732A98 Visualization
Division of Statistics and Machine Learning
Department of Computer and Information Science

Laboratory work 1

Instructions

- Be concise and do not include unnecessary printouts and figures produced by the software and not required in the assignments.
- **Include all your codes as an appendix into your report; you are also recommended to show parts of the codes in the flowing text of the report.**
- A typical lab report should 2-4 pages of text plus some amount of figures plus appendix with codes.
- Create a report to the lab solutions in RMarkdown. Make sure that it is can be compiled to HTML and that all paths in RMD file are relative to the current directory where the RMD file is located. **Reports that can not be compiled are returned without revision.**
- Put the RMD file and all supporting files into one ZIP archive when you submit it to LISAM.
- The lab report should be submitted via LISAM before the deadline.

Assignment 1

Sometimes it is necessary to adjust visualizations obtained by complicated R packages and it is difficult to do it programmically. File **tree.pdf** shows a decision tree created by **party** package. Use Inkscape to produce a publication quality graph which is like the one shown in Figure 1 (you may make more improvements if you like!). Note that:

- Terminal nodes are renumbered
- Node numbers and p-values are removed from the non-terminal nodes, title is removed
- Percentages are explicitly added below each terminal node and their positions are adjusted appropriately
- Colors are adjusted

Add the resulting PDF picture to your report.

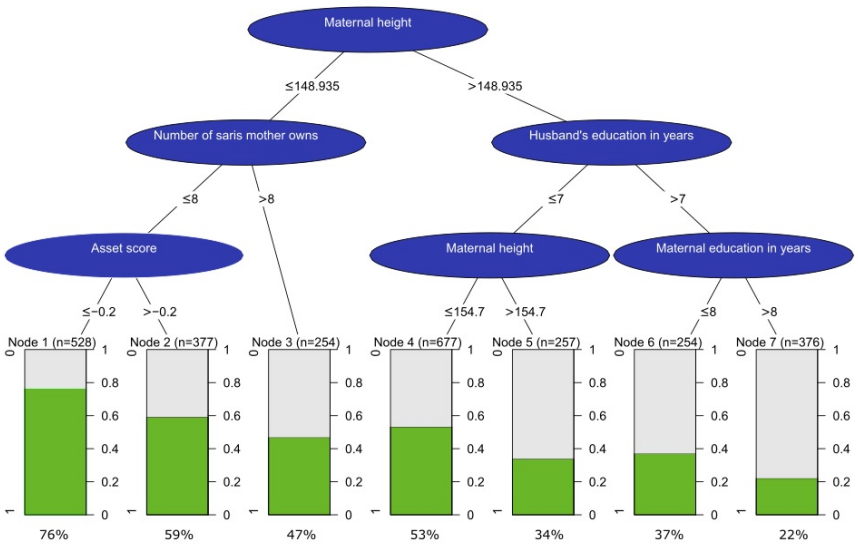

Figure 1. A tree from assignment 1.

732A98 Visualization
Division of Statistics and Machine Learning
Department of Computer and Information Science

Assignment 2

Data set SENIC describes the results of measurements taken at different US hospitals. The description of the variables is given in the accompanying document **SENIC.pdf.**

1. Read data from SENIC.txt into R.
2. Create a function that for a given column (vector) X does the following:
    a. It computes first and third quantiles Q1 and Q3 with quantiles()
    b. It returns indices of outlying observations, i.e. observation with X-values greater than Q3+1.5(Q3-Q1) or less than Q1-1.5(Q3-Q1).
3. Use **ggplot2** and the function from step 2 to create a density plot of *Infection risk* in which outliers are plotted as a diamond symbol (◊) . Make some analysis of this graph.
4. Produce graphs of the same kind as in step 3 but for all other quantitative variables in the data (aes_string() can be useful here). Put these graphs into one (hint: arrangeGrob() in gridExtra package can be used) and make some analysis.
5. Create a **ggplot2** scatter plot showing the dependence of *Infection risk* on the *Number of Nurses* where the points are colored by *Number of Beds*. Is there any interesting information in this plot that was not visible in the plots in step 4? What do you think is a possible danger of having such a color scale?
6. Convert graph from step 3 to **Plotly** with *ggplotly* function. What important new functionality have you obtained compared to the graph from step 3? Make some additional analysis of the new graph.
7. Use data plot-pipeline and the pipeline operator to make a histogram of *Infection risk* in which outliers are plotted as a diamond symbol (◊) . Make this plot in the **Plotly** directly (i.e. without using ggplot2 functionality). **Hint**: select(), filter() and is.element() functions might be useful here.
8. Write a **Shiny** app that produces the same kind of plot as in step 4 but in addition include:
    a. Checkboxes indicating for which variables density plots should be produced
    b. A slider changing the bandwidth parameter in the density estimation ('bw' parameter)

Comment how the graphs change with varying bandwidth and which bandwidth value is optimal from your point of view.

Submission procedure

Assume that X is the current lab number, Y is your group number.

If you are neither speaker nor opponent for this lab,

- Make sure that you or your group comrade submits the group report using *Lab X* item in the *Submissions* folder before the deadline

If you are a speaker for this lab,

- Make sure that you or your group comrade does the following before the deadline:
  - submits the group report using *Lab X* item in the *Submissions* folder before the deadline
  - Goes to Study room *Group Y→Documents* and opens file *Password X.txt*. Then the student should put your group report into ZIP file *Lab X_Group Y.zip* and protect it with a password you found in Password X.txt
  - Uploads the file to *Collaborative workspace →Lab X* folder

If you are opponent for this lab,

- Make sure that you or your group comrade submits the group report using *Lab X* item in the *Submissions* folder before the deadline
- After the deadline for the lab has passed, go to Collaborative workspace→*Lab X* folder and download the appropriate ZIP file. Open the RMD file in this ZIP file by using the password available in *Course Documents →Password X.txt,* compile it, read it carefully and prepare (in cooperation with your group comrade) **at least three questions/comments/improvement suggestions per lab assignment** in order to put them at the seminar.