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

Laboratory work 1

Instructions

- Create a report to the lab solutions in PDF.
- 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.**
- A typical lab report should 2-4 pages of text plus some amount of figures plus appendix with codes.
- The lab report should be submitted via LISAM before the deadline.

Assignment 1

Data set “jobs.txt” contains information about the percentage employed in different industries in European countries during 1979. The variables are:

- `Country: Name of country`
- `Agr: Percentage employed in agriculture`
- `Min: Percentage employed in mining`
- `Man: Percentage employed in manufacturing`
- `PS: Percentage employed in power supply industries`
- `Con: Percentage employed in construction`
- `SI: Percentage employed in service industries`
- `Fin: Percentage employed in finance`
- `SPS: Percentage employed in social and personal services`
- `TC: Percentage employed in transport and communications`

1. The file “jobs.txt” is tab-separated. Add one extra column “Group” in the file that shows “Sc” for Scandinavian countries, “S” for southern Europe countries, “W” for other western Europe and “E” for the eastern Europe (former Soviet block)
2. Import the file to R and create a tree map where rectangles are grouped by Group, rectangle sizes show SI and colors show Fin. The map should show both observation names and group names. Analyze the plot and comment whether the results you have discovered seem to be reasonable.
3. Make plots of all quantitative variables by using Chernoff faces. Analyze the plot. Interpret your results.
4. Reorder data with the help of the single link hierarchical clustering (order.single()). Repeat step 3 by plotting a segment chart of the reordered data instead. Compare the findings in steps 3 and 4.
5. Create a parallel coordinate plot of the quantitative variables. Can you distinguish clusters or outliers in the plot? Are there variables that seem to be correlated?
6. Create a parallel coordinate plot of the quantitative variables by:
    a. Computing the distance between any pair of variables X1 and X2 as 1-correlation(X1,X2)

   - **b.** Solving a corresponding travelling salesman problem
   - **c.** Creating a parallel coordinate plot with the permuted columns

   Can you see clusters more distinctly compared to the plot from step 5?

7. Create a coordinate plot by using the same distance measure as in step 6 but use hierarchical clustering seriation for making permutations (**HINT:** use function seriate() to do seriation and function get_order() to extract the actual permutation vector). Identify the most severe outliers in the plot and identify with respect to which variables they are outlying. Select the most obvious clusters in the plot by setting a color for plotting of each of them. Which variables are important for defining each of these clusters? Comment how the clusters you have found are related to the Group variable. Interpret the results.
8. Which parallel coordinate plot was the easiest for you to analyze? Were the clusters defined by parallel coordinates similar to those found by Chernoff and segment plots?

Assignment 2

In this assignment, you will analyze the prices of the real estate in Sweden. Go to http://www.scb.se and choose “English” language. Search for “Sold agricultural real estate by region and buildings condition” and open the corresponding page. Switch to the tab “Statistical Database” and click the corresponding item that will appear. Select “Purchase price, average in 1000 SEK”, all counties (except of “Sweden”), buildings condition=”all …” and the years 2005 and 2010. Download the Excel file.

1. Prepare your data: clean unnecessary text, covert the county names so they contain only the name of the county, i.e. “Stockholm” instead of “01 Stockholm county” and then set the proper variable names.
2. Download the relevant map of Sweden from http://gadm.org/country and load it into R
3. Import your data file to R and merge it with the map. Create choropleth maps showing the average purchase price for 2005 and for 2010. Analyze each map and then make the comparative analysis of the two maps. Interpret the results.

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,

- Submit your report using *Lab X* item in the *Submissions* folder before the deadline.
- Make sure that you or some of your group members submits the group report using *Lab X group report* in the *Submissions* folder before the deadline

If you are a speaker for this lab,

- Submit your report using *Lab X* item in the *Submissions* folder before the deadline.
- Make sure that you or some of your group members does the following before the deadline:
    - submits the group report using *Lab X group report* 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,

- Submit your report using *Lab X* item in the *Submissions* folder before the deadline.
- Make sure that you or some of your group members submits the group report using *Lab X group report* 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 PDF in this ZIP file by using the password available in *Course Documents→Password X.txt,* read it carefully and prepare (in cooperation with other group members) **at least three questions/comments/improvement suggestions per lab assignment** in order to put them at the seminar.