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. Equirectangular 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

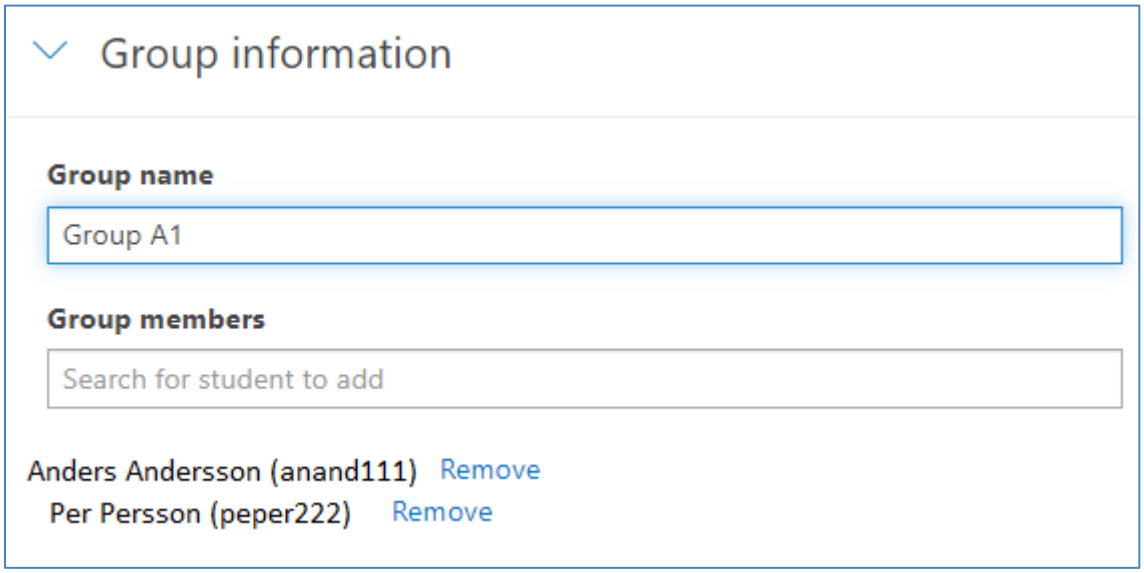
In this assignment, you will analyze the mean incomes of the Swedish households. Go to http://www.scb.se and choose "English" language, and in the "Search" menu type "Disposable income for households by region, type of households and age", click "Search" and then click at "Statistical Database". Select "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 in Plotly 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 in Plotly 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 plot_geo function with trace "choropleth" 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

When submitting the report, remember to specify in LISAM your group name and all group members that wrote the report!



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

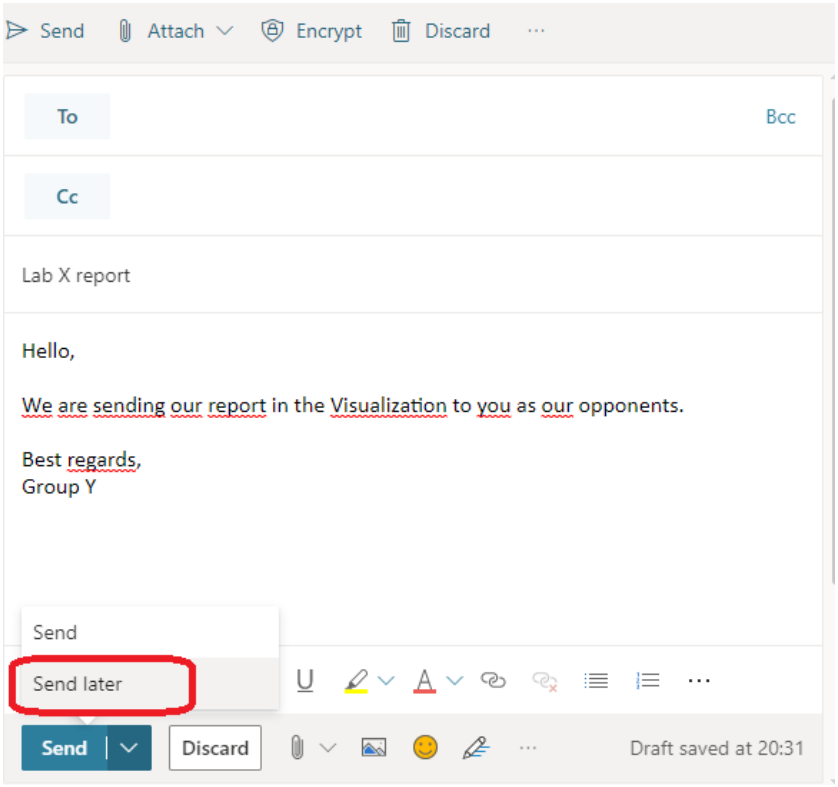
If you are neither speaker nor opponent for this lab,

- 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 below)**



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.