Coursera Capstone Report

IBM Data Scientist Professional

MAY 21, **2021**

Authored by: Reza Sadeghi Jafari

Coursera Capstone Report

IBM Data Scientist professional

This Report developed based on Coursera capstone project which represent the business problem, data gathering, data analysis, ML solutions, the result, discussion and the conclusion.

Criteria
THIS CAPSTONE PROJECT WILL BE GRADED BY YOUR PEERS. THIS CAPSTONE PROJECT IS WORTH 70% OF YOUR TOTAL GRADE. THE PROJECT WILL BE COMPLETED OVER THE COURSE OF 2 WEEKS. WEEK 1 SUBMISSIONS WILL BE WORTH 30% WHEREAS WEEK 2 SUBMISSIONS WILL BE WORTH 40% OF YOUR TOTAL GRADE.
FOR THIS WEEK, YOU WILL REQUIRED TO SUBMIT THE FOLLOWING:
A DESCRIPTION OF THE PROBLEM AND A DISCUSSION OF THE BACKGROUND. (15 MARKS) **A DESCRIPTION OF THE DATA AND HOW IT WILL BE USED TO SOLVE THE PROBLEM. (15 MARKS)** **FOR THE SECOND WEEK, THE FINAL DELIVERABLES OF THE PROJECT WILL BE:**
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)

Background:

Health and security facilities in each part of city is one of the most important items when residents want to choose a location to start businesses or location home, so the government team mostly check the distribution of these venues all over the city to be ensure people had access to facilities based on and predefined standards.

Business Problem:

Most cities have a committee for planning the city development plan, they need to be ensured all the essential needs of the city's residents are well available in all parts of the city. Concentrating facilities in a part of the city will cause complex problems for urban management such as unwanted traffic on morning and evening, caricatural growing up of city and their facilities, there is some unhappy citizens and so much. Therefore, in municipalities, a sector was generally considered as urban development, and this team needs to control the distribution of facilities in different urban sectors. Every year they need some statistics data to compare different neighborhoods, when they cluster data to identify which cluster need to priories in next year development plan.

Interests:

In this project, the city development team who play the role, as a customer, has requested that the necessary facilities be explored as follows throughout the city of **Toronto**, and that different parts of the city be segmented accordingly.

1- Medical centers
2- Police stations
3- Fire stations

So, Sectors with fewer facilities, that will need more development, will be a priority in next year's development plans.

Data acquisition and cleaning:

To achieve this goal, we need data from different urban areas as follows

1- Urban areas
2- Facilities available in each area
3- the population on each area

To provide this information, we fetch urban areas from the information provided on the web - Wikipedia - as well as urban venues information from the foursquare.com based on the geographical location of the areas. At last, we need the population of Toronto on each area, after some google search, I found it at `https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/pd-pl/Table.cfm?Lang=Eng&T=1201&SR=1&S=22&O=A&RPP=9999&PR=0`

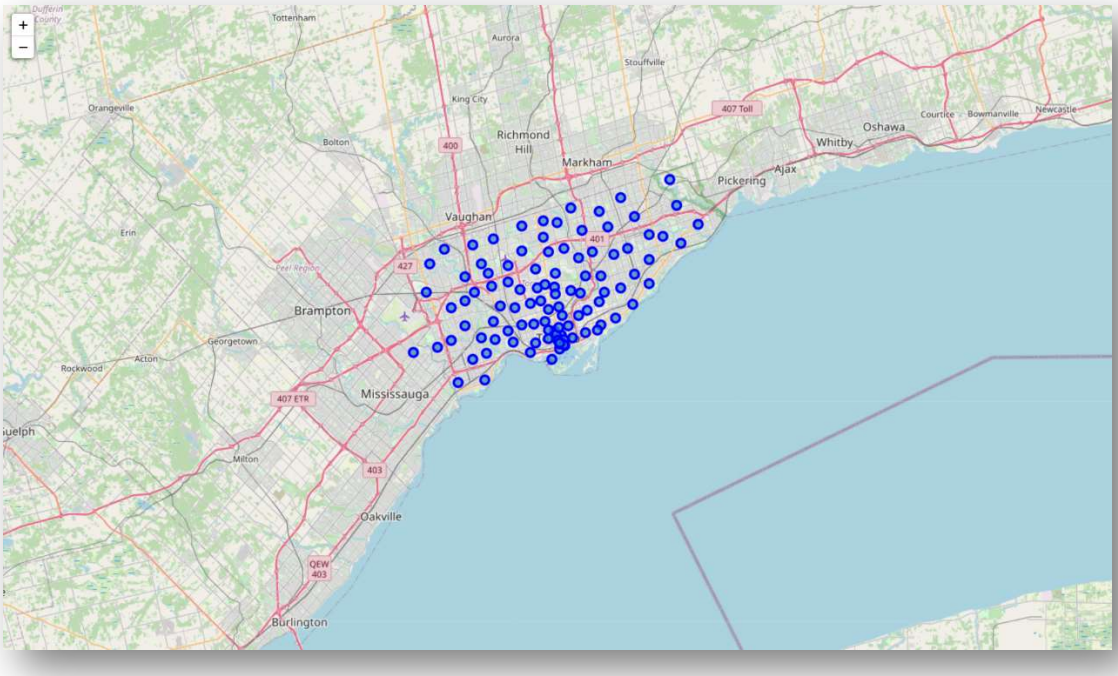
Data Cleaning:

The data of Urban areas which provided on Wikipedia page had some invalid values that should be removed,

We removed “Not assigned” borough from or data , then we need to group by and remove duplicate records.

Data Exploratory and Analysis:

Then, the geographical information of urban areas will be added to base data, you can see the neighbors in the following map:



The next step is, adding the information of the venues which specified in the program for each area. It's a bit difficult because I need to search (special query) for each neighbor from **Foursquere.com** web site with free account which is limited number of requests, so after fetching successfully I save it to a file for future use when needed to fulfill the dataset at development process.
we had some data which represent that each neighbor and count of nearby health, security venues so I use the search method which required a query, the query is venues categories. I need to add population data for each neighbor to calculate per capita, which is the key to identify which neighbor need more attention in development plan.

Solution:
By clustering the data, we can group similar neighbors, I divide it to 5 groups of cluster which represent 0, 0-25%, 25%-50%, 50%-75%, 75%-100% based on describing of dataframe,

```
result_grouped.describe()
```

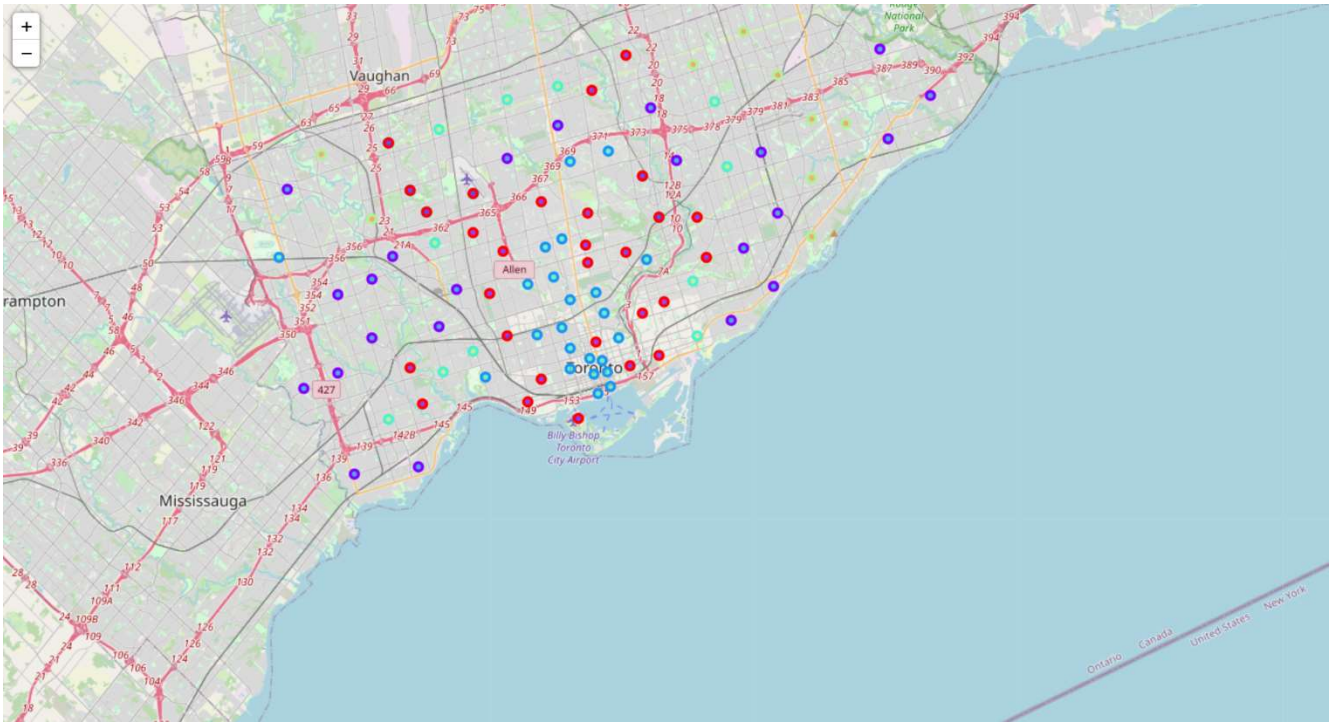
	Medical_Venue	Security_Venue
count	96.000000	96.000000
mean	0.000373	0.000713
std	0.000606	0.001284
min	0.000025	0.000025
25%	0.000098	0.000141
50%	0.000197	0.000312
75%	0.000361	0.000802
max	0.004489	0.009975

each neighbor will belong to a group number in each feature,

	Neighbourhood	Medical_Venue	Security_Venue	med_level	sec_level
0	Parkwoods	0.000087	0.000202	1	2

We use k-means model to cluster this 2 venues category based on med_level and sec_level features, to find simular neighbors and identify them. This help find the neighbors which are in same level of access to facilities.

Now the team can examine the distribution of facilities in each part of the city and prioritize the facilities that need further development based on each of the needs.



Conclusion:

By applying this project on these data, we can identify similar neighbors, and optimize city development plan for future and prevent from concentration. Clustering data help us to cluster urbans to different level. Each level represent