Team 7: Final Report

Team Members: Neil Balazon, Mei Chang, Andrew Ivanov, Sam Lee, Skye Morgan

Introduction

Our main objective was to study Toronto's crimes in hopes of learning how Neighborhoods were being affected. Throughout our analysis, we made sure to take multiple types of crimes into account; Assault, Auto Theft, Break and Enter,Robbery and Theft Over. From doing this, it allowed us to better capture how Neighbourhoods were being impacted by crime.

Motivation

Crime statistics can be useful in helping criminal justice professionals anticipate increased risk of crime in specific neighborhoods. By taking a closer look at the statistics for certain neighbourhoods as well as the hours of the day crimes are committed, we can gain a deeper understanding of crime, why it occurs, and when it is most likely to occur. From the data we can find out what type of crime is most common in which neighborhoods, which can then allow criminal justice professionals to look into why certain types of crime are occurring in some neighborhoods. Furthermore, predictive policing could be used to anticipate the type and the risk of various crimes in a particular neighborhood, which could ultimately lead to better budget formation and resource allocation by law enforcement agencies.

Government agencies and other private entities such as real estate developers could also use statistics gathered within the field of urban planning and development in order to reduce crime. The initiatives could then be further studied in order to determine whether they are working for certain neighborhoods.

Dataset Description

The dataset we chose was shared by Toronto Police Service and posted on public safety portal website (http://data.torontopolice.on.ca/datasets/mci-2014-to-2018). There are 167,525 records and 27 unique variables in the dataset. We were particularly interested in Major Crime Indicator (MCI) and Neighbourhood. The MCI categories were assault, auto theft, break and enter, murder, robbery, sexual violation, and theft over. As such we grouped the data so we could study how MCI affected individual Neighbourhoods (See Figure 2).

Summary Table Explanation

The summary of our dataset showed that the mean for Assault was significantly higher than the other crimes. Furthermore it showed that Assault is clearly the most prevalent type of crime, and has the largest impact on Neighbourhoods. Ranking the type of crimes by most prevalent (mean as the measurement) we have Assault First, Break and Enter as second, Auto theft as third, Robbery fourth and Theft over as fifth. Considering that the dataset counts the successful attempts of these crimes it goes to show that Assault has the most influence in the crime rate for Neighbourhoods in Toronto.

K-means Clustering

Setting out, we wanted to gain some insight into the nature of how crime was affecting the Neighbourhoods in Toronto. That would involve looking into the individual crime statistics for each Neighbourhood. We quickly recognized that Clustering would be a suitable Data Mining Method because it would provide a structure in our collection of data. Additionally, k-means was selected because it proved to be relatively simple to implement and understand. However, before we could run K-means clustering there were some important steps that had to be performed on the dataset.

We removed the first column (Neighbourhood) as it wasn't quantitative. Then we made sure to

remove any missing values in the dataset as well as to scale the dataset (See figure 2).

Our K-means clustering using k=2 results were such that the first cluster had 126 Neighbourhoods while the second one had 15 Neighbourhoods. Cluster 1 had neighbourhoods with low assault,low auto theft,low break and enter,low robbery and low theft over. In contrast, Cluster 2 had neighbourhoods with high assault,high auto theft,high break and enter,high robbery and high theft over.

We were not satisfied with this particular model. Firstly, from our Within Cluster Sum of Squares we could see that only 47.5% of the total variance in the dataset was explained by clustering. Furthermore, the Total Within Sum of Squares was ~367. This value was quite high and raised concerns as we aimed to minimize the within group dispersion and maximize the between-group dispersion. Taking these concerns into consideration, we decided to perform k-means clustering with k=3 and compare our results.

Our K-means clustering using k=3 results were such that the first cluster had 10 Neighbourhoods, the second one had 41 and the third cluster had 90 Neighbourhoods. Cluster 1 had neighbourhoods with high assault,high auto theft,high break and enter,high robbery and high theft over. Cluster 2 had neighbourhoods with medium assault,medium auto theft,medium break and enter,medium robbery and medium theft over. Cluster 3 had neighbourhoods with low assault,low auto theft,low break and enter,low robbery and low theft over.

Our results indicated that this model was quite an improvement over our earlier one. Within Sum of Squares was now 62.2% while Total Within Sum of Squares was ~265. This indicated that 62.2% of the total variance in our data was explained by clustering. Moreover, our Total Within Sum of Squares was now lower than our previous model. Comparing with our results from k=2,

Within Sum of Squares increased by ~15% and we had managed to minimize our Total Within Sum of Squares. Taking these factors into consideration, we settled with staying with this model and used the results to form our conclusion.

We noticed that two data points (125,33) which were grouped into cluster 2 (High Crime) when we performed k=2 were now grouped into the cluster 2 (Medium Crime) upon using k=3. (See figure 7 and Figure 10). This was taken as an indication that the model had improved. Additionally when we looked at 4 sample Neighbourhoods from each cluster we saw that the crime occurrence between Neighbourhoods in the clusters were very similar and that across clusters they were different. This further reassured us that choosing this model was a good decision. (See Figure 11)

The results of the k-means clustering using k=3 allowed us to isolate 10 Neighbourhoods where crime occurrence was high. These Neighbourhood were: Church-Yonge Corridor, Moss Park, Kensington-Chinatown, Bay Street Corridor, Waterfront Communities-The Island, Woburn, Annex, York University Heights, Islington-City Centre West and West Humber-Clairville.

Additionally, we decided to forgo Hierarchical clustering as a Data Mining Method because it was too complicated and difficult to interpret. We attributed this to having 141 Neighbourhoods which made the Cluster Dendrogram very messy. As such, we did not move forward with it as a means of analyzing the dataset.

Conclusion

In conclusion, we were able to identify 10 Neighbourhoods where the amount of crime being committed was higher than anywhere else in Toronto. Furthermore, we were able to identify 3

seperate groups of Neighbourhoods where the occurrence of crime was generally low, medium or high. That said, most Neighbourhoods in Toronto are generally safe. However, there are a few exceptions. This is not surprising as most cities tend to have similar statistics. Nonetheless, this will remain useful to both law enforcement as well as residents/tourists to the area

Practical Implication

Crime statistics can be useful in helping criminal justice professionals anticipate increased risk of crime in specific neighborhoods. Two applications would be Budget Formation and Resource Allocation. Additionally, Government agencies and other private entities such as real estate developers could also use statistics gathered within the field of urban planning and development in order to reduce crime. Publicly Funded Initiatives also offers a window of opportunity in decreasing crime in the Neighborhoods where crime is rampant.

Appendix

Variables of Interest	Description
Neighbourhood	The different neighborhoods (141) in the city of Toronto
Assault	Direct or indirect application of force to another person, or the attempt or threat to apply force to another person, without that person's consent
Auto Theft	Act of taking or another person's vehicle (not including attempts). Auto Theft figures represent the number of vehicles stolen.
Break and Enter	Act of entering a place with the intent to commit an indictable offence therein.
Robbery	Act of taking property from another person or business by the use of force or intimidation in the presence of the victim.
Theft Over	Act of stealing property in excess of $5,000 (excluding auto theft)

Figure 1: Key Variables and Description

	Neighbourhood	Assault	Auto Theft	Break and Enter	Robbery	Theft Over
1	Agincourt North (129)	392	133	294	162	27
2	Agincourt South-Malvern West (128)	580	159	381	148	63
3	Alderwood (20)	182	84	124	36	34
4	Annex (95)	1523	115	782	314	174
5	Banbury-Don Mills (42)	398	84	341	70	45
6	Bathurst Manor (34)	417	174	240	69	51
7	Bay Street Corridor (76)	2446	109	538	286	185
8	Bayview Village (52)	415	88	217	38	40
9	Bayview Woods-Steeles (49)	180	61	157	19	11
10	Bedford Park-Nortown (39)	270	220	513	75	57
11	Beechborough-Greenbrook (112)	329	62	119	94	11
12	Bendale (127)	1246	257	357	347	70
13	Birchcliffe-Cliffside (122)	827	81	377	98	28
14	Black Creek (24)	1302	258	144	252	50
15	Blake-Jones (69)	261	22	83	56	3
16	Briar Hill-Belgravia (108)	613	114	197	167	26
17	Bridle Path-Sunnybrook-York Mills (41)	113	41	237	6	23
18	Broadview North (57)	180	20	70	24	5
19	Brookhaven-Amesbury (30)	422	145	140	93	29

Figure 2: Formatted Dataset

Variables	Mean	Median	Min	Max
Assault	644.5	450	4	4061
Auto Theft	128.9	89	4	1684
Break and Enter	247.6	200	9	1074
Robbery	128.6	95	3	904
Theft Over	38.51	24	3	245

Figure 3: Summary Table

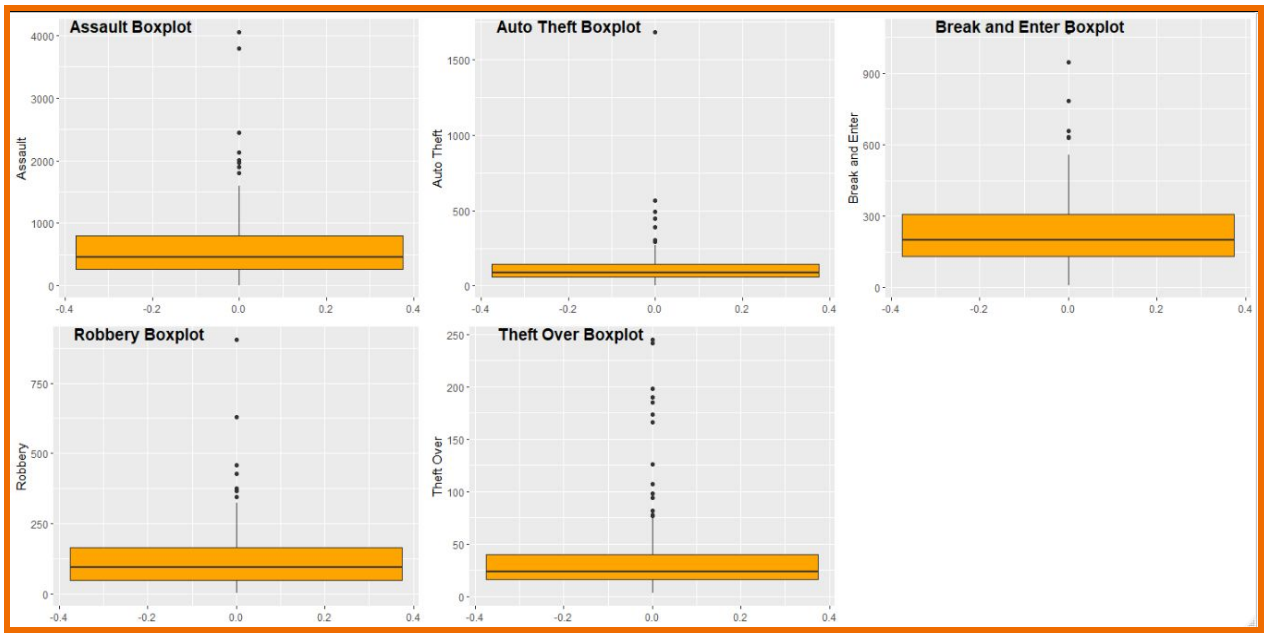

Figure 4: Boxplot of Key Variables

	row.names	Assault	Auto Theft	Break and Enter	Robbery	Theft Over
1	1	-0.4129285	0.02547666	0.2711717	0.2743428	-0.2640418
2	2	-0.105511	0.187907	0.7795721	0.159461	0.5617598
3	3	-0.7563204	-0.280642	-0.7222543	-0.7595932	-0.1034692
4	4	1.436482	-0.08697509	3.122889	1.521631	3.107981
5	5	-0.4031173	-0.280642	0.5458248	-0.4805946	0.148859
6	6	-0.3720485	0.2816168	-0.04438712	-0.4888005	0.2864926
7	7	2.945771	-0.124459	1.69703	1.291867	3.36031
8	8	-0.3753189	-0.2556527	-0.1787918	-0.7431815	0.03416437
9	9	-0.7595908	-0.4243303	-0.5294128	-0.8990925	-0.6310647
10	10	-0.6124228	0.5689935	1.550938	-0.4395654	0.4241262
11	11	-0.515946	-0.418083	-0.7514727	-0.2836544	-0.6310647
12	12	0.9835319	0.8001443	0.6393237	1.792424	0.7223324
13	13	0.2983833	-0.299384	0.7561974	-0.250831	-0.2411028
14	14	1.075103	0.8063916	-0.6053806	1.012869	0.2635537
15	15	-0.6271396	-0.6679758	-0.9618453	-0.5954764	-0.8145762
16	16	-0.05154938	-0.09322241	-0.2956655	0.315372	-0.2869807
17	17	-0.8691491	-0.5492767	-0.06191817	-1.005768	-0.3557975
18	18	-0.7595908	-0.6804705	-1.037813	-0.8580633	-0.7686983
19	19	-0.3638725	0.1004445	-0.6287554	-0.2918603	-0.2181639

Figure 5: Scaled Dataset

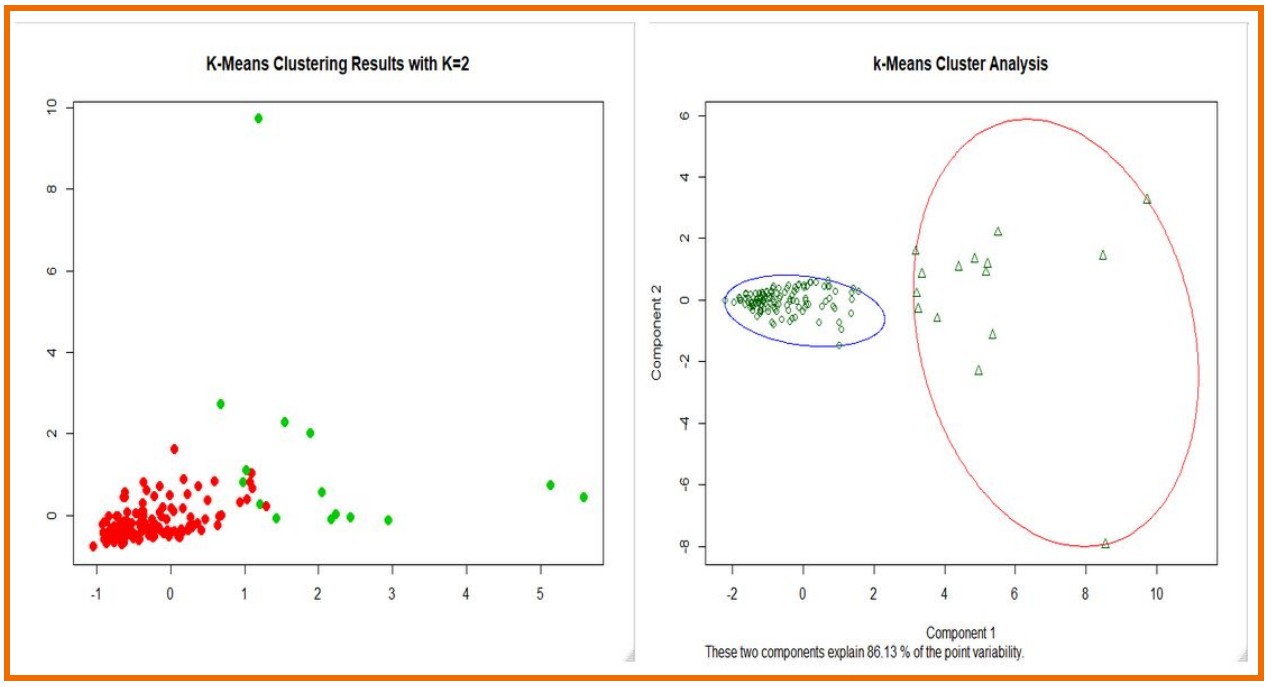

Figure 6: K-means clustering results with k=2

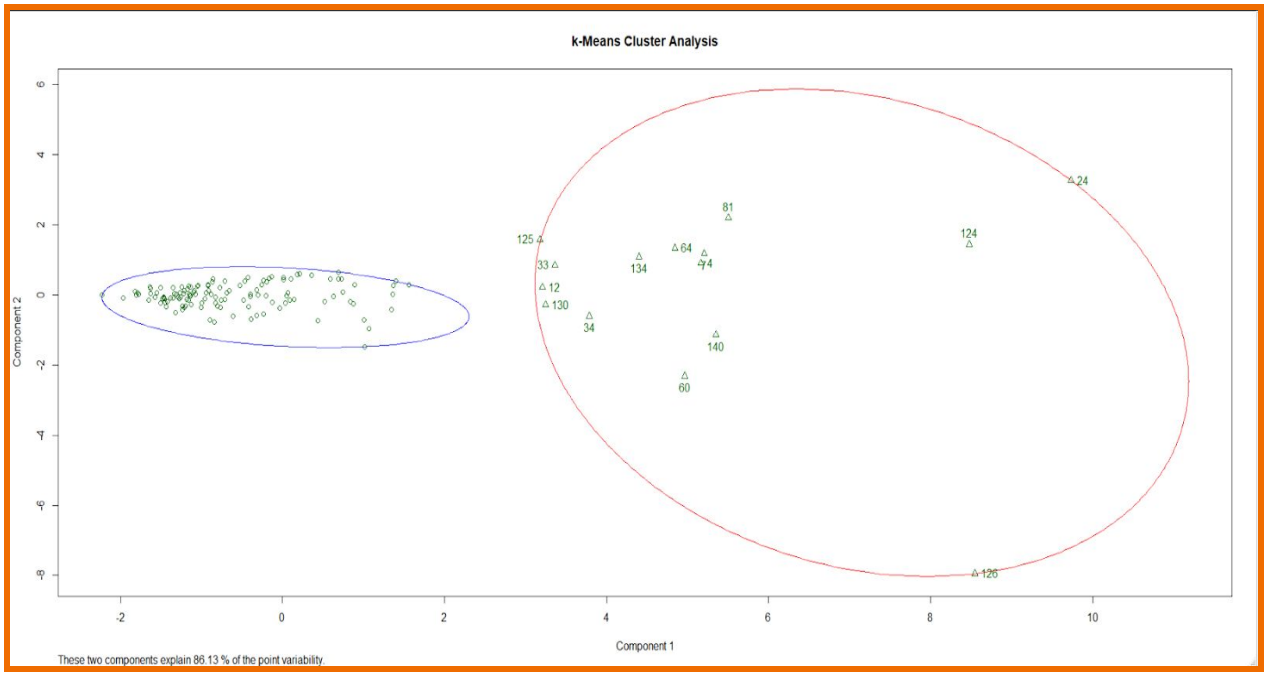

Figure 7: K-means clustering results with k=2 (With Labels)

```
> km.out
K-means clustering with 2 clusters of sizes 126, 15

Cluster means:
     Assault Auto Theft Break and Enter    Robbery Theft Over
1 -0.2582983 -0.1608521       -0.245113 -0.2452954 -0.2607648
2  2.1697057  1.3511579        2.058949  2.0604810  2.1904241

Within cluster sum of squares by cluster:
[1] 167.3139 200.0593
 (between_SS / total_SS =  47.5 %)

> km.out$tot.withinss
[1] 367.3733
```

Figure 8: Results using k=2

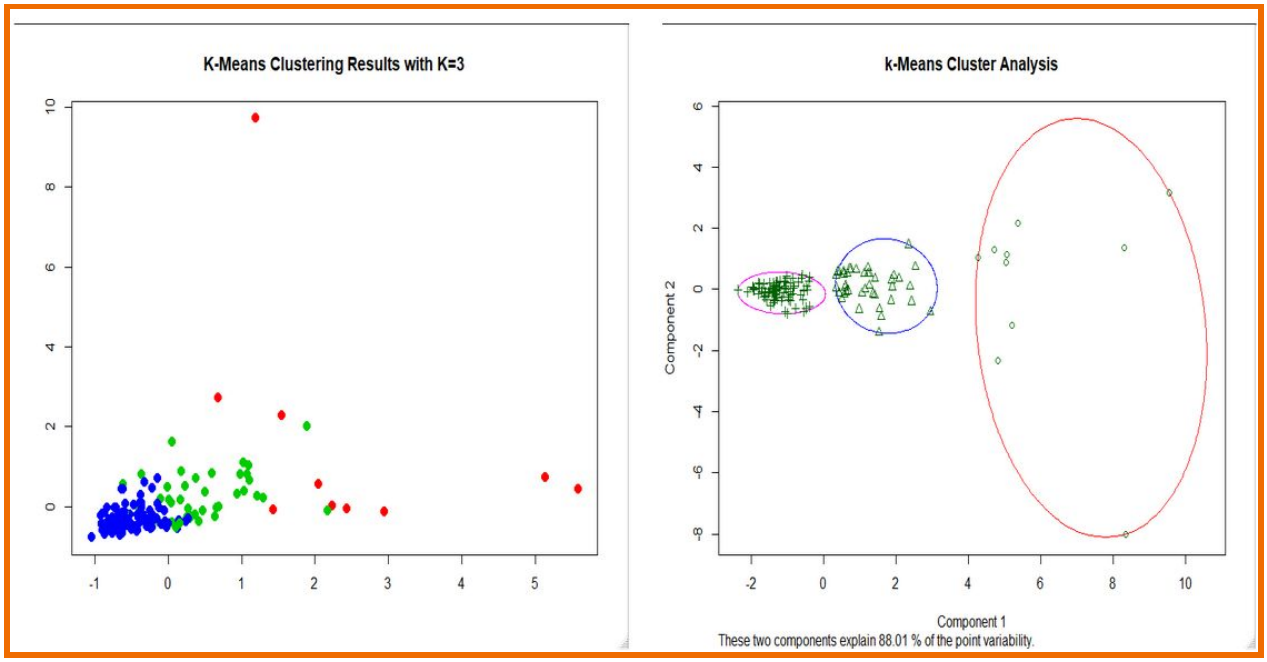

Figure 9: K-means clustering results with k=3

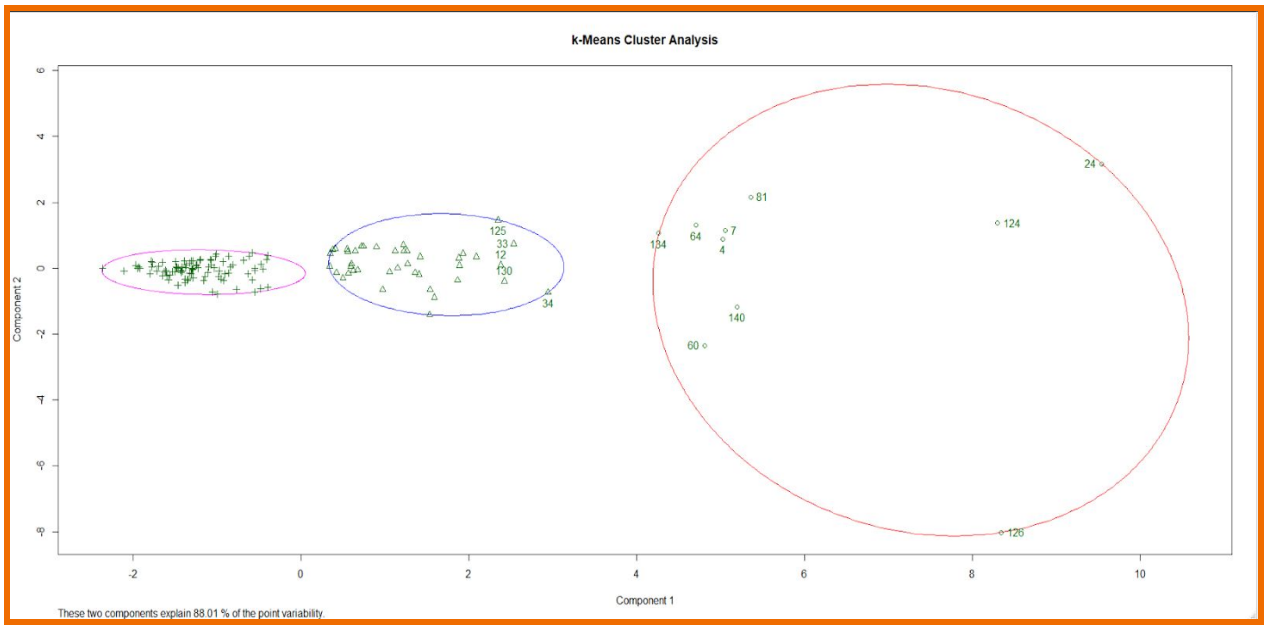

Figure 10: K-means clustering results with k=3 (With Labels)

```
> km.out
K-means clustering with 3 clusters of sizes 10, 41, 90

Cluster means:
     Assault Auto Theft Break and Enter    Robbery Theft Over
1  2.5265062  1.6191679       2.5654014  2.3963735  3.0070501
2  0.4569977  0.2643985       0.5438294  0.4934988  0.2562804
3 -0.4889107 -0.3003558      -0.5327891 -0.4910799 -0.4508666

Within cluster sum of squares by cluster:
[1] 157.01345  68.48119  39.38121
 (between_SS / total_SS =  62.2 %)

> km.out$tot.withinss
[1] 264.8759
```

Figure 10: Results using k=3

Comparison of Clustered Neighbourhoods for k=3

Cluster 1 - High Crime

row.names	Neighbourhood	Assault	Auto Theft	Break and Enter	Robbery	Theft Over
24	Church-Yonge Corridor (75)	4061	200	946	904	190
81	Moss Park (73)	2134	119	632	630	94
64	Kensington-Chinatown (78)	2013	130	627	367	126
7	Bay Street Corridor (76)	2446	109	538	286	185

Cluster 2 - Medium Crime

row.names	Neighbourhood	Assault	Auto Theft	Break and Enter	Robbery	Theft Over
125	West Hill (136)	1971	112	416	323	35
33	Dovercourt-Wallace Emerson-Junction (93)	1386	173	512	318	61
12	Bendale (127)	1246	257	357	347	70
130	Wexford/Maryvale (119)	1270	306	494	202	78

Cluster 3 - Low Crime

row.names	Neighbourhood	Assault	Auto Theft	Break and Enter	Robbery	Theft Over
98	Pelmo Park-Humberlea (23)	254	197	128	91	22
133	Willowridge-Martingrove-Richview (7)	450	227	231	151	22
115	Stonegate-Queensway (16)	417	148	253	88	36
57	Humbermede (22)	510	203	129	97	29

Figure 11: Comparison of clustered Neighbourhoods for k=3