Worksheet *Analytic 2*

Sven Garbade
Centre for Pediatric and Adolescent Medicine
Division for Neuropediatrics and Metabolic Medicine
University Hospital of Heidelberg
sven.garbade@med.uni-heidelberg.de

April 2019

Instructions

1. Use R to solve all Problems. Prepare your results as a short presentation.
2. Use all sources of help (slides, internet, etc.) you need to solve the problems.

Problem 1

The data set `housing2.csv` is a case by case representation of the R data set *housing*.

(a) Make yourself familiar with the data set by reading the R help page for *housing* data.

Sample solution:

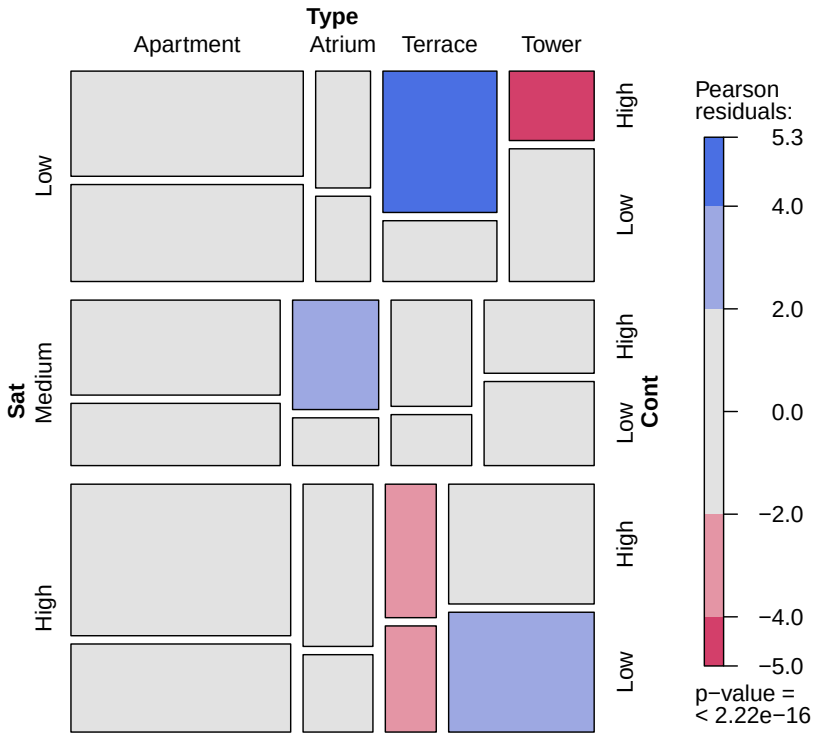
```
> ## data is in package mlbench
> ## not run
> ## require(mlbench)
> ## ?housing
> ## or use Google
```

(b) Load data set `housing2.csv`. Analyze *Sat*, *Type* and *Cont*. Conclusions?

Sample solution:

```
> d <- read.csv("./data/housing2.csv")
> d$Sat <- factor(d$Sat, levels=c("Low", "Medium", "High"))
> f <- xtabs( ~ Sat + Type + Cont, data=d)
> f

, , Cont = High

        Type
Sat      Apartment Atrium Terrace Tower
  Low          141     37      93    34
  Medium       116     55      50    47
  High         191     65      39   100

, , Cont = Low

        Type
Sat      Apartment Atrium Terrace Tower
  Low          130     27      40    65
  Medium        76     24      24    54
  High         111     31      31   100

> ## chi^2 analysis
> summary(f)

Call: xtabs(formula = ~Sat + Type + Cont, data = d)
Number of cases in table: 1681
Number of factors: 3
Test for independence of all factors:
        Chisq = 120.03, df = 17, p-value = 1.523e-17
```

```
> require(vcd)
> ## Note: vcd uses p-value from likelihood ratio test
> mosaic(f, shade=TRUE)
```



Problem 2

Clean the data set in file `Data_to_clean.csv`. Solve the following problems:

(a) Convert column `date` into ISO-format `Years-month-day`, e. g. `1980-06-20` for 20th of June 1980. You can use `as.Date()` for converting strings to date objects.

(b) In column `sex`, use `f` and `m` for coding *female* and *male*.

(c) Write code to filter extreme values with one of the following rules:

- Filter 1: extreme value = mean $\pm$ 1.5 standard deviation.
- Filter 2: extreme value: value is above 98th percentile.

Problem 3

The data set `credit.csv` stores information about persons asking for a credit. Following attributes are recorded: age (years), income, number of children, car (0=no, 1=yes) and creditworthy (yes/no).

According to the collected data, are the following persons creditworthy?

	age	income	children	car
1	41.00	2500.00	1.00	1.00
2	43.00	5000.00	3.00	1.00

(a) Use all statistical methods you know to solve this problem. Are there any differences between the models?

Problem 4

The internet repository `https://archive.ics.uci.edu/ml/index.php` has about 470 data sets for machine learning and statistics.

(a) Build groups and choose a data set and draft a research problem. Use the CRIPS-DM to organise your steps.

(b) Analyse the data.

(c) Prepare a short presentation and discuss your results in our course.