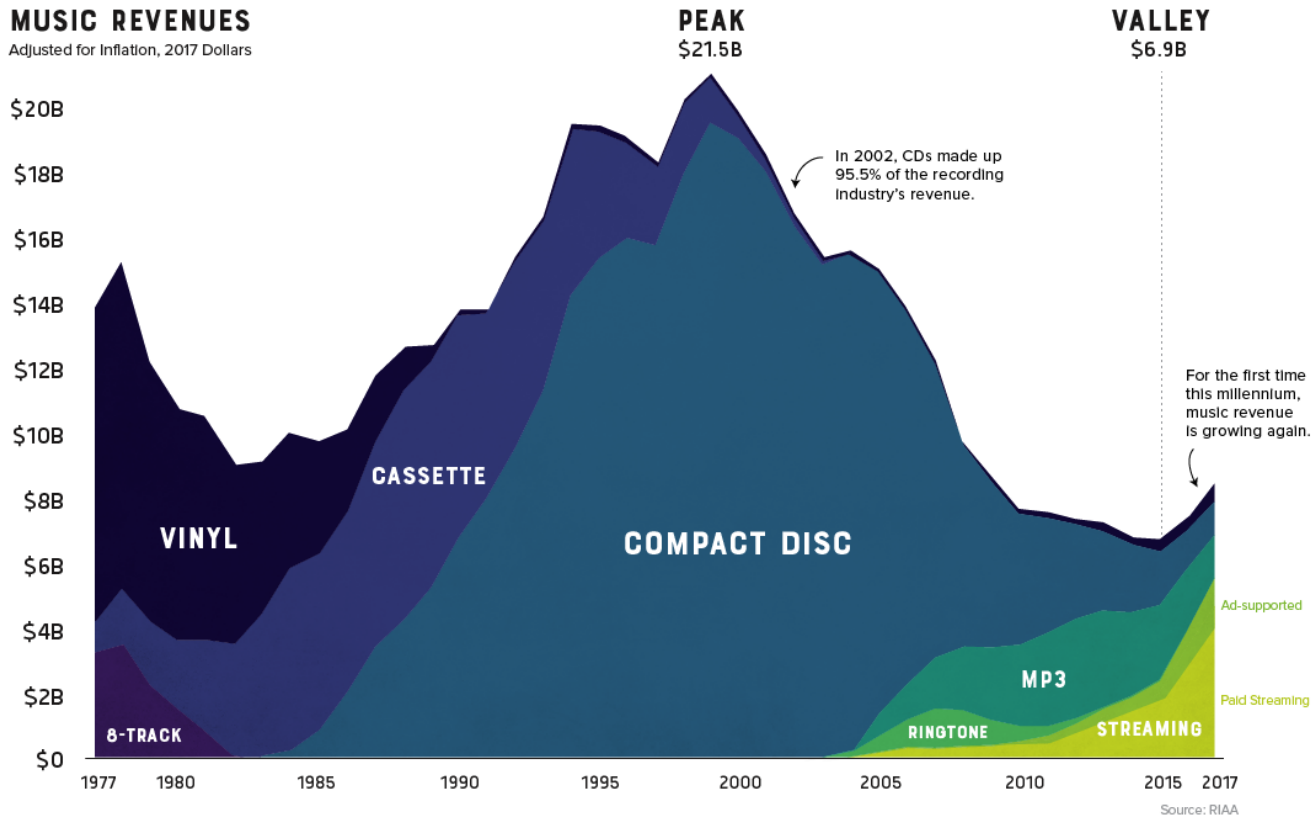
University of Trento --- M.Sc. Data Science

Data Visualization Lab Exam
12 July 2019

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 three elements of graph understanding (perceiving/interpreting/comprehending), first in general and then applying these principles to the reading of the following plot



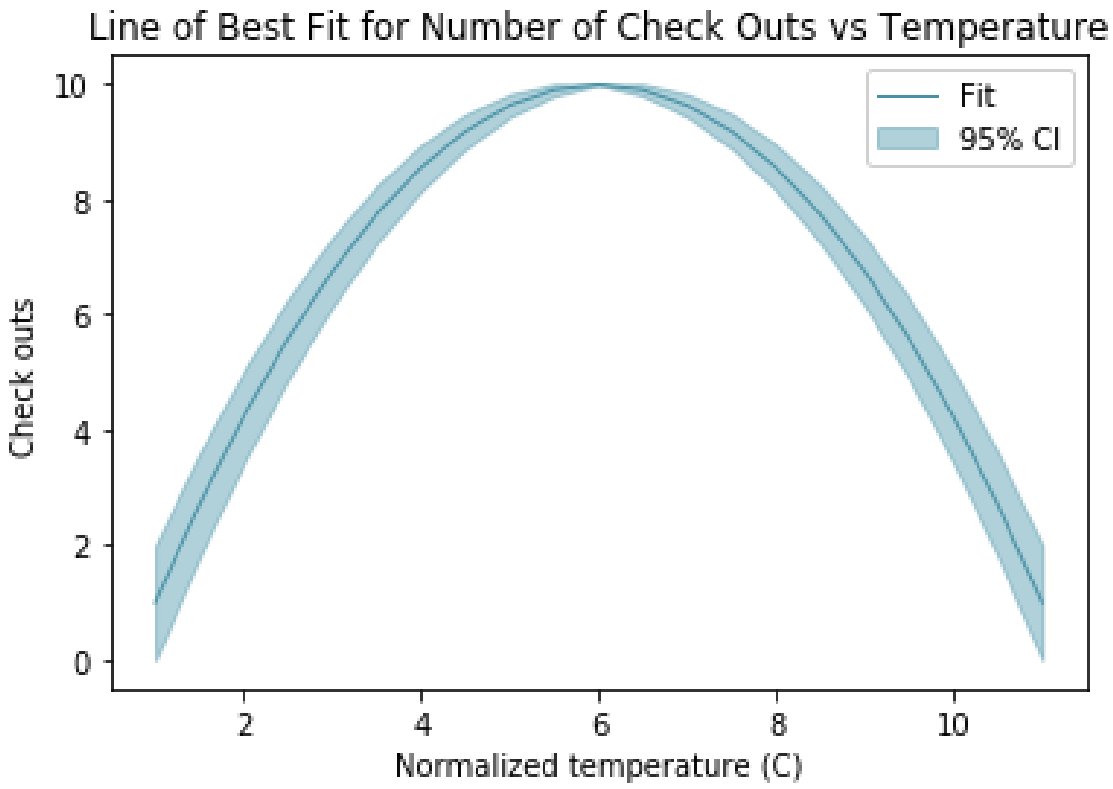
2. [0-5 points] Describe in detail at least two chart types where a statistical quantity is graphically represented on a geographical map, clearly listing the differences between the two charts. For each of them, show an example.

3. [0-5 points] Describe the UMAP algorithm, highlighting its pros and cons; find a case study where UMAP is more appropriate than PCA.

4. [0-7 points] Find a data visualization (one or more plots) for the data in http://bit.ly/2FZQybB (or a subset of the data); justify your choice and find/comment one or more non-trivial relations among the data visible in the plot(s).

5. [0-7 points] The Wisconsin Breast Cancer dataset http://bit.ly/2xAuP5t includes data of 569 breast cancer patients. In detail, first column is PatientID, second column is the Diagnosis (Malignant/Begnin), while the remaining column 3-32 consists of 3 groups of 10 histological descriptors of tumoral cells; for each descriptor, its Average, StandardError and Worst Case are listed. Using the 30 descriptor columns, project using t-SNE the 569 patients on a 2D space and color them according to their diagnosis. Are the two different diagnoses well separated in the 2D space?
   Repeat the same procedure using only the last 10 Worst Case columns: is the separation between diagnoses improving or worsening?

6. [0-7 points] Using the datafile http://bit.ly/2JuGylu , replicate the following plot



Email the notebook(s) to *giuseppe.jurman@unitn.it* and please wait for correct reception 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.