Humboldt-Universität zu Berlin

Ladislaus von Bortkiewicz Chair for Statistics

Data analysis I

Sigbert Klinke

April 25, 2016

- Quick-R: `http://www.statmethods.net`
- UCLA Institute for digital research and education

Choose R, SPSS or any other statistical package to fulfill the following tasks. If you do not know the software good enough then use the help of your software or search for help in the internet.

1. Read in the `BOSTONH` data set (Boston Housing data). The data contain the Housing data for 506 census tracts of Boston from the 1970 census; see Harrison, D. and Rubinfeld, D.L. (1978), *Hedonic prices and the demand for clean air*, Journal of Environmental Economics and Management, 5, 81–102.

`crim`	per capita crime rate by town
`zn`	proportion of residential land zoned for lots over 25,000 sq.ft
`indus`	proportion of non-retail business acres per town
`chas`	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
`nox`	nitric oxides concentration (parts per 10 million)
`rm`	average number of rooms per dwelling
`age`	proportion of owner-occupied units built prior to 1940
`dis`	weighted distances to five Boston employment centres
`rad`	index of accessibility to radial highways
`tax`	full-value property-tax rate per USD 10,000
`ptratio`	pupil-teacher ratio by town
`b`	$1000(B - 0.63)^2$ where B is the proportion of blacks by town
`lstat`	percentage of lower status of the population
`medv`	median value of owner-occupied homes in USD 1000's

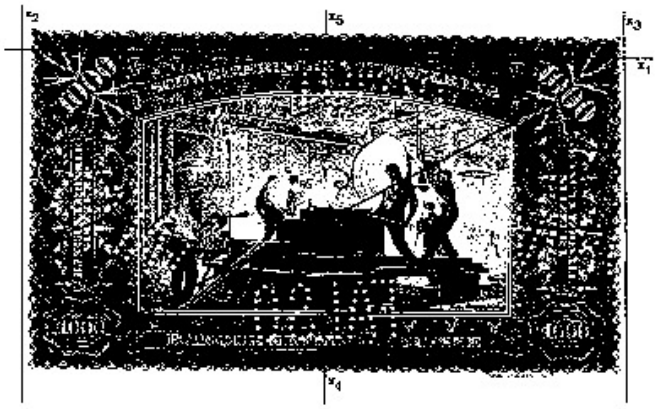
   Hint for R users: The Boston Housing data are available in R directly:

```
library("MASS")
Boston
```

2. Create a table plot for the Boston Housing data. Choose another sorting variable as in the lecture. What do you learn from the plot?

3. Read in the `BANK2` data set. The data are taken from the book of Flury, B. and Riedwyl, H. (1988). *Multivariate Statistics: A practical approach*. London: Chapman & Hall. They are about the measurement of 200 genuine and forged old Swiss bank notes.

WIDTH (X1)	Width of the bank note
LEFT (X2)	Height of the bank note, measured on the left
RIGHT (X3)	Height of the bank note, measured on the right
LOWER (X4)	Distance of inner frame to the lower border
UPPER (X5)	Distance of inner frame to the upper border
DIAGONAL (X6)	Length of the diagonal



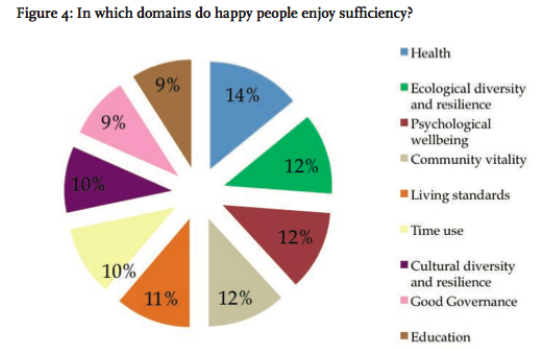
4. Create a table plot for the Swiss Banknote data. Choose the diagonal as sorting variable. What do you learn from the plot?

5. Explain why the graphics in the next page are misleading for the viewer?

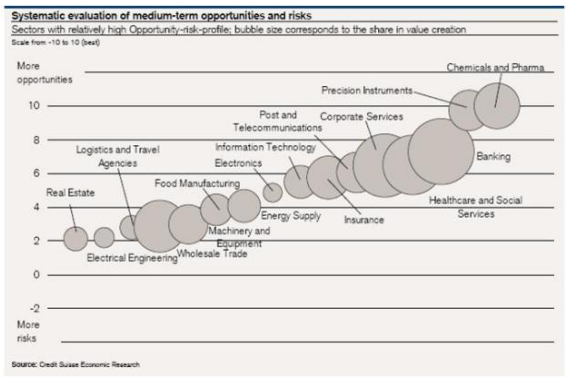
looks like one-dimension, although there are three and therefore looks like there are no risks at all
- description is too small but highly necessary to understand the meaning

Trivial research
- Happy people enjoy sufficiency

- nicht 100%
- Unterschiede zwischen Tortenstücke schwer zu sehen (Tortendiagramme in Statistik unerwünscht)

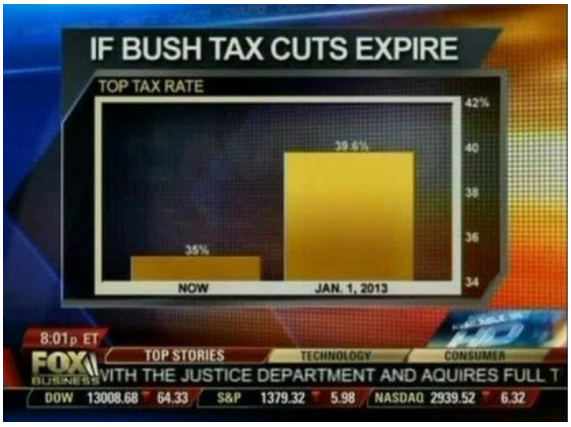


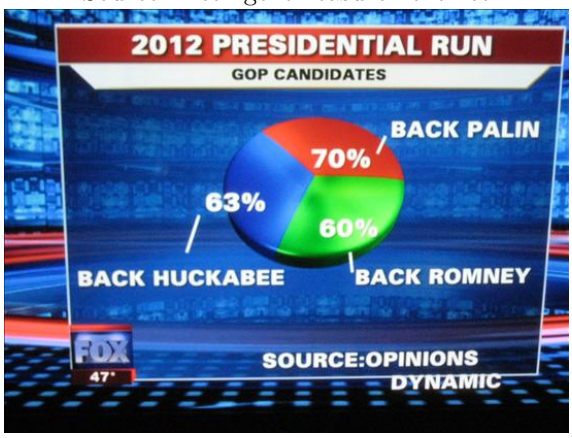
Source: andrewgelman.com



Source: intelligentmeasurement.net

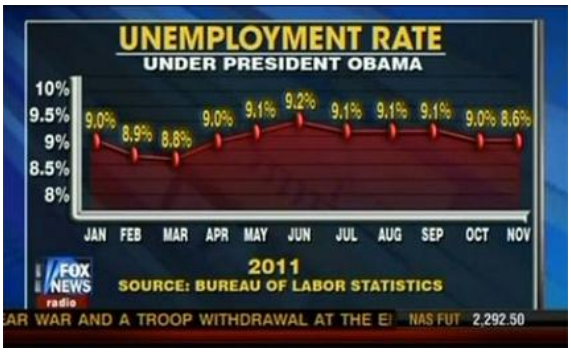
Visual effect:
- tax increase of 4.6% appears tremendeously high





over 100%
- Mehrfachantwort war möglich
- Kreisdiagramm sehr schlecht geeignet

8.9 is lower ranked than 8.6



Accumulated - misleading
- Quarter komisch gewählt
- Ziel war linearen Trend abzubilden

Source: simplystatistics.org (used in Fox news)

Selektive Auswahl bestimmter Daten

- Grundgesamtheit der Waffenbesitzter und Grundgesamtheit der Nutzer von Heimwerker-Tools

Skala ist nicht richtig

Source: www.johnspens.net

Source: therandomtexan.files.wordpress.com