Statistical Computing

Master Data Science

Winter Semester 2017/18

BEUTH HOCHSCHULE FÜR TECHNIK BERLIN
University of Applied Sciences

Prof. Tim Downie

Workshop 5
Data Handling, Exploratory Data Analysis and Graphics: Exercises

1 Introduction

The aim of this workshop is to consolidate what you have learnt in Workshops 1 to 4. You are not expected to have remembered all the commands you have learnt so far; you will certainly need to refer to your notes from the first four weeks and to the R help files.

Exercise 1 concentrates on data manipulation and Exercise 2 concentrates on exploratory data analysis.

Your script file should be properly commented and your answers to statistical questions should be written up in a word file or similar.

1.1 Preliminaries

Download the data files from the course Moodle page into your `H:\\StatComp` directory. The files are: `SmartiesB1.Rda`, `SmartiesB2.Rda`, `SmartiesB3.Rda` and `MunichAccomm.dat`

Start RStudio using the Icon on the desktop

Strg+Shift+N opens a new R script.

Type the comment

```
#Statistical Computing: Workshop 4
#Basic Statistics and exploratory data II
```

as the title.

Save the file in your `H:\\StatComp` folder with the name `Workshop4.R`.

Set your *working directory* to be `H:\\StatComp`. The code to do this is

```
> setwd("H://StatComp")
```

By now it is possible that you have accidentally saved your workspace when leaving RStudio. If your Environment Window is displaying any objects then clear the workspace using *Session > clear workspace*.

Open a word processing program for you to write up your results.

2 The Smarties Data

A group of statistics students were each given a box of Mini-Smarties in return for recording the number of Smarties of each colour in that box. The Smarties came in three plastic bags, each bag containing 11 boxes.

The data for each bag is stored in a separate file: `SmartiesB1.Rda`, `SmartiesB2.Rda` and `SmartiesB3.Rda`

(a) Load the 3 Smarties files (the data files are in .Rda format).

(b) Inspect the data from one file and make sure that you understand the format.

(c) Create one data frame containing the data for all three bags in "long format" (i.e. one row for every box). This data frame should include an identification variable called `BagNumber`.

(d) How many boxes were in each bag?

(e) Create a variable which contains the number of Smarties in each box. Hint: The function `apply(mat,index,fun)` takes the matrix or data frame `mat` and applies the function `fun` to each row (when `index=1`) or column (when `index=2`).

(f) Display the number of Smarties in each box in a bar chart. Give the diagram a title and sensible axis labels.

(g) Find the mean number of Smarties per box for each colour (e.g mean number of red Smarties per box, mean number of green Smarties per box etc.) Use the R function `round()` to give these answers to 1 decimal place.

(h) Write this dataset to a csv file.

(i) Create another data frame consisting of 3 rows one for each bag. Each row contains the mean values for total Smarties and Smarties of each colour.

3 Munich Accommodation data

The company Infratest carried out a representative survey of rental apartments and houses in Munich in 2003. The data are available in the file `MunichAccomm.dat`. Each row represents one apartment or house. For convenience all rows will be referred to as apartments. The dataset contains the following variables for each apartment:

`Rent`	Monthly rent (Euros)
`Rentpm2`	Monthly rent per square meter (Euros/m^2)
`Area`	Official floor area (m^2)
`NumRooms`	Number of rooms (excluding kitchen and bathroom)
`Year`	Year the accommodation was built
`Borough`	Numeric code for borough where the apartment is situated
`GoodArea`	Is the Area considered a good area? (0/1)
`GreatArea`	Is the Area considered an exceptionally good area? (0/1)
`HotWater`	Does the apartment have running hot water? (0/1)
`CentralHeating`	Does the apartment have central heating? (0/1)
`TiledBathrm`	Does the apartment have a tiled bathroom? (0/1)
`QualityBathrm`	Is the bathroom above a certain standard? (0/1)
`QualityKitch`	Is the kitchen above a certain standard? (0/1)

(a) The datafile is in ASCII format. Each field is separated by a space and the first row contains the variable names. Read the data into R.

(b) How many apartments are included in the study.

(c) How many apartments have a "quality Kitchen", and what proportion of appartments is this?

(d) What is the mean and median for the variables *rent* and *rent per square meter*?

(e) What proportion of apartments have 2 rooms.

(f) Display the *build year* in a histogram with on bar for each decade. What do you notice is odd about this diagram?

(g) Plot the rent *against* area (rent on y-axis *against* area on x-axis) in a scatter plot.

(h) Plot the rent per square meter *against* area in a scatter plot.

(i) Obtain the correlation coefficient for

  (i) rent and area, and

(ii) rent per square meter and area

(j) Use the following command to define a new variable called `Old`. The variable is a factor variable with `"yes"` if the build year is before 1950, and `"no"` otherwise.

```
> Accom$Old<-factor(Accom$Year<1950,labels=c("no","yes"))
```

(k) What proportion of apartments are “old” apartments?

(l) Produce a bar chart of number of rooms split by Old/New Status.

(m) What is the mean rental price per square meter for old and new apartments? Round your answer to 2 decimal places.

(n) Plot the mean rental price per square meter in a box plot with separate boxes for old and new apartments.

(o) Carry out a t-Test so see if there is a statistically significant difference in mean rental price per square meter for old compared to new apartments?

4 Tidying up

► Make sure your script file has sensible comments.

► Save the script file (source file) again: Strg + S or *File > Save as*.

► Leave RStudio by typing the command:

```
> q()
```

When R asks you *Save workspace image ...?*, click on **Don’t save**!

► Feierabend!