**Battle Of Healthcare Accessibility**

**between**

**Toronto and New York City**

**Michael Tang**

1. **INTRODUCTION:**
   1. **BACKGROUND INFORMATION:**

Each year, both the U.S and Canada invest a lot of money in their healthcare systems with the goal to achieve better patient health outcomes. Despite the cultural similarity, there is a distinct difference in the healthcare system between the two countries. Concretely, Canada advocates a single-payer system where the government is the only entity paying for the healthcare coverage, whereas in the U.S, most healthcare facilities are private and operate on a multi-payer system. Even though they each offers its own pros and cons, many researches have suggested that Canada has a more superior healthcare system in terms of two common health outcome measures, the infant mortality rate and life expectancy. Additionally, Canada spends 10.4% of its GDP on healthcare in comparison to the 16% spent by the U.S. It seems conspicuous that the Canadian system is doing more for less.

* 1. **PROBLEM AND INTEREST**

Based on the aforementioned information, it would be of a special interest to the U.S politicians to determine if the single-payer system is a potential alternative to adopt. Ultimately, it can also help both parties determine the problems that exist in the opposite party. In the project, I will be looking at the healthcare accessibility of healthcare facilities in both countries, which could indirectly reflect the performance of a country’s healthcare system. More specifically, healthcare accessibility can be roughly divided into two factors – wait time and affordability. A longer wait time and a less affordability can both result in unmet medical needs, which are indicators of potential deficiency in healthcare resources. Finally, as a Torontonian, I will analyze the distribution of healthcare resources in Toronto. I hope the result can assist physicians in determining the best type of healthcare facilities to open at given location.

1. **DATA ACQUISITION & CLEANING**

In this section, I will discuss the data sources and data wrangling for this analysis. First of all, I will narrow down the targets from a national level to a municipal level - Toronto and Boston. Precisely, while the healthcare system in these two cities might not best represent their own countries (Intuitively, Toronto and Boston would most likely have more healthcare practitioners than Winnipeg and Phoenix, thus can offer greater accessibilities), both cities are similar in many other aspects. For instance, both Toronto and Boston have a large yet diverse population, and are the educational capital of their respective countries. Therefore, Toronto and Boston are selected for the purpose of this comparison.

* 1. **DATA SOURCES**

The first data required for this project are the demographics information for Toronto and Boston. For the ease of comparison and data acquisition, I thought about dividing each city into smaller and more manageable districts. This has resulted in data collection for the 25 wards in Toronto and the 26 neighborhoods in Boston. The following specific information are collected for each ward/neighborhood:

* Geological information (Latitude & Longitude, Area)
* Demographics information (Number of households & Median household income)

The sources for which the data are acquired from are listed here:

* 24 Toronto wards:
  + [Demographics and area data](https://open.toronto.ca/dataset/ward-profiles-2018-25-ward-model/) (Downloaded in .excel format)
  + [Latitude and longitude](https://open.toronto.ca/dataset/city-wards/) (Downloaded in .csv format)
* 26 Boston neighborhoods:
  + [Demographics data](https://data.boston.gov/dataset/neighborhood-demographics) (Downloaded in .excel format)
  + [Area data](https://boston.opendatasoft.com/explore/dataset/boston-neighborhoods/information/?dataChart=eyJxdWVyaWVzIjpbeyJjb25maWciOnsiZGF0YXNldCI6ImJvc3Rvbi1uZWlnaGJvcmhvb2RzIiwib3B0aW9ucyI6e319LCJjaGFydHMiOlt7ImFsaWduTW9udGgiOnRydWUsInR5cGUiOiJjb2x1bW4iLCJmdW5jIjoiQVZHIiwieUF4aXMiOiJvYmplY3RpZCIsInNjaWVudGlmaWNEaXNwbGF5Ijp0cnVlLCJjb2xvciI6IiM2NmMyYTUifV0sInhBeGlzIjoib2JqZWN0aWQiLCJtYXhwb2ludHMiOjUwLCJzb3J0IjoiIn1dLCJ0aW1lc2NhbGUiOiIiLCJkaXNwbGF5TGVnZW5kIjp0cnVlLCJhbGlnbk1vbnRoIjp0cnVlfQ%3D%3D&sort=-neighborho) (Downloaded in .json format)
  + [Latitude and longitude](https://data.boston.gov/dataset/boston-neighborhoods) (Downloaded in .csv format)

Healthcare affordability can be derived from the average healthcare expenditure. The healthcare expenditure is a little bit tricky to find since there was no readily available data online. In the case of Toronto, I am using $12935 , which is the average healthcare paid by a Canadian family according to [local news](https://globalnews.ca/news/4364344/cost-health-care-canadian-families/) (For the purpose of this project, a family is equivalent to a household). For Boston, each family is paying over $21085 according to the [Cost Trends Report](https://www.mass.gov/files/documents/2018/03/28/Cost%20Trends%20Report%202017.pdf) published by the Massachusetts government.

Last but not the least, I will be using the location data provided by Foursquare. Concretely, Foursquare provides a list of venue categories with their unique category ID. And in this case, I will be using the category ’Medical Center’ because I am interested in locating all healthcare related facilities in a specified neighborhood. Furthermore, the ‘Medical Center’ category encapsulates several sub-categories, such as ‘Dentist’s Office’, ‘Emergency Room’, ‘Mental Health Office’, ‘Rehab Center’, etc. (a complete list can be found on [here](https://developer.foursquare.com/docs/resources/categories)). This information will be helpful in determining the ideal neighborhood for opening different types of healthcare facilities.

* 1. **DATA CLEANING**

After downloading and reading in the data. There are 3 dataframes for Toronto and 2 dataframes for Boston. The first step is to remove all the unwanted rows and columns. For example, Figure 1 below shows the data frame that contains the location information (latitude and longitude) of each Toronto ward:

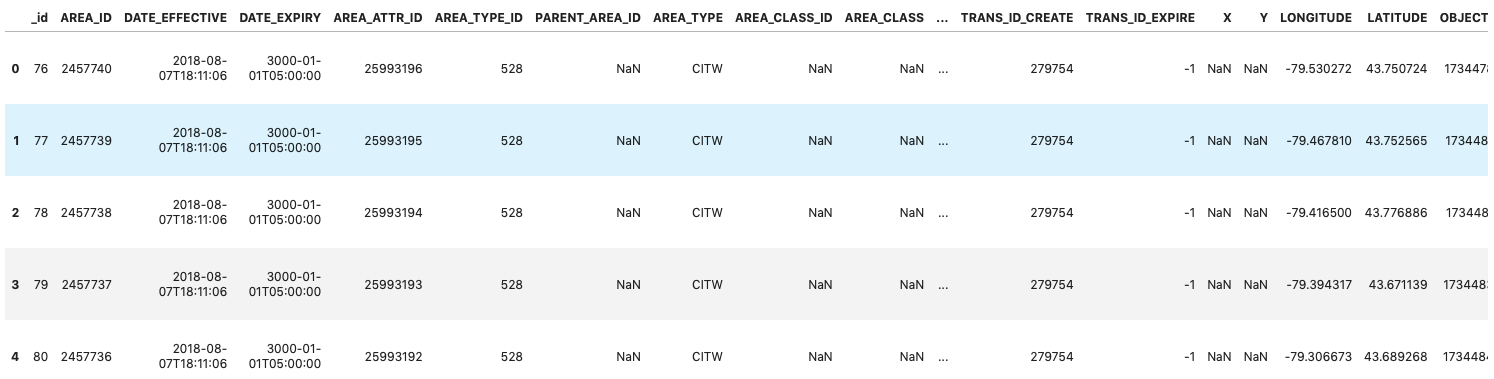
To clean this dataframe, several columns need to be dropped and renamed. The cleaned dataframe looks like the one below:

Figure 1. Toronto Ward Location Dataframe

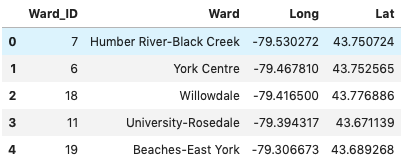
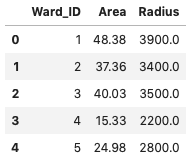
In this case, the columns ‘Ward\_ID’ and ‘Ward’ are used as the reference keys for performing dataframe merging later on.

Figure 2. Cleaned Toronto Ward Location Dataframe

As I have mentioned in the previous section, the area information has also been collected. This is used for calculating the radius of each ward, assuming that the shape of each ward is circular. The same assumption is also made for each Boston neighborhood. The motivation behind this assumption is to increase the accuracy of Foursquare API calls. While it is not possible to define the actual boundary of each ward/neighborhood for performing venue searching, the radius for each search area could be set a variable to account for different ward/neighborhood areas. Therefore, the area information is then used to produce a ‘Radius’ column as shown below:

Figure 3. Toronto Ward Area Dataframe (With Radius)

Additionally, unit conversion is also performed since Boston data are in customary units. For example, the area for each Boston neighborhood is recorded in Square-Miles and is thus converted to Square-Kilometers for consistency.

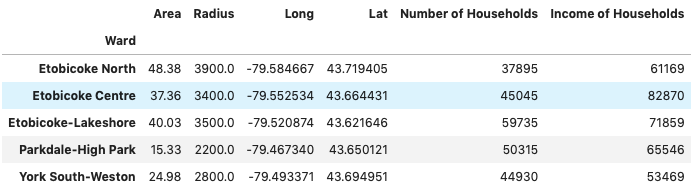
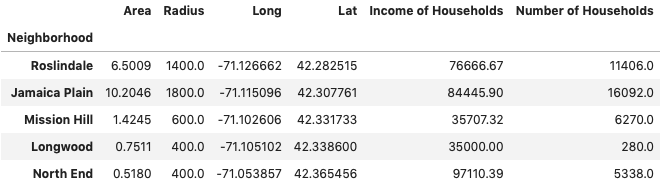
Finally, dataframe are merged together to produce the final datasets for analysis, as shown in Figure 4 for Toronto and Figure 5 to Boston:

Figure 4. Boston Neighborhood Dataframe

Figure 5. Toronto Ward Dataframe

* 1. **VENUE CATEGORY SELECTION**

In order to complete the objective of this project, which is to determine the wait time for each healthcare facility. This section focuses on Foursquare categories that I used. As mentioned before, Foursquare provides an extensive list of categories. The ‘Medical Center’ categories consists of a total of 16 different subcategories. For the purpose of this project, I define a healthcare facility as a place that provides direct healthcare treatments or intervention to a person, or services that are generally included in a healthcare plan (e.g. OHIP) or insurance. Therefore, I excluded ‘Alternative Healer’, ‘Nutritionist’, ‘Veterinarian’ from the list of healthcare facilities used for analysis.

To perform an API call on Foursquare, the unit Category ID of each healthcare facility needs to be provided. I have therefore constructed a dictionary for storing these information:

Figure 6. Healthcare Facility Dictionary

1. **METHODOLOGY & RESULTS**

To understand the difference in healthcare system between the U.S and Canada, most researchers would dive into mainstream measurement such infant mortality rate or life expectancy. However, besides the intricate diagnostic procedures and treatments each country is able to provide, these measurements have large correlation with other factors such as smoking and accidents, as suggested in the [study](https://www.nber.org/papers/w13429) conducted by June O'Neill and Dave M. O'Neill. The other side of healthcare is accessibility, which has a huge impact on medical needs. As I have mentioned in the introduction section. Healthcare accessibility can be translated to wait time and affordability. Furthermore, because each country’s healthcare system operates on a different model, it is expected that Toronto would have a longer wait time while Boston would be less affordable. The first two parts of this section will validate this hypotheses. The last part of this section focuses on the method used to determine the ideal type of facility to open at a given location in Toronto.

* 1. **PATIENT WAIT TIME**

A prolonged wait time in any kind of healthcare facility not only results in patients dissatisfaction, but can also result in severe outcomes in the extreme cases. However, there are no reliable resources on this information, which suggests an alternative approach to assess this problem. Namely, wait time largely depends on the limited healthcare resources. A smaller number of hospitals means that each hospital is ‘assigned’ more patients and would therefore cause a longer wait time. Of course, there are other complications such as facility rating, cost, readmission that would affect the number of visiting patients and the wait time. But for the simplicity of this project, we assume that distribution of patient at each type of healthcare facility is equal. In this light, the wait time can roughly translate to the following two measurements:

* The number of households ‘assigned’ to each type of facility
* The area ‘covered’ by each type of facility

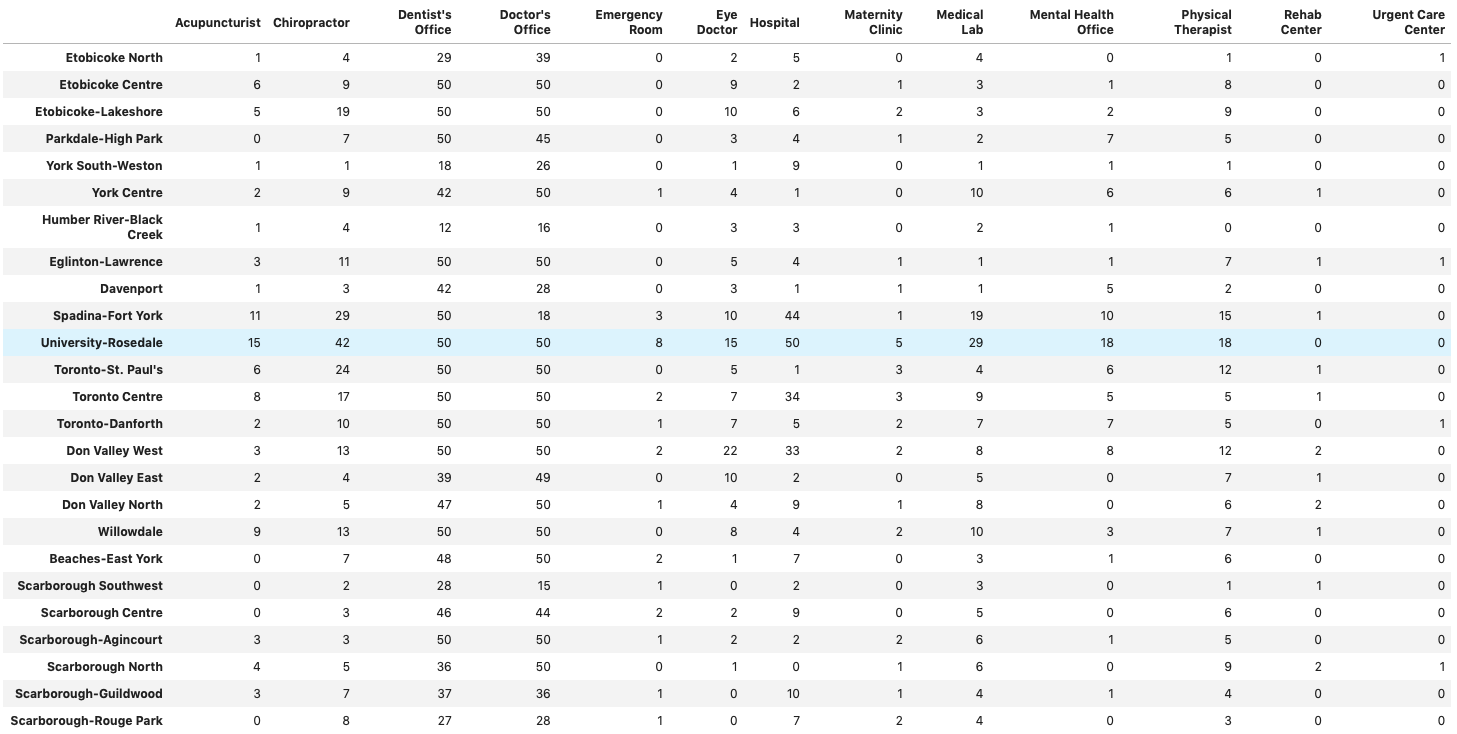
To derive these two items, I first used Foursquare to obtain the count of each type of healthcare facilities in each Toronto ward, as shown in the figure below:

Figure 7. Toronto Ward Healthcare Facility Count

There are 3 things worth noting here. First, the numbers are retrieved by searching for venues with specified category ID within a specified circular area. Therefore, certain wards might produce overlapping searching area, which would result in duplicated count. To minimize this error, I decreased the size of the circular searching area by rounding the radius of each ward/neighborhood down to the nearest hundreds. The second point is about the healthcare facility type. Foursquare platform allows platform users or business owner’s to categorize their own facility. Therefore, some facilities could get mis-categorized (e.g. a dentist’s office gets categorized as a doctor’s office). However, we can reasonably make an assumption that the mis-categorization rate is equal across all different types of healthcare facilities. The last point is also related to Foursquare API. Foursquare limits the maximum number of returns for each call to 50. This means that a ward/neighborhood might have more than 50 of a particular type of healthcare facility but only shows 50 in the dataframe above. However, this mostly only affects ‘Doctor’s office’ and ‘Dentist’s office’, which are intuitively the two most common facilities in both cities. Thus, this offset should not significantly impact the final observation.

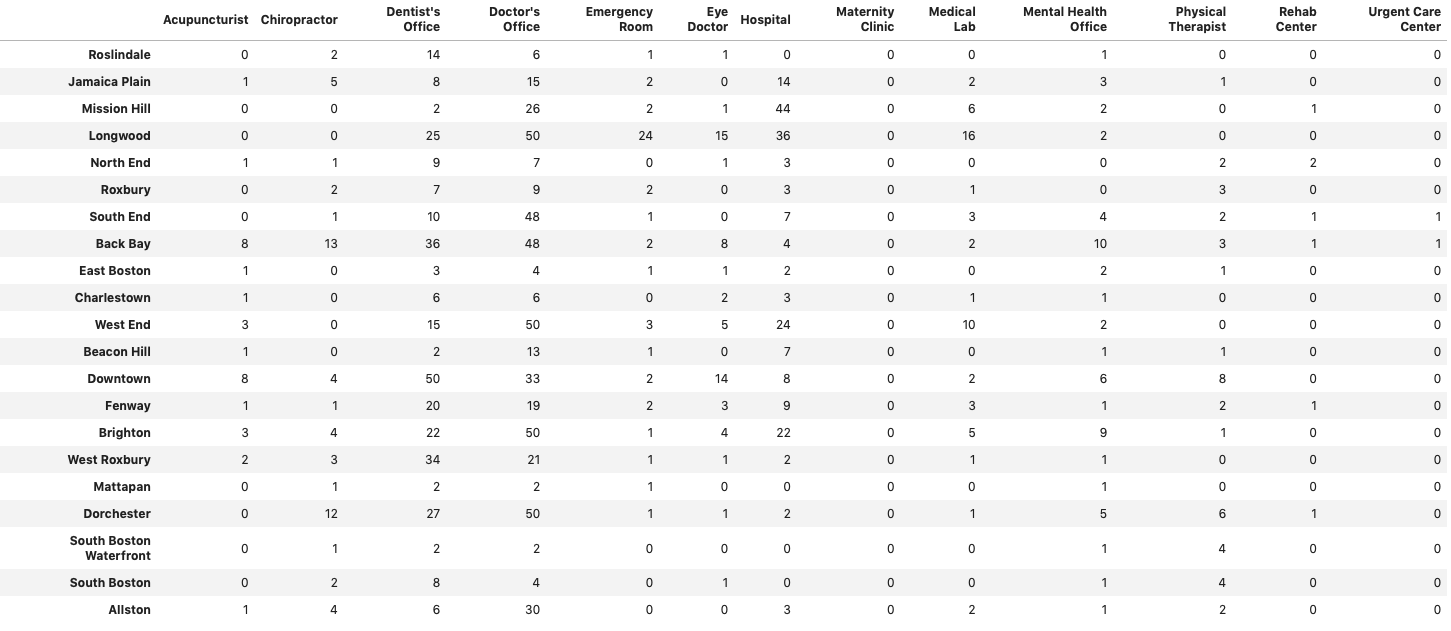
The dataframe shown in Figure 7 and Figure 8 are summed across each row to produce an approximation of the count of each type of healthcare facility in each city. Then, this information is divided from the total number of households and the total area of each city, which equal to the two measurements mentioned above. Note that count of ‘Maternity Clinic’ for Boston is 0 across all of its neighborhoods. This might be caused by the mis-categorization error I mentioned earlier. This means that the two measurements we just calculated for ‘Maternity Clinic’ would be infinite. In this case, I have marked them as NaN (Not a Number):

Figure 8. Boston Neighborhood Healthcare Facility Count

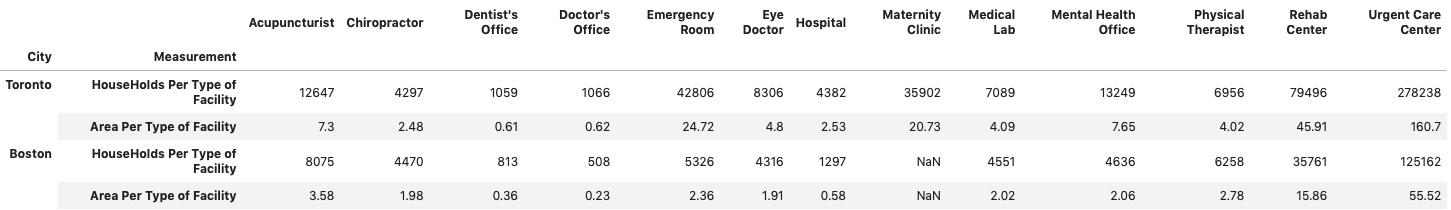
It can be noticed that both measurements are higher in Toronto than they are in Boston, except for ‘Chiropractor’, which has 4297 Households in Toronto as opposed to the 4470 households in Boston.

Figure 9. Households and Area per Type of Facility for Toronto and Boston

I then plotted bar graph for better comparison and visualization:

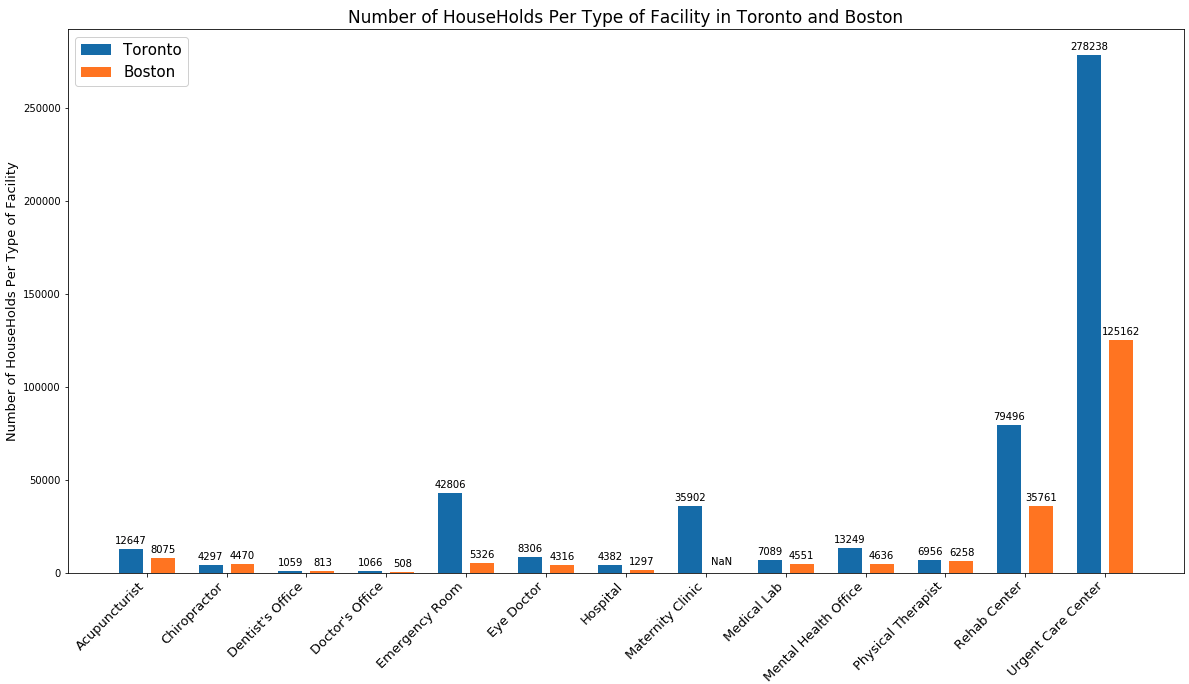
Figure 10. can be interpreted as the number of households that each type of healthcare facility is assigned with. It is pretty obvious that almost every type of healthcare facility in Toronto is more crowded than they are in Boston. More specifically, the number of households assigned to each ‘Emergency Room’ is almost 7 times more higher in Toronto. Even with the less risky ‘Doctor’s office’, Toronto has twice as many households than Boston. This could potentially translate to twice the wait time for a regular family doctor visit. The same trend is observed in Figure 11., where the area covered by each type of healthcare facility is also much higher in Toronto. This suggests a relatively sparse distribution of healthcare sources in Toronto, which could also impact the wait time.

Figure 10. Number of Households per Type of Facility in Toronto and Boston

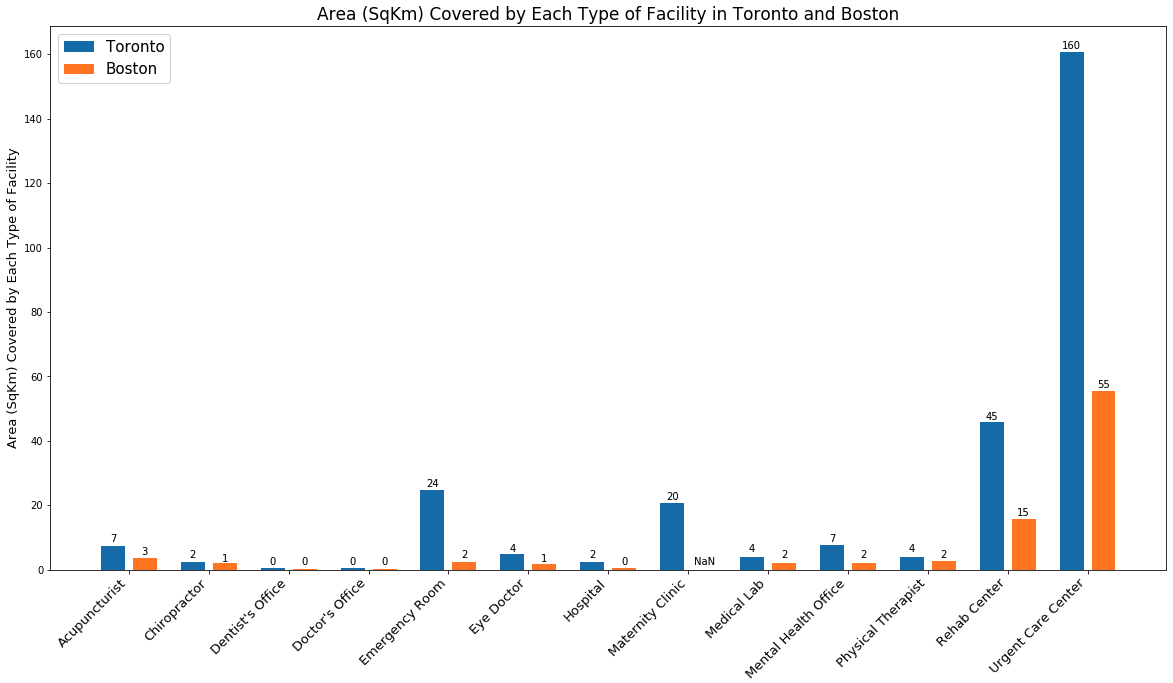


Figure 11. Area Covered by Each Type of Facility in Toronto and Boston

* 1. **AFFORDABILITY**

Affordability is more straightforward in this case. Due to the lack of data, I have used the national and provincial healthcare expenditure for Toronto and Boston respectively. These information are then combined with median household income to produce the following simple dataframe:

Figure 12. Expenditure Distribution in Toronto and Boston

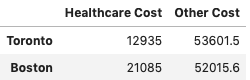
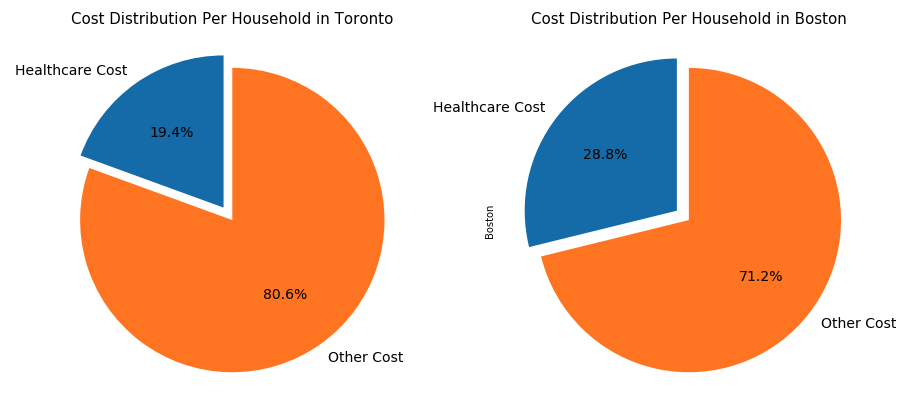
It can be seen that the healthcare cost in Boston in almost 1.5 times the healthcare cost in Toronto, while the cost spent on other categories are around the same. Based on this information, the following pie chart is produced to help with visualization:

Figure 13. Cost Distribution in Toronto and Boston - Pie Chart

As expected, the healthcare affordability behaves the complete opposite of wait time. Concretely, healthcare cost takes up 28.8% of the overall household income in Boston, where it only uses 19.4% in Toronto.

* 1. **K-MEANS CLUSTERING**

For this part of the project, I am using k-means clustering to cluster the neighborhoods in Toronto, using the distribution of healthcare facilities. The goal is to determine if there are any similarities within each cluster or differences among the clusters. This can in turn determine if a particular type of healthcare facility is saturated within a cluster of wards. And of course, saturation could mean two things, competitiveness or higher demands.

First of all, I calculated the percentage of each type of healthcare facility in each Toronto Ward. This information is then transformed into a map, where each ward is marked and labeled with the most common type of healthcare facilities:

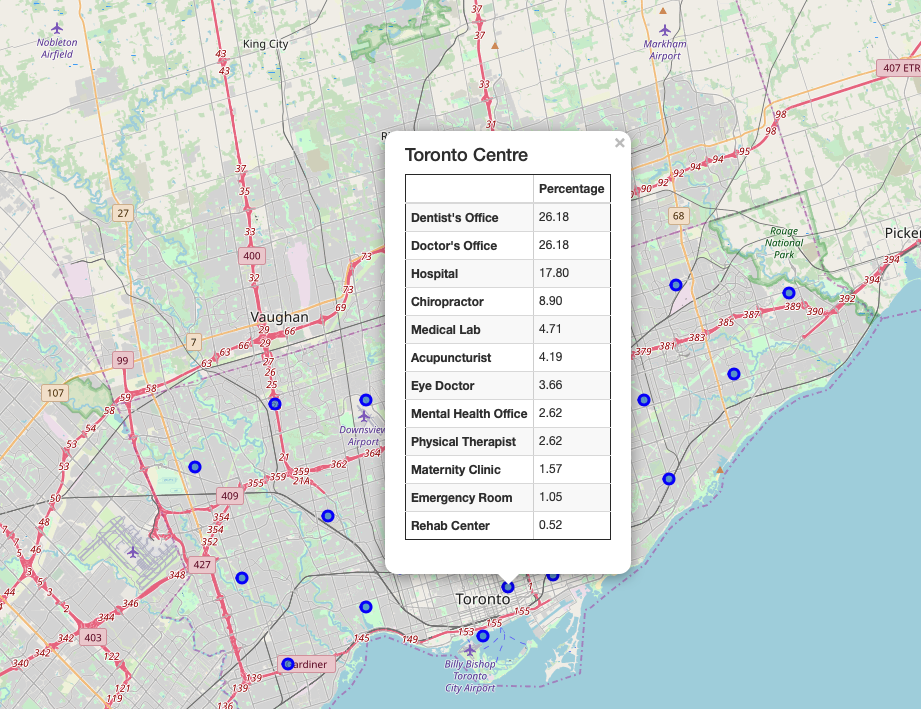
We can see that Toronto Centre contains almost every type of healthcare facilities. And the two most common types are ‘Dentist’s Office’ and ‘Doctor’s Office.

Figure 14. Toronto Map – Healthcare Facility Distribution – Toronto Centre

The dataframe that contains facility count information for each Toronto ward is normalized before clustering. To determine the number of clusters to be used, I looped through different values of clusters (1 to 11) and used the SSE (Sum of Square error) to determine the best K values. Normally, when SSE is plotted against the number of clusters, a curve with an elbow point should be observed. This point indicates the K values that results in the sharpest decrease in SSE and is used for actual use. However, as shown in Figure 15. There is no obvious elbow point to choose from. This suggests that there might not that many intrinsic differences between each cluster.

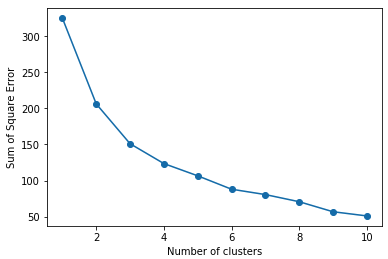
In this case, I processed with a k value of 6 as it has a relatively shaper change. After determining the k value, I performed k-means clustering and yielded the results shown in figure 16, where the color of each marker represents its cluster.

Figure 15. SSE vs. Number of Clusters

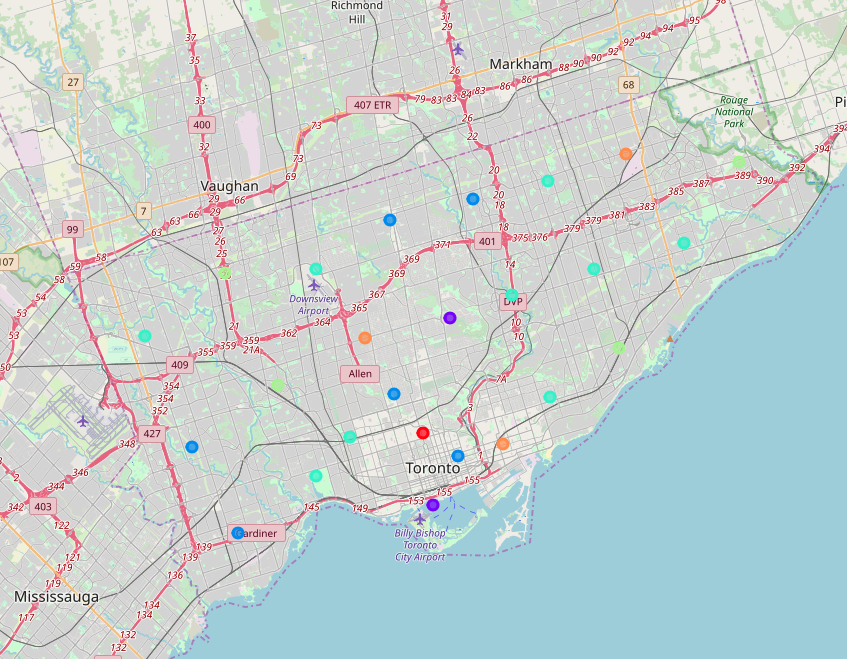
The detail of each clusters are shown in the figures below:

Figure 16. Map of Clustered Toronto Wards



Figure 17. Cluster 1

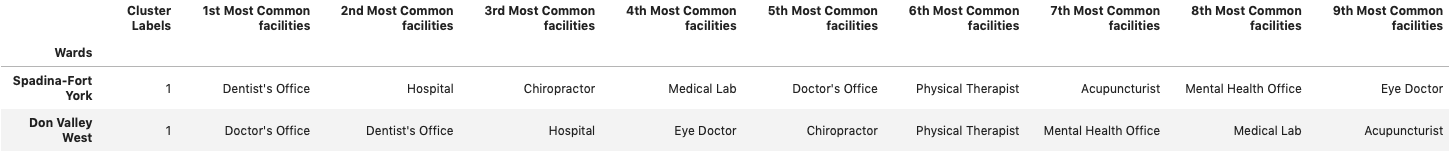


Figure 18. Cluster 2

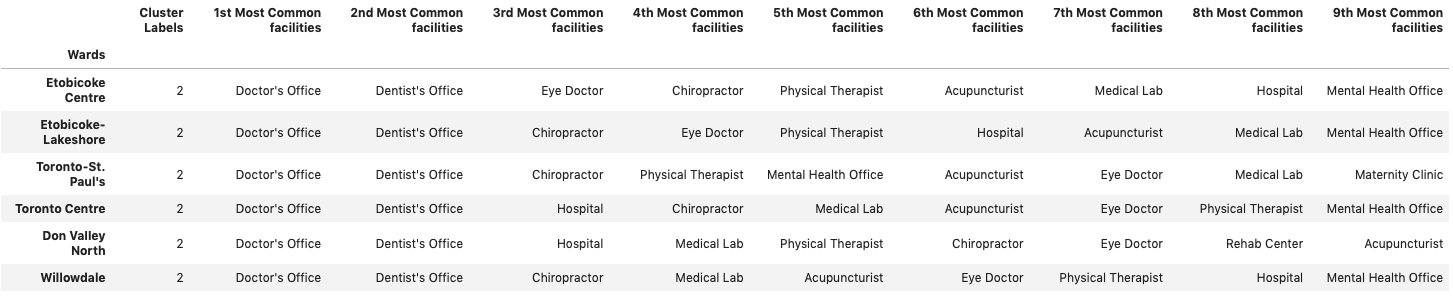


Figure 19. Cluster 3



Figure 20. Cluster 4

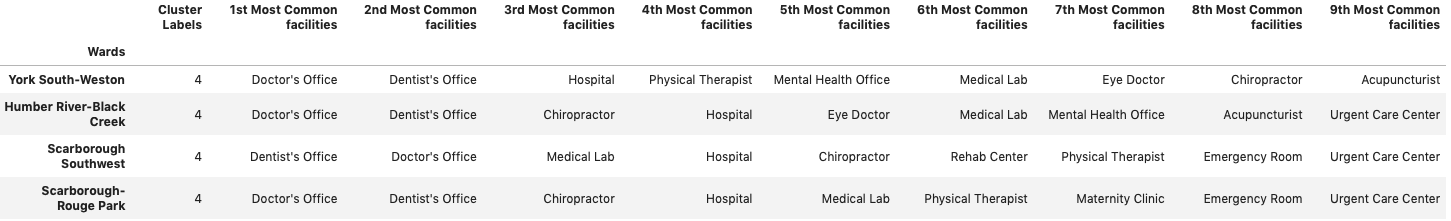


Figure 21. Cluster 5

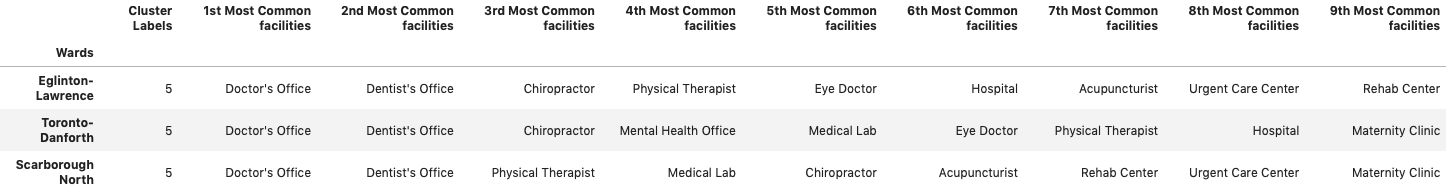


Figure 22. Cluster 6

The figures above (Figure 17. – Figure 22.) show the wards within each cluster as well the top 9 most common types of healthcare facilities. It can be seen that Cluster 1 only has one ward, this is because its most common type of facility is ‘Hospital’. And as expected, ‘Dentist’s Office’ and ‘Doctor’s Office’ have again appeared in all other wards. Thus their contribution to each cluster might be very little.

From a business perspective, this person should try to avoid competition while targeting a ward with high demands, he or she would probably want to look at the top 3 most common types of facility among 4th to the 6th most common type of facility in each ward of a cluster. This person would also want to only look at the type of facilities that are generally privately owned and operated, which excludes: Emergency room, hospital, urgent care center. This can be summarized in the table below:

|  |  |  |
| --- | --- | --- |
| Cluster 1 | 1st | Medical Lab |
| 2nd | Physical therapist |
| 3rd | Chiropractor |
| Cluster 2 | 1st | Physical therapist |
| 2nd | Medical Lab, Doctor’s office, Eye doctor, Chiropractor |
| Cluster 3 | 1st | Physical therapist |
| 2nd | Acupuncturist |
| 3rd | Chiropractor |
| Cluster 4 | 1st | Chiropractor |
| 2nd | Physical therapist |
| 3rd | Medical lab |
| Cluster 5 | 1st | Medical lab |
| 2nd | Physical therapist |
| 3rd | Eye doctor, Rehab center, Mental health office, chiropractor |
| Cluster 6 | 1st | Eye doctor |
| 2nd | Physical therapist, Mental health office, Medical lab, Chiropractor, Acupuncturist |

This table depicts the type of facilities that a person should open in the wards contains within each cluster. Of course there are many other factors to consider. However, for the simplicity of this project and illustration of clustering, the decision is solely made based on the count of each type of facilities. While the driver is to avoid saturation as well as capturing the high demands from the crowd.

1. **Conclusion**

In this project, I analyzed the difference in healthcare systems between the U.S and Canada. The study was performed on a municipal scale between Boston and Toronto. The data collected for this analysis included local demographic data and location data. Additionally, Foursquare API was employed to help determine the distribution of different types of healthcare facilities in both cities. The result matched my initial prediction, which stated that the wait time in Toronto is much longer, while the affordability in Boston is lower. This suggested that there really was no winner in determining which country offers better accessibility with its own healthcare system. There is always a tradeoff, pros and cons within each model. A further study needs to be conducted to take other measurements, such as other treatment efficacy, readmission rate into account.

For the final part, I used unsupervised machine learning technique to cluster Toronto wards into 6 different clusters. However, it can be seen that the distribution of different types of healthcare facilities is not the best features to perform k-means clustering on. This might be caused by the lack of similarities within each cluster and differences among cluster. Further study should include other factors such as age distribution, income distribution, which potentially impact the business aspects of a healthcare facility.