Becoming Tidier Over Time: Studying #tidytuesday as a Social Media-Based Context for Learning to Visualize Data

Joshua M. Rosenberg, University of Tennessee, Knoxville, jmrosenberg@utk.edu
Anthony Schmidt, University of Tennessee, Knoxville, aschmi11@utk.edu
Aaron Rosenberg, Thermo-Fisher Scientific, aaron.matthew.rosenberg@gmail.com
Jennifer Longnecker, University of Tennessee, Knoxville, jlongne1@vols.utk.edu
Michael Mann, University of Tennessee, Knoxville, mmann14@vols.utk.edu

Abstract: This poster explores a new context and design to learn to visualize data, the social media-based #tidytuesday weekly challenge. We use a novel data source for understanding learning —tweets from participating individuals over more than one year—and a combination of qualitative and computational research methods. The content individuals shared explained, qualified, or highlighted the substance of the visualizations they created, and participation over time (and longer code) was related to being recognized more by others.

Background

This poster explores a new context to learn data visualization, a core capability for statistics and data science learners and professionals (Laina & Wilkerson, 2016), the social media-based #tidytuesday weekly challenge. #tidytuesday developed from a broader community that was created to help novice data scientists to learn to use the programming language *R* (Maegan, 2019), with an emphasis on the *tidyverse* suite of R packages (Wickham et al., 2019). Every week since 2018, #tidytuesday moderators have provided a weekly challenge focused on the creation of data visualizations based around interesting, timely, or socially-important topics, such as those with data about the Women's World Cup, incarceration trends, and school diversity in the United States.

While data visualization has been described as an important (and perhaps accessible to beginners) capability for statistics and data science learners (and for science and math education learners), it has been the focus of less research than other aspects of work with data, such as making inferences from models. The purpose of this study is to explore #tidytuesday as a social media-based design for beginning data scientists' development of data visualization skills. Studying beginning data scientists' development in the context of #tidytuesday can speak to how a community organized on and through social media can, with the right design features, be a potentially powerful setting through which to develop new capabilities. Lastly, this work shows how learners—through social media—can develop both *technical* (i.e., how to create useful and accurate figures) and *social* (i.e., how to share one's work) capabilities related to visualizing data. Accordingly, our guiding research question in the context of a prominent social media-based context for learning to visualize data, how do individuals' social and technical contributions to a relate to the other?

Method

We first accessed data on #tidytuesday through the *tidytuesday.rocks* (1) interactive web application, which was created to provide a platform to highlight individuals' weekly contributions to the hashtag. The particular data we collected was the content of 4,418 total tweets contributed by 1,231 individuals over a one-year period. From this data, we determined which tweets included links to code (on GitHub) and accessed the code. To understand the technical skills that individuals evidenced through their data visualizations, we used the *tidycode* R package to classify the proportion of the code belonging to one of nine categories, including cleaning data, using a statistical model, and visualizing data. To understand individuals' social involvement, we looked at the consistency of individuals' participation and determined how many recognitions (favorites or retweets), on average, tweets received. We also carried out a qualitative content analysis of randomly-sampled tweets. Lastly, to explore how the amount and type of code shared relate to social aspects of participation in #tidytuesday, we explored how receiving recognitions was correlated to the timing of the tweet (early or later in an individual's participation) and the length of the code individuals shared.

Findings

We found that—on average—individuals' code included 55.73 functions, and though #tidytuesday contributions were focused on data visualization, the plurality of individuals' code (41.19%) belonged to the *data cleaning* category, followed, less surprisingly, by *visualization* (33.88%) and then setup (including loading other tools, primarily those available in R to carry out different analyses that are more sophisticated than those available in the 'base' R software; 8.09%) and *import* of data (6.94%). Individuals' tweets were highly-recognized by others,

receiving, on average, 31.93 favorites or retweets. Regarding the tweets' content, 58% involved individuals' explanations of their visualizations; 18% focused on substantive findings with respect to the selected dataset; 13% expressed some emotion related to learning (i.e., patience); and, 9% of the sampled tweets expressed humility about their contribution. Finally, our exploration of the relationship between the technical and social aspects of #tidytuesday revealed a complex portrait we are beginning to understand and explore further (Figure 1). Preliminary analyses indicated that, over time, individuals wrote longer code (r = .36, p < .001) which was associated with receiving more recognitions (r = .12, p = .017); although statistically significant, this relationship is small, and therefore suggestive of further scrutiny.

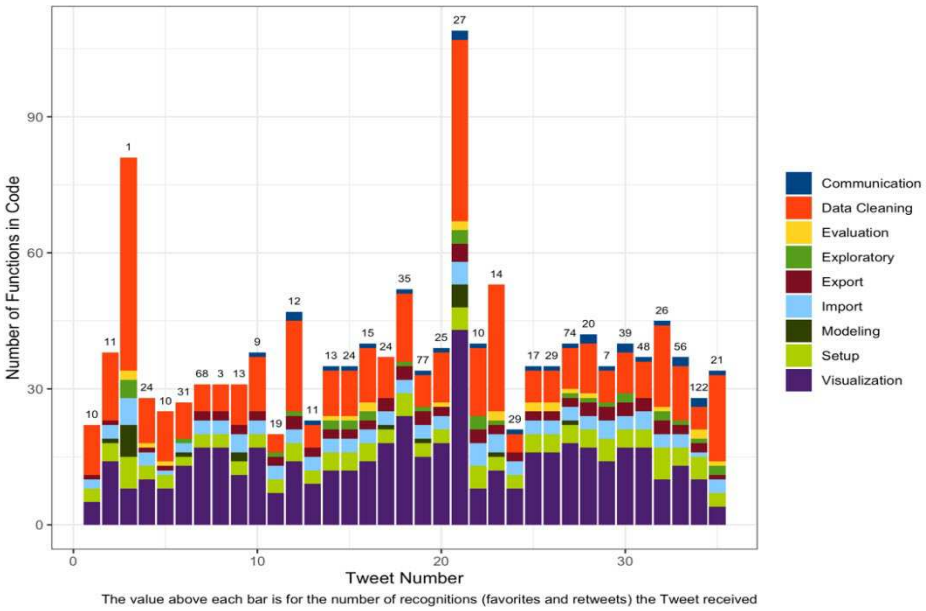

Figure 1. Example of how being recognized related to contributions of code (from one #tidytuesday user).

Discussion

This poster will interpret these early-stage findings in light of the ways in which individuals appear to be interacting *socially* through Twitter to share their *technical* contributions in ways that are shared with and also distinct from other informal networks and learning communities, even those that are social media-based (Lantz-Andersson et al., 2019; Macià & Garcia, 2016). Accordingly, we will present the results for each of our research questions as well as a description of what design features—being informal and open yet also moderated and focused on relevant datasets as a part of weekly challenges—may contribute to the intertwined technical and social dimensions of #tidytuesday. We are also exploring other means of studying involvement in #tidytuesday and what outcomes individuals achieve through participation, including interviews with (or surveys of) moderators and participants, and social network analytic methods to understand who is interacting with whom and how interactions might influence individuals' interwoven technical and social contributions and learning over time.

Endnotes

(1) https://nsgrantham.shinyapps.io/tidytuesdayrocks/

References

Laina, V., & Wilkerson, M. (2016). *Distributions, trends, and contradictions: A case study in sensemaking with interactive data visualizations*. Singapore: International Society of the Learning Sciences.

Lantz-Andersson, A., Lundin, M., & Selwyn, N. (2018). Twenty years of online teacher communities: A systematic review of formally-organized and informally-developed professional learning groups. *Teaching and Teacher Education, 75,* 302-315.

Macià, M., & García, I. (2016). Informal online communities and networks as a source of teacher professional development: A review. Teaching and teacher education, 55, 291-307.

Maegan, J. (2019). *R4DS online learning community Improvements to self-taught data science & the critical need for diversity, equity, and inclusion.* Presentation at rstudio::conf(2019).

Wickham et al., (2019). Welcome to the tidyverse. *Journal of Open Source Software, 4*(43), 1686,