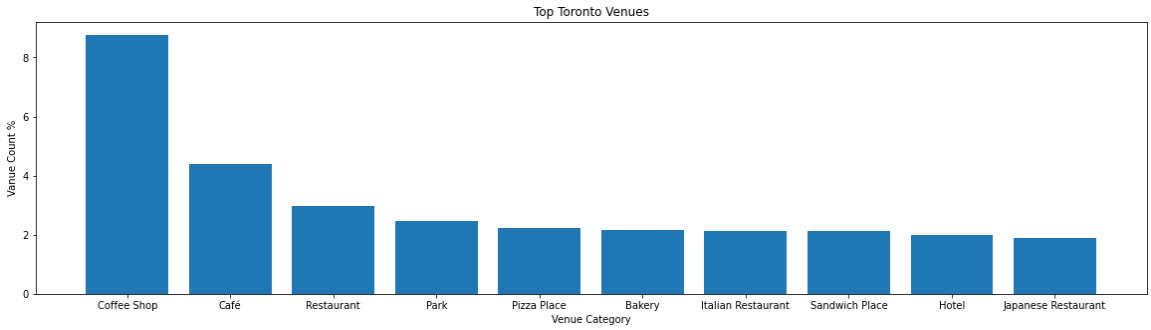
```
df_cluster = toronto_merged[(toronto_merged['Cluster Labels'] ==2)]
df_cluster['1st Most Common Venue'].value_counts()
```

```
Coffee Shop              18
Grocery Store             7
Café                      7
Park                      6
Clothing Store            3
Pizza Place               3
Trail                     3
Gym                       3
```

We can continue this analysis on different clusters.

And if we can also look into the overall statistics by comparing the top venue quantity (to make apple to apple comparison, we convert the quantity to %)





We can definitely say: In New York, Pizza Place is much more popular than in Toronto, but Toronto has more Coffee Shops and Parks.

We can help put client ***ABC*** company to look into more details of similarity and differences between New York and Toronto. Here we just brief some possibilities.

Business Suggestion part:

We can combine tables for Toronto of “Post Code - Neighborhood” and “Post Code – Latitude - Longitude”

And I also use the Statistics Canada Census data breaking down into post codes. There are 2257 lines of data for each post code which brings huge potential for various useful business analysis. But here we only focus on “Average Number of Rooms” and “Average Dwelling Value”.

So, the key information is “Post Code – Average Dwelling Value – Average Number of Rooms”

The information will help ***ABC*** company to understand how big houses/condos they should build and for what price.

After removing some dummy post codes reserved for governments, Canada Post, there are 96 post codes to analysis.

By combining the data together, we can say:

- Toronto average house price is $806,000

- Toronto average dwelling size has 5.16 room

(Note: this is simple average among all post codes, not weighted average, but may very close to reflect the overall picture)

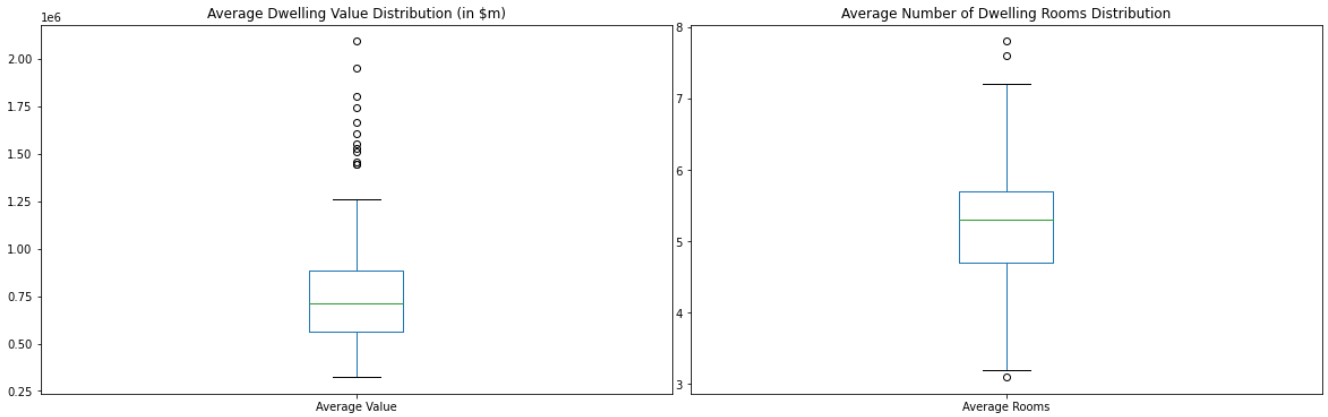
```
tr_data['Average Value'].describe()
```

```
]: count    9.600000e+01
   mean     8.069142e+05
   std      3.701061e+05
   min      3.245700e+05
   25%      5.597500e+05
   50%      7.121075e+05
   75%      8.859228e+05
   max      2.090328e+06
   Name: Average Value, dtype: float64
```

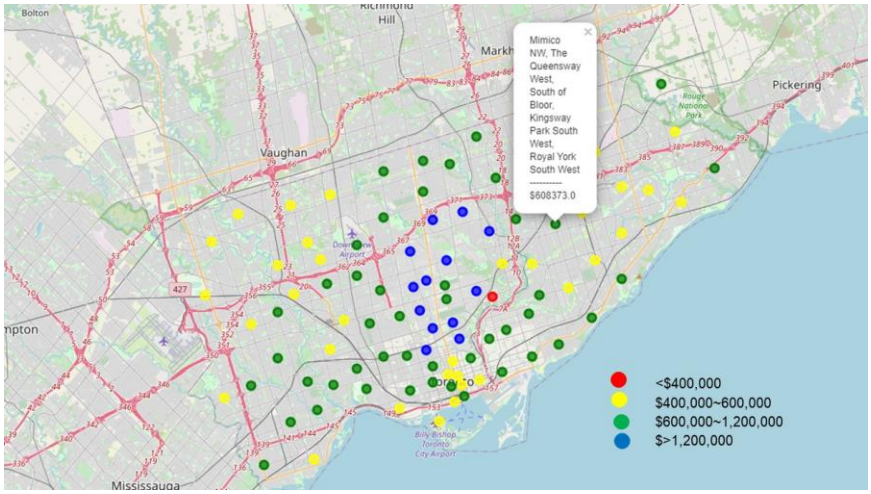
```
[33]: tr_data['Average Rooms'].describe()
```

```
Out[33]: count    96.000000
         mean      5.167708
         std       1.001525
         min       3.100000
         25%       4.700000
         50%       5.300000
         75%       5.700000
         max       7.800000
         Name: Average Rooms, dtype: float64
```

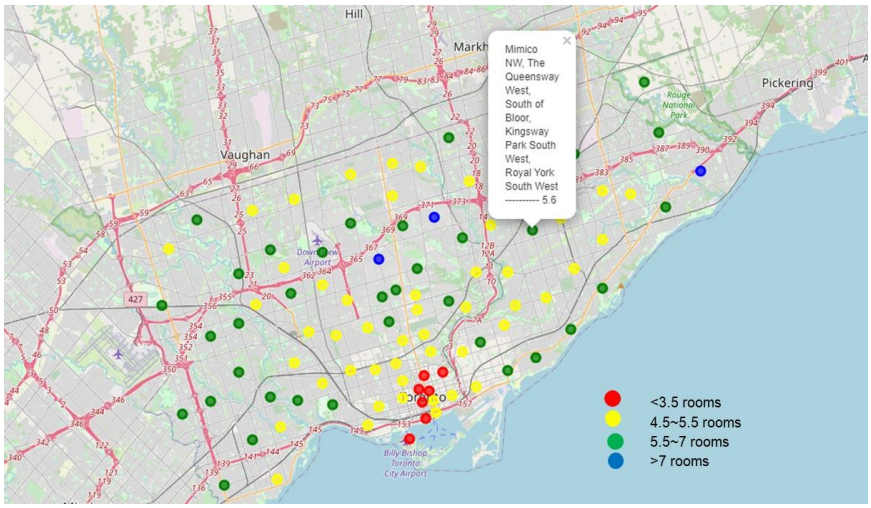
And if we put all the data into box plot, we will get this



And we broke the neighborhoods by Average Dwelling Value and marked them in different colors



And we broke the neighborhoods by Average Number of Rooms and marked them in different colors.



By hovering the mouse cursor onto each neighborhood, the specific information will popup.

5. Observations and Discussions

In general, I would give the information to ***ABC*** company:

- Toronto is a city with similarity, but with different combinations of venues in neighborhood (e.g. more coffee shops, parks)
- We should focus on the high value areas with smaller number of rooms per dwelling

- There are many other possibilities we can do with the venue data (different queries and comments) and Statistic Canada data (2247 lines per post code).
- Most importantly, this report is to help ABC company to understand the power of data science and what potential data scientist can unlock for them to make business in Toronto.

- There are many more data sources, e.g. paid subscription to Statistics Canada for many more data combination.

Limitation for this analysis:

- Ideally, it is better to use map.choropleth to use the depth of color to reflect the “Average Dwelling Value” and “Number of Rooms”. But there is a limitation to manipulate the Geojson file for Toronto. The version I found in internet, its breakdown does not match post code well and there was some problem to fetch the streambody object. This can be done later
- The Statistic Canada data is 2016, maybe too old to reflect the current situation

6. Conclusion

This report and analysis give the brief to ***ABC*** company about Toronto neighborhood, what kind of houses they can build in Toronto for what price.

It is helpful to ABC company to setup their business in Toronto. And when business setup and operation goes on, we can help to discover and unlock more possibilities and potentials.