Exploratory data analysis

Visualize! Visualize! Visualize! Oh, and don't waste ink!

Guy Maskall Follow
Jul 15 · 5 min read

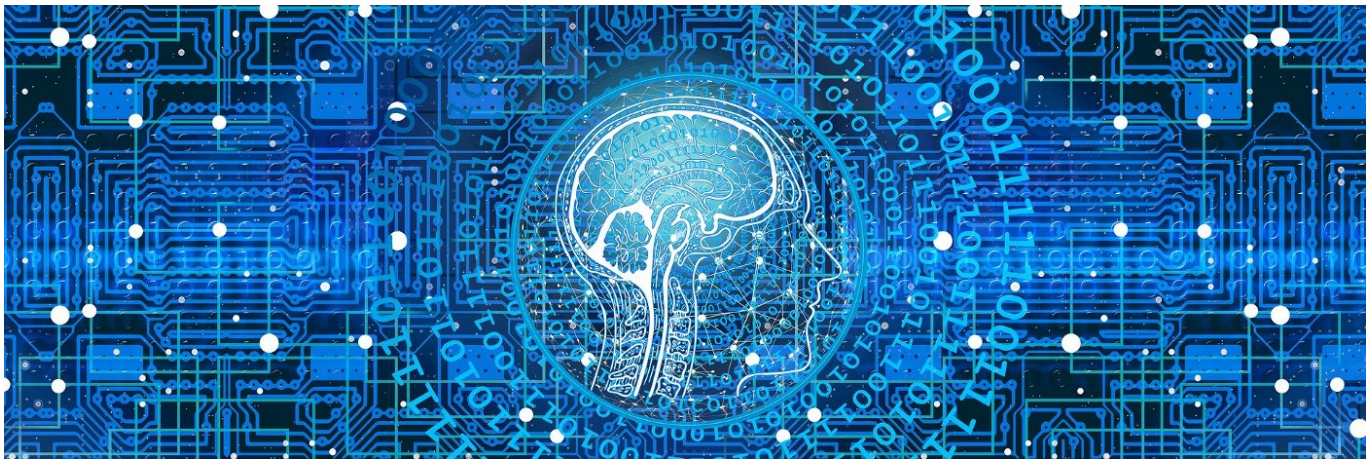

Image by Gerd Altmann from Pixabay

Data quality

We previously wrote about issues in data quality. Certainly, many issues can be spotted quickly and clearly right at the start, those missing river level values for example (even if you're not immediately sure how to handle them). But many issues will only come to light once you dig a little deeper into the data and start to consider the distributions and relationships within and between variables. A person's height, for example, might very reasonably be somewhere between 50 cm (toddler) and 1.80 m. Their weight may also very reasonably be anywhere between 10 kg and 100 kg. But only when considering both together could you realise there's a "toddler" weighing 100 kg or a tall adult weighing 15 kg. You suddenly have some outliers on your hands! Thus, exploratory data analysis (EDA) is a step when data quality issues are still very much being explored.

Complex relationships

We just brought up an example where you need to compare two variables in unison in order to spot outliers. By implication this was easy; you'd just plot one variable on the x-axis and the other on the y-axis and see what "sticks out". What if you had age as well? Well, there are graphing tools that allow you to define a z-axis as well, and even rotate the axes

dynamically to visualize the data from different directions. What if you had salary as well? It's getting tricky now, isn't it? You could start plotting the different combinations of variables as x and y over multiple graphs. But even that's only considering pairwise relationships.

What if your data had 10 variables? 20? 1000? Fortunately, there are a number of techniques to reduce the dimensionality of your data and visualize patterns. One common technique is principle components analysis (PCA), although there are many others.

Visualizations

A skill we can't emphasise enough when it comes to EDA is data visualization. And this doesn't always have to be a figure! If it's a reasonably small number of summary statistics, then a table can often be the most effective, and efficient, way to communicate. We don't tend to print out so much these days, but imagine how much ink you'd use printing out a moderate sized bar chart that merely conveyed three data values. Perhaps that would be a good candidate to present in a simple table.

There are more and more plot types readily available in packages now, offering many and varied ways to represent your data. But just because they exist, doesn't mean you need to use them. Don't be the datavis equivalent of that guy who gives a powerpoint presentation using every single animation and sound effect he could find just because they were there! Think carefully about what message you want to draw out in your figure and craft it accordingly. Some basic types to get you started:

- scatter (or point) plots — great for showing relationships between two variables. Effective use can be made of point colour, size, and even shape. Give some consideration to which features you map to which aesthetic. Don't map a feature with 20 distinct values to shape unless you're confident your reader has a very sharp eye! Do you even want to have 20 different values represented anyway? Perhaps just colour the top five categories and lump the rest into 'Other'. Distinct colours can be informative for discrete (categorical) features. Beware having too many! Tone (of a single colour, e.g. black through to ever lighter shades of grey) can be useful for adding an additional continuous variable to your figure. Many graphics libraries will choose between colours or tones according to data type. Take care what data type your plotting function thinks a feature is! A 1/0 for True/False is a classic example that can appear represented as a

tone, when it would be clearer as distinct colours. Above all, explore different options that bring out *your* message effectively and without redundancy. Play around with the alpha value if you have a dense mass of points, and ask yourself whether you really need to put every data point on the page — could you show the same message about the data distribution by plotting one in every thousand points?

- bar plot —a useful favourite for showing a set of discrete values, e.g. counts of categories. Don't colour each category separately if the x-axis is also the category just because it makes a pretty rainbow! For example, say the x-axis is month and the y-value is butterfly count. Don't colour the bars by month. But do consider colouring by season, perhaps. That could add something. Or perhaps it would be relevant to colour according to whether the month was particularly wet or dry. That would be an effective use of colour. If you have so many categories (on the x-axis) that the labels are illegible, ask yourself whether you really need to show all of them. Perhaps just a figure showing the top twenty categories gets your point across? And here we come to the "waste of ink" point. If you have two categories, say male and female, then perhaps just put them in a table.

- histogram — great for summarising distributions. Do explore different bin widths to pull out just the right amount of structure in the data that makes sense. No, I can't tell you what that is.

- line plots — great for plotting a continuous sequence of values. Avoid them if intermediate x-values don't make sense, such as when dealing with categories. The eye will naturally want to assume the trail of ink joining points is a valid interpolation. Reconsider their use if you don't have many x-values. Line plots comprising three vertices are not very informative. A personal bugbear is lines that comprise many values (points), but they form something very, very close to a straight line. Does that line plot comprise two points only, one at each end, in which case it's a poor plot that gives an illusion of a high correlation, or are there 1000 points, which is much stronger evidence for (a very suspicious) correlation. Would a scatter plot be better here? Or at least a plot with points and lines! A single plot with a number of lines coloured by different categories can be an effective way to highlight different trends between the categories.

This is an incredibly abridged selection of plot types, but they are good stalwarts that, with some careful thought, you can make effective use of time and time again. Above all, and this

cannot be stressed enough, **think about what it is you're trying to communicate in your plot** and try to effectively convey it, and only it, with a minimum of ink and extraneous fuss and bother.

About this article

This is the fourth article of a linked series written to provide a straightforward introduction to getting started with the data science process. You can find the introduction here, the previous article here, and the next article in the series here.