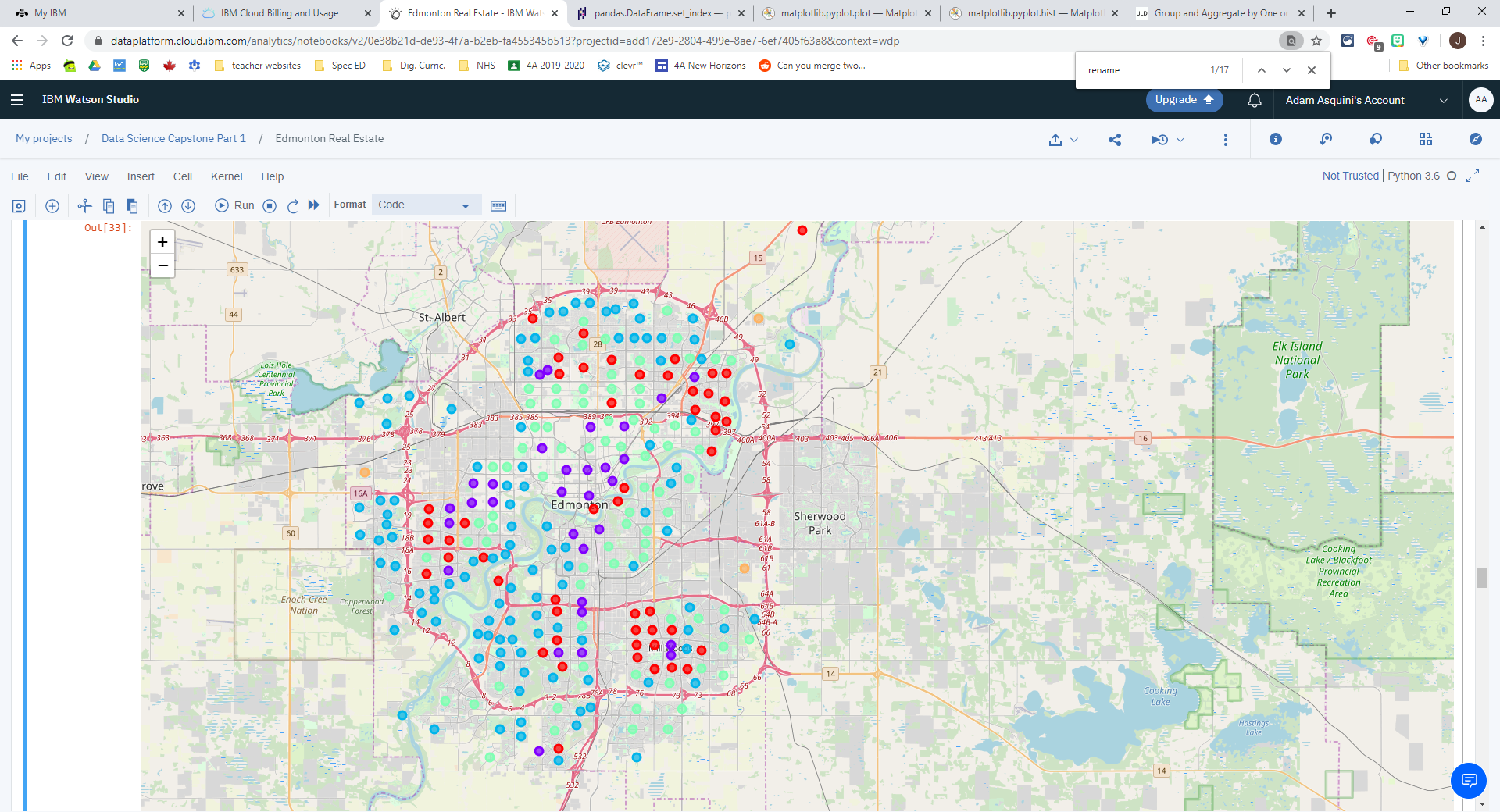
Edmonton Neighbourhood Selection



**Report Prepared By:**

Adam Asquini, Coursera / IBM Applied Data Science Professional Certificate Capstone Project

Part 2: March 22, 2020

# Introduction

## Background Context

Edmonton is a major Canadian city with have a population of over 1M people (including surrounding suburban communities). Edmonton is located in the northern half of the Canadian province of Alberta, Canada and is Canada’s most northerly major city[[1]](#footnote-1). Often referred to as the gateway to the north, Edmonton serves as a major economic centre for the oil and gas industry and provides a logistical and transportation hub for communities further north in the Alberta oil sands such as Fort McMurray. Edmonton also has a large public sector workforce and is growing its technology sector. As a large urban centre, Edmonton has thousands of real estate transactions across its hundreds of neighbourhoods each year. Having knowledge of these neighbourhoods would be a benefit to buyers, sellers and real estate professionals

## Business Problem

According to the Alberta Real Estate Association, approximately 19,000 real estate transactions occur in the Edmonton region each year[[2]](#footnote-2). Given that many real estate transactions see hundreds of thousands (or millions) of dollars exchange hands, real estate purchases require careful consideration and can have profound impacts on both buyers and sellers. With thousands of real estate transactions occurring annually, both customers and real estate professionals would benefit from data-driven analysis to aide in these important investments.

This report examines a broad set of publicly available, online sources of data to aide customers and real estate professionals in making decisions around real estate transactions. Without this type of data-driven analysis, stakeholders are often left to more traditional and limited data analysis techniques such as looking at various sources of data independently or are reliant on the advice of people such as friends, advisors and workplace companions, which can be rife with bias.

The analysis completed in this report will attempt to segment and cluster neighbourhoods based on their similarities to other neighbourhoods (not including location). Data analyzed about neighbourhoods to conduct this grouping will include attributes such as amenities, crime statistics, demographics and property assessment values. The output of this analysis will be tool that can be used to better understand the general characteristics of Edmonton’s neighbourhoods, which could then be mapped to a client’s unique needs and preferences. For example, a client with a budget of $500K and a desire to live in an area with low crime rates could use the tool to find a cluster of neighbourhoods across the city that best meet this criteria. From here, other personal preferences such as the client’s commute to work and the location of family and friends could be applied to further refine the areas the client may look for a new home.

While some tools already exist to complete similar analysis, such as the Multiple Listing Service (MLS), this tool will integrate other non-traditional sources of data from diverse sources to provide a more thorough analysis of neighbourhoods in Edmonton.

## Audience for this Report

The primary intended audience of this report are the clients (both buyers and sellers) that are interested in participating in a real estate transaction. The buyer’s stake in this report is having more knowledge of neighbourhoods they may move to so that they can make a more informed choice about their potential home purchase. The seller’s stake in this report is that they may be able to better understand the characteristics of the neighbourhood in which their property is located, which may help them with pricing and promoting positive attributes of their neighbourhood (i.e. location to parks or amenities).

A secondary audience of this report is real estate professionals. While Realtors and other real estate professionals may come to similar conclusions about neighbourhoods through their experience, this report may be useful when discussing options with clients as the report provides a less biased view of certain neighbourhoods in Edmonton.

# Data Sources and Acquisition

The following data sources and their respective description have been utilized in this report.

|  |  |
| --- | --- |
| **Data Source (hyperlinked)** | **Description** |
| [City of Edmonton Open Data Portal – City of Edmonton Neighbourhoods](https://data.edmonton.ca/City-Administration/City-of-Edmonton-Neighbourhoods/65fr-66s6) | List of all neighbourhoods and their respective neighbourhood id in the city of Edmonton. This list forms the basis for comparison across neighbourhoods. |
| [City of Edmonton Open Data Portal – Neighbourhood (Centroid Point)](https://data.edmonton.ca/City-Administration/City-of-Edmonton-Neighbourhoods-Centroid-Point-/3b6m-fezs) | List of the latitude and longitude of the centre point (centroids) of each Edmonton neighbourhood. Used as a coordinate point to determine nearby amenities. |
| [City of Edmonton Open Data Portal – Property Assessment Data (Current Calendar Year)](https://data.edmonton.ca/City-Administration/Property-Assessment-Data-Current-Calendar-Year-/q7d6-ambg) | Property Assessment values for all homes in the city of Edmonton. The average assessed values of homes in each respective neighbourhood will be used as a comparator. |
| [City of Edmonton Open Data Portal – Edmonton Police Service Neighbourhood Criminal Occurrences](https://dashboard.edmonton.ca/dataset/EPS-Neighbourhood-Criminal-Occurrences/xthe-mnvi) | Statistics on crime reported to the Edmonton Police Service by neighbourhood in Edmonton in the year 2009 – 2019. Aggregate crime rates by neighbourhood will be calculated as an indicator of the level of crime in the respective neighbourhood |
| [City of Edmonton Open Data Portal – 2016 Census Population by Household Income (Neighbourhood / Ward)](https://data.edmonton.ca/Census/2016-Census-Population-by-Household-Income-Neighbo/jkjx-2hix) | List of the number of people belonging to each income bracket in each neighbourhood in the city of Edmonton. Used to determine the average household income in each neighbourhood. |
| [City of Edmonton Open Data Portal – 2016 Census - Population by Age (Neighbourhood / Ward)](https://data.edmonton.ca/Census/2016-Census-Population-by-Age-Range-Neighbourhood-/phd4-y42v) | List of the number of people belonging to each age bracket in each neighbourhood in the city of Edmonton. Used to determine the average age in each neighbourhood. |
| [City of Edmonton Open Data Portal – 2016 Census – Population by Structure Type (Neighbourhood/Ward)](https://data.edmonton.ca/Census/2016-Census-Population-by-Structure-Type-Neighbour/68uk-rvf5) | List of the number of each home type (single-family dwelling, apartment etc..) in each neighbourhood in the city of Edmonton. Used to determine the rates of each structure type in each neighbourhood. |
| Foursquare location data  (No hyperlink available. Data pulled through Foursquare API) | List of nearby (within 500 metres) venues from each neighbourhood centroid. |

# Methodology

## Exploratory Data Analysis

To conduct my analysis, I created a Jupyter Notebook on IBM Watson studio. A link to the completed notebook can be found on [GitHub here](https://github.com/aasquini/Data-science-project/blob/master/Edmonton%20Neighbourhood%20Selection.ipynb).

**Data Wrangling and Exploration:** I began my analysis by importing all the datasets listed in the data section above. Specific data exploration and wrangling activities are as follows:

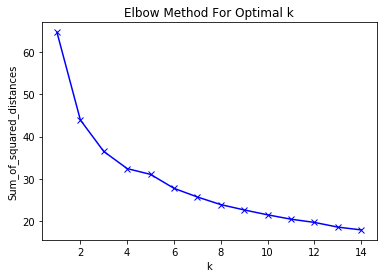
* **Neighbourhoods** – No data wrangling required. The shape of this dataset was 400 rows by 2 columns, meaning there are 400 neighbourhoods in Edmonton
* **Neighbourhood Centroids** - No data wrangling required. The shape of this dataset was 400 rows by 2 columns, meaning there was centroid data (latitude and longitude) for all 400 neighbourhoods in Edmonton
* **Property Assessment Values** – All commercial and residential properties were included. Originally the dataset contained 406,406 rows and 12 columns. One row was removed for having no values. I also removed all commercial properties from the dataset. Finally I grouped the dataset by neighbourhood. Of the 400 neighbourhoods in Edmonton, only 352 had residential properties.
* **Criminal Occurrences** – All criminal occurrences from 2009 to 2019 were included. Resulting in a dataset of 113,077 rows and six columns. I decided to keep the crime data from 2019 only for this analysis and group the dataset by neighbourhood. Once grouped, I was able to determine that 376 of the 400 rows had criminal occurrences in 2019.
* **Population by Household Income** – This dataset was already aggregated by neighbourhood. It had 388 neighbourhoods analyzed. The total number of households within income brackets such as $0-<$30,000, $30,000-$60,000 etc.. were provided.
* **Population by Age** – Similar to the Population by Household Income Dataset, this dataset was already aggregated by neighbourhood.
* **Population by Housing Structure Type** – Similar to the Population by Household Income Dataset, this dataset was already aggregated by neighbourhood.
* **Foursquare data** – The location of venues within 2km (2,000 metres) was collected by calling the foursquare explore venues API for each neighbourhood. 11,521 records were identified through this call to the Foursquare API. Of note, it was accepted that not all venues may be unique. If a given venue is within 2km of two (or more) neighbourhood centroids, this may cause the same venue to appear more than once. This was accepted because a venue within 2km was deemed to be accessible even if it wasn’t within a given neighbourhood’s boundaries. Without this determination, residential suburbs may be significantly impacted where venues such as grocery stores and drug stores are generally more spread out and serve multiple neighbourhoods in a ward or region. After collecting the 11,521 records, the dataset was sliced to only include venues that are traditionally seen as desirable for real estate. The types of venues included were: Bank, Café, Gas Station, Grocery Store, Gym, Gym / Fitness Center, Library, Playground, Pool, Recreation Center, School, Supermarket.

**Merging datasets and normalization:** Once all data was brought into the notebook, I created a dataframe to merge the results. Because the numerical value of a property assessment is significantly larger than the number of criminal occurrences, which is larger than the number of schools or coffee shops in a neighbourhood, I normalized the datasets so that all values were between 0 and 1. To do this I applied the following rules:

* For Property Assessment Values. I first grouped the property values by neighbourhood and calculated the mean average by neighbourhood. I then applied Mix Max Scaling using the SKLearn preprocessing package to translate mean property value by neighbourhood to a number between 0 and 1.
* For Criminal Occurrence Data. I first grouped the criminal occurrence data for 2019 by neighbourhood and took the total count by neighbourhood. I then applied Min Max Scaling to each mean neighbourhood value.
* For the population by household income, population by age and population by construction type, I took the mean value of each column for each neighbourhood. For example, if a neighbourhood had 10 out of 100 households in the $30,000 - $60,000 income bracket, the value of 10 was converted to 0.1.
* For Foursquare Data: After collecting the venues, one hot encoding was used to give each venue type its own column. Once this was completed, the relative percentage of the venue type for each neighbourhood was taken.

## Modeling Techniques

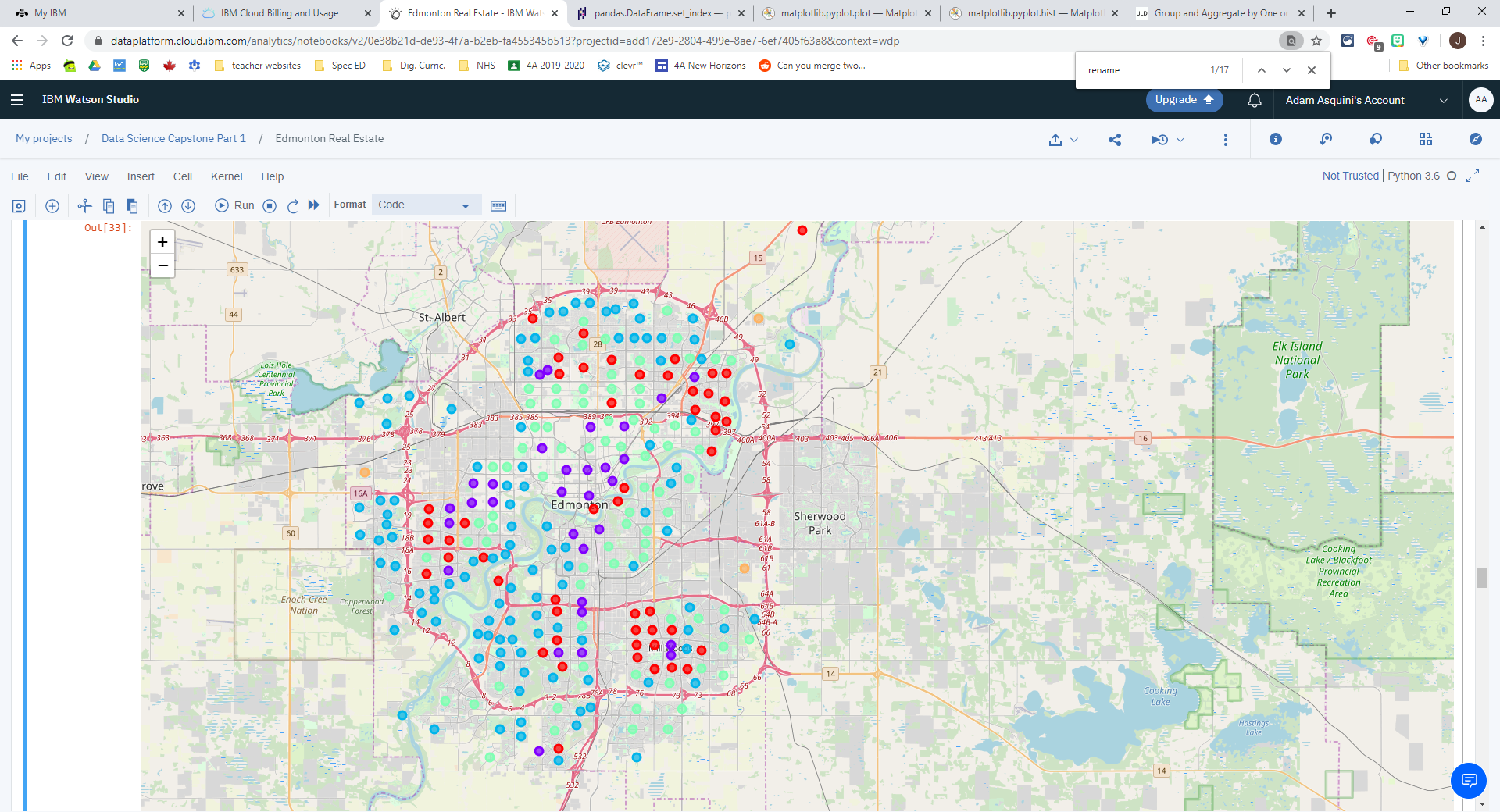
Once all data was normalized and merged, the K-Means Neighbours algorithm was applied using the SKLearn package. The merged dataset was fit with k values from 1-15 and the elbow method to determine the optimal K value was utilized. The results of this analysis are below:



Based on this diagram, I selected a K value of 5 and applied the KMeans algorithm.

# Results

The resulting five clusters of neighbourhoods were plotted on a map using the Folium package. The resulting map is show below.



Two dataframes were also created in the Jupyter Notebook to show the normalized mean results for each category by cluster as well as the actual mean results. While the dataframes are too large to view in a word document, the following facts from each cluster were observed:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Cluster** | **Cluster Size** | **Avg Assessment Value** | **Avg Criminal Occurrences / Neighbourhood** | **Income** | **Age Distribution** | **Structure** | **Venue / Amenities** |
| **0** **Red** | 51 | $282K – 3rd | 30 – 3rd | 68% between $0-$100K | Results similar in all clusters. However, cluster 2 has more 20-30 year-olds than the other clusters. | 43% single detached house | Highest # of recreation and supermarkets |
| **1 Purple** | 31 | $188K – 4th | 48 – 1st | 79% between $0-$100K | 20% single detached house | 4-5 times more cafes than other clusters |
| **2** **Blue** | 106 | $468K – 1st | 16 – 5th | 42% between $0-$100K | 84% single detached house | Highest number of playgrounds  2nd lowest $ of amenities |
| **3 Green** | 79 | $303 – 2nd | 36 – 4th | 65% between $0-$100K | 67% single detached house | 2nd Highest # of recreation centres and supermarkets |
| **4 Orange** | 3 | $51K – 5th | 20 – 2nd | 85% between $0-$100K | 0% single detached house  99.8% mobile home | Gas station only noted amenity |

# Discussion

Based on the results of the clustering analysis, I believe that buyers and sellers would be able to utilize this tool.

A general description of each cluster and the target customer is presented below:

| **Cluster** | **Locations Observed** | **Description of Cluster** | **Recommended Target Customer** |
| --- | --- | --- | --- |
| **0 Red** | Dispersed, but 3 pockets of neighbourhood observed in the NE, W and SE | Mix of single detached homes, row houses and low-rise condos. The cluster has moderate crime levels and a diverse spread of incomes and age brackets. This cluster has the most recreation centres and super markets. Many of the neighbourhoods are mature in this area. | First time home-buyers, families with modest income looking for a single family home, townhouse or condo. |
| **1**  **Purple** | Mostly inner city with small pockets in the W and SE | This cluster hosts most of the downtown neighbourhoods. There is a high number of condos & apartments. Crime is higher in this cluster. However, this may be expected with a cluster hosting many downtown neighbourhoods. The cluster offers the most entertainment and amenity options. | Single adults or couples without kids. Anyone looking for a downtown / urban experience. This cluster may also be attractive for investors looking to rent apartments to business people or students. |
| **2**  **Blue** | Dispersed throughout the perimeter of the N, NW, W and SW portions of the city | Mainly suburban neighbourhoods, many of which are newer explaining their locations closer to the perimeter of the city. This cluster has the highest average household income, highest percentage of single family homes, highest average property assessment value and lowest crime rates | Families looking to expand into a larger home. May also cater to individuals or family that work away from downtown and are not affected as much by commutes into downtown for work. |
| **3**  **Green** | Dispersed further towards the centre or the N, NW, W and SW portions of the city | Mainly suburban, mature neighbourhoods. This cluster is made up of neighbourhoods with predominantly single detached homes and low-rise condos. Like cluster 0, there are lots of family amenities close by in these neighbourhoods. | First time home-buyers, families with modest income, but are more interested in a single family home. Also may be an option for seniors given the proximity of neighbourhoods to services and transportation. May be an option for a customer that does not want to drive given the relative proximity to downtown. |
| **4**  **Orange** | Dispersed on the perimeter in rural areas | This is a small cluster of mobile homes. | Any customer interested in living in a mobile home. |

As noted in the problem description, the intent of forming the clusters was to give real estate customers a start point to analyze similar neighbourhoods. I believe the clusters would give a great start point for potential buyers to understand where in the city they could or should look. I also believe that sellers could use this information to understand characteristics of their own cluster compared to others, or just as importantly how their neighbourhood compares to other similar neighbourhoods in the same cluster.

# Conclusion

In this analysis, the neighbourhoods of Edmonton were compared and grouped into clusters based on attributes such as average assessed value, criminal occurrences in each neighbourhood, census data and crowd-sourced data about local venues from Foursquare. The report provided a comprehensive analysis of neighbourhoods in the city of Edmonton to aide potential buyers, sellers and real estate professionals with real estate transactions. While it is believed this analysis has given a fairly good picture of Edmonton neighbourhoods and how they cluster and would be useful, its not enough to make real estate decisions on independently. Further analysis of Edmonton neighbourhoods using more data sets, different clustering techniques and different preprocessing techniques should also be completed to validate the results in this report.

1. <https://globalnews.ca/news/5861019/census-edmonton-population-2019/> [↑](#footnote-ref-1)
2. <https://albertarealtor.ca/page/data-explorer> [↑](#footnote-ref-2)