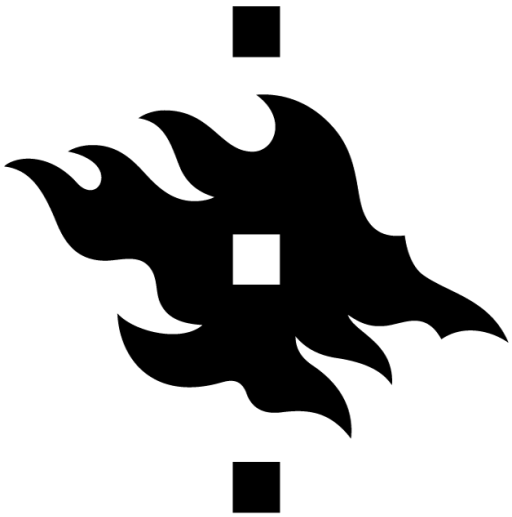

Analyzing Visualizations

Assignment 1
Interactive Data Visualization

Enrico Buratto

University of Helsinki

Faculty of Science

Academic Year 2021-2022

Contents

1 First visualization

The first visualization I chose for it to be analyzed is a plot of the famous IRIS dataset[1], retrieved from coderzoclumn.com[2]; the plot I'm referring to is displayed in Figure 1. The IRIS dataset is a multivariate dataset containing 50 samples from each of three species of Iris, a flower; every sample consists of four features: sepal length, sepal width, petal length and petal width.

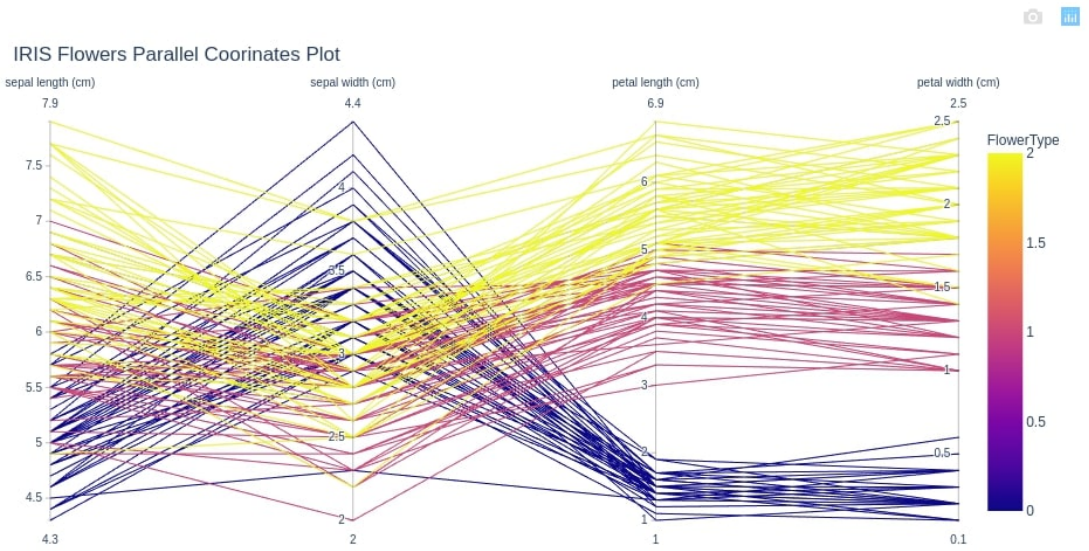

Figure 1: Parallel coordinates plot for IRIS dataset.

1.1 Brief explanation

This plot is a **parallel coordinates** visualization, a particular type of statistical distribution plot; this means that, instead of plotting every pair of features in two dimensions, we plot all of them into parallel axis and we connect them with polylines. Each of these polylines represent one sample of the database; if these lines cross each other, this usually means that some type of inverse correlation exists between the two features. Vice versa, when there are no intersections we will have some type of direct correlation or not correlation at all.

1.2 Purposes

The main purpose of this plot is to highlight relations which occur between the four features on the samples of the database, like every other parallel coordinates plot. In particular, with this plot the author probably wanted to find a pattern in the data, *i.e.* if the samples from the three types of flower have some feature value in common or if the data is just random. The author probably is also searching for inverse relations between the different features, for instance if the sepal length is

inversely proportional with the sepal width. However, these considerations are done in the next subsection.

1.3 Inferred information

Even if this particular plot is a little bit cluttered and the colors are not the best that could be chosen (as we'll see later in this section), some information can be inferred.

First of all, we can notice that the variable values are quite sparse for some features but more concentrated on others. We can observe these phenomena, for example, in the sepal width and petal length axes: in the former, the distribution of values is quite sparse, with values of "flower type 0" which vary from 2 cm to 4.4 cm on a scale that goes from 2 cm to 4.4 cm. In the latter, however, we can notice a pretty concentrated distribution, with feature values for the same flower which vary from 1 cm to 2 cm on a scale that goes from 1 cm to 6.9 cm.

We can also notice some outliers: for instance, there is a sample of "flower type 0" that has a sepal width of 2.3 cm, while the average is around 3.5 cm. Other outliers are visible for "flower type 2", though are less clear due to color mixing.

Besides this, we can still notice some patterns in the data: for instance, we can see that flowers of type 0 have usually smaller sepal length than the others, while they have a higher sepal width. Remaining on this, we can also notice that flowers of this type have both smaller petal length and petal width of the others. Similar information can be inferred also for the other types of flower.

As a last remark, we can notice that sepal length and sepal width are in inverse correlation with each other, and the same applies for sepal width and petal length; we can notice this from the crossing polylines. For the same reason, we can say that petal length and petal width are in direct correlation, since the groups of lines (lines for type of flower) don't intersect.

1.4 Benefits and drawbacks

The main benefit of this visualization in general is the high dimensionality it can represent: in fact, if we had to represent these correlations with other two-dimensional plots (for instance with index charts), we would have needed a graph for each couple of features. This might could have worked with this particular dataset, since it's composed by only four features, but with higher dimensional datasets this would be a total nightmare.

Other advantages that derive from this are the high scalability and the compactness of the visualization: if we had to add another feature, in fact, the graph size would stay almost the same, and overall we can represent a high amount of data in a relatively small space.

The main drawback of this visualization is the overlaying of data lines for common data values: as we can also see from this particular plot, it is sometimes difficult to infer some particular pattern, and this is due to the clutterness that comes with a high amount of lines.
Another cause of this are, in my opinion, the colors: these have to be chosen wisely, because too similar color are likely to confuse the reader. However, two solutions can be implemented to solve this problems: changing the axes order and choosing the colors so that they are different from each other. The latter is trivial, and the former ensures that lines don't cross too many times.

A final problem that regards only this particular plot is the legend: even if it's technically correct, it's not trivial to understand which type of flower corresponds to which color. Personally, I would improve this graph modifying the legend on the right from a continuous spectrum of colors to a more clear legend that states that blue is flower 0, magenta is flower 1 and yellow is flower 2.

2 Second visualization

The second visualization I chose for it to be analyzed is a map showing some information about the 2020 presidential elections in the USA, retrieved from viewsoftheworld.net[3]; the map is displayed in Figure 2.

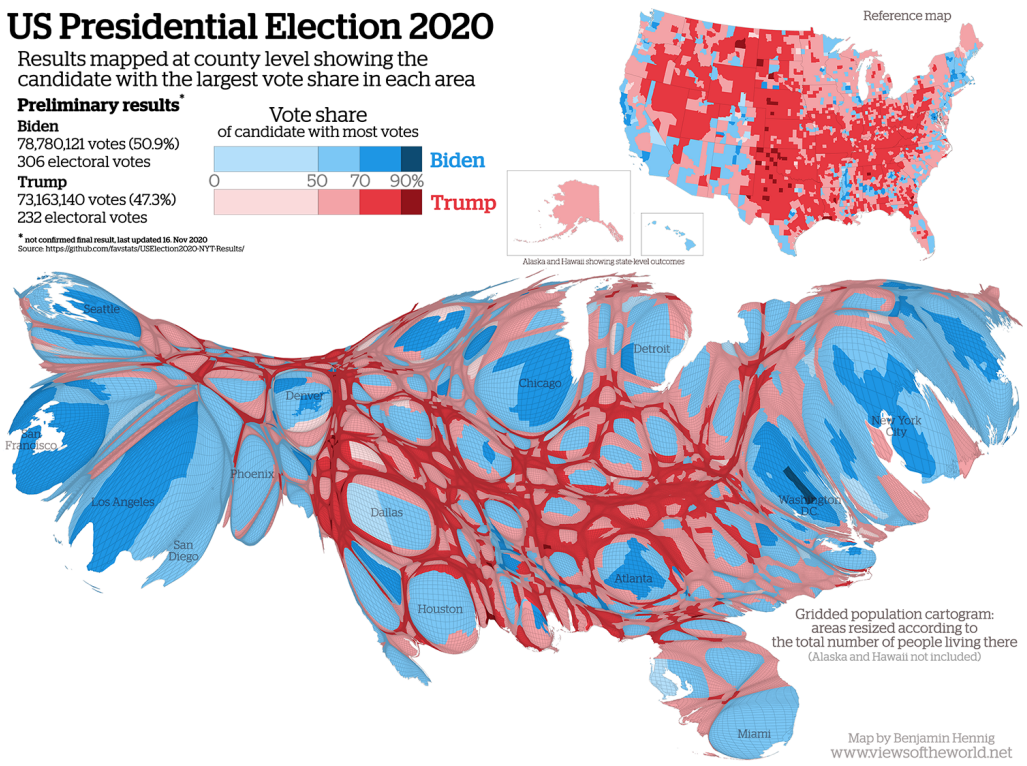

Figure 2: Cartogram for USA 2020 presidential elections.

2.1 Brief explanation

This visualization is a **cartogram**, a particular type of map; like other types of maps, it is meant to represent geography-related data and it relies on altering the geographic size to be proportional to a selected variable, for instance time and population. The main peculiarity of this type of visualization is, then, that it distorts the shape of geographic regions in order to encode a data variable.

There are many techniques to draw cartograms, and this one in particular is a *continuous cartogram* also known as *Gastner-Newman cartogram*. This technique allows the size of the area to be equalized so that the regions are visually comparable[4]; in this visualizations the shapes are usually well preserved, though the areas result often distorted.

However, a second type of visualization is combined with the cartogram in order to represent another axis of data: the color is, in fact, used to represent the vote share (as we will see in the next subsections). This makes this visualization an hybrid of cartogram and **choropleth** map, which is defined as a type of map that uses intensity of color to correspond with an aggregate summary of a geographic characteristic withing spatial enumeration units.[5]

2.2 Purposes

The main purpose of this plot is to represent two different information: the first, which is actually represented by the cartogram itself, is the number of votes that a candidate received for each state of the United States of America during the presidential elections of 2020: a higher size of a state area corresponds to a higher number of voters for the candidate who won on that state.

The second information is represented by the choropleth, which is combined with the cartogram. In every state, in fact, different intensities of colors are used to represent different vote shares.

2.3 Inferred information

Many information can be inferred from this map, and not only from the cartogram part of it.

First of all, it is quite easy to distinguish the areas where Biden and Trump won, and this is also one of the purposes of the plot itself: the blue states are the ones where Biden won, and the red states are the ones where Trump won. However, some previous knowledge is needed, since there are no state names but only important cities: for instance, there is no name on California, but Los Angeles, San Francisco and San Diego are reported instead. Even ignoring this, it is still quite hard to get an idea of where a candidate won or lost, and this is mostly true for the central states where the population size is lower. However, this last consideration is an information

itself: from the cartogram we can state that the population, *i.e.* number of voters, is more concentrated on the coasts.

It is not easy to understand how many people voted for the winning candidate in each state; however, this is not the main purpose of a cartogram, while it is more important to get an intuition of the differences in population size between the states, where by population we mean the voters for the winning candidate. As stated above, we can easily infer that more central states are less populated or less likely to vote for a candidate.

Another information, which comes this time from the superimposed choropleth, is the one regarding the vote share of candidates. For instance we can easily state that, in California, between 70% and 90% of the voters voted for Biden, and in Washington D.C. more than 90% voted for him. The same does not apply for republican voters, since the population size is smaller and therefore harder to interpret; this problem, however, is addressed in the next subsection.

For the sake of completeness, as anticipated, there are also other information that can be inferred from this visualization: on the top left of the image the preliminary results are reported, including number of votes for both candidates, percentage of them and number of electoral votes. On the top right side of the image, a reference map is reported; this map consists in only a choropleth from which we can infer for example that Trump won in most of the central states, while Biden won in the coastal ones.

2.4 Benefits and drawbacks

The main benefit of cartograms is that they are easy to understand and make a strong impact on the reader: the size, in fact, is often the most intuitive visual variable for representing a quantitative feature. However, some precautions must be taken: in this particular cartogram, for instance, is really difficult to understand the boundaries between the states and the states themselves; if I may give an opinion, I would add also the state names to the map, and I would emphasize the border with (for example) a thicker line between one and the other.

Moreover, cartograms in general have also other drawbacks: the main disadvantage is that it changes the visual representation of geography. With this type of visualization, in fact, geographical accuracy is often almost wrong, due to the design itself of the technique.

Coming to the choroplets, which are not the main focus of this analysis but should be taken into consideration since the visualization contains also two of them, the main advantages are that they are usually easy to prepare and really easy to understand: even if the size is often more effective, in fact, also colors play an important role in instantaneus data visualization.

However, also choroplets have their disadvantages; the main one is that they are

more likely to give a false impression of sudden change at the boundaries of shaded units, and it can be overall difficult to distinguish between shades (in the case of this particular choropleth, this is particularly true with the republican states in the center).

References

[1] Iris Flower Dataset, retrieved from https://en.wikipedia.org/wiki/Iris_flower_data_set

[2] How to plot parallel coordinates plot in python, retrieved from https://coderzcolumn.com/tutorials/data-science/how-to-plot-parallel-coordinates-plot-in-python-matplotlib-plotly

[3] US Presidential Election 2020, retrieved from https://www.viewsoftheworld.net/?p=5777

[4] Gastner-Newman Cartogram, retrieved from https://www.arcgis.com/home/item.html?id=4b8c9ce99a5749e298bb96366692f35d

[5] Choropleth Map, retrieved from https://en.wikipedia.org/wiki/Choropleth_map