Assignment 1: Analysing Visualizations

Course: Interactive Data Visualization (Spring 2020)

Student: Diana Crowe (student nr.: 012056152)

Description: In this assignment, you will select and analyse existing visualization designs. The goal is for you to start thinking about the purpose of a visualization and look at the visualization critically.

Tasks

1. Read the paper -- Heer, J., Bostock, M. and Ogievetsky, V., 2010. A tour through the visualization zoo. Queue, 8(5), p.20.
2. Pick two types of visualizations from the paper. Each type visualizes a different type of data. For instance, you can pick stacked graphs for time-series data, and flow maps for geographical data;
3. Find from the web or literature one example of each type of the visualizations you picked;
4. Write a 3-8 pages report to discuss each example from the following perspectives:
    - Purposes of the visualization; (1 point)
    - What information you have learned from the visualization; (1 point)
    - What information is hidden or not easy to tell from the visualization; (1 point)
    - Benefits and drawbacks of this type of visualization. (2 points)

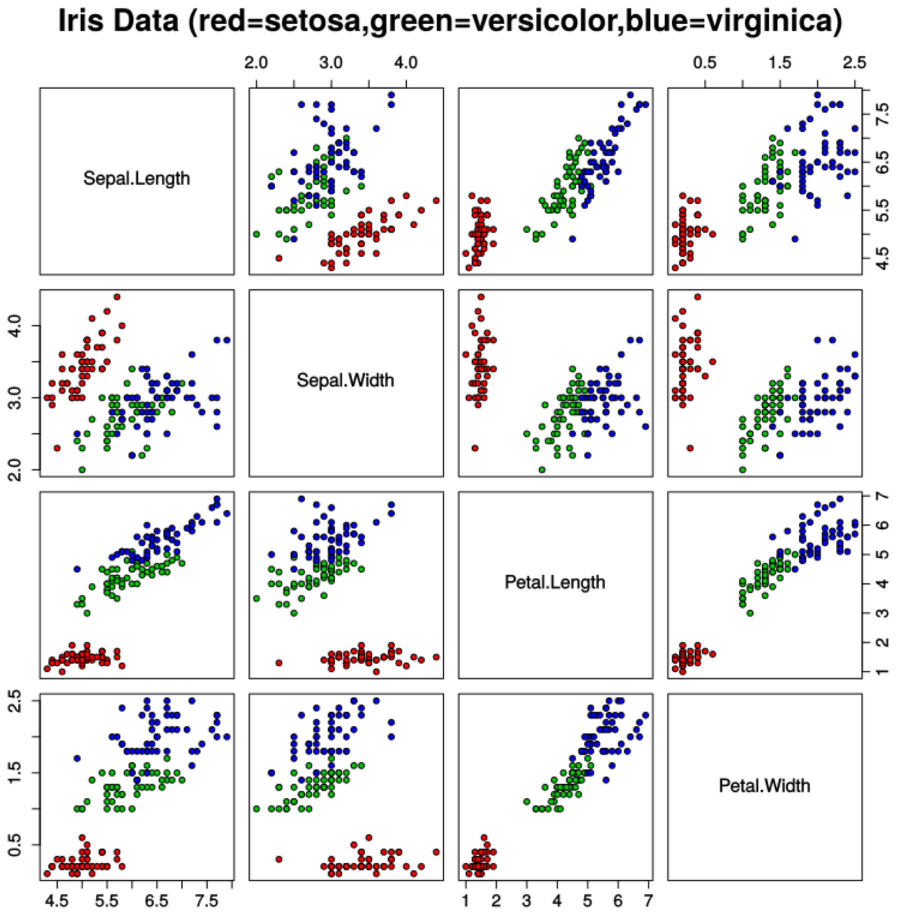
SPLOM (Scatter Plot Matrix)

In a scatter plot matrix (SPLOM) we have a matrix with independent variable being studied. Each column/row index stands for a variable and the matrix elements are scatter plots of the variable index combination. This way we visualize pairwise relations between variables and can immediately see correlations. This is preferable to having a single multi-dimensional plot which would be much more difficult to interpret (or which couldn't even be represented due to not enough dimensions available).

Extra interaction techniques can be used to explore the data, such as brushing-and-linking—in which a selection of points on one graph highlights the same points on all the other graphs.

Example: scatter plot matrix of the Iris flower data set or Fisher's Iris data set

https://en.wikipedia.org/wiki/Iris_flower_data_set



1. Purpose of the visualization

The visualization tries to help distinguish (make differences obvious) between 3 species of Iris flowers (setosa, versicolor and virginica) by comparing pairwise 4 different measured characteristics (sepal length and width, and petal length and width).

2. What information have we learned from the visualization

- Independent variables: Sepal length, sepal width, petal length, petal width
- Dependent variables: species (setosa, versicolor, virginica)
- Instead of three distinct clusters (corresponding to the 3 species), we can only really distinguish two clusters – one with Iris setora, and the other with the Iris versivolor and Iris virginica.
- There is a positive linear correlation between the sepal width and length in all 3 species
- There is a very strong linear correlation between the petal width and length in all 3 species
- Setosa is the smallest species of Iris (it has the smallest length and width of petals)
- Iris Versicolor and Iris virginica show a positive linear correlation between the widths and lengths of sepals versus petals, but that correlation is absent in the case of Iris setosa

3. What information is hidden or not easy to tell from the visualization

It is quite difficult to distinguish between Iris versicolor and Iris virginica.

4. Benefits and drawbacks of this type of visualization

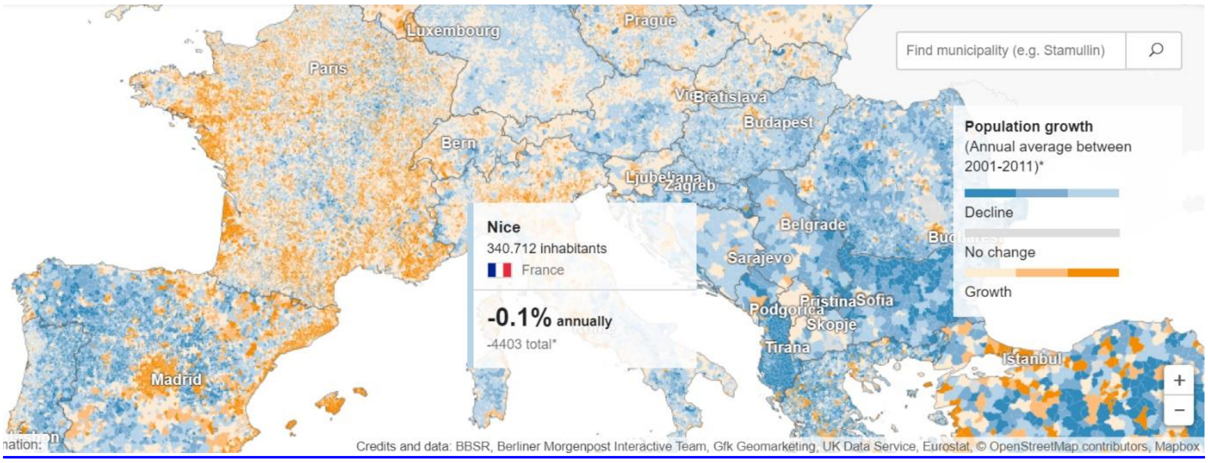
- A scatterplot is convenient: we can look at all pairwise correlations in one place
- Most take care if choosing the colours (high contrast is good; pay attention to colours that can be confused by colour-blind people)
- We have to make sure that the variables being plotted are independent, else we shall see false correlations
- It illustrates correlations between variables (positive, negative or null).
- It is useful to have the visual representation to confirm correlations. It is possible to numerically obtain (for example) a Pearsson's coefficient value indicating a strong linear correlation, only to see the graph and realise that such is not really the case.
- There is limited information that we can obtain from the graphic – and it is mostly qualitative.

Choropleth Maps

In a choropleth map the data is organized in a geographical area display. Often colour-coding is used.

Example: Interactive Europe map from the Berliner Morgenpost**:**

https://interaktiv.morgenpost.de/europakarte/#6/43.716/6.108/en



Nice
340.712 inhabitants
France
-0.1% annually
-4403 total*

Credits and data: BBSR, Berliner Morgenpost Interactive Team, Gfk Geomarketing, UK Data Service, Eurostat, © OpenStreetMap contributors, Mapbox

Purposes of the visualization

The map shows where the population in Europe is growing (blue areas), where there is no change (grey) and where it is declining (yellow to orange areas)**.**

The map view starts with all of Europe displayed and it is possible to zoom in to see greater detail in our area of interest. Panning the mouse over the map gives extra information. In the screenshot above I hovered the mouse over Nice to get: Name of the town/city, number of inhabitants, country it belongs to, percentual growth and how many inhabitants that percentage corresponds to. One can also type in the name of a location (upper right corner) to find its information on the map (upon testing, that feature doesn't actually work!).

What information you have learned from the visualization

In the area that I selected from the map:

- Good UI design – the detailed information box always pops where the mouse is (which is presumably where our eyes are looking at)
- The diverging colour scheme emphasizes the extremes of population growth and decline
- Choosing small units allows us to see regional patterns
- Most of Portugal and Spain are experiencing population decline. There is however growth in the capital areas of both countries (greater Madrid area and Lisbon area), the coastal areas of Spain that have the nicest beaches, and the Spanish Balearic islands in the Mediterranean (Majorca and Minorca).

Diana Crowe (student nr.: 012056152)

Hypothesising: People might be moving into the big cities and capitals in order to find better jobs/a better life, and people might be moving into the east-coast areas and islands for the lifestyle (retirees) or job opportunities (mostly tourism related).

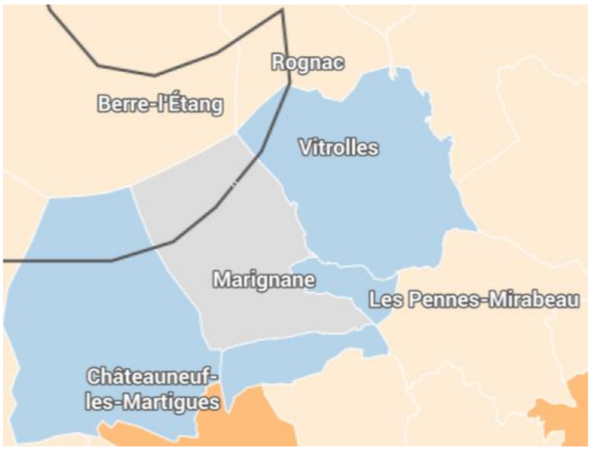
- France in general shows a lot of population growth except for smaller areas in the interior. Coastal areas are again very popular.
- Most eastern countries show a decline in population (emigration? Lower birth rates?)
- Germany shows a lot of decline in population.
- Italy shows mostly decline, with some growth around the capital Rome and in big-city or touristy areas (Florence, Venice, Milan, Bologna…)

Side note: the maps covers the time period 2001-2011, which is before the European Migrant Crisis (which was officially 2015-2019).

What information is hidden or not easy to tell from the visualization

- In order to see smaller countries/locations (Monaco, San Marino, or even pinpoint a city like Nice) one needs to zoom in quite a lot. The “search” functionality in the upper right corner doesn’t actually work…
- There are country boundaries but no regional boundaries making regional generalizations difficult (unless you are acquainted with the geography of the country).
- The labels are not always helpful. For example, there are labels for the cities but not for the countries or regions. At a glance, for generalization purposes, it would be mor useful to have country and region names.
- We may tend to overestimate growth areas by misinterpreting grey as light yellow.

  The colour grey is used for “no change” in population. This is much closer in shade to the smaller “growth in population” than it is to the light blue of the smaller “decline in population”. In the detail below (zoomed-in screenshot) Marignane had net 0% growth which in total numbers was minus 38 people/year and yet looks more similar to a growth area in colour.



Benefits and drawbacks of this type of visualization

- Good visualization for generalizations, but not so great for details or subtle differences. The colours can be tricky to tell apart when they are blended in.
- They can give a false impression of abrupt change at the boundaries of shaded units
- Colours:
  - It is not obvious how the intervals between the colours and the intervals between the values in the data are related. When grouping data intervals to attribute to a discrete colour-step, detail is lost. Patterns in the map become easier to recognize, but it is harder to compare the exact values of regions with each other. Colour gradients give more precision but are not as easy to "read". One has to sacrifice either quick readability or precision.
  - Must choose the right colour scheme. In maps we can have three different kinds of colour schemes: Sequential (e.g. from bright blue to dark blue), diverging (e.g. from red via white to blue) and qualitative/categorical (e.g. one green colour, one blue colour). *Sequential* colour schemes work best for emphasizing high values, e.g. for unemployment rates. D*iverging* schemes emphasize more both extremes of the scale, e.g. too show the difference in votes between two competing parties.
  - Must use colour-blind-friendly colours.
  - If too many colours are used, it gets very confusing to read the map
- Sometimes some regions are too small to be shown in a map
- One has to use normalized values (proportions, etc) instead of raw total data values (such as population) to produce a density map. This is extra work.
- The perception of the shaded value can be affected by the underlying shade of the area of the geographic region.
- Good when displaying just one variable, which can also be the difference between two variables (eg. the rate of populations growth in a time interval, or the change of the unemployment rate from last year to this year).
- Not good for seeing correlation between values.
- Good for relative data (eg. better to show the rate of unemployment rather than displaying the total number of unemployed people without knowing what is the total population of the area)
- You need to have good labels since the reader might not have intricate geographical knowledge of the area displayed
- It's a balancing act between choosing the smallest units possible (so as to not lose too much data detail) and not get the display too messy with lots of segmented colour ("does this region have more orange or blue? It's all checkered!")