Examination

Linköping University, Department of Computer and Information Science, Statistics

Course code and name	732A98 Visualization
Date and time	2016-10-21, 08.00-12.00
Assisting teacher	Oleg Sysoev
Allowed aids	Book “Visualize this” by N. Yau
Grades:	A=19-20 points
	B=16-18 points
	C=11-15 points
	D=9-10 points
	E=7-8 points
	F=0-6 points

Provide a detailed report that includes necessary code, plots, conclusions and interpretations.

All R plots in your report should contain necessary graphical elements (title, axis,…) and some elements of design (color). If necessary or if it is stated in the assignment, use Inkscape for improving the plot (don’t spend much time on this).

In GGobi plots, background color should be changed to white and the default accent color should be black (see Color&Glyph settings), and the resulting plot can be copied into your report.

When plotting maps, no Inkscape processing is required.

Assignment 1 (11p)

A psychological experiment was performed to investigate the influence of scents on the ability of solving a task. File ***scents.csv*** contains information about time subjects required to complete a pencil and paper labyrinth when they were smelling a floral scent and when they were not. The variables are:

- ID
- Sex: M=male, F=female

- Smoker: Y if subject smoked, N if did not
- Opinion: "pos" if subject found the odor inherently positive, "indiff" if indifferent, "neg" if inherently negative
- Order: 1 if did unscented trials first, 2 if did scented trials first
- U-Trial 1: length of time required for first unscented trial
- U-Trial 2 : length of time required for second unscented trial
- U-Trial 3: length of time required for third unscented trial
- S-Trial 1 : length of time required for first scented trial
- S-Trial 2 : length of time required for second scented trial
- S-Trial 3: length of time required for third scented trial

1. Use ggplot to create a publication quality scatterplot that shows connection between U-Trial 2 and U-Trial 3. Perform a detailed analysis of the plot and interpret your findings. **(2p)**
2. Researchers have also got access to the file **pie.pdf** showing the distribution of the ages for the participants of the experiment. Improve this plot further in Inkscape by
    a. Changing the white pie slice color to green and edge colors to white
    b. Shortening the categories and placing them inside of the pie slices
        i. You can shorten "[31.7-48.3]" as "32-48"
    c. Adding necessary text elements
    d. Rotating the pie so that the edge between the blue and the green pie slices is vertical

    Save the resulting file as PDF and put it in the solutions folder together with your report.

    Report the main reason of why the original picture **pie.pdf** can be misleading. **(3p)**

3. Present all time variables by means of a heatmap in which the rows and columns are permuted by employing the single-link hierarchical clustering. Describe the clusters, report which variables are important in defining these clusters and make interpretations of the clusters. Find and report outliers. Create a variable Cluster_ID in your data that contains the cluster number (according to your visual findings) or "-1" if the observation is an outlier. **(3p)**
4. Visualize these new data in GGobi, create a 2D-tour involving all time variables and find a projection that shows clusters (at least two, do not spend much time on this). Brush the observations by Cluster_ID and check how this coloring is related to 2D-tour clusters and to the histograms of the variables Sex, Smoking and Opinion. Interpret your findings. Which clustering strategy (heatmap or 2D-tour) do you think is more reliable and why? **(3p)**

Assignment 2 (9p)

File **rates.csv** contains information about the currency exchange rates (standardized to [0,1]) versus Euro at various time points.

1. Compute a new variable MonthC as Month+Day/32 and subset the data to first four months. Make a shingle with 5% overlap based on MonthC and the reduced data. Use

this shingle and the reduced data in order to produce a trellis plot with 4 panels where each panel shows a boxplot of the DKK exchange rate. Analyze this plot and interpret your findings. **(3p)**

2. Produce a similar plot to step 1 where you condition on the Month instead of the shingle variable. Compare the results with step 1 and comment what advantages there are in using shingles. **(1p)**
3. File **rates2.csv** contains interpolated values of the exchange rates. Make a plot containing the original values of SEK rates (from rates.csv) and the interpolated values of SEK versus time. By looking at the plot, provide some positive and negative arguments of using interpolated values instead of the original ones for doing an animation. For which time period can the interpolation be misleading? **(2p)**
4. Use all available exchange rate values from rate2.csv to produce an animation where the exchange rates for a specific time point are visualized as a bar chart. Watch the animation then and reorder the bars so that the final video corresponds to the standards of the good animation. Motivate your choice. List also at least three noteworthy features from the resulting animation. **(3p)**

Put the resulting mp4 file in the solutions folder together with your report.