Data Preparation & Outliers removal

2024-02-01

Outliers removal

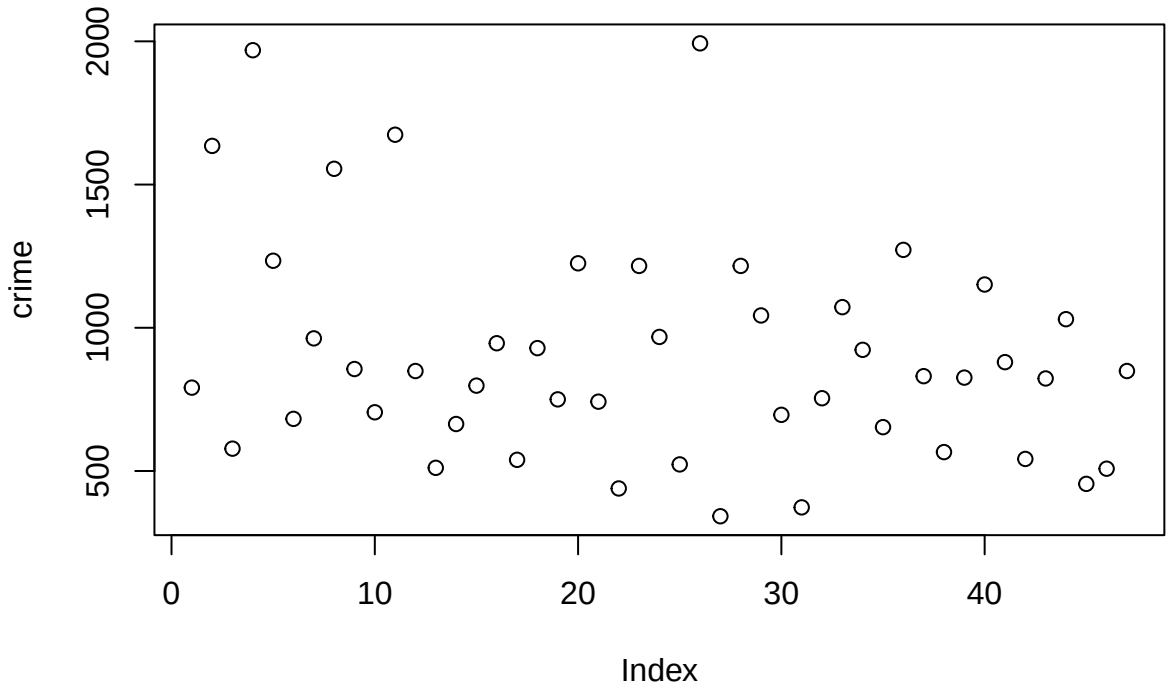
```
#load library outliers.
library (outliers)

#load library corrplot.
library (corrplot)
```

```
## corrplot 0.92 loaded
```

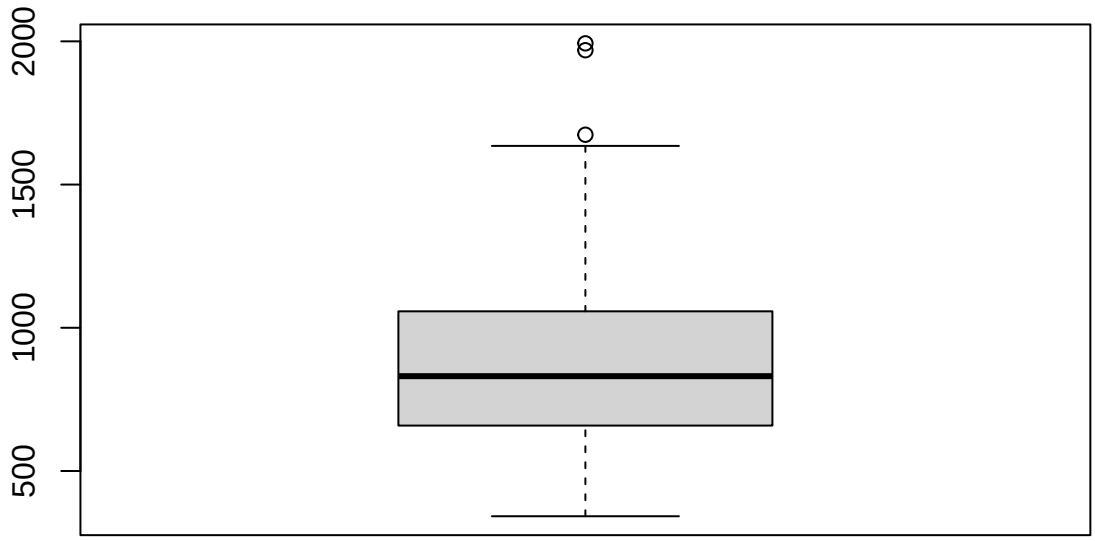
```
#import data 'uscrime.txt' into table with headers.
uscrime <- read.table("uscrime.txt", stringsAsFactors = FALSE, header = TRUE)

#head data from table, view first 6 data points.
head(uscrime)
```

```
##      M So   Ed  Po1  Po2    LF   M.F Pop   NW    U1  U2 Wealth Ineq     Prob
## 1 15.1  1  9.1  5.8  5.6 0.510  95.0  33 30.1 0.108 4.1   3940 26.1 0.084602
## 2 14.3  0 11.3 10.3  9.5 0.583 101.2  13 10.2 0.096 3.6   5570 19.4 0.029599
## 3 14.2  1  8.9  4.5  4.4 0.533  96.9  18 21.9 0.094 3.3   3180 25.0 0.083401
## 4 13.6  0 12.1 14.9 14.1 0.577  99.4 157  8.0 0.102 3.9   6730 16.7 0.015801
## 5 14.1  0 12.1 10.9 10.1 0.591  98.5  18  3.0 0.091 2.0   5780 17.4 0.041399
## 6 12.1  0 11.0 11.8 11.5 0.547  96.4  25  4.4 0.084 2.9   6890 12.6 0.034201
##      Time Crime
## 1 26.2011   791
## 2 25.2999  1635
## 3 24.3006   578
## 4 29.9012  1969
## 5 21.2998  1234
## 6 20.9995   682
```

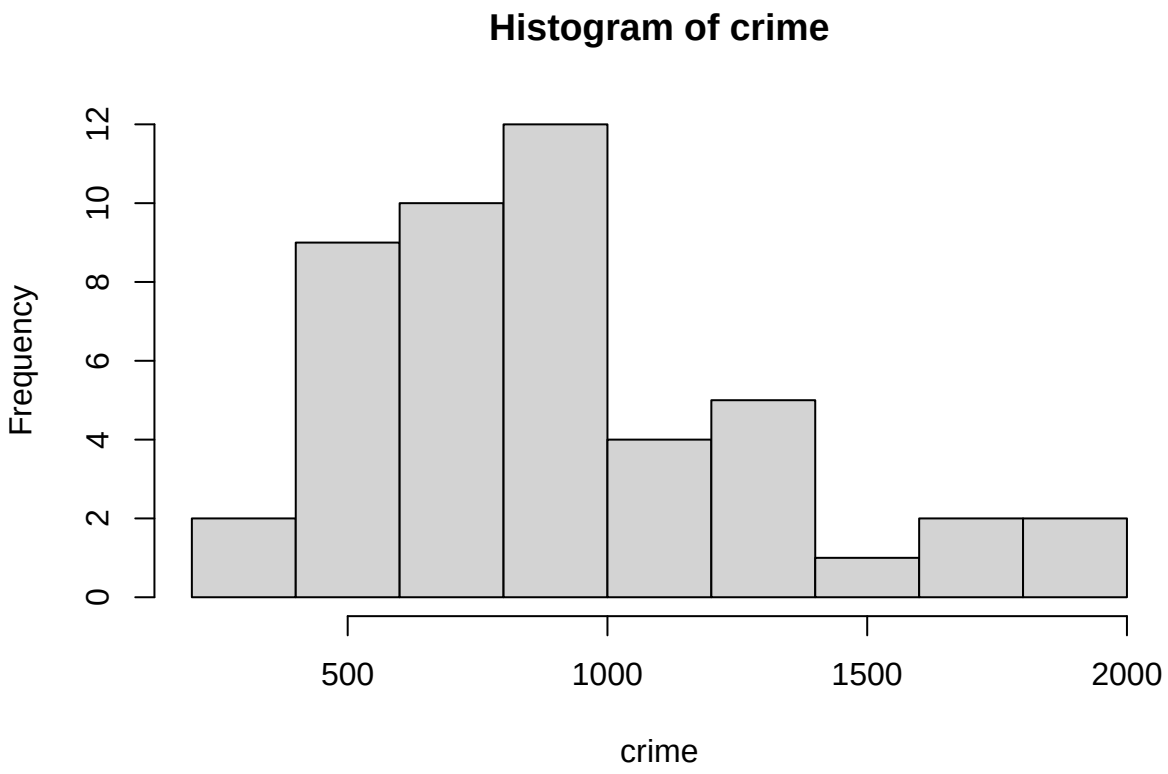
```
#Put data from column 'Crime' into a vector called 'crime'.
crime <- uscrime[,"Crime"]

#plot values of vector 'crime' to see if any outliers.
plot(crime)
```




```
#plot values of vector 'crime' into a boxplot to statistically analyze if outliers exist.
boxplot(crime)
```



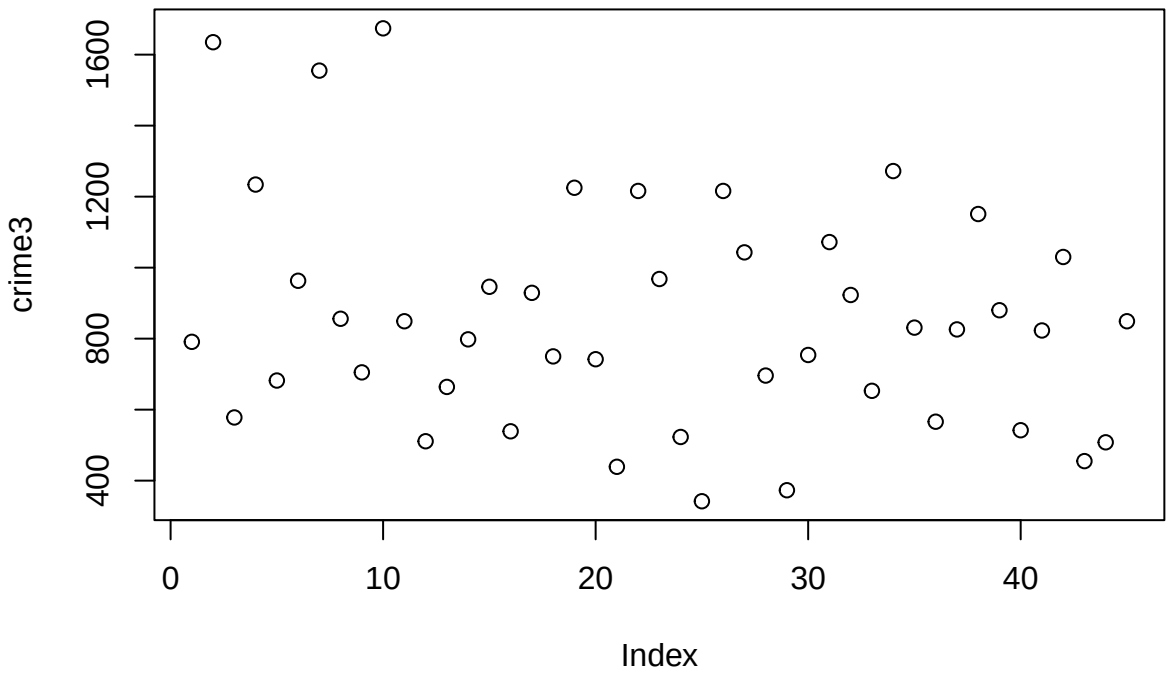

```
#plot values of vector 'crime' into a histogram to view distribution of data.
hist(crime)
```



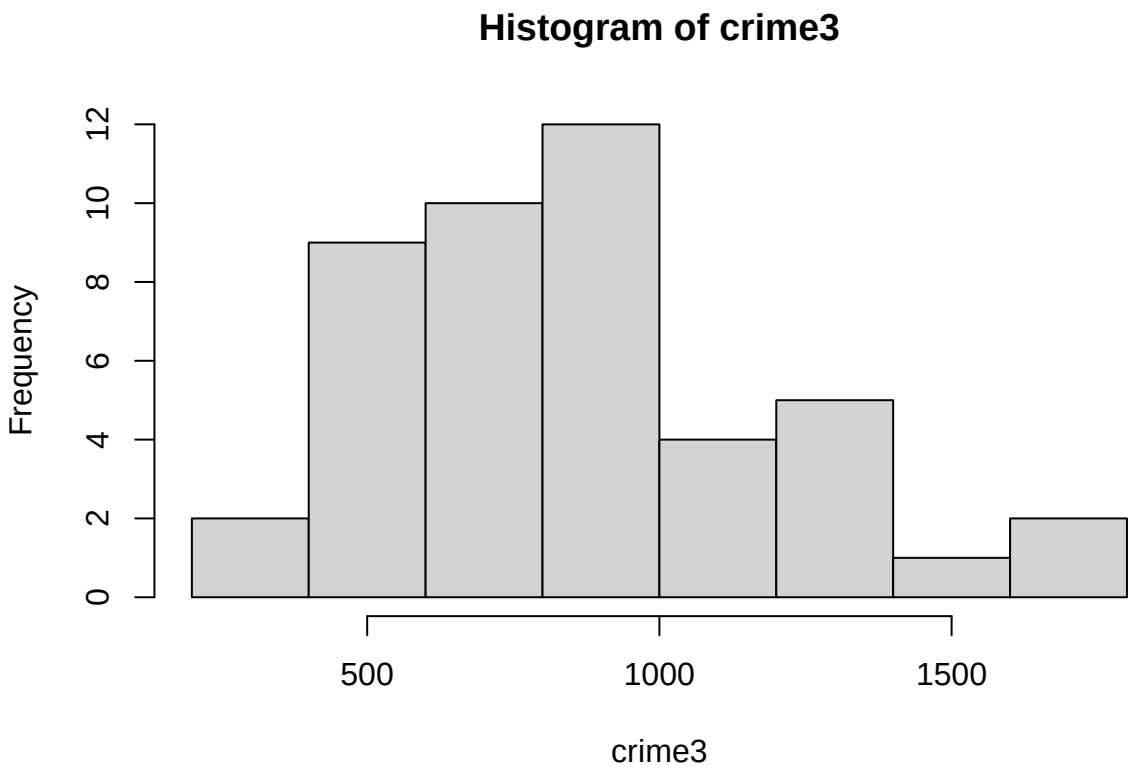

```
#find the mean value of vector crime.
m <- mean(crime)

#find standard deviation of vector crime.
std <- sqrt(var(crime))

#grubs test to identify outlier
grubbs.test(crime, type =10)
```

```
##
##  Grubbs test for one outlier
##
## data:  crime
## G = 2.81287, U = 0.82426, p-value = 0.07887
## alternative hypothesis: highest value 1993 is an outlier
```

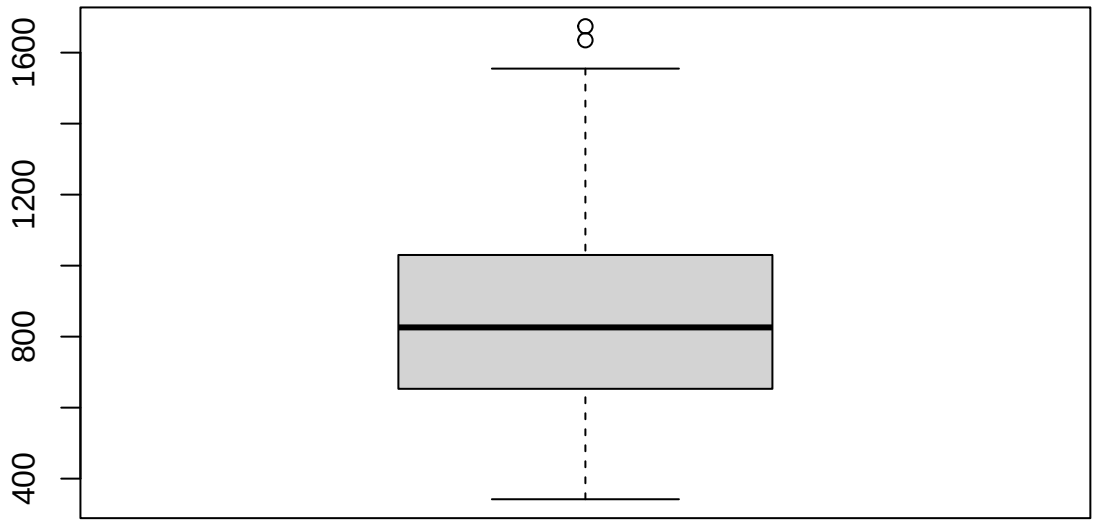
```
# outliers (2) removed as they were far out and above the rest of the data points.
crime2 <-uscrime[-26,16]
crime3 <-crime2[-4]

#plot new data with removed outliers to view spread.
plot(crime3)
```




```
#plot new data with removed outliers in histogram to view bell curve.
hist(crime3)
```




```
#View final data with removed outliers in boxplot to check for more outliers.
boxplot(crime3)
```




```
#Check via grubs test
grubbs.test(crime3, type =10)
```

```
##
##  Grubbs test for one outlier
##
## data:  crime3
## G = 2.56457, U = 0.84712, p-value = 0.1781
## alternative hypothesis: highest value 1674 is an outlier
```

In this analysis, I utilized the "uscrime.txt" data file to examine the data for any outliers using the grubbs.test function.

I employed various plots, histograms, and boxplots to more clearly visualize the data and uncover details that are not immediately apparent from the dataset as a whole. My focus was on the "Crime" column, which represents crimes committed per 100,000 people.

Following my code and comments, you'll find that I identified an outlier, 1993, from the crime vector using the grubbs.test. I also chose to remove the point 1969, as it was nearly as far from the mean and standard deviation of the data. After removing these two outliers, I re-plotted the data with a plot, histogram, and boxplot. To me, the boxplot suggested the presence of two additional outliers. However, after applying "grubbs.test(crime3, type = 10)," the results indicated a p-value of .1781. This suggests a 17.8% probability of obtaining a test statistic as extreme as the one observed, which is relatively high. Therefore, I decided not to remove these two so-called outliers. The data is now considered valid, clean, and ready for use.