Leonardo Patricelli Sep 25, 2020 4 min read

Study on Toronto's bicycle thefts

Updated: Sep 26, 2020

Hello Toronto!

My bike got stolen recently, so I decided to go deeper into this the only way I know, with statistics. Moreover, I wanted to try an XGBoost model with R. So here I am.

I will not discuss in here the math and the code used to generate this info, but I'll post an attachment with the R script I wrote with my notes. The package used for data visualization is Ggplot2.

1. Getting ready

First of all, let's download the data from Toronto's open data portal. Here is the link:

https://open.toronto.ca/

This website posts a lot of free datasets about the city of Toronto. Most of the data are police records and, usually, all the data have a field containing the neighbourhood. In this website, I found data about byke thefts from 2014 to 2019. Here is the link with all the details.

https://open.toronto.ca/dataset/bicycle-thefts/

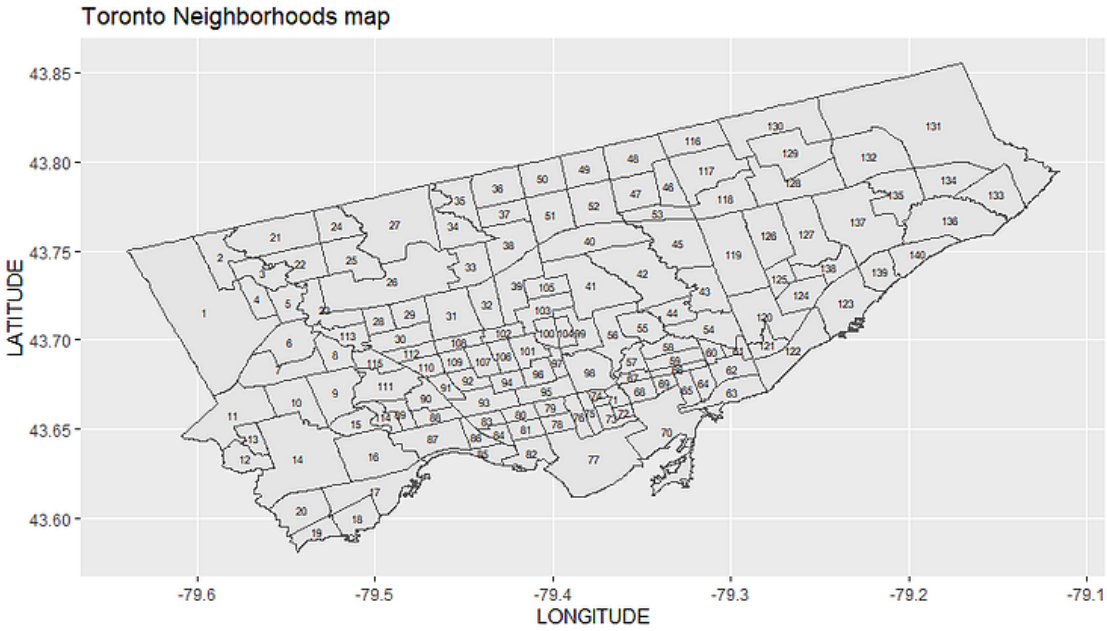
From the Toronto open data website, you can also download the city shapefile (a kind of map that you can use to plot statistics). Here is the link.

The most important features used in this analysis are:

- The date on which the theft occurred;
- The neighborhood of the theft.

It could be really interesting and useful to use the information about the bike model contained in the dataset, but the variables have a serious problem with consistency. In fact, there is no info about the labels used and there are too many missing values to use the variable in the analysis.

Let's start our exploratory analysis.
The map below shows Toronto's neighborhoods. Inside each neighborhood, I wrote the relative ID number. The table below shows the names.



Here it is a brief description :

1. **ID** is the ID number displayed in the map above;
2. **Area_name** is the name of the neighborhood;
3. **Theft Risk** is divided into 4 categories, fro low risk to high risk, and determines the chance of theft;
4. **The year** says the number of thefts occurred that year in a specific neighborhood;
5. **Total** is the total of bike stolen in a neighborhood in the latest 6 years.

ID	AREA_NAME	THEFT RISK	2014	2015	2016	2017	2018	2019	Total
1	West Humber-Clairville (1)	Low risk	4	18	21	10	9	8	70
2	Mount Olive-Silverstone-Jamestown (2)	Low risk	8	5	10	9	3	5	40
3	Thistletown-Beaumond Heights (3)	Low risk	1	5	3	0	0	0	9
4	Rexdale-Kipling (4)	Low risk	0	1	3	2	1	0	7
5	Elms-Old Rexdale (5)	Low risk	1	4	1	3	2	1	12
6	Kingsview Village-The Westway (6)	Low risk	0	6	2	4	6	4	22
7	Willowridge-Martingrove-Richview (7)	Low risk	2	11	2	3	3	2	23
8	Humber Heights-Westmount (8)	Low risk	3	3	1	1	3	1	12
9	Edenbridge-Humber Valley (9)	Low risk	3	14	4	11	3	1	36
10	Princess-Rosethorn (10)	Low risk	3	4	3	3	6	5	24
11	Eringate-Centennial-West Deane (11)	Low risk	6	2	4	4	2	3	21
12	Markland Wood (12)	Low risk	31	6	1	3	5	9	55
13	Etobicoke West Mall (13)	Low risk	11	1	1	2	2	3	20
14	Islington-City Centre West (14)	Low risk	19	24	26	20	15	23	127
15	Kingsway South (15)	Low risk	12	14	14	8	7	7	62
16	Stonegate-Queensway (16)	Low risk	10	20	11	8	11	14	74
17	Mimico (includes Humber Bay Shores) (17)	Mid-low risk	27	33	40	62	101	94	357
18	New Toronto (18)	Low risk	7	12	6	9	35	8	77
19	Long Branch (19)	Low risk	14	8	10	17	35	13	97
20	Alderwood (20)	Low risk	5	2	3	4	9	5	28
21	Humber Summit (21)	Low risk	7	3	6	3	2	7	28
22	Humbermede (22)	Low risk	2	5	3	3	2	1	16
23	Pelmo Park-Humberlea (23)	Low risk	3	1	1	0	4	1	10
24	Black Creek (24)	Low risk	9	9	4	2	10	7	41
25	Glenfield-Jane Heights (25)	Low risk	10	28	10	13	8	8	77
26	Downsview-Roding-CFB (26)	Low risk	12	12	16	5	21	14	80
27	York University Heights (27)	Low risk	16	28	27	35	33	20	159
28	Rustic (28)	Low risk	0	2	3	0	0	2	7
29	Maple Leaf (29)	Low risk	0	1	1	2	1	0	5
30	Brookhaven-Amesbury (30)	Low risk	2	6	3	7	4	4	26
31	Yorkdale-Glen Park (31)	Low risk	13	16	11	13	14	22	89
32	Englemount-Lawrence (32)	Low risk	6	29	17	6	12	9	79
33	Clanton Park (33)	Low risk	7	5	2	5	10	8	37
34	Bathurst Manor (34)	Low risk	3	5	4	7	9	3	31
35	Westminster-Branson (35)	Low risk	8	11	15	11	4	12	61
36	Newtonbrook West (36)	Low risk	22	18	30	30	13	17	130
37	Willowdale West (37)	Low risk	8	14	49	22	9	13	115
38	Lansing-Westgate (38)	Low risk	5	5	16	7	4	8	45
39	Bedford Park-Nortown (39)	Low risk	11	7	21	10	20	20	89
40	St.Andrew-Windfields (40)	Low risk	8	1	7	5	7	4	32
41	Bridle Path-Sunnybrook-York Mills (41)	Low risk	3	6	6	5	3	6	29
42	Banbury-Don Mills (42)	Low risk	3	12	5	10	13	2	45
43	Victoria Village (43)	Low risk	3	3	1	1	3	1	12
44	Flemingdon Park (44)	Low risk	13	10	4	5	2	7	41
45	Parkwoods-Donalda (45)	Low risk	2	7	7	18	6	6	46
46	Pleasant View (46)	Low risk	1	0	2	2	1	4	10
47	Don Valley Village (47)	Low risk	7	7	2	4	9	4	33
48	Hillcrest Village (48)	Low risk	4	4	3	7	8	9	35
49	Bayview Woods-Steeles (49)	Low risk	2	2	1	2	1	2	10
50	Newtonbrook East (50)	Low risk	6	8	6	6	5	5	36
51	Willowdale East (51)	Low risk	22	29	48	33	23	25	180
52	Bayview Village (52)	Low risk	8	10	24	11	14	9	76
53	Henry Farm (53)	Low risk	4	2	2	3	1	6	18
54	O'Connor-Parkview (54)	Low risk	9	11	7	11	8	5	51
55	Thorncliffe Park (55)	Low risk	9	5	15	10	3	8	50
56	Leaside-Bennington (56)	Low risk	24	23	38	17	23	26	151
57	Broadview North (57)	Low risk	4	10	13	11	2	12	52
58	Old East York (58)	Low risk	4	2	1	5	3	0	15

61	Taylor-Massey (61)	Low risk	10	11	9	6	16	4	56
62	East End-Danforth (62)	Mid-low risk	43	43	51	38	45	27	247
63	The Beaches (63)	Mid-low risk	51	18	93	25	34	34	255
64	Woodbine Corridor (64)	Low risk	5	15	27	19	19	24	109
65	Greenwood-Coxwell (65)	Low risk	16	23	36	28	28	40	171
66	Danforth (66)	Low risk	9	10	24	10	22	28	103
67	Playter Estates-Danforth (67)	Low risk	21	11	19	15	12	12	90
68	North Riverdale (68)	Mid-low risk	28	27	30	26	51	36	198
69	Blake-Jones (69)	Low risk	13	14	12	16	6	18	79
70	South Riverdale (70)	Mid-high risk	70	93	128	116	139	70	616
71	Cabbagetown-South St.James Town (71)	Mid-low risk	59	67	65	49	84	58	382
72	Regent Park (72)	Mid-low risk	26	25	37	38	40	42	208
73	Moss Park (73)	Mid-high risk	102	111	118	138	133	117	719
74	North St.James Town (74)	Mid-low risk	45	55	61	66	35	53	315
75	Church-Yonge Corridor (75)	High risk	214	210	198	268	263	259	1412
76	Bay Street Corridor (76)	High risk	327	267	320	337	380	325	1956
77	Waterfront Communities-The Island (77)	High risk	298	259	416	437	462	399	2271
78	Kensington-Chinatown (78)	Mid-high risk	90	105	129	143	148	115	730
79	University (79)	Mid-high risk	105	111	135	145	135	76	707
80	Palmerston-Little Italy (80)	Mid-low risk	48	53	55	54	65	47	322
81	Trinity-Bellwoods (81)	Mid-low risk	41	42	69	65	62	59	338
82	Niagara (82)	Mid-high risk	95	122	124	144	182	209	876
83	Dufferin Grove (83)	Mid-low risk	24	42	42	46	30	46	230
84	Little Portugal (84)	Mid-low risk	38	40	41	83	54	65	321
85	South Parkdale (85)	Mid-low risk	26	27	31	42	48	75	249
86	Roncesvalles (86)	Mid-low risk	26	42	26	71	68	49	282
87	High Park-Swansea (87)	Mid-low risk	20	25	55	37	44	38	219
88	High Park North (88)	Mid-low risk	32	36	46	63	51	51	279
89	Runnymede-Bloor West Village (89)	Low risk	22	16	18	19	5	10	90
90	Junction Area (90)	Low risk	25	18	25	22	27	20	137
91	Weston-Pellam Park (91)	Low risk	16	13	4	6	0	9	48
92	Corso Italia-Davenport (92)	Low risk	13	7	9	12	5	12	58
93	Dovercourt-Wallace Emerson-Junction (93)	Mid-high risk	66	77	107	107	94	97	548
94	Wychwood (94)	Low risk	11	35	37	21	15	22	141
95	Annex (95)	Mid-high risk	96	112	122	137	126	150	743
96	Casa Loma (96)	Low risk	18	19	13	10	15	19	94
97	Yonge-St.Clair (97)	Low risk	10	19	17	12	14	24	96
98	Rosedale-Moore Park (98)	Mid-low risk	46	61	46	50	66	67	336
99	Mount Pleasant East (99)	Low risk	7	34	20	11	22	19	113
100	Yonge-Eglinton (100)	Low risk	14	23	21	25	15	17	115
101	Forest Hill South (101)	Low risk	9	10	9	8	16	8	60
102	Forest Hill North (102)	Low risk	17	41	13	9	7	20	107
103	Lawrence Park South (103)	Low risk	6	11	12	18	16	11	74
104	Mount Pleasant West (104)	Mid-low risk	20	31	36	44	45	63	239
105	Lawrence Park North (105)	Low risk	8	9	14	19	12	23	85
106	Humewood-Cedarvale (106)	Low risk	26	52	15	13	19	13	138
107	Oakwood Village (107)	Low risk	16	20	18	14	7	11	86
108	Briar Hill-Belgravia (108)	Low risk	19	9	7	7	12	3	57
109	Caledonia-Fairbank (109)	Low risk	6	3	5	3	3	0	20
110	Keelesdale-Eglinton West (110)	Low risk	3	4	4	5	3	2	21
111	Rockcliffe-Smythe (111)	Low risk	12	12	9	12	6	5	56
112	Beechborough-Greenbrook (112)	Low risk	1	5	1	0	1	1	9
113	Weston (113)	Low risk	7	5	7	8	16	7	50
114	Lambton Baby Point (114)	Low risk	7	13	6	4	7	10	47
115	Mount Dennis (115)	Low risk	3	1	3	5	2	2	16
116	Steeles (116)	Low risk	1	4	1	6	0	0	12
117	L'Amoreaux (117)	Low risk	2	3	7	9	11	7	39
118	Tam O'Shanter-Sullivan (118)	Low risk	2	2	2	5	4	6	21
119	Wexford/Maryvale (119)	Low risk	18	10	8	5	6	6	53
120	Clairlea-Birchmount (120)	Low risk	23	20	7	16	10	6	82

123	Cliffcrest (123)	Low risk	6	1	7	9	1	4	28
124	Kennedy Park (124)	Low risk	1	6	5	9	12	3	36
125	Ionview (125)	Low risk	0	3	0	3	3	4	13
126	Dorset Park (126)	Low risk	6	4	4	3	2	1	20
127	Bendale (127)	Low risk	11	9	11	25	2	4	62
128	Agincourt South-Malvern West (128)	Low risk	1	5	3	8	5	3	25
129	Agincourt North (129)	Low risk	5	6	2	0	1	0	14
130	Milliken (130)	Low risk	2	1	0	6	2	3	14
131	Rouge (131)	Low risk	10	5	13	7	22	8	65
132	Malvern (132)	Low risk	7	8	8	8	4	7	42
133	Centennial Scarborough (133)	Low risk	1	2	6	4	6	3	22
134	Highland Creek (134)	Low risk	14	11	15	2	0	3	45
135	Morningside (135)	Low risk	5	3	3	0	1	1	13
136	West Hill (136)	Low risk	16	16	26	13	7	11	89
137	Woburn (137)	Low risk	12	8	17	17	8	6	68
138	Eglinton East (138)	Low risk	1	3	2	5	1	5	17
139	Scarborough Village (139)	Low risk	2	2	2	7	0	2	15
140	Guildwood (140)	Low risk	3	2	6	2	1	2	16

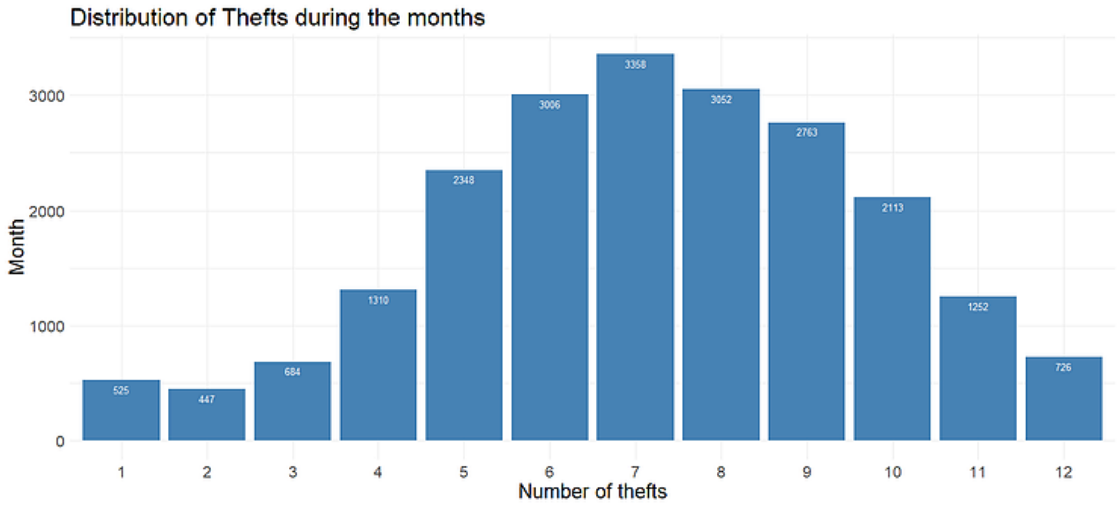
2. Top 20 neighborhoods

In this table, the neighborhoods are sorted in descending order of the number of stolen bikes.
The majority of thefts happen downtown, in the entertainment ***district*** and close to ***Yonge-Dundas Square.*** In addition, it seems that the more you move away from these zones, the more the number of thefts decreases.
Neighborhoods close to the ***entertainment district*** have a ***mid-high risk*** associated, while further neighborhoods just low or mid-low.

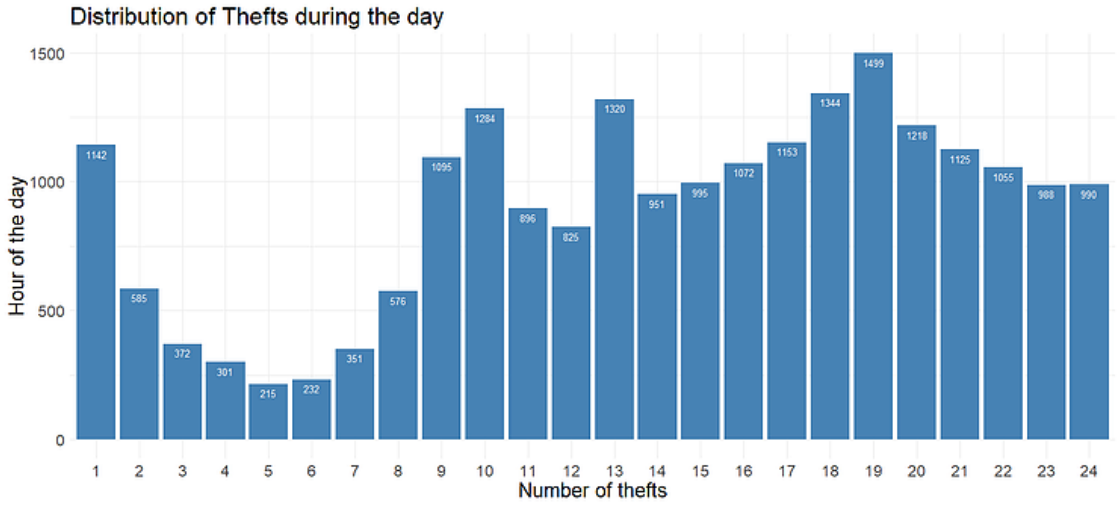
ID	AREA_NAME	THEFT RISK	2014	2015	2016	2017	2018	2019	Total
77	Waterfront Communities-The Island (77)	High risk	298	259	416	437	462	399	2271
76	Bay Street Corridor (76)	High risk	327	267	320	337	380	325	1956
75	Church-Yonge Corridor (75)	High risk	214	210	198	268	263	259	1412
82	Niagara (82)	Mid-high risk	95	122	124	144	182	209	876
95	Annex (95)	Mid-high risk	96	112	122	137	126	150	743
78	Kensington-Chinatown (78)	Mid-high risk	90	105	129	143	148	115	730
73	Moss Park (73)	Mid-high risk	102	111	118	138	133	117	719
79	University (79)	Mid-high risk	105	111	135	145	135	76	707
70	South Riverdale (70)	Mid-high risk	70	93	128	116	139	70	616
93	Dovercourt-Wallace Emerson-Junction (93)	Mid-high risk	66	77	107	107	94	97	548
71	Cabbagetown-South St.James Town (71)	Mid-low risk	59	67	65	49	84	58	382
17	Mimico (includes Humber Bay Shores) (17)	Mid-low risk	27	33	40	62	101	94	357
81	Trinity-Bellwoods (81)	Mid-low risk	41	42	69	65	62	59	338
98	Rosedale-Moore Park (98)	Mid-low risk	46	61	46	50	66	67	336
80	Palmerston-Little Italy (80)	Mid-low risk	48	53	55	54	65	47	322
84	Little Portugal (84)	Mid-low risk	38	40	41	83	54	65	321
74	North St.James Town (74)	Mid-low risk	45	55	61	66	35	53	315
86	Roncesvalles (86)	Mid-low risk	26	42	26	71	68	49	282
88	High Park North (88)	Mid-low risk	32	36	46	63	51	51	279

3. Exploratory analysis

Let's start plotting the distribution of the thefts in each month. Maybe you didn't know it, but Toronto is really cold in the winter, so I think they are more likely to happen in summer rather than in winter.
Let's have a look. In each bar, I wrote the number of thefts that occurred during the relative month.

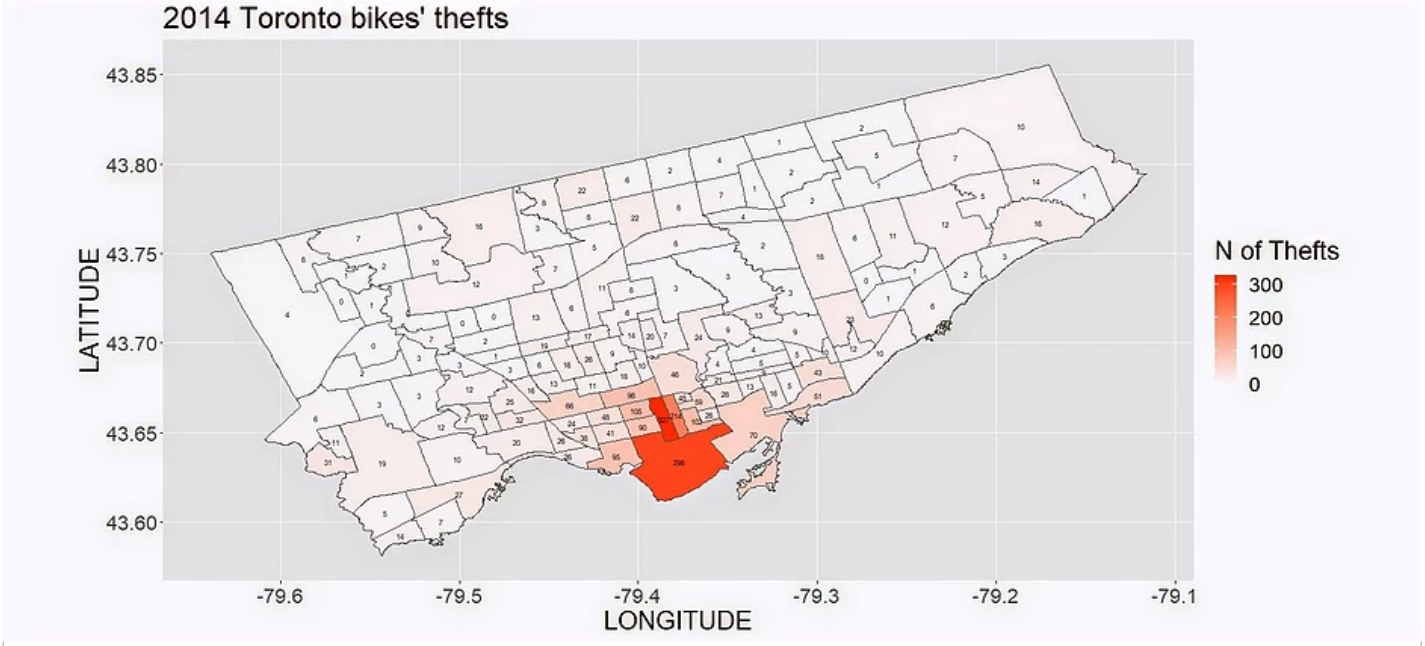


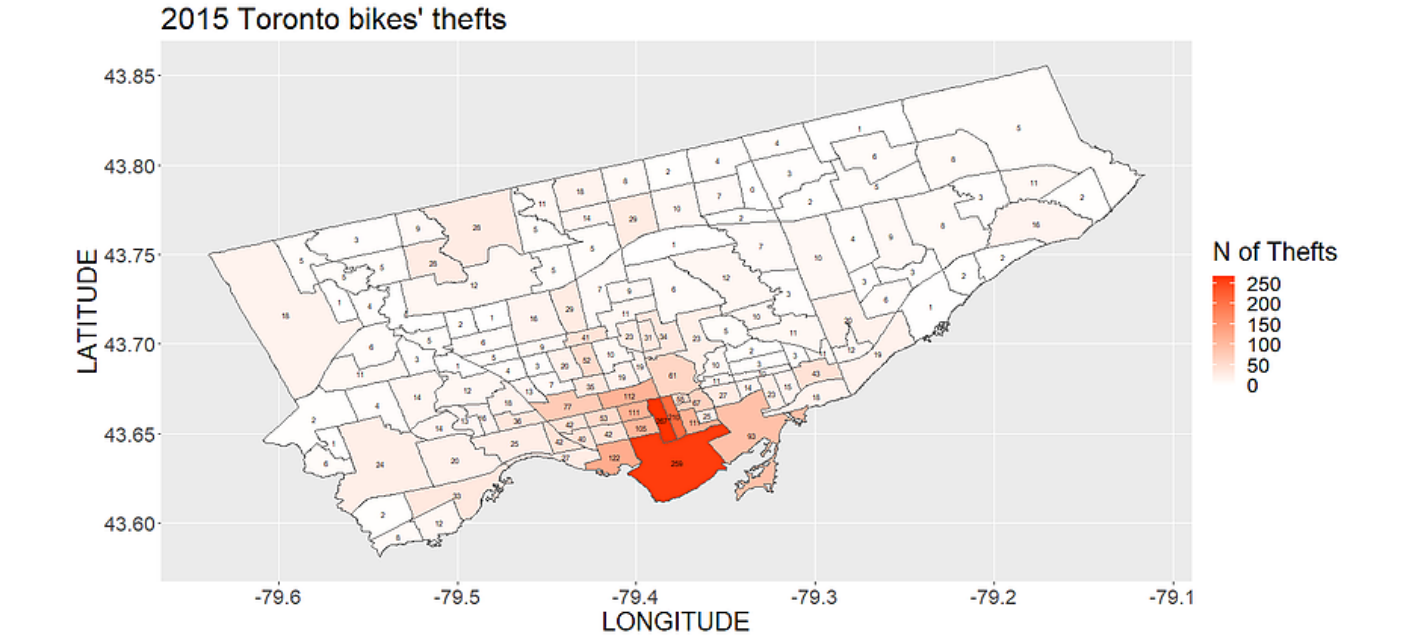
It seems that the thefts have a normal distribution and that most crimes happen during the summer (when it's hot and people are more likely to go out with their own bike for a ride). The number on top of the bar is the number of thefts that occurred during that specific hour in the last 6 years.

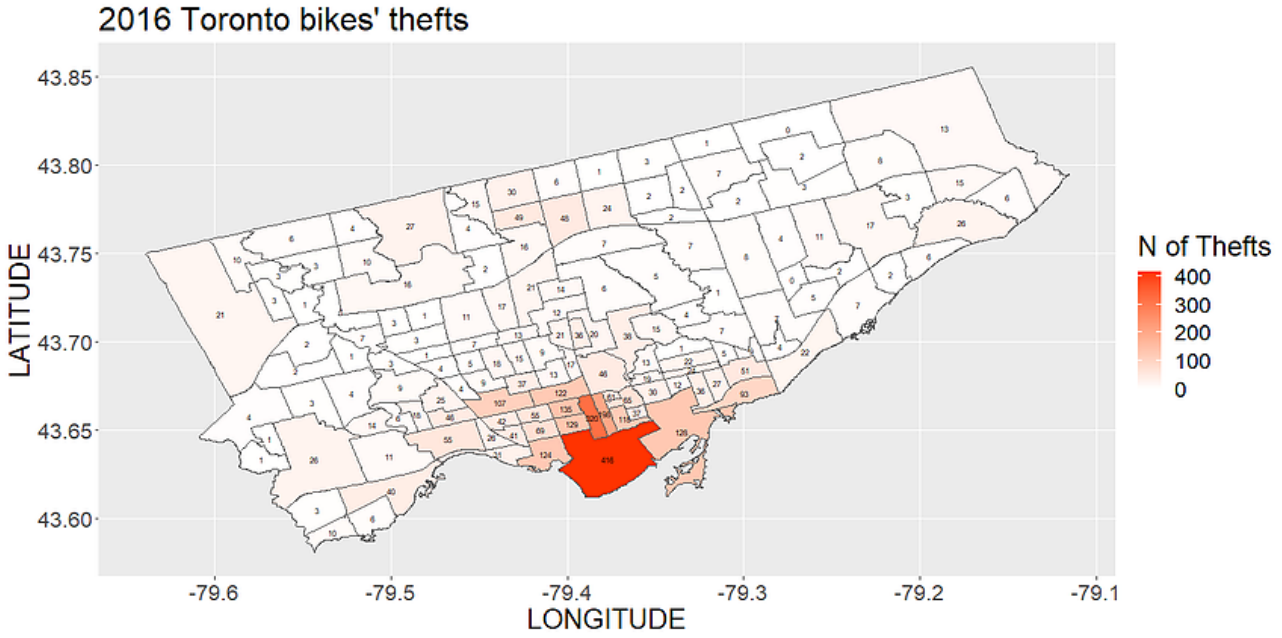


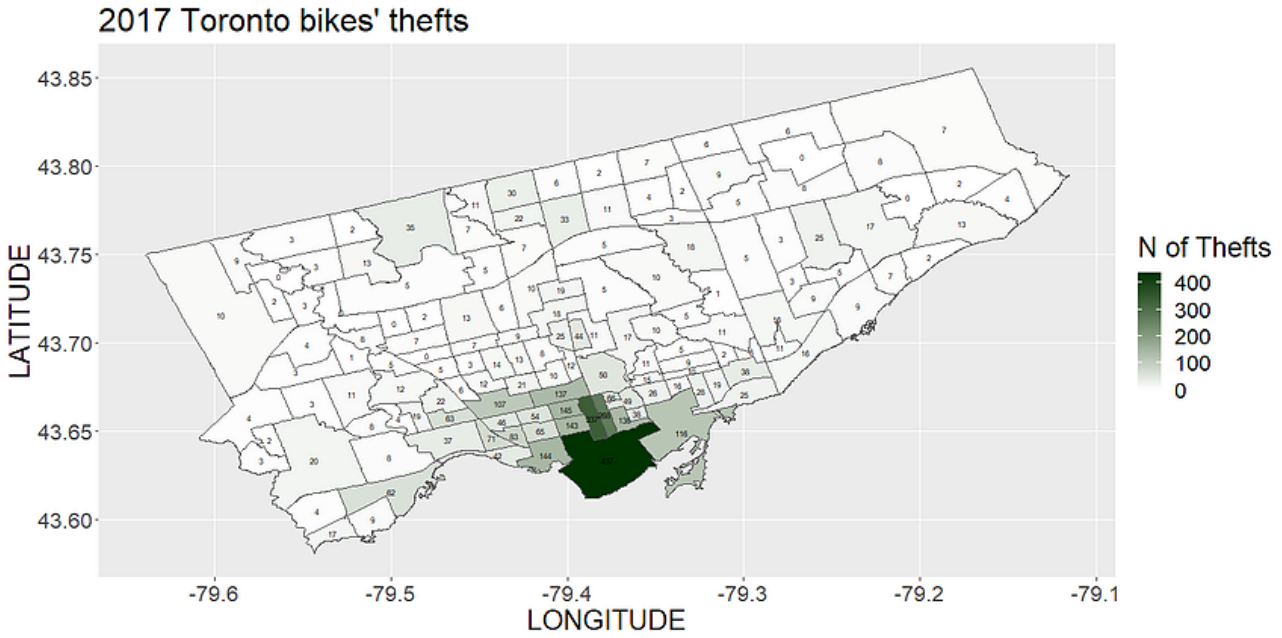
The graph of the hourly distribution shows that most thefts happen in the evening.

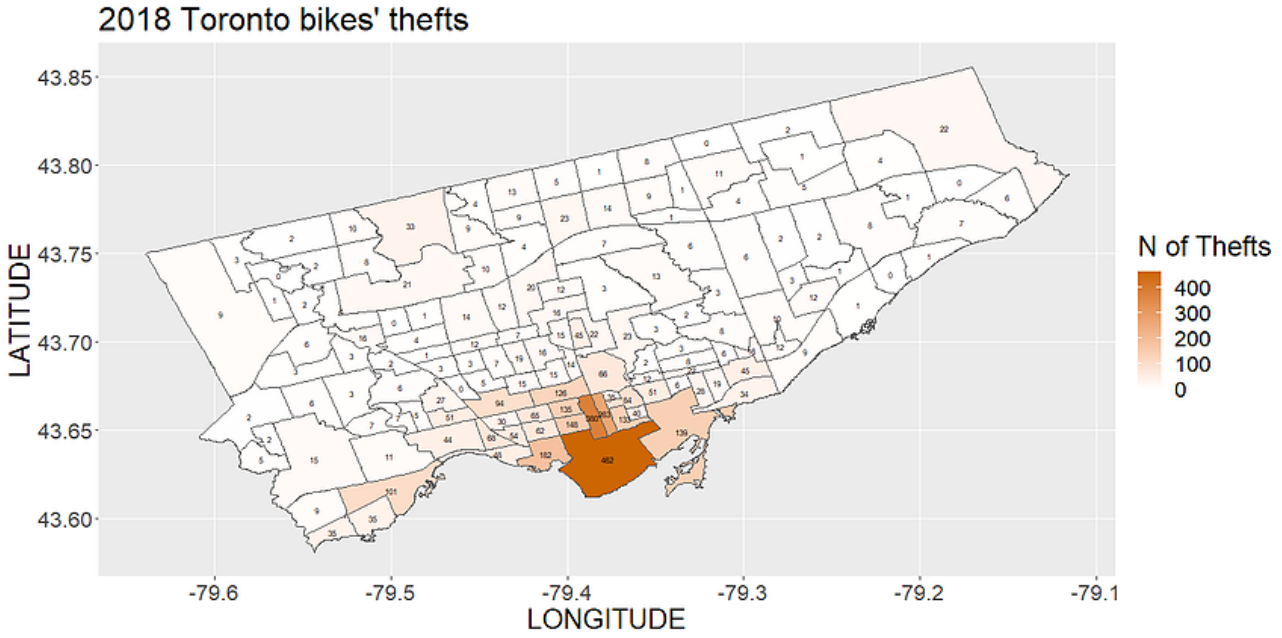
Let's have a look at the number of thefts per neighborhood from 2014 to 2019.
I wrote the number of thefts that occurred each year in each neighborhood area.
Click on the images to zoom.

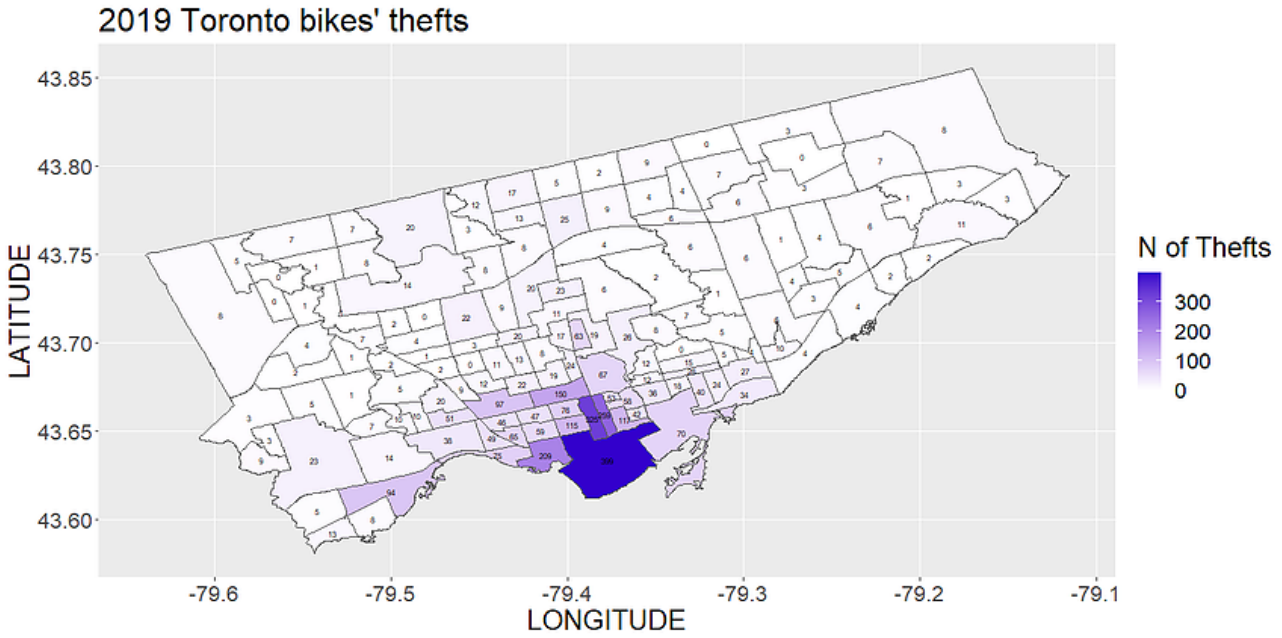








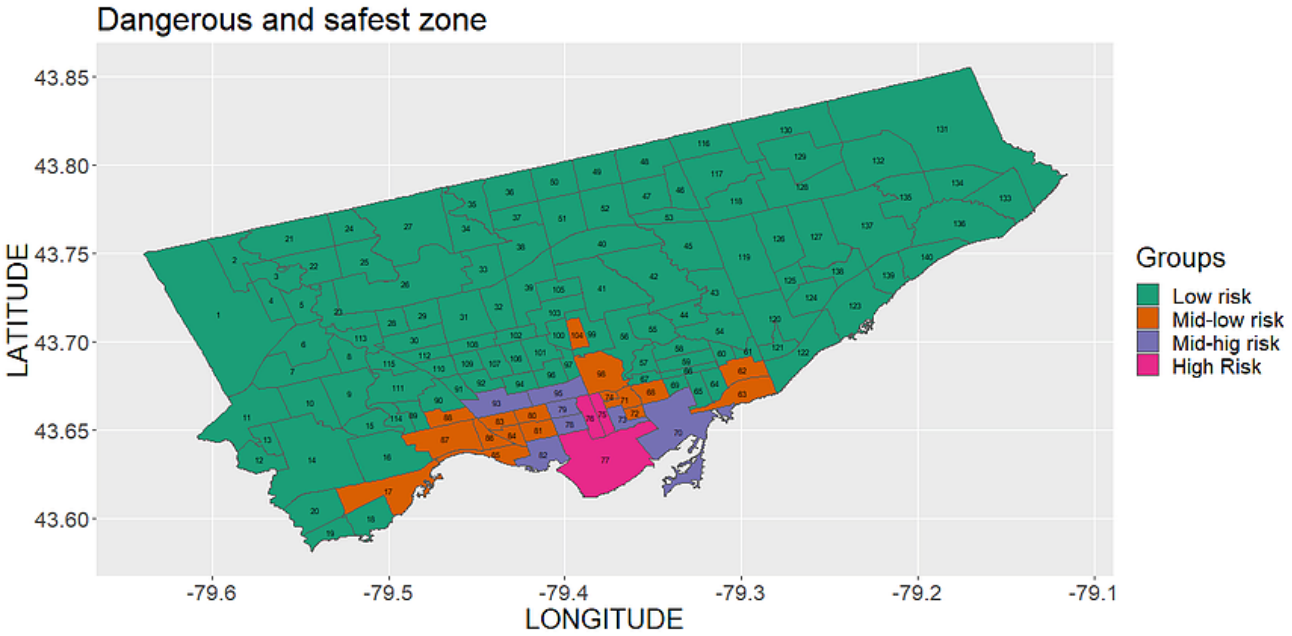


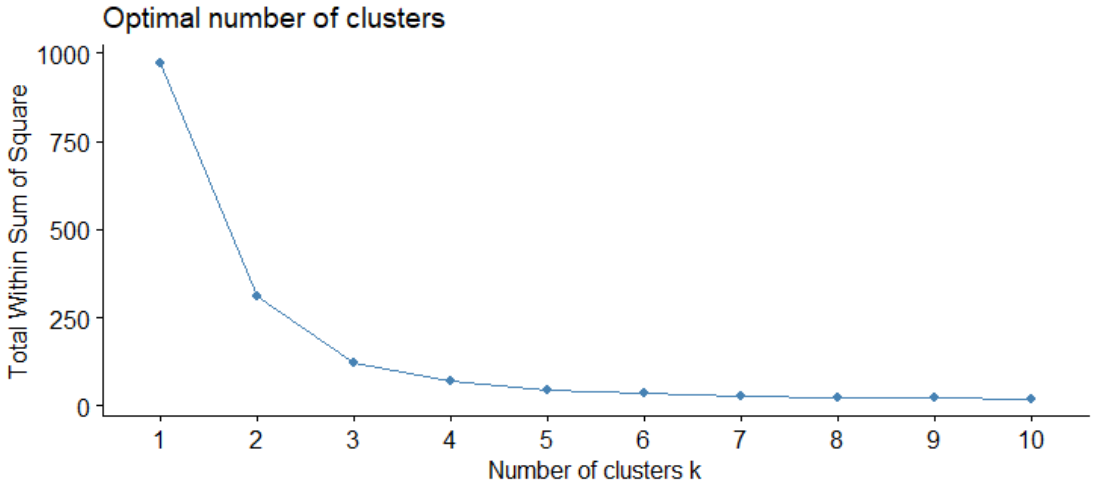


You can notice that most of them happen downtown and in the university/campus area. So if you leave your bike there, watch out!

4. Clustering

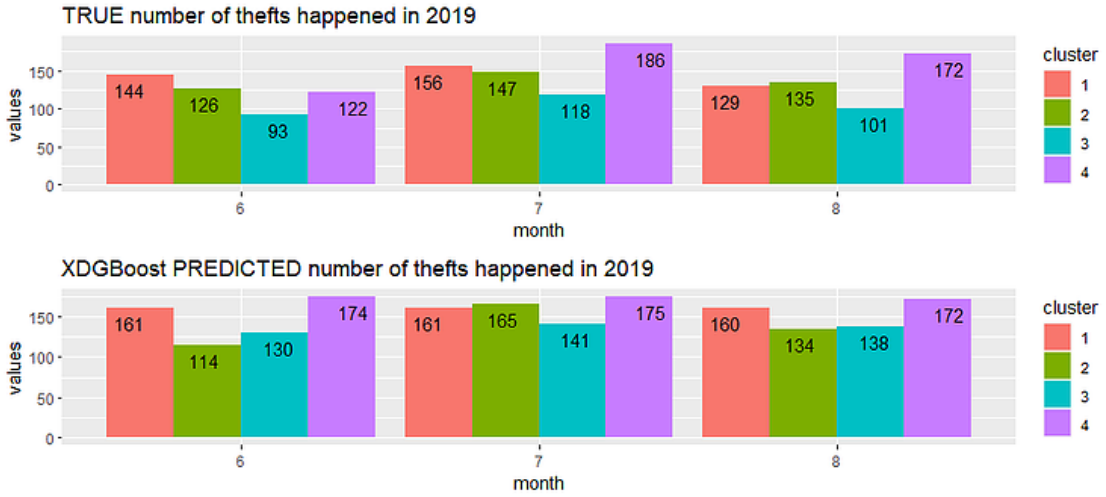
We can notice that there are areas in which thefts happen more frequently. We can create categories according to the number of thefts per year.
To do so we are going to use an unsupervised machine learning method called "hierarchical clustering". The elbow method shows us the optimal number of clusters is 4.
Here are the results.





5. XGBoost model predictions

Now let's try to predict the number of thefts that will happen in summer 2019. Unfortunately, we don't have 2020 data, so we cannot predict them. in addition, due to COVID-19, the number of riders around the town will be way less than before, so it means also fewer thefts.
Let's see the performance of the model.



Here let's compare the prediction with the true values.
In the last row of each table you can notice the difference in percentage.
When the number is positive, it means the model is overestimating the thefts, viceversa if negative it is underestimating. In the bottom right cell there is the total comparison.

JUNE	June CL 1	June CL 2	June CL 3	June CL 4	
PREDICTED	161	114	130	174	579
TRUE	144	126	93	122	485
DIFF	17	-12	37	52	94
%	3.51%	-2.47%	7.63%	10.72%	19.38%

JULY	July CL 1	July CL 2	July CL 3	July CL 4	
PREDICTED	161	165	141	175	642
TRUE	156	147	118	186	607
DIFF	5	18	23	-11	35
%	1.03%	3.71%	4.74%	-2.27%	7.22%

AUGUST	August CL 1	August CL 2	August CL 3	August CL 4	
PREDICTED	160	134	138	172	604
TRUE	129	135	101	172	537
DIFF	31	-1	37	0	67
%	6.39%	-0.21%	7.63%	0.00%	13.81%

SUMMER	CL 1	CL 2	CL 3	CL 4	
PREDICTED	482	413	409	521	1825
TRUE	429	408	312	480	1629
DIFF	53	5	97	41	196
%	3.25%	0.31%	5.95%	2.52%	12.03%

June has a high bias in estimations (almost 20%), while July has only a 7% error rate.
The overall model performance is not so great, with an error rate of 12.03 %.
To improve this result we could use some feature engineering and add, for example, a variable for the season or related to the major crimes (there is a dataset on the Toronto open data portal about it), but we are not going to do that.

Hope you enjoyed it! And watch out your bike!

Leonardo,
The pizza Statistician

P.s. Here you can download the script

R_bikes_script.txt
TXT • 21KB

bike thefts | R | ggplot2 | XGBoost | machine learning

R

112 views 0 comments 15

The Pizza Statistician