Schweizer Paraplegiker Forschung

Exercise: Data visualization

Workshop – Introduction to R

SOLUTION

1 Preparing the NHANES data

a) In this exercise, we will work with the `NHANES` data which is survey data collected by the US National Center for Health Statistics. It can be used by loading the `NHANES` package and then calling the data set by its name (`NHANES`). We first want to reduce the size of the data set. Run the commands below to prepare the data accordingly:

```
### Install and load NHANES package:
install.packages("NHANES")
library(NHANES)
```

```
### Store data under new name:
d <- NHANES
### Take only data from one year:
d <- d[d$SurveyYr=="2009_10", ]
### Remove duplicated rows:
d <- d[!duplicated(d),]
### Extract wanted variables:
d <- d[, c("HHIncomeMid", "BMI", "Age", "Gender",
           "Education", "HealthGen", "PhysActive")]
```

b) Take a look at the preprocessed data to get an impression of its variables. You can find an explanation of the variables' meaning on the help page (`?NHANES`).

```
head(d)
```

```
##   HHIncomeMid   BMI Age Gender    Education HealthGen PhysActive
## 1       30000 32.22  34   male  High School      Good         No
## 4       22500 15.30   4   male         <NA>      <NA>       <NA>
## 5       40000 30.57  49 female Some College      Good         No
## 6       87500 16.82   9   male         <NA>      <NA>       <NA>
## 7       60000 20.64   8   male         <NA>      <NA>       <NA>
## 8       87500 27.24  45 female College Grad     Vgood        Yes
```

```
str(d)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame': 3568 obs. of  7 variables:
##  $ HHIncomeMid: int  30000 22500 40000 87500 60000 87500 30000 100000 70000 NA ...
##  $ BMI        : num  32.2 15.3 30.6 16.8 20.6 ...
##  $ Age        : int  34 4 49 9 8 45 66 58 54 10 ...
##  $ Gender     : Factor w/ 2 levels "female","male": 2 2 1 2 2 1 2 2 2 1 ...
##  $ Education  : Factor w/ 5 levels "8th Grade","9 - 11th Grade",..: 3 NA 4 NA NA 5 4 5 2 NA ...
##  $ HealthGen  : Factor w/ 5 levels "Excellent","Vgood",..: 3 NA 3 NA NA 2 2 2 4 NA ...
##  $ PhysActive : Factor w/ 2 levels "No","Yes": 1 NA 1 NA NA 2 2 2 2 NA ...
```

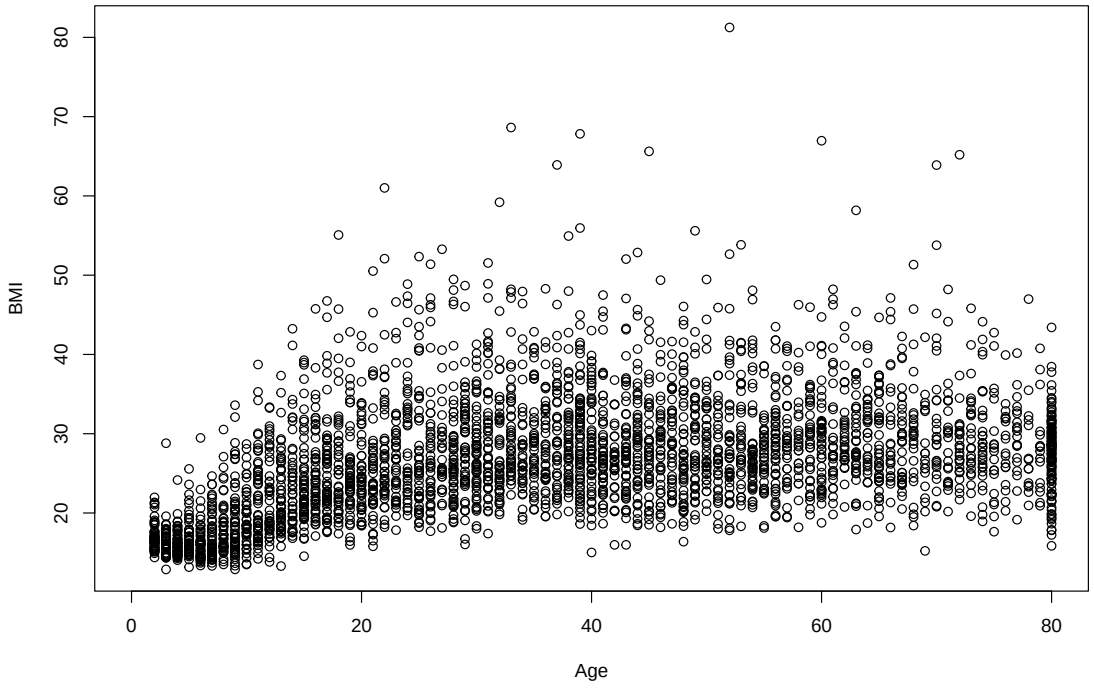
```
summary(d)
```

```
##   HHIncomeMid          BMI              Age           Gender
##  Min.   :  2500   Min.   :12.88   Min.   : 0.00   female:1799
##  1st Qu.: 22500   1st Qu.:21.42   1st Qu.:16.00   male  :1769
##  Median : 50000   Median :26.05   Median :35.00
##  Mean   : 54419   Mean   :26.76   Mean   :35.79
##  3rd Qu.: 87500   3rd Qu.:30.94   3rd Qu.:54.00
##  Max.   :100000   Max.   :81.25   Max.   :80.00
##  NA's   :322      NA's   :168
##           Education        HealthGen    PhysActive
##  8th Grade     : 196   Excellent: 268   No  :1344
##  9 - 11th Grade: 356   Vgood    : 797   Yes :1543
##  High School   : 572   Good     :1025   NA's: 681
##  Some College  : 728   Fair     : 401
##  College Grad  : 622   Poor     :  83
##  NA's          :1094   NA's     : 994
##
```

2 The `plot` function

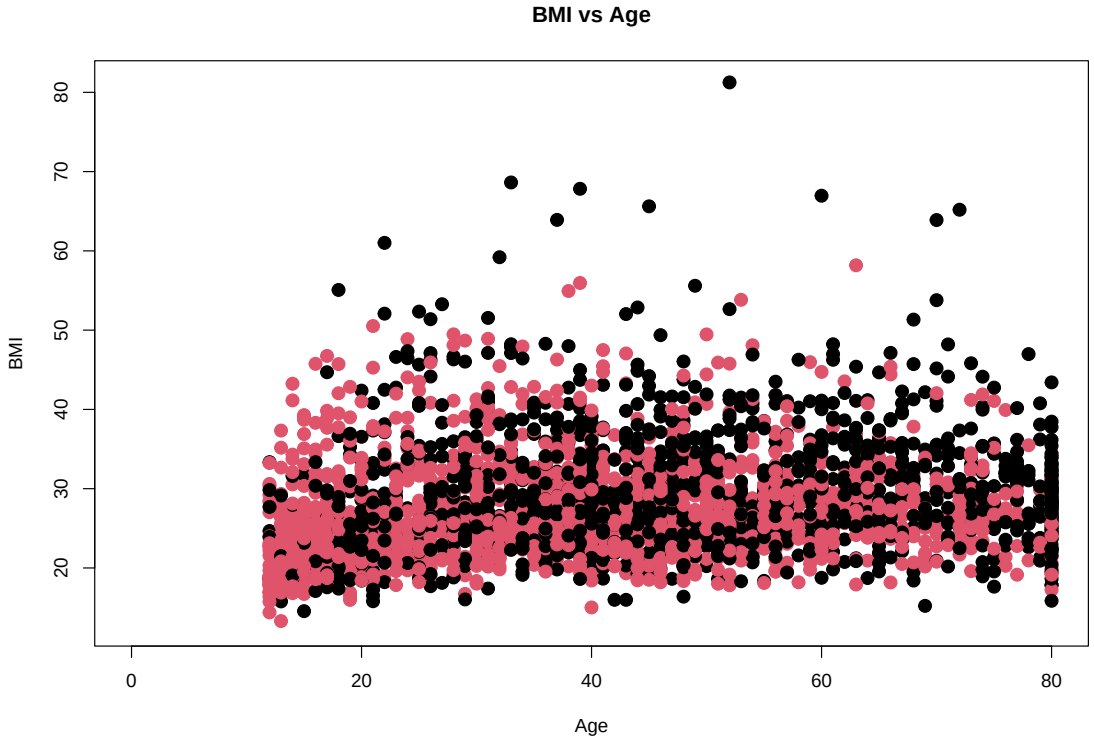
a) Create a simple scatter plot with the prepared `NHANES` data in which you plot the age of participants on the x-axis vs. their BMI on the y-axis.

```
plot(BMI ~ Age, data=d)
```



b) Extend the plot command with further arguments to change the appearance of the figure. Add a title, change the used plotting symbol. Additionally, color the points according to whether participants are physically active or not (`PhysActive` variable).

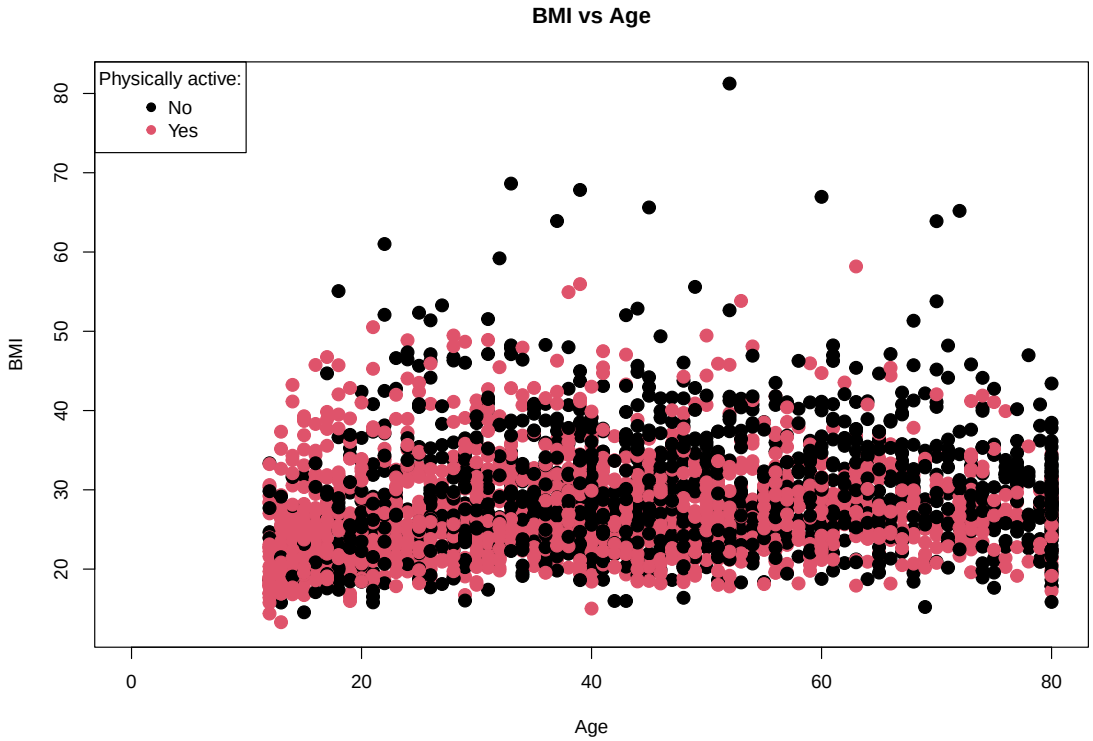
```
plot(BMI ~ Age, data=d, col=d$PhysActive, pch=19, cex=1.5, main="BMI vs Age")
```



Sidenote: If we compare this figure with the figure from the previous exercise we can see that the points for people below an age of 18 are missing. The reason is that the PhysActive variable was only collected from the age of 18 and, therefore, has a missing value for younger people. Using a missing value as the value for the col argument in the plot function will result in the point not having any color, thus not being visible in the final figure. We could create our own version of the PhysActive variable in which we add a third category for people under 18 so that they would get their own color in the plot.

c) **Extra:** The points are now colored according to the physical activity of the people. More specifically, the points are colored according to the level numbers of the `PhysActive` factor (see slides). We can use the `legend` function to manually add a legend to the plot. In the `legend` function we can specify the legend text with the `legend` argument and the legend symbols and colors with the `pch` and `col` arguments, respectively. The commands below create the figure again and add a legend with the correct color assignment. Run the code and try to understand the individual commands inside the function call.

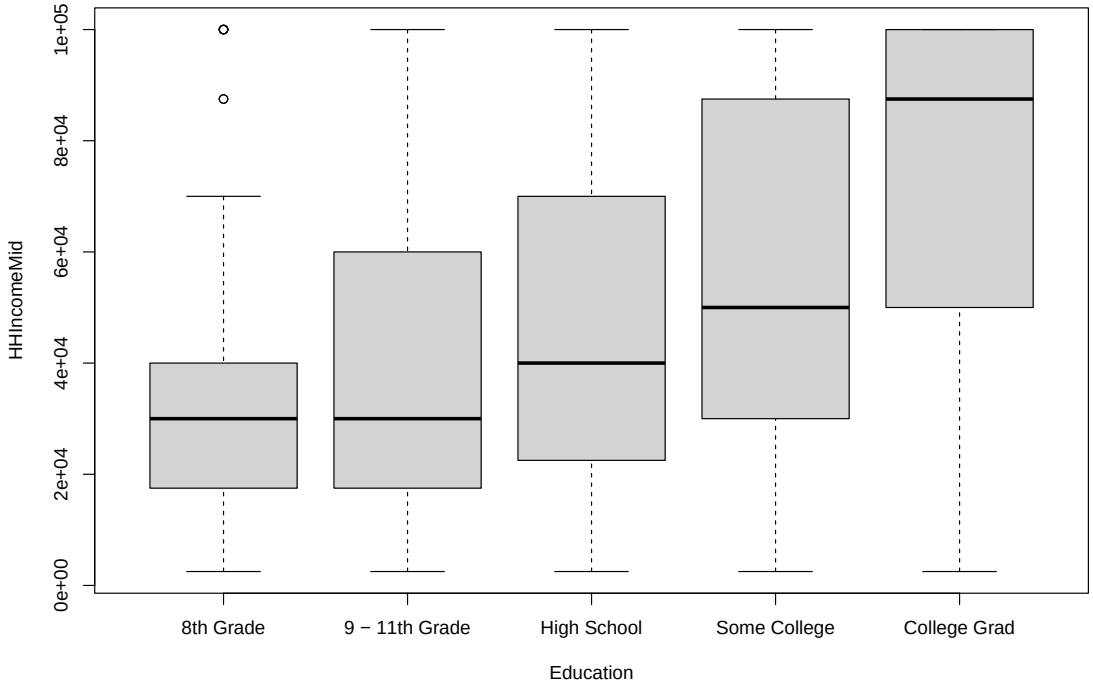
```
plot(BMI ~ Age, data=d, col=d$PhysActive, pch=19, cex=1.5, main="BMI vs Age")
legend('topleft',
       title = 'Physically active:',
       legend = levels(d$PhysActive),
       col = 1:nlevels(d$PhysActive),
       pch=19)
```



3 Creating a boxplot

a) Using the `NHANES` data, create a boxplot in which you plot the distribution of the income (`HHIncomeMid` variable) versus the level of education (`Education` factor). Does the figure suggest a relation between the two variables?

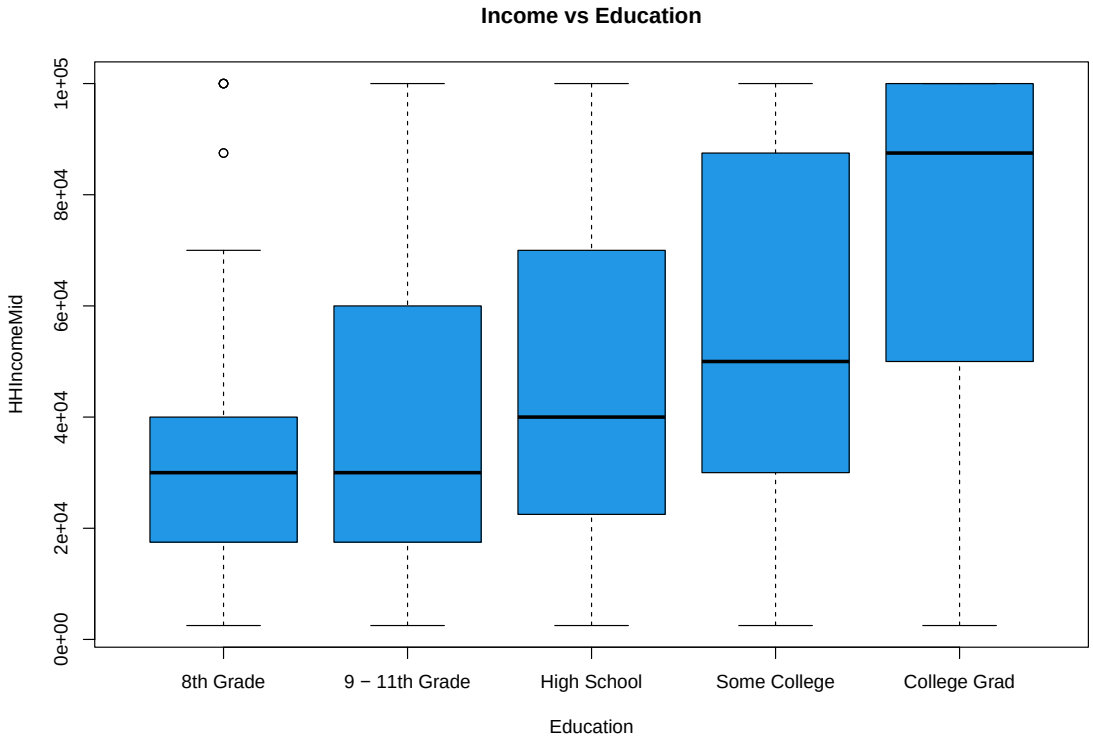
```
boxplot(HHIncomeMid ~ Education, data = d)
```



The figure indicates a strong positive relation between the level of education and the income.

b) We can again use additional arguments to adapt the appearance of the figure. Try for example to change the color of the boxplots and add a title.

```
boxplot(HHIncomeMid ~ Education, data = d, col=4, main='Income vs Education')
```



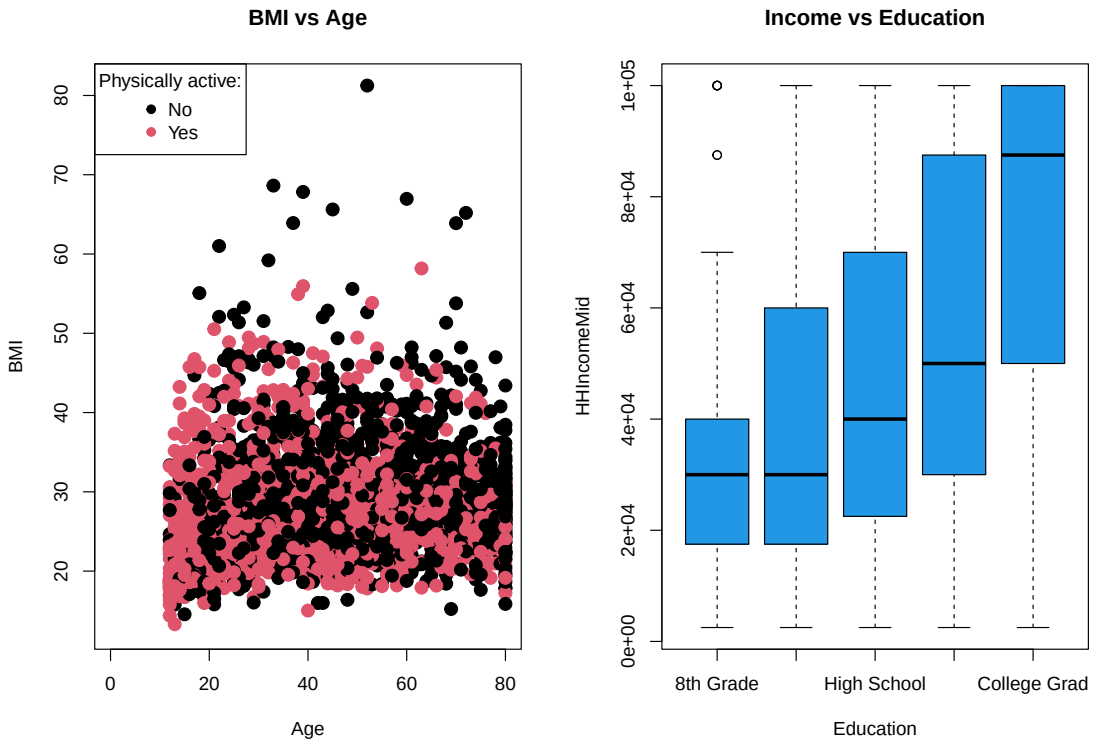
4 Setting graphical parameters

a) In R, we can change graphical parameters to further control R's behavior when creating figures. For example, the `mfrow` parameter determines how many figures are printed inside a plot window. Normally, it is set to a value of `c(1,1)` which means that in a plot window there will be 1 row and 1 column, resulting in only one figure in the window. To plot multiple figures next to each other we can change the value of the `mfrow` parameter. To do this, we have to use the `par` function which controls all graphical parameters (see `?par`). With the following command we can change the `mfrow` value to 1 row and 2 columns:

```
par(mfrow=c(1, 2))
```

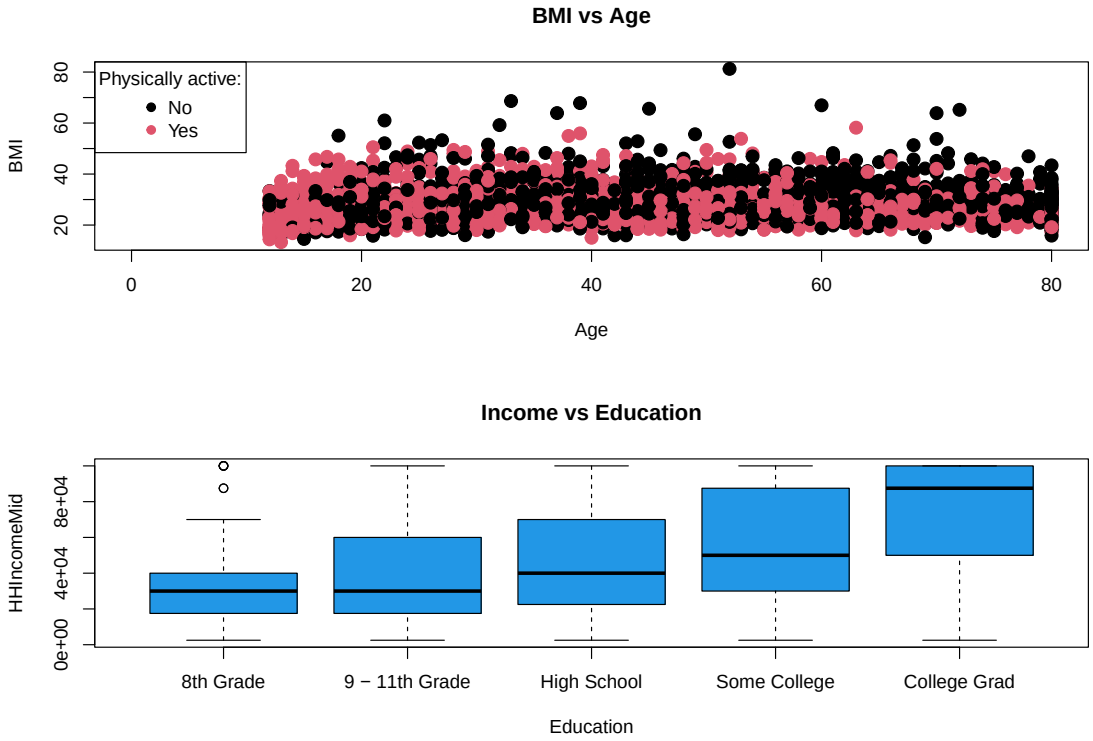
Run the command and create again the scatter plot and the boxplot from the previous exercise. How are they now placed inside the plot window? Try out different values for the `mfrow` parameter and observe its effect.

```
par(mfrow=c(1, 2))
### Scatter plot:
plot(BMI ~ Age, data=d, col=d$PhysActive, pch=19, cex=1.5, main="BMI vs Age")
legend('topleft',
       title = 'Physically active:',
       legend = levels(d$PhysActive),
       col = 1:nlevels(d$PhysActive),
       pch=19)
### Boxplot:
boxplot(HHIncomeMid ~ Education, data = d, col=4, main='Income vs Education')
```



Plotting one figure above other:

```
par(mfrow=c(2, 1))
### Scatter plot:
plot(BMI ~ Age, data=d, col=d$PhysActive, pch=19, cex=1.5, main="BMI vs Age")
legend('topleft',
       title = 'Physically active:',
       legend = levels(d$PhysActive),
       col = 1:nlevels(d$PhysActive),
       pch=19)
### Boxplot:
boxplot(HHIncomeMid ~ Education, data = d, col=4, main='Income vs Education')
```



In the end, one might want to reset `mfrow` to its original value:

```
par(mfrow=c(1, 1))
```

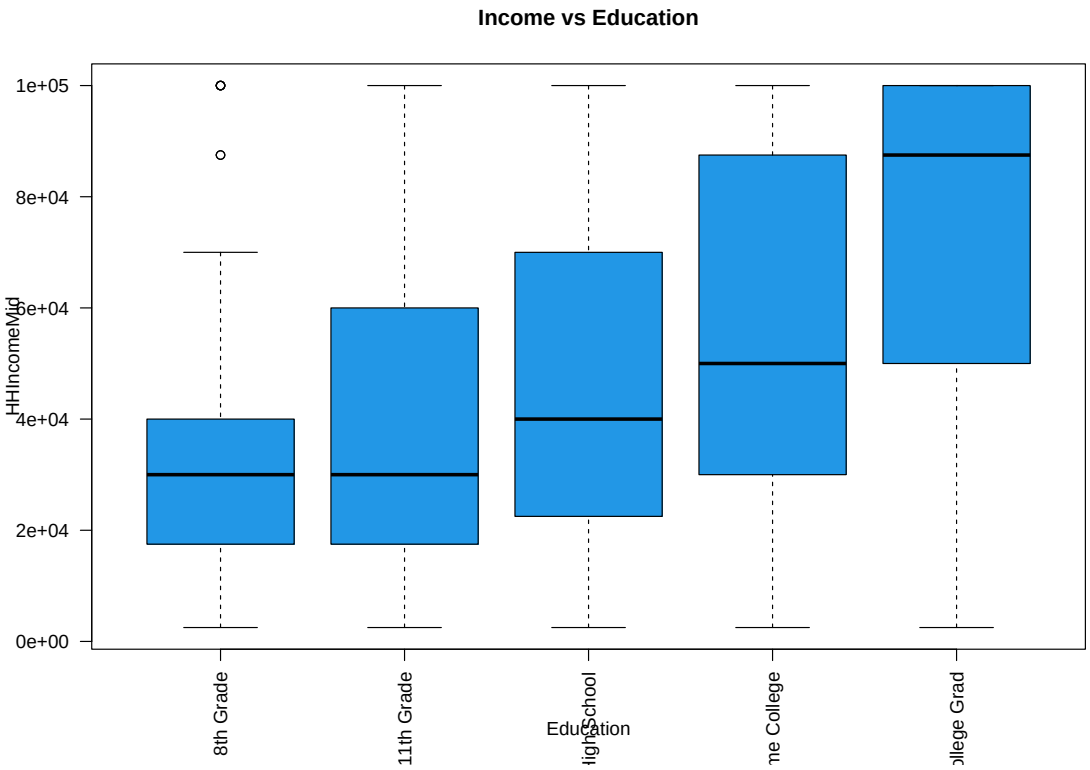
b) **Extra:** Another graphical parameter that is often used is the `mar` parameter. It defines the sizes of the margins around a figure. With the following code we can check its current value:

```
par()$mar

## [1] 5.1 4.1 4.1 2.1
```

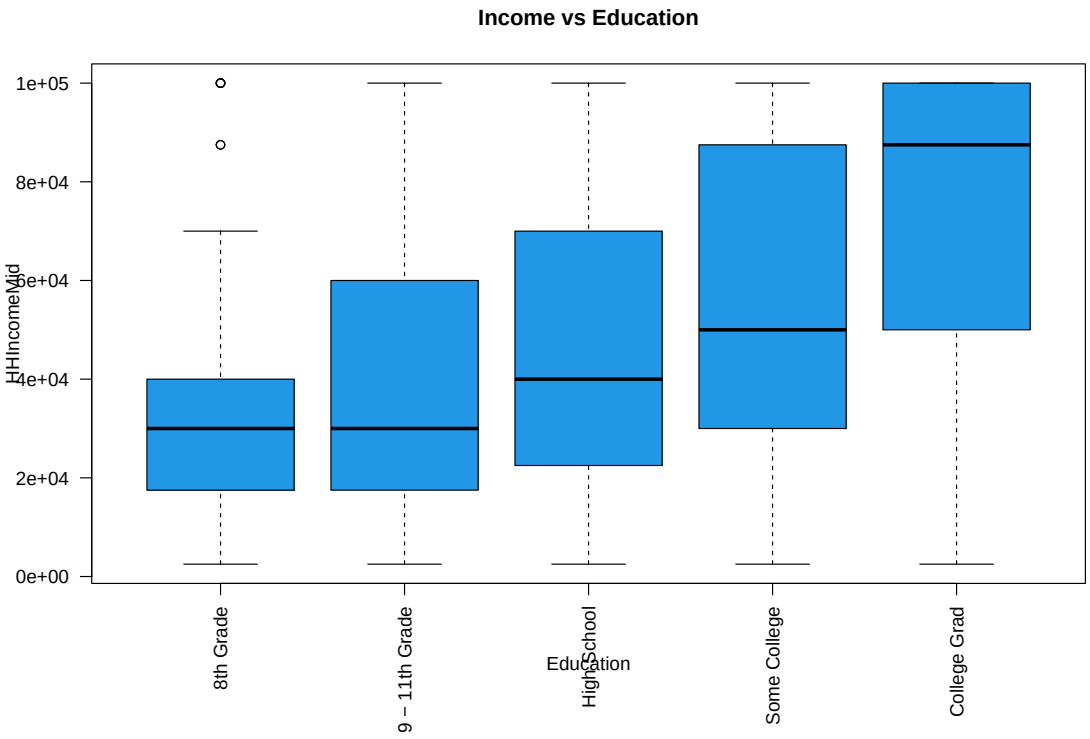
As can be read in the `par` helppage (`?par`) `mar` determines the sizes of the bottom, left, top and right figure margins (in this order). This can for example be useful when we have text in a figure that is too large for the figure margins. The following command creates again the boxplot from the previous exercise but uses the `las` argument to rotate the axis labels by 90 degrees (sometimes necessary with very large labels). Run the command and look at the resulting figure.

```
plot(HHIncomeMid ~ Education, data = d, col=4, main='Income vs Education', las=2)
```



As we can see, some of the labels are cut of by the bottom margin (and they also overlay with the axis titles, but we'll ignore this here). Try to change the `mar` parameter to increase the size of the bottom margin, making all axis labels readable.

```
par(mar=c(8, 4.1, 4.1, 2.1))
plot(HHIncomeMid ~ Education, data = d, col=4, main='Income vs Education', las=2)
```




```
par(mar=c(5.1, 4.1, 4.1, 2.1))   # Set back to original value
```

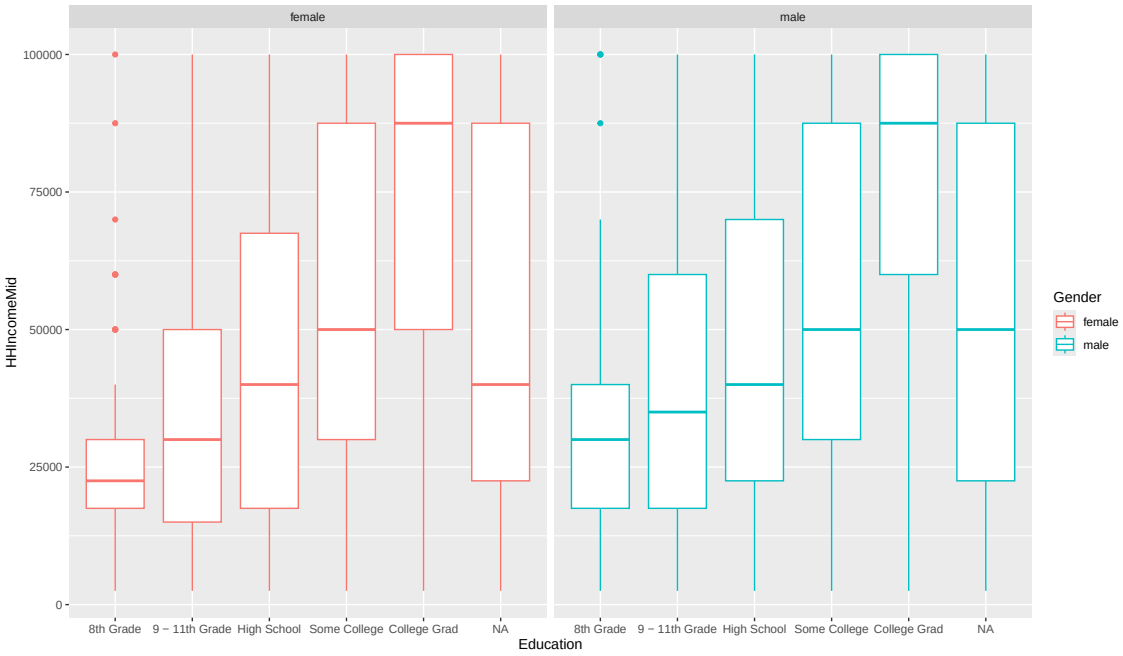
All labels are now readable.

5 Plotting multiple settings

a) As we saw in the lecture, `ggplot2` is a handy R package to create complex figures. Try to use the `ggplot2` package to create again boxplots of the income vs the education level but separated into two figures, one only showing data from men and the other showing data from women.

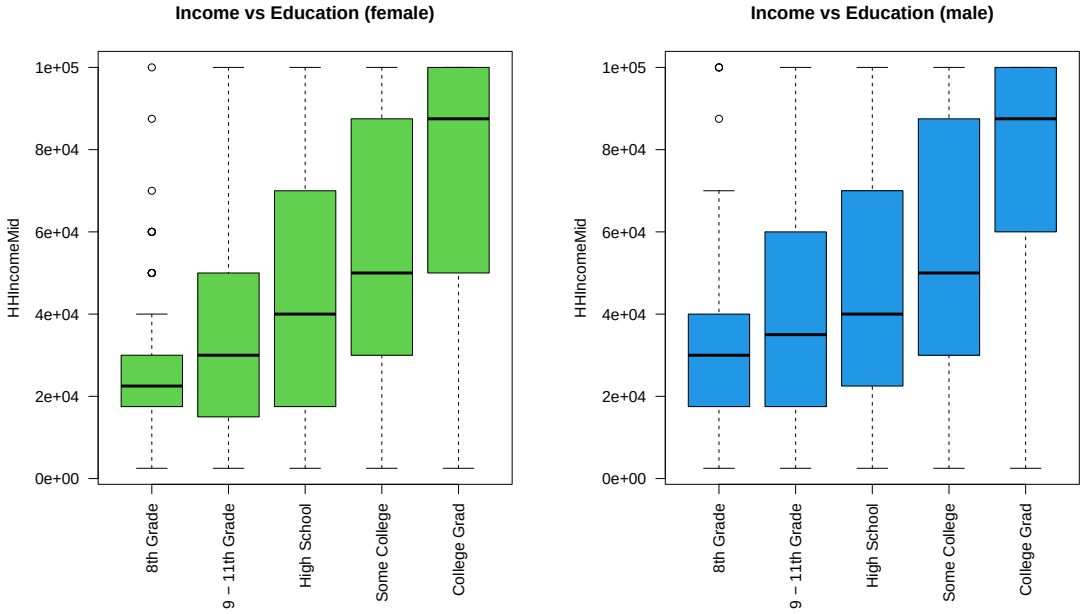
```
library(ggplot2)
ggplot(data = d, mapping = aes(x = Education, y=HHIncomeMid, colour = Gender))+
  facet_wrap(~Gender) +
  geom_boxplot()

## Warning:  Removed 322 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```



b) **Extra:** Although a bit more effortful, we now have all the tools to create such a figure ourselves without the help of `ggplot2`. Try to recreate the figure by 1) Splitting the data into two data frames, one containing the men and one the women data and 2) use the `plot` function together with the `mfrow` argument to plot the two figures next to each other like in the ggplot version.

```
### Split data:
dfm <- d[d$Gender=='female', ]
dm <- d[d$Gender=='male',]
### Create plot:
par(mfrow=c(1,2), mar=c(8, 6, 4.1, 2.1))
par(mgp=c(4, 1, 0))    # mgp can be used to change the
                       # position of the overal axis labels (to prevent overlap)
plot(HHIncomeMid ~ Education, data = dfm, col=3,
     main='Income vs Education (female)',
     las=2, xlab='')
plot(HHIncomeMid ~ Education, data = dm, col=4,
     main='Income vs Education (male)',
     las=2, xlab='')
```



One thing that is different compared to the ggplot version is that ggplot automatically created a boxplot for the values of people with a missing education value (boxplot at the most right side).