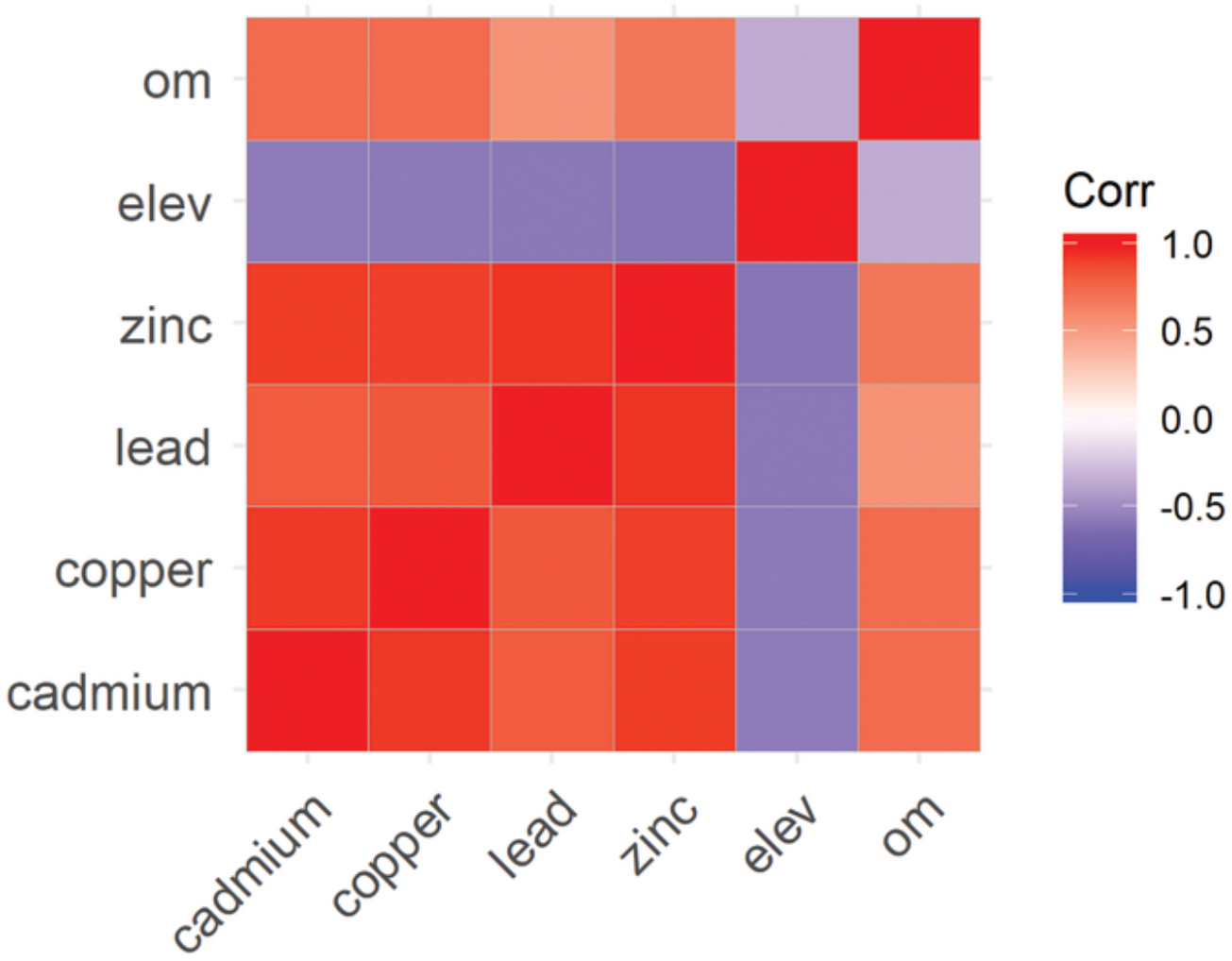
Science Societies

Tips for initial data exploration and visualization

By Emma Grace Matcham Ph.D. candidate, Lauren Schwarck M.S. student

July 29, 2021



Plot of correlation values between topsoil heavy metal concentrations, organic matter, and elevation of soil sample locations in the floodplain of the river Meuse

---

One of the most difficult tasks after collecting measurements in the field or lab is finding your starting point for data analysis. While strategies for data visualization can differ based on personal preferences and data types, we hope you find some of the suggestions below helpful in your initial data exploration. The example figures in this article were made using R, but the general concepts for selecting and interpreting graphs apply regardless of software.

Explore the Data

Before digging into your dataset, think about organization and whether your data is effectively structured for analysis. This could include moving all necessary data into the same file, clearly labeling columns, and storing necessary metadata. For further details regarding effective data management, refer to the CSA News article from April 2021, “Tips for Effectively Managing Your Data.”

After determining your data structure, move on to evaluating your data set for completeness. Take inventory of any missing data points or potential outliers. Depending on your experimental design and field of investigation, tactics for handling missing values or outliers may vary. When addressing either missing values or outliers, reach out to experts or statisticians in your field, investigate common methods for handling these situations, and always document your procedures for transparency and consistency.

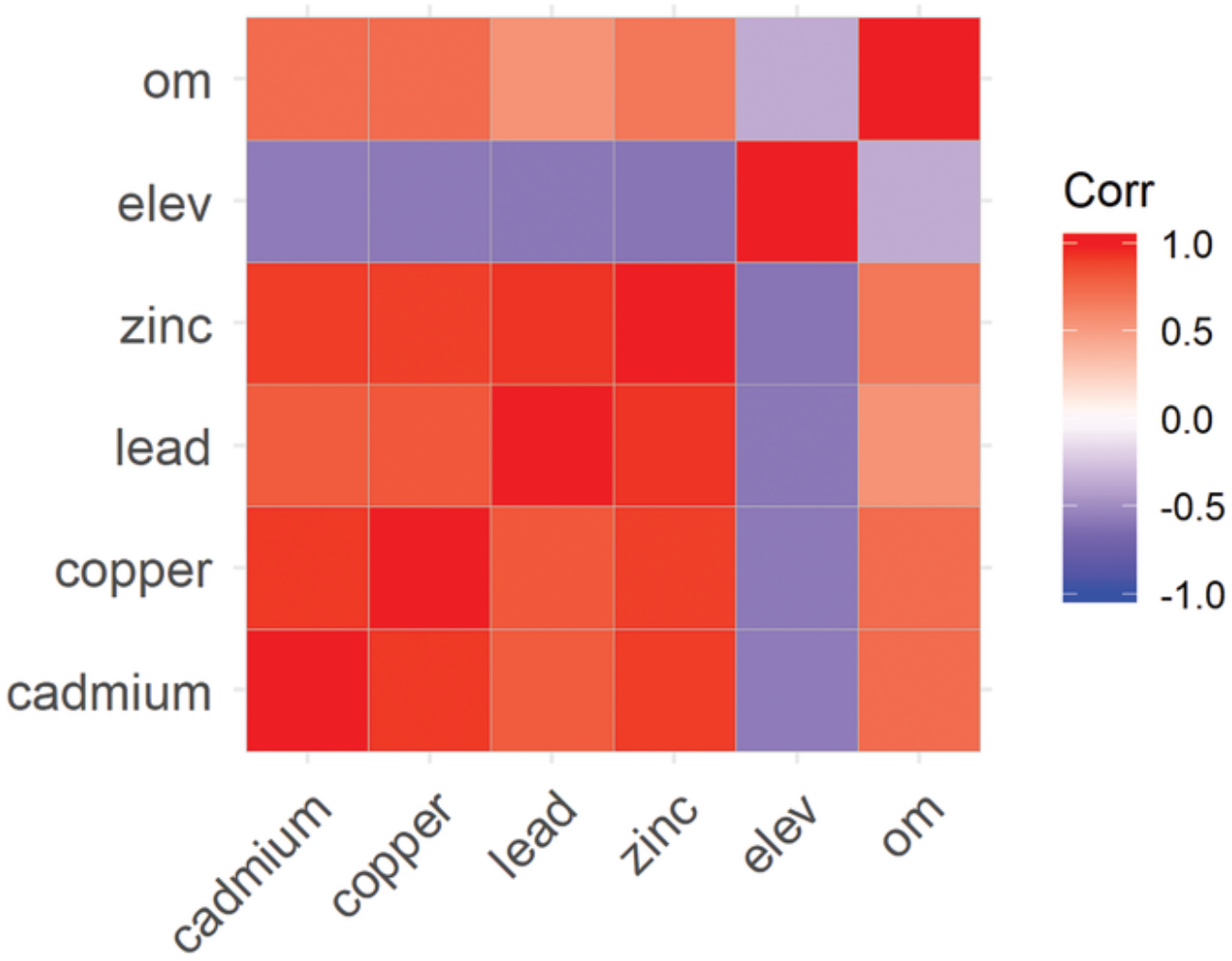
Focus on Variables of Interest

Depending on your research questions and study design, you may or may not know what your variables of interest are when you begin your analysis. In a traditional experiment that applies various treatments to measure outcomes, you probably identified variables of interest when you designed your experiment and planned out measurements to collect. If you already have your variables of interest identified, you can move on to plotting your data (see next section).

Alternatively, you may be performing an observational study or analyzing surveys. Compared to experiments, observational and survey studies tend to have variables with differing levels of importance to addressing the research question. You might know some of the variables in the data set are likely to be important or interesting for answering your research questions, but others might simply be extraneous.

Correlation tables can provide insight into which variables might be useful in answering your research question. A correlation value gives you a measure of the strength of the linear relationship between two variables in your data set, and many statistical software packages compute these values easily for all pairs of variables in your data set. You can also plot correlation tables to visually assess patterns—we like making correlation table plots using the R package *ggcorrplot*, but there are many options. When you read your correlation table, start by identifying variables that are highly correlated with the variable(s) you might want to predict in your model.

For example, this correlation plot (Figure **1**) was made using the *meuse* data set in R, originally published by Burrough & McDonnell in **1998** and available in the R package *sp.* Elevation (elev) is negatively correlated with most variables in the data set, and the heavy metals are positively correlated with each other and with organic matter (om) although the correlation with organic matter is less strong.



Plot of correlation values between topsoil heavy metal concentration, organic matter, and elevation of soil sample locations in the floodplain of the river Meuse.

Take note of any variables that are highly correlated with each other since most multivariate linear modeling methods require predictor variables to be independent of each other. Highly correlated predictor variables in the same model causes multicollinearity, which can reduce the power of your final model by lowering the precision of the coefficient estimates.

For some research questions (such as, “How accurately can I predict my dependent variable with all currently available data?”), using correlation tables to limit the number of input variables you consider may be less helpful. You're probably considering building a model that includes every single variable that improves the predictive

strength and might consider methods like random forest analysis or other black-box machine-learning techniques that can provide predictions without multicollinearity concerns, but don't specify relationships between variables in the model. Even so, correlation tables can still be a useful tool for assessing whether the model output makes sense. You would expect variables that have higher correlation with your dependent variable to improve model fit to a higher degree than variables with low correlation—you would want to further investigate variables that deviate from this trend.

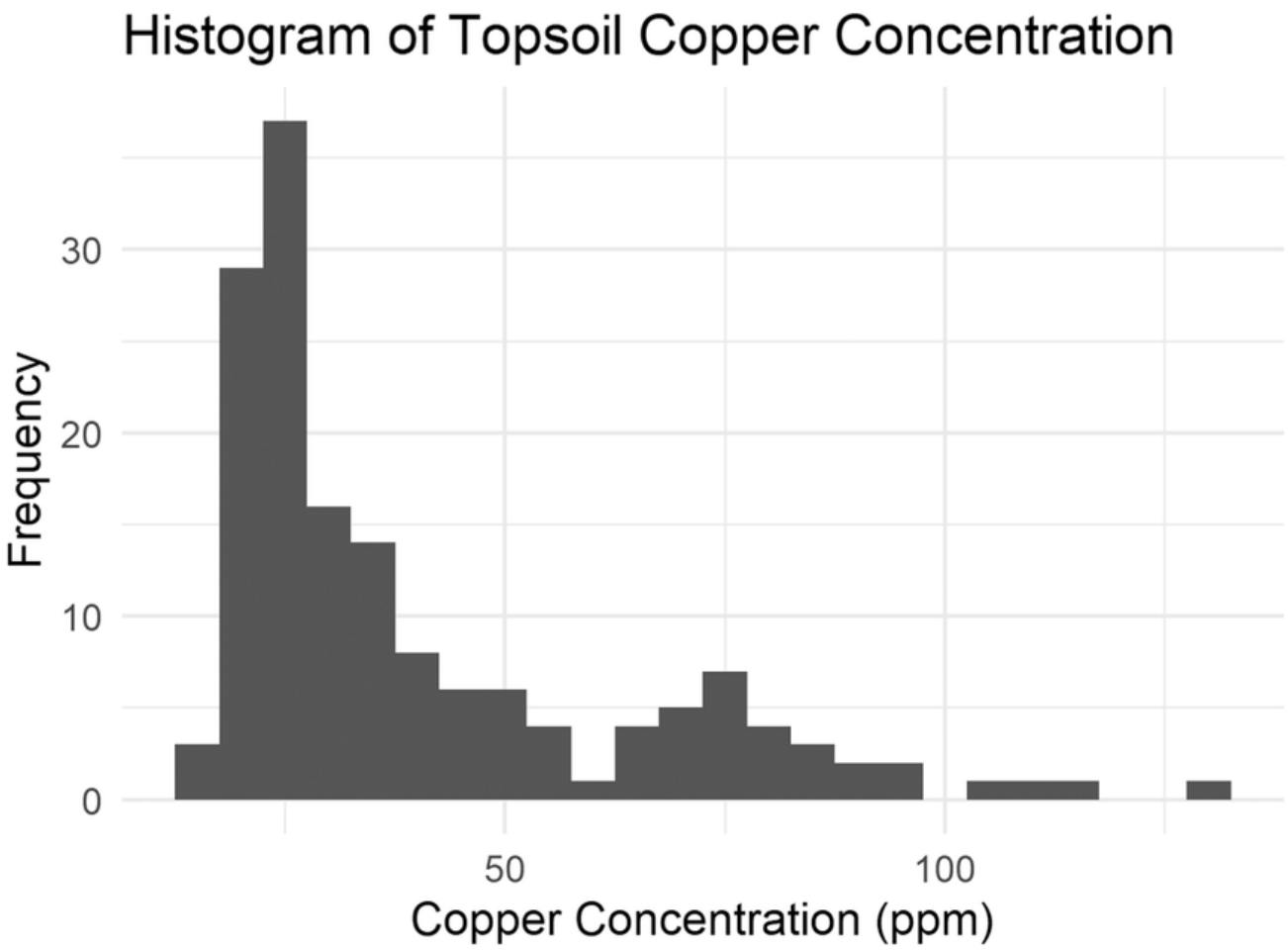
Using Graphics to Explore Relationships

Graphics can be a great way to look at relationships between and among variables of interest. Number and type of variables influence the type of graphics you can use to visualize information. Remember, when initially visualizing data, graphics do not have to be extraordinarily complicated or be ready for use in a publication. Below are ideas for graphic types along with variable types and combinations they include.

“ Remember, when initially visualizing data, graphics do not have to be extraordinarily complicated or be ready for use in a publication. ”

Histogram

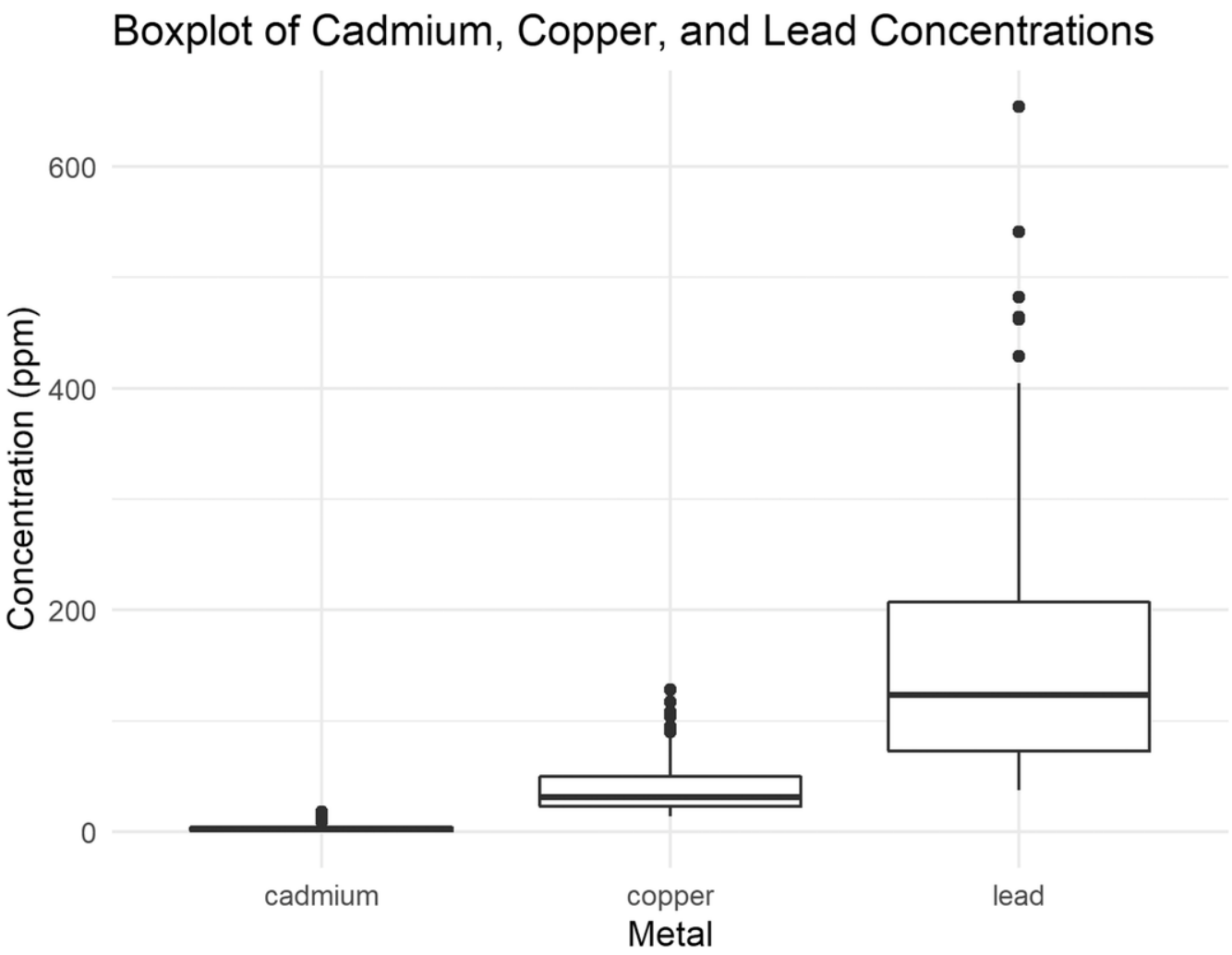
Single continuous variables are commonly visualized using histograms. Histograms allow you to see the arrangement of the data (i.e. skew, distribution, etc.). With this type of graphic, you are able to check the normality of your data distribution. For initial data visualization, the hist() function in R or the geom_histogram within the package *ggplot2* can be useful. We have continued using the *meuse* data set for these examples, which include topsoil heavy metal concentrations in the floodplain of the river Meuse as well as elevation and land use observations. The copper concentration displayed in Figure **2** would be described as unimodal and skewed right. If you were going to continue analysis of this data further, it may require transformation before pursuing statistical methods that assume normal data distributions such as t-tests.



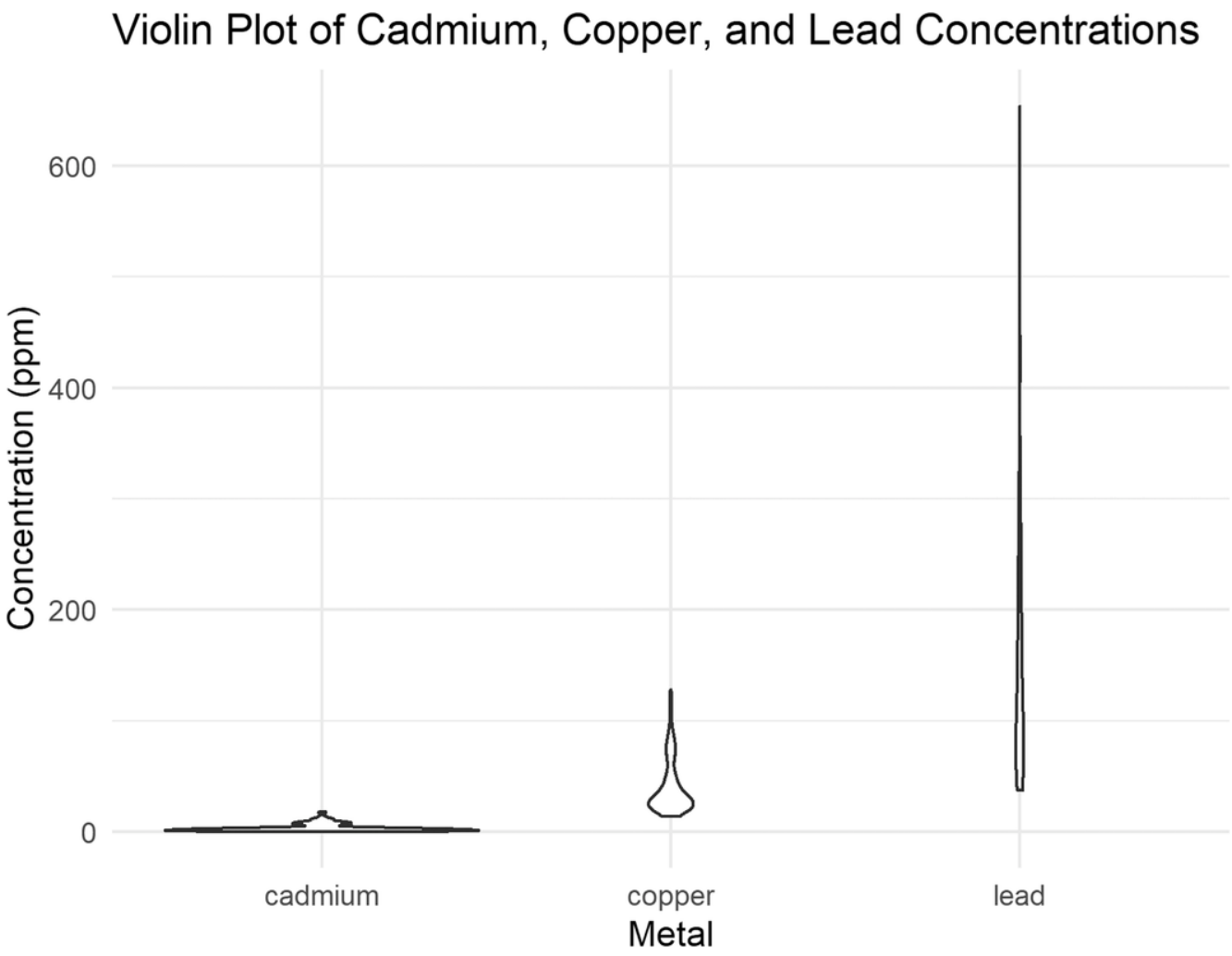
Histogram of topsoil copper concentration in parts per million.

Box Plot/Violin Plot

When a dataset has a continuous variable and a categorical variable, a box plot (Figure 3) or violin plot (Figure 4) might be appropriate. Violin plots can assess the normality of individual categories (similarly to a histogram) while box plots make it easy to compare the spread of your data across categories. In the example graphics shown below, both the violin plot and box plot of the topsoil concentration of heavy metals show that lead has the largest spread.



Box plot of topsoil cadmium, copper, and lead concentrations in parts per million.



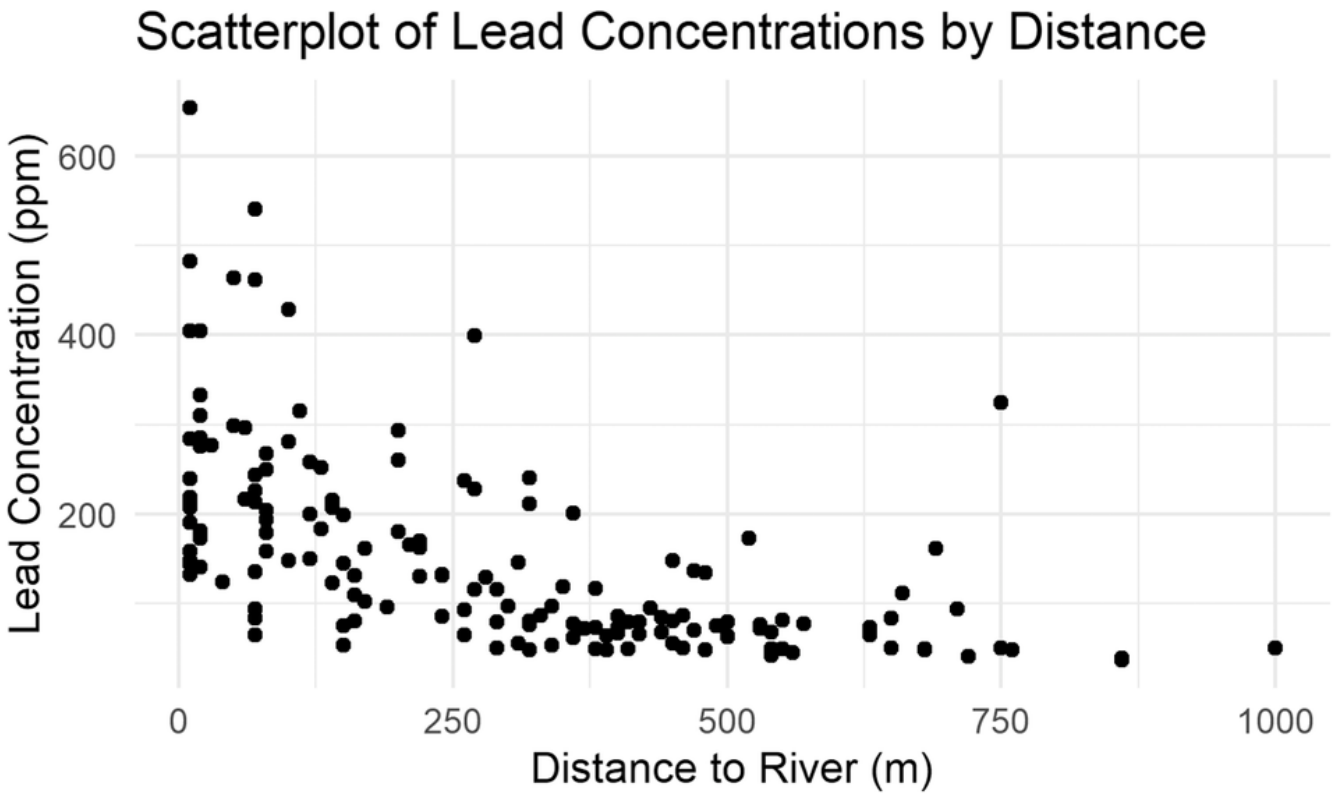
*Violin plot of topsoil cadmium, copper, and lead concentrations in parts per million. The same data were used for both Figures **3** and **4**.*

Depending on the structure of your data set, you may have multiple categories of data you're interested in comparing. Note in Figure **4** that the violin for cadmium is much wider than the violin for lead, which indicates there are many samples with the same concentration of cadmium as each other. Figure **4** shows few samples with the same concentration of lead while Figure **3** does not show any information regarding frequency. Both Figures **3** and **4** show the range of lead concentrations is much wider

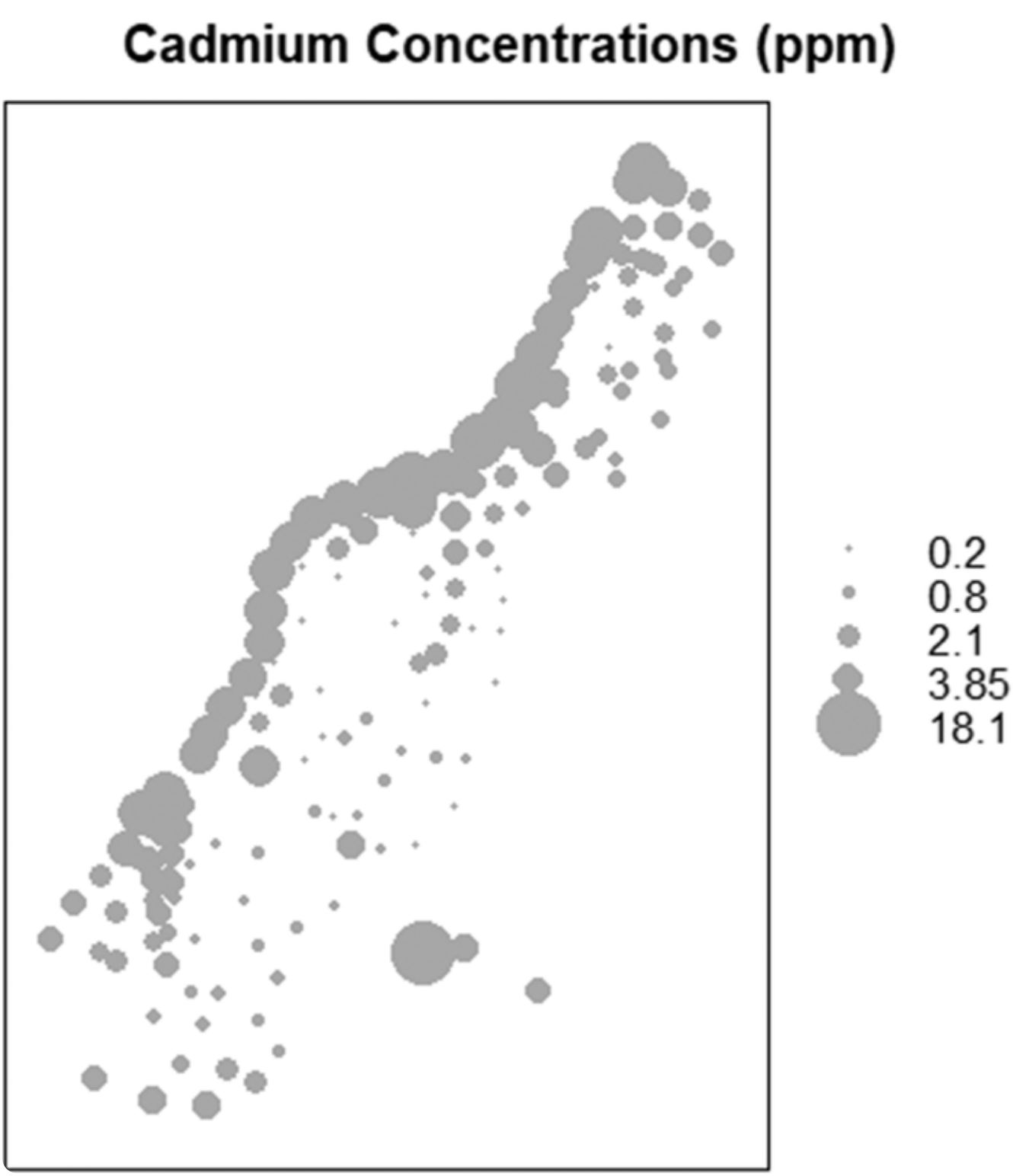
than the range of cadmium or copper concentrations. Although not shown in the box plot or violin plot examples, two categorical variables can be displayed in both violin plots and box plots by grouping boxes using different colors if desired.

Scatterplot

In a scatterplot, two continuous variables are visualized against one another. These plots can display a variety of relationships (e.g., linear, exponential, logarithmic, etc.) that might inform the next steps in evaluating your data or the analysis that you should use. Figure **5** plots the lead concentration in topsoil at a variety of distances from the river Meuse. If you were analyzing this data set, a scatterplot would be helpful for recognizing that the relationship between these variables is not linear. Variables with nonlinear relationships will have Pearson's R values that underestimate the strength of the relationship, which you may need to transform before including them in multivariate linear modeling.



Scatterplot of topsoil lead concentrations at various distances from the river Meuse.



A bubble plot of soil cadmium concentrations in the southeastern floodplain of the river Meuse. Larger bubbles denote higher cadmium concentrations (in ppm), and most high concentrations were located close to the river.

Maps

Visualizing spatial data is also important for assessing patterns. Plotting point data over a map of field boundaries, counties, or other features appropriate for the scale of your study is a useful way to assess the spatial distribution of your data. If you have data that has both a spatial and numeric or categorical component, you can plot data using colored markers to help identify trends in both data location and value. Plotting continuous data using a range of marker sizes instead of a gradient of marker color can also be a great way to identify locations in your study region with clusters of high or low values. This example of soil cadmium concentrations in the floodplain on the southeast side of the river Meuse was generated using data from the R package *sp* and the bubble function, also in *sp*. The large, overlapping bubbles help visualize the high cadmium concentrations along the river bank.

Polygon or raster data can be similarly explored using color gradients. If you have multiple layers of polygon or raster data, using transparent layers in different primary color gradients can make it easier to assess combinations of data layers.

Data Visualization Webinar

In 2020, Dr. Shantel Martinez, a post-doctoral researcher at Cornell University in Plant Breeding and Genetics, hosted a webinar focusing on key concepts for data visualization. During the webinar, Martinez explained an overview of data visualization and how formatting a visual can influence your message. This webinar provided considerable resources to students who are both new and well versed in data visualization. To access resources from this presentation and learn more about the basics of data visualization discussed in this webinar, please visit **https://shantel-martinez.github.io/DataViz2020.html**.

In conclusion, there are many ways to visually assess your data. Graphs, maps, correlation plots, and other tools do not need to be complex or publication worthy to help illuminate patterns that can help inform your further analysis.

In conclusion, there are many ways to visually assess your data. Graphs, maps, correlation plots, and other tools do not need to be complex or publication worthy to help illuminate patterns that can help inform your further analysis.

Reference

- Burrough, P.A., & McDonnell, R.A. (1998). *Principles of geographical information Systems* (2nd edition). Oxford University Press. Google Scholar

More careers & education

Back to issue

Back to home