APPLIED DATA SCIENCE CAPSTONE

Project Title:
More than walls: Choosing a place to call home

Author:
Francis Frimpong

Table of Contents

1. Introduction

Over the past few decades, the world has seen major signs of growth and development. In general terms, people alive today are living better and at the same time, are undergoing a various forms evolution, from economic, to social, and cultural evolution. In spite of these monstrous changes, the basic needs of humans have remained more or less the same i.e. food, clothing, and housing. In this project, we will be talking about housing and the different housing choices.

For most people, a house is more than walls with a roof. Additional elements are taken into consideration to make a house a home. At the individual and family level, people's choice of a home will take into consideration factors such as the location, proximity to amenities such as schools, hospitals, market centers, and security. At the business level, investors will want properties in neighbourhoods they believe families will want to live now and in the future. Given the considerations above, both families and investors will look out for many specific characteristics when deciding on investing in housing and real estate. For a family or individual, they may lose on their property when 'unfavourable characteristics' affect the value of their home. For real estate investors, they may experience similar consequences and loose on their investments if renters and buyers shy away from properties in neighbourhoods affected by these so called 'unfavourable characteristics'.

In this project, we want to find out what will make a neighbourhood more favourable to invest in housing. How should people decide on which neighbourhoods to live in and call their homes? Which communities should a real estate investor put their investments in order not to make eventual losses? That is the business problem, and we are going to solve this with data and data science tools.

2. Data

To solve our business problem, we will use data and data science tools find out how people and investors make their choices regarding which communities to invest in real estate. This project will use data from the City of Edmonton, the capital of the Canadian province of Alberta.

This data is publicly available and can be sourced by navigating the City of Edmonton's website (www.edmonton.ca). Specifically, we use the 2016 Census Data and Crime Occurrence data tabulated by law enforcement in the City of Edmonton. The census data is a cross sectional data showing characteristics of the neighbourhoods in Edmonton that we will find relevant in solving our business problem. The latter is a dataset of aggregated crime occurrences in each neighbourhood between 2009 and 2019. The links below will give access to the data used and detailed description of the data content and how it was collected.

https://data.edmonton.ca/Census/2016-Census-Population-by-Age-Range-Neighbourhood-/phd4-y42v

https://data.edmonton.ca/City-Administration/City-of-Edmonton-Neighbourhoods-Centroid-Point-/3b6m-fezs.

Our focus, working with this dataset, will be on neighbourhoods with population densities greater than the average within the city. Further, we explore various neighbourhoods for specific features and amenities that are universally considered pleasant and hence, attract home ownership and real estate investment. To achieve this purpose location data is retrieved online from the Foursquare application. Some of the amenities which will be of interest will include universities and colleges, supermarkets and shopping malls, public transit stations, and recreational facilities such as hockey rinks. Edmonton is a major hockey city in North America, with a professional hockey franchise in the NHL, hence, this is particularly important for kids and adults alike.

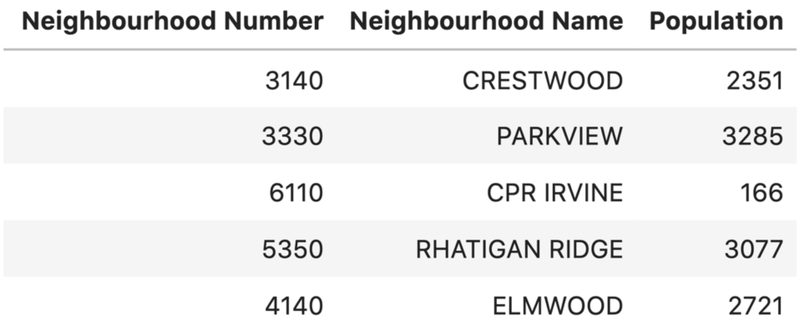
3. Methodology

The main data science analytic tool to be used in this exercise is Python (version 3.8). Microsoft Excel is also used for some functions on selected datasets for convenience and to streamline the data for the objectives.

3.1 Data wrangling procedures

First, we import the data into Python and explore its attributes. From the Census data, we can visualize the relevant demographic characteristics.

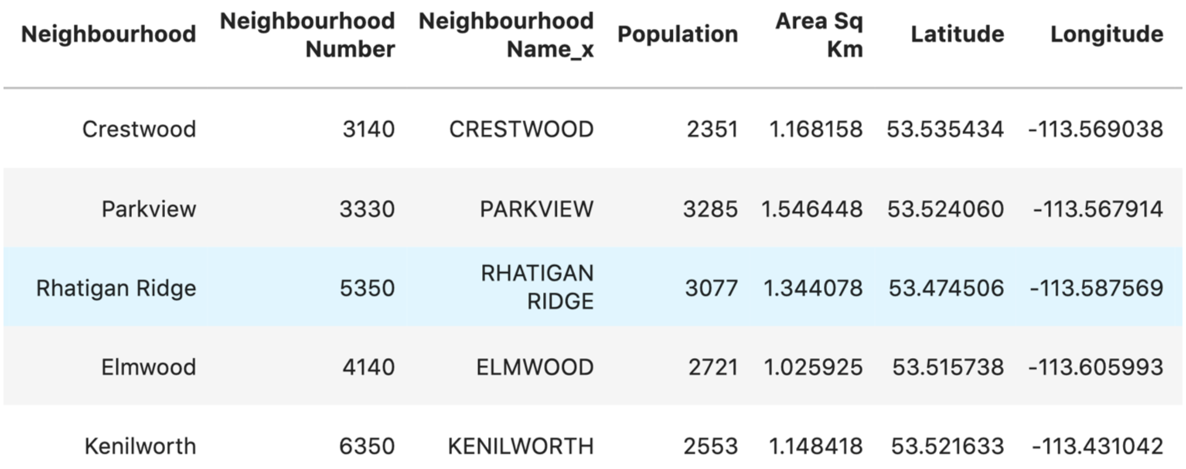
Table 1 – Demography: Neighbourhood and population



Neighbourhood Number	Neighbourhood Name	Population
3140	CRESTWOOD	2351
3330	PARKVIEW	3285
6110	CPR IRVINE	166
5350	RHATIGAN RIDGE	3077
4140	ELMWOOD	2721

Now the demographic data is merged with location data from the City of Edmonton. This allows to present the data with the precise geographical coordinates as show below.

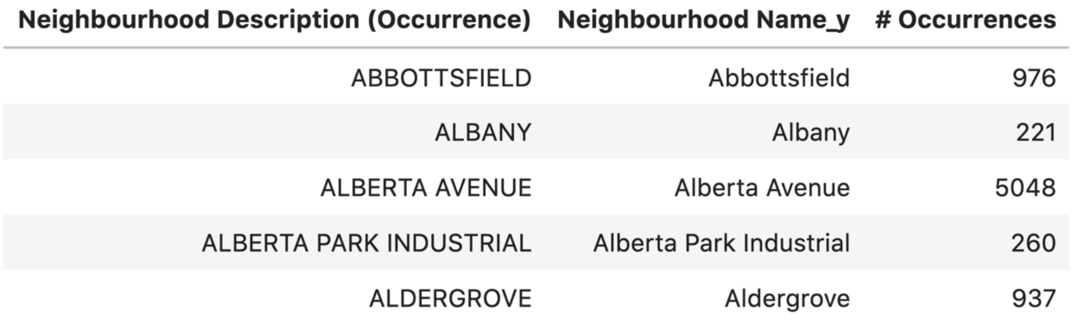
Table 2 – Neighbourhoods and Location coordinates



Neighbourhood	Neighbourhood Number	Neighbourhood Name_x	Population	Area Sq Km	Latitude	Longitude
Crestwood	3140	CRESTWOOD	2351	1.168158	53.535434	-113.569038
Parkview	3330	PARKVIEW	3285	1.546448	53.524060	-113.567914
Rhatigan Ridge	5350	RHATIGAN RIDGE	3077	1.344078	53.474506	-113.587569
Elmwood	4140	ELMWOOD	2721	1.025925	53.515738	-113.605993
Kenilworth	6350	KENILWORTH	2553	1.148418	53.521633	-113.431042

From this point, the crime occurrence data for the various neighbourhoods is imported and explored. The data is merged to Table 2 above to provide a new dataset, as show below

Table 3 – Neighbourhood and Crime occurences

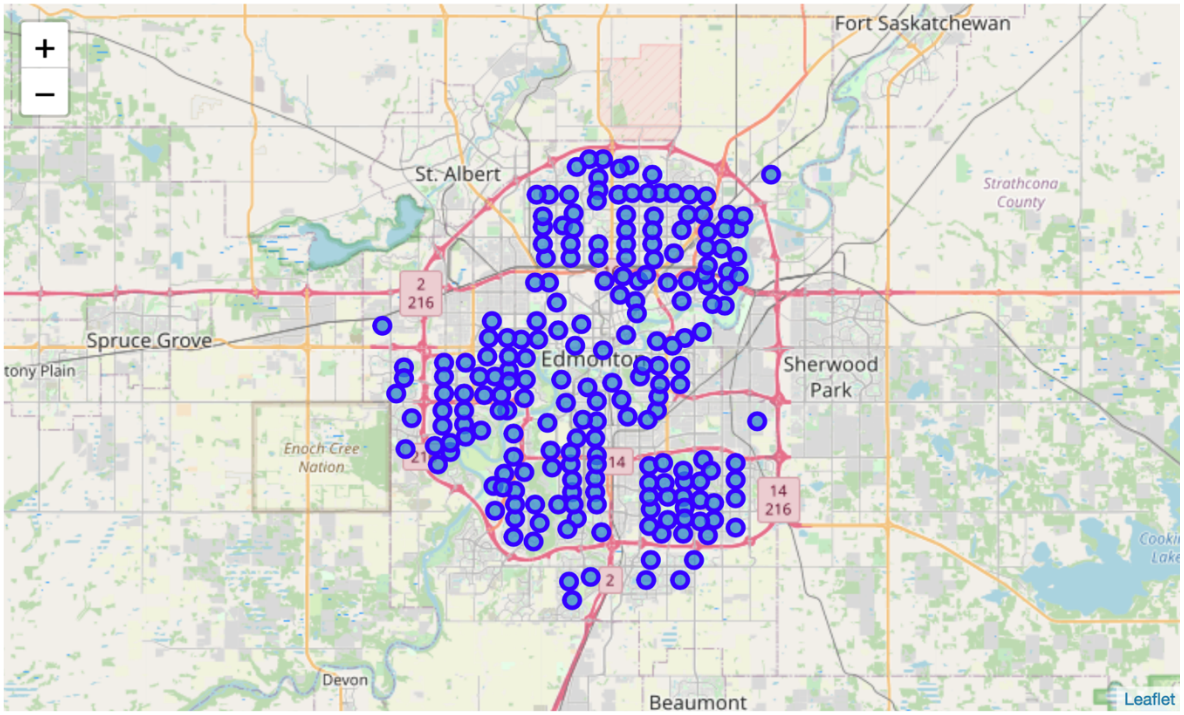


Neighbourhood Description (Occurrence)	Neighbourhood Name_y	# Occurrences
ABBOTTSFIELD	Abbottsfield	976
ALBANY	Albany	221
ALBERTA AVENUE	Alberta Avenue	5048
ALBERTA PARK INDUSTRIAL	Alberta Park Industrial	260
ALDERGROVE	Aldergrove	937

3.2 Exploratory analyses

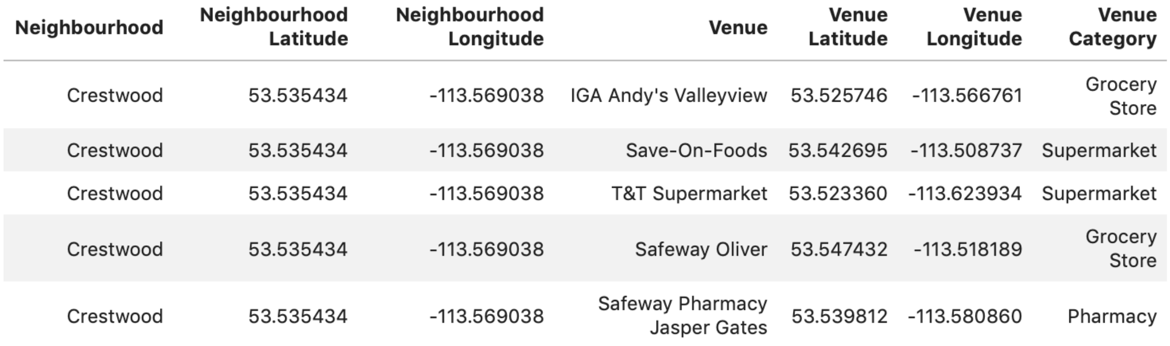
Python tools are used to further explore the data to gain more deeper insights. From the *folium* package, the library *geopy* is applied to generate the coordinates of various neighbourhoods within the city that have above average population densities. The reason for this is that, people are attracted to areas where other people live, and such communities are more likely to attract business and popular amenities for economic and social purposes. A picture of the map is show below

Figure 1 – Map of Edmonton



After visualizing the City of Edmonton and the location of neighbourhoods of interest, we proceed to identify what amenities are located in these neighbourhoods which may influence the decision of people to buy homes or invest in real estate development. For this, we use location data from the Foursquare application. A snapshot of what particular amenities of interest are in which neighbourhoods is show below.

Table 4 – Neighbourhoods and amenities

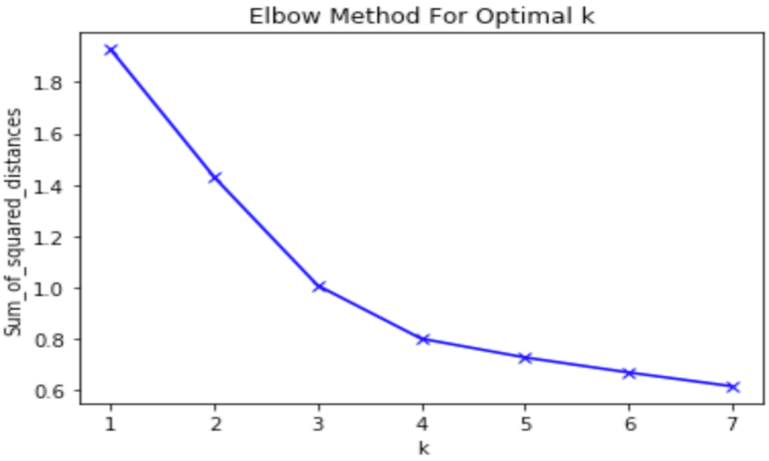


Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Crestwood	53.535434	-113.569038	IGA Andy's Valleyview	53.525746	-113.566761	Grocery Store
Crestwood	53.535434	-113.569038	Save-On-Foods	53.542695	-113.508737	Supermarket
Crestwood	53.535434	-113.569038	T&T Supermarket	53.523360	-113.623934	Supermarket
Crestwood	53.535434	-113.569038	Safeway Oliver	53.547432	-113.518189	Grocery Store
Crestwood	53.535434	-113.569038	Safeway Pharmacy Jasper Gates	53.539812	-113.580860	Pharmacy

3.3 Model development

The data science model which will be used for this project is the *machine learning* model know as K-Nearest Neighbour (KNN). The *k-means* algorithm is employed for clustering. By using clustering, it will be possible using data science to identify which specific neighbourhoods are most preferred for people to choose housing or invest in real estate. TO identify the optimal number of clusters, the *elbow method* was used to find the best *k*. As shown in the figure below, the best value of *k* was 3.

Figure 2 – The elbow method for selecting best k



At this point, using the best *k*, the algorithm is trained to yield three clusters of neighbourhoods as show in the diagram with different colour marks. For ease of analysis, amenities are grouped based on functionality, arbitrary scores are assigned and weighted to make sure that the size of the scores have no real statistical impact on the analysis. Lastly, the crime data is merged to the data set. The crime frequency is weighted and summed up to generate an aggregate score for each neighbourhood. This is presented below.

Figure 3 – Neighbourhood clusters

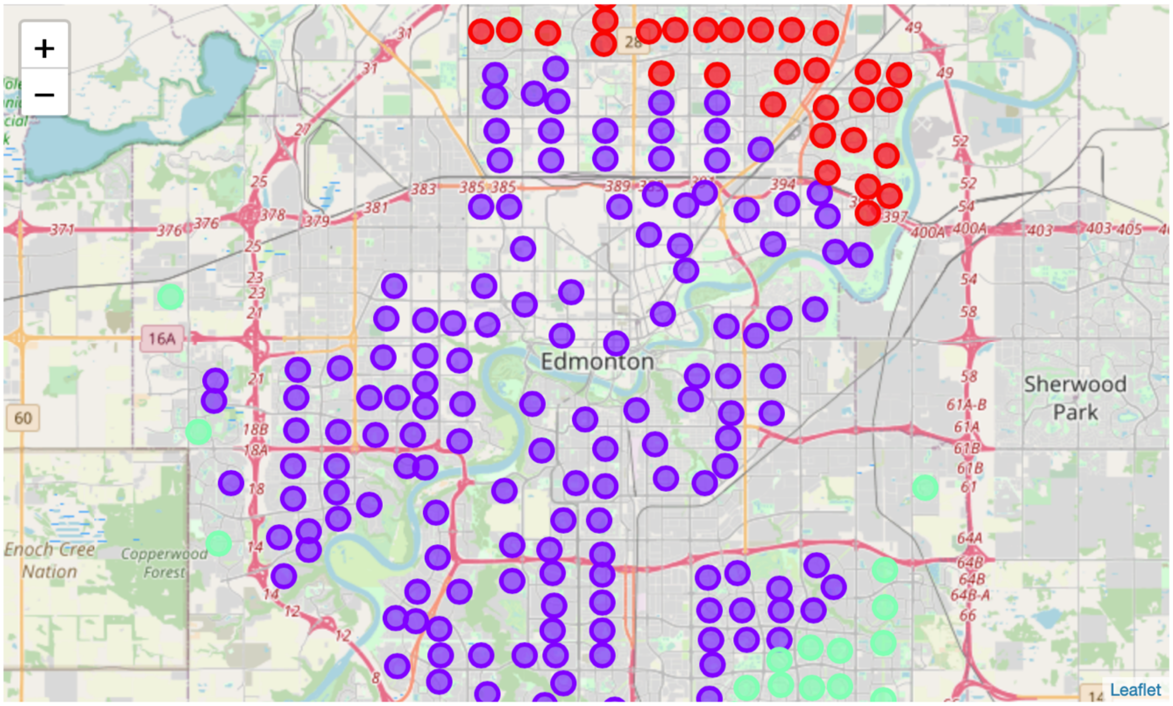
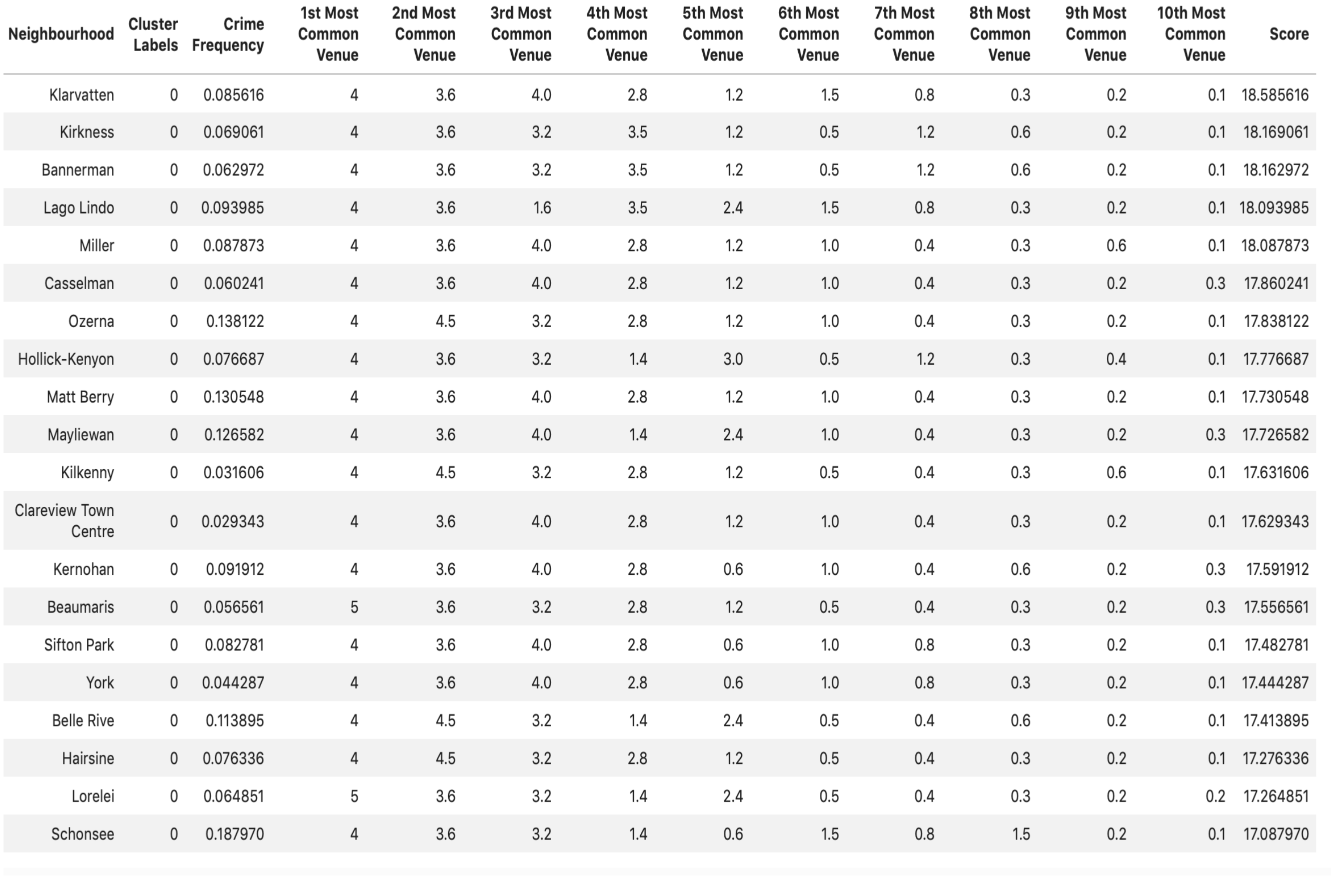


Table 5 – Scores: ratings of neighbourhood preference – ranked in descending order



Neighbourhood	Cluster Labels	Crime Frequency	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	Score
Klarvatten	0	0.085616	4	3.6	4.0	2.8	1.2	1.5	0.8	0.3	0.2	0.1	18.585616
Kirkness	0	0.069061	4	3.6	3.2	3.5	1.2	0.5	1.2	0.6	0.2	0.1	18.169061
Bannerman	0	0.062972	4	3.6	3.2	3.5	1.2	0.5	1.2	0.6	0.2	0.1	18.162972
Lago Lindo	0	0.093985	4	3.6	1.6	3.5	2.4	1.5	0.8	0.3	0.2	0.1	18.093985
Miller	0	0.087873	4	3.6	4.0	2.8	1.2	1.0	0.4	0.3	0.6	0.1	18.087873
Casselman	0	0.060241	4	3.6	4.0	2.8	1.2	1.0	0.4	0.3	0.2	0.3	17.860241
Ozerna	0	0.138122	4	4.5	3.2	2.8	1.2	1.0	0.4	0.3	0.2	0.1	17.838122
Hollick-Kenyon	0	0.076687	4	3.6	3.2	1.4	3.0	0.5	1.2	0.3	0.4	0.1	17.776687
Matt Berry	0	0.130548	4	3.6	4.0	2.8	1.2	1.0	0.4	0.3	0.2	0.1	17.730548
Mayliewan	0	0.126582	4	3.6	4.0	1.4	2.4	1.0	0.4	0.3	0.2	0.3	17.726582
Kilkenny	0	0.031606	4	4.5	3.2	2.8	1.2	0.5	0.4	0.3	0.6	0.1	17.631606
Clareview Town Centre	0	0.029343	4	3.6	4.0	2.8	1.2	1.0	0.4	0.3	0.2	0.1	17.629343
Kernohan	0	0.091912	4	3.6	4.0	2.8	0.6	1.0	0.4	0.6	0.2	0.3	17.591912
Beaumaris	0	0.056561	5	3.6	3.2	2.8	1.2	0.5	0.4	0.3	0.2	0.3	17.556561
Sifton Park	0	0.082781	4	3.6	4.0	2.8	0.6	1.0	0.8	0.3	0.2	0.1	17.482781
York	0	0.044287	4	3.6	4.0	2.8	0.6	1.0	0.8	0.3	0.2	0.1	17.444287
Belle Rive	0	0.113895	4	4.5	3.2	1.4	2.4	0.5	0.4	0.6	0.2	0.1	17.413895
Hairsine	0	0.076336	4	4.5	3.2	2.8	1.2	0.5	0.4	0.3	0.2	0.1	17.276336
Lorelei	0	0.064851	5	3.6	3.2	1.4	2.4	0.5	0.4	0.3	0.2	0.2	17.264851
Schonsee	0	0.187970	4	3.6	3.2	1.4	0.6	1.5	0.8	1.5	0.2	0.1	17.087970

4. Results

From our analyses, neighbourhoods in the first cluster is dominated mostly by grocery shops and supermarkets. There is easy access to transit stations and hospital facilities. However, schools are not very common as compared to neighbourhoods in other clusters.

In the second cluster, we realize there are a significant amount of grocery shops but not large supermarkets. Although there is relatively a fewer number of transit stations, hospitals and medical facilities are quite common.

Finally, within neighbourhoods in the third cluster, schools and grocery shop amenities are dominant significantly as compared to their counterparts in other clusters.

5. Discussion

The analysis indicated that neighbourhoods in the first cluster have the highest rated score, showing that, such neighbourhoods are most preferred for people who are seeking housing or looking for opportunities to invest in real estate. Reviewing the map, it is clear that these neighbourhoods are concentrated within and around the centre of the city. It is of no surprise, as the University of Alberta, colleges, the Alberta Hospital, and many light rail transit (LRT) stations are found in these areas. The neighbourhood of Klavartten will be recommended by the model as the best place to choose a home or invest in real estate as it has the best rating score.

6. Conclusion

In this project, we set out to explore what influences the decision of people when choosing homes or investors when finding areas to invest in real estate development. The data we used was from the City of Edmonton. After the analyses, we found out that neighbourhoods with the highest ratings as determined by the model were the most preferred. These neighbourhoods were mostly found in the first cluster, which was concentrated in and around the city centre.