2021

Proposal for Clothing store Build site

PART I OF THE CAPSTONE PROJECT

ALEC PARISE

JANUARY 8, 2021 |

Contents

Table of Figures

I. Introduction/ Business Problem

In recent decades, the Waterloo region has experienced linear growth patterns in population since 2011. According to the 2016 Census of Canada, Waterloo Region is home to 535,154 people, which represents 4.0% of the total population of Ontario (Smale, 2017). In a country which has annual growth of 5% each year, the growth rate of Waterloo region ranks it amounts the highest in Canada. Furthermore, corporations like Amazon and Google are paving the way for opportunity in the region by creating jobs in the tech and retail industry. As appealing as this would seem for new business owners in the region, would it be economically viable, given the increase property value, to open a business within the region? If so where?

My client is wanting to expand their apparel company into the region, and tasked me with determining the location, more specifically the neighborhood, that provides the most potential for their brand. This report is separated in IV sections where, the Introduction/Problem is discussed in Section I, Description of the Data in Section II, the Methodology in Section III, the Results in Section IV, the Discussion of the results in Section V. Finally, in Section VI the Conclusion.

II. Description of the Data

The data will consist of three types of data which will be required for analysis for this contract.

1. Names of the Neighborhoods within the region
2. Geographic Coordinates of the Neighborhoods
3. The venues and foot traffic for the Waterloo Region

Each data type will be merged with the ladder creating one dataframe containing the Neighborhoods, Geographic coordinates, and the venues foot traffic.

Names of the Neighborhoods

This dataset will contain the names of the different neighborhoods in the Waterloo Region. The data was obtained from the Waterloo Region open-source website (link: https://rowopendata-rmw.opendata.arcgis.com/datasets/d1656d3c7abb4c7da20fab83c77caec7_0/data?page=32). The records provide outdated population counts, so for the purpose of this study they were omitted from this dataset but will be included later on in the Foursquare dataset.

Geographic Coordinates

The geographic coordinates of each location were added to the dataset using the *Nominatim* module from the Python Geopy Library. This allowed to assigned Latitude and Longitude coordinates to each Neighborhood in the Waterloo Region.

Venues and foot traffic

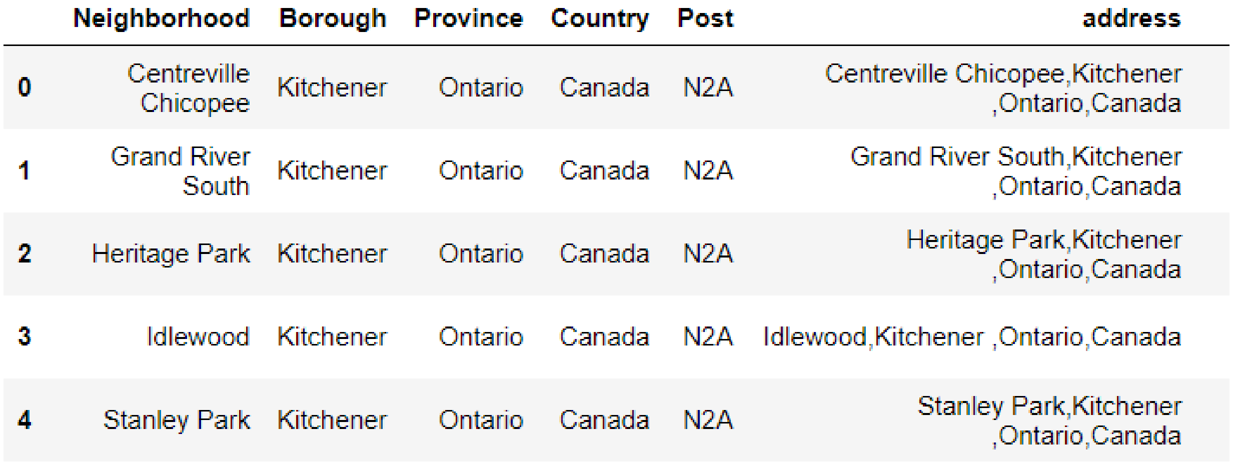
The Foursquare API was used to retrieve information about the most popular spots in berlin. The popular spots returned depends on the highest foot traffic at the time when the call is made. We may get different popular venues at different times of the day.

III. Methodology

The methodology was broken up into three subsections, Data Cleaning Data exploration phase, Building the KNN model clustering model.

Data Cleaning

Since there was no data readily on a website, I had to manually insert the coordinates to into a csv file which included 5 columns: The neighborhood, the City in which it belonged, and the postal code, the province, and the Country. I then imported .csvfile into Jupyter notebook and began pre-processing the data. Initially to retrieve the foursquare data the Neighborhood column did not provide enough context to retrieve the proper Latitudde and Longitude coordinates. So to fix this I concatenated the five columns to form an address column, as seen in the Table 1 below.



	Neighborhood	Borough	Province	Country	Post	address
0	Centreville Chicopee	Kitchener	Ontario	Canada	N2A	Centreville Chicopee,Kitchener ,Ontario,Canada
1	Grand River South	Kitchener	Ontario	Canada	N2A	Grand River South,Kitchener ,Ontario,Canada
2	Heritage Park	Kitchener	Ontario	Canada	N2A	Heritage Park,Kitchener ,Ontario,Canada
3	Idlewood	Kitchener	Ontario	Canada	N2A	Idlewood,Kitchener ,Ontario,Canada
4	Stanley Park	Kitchener	Ontario	Canada	N2A	Stanley Park,Kitchener ,Ontario,Canada

By doing this I was able to retrieve the proper geographic coordinates for each venue in the neighborhood. The next section will entail the procedures for exploring the data.

Data Exploration

As described in the previous section, the final data set consisted of all of the neighborhoods in the Waterloo region. To create the final dataframe, the total number of venues per neighborhood was taken in a radium of one kilometer. From the image below (Figure 1) there was substantial overlap between neighborhoods. This was due to the sparsity of certain neighborhoods in the outskirts of the region. To compensate for the overlap in city core regions the duplicate venues were dropped based on the proximity of the closest neighborhoods. One good thing to note however that some venues had coverage in up to 5 neighborhoods in the city of Kitchener, a city within the Waterloo region.

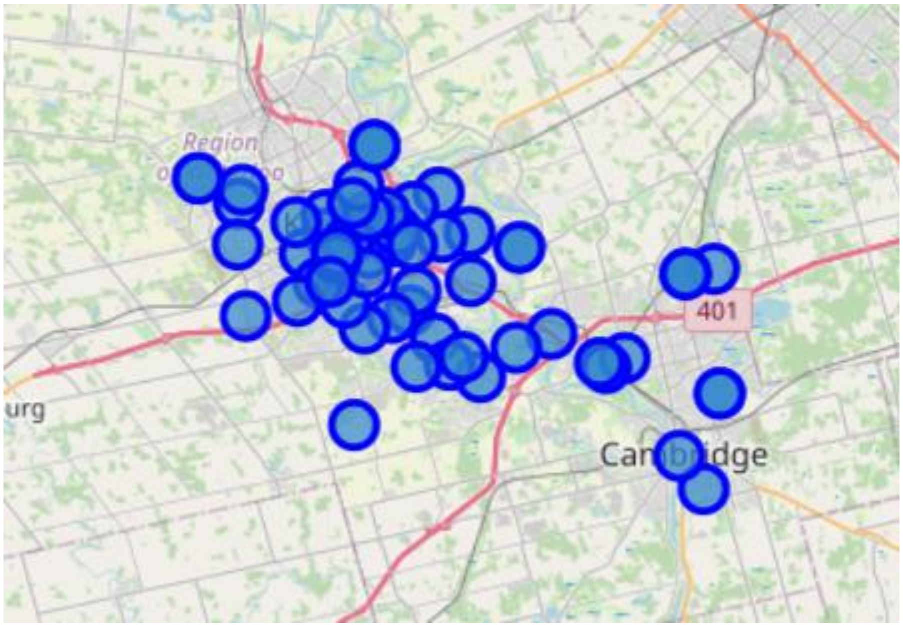


Figure 1: Neighborhoods in the Waterloo Region.

To understand the density of venues withing the region, a heat map was created. Heat maps are great visuals for interpreting locations of high density. In the case venue density. This figure helped me understand the areas which had more foot traffic, which resulted in more established businesses. Figure 2 below show the heat map.

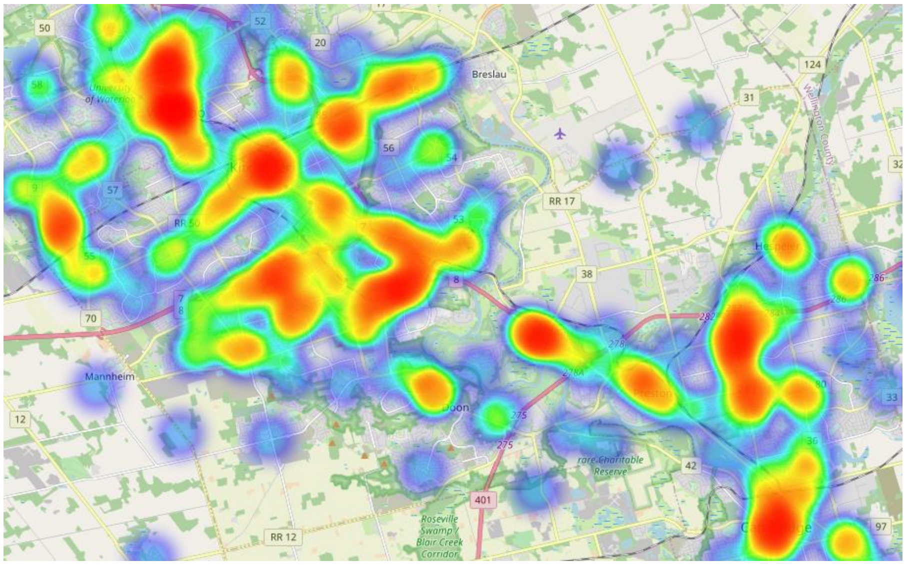


Figure 2: Heat Map showing venue density.

This section allowed me to understand my data and proves my hypothesis that Neighborhoods close to the core would have the highest density of venues and population. This will further show in the next section; Clustering.

Clustering

To cluster the data, the KNN method was used. This method was selected because of its versatility and ability to cluster location based categorical data. However, for this dataset there was no ordinal relationship between categories and integers, therefore, to be able to cluster categorical data you must normalize it so that the data can be interpreted by the algorithm. One way to do this is to use Onehot encoding. Therefore, prior to creating the clustering algorithm the dataset was input into a One hot encoding algorithm which converted categorical data into integer values. Then to select the optimal value for k (an essential process for using k-means) the elbow method was selected. Figure 3 shows the result for the elbow method

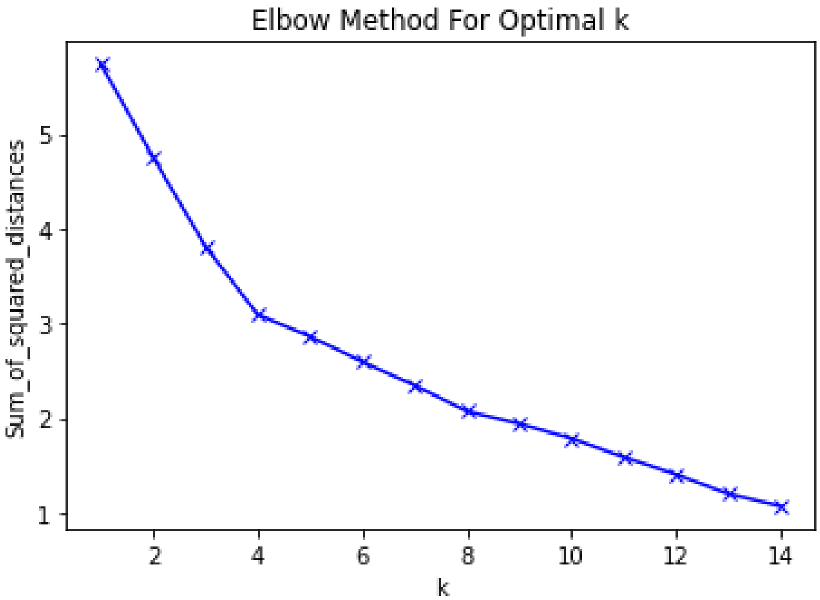


Figure 3: Elbow method for KNN clustering.

From the figure above we see that there is a slight bend at k = 4. This value was selected as the k values, which told the KNN algorithm how many clusters were to be made. The output of the final clustering can be seen in Figure 4. Note that we see four main groups with 2 being outliers with single members, seen in table . The larger groups, cluster 1, cluster 2 and cluster 3 show the number of members in that region.

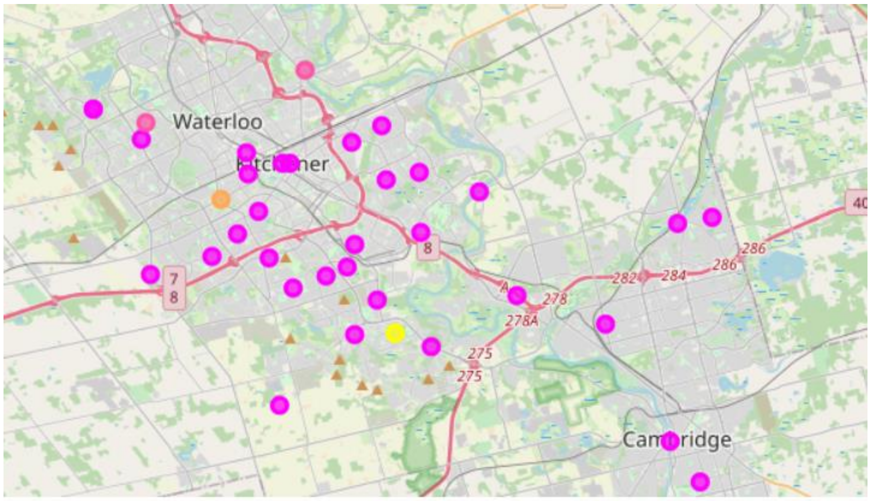


Figure 4: Map of 4 clusters made by the KNN model.

To dig a little deeper in the specifics of clustering, refer to Figure 5 , where the frequency of each venue type can be seen for the 4 clusters. The venues are plotted in alphabetical order on the x-axis from left to right. This plot allows a high level of visualization of the difference and similarities for the cluster groups. Some key findings from this visual is the two largest peak seen cluster 0 and cluster 3. These are areas on the outskirts of the city and only have 1 venue within its vicinity, a hocket arena and farmer's market.

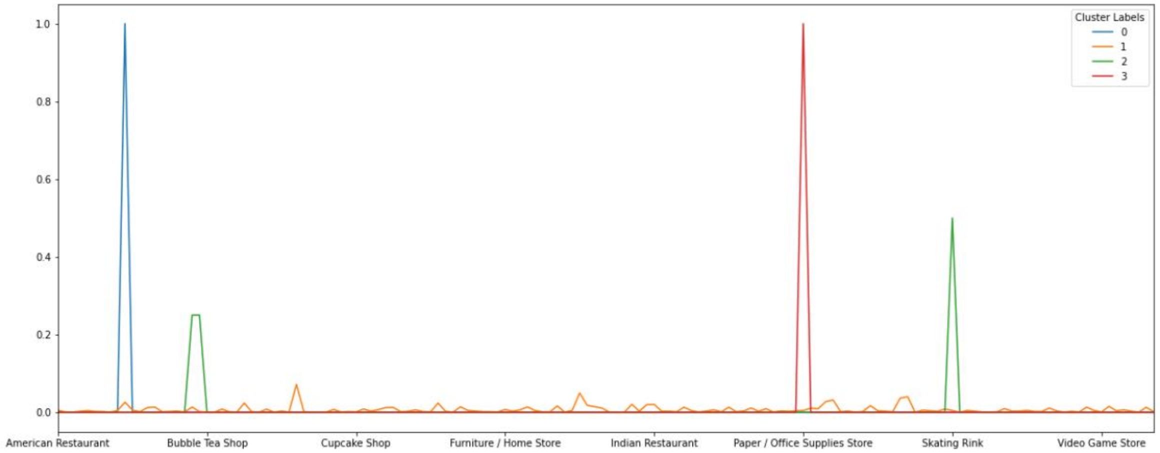


Figure 5: Distribution of venues and associated clusters.

Furthermore, since the visualization seen above is limited to the number of venues that can be included. I have ranked the top 10 venues in the most popular neighborhoods. Based on this, the area types for clusters can be characterized by venues close to them.

IV. Results

In this project we have explored the top venues and locations of each venue in the Waterloo Region. Using clustering based on venues within walking distance, defined as 1000m radius we have clustered the venues into four groups.

Comparing the three largest clusters (clusters 2, 3 and 1 in descending order) we note the following:

The second and third largest clusters only contain approximately 6 venues and the biggest cluster (Cluster 1) contained the majority of venues located in the biggest neighborhoods. Cluster 1 contained all of the venues like shopping malls, coffee shops and cafes. To be more specific about the type of each area beyond these shared features is difficult but we note a few differences using the top ten venue types for each cluster: Cluster 1 (label 1) has the "core" venue types one might except and has many types of restaurants and grocery stores. Cluster 2 (label 2) has more skating rinks and breakfast spots compared to other clusters. Cluster 3 (label 3) shows local restaurants and some fast-food restaurants. however very few. Cluster 0 (label 0) Contains venues like farmers markets and breweries.

V. Discussion

Based on the results we have identified three types of areas unique venues appear. As assumed at the beginning of this experiment, the location of establishing a business will limited to more popular neighborhoods, specifically those located at the city core. This assumption would need to be validated with data obtained from the waterloo region. The metrics from this experiment proved that areas of densely populated regions contain shopping malls, cafes, and restaurants, whereas lower populated areas had venues like local restaurants, schools, skating arenas and farmers markets. Other potential metrics to measure the viability of establishing a business would be establishment and maintenance costs (e.g., how much would property tax be in a region, the growth rate of each region and expansion trends, and finally what venues are nearby).

In theory, using these results, a clothing brand could characterise other areas where they might be considering opening a business. However, a good starting point would be to find means within their budget but will currently have significant foot traffic. One thing noted during the exploration phase we created a heat map which was a clear indicator of foot traffic density and how many venues were within a 1km radius. In further work it would be idea if we could create a time lapse and see the growth rate over time to better understand the trends of growth within the waterloo region.

IV. Conclusion

The aim of this project was to explore the following question:

Where would be the best location to build a business as a clothing brand within the waterloo region? What are the most popular neighborhoods and what venues are situated within? Alternatively, is there a particular type or number of types of neighbourhoods in which a clothing brand should be established. We have been able to show that the most popular venues fall into three main types of neighbourhood, city core, outer city, and outskirts. This analysis might allow businesses to understand what is in the region and what kind of impact that may or may not have on their business. Key things to explore further would be how to implement a time series analysis to determine the trends of city expansion, what are the most successful brands and what kind of services are nearby, (for example commuter train stations and key roads).