Relationship between Neighbourhoods and Crime Rates in Vancouver City

Capstone Project for IBM Data Science on Coursera

Huabei You (huabeiyau@gmail.com)

I. Introduction

Background

RECENT studies in social science suggest that the local crime rate may be correlated with the characteristics of the neighbourhood [1]–[3]. On the other hand, neighbourhoods can be categorized by the types of venues located within them. Since high crime rates can have a negative impact on the property value [4], it is in stakeholders' best interests to invest in areas with low crime rate not only at the moment but also in the future.

Problem

This study aims to discover the relationship, if any, between popular venues in the neighbourhood and local crime rates. Vancouver city was selected as the real life example and a recommendation of the best neighbourhood in Vancouver is made based on results of the analysis.

Target Audience

Real estate companies and individual house buyers will certainly benefit from the insight given by this report.

II. Materials and Methods

A. Data Sources

The geographic data of Vancouver city can be downloaded from Vancouver city open data portal [5] and can be used under the terms of the Open Government Licence – Vancouver [6]. The crime data is available in Vancouver police department website [7]. Finally, the venue information is provided by Foursquare Places API [8].

B. Methodology

The geographic data set contains the boundaries of 22 neighbourhoods in Vancouver city. Their names and center coordinates are extracted and visualized in Fig. 1, which is identical to the official map of Vancouver city (See Fig. 2).

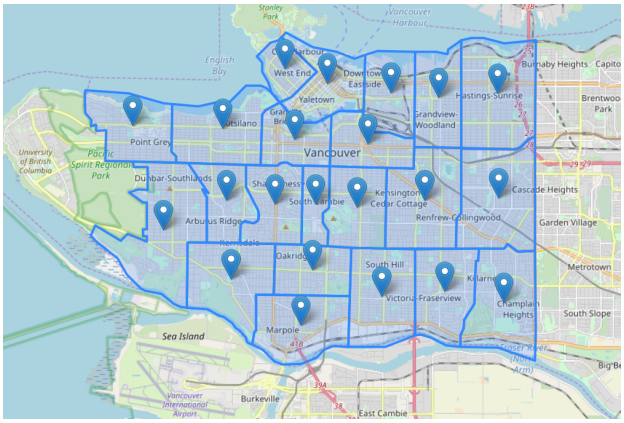

Fig. 1. Neighbourhood divisions of Vancouver City based on the geographic data set. (Centrals of the respective neighbourhoods are marked by blue circles.)

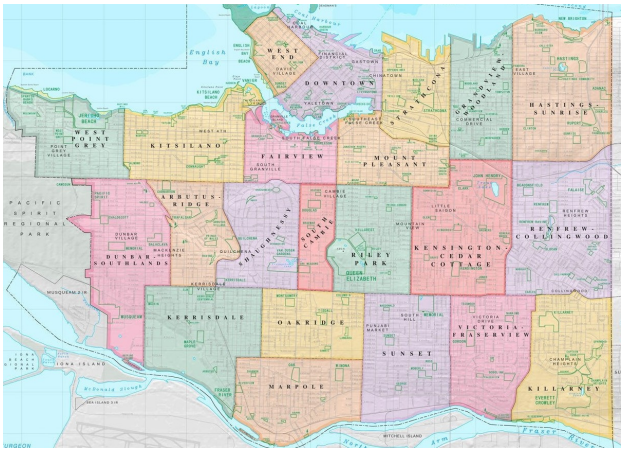

Fig. 2. Official Neighbourhood Map of Vancouver City.

The crime date set consists of 627174 records from year 2003 to 2019 with information about the type, date, time, block name, neighbourhood name, and coordinates of the crime. As shown in Fig. 3, "Theft from Vehicle" is the most prevalent type of crime, following by "Mischief" and "Break and Enter Residential/Other". On the other hand, the crimes that are least likely to occur are homicide and vehicle collisions.

Fig. 4 shows the top 10 dangerous blocks based on records in the data-set. "OFFSET TO PROTECT PRIVACY" is the most dangerous block. Further inspection shows that all crime records with this block name have missing neighbourhood and coordinate data. This is because Vancouver police removed location information for all crimes labeled "Offence Against a

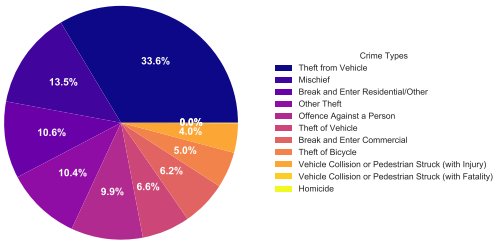

Fig. 3. Crime Type Distributions in Vancouver City.

Person" or "Homicide" to protect privacy. As a result, crime types analysed in this study do not include "Offence Against a Person" and "Homicide" for the lack of location data.

Also, The coordinates in the original data set are Universal

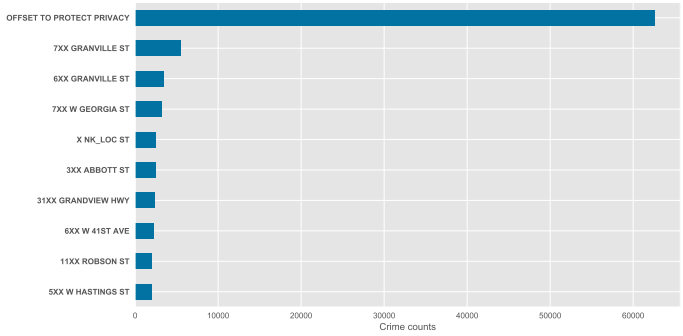

Fig. 4. Top 10 Dangerous Blocks in Vancouver City.

Transverse Mercator (UTM) coordinates and are converted to latitude and longitude for plotting. Finally, for the purpose of this study, time information is not important. Therefore, columns that represent the date and time of the crime are dropped. The first five rows of the data set after cleaning are shown in Table I.

Type	Neighbourhood	Latitude	Latitude
Break and Enter Commercial	Fairview	49.26667	-123.12902
Break and Enter Commercial	West End	49.28525	-123.12364
Break and Enter Commercial	West End	49.28518	-123.12353
Break and Enter Commercial	West End	49.28513	-123.12346
Break and Enter Commercial	West End	49.28513	-123.12346

TABLE I
FIRST FIVE ROWS OF CRIME DATA SET.

The last data set used in this study is the venue data set obtained from Foursquare Place API. Foursquare API returns a list of up to 50 venues around the specific location within the user defined radius. In this study, the searching radius is set to 1250 meters and the coordinates used are the coordinates in the neighbourhood data set. As shown in Fig 5, not all queries have received the required amount of information. This can be caused by either an untuned searching radius or the lack of data in the foursquare database. In order to eliminate the influence of searching radius, venue locations in five neighbourhoods with the least amount of data, namely Arbutus Ridge, Killarney, Victoria-Fraserview, Dunbar-Southlands, and West Point Grey, are plotted in Fig. 6. The visualization shows that a radius of 1250 meters is enough to cover the entire area without data overlap. Therefore, the number of venues is limited by the available data in foursquare database.

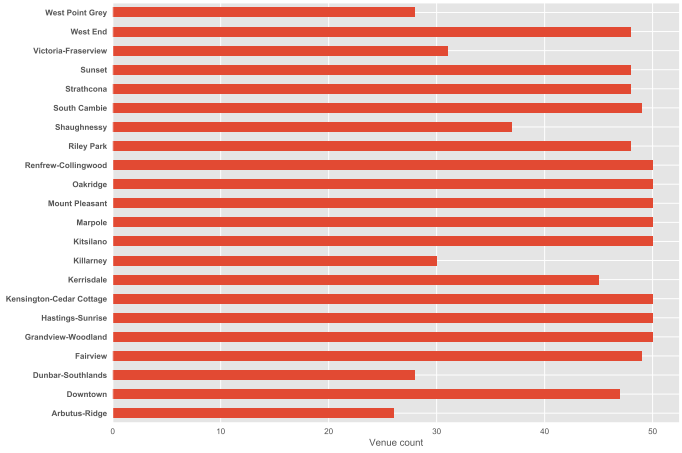

Fig. 5. Number of Venues in Each Neighbourhood Returned by Foursquare API.

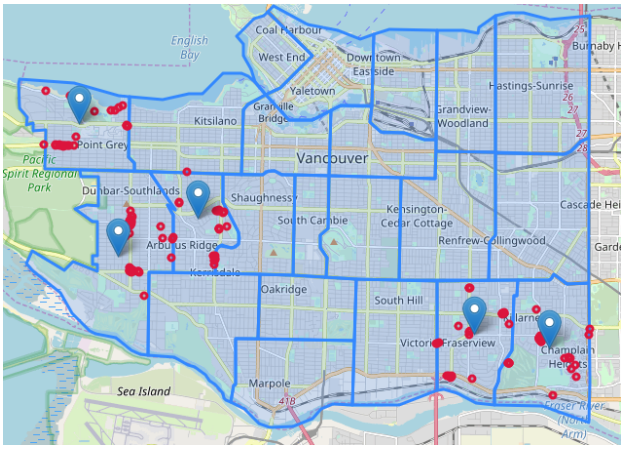

Fig. 6. Venue Distributions in Neighbourhoods with Less Data.

In order to discover the relationship between neighbourhood and crime rates, this study firstly find neighbourhoods with high similarities and cluster them in subgroups. After identifying all subgroups in the data set, the correlation coefficients between venue types and crime counts are calculated. Finally, the neighbourhood cluster with venues that are negatively correlated to the occurrence of crimes would be the ideal area for real estate investments.

K-means clustering

K-means algorithm is one of the most popular clustering algorithms because it is straight-forward and effective. The algorithm aims to partition the data set into k clusters in which a data point belongs to the cluster where the sum of squares distance between within-cluster points to the centroid of the

cluster is minimized. The objective function can be expressed as follows:

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{x \in S_i} \| x - \mu_i \|^2 \quad (1)$$

where S is the cluster of points; k is the total number of clusters; i is the cluster index; x is a data point within cluster i; and μ_i is the the mean of points in S_i or the centroid of cluster i.

Correlation Coefficient

The correlation coefficient, R, is a statistical measurement that calculates the strength of the relationship between two variables. The value of R ranges between -1.0 and 1.0. If $R = -1.0$, the two variables are negatively linearly correlated, meaning that one goes up when the other goes down and vice versa, whereas $R = 1.0$ indicates a perfect positive linear correlation, meaning that values of the two variables change in the same direction. A correlation of 0.0 means that the two variables are not correlated.

III. Results and Discussion

The Elbow Method is used to determine the optimal value of k into which the data may be clustered in k-means algorithm. As shown in Fig. 7, the optimal k is 5 based on the Calinski-Harabasz score. The result of k-means clustering using a k value of 5 is shown in Fig. 8 and Table II. The Top 10 common venues in each cluster are shown in Fig. 9. All clusters have different venue compositions, especially Cluster 4 which corresponds to the Downtown area.

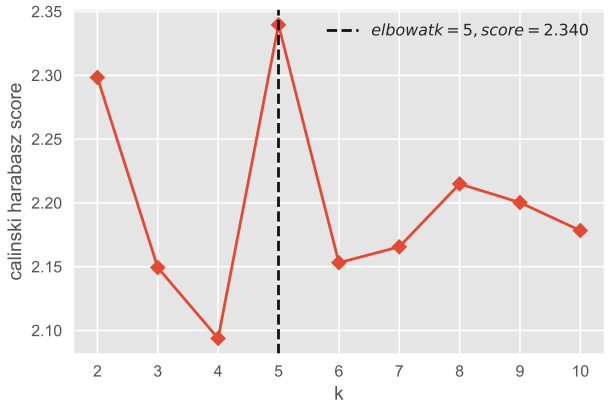

Fig. 7. The Calinski Harabasz Score Based Elbow for K-means Algorithm.

A choropleth map that colors the neighbourhoods in different shade based on the total number of crimes in that respective area is shown in Fig. 10. Areas with the highest number of crime counts are colored in red, whereas neighbourhoods with low crime rates are colored in yellow. Among all the crime types in the data set, Break and Enter Residential/Others has the largest impact on the real estate values. Therefore, a choropleth map based on the number of occurrence of Break and Enter Residential/Others is also presented (See Fig. 11). Compared to the choropleth map of overall crime counts, the one for Break and Enter Residential/Others has different trends. For example, Downtown area is the most dangerous area overall, but houses and apartments there are generally safer to live in than places in Kitsilano, Grandview-Woodland, Kensington-Cedar Cottage, and Renfrew-Collingwood. Furthermore, despite having low crime rates overall, about half of the neighbourhoods in Cluster 1 and 3 have higher rates for Break and Enter Residential/Others to occur. On the other hand, all neighbourhoods in Cluster 2 remains on the lower end of the spectrum.

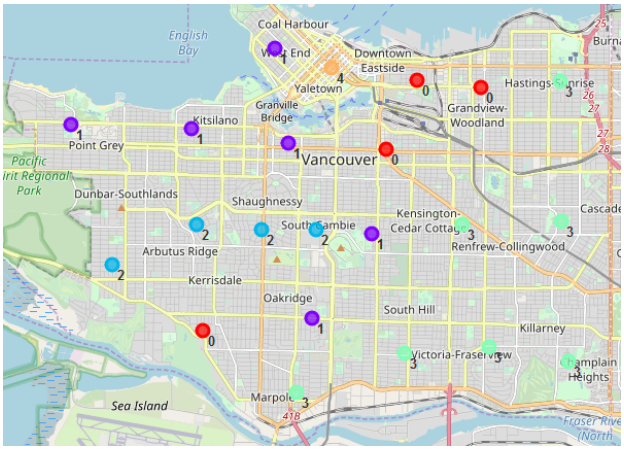

Fig. 8. Clustering result of Neighbourhoods in Vancouver City.

Cluster	Neighbourhoods
0	Grandview-Woodland, Kerrisdale, Mount Pleasant, Strathcona
1	Fairview, Kitsilano, Oakridge, Riley Park, West End, West Point Grey
2	Arbutus-Ridge, Dunbar-Southlands, Shaughnessy, South Cambie
3	Hastings-Sunrise, Kensington-Cedar Cottage, Killarney, Marpole, Renfrew-Collingwood, Sunset, Victoria-Fraserview
4	Downtown

TABLE II
Neighbourhoods in Each Cluster

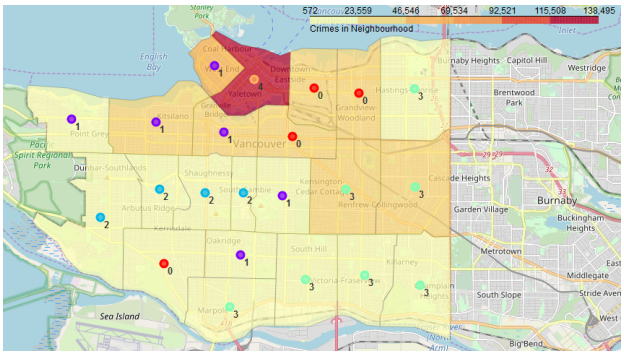

Fig. 10. The Crime Choropleth Map.

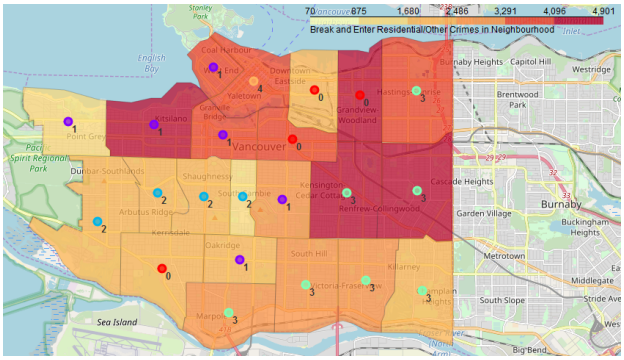

Fig. 11. The Choropleth Map of Break and Enter Residential/Others.

Finally, the correlation coefficients between all crime types and common venue categories are calculation and the correlation matrix is presented in Fig. 12 where green indicates the positive correlation and red indicates the negative correlation. If correlation coefficient is close to zero, the tile is colored yellow, meaning the two variables are not linearly correlated. Generally, if the absolution value of correlation coefficient is below 0.6, the collinearity among the predictor variables is considered weak. Using this threshold, the only venue category that has a strong relationship with crimes is Hotel which is positively correlated to the occurrence of all types of crime analysed in this study except for Break and Enter Residential/Other and Theft of Vehicle.

IV. Summary

In the present work, neighbourhoods in Vancouver are divided to five clusters based on the types of venues located within them. Combining the clustering results with Vancouver crime data, neighbourhoods in Cluster 2, namely Arbutus-Ridge, Dunbar-Southlands, Shaughnessy, and South Cambie, are safer areas to live and therefore, would be recommended to real estate companies and individual house buyers. On the other hand, there is little to no linear correlation between venue categories and crime types with one exception being Hotel. However, the results of this study are strongly dependent on the quality of venue and crime data set which is inadequate for some neighbourhoods and some types of crimes, respectively. Furthermore, Foursquare venue database is being updated frequently and the changes can affect the result of clustering. Therefore, future work should focus on improving venue and crime data set, possibly by combining the current data with data from other sources.

References

[1] R. J. Sampson, S. W. Raudenbush, and F. Earls, "Neighborhoods and violent crime: A multilevel study of collective efficacy," *Science*, vol. 277, no. 5328, pp. 918–924, 1997.

[2] R. D. Dietz, "The estimation of neighborhood effects in the social sciences: An interdisciplinary approach," *Social science research*, vol. 31, no. 4, pp. 539–575, 2002.

[3] J. R. Kling, J. Ludwig, and L. F. Katz, "Neighborhood effects on crime for female and male youth: Evidence from a randomized housing voucher experiment," *The Quarterly Journal of Economics*, vol. 120, no. 1, pp. 87–130, 2005.

[4] K. Ihlanfeldt and T. Mayock, "Panel data estimates of the effects of different types of crime on housing prices," *Regional Science and Urban Economics*, vol. 40, no. 2-3, pp. 161–172, 2010.

[5] *Local area boundary*, City of Vancouver. [Online]. Available: https://opendata.vancouver.ca/explore/dataset/local-area-boundary/information/.

[6] *Open government licence – vancouver*, version 2.0, City of Vancouver. [Online]. Available: https://opendata.vancouver.ca/pages/licence/.

[7] *Vancouver police department (vpd) crime data*, Vancouver Police Department. [Online]. Available: https://geodash.vpd.ca/opendata/.

[8] *Places api*, Foursquare. [Online]. Available: https://developer.foursquare.com/places.

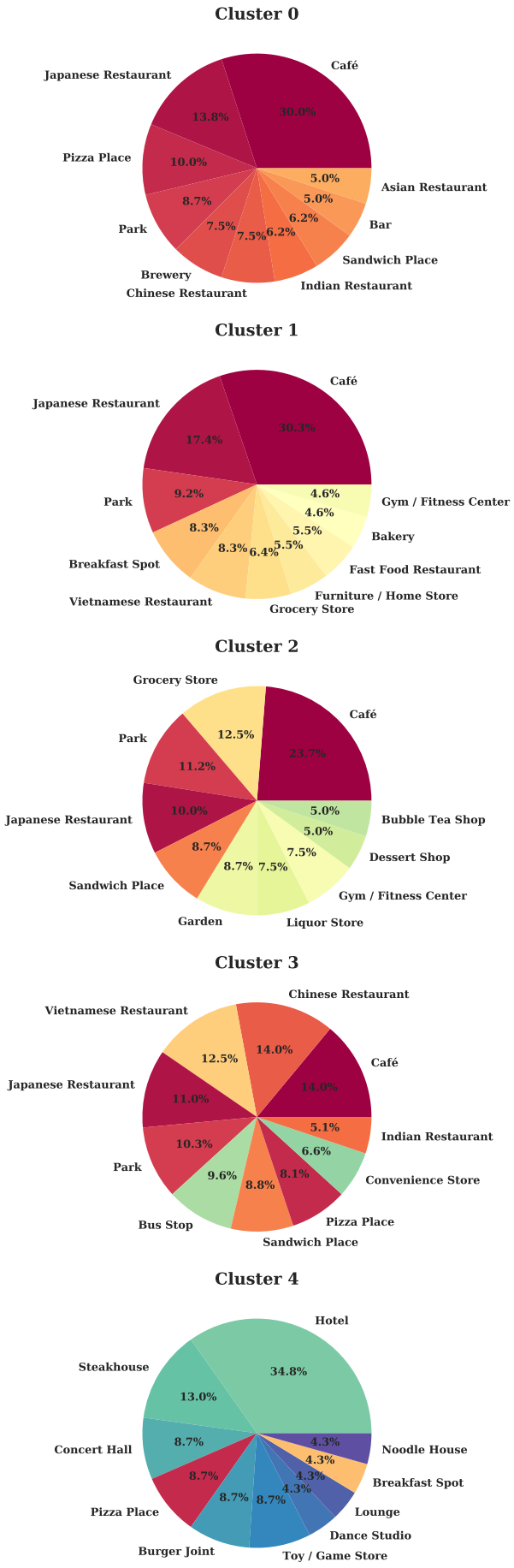

Fig. 9. Top 10 Common Venues in Each Cluster.

Fig. 12. Correlation Coefficients between Crime Types and Common Venue Categories.