Exercise 2 (R)

Analyse the esoph dataset. Can you derive some useful statements from it? Use data() to see all available datasets.

In [1]:

```
#Zuerst muss das esoph dataset geladen werden
data("esoph")

#Jetzt muss ich wissen was für einen Datentyp das Dataset hat
class(esoph)
```

'data.frame'

In [2]:

```
#Das esoph-Dataset ist ein data.frame. Ich kann dann die Größe/Dimension ermitteln.
dim(esoph) #Das esoph-Dataset hat 88 Zeilen und 5 Spalten
```

88 5

In [3]:

```
#Jetzt werden die Namen der Spalten angezeigt
colnames(esoph) #[1] "agegp"     "alcgp"     "tobgp"     "ncases"    "ncontrols"
```

'agegp' 'alcgp' 'tobgp' 'ncases' 'ncontrols'

Um mehr Informationen über die Daten zu haben, kann man auch die Funktion: str(esoph). Das Ergebnis zeigt, dass die zwei letzten Spalten numerische Werte haben, während die 3 ersten Spalten Strings enthalten.

In [14]:

```
str(esoph)
```

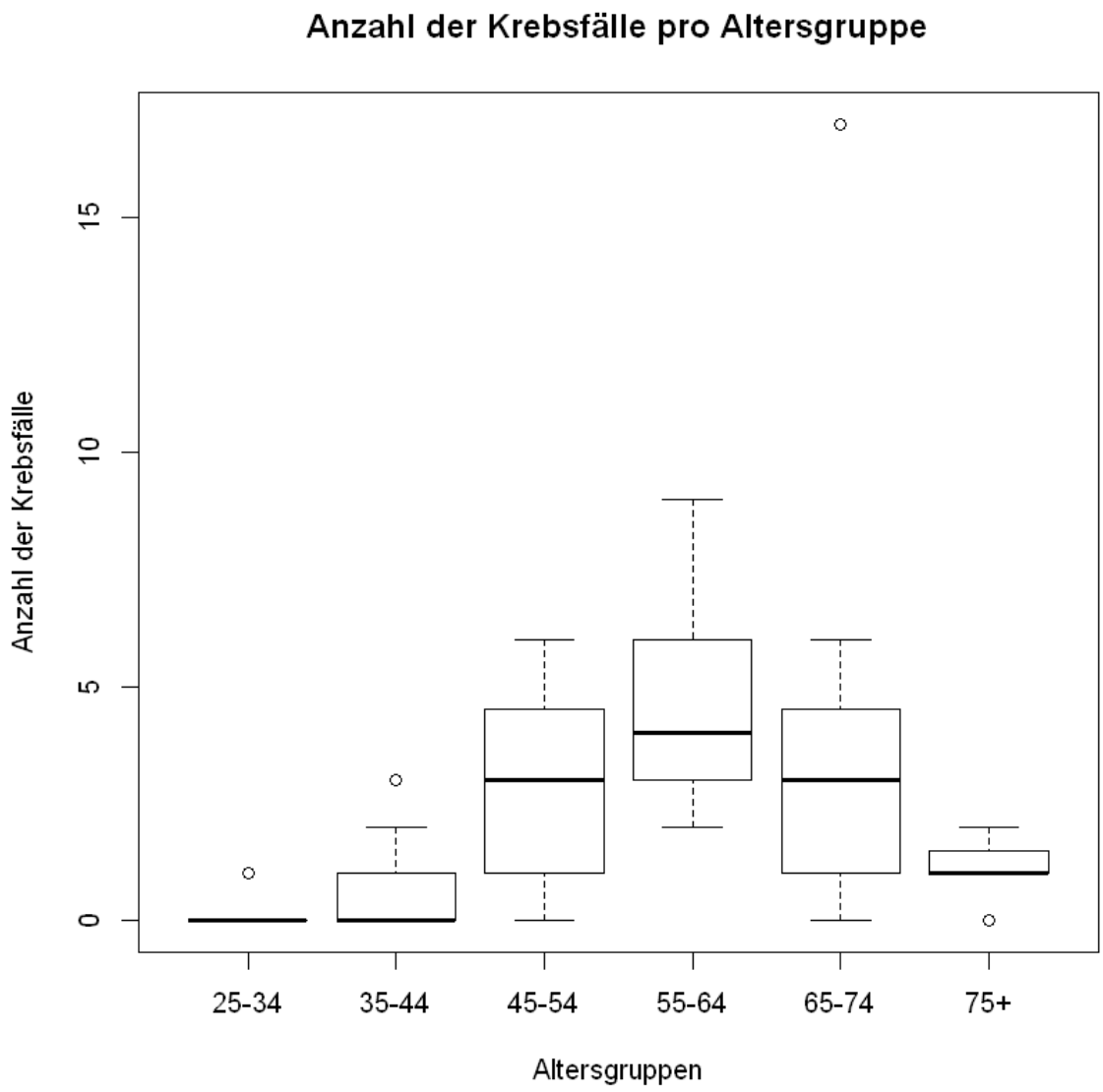
```
'data.frame':   88 obs. of  5 variables:
 $ agegp    : Ord.factor w/ 6 levels "25-34"<"35-44"<..: 1 1 1 1 1 1 1 1 1 1
...
 $ alcgp    : Ord.factor w/ 4 levels "0-39g/day"<"40-79"<..: 1 1 1 1 2 2 2 2
3 3 ...
 $ tobgp    : Ord.factor w/ 4 levels "0-9g/day"<"10-19"<..: 1 2 3 4 1 2 3 4
1 2 ...
 $ ncases   : num  0 0 0 0 0 0 0 0 0 0 ...
 $ ncontrols: num  40 10 6 5 27 7 4 7 2 1 ...
```

Ich kann die Anzahl der Krebsfälle pro Altersgruppe als Chart angeizen. Das Ergebnis zeigt, dass die meisten

Fälle in dem Bereich zwischen 45 und 74 Jahren zu finden sind.

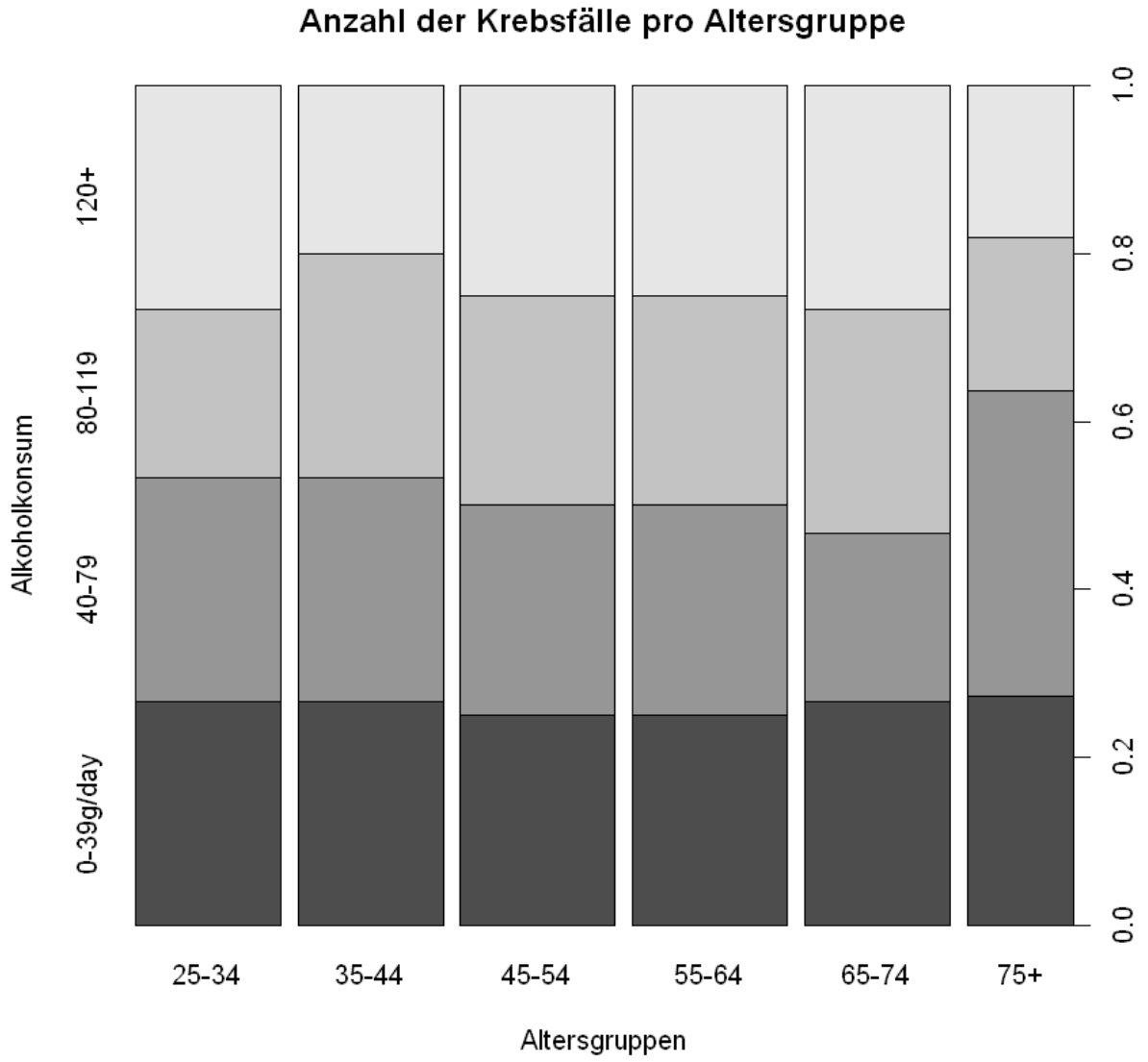
In [11]:

```
#Jetzt haben wir die Namen der Spalten und ich kann als Chart anzeigen wie viele
#Krebsfälle es pro Altersgruppe gibt.
plot(esoph$agegp, esoph$ncases, main="Anzahl der Krebsfälle pro Altersgruppe",
     xlab="Altersgruppen", ylab="Anzahl der Krebsfälle")
```



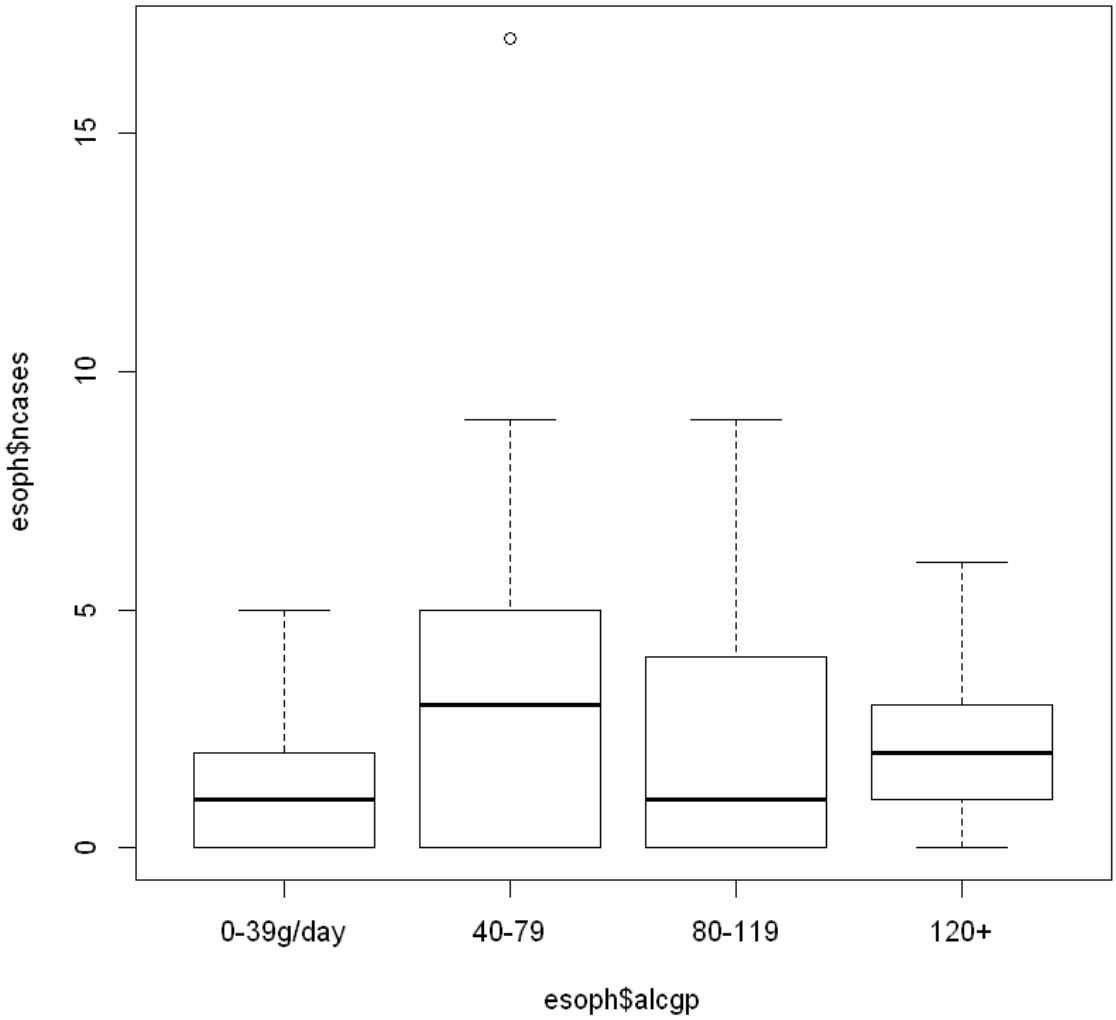
In [10]:

```
plot(esoph$agegp, esoph$alcgp, main="Anzahl der Krebsfälle pro Altersgruppe",
     xlab="Altersgruppen", ylab="Alkoholkonsum")
```



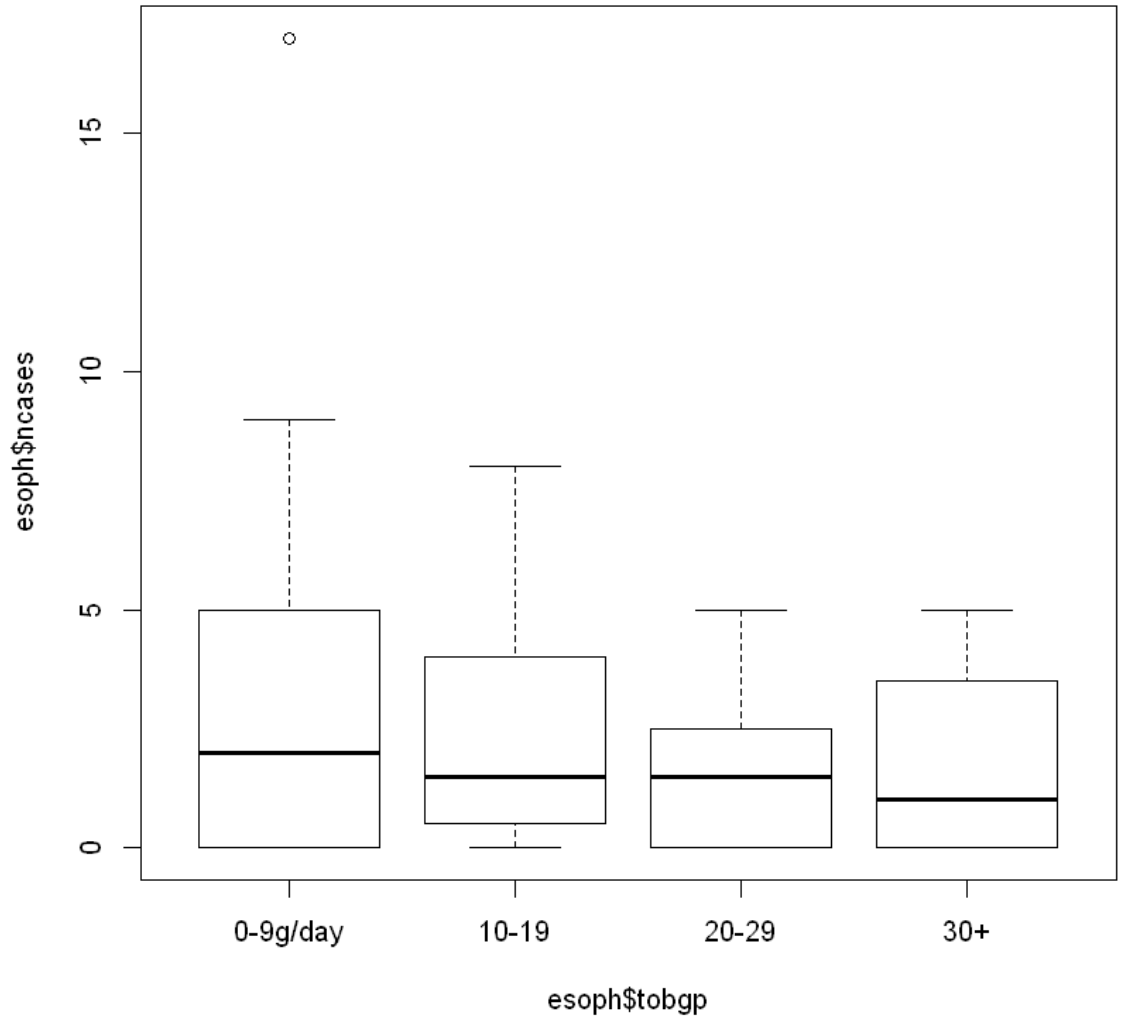
In [6]:

```
#Ich kann dann die Anzahl der Krebsfällen im Verhähltnis zu dem Alkoholkonsum
boxplot(esoph$ncases ~ esoph$alcgp)
```



In [7]:

```
1 boxplot(esoph$ncases ~ esoph$tobgp)
```



In R gibt es eine Funktion „summary()“, die statistischen Eigenschaften der Spalten berechnet.

In [13]:

```
1 summary(esoph)
```

```
  agegp          alcgp          tobgp        ncases          ncontrols     
 25-34:15   0-39g/day:23   0-9g/day:24   Min.   : 0.000   Min.   : 1.00  
 35-44:15   40-79    :23   10-19   :24   1st Qu.: 0.000   1st Qu.: 3.00  
 45-54:16   80-119   :21   20-29   :20   Median : 1.000   Median : 6.00  
 55-64:16   120+     :21   30+     :20   Mean   : 2.273   Mean   :11.08  
 65-74:15                                3rd Qu.: 4.000   3rd Qu.:14.00  
 75+  :11                                Max.   :17.000   Max.   :60.00  
```

Das Package Hmisc bringt die Funktion "describe()", die ein bisschen mehr Informationen liefert, als "summary()"

In [24]:

```
library(Hmisc)

describe(esoph)
```

```
esoph

 5  Variables      88  Observations
--------------------------------------------------------------------------------
agegp
       n  missing distinct
      88        0        6

Value      25-34 35-44 45-54 55-64 65-74   75+
Frequency     15    15    16    16    15    11
Proportion 0.170 0.170 0.182 0.182 0.170 0.125
--------------------------------------------------------------------------------
alcgp
       n  missing distinct
      88        0        4

Value      0-39g/day     40-79    80-119      120+
Frequency         23        23        21        21
Proportion     0.261     0.261     0.239     0.239
--------------------------------------------------------------------------------
tobgp
       n  missing distinct
      88        0        4

Value      0-9g/day    10-19    20-29      30+
Frequency        24       24       20       20
Proportion    0.273    0.273    0.227    0.227
--------------------------------------------------------------------------------
ncases
       n  missing distinct     Info     Mean      Gmd      .05      .10
      88        0       10    0.954    2.273    2.707      0.0      0.0
     .25      .50      .75      .90      .95
     0.0      1.0      4.0      5.3      6.0

Value          0     1     2     3     4     5     6     8     9    17
Frequency     29    16    11     9     8     6     5     1     2     1
Proportion 0.330 0.182 0.125 0.102 0.091 0.068 0.057 0.011 0.023 0.011
--------------------------------------------------------------------------------
ncontrols
       n  missing distinct     Info     Mean      Gmd      .05      .10
      88        0       30    0.994    11.08    12.23      1.0      1.0
     .25      .50      .75      .90      .95
     3.0      6.0     14.0     29.1     40.0

lowest :  1  2  3  4  5, highest: 40 46 48 49 60
--------------------------------------------------------------------------------
----
```