In an effort to be more transparent, cities are now taking part in an OpenData initiative, resulting in publicly available crime and census data. All data used in this report was collected from the City of Toronto’s Open Data portal. The two main datasets used were the 2016 census data from the city of Toronto and crime data set released by Toronto police. Prior to performing any analysis, the dataset had to be gathered, filtered and formatted in a manner that could be used in R. The crime dataset required minimal preprocessing before it was ready to be used, however, the census data had to be filtered and formatted significantly. This entailed creating a function to gather variables of interest from the census dataset and join them into a single data frame. Over 50 variables where collected from the census data, relating to topics such as income, demographics, population, education and housing characteristics, etc.

Put getdata function here (Merging 63-69)

Table of variables?

**Feature Engineering**

In addition to the variables that were collected directly from census data, additional variables were also created. The crime statistics provided by the in the crime dataset simply state the number of occurrences of a crime in a neighbourhood, however this number does not take into account the population of the neighbourhood. To make crime statistics more comparable across neighbourhoods with different populations each crime statistic was converted into a crime rate per 10 000, similar to the commonly used crime rate per 100 000.

Put crime rate function here (Merging.R 460-463)

In an effort to gain insights about the economic wealth of a neighbourhood we used the average income of each neighbourhood, obtained from the Toronto census data. The average, though, provides an incomplete picture, as it can be heavily skewed by a few high-earning individuals. It ignores the distribution of incomes and income inequality in each neighbourhood, which is considered to be positively correlated with violent crime rates around the world (Fajnzylber, Lederman, & Loayza, 2002). To get a more complete picture of the economic wealth of each neighbourhood it became pertinent to look at the median income in each neighbourhood as well. The median income data was not readily available from the census data, but the distribution of incomes in each neighbourhood was. Below is a sample of the census data that showed the distribution of incomes in each neighbourhood:

\begin{table}[ht]

\centering

\begin{tabular}{rlrrr}

\hline

& Characteristic & Agincourt.North & Agincourt.South.Malvern.West & Alderwood \\

\hline

970 & Under \$10,000 (including loss) & 5170.00 & 4535.00 & 1365.00 \\

971 & \$10,000 to \$19,999 & 6325.00 & 4505.00 & 1505.00 \\

972 & \$20,000 to \$29,999 & 3520.00 & 2715.00 & 1360.00 \\

973 & \$30,000 to \$39,999 & 2465.00 & 2020.00 & 1095.00 \\

974 & \$40,000 to \$49,999 & 1895.00 & 1560.00 & 950.00 \\

975 & \$50,000 to \$59,999 & 1265.00 & 1125.00 & 825.00 \\

976 & \$60,000 to \$69,999 & 865.00 & 825.00 & 690.00 \\

977 & \$70,000 to \$79,999 & 655.00 & 570.00 & 530.00 \\

978 & \$80,000 to \$89,999 & 435.00 & 435.00 & 395.00 \\

979 & \$90,000 to \$99,999 & 365.00 & 315.00 & 370.00 \\

981 & \$100,000 to \$149,999 & 530.00 & 525.00 & 620.00 \\

982 & \$150,000 and over & 135.00 & 165.00 & 225.00 \\

\hline

\end{tabular}

\end{table}

Using a method of calculating the median from a set of grouped intervals from Statistics Canada (<https://www.statcan.gc.ca/edu/power-pouvoir/ch11/median-mediane/5214872-eng.htm>) and the census information, a function was created to find the median income of each neighbourhood, resulting in the median income of each neighbourhood.

Put grouped median function here (Merging.R 208-240)

\begin{table}[ht]

\centering

\begin{tabular}{rllrr}

\hline

& Neighbourhood & Hood\\_ID & avg.income & median.income \\

\hline

1 & Agincourt.North & 129 & 30414.00 & 20901.90 \\

2 & Agincourt.South.Malvern.West & 128 & 31825.00 & 22237.35 \\

3 & Alderwood & 20 & 47709.00 & 36711.66 \\

4 & Annex & 95 & 112766.00 & 43035.87 \\

5 & Banbury.Don.Mills & 42 & 67757.00 & 42493.22 \\

6 & Bathurst.Manor & 34 & 45936.00 & 29612.30 \\

7 & Bay.Street.Corridor & 76 & 56526.00 & 27511.69 \\

8 & Bayview.Village & 52 & 52035.00 & 32818.51 \\

\hline

\end{tabular}

\end{table}

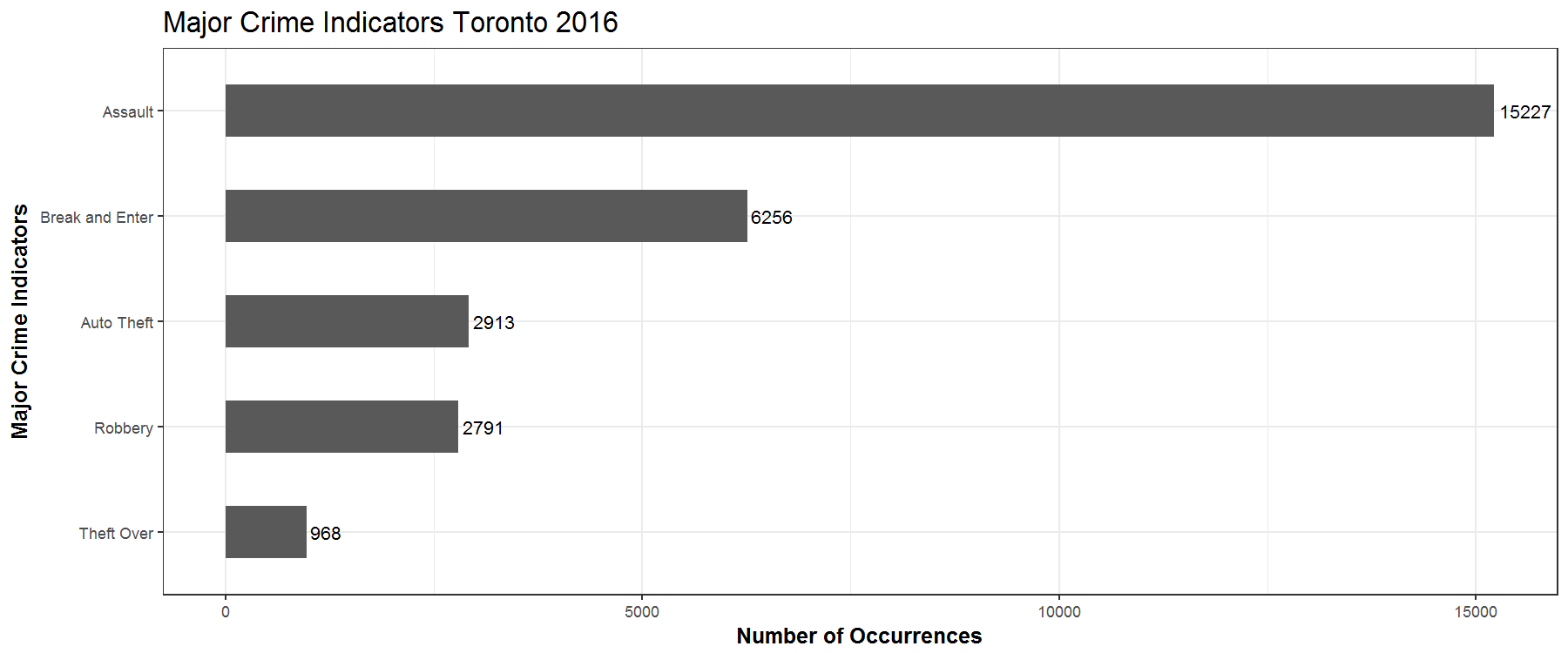
Looking at the above table, it becomes readily apparent that certain neighbourhoods have a much higher average income than median income, indicating the presence of a few high earners and income inequality.

With all variables collected into a single data frame, analysis could begin.

**Exploratory Data Analysis**

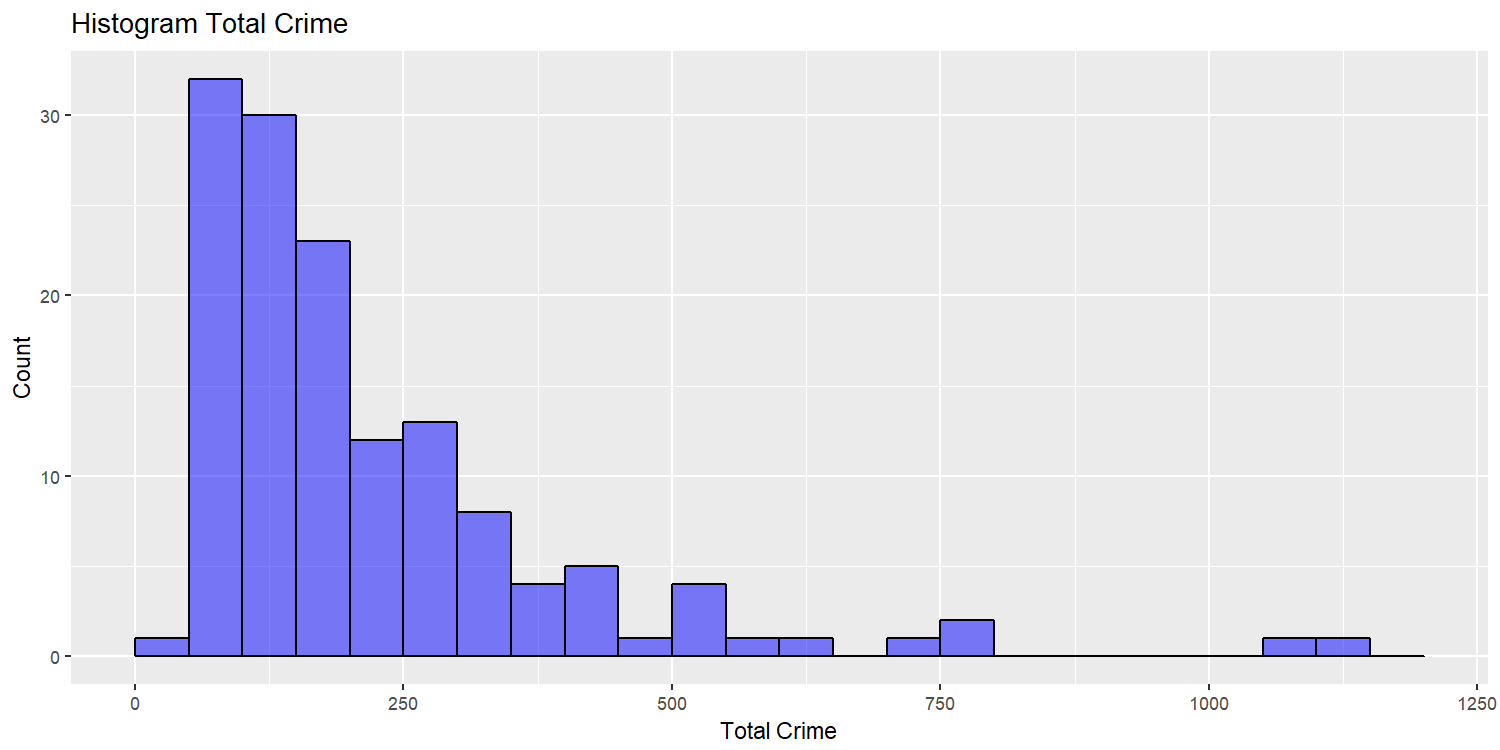
A first step to analyzing crime in Toronto was a look at the prevalence of different crime types in Toronto. Using the ggplot and dplyr packages in r to aggregate the data by crime type and plot it, one can see the most common type of crime in the city is assault, which overshadows every other type of crime.

Put ggplot code here (initial exploratory analysis 39-54)



To see the distribution of our crime statistics a function was created to plot a histogram. Since the shape of a histogram can be heavily influenced by the bin widths used to divide a continuous variable, an additional argument allowing the user to specify bin widths was added to the function in order to add flexibility.

Put histogram function here(initial data analysis 1-12)

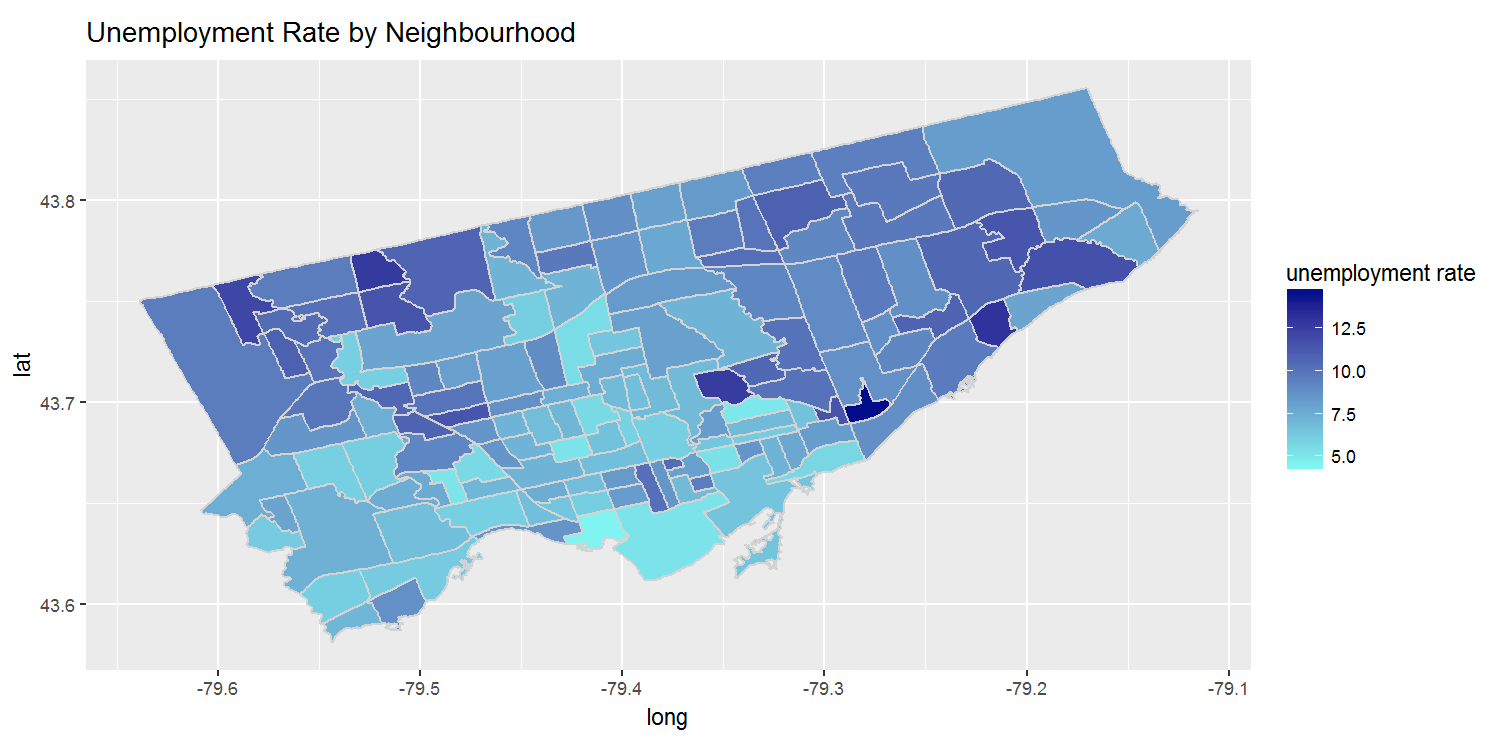
Using this function on the different crime variables it is apparent that the distribution of crime in Toronto is left-tailed, with many neighbourhoods having a low level of crime. However, there are a few neighbourhoods that are on the far-right of the distribution, indicating a significantly high number of crime occurrences.

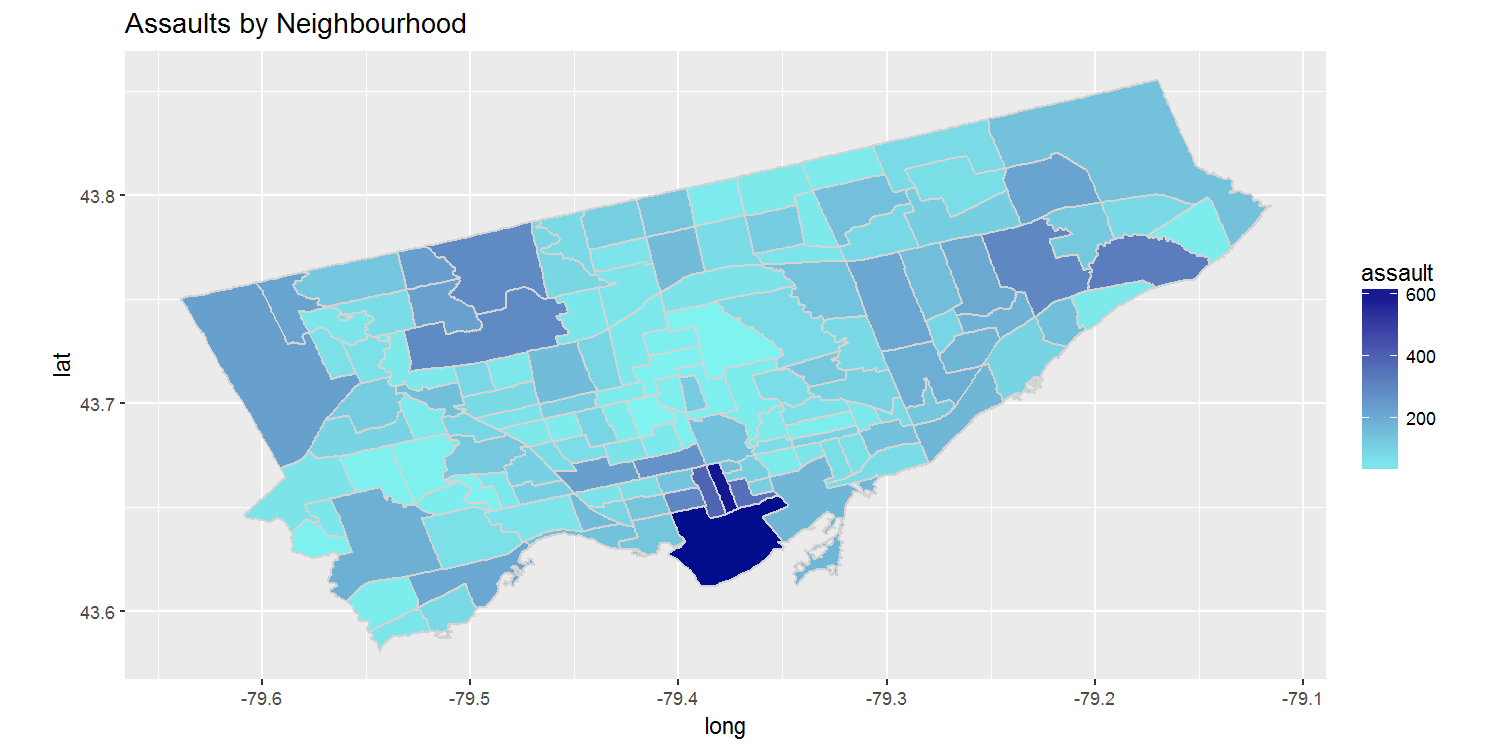
To gain a better understanding of the differences between neighbourhoods it was pertinent to see the data plotted on a heat map of the city. With a quick visual inspection we can see how each neighbourhood differs.

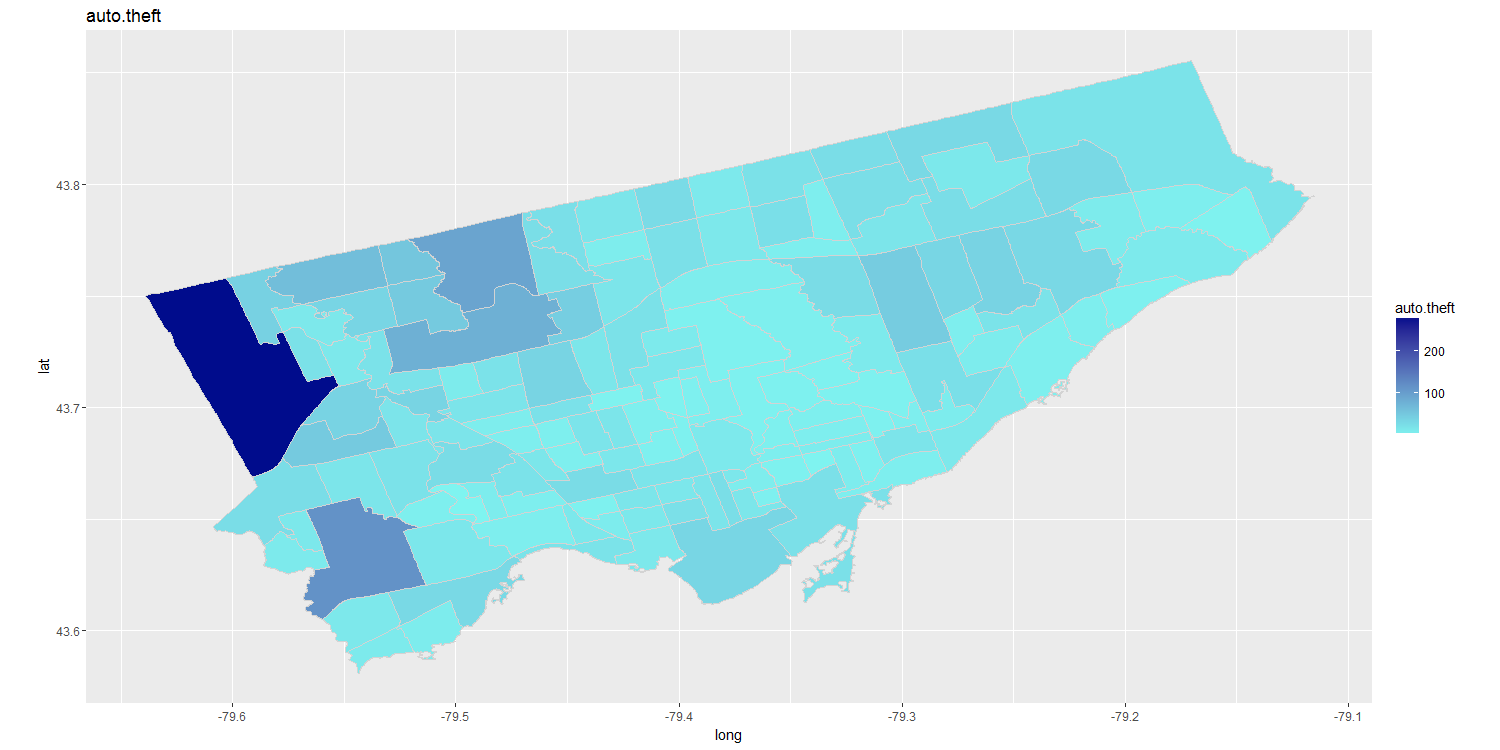
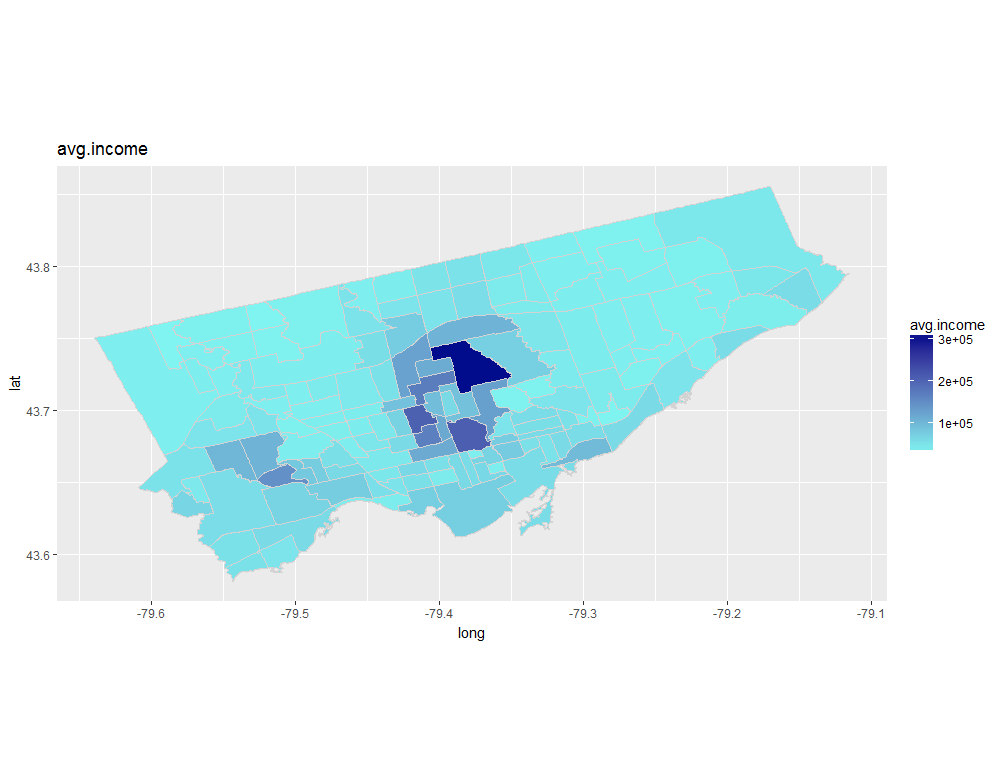
The process of creating heat maps involved first reading the necessary shapefile (obtained from the City of Toronto Open Data Catalogue) and joining it to the retrieved census data.

Insert initial exploratory analysis 241-257

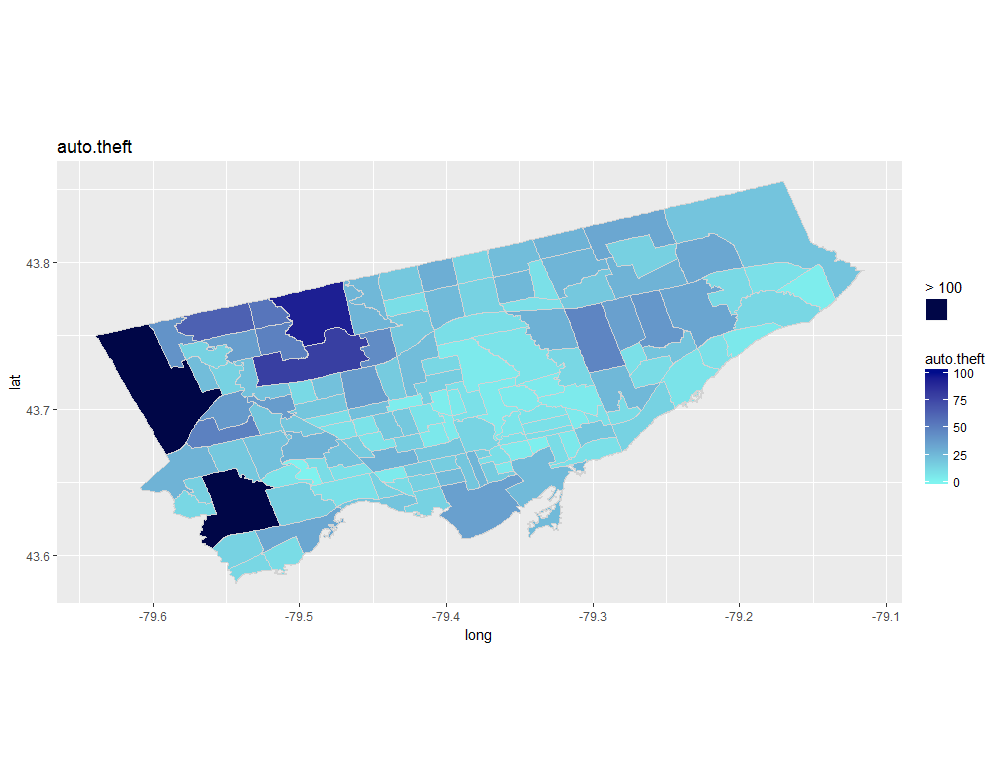
With the data in a usable format, a function was then created to plot each variable according to each neighbourhood.

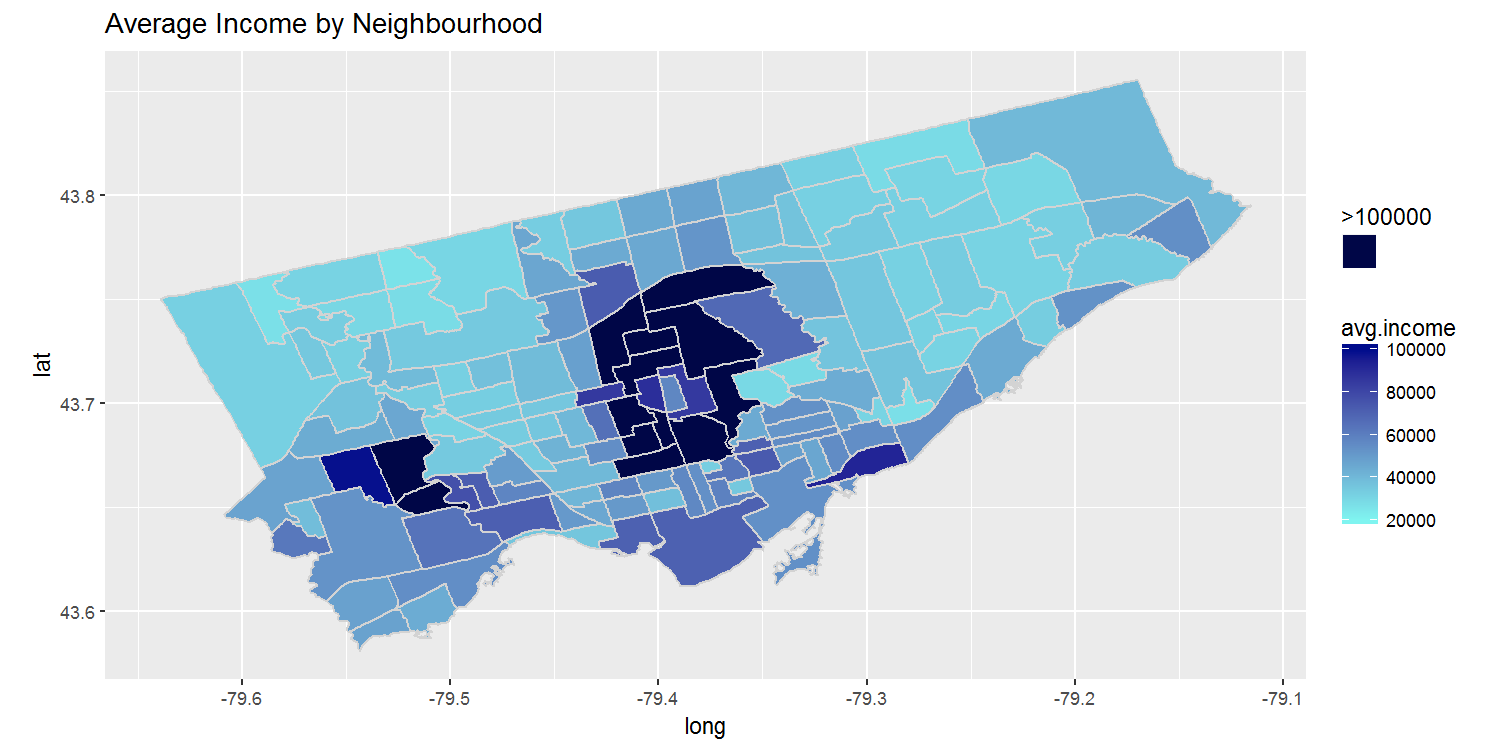
Insert heatMap function (data analysis 260-270)



Several problems occurred when plotting variables using this manner, particularly the presence of outliers. The presence of an outlier in the data resulted in a scale that made it difficult to visualize the difference between the majority of neighbourhoods. In the case of Toronto, this is evident in the occurrence of auto thefts by neighbourhood. One neighbourhood had a significantly higher number of auto thefts than the rest, which made it impossible to see any differences between other neighbourhoods.

The solution to this problem was to create a function for plotting a heat map that had an upper limit. Any number above this limit would appear as one colour and the scale of the heat map would be preserved, resulting in a more readable and insightful heat map.

Insert heat\_map\_limit function (data analysis 283-299)



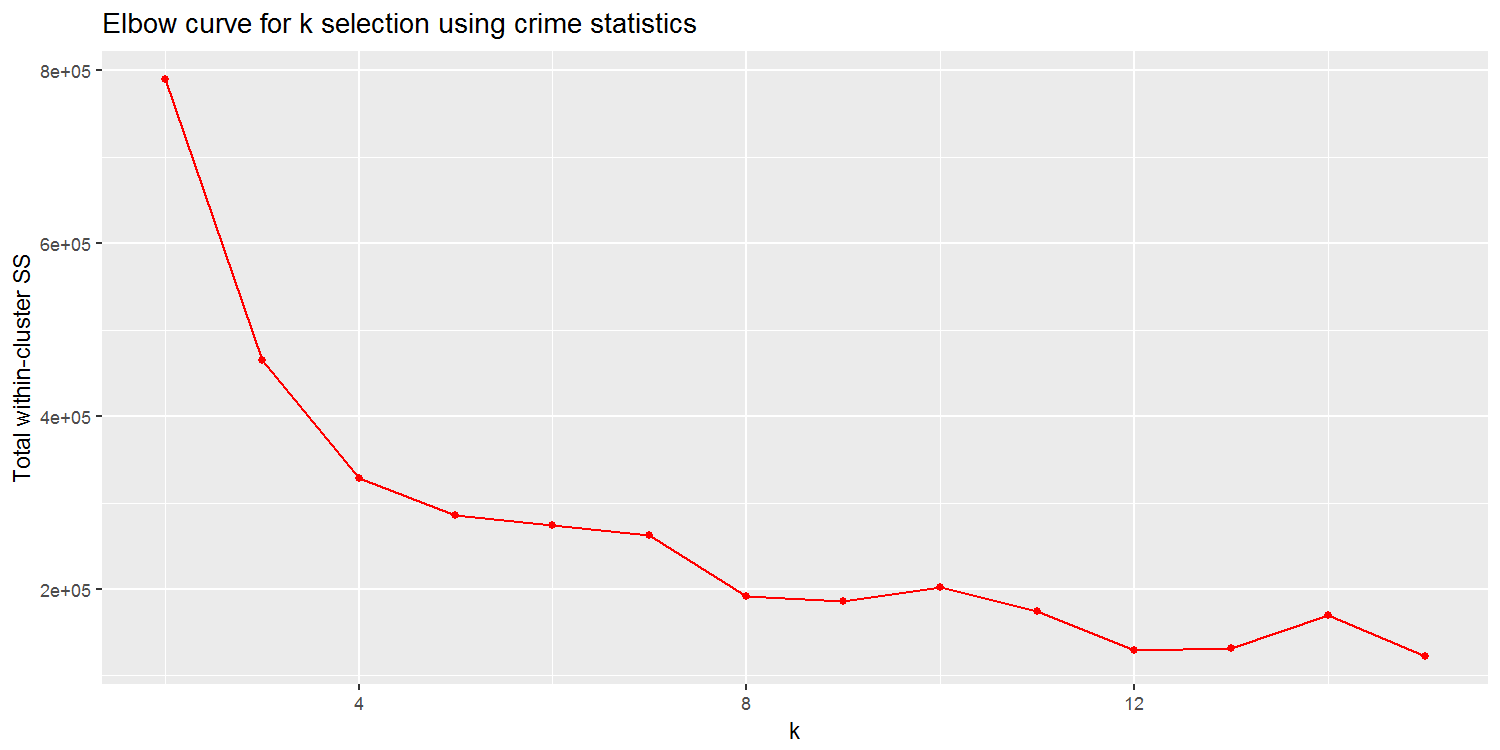
**Cluster Analysis**

With 140 neighbourhoods in the city, it may be helpful to group neighbourhoods that exhibit similar characteristics together. One popular and often used method of clustering is the k means clustering algorithm. The k means algorithm is used to assign each observation to a cluster, based on similar characteristics in order to simply the number of groups to look at. The algorithm first places a specified number of centroids and along the specified features. Each centroid defines a cluster. Each data point is then assigned to its nearest centroid, based on the squared Euclidean distance. The algorithm then moves each centroid towards the mean of all data points assigned to that centroid. The new clusters are calculated and the process is repeated until some stopping criteria is met. The goal being to minimize the sum of the distance between points within each cluster.

Insert K means formula here

The first thing to be decided was the number of clusters to initially specify. Too few clusters would ignore nuances between groups and too many clusters would not simplify anything. A common method of determining the optimal number of clusters involves testing a varying number of clusters and looking at the sum of squares within clusters. The optimal number is the point at which the sum of squares decreases most sharply and any further increases in the number of clusters results in only a small decrease in the sum of squares within clusters. The easiest method of visualizing this is plotting the total sum of squares within clusters with the number of clusters and finding the “elbow” of the curve.

Insert k means code (data analysis 147-170, 211-215)



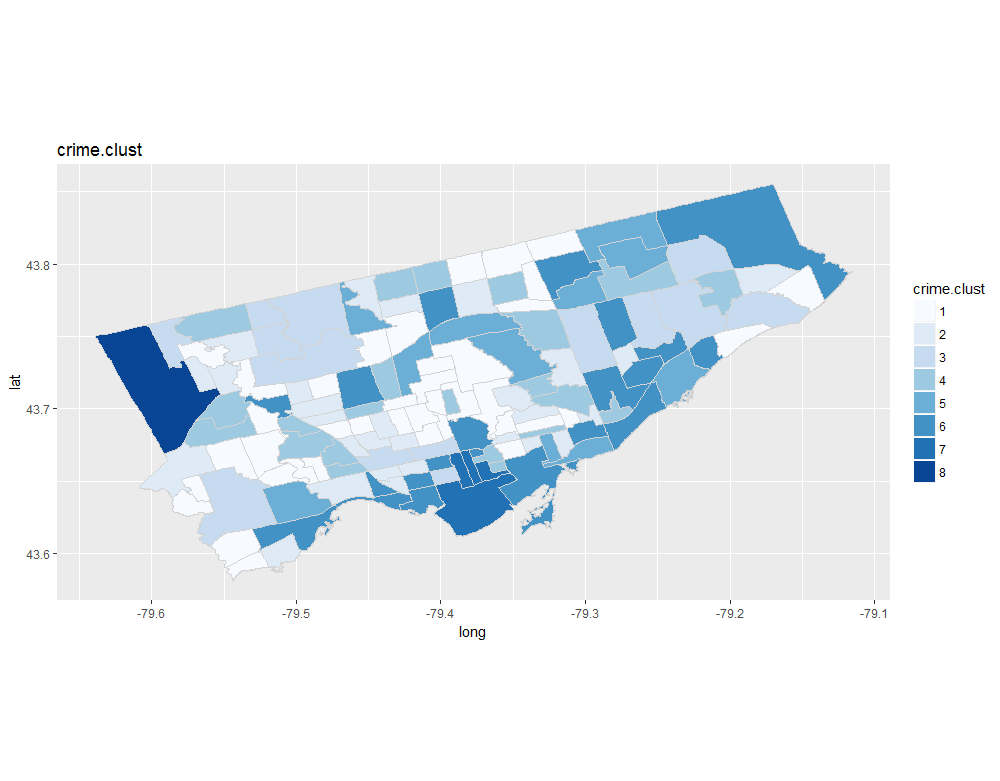
Above we can see the optimal number of clusters to be 8, based on using the different crime statistics of each neighbourhood. Plotting a cluster plot shows us the neighbourhoods that belong to each cluster.



However, it is not possible to see exactly which neighbourhood belongs to each cluster. To see this, a new variable with each neighbourhood’s cluster number was added to the aggregated dataset and plotted using a new heat map function.

Insert clust\_func (data analysis 172-179, 219)

Insert heat\_map\_clust function (data analysis 272-281)

From the figure above, it is apparent that the physical location of each neighbourhood does not determine which cluster the neighbourhood belongs to. Although we see several neighbourhoods that are near each other belonging to the same cluster, there are several neighbourhoods that are on opposite ends of the city but still belong to the same cluster.