
Coursera - Capstone Project
Segmenting and Clustering Neighborhoods in DC

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

Introduction to the opportunity

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.

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/

1. 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.

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.

Slide Type Slide

Examine Crime Types

In [25]:

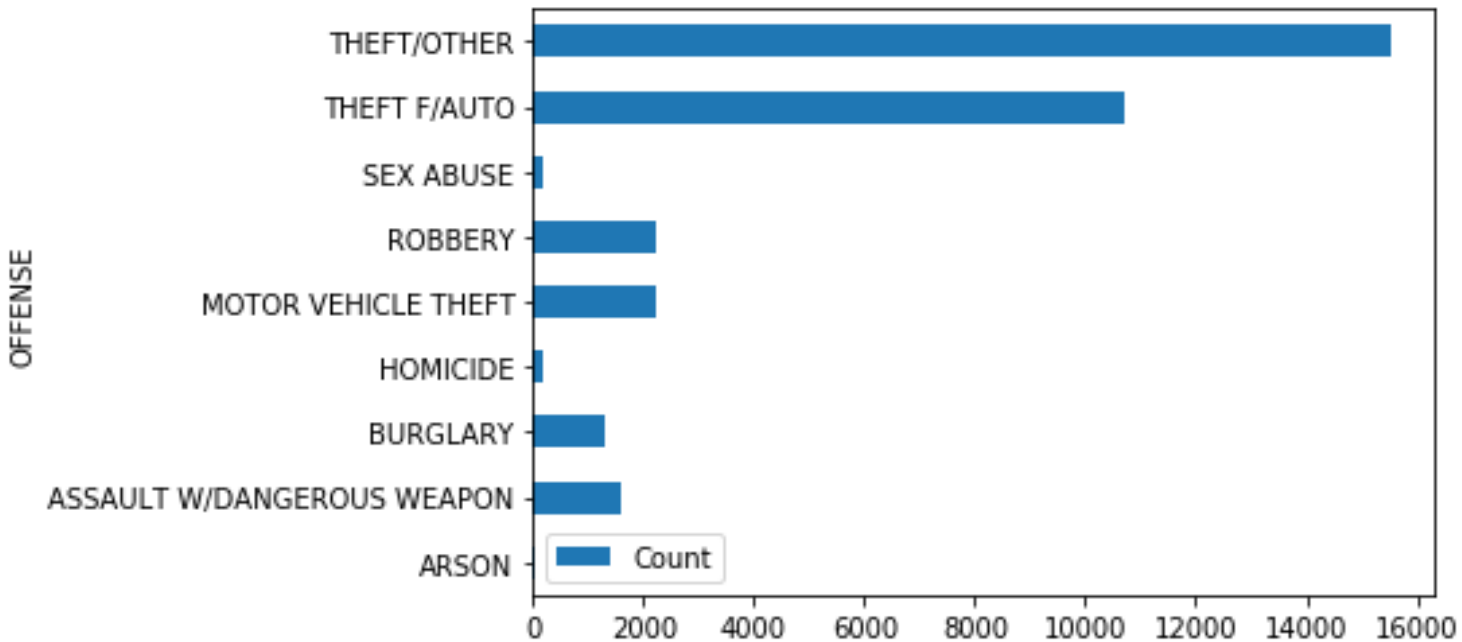
Slide Type Fragment

```
crimetype_data = crime_df.groupby(['OFFENSE']).size().to_frame(name='Count').reset_index()
crimetype_data
```

Out[25]:

	OFFENSE	Count
0	ARSON	8
1	ASSAULT W/DANGEROUS WEAPON	1568
2	BURGLARY	1273
3	HOMICIDE	165
4	MOTOR VEHICLE THEFT	2205
5	ROBBERY	2234
6	SEX ABUSE	193
7	THEFT F/AUTO	10711
8	THEFT/OTHER	15550

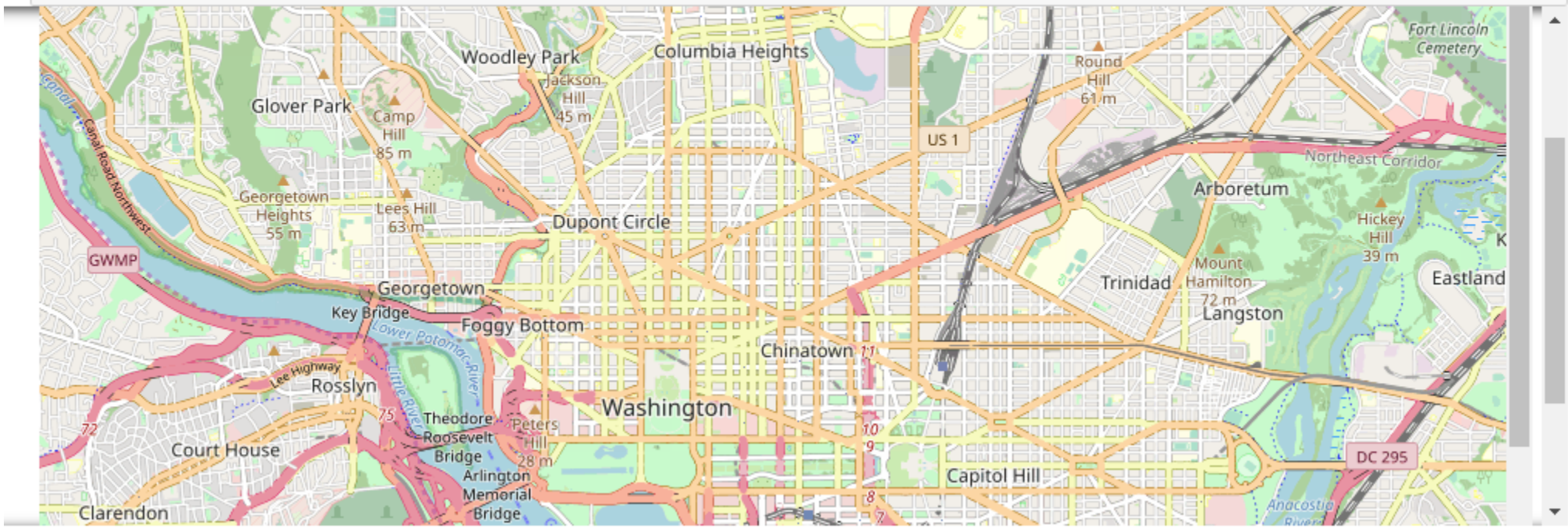
Examine Type Crime



```
world_geo = r'world_countries.json' # geojson file

dc_1_map = folium.Map(location=[38.89, -76.98], width=1000, height=750,zoom_start=12)

dc_1_map
```



In [26]:

Slide Type Sub-Slide ▾

```
crimepivot = crime_df.pivot_table(index='NEIGHBORHOOD_CLUSTER', columns='OFFENSE', aggfunc=pd.Series.count, fill_value=0)
crimepivot
```

Out[26]:

	ANC									BID	
OFFENSE	ARSON	ASSAULT W/DANGEROUS WEAPON	BURGLARY	HOMICIDE	MOTOR VEHICLE THEFT	ROBBERY	SEX ABUSE	THEFT F/AUTO	THEFT/OTHER	ARSON	ASSAULT W/DANGEROUS WEAPON
NEIGHBORHOOD_CLUSTER											
Cluster 1	1	21	41	1	52	32	2	189	371	0	14
Cluster 10	0	2	12	0	10	5	0	154	94	0	0
Cluster 11	0	5	23	1	18	20	1	212	328	0	0
Cluster 12	0	5	14	0	11	7	2	58	138	0	0
Cluster 13	0	1	18	1	17	4	3	126	54	0	0
Cluster 14	0	4	9	1	24	7	3	70	184	0	0
Cluster 15	0	8	15	0	19	9	1	98	279	0	0
Cluster 16	0	4	9	0	13	5	0	161	86	0	0
Cluster 17	0	15	17	0	54	59	6	304	425	0	0
Cluster 18	0	43	82	5	117	91	5	400	595	0	0
Cluster 19	0	10	18	1	32	31	7	223	192	0	0
Cluster 2	0	99	64	10	119	215	10	704	1201	0	0
Cluster 20	0	10	21	0	23	26	3	165	139	0	0
Cluster 21	0	82	93	6	129	114	5	380	425	0	16
Cluster 22	0	59	35	7	78	62	1	392	645	0	0
Cluster 23	1	121	69	8	88	107	9	502	568	0	0
Cluster 24	0	17	22	1	56	28	1	246	204	0	0

Results

- The analysis enabled us to discover and describe visually and quantitatively:
- Neighbourhoods in DC
- Crime freqency by neighbourhood
- Crime type frequency and statistics. The mean crime count in the City of DC is 22.
- 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.

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.