Лабораторна робота №1. Підготовка даних

студентки групи ЕП-61

Лузанової О.С.

Мета роботи – підготовка даних, а саме: збір даних, кодування, обчислення статистик, видалення викидів, шкалювання даних, обробка відсутніх значень, вибір ознак, перехресна перевірка.

Після виконання роботи студент повинен:

ЗНАТИ сутність підготовки та первинної обробки даних.

УМІТИ організувати та провести зібр даних, кодування, обчислення та інтерпретування статистик, видалити викиди та провести шкалювання даних, обробити відсутні значення.

Завдання:

1. Провести збір даних.
2. Зробити кодування.
3. Обчислити та оцінити статистики.
4. Дослідити та видалити викиди.
5. Розробити шкалювання даних.
6. Провести обробку відсутніх значень.
7. Проаналізувати та зробити вибір ознак.
8. Провести перехресну перевірку.

Для виконання Лабораторної роботи №1 було обрано тему – збір та перевірка даних о кількості вбивств та економічних факторів, що впливають на кількість вбивств. Дані зібрані по країнах за 2017 рік.

Щоб виявити економічні фактори, що впливають на кількість вбивст, скоєних у країні, було проведено аналіз наукової літератури. Обрано такі показники, як:

1. Кількість населення (Population).
2. ВВП у доларах США (GDP).
3. ВВП на душу населення у доларах США (GDP\_per\_capita)
4. ВНП у доларах США (GNI).
5. Економічне активне населення (Labor\_force).
6. Рівень безробіття (Unempl).
7. Дохід від с/г, лісного та рибного господарства (Agri\_fore\_fish\_GDP).
8. Ймовірна тривалість життя (Life\_expect).
9. Рівень смертності (Death\_rate).
10. Рівень народжувальності (Birth\_rate).
11. % зайнятих віком від 15 років (Empl).
12. % зайнятих у сфері послуг (Empl\_.serv)
13. % зайнятих у сфері с/г (Empl\_agri).
14. % зайнятих у сфері виробництва (Empl\_indu).
15. Державні видатки у сферу освіти (educ\_expen).
16. Державні видатки у сферу збройних сил (Mil\_expen).
17. Кільксть воєнослужбовців (Arm\_forc\_person).
18. Регіон, до якого входить країна (Region).
19. Підрегіон, до якого входить країна (Subregion).

DATA PREPARATION

# Download data

#Set Working Directory  
#setwd("D:/luzanova")  
#OR Choose your Directory in 'Files' and click on 'More' -> 'Set as Working Directory'   
#Download file to the table. Source file is 'flats.csv'  
f <- read.csv2('crime.csv', header = TRUE, encoding = 'UNICOD')  
#Connect library  
library (dplyr)

##   
## 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

#Have a look at the data  
glimpse(f)

## Observations: 113  
## Variables: 21  
## $ Territory <fct> Kenya, Mauritius, Uganda, Cameroon, Morocco, Esw...  
## $ Homicide <int> 2466, 33, 4735, 341, 761, 130, 20355, 126, 240, ...  
## $ Population <int> 50221473, 1264613, 41162465, 24566045, 35581294,...  
## $ GDP <dbl> 7.896500e+10, 1.325935e+10, 3.075647e+10, 3.5009...  
## $ GDP\_per\_capita <int> 1572, 10485, 747, 1425, 3036, 3953, 6132, 1137, ...  
## $ GNI <dbl> 7.745641e+10, 1.457267e+10, 3.003276e+10, 3.4347...  
## $ Labor\_force <int> 22401022, 608131, 15285775, 10703273, 11742989, ...  
## $ Unempl <int> 3, 7, 2, 3, 9, 23, 27, 2, 6, 12, 4, 3, 10, 10, N...  
## $ Agri\_fore\_fish\_GDP <dbl> 27504481636, 410482173, 7215108857, 5031606455, ...  
## $ Life\_expect <int> 66, 75, 63, 59, 76, 58, 64, 61, 61, 73, 63, 58, ...  
## $ Death\_rate <int> 6, 8, 7, 10, 5, 10, 9, 9, 8, 6, 7, 10, 7, 9, NA,...  
## $ Birth\_rate <int> 29, 11, 39, 36, 19, 27, 21, 37, 38, 20, 30, 36, ...  
## $ Empl <int> 73, 55, 69, 74, 41, 40, 41, 69, 62, 53, 65, 70, ...  
## $ Empl\_serv <int> 37, 68, 20, 40, 42, 63, 71, 41, 40, 64, 48, 24, ...  
## $ Empl\_agri <int> 56, 6, 73, 45, 36, 13, 5, 40, 27, 14, 32, 69, 2,...  
## $ Empl\_indu <int> 7, 26, 7, 15, 22, 24, 23, 19, 33, 22, 20, 7, 15,...  
## $ educ\_expen <dbl> 3826851211, 695116271, 602376430, 920524862, 559...  
## $ Mil\_expen <dbl> 1015384782, 23823068, 346762884, 408157575, 3461...  
## $ Arm\_forc\_person <int> 29000, 3000, 46000, 24000, 246000, NA, 80000, 12...  
## $ Region <fct> Africa, Africa, Africa, Africa, Africa, Africa, ...  
## $ Subregion <fct> Eastern Africa, Eastern Africa, Eastern Africa, ...

head(f)

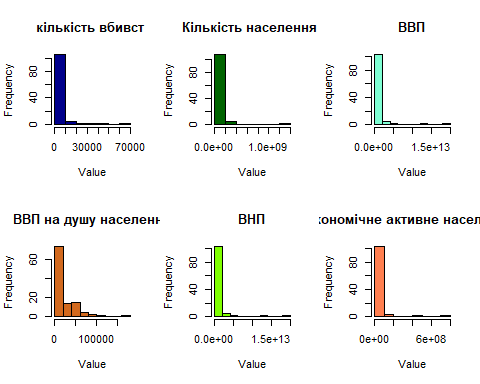
## Territory Homicide Population GDP GDP\_per\_capita GNI  
## 1 Kenya 2466 50221473 78965004656 1572 77456412235  
## 2 Mauritius 33 1264613 13259351418 10485 14572668163  
## 3 Uganda 4735 41162465 30756466548 747 30032761102  
## 4 Cameroon 341 24566045 35009262788 1425 34347942597  
## 5 Morocco 761 35581294 110000000000 3036 108000000000  
## 6 Eswatini 130 1124753 4446248676 3953 4185340308  
## Labor\_force Unempl Agri\_fore\_fish\_GDP Life\_expect Death\_rate Birth\_rate Empl  
## 1 22401022 3 27504481636 66 6 29 73  
## 2 608131 7 410482173 75 8 11 55  
## 3 15285775 2 7215108857 63 7 39 69  
## 4 10703273 3 5031606455 59 10 36 74  
## 5 11742989 9 13559533636 76 5 19 41  
## 6 360335 23 370949317 58 10 27 40  
## Empl\_serv Empl\_agri Empl\_indu educ\_expen Mil\_expen Arm\_forc\_person Region  
## 1 37 56 7 3826851211 1015384782 29000 Africa  
## 2 68 6 26 695116271 23823068 3000 Africa  
## 3 20 73 7 602376430 346762884 46000 Africa  
## 4 40 45 15 920524862 408157575 24000 Africa  
## 5 42 36 22 5593342551 3461461531 246000 Africa  
## 6 63 13 24 284222939 88678494 NA Africa  
## Subregion  
## 1 Eastern Africa  
## 2 Eastern Africa  
## 3 Eastern Africa  
## 4 Middle Africa  
## 5 Northern Africa  
## 6 Southern Africa

#Висновок: кількість спостережень – 113, кількість змінних – 21.

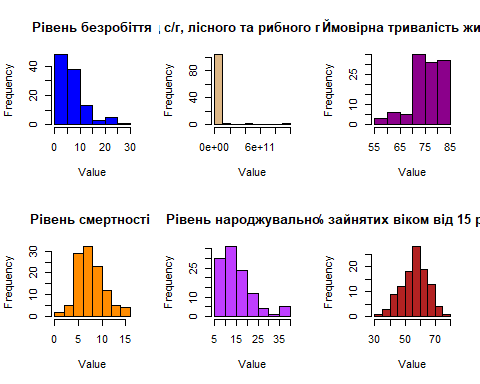
# Visualising

## Histogram

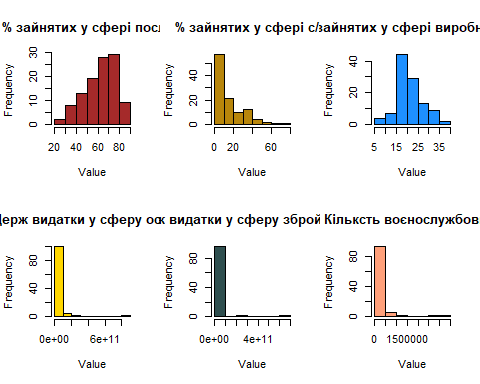
library(ggplot2)  
par(mfrow = c(2, 3))   
hist(f$Homicide, col = 'dark blue', main = 'кількість вбивст ', xlab = 'Value')  
hist(f$Population, col = 'dark green', main = 'Кількість населення ', xlab = 'Value')  
hist(f$GDP, col = 'aquamarine', main = 'ВВП', xlab = 'Value')  
hist(f$GDP\_per\_capita, col = 'chocolate', main = 'ВВП на душу населення', xlab = 'Value')  
hist(f$GNI, col = 'chartreuse', main = 'ВНП', xlab = 'Value')  
hist(f$Labor\_force, col = 'coral', main = 'Економічне активне населення ', xlab = 'Value')



hist(f$Unempl, col = 'blue', main = 'Рівень безробіття', xlab = 'Value')  
hist(f$Agri\_fore\_fish\_GDP, col = 'burlywood', main = 'Дохід від с/г, лісного та рибного господарства', xlab = 'Value')  
hist(f$Life\_expect, col = 'darkmagenta', main = 'Ймовірна тривалість життя', xlab = 'Value')  
hist(f$Death\_rate, col = 'darkorange', main = 'Рівень смертності ', xlab = 'Value')  
hist(f$Birth\_rate, col = 'darkorchid1', main = 'Рівень народжувальності', xlab = 'Value')  
hist(f$Empl, col = 'firebrick', main = '% зайнятих віком від 15 років', xlab = 'Value')



hist(f$Empl\_serv, col = 'brown', main = '% зайнятих у сфері послуг ', xlab = 'Value')  
hist(f$Empl\_agri, col = 'darkgoldenrod', main = '% зайнятих у сфері с/г ', xlab = 'Value')  
hist(f$Empl\_indu, col = 'dodgerblue', main = '% зайнятих у сфері виробництва', xlab = 'Value')  
hist(f$educ\_expen, col = 'gold', main = 'Держ видатки у сферу освіти ', xlab = 'Value')  
hist(f$Mil\_expen, col = 'darkslategray', main = 'Держ видатки у сферу збройних сил', xlab = 'Value')  
hist(f$Arm\_forc\_person, col = 'lightsalmon', main = 'Кільксть воєнослужбовців', xlab = 'Value')



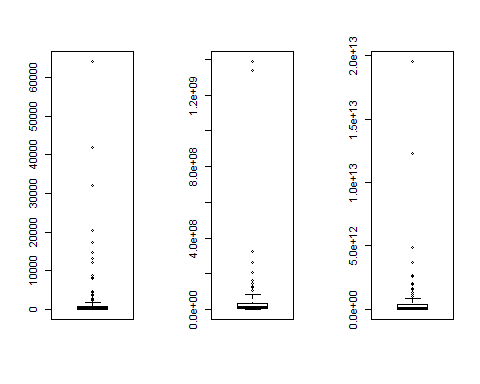
#Висновок: розподіл змінних Homicide, Population, GDP, GNI, Labor\_force, Agri\_fore\_fish\_GDP, educ\_expen, Mil\_expen, Arm\_forc\_person мають довгий хвіст.

## Box-plot

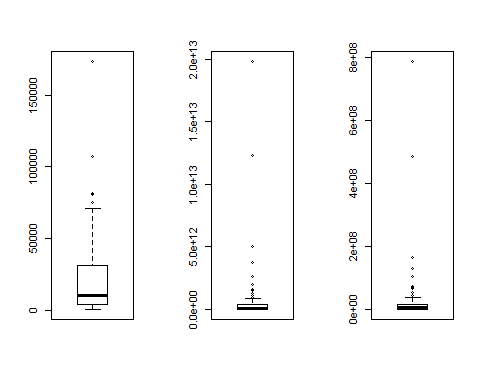
par(mfrow = c(1, 3))  
boxplot(f$Homicide)  
boxplot(f$Population)

## Warning in x[floor(d)] + x[ceiling(d)]: целочисленное переполнение привело к  
## созданию NA

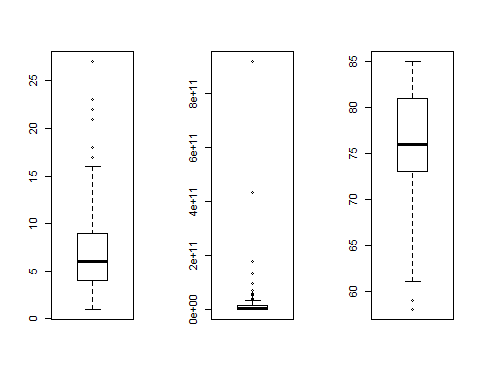
boxplot(f$GDP)



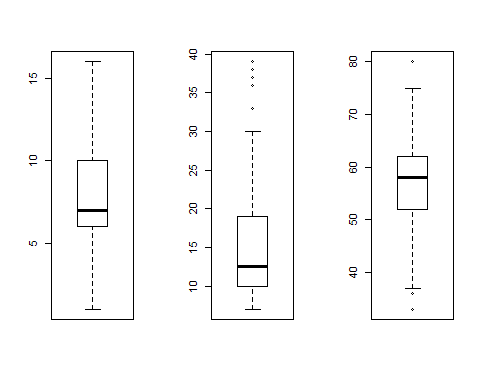
boxplot(f$GDP\_per\_capita)  
boxplot(f$GNI)  
boxplot(f$Labor\_force)



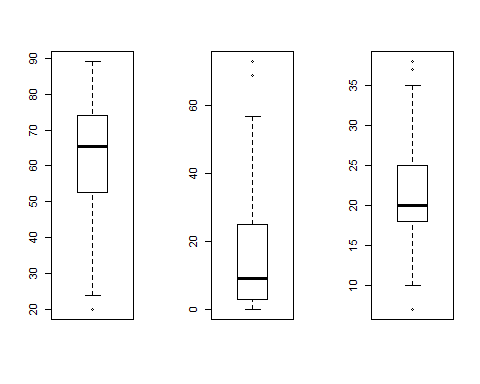
boxplot(f$Unempl)  
boxplot(f$Agri\_fore\_fish\_GDP)  
boxplot(f$Life\_expect)



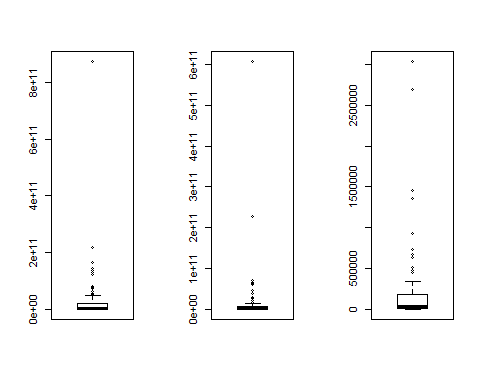
boxplot(f$Death\_rate)  
boxplot(f$Birth\_rate)  
boxplot(f$Empl)



boxplot(f$Empl\_serv)  
boxplot(f$Empl\_agri)  
boxplot(f$Empl\_indu)



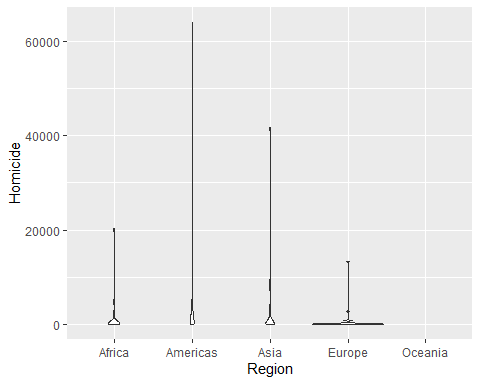
boxplot(f$educ\_expen)  
boxplot(f$Mil\_expen)  
boxplot(f$Arm\_forc\_person)



#Висновок: всі змінні мають викиди, окрім Death\_rate. Малу кількість викидів мають GDP\_per\_capita, Unempl, Life\_expectancy, Death\_rate, Birth\_rate, Empl, Empl\_.serv, Empl\_agri, Empl\_indu

## Violin

qplot(data = f,   
 x = Region,   
 y = Homicide,   
 geom = "violin")



#Висновок: більшість нетипових значень змінної Homicide належать до країн Americas Region

# Statistics

## Descriptive statistics

library (psych)

## Warning: package 'psych' was built under R version 3.6.3

##   
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':  
##   
## %+%, alpha

describe(f)

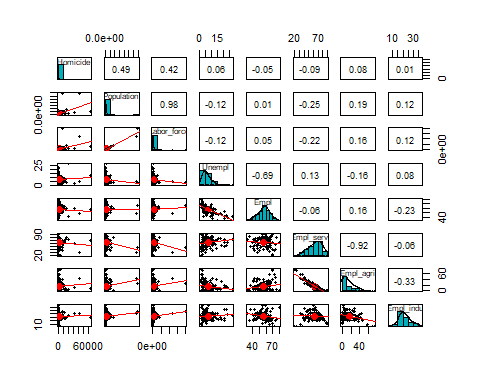
## vars n mean sd median trimmed  
## Territory\* 1 113 5.700000e+01 3.276000e+01 5.700000e+01 5.700000e+01  
## Homicide 2 113 2.657780e+03 8.286660e+03 2.010000e+02 6.524700e+02  
## Population 3 113 5.318512e+07 1.844720e+08 9.429013e+06 1.775540e+07  
## GDP 4 111 6.890122e+11 2.265672e+12 7.896500e+10 2.348678e+11  
## GDP\_per\_capita 5 111 2.113142e+04 2.626334e+04 1.048500e+04 1.619009e+04  
## GNI 6 111 6.901319e+11 2.292120e+12 7.011124e+10 2.306630e+11  
## Labor\_force 7 108 2.590930e+07 9.025895e+07 4.763391e+06 9.025446e+06  
## Unempl 8 108 7.380000e+00 5.080000e+00 6.000000e+00 6.520000e+00  
## Agri\_fore\_fish\_GDP 9 108 2.591719e+10 9.916110e+10 4.116603e+09 8.642335e+09  
## Life\_expect 10 112 7.585000e+01 6.080000e+00 7.600000e+01 7.658000e+01  
## Death\_rate 11 112 8.040000e+00 2.890000e+00 7.000000e+00 7.840000e+00  
## Birth\_rate 12 112 1.557000e+01 7.320000e+00 1.250000e+01 1.441000e+01  
## Empl 13 108 5.679000e+01 9.010000e+00 5.800000e+01 5.699000e+01  
## Empl\_serv 14 108 6.294000e+01 1.482000e+01 6.550000e+01 6.388000e+01  
## Empl\_agri 15 108 1.542000e+01 1.564000e+01 9.000000e+00 1.306000e+01  
## Empl\_indu 16 108 2.167000e+01 6.130000e+00 2.000000e+01 2.153000e+01  
## educ\_expen 17 107 2.820861e+10 9.092148e+10 3.826851e+09 1.133980e+10  
## Mil\_expen 18 98 1.646204e+10 6.598654e+10 1.903126e+09 4.749618e+09  
## Arm\_forc\_person 19 102 1.961471e+05 4.541384e+05 4.150000e+04 9.124390e+04  
## Region\* 20 113 2.930000e+00 1.060000e+00 3.000000e+00 3.010000e+00  
## Subregion\* 21 113 1.118000e+01 5.620000e+00 1.200000e+01 1.138000e+01  
## mad min max range skew  
## Territory\* 4.151000e+01 1 1.130000e+02 1.120000e+02 0.00  
## Homicide 2.624200e+02 1 6.407800e+04 6.407700e+04 5.12  
## Population 1.283011e+07 37800 1.386395e+09 1.386357e+09 6.38  
## GDP 1.118026e+11 519837037 1.950000e+13 1.949948e+13 6.45  
## GDP\_per\_capita 1.067175e+04 556 1.733560e+05 1.728000e+05 2.55  
## GNI 9.879861e+10 508221815 1.980000e+13 1.979949e+13 6.48  
## Labor\_force 6.223672e+06 98287 7.848395e+08 7.847412e+08 6.76  
## Unempl 2.970000e+00 1 2.700000e+01 2.600000e+01 1.66  
## Agri\_fore\_fish\_GDP 5.450315e+09 8632071 9.190000e+11 9.189914e+11 7.45  
## Life\_expect 5.930000e+00 58 8.500000e+01 2.700000e+01 -0.95  
## Death\_rate 2.970000e+00 1 1.600000e+01 1.500000e+01 0.51  
## Birth\_rate 5.190000e+00 7 3.900000e+01 3.200000e+01 1.36  
## Empl 7.410000e+00 33 8.000000e+01 4.700000e+01 -0.21  
## Empl\_serv 1.557000e+01 20 8.900000e+01 6.900000e+01 -0.58  
## Empl\_agri 1.038000e+01 0 7.300000e+01 7.300000e+01 1.40  
## Empl\_indu 4.450000e+00 7 3.800000e+01 3.100000e+01 0.22  
## educ\_expen 5.088062e+09 15655251 8.720000e+11 8.719843e+11 7.64  
## Mil\_expen 2.652582e+09 0 6.060000e+11 6.060000e+11 7.62  
## Arm\_forc\_person 5.411490e+04 0 3.031000e+06 3.031000e+06 4.44  
## Region\* 1.480000e+00 1 5.000000e+00 4.000000e+00 -0.31  
## Subregion\* 7.410000e+00 1 1.900000e+01 1.800000e+01 -0.28  
## kurtosis se  
## Territory\* -1.23 3.080000e+00  
## Homicide 29.90 7.795400e+02  
## Population 42.08 1.735367e+07  
## GDP 46.18 2.150480e+11  
## GDP\_per\_capita 9.43 2.492810e+03  
## GNI 46.62 2.175584e+11  
## Labor\_force 49.52 8.685172e+06  
## Unempl 2.67 4.900000e-01  
## Agri\_fore\_fish\_GDP 60.65 9.541782e+09  
## Life\_expect 0.72 5.700000e-01  
## Death\_rate 0.09 2.700000e-01  
## Birth\_rate 1.49 6.900000e-01  
## Empl -0.06 8.700000e-01  
## Empl\_serv -0.26 1.430000e+00  
## Empl\_agri 1.75 1.510000e+00  
## Empl\_indu 0.25 5.900000e-01  
## educ\_expen 66.64 8.789711e+09  
## Mil\_expen 63.11 6.665647e+09  
## Arm\_forc\_person 22.00 4.496640e+04  
## Region\* -1.03 1.000000e-01  
## Subregion\* -1.27 5.300000e-01

#Висновок: аналіз основних показників описової статистики за кожною змінною показав, що є пропущені значення в усіх змінних, окрім – Homicide, Population, Region, Subregion. ЯК ПРОВЕСТИ АНАЛІЗ ВИКИДІВ?

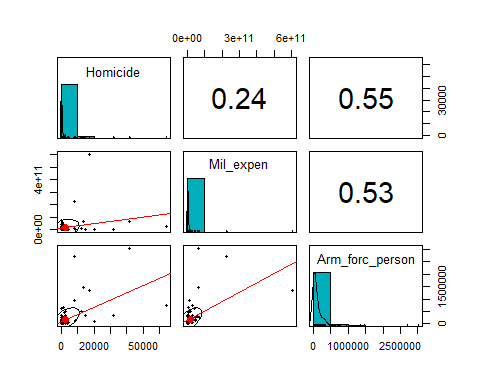
## Correlations

#Побудуемо кореляційну матрицю. Попередньо зробимо групування показників з додаванням показника Homicide у кожну. 1-а група - Кількість населення, Економічне активне населення, Рівень безробіття, % зайнятих віком від 15 років, % зайнятих у сфері послуг, % зайнятих у сфері с/г, % зайнятих у сфері виробництва; 2-а група - Державні видатки у сферу збройних сил, Кільксть воєнослужбовців; 3-я група -Ймовірна тривалість життя, Рівень смертності, Рівень народжувальності; Державні видатки у сферу освіти 4-а група - ВВП, ВВП на душу населення, ВНП, Дохід від с/г, лісного та рибного господарства

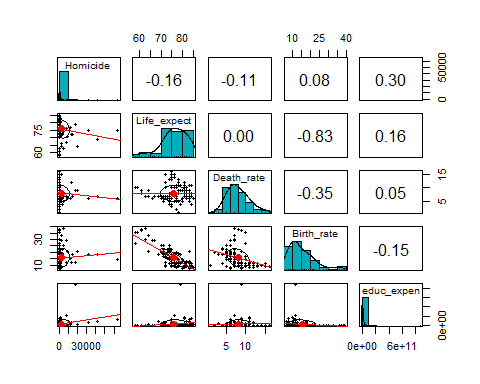
f1 <- c(2,3,7,8,13,14,15,16)  
c\_f1 <- f[f1]  
pairs.panels(c\_f1, lm=TRUE, # linear fit  
 method = "pearson", # correlation method  
 hist.col = "#00AFBB"  
 )



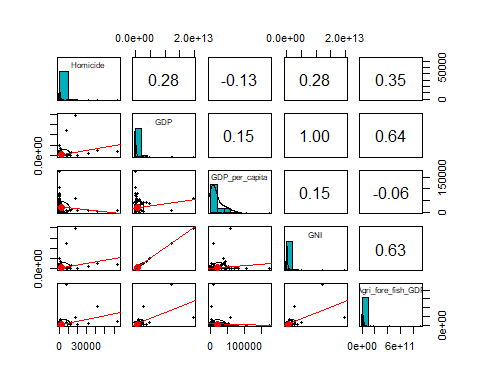
f2 <- c(2,18,19)  
c\_f2 <- f[f2]  
pairs.panels(c\_f2, lm=TRUE, # linear fit  
 method = "pearson", # correlation method  
 hist.col = "#00AFBB"  
 )



f3 <- c(2,10,11,12,17)  
c\_f3 <- f[f3]  
pairs.panels(c\_f3, lm=TRUE, # linear fit  
 method = "pearson", # correlation method  
 hist.col = "#00AFBB"  
 )

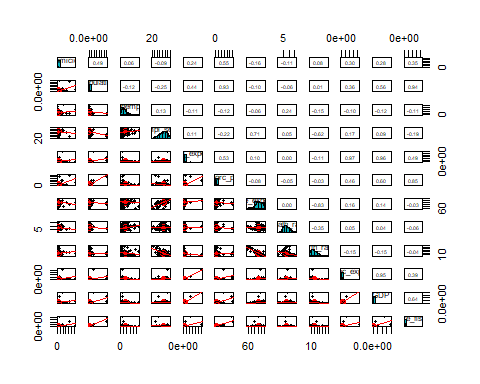


f4 <- c(2,4,5,6,9)  
c\_f4 <- f[f4]  
pairs.panels(c\_f4, lm=TRUE, # linear fit  
 method = "pearson", # correlation method  
 hist.col = "#00AFBB"  
 )



#Змінні, що мають незначний вплив на Homicide (з рівнем менш за 0,05) та ті, що корелюють між собою не вкдючати до кореляційного аналізу

f5 <- c( 2,3,8,14,18,19,10,11,12,17,4,9)  
c\_f5 <- f[f5]  
pairs.panels(c\_f5, lm=TRUE, # linear fit  
 method = "pearson", # correlation method  
 hist.col = "#00AFBB"  
 )



#Тісний зв’язок виявлено між змінними Кількість населення та Кільксть воєнослужбовців; Кількість населення та Дохід від с/г, лісного та рибного господарства; Державні видатки у сферу освіти та Державні видатки у сферу збройних сил; Державні видатки у сферу збройних сил та ВВП; Кільксть воєнослужбовців та Дохід від с/г, лісного та рибного господарства; Ймовірна тривалість життя та Рівень народжувальності; Державні видатки у сферу освіти та ВВП. Рекомендовано, з моделі видалити змінні Кількість населення, Державні видатки у сферу збройних сил, Рівень народжувальності; ВВП.

# Missing data

## Delete N/A (not recommended strategy)

f\_reduce <- tidyr::drop\_na(f)  
cat('there are',nrow(f\_reduce),'rows in the f\_reduce')

## there are 92 rows in the f\_reduce

#Після видалення пропущених значень в базі залишилося би 92 рядка.

## Fill n/a

#1 with neighboring values  
f\_fill1 <- tidyr::fill(f, Population,GDP,GDP\_per\_capita,GNI,Labor\_force,Unempl,Agri\_fore\_fish\_GDP,Life\_expect,Death\_rate,Birth\_rate,Empl,Empl\_serv,Empl\_agri,Empl\_indu,educ\_expen,Mil\_expen,Arm\_forc\_person, .direction = 'down')   
  
head(f\_fill1)

## Territory Homicide Population GDP GDP\_per\_capita GNI  
## 1 Kenya 2466 50221473 78965004656 1572 77456412235  
## 2 Mauritius 33 1264613 13259351418 10485 14572668163  
## 3 Uganda 4735 41162465 30756466548 747 30032761102  
## 4 Cameroon 341 24566045 35009262788 1425 34347942597  
## 5 Morocco 761 35581294 110000000000 3036 108000000000  
## 6 Eswatini 130 1124753 4446248676 3953 4185340308  
## Labor\_force Unempl Agri\_fore\_fish\_GDP Life\_expect Death\_rate Birth\_rate Empl  
## 1 22401022 3 27504481636 66 6 29 73  
## 2 608131 7 410482173 75 8 11 55  
## 3 15285775 2 7215108857 63 7 39 69  
## 4 10703273 3 5031606455 59 10 36 74  
## 5 11742989 9 13559533636 76 5 19 41  
## 6 360335 23 370949317 58 10 27 40  
## Empl\_serv Empl\_agri Empl\_indu educ\_expen Mil\_expen Arm\_forc\_person Region  
## 1 37 56 7 3826851211 1015384782 29000 Africa  
## 2 68 6 26 695116271 23823068 3000 Africa  
## 3 20 73 7 602376430 346762884 46000 Africa  
## 4 40 45 15 920524862 408157575 24000 Africa  
## 5 42 36 22 5593342551 3461461531 246000 Africa  
## 6 63 13 24 284222939 88678494 246000 Africa  
## Subregion  
## 1 Eastern Africa  
## 2 Eastern Africa  
## 3 Eastern Africa  
## 4 Middle Africa  
## 5 Northern Africa  
## 6 Southern Africa

#2 with average or the most frequent  
f\_fill2 <- f  
##with average for integer vars   
f\_fill2$Population <- ifelse(is.na(f$Population),round(mean(f$Population,na.rm = TRUE)),f$Population)  
f\_fill2$GDP <- ifelse(is.na(f$GDP),round(mean(f$GDP,na.rm = TRUE)),f$GDP)  
f\_fill2$GDP\_per\_capita <- ifelse(is.na(f$GDP\_per\_capita),round(mean(f$GDP\_per\_capita,na.rm = TRUE)),f$GDP\_per\_capita)  
f\_fill2$GNI <- ifelse(is.na(f$GNI),round(mean(f$GNI,na.rm = TRUE)),f$GNI)  
f\_fill2$Labor\_force <- ifelse(is.na(f$Labor\_force),round(mean(f$Labor\_force,na.rm = TRUE)),f$Labor\_force)  
f\_fill2$Unempl <- ifelse(is.na(f$Unempl),round(mean(f$Unempl,na.rm = TRUE)),f$Unempl)  
f\_fill2$Agri\_fore\_fish\_GDP <- ifelse(is.na(f$Agri\_fore\_fish\_GDP),round(mean(f$Agri\_fore\_fish\_GDP,na.rm = TRUE)),f$Agri\_fore\_fish\_GDP)  
f\_fill2$Life\_expect <- ifelse(is.na(f$Life\_expect),round(mean(f$Life\_expect,na.rm = TRUE)),f$Life\_expect)  
f\_fill2$Death\_rate <- ifelse(is.na(f$Death\_rate),round(mean(f$Death\_rate,na.rm = TRUE)),f$Death\_rate)  
f\_fill2$Birth\_rate <- ifelse(is.na(f$Birth\_rate),round(mean(f$Birth\_rate,na.rm = TRUE)),f$Birth\_rate)  
f\_fill2$Empl <- ifelse(is.na(f$Empl),round(mean(f$Empl,na.rm = TRUE)),f$Empl)  
f\_fill2$Empl\_serv <- ifelse(is.na(f$Empl\_serv),round(mean(f$Empl\_serv,na.rm = TRUE)),f$Empl\_serv)  
f\_fill2$Empl\_agri <- ifelse(is.na(f$Empl\_agri),round(mean(f$Empl\_agri,na.rm = TRUE)),f$Empl\_agri)  
f\_fill2$Empl\_indu <- ifelse(is.na(f$Empl\_indu),round(mean(f$Empl\_indu,na.rm = TRUE)),f$Empl\_indu)  
f\_fill2$educ\_expen <- ifelse(is.na(f$educ\_expen),round(mean(f$educ\_expen,na.rm = TRUE)),f$educ\_expen)  
f\_fill2$Mil\_expen <- ifelse(is.na(f$Mil\_expen),round(mean(f$Mil\_expen,na.rm = TRUE)),f$Mil\_expen)  
f\_fill2$Arm\_forc\_person<- ifelse(is.na(f$Arm\_forc\_person),round(mean(f$Arm\_forc\_person,na.rm = TRUE)),f$Arm\_forc\_person)  
  
  
head(f\_fill2)

## Territory Homicide Population GDP GDP\_per\_capita GNI  
## 1 Kenya 2466 50221473 78965004656 1572 77456412235  
## 2 Mauritius 33 1264613 13259351418 10485 14572668163  
## 3 Uganda 4735 41162465 30756466548 747 30032761102  
## 4 Cameroon 341 24566045 35009262788 1425 34347942597  
## 5 Morocco 761 35581294 110000000000 3036 108000000000  
## 6 Eswatini 130 1124753 4446248676 3953 4185340308  
## Labor\_force Unempl Agri\_fore\_fish\_GDP Life\_expect Death\_rate Birth\_rate Empl  
## 1 22401022 3 27504481636 66 6 29 73  
## 2 608131 7 410482173 75 8 11 55  
## 3 15285775 2 7215108857 63 7 39 69  
## 4 10703273 3 5031606455 59 10 36 74  
## 5 11742989 9 13559533636 76 5 19 41  
## 6 360335 23 370949317 58 10 27 40  
## Empl\_serv Empl\_agri Empl\_indu educ\_expen Mil\_expen Arm\_forc\_person Region  
## 1 37 56 7 3826851211 1015384782 29000 Africa  
## 2 68 6 26 695116271 23823068 3000 Africa  
## 3 20 73 7 602376430 346762884 46000 Africa  
## 4 40 45 15 920524862 408157575 24000 Africa  
## 5 42 36 22 5593342551 3461461531 246000 Africa  
## 6 63 13 24 284222939 88678494 196147 Africa  
## Subregion  
## 1 Eastern Africa  
## 2 Eastern Africa  
## 3 Eastern Africa  
## 4 Middle Africa  
## 5 Northern Africa  
## 6 Southern Africa

#let's work with "filled with ave" data   
f <- f\_fill2

#Для заповнення пропусків обрано варіант заповнення середніми для кількісних змінних

# Ejections (outside the three sigma)

## Remove the ejections (not recommended strategy)

f\_ej1 <- f[f$Population < mean(f$Population)+sd(f$Population)\*3, ]  
f\_ej1 <- f[f$GDP < mean(f$GDP)+sd(f$GDP)\*3, ]  
f\_ej1 <- f[f$GDP\_per\_capita < mean(f$GDP\_per\_capita)+sd(f$GDP\_per\_capita)\*3, ]  
f\_ej1 <- f[f$GNI < mean(f$GNI)+sd(f$GNI)\*3, ]  
f\_ej1 <- f[f$Labor\_force < mean(f$Labor\_force)+sd(f$Labor\_force)\*3, ]  
f\_ej1 <- f[f$Unempl < mean(f$Unempl)+sd(f$Unempl)\*3, ]  
f\_ej1 <- f[f$Agri\_fore\_fish\_GDP < mean(f$Agri\_fore\_fish\_GDP)+sd(f$Agri\_fore\_fish\_GDP)\*3, ]  
f\_ej1 <- f[f$Life\_expect < mean(f$Life\_expect)+sd(f$Life\_expect)\*3, ]  
f\_ej1 <- f[f$Death\_rate < mean(f$Death\_rate)+sd(f$Death\_rate)\*3, ]  
f\_ej1 <- f[f$Birth\_rate < mean(f$Birth\_rate)+sd(f$Birth\_rate)\*3, ]  
f\_ej1 <- f[f$Empl < mean(f$Empl)+sd(f$Empl)\*3, ]  
f\_ej1 <- f[f$Empl\_serv < mean(f$Empl\_serv)+sd(f$Empl\_serv)\*3, ]  
f\_ej1 <- f[f$Empl\_agri < mean(f$Empl\_agri)+sd(f$Empl\_agri)\*3, ]  
f\_ej1 <- f[f$Empl\_indu < mean(f$Empl\_indu)+sd(f$Empl\_indu)\*3, ]  
f\_ej1 <- f[f$educ\_expen < mean(f$educ\_expen)+sd(f$educ\_expen)\*3, ]  
f\_ej1 <- f[f$Mil\_expen < mean(f$Mil\_expen)+sd(f$Mil\_expen)\*3, ]  
f\_ej1 <- f[f$Arm\_forc\_person < mean(f$Arm\_forc\_person)+sd(f$Arm\_forc\_person)\*3, ]  
describe(f\_ej1)

## vars n mean sd median trimmed  
## Territory\* 1 111 5.733000e+01 3.291000e+01 58 5.742000e+01  
## Homicide 2 111 2.256820e+03 7.452880e+03 193 5.643800e+02  
## Population 3 111 2.959337e+07 5.390914e+07 8797566 1.631441e+07  
## GDP 4 111 5.667421e+11 1.964086e+12 78965004656 2.177670e+11  
## GDP\_per\_capita 5 111 2.141432e+04 2.617219e+04 10720 1.654291e+04  
## GNI 6 111 5.681523e+11 1.994890e+12 70111244439 2.142615e+11  
## Labor\_force 7 111 1.494306e+07 2.563936e+07 4838602 8.974012e+06  
## Unempl 8 111 7.410000e+00 4.990000e+00 6 6.550000e+00  
## Agri\_fore\_fish\_GDP 9 111 1.419498e+10 2.571556e+10 4140506133 8.692822e+09  
## Life\_expect 10 111 7.591000e+01 6.070000e+00 76 7.666000e+01  
## Death\_rate 11 111 8.050000e+00 2.900000e+00 8 7.870000e+00  
## Birth\_rate 12 111 1.559000e+01 7.340000e+00 13 1.442000e+01  
## Empl 13 111 5.680000e+01 8.800000e+00 57 5.702000e+01  
## Empl\_serv 14 111 6.339000e+01 1.419000e+01 64 6.436000e+01  
## Empl\_agri 15 111 1.504000e+01 1.514000e+01 10 1.261000e+01  
## Empl\_indu 16 111 2.159000e+01 6.000000e+00 21 2.142000e+01  
## educ\_expen 17 111 2.600677e+10 8.718745e+10 4111671813 1.081653e+10  
## Mil\_expen 18 111 1.412298e+10 5.836334e+10 3138359204 5.664914e+09  
## Arm\_forc\_person 19 111 1.480956e+05 2.391929e+05 49000 9.540019e+04  
## Region\* 20 111 2.930000e+00 1.070000e+00 3 3.010000e+00  
## Subregion\* 21 111 1.119000e+01 5.640000e+00 12 1.140000e+01  
## mad min max range skew  
## Territory\* 4.151000e+01 1 1.130000e+02 1.120000e+02 -0.02  
## Homicide 2.505600e+02 1 6.407800e+04 6.407700e+04 6.01  
## Population 1.193789e+07 37800 3.249855e+08 3.249477e+08 3.19  
## GDP 1.118026e+11 519837037 1.950000e+13 1.949948e+13 8.22  
## GDP\_per\_capita 1.126331e+04 556 1.733560e+05 1.728000e+05 2.55  
## GNI 9.879861e+10 508221815 1.980000e+13 1.979949e+13 8.23  
## Labor\_force 6.330506e+06 98287 1.642681e+08 1.641698e+08 3.33  
## Unempl 2.970000e+00 1 2.700000e+01 2.600000e+01 1.69  
## Agri\_fore\_fish\_GDP 5.530134e+09 8632071 1.790000e+11 1.789914e+11 3.77  
## Life\_expect 5.930000e+00 58 8.500000e+01 2.700000e+01 -0.98  
## Death\_rate 2.970000e+00 1 1.600000e+01 1.500000e+01 0.49  
## Birth\_rate 5.930000e+00 7 3.900000e+01 3.200000e+01 1.36  
## Empl 5.930000e+00 33 8.000000e+01 4.700000e+01 -0.22  
## Empl\_serv 1.483000e+01 20 8.900000e+01 6.900000e+01 -0.62  
## Empl\_agri 1.038000e+01 0 7.300000e+01 7.300000e+01 1.51  
## Empl\_indu 4.450000e+00 7 3.800000e+01 3.100000e+01 0.25  
## educ\_expen 5.674576e+09 15655251 8.720000e+11 8.719843e+11 8.33  
## Mil\_expen 4.440111e+09 0 6.060000e+11 6.060000e+11 9.41  
## Arm\_forc\_person 6.671700e+04 0 1.454000e+06 1.454000e+06 3.28  
## Region\* 1.480000e+00 1 5.000000e+00 4.000000e+00 -0.30  
## Subregion\* 7.410000e+00 1 1.900000e+01 1.800000e+01 -0.29  
## kurtosis se  
## Territory\* -1.24 3.120000e+00  
## Homicide 42.58 7.074000e+02  
## Population 11.30 5.116829e+06  
## GDP 75.09 1.864228e+11  
## GDP\_per\_capita 9.49 2.484150e+03  
## GNI 75.14 1.893466e+11  
## Labor\_force 13.20 2.433580e+06  
## Unempl 2.86 4.700000e-01  
## Agri\_fore\_fish\_GDP 17.79 2.440813e+09  
## Life\_expect 0.80 5.800000e-01  
## Death\_rate 0.07 2.800000e-01  
## Birth\_rate 1.46 7.000000e-01  
## Empl 0.12 8.400000e-01  
## Empl\_serv 0.00 1.350000e+00  
## Empl\_agri 2.30 1.440000e+00  
## Empl\_indu 0.43 5.700000e-01  
## educ\_expen 77.01 8.275466e+09  
## Mil\_expen 92.31 5.539603e+09  
## Arm\_forc\_person 12.84 2.270319e+04  
## Region\* -1.07 1.000000e-01  
## Subregion\* -1.27 5.400000e-01

#Висновок: після видалення викидів змінної price в базі залишилося би 111 рядків.

## Replace with max

f\_ej2 <- f  
f\_ej2$Population <- ifelse(f$Population < mean(f$Population)+sd(f$Population)\*3,f$Population,mean(f$Population)+sd(f$Population)\*3)  
f\_ej2$GDP <- ifelse(f$GDP < mean(f$GDP)+sd(f$GDP)\*3,f$GDP,mean(f$GDP)+sd(f$GDP)\*3)  
f\_ej2$GDP\_per\_capita <- ifelse(f$GDP\_per\_capita < mean(f$GDP\_per\_capita)+sd(f$GDP\_per\_capita)\*3,f$GDP\_per\_capita,mean(f$GDP\_per\_capita)+sd(f$GDP\_per\_capita)\*3)  
f\_ej2$GNI <- ifelse(f$GNI < mean(f$GNI)+sd(f$GNI)\*3,f$GNI,mean(f$GNI)+sd(f$GNI)\*3)  
f\_ej2$Labor\_force <- ifelse(f$Labor\_force < mean(f$Labor\_force)+sd(f$Labor\_force)\*3,f$Labor\_force,mean(f$Labor\_force)+sd(f$Labor\_force)\*3)  
f\_ej2$Unempl <- ifelse(f$Unempl < mean(f$Unempl)+sd(f$Unempl)\*3,f$Unempl,mean(f$Unempl)+sd(f$Unempl)\*3)  
f\_ej2$Agri\_fore\_fish\_GDP <- ifelse(f$Agri\_fore\_fish\_GDP < mean(f$Agri\_fore\_fish\_GDP)+sd(f$Agri\_fore\_fish\_GDP)\*3,f$Agri\_fore\_fish\_GDP,mean(f$Agri\_fore\_fish\_GDP)+sd(f$Agri\_fore\_fish\_GDP)\*3)  
f\_ej2$Life\_expect <- ifelse(f$Life\_expect < mean(f$Life\_expect)+sd(f$Life\_expect)\*3,f$Life\_expect,mean(f$Life\_expect)+sd(f$Life\_expect)\*3)  
f\_ej2$Death\_rate <- ifelse(f$Death\_rate < mean(f$Death\_rate)+sd(f$Death\_rate)\*3,f$Death\_rate,mean(f$Death\_rate)+sd(f$Death\_rate)\*3)  
f\_ej2$Birth\_rate <- ifelse(f$Birth\_rate < mean(f$Birth\_rate)+sd(f$Birth\_rate)\*3,f$Birth\_rate,mean(f$Birth\_rate)+sd(f$Birth\_rate)\*3)  
f\_ej2$Empl<- ifelse(f$Empl < mean(f$Empl)+sd(f$Empl)\*3,f$Empl,mean(f$Empl)+sd(f$Empl)\*3)  
f\_ej2$Empl\_serv <- ifelse(f$Empl\_serv < mean(f$Empl\_serv)+sd(f$Empl\_serv)\*3,f$Empl\_serv,mean(f$Empl\_serv)+sd(f$Empl\_serv)\*3)  
f\_ej2$Empl\_agri <- ifelse(f$Empl\_agri < mean(f$Empl\_agri)+sd(f$Empl\_agri)\*3,f$Empl\_agri,mean(f$Empl\_agri)+sd(f$Empl\_agri)\*3)  
f\_ej2$Empl\_indu <- ifelse(f$Empl\_indu < mean(f$Empl\_indu)+sd(f$Empl\_indu)\*3,f$Empl\_indu,mean(f$Empl\_indu)+sd(f$Empl\_indu)\*3)  
f\_ej2$educ\_expen <- ifelse(f$educ\_expen < mean(f$educ\_expen)+sd(f$educ\_expen)\*3,f$educ\_expen,mean(f$educ\_expen)+sd(f$educ\_expen)\*3)  
f\_ej2$Mil\_expen <- ifelse(f$Mil\_expen < mean(f$Mil\_expen)+sd(f$Mil\_expen)\*3,f$Mil\_expen,mean(f$Mil\_expen)+sd(f$Mil\_expen)\*3)  
f\_ej2$Arm\_forc\_person <- ifelse(f$Arm\_forc\_person < mean(f$Arm\_forc\_person)+sd(f$Arm\_forc\_person)\*3,f$Arm\_forc\_person,mean(f$Arm\_forc\_person)+sd(f$Arm\_forc\_person)\*3)  
  
describe(f\_ej2)

## vars n mean sd median trimmed  
## Territory\* 1 113 5.700000e+01 3.276000e+01 57 5.700000e+01  
## Homicide 2 113 2.657780e+03 8.286660e+03 201 6.524700e+02  
## Population 3 113 3.980590e+07 9.324374e+07 9429013 1.775540e+07  
## GDP 4 113 5.390134e+11 1.210805e+12 79997975622 2.448490e+11  
## GDP\_per\_capita 5 113 2.040085e+04 2.262935e+04 10514 1.629868e+04  
## GNI 6 113 5.388897e+11 1.225614e+12 76204178959 2.407613e+11  
## Labor\_force 7 113 1.982147e+07 4.447771e+07 4919831 9.672367e+06  
## Unempl 8 113 7.310000e+00 4.790000e+00 6 6.510000e+00  
## Agri\_fore\_fish\_GDP 9 113 1.954877e+10 4.748145e+10 4274025159 9.348282e+09  
## Life\_expect 10 113 7.585000e+01 6.050000e+00 76 7.657000e+01  
## Death\_rate 11 113 8.040000e+00 2.880000e+00 7 7.850000e+00  
## Birth\_rate 12 113 1.556000e+01 7.230000e+00 13 1.443000e+01  
## Empl 13 113 5.680000e+01 8.810000e+00 57 5.701000e+01  
## Empl\_serv 14 113 6.294000e+01 1.449000e+01 64 6.390000e+01  
## Empl\_agri 15 113 1.523000e+01 1.470000e+01 10 1.302000e+01  
## Empl\_indu 16 113 2.168000e+01 5.990000e+00 21 2.153000e+01  
## educ\_expen 17 113 2.308973e+10 4.548761e+10 4398152687 1.190058e+10  
## Mil\_expen 18 113 1.263353e+10 2.931955e+10 3429968212 6.135345e+09  
## Arm\_forc\_person 19 113 1.718449e+05 2.962678e+05 52000 1.008859e+05  
## Region\* 20 113 2.930000e+00 1.060000e+00 3 3.010000e+00  
## Subregion\* 21 113 1.118000e+01 5.620000e+00 12 1.138000e+01  
## mad min max range skew  
## Territory\* 4.151000e+01 1 1.130000e+02 1.120000e+02 0.00  
## Homicide 2.624200e+02 1 6.407800e+04 6.407700e+04 5.12  
## Population 1.283011e+07 37800 6.066012e+08 6.065634e+08 4.47  
## GDP 1.138458e+11 519837037 7.425066e+12 7.424546e+12 3.97  
## GDP\_per\_capita 1.095790e+04 556 9.921478e+04 9.865878e+04 1.55  
## GNI 1.079271e+11 508221815 7.504819e+12 7.504311e+12 3.99  
## Labor\_force 6.388150e+06 98287 2.905730e+08 2.904747e+08 4.51  
## Unempl 2.970000e+00 1 2.225000e+01 2.125000e+01 1.55  
## Agri\_fore\_fish\_GDP 5.728089e+09 8632071 3.166844e+11 3.166758e+11 4.83  
## Life\_expect 5.930000e+00 58 8.500000e+01 2.700000e+01 -0.96  
## Death\_rate 2.970000e+00 1 1.600000e+01 1.500000e+01 0.51  
## Birth\_rate 5.930000e+00 7 3.743000e+01 3.043000e+01 1.34  
## Empl 7.410000e+00 33 8.000000e+01 4.700000e+01 -0.22  
## Empl\_serv 1.483000e+01 20 8.900000e+01 6.900000e+01 -0.59  
## Empl\_agri 1.038000e+01 0 6.126000e+01 6.126000e+01 1.23  
## Empl\_indu 4.450000e+00 7 3.800000e+01 3.100000e+01 0.22  
## educ\_expen 6.115390e+09 15655251 2.935663e+11 2.935506e+11 3.39  
## Mil\_expen 4.871209e+09 0 2.006891e+11 2.006891e+11 4.89  
## Arm\_forc\_person 7.116480e+04 0 1.489929e+06 1.489929e+06 3.10  
## Region\* 1.480000e+00 1 5.000000e+00 4.000000e+00 -0.31  
## Subregion\* 7.410000e+00 1 1.900000e+01 1.800000e+01 -0.28  
## kurtosis se  
## Territory\* -1.23 3.080000e+00  
## Homicide 29.90 7.795400e+02  
## Population 22.49 8.771633e+06  
## GDP 17.68 1.139030e+11  
## GDP\_per\_capita 1.85 2.128790e+03  
## GNI 17.80 1.152961e+11  
## Labor\_force 22.81 4.184111e+06  
## Unempl 2.07 4.500000e-01  
## Agri\_fore\_fish\_GDP 25.66 4.466679e+09  
## Life\_expect 0.76 5.700000e-01  
## Death\_rate 0.12 2.700000e-01  
## Birth\_rate 1.37 6.800000e-01  
## Empl 0.07 8.300000e-01  
## Empl\_serv -0.13 1.360000e+00  
## Empl\_agri 0.91 1.380000e+00  
## Empl\_indu 0.40 5.600000e-01  
## educ\_expen 13.41 4.279114e+09  
## Mil\_expen 27.40 2.758152e+09  
## Arm\_forc\_person 10.03 2.787053e+04  
## Region\* -1.03 1.000000e-01  
## Subregion\* -1.27 5.300000e-01

#let's work with f\_eg2  
f <- f\_ej2

#для корекції викидів обраний варіант заповнення граничними значеннями.

# Splitting the dataset into the TRAIN set and TEST set

set.seed(123)  
library(caTools)

## Warning: package 'caTools' was built under R version 3.6.3

split = sample.split(f$Territory, SplitRatio = 0.8)  
f\_train = subset(f, split == TRUE)  
f\_test = subset(f, split == FALSE)  
#Write prepared data to the file  
write.csv2(f\_train, file = "crime\_train.csv")  
write.csv2(f\_test, file = "crime\_test.csv")

#датасет розподілений на навчальну та тестову вибірки. Результати збережені в окремих файлах.

Висновок: поставлені завдання за лабораторною роботою виконано, а саме: проведено збір даних, зроблено кодування, обчислено та оцінено статистики, досліджено та видалено викиди, розроблено шкалювання даних, проведено обробку відсутніх значень, проаналізовано та зроблено вибір ознак, проведено перехресну перевірку.