Big Data Analytics with R - Solutions to Coursework 1

1. Basic Statistics

(a) We first create a vector to hold the data on number of siblings:

```
siblings <- c(2, 3, 0, 5, 2, 1, 1, 0, 3, 3)
```

For this data we have:

(i) Mean: 2

```
mean(siblings)
```

```
## [1] 2
```

(ii) Median: 2

```
median(siblings)
```

```
## [1] 2
```

(iii) Mode: 3

```
names(sort(-table(siblings)))[1]
```

```
## [1] "3"
```

(iv) Variance: 2.4444444

```
var(siblings)
```

```
## [1] 2.444444
```

(v) Standard deviation: 1.5634719

```
sd(siblings)
```

```
## [1] 1.563472
```

(b) Next, create a vector to hold the ages of the students:

```
ages <- c(23, 25, 18, 45, 30, 21, 22, 19, 29, 35)
```

Then we have:

(i) Covariance: 11.8888889

```r
cov(siblings, ages)
```

```
## [1] 11.88889
```

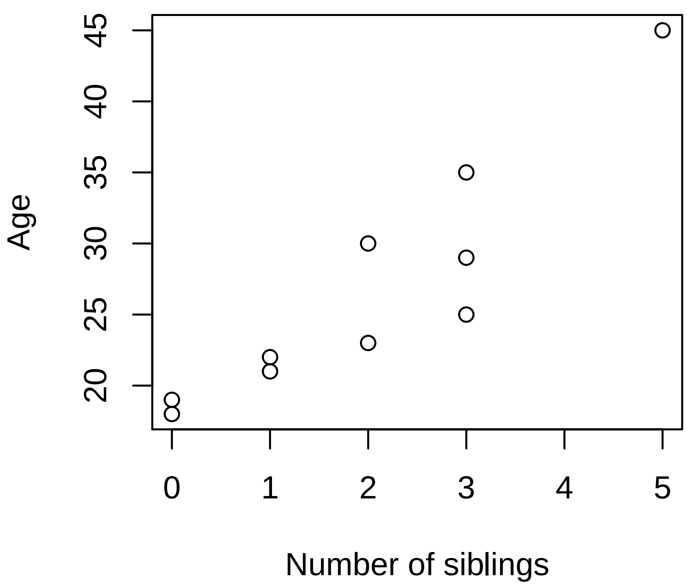
Correlation: 0.9116971

```r
cor(siblings, ages)
```

```
## [1] 0.9116971
```

(ii) From (i) there appears to be a strong positive correlation between the number of siblings and age, which is clear in the following plot:

```r
plot(siblings, ages, xlab="Number of siblings", ylab="Age")
```



(iii) It is unlikely that there is direct causal relationship between the number of siblings and age, since the age of a student should have no influence on how many siblings they have and vice versa.

2. Getting familiar with R

(a) Load the data

```r
library(MASS)
```

The number of rows in the `Boston` data set is

```r
nrow(Boston)
```

```
## [1] 506
```

The number of columns is

```r
ncol(Boston)
```

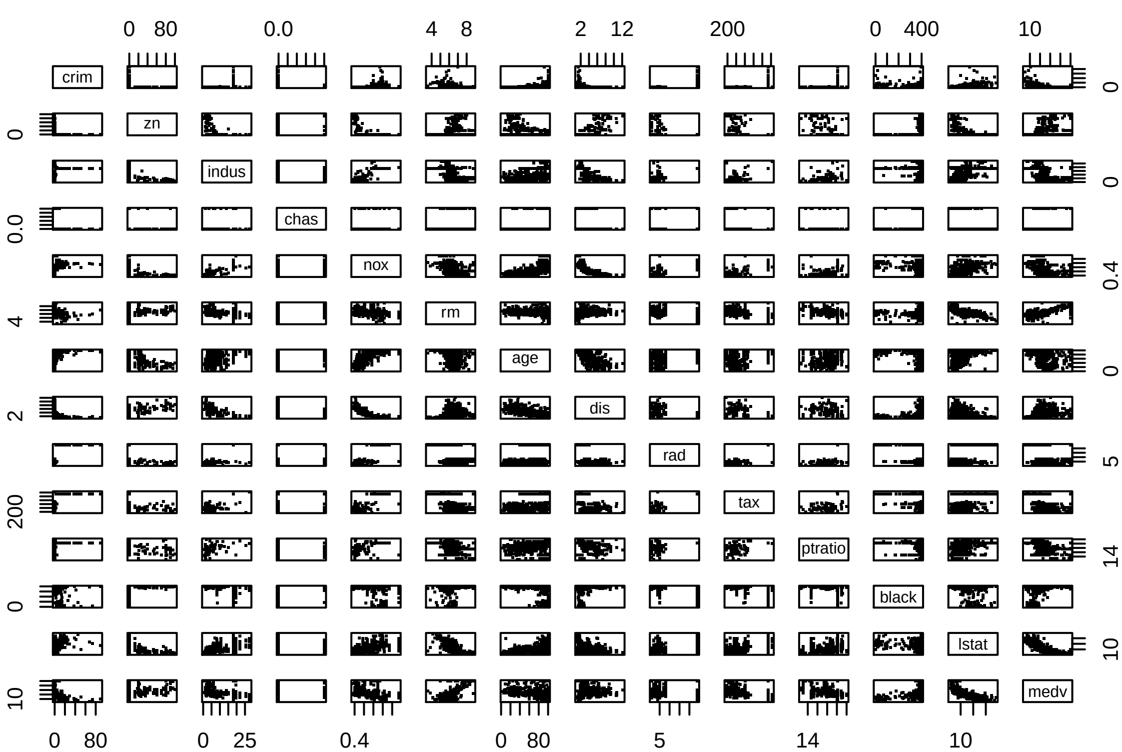
```
## [1] 14
```

From the package documentation we see that the rows are neighbourhoods of the Boston area and the columns are 14 predictor variables, including:

- `crim`: per capita crime rate by town
- `zn`: proportion of residential land zoned for lots over 25,000 sq.ft
- `indus`: proportion of non-retail business acres per town
- `chas`: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
- `nox`: nitrogen oxides concentration (parts per 10 million)
- `rm`: average number of rooms per dwelling
- `age`: proportion of owner-occupied units built prior to 1940
- `dis`: weighted mean of distances to five Boston employment centres
- `rad`: index of accessibility to radial highways
- `tax`: full-value property-tax rate per $10,000
- `ptratio`: pupil-teacher ratio by town
- `black`: 1000(Bk - 0.63)^2, where Bk is the proportion of blacks by town
- `lstat`: lower status of the population (percent)
- `medv`: median value of owner-occupied homes in $1000s

(b) A plot of pairwise scatterplots of all variables is given below:

```r
plot(Boston, pch='.')
```



From the plot we see that some pairs of predictors appear to be related, including:

- `medv` and `rm` (postive)
- `medv` and `lstat` (negative)
- `nox` and `dis` (negative)
- `nox` and `age` (positive)
- `crim` and `lstat` (positive)
- `crim` and `indus` (positive)
- `crim` and `rad` (positive)
- `crim` and `tax` (positive)
- `lstat` and `rm` (negative)
- `medv` and `nox` (negative)

(c) To check if any predictors are related to `crim` we can determine the correlation between `crim` and each of the other variables.

```
crim_cor <- function(predictor, predictor2 = Boston$crim) {return(cor(predictor, Boston$crim))}
print(sort(sapply(Boston, crim_cor)))
```

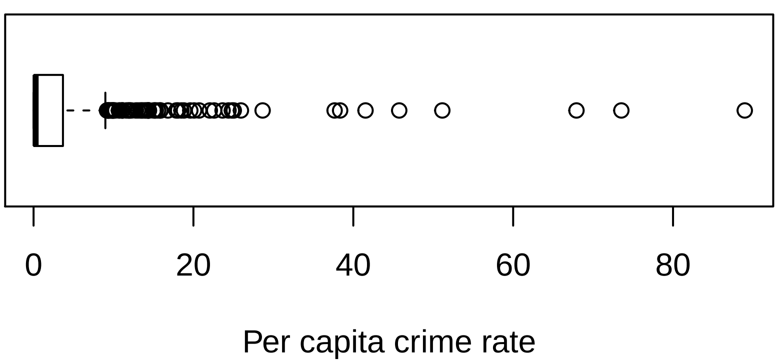
```
##        medv       black         dis          rm          zn        chas
## -0.38830461 -0.38506394 -0.37967009 -0.21924670 -0.20046922 -0.05589158
##     ptratio         age       indus         nox       lstat         tax
##  0.28994558  0.35273425  0.40658341  0.42097171  0.45562148  0.58276431
##         rad        crim
##  0.62550515  1.00000000
```

This suggests that `crim` is most associated with:

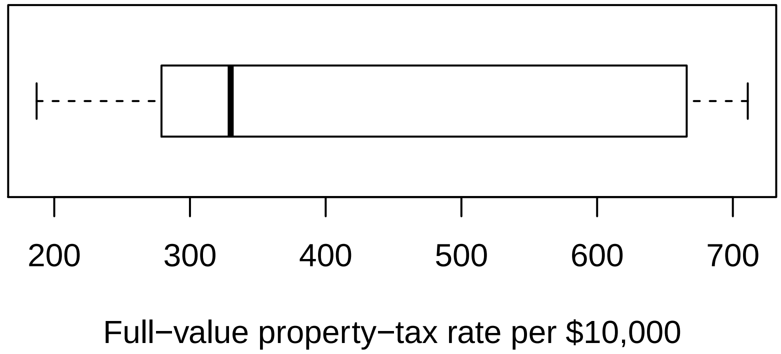
- `rad` (positive)
- `tax` (positive)
- `lstat` (positive)
- `nox` (positive)
- `indus` (positive)
- `black` (negative)
- `medv` (negative)

(d) Plotting boxplots for `crim`, `tax` and `ptratio` we see

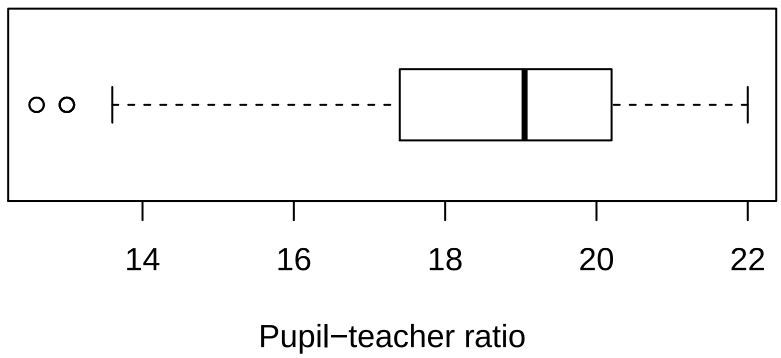
```
boxplot(Boston[c("crim")], horizontal=TRUE, par(pin=c(4,1)), xlab="Per capita crime rate")
```



```r
boxplot(Boston[c("tax")], horizontal=TRUE, xlab="Full-value property-tax rate per $10,000")
```




```r
boxplot(Boston[c("ptratio")], horizontal=TRUE, xlab="Pupil-teacher ratio")
```




```r
t(sapply(Boston[c("crim", "tax", "ptratio")], range))
```

```
##              [,1]     [,2]
## crim    6.32e-03  88.9762
## tax     1.87e+02 711.0000
## ptratio 1.26e+01  22.0000
```

“‘

The per-capita crime rate is highly positively skewed and with a wide range (0.00632-88.98). Most neigbourhoods having low crime rates around 0 and 9 neighbourhoods being extreme outliers, with crime rates above 30%.

The distribution of tax rates is also positively skewed (range 187-711) but with no extreme outliers.

The distribution of pupil-teacher ratios is more even and with a relatively narrow range (12.6-22.0), with only two extreme outliers on the low end.

(e) The number of neighbourhoods that bound the Charles River is:

```r
sum(Boston$chas)
```

```
## [1] 35
```

(f) The median pupil-teacher ratio among the neighbourhoods is:

```r
median(Boston$ptratio)
```

```
## [1] 19.05
```

(g) The neighbourhood with the lowest median value of owner occupied homes is:

```r
which.min(Boston$medv)
```

```
## [1] 399
```

The predictor variable values for this neighbourhood are:

```r
t(Boston[which.min(Boston$medv), ])
```

```
##              399
## crim     38.3518
## zn        0.0000
## indus    18.1000
## chas      0.0000
## nox       0.6930
## rm        5.4530
## age     100.0000
## dis       1.4896
## rad      24.0000
## tax     666.0000
## ptratio  20.2000
## black   396.9000
## lstat    30.5900
## medv      5.0000
```

The range for all 14 predictor variables are:

```r
t(sapply(Boston, range))
```

```
##              [,1]     [,2]
## crim      0.00632  88.9762
## zn        0.00000 100.0000
## indus     0.46000  27.7400
## chas      0.00000   1.0000
## nox       0.38500   0.8710
## rm        3.56100   8.7800
## age       2.90000 100.0000
## dis       1.12960  12.1265
## rad       1.00000  24.0000
## tax     187.00000 711.0000
## ptratio  12.60000  22.0000
## black     0.32000 396.9000
## lstat     1.73000  37.9700
## medv      5.00000  50.0000
```

Comparing the values for neighbourhood 399 with the ranges across the whole data set, we see the following predictors as standing out: `zn` (low), `indus` (high), `age` (high), `dis` (low), `rad` (high), `tax` (high), `ptratio` (high), `black` (high), and `medv` (low).

The values suggest that this is an, old, highly industrial, low-income, inner-city neighbourhood, with a high proportion of black people and with schools having a large number of pupils per teacher.

(h) The number of neighbourhoods averaging more than seven rooms per dwelling is

```
nrow(Boston[Boston$rm > 7, ])
```

```
## [1] 64
```

The number of neighbourhoods averaging more than eight rooms per dwelling is

```
nrow(Boston[Boston$rm > 8, ])
```

```
## [1] 13
```

The predictor values for those neighbourhoods averaging more than eight rooms per dwelling are:

```
Boston[Boston$rm > 8, ]
```

```
##        crim zn indus chas    nox    rm  age    dis rad tax ptratio  black
## 98  0.12083  0  2.89    0 0.4450 8.069 76.0 3.4952   2 276    18.0 396.90
## 164 1.51902  0 19.58    1 0.6050 8.375 93.9 2.1620   5 403    14.7 388.45
## 205 0.02009 95  2.68    0 0.4161 8.034 31.9 5.1180   4 224    14.7 390.55
## 225 0.31533  0  6.20    0 0.5040 8.266 78.3 2.8944   8 307    17.4 385.05
## 226 0.52693  0  6.20    0 0.5040 8.725 83.0 2.8944   8 307    17.4 382.00
## 227 0.38214  0  6.20    0 0.5040 8.040 86.5 3.2157   8 307    17.4 387.38
## 233 0.57529  0  6.20    0 0.5070 8.337 73.3 3.8384   8 307    17.4 385.91
## 234 0.33147  0  6.20    0 0.5070 8.247 70.4 3.6519   8 307    17.4 378.95
## 254 0.36894 22  5.86    0 0.4310 8.259  8.4 8.9067   7 330    19.1 396.90
## 258 0.61154 20  3.97    0 0.6470 8.704 86.9 1.8010   5 264    13.0 389.70
## 263 0.52014 20  3.97    0 0.6470 8.398 91.5 2.2885   5 264    13.0 386.86
## 268 0.57834 20  3.97    0 0.5750 8.297 67.0 2.4216   5 264    13.0 384.54
## 365 3.47428  0 18.10    1 0.7180 8.780 82.9 1.9047  24 666    20.2 354.55
##     lstat medv
## 98   4.21 38.7
## 164  3.32 50.0
## 205  2.88 50.0
## 225  4.14 44.8
## 226  4.63 50.0
## 227  3.13 37.6
## 233  2.47 41.7
## 234  3.95 48.3
## 254  3.54 42.8
## 258  5.12 50.0
## 263  5.91 48.8
## 268  7.44 50.0
## 365  5.29 21.9
```

Looking at just the 13 neighbourhoods averaging more than eight rooms per dwelling, we see several similiarities, including a low crime rate (`crim`), low industrialisation (`indus`), high age (`age`), a high proportion of black residents (`black`), and a high median value (`medv`).

3. Linear Regression

(a) We first set the seed for the session and then create a vector `x` that includes 100 draws from the standard normal distribution.

```
set.seed(1)
x <- rnorm(100, mean=0, sd=1)
```

(b) Next create the vector `eps`, containing 100 draws from a normal distribution with mean 0 and standard deviation 0.5 (i.e. variance = 0.25).

```
eps <- rnorm(100, mean=0, sd=0.5)
```

(c) Now generate the `y` outcome vector, using the formula $\mathbf{Y} = -1 + 0.5\mathbf{X} + \epsilon$.

```
y <- -1 + 0.5 * x + eps
```

The vector `y` has length

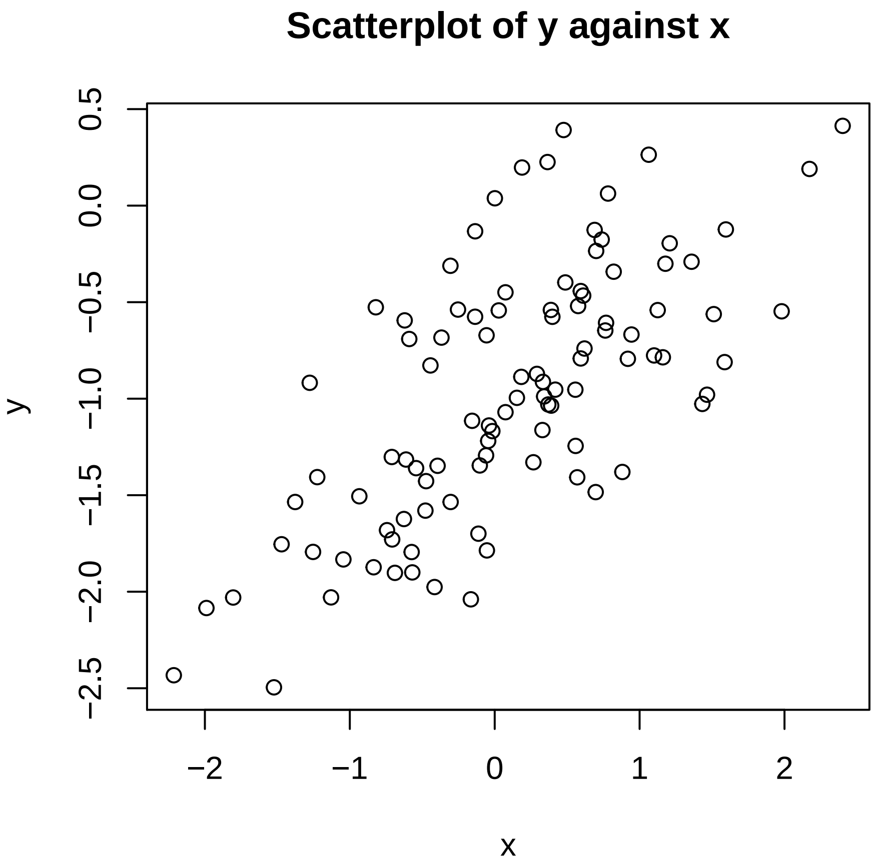
```
length(y)
```

```
## [1] 100
```

In this model we have $\beta_0 = -1$ and $\beta_1 = 0.5$.

(d) We can create a scatterplot of `y` against `x`

```
plot(x, y, main="Scatterplot of y against x")
```



The plot shows a positive relationship between `x` and `y`.

(e) Regressing `y` on `x` we get

```
fit = lm(y~x)
summary(fit)
```

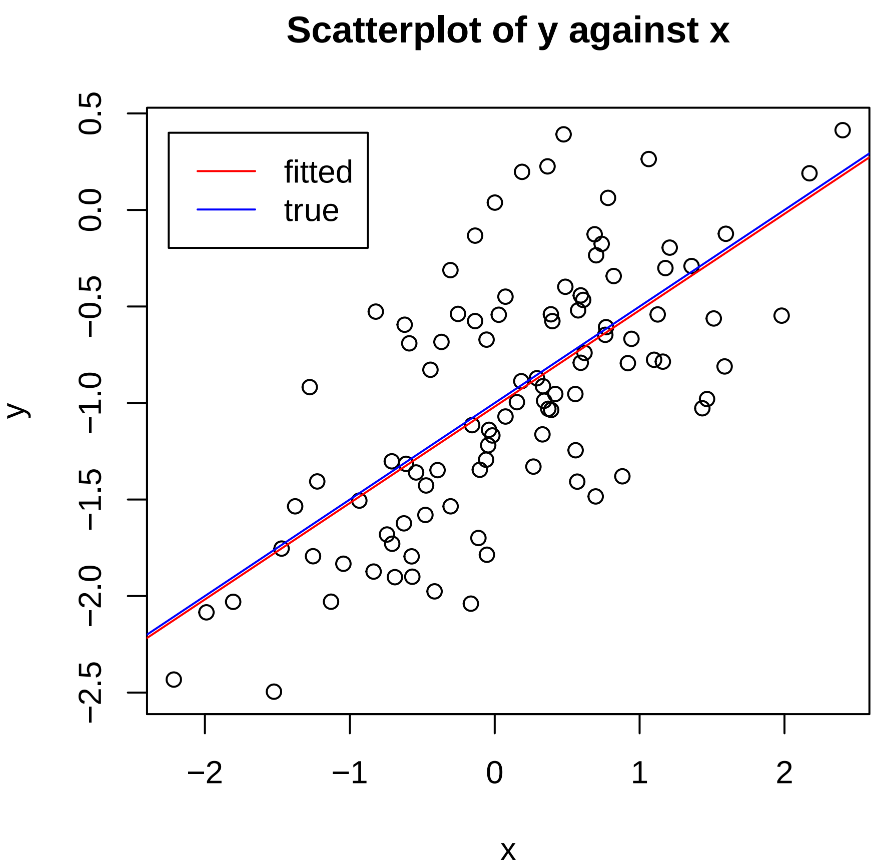
```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.93842 -0.30688 -0.06975  0.26970  1.17309
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.01885    0.04849 -21.010  < 2e-16 ***
## x            0.49947    0.05386   9.273 4.58e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4814 on 98 degrees of freedom
## Multiple R-squared:  0.4674, Adjusted R-squared:  0.4619
## F-statistic: 85.99 on 1 and 98 DF,  p-value: 4.583e-15
```

The model has R^2 = 0.4673515 suggesting that the fit is only reasonably good.

We see that $\hat{\beta}_0$ = -1.0188463 and $\hat{\beta}_1$ = 0.4994698, which are very close to the true values of $\beta_0 = -1$ and $\beta_1 = 0.5$. Both estimates are significantly different from 0.

(f) We can replot the data and now draw on the estimated regression line (blue) and the true population regression line (red).

```
plot(x, y, main="Scatterplot of y against x")
abline(fit, col="red")
abline(-1, 0.5, col="blue")
legend(-2.25, 0.4, legend=c("fitted", "true"), col=c("red", "blue"), lty=1, lwd=1)
```



From the plot, it's clear that the lines are nearly identical.