Predicting Neighborhood Crime Rate from Venue Types

Dennis Menze

May 13, 2019

1. Introduction

1.1 Description & Discussion of the Background

Living in large cities has a lot of benefits, but also some disadvantages: whereas one of the advantages is the abundance of diverse venues promising a variety of activities at all times, a clear disadvantage is the higher crime rate common in metropolises.

Crime rates are not equally distributed among all neighborhoods of a city. As well as venues, which cluster in city centers and certainly are of different kinds and rarer in suburbs.

Venues have a lot to tell about the neighborhoods they are located in. But what do they say about the inhabitants and the visitors of the neighborhoods?

Is there a correlation between types of crimes and types of venues? Can you predict the crime rate of a neighborhood from the types of its venues? To approach an answer to those questions, in this paper, as an example, the city of Toronto, its neighborhoods and venues are analyzed with respect to its crime rates.

Different clients or groups of people could potentially be interested in the relationship of crime rate and venues in a neighborhood: sociologists, investors, journalists, politicians, police departments, and of course people who live in cities or think about moving there or in another neighborhood.

In the next section, we will describe which data we used as a source for our analysis.

1.2 Data Description

- Wikipedia "The list of postal code of Canada" [1] for a list of the neighborhoods of Toronto

	Postcode	Borough	Neighbourhood
0	M1B	Scarborough	Rouge, Malvern
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union
2	M1E	Scarborough	Guildwood, Morningside, West Hill
3	M1G	Scarborough	Woburn
4	M1H	Scarborough	Cedarbrae
5	M1J	Scarborough	Scarborough Village
6	M1K	Scarborough	East Birchmount Park, Ionview, Kennedy Park
7	M1L	Scarborough	Clairlea, Golden Mile, Oakridge
8	M1M	Scarborough	Cliffcrest, Cliffside, Scarborough Village West
9	M1N	Scarborough	Birch Cliff, Cliffside West
10	M1P	Scarborough	Dorset Park, Scarborough Town Centre, Wexford ...

- Foursquare API to get the most common venues of given borough of Toronto [2]
- Geospatial coordinates of the boroughs and neighborhoods of Toronto [3]

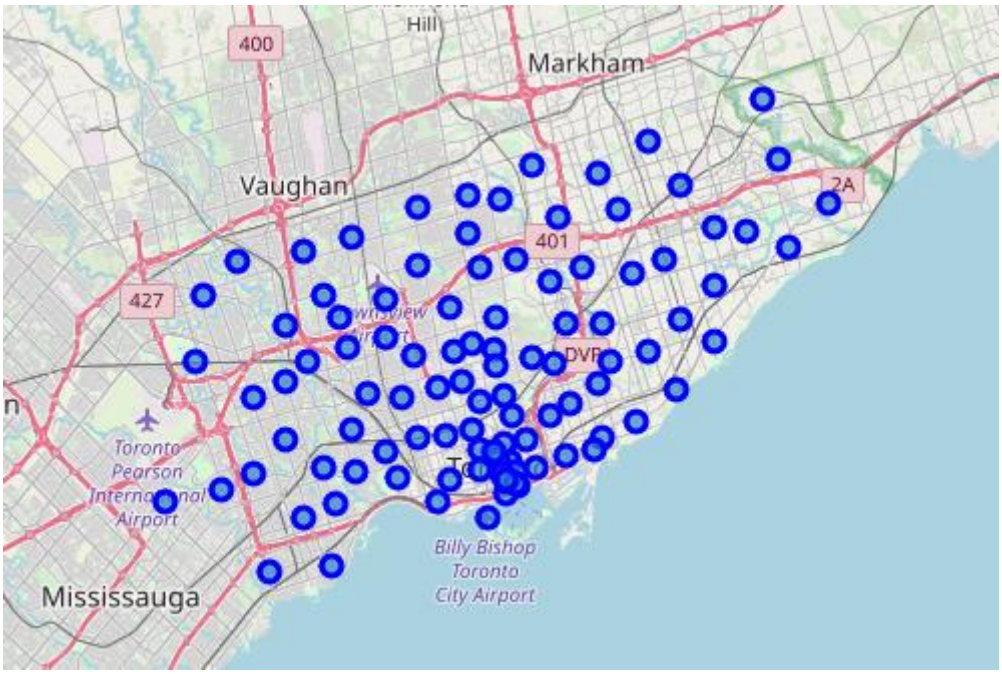
	Postal Code	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917
4	M1H	43.773136	-79.239476
5	M1J	43.744734	-79.239476
6	M1K	43.727929	-79.262029
7	M1L	43.711112	-79.284577
8	M1M	43.716316	-79.239476
9	M1N	43.692657	-79.264848
10	M1P	43.757410	-79.273304

- Crime data for Toronto per neighborhood for 2008 and 2011 [4]

Neighbourhood	Neighbourhood Id	Arsons	Assaults	Break & Enters	Drug Arrests	Fire Medical Calls	Fire Vehicle Incidents	Fires & Fire Alarms	Hazardous Incidents	Murders	Robberies	Sexual Assaults	Thefts	In
West Humber-Clairville	1	4	390	175	62	1321	502	705	210	0	82	68	54	
Mount Olive-Silverstone-Jamestown	2	3	316	61	90	1016	59	361	176	1	78	75	7	
Thistletown-Beaumond Heights	3	0	85	36	16	323	48	90	34	0	17	24	2	
Rexdale-Kipling	4	0	59	32	15	305	34	94	55	1	16	20	3	

2. Methodology

First, let's visualize the neighborhoods of Toronto on a map with the folium library:



Then, with the Foursquare API, we get the data of 100 venues in the radius of 500 meters for all 103 neighborhoods of Toronto. The first five of 2244 rows with neighborhood name, venue name, category, latitude and longitude information look like this:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Rouge, Malvern	43.806686	-79.194353	Wendy's	43.807448	-79.199056	Fast Food Restaurant
1	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497	Royal Canadian Legion	43.782533	-79.163085	Bar
2	Guildwood, Morningside, West Hill	43.763573	-79.188711	Swiss Chalet Rotisserie & Grill	43.767697	-79.189914	Pizza Place
3	Guildwood, Morningside, West Hill	43.763573	-79.188711	G & G Electronics	43.765309	-79.191537	Electronics Store
4	Guildwood, Morningside, West Hill	43.763573	-79.188711	Marina Spa	43.766000	-79.191000	Spa

In sum, there are 274 unique categories of venues.

We can group the venues by neighborhood to check how many venues there are in each:

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Adelaide, King, Richmond	100	100	100	100	100	100
Agincourt	4	4	4	4	4	4
Agincourt North, L'Amoreaux East, Milliken, Steeles East	3	3	3	3	3	3
Albion Gardens, Beaumond Heights, Humbergate, Jamestown, Mount Olive, Silverstone, South Steeles, Thistletown	9	9	9	9	9	9

So, “Adelaide, King, Richmond” reached the limit of 100 venues, whereas there are only four venues in Agincourt.

We could further analyze our neighborhoods by determining the most common venue categories per neighborhood:

	Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Adelaide, King, Richmond	Coffee Shop	Café	Steakhouse	Thai Restaurant	American Restaurant
1	Agincourt	Breakfast Spot	Lounge	Sandwich Place	Skating Rink	Donut Shop
2	Agincourt North, L'Amoreaux East, Milliken, St...	Park	Playground	Coffee Shop	Yoga Studio	Donut Shop
3	Albion Gardens, Beaumond Heights,	Grocery Store	Pharmacy	Coffee Shop	Sandwich Place	Fast Food Restaurant

So, in Agincourt for instance, the most common venues are breakfast spots.

Most machine learning techniques need numerical data for their predictors. Right now, our venue data mostly consists of text data. We can produce numerical data from it by grouping it by neighborhood and by taking the mean of the frequency of occurrence of each category. In this way, our data also will be normalized.

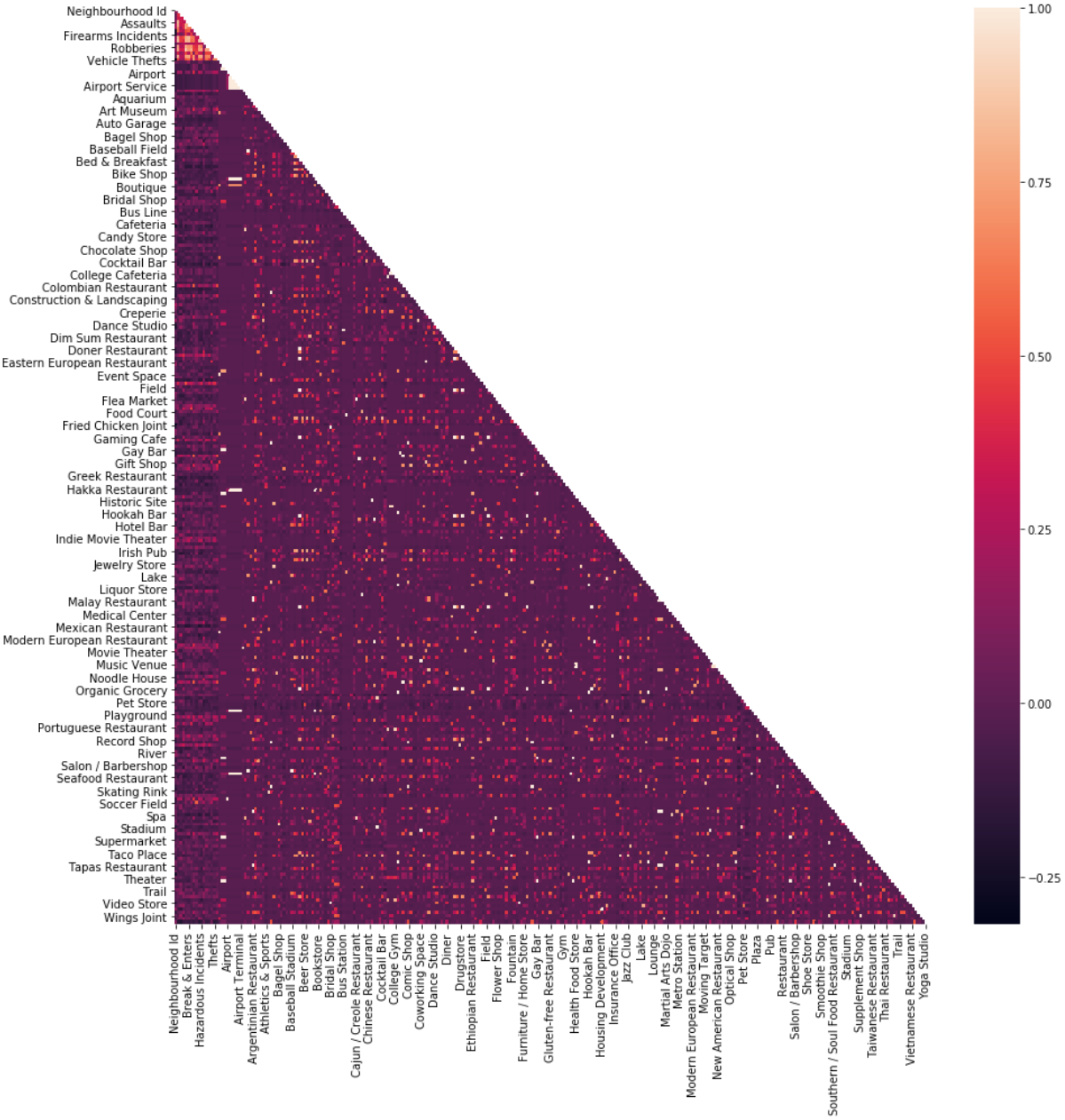
The resulting table looks like the following:

	Neighbourhood	Accessories Store	Adult Boutique	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service
0	Adelaide, King, Richmond	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	Agincourt	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	Agincourt North, L'Amoreaux East, Milliken, St...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	Albion Gardens, Beaumond Heights, Humbergate, ...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	Alderwood, Long Branch	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

3. Results

In the next step, we join venue and safety (crime) data to prepare our training data for the following machine learning steps. For this experiment, we assume that all venues exist at least since 2008.

Before that, we can calculate the correlations between venue frequency and crime data. Visualizing those correlations, we get the following heatmap.



As this diagram is a little crowded, we pick out some interesting correlations:

- Airports and fire, fire alarms and fire vehicle incidents are correlated.

```
Fire Vehicle Incidents    0.401650
Fires & Fire Alarms       0.271274
Thefts                    0.255725
Vehicle Thefts            0.245793
Ambulance Calls           0.205429
Name: Airport, dtype: float64
```

- American restaurants and Vehicle Thefts, Murders, Drug arrests and assaults are correlated (as well as BBQ joints, bagel shops, bars).

```
Vehicle Thefts            0.327264
Murders                   0.265912
Drug Arrests              0.240028
Fire Vehicle Incidents    0.233083
Assaults                  0.224617
Name: American Restaurant, dtype: float64
```

- Strangely, art museums and murders are correlated.

```
Fire Vehicle Incidents    0.396565
Vehicle Thefts            0.394308
Murders                   0.348079
Name: Art Museum, dtype: float64
```

- During concerts, vehicles get stolen.

```
Vehicle Thefts            0.382650
Fire Vehicle Incidents    0.353905
Ambulance Referrals       0.178992
Name: Concert Hall, dtype: float64
```

- Drug stores seem to be dangerous places.

```
Arsons                 0.401736
Murders                0.381824
Ambulance Referrals    0.307636
Sexual Assaults        0.305026
Firearms Incidents     0.264646
Hazardous Incidents    0.252290
Fires & Fire Alarms    0.216097
Drug Arrests           0.188385
Break & Enters         0.184199
Name: Drugstore, dtype: float64
```

- Fast food restaurants seem to be the most dangerous places.

```
Assaults                  0.513790
Thefts                    0.407563
Hazardous Incidents       0.375351
Ambulance Calls           0.374421
Sexual Assaults           0.368571
Drug Arrests              0.321326
Break & Enters            0.282744
TCHC Safety Incidents     0.228861
Fires & Fire Alarms       0.224875
Fire Vehicle Incidents    0.195933
Robberies                 0.169660
Arsons                    0.146332
Ambulance Referrals       0.145746
Name: Fast Food Restaurant, dtype: float64
```

- Many drug arrests seem to happen in front of food shops.

```
Drug Arrests    0.48263
Name: Food & Drink Shop, dtype: float64
```

- Why are gardens correlated with ambulance calls?? Are they near hospitals or old people's homes?

```
Ambulance Calls    0.388750
Thefts             0.331781
Robberies          0.323973
Assaults           0.314579
Name: Garden, dtype: float64
```

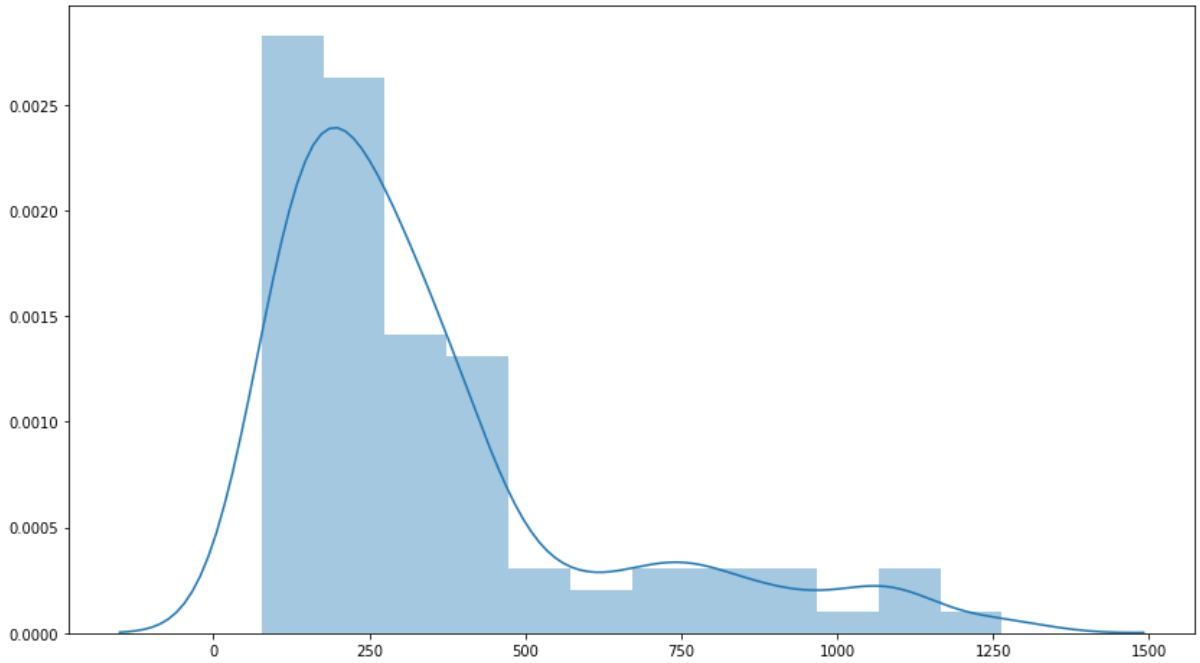
- Opera houses are highly correlated with vehicle thefts and fire vehicle indicents.

```
Fire Vehicle Incidents    0.532798
Vehicle Thefts            0.498478
Ambulance Referrals       0.273625
Thefts                    0.252017
Break & Enters            0.245656
Robberies                 0.202323
Fires & Fire Alarms       0.197003
Arsons                    0.190851
Name: Opera House, dtype: float64
```

We want to train a Support Vector Machine (SVM) to predict crime data from venue data. As SVM's are classifier, we need to transform the crime data from numerical to categorical data.

To make our classification task easier, we do not try to predict crime in each category separately (arsons, assaults, murders, etc.), but try to predict their sum per neighborhood (target variable).

To find two categories (binary classification) which split our target variable equally, we look at its distribution:



Thus, the value ranges for our target variable are: 75.678 <= 738 < 1399

I.e., values below or equal 738 belong to category 0, values above 738 belong to category 1.

We split our dataset in the following way: for the training of the SVM, we use crime data for 2008, for testing, we use crime data for 2011.

When evaluating our trained SVM, as a result, we get a balanced accuracy score of **0.694.**

4. Conclusion

For the city of Toronto, it is possible to predict crime rates with an accuracy of about 70 percent per neighborhood only by using venue data.

These results could be very easily enhanced. We only used crime data for 2008 for training and 2011 for testing, and only for Toronto. If we used more data from other years, results could probably be better. It would also be interesting to see if the results could be reproduced or enhanced with data of other cities.

A further investigation could try to find other explanatory variables besides venue data to reach an accuracy above 80 or 90 percent.

5. References

[1] „List of postal codes of Canada - Wikipedia,“ [Online]. Available: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M.

[2] „Foursquare API,“ [Online]. Available: https://developer.foursquare.com/.

[3] „Geospatial coordinates of the boroughs and neighborhoods of Toronto,“ [Online]. Available: https://geo.nyu.edu/catalog/nyu_2451_34572.

[4] „Crime data for Toronto per neighborhood for 2008 and 2011,“ [Online]. Available: http://opendata.toronto.ca/social.development/wellbeing/WB-Safety.xlsx.