Data cleaning

Payalben Siddharthsinh chavda

June 9, 2020

library(dplyr)

## Warning: package 'dplyr' was built under R version 3.5.3

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.5.3

library(plyr)

## Warning: package 'plyr' was built under R version 3.5.3

## -------------------------------------------------------------------------

## You have loaded plyr after dplyr - this is likely to cause problems.  
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:  
## library(plyr); library(dplyr)

## -------------------------------------------------------------------------

##   
## Attaching package: 'plyr'

## The following objects are masked from 'package:dplyr':  
##   
## arrange, count, desc, failwith, id, mutate, rename, summarise,  
## summarize

## Data cleaning

## Read data file into R

df <- read.csv("C:/Users/pchav/Downloads/MCI\_2014\_to\_2019.csv")

## Drop irrelevant attributes

crime\_data <- df[,c(5,15:21,24:26)]

## Let’s check the structure of dataset:

str(crime\_data)

## 'data.frame': 206435 obs. of 11 variables:  
## $ premisetype : Factor w/ 5 levels "Apartment","Commercial",..: 3 1 1 1 3 5 1 3 1 2 ...  
## $ occurrenceyear : int 2014 2014 2014 2014 2014 2014 2014 2014 2014 2014 ...  
## $ occurrencemonth : Factor w/ 13 levels "","April","August",..: 5 2 2 2 2 9 9 5 6 6 ...  
## $ occurrenceday : int 27 1 1 1 1 16 16 15 1 1 ...  
## $ occurrencedayofyear: int 58 91 91 91 91 75 75 46 1 1 ...  
## $ occurrencedayofweek: Factor w/ 8 levels "","Friday ",..: 6 7 7 7 7 5 5 4 8 8 ...  
## $ occurrencehour : int 16 16 16 16 12 7 2 1 11 3 ...  
## $ MCI : Factor w/ 5 levels "Assault","Auto Theft",..: 5 1 1 3 5 4 1 1 1 4 ...  
## $ Neighbourhood : Factor w/ 140 levels "Agincourt North (129)",..: 45 92 92 92 6 110 128 106 25 24 ...  
## $ Long : num -79.4 -79.3 -79.3 -79.3 -79.5 ...  
## $ Lat : num 43.7 43.7 43.7 43.7 43.8 ...

print("All variables have correct data type. There are 5 categorical and 6 numerical attributes.")

## [1] "All variables have correct data type. There are 5 categorical and 6 numerical attributes."

## summary of dataset

summary(crime\_data)

## premisetype occurrenceyear occurrencemonth occurrenceday   
## Apartment :49996 Min. :2000 October :18501 Min. : 1.00   
## Commercial:41081 1st Qu.:2015 July :18329 1st Qu.: 8.00   
## House :37927 Median :2017 November:17983 Median :16.00   
## Other :23178 Mean :2017 June :17960 Mean :15.51   
## Outside :54253 3rd Qu.:2018 May :17914 3rd Qu.:23.00   
## Max. :2019 August :17905 Max. :31.00   
## NA's :59 (Other) :97843 NA's :59   
## occurrencedayofyear occurrencedayofweek occurrencehour   
## Min. : 1.0 Friday :31387 Min. : 0.00   
## 1st Qu.: 99.0 Saturday :30781 1st Qu.: 7.00   
## Median :188.0 Sunday :29718 Median :14.00   
## Mean :186.5 Thursday :28885 Mean :12.55   
## 3rd Qu.:276.0 Monday :28806 3rd Qu.:19.00   
## Max. :366.0 Wednesday :28654 Max. :23.00   
## NA's :59 (Other) :28204   
## MCI Neighbourhood   
## Assault :111423 Waterfront Communities-The Island (77): 7747   
## Auto Theft : 23380 Bay Street Corridor (76) : 6817   
## Break and Enter: 43302 Church-Yonge Corridor (75) : 6232   
## Robbery : 21543 West Humber-Clairville (1) : 5702   
## Theft Over : 6787 Moss Park (73) : 4786   
## York University Heights (27) : 3989   
## (Other) :171162   
## Long Lat   
## Min. :-79.64 Min. :43.59   
## 1st Qu.:-79.47 1st Qu.:43.66   
## Median :-79.39 Median :43.70   
## Mean :-79.39 Mean :43.71   
## 3rd Qu.:-79.32 3rd Qu.:43.75   
## Max. :-79.12 Max. :43.85   
##

print("summary shows that there is missing data in occurrenceyear, occurrencedayofyear, occurrenceday has missing values.  
")

## [1] "summary shows that there is missing data in occurrenceyear, occurrencedayofyear, occurrenceday has missing values.\n"

## analyze categorical variables by frequency table

count(crime\_data, "occurrencemonth")

## occurrencemonth freq  
## 1 59  
## 2 April 16473  
## 3 August 17905  
## 4 December 16769  
## 5 February 14495  
## 6 January 16347  
## 7 July 18329  
## 8 June 17960  
## 9 March 16049  
## 10 May 17914  
## 11 November 17983  
## 12 October 18501  
## 13 September 17651

count(crime\_data,"occurrencedayofweek")

## occurrencedayofweek freq  
## 1 59  
## 2 Friday 31387  
## 3 Monday 28806  
## 4 Saturday 30781  
## 5 Sunday 29718  
## 6 Thursday 28885  
## 7 Tuesday 28145  
## 8 Wednesday 28654

count(crime\_data,"MCI")

## MCI freq  
## 1 Assault 111423  
## 2 Auto Theft 23380  
## 3 Break and Enter 43302  
## 4 Robbery 21543  
## 5 Theft Over 6787

count(crime\_data,"Neighbourhood")

## Neighbourhood freq  
## 1 Agincourt North (129) 1157  
## 2 Agincourt South-Malvern West (128) 1650  
## 3 Alderwood (20) 545  
## 4 Annex (95) 2917  
## 5 Banbury-Don Mills (42) 1205  
## 6 Bathurst Manor (34) 727  
## 7 Bay Street Corridor (76) 6817  
## 8 Bayview Village (52) 927  
## 9 Bayview Woods-Steeles (49) 539  
## 10 Bedford Park-Nortown (39) 1240  
## 11 Beechborough-Greenbrook (112) 605  
## 12 Bendale (127) 2552  
## 13 Birchcliffe-Cliffside (122) 1730  
## 14 Black Creek (24) 2069  
## 15 Blake-Jones (69) 679  
## 16 Briar Hill-Belgravia (108) 930  
## 17 Bridle Path-Sunnybrook-York Mills (41) 560  
## 18 Broadview North (57) 593  
## 19 Brookhaven-Amesbury (30) 1162  
## 20 Cabbagetown-South St.James Town (71) 1223  
## 21 Caledonia-Fairbank (109) 678  
## 22 Casa Loma (96) 480  
## 23 Centennial Scarborough (133) 508  
## 24 Church-Yonge Corridor (75) 6232  
## 25 Clairlea-Birchmount (120) 2711  
## 26 Clanton Park (33) 1117  
## 27 Cliffcrest (123) 1217  
## 28 Corso Italia-Davenport (92) 918  
## 29 Danforth (66) 841  
## 30 Danforth East York (59) 666  
## 31 Don Valley Village (47) 1495  
## 32 Dorset Park (126) 2109  
## 33 Dovercourt-Wallace Emerson-Junction (93) 2639  
## 34 Downsview-Roding-CFB (26) 3974  
## 35 Dufferin Grove (83) 1049  
## 36 East End-Danforth (62) 1895  
## 37 Edenbridge-Humber Valley (9) 634  
## 38 Eglinton East (138) 1865  
## 39 Elms-Old Rexdale (5) 593  
## 40 Englemount-Lawrence (32) 1248  
## 41 Eringate-Centennial-West Deane (11) 787  
## 42 Etobicoke West Mall (13) 515  
## 43 Flemingdon Park (44) 1163  
## 44 Forest Hill North (102) 565  
## 45 Forest Hill South (101) 494  
## 46 Glenfield-Jane Heights (25) 2776  
## 47 Greenwood-Coxwell (65) 1155  
## 48 Guildwood (140) 411  
## 49 Henry Farm (53) 787  
## 50 High Park-Swansea (87) 1233  
## 51 High Park North (88) 1070  
## 52 Highland Creek (134) 716  
## 53 Hillcrest Village (48) 810  
## 54 Humber Heights-Westmount (8) 509  
## 55 Humber Summit (21) 1670  
## 56 Humbermede (22) 1521  
## 57 Humewood-Cedarvale (106) 645  
## 58 Ionview (125) 859  
## 59 Islington-City Centre West (14) 3287  
## 60 Junction Area (90) 1119  
## 61 Keelesdale-Eglinton West (110) 785  
## 62 Kennedy Park (124) 2043  
## 63 Kensington-Chinatown (78) 3823  
## 64 Kingsview Village-The Westway (6) 1255  
## 65 Kingsway South (15) 496  
## 66 L'Amoreaux (117) 1866  
## 67 Lambton Baby Point (114) 353  
## 68 Lansing-Westgate (38) 929  
## 69 Lawrence Park North (105) 513  
## 70 Lawrence Park South (103) 627  
## 71 Leaside-Bennington (56) 582  
## 72 Little Portugal (84) 1015  
## 73 Long Branch (19) 725  
## 74 Malvern (132) 2624  
## 75 Maple Leaf (29) 410  
## 76 Markland Wood (12) 413  
## 77 Milliken (130) 1799  
## 78 Mimico (includes Humber Bay Shores) (17) 2568  
## 79 Morningside (135) 1245  
## 80 Moss Park (73) 4786  
## 81 Mount Dennis (115) 953  
## 82 Mount Olive-Silverstone-Jamestown (2) 2591  
## 83 Mount Pleasant East (99) 509  
## 84 Mount Pleasant West (104) 1293  
## 85 New Toronto (18) 996  
## 86 Newtonbrook East (50) 853  
## 87 Newtonbrook West (36) 1693  
## 88 Niagara (82) 2465  
## 89 North Riverdale (68) 888  
## 90 North St.James Town (74) 1721  
## 91 O'Connor-Parkview (54) 1420  
## 92 Oakridge (121) 1660  
## 93 Oakwood Village (107) 1261  
## 94 Old East York (58) 479  
## 95 Palmerston-Little Italy (80) 1053  
## 96 Parkwoods-Donalda (45) 1752  
## 97 Pelmo Park-Humberlea (23) 928  
## 98 Playter Estates-Danforth (67) 707  
## 99 Pleasant View (46) 570  
## 100 Princess-Rosethorn (10) 563  
## 101 Regent Park (72) 1052  
## 102 Rexdale-Kipling (4) 810  
## 103 Rockcliffe-Smythe (111) 1558  
## 104 Roncesvalles (86) 1138  
## 105 Rosedale-Moore Park (98) 1276  
## 106 Rouge (131) 2042  
## 107 Runnymede-Bloor West Village (89) 573  
## 108 Rustic (28) 607  
## 109 Scarborough Village (139) 1528  
## 110 South Parkdale (85) 2121  
## 111 South Riverdale (70) 2726  
## 112 St.Andrew-Windfields (40) 987  
## 113 Steeles (116) 806  
## 114 Stonegate-Queensway (16) 1148  
## 115 Tam O'Shanter-Sullivan (118) 1371  
## 116 Taylor-Massey (61) 1147  
## 117 The Beaches (63) 1116  
## 118 Thistletown-Beaumond Heights (3) 692  
## 119 Thorncliffe Park (55) 903  
## 120 Trinity-Bellwoods (81) 1572  
## 121 University (79) 1455  
## 122 Victoria Village (43) 1158  
## 123 Waterfront Communities-The Island (77) 7747  
## 124 West Hill (136) 3497  
## 125 West Humber-Clairville (1) 5702  
## 126 Westminster-Branson (35) 1218  
## 127 Weston-Pellam Park (91) 858  
## 128 Weston (113) 1942  
## 129 Wexford/Maryvale (119) 2419  
## 130 Willowdale East (51) 1904  
## 131 Willowdale West (37) 1035  
## 132 Willowridge-Martingrove-Richview (7) 1252  
## 133 Woburn (137) 3798  
## 134 Woodbine-Lumsden (60) 377  
## 135 Woodbine Corridor (64) 877  
## 136 Wychwood (94) 801  
## 137 Yonge-Eglinton (100) 823  
## 138 Yonge-St.Clair (97) 412  
## 139 York University Heights (27) 3989  
## 140 Yorkdale-Glen Park (31) 1998

count(crime\_data,"premisetype")

## premisetype freq  
## 1 Apartment 49996  
## 2 Commercial 41081  
## 3 House 37927  
## 4 Other 23178  
## 5 Outside 54253

print("Frequency tabble shows that occurrencemonth and occurrencedayofweek has missing values.")

## [1] "Frequency tabble shows that occurrencemonth and occurrencedayofweek has missing values."

## check records with missing values

missing\_records <- crime\_data[is.na(crime\_data$occurrenceyear),]  
print("AS we look at those 59 records it contains 5 variable with missing values. So it’s better to remove those records as they won’t provide any useful information.")

## [1] "AS we look at those 59 records it contains 5 variable with missing values. So it’s better to remove those records as they won’t provide any useful information."

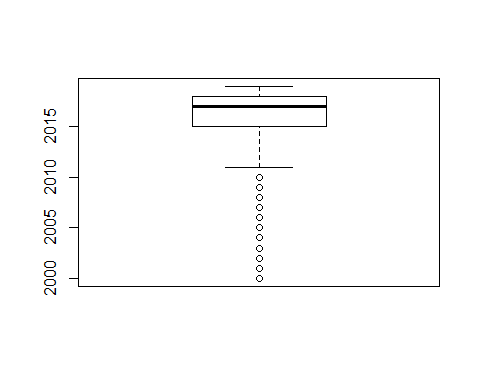
## Dealing with missing values

crime\_data <- crime\_data[!is.na(crime\_data$occurrenceyear),]  
sum(is.na(crime\_data))

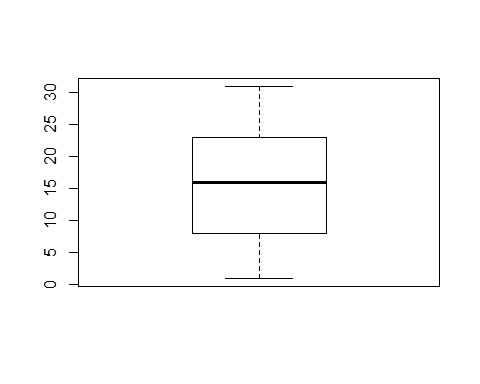
## [1] 0

## Outlier detection using boxplot:

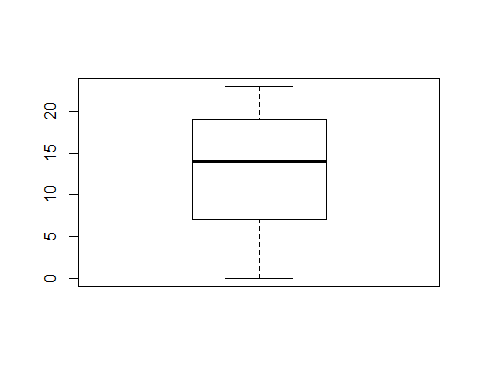
boxplot(crime\_data$occurrenceyear)



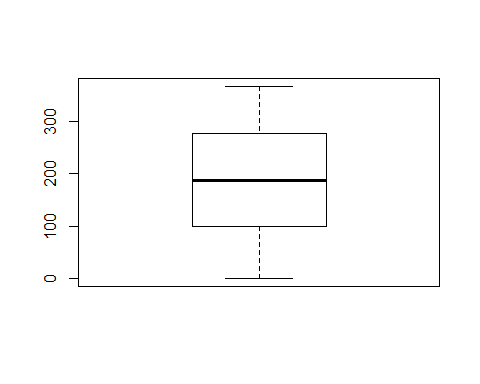
boxplot(crime\_data$occurrenceday)



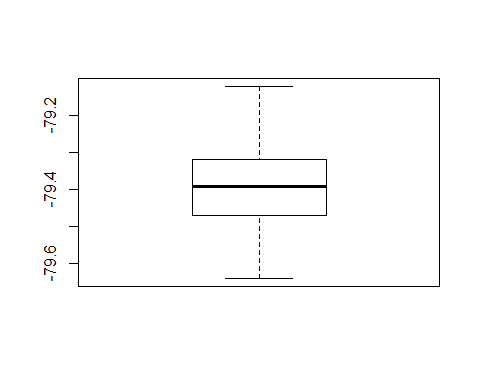
boxplot(crime\_data$occurrencehour)



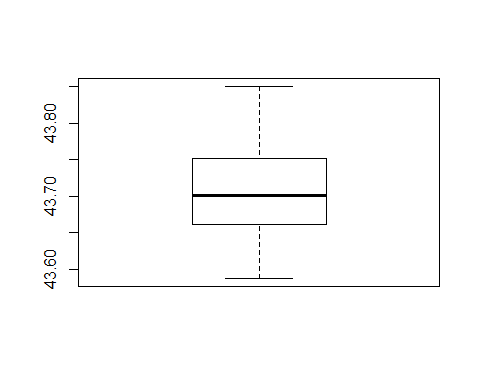
boxplot(crime\_data$occurrencedayofyear)



boxplot(crime\_data$Long)



boxplot(crime\_data$Lat)



print("This shows that only “occurrenceyear” has outliers.As project aim to analyse 2014-2019 data, we will remove records with of 2000 to 2013.")

## [1] "This shows that only “occurrenceyear” has outliers.As project aim to analyse 2014-2019 data, we will remove records with of 2000 to 2013."

## Removing records from year 2000 to 2013:

crime\_data <- crime\_data[(crime\_data$occurrenceyear > 2013),]

## Remove duplicates from dataset

crime\_data <- crime\_data[!duplicated(crime\_data), ]  
  
print("This is our consistence dataset.I will use this dataset for futher analysis.")

## [1] "This is our consistence dataset.I will use this dataset for futher analysis."

## Dimension of consistence dataset

dim(crime\_data)

## [1] 180110 11

## Imbalance data

prop.table(table(crime\_data$MCI))

##   
## Assault Auto Theft Break and Enter Robbery   
## 0.51268114 0.11929376 0.23487313 0.09631892   
## Theft Over   
## 0.03683305

print("This shows data is imbalance, more than 50 percent of data belongs to Assault class.")

## [1] "This shows data is imbalance, more than 50 percent of data belongs to Assault class."