Data Science Capstone - Report

Anna C

Introduction/Business Problem

The target audience of the problem addressed in this project is people interested in buying a house in Toronto, who would like to understand the factors that impact housing prices, and in particular with a particular focus on what kind of venues are available in the neighborhood and how these venues may affect the price of a house.

This problem is of interest because the housing market in Toronto is very competitive and is clearly a sellers market (even with the effects of the Covid-19 crisis). Houses with multiple offers often even sell for significantly above the asking price. Therefore, individuals with a limited budget who are looking to buy a house may have to make some compromises, for example with regards to the location and the venues available in close proximity to the house.

It will be important to first try to normalize the price of houses with regards to key characteristics such as the area or the number of bedrooms, bathrooms, and type of house. Then we will be able to better evaluate the impact of the neighborhood on the price of the house.

Data

The first dataset required is the list of Toronto Neighborhoods with their coordinates, created in the Capstone course assignment.

The second dataset is a list of houses in Toronto with their selling price and key characteristics, such as the address, the living area, number of bedrooms, and so on. Using the address, the coordinates will be added to the dataset and then each house will be assigned a cluster corresponding to the Toronto neighbourhood.
The dataset was handcrafted in another project on GitHub by searching through a real estate website's data and extracting it into a dataframe.
It will include:

- House sale price
- House listing price
- Address
- Number of bedrooms
- Number of bathrooms
- Number of parking spots
- Type (detached, semi-detached…)
- Latitude & longitude coordinates
- Neighborhood

The third dataset is the Foursquare venues data, which will be used to characterise the venues in close proximity to the house / its neighborhood. The objective will be to group the venues into a smaller number of categories, such as restaurants, bars, schools, parks, shops, grocery stores and so on that may add to the value of the house if in close proximity.
It will include:

- Venue name
- Venue type
- Venue category

The final dataset that may be necessary is a list of Toronto subway stations with their coordinates, as housing prices are also affected by proximity to public transport.