A Data-driven Approach to Eliminating Pedestrian Collisions in Toronto

Jason Kim

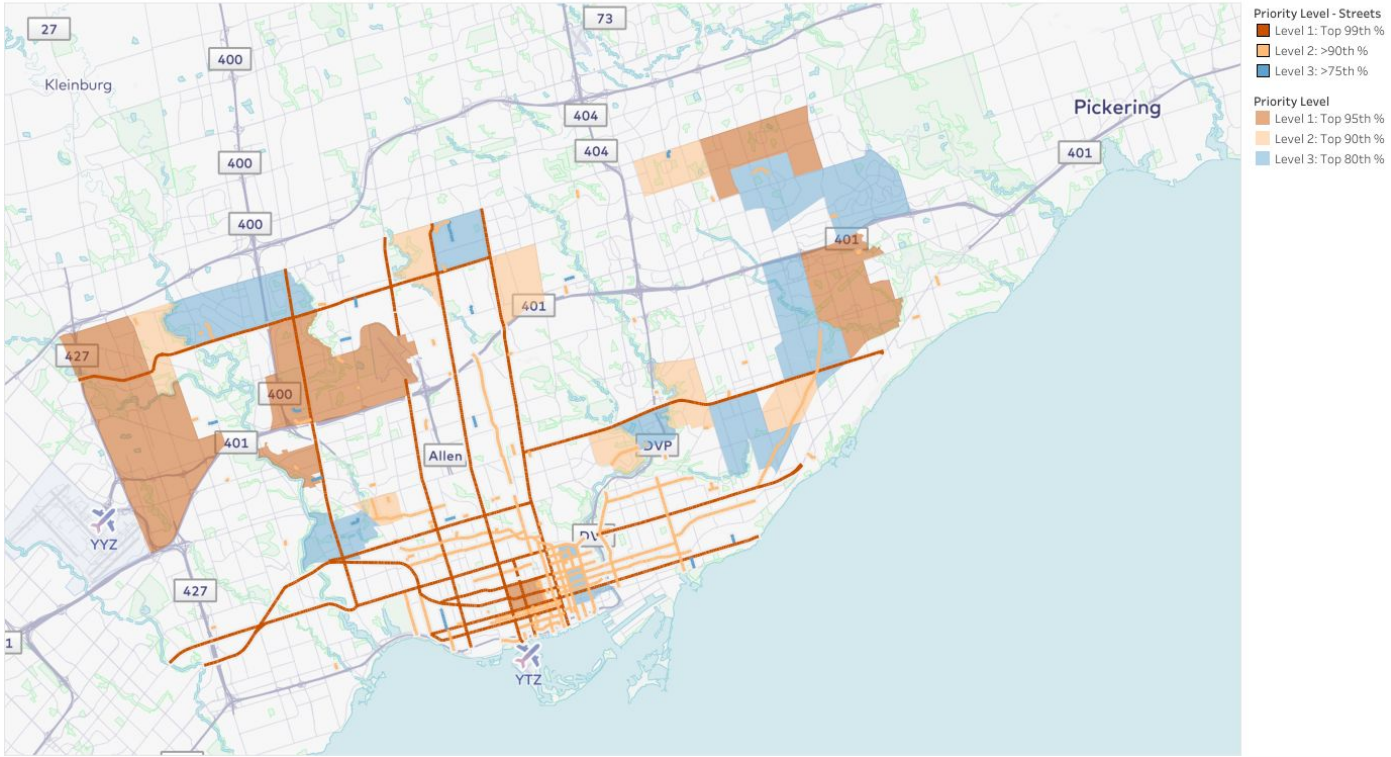
jason2.kim@ryerson.ca
https://github.com/jasonukim/capstone-repo

Context

- Just a few months ago, the City of Toronto approved $22 million in additional funding for accelerating its Vision Zero Road Safety Plan
- The goal of this plan is to eliminate pedestrian deaths from car collisions by 2021
- 2018 is one of the worst years for pedestrian and cyclist fatalities, exceeding 2017’s peak year of 162 deaths or major hospitalizations
- Politically sensitive topic due to perceptions:
  - “War against the car”
  - Collisions are caused by irresponsible pedestrians rather than drivers

Previous Work

Toronto High Injury Network



Research Question

- To identify and prioritize zones within Toronto that have a high risk of pedestrian collisions so that safety improvements could be more efficiently implemented by the City

Methodology

- Joined multiple datasets to the collision dataset to increase level of detail about factors that affect collision density
- Used various clustering algorithms in order to identify high collision zones
  - Kernel Density Estimation (KDE)
  - K-means Clustering
  - Density-based Spatial Clustering of Applications with Noise (DBSCAN)
- Used Average Silhouette to test the “goodness of fit” for clusters
- Random Forest to classify unseen collisions into clusters and check Variable Importance

Tools Used

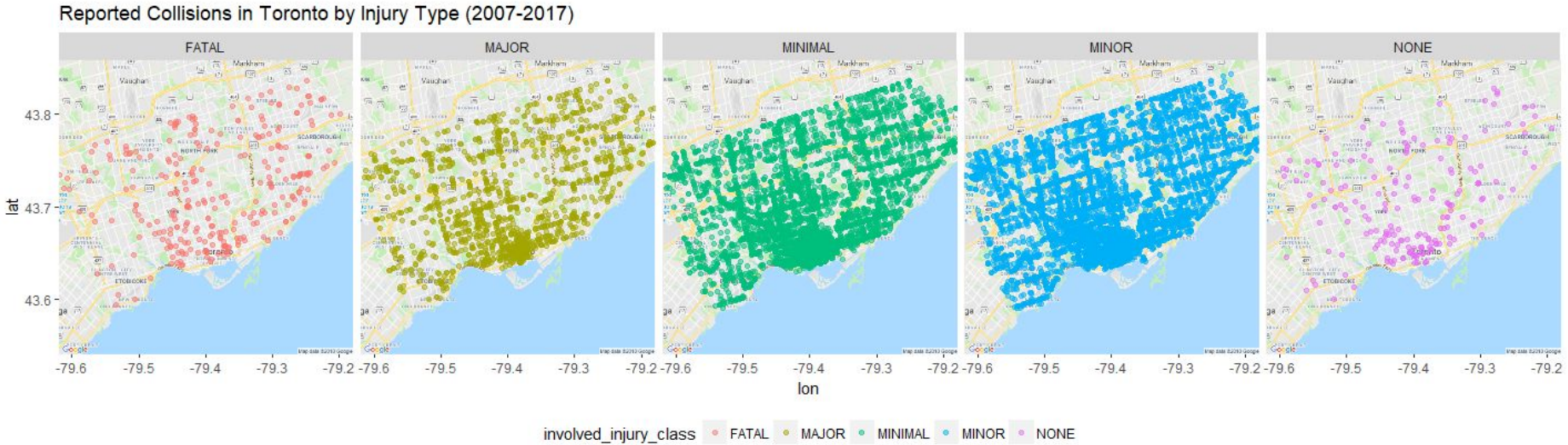
- R
  - General: Caret, corrplot, sqldf, mlbench, MASS, sp
  - Mapping: Maptools, ggmap, rgeos
  - Visualization: Ggplot2, knitr
  - Clustering, Silhouette, and Classification: Factoextra, NbClust, dbscan, cluster, RandomForest
  - Parallel Processing: doSNOW
- QGIS
  - For spatial joins on shapefiles (a special kind of dataset containing geo data)

Dataset Details

- Over 10,000 rows of collisions
- 88 attributes from 9 different open data sources
- Geospatial, demographic, economic, health, language, SES, etc.
- Full data dictionary:

https://github.com/jasonukim/capstone-repo/blob/master/Datasets/Data%20Dictionary.md

Exploratory Data Analysis


Reported Collisions in Toronto by Injury Type (2007-2017)
FATAL
MAJOR
MINIMAL
MINOR
NONE
lat
43.8
43.7
43.6
-79.6
-79.5
-79.4
-79.3
-79.2
lon
involved_injury_class
FATAL
MAJOR
MINIMAL
MINOR
NONE

Neighbourhood	Collisions
Waterfront Communities-The Island (77)	608
Bay Street Corridor (76)	567
Church-Yonge Corridor (75)	367
Downsview-Roding-CFB (26)	286
Islington-City Centre West (14)	286

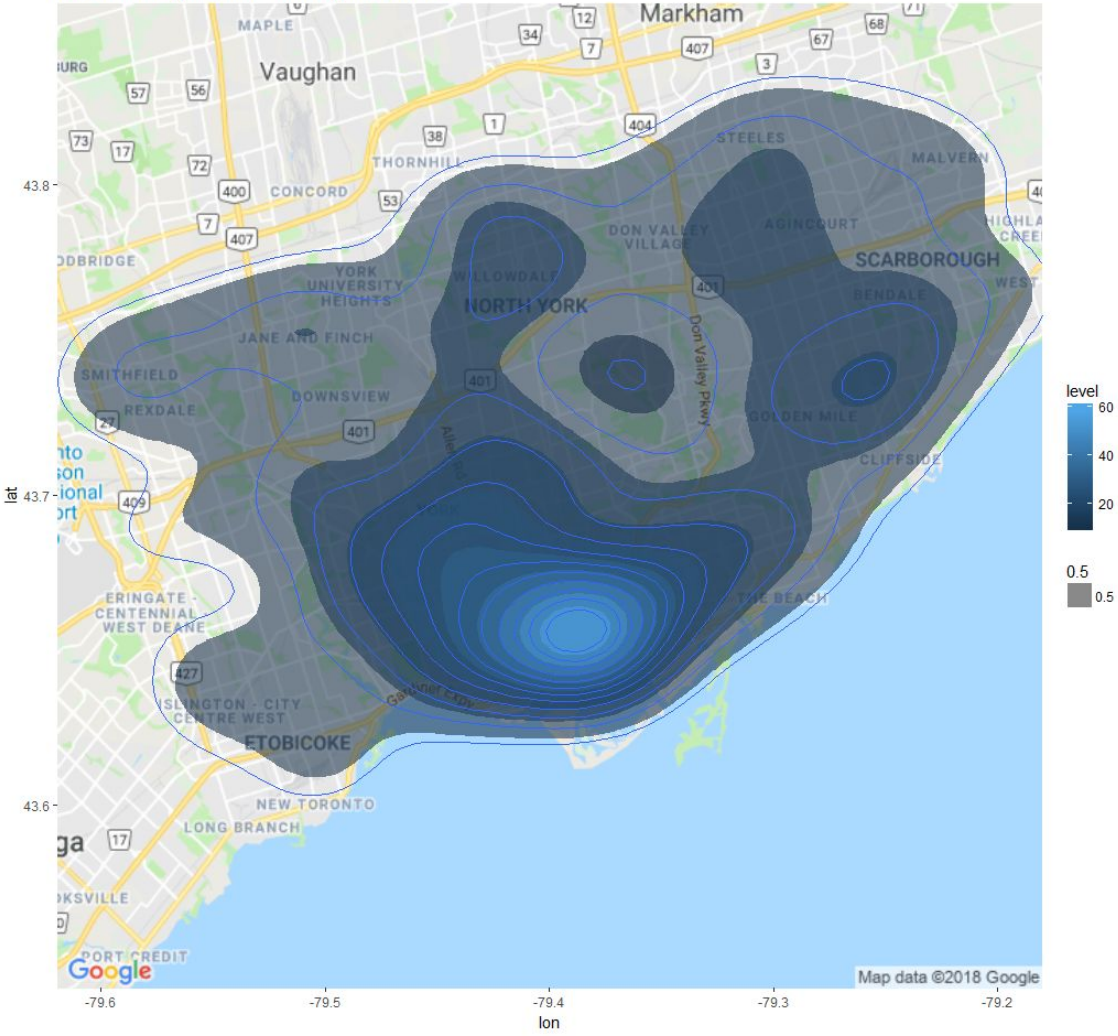
Street	Collisions
YONGE ST	555
DUNDAS ST W	371
BATHURST ST	366
BLOOR ST W	326
EGLINTON AVE E	319

Road Type	Collisions
MAJOR ARTERIAL	10318
MINOR ARTERIAL	2645
COLLECTOR	1294
LOCAL	1276

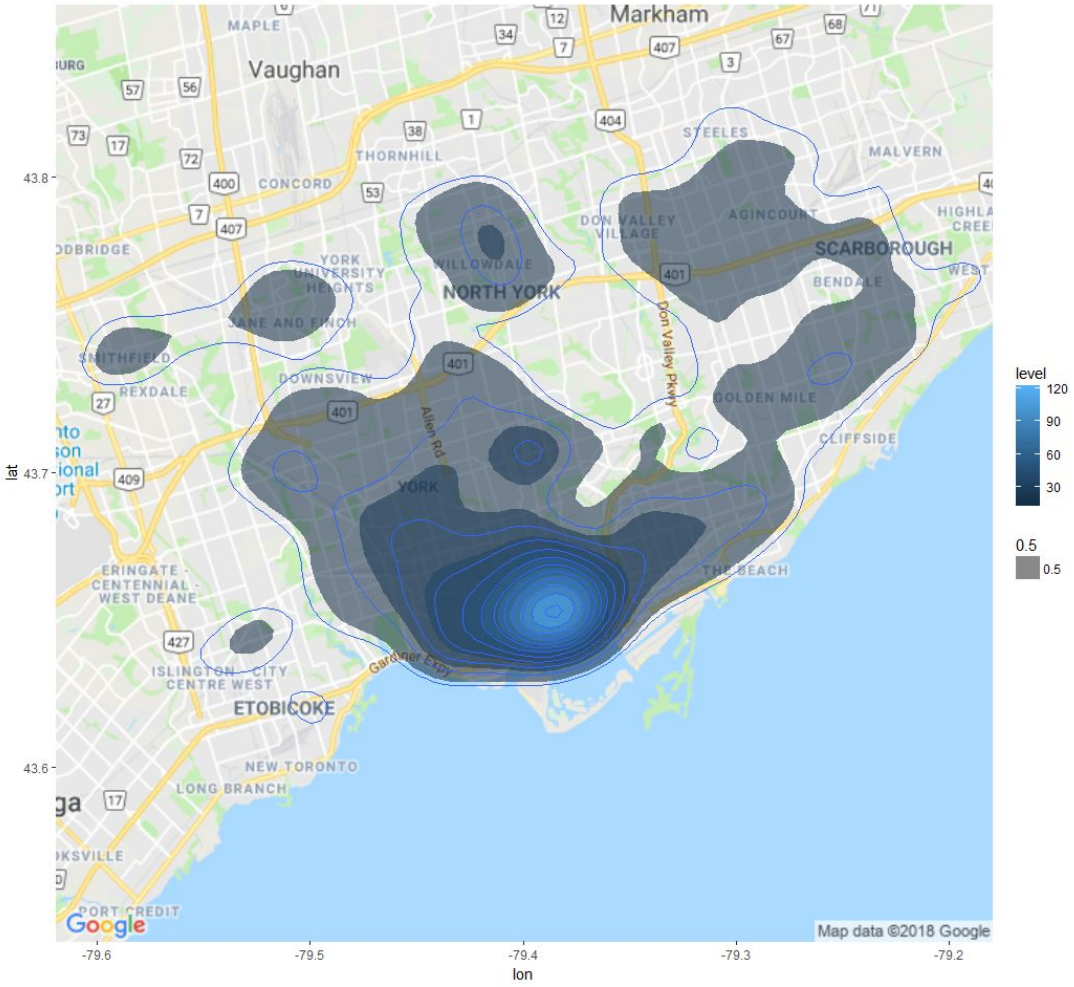
Location	Collisions
INTERSECTION	12186
MID-BLOCK	3304

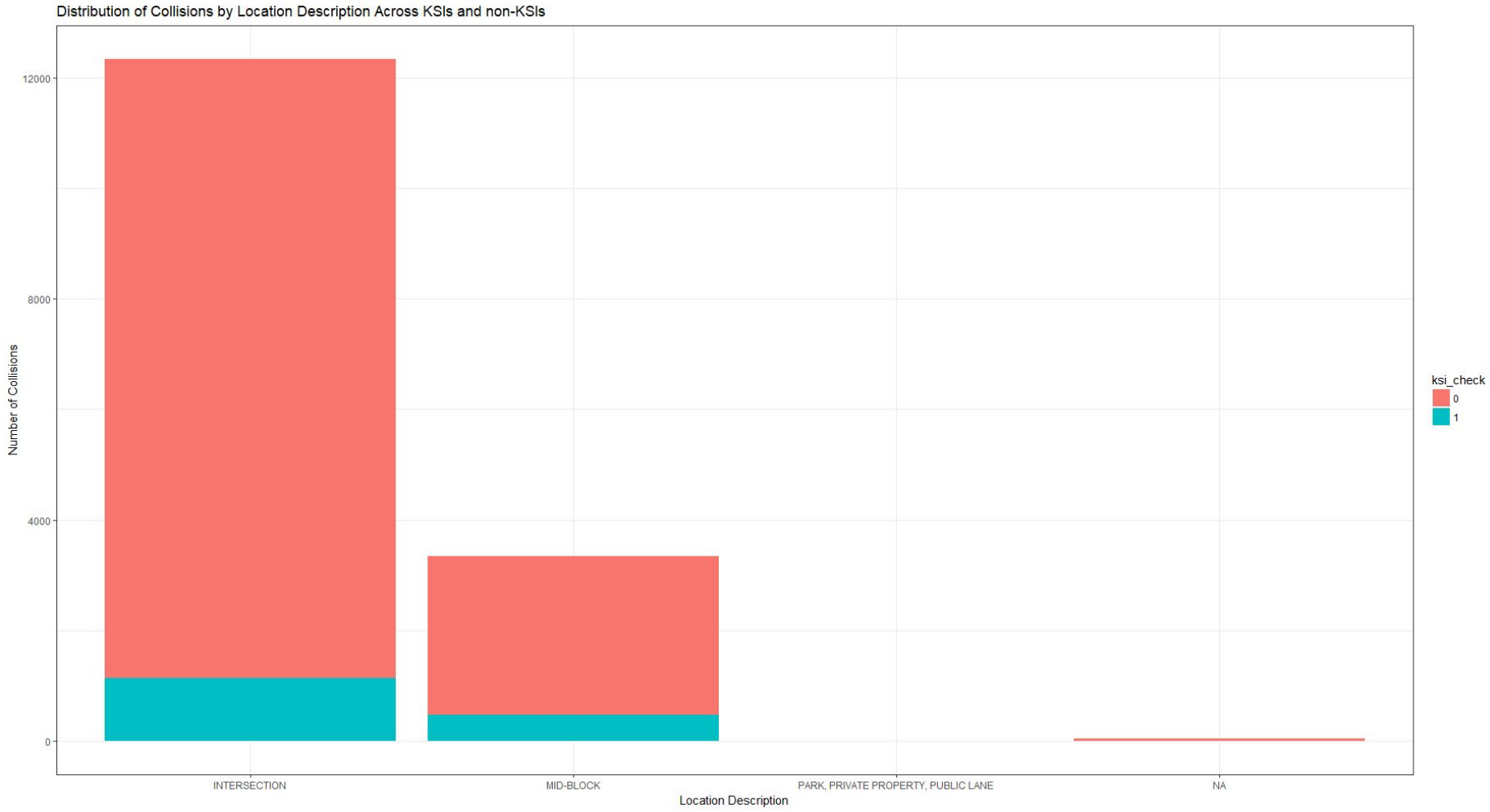
Visibility Condition	Collisions
CLEAR	12311
RAIN	2563
SNOW	457
OTHER	72
FREEZING RAIN	41
FOG, MIST, SMOKE, DUST	40
DRIFTING SNOW	29

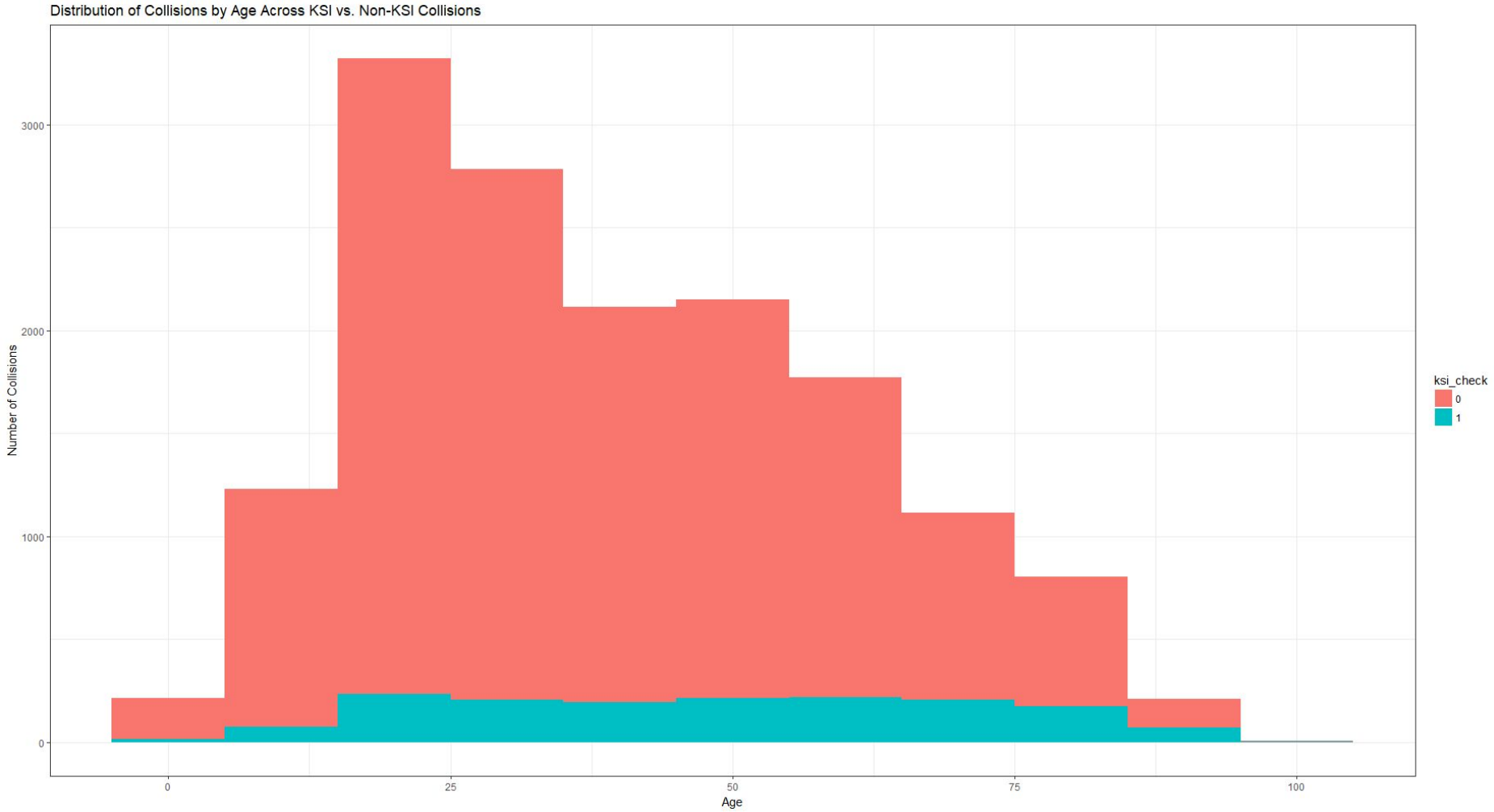
Kernel Density Estimation of High KSI Zones



Kernel Density Estimation of High Non-KSI Zones

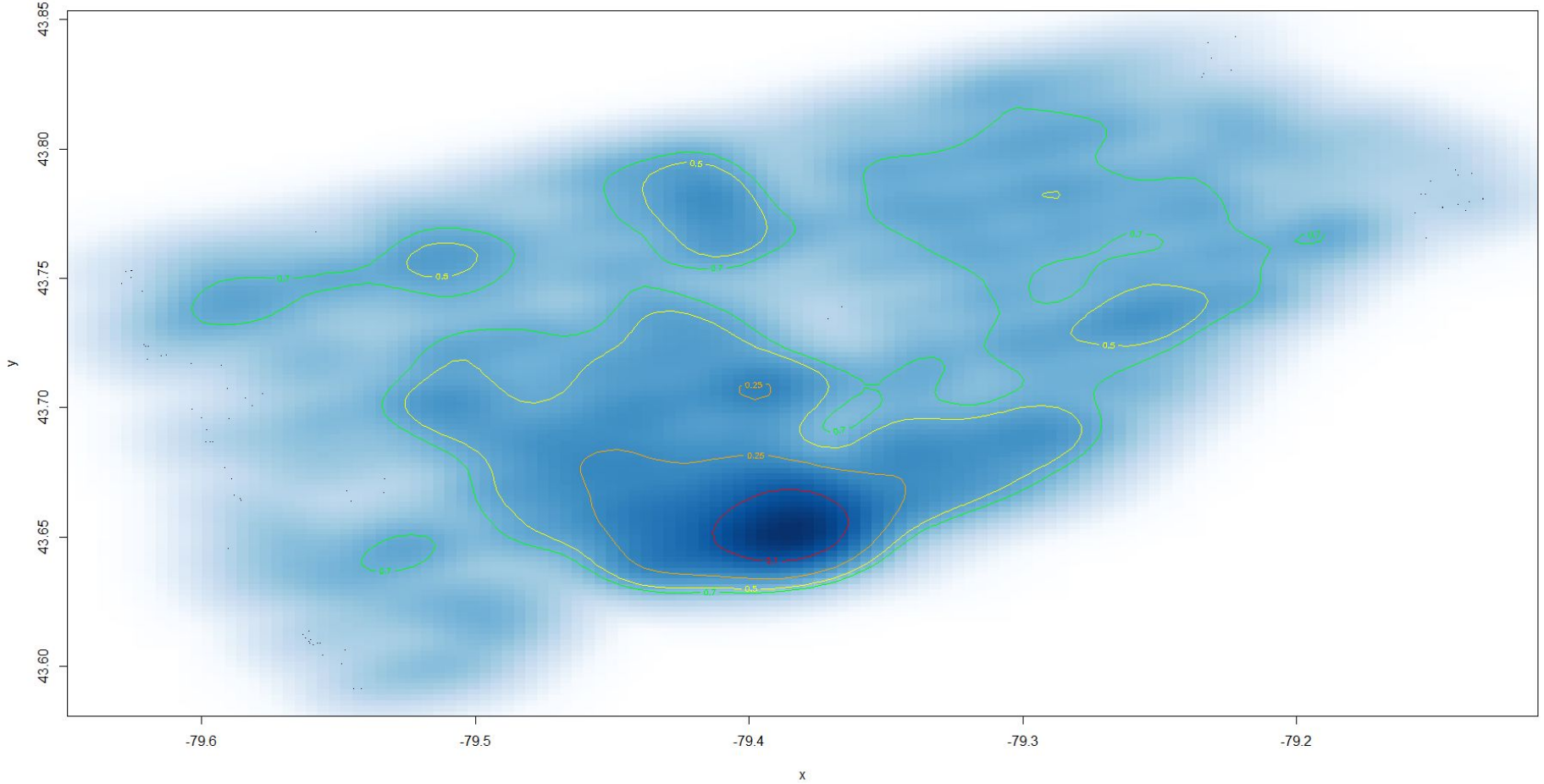



Distribution of Collisions by Location Description Across KSIs and non-KSIs
12000
8000
4000
0
Number of Collisions
INTERSECTION
MID-BLOCK
PARK, PRIVATE PROPERTY, PUBLIC LANE
NA
Location Description
ksi_check
0
1

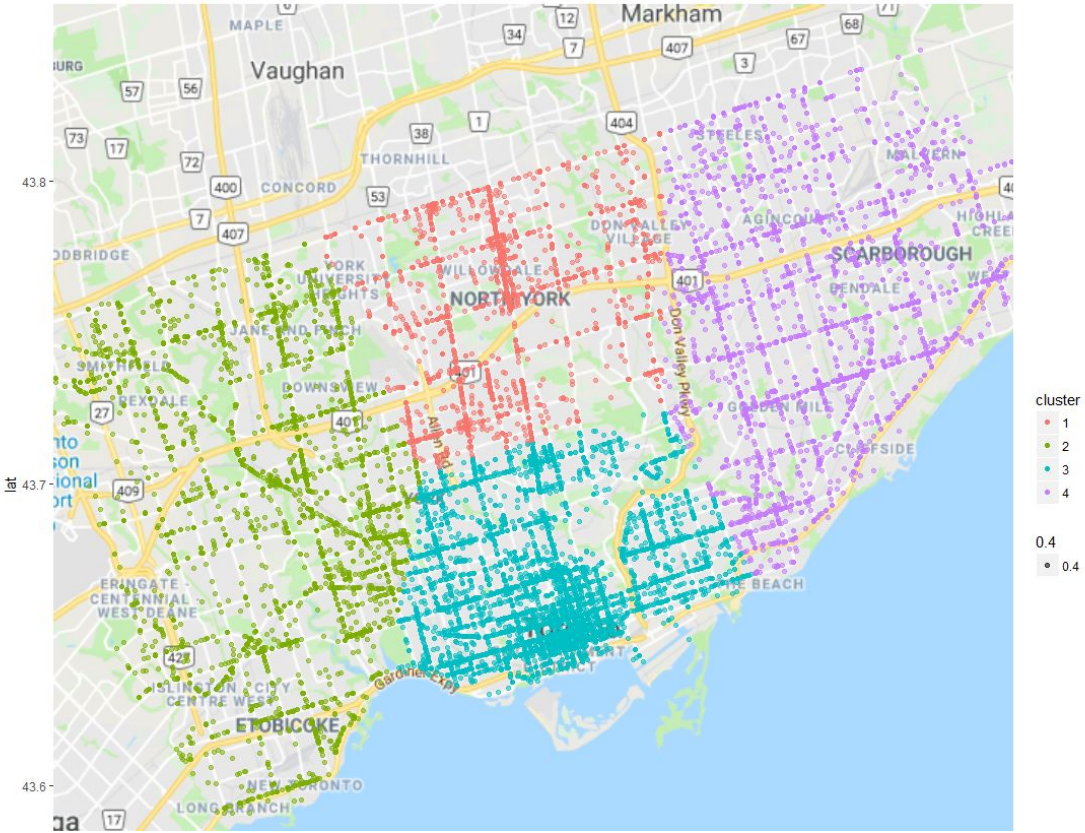

Distribution of Collisions by Age Across KSI vs. Non-KSI Collisions
Number of Collisions
3000
2000
1000
0
0
25
50
75
100
Age
ksi_check
0
1

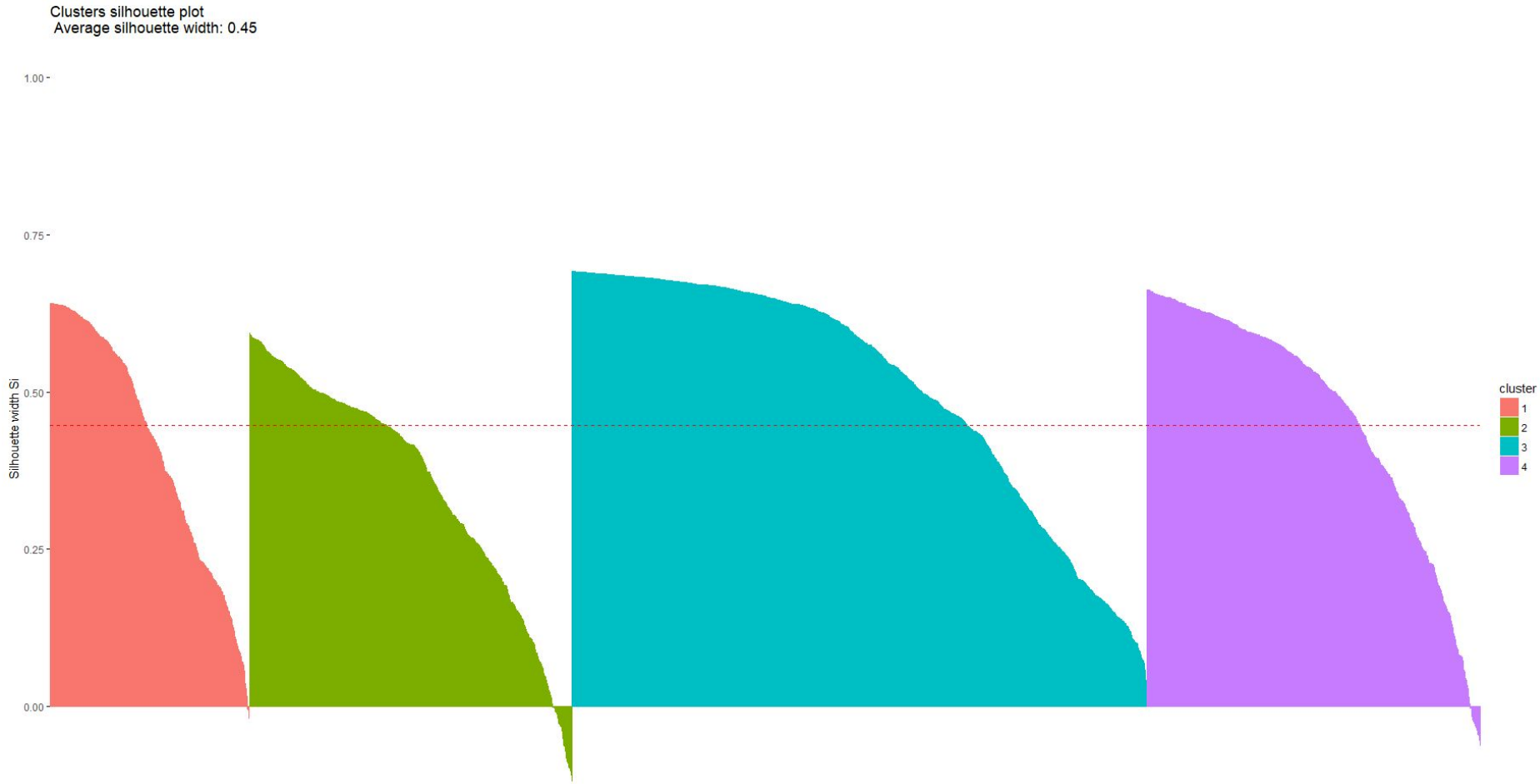
Clustering by Collision Location Only

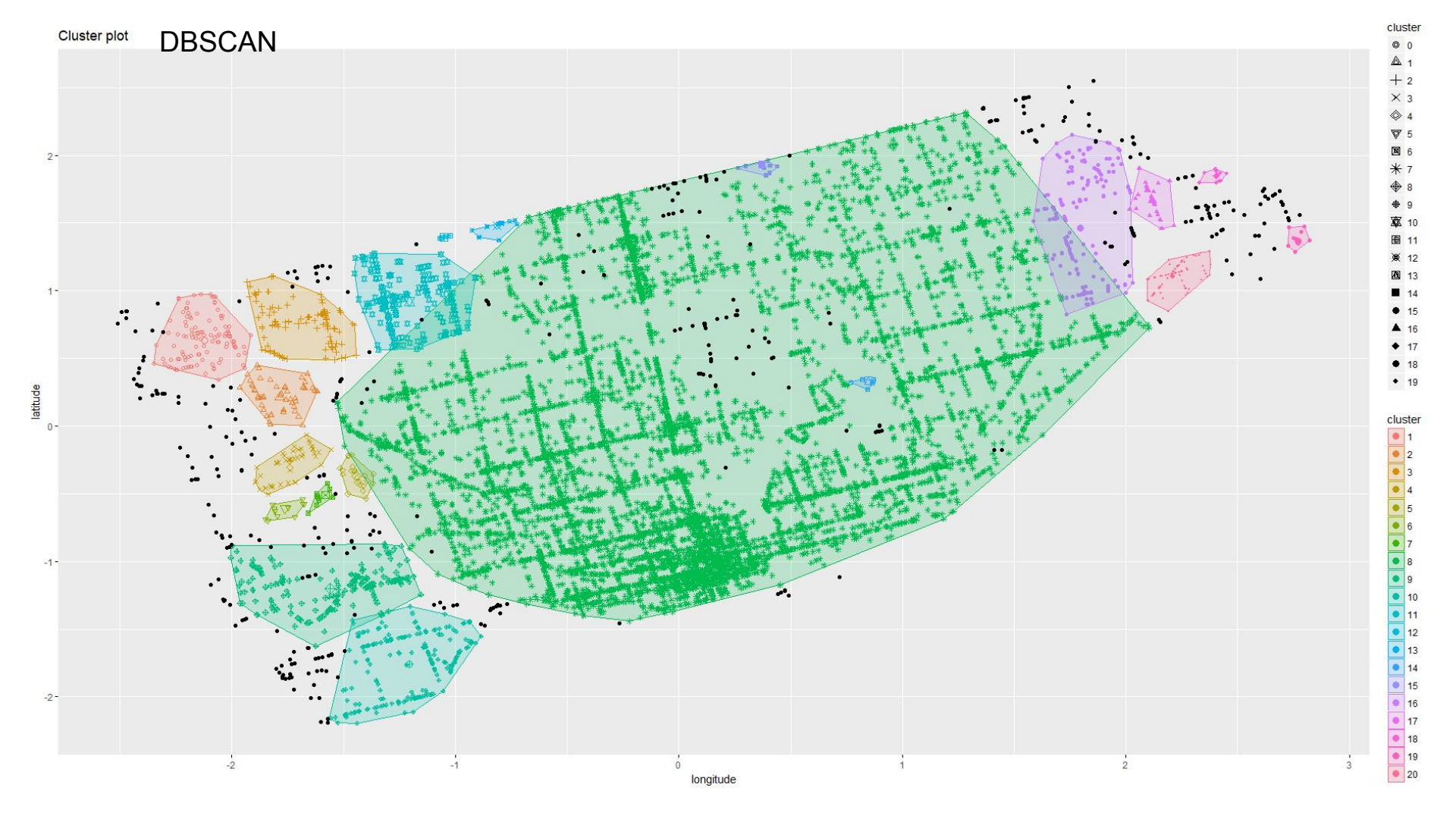
KDE

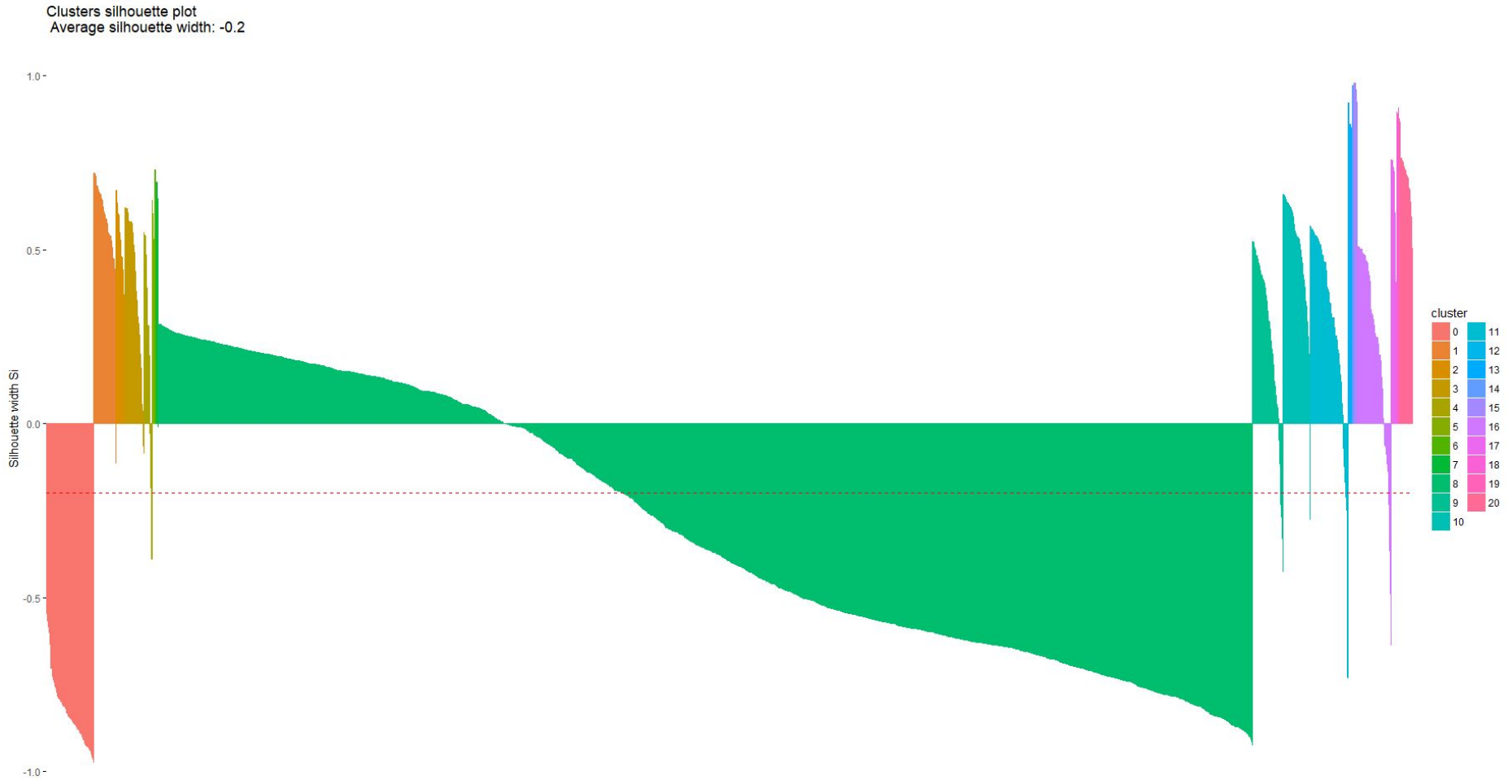


K-Means Clustering



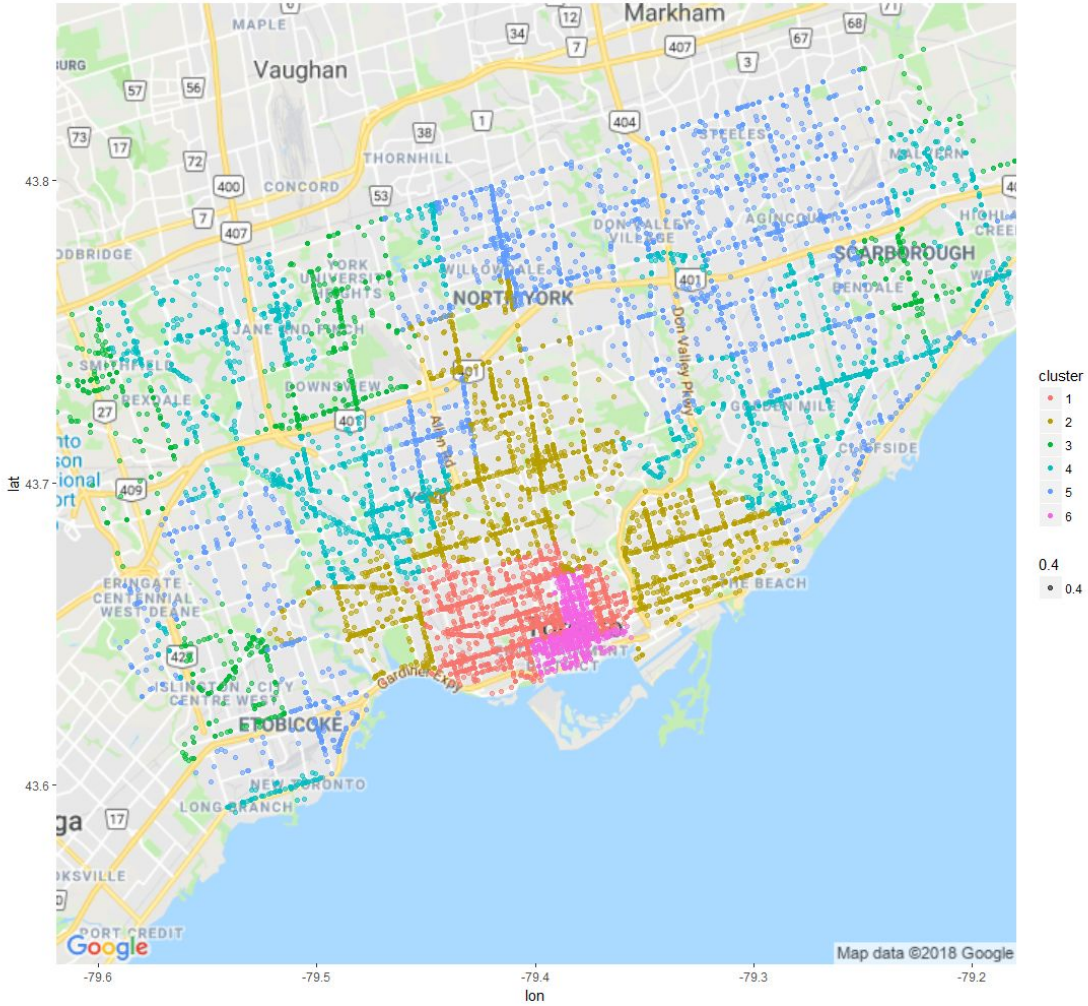

Clusters silhouette plot
Average silhouette width: 0.45
1.00
0.75
0.50
0.25
0.00
Silhouette width Si
cluster
1
2
3
4

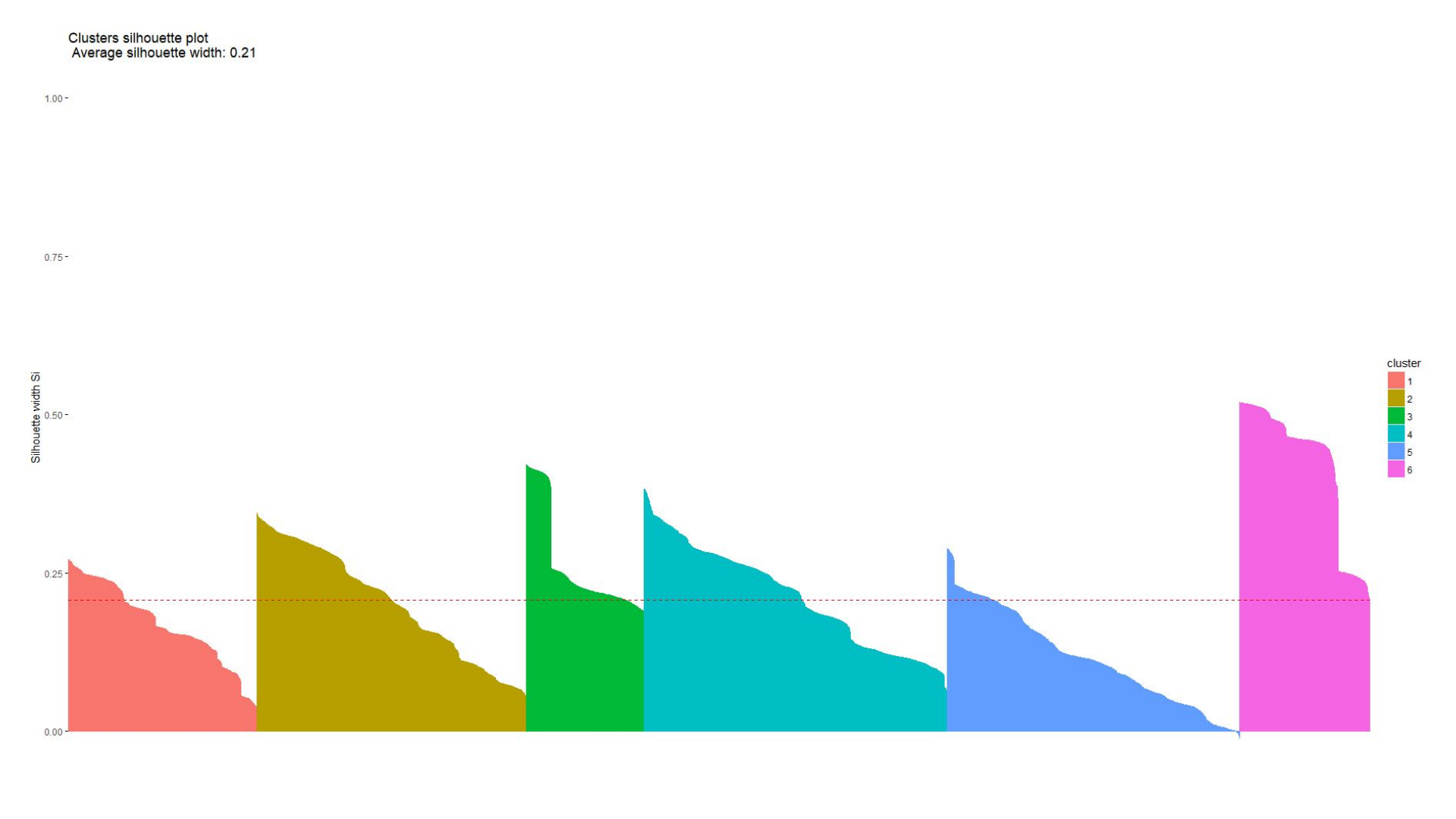

Cluster plot
DBSCAN
cluster
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
cluster
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
latitude
2
1
0
-1
-2
longitude
-2
-1
0
1
2
3


Clusters silhouette plot
Average silhouette width: -0.2
Silhouette width Si
1.0
0.5
0.0
-0.5
-1.0
cluster
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20

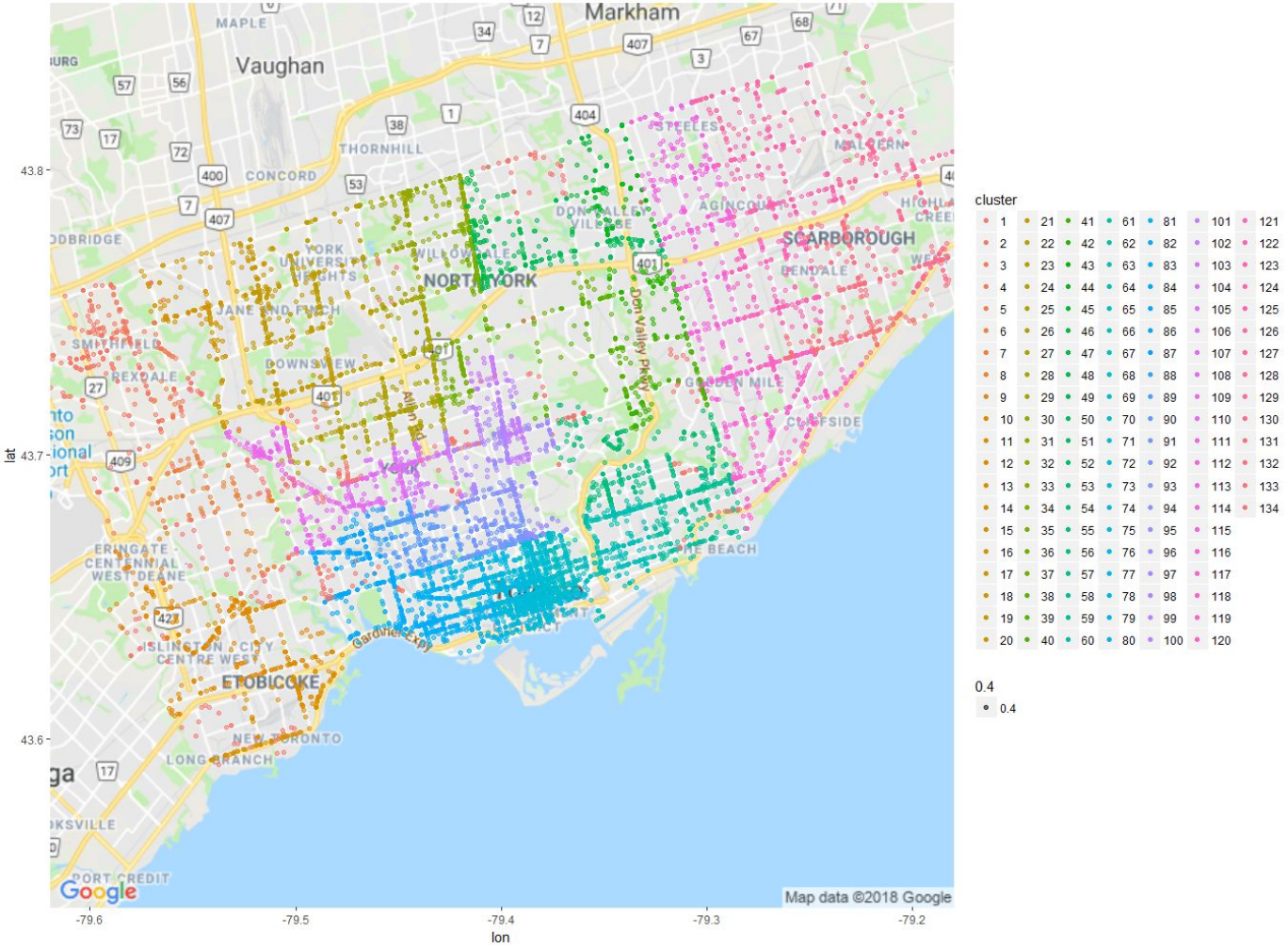
Clustering by Normalized Numerical Variables

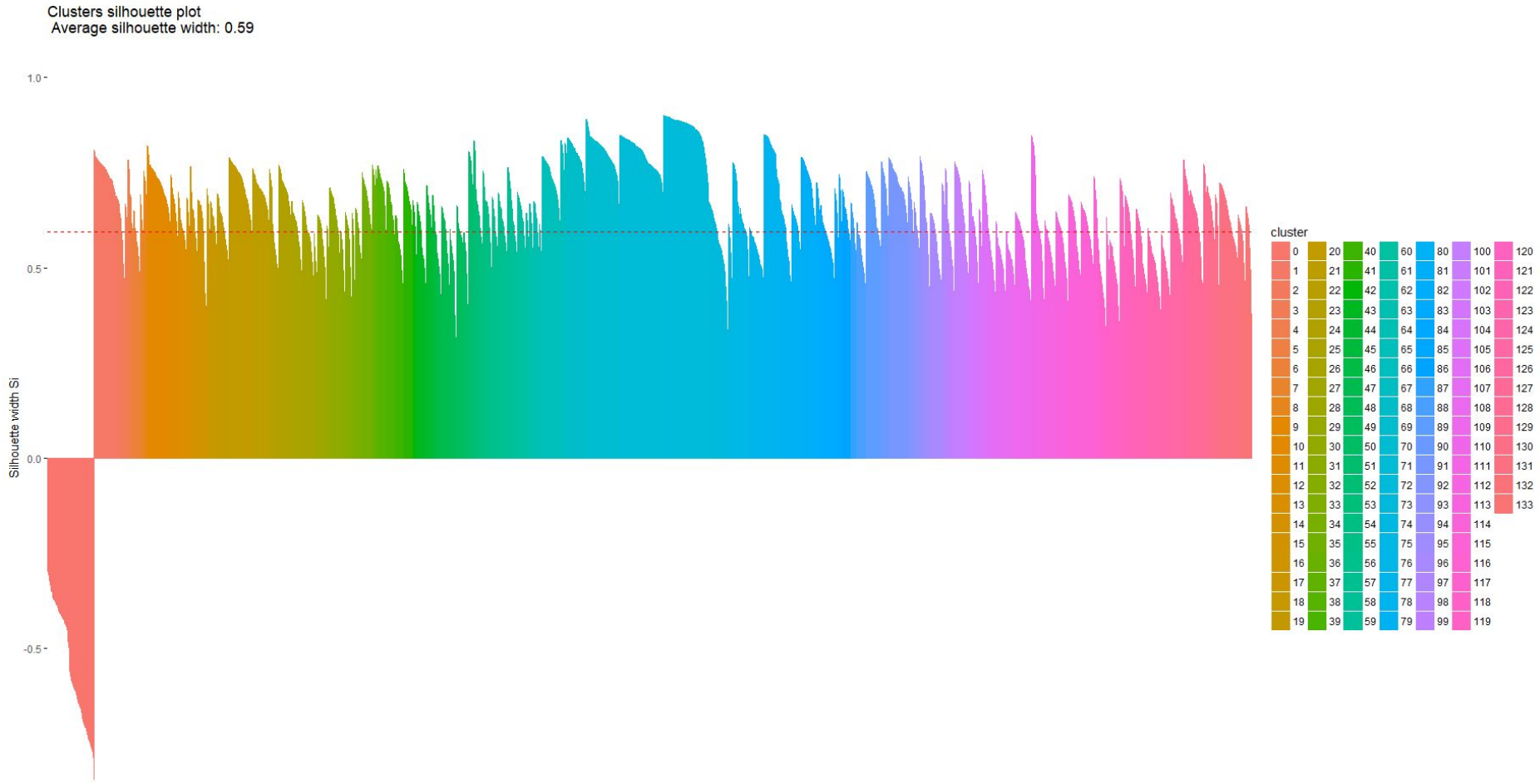
K-means

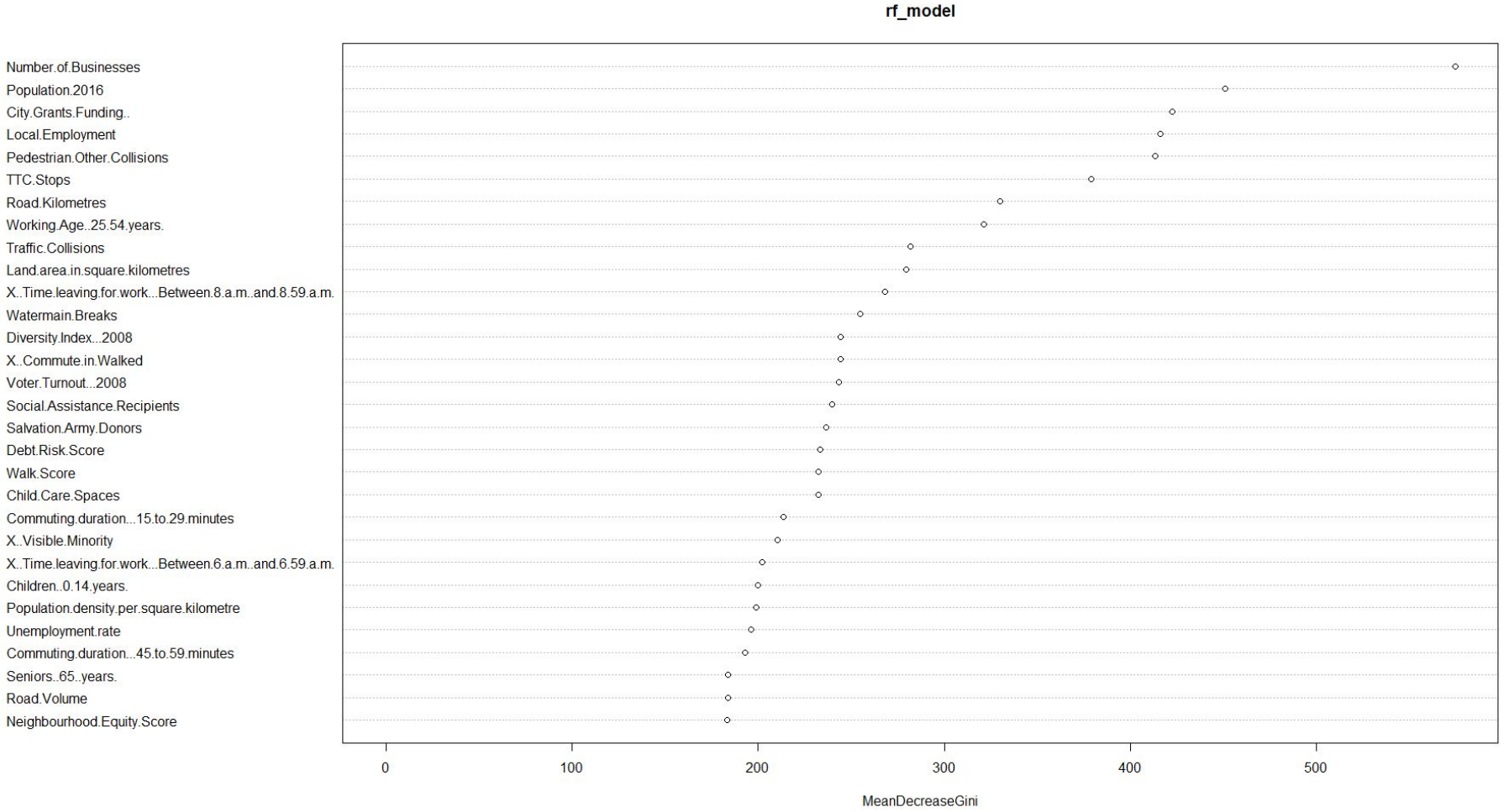

Markham
Vaughan
THORNHILL
CONCORD
NORTH YORK
SCARBOROUGH
Don Valley Pkwy
Allen Rd
Gardiner Expy
ROBICOKE
NEW TORONTO
LONG BRANCH
PORT CREDIT
Google
Map data ©2018 Google
lat
43.8
43.7
43.6
lon
-79.6
-79.5
-79.4
-79.3
-79.2
cluster
1
2
3
4
5
6
0.4
0.4


Clusters silhouette plot
Average silhouette width: 0.21
1.00
0.75
0.50
0.25
0.00
Silhouette width Si
cluster
1
2
3
4
5
6

DBSCAN


Markham
Vaughan
MAPLE
THORNHILL
CONCORD
NORTH YORK
SCARBOROUGH
AGINCOURT
DOWNSVIEW
ETOBICOKE
LONG BRANCH
NEW TORONTO
PORT CREDIT
Map data ©2018 Google
Google
lat
43.8
43.7
43.6
lon
-79.6
-79.5
-79.4
-79.3
-79.2
cluster
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40
41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60
61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80
81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100
101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120
121 122 123 124 125 126 127 128 129 130 131 132 133 134
0.4
0.4


Clusters silhouette plot
Average silhouette width: 0.59
1.0
0.5
0.0
-0.5
Silhouette width Si
cluster
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39
40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59
60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79
80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99
100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119
120 121 122 123 124 125 126 127 128 129 130 131 132 133


rf_model
Number.of.Businesses
Population.2016
City.Grants.Funding..
Local.Employment
Pedestrian.Other.Collisions
TTC.Stops
Road.Kilometres
Working.Age..25.54.years.
Traffic.Collisions
Land.area.in.square.kilometres
X..Time.leaving.for.work...Between.8.a.m..and.8.59.a.m.
Watermain.Breaks
Diversity.Index...2008
X..Commute.in.Walked
Voter.Turnout...2008
Social.Assistance.Recipients
Salvation.Army.Donors
Debt.Risk.Score
Walk.Score
Child.Care.Spaces
Commuting.duration...15.to.29.minutes
X..Visible.Minority
X..Time.leaving.for.work...Between.6.a.m..and.6.59.a.m.
Children..0.14.years.
Population.density.per.square.kilometre
Unemployment.rate
Commuting.duration...45.to.59.minutes
Seniors..65..years.
Road.Volume
Neighbourhood.Equity.Score
0
100
200
300
400
500
MeanDecreaseGini

Conclusion

- KDE is very useful for clustering densities, but not much else; KDE is the dominant method used in traffic research
- DBSCAN is incredibly strong at clustering collisions with shared characteristics that don’t follow a circular pattern
- Although RF identified which variables are important as a whole, not sufficient to tell us about each unique cluster
- Unsupervised learning makes supervised learning even more powerful and vice versa
- The City should adopt a clusters-based approach rather than street or neighbourhood-level due to their size

A Data-driven Approach to Eliminating Pedestrian Collisions in Toronto

Jason Kim

jason2.kim@ryerson.ca
https://github.com/jasonukim/capstone-repo