

CS3205/CS6426 Data Visualization for Analytics Applications

Rosane Minghim

FEBRUARY 6, 2022

Assignment 1

Visual Exploration of Survey Data

Introduction to Multidimensional Projections

Lab Report - Submission date Feb 21st 2022

There are two tasks below, divided in sub-tasks, related to survey data visualizations and visualization of multidimensional data via projections. The lab report should include samples of the visualizations employed as well as your interpretation of the data through them. The [format](#) for the Lab Report is presented at the end of this document.

The tasks you should perform for this report are:

1. **Exploring Human Development Report (HDR) Data** In this task you should access a data visualization tool to explore HDR 2020 (HDR data up to 2019).

Data Sets.





- (a) Collect the data from [HDR data site](#). Choose 8 to 10 variables you wish to observe (besides the HDI index) and add a **Label** column with the level of development mapped to numbers (very high - 1; high - 2, medium - 3, low - 4). Alternatively, you may use the data set provided in Canvas, Lab section, but do not forget to mention in your report that this is the one you chose to use.
- (b) HDI series. Download from Canvas - also Lab section, the HDI series data set (hdi.csv). It contains the full series of Human Development Indices. Write code to add to the file the Human development Level for each year. That means each HDI column will have a added column with the level (very high - 1; high - 2, medium - 3, low - 4), named appropriately. This data process can be shared between colleagues, but you must mention in your lab report where you your data that from.
HDI is divided into four tiers: very high human development (0.8-1.0), high human development (≥ 0.7 and < 0.8), medium human development (≥ 0.55 and < 0.7), and low human development (below 0.55).

For the above data sets, There are two sets of tasks - free exploration and specific observations, described below:

- (a) **Free Exploration.** Generate Visualizations and use them to have a global panorama of the data. Make your own free observations as to the general distribution of HDI and of your main variables of interest. You are free to explore any aspects of the data in this part. For the report, the requirements are:



- i. At least 4 different visualizations. If you wish, it can be 2 different visualizations, each of them with two different visual mappings (color, texture, etc.). Present those visualizations together with their explanations and your interpretation.

- ii. Report of at least four patterns found, together with a description of how you came to that conclusion. Examples of patterns are (but definitely not limited to): "In general, there seems to be a distribution of quality of life in the world in this way: ****"; "Country [X] and [Y] present a different panorama from their neighbours in regard to [your observed variable]. I hypothesise that this might be due to [your hypothesis]".
- (b) **Specific Observations.** Plot variables in your data set against country - for the first task below, and against each other for the remaining tasks, using any visualization you deem fit (scatter plots, histograms, curves, heatmaps, etc.).
 -  • If you order a particular variable of interest by country, do you see spurious values (for instance, country or countries you expected that variable to be higher or lower than it is? - Explain. Do that for at least two variables of the data set.
 -  • Observe possible correlations between different indicators in HDR (attributes, variables), employing comparative visualizations for all countries. Report on two visible correlations and two seemingly un-correlated variables. Requirements: employ at least 5 different variables for these four plots. Employ HDI as a variable only **once**, as its value is naturally dependent on many other variables.
 -  • Find two expected patterns and two unexpected patterns in correlations, using previous visualizations and new ones if needed. Once you find them, draw hypotheses for the patterns found. Here are some examples, but you can draw your own: which countries report values different (higher or lower) than you expected? Which countries deviate from the correlations found? Which countries lead in value in chosen variables? Which countries present a relation between two variables that do not match the general trend? Which countries have low values of your chosen variables? Requirements for the report: For this task, present at least 5 visualizations and their explanations, as well as the patterns found and your hypotheses.
 -  • The usefulness (or not) of the various available visualizations. While at it, change at least two of the visualizations you used in the previous task and verify what types of insight they can provide for this kind of data. Are they better or worse than the ones used previously? What new insight they provide? What previous insight is not as clear now?

2. **Employing projections to visualize multidimensional numerical data.** Use VisPipeline to visualize a collection of images and a collection of documents. Employ LSP, T-sne and a third projection of your choice. Choose one that allows for document upload. From this exercise, try to draw the following categories of observations:

- (a) **Exploration of an image collection.** The data set Corel in the data archive is a vector space composed of image features, for a collection of photographs and drawings. There are 10 labels and 1000 items in the data set. Apply the different projections with different parameters to that data set, making observation in regards to segregation of labels and the differences between projections. Give examples of images that seem to be difficult to discriminate from other labels.
- (b) **Exploration of a document collection.** The data named CBR is a collection of article abstracts from certain areas of knowledge. It comprises a vector space model extracted from 600+ documents. Load the data for CBR and explore neighborhoods and labels in all three projections. Try changing projections and parameters where possible and make observations after changes. Is there gain or loss for the task of segregation? How would you compare the separation between labels in your explorations? Register your choices for each set of observations.
- (c) **Other Data Sets.** Explore two other data sets and draw similar observations from the ones above.
- (d) **Projections of HDR.** Remove from HDR 2020 data the columns related to HDI. Using the data file description in the Lab session of the Module, build a '.data' file. In the file,

use the country name as **ID** and the level of development as **class** in the file. Show the visualizations using the three choices of projections, and explain what you see.

Report Format

The Report will have an Introduction and a Conclusion, as well as one section for each of the tasks above. In each section, you should make clear what your actions were, what the observations on the data are, and present pictures to clarify how you got to those observations. You should use minimum of half a page and maximum two pages for each sub-task (a), (b),..., font size 11, excluding pictures. Please keep pictures small. You may add 'live' links to larger pictures on a repository on the web. You should upload one '.zip' file containing: the report in **pdf**, all and only data sets USED in the report, all workbooks, code or link to tools and visualizations employed in the report. In short, all materials needed to reproduce your tasks should be included in the submission file. Please take a look at the 'writing_tips.pdf' file uploaded with this specification and follow those that apply to your report.

The format for the report is presented below:

Identification (Your name, Module Identification, Report Title, Year)

1. **Introduction**

Explain the problem, the data sets, and the goal of the report and your understanding of the use of visualization for both survey and multidimensional numerical data.

2. **Data Description**

Describe all the data sets used in the report as well as the transformations you performed on them. Employ the name of the file when referring to them. When submitting the report, include all the files used **before and after** transformation.

3. **Task 1**

Results of Task 1

4. **Task 2**

Results of Task 2

5. **Conclusions**

Explain your conclusions on the data sets, on the exercises and on the use of visualization for these tasks.

6. **References**