Tommy Neubert ()
Benjamin Reissaus(749897)   Data Mining Exercise 1   21.04.2015
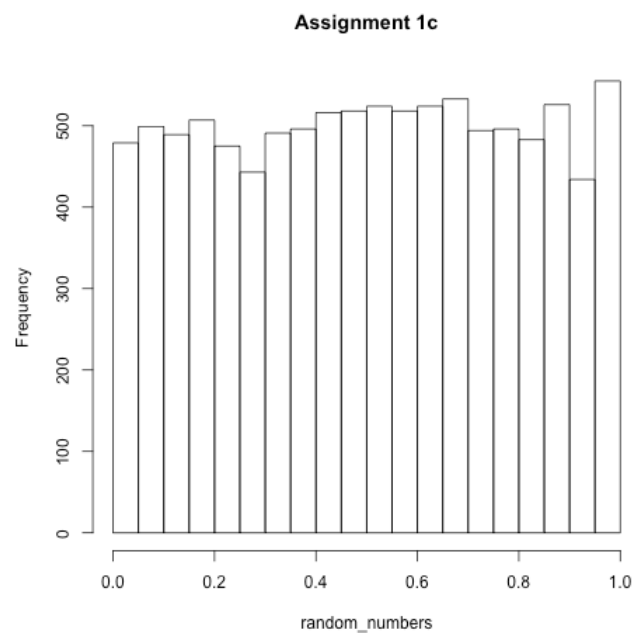
Assignment 1

a)
mean: 0.508
median: 0.545
variance: 0.1415956

b)
mean = 0.9982699
median = 0.4995539
var = 0.08299303

c)

Assignment 2

min(faithful\$eruptions) = 1.6
max(faithful\$eruptions) = 5.1
median(faithful\$eruptions) = 4.0
mean(faithful\$eruptions) = 3.488
difference between mean and median: 0.512

min(faithful\$waiting) = 43
max(faithful\$waiting) = 96
median(faithful\$waiting) = 76
mean(faithful\$waiting) = 70.89706
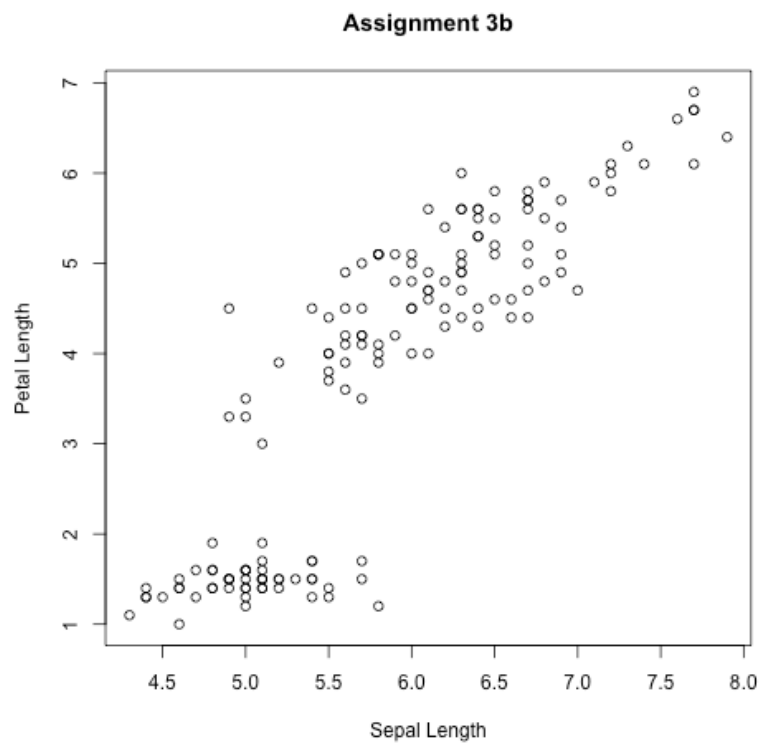difference between mean and median: 5,10294

Assignment 3
a)
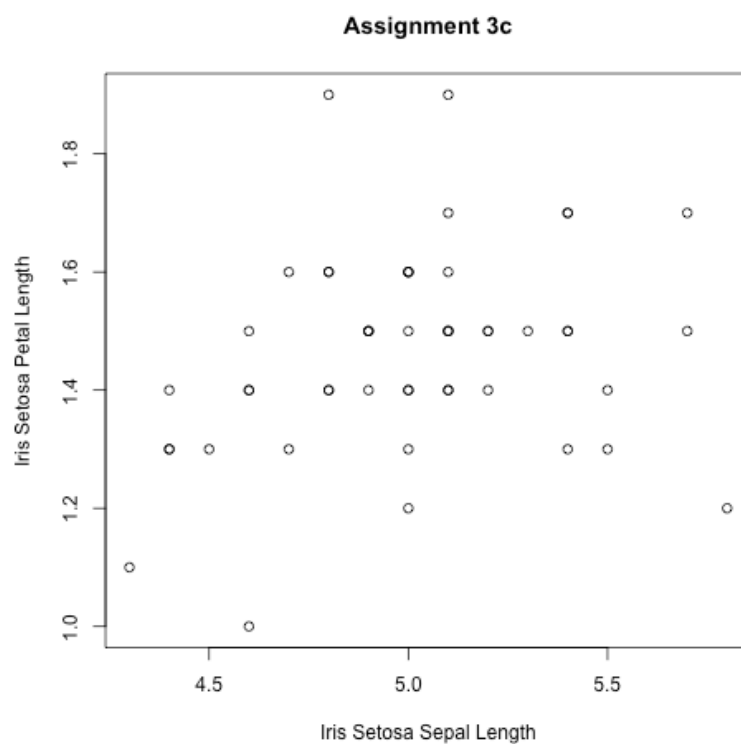Unique Values and how often they occur:
Iris-setosa : 50
Iris-versicolor : 50
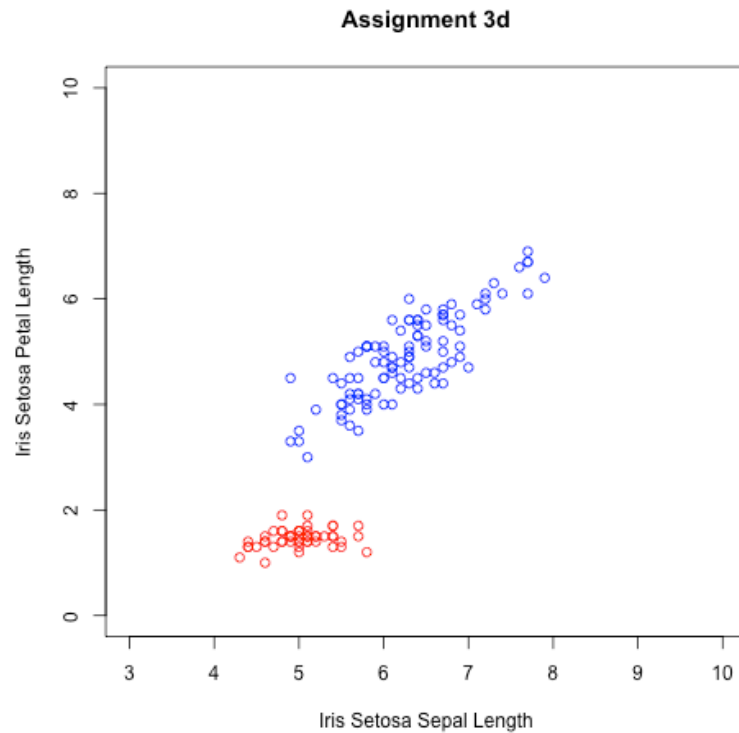Iris-virginica : 50

b)

c)

The Petal Length and Sepal Length of Iris Virginica and Versicolor are also correlated positively.
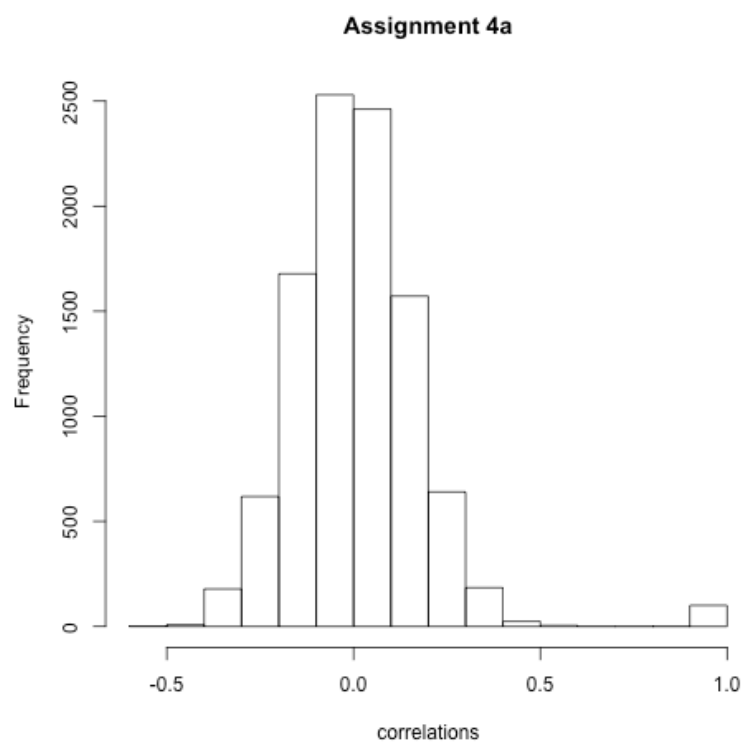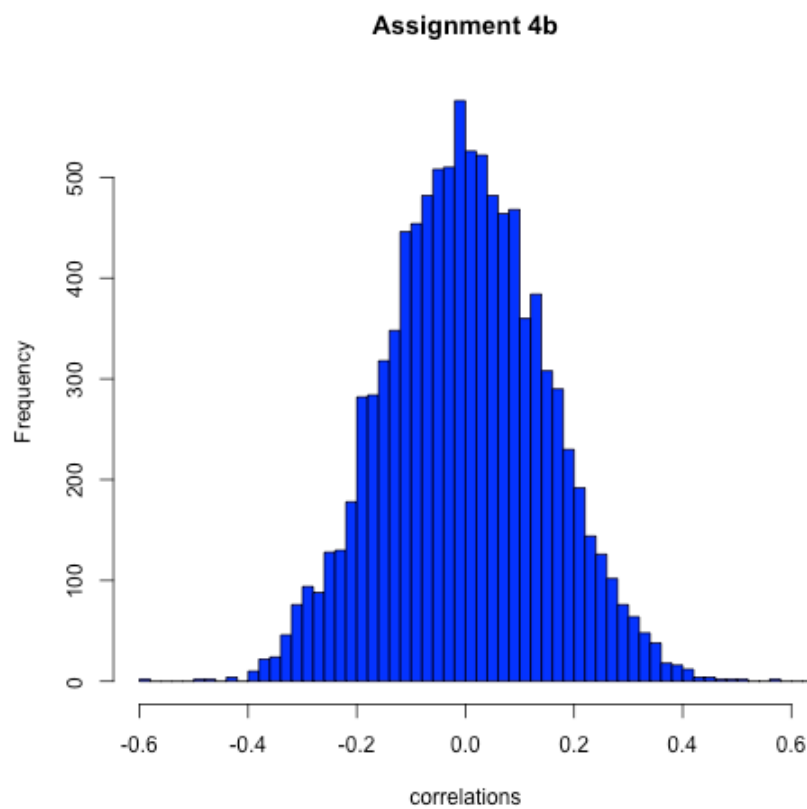The Petal Length and Seapal Length of Iris-Setosa is not correlated.

d)

Assignment 4
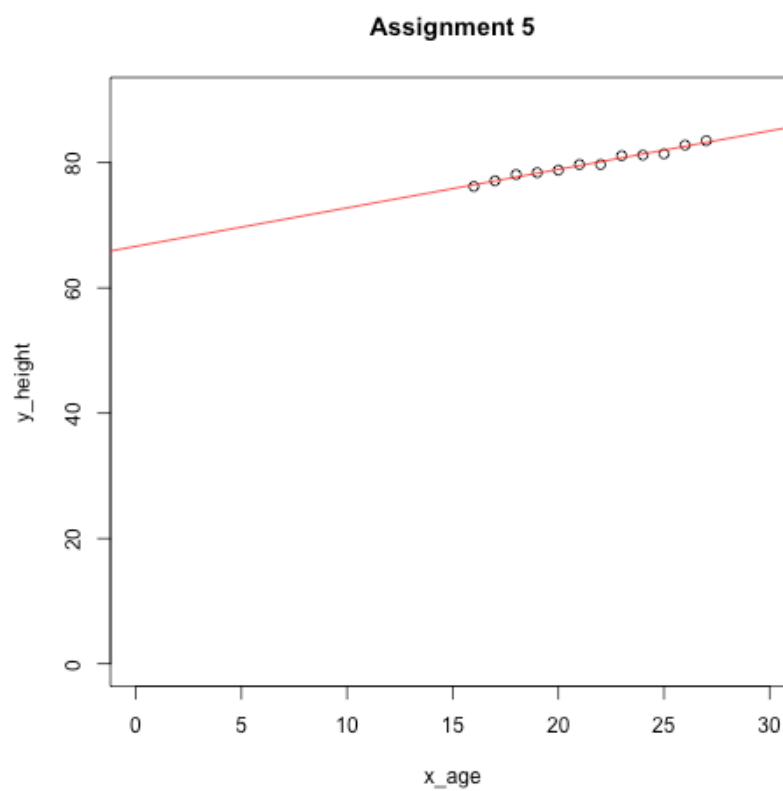a)

b)

Assignment 4b

Assignment 5

Assignment 5

Assignment 6

High accuracy:
Detecting Cancer is something that needs high accuracy. People who do not have cancer
should not be treated with chemo therapy and all people who have cancer should be
classified as having cancer.

High Precision:
An Antivirus progamm is an example which should have a high recall. It is not
problematic to accuse a 'program' of being malicious when it is not but highly important
to find all virus programs.

Assignment 7

maximum precision: score > 0.93
This way only the first element with a score of 0.05 would be classified into class C
which would be correct and lead to a precision of 100%.

maximum recall: score >= 0.48
This way all items get classified into class C. This leads to a recall of 100% because all
items belonging to C are found.

maximum accuracy: score > 0.93

Bonus Assignment 1

Bonus Assignment 1