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)

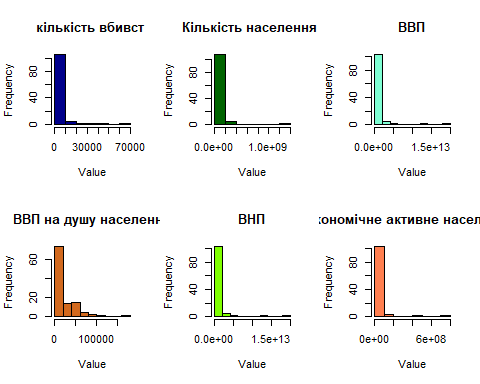
## Territory Homicide Population GDP GDP\_per\_capita GNI  
## 1 Kenya 2466 50221473 78965004656 1572 77456412235  
## 2 Mauritius 33 1264613 13259351418 10485 14572668163  
## 3 Uganda 4735 41162465 30756466548 747 30032761102  
## 4 Cameroon 341 24566045 35009262788 1425 34347942597  
## 5 Morocco 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.

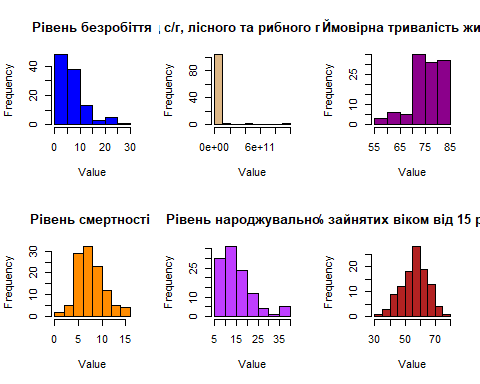
# Visualising

## Histogram

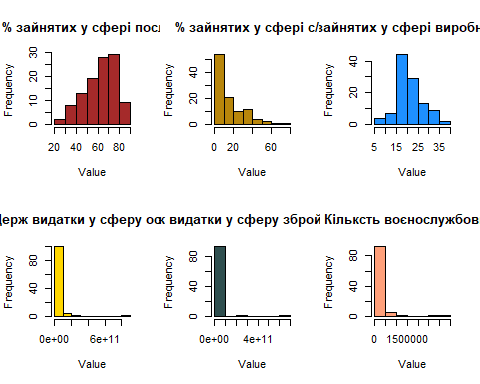
library(ggplot2)  
par(mfrow = c(2, 3))   
hist(f$Homicide, col = 'dark blue', main = 'кількість вбивст ', xlab = 'Value')  
hist(f$Population, col = 'dark green', main = 'Кількість населення ', xlab = 'Value')  
hist(f$GDP, col = 'aquamarine', main = 'ВВП', xlab = 'Value')  
hist(f$GDP\_per\_capita, 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.

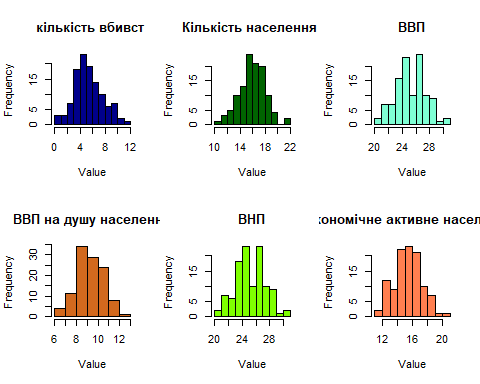
#проведемо логарифмування даних #как логарифмировать, если у значение 0

f\_log <- f  
f\_log$Homicide <- log(f$Homicide, )  
f\_log$Population <- log(f$Population, )  
f\_log$GDP <- log(f$GDP, )  
f\_log$GDP\_per\_capita <- log(f$GDP\_per\_capita, )  
f\_log$GNI <- log(f$GNI, )  
f\_log$Labor\_force <- log(f$Labor\_force, )  
f\_log$Unempl <- log(f$Unempl, )  
f\_log$Agri\_fore\_fish\_GDP<- log(f$Agri\_fore\_fish\_GDP, )  
f\_log$Life\_expect <- log(f$Life\_expect, )  
f\_log$Death\_rate <- log(f$Death\_rate )  
f\_log$Birth\_rate <- log(f$Birth\_rate, )  
f\_log$Empl <- log(f$Empl, )  
f\_log$Empl\_serv <- log(f$Empl\_serv, )  
f\_log$Empl\_agri <- log(f$Empl\_agri, )  
f\_log$Empl\_indu <- log(f$Empl\_indu, )  
f\_log$educ\_expen <- log(f$educ\_expen, )  
f\_log$Mil\_expen <- log(f$Mil\_expen, )  
f\_log$Arm\_forc\_person <- log(f$Arm\_forc\_person, )  
  
head(f\_log)

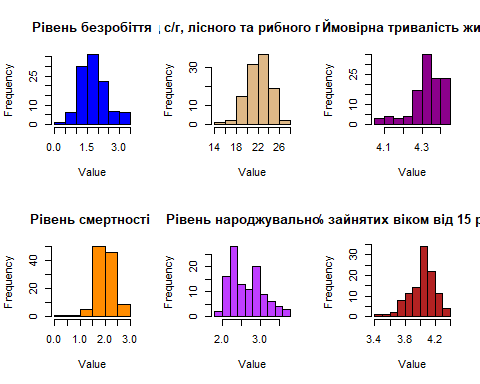
## Territory Homicide Population GDP GDP\_per\_capita GNI Labor\_force  
## 1 Kenya 7.810353 17.73195 25.09227 7.360104 25.07298 16.92462  
## 2 Mauritius 3.496508 14.05028 23.30797 9.257701 23.40241 13.31815  
## 3 Uganda 8.462737 17.53304 24.14937 6.616065 24.12555 16.54243  
## 4 Cameroon 5.831882 17.01688 24.27888 7.261927 24.25981 16.18606  
## 5 Morocco 6.634633 17.38733 25.42375 8.018296 25.40540 16.27877  
## 6 Eswatini 4.867534 13.93307 22.21533 8.282230 22.15485 12.79479  
## Unempl Agri\_fore\_fish\_GDP Life\_expect Death\_rate Birth\_rate Empl  
## 1 1.0986123 24.03761 4.189655 1.791759 3.367296 4.290459  
## 2 1.9459101 19.83284 4.317488 2.079442 2.397895 4.007333  
## 3 0.6931472 22.69944 4.143135 1.945910 3.663562 4.234107  
## 4 1.0986123 22.33901 4.077537 2.302585 3.583519 4.304065  
## 5 2.1972246 23.33036 4.330733 1.609438 2.944439 3.713572  
## 6 3.1354942 19.73158 4.060443 2.302585 3.295837 3.688879  
## Empl\_serv Empl\_agri Empl\_indu educ\_expen Mil\_expen Arm\_forc\_person Region  
## 1 3.610918 4.025352 1.945910 22.06531 20.73853 10.275051 Africa  
## 2 4.219508 1.791759 3.258097 20.35959 16.98616 8.006368 Africa  
## 3 2.995732 4.290459 1.945910 20.21639 19.66415 10.736397 Africa  
## 4 3.688879 3.806662 2.708050 20.64045 19.82716 10.085809 Africa  
## 5 3.737670 3.583519 3.091042 22.44484 21.96496 12.413087 Africa  
## 6 4.143135 2.564949 3.178054 19.46527 18.30053 NA Africa  
## Subregion  
## 1 Eastern Africa  
## 2 Eastern Africa  
## 3 Eastern Africa  
## 4 Middle Africa  
## 5 Northern Africa  
## 6 Southern Africa

f<- f\_log

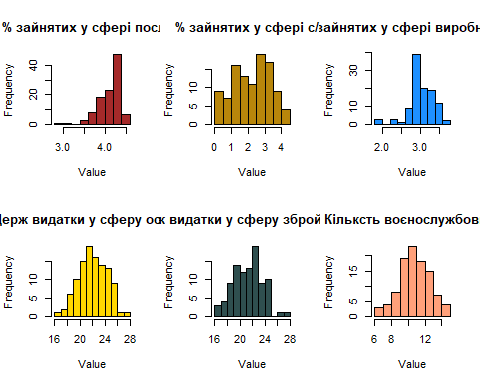
par(mfrow = c(2, 3))   
hist(f$Homicide, col = 'dark blue', main = 'кількість вбивст ', xlab = 'Value')  
hist(f$Population, col = 'dark green', main = 'Кількість населення ', xlab = 'Value')  
hist(f$GDP, col = 'aquamarine', main = 'ВВП', xlab = 'Value')  
hist(f$GDP\_per\_capita, col = 'chocolate', 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')



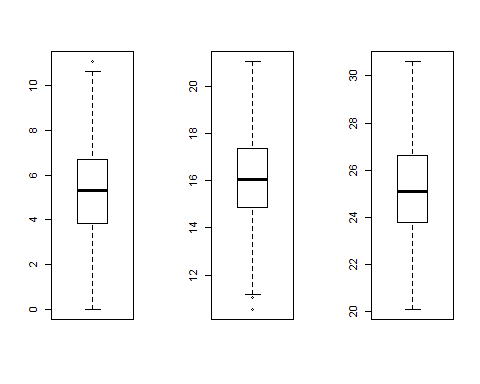
hist(f$Unempl, col = 'blue', main = 'Рівень безробіття', xlab = 'Value')  
hist(f$Agri\_fore\_fish\_GDP, col = 'burlywood', main = 'Дохід від с/г, лісного та рибного господарства', xlab = 'Value')  
hist(f$Life\_expect, col = 'darkmagenta', main = 'Ймовірна тривалість життя', xlab = 'Value')  
hist(f$Death\_rate, 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')



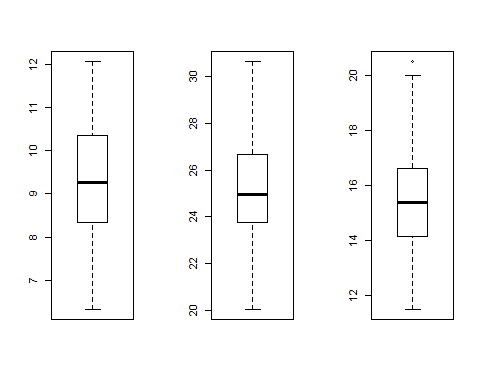
hist(f$Empl\_serv, col = 'brown', main = '% зайнятих у сфері послуг ', xlab = 'Value')  
hist(f$Empl\_agri, col = 'darkgoldenrod', main = '% зайнятих у сфері с/г ', xlab = 'Value')  
hist(f$Empl\_indu, col = 'dodgerblue', main = '% зайнятих у сфері виробництва', xlab = 'Value')  
hist(f$educ\_expen, 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')

 ## Box-plot

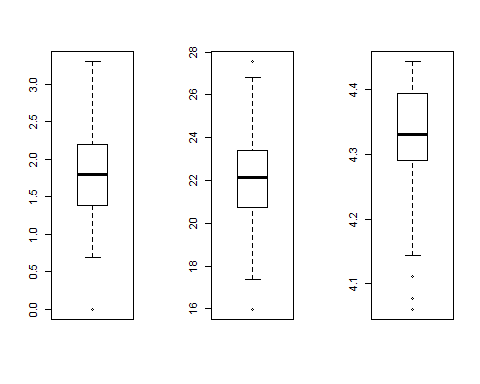
par(mfrow = c(1, 3))  
boxplot(f$Homicide)  
boxplot(f$Population)  
boxplot(f$GDP)



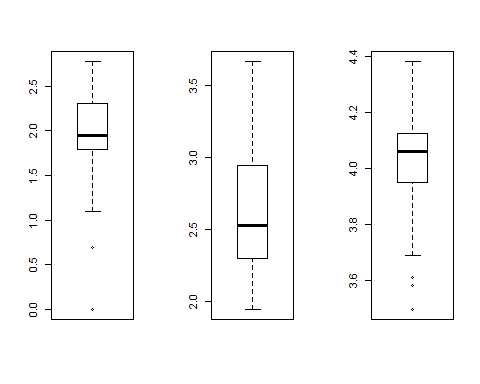
boxplot(f$GDP\_per\_capita)  
boxplot(f$GNI)  
boxplot(f$Labor\_force)



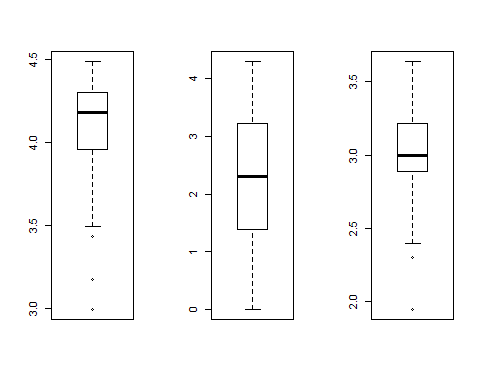
boxplot(f$Unempl)  
boxplot(f$Agri\_fore\_fish\_GDP)  
boxplot(f$Life\_expect)



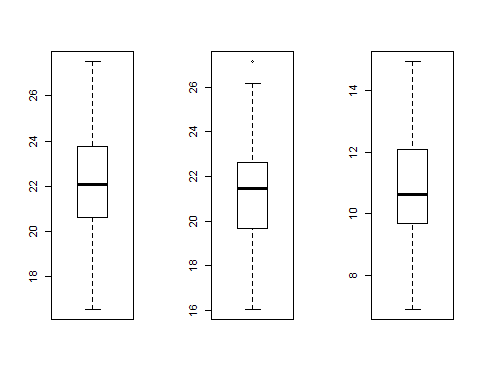
boxplot(f$Death\_rate)  
boxplot(f$Birth\_rate)  
boxplot(f$Empl)



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



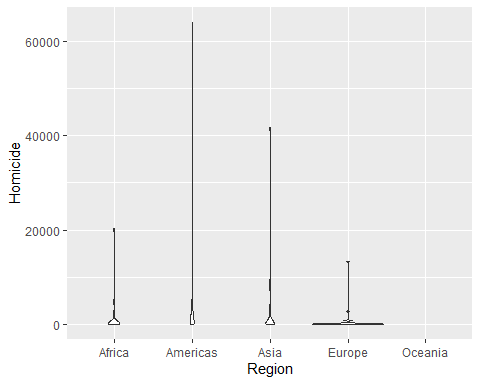
boxplot(f$educ\_expen)  
boxplot(f$Mil\_expen)  
boxplot(f$Arm\_forc\_person)



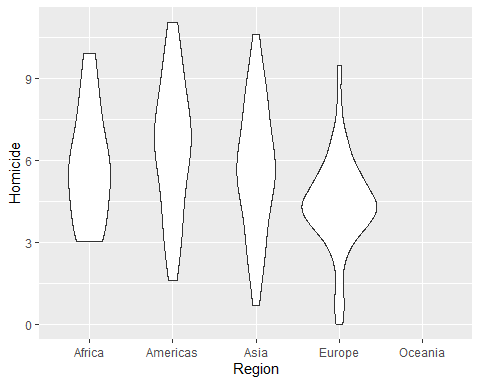
#Висновок: всі змінні мають викиди, окрім GDP, GDP\_per\_capita, GNI, Birth\_rate, educ\_expen, Arm\_forc\_person

## Violin

qplot(data = read.csv2('crime.csv', header = TRUE, encoding = 'UNICOD'),   
 x = Region,   
 y = Homicide,   
 geom = "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)

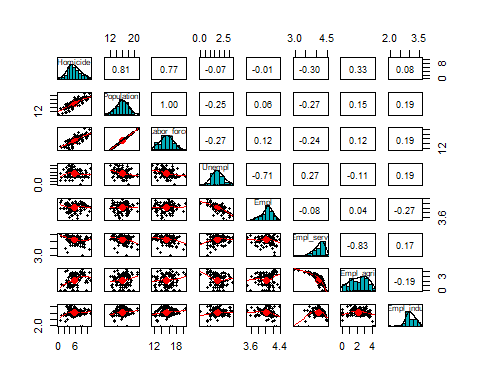
## vars n mean sd median trimmed mad min max  
## Territory\* 1 113 57.00 32.76 57.00 57.00 41.51 1.00 113.00  
## Homicide 2 113 5.45 2.27 5.30 5.39 2.19 0.00 11.07  
## Population 3 113 15.94 2.02 16.06 15.98 1.83 10.54 21.05  
## GDP 4 111 25.23 2.15 25.09 25.24 2.21 20.07 30.60  
## GDP\_per\_capita 5 111 9.26 1.26 9.26 9.30 1.42 6.32 12.06  
## GNI 6 111 25.20 2.16 24.97 25.21 2.35 20.05 30.62  
## Labor\_force 7 108 15.37 1.81 15.38 15.35 1.82 11.50 20.48  
## Unempl 8 108 1.80 0.62 1.79 1.79 0.60 0.00 3.30  
## Agri\_fore\_fish\_GDP 9 108 22.03 2.05 22.14 22.08 1.99 15.97 27.55  
## Life\_expect 10 112 4.33 0.08 4.33 4.34 0.08 4.06 4.44  
## Death\_rate 11 112 2.01 0.41 1.95 2.03 0.37 0.00 2.77  
## Birth\_rate 12 112 2.65 0.42 2.52 2.62 0.49 1.95 3.66  
## Empl 13 108 4.03 0.17 4.06 4.04 0.13 3.50 4.38  
## Empl\_serv 14 108 4.11 0.28 4.18 4.14 0.23 3.00 4.49  
## Empl\_agri 15 105 2.20 1.17 2.30 2.25 1.36 0.00 4.29  
## Empl\_indu 16 108 3.03 0.31 3.00 3.05 0.24 1.95 3.64  
## educ\_expen 17 107 22.06 2.20 22.07 22.09 2.40 16.57 27.49  
## Mil\_expen 18 95 21.28 2.23 21.45 21.27 2.31 16.05 27.13  
## Arm\_forc\_person 19 101 10.75 1.79 10.65 10.77 1.87 6.91 14.92  
## Region\* 20 113 2.93 1.06 3.00 3.01 1.48 1.00 5.00  
## Subregion\* 21 113 11.18 5.62 12.00 11.38 7.41 1.00 19.00  
## range skew kurtosis se  
## Territory\* 112.00 0.00 -1.23 3.08  
## Homicide 11.07 0.23 -0.24 0.21  
## Population 10.51 -0.20 0.10 0.19  
## GDP 10.53 -0.02 -0.44 0.20  
## GDP\_per\_capita 5.74 -0.13 -0.69 0.12  
## GNI 10.57 -0.01 -0.44 0.20  
## Labor\_force 8.99 0.13 -0.17 0.17  
## Unempl 3.30 0.11 -0.07 0.06  
## Agri\_fore\_fish\_GDP 11.58 -0.18 0.17 0.20  
## Life\_expect 0.38 -1.19 1.36 0.01  
## Death\_rate 2.77 -1.26 4.56 0.04  
## Birth\_rate 1.72 0.54 -0.53 0.04  
## Empl 0.89 -0.69 0.41 0.02  
## Empl\_serv 1.49 -1.34 2.17 0.03  
## Empl\_agri 4.29 -0.28 -0.96 0.11  
## Empl\_indu 1.69 -1.05 2.36 0.03  
## educ\_expen 10.93 -0.09 -0.45 0.21  
## Mil\_expen 11.08 0.05 -0.44 0.23  
## Arm\_forc\_person 8.02 -0.01 -0.42 0.18  
## Region\* 4.00 -0.31 -1.03 0.10  
## Subregion\* 18.00 -0.28 -1.27 0.53

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

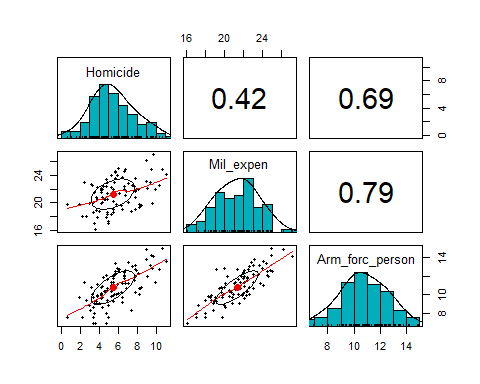
## Correlations

#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

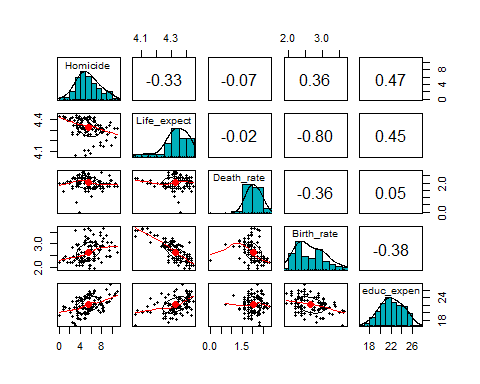
f1 <- c(2,3,7,8,13,14,15,16)  
c\_f1 <- f[f1]  
pairs.panels(c\_f1, lm=FALSE, # linear fit  
 method = "pearson", # correlation method  
 hist.col = "#00AFBB"  
 )



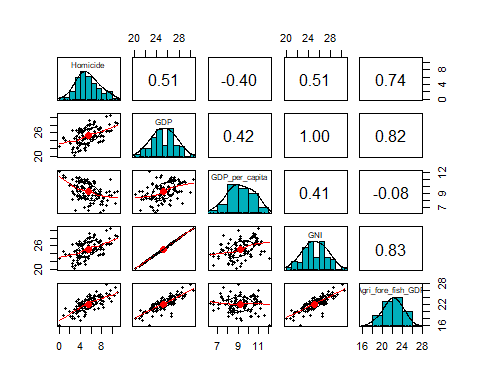
f2 <- c(2,18,19)  
c\_f2 <- f[f2]  
pairs.panels(c\_f2, lm=FALSE, # linear fit  
 method = "pearson", # correlation method  
 hist.col = "#00AFBB"  
 )



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



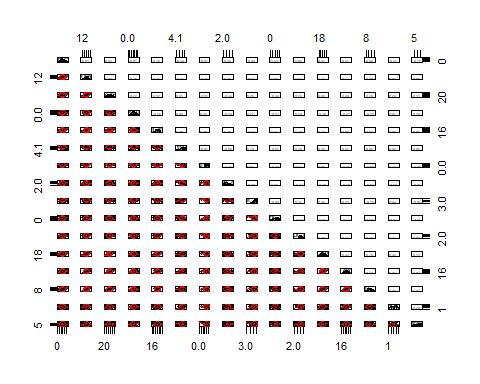
#заміна lm=TRUE на lm=FALSE

#Змінні, що мають незначний вплив на Homicide (з рівнем менш за 0,05) та ті, що корелюють між собою рекомендовано не вкдючати до подальшого аналізу (GDP\_per\_capita, GNI, Labor\_force, Empl)

f <- f[ ,-c(5,6,7,13)]  
head(f)

## Territory Homicide Population GDP Unempl Agri\_fore\_fish\_GDP  
## 1 Kenya 7.810353 17.73195 25.09227 1.0986123 24.03761  
## 2 Mauritius 3.496508 14.05028 23.30797 1.9459101 19.83284  
## 3 Uganda 8.462737 17.53304 24.14937 0.6931472 22.69944  
## 4 Cameroon 5.831882 17.01688 24.27888 1.0986123 22.33901  
## 5 Morocco 6.634633 17.38733 25.42375 2.1972246 23.33036  
## 6 Eswatini 4.867534 13.93307 22.21533 3.1354942 19.73158  
## Life\_expect Death\_rate Birth\_rate Empl\_serv Empl\_agri Empl\_indu educ\_expen  
## 1 4.189655 1.791759 3.367296 3.610918 4.025352 1.945910 22.06531  
## 2 4.317488 2.079442 2.397895 4.219508 1.791759 3.258097 20.35959  
## 3 4.143135 1.945910 3.663562 2.995732 4.290459 1.945910 20.21639  
## 4 4.077537 2.302585 3.583519 3.688879 3.806662 2.708050 20.64045  
## 5 4.330733 1.609438 2.944439 3.737670 3.583519 3.091042 22.44484  
## 6 4.060443 2.302585 3.295837 4.143135 2.564949 3.178054 19.46527  
## Mil\_expen Arm\_forc\_person Region Subregion  
## 1 20.73853 10.275051 Africa Eastern Africa  
## 2 16.98616 8.006368 Africa Eastern Africa  
## 3 19.66415 10.736397 Africa Eastern Africa  
## 4 19.82716 10.085809 Africa Middle Africa  
## 5 21.96496 12.413087 Africa Northern Africa  
## 6 18.30053 NA Africa Southern Africa

f5 <- c( 2:17)  
c\_f5 <- f[f5]  
pairs.panels(c\_f5, lm=FALSE, # linear fit  
 method = "pearson", # correlation method  
 hist.col = "#00AFBB"  
 )



#Тісний зв’язок виявлено між змінними Кількість населення та Дохід від с/г, лісного та рибного господарства; Кількість населення та кількості воєнослужбовців; Державні видатки у сферу освіти та ВВП. Рекомендовано, в моделі залишити усі змінні.

# 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 88 rows in the f\_reduce

#Після видалення пропущених значень в базі залишилося би 88 рядок.

## Fill n/a

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

## Territory Homicide Population GDP Unempl Agri\_fore\_fish\_GDP  
## 1 Kenya 7.810353 17.73195 25.09227 1.0986123 24.03761  
## 2 Mauritius 3.496508 14.05028 23.30797 1.9459101 19.83284  
## 3 Uganda 8.462737 17.53304 24.14937 0.6931472 22.69944  
## 4 Cameroon 5.831882 17.01688 24.27888 1.0986123 22.33901  
## 5 Morocco 6.634633 17.38733 25.42375 2.1972246 23.33036  
## 6 Eswatini 4.867534 13.93307 22.21533 3.1354942 19.73158  
## Life\_expect Death\_rate Birth\_rate Empl\_serv Empl\_agri Empl\_indu educ\_expen  
## 1 4.189655 1.791759 3.367296 3.610918 4.025352 1.945910 22.06531  
## 2 4.317488 2.079442 2.397895 4.219508 1.791759 3.258097 20.35959  
## 3 4.143135 1.945910 3.663562 2.995732 4.290459 1.945910 20.21639  
## 4 4.077537 2.302585 3.583519 3.688879 3.806662 2.708050 20.64045  
## 5 4.330733 1.609438 2.944439 3.737670 3.583519 3.091042 22.44484  
## 6 4.060443 2.302585 3.295837 4.143135 2.564949 3.178054 19.46527  
## Mil\_expen Arm\_forc\_person Region Subregion  
## 1 20.73853 10.275051 Africa Eastern Africa  
## 2 16.98616 8.006368 Africa Eastern Africa  
## 3 19.66415 10.736397 Africa Eastern Africa  
## 4 19.82716 10.085809 Africa Middle Africa  
## 5 21.96496 12.413087 Africa Northern Africa  
## 6 18.30053 12.413087 Africa 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$Unempl <- ifelse(is.na(f$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$Empl\_serv <- ifelse(is.na(f$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 Unempl Agri\_fore\_fish\_GDP  
## 1 Kenya 7.810353 17.73195 25.09227 1.0986123 24.03761  
## 2 Mauritius 3.496508 14.05028 23.30797 1.9459101 19.83284  
## 3 Uganda 8.462737 17.53304 24.14937 0.6931472 22.69944  
## 4 Cameroon 5.831882 17.01688 24.27888 1.0986123 22.33901  
## 5 Morocco 6.634633 17.38733 25.42375 2.1972246 23.33036  
## 6 Eswatini 4.867534 13.93307 22.21533 3.1354942 19.73158  
## Life\_expect Death\_rate Birth\_rate Empl\_serv Empl\_agri Empl\_indu educ\_expen  
## 1 4.189655 1.791759 3.367296 3.610918 4.025352 1.945910 22.06531  
## 2 4.317488 2.079442 2.397895 4.219508 1.791759 3.258097 20.35959  
## 3 4.143135 1.945910 3.663562 2.995732 4.290459 1.945910 20.21639  
## 4 4.077537 2.302585 3.583519 3.688879 3.806662 2.708050 20.64045  
## 5 4.330733 1.609438 2.944439 3.737670 3.583519 3.091042 22.44484  
## 6 4.060443 2.302585 3.295837 4.143135 2.564949 3.178054 19.46527  
## Mil\_expen Arm\_forc\_person Region Subregion  
## 1 20.73853 10.275051 Africa Eastern Africa  
## 2 16.98616 8.006368 Africa Eastern Africa  
## 3 19.66415 10.736397 Africa Eastern Africa  
## 4 19.82716 10.085809 Africa Middle Africa  
## 5 21.96496 12.413087 Africa Northern Africa  
## 6 18.30053 11.000000 Africa 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$Homicide < mean(f$Homicide)+sd(f$Homicide)\*3, ]  
f\_ej1 <- f[f$Population < mean(f$Population)+sd(f$Population)\*3, ]  
f\_ej1 <- f[f$GDP < mean(f$GDP)+sd(f$GDP)\*3, ]  
f\_ej1 <- f[f$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$Empl\_serv < mean(f$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 mad min max  
## Territory\* 1 113 57.00 32.76 57.00 57.00 41.51 1.00 113.00  
## Homicide 2 113 5.45 2.27 5.30 5.39 2.19 0.00 11.07  
## Population 3 113 15.94 2.02 16.06 15.98 1.83 10.54 21.05  
## GDP 4 113 25.22 2.13 25.00 25.24 2.31 20.07 30.60  
## Unempl 5 113 1.81 0.61 1.79 1.80 0.60 0.00 3.30  
## Agri\_fore\_fish\_GDP 6 113 22.03 2.00 22.09 22.08 1.91 15.97 27.55  
## Life\_expect 7 113 4.32 0.09 4.33 4.34 0.08 4.00 4.44  
## Death\_rate 8 113 2.01 0.41 1.95 2.03 0.37 0.00 2.77  
## Birth\_rate 9 113 2.66 0.42 2.56 2.63 0.48 1.95 3.66  
## Empl\_serv 10 113 4.10 0.27 4.16 4.14 0.24 3.00 4.49  
## Empl\_agri 11 113 2.19 1.13 2.08 2.23 1.28 0.00 4.29  
## Empl\_indu 12 113 3.03 0.31 3.00 3.05 0.25 1.95 3.64  
## educ\_expen 13 113 22.05 2.14 22.00 22.09 2.21 16.57 27.49  
## Mil\_expen 14 113 21.23 2.05 21.00 21.21 1.86 16.05 27.13  
## Arm\_forc\_person 15 113 10.78 1.69 10.97 10.80 1.51 6.91 14.92  
## Region\* 16 113 2.93 1.06 3.00 3.01 1.48 1.00 5.00  
## Subregion\* 17 113 11.18 5.62 12.00 11.38 7.41 1.00 19.00  
## range skew kurtosis se  
## Territory\* 112.00 0.00 -1.23 3.08  
## Homicide 11.07 0.23 -0.24 0.21  
## Population 10.51 -0.20 0.10 0.19  
## GDP 10.53 -0.01 -0.39 0.20  
## Unempl 3.30 0.07 0.03 0.06  
## Agri\_fore\_fish\_GDP 11.58 -0.18 0.32 0.19  
## Life\_expect 0.44 -1.32 1.82 0.01  
## Death\_rate 2.77 -1.27 4.63 0.04  
## Birth\_rate 1.72 0.52 -0.55 0.04  
## Empl\_serv 1.49 -1.31 2.25 0.03  
## Empl\_agri 4.29 -0.25 -0.82 0.11  
## Empl\_indu 1.69 -1.07 2.59 0.03  
## educ\_expen 10.93 -0.08 -0.31 0.20  
## Mil\_expen 11.08 0.11 0.04 0.19  
## Arm\_forc\_person 8.02 -0.06 -0.12 0.16  
## Region\* 4.00 -0.31 -1.03 0.10  
## Subregion\* 18.00 -0.28 -1.27 0.53

#Висновок: після видалення викидів в базі залишилося би 113 рядків, тобто модель не має максимальніх викидів (можна не виконувати пункт Replace with max )

## Replace with max

f\_ej2 <- f  
f\_ej2$Homicide <- ifelse(f$Homicide < mean(f$Homicide)+sd(f$Homicide)\*3,f$Homicide,mean(f$Homicide)+sd(f$Homicide)\*3)  
f\_ej2$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$GDP)\*3)  
f\_ej2$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$Empl\_serv <- ifelse(f$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 mad min max  
## Territory\* 1 113 57.00 32.76 57.00 57.00 41.51 1.00 113.00  
## Homicide 2 113 5.45 2.27 5.30 5.39 2.19 0.00 11.07  
## Population 3 113 15.94 2.02 16.06 15.98 1.83 10.54 21.05  
## GDP 4 113 25.22 2.13 25.00 25.24 2.31 20.07 30.60  
## Unempl 5 113 1.81 0.61 1.79 1.80 0.60 0.00 3.30  
## Agri\_fore\_fish\_GDP 6 113 22.03 2.00 22.09 22.08 1.91 15.97 27.55  
## Life\_expect 7 113 4.32 0.09 4.33 4.34 0.08 4.00 4.44  
## Death\_rate 8 113 2.01 0.41 1.95 2.03 0.37 0.00 2.77  
## Birth\_rate 9 113 2.66 0.42 2.56 2.63 0.48 1.95 3.66  
## Empl\_serv 10 113 4.10 0.27 4.16 4.14 0.24 3.00 4.49  
## Empl\_agri 11 113 2.19 1.13 2.08 2.23 1.28 0.00 4.29  
## Empl\_indu 12 113 3.03 0.31 3.00 3.05 0.25 1.95 3.64  
## educ\_expen 13 113 22.05 2.14 22.00 22.09 2.21 16.57 27.49  
## Mil\_expen 14 113 21.23 2.05 21.00 21.21 1.86 16.05 27.13  
## Arm\_forc\_person 15 113 10.78 1.69 10.97 10.80 1.51 6.91 14.92  
## Region\* 16 113 2.93 1.06 3.00 3.01 1.48 1.00 5.00  
## Subregion\* 17 113 11.18 5.62 12.00 11.38 7.41 1.00 19.00  
## range skew kurtosis se  
## Territory\* 112.00 0.00 -1.23 3.08  
## Homicide 11.07 0.23 -0.24 0.21  
## Population 10.51 -0.20 0.10 0.19  
## GDP 10.53 -0.01 -0.39 0.20  
## Unempl 3.30 0.07 0.03 0.06  
## Agri\_fore\_fish\_GDP 11.58 -0.18 0.32 0.19  
## Life\_expect 0.44 -1.32 1.82 0.01  
## Death\_rate 2.77 -1.27 4.63 0.04  
## Birth\_rate 1.72 0.52 -0.55 0.04  
## Empl\_serv 1.49 -1.31 2.25 0.03  
## Empl\_agri 4.29 -0.25 -0.82 0.11  
## Empl\_indu 1.69 -1.07 2.59 0.03  
## educ\_expen 10.93 -0.08 -0.31 0.20  
## Mil\_expen 11.08 0.11 0.04 0.19  
## Arm\_forc\_person 8.02 -0.06 -0.12 0.16  
## Region\* 4.00 -0.31 -1.03 0.10  
## Subregion\* 18.00 -0.28 -1.27 0.53

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

#в моделі наявні мінімільні викиди. Було запропоновано однією обробкою замінити і максимальні (якщо б вони були), і мінімальні викиди.

qn = quantile(f$Population, c(0.05, 0.95), na.rm = TRUE)  
f = within(f, { Population = ifelse(Population < qn[1], qn[1], Population)  
 Population = ifelse(Population > qn[2], qn[2], Population)})  
qn = quantile(f$Unempl, c(0.05, 0.95), na.rm = TRUE)  
f = within(f, { Unempl= ifelse(Unempl < qn[1], qn[1], Unempl)  
 Unempl = ifelse(Unempl > qn[2], qn[2], Unempl)})  
qn = quantile(f$Agri\_fore\_fish\_GDP, c(0.05, 0.95), na.rm = TRUE)  
f = within(f, { Agri\_fore\_fish\_GDP= ifelse(Agri\_fore\_fish\_GDP < qn[1], qn[1], Agri\_fore\_fish\_GDP)  
 Agri\_fore\_fish\_GDP = ifelse(Agri\_fore\_fish\_GDP > qn[2], qn[2], Agri\_fore\_fish\_GDP)})  
qn = quantile(f$Life\_expect, c(0.05, 0.95), na.rm = TRUE)  
f = within(f, { Life\_expect= ifelse(Life\_expect < qn[1], qn[1], Life\_expect)  
 Life\_expect = ifelse(Life\_expect > qn[2], qn[2], Life\_expect)})  
qn = quantile(f$Death\_rate, c(0.05, 0.95), na.rm = TRUE)  
f = within(f, { Death\_rate= ifelse(Death\_rate < qn[1], qn[1], Death\_rate)  
 Death\_rate = ifelse(Death\_rate > qn[2], qn[2], Death\_rate)})  
qn = quantile(f$Empl\_serv, c(0.05, 0.95), na.rm = TRUE)  
f = within(f, { Empl\_serv= ifelse(Empl\_serv < qn[1], qn[1], Empl\_serv)  
 Empl\_serv = ifelse(Empl\_serv > qn[2], qn[2], Empl\_serv)})  
qn = quantile(f$Empl\_indu, c(0.05, 0.95), na.rm = TRUE)  
f = within(f, { Empl\_indu= ifelse(Empl\_indu < qn[1], qn[1], Empl\_indu)  
 Empl\_indu = ifelse(Empl\_indu > qn[2], qn[2], Empl\_indu)})  
qn = quantile(f$Mil\_expen, c(0.05, 0.95), na.rm = TRUE)  
f = within(f, { Mil\_expen= ifelse(Mil\_expen < qn[1], qn[1], Mil\_expen)  
 Mil\_expen = ifelse(Mil\_expen > qn[2], qn[2], Mil\_expen)})  
  
describe(f)

## vars n mean sd median trimmed mad min max  
## Territory\* 1 113 57.00 32.76 57.00 57.00 41.51 1.00 113.00  
## Homicide 2 113 5.45 2.27 5.30 5.39 2.19 0.00 11.07  
## Population 3 113 15.95 1.79 16.06 15.98 1.83 12.67 18.99  
## GDP 4 113 25.22 2.13 25.00 25.24 2.31 20.07 30.60  
## Unempl 5 113 1.81 0.58 1.79 1.80 0.60 0.69 2.95  
## Agri\_fore\_fish\_GDP 6 113 22.01 1.79 22.09 22.08 1.91 18.43 24.87  
## Life\_expect 7 113 4.33 0.08 4.33 4.34 0.08 4.13 4.42  
## Death\_rate 8 113 2.03 0.33 1.95 2.03 0.37 1.39 2.59  
## Birth\_rate 9 113 2.66 0.42 2.56 2.63 0.48 1.95 3.66  
## Empl\_serv 10 113 4.11 0.23 4.16 4.14 0.24 3.63 4.41  
## Empl\_agri 11 113 2.19 1.13 2.08 2.23 1.28 0.00 4.29  
## Empl\_indu 12 113 3.04 0.26 3.00 3.05 0.25 2.45 3.48  
## educ\_expen 13 113 22.05 2.14 22.00 22.09 2.21 16.57 27.49  
## Mil\_expen 14 113 21.23 1.87 21.00 21.21 1.86 17.95 24.67  
## Arm\_forc\_person 15 113 10.78 1.69 10.97 10.80 1.51 6.91 14.92  
## Region\* 16 113 2.93 1.06 3.00 3.01 1.48 1.00 5.00  
## Subregion\* 17 113 11.18 5.62 12.00 11.38 7.41 1.00 19.00  
## range skew kurtosis se  
## Territory\* 112.00 0.00 -1.23 3.08  
## Homicide 11.07 0.23 -0.24 0.21  
## Population 6.32 -0.16 -0.83 0.17  
## GDP 10.53 -0.01 -0.39 0.20  
## Unempl 2.26 0.10 -0.53 0.05  
## Agri\_fore\_fish\_GDP 6.44 -0.27 -0.67 0.17  
## Life\_expect 0.29 -0.94 0.33 0.01  
## Death\_rate 1.21 -0.09 -0.81 0.03  
## Birth\_rate 1.72 0.52 -0.55 0.04  
## Empl\_serv 0.78 -0.65 -0.62 0.02  
## Empl\_agri 4.29 -0.25 -0.82 0.11  
## Empl\_indu 1.03 -0.30 -0.09 0.02  
## educ\_expen 10.93 -0.08 -0.31 0.20  
## Mil\_expen 6.72 0.07 -0.70 0.18  
## Arm\_forc\_person 8.02 -0.06 -0.12 0.16  
## Region\* 4.00 -0.31 -1.03 0.10  
## Subregion\* 18.00 -0.28 -1.27 0.53

#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")

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.