Assigment 1 - Descriptive Statistics

Vitoria Barbosa Ferreira

10 4 2021

Introduction

This assignment aims to dig into descriptive statics and descriptive spatial statics. The data set used in this exercise is the USAArrests, which contains statiscs about arrests, assault, murder and rape in 50 US states in 1973. All crime variables represents the value per 100.000 habitants.

1) Histogram and Density Plotsa

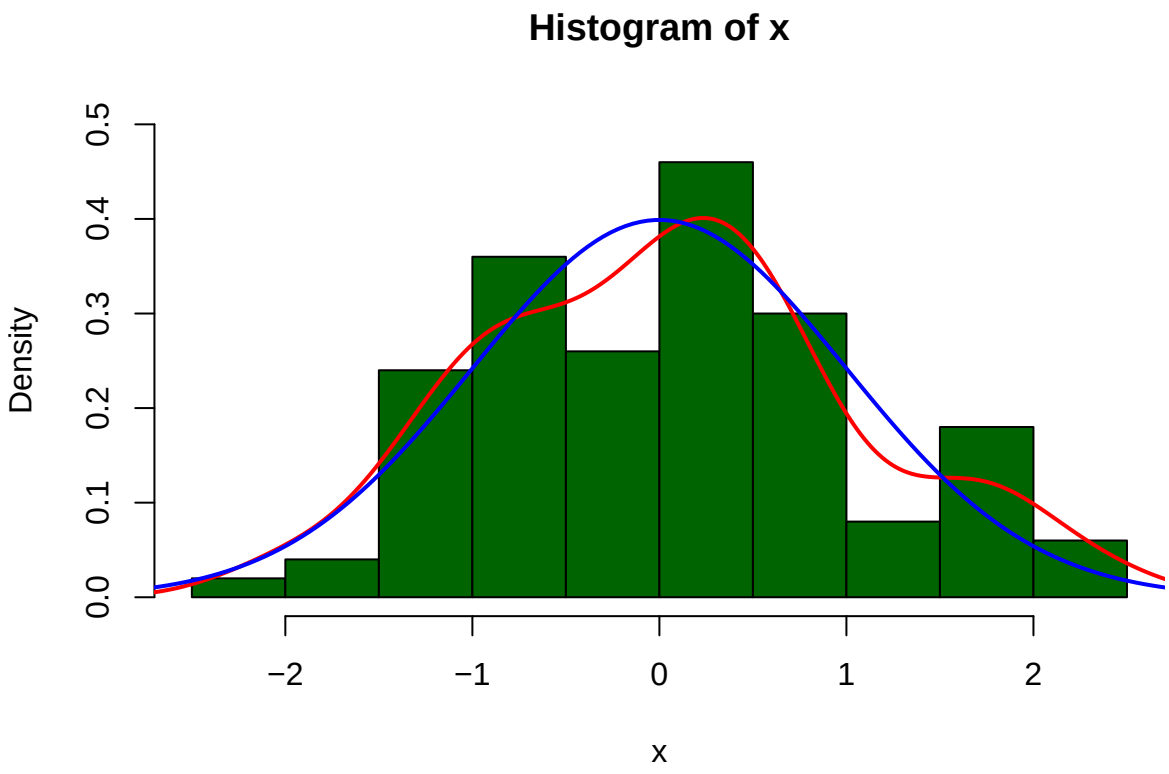
The plot below illustrates a histogram of the murders in the USA.

- rnorm () is a distribution function
- seq () is function to generate regular sequences
- hist() is a function to plot a histogram, based on given data values.

Arguments of hist

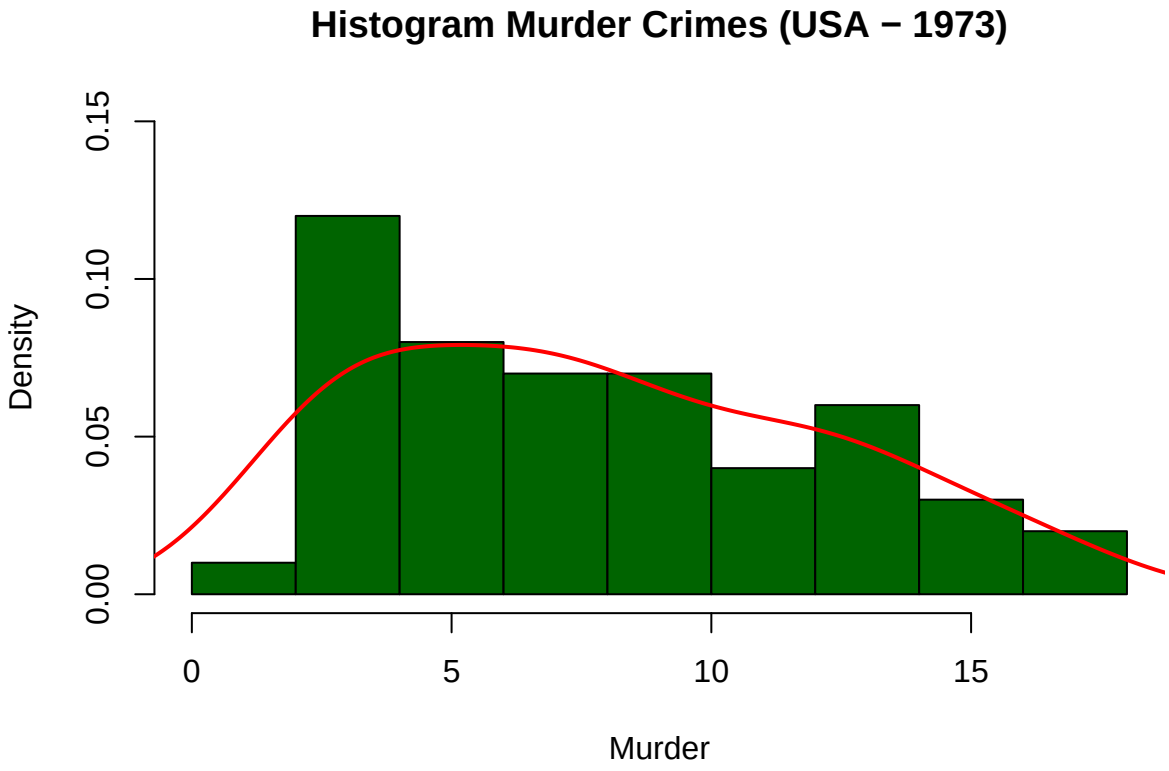
- freq - if set to False it shows probabilities densities, otherwise it shows the count of each frequency
- ylim - defines the range of y-axis

```
x <- rnorm(100)
y <- seq(-4,4,length.out=200)

hist(x,freq=F,ylim=c(0,0.5), col="darkgreen")

lines(density(x),col="red",lwd=2)
lines(y,dnorm(y),col="blue",lwd=2)
```



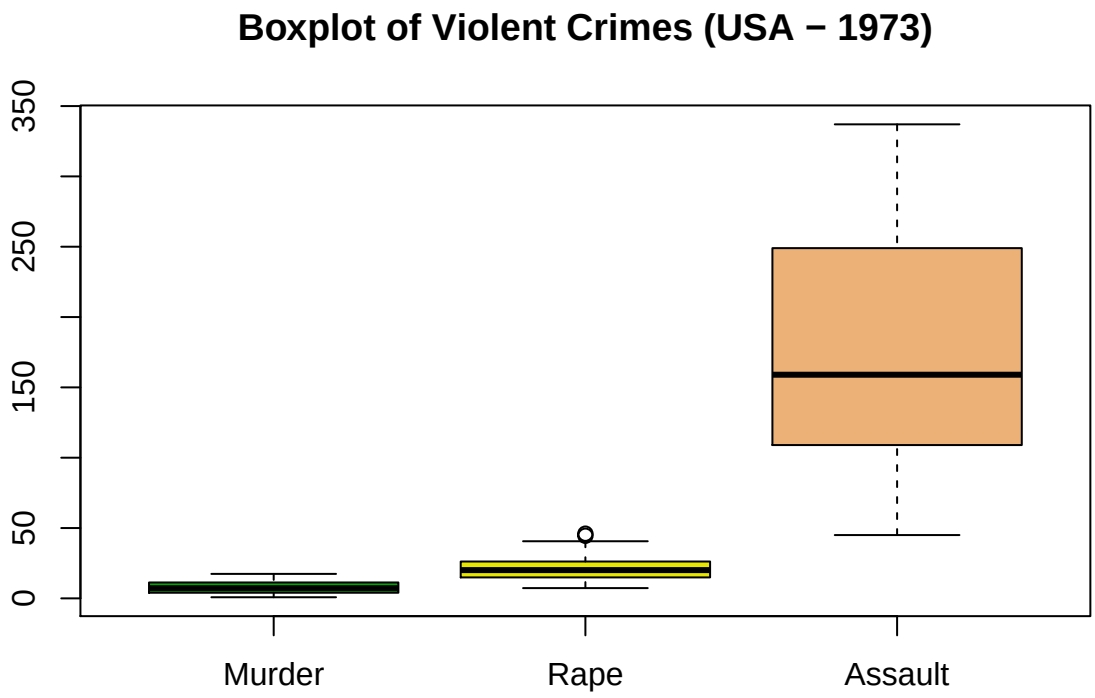
The breaks argument in the following plot informs the size of the break between the bars.

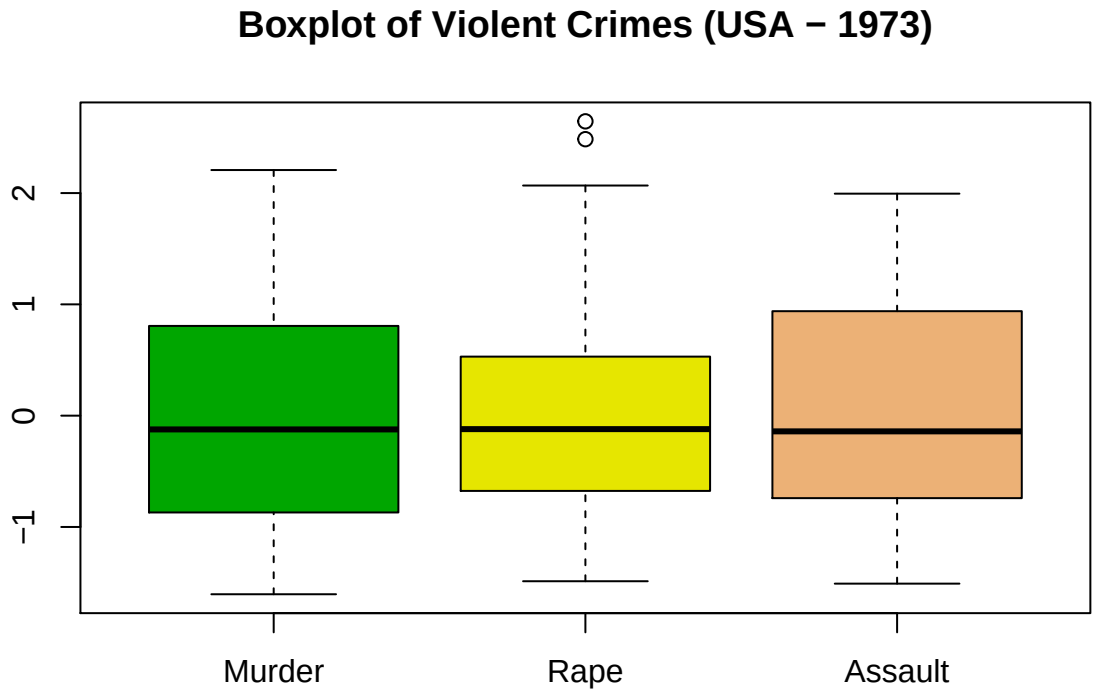
```
data("USArrests")

hist(USArrests$Murder,
     freq=F,
     ylim=c(0,0.15),
     breaks = 6,
     main = "Histogram Murder Crimes (USA - 1973)",
     xlab = "Murder",
     col = "darkgreen")

lines(density(USArrests$Murder),
      col="red",
      lwd=2)
```



2) Exploring Plots

a) Boxplots In descriptive statistics, the boxplot is a very common way of representing the distribution of variables in a data set. It is very useful to compare the distribution among distinct variables, which may lead to further questions/conclusions. In this regard, the first plot illustrates the distribution of 3 variables: Murder, Rape and Assault. As they have a very different scale, it is hard to actually understand the distribution of murder and rape. One way to overcome this situation is to scale the boxplot. The last figure shows the boxplot of all variables scaled from -2 to 2. As a result, now it is possible to have a better visualization of the results, but it looses the association with the real numbers.


Boxplot of Violent Crimes (USA − 1973)
0
50
150
250
350
Murder
Rape
Assault



b) Stemplots Stemplot is another way to graphically visualize the distribution of the data. The following tables show the distribution of the data for each stem (left values). Taking murder as an example, it shows that the decimal point falls at the |, in this regard the first line represents that there is one value starting with 0. followed by 8, so: 0.8. In the next line it shows that there are several values starting with 2., like 2.1, 2.2 and so on. With the stemplot we can quickly see that the scale of the variables diverge among them.

```
#stemplot

print("Murder",stem(USArrests$Murder))
```

```
##
##   The decimal point is at the |
##
##    0 | 8
##    2 | 11226672348
##    4 | 0349379
##    6 | 003682349
##    8 | 158007
##   10 | 04134
##   12 | 127022
##   14 | 444
##   16 | 14
##
## [1] "Murder"
```

```
print("Assault",stem(USArrests$Assault))
```

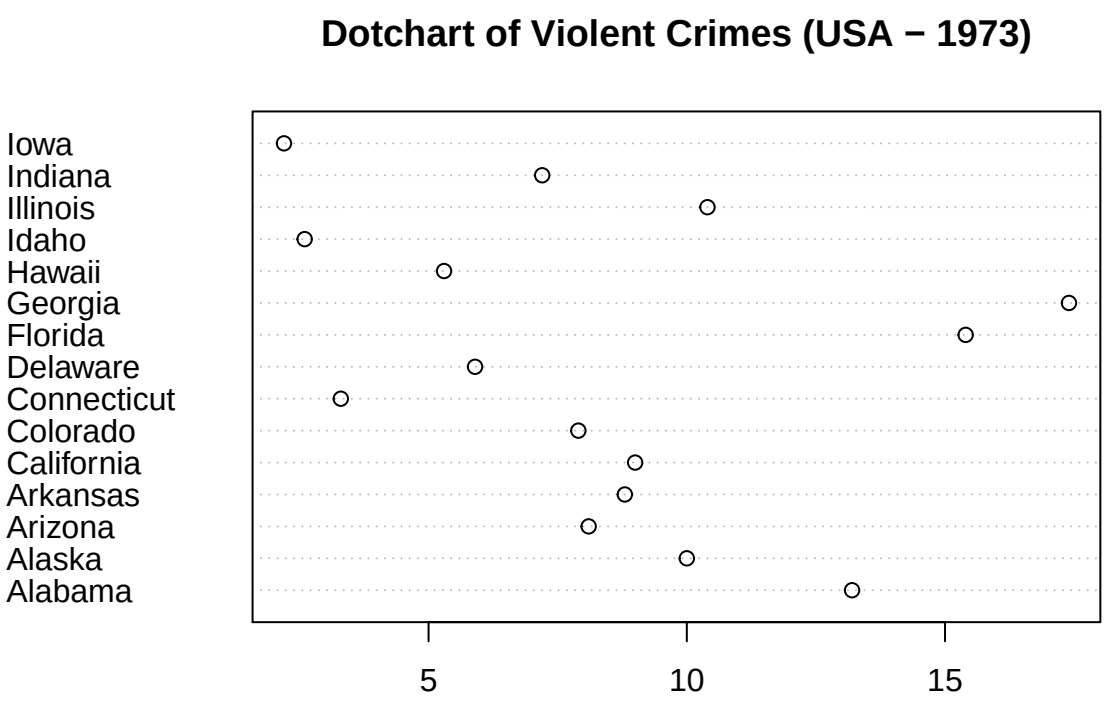
```
##
##   The decimal point is 2 digit(s) to the right of the |
##
##   0 |
##   0 | 5555667889
##   1 | 0111112222
##   1 | 55566667899
##   2 | 00144
##   2 | 55556668899
##   3 | 044
##
## [1] "Assault"
```

```
print("Rape",stem(USArrests$Rape))
```

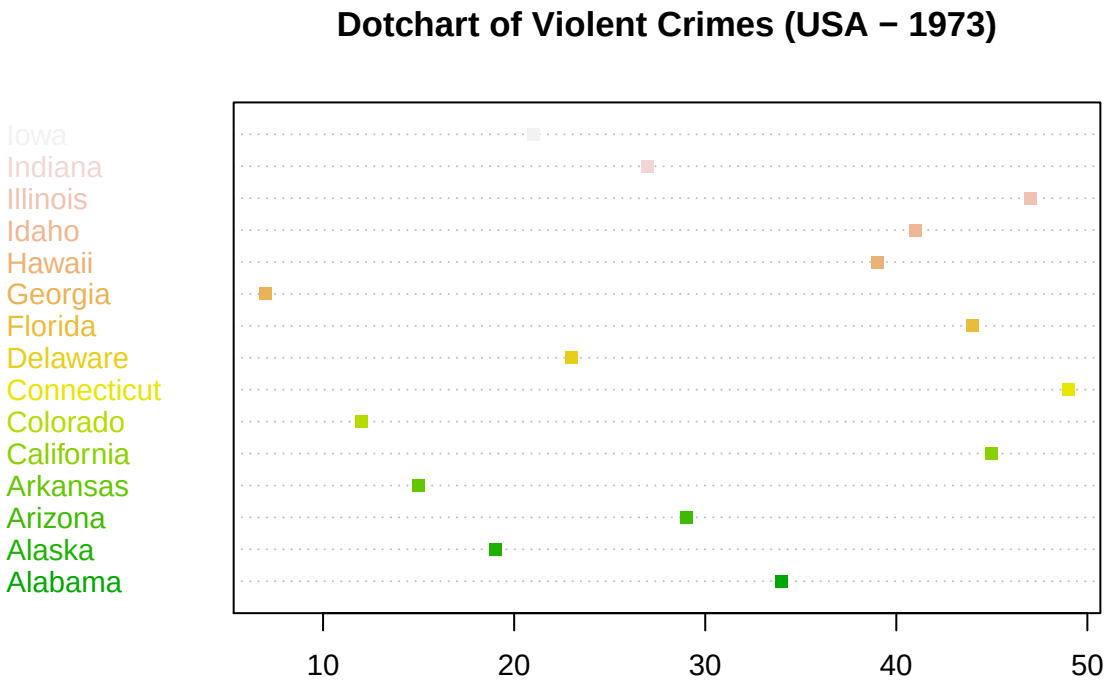
```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   0 | 7889
##   1 | 0111134
##   1 | 556666667789
##   2 | 00011112334
##   2 | 66667889
##   3 | 122
##   3 | 59
##   4 | 1
##   4 | 56
##
## [1] "Rape"
```

c) Dotcharts The next plot is a dotchart also representing the distribution of the data. In this example we are using the variable murder per US state. Only the first 15 observations are shown so visibility is not compromised.

```
#dotchart on variable in R
first_obs <- head(USArrests$Murder,15)

dotchart(first_obs,
         labels = row.names(USArrests),
         main="Dotchart of Violent Crimes (USA - 1973)")
```

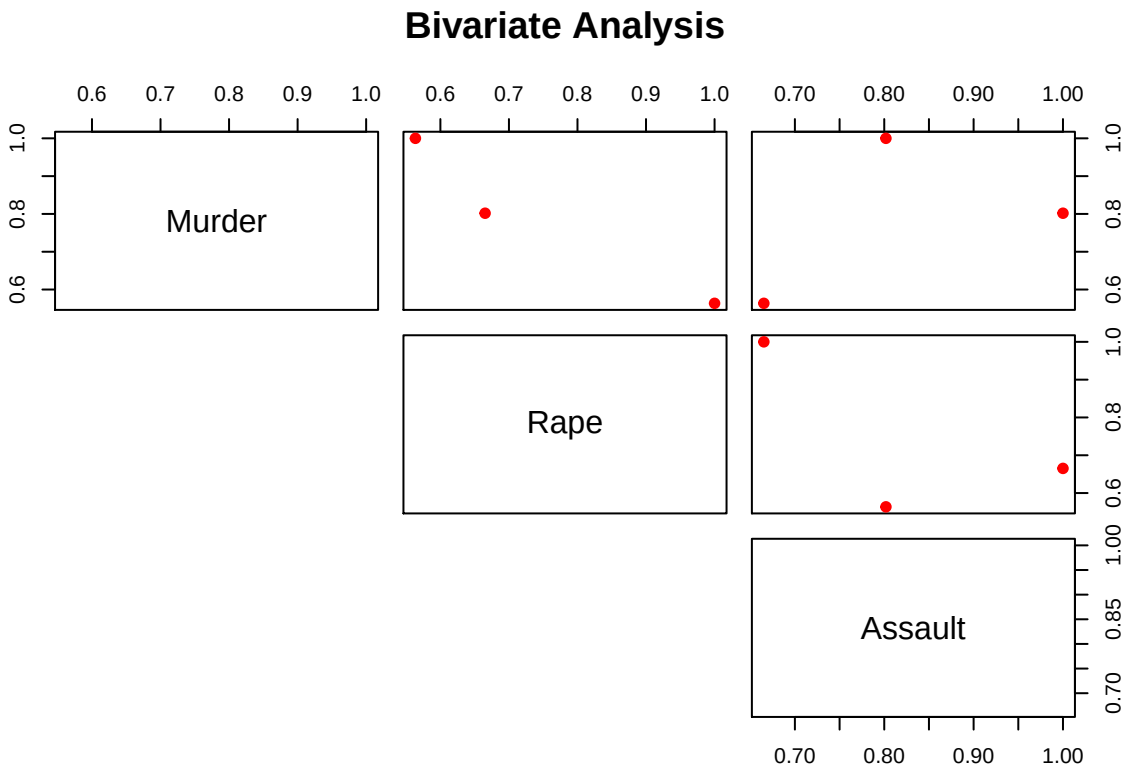



```r
#ordering the values
order <- order(USArrests$Murder)

dotchart(head(order, 15),
         labels = row.names(USArrests),
         main="Dotchart of Violent Crimes (USA - 1973)",
         color = terrain.colors(15),
         cex = 0.9,  pch = 15)
```



3) Bivariate Analysis The next chart shows the correlation among the variables.In plot using pairs function is possible to see that Murder and Assault have a high positive correlation. In fact, the Pearson Correlation Coefficient for these two variables is 0.80, which indicates very high correlation.

```
#calculates the correlation coefficient
cor <- cor(subset)

pairs(cor,
      cex.labels = 1.5,
      lower.panel = NULL,
      pch = 19,
      col="red",
      main="Bivariate Analysis"
      )
```




```
cor
```

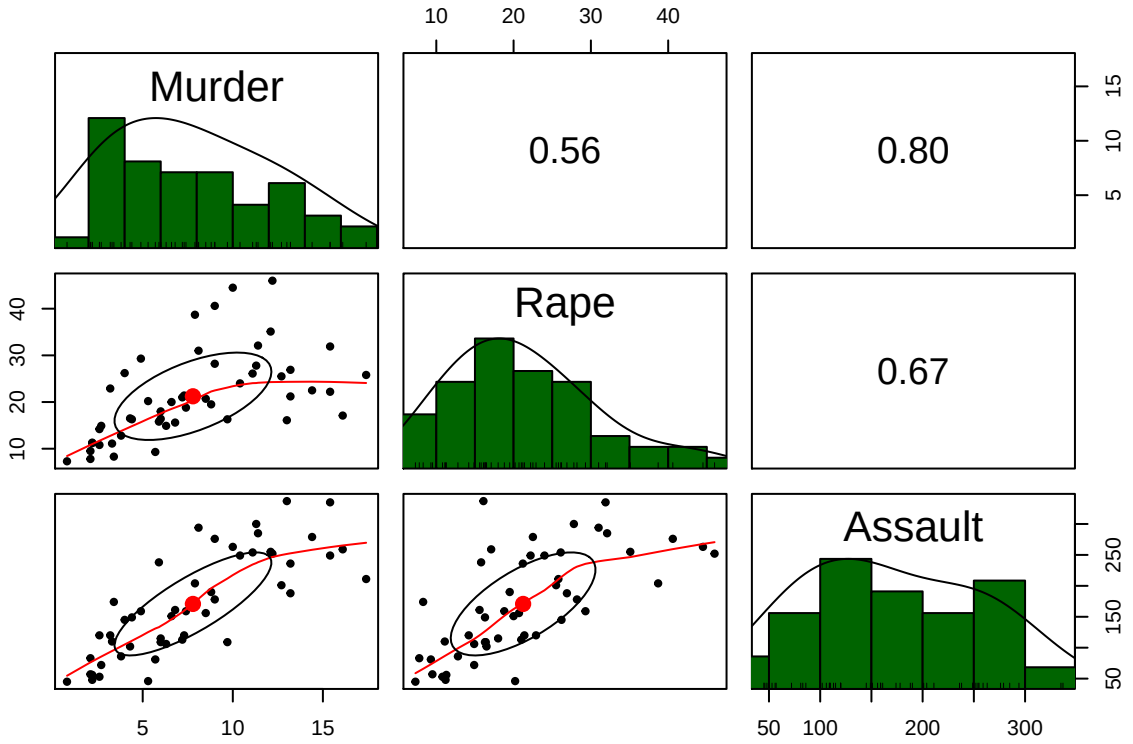
```
##             Murder      Rape   Assault
## Murder  1.0000000 0.5635788 0.8018733
## Rape    0.5635788 1.0000000 0.6652412
## Assault 0.8018733 0.6652412 1.0000000
```

Another way of comparing the correlation of all variables against each other is by using the library psych. In the plots below we can see the Pearson Correlation Coefficient among all variables, the distribution of the variables and the regression line. It is possible to conclude that the crime variables are positively correlated, being murder and assault the most correlated among each other.

```
library(psych)
```

```
## Warning: package 'psych' was built under R version 4.0.5
```

```
pairs.panels(subset,
             method = "pearson", # correlation method
             hist.col = "darkgreen",
             density = TRUE,  # show density plots
             ellipses = TRUE ,# show correlation ellipses
             cex = 0.45)
```



4) Descriptive Statistics with different functions

The last chart uses the ggplot2 library to plot a histogram and its normal distribution.One of the main advantage of working with ggplot2 is the possibility to better handle the variables by integrating dplyr functions.

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```
##
## Attaching package: 'ggplot2'
```

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

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.0.5
```

```
##
## 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
```

```
USArrests %>%
  arrange(Murder) %>%
  ggplot(aes(x = Murder)) +
  geom_histogram(aes(y =..density..), binwidth = 2.5, fill="darkgreen") +
  theme_bw()+
  ggtitle("Histogram Murder Crimes (USA - 1973)") +
  stat_function(
    fun = dnorm,
    args = list(
      mean = USArrests %>% pull(Murder) %>% mean(),
      sd = USArrests %>% pull(Murder) %>% sd()
    ),
    colour = "red", size = 1
  )
```