The Capstone Project Report

1. Business Problem

Assuming one wants to move to Vancouver and they are contemplating on which neighborhood to move to, they might want to know which neighborhood performs better in terms of crime rate. What would be more interesting rather is which neighborhoods has low **break and enter** crime rate or **vehicle accidents**. It all depends on what one really is moving there for.

All other factors held constant, an investor who wants to construct a business infrastructure might be interested in neighborhoods that have low **commercial break and enter** neighborhoods while a couple trying to raise a family might more keen on neighborhoods with low **residential break and enter** crime rate. This information is essential for such a large group of individuals given their reasons of moving to Vancouver or given the data, any other location in the world.

We will use the **2019 Vancouver crime dataset** to cluster which neighborhoods are ideal either to live in or do business in depending on the rate of specific crime rate in those neighborhoods. Further, we will transform the Foursquare neighborhoods data into a dataframe that can be merged with the clusters to find if there is high-view relation between the number of crimes and specific number of venues.

2. Datasets

2.1 Justification and Exploration

For this exercise, I have obtained two sets of datasets:

1. Vancouver crime report data, the csv version of which the raw data is from here
2. *Foursquare location data*, to see how the Foursquare fair into our crime clusters

As we will be comparing with the venues dataset, we don't have information about the date when the each venue was constructed enough to related a venue with a crime occurrence but we know that the venue exist now (2019), therefore, we will consider 2019 crime dataset only. The data is stored on github.

2.2 Data Cleaning

The raw crime data is a large dataset containing over 600,000 records of crime from the year 2003 to current (2019). Our focus is on the 2019 data because we are going to eventually make a comparison to the number of venues relate to the crime, it would be far too erroneous to use all the data as we do not have the data about the time when the venues were constructed. The 2019 crime data has 20831 records of which its dataframe preview as below:

	TYPE	YEAR	MONTH	DAY	HOUR	MINUTE	HUNDRED_BLOCK	NEIGHBOURHOOD	X	Y
0	Theft from Vehicle	2019	4	2	16.0	0.0	5XX CARRALL ST	Central Business District	492408.03	5458534.62
1	Theft from Vehicle	2019	2	20	9.0	47.0	5XX CARRALL ST	Central Business District	492403.16	5458628.36
2	Other Theft	2019	3	6	15.0	37.0	7XX GRANVILLE ST	Central Business District	491396.06	5458846.22
3	Other Theft	2019	2	22	19.0	15.0	11XX ROBSON ST	West End	490910.24	5459118.24
4	Offence Against a Person	2019	4	3	NaN	NaN	OFFSET TO PROTECT PRIVACY	NaN	0.00	0.00

From the preview, we note that there are some records that have null values. These constitutes about 8.48% of the dataset of which 8.22% is **Offence Against a Person** and has been omitted to protect the privacy of the individual. Due the sensitivity of this information, we get rid of all records with **Offence Against a Person** and the less than 0.26% of the other null values as it is too small to affect our results.

2.3 Feature Selection

Furthermore, we remove the **TIME** and **HUNDRED_BLOCK** columns as our analysis will involve frequency of crime in a neighborhood. The resulting dataframe preview is as follows:

	CRIME_TYPE	NEIGHBOURHOOD	X	Y
0	Theft from Vehicle	Central Business District	492408.03	5458534.62
1	Theft from Vehicle	Central Business District	492403.16	5458628.36
2	Other Theft	Central Business District	491396.06	5458846.22
3	Other Theft	West End	490910.24	5459118.24
5	Theft from Vehicle	West End	490848.76	5458857.79

Multiple dataframe transformations including changing the **XY-coordinate** to **latitude-longitude** coordinates, one-hot encoding of the **crime_type** to deal with categorical data then grouping by neighborhood. The resulting dataframe preview is below:

	neighborhood	Break and Enter Commercial	Break and Enter Residential/Other	Mischief	Other Theft	Theft from Vehicle	Theft of Bicycle	Theft of Vehicle	Vehicle Collision or Pedestrian Struck (with Fatality)	Vehicle Collision or Pedestrian Struck (with Injury)
0	Arbutus Ridge	0.027211	0.183673	0.149660	0.108844	0.401361	0.040816	0.034014	0.000000	0.054422
1	Central Business District	0.049656	0.011853	0.153292	0.201185	0.517059	0.039244	0.012975	0.000160	0.014576
2	Dunbar-Southlands	0.011905	0.166667	0.202381	0.077381	0.416667	0.029762	0.023810	0.000000	0.071429
3	Fairview	0.081106	0.045161	0.122581	0.177880	0.437788	0.094931	0.017512	0.000000	0.023041
4	Grandview-Woodland	0.055843	0.084798	0.202689	0.134436	0.377456	0.055843	0.057911	0.001034	0.029990

2.4 Venues Frequency

Below is a dataframe that will be used to merge the most common crimes with the most common venues:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Arbutus Ridge	Bakery	Pet Store	Grocery Store	Nightlife Spot	Yoga Studio
1	Central Business District	Coffee Shop	Hotel	Steakhouse	Theater	Poke Place
2	Dunbar-Southlands	Sushi Restaurant	Italian Restaurant	Coffee Shop	Liquor Store	Fast Food Restaurant
3	Fairview	Coffee Shop	Park	Asian Restaurant	Sushi Restaurant	Korean Restaurant
4	Grandview-Woodland	Italian Restaurant	Pizza Place	Coffee Shop	Sushi Restaurant	Indian Restaurant

3. Methodology

Since the data is unlabeled and the whole point in this project is to cluster the neighborhoods-based crime occurrence, we use KMeans. This model is chosen because of its simplicity and it serves the purpose of this exercise well enough because all we want to see is how the neighborhoods are similar in terms of crime rate.

As stated earlier, our analysis will involve frequency of the crimes in the neighborhoods. The first place to start with is see how many crimes happened in each neighborhood in the year 2019, grouping the number of crimes per neighborhood, we get the following table:

neighborhood	crime_type
Arbutus Ridge	147
Central Business District	6243
Dunbar-Southlands	168
Fairview	1085
Grandview-Woodland	967
Hastings-Sunrise	723
Kensington-Cedar Cottage	800
Kerrisdale	209
Killarney	277
Kitsilano	777
Marpole	452
Mount Pleasant	1215
Musqueam	8
Oakridge	296
Renfrew-Collingwood	875
Riley Park	486
Shaughnessy	158
South Cambie	204
Stanley Park	104
Strathcona	1108
Sunset	573
Victoria-Fraserview	343
West End	1625
West Point Grey	166

The table shows that Central Business District had the highest number of crimes in 2019 while Musqueam has the lowest number.

And what are the most occurring crimes in Vancouver? Below is a sneak-peak of the most occurring crimes:

	neighborhood	1st Most Common Crime	2nd Most Common Crime	3rd Most Common Crime	4th Most Common Crime	5th Most Common Crime
0	Arbutus Ridge	Theft from Vehicle	Break and Enter Residential/Other	Mischief	Other Theft	Vehicle Collision or Pedestrian Struck (with I...
1	Central Business District	Theft from Vehicle	Other Theft	Mischief	Break and Enter Commercial	Theft of Bicycle
2	Dunbar-Southlands	Theft from Vehicle	Mischief	Break and Enter Residential/Other	Other Theft	Vehicle Collision or Pedestrian Struck (with I...
3	Fairview	Theft from Vehicle	Other Theft	Mischief	Theft of Bicycle	Break and Enter Commercial
4	Grandview-Woodland	Theft from Vehicle	Mischief	Other Theft	Break and Enter Residential/Other	Theft of Vehicle

Fascinating enough, **Theft from Vehicle** is such a big problem in Vancouver that all the neighborhoods has it as the most common Crime. This is an interesting observation as it will help up to interpret the result of our model from a different perspective.

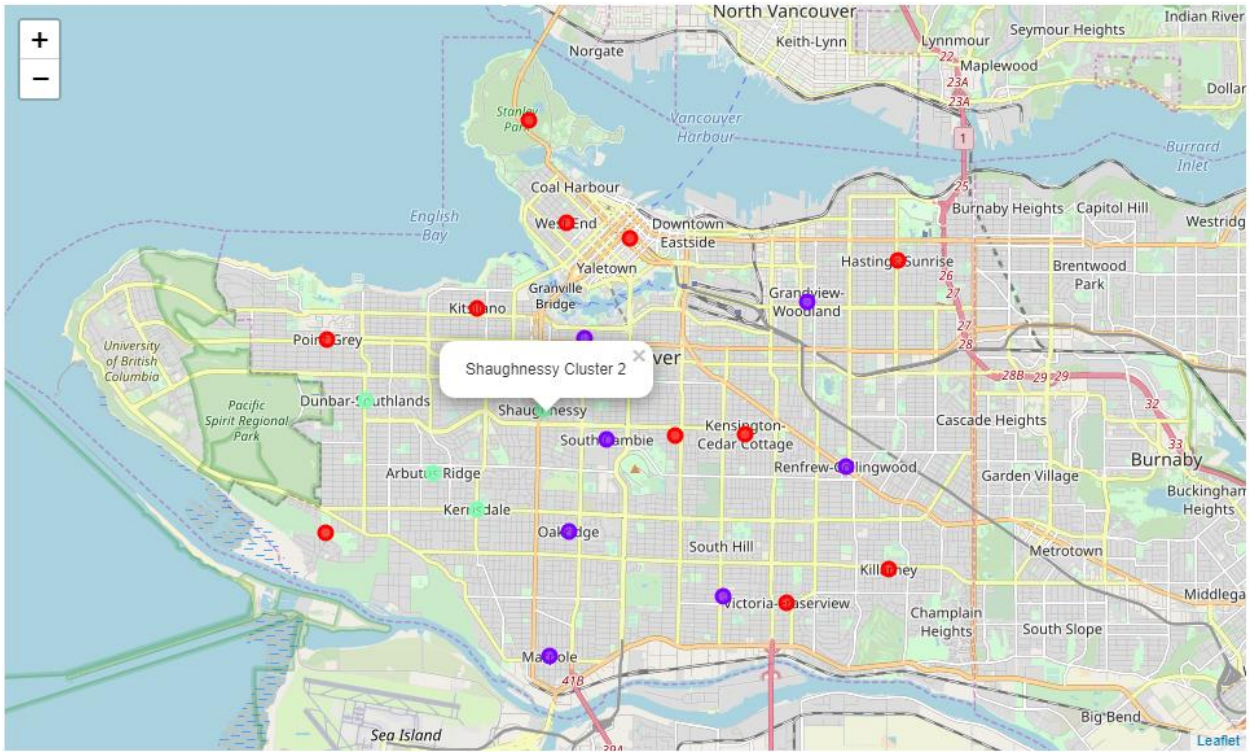
After running the KMeans algorithm with Kclusters = 3, we get the following dataframe with **cluster_labels** column. The neighborhoods have been clustered into 3 clusters (0,1,2)

	neighborhood	latitude	longitude	cluster_labels	1st Most Common Crime	2nd Most Common Crime	3rd Most Common Crime	4th Most Common Crime	5th Most Common Crime
0	Central Business District	49.281466	-123.114711	0	Theft from Vehicle	Other Theft	Mischief	Break and Enter Commercial	Theft of Bicycle
1	West End	49.284131	-123.131795	0	Theft from Vehicle	Other Theft	Mischief	Break and Enter Commercial	Theft of Bicycle
2	Riley Park	49.247438	-123.102966	0	Theft from Vehicle	Break and Enter Residential/Other	Mischief	Other Theft	Theft of Vehicle
3	Kerrisdale	49.234673	-123.155389	2	Theft from Vehicle	Break and Enter Residential/Other	Mischief	Vehicle Collision or Pedestrian Struck (with I...	Other Theft
4	Marpole	49.209223	-123.136150	1	Theft from Vehicle	Other Theft	Mischief	Break and Enter Residential/Other	Break and Enter Commercial

4. Cluster Results

4.1 Geographical Cluster Results

Let's look at how our clusters are fairing geographically on a map:



The cluster has cluster 0 with neighborhoods around the city, while cluster 1 and 2 are fairly within the city with cluster 1 spread to the west side of the city while cluster 2 is to the east.

4.2 Cluster 0

	neighborhood	1st Most Common Crime	2nd Most Common Crime	3rd Most Common Crime	4th Most Common Crime	5th Most Common Crime
0	Central Business District	Theft from Vehicle	Other Theft	Mischief	Break and Enter Commercial	Theft of Bicycle
1	West End	Theft from Vehicle	Other Theft	Mischief	Break and Enter Commercial	Theft of Bicycle
2	Riley Park	Theft from Vehicle	Break and Enter Residential/Other	Mischief	Other Theft	Theft of Vehicle
5	Kensington-Cedar Cottage	Theft from Vehicle	Mischief	Other Theft	Break and Enter Residential/Other	Theft of Vehicle
6	Stanley Park	Theft from Vehicle	Mischief	Theft of Bicycle	Vehicle Collision or Pedestrian Struck (with I...	Other Theft
10	Strathcona	Theft from Vehicle	Mischief	Break and Enter Commercial	Other Theft	Break and Enter Residential/Other
11	Kitsilano	Theft from Vehicle	Mischief	Other Theft	Break and Enter Residential/Other	Theft of Bicycle
13	Hastings-Sunrise	Theft from Vehicle	Mischief	Break and Enter Residential/Other	Other Theft	Theft of Vehicle
15	Victoria-Fraserview	Theft from Vehicle	Mischief	Break and Enter Residential/Other	Vehicle Collision or Pedestrian Struck (with I...	Theft of Vehicle
16	West Point Grey	Theft from Vehicle	Mischief	Break and Enter Residential/Other	Theft of Bicycle	Theft of Vehicle
19	Killarney	Theft from Vehicle	Mischief	Other Theft	Break and Enter Residential/Other	Theft of Vehicle
23	Musqueam	Theft from Vehicle	Vehicle Collision or Pedestrian Struck (with I...	Theft of Vehicle	Mischief	Break and Enter Residential/Other

As we will observe for the other clusters, the 1st Most Common Crime doesn't give us any relevant information apart from the fact the Theft from Vehicle is a big problem all over Vancouver. All the neighborhoods are pretty much infested with the problem. But if we take a look at the 2nd, 3rd and 4th Most Common Crimes in Vancouver, we will notice that this cluster is has a problem of Mischief defined as "willfully causing malicious destruction, damage, or defacement of property including any public mischief towards another person" by the Vancouver Police here and Break and Enter be it commercial or residential.

Cluster 0 merged with Events dataframe:

	neighborhood	1st Most Common Crime	2nd Most Common Crime	3rd Most Common Crime	4th Most Common Crime	5th Most Common Crime	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Central Business District	Theft from Vehicle	Other Theft	Mischief	Break and Enter Commercial	Theft of Bicycle	Coffee Shop	Hotel	Steakhouse	Theater	Poke Place
1	West End	Theft from Vehicle	Other Theft	Mischief	Break and Enter Commercial	Theft of Bicycle	Gay Bar	Bakery	Ramen Restaurant	Restaurant	Café
2	Riley Park	Theft from Vehicle	Break and Enter Residential/Other	Mischief	Other Theft	Theft of Vehicle	Arts & Crafts Store	Vegetarian / Vegan Restaurant	Restaurant	Japanese Restaurant	Grocery Store
3	Kensington-Cedar Cottage	Theft from Vehicle	Mischief	Other Theft	Break and Enter Residential/Other	Theft of Vehicle	Coffee Shop	Vietnamese Restaurant	Bus Stop	Indian Restaurant	Filipino Restaurant
4	Stanley Park	Theft from Vehicle	Mischief	Theft of Bicycle	Vehicle Collision or Pedestrian Struck (with I...	Other Theft	Trail	Lake	Park	Yoga Studio	Filipino Restaurant
5	Kitsilano	Theft from Vehicle	Mischief	Other Theft	Break and Enter Residential/Other	Theft of Bicycle	Bakery	Coffee Shop	Tea Room	Yoga Studio	Grocery Store
6	Hastings-Sunrise	Theft from Vehicle	Mischief	Break and Enter Residential/Other	Other Theft	Theft of Vehicle	Vietnamese Restaurant	Liquor Store	Beer Store	Park	Coffee Shop
7	Victoria-Fraserview	Theft from Vehicle	Mischief	Break and Enter Residential/Other	Vehicle Collision or Pedestrian Struck (with I...	Theft of Vehicle	Convenience Store	Pizza Place	Sandwich Place	Fast Food Restaurant	Yoga Studio
8	West Point Grey	Theft from Vehicle	Mischief	Break and Enter Residential/Other	Theft of Bicycle	Theft of Vehicle	Pool	Yoga Studio	Cuban Restaurant	Dessert Shop	Dim Sum Restaurant
9	Killarney	Theft from Vehicle	Mischief	Other Theft	Break and Enter Residential/Other	Theft of Vehicle	Pool	Soccer Field	Italian Restaurant	Gym	Deli / Bodega

The variation of *Most Common Venues* in these neighborhoods might imply that our model has identifies these neighborhoods as some of the most active. These are probably where business is mostly happening, and these kinds of crimes are prominent.

4.3 Cluster 1

	neighborhood	1st Most Common Crime	2nd Most Common Crime	3rd Most Common Crime	4th Most Common Crime	5th Most Common Crime
4	Marpole	Theft from Vehicle	Other Theft	Mischief	Break and Enter Residential/Other	Break and Enter Commercial
7	Mount Pleasant	Theft from Vehicle	Other Theft	Mischief	Break and Enter Commercial	Theft of Bicycle
8	Renfrew-Collingwood	Theft from Vehicle	Other Theft	Mischief	Break and Enter Residential/Other	Theft of Vehicle
12	Grandview-Woodland	Theft from Vehicle	Mischief	Other Theft	Break and Enter Residential/Other	Theft of Vehicle
14	Sunset	Theft from Vehicle	Other Theft	Mischief	Break and Enter Commercial	Theft of Vehicle
17	Oakridge	Theft from Vehicle	Other Theft	Break and Enter Residential/Other	Mischief	Theft of Vehicle
18	Fairview	Theft from Vehicle	Other Theft	Mischief	Theft of Bicycle	Break and Enter Commercial
20	South Cambie	Theft from Vehicle	Other Theft	Mischief	Break and Enter Residential/Other	Vehicle Collision or Pedestrian Struck (with I...

Again, we will ignore the *Theft from Vehicle* and move on to the 2nd and 3rd Most common Crimes. We have *Other Theft* followed by *Mischief* as the Most Common Crimes in these neighborhoods. *Other Theft* is defined as "theft of property that includes personal items (purse, wallet, cellphone, laptop, etc.), bicycle, etc" here

Cluster 1 merged with Events dataframe:

	neighborhood	1st Most Common Crime	2nd Most Common Crime	3rd Most Common Crime	4th Most Common Crime	5th Most Common Crime	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Marpole	Theft from Vehicle	Other Theft	Mischief	Break and Enter Residential/Other	Break and Enter Commercial	Sushi Restaurant	Pizza Place	Chinese Restaurant	Vietnamese Restaurant	Grocery Store
1	Renfrew-Collingwood	Theft from Vehicle	Other Theft	Mischief	Break and Enter Residential/Other	Theft of Vehicle	Vietnamese Restaurant	Chinese Restaurant	Pharmacy	Shanghai Restaurant	Café
2	Grandview-Woodland	Theft from Vehicle	Mischief	Other Theft	Break and Enter Residential/Other	Theft of Vehicle	Italian Restaurant	Pizza Place	Coffee Shop	Sushi Restaurant	Indian Restaurant
3	Sunset	Theft from Vehicle	Other Theft	Mischief	Break and Enter Commercial	Theft of Vehicle	Indian Restaurant	Dessert Shop	Ski Area	Filipino Restaurant	Deli / Bodega
4	Oakridge	Theft from Vehicle	Other Theft	Break and Enter Residential/Other	Mischief	Theft of Vehicle	Convenience Store	Fast Food Restaurant	Vietnamese Restaurant	Sandwich Place	Sushi Restaurant
5	Fairview	Theft from Vehicle	Other Theft	Mischief	Theft of Bicycle	Break and Enter Commercial	Coffee Shop	Park	Asian Restaurant	Sushi Restaurant	Korean Restaurant
6	South Cambie	Theft from Vehicle	Other Theft	Mischief	Break and Enter Residential/Other	Vehicle Collision or Pedestrian Struck (with I...	Coffee Shop	Park	Liquor Store	Bus Stop	Café

Most of these venues are food places where people are most likely to forget their items that are mentioned as examples for *Other Theft* and *Mischief* is likely to happen.

4.4 Cluster 2

	neighborhood	1st Most Common Crime	2nd Most Common Crime	3rd Most Common Crime	4th Most Common Crime	5th Most Common Crime
3	Kerrisdale	Theft from Vehicle	Break and Enter Residential/Other	Mischief	Vehicle Collision or Pedestrian Struck (with I...	Other Theft
9	Dunbar-Southlands	Theft from Vehicle	Mischief	Break and Enter Residential/Other	Other Theft	Vehicle Collision or Pedestrian Struck (with I...
21	Shaughnessy	Theft from Vehicle	Break and Enter Residential/Other	Mischief	Vehicle Collision or Pedestrian Struck (with I...	Break and Enter Commercial
22	Arbutus Ridge	Theft from Vehicle	Break and Enter Residential/Other	Mischief	Other Theft	Vehicle Collision or Pedestrian Struck (with I...

As usual, looking at the 2nd and 3rd Most Common Crimes one would deduce that these are mostly residential areas as Break and Enter Residential/Other is prominent.

Cluster 2 merged with Events dataframe:

	neighborhood	1st Most Common Crime	2nd Most Common Crime	3rd Most Common Crime	4th Most Common Crime	5th Most Common Crime	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Kerrisdale	Theft from Vehicle	Break and Enter Residential/Other	Mischief	Vehicle Collision or Pedestrian Struck (with I...	Other Theft	Coffee Shop	Pharmacy	Tea Room	Chinese Restaurant	Sandwich Place
1	Dunbar-Southlands	Theft from Vehicle	Mischief	Break and Enter Residential/Other	Other Theft	Vehicle Collision or Pedestrian Struck (with I...	Sushi Restaurant	Italian Restaurant	Coffee Shop	Liquor Store	Fast Food Restaurant
2	Shaughnessy	Theft from Vehicle	Break and Enter Residential/Other	Mischief	Vehicle Collision or Pedestrian Struck (with I...	Break and Enter Commercial	Park	Bus Stop	French Restaurant	Yoga Studio	Filipino Restaurant
3	Arbutus Ridge	Theft from Vehicle	Break and Enter Residential/Other	Mischief	Other Theft	Vehicle Collision or Pedestrian Struck (with I...	Bakery	Pet Store	Grocery Store	Nightlife Spot	Yoga Studio

These look like primarily residential areas as most of these are quick needs venues like Pharmacy, Grocery Store, Pet Store and food places

5. Discussion

As an observation, it can be stated that the KMeans cluster has segmented our neighborhoods into three parts: **the business-oriented cluster, the food cluster** and **the residential cluster.**

The business-oriented cluster experiences a lot of mischief and break and enter be it commercial or residential. Noting the wide variety of most frequent venues stresses the point that different activities are operating in these neighborhoods. The high frequency of *Break and Enter Residential* also tells us how the population is business oriented. These are highly competitive business residentials and an individual looking to do business might want to consider the other two clusters.

The Food Cluster has a lot of food venues. They are associated with *Other theft,* which according to the source consists of theft of items such as phones, laptops and purses. As a distant observer, these neighborhoods might have a lot of well-established companies where people work regular jobs hence the frequency of food venues. You might want to hold on tight to your mobile luggage if you are moving of visiting neighborhoods in this cluster.

The Residential Cluster has *Break and Enter Residential* as the most frequently occurring crime. There is also an indication a lot of quick-needs venues like restaurants, pharmacy and Grocery Stores ideal for residential areas. If one is looking for a completely new space to start a business without a competition, these areas would be ideal.

6. Conclusion

The whole point of this analysis was to give an overview of the neighborhoods in Vancouver based on the frequency of crime type that they experience as well as how those crimes relate to the venues that are in those neighborhoods. As such, we looked at the data provided by the Vancouver police and analyzed their frequency in 2019. We extracted the features of the 5 top most occurring crimes for each neighborhood. We then used KMeans to cluster those features and merged them to the corresponding events for those neighborhoods.

As an improvement to this analysis, it would be worth getting time data about the venues, when they were constructed, how long they have been operating etc. Then an analysis across time related to crimes would be performed and see if there is any relation between specific venues and crimes.