Skills for Impact - Data Visualization

Data Visualization: Principles and Practice

Instructor:

Constantin Manuel BOSANCIANU, PhD

“Institutions and Political Inequality” research unit

Berlin Social Science Center (WZB)

Reichpietschufer 50, C7, 10785 Berlin

Phone: +49 30 25491 382

Email: BosancianuC@spp.ceu.edu

Web: https://cmbosancianu.github.io

Course Description

Over the last decade data visualization has become a topic of increasing focus and importance for a range of institutions and professions. Presenting ideas and insights using graphical tools has always been a core area of academic work. What we see in recent years, though, is the expansion of this practice in the routine work of international organizations, private-sector companies, and media outlets. A number of trends have driven this expansion: (1) an exponential increase in the public availability of data; (2) a considerable push in both the public and private sectors for evidence-based decision-making and quantifiable metrics for measuring success; (3) an increasing degree of familiarity with data-based arguments on the part of a greater share of the public; and (4) a greater need for public institutions, and even companies, to exhibit transparency toward stakeholders.

With rapid growth in any field, though, comes the difficulty of maintaining standards of quality. This is what this course tries to address. One of its main goals is to present students with a set of standards of good practice when assessing or creating data visualizations. At the same time, it is also intended to give students the tools with which to easily create such visualizations of their own. It will be a *hands-on*, practical course in *how to evaluate and create data visualizations*, based on *open-source* software and up-to-date data sets.

Learning Outcomes

By the end of the course, participants should be capable of (1) judging the deficiencies and strengths of particular plots with respect to how well they transmit useful information; (2) acquire and use a vocabulary of terms with which to effectively communicate about graphical displays of information; and (3) use R and the ggplot2 package to produce high-quality, publication-ready plots based on their own data sources.

Course Requirements

Students should expect to obtain a maximum grade for the course if they:

1. Participate actively during class discussions (20% of the final grade).
2. Submit a final assignment for the course (80% of the final grade). The final assignment is based on a paper that participants have written in the past, and which uses some form of quantitative analysis. The assignment is to replicate all the tables (either of descriptive statistics, bivariate associations, or regression results) using only plots. If any plots are present in the paper, participants should evaluate whether they are suitable for the idea they try to convey; if not, they should improve on them using the knowledge gained in the course.

Due to time constraints, **please send me the paper that you would like to improve on by the day before the class begins** (by April 24, 11:59 PM CET, at BosancianuC@spp.ceu.edu). If you do not have a paper of your own, please send me the title of the published paper you would like to improve upon, as well as a link to where the replication materials are.

Course Schedule

Most of the readings for the course have been assigned from a single textbook: Kieran Healy, *Data Visualization: A Practical Introduction* (Princeton University Press, 2018). For the purposes of a short course like ours, this should be more than sufficient. For those of you who are in need of additional information or clarifications, these can be found in the optional readings assigned in the next section. For assistance with ggplot2 code I particularly recommend Hadley Wickham's *ggplot2: Elegant Graphics for Data Analysis* (Springer, 2009), as well as the very informative online resource he maintains: https://ggplot2.tidyverse.org/.

All readings assigned in this section for their respective sessions are mandatory. Additional sources are listed in the next section.

Additional readings

For more advanced topics related to using R for producing graphs, more sophisticated types of data visualizations, as well as the specific challenges pertaining to graphically displaying very large quantities of information, please consult one of the relevant references below.

- Chen, Chun-houh, Wolfgang Härdle, and Antony Unwin. 2008. *Handbook of Data Visualization*. New York: Springer.
- Cleveland, William S. 1985. *The Elements of Graphing Data*. Monterey, CA: Wadsworth Advanced Books and Software.
- Tukey, John W. 1977. *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.
- Unwin, Antony, Martin Theus, and Heike Hofmann. 2006. *Graphics of Large Datasets: Visualizing a Million*. New York: Springer.
- Wickham, Hadley. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Second edition. New York: Springer.
- Wilkinson, Leland. 2005. *The Grammar of Graphics*. Second edition. New York: Springer.
- Cook, Dianne, Eun-Kyung Lee, and Mahbubul Majumder. 2016. "Data Visualization and Statistical Graphics in Big Data Analysis". *Annual Review of Statistics and Its Application* **3**: 133–159.
- Evergreen, Stephanie, and Chris Metzner. 2013. "Design Principles for Data Visualization in Evaluation." *Data Visualization, part II: New Directions for Evaluation* **140**: 5–20.
- Frees, Edward W., and Robert B. Miller. 1998. "Designing Effective Graphs." *North American Actuarial Journal* **2**(2):

53–76.

- Zeileis, Achim, Kurt Hornik, and Paul Murrell. 2009. "Escaping RGBland: Selecting colors for statistical graphics." *Computational Statistics & Data Analysis* **53**(9): 3259–3270.

There is also a vibrant community devoted to interactive data visualizations. Although we do not have the time in this course to cover this topic as well, if interested you can find out more about this in the following places:

- https://master.bioconductor.org/help/course-materials/2015/CSAMA2015/lab/shiny.html
- https://seankross.com/developing-data-products/shiny.html
- http://shiny.rstudio.com/tutorial/
- http://shiny.rstudio.com/images/shiny-cheatsheet.pdf
- http://shiny.rstudio.com/articles/

Course policies

During class

During the course participants are expected to conduct themselves as befits a member of the CEU community, particularly in terms of showing respect and tolerance of others' argued opinions. Physical or verbal threats, along with harassment of any nature will be swiftly dealt with and reported to university officials. The same procedure will be followed in case of discrimination based on national origin, gender, racial or ethnic background, sexual orientation, religious beliefs, or cultural symbols (including clothing).

Attendance policy

Attendance throughout SFI courses is mandatory. Absence without a medical note will result in a "fail" for the class. Please notify the instructor via email regarding these instances as soon as reasonably possible.

Incomplete grades and late assignments

No incomplete grades will be assigned for this course. Late assignments are allowed, but with a penalty of 1 point (assuming a 0-10 grading scale) for *each day of delay*. The penalty is applied the moment at which the new day starts, i.e. the minute after the deadline has passed.

In case of a personal emergency we can negotiate a new deadline for submission on a case-by-case basis. Please notify the instructor regarding these instances as soon as reasonably possible.

Academic integrity and honesty

All participants are responsible for being aware and following the guidelines pertaining to academic integrity currently in force at the Central European University. The main documents that pertain to this are the university's Code of Ethics (https://documents.ceu.edu/file/2660/download?token=Mbo1tSZC) and the Policy on Plagiarism (https://documents.ceu.edu/file/2659/download?token=u64ukEwZ), as well as the Policy on Harassment (https://documents.ceu.edu/file/1708/download?token=U5RoURPM). Please make sure you are aware of the provisions in these documents before the class begins. The behaviors these documents cover extend to cheating, plagiarism, appropriating another person's work, submitting material that was already submitted for grade in another course, threatening behavior, harassment, and others.

As a rule, **plagiarism will result in a "fail"** and the case being brought up in front of the Departmental Committee on Academic Dishonesty. If you are unsure about whether an action you're about to take violates these standards, please ask the instructor. With respect to whether a particular practice constitutes plagiarism or not, a good rule of thumb is: **when in**

doubt, cite!

Email policy

I will reply students' emails within two business days.

Main Course Forum

Add a new topic to this forum

April 25, O/Blue Lab

Principles of good data visualization (April 25, Session 1)

We start by delving into a few of the guidelines and principles for how to create effective data visualizations.

Readings:

1. Tufte, Edward R. 2001. *The Visual Display of Quantitative Information*. Second edition. Cheshire, CT: Graphics Press. Chapters 4 and 6.
2. Healy, Kieran. 2018. *Data Visualization: A Practical Introduction*. Princeton, NJ: Princeton University Press. Chapter 1.

Crimes against graphs (April 25, Session 2)*

In this session we cover a few prominent examples of ineffective visualizations and try to better understand which of the guidelines mentioned in the previous session they violate.

Readings:

1. Tufte, Edward R. 2001. *The Visual Display of Quantitative Information*. Second edition. Cheshire, CT: Graphics Press. Chapters 2 and 5.

Univariate and multivariate graphs (April 25, Session 3)

We start the applied portion of the class by going over the most typical types of graphs for univariate and multivariate data, and how to implement them in ggplot2. We cover histograms, bar charts, area plots, box-and-whisker plots, line charts, and scatterplots.

Readings:

1. Healy, Kieran. 2018. *Data Visualization: A Practical Introduction*. Princeton, NJ: Princeton University Press. Chapters 3 and 4.
2. Unwin, Antony. 2015. *Graphical Data Analysis with R*. Boca Raton, FL: CRC Press. Chapters 3, 4 and 5.

* The name of the session is inspired by a 2016 presentation by Andrew Gelman, titled “Crimes against data”.

April 26, O/Blue Lab

Customizing graphs (April 26, Session 4)

Most graphs intended for public release require heavy customization. In this session we advance by going beyond the ggplot2 defaults: (1) using colors, shapes and sizes to encode more information into the plot; (2) faceting; (3) highlighting specific data points; and (4) “cleaning up” a plot to highlight the information we want conveyed.

Readings:

1. Healy, Kieran. 2018. *Data Visualization: A Practical Introduction*. Princeton, NJ: Princeton University Press. Chapters 5 and 8.

Maps with ggmap (April 26, Session 5)

Presenting spatial data and displaying it under the form of maps has recently been made much easier by the launch of the ggmap package. In this session we go over how to create maps with the ggmap package, and how to present information at different levels of aggregation.

Readings:

1. Healy, Kieran. 2018. *Data Visualization: A Practical Introduction*. Princeton, NJ: Princeton University Press. Chapter 7.
2. Kahle, David, and Hadley Wickham. 2013. “ggmap: Spatial Visualization with ggplot2.” *The R Journal* **5**(1): 144–161.

Graphs derived from model output (April 26, Session 6)

In the final session, we cover the creation of graphs from statistical model output in order to convey our model-based insights to a wider audience. We learn how to plot regression coefficients, predictions from the data, as well as how to present uncertainty about our conclusions.

Readings:

1. Healy, Kieran. 2018. *Data Visualization: A Practical Introduction*. Princeton, NJ: Princeton University Press. Chapter

6.

2. Kastellec, Jonathan P., and Eduardo L. Leoni. 2007. “Using Graphs Instead of Tables in Political Science.” *Perspectives on Politics* **5**(4): 755–771.