FACULTY OF ENGINEERING AND APPLIED SCIENCE

SOFE 3720U

Introduction to Artificial Intelligence.

Course Instructor: Dr. Masoud Makrehchi

Submission date: April 11th, 2022

Names	Student ID
Vidurshan Sribalasuhabiramam	100558257
Tasfia Alam	100584647
Sarah Long	100067484

Github: https://github.com/SOFE3720Grp11/AI-Project

(The Jupyter Notebook file is on github^)

Introduction: Background and Business Problem

In this project, we decided to create a program to analyze the data patterns between the number of crimes related to theft and income levels in the various neighborhoods in Toronto. We will be finding the information to correlate the data with the income of households and the crime rates in those neighborhoods in regard to robberies, automobile theft, and theft over $5000 in those neighborhoods. We would like to explore the correlation between household income levels and crimes related to theft in order to determine whether lower income areas are more prone to theft and robberies than higher income areas.

The results of this data analysis can be used by governmental organizations and businesses to drive business decisions such as the level of security needed in an area or what frameworks can be placed in order to reduce crime and protect the public.

Data explanation and data sources

The data used for the project was collected from numerous sources as mentioned below. These datasets were used because they correspond to the business problem we are exploring.

Neighborhood Data:

The data for the postal codes, neighbourhood names, and borough names were collected from the Wikipedia page: List_of_postal_codes_of_Canada:_M. This data was used to parse into Foursquare API and used to retrieve the longitude and latitude of each postal code.

Wikipedia: List_of_postal_codes_of_Canada:_M

- Postal codes
- Neighborhood names
- Boroughs

Income Statistics for Households:

The household income after taxes for the year of 2015 was acquired from the Open-Data Toronto's Neighborhood Profiles. The following characteristics from the Neighbourhood Profiles were considered in terms of income statistics for each neighbourhood.

Under $5,000
$5,000 to $9,999
$10,000 to $14,999
$15,000 to $19,999
$20,000 to $24,999
$25,000 to $29,999
$30,000 to $34,999
$35,000 to $39,999
$40,000 to $44,999
$45,000 to $49,999
$50,000 to $59,999
$60,000 to $69,999
$70,000 to $79,999
$80,000 to $89,999
$90,000 to $99,999
$100,000 to $124,999
$125,000 to $149,999

Crime Statistics:

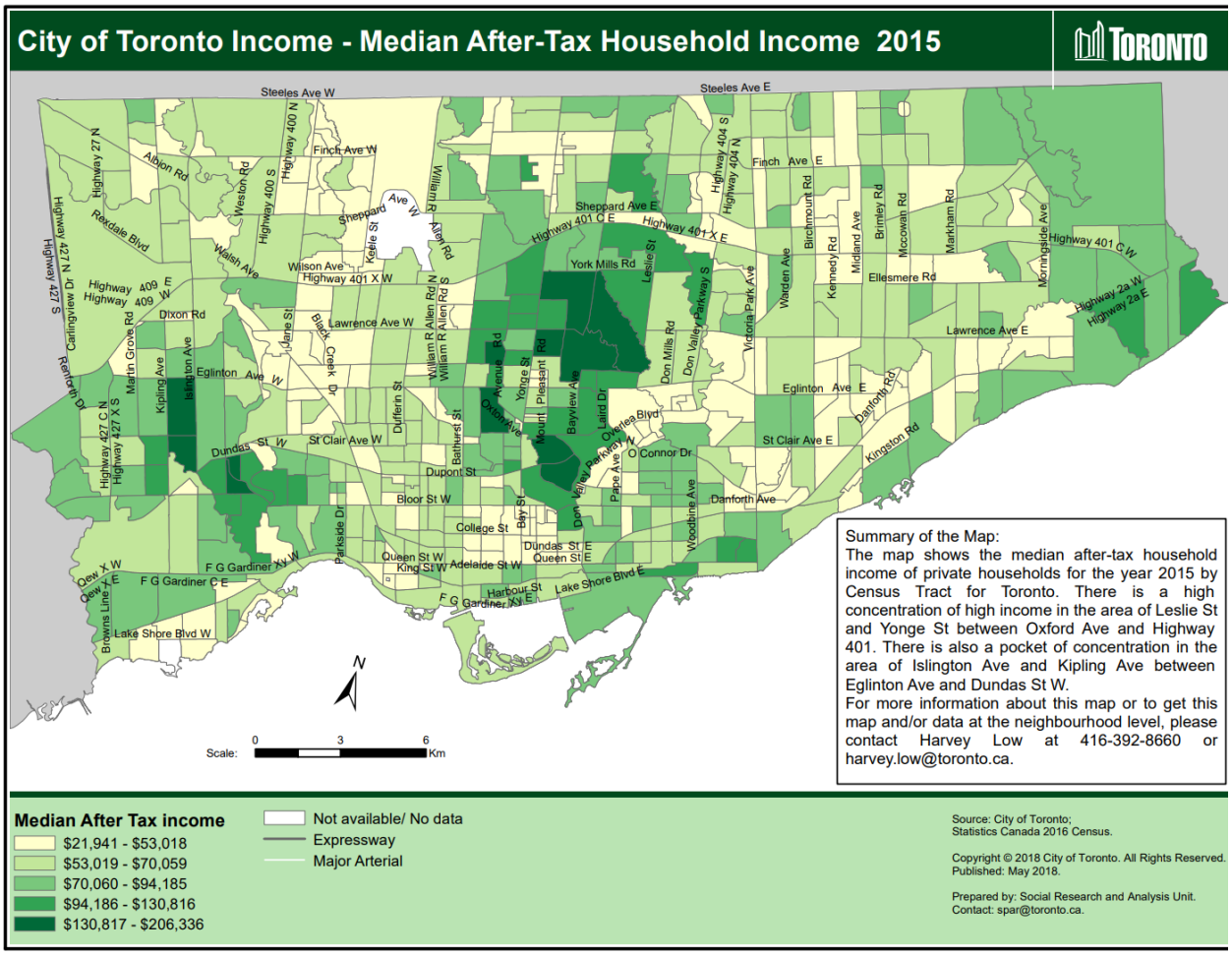
The crime statistics were collected from the Neighbourhood Crime Rates from the Toronto Police data portal. The following three crime stats for the year 2015 were used in our data analysis.

- Robberies
- Auto theft
- Theft over $5000

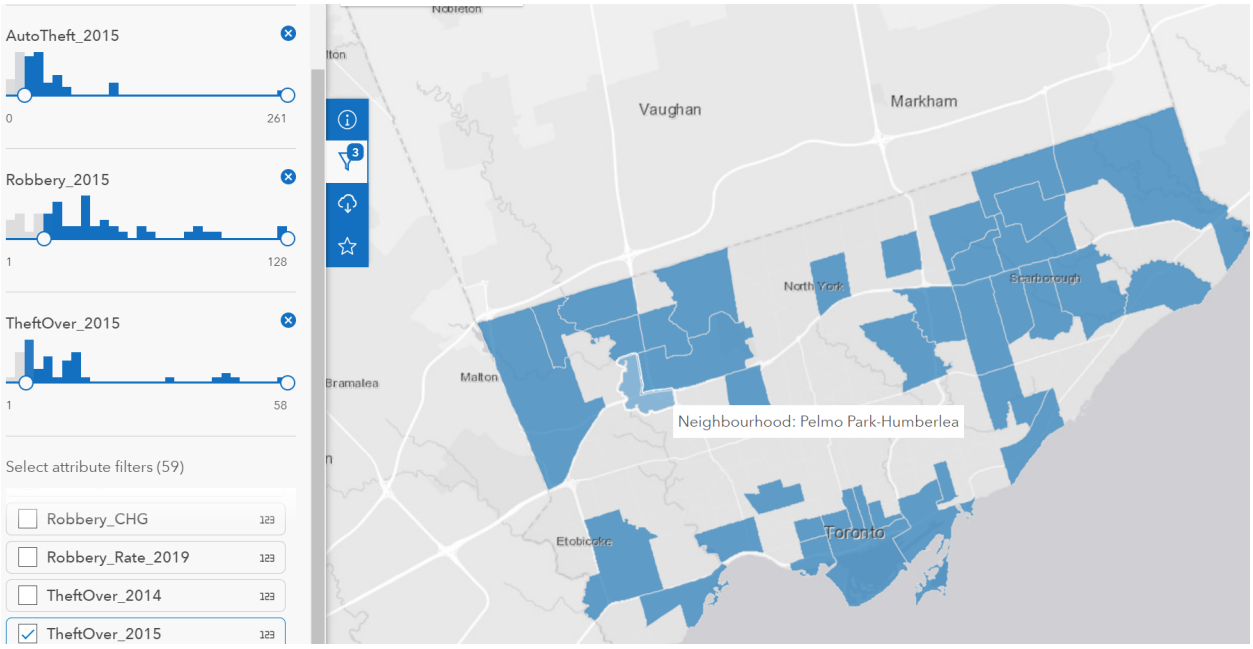
These three crime statistics were specifically chosen because they have monetary value, which can be related to income levels of households.

Methodology

We determined that k-means clustering would be the best method to cluster neighbourhoods in terms of median after-tax household income levels and crime rates. Below is a sample heatmap of the median after-tax household income for 2015 that shows what neighborhoods are in which income bracket.



The crime frequency for auto-theft, robberies, and theft over $5000 are shown below in a sample heatmap. The highlighted neighbourhoods are those with a frequency of any of the three crimes above the city of Toronto's median crime rate for each category.

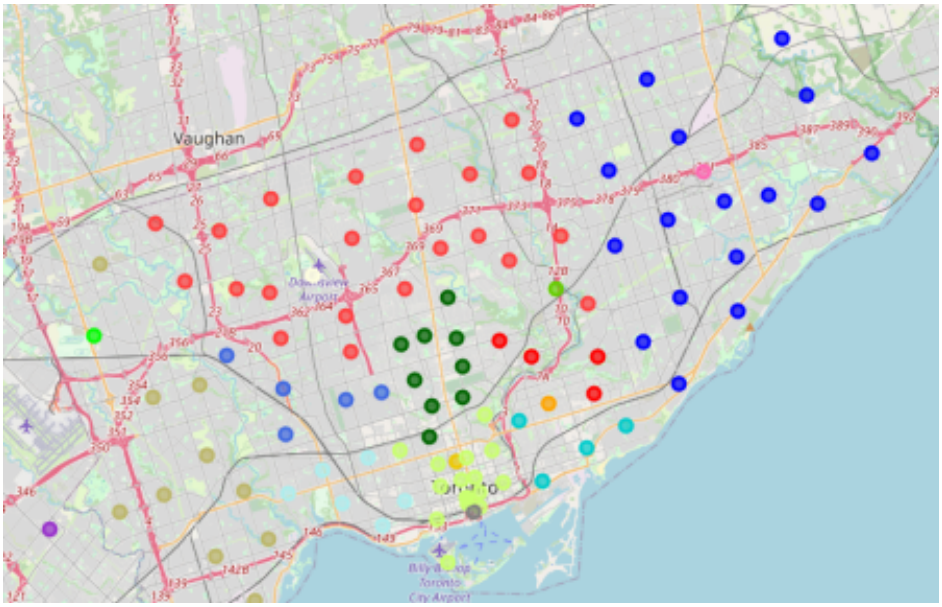


Our two variables for the K-means clustering would be the median after-tax income level for each neighbourhood and the sum of the crime rates for each neighbourhood would be our other variable. Each neighbourhood's data point in relation to the next would be determined using the Euclidean distance between each other. K-means clustering would cluster the data points based on the minimum variance between points, so as to minimize the within-cluster sum of squares (WCSS).

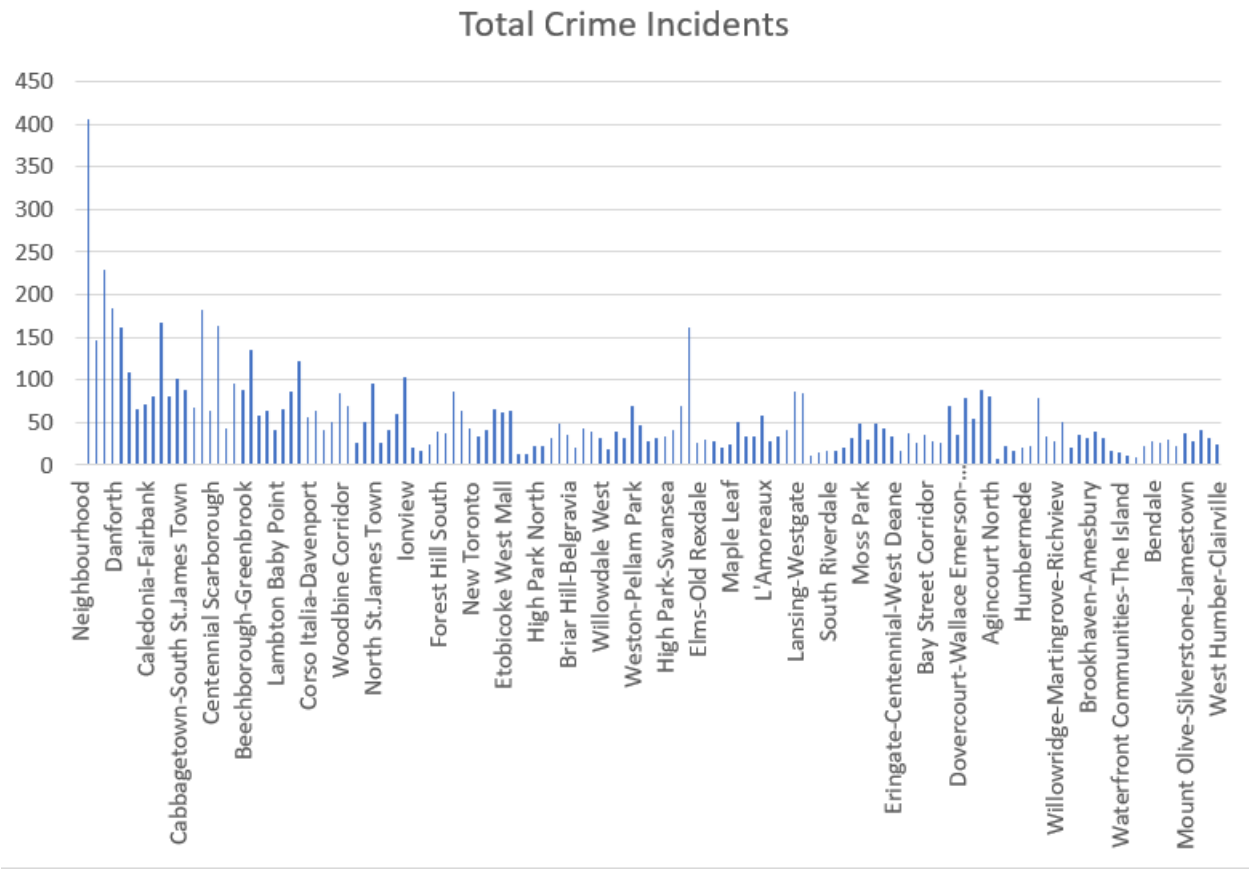
$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2 = \arg\min_{\mathbf{S}} \sum_{i=1}^{k} |S_i| \operatorname{Var} S_i$$

Results

Firstly, we clustered the neighborhood data with different colored dots representing each borough in Toronto. Below is a chart depicting which colors refer to which boroughs.



Color	Borough
Brown	Central Toronto
Lime	York
Grey	Downtown TorontoStn/ The Esplanade
Teal	Scarborough
Pink	East Toronto
Red	Etobicoke Northwest
Orange	North York(Don Mills)South
Tan	Mississauga/Canada Post Gateway Processing Centre
Bright Green	West Toronto
Purple	Etobicoke
Pale Red	Downtown Toronto
Lime Green	Queen's Park
Pale Yellow	East York
Yellow-Orange	North York
Royal Blue	East Toronto/Business reply mail Processing Centre 969 Eastern
Periwinkle	East YorkEast Toronto
Dark Green	North York(Downsview)East



Graph of total crime incidents for all Neighborhoods

Since we plan to use both crime statistics and income sources, we will be using K-means clustering to analyze our data. Our K-value is 4 as we determined that is the best number of clusters to segment the neighbourhoods crime data. Based on the incident rates shown above, we can see that there are distinct variances in crime rates and income levels.

Discussion and Conclusion

In conclusion, neighbourhoods with lower median after-tax income had higher rates of crime. Even though it would be expected that higher income neighborhoods would be the victim of theft and robbery crimes, the analysis proved otherwise. K-means clustering provided the means to analyze the data and explore the correlation between income levels and crime rates.