Introduction to Data Visualization

Hasso-Plattner-Institut
Fachgebiet Computergrafische Systeme
Prof. Dr. Jürgen Döllner

Winter Term 2021/2022 – Assignment Sheet # 3

Objectives

With these assignments you will learn how to

- calculate basis statistics measures;
- visualize based on parallel plots and stacked charts;
- evaluate sample visualizations.

For your programming tasks, you will use JavaScript for preprocessing and visualization. You are expected to comment and document your code in a focused and clear manner.

Your solutions must be uploaded via Moodle by **January 06, 2022, 9am (UTC+1)** as one ZIP file that contains all answers and source files. The naming convention for this ZIP file is **sheet3_<group_name>.zip**.

Instructions

Implementation As visualization framework we use *D3.js* and its respective language JavaScript; for details we refer to its documentation[1].

The implementations with D3.js should be implemented as a web page based using the framework. Make sure that relative paths are used. Supplementary exercise-specific data should be organized in a sub-folder called "data". Color schemes, in particular, can be explored and loaded via colorbrewer2.org[2]. Local provisioning for debugging and testing can be done using a local web server such as the http-server by node[3]. The server can be installed with the following command:

- npm install http-server -g

In order to load e.g. CSV files you need to start a local server by

- http-server -a 127.0.0.1 -o

in your working directory containing both the sources and the data.

Pair Programming On these assignments, you are encouraged (not required) to work with a partner provided you practice pair programming. Pair programming "is a practice in which two programmers work side-by-side at one computer, continuously collaborating on the same design, algorithm, code, or test." One partner is driving (designing and typing the code) while the other is navigating (reviewing the work, identifying bugs, and asking questions). The two partners switch roles every 30–40 minutes and, on demand, brainstorm.

Violation of Rules A violation of rules results in grading the affected assignments with 0 points.

- Writing code with a partner without following the pair programming instructions listed above (e.g., if one partner does not participate in the process) is a serious violation of the course collaboration policy.
- Plagiarism represents a serious violation of the course policy.

[1] D3.js documentation
[2] COLORBREWER 2.0
[3] Node.js®

Exercise 3.1: Basic Descriptive Statistics (4 Points, Theory)

For the given data set *Obesity.csv* regarding age, height, and weight of a group of persons, calculate the following basic statistical measures and derive insights:

a) Mean value for each data dimension

b) Standard deviation for each dimension

c) Co-variance matrix for all combinations of dimensions

d) Three derived written statements based on analytically looking at the numbers about the insights the data and numbers tell us.

Exercise 3.2: Normalized Area Chart for Energy Consumption (8 Points)

The data set *EnergyConsumption.csv* contains a snapshot of data taken from a solar plant system installed for a building. The data is sampled in 10 min intervals (time series) and values are measured in kW. Conceptually, we consider four key values:

- E_C: current total energy consumption of the building ($E_C \geq 0$)
- E_S: current energy production by the solar plant ($E_S \geq 0$)
- E_B: current energy taken from battery (i.e., $E_B < 0$) or charged in battery (i.e., $E_B > 0$)
- E_N: current energy demand from ($E_N > 0$) or supply into general power network ($E_N < 0$)

The equation $E_N + E_S = E_C + E_B$ describes an idealized energy balance. Examples: During daytime the solar plant generates 5 kW, the building consumes 0.5 kW, the battery receives 3.0 kW, and 1.5 kW is fed into in the general power grid, i.e., $-1.5 + 5 = 0.5 + 3.0$. At night the building consumes 0.5 kW and takes the whole 0.5 kW from the battery, i.e., $0.0 + 0.0 = 0.5 - 0.5$.

Create a normalized stacked area chart that visualizes these values. Ensure to visualize the total amount of energy consumed by the building, and the proportions that represent energy production and consumption.

Notes: (a) If the solar plant generates more energy than the building consumes, the battery gets charged. Since its charging capacity is limited, the remaining energy is simultaneously fed into the general power grid. (b) When the building consumes more energy than is delivered by the solar plant, it intends to get energy from the battery (if it is charged enough) and, if needed, from the general power grid (e.g., during peak consumption). (c) If the battery is not charged and no solar energy can be utilized, energy is drawn exclusively from the general power grid.

Exercise 3.3: Interactive Parallel Coordinates Plot (8 Points)

The data set *Wine.csv*[4] contains 178 different wines described by thirteen numerical characteristics. Furthermore, the data points are classified in three categories according to their quality.

Create a Parallel Coordinates Plot (PCP) that shows the relationships between dataset-specific characteristics. Each data element (wine) should be represented as a polyline. The wine quality class should be mapped to the line color.

To improve readability, the scaling and arrangement of the axes should be adequately chosen. In addition, an interaction technique shall be implemented that allows the user to filter the data points by specifying a selected range of values for the attribute values of an attribute.

Notes: There are D3.js-based examples on the internet for PCPs that will be helpful in implementing this exercise.

Exercise 3.4: Evaluation of Visualization Examples (2 Points)

In the zip file of this exercise sheet you will find four screenshots of visualization examples from different applications and domains. Describe and explain a selected aspect that conceptually represents a weak point in the respective visualization example.

[4] scikit-learn developers, wine_dataset