University of Trento --- M.Sc. Data Science

Data Visualization Lab Exam
11 September 2020

Create a Jupyter and/or R Notebook, named
name_surname.{ipynb,Rmd}

Answer the questions (in a Markdown cell/ as plain text) and solve the exercises listed hereafter:

1. [0-5 points] Describe in detail the meaning of the visual encoding elements in the following infographic describing the Average Yearly Rainfall in Costa Rica. Indicate also pros and cons of the infographic itself in terms of data visualization aspects.

2. [0-5 points] Discuss in detail the issue of *chart value sorting,* providing graphical examples for some of the discussed points (different from those included in the lecture notes).

3. [0-5 points] Describe the differences between linear and non-linear algorithms in dimensionality reduction, and provide an example highlighting such differences.

4. [0-7 points] Using the datafiles economy.csv and demography.csv, prepare a choropleth map of the italian regions including at least one data column from both datafiles.

5. [0-7 points] Consider the synthetic dataset madelon.csv, collecting 2000 samples (rows) described by 500 features (columns) and belonging to two classes (1 or -1) defined by the last column. Using the PCA dimensionality reduction planar (2D) projection on the 500 describing features, color the samples according to their class and discuss if the patients and controls can be well separated into two distinct groups in the projection. Repeat the same process with the UMAP algorithm and compare the two results.

6. [0-7 points] Using the datafile sankey.csv, try to replicate the following Sankey plot.
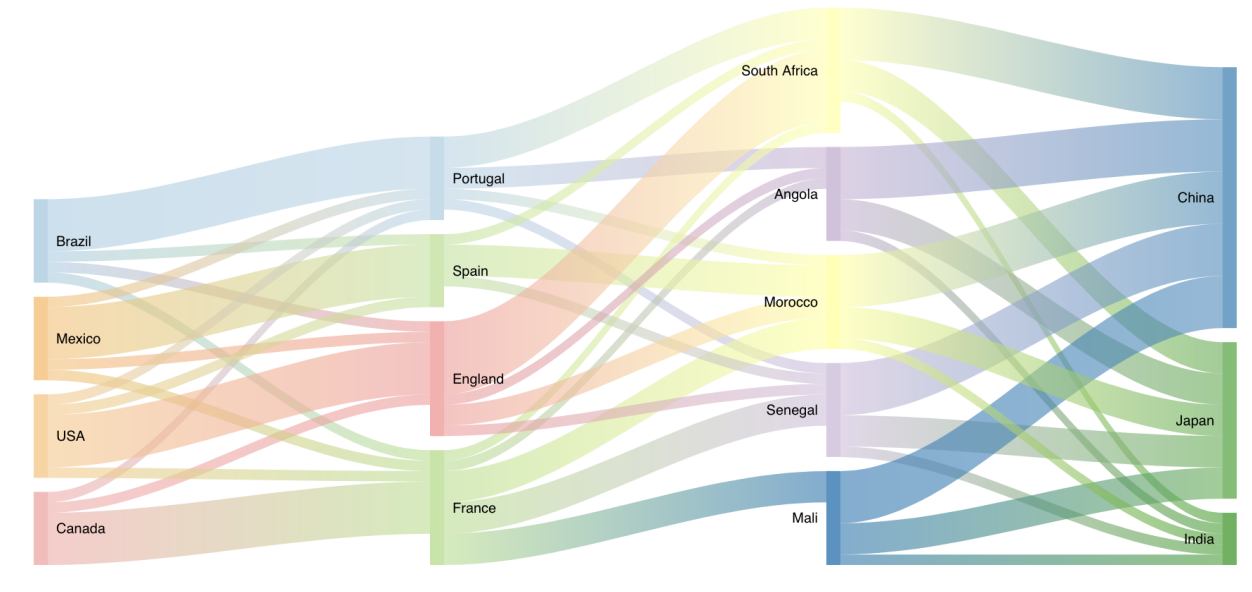
Email the notebook(s) to *giuseppe.jurman@unitn.it* and please **wait for confirmation of correct receipt of the files before leaving the room**.

Notes:
- Exam is passed when at least 18 points are earned.
- If more than 30 points are achieved, the corresponding mark will be "30 cum laude"
- Use of the internet is allowed, but the candidate is expected to work individually.