Exercise 1 (Univariate descriptives)

In this exercise we consider a slightly modified version of the same dataset as last week. It contains a survey of school children and it's stored in CSV format (`survey.csv`). The dataset can be downloaded from the course webpage.

(a) Read in the data (**R-Hint**: Use `read.table(..., sep=";", header=TRUE)` to read in your file. `getwd()` shows you the currect working directory, where R searchs for the file. With `setwd()` you can change this directory. Alternatively you can specify the complete path to your file in the `read.table()` function.).

```
# read in the data. Paste0 combines the string stored in "dir"
# with "data/survey" to get the  complete path to the file
dat <- read.table(file = paste0(dir,"data/survey.csv"),
                  sep = ";", header = TRUE)
```

(b) To gain an overview over the data calculate some characteristic measures of the distribution:

- Determine the mean and the median of `Arm.span` (**R-Hint**: `mean(), median()`).

```
# mean
mean(dat$Arm.span)

## [1] 178.8333

#median
median(dat$Arm.span)

## [1] 180
```

- Determine the 10% quantile of `Arm.span` (**R-Hint**: `quantile()`).

```
# 10-quantile
quantile(dat$Arm.span, probs = c(0.1))

##   10%
## 171.3
```

- Calculate the range, variance, standard deviation and interquartile range of `Arm.span` (**R-Hint**: `range(), var(), sd(), IQR()`).

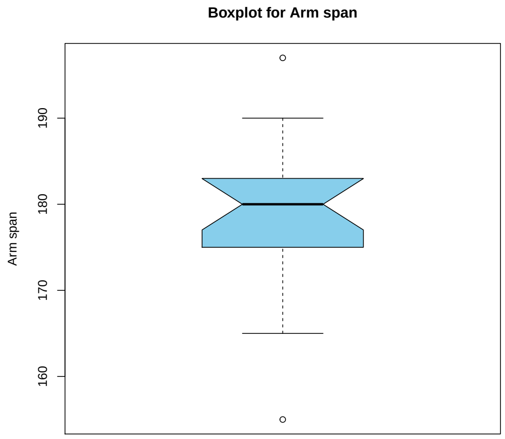
```
# range
range(dat$Arm.span)

## [1] 155 197
```

```
# Variance
var(dat$Arm.span)

## [1] 81.20588

# Standard deviation
sqrt(var(dat$Arm.span))

## [1] 9.011431

# or sd(dat$Arm.span)

# IQR
IQR(dat$Arm.span)

## [1] 7.5
```

(c) Now we would like to visualize the distribution of the variable `Arm.span`:

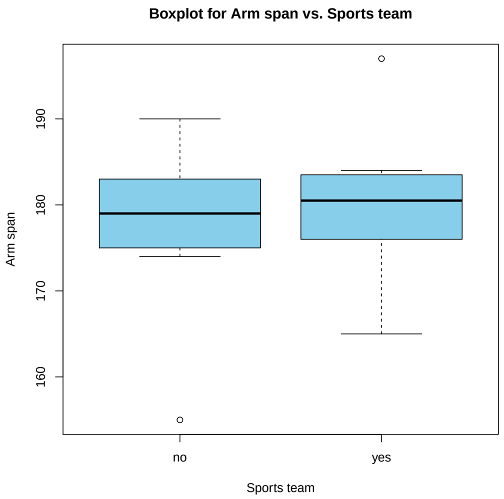
- Visualize the data as a boxplot and add the notches (**R-Hint**: `boxplot(...,notch=TRUE)`).

```
boxplot(dat$Arm.span, notch=TRUE,
        main="Boxplot for Arm span",
        ylab="Arm span",
        col='skyblue')
```



- Visualize the difference between students who take part in a sports team (`Sports.team`) and those who don't (**R-Hint**: `boxplot(Arm.span ~ ...)`).

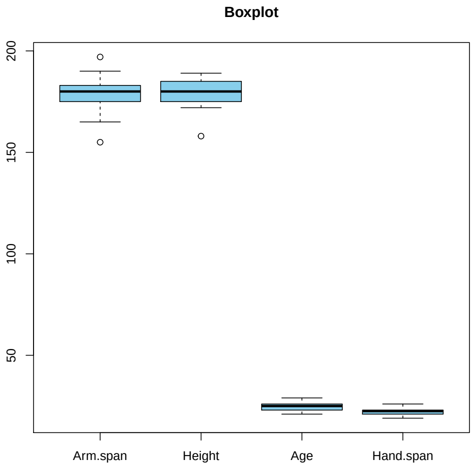
```
boxplot(Arm.span~Sports.team,
        main="Boxplot for Arm span vs. Sports team",
        xlab="Sports team",
        ylab="Arm span",
        col='skyblue',
        data=dat)
```



```
# The medians are quite similar between the two groups, however
# the range of the whiskers is different.
```

- Visualize the four variables `Arm.span`, `Height`, `Age`, `Hand.span` in one boxplot. Does this visualization make sense?
(**R-Hint**: `boxplot(dat[,c("Arm.span","Height","Age","Hand.span")])`)

```
boxplot(dat[,c("Arm.span","Height","Age","Hand.span")],
        main="Boxplot",
        col='skyblue')
```



```
# It doesn't make sense to plot the variables together. They are
# on different scales (age in years, height in cm, etc)!
```

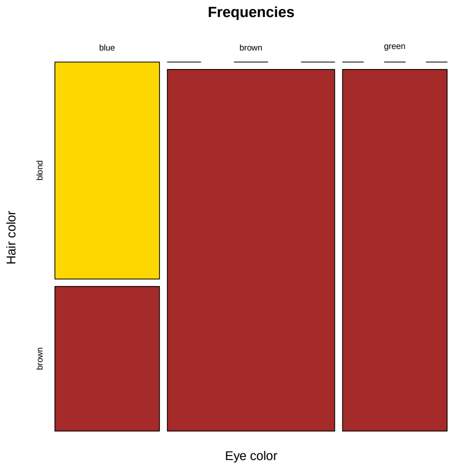
(d) Now, we want to compare two variables to each other:

- Determine the contingency table between `Eye.color` and `Hair.color` (**R-Hint**: `table()` or `xtabs( ...)`).

```
table(dat$Eye.color, dat$Hair.color)

##
##          blond brown
##   blue       3     2
##   brown      0     8
##   green      0     5
```

- Display the frequencies of the contingency table as mosaic plot (**R-Hint**: `mosaicplot()`). What do you observe?

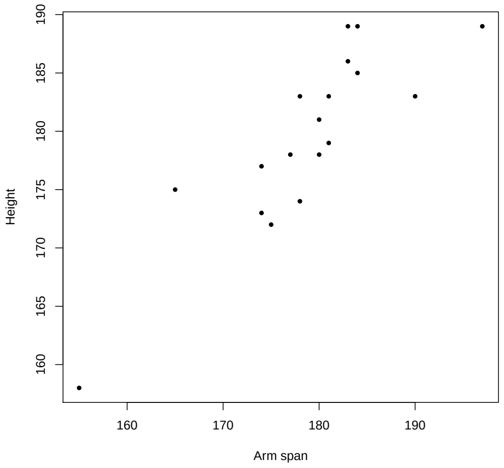
```
mosaicplot(table(dat$Eye.color, dat$Hair.color),
           col=c('gold','brown'),
           xlab="Eye color",
           ylab="Hair color",
           main='Frequencies')
```



```
# All blondes in the dataset have blue eyes while brown haired
# people have blue, brown and green eyes.
```

- Visualize the relation between `Arm.span` and `Height` in a scatterplot (**R-Hint**: `plot()`). What do you observe?

```
plot(dat$Arm.span, dat$Height,
     xlab='Arm span',
     ylab="Height",
     pch=20)
```



```
# There seems to be a relationship between height and arm span.
# The larger the arm span, the taller the student.
```

Exercise 2 (Descriptives)

The dataset of this exercise is from a study on guinea pigs. The study investigates the effects of Vitamin C consumption on the length of the teeth growth. Therefore, the guinea pigs were fed by orange juice (`OJ`) or ascorbic acid (`VC`) using different doses of Vitamin C (0.5, 1.0, 2.0). The data contains the following variables:

`len`	mean of teeth length
`supp`	supplement type (OJ or VC)
`dose`	vitamin C dose in mg

In order to access the data, you can use the following code:

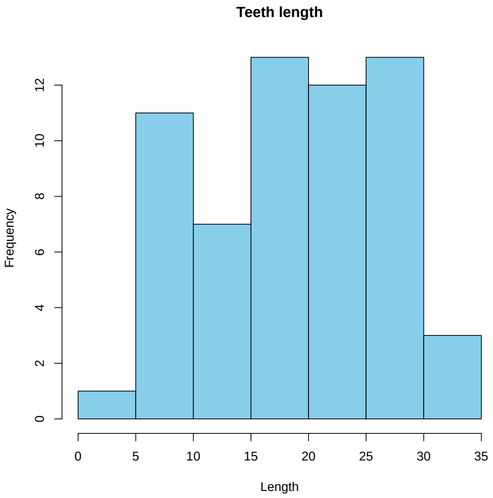
```
# The data is contained in the R package datasets. With data(),
# the data is loaded into the workspace.
data("ToothGrowth")

# then we can rename the dataset and store it in dat
# (easier for coding purposes)
dat <- ToothGrowth

# Consider the first few lines of dat
head(dat)

##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
## 4  5.8   VC  0.5
## 5  6.4   VC  0.5
## 6 10.0   VC  0.5
```

(a) How many guinea pigs have been included into this study?

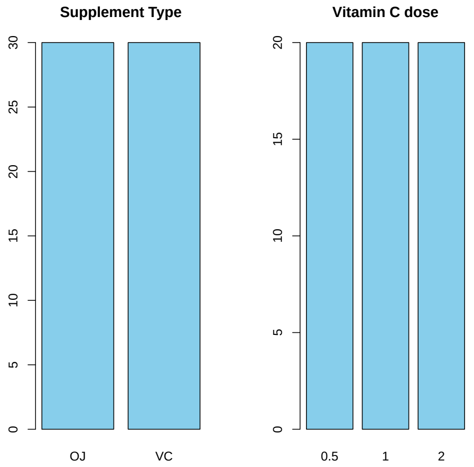
```
# Each row contains the information of one guinea pig since there are
# 60 rows, 60 guinea pigs have been included
dim(dat)

## [1] 60  3
```

(b) Investigate the three variables of the dataset by appropriate plots (**R-Hint**: `hist()`, `barplot()`).

```r
# Since the teeth length is a continuous measure we use a histogram
hist(dat$len,
     col="skyblue",
     xlab='Length',
     main='Teeth length')
```



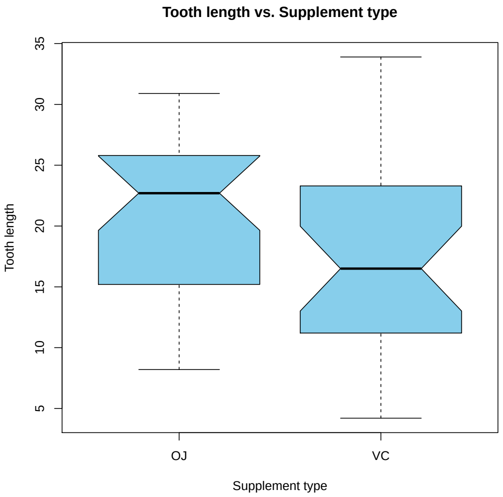
```r
# The variables supp and dose are categorical, i.e. we can  visualize them
# using barplots (We first have to calculate the frequencies with table()).
par(mfrow=c(1,2)) # we want two plots next to each other
barplot(table(dat$supp),
        main='Supplement Type',
        col="skyblue")
barplot(table(dat$dose),
        main='Vitamin C dose',
        col="skyblue")
```



```
# We see that 30 guines pigs were fed with OJ, 30 with VC
# and each dose was given to 20 guinea pigs.
```

(c) Does the distribution of the tooth length depend on the supplement type? Illustrate your answer with an appropriate plot (**R-Hint**: `boxplot()`).

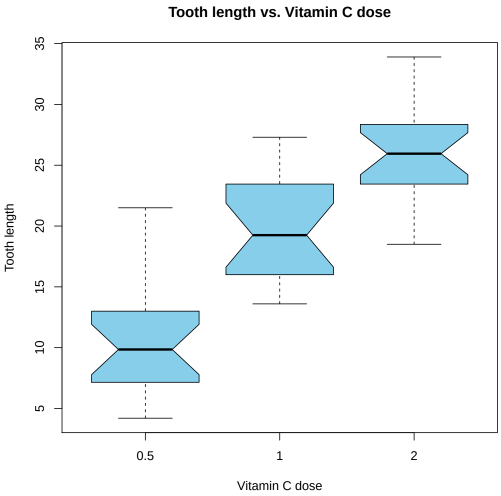
```
# The tooth length is continuous while the supplement type is categorical.
# We therefore choose a boxplot to investigate the association.
boxplot(dat$len~dat$supp, notch=TRUE,
        xlab="Supplement type",
        ylab="Tooth length",
        main="Tooth length vs. Supplement type",
        col="skyblue")
```



```
# OJ seems to have a higher effect on tooth growth than VC.
```

(d) Does the distribution of the tooth length depend on the Vitamin C dose? Illustrate your answer with an appropriate plot (**R-Hint**: `boxplot()`). What's the percentage of guinea pigs in group 3 (2mg Vitamin C) that has longer teeths than 75% of the guinea pigs in group 2 (1mg Vitamin C)?

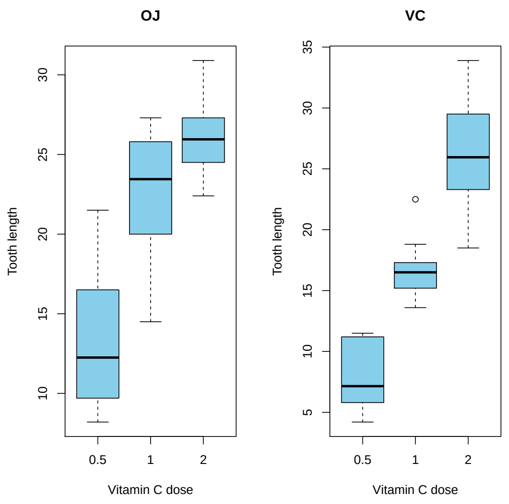
```
boxplot(dat$len~dat$dose, notch=TRUE,
        xlab="Vitamin C dose",
        ylab="Tooth length",
        main="Tooth length vs. Vitamin C dose",
        col="skyblue")
```



```
# Obviously, the higher the Vitamin C dose, the larger the tooth groth.
# From the boxplot we can see that the 75% quantile of dose 2 group is
# on the same value as the 25% of dose 3 group. Therefore, 75% of the
# guinea pigs in group 3 have larger teeths than 75% in group 2.
```

(e) Find out if the influence of the Vitamin C dose is different for the two delivery types. (**R-Hint**: Take subsets of the data using e.g. `dat_oj ← subset(dat, supp=="OJ")`, `dat_vc ← subset(dat, supp=="VC")` and do boxplots for the two subsets.)

```
# Take subsets of the data

# consider guinea pigs which were fed with OJ
dat_oj <- subset(dat, supp=="OJ")
dat_oj[1:3,]

##     len supp dose
## 31 15.2   OJ  0.5
## 32 21.5   OJ  0.5
## 33 17.6   OJ  0.5

# consider guinea pigs which were fed with VC
dat_vc <- subset(dat, supp=="VC")
dat_vc[1:3,]

##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5

# plot the tooth length in both groups
par(mfrow=c(1,2))
boxplot(len~dose,
        xlab="Vitamin C dose",
        ylab="Tooth length",
        main="OJ",
        col="skyblue",
        data=dat_oj)
boxplot(len~dose,
        xlab="Vitamin C dose",
        ylab="Tooth length",
        main="VC",
        col="skyblue",
        data=dat_vc)
```



```
# In both groups, an increase in Vitamin C dose leads to an increase
# in the tooth length. In The VC group, the effect seems to be slightly
# higher (larger maximum) while in the OJ group, a Vitamin C dose of 1
# results in a better tooth growth than in group VC.
```