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.

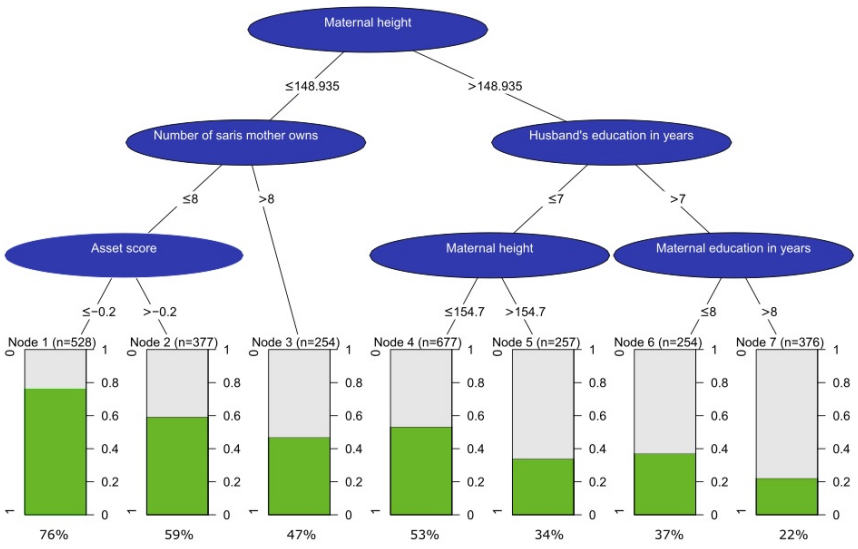

Figure 1. A tree from assignment 1.

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.

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 Linolenic. Create also a similar scatter plot in which you divide Linolenic 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 Linolenic 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 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.

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

Laboratory work 3

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. Visualization of mosquito's populations

File **aegypti_albopictus.csv** shows information about the location and detection time of two types of mosquitoes. Both *Aedes aegypti* and *Aedes albopictus* mosquitoes may spread viruses like Zika, dengue, chikungunya and other viruses but *Aedes aegypti* are more likely to spread these viruses (and therefore are more dangerous). The data file contain the following variables:

VECTOR: Identifying the species; Ae. aegypti or Ae. albopictus
LOCATION_TYPE: Whether the record represents a point or a polygon location.
POLYGON_ADMIN: Admin level or polygon size which the record represents when the location type is a polygon. -999 when the location type is a point (5 km x 5 km).
X: The longitudinal coordinate of the point or polygon centroid (WGS1984 Datum).
Y: The latitudinal coordinate of the point or polygon centroid (WGS1984 Datum).
YEAR: The year of the occurrence.
COUNTRY: The name of the country within which the occurrence lies.
COUNTRY_ID: ISO alpha-3 country codes. .
GAUL_AD0: The country-level global administrative unit layer (GAUL) code (see http://www.fao.org/geonetwork) which identifies the Admin-0 polygon within which any smaller polygons and points lie.
STATUS: Established vs. transient populations.

1. Use MapBox interface in Plotly to create two dot maps (for years 2004 and 2013) that show the distribution of the two types of mosquitos in the world (use color to distinguish between mosquitos). Analyze which countries and which regions in these countries had high density of each mosquito type and how the situation changed between these time points. What perception problems can be found in these plots?
2. Compute Z as the numbers of mosquitos per country detected during all study period. Use plot_geo() function to create a choropleth map that shows Z values. This map should have an Equirectangular projection. Why do you think there is so little information in the map?

3. Create the same kind of maps as in step 2 but use
    a. Equidistant projection with choropleth color $\log(Z)$
    b. Conic equal area projection with choropleth color $\log(Z)$

    Analyze the map from step 3a and make conclusions. Compare maps from 3a and 3b and comment which advantages and disadvantages you may see with both types of maps.
4. In order to resolve problems detected in step 1, use data from 2013 only for Brazil and
    a. Create variable X1 by cutting X into 100 piecies (use cut_interval())
    b. Create variable Y1 by cutting Y into 100 piecies (use cut_interval())
    c. Compute mean values of X and Y per group (X1,Y1) and the amount of observations N per group (X1,Y1)
    d. Visualize mean X,Y and N by using MapBox

    Identify regions in Brazil that are most infected by mosquitoes. Did such discretization help in analyzing the distribution of mosquitoes?

Assignment 2. Visualization of income in Swedish households

In this assignment, you will analyze the mean incomes of the Swedish households. Go to http://www.scb.se and choose "English" language, and in the "Search" menu type "Disposable income for households by region, type of households and age", click "Search" and then click at "Statistical Database". Select "Mean value, SEK thousands", all counties, age groups 18-29, 30-49 and 50-64, and year 2016. Download the "Comma delimited without heading" file.

1. Download a relevant map of Swedish counties from http://gadm.org/country and load it into R. Read your data into R and process it in such a way that different age groups are shown in different columns. Let's call these groups Young, Adult and Senior.
2. Create a plot containing three violin plots showing mean income distributions per age group. Analyze this plot and interpret your analysis in terms of income.
3. Create a surface plot showing dependence of Senior incomes on Adult and Young incomes in various counties. What kind of trend can you see and how can this be interpreted? Do you think that linear regression would be suitable to model this dependence?
4. Use Plotly and add_sf() function to visualize incomes of Young and Adults in two choropleth maps. Analyze these maps and make conclusions. Is there any new information that you could not discover in previous statistical plots?
5. Use GPVisualizer http://www.gpsvisualizer.com/geocoder/ and extract the coordinates of Linköping. Add a red dot to the choropleth map for Young from step 4 in order to show where we are located :)

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

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.

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

Laboratory work 4

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. High-dimensional visualization of economic data

File **prices-and-earnings.txt** shows a UBS's (one of the largest banks in the world) report comparing prices, wages, and other economic conditions in cities around the world. Some of the variables measured in 73 cities are Cost of Living, Food Costs, Average Hourly Wage, average number of Working Hours per Year, average number of Vacation Days, hours of work (at the average wage) needed to buy an iPhone, minutes of work needed to buy a Big Mac, and Women's Clothing Cost.

1. For further analysis, import data to R and keep only the columns with the following numbers: 1,2,5,6,7,9,10,16,17,18,19. Use the first column as labels in further analysis.
2. Plot a heatmap of the data without doing any reordering. Is it possible to see clusters, outliers?
3. Compute distance matrices by a) using Euclidian distance and b) as one minus correlation. For both cases, compute orders that optimize Hamiltonian Path Length and use Hierarchical Clustering (HC) as the optimization algorithm. Plot two respective heatmaps and state which plot seems to be easier to analyse and why. Make a detailed analysis of the plot based on Euclidian distance.
4. Compute a permutation that optimizes Hamiltonian Path Length but uses Traveling Salesman Problem (TSP) as solver. Compare the heatmap given by this reordering with the heatmap produced by the HC solver in the previous step – which one seems to be better? Compare also objective function values such as Hamiltonian Path length and Gradient measure achieved by row permutations of TSP and HC solvers (Hint: use criterion() function)
5. Use Ploty to create parallel coordinate plots from unsorted data and try to permute the variables in the plot manually to achieve a better clustering picture. After you are ready with

this, brush clusters by different colors and comment about the properties of the clusters: which variables are important to define these clusters and what values of these variables are specific to each cluster. Can these clusters be interpreted? Find the most prominent outlier and interpret it.

6. Use the data obtained by using the HC solver and create a radar chart diagram with juxtaposed radars. Identify two smaller clusters in your data (choose yourself which ones) and the most distinct outlier.
7. Which of the tools you have used in this assignment (heatmaps, parallel coordinates or radar charts) was best in analyzing these data? From which perspective? (e.g. efficiency, simplicity, etc.)

Assignment 2. Trellis plots for population analysis

File **adult.csv** shown data collected in a population census in 1994. The following metrics are available:

1. age: continuous.
2. workclass: Private, Self-emp-not-inc, etc.
3. fnlwgt: a population index.
4. education: Bachelors, Some-college, etc.
5. education-num: ordered Education variable.
6. marital-status: Married-civ-spouse, Divorced, etc.
7. occupation: Tech-support, Craft-repair, etc.
8. relationship: Wife, Own-child, etc.
9. race: White, Asian-Pac-Islander etc.
10. sex: Female, Male.
11. capital-gain: continuous.
12. capital-loss: continuous.
13. hours-per-week: continuous.
14. native-country: United-States, Cambodia etc.
15. Income level

1. Use ggplot2 to make a scatter plot of Hours per Week versus age where observations are colored by Income level. Why it is problematic to analyze this plot? Make a trellis plot of the same kind where you condition on Income Level. What new conclusions can you make here?
2. Use ggplot2 to create a density plot of age grouped by the Income level. Create a trellis plot of the same kind where you condition on Marital Status. Analyze these two plots and make conclusions.
3. Filter out all observations having Capital loss equal to zero. For the remaining data, use Plotly to create a 3D-scatter plot of Education-num vs Age vs Captial Loss. Why is it difficult to analyze this plot? Create a trellis plot with 6 panels in ggplot2 in which each panel shows a raster-type 2d-density plot of Capital Loss versus Education-num conditioned on values of Age (use cut_number()) . Analyze this plot.

4. Make a trellis plot containing 4 panels where each panel should show a scatter plot of Capital Loss versus Education-num conditioned on the values of Age by a) using cut_number() b) using Shingles with 10% overlap. Which advantages and disadvantages you see in using Shingles?

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.

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

Laboratory work 5

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. Text Visualization of Amazon reviews

In this assignment you will analyze feedbacks given by customers for watches Casio AMW320R-1EV bought at www.amazon.com . Files Five.txt and OneTwo.txt contain feedbacks of the customers who were pleased and not pleased with their buy, respectively.

1. Use R tools to create a word cloud corresponding to Five.txt and OneTwo.txt and adjust the colors in the way you like. Analyze the graphs.
2. Run Phrase Nets as follows:
    a. Download the software from https://www.cg.tuwien.ac.at/courses/InfoVis/HallOfFame/2011/Gruppe08/Homepage/
    b. Use the file phrase-net.zip that you have downloaded and unpack it somewhere
        i. If you are using Windows
            1. Launch the command line environment in Windows by opening the start menu in Windows and typing *cmd* in the search field and change the current directory to the "phrase-nets" directory with this kind of commands:
            ```
            z:
            cd lab5\phrase-nets
            ```
            2. In the command line environment, copy and paste the contents of run.txt
        ii. If you are using Linux
            1. Change your working directory to "phrase-nets" directory using Terminal
            2. Run "run.sh"

Note that you need Java to run this software:
https://www.java.com/en/download/

Create the phrase nets for Five.Txt and One.Txt with connector words

- am, is, are, was, were
- a, the
- at
- of

where you choose ”Filter stopwords” option.

3. When you find an interesting connection between some words, use Word Trees https://www.jasondavies.com/wordtree/ to understand the context better. Note that this link might not work properly in Microsoft Edge (if you are using Windows 10) so use other browsers.

Analyze the graphs obtained and comment on the most interesting findings, like:

- Which properties of this watch are mentioned mostly often?
- What are satisfied customers talking about?
- What are unsatisfied customers talking about?
- What are good and bad properties of the watch mentioned by both groups?
- Can you understand watch characteristics (like type of display, features of the watches) by observing these graphs?

Assignment 2. Interactive analysis of Italian olive oils.

In this assignment, you will continue analyzing data **olive.csv** that you started working with in lab 2. These data contain 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

ATTN: All diagrams that support your judgments should be included to the report

In this assignment, you are assumed to use Plotly without Shiny.

1. Create an interactive scatter plot of the eicosenoic against linoleic. You have probably found a group of observations having unusually low values of eicosenoic. Hover on these observations to find out the exact values of eicosenoic for these observations.
2. Link the scatterplot of (eicosenoic, linoleic) to a bar chart showing Region and a slider that allows to filter the data by the values of stearic. Use persistent brushing to identify

the regions that correspond unusually low values of eicosenoic. Use the slider and describe what additional relationships in the data can be found by using it. Report which interaction operators were used in this step.

3. Create linked scatter plots eicosenoic against linoleic and arachidic against linolenic. Which outliers in (arachidic, linolenic) are also outliers in (eicosenoic, linoleic)? Are outliers grouped in some way? Use brushing to demonstrate your findings.
4. Create a parallel coordinate plot for the available eight acids, a linked 3d-scatter plot in which variables are selected by three additional drop boxes and a linked bar chart showing Regions. Use persistent brushing to mark each region by a different color. Observe the parallel coordinate plot and state which three variables (let's call them influential variables) seem to be mostly reasonable to pick up if one wants to differentiate between the regions. Does the parallel coordinate plot demonstrate that there are clusters among the observations that belong to the same Region? Select the three influential variables in the drop boxes and observe in the 3d-plot whether each Region corresponds to one cluster.
5. Think about which interaction operators are available in step 4 and what interaction operands they are be applied to. Which additional interaction operators can be added to the visualization in step 4 to make it even more efficient/flexible? Based on the analysis in the previous steps, try to suggest a strategy (or, maybe, several strategies) that would use information about the level of acids to discover which regions different oils comes from.

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.

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

Laboratory work 6

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. Network visualization of terrorist connections

Files **trainData.dat** and **trainMeta.dat** contain information about a network of the individuals involved in the bombing of commuter trains in Madrid on March 11, 2004. The names included were of those people suspected of having participated and their relatives.

File **trainMeta.dat** contains the names of individuals (first column) and Bombing group (second column) which shows "1" if person participated in placing the explosives and "0" otherwise. According to the order in this file, persons were enumerated 1-70.

File **trainData.dat** contains information about connections between the individuals (first two columns) and strength of ties linking (from one to four):

1. Trust--friendship (contact, kinship, links in the telephone center).
2. Ties to Al Qaeda and to Osama Bin Laden.
3. Co-participation in training camps and/or wars.
4. Co-participation in previous terrorist Attacks (Sept 11, Casablanca).

1. Use visNetwork package to plot the graph in which
    a. you use strength of links variable
    b. nodes are colored by Bombing Group.
    c. size of nodes is proportional to the number of connections (function strength() from IGRAPH might be useful here)
    d. you use a layout that optimizes repulsion forces (visPhysics(solver="repulsion")).
    e. all nodes that are connected to a currently selected node by a path of length one are highlighted

    Analyse the obtained network, in particular describe which clusters you see in the network.

2. Add a functionality to the plot in step 1 that highlights all nodes that are connected to the selected node by a path of length one or two. Check some amount of the largest nodes and comment which individual has the best opportunity to spread the information in the network. Read some information about this person in Google and present your findings.
3. Compute clusters by optimizing edge betweenness and visualize the resulting network. Comment whether the clusters you identified manually in step 1 were also discovered by this clustering method.
4. Use adjacency matrix representation to perform a permutation by Hierarchical Clustering (HC) seriation method and visualize the graph as a heatmap. Find the most pronounced cluster and comment whether this cluster was discovered in steps 1 or 3.

Assignment 2. Animations of time series data.

The data file ***Oilcoal.csv*** provides time series about the consumption of oil (million tonnes) and coal (million tonnes oil equivalents) in China, India, Japan, US, Brazil, UK, Germany and France. Marker size shows how large a country is (1 for China and the US, 0.5 for all other countries).

1. Visualize data in Plotly as an animated bubble chart of Coal versus Oil in which the bubble size corresponds to the country size. List several noteworthy features of the investigated animation.
2. Find two countries that had similar motion patterns and create a motion chart including these countries only. Try to find historical facts that could explain some of the sudden changes in the animation behavior.
3. Compute a new column that shows the proportion of fuel consumption related to oil: $Oil_p = \frac{Oil}{Oil+Coal} * 100$. One could think of visualizing the proportions Oil_p by means of animated bar charts; however smooth transitions between bars are not yet implemented in Plotly. Thus, use the following workaround:
    a. Create a new data frame that for each year and country contains two rows: one that shows Oil_p and another row containing 0 in Oil_p column
    b. Make an animated line plot of Oil_p versus Country where you group lines by Country and make them thicker

    Perform an analysis of this animation. What are the advantages of visualizing data in this way compared to the animated bubble chart? What are the disadvantages?
4. Repeat the previous step but use "elastic" transition (easing). Which advantages and disadvantages can you see with this animation? Use information in https://easings.net/ to support your arguments.
5. Use Plotly to create a guided 2D-tour visualizing Coal consumption in which the index function is given by Central Mass index and in which observations are years and variables are different countries. Find a projection with the most compact and well-separated clusters. Do clusters correspond to different Year ranges? Which variable has the largest contribution to this projection? How can this be interpreted? (Hint: make a time series plot for the Coal consumption of this country)

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

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.