Example

Speed dating dataset

In the waves other than 6-9 the participants were asked to answer the following question:

You have 100 points to distribute among the following attributes – give more points to those attributes that are more important in a potential date, and fewer points to those attributes that are less important in a potential date. Total points must equal 100.

- attractive
- sincere
- intelligent
- fun
- ambitious
- shared interests

In the following we will write a function that

1. loads the speed dating data set,
2. subselects the data from the first speed date of each individual in waves other than 6-9,
3. extracts the importance ratings of **attractiveness** by the participants in those dates,
4. produces a boxplot of the importance ratings grouped by gender.

Subsetting the data

To load the data we can use the R command:

```
SpeedRawData <- read.csv("SpeedDatingRawData.csv")
```

```
dim(SpeedRawData)
```

```
## [1] 8378  195
```

The raw data has 8378 rows and 195 columns. Then knit the document again and see what happens.

The data is recorded twice for each speed date, hence the ratings of an individual who took part in n speed dates are recorded n times. To reduce redundancy we select the ratings from the first date only. We first identify the rows that record the first speed date for each individual in the waves of interest. We then select the attractiveness ratings as well as the unique id and the gender of the individual. For display purposes we also turn the binary gender column into a factor with labels `female` and `male`.

```
ind <- (SpeedRawData$partner == 1) & (SpeedRawData$wave %in% c(1:5, 10:21))
SpeedData <- SpeedRawData[ind, c("iid", "gender", "attr1_1")]
SpeedData$gender <- factor(SpeedData$gender, labels = c("female", "male"))
```

We save the data frame `SpeedData` in a csv file as follows.

```
write.table(SpeedData, file = "SpeedData.csv", sep = ",", row.names = FALSE)
```

Plots

We use R code chunks to embed plots such as a boxplot.

```
boxplot(attr1_1 ~ gender, data = SpeedData, main = "Importance of Attractiveness",
        col = (c("pink", "lightblue")), xlab = "Gender", ylab = "Percentage points")
```

Functions

The use of R functions ensures reproducibility and avoids cluttering the workspace with variables. Our function will take as input the raw data, the relevant waves and the attribute of interest. It then produces a boxplot of the attribute grouped by gender and saves the selected data in a csv file.

Here is the function that selects the data of interest.

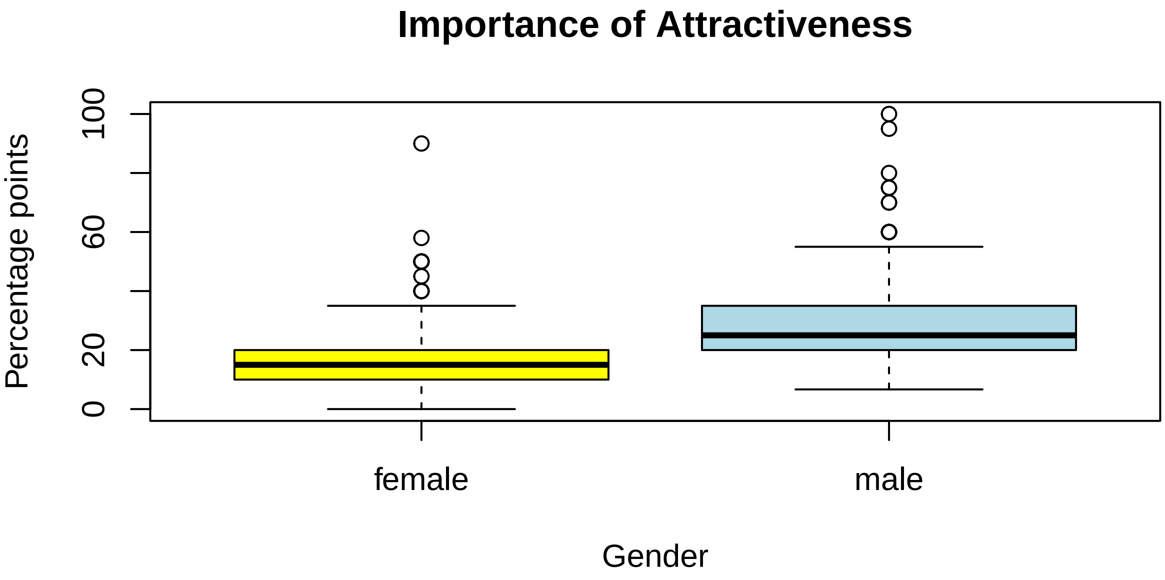
```
SelectData <- function(file, waves = c(1:5), attribute = "attr1_1")
{
  SpeedRawData <- read.csv(file)
  ind <- (SpeedRawData$partner == 1) & (SpeedRawData$wave %in% waves)
  SpeedData <- SpeedRawData[ind, c("iid", "gender", attribute)]
  SpeedData$gender <- factor(SpeedData$gender, labels = c("female", "male"))
  return(SpeedData)
}
```

Below is the function that plots and then saves the data.

```
PlotAndSave <- function(inputfile = "input.csv", waves = c(1:5), attribute = "attr1_1",
                        attributename = "Attribute", outputfile = "output.csv")
{
  data <- SelectData(file = inputfile, waves = waves, attribute = attribute)
  boxplot(data[,3] ~ data$gender, main = paste0("Importance of ", attributename),
          col = (c("yellow", "lightblue")), xlab = "Gender", ylab = "Percentage points")
  write.table(data, file = outputfile, sep = ",", row.names = FALSE)
}
```

Let us try out the function.

```
PlotAndSave("SpeedDatingRawData.csv", waves = c(1:5, 10:21), attribute = "attr1_1",
            attributename = "Attractiveness", outputfile = "SpeedData.csv")
```



This website has a nice tutorial on R Markdown that has more details and also links to the reference guide and to cheat sheets for R Markdown.