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

Laboratory work 2

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. Perception in Visualization

File **olive.csv** contains information about contents of olive oils coming from different regions in Italy. Each observation contains information about

- Region (1=North, 2=South, 3= Sardinia island)
- Area (different Italian regions)

Different acids:

- Palmitic
- …
- Eicosenoic

1. Create a scatterplot in Ggplot2 that shows dependence of Palmitic on Oleic in which observations are colored by Linoleic. Create also a similar scatter plot in which you divide Linoleic variable into fours classes (use cut_interval()) and map the discretized variable to color instead. How easy/difficult is it to analyze each of these plots? What kind of perception problem is demonstrated by this experiment?
2. Create scatterplots of Palmitic vs Oleic in which you map the discretized Linoleic with four classes to:
    a. Color
    b. Size
    c. Orientation angle (use geom_spoke())

    State in which plots it is more difficult to differentiate between the categories and connect your findings to perception metrics (i.e. how many bits can be decoded by a specific aesthetics)
3. Create a scatterplot of Oleic vs Eicosenoic in which color is defined by numeric values of Region. What is wrong with such a plot? Now create a similar kind of plot in which Region is a categorical variable. How quickly can you identify decision boundaries? Does preattentive or attentive mechanism make it possible?

4. Create a scatterplot of Oleic vs Eicosenoic in which color is defined by a discretized Linoleic (3 classes), shape is defined by a discretized Palmitic (3 classes) and size is defined by a discretized Palmitoleic (3 classes). How difficult is it to differentiate between 27=3*3*3 different types of observations? What kind of perception problem is demonstrated by this graph?
5. Create a scatterplot of Oleic vs Eicosenoic in which color is defined by Region, shape is defined by a discretized Palmitic (3 classes) and size is defined by a discretized Palmitoleic (3 classes). Why is it possible to clearly see a decision boundary between Regions despite many aesthetics are used? Explain this phenomenon from the perspective of Treisman's theory.
6. Use Plotly to create a pie chart that shows the proportions of oils coming from different Areas. Hide labels in this plot and keep only hover-on labels. Which problem is demonstrated by this graph?
7. Create a 2d-density contour plot with Ggplot2 in which you show dependence of Linoleic vs Eicosenoic. Compare the graph to the scatterplot using the same variables and comment why this contour plot can be misleading.

Assignment 2. Multidimensional scaling of a high-dimensional dataset

The data set ***baseball-2016.xlsx*** contains information about the scores of baseball teams in USA in 2016, such as:

> Games won, Games Lost, Runs peer game, At bats, Runs, Hits, Doubles, Triples, Home runs, Runs batted in, Bases stolen, Time caught stealing, Bases on Balls, Strikeouts, Hits/At Bats, On Base Percentage, Slugging percentage, On base+Slugging, Total bases, Double plays grounded into, Times hit by pitch, Sacrifice hits, Sacrifice flies, Intentional base on balls, and Runners Left On Base.

1. Load the file to R and answer whether it is reasonable to scale these data in order to perform a multidimensional scaling (MDS).
2. Write an R code that performs a non-metric MDS with Minkowski distance=2 of the data (numerical columns) into two dimensions. Visualize the resulting observations in Plotly as a scatter plot in which observations are colored by League. Does it seem to exist a difference between the leagues according to the plot? Which of the MDS components seem to provide the best differentiation between the Leagues? Which baseball teams seem to be outliers?
3. Use Plotly to create a Shepard plot for the MDS performed and comment about how successful the MDS was. Which observation pairs were hard for the MDS to map successfully?
4. Produce series of scatterplots in which you plot the MDS variable that was the best in the differentiation between the leagues in step 2 against all other numerical variables of the data. Pick up two scatterplots that seem to show the strongest (positive or negative)

connection between the variables and include them into your report. Find some information about these variables in Google – do they appear to be important in scoring the baseball teams? Provide some interpretation for the chosen MDS variable.

Submission procedure

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

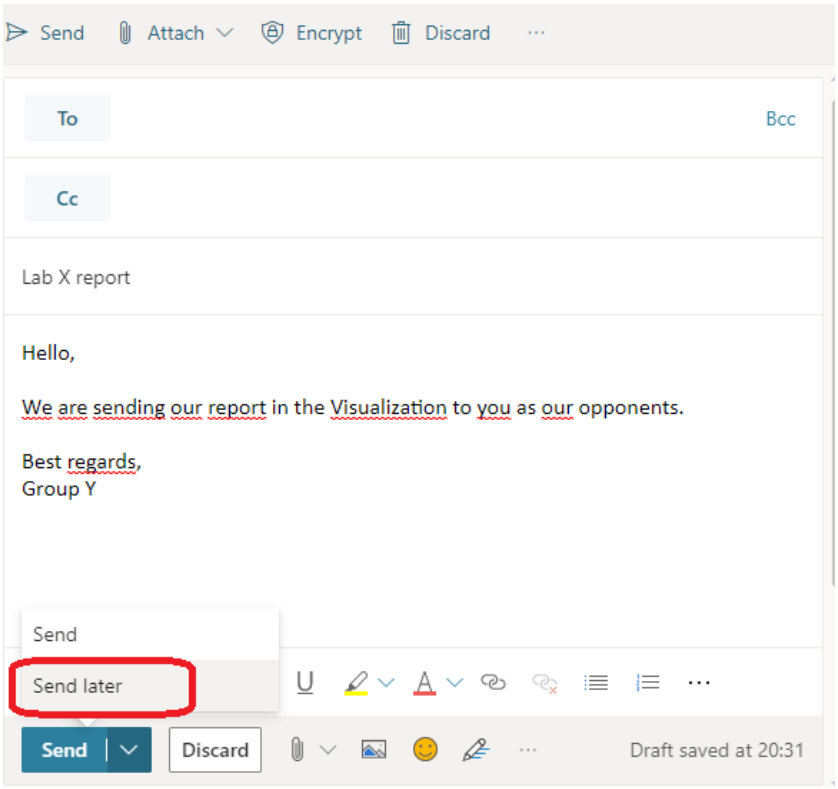
If you are neither speaker nor opponent for this lab,

- Make sure that you or your group mate submits the group report using *Lab X* item in the *Submissions* folder before the deadline. Make sure that the report contains the Statement Of Contribution describing how each group member has contributed into the group report.

If you are a speaker for this lab,

- Make sure that you or your group mate does the following before the deadline:
    1. submits the group report using *Lab X* item in the *Submissions* folder before the deadline. Makes sure that the report contains the Statement Of Contribution describing how each group member has contributed into the group report.
    2. Goes to LISAM→Course Documents→Deadlines.PDF, finds the deadline (date and time) for the current lab.
    3. Goes to LISAM→Course Documents→Seminars.PDF and find the group number of your opponent group
    4. Goes to LISAM→Course Documents→Groups.PDF and finds email addresses of the students in the opponent group
    5. Go to LISAM→Outlook app and in the Outlook web client creates a new message where you
        - Specify Lab X report as a title (X is lab number)
        - Specify email addresses of the opponents in the "To:" field
        - Attach your RMD report and accompanying data files (Note: NOT HTML!)
        - **Important: Click on arrow next to "Send" button, choose "Send Later" and specify the lab deadline as the message delivery time stamp (see figure)**

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



If you are opponent for this lab,

- Make sure that you or your group mate submits the group report using *Lab X* item in the *Submissions* folder before the deadline. Make sure that the report contains the Statement Of Contribution describing how each group member has contributed into the group report.
- After the deadline for the lab has passed you should be able to receive the RMD report of the speakers per email. 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.