TRAINING Computer exam BFVH15DAVUR

Data Analysis and Visualization using R

YOUR NAME (YOUR STUDENT NUMBER)

June 2016

Test header

- **Teacher** Michiel Noback (NOMI), to be reached at +31 50 595 4691
- **Test size** 4 pages; 7 questions
- **Aiding materials** Computer on the BIN network
- **Data files**
  - `food_constituents.txt`
- **Supplementary materials**
  - `TRAINING_EXAM.pdf` This test as pdf
  - `TRAINING_EXAM.Rmd` This test as R markdown
  - `R_cheatsheet.pdf` Lists all R functions that may be used
  - `rmarkdown-reference.pdf` R markdown reference document

Instructions

In the real test, you should be logged in as guest (username = "gast", password = "gast"). On your desktop you will find all supplied data and supplements, as well as the submit script `submit_your_work`. For this training test, simply quit your browser and time your work; in the real exam, you will have two hours to solve a set of similar questions. Use the supplied R markdown file `TRAINING_EXAM.Rmd` to solve and answer the questions of this test. Fill in your name and student number in the header of this document. **Note: never use `echo = False` in your code chunk headers.**
All questions have the possible number of points to be scored indicated. your grade will be calculated as
$Grade = 1 + (\frac{PointsScored}{MaximumScore} * 9)$

After finishing, `knit` the result into a pdf document and rename it to `TRAINING_EXAM_YOUR_NAME.pdf`.

Data description

This test explores a dataset containing measurements of several food constituents in a variety of foods, categorized over several groups.

Code "Book"

These are the columns, and their descriptions, included in the data file `food_constituents.txt`:
id.nr Type kcal protein carb.total carb.sugar carb.other fat.total fat.sat fat.unsat fiber Na 2 chocolate 442 5.00 67.40 64.60 2.80 15.50 9.00 6.50 6.60 0.100

1. **id.nr** simple measurement counter
2. **Type** food group
3. **kcal** energy contents in kcal/100g product
4. **protein** protein content in g/100g product
5. **carb.total** total carbohydrate content in g/100g product
6. **carb.sugar** sugar carbohydrates in g/100g product
7. **carb.other** other carbohydrates in g/100g product
8. **fat.total** total fat content in g/100g product
9. **fat.sat** saturated fats in g/100g product
10. **fat.unsat** unsaturated fats in g/100g product
11. **fiber** fiber contents in g/100g product
12. **Na** Sodium content in g/100g product

Here starts the actual test

Part 1: Data loading and cleaning

Question 1 (10 points) Load the data from file `food_constituents.txt` and assign it to a variable called `foods`. Take special care with missing/invalid fields, and also make sure the columns are loaded in the right data type.

```
#your code here
```

If you fail to load the data as instructed above, you may load the pre-processed file using the following code chunk (uncomment the R code). Make sure your working directory is set appropriately! You will not get any points for this question, however.

```
## Uncomment this line to load pre-processed data
#load("./foods_raw.Rdata")
```

Question 2 (5 points) There are several rows with missing data. Report these and also remove these from the `foods` dastaset. Hint: use the function `complete.cases()` to achieve this.

```
#your code here
```

Part 2: Data exploration

Question 3 (6 points) **Question 3 a (2 points)** What is the average caloric value of this food listing?

```
#your code here
```

Question 3 b (2 points) Tabulate the frequencies of the different food categories (e.g. Type)

```
#your code here
```

Question 3 c (2 points) Show the "6-number summary" for -only- the fat measurements.

```
#your code here
```

Question 4 (12 points) **Question 4 a (4 points)** Create a new column called `fat.cat` that divides the foods into 3 food categories based on total fat content: `high.fat`, `medium.fat` and `low.fat`. Take into account that this is an ordinal scale!.

```
#your code here
```

If you are not able to create this factor, load it from file and attach it to your foods dataframe. You will not get points for this question of course.

```
##uncomment this if you could not create the factor yourself
#load("foods_fat_cat.RData")
```

Question 4 b (4 points) Calculate mean energy content for each fat.cat category.

```
#your code here
```

Question 4 c (8 points) -Challenge question- Report which foods from each fat.cat group have the largest fraction of saturated fat relative to total fat.

```
#your code here
```

Is there anything funny in these results? Discuss/explain these!

Question 5 (8 points) Sort (and list) the Pasta foods by energy content, from high to low.

```
#your code here
```

Part 3: Visualization

Question 6 (8 points) Create a -well annotated- box plot showing distributions of total total carbohydrate content for the three fat categories (low.fat, medium.fat and high.fat).

```
#your code here
```

Question 7 (15 points) Create a -well annotated- scatter plot exploring the total carbohydrate content relative to energy content. You should add a linear regression line to emphasise the relationship.

```
#your code here
```

Is there a clear relationship as you would expect? If not, can you explain?