Segmenting and Clustering Neighborhoods in DC, NB

Applied Data Science Capstone Week 5

Hernan Labastie

Introduction to the opportunity

Washington, D.C., formally the District of Columbia and commonly referred to as D.C., Washington, or The District, is the capital of the United States .The U.S. Census Bureau estimates that the District's population was 705,749 as of July 2019, an increase of more than 100,000 people since the 2010 United States Census.This continues a growth trend since 2000, following a half-century of population decline.The city was the 24th most populous place in the United States as of 2010.

According to data from 2010, commuters from the suburbs increase the District's daytime population to over a million.

Crime in Washington, D.C., is concentrated in areas associated with poverty, drug abuse, and gangs. A 2010 study found that 5% percent of city blocks accounted for more than 25% of the District's total crimes

Developers, investors, policy makers and/or city planners have an interest in answering the following questions as the need for additional services and citizen protection:

1. What neighbourhoods have the highest crime?

2. Is population density correlated to crime level?

3. Using Foursquare data, what venues are most common in different locations within the city?

4. Where really need a coffee shop?

Does the Open Data project have specific enough or thick enough data to empower decisions to be made or is it too aggregate to provide value in its current detail? Let's find out.

In [1]:

Slide Type Slide

```
# from PIL import Image
import requests
from PIL import Image

url = 'https://encrypted-tbn0.gstatic.com/images?q=tbn%3AANd9GcTb-Z15PkiR2QFm6NSI1ZnFx_fgzR3DYcQbbT6AmW4IG7VL02c3'
im = Image.open(requests.get(url, stream=True).raw)
im
```

Out[1]:

Data

To understand and explore we will need the following District of Columbia Open Data:

1. Open Data Site: https://dcatlas.dcgis.dc.gov/crimecards/

2. Neighbourhoods: hhttps://en.wikipedia.org/wiki/Washington,_D.C.#Demographics

3. 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. Locations of interest will then allow us to cluster and quantitatively understand the venues most common to that location.

Methodology

All steps are referenced beleow 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 DC Open Data site. This dataset uses block polygon shape coordinates which are better for visualization and comparison.

The other dataset, downloaded from the City of DC 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 neighbourhood gives us the first glimpse into the distribution.

One note is the possibility neighbourhoods names could change at different times. 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.

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 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 "DC 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 DC
2. Crime freqency by neighbourhood
3. Crime type frequency and statistics. The mean crime count in the City of DC 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 DC, 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 DC 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.

Errors exist in the current open data. An error was 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".

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 DC.

Conclusion

Using a combination of datasets from the City of DC 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 DC 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 DC are able to share their data as they wish for detailed analysis that enables the creation of valued services.