Introduction to the opportunity

Toronto, the capital of the province of Ontario, is a major Canadian city along Lake Ontario's northwestern shore. It's a dynamic metropolis with a core of soaring skyscrapers, all dwarfed by the iconic, free-standing CN Tower. Toronto also has many green spaces, from the orderly oval of Queen's Park to 400-acre High Park and its trails, sports facilities and zoo.

As the city rapidly grows and develops in all sectors, it becomes most important to examine and understand it's growth in all the directions with all the possible possibilities. The AAP government of Delhi City of provides free education, medical facilities and some electricity to ever family and encourages entrepreneurial use to develop services for the benifit of its citizens.

But all the developers, investors, policy makers and/or city planners are always keen to know answers of the following questions:

1. What is crime rate?
2. What is population density correlated to crime level?
3. Can we get the data like what venues are most common in different locations within the city?
4. Does city has trained man power for their business?
5. Whether coffee shop is really needed in the Knowledge Park really? Etc…

Data

To understand and explore we will need the following Data of Toronto City

1. Open Data Site: https://open.toronto.ca/catalogue/?sort=score%20desc

   Here we get all the required data about TORONTO City.

2. Toronto Neighbourhood:

   Population Total (2016 Census Data)

   The data refers to Total Population from the 2016 Census, aggregated by the City of Toronto to the City's 140 Neighbourhood Planning Areas. Although Statistics Canada makes a great effort to count every person, in each Census a notable number of people are left out for a variety of reasons. For Census 2016: Population and Dwellings example, people may be travelling, some dwellings are hard to find, and some people simply refuse to participate. Statistics Canada takes this into account and for each Census estimates a net 'undercoverage' rate for the urban region, the Toronto Census Metropolitan Area (CMA), but not for the city. The 2011 rate for the Toronto CMA was 3.72% plus or minus 0.53%. The 2016 rate is not yet available.

   Data is present into csv file named : Neighbourhoods.csv

   Web Address to download CSV file:
   https://open.toronto.ca/dataset/neighbourhoods/

3. Toronto Crime by Neighbourhood:

   This dataset contains three worksheets. The full description for each column of data in worksheets two and three is available in the first worksheet called "IndicatorMetaData". The data was provided by Toronto's Police Services, Fire Services, Paramedic Services (formerly EMS) and Community Housing Corporation. Refer to the descriptions in worksheet 1 for more information.

Users should note that the data for each neighbourhood are based on the mathematical aggregation of smaller sub-areas (in this case Census Tracts) that when combined, define the entire neighbourhood. Since smaller areas may have their values rounded or suppressed (to abide by Statistics Canada privacy standards), the overall total may be undercounted.

Data is present into xlsx file named : wellbeing-toronto-safety.xlsx

Web Address to download CSV file : https://open.toronto.ca/dataset/wellbeing-toronto-safety/

4. Toronto Census Tract Demographics:

The Census of Population is held across Canada every 5 years and collects data about age and sex, families and households, language, immigration and internal migration, ethnocultural diversity, Aboriginal peoples, housing, education, income, and labour. City of Toronto Neighbourhood Profiles use this Census data to provide a portrait of the demographic, social and economic characteristics of the people and households in each City of Toronto neighbourhood. The profiles present selected highlights from the data, but these accompanying data files provide the full data set assembled for each neighbourhood.

For more information, visit the Neighbourhood Profiles webpage.

In these profiles, "neighbourhood" refers to the City of Toronto's 140 social planning neighbourhoods. These social planning neighbourhoods were developed by the City of Toronto to help government and community organizations with local planning by providing socio-economic data at a meaningful geographic area. The boundaries of these social planning neighbourhoods are consistent over time, allowing for comparison between Census years. Neighbourhood level data from a variety of other sources are also available through the City's Wellbeing Toronto mapping application and here on the Open Data portal.

Each data point in this file is presented for the City's 140 neighbourhoods, as well as for the City of Toronto as a whole. The data is sourced from a number of Census tables released by Statistics Canada. The general Census Profile is the main source table for this data, but other Census tables have also been used to provide additional information.

For definitions of terms and concepts referenced in this data set, users should consult the reference materials produced by Statistics Canada for the 2016 Census, available online at: http://www12.statcan.gc.ca/census-recensement/2016/ref/index-eng.cfm.

PLEASE NOTE: Statistics Canada does not release data at the level of Toronto's social planning neighbourhoods. Neighbourhood level data for 2016 are initially calculated by summing data for the Census Tracts which comprise each neighbourhood. Statistics Canada's random rounding reporting practices may have a compounded effect on the totals. These figures should be interpreted with caution where there are relatively few observations, such as for 5 year age groups by sex, or language groups with a small number of speakers. Certain values such as median values (and in some cases percentages and means) cannot yet be calculated at the neighbourhood level from available data. Where possible, these data will be updated when custom data at the neighbourhood level has been acquired from Statistics Canada.

Data is present into csv file named : neighbourhood-profiles-2016.csv

Web Address to download CSV file : https://open.toronto.ca/dataset/neighbourhood-profiles/

5. Toronto locations of interest:

Provide a listing of attractions in Toronto, as determined by Visitor Services. This is not necessarily a comprehensive listing, and will continue to expand over time. This information has been gathered through research and day-to-day operations of the Visitor Services Department

To suggest inclusion of an additional locations or for any comments or update requests to existing comment please email visitorservices@toronto.ca

Please note: This data was originally accompanied by a tabular file which needed to be removed as it included information intended for internal discussion purposes only and was created by the subjective opinion of City of Toronto staff. As this discussion information did not use scientific/methodical evidence to determine the values, the publishing of it would mislead the reader.

Data is present into csv file named : Places of Interest and Attractions.csv

Web Address to download CSV file : https://open.toronto.ca/dataset/places-of-interest-and-toronto-attractions/

6. Foursquare Developers Access to venue data: https://foursquare.com/
(https://foursquare.com/)

Using this data will allow exploration and examination to answer the questions. The neighbourhood data will enable us to properly group crime by neighbourhood. The Census data will enable us to then compare the population density to examine if areas of highest crime are also most densely populated. Toronto locations of interest will then allow us to cluster and quantitatively understand the venues most common to that location.

Methodology

All steps are referenced below in the Appendix:
Analysis section.

The methodology will include:

1. Loading each data set

2. Examine the crime frequency by neighbourhood
3. Study the crime types and then pivot analysis of crime type frequency by neighbourhood
4. Understand correlation between crimes and population density
5. Perform k-means statisical analysis on venues by locations of interest based on findings from crimes and neighbourhood
6. Determine which venues are most common statistically in the region of greatest crime count then in all other locations of interest.
7. Determine if an area, such as the Knowledge Park needs a coffee shop.

Loading the data

After loading the applicable libraries, the referenced geojson neighbourhood data was loaded from the City of Toronto Open Data site. This dataset uses block polygon shape coordinates which are better for visualization and comparison. The City also uses Ward data but the Neighbourhood location data is more accurate and includes more details.

The crime dataset, an excel file, "Crime by Neighbourhood" downloaded from the City of Toronto Open Data site is found under the Public Safety domain. This dataset was then uploaded for the analysis. It's interesting to note the details of this dataset are aggregated by neighbourhood. It is not an exhaustive set by not including all crimes (violent offenses) nor specific location data of the crime but is referenced by neighbourhood.

This means we can gain an understanding of the crime volume by type by area but not specific enough to understand the distribution properties. Valuable questions such as, "are these crimes occuring more often in a specific area and at a certain time by a specific demographic of people?" cannot be answered nor explored due to what is reasonably assumed to be personal and private information with associated legal risks.

There is value to the city to explore the detailed crime data using data science to predict frequency, location, timing and conditions to best allocated resources for the benefit of its citizens and it's police force. However, human behaviour is complex requiring thick profile data by individual and the conditions surrounding the event(s). To be sufficient for reliable future prediction it would need to demonstrate validity, currency, reliability and sufficiency.

Exploring the data

Exploring the count of crimes by neighborhood gives us the first glimpse into the distribution.

One note is the possibility neighborhoods names could change at different times. The crime dataset did not mention which specific neighborhood naming dataset it was using but we assumed the neighbourhood data provided aligned with the neighbourhoods used in the crime data. It may be beneficial for the City to note and timestamp neighbourhood naming in the future or simply reference with neighbourhood naming file it used for the crime dataset.

An example of data errors: There was an error found in the naming of the neighbourhood "Platt". The neighbourhood data stated "Plat" while the crime data stated "Platt". Given the crime dataset was most simple to manipulate it was modified to "Plat". The true name of the neighbourhood is "Platt".

First Visualization of Crime

Once the data was prepared, a choropleth map was created to view the crime count by neighbourhood. As expected the region of greatest crime count was found in the downtown and Platt neighbourhoods.

Examining the crime types enables us to learn the most frequent occuring crimes which we then plot as a bar chart to see most frequenty type.

Theft from motor vehicles is most prevalent in the same area as the most frequent crimes. It's interesting to note this area is mostly residential and most do not have garages. It would be interesting to further examine if surveillance is a deterant for motor vehicle crimes in the downtown core compared to low surveillance in the Platt neighbourhood.

Examining 2nd most common crime given it is specific: theft from vehicles

After exploring the pivot table showing Crime_Type by Neighbourhood, we drill into a specific type of crime, theft from vehicles and plot the choropleth map to see which area has the greatest frequency.

Again, the Platt neighbourhood appears as the most frequent.

Is this due to population density?

Introducing the Census data to explore the correlation between crime frequency and population density.

Visualising the population density enables us to determine that the Platt neighbourhood has lower correlation to crime frequency than I would have expected.

It would be interesting to further study the Census data and if this captures the population that is renting or more temporary/transient poplution, given the City is a University hub.

Look at specific locations to understand the connection to venues using Foursquare data

Loading the "Toronto Locations" data enables us to perform a statistical analysis on the most common venues by location.

We might wonder if the prevalence of bars and clubs in the downtown region has something to do with the higher crime rate in the near Platt region.

Plotting the latitude and longitude coordinates of the locations of interest onto the crime choropleth map enables us to now study the most common venues by using the Foursquare data. Analysing each Location

Grouping rows by location and the mean of the frequency of occurance of each category we venue categories we study the top five most common venues.

Putting this data into a pandas dataframe we can then determine the most common venues by location and plot onto a map.

Results

The analysis enabled us to discover and describe visually and quantitatively:

1. Neighbourhoods in Toronto
2. Crime freqency by neighbourhood
3. Crime type frequency and statistics. The mean crime count in the City of Toronto is 22.
4. Crime type count by neighbourhood.

Theft from motor vehicles is most prevalent in the same area as the most frequent crimes. It's interesting to note this area is mostly residential and most do not have garages. It would be interesting to further examine if surveillance is a deterant for motor vehicle crimes in the downtown core compared to low surveillance in the Platt neighbourhood.

1. Motor Vehicle crimes less than $5000 analysis by neighbourhood and resulting statistics. The most common crime is Other Theft less than 5k followed by Motor Vehicle Theft less than 5k. There is a mean of 6 motor vehicle thefts less than 5k by neighbourhood in the City.

2. That population density and resulting visual correlation is not strongly correlated to crime frequency. Causation for crime is not able to be determined given lack of open data specificity by individual and environment.

3. Using k-menas, we were able to determine the top 10 most common venues within a 1 km radius of the centroid of the highest crime neighbourhood. The most common venues in the highest crime neighbourhood are coffee shops followed by Pubs and Bars.

While, it is not valid, consistent, reliable or sufficient to assume a higher concentration of the combination of coffee shops, bars and clubs predicts the amount of crime occurance in the City of Fredericton, this may be a part of the model needed to be able to in the future.

1. We were able to determine the top 10 most common venues by location of interest.
2. Statisically, we determined there are no coffee shops within the Knowledge Park clusters.

Discussion and Recommendations

The City of Toronto Open Data enables us to gain an understanding of the crime volume by type by area but not specific enough to understand the distribution properties. Valuable questions such as, "are these crimes occuring more often in a specific area and at a certain time by a specific demographic of people?" cannot be answered nor explored due to what is reasonably assumed to be personal and private information with associated legal risks.

There is value to the city to explore the detailed crime data using data science to predict frequency, location, timing and conditions to best allocated resources for the benefit of its citizens and it's police force. However, human behaviour is complex requiring thick profile data by individual and the conditions surrounding the event(s). To be sufficient for reliable future prediction it would need to demonstrate validity, currency, reliability and sufficiency.

A note of caution is the possibility neighbourhoods names could change. The crime dataset did not mention which specific neighbourhood naming dataset it was using but we assumed the neighbourhood data provided aligned with the neighbourhoods used in the crime data. It may be beneficial for the City to note and timestamp neighbourhood naming in the future or simply reference with neighbourhood naming file it used for the crime dataset.

Theft from motor vehicles is most prevalent in the same area as the most frequent crimes. It is interesting to note this area is mostly residential and most do not have garages.
It would be interesting to further examine if surveillance is a deterant for motor vehicle crimes in the downtown core compared to low surveillance in the Platt neighbourhood. It would be interesting to further study the Census data and if this captures the population that is renting or more temporary/transient poplution, given the City is a University hub.

Given the findings of the top 10 most frequent venues by locations of interest, the Knowledge Park does not have Coffee Shops in the top 10 most common venues as determined from the Foursquare dataset. Given this area has the greatest concentration of stores and shops as venues, it would be safe to assume a coffee shop would be beneficial to the business community and the citizens of Fredericton.

Conclusion

Using a combination of datasets from the City of Toronto Open Data project and Foursquare venue data we were able to analyse, discover and describe neighbhourhoods, crime, population density and statistically describe quantitatively venues by locations of interest.

While overall, the City of Toronto Open Data is interesting, it misses the details required for true valued quantitiatve analysis and predictive analytics which would be most valued by investors and developers to make appropriate investments and to minimize risk.

The Open Data project is a great start and empowers the need for a "Citizens Like Me" model to be developed where citizens of digital Toronto are able to share their data as they wish for detailed analysis that enables the creation of valued services.