Linear Models with R - Julian J. Faraway

Ananda Biswas

Chapter 1 - Exercise 1

- loading the library

```
library(faraway)

## Warning:  package 'faraway' was built under R version 4.2.3
```

- loading the data set

```
data("teengamb")

my_data <- teengamb
```

- description of the data

```
`?`(teengamb)

## starting httpd help server ...  done
```

- doing sanity checks

```
dim(my_data)

## [1] 47  5
```

```
names(my_data)

## [1] "sex"    "status" "income" "verbal" "gamble"
```

```
head(my_data)

##   sex status income verbal gamble
## 1   1     51   2.00      8    0.0
## 2   1     28   2.50      8    0.0
## 3   1     37   2.00      6    0.0
## 4   1     28   7.00      4    7.3
## 5   1     65   2.00      8   19.6
## 6   1     61   3.47      6    0.1
```

```
tail(my_data)

##    sex status income verbal gamble
## 42   0     61  15.00      9   69.7
## 43   0     75   3.00      8   13.3
## 44   0     66   3.25      9    0.6
## 45   0     62   4.94      6   38.0
## 46   0     71   1.50      7   14.4
## 47   0     71   2.50      9   19.2
```

- From the description we see that, sex is a categorical variable.

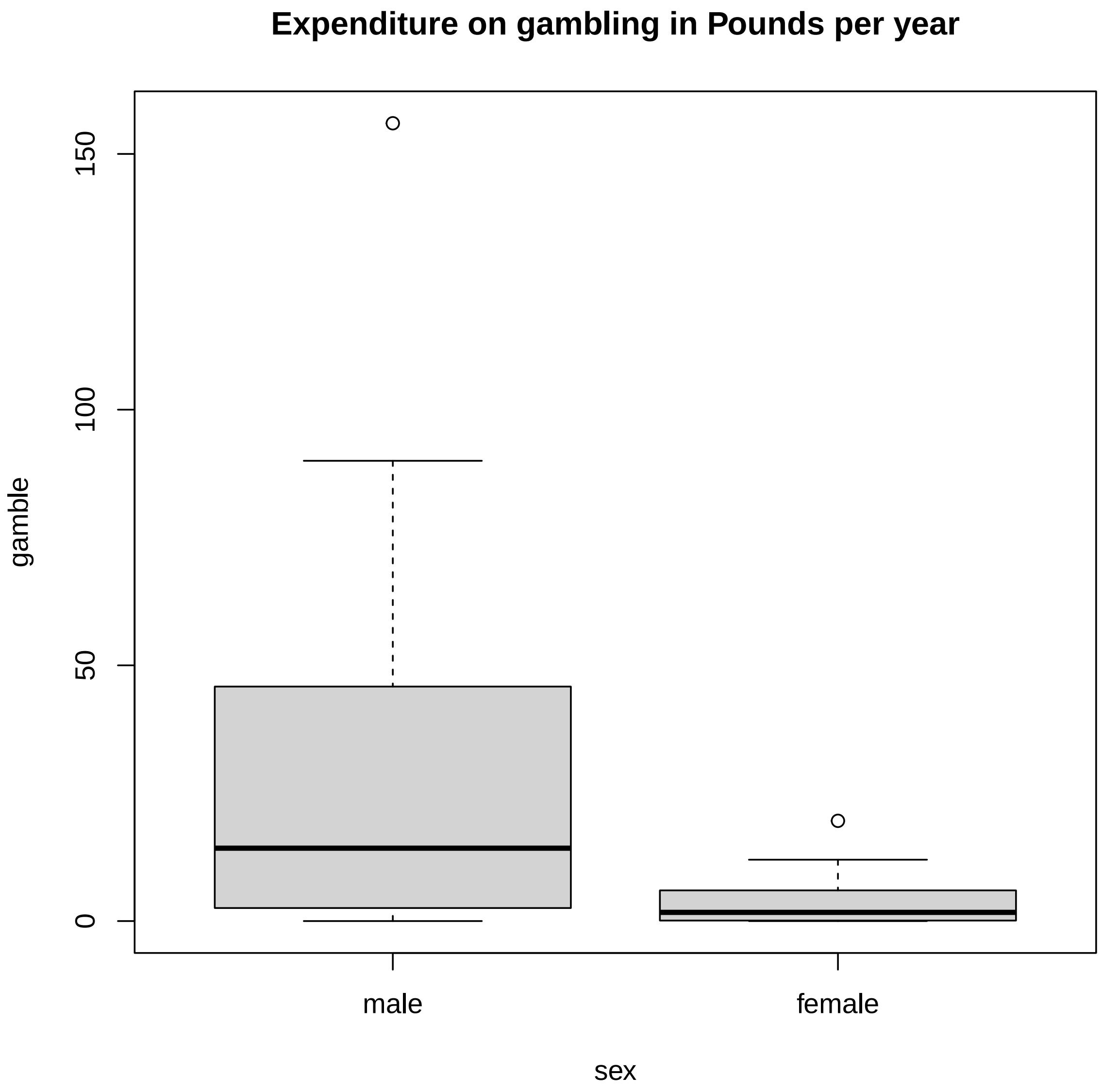
```
my_data$sex <- factor(my_data$sex)

levels(my_data$sex) <- c("male", "female")
```

```
class(my_data$sex)

## [1] "factor"

class(my_data$status)

## [1] "integer"

class(my_data$income)

## [1] "numeric"

class(my_data$verbal)

## [1] "integer"

class(my_data$gamble)

## [1] "numeric"
```

```
summary(my_data)

##      sex         status          income           verbal          gamble     
##  male  :28   Min.   :18.00   Min.   : 0.600   Min.   : 1.00   Min.   :  0.0  
##  female:19   1st Qu.:28.00   1st Qu.: 2.000   1st Qu.: 6.00   1st Qu.:  1.1  
##              Median :43.00   Median : 3.250   Median : 7.00   Median :  6.0  
##              Mean   :45.23   Mean   : 4.642   Mean   : 6.66   Mean   : 19.3  
##              3rd Qu.:61.50   3rd Qu.: 6.210   3rd Qu.: 8.00   3rd Qu.: 19.4  
##              Max.   :75.00   Max.   :15.000   Max.   :10.00   Max.   :156.0  
```

- Now we shall do a box-plot for gambling expenditure and sex.

```
boxplot(gamble ~ sex, my_data, main = "Expenditure on gambling in Pounds per year")
```



☞ We see that expenditure on gambling by men is much higher than that by women.

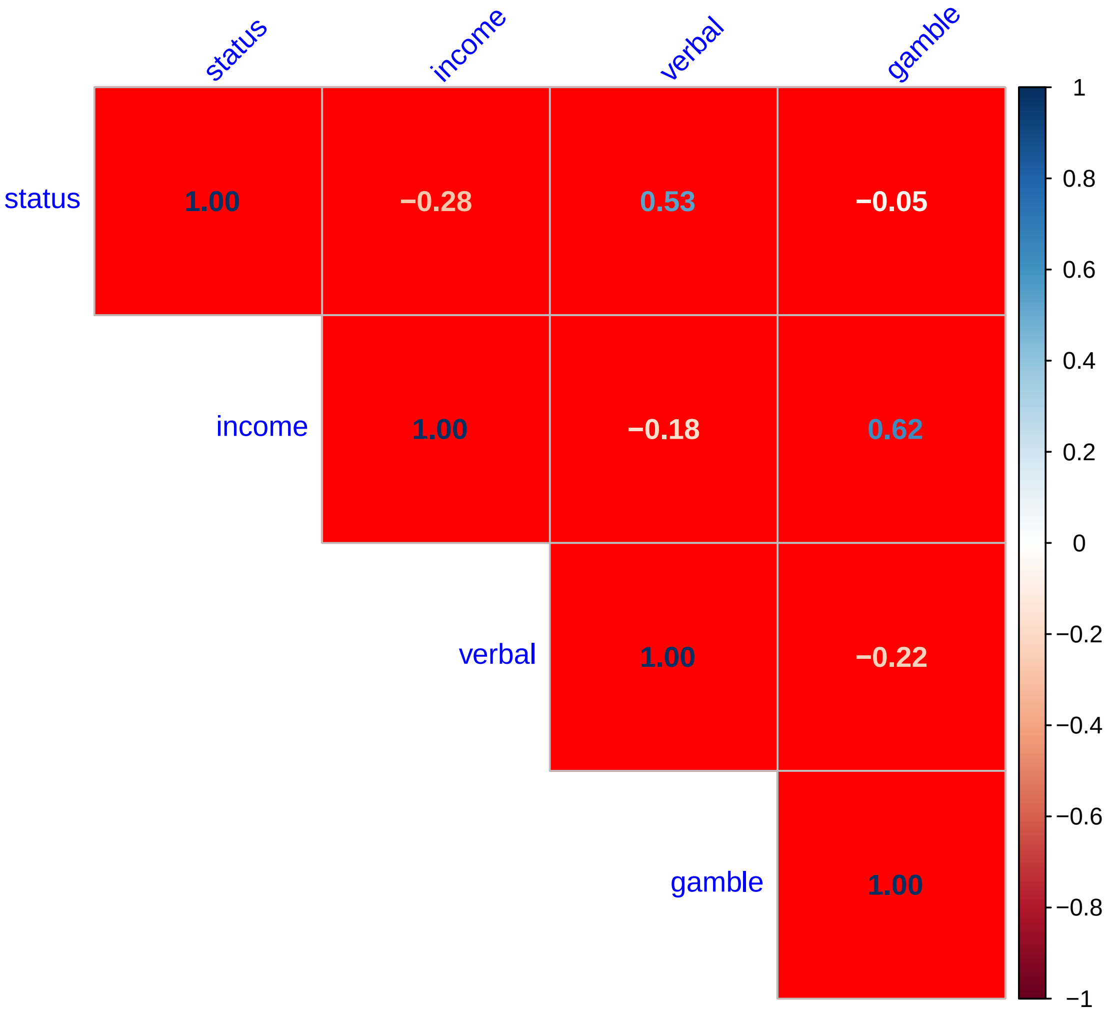
- Now we shall see the correlation matrix.

```
cor(my_data[-1])

##             status     income     verbal      gamble
## status  1.00000000 -0.2750340  0.5316102 -0.05042081
## income -0.27503402  1.0000000 -0.1755707  0.62207690
## verbal  0.53161022 -0.1755707  1.0000000 -0.22005619
## gamble -0.05042081  0.6220769 -0.2200562  1.00000000
```

A helpful visualization of the correlation matrix is as follows :

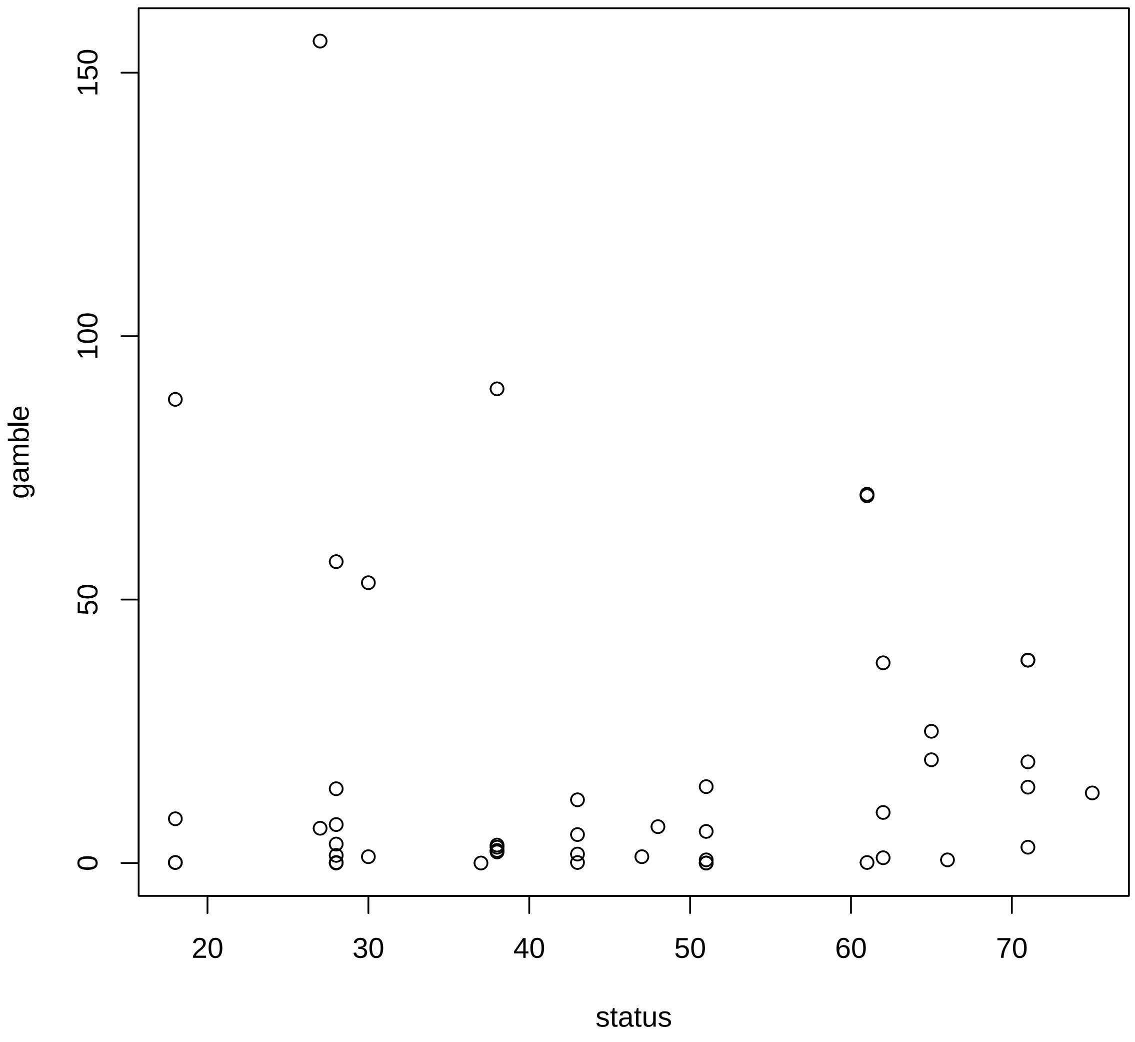
```
library(corrplot)

## Warning:  package 'corrplot' was built under R version 4.2.3
## corrplot 0.92 loaded
```

```
corrplot(cor(my_data[-1]), method = "number",
    type = "upper", tl.srt = 45, bg = "red",
    outline = TRUE, tl.col = "blue")
```


status
income
verbal
gamble
status
1.00
−0.28
0.53
−0.05
income
1.00
−0.18
0.62
verbal
1.00
−0.22
gamble
1.00
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1

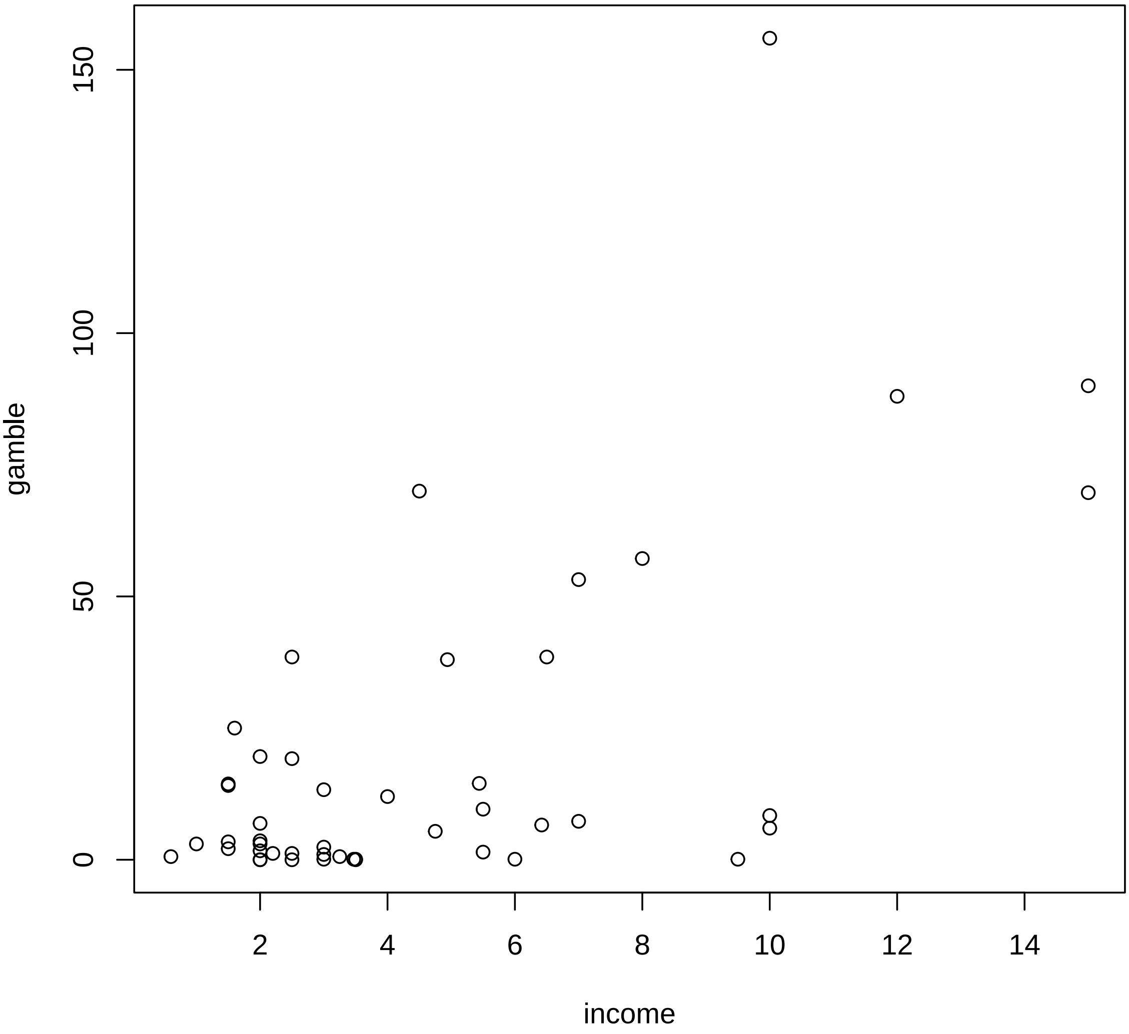
We see that **status** and **gamble** have a slightly negative correlation. Let's see their scatterplot.

```
plot(gamble ~ status, data = my_data)
```



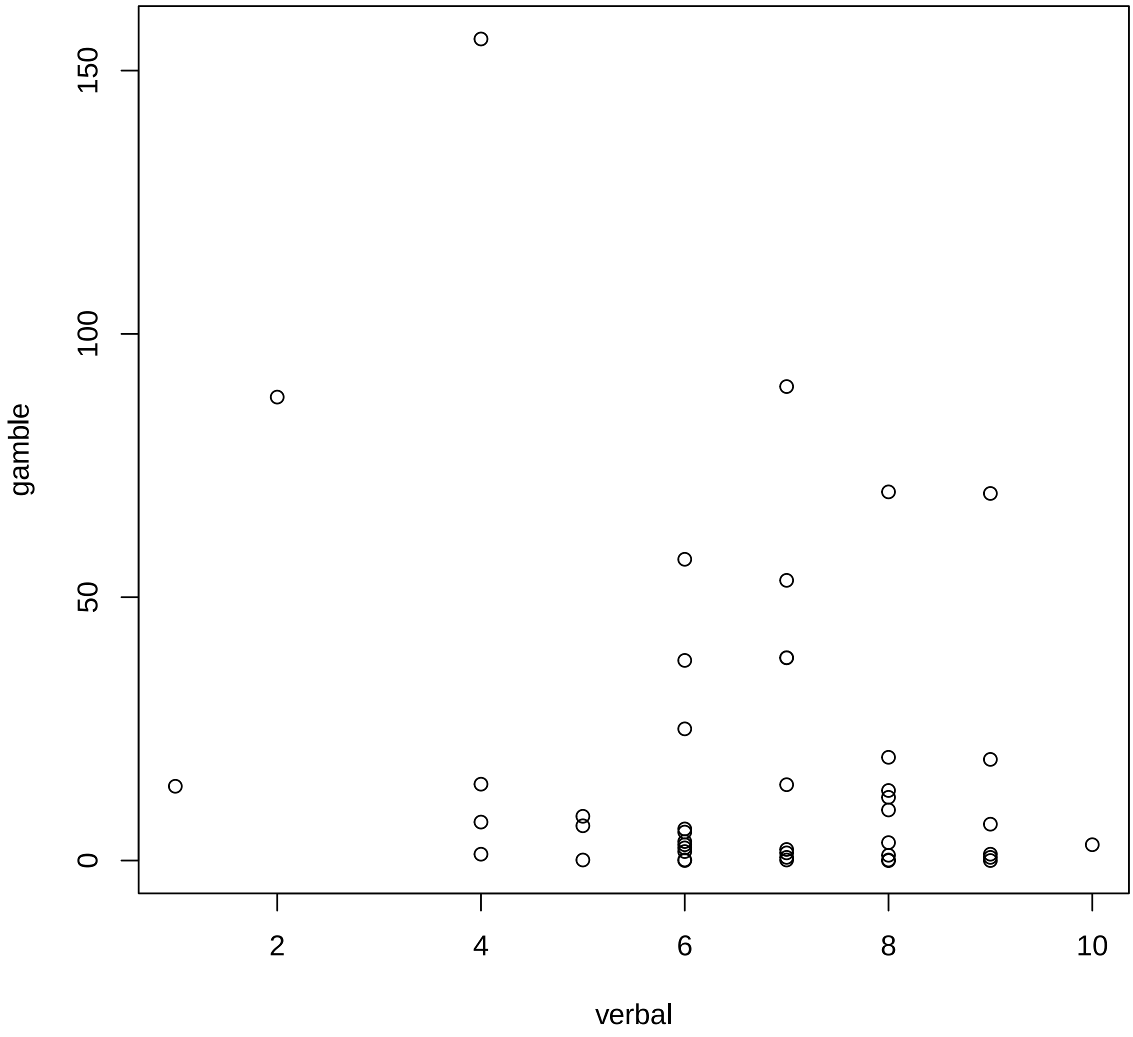
☞ We see that **income** and **gamble** have a correlation of 0.62, a pretty decent correlation, implying that, people with higher income spend more on gambling. Let's see their scatterplot.

```
plot(gamble ~ income, data = my_data)
```



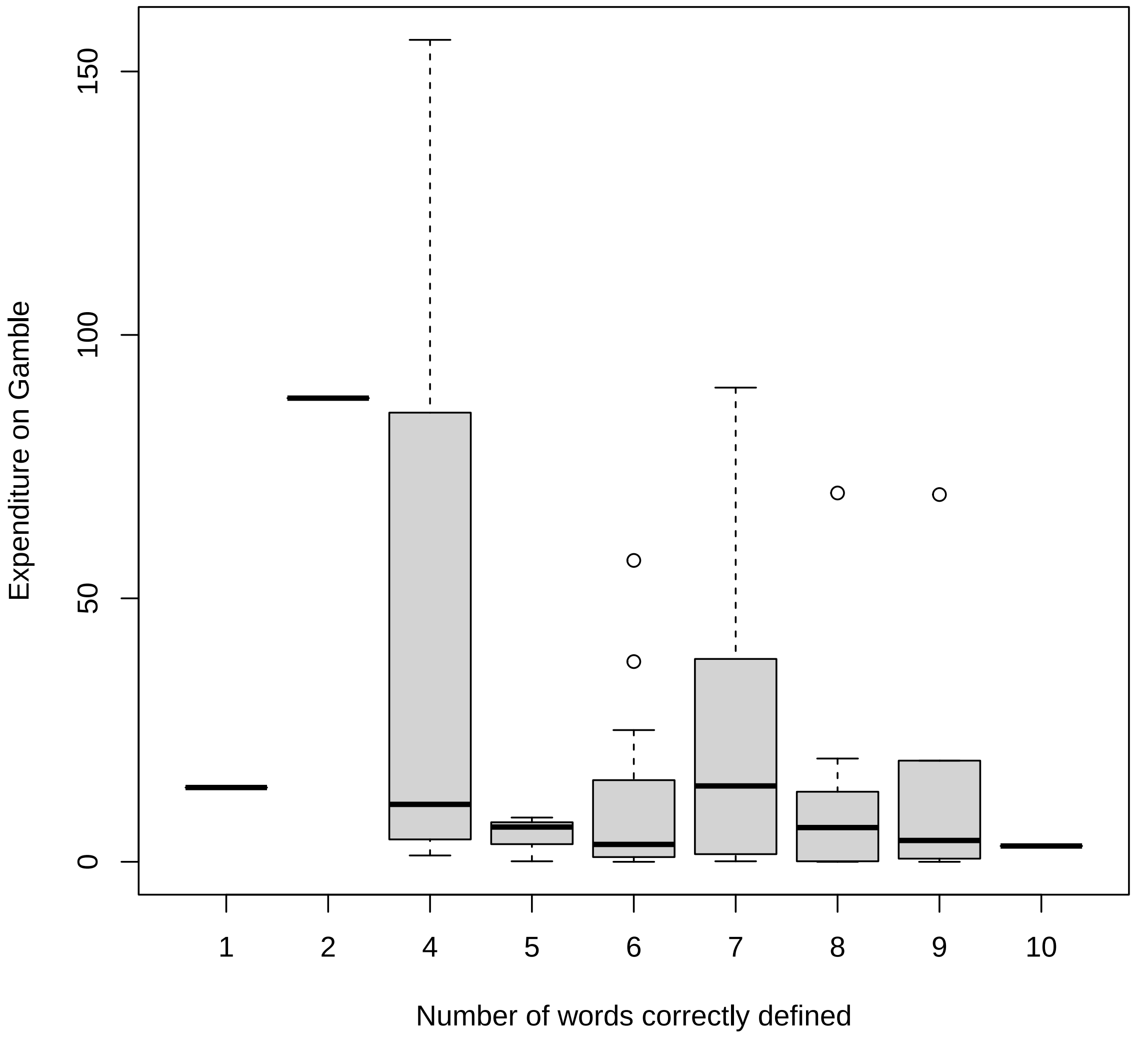
☞ **verbal** and **gamble** have a negative correlation. The variable *verbal* is the score in words out of 12 correctly defined. This may be interpreted as a measure of literacy. That means literacy and expenditure in gambling are negatively correlated. Literate people tend to spend less on gambling, while less literate people end up spending more on gambling. Let's see their scatterplot.

```
plot(gamble ~ verbal, data = my_data)
```



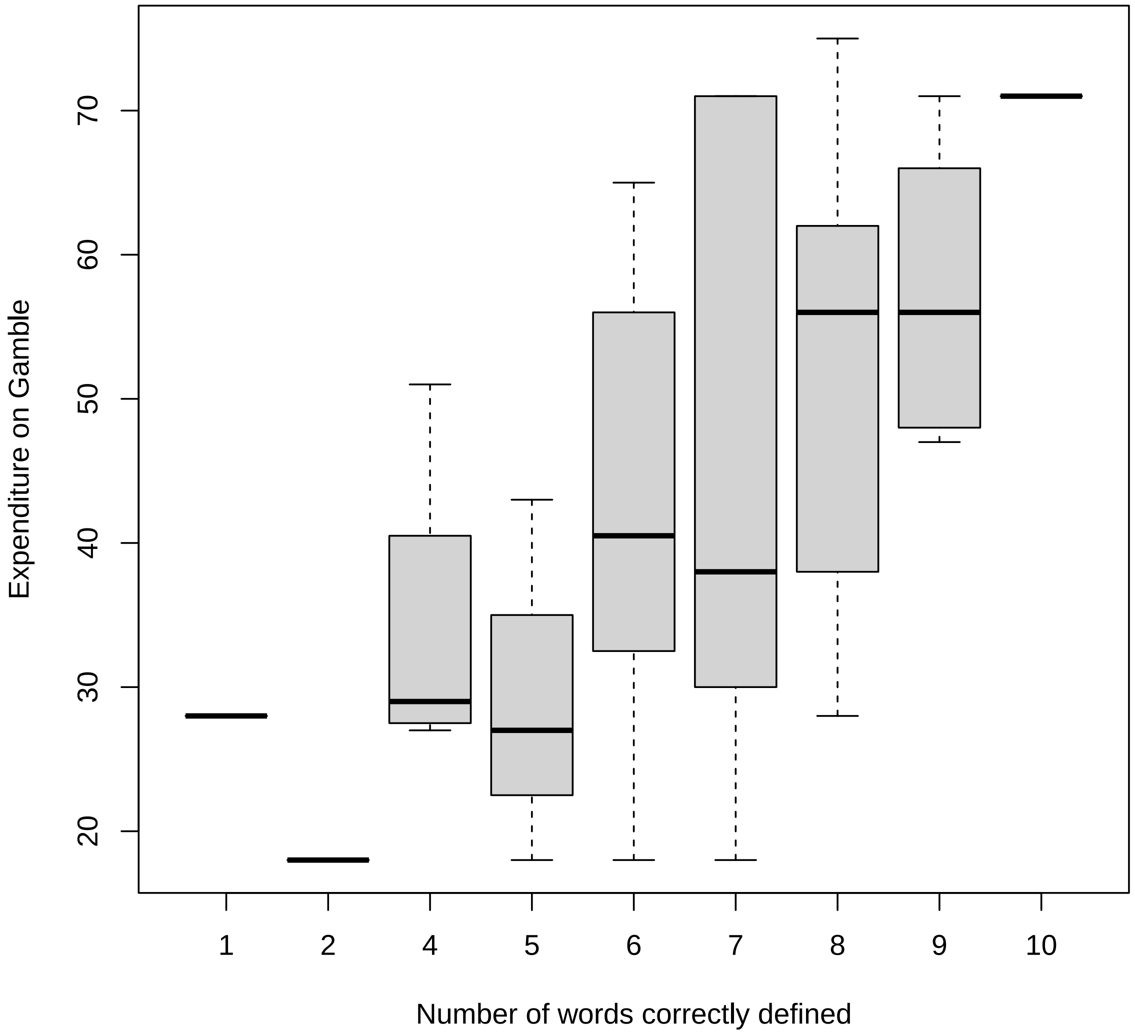
The variable *verbal* is a discrete valued variable. A box-plot may be helpful to see the relation between *verbal* and *gamble*.

```
boxplot(gamble ~ factor(verbal), data = my_data,
    xlab = "Number of words correctly defined",
    ylab = "Expenditure on Gamble")
```



We also see a positive correlation between *status* and *verbal*.

```
boxplot(status ~ factor(verbal), data = my_data,
    xlab = "Number of words correctly defined",
    ylab = "Expenditure on Gamble")
```



- Average spending on gambling by men is :

```
mean(my_data$gamble[which(my_data$sex ==
    "male")])

## [1] 29.775
```

- Average spending on gambling by women is :

```
mean(my_data$gamble[which(my_data$sex ==
    "female")])

## [1] 3.865789
```