Lecture 1

Introduction to Data Visualization

1

About the course

Course structure

- 7 lectures (presentations)
- 6 labs, work in groups 2 persons
- 3 seminars
- Star-marked assignments in 3 occasions – to be solved individually, optional.

Examination

- Submission of lab reports
- Presentation of lab reports and opposition
- Computer-based written exam
- **Star-marked assignments passed+ earned at least 14 points at the exam =get 2 points more**

2

About the course

Information & Lab reporting

- LISAM is used
- Good lab practices
  - Supervision time is limited (2h)
  - Lab is normally put at LISAM a day before the lab supervision session
  - Start doing lab before the supervision session
  - Possible strategy: one individual in the group works with assignment 1, one with assignment 2 during the supervision time, then help each other later
- Deadlines
- Seminars are obligatory – speakers and opponents selected randomly

3

About the course

Course literature:

- **"Interactive Data Visualization"** by M.O. Ward et al., Second Edition.
- Papers, software documentation & manuals
- Decide groups
  - https://docs.google.com/spreadsheets/d/1GbN6K4dZp2MtgTX5QiKc53QHvYygzd7Iyb2zy3Tls4o/edit?usp=sharing

4

Introduction

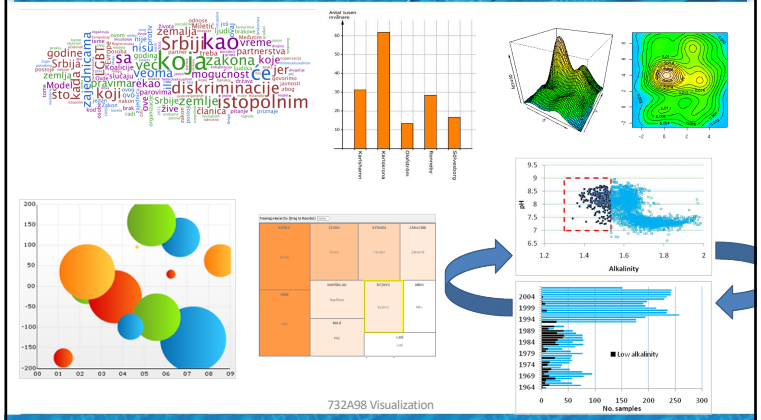
Visualization in Statistics and Machine Learning…

… is a methodology that allows for discovering or confirming a useful information about the data by constructing and examining the graphical output

Course contents

- **Topic 1**: Introduction to Data Visualization. Introduction to Ggplot2, Plotly, Shiny.
- **Topic 2**: Perception and Visualization. Data preprocessing.
- **Topic 3**: Basic graphs. Geospatial visualization.
- **Topic 4**: Multivariate data visualization.
- **Topic 5**: Interactive visualization. Text visualization.
- **Topic 6**: Graph visualization. Animation.

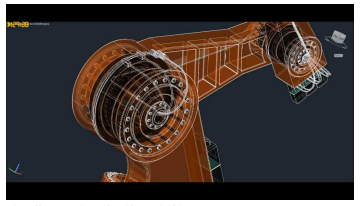
Viualizations



Different types of visualization

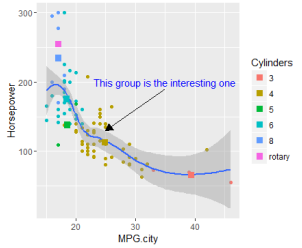
- In this course, we focus on **visualization=information visualization**
  - Data→Visualization→Analysis
- Related concepts
  - **Computer graphics**: Data is not necessary present, analysis is not normally assumed
    - Example: Computer games
  - **Scientific visualization**: similar to information visualization, often engineering data, statistical/machine learning analysis is normally not assumed
    - Example: Industrial robots

Different types of visualization



https://www.youtube.com/watch?v=mpechGIfPbw

Scientific visualization



Information visualization

Challenges in information visualization

- Which graphs can be used for analysis of my data?
- How to create these graphs?
- How should these graphs be analysed?
- How to make these graphs looking good enough for publication or presentation?

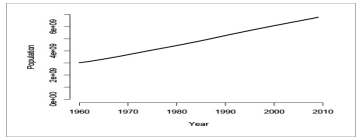
9

Why is visualization important?

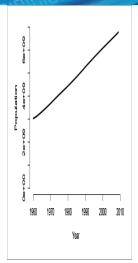
- Human sight = primary resource for information understanding
- Visualization is often the **quickest** way for data understanding
- The way of data visualization may affect decision making dramatically

10

Why is visualization important?



Decision here: population does not increase so much, no intervention needed

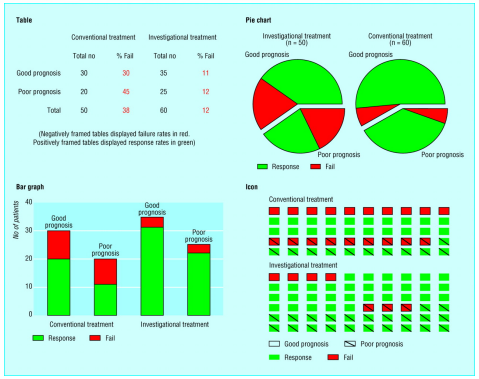


Decision here: population increases quickly, intervention is required

Visual perception problem

11

Why is visualization important?



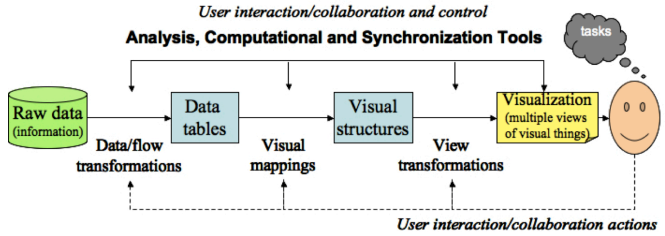
Source: Elting Linda S, Martin Charles G, Cantor Scott B, Rubenstein Edward B. Influence of data display formats on physician investigators' decisions to stop clinical trials: prospective trial with repeated measures *BMJ* 1999; 318 :1527

12

Visualization aims

- Visualization for **exploration**
  - Clusters
  - Trends
  - Anomalies
  - …
- **Confirmatory** visualization
  - Example 1: Perform linear regression, analyse residuals→ was linear regression reasonable
  - Example 2: Discover clusters by K-means, visualize clusters→ are they clusters actually?
- Visualization for **presentation**

Visualization pipeline



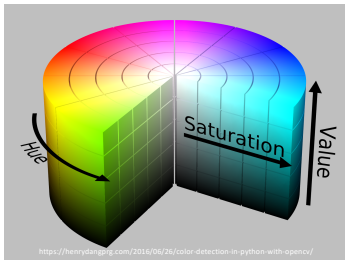
Key ingredient: mapping data columns to visual structures (aesthetics)

The role of perception

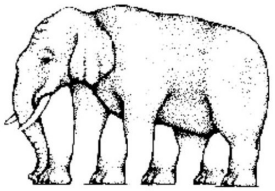
- Human visual system has limitations
- These limitations may lead to wrong/incomplete analysis of graphs
- Understanding how we see→ better displays
- Misleading graphics needs to be avoided

Colors

- Color= hue + saturation + value (lightness)
- 8% of males are color deficient→ what are good colors?

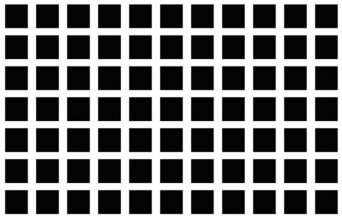


Illusions



How many legs does this elephant have?

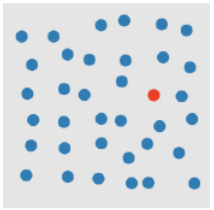
http://www.illusa.com/gallery/elephant-illusion.jpg



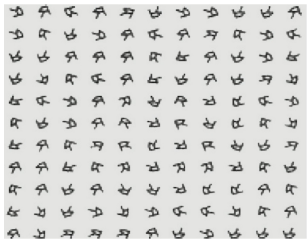
17

Preattentive processing

- Certain aesthetics are fast to process



How quickly can you identify a red dot?

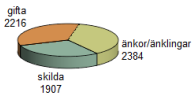


How quickly can you identify a square of right-handed Rs?

18

The role of perception

- How can this affect analysis?



19

Data preprocessing

- Viewing raw data is often prefered
- Sometimes some preprocessing is needed

- Missing values and Data cleaning
  - Discard the bad record →may remove almost all data
  - Assign sentinel value
  - Column mean imputation
  - Nearest neighbor imputation
  - Other imputations

20

Data preprocessing

- Normalization
  - Converting column to range [0,1]. Useful in for ex. color mapping
  - Centering and scaling 0/1
  - Nonlinear transformations: log, sqrt
- Segmentation
  - Split data according to some column

21

Data preprocessing

- Sampling, subsetting and expanding
  - Random sampling reduces size of data and facilitates overplotting (for ex. scatterplots)
  - Interpolation: linear (one dimension), bilinear (two dimensions), nonlinear. Select necessary amount of intrepolation points.
- Dimension reduction
  - PCA
  - MDS
  - Other techniques (ex. ICA, Autoencoders), welcome to **Machine Learning** course..

22

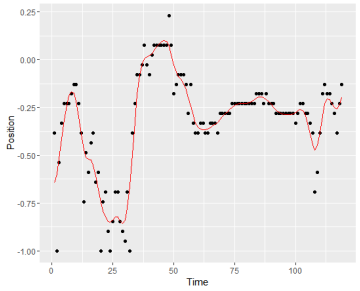
Data preprocessing

- Mapping nominal dimensions to numbers
  - Random mapping should never be done unless intrinsic ordering is present
  - Use other numeric variables to measure in the data to measure “closeness” of values in the nominal variable
  - Correspondence analysis
- Aggregation and summarization
  1. Grouping observations
  2. Computing summary statistics per group

23

Data preprocessing

- Smoothing and filtering
  - Replace original values with a smoothed versions



24

Software

Commercial:

- SAS and SAS JMP – environment. Special visual tools are available (JMP), require separate license. Well documented. Good even for large sets. **SAS Enterprise Guide** has many visual static tools
- Spotfire – Many static and interactive visualization tools
- Tableau– Many static and interactive visualization tools
- InfoScope – visualizing maps, interactive visualizations

Free:

- R – programming language. Set of packages is constantly updated. A lot of statistical tools (even the newest methods) Badly documented
- Plotly – a tool for interactive and dynamic graphics, R interface available
- Shiny - – a tool for R-based web applications using graphics
- GraphViz – visualization of graph data, coding needed
- Jigsaw – Text analysis

25

Software

Tools for the web (used by web designers):

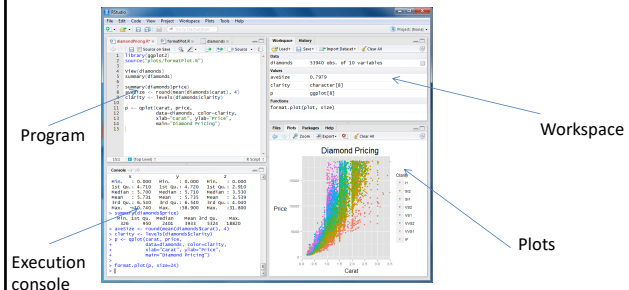
- ActionScript
- JavaScript
- Prefuse
- VTK

… much more references given in the course book

26

Course software

- Install R: http://www.r-project.org/
- Install RStudio: http://rstudio.org/

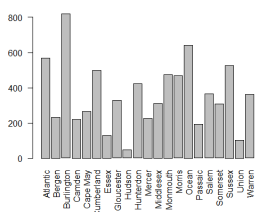


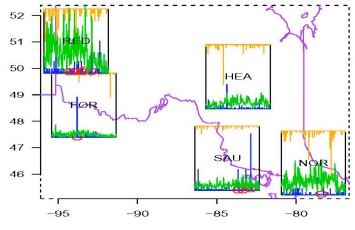
27

Graphical tools in R

- R base: basic plotting, not publication quality

Grob package: low-level plotting tools, new types of plots





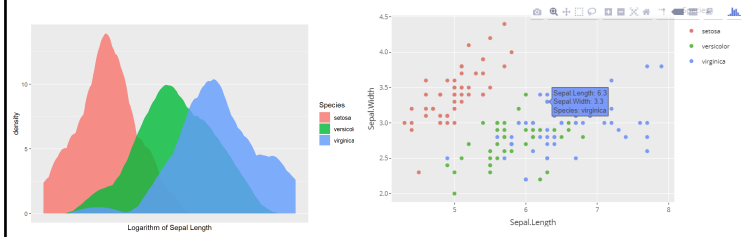
Zhou, Lutong, and W. John Braun. "Fun with the r grid package." *Journal of Statistics Education* 18.3 (2010).

28

Graphical tools in R

Ggplot2 package: based on **grammar of graphics**, close to publication quality

Plotly package: Ggplot2 + interactivity



29

Graphical procedures

Base R graphical procedures:

- plot(x,..) plots time series
- plot(x,y) scatter plot
- plot(x,y) followed by points(x,y) plots several scatterplots in one coordinate system
- hist(x,..) plots a hitogram
- persp(x,y,z,...) creates surface plots
- cloud(formula,data..) creates 3D scatter plot

```
x<-c(1,4,7 8, 12);
y<-c(4,3,1,6,9);

plot(x,y, type="l", col="orange",
    main="My plot", xlab="X1", ylab="X2");
points(x, y/2+2, type="b", col="blue");
```

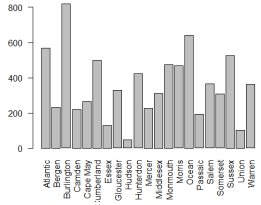
30

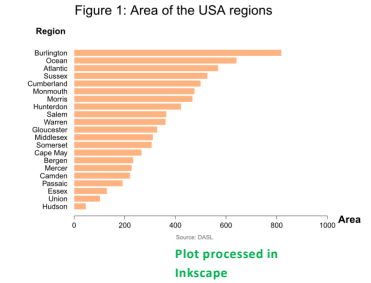
Publication quality graphics

- Visualization for exploration
  - Default settings
- Visualization for presentation for publication
  - Higher quality graphics is required
    - Improve the graph quality in the software (often requires quite a bit of programming)
    - Use postprocessing tools, such as Inkscape or Adobe Illustrator

31

Publication quality graphics





Plot processed in Inkscape

Example: Compare two plots and state what is improved in the second plot.

32

Making publication quality graphics

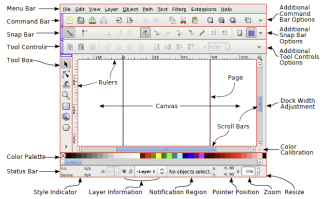
- Install **Inkscape**
  - http://inkscape.org/
  - Inkscape is an open source, *SVG*-based vector drawing program
  - file format that Inkscape uses is compact and quickly transmittable over the Internet.
  - Vector graphics: image is defined in terms of lines, not pixels
    - Benefit: can be enlarged without loss of picture quality

1. Save your R plot as PDF and import it to Inkscape
2. Make changes and export your plot as a PNG-file or save it as PDF.

Bitmap image and vector image (enlarged)

Inkscape

- Menu bar **(Important:** File, Object, Path, Text, View→Grid)
- Command bar (Zoom to fit page, Edit object's colors)
- Tool controls
- Tool box (Select and transform, zoom, Text, Eraser, Fill)
- Color palette
- Status bar



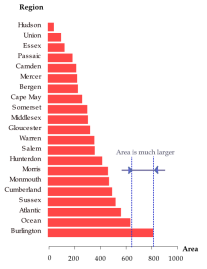


Figure 1: There seems to be one outlying region

Home reading

- Course book, chapters 1.1, 1.3-1.8 and 2
- Manual to InkScape: http://tavmjong.free.fr/INKSCAPE/MANUAL/html/index.php