1. Introduction

1.1 Background

Toronto is the capital city of the Canadian province of Ontario. With a recorded population of 2,731,571 in 2016 it is the most populous city in Canada and the fourth most populous city in North America. Toronto is an international centre of business, finance, arts, and culture, and is recognized as one of the most multicultural and cosmopolitan cities in the world. Even life in Toronto sounds exciting but for adults with children it could be really challenging. Traffic, long distances between places is not comfortable for all people but for people with childrens it creates more difficulties.Therefore for people with children is really important if neighborhood will be comfortable for life with kids, they prefer that in their chosen neighborhood would be more schools, parks and playgrounds, easy access to hospitals and also that grocery stores would be near home.

1.2 Problem

Data that might contribute to determining which Toronto neighborhood is more suitable for life with childrens includes schools amount, parks and playground amount, hospitals amount and grocery store amount around neighborhoods. This project aims to cluster Toronto neighborhoods in clusters which could help people with children to choose neighborhoods for living.

1.3 Interest

People with childrens who is thinking to start life in Toronto or people with childrens which want to change Toronto neighborhood in similar Toronto neighborhood or to change Toronto neighborhood to different Toronto neighborhood to improve they life.

2. Data acquisition and cleaning

2.1 Data sources

Data of the Toronto neighborhood was used from Foursquare API. To get data from Foursquare several parameters should be described.

1. Latitude and Longitude. First of all, to get Toronto neighborhoods Wikipedia page https://en.wikipedia.org/_wiki/List_of_postal_codes_of_Canada:_ was scraped. Then Latitude and Longitude were added from 'GeoSpatial Dataset.csv' which was provided in IBM Applied Data Science Capstone course.

2. Date of data or version. In this study it was used 2021 - 07 - 21.
3. Search query. It was performed 5 Foursquare API calls, where search parameters were: 1 - school, 2 - park, 3 - playground, 4 - hospital, 5 - grocery stores.
4. Radius. A radius limitation was added: for schools, hospitals and grocery stores it was searched 2 km around the neighborhood latitude and longitude and for parks and playgrounds - 3 km.

2.2 Data cleaning and Feature selection

Data from Foursquare were combined in tables. In tables were a lot of not relevant information because Foursquare returns all venues in which name or category it found a search query word. For example: in received Park data it was included categories of areas like bus stops, parking, etc. Therefore data cleaning was performed:

- For schools data it was left just data which category is equal to ‘School’, ‘Elementary School’, ‘High School’, ‘Daycare’, ‘Middle School’. Because clustering will use just one feature of schools, schools data categories were renamed to ‘schools’.
- For Parks and Playground data it was left which category is equal to ‘Park’ or ‘Playground’. Because clustering will use just one feature of parks/playgrounds, Parks and Playground data categories were renamed to ‘parks/playgrounds’.
- For Hospital data it was left data which category is equal to ‘Hospital’, ‘Medical Center’, ‘Doctor's Office’, ‘Emergency Room’. Because clustering will use just one feature of hospitals, Hospitals data categories were renamed to ‘hospitals’.
- For Grocery store data it was left data which category is equal to 'Convenience Store', 'Miscellaneous Shop', 'Grocery Store', 'Supermarket', 'Food & Drink Shop', 'Market'. Because clustering will use just one feature of grocery stores, Grocery store data different categories were renamed to ‘grocery store’.

In studies it was decided to use 4 features for clustering - schools rate per neighborhood, parks/playgrounds rate per neighborhood, hospitals rate per neighborhood, grocery stores rate per neighborhood. Therefore features transformation where performed. After calculating category rates per neighborhood all information was combined in one table. Because some neighborhoods do not have parks/playgrounds or hospitals in the table was found 15 NaN values, these NaN values were filled in with 0.