Faculty of Engineering & Applied Science

SOFE 3720U

Introduction to Artificial Intelligence

Final Project

GROUP 35

Submission date: April 11, 2021

Name	**Student ID**
Ashad Ahmed	100745913
Milan Saju Samuel	100757350
Shahroze Butt	100701891

Introduction

Our group is looking into the subject of crime rates and revenue sources because we wanted to discover whether there was a link between the two. Both of these variables have had a significant impact on society, owing to the fact that money is a prevalent issue. When people are short on cash, they may resort to a variety of methods to generate revenue. We wanted to determine which districts in Toronto had the greatest crime rates in relation to the money or income of the neighborhoods. The information we gathered came from a variety of sources that we discovered.

The amount of money given to people has a direct impact on their quality of life and happiness. When the root of the problem is money, and how money is a common factor in problems. It then generates problems inside families and children as they try to find a means to make money, or it produces problems within households. When children are continuously confronted with situations that they should not be exposed to, they grow quicker than they should in order to catch up and help around the house. As a result, children have challenges with not knowing what they're doing and why they're doing it, as well as being money hungry. When youngsters are finally shown a fraction of what a healthy and happy existence looks like, they will desire to achieve it in some way. When everything that surrounds them is incessant criminality and no acceptable role models, children are forced to make a choice between two options, one of which is clearly more viable than the other.

When we analyzed the data from various sources, we discovered a link between high crime rates and low-income neighborhoods. There were higher crime rates in regions where there was a lot of government housing because the residents' income was low, lowering their quality of life. However, there are many more factors at play than just money when it comes to crime rates. They go hand in hand, but there are many more ancillary concerns that result in higher crime rates in some places than others. However, we are only concerned with the effects of the crime rate on revenue sources.

Data explanation and data source

To handle the data for the case we're working on, we had to install the necessary libraries, such as pandas. It gives the system the ability to read longitudinal and latitude coordinates as well as the postal code. Then we had to import the names of the neighborhoods, their boroughs, and their postal codes. However, integrating the two datasets and having only one column for postal code instead of the two differing postal codes, was more of a challenge. Furthermore, we began to implement the sources holding the data we need for the neighborhood crime data collection. The table had gained more information as a result of this; however, it had to be integrated with the neighborhood in order for the information for each selected neighborhood to be displayed.

We investigated: Total - Income Statistics for the Population Aged 15 and Over in Private Households in 2015. Then we had to define instances for what we would consider to be inside which category, as well as create similarity limits for the API to create clusters and label them under. We wanted to demonstrate where robberies have happened in the past and how likely they are to happen again. We took the robbery value from 2015 and divided it into five categories: Very Unlikely, Unlikely, Maybe, Likely, and Very Likely. We created the limitations after tracking the values:

Very Likely: 90-100+

Likely: 61-90

Maybe: 41-60

Unlikely: 11-40

Very Unlikely: 0-10

Following the creation of the map, it was discovered that there was no significant relationship between the cause and effect of robbery and household income in Toronto.

Methodology

Many external libraries were utilized to create and augment the data for this Python project. Since it turned raw data into a simple and usable data frame, Pandas was the most crucial library. BeautifulSoup was used to scrape and extract information from the html website because much of the data was provided as a comma-separated values file. In addition, the Folium Library was used to produce coloured markers that denoted the severity of each neighborhood's robberies. The colors used were red, green, yellow, blue, and black, with red, green, yellow, blue, and black representing increasing levels of robbery.

Results

Installing and importing necessary libraries

Libraries that were installed to run the code

- It gives the system the ability to read longitudinal and latitude coordinates as well as the postal code.

```
!pip install beautifulsoup4
!pip install lxml
!pip install folium

#it is a library for data analysis
import pandas as pd
#it tranforms the json file into a panda dataframe library
from pandas.io.json import json_normalize
#library that handles all the data within vectorized manner
import numpy as np
import json
#libraray that handkes requests

import requests
#it helps plot libararies
import folium
import requests
#it is a library that displays images
from IPython.display import display_html
from IPython.core.display import HTML
from bs4 import BeautifulSoup
```

```
Requirement already satisfied: beautifulsoup4 in c:\programdata\anaconda3\lib\site-packages (4.10.0)
Requirement already satisfied: soupsieve>1.2 in c:\programdata\anaconda3\lib\site-packages (from beautifulsoup4) (2.2.1)
Requirement already satisfied: lxml in c:\programdata\anaconda3\lib\site-packages (4.6.3)
Requirement already satisfied: folium in c:\programdata\anaconda3\lib\site-packages (0.12.1.post1)
Requirement already satisfied: jinja2>=2.9 in c:\programdata\anaconda3\lib\site-packages (from folium) (2.11.3)
Requirement already satisfied: branca>=0.3.0 in c:\programdata\anaconda3\lib\site-packages (from folium) (0.4.2)
Requirement already satisfied: requests in c:\programdata\anaconda3\lib\site-packages (from folium) (2.26.0)
Requirement already satisfied: numpy in c:\programdata\anaconda3\lib\site-packages (from folium) (1.20.3)
Requirement already satisfied: MarkupSafe>=0.23 in c:\programdata\anaconda3\lib\site-packages (from jinja2>=2.9->folium) (1.
1.1)
Requirement already satisfied: charset-normalizer~=2.0.0 in c:\programdata\anaconda3\lib\site-packages (from requests->foliu
m) (2.0.4)
Requirement already satisfied: idna<4,>=2.5 in c:\programdata\anaconda3\lib\site-packages (from requests->folium) (3.2)
Requirement already satisfied: certifi>=2017.4.17 in c:\programdata\anaconda3\lib\site-packages (from requests->folium) (202
1.10.8)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in c:\programdata\anaconda3\lib\site-packages (from requests->folium)
(1.26.7)
```

Reading the geospatial coordinates CSV file's data and

Postcode, Latitude, and Longitude

-It reads the Geospatial_Coordinates and outputs the postal code, along with the longitude and the latitutide.

In [5]:
```
df1 = pd.read_csv('Geospatial_Coordinates.csv')
df1
```

Out[5]:

	Postcode	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917
4	M1H	43.773136	-79.239476
...	...	...	...
98	M9N	43.706876	-79.518188
99	M9P	43.696319	-79.532242
100	M9R	43.688905	-79.554724
101	M9V	43.739416	-79.588437
102	M9W	43.706748	-79.594054

103 rows × 3 columns

Scraping the Wikipedia source page

Using the Wikipedia Page to create Data

-Using the Wikipedia page provided to create a dataset while assuring that values aren't being duplicated but rather removed if there is a duplicate available.

In [6]:
```
source = requests.get('https://en.wikipedia.org/w/index.php?title=List_of_postal_codes_of_Canada:_M&oldid=945633050').text
soup = BeautifulSoup(source,'lxml')
table = str(soup.table)
dfs = pd.read_html(table)
df = dfs[0]

# Removing the rows where Neighbourhood is 'Not assigned'
df0 = df[df.Neighbourhood != 'Not assigned']

#Drop neighbourhood names with same Postcode
df2 = df0.drop_duplicates(subset=['Postcode'])
df2.reset_index(inplace=True)
df2 = df2.drop('index', 1)
df2
```

Out[6]:

	Postcode	Borough	Neighbourhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Harbourfront
3	M6A	North York	Lawrence Heights
4	M7A	Downtown Toronto	Queen's Park
...	...	...	...
98	M8X	Etobicoke	The Kingsway
99	M4Y	Downtown Toronto	Church and Wellesley
100	M7Y	East Toronto	Business Reply Mail Processing Centre 969 Eastern
101	M8Y	Etobicoke	Humber Bay
102	M8Z	Etobicoke	Kingsway Park South West

Merging the two datasets

Datasets being Merged

-After having the Geospatial_coordinates we combine both the tables that we had created along with the postal codes and created the table below.

In [7]:
```
df3 = pd.merge(df1,df2,on = 'Postcode')
df3
```

Out[7]:

	Postcode	Latitude	Longitude	Borough	Neighbourhood
0	M1B	43.806686	-79.194353	Scarborough	Rouge
1	M1C	43.784535	-79.160497	Scarborough	Highland Creek
2	M1E	43.763573	-79.188711	Scarborough	Guildwood
3	M1G	43.770992	-79.216917	Scarborough	Woburn
4	M1H	43.773136	-79.239476	Scarborough	Cedarbrae
...	...	...	...	...	...
98	M9N	43.706876	-79.518188	York	Weston
99	M9P	43.696319	-79.532242	Etobicoke	Westmount
100	M9R	43.688905	-79.554724	Etobicoke	Kingsview Village
101	M9V	43.739416	-79.588437	Etobicoke	Albion Gardens
102	M9W	43.706748	-79.594054	Etobicoke	Northwest

103 rows × 5 columns

Reading and Merging Neighborhood Crime Rates CSV file data

Merging the Neighbourhood Crime Rate Dataset

-They called for the csv link for the Neighbourhood Crime Rate and merging it with the table.

[8]:
```
#calling for the file neighbourhood crime rate to then analyze and read the dataset
df4 = pd.read_csv('Neighbourhood_Crime_Rates.csv')
#we had to merge the dataset after the continuation with the neighbourhood
df5 = pd.merge(df3,df4,on = 'Neighbourhood')
df5
```

Out[8]:

rhood	OBJECTID	Hood_ID	Population	Assault_2014	Assault_2015	...	TheftOver_2015	TheftOver_2016	TheftOver_2017	TheftOver_2018	TheftOver_2019
Rouge	98	131	46496	177	167	...	8	13	11	16	13
Creek	72	134	12494	49	50	...	5	1	1	3	1
dwood	37	140	9917	52	38	...	2	1	4	1	1
Voburn	112	137	53485	352	395	...	13	14	23	13	8
orough Village	15	139	16724	161	153	...	2	2	4	3	2
iffcrest	25	123	15935	79	97	...	5	3	7	3	3
et Park	64	126	25003	135	168	...	10	12	4	15	8
t North	80	129	29113	67	76	...	5	5	11	4	2
Village	62	48	16934	63	59	...	5	2	4	7	6
Village	114	52	21396	79	104	...	8	5	7	7	13
e West	97	37	16936	86	116	...	7	4	4	6	10
ingdon Park	26	44	21933	128	147	...	3	3	4	1	3
Manor	86	34	15873	42	38	...	2	4	4	3	5
Village	39	43	17510	118	138	...	6	4	5	4	5
eaches	7	63	21567	83	108	...	5	3	7	5	9
e Park	9	55	21108	86	105	...	6	11	9	3	5
l North	34	102	12806	24	28	...	4	2	3	0	2
ewood-arvale	11	106	14365	43	52	...	3	5	6	7	4

Reading and merging Neighborhood Profiles CSV file data

Neighbourhood Data

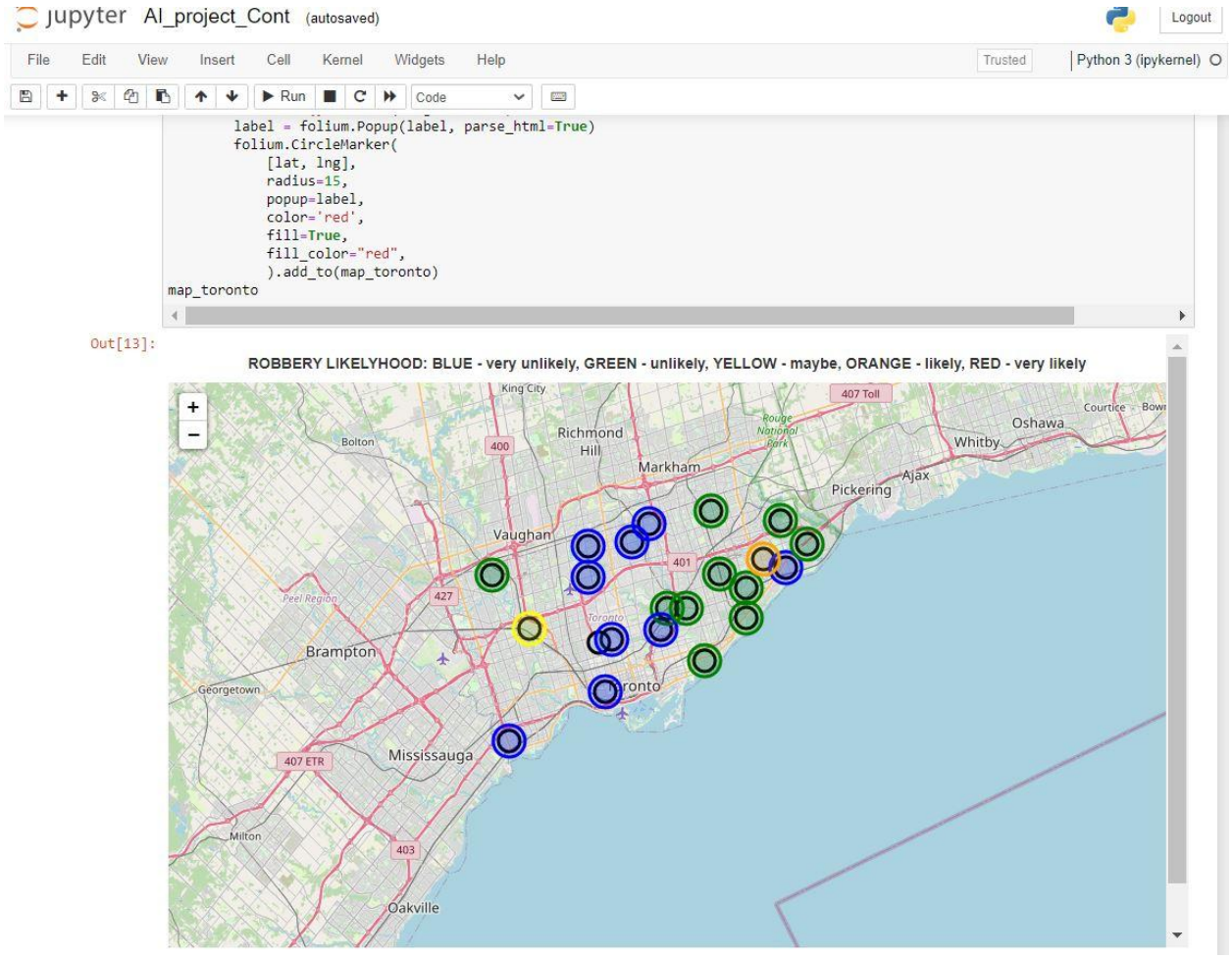
-It provides detailed information about each neighborhood within Toronto.

In [8]:
```
# it calls and reads for the neighbourhood profile
df6 = pd.read_csv('neighbourhood-profiles.csv')
#we had to merge the dataset after the continuation with the neighbourhood
df7 = pd.merge(df6,df5,on = 'Neighbourhood')
df7['Total - Income statistics in 2015 for the population aged 15 years and over in private households'] = pd.to_numeric(df7[
df7
```

Out[8]:

	Neighbourhood	Neighbourhood Number	TSNS2020 Designation	Population, 2016	Population, 2011	Population Change 2011-2016	Total private dwellings	Private dwellings occupied by usual residents	Population density per square kilometre	Land area in square kilometres	...	TheftOver_2015
0	Agincourt North	129	No Designation	29,113	30,279	-3.90%	9,371	9,120	3,929	7.41	...	5
1	Alderwood	20	No Designation	12,054	11,904	1.30%	4,732	4,616	2,435	4.95	...	3
2	Bathurst Manor	34	No Designation	15,873	15,434	2.80%	6,418	6,089	3,377	4.7	...	2
3	Bayview Village	52	No Designation	21,396	17,671	21.10%	10,111	9,532	4,195	5.1	...	8
4	Cliffcrest	123	No Designation	15,935	15,703	1.50%	6,094	5,902	2,273	7.01	...	5
5	Dorset Park	126	Emerging Neighbourhood	25,003	24,363	2.60%	8,995	8,777	4,146	6.03	...	10
6	Flemingdon Park	44	NIA	21,933	22,168	-1.10%	7,964	7,830	9,026	2.43	...	3
7	Forest Hill North	102	No Designation	12,806	12,474	2.70%	5,784	5,446	8,054	1.59	...	4
8	Guildwood	140	No Designation	9,917	9,816	1.00%	4,044	3,991	2,673	3.71	...	2
9	Highland Creek	134	No Designation	12,494	13,097	-4.60%	3,907	3,700	2,403	5.2	...	5
10	Hillcrest Village	48	No Designation	16,934	17,656	-4.10%	6,642	6,398	3,148	5.38	...	5
11	Humber Summit	21	NIA	12,416	12,525	-0.90%	4,288	3,897	1,570	7.91	...	16
12	Humewood-Cedarvale	106	No Designation	14,365	14,108	1.80%	6,865	6,566	7,682	1.87	...	3
13	Little Portugal	84	No Designation	15,559	12,050	29.10%	8,095	7,427	12,859	1.21	...	1
14	Rouge	131	No Designation	46,496	45,912	1.30%	13,730	13,389	1,260	36.89	...	8
15	Scarborough	139	NIA	16,724	16,609	0.70%	6,133	5,923	5,395	3.1		2

Visualizing complete clustered dataset with Foilum map



Discussion and Conclusion

Overall, we have presented and understood how the web-scraping and clustering works and how it assists users in determining latitude, longitude, and venues as a group. We overcome numerous challenges in getting everything to function and comprehending how it all works together. We gained a deeper and more in-depth grasp of how clustering works and how it aids consumers in comprehending what's going on. We learned and comprehended how the association of robberies and income correlates with one another using the numerous approaches we learned. We delved deeper into Toronto and what was going on in each city during the year of 2015.

We were also able to supply the cities, as well as the postal codes, city names, crime statistics, and other information. We were able to allow our system to interpret and understand the data in order to produce various datasets, merge them, and create a map. With all of the data we gathered, we determined there isn't much of a link because there are so many other variables that could influence a decision. To have a lot of robberies, it would take more than simply a high household income, but it could be a contributing cause. That is why, by learning how to use this tool, we will be able to speed our notion and discover the broader logic for why robberies occur rather than focusing on a single incident.