Remote, but Connected: How #TidyTuesday Provides an Online Community of Practice for Data Scientists

NISCHAL SHRESTHA, North Carolina State University, USA
TITUS BARIK, Microsoft, USA
CHRIS PARNIN, North Carolina State University, USA

Data science practitioners face the challenge of continually honing their skills such as data wrangling and visualization. As data scientists seek online spaces to network, learn and share resources with one another, each individual has to employ their own ad-hoc strategy to practice their data science skills. Given these disjointed efforts, it is crucial to ask: how can we build an inclusive, welcoming online community of practice that unites data scientists in their collective efforts to become experts? Daily hashtags on Twitter are used on specific days and have shown promise in forming a community of practice (CoP) in social networking sites like Twitter, but how do they benefit the community and its members? To understand how daily hashtags benefit data scientists and form an online CoP, we conducted a qualitative study on #TidyTuesday—a daily hashtag project for data scientists using R—using the framework of CoP as a lens for analysis. We conducted semi-structured interviews with 26 participants and uncovered motivations behind their participation in #TidyTuesday, how the project benefited them, and how it cultivated an online CoP. Our findings contribute to the CSCW research on community of practices by providing design trade-offs of using daily hashtags on Twitter, and guidelines on growing and sustaining an online community of practice for data scientists.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**; **Empirical studies in collaborative and social computing**.

Additional Key Words and Phrases: community of practice, data science, hashtag movements, online social movements

ACM Reference Format:
Nischal Shrestha, Titus Barik, and Chris Parnin. 2021. Remote, but Connected: How #TidyTuesday Provides an Online Community of Practice for Data Scientists. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 52 (April 2021), 31 pages. https://doi.org/10.1145/3449126

1 INTRODUCTION

"There must be a better way," Thomas thought. There were 20 tabs open on his browser: a scattered collection of blogs, tutorials, and videos. Two books laid open on his desk on how to use a graphics library—good start, but not enough. The cursor on his editor blinked waiting for his next move. Thomas began to reflect on his process for practicing data science. The resources he reached for did not provide practice on real-world datasets, and his neurobiology research datasets were not suitable for showcasing his full range of visualization skills to potential employers. It also became

Authors' addresses: Nischal Shrestha, nshrest@ncsu.edu, North Carolina State University, Campus Box 8206, 890 Oval Dr., Raleigh, North Carolina, 27695, USA; Titus Barik, titus.barik@microsoft.com, Microsoft, One Microsoft Way, BRAVERN-2/18215, Redmond, Washington, 98052, USA; Chris Parnin, cjparnin@ncsu.edu, North Carolina State University, Campus Box 8206, 890 Oval Dr., Raleigh, North Carolina, 27695, USA.

2573-0142/2021/4-ART52 $15.00
https://doi.org/10.1145/3449126

harder and harder to motivate himself and stay consistent with his practice. In a nutshell, his approach to practice felt ad-hoc. Something was missing, or maybe *someone*. Thomas was alone. But, he would soon start a movement to bring together countless data scientists.

Data scientists are increasingly prevalent in online spaces. They are a group of people who are distributed across various parts of the world with diverse backgrounds, from statistics to bioinformatics to graphics. Data scientists also differ from traditional programmers and are typically end-users without a formal background in programming [50]. To build up their expertise, they are faced with the constant challenge of practicing skills like acquiring, cleaning, wrangling, visualizing, and presenting data. To expedite their learning process, data scientists are becoming dependent on faster, more accessible resources which are typically found online like tutorials, documentation, or Q&A sites [87, 88]. However, data scientists can get socially isolated in their efforts for practice without a community of practice, which can negatively impact motivation for consistent practice. Without a community to grow in, data scientists also miss out on tacit knowledge like best practices and techniques not captured in online resources. As data scientists seek help in online spaces, it is crucial to ask: how can we build an inclusive, welcoming online community of practice that unites data scientists all over the world in their collective efforts in becoming experts?

Daily hashtags have been used by several online communities on Twitter to organize discussions around a topic, activity, or event. Twitter hashtags (#) have been previously used for trending "tweet chats" around activism such as challenging engineering stereotypes [61] (#ILookLikeAnEngineer), or exchanging knowledge during breaking news such as pandemics [55] (#SwineFlu). A daily hashtag is a different type of hashtag which is used periodically. For instance, #AdventOfCode is a popular daily hashtag where in the month of December, each day presents a new programming puzzle. Programmers then post a tweet and share their thoughts about their own approaches in solving the puzzles. Daily hashtags can help build a community of practice (CoP) by allowing programmers of all skill levels to practice solving programming puzzles and network or exchange knowledge on Twitter. But, can daily hashtags provide an online CoP for data scientists? Thus, our research questions for this study are:

- **RQ1**: Who participates in Tidy Tuesday and what are their motivations and goals?
- **RQ2**: What do participants gain by participating in Tidy Tuesday?
- **RQ3**: How does social activity around Tidy Tuesday cultivate a community of practice?

To investigate our research questions, we conducted a qualitative case study on #TidyTuesday—a daily hashtag project for data scientists to practice their data wrangling and visualization skills using R. #TidyTuesday provides data scientists access to a curated dataset in a GitHub repository every Tuesday. Participants perform their analysis of the dataset and produce plots answering exploratory questions of their own. They are encouraged to share a tweet with the hashtag including a link to their code and the plot they produced. #TidyTuesday can be characterized using the three main components of a CoP: the domain (data science), the people (data scientists), and the practice (data analysis and visualization). To understand the motivations and goals of #TidyTuesday participants and the social interactions that help form and sustain an online CoP, we conducted semi-structured interviews with 26 data scientists. The participants were from diverse backgrounds with varying skill levels from beginners to veterans, some of whom are widely known in the larger data science community. These characteristics provided us with the opportunity to explore a broad range of experiences which provide insights into why data scientists use #TidyTuesday, what the benefits are, and how it is used to cultivate an online CoP using [92].

From our qualitative analysis of #TidyTuesday, we found several motivations behind participation, and the ways in which the project successfully grows an online community of practice, which both corroborate previous findings and extend them. Participants' main motivation was to hone their data science skills with the help of weekly-released, curated datasets and a community of practice. #TidyTuesday participants underwent transformative experiences such as discovering numerous R packages and tools from others, improving data wrangling and visualization skills, building data visualization portfolios for the job market, and supporting offline events like workshops and "hacky hours". We discuss how #TidyTuesday enabled these experiences by relating the project to constructs of a CoP and its design components [92] such as providing a rhythm, and having a loose and flexible structure to fit each individual's needs. However, we also identified barriers of entry for newcomers such as not knowing how to start and a general lack of constructive feedback and mentorship. To our knowledge, this is the first paper to explore the R community in cultivating an online CoP through the use of daily hashtags on Twitter. The key contributions of this paper are:

- The first qualitative study of #TidyTuesday, a daily hashtag that formed an online CoP for data scientists using R, through semi-structured interviews with 26 data scientists.
- An analysis of the intrinsic and extrinsic motivations behind participation in #TidyTuesday, the benefits gained by participants, and the social interactions that helped grow and sustain the project.
- A discussion of the design trade-offs of using daily hashtags on Twitter and a set of guidelines to successfully grow and sustain an online CoP for data scientists, and overcome learning and social barriers.

2 RELATED WORK

In the following subsections, we present findings in the CSCW and HCI literature with regards to community of practice, data scientists and the R community, as well as daily hashtags on Twitter. We highlight the gaps in knowledge with how to foster an online community of practice through the use of Twitter for data scientists.

2.1 Communities of Practice

Communities of practice (CoP) are groups of people who share a concern, a set of problems, or a passion about a topic, and who deepen their knowledge and expertise in this area by interacting on an ongoing basis [93]. We study a nascent **R community** forming on Twitter around the #TidyTuesday project, designed to provide an online CoP for **data scientists** to **practice data wrangling and visualization**. The relevant literature we examined for studying #TidyTuesday include the design components necessary for cultivating a CoP developed by Wenger et al. [92], the idea of Twitter as an imagined community [15] by Gruzd et al. [38], and sense of community (SoC) theory by McMillan and Chavis [67]. Wenger et al. [92] describe how a CoP is different from organizational design which focuses on fixed goals and elements, where optimizing for *aliveness* is emphasized because a community has to invite interactions to keep it alive and growing. The authors suggest creating rhythm for the community, which daily hashtags like #TidyTuesday is designed to provide. Gruzd et al. [38]'s study of a single member's Twitter network is also important because they found that Twitter can meet Jones [49]'s minimum requirements for a virtual settlement like interactivity, or sustained membership over time. The authors found that a single individual can form their own personal community on Twitter through network analysis, while we study many individuals who are joining a growing community of practice around #TidyTuesday using a qualitative approach to gain rich insights into motivations, and social interactions that help cultivate an online CoP. Finally, McMillan and Chavis [67]'s "Sense of Community" (SoC) theory outlining characteristics of a community like the fulfillment of needs is also relevant to how well #TidyTuesday meets the R community needs. We extend the literature on online CoPs by applying the framework on the #TidyTuesday project on Twitter and contribute new perspectives on how a data science CoP can be formed and sustained.

The CSCW and HCI communities have used the CoP framework as a lens for studying various groups in social media sites. For example, Marlow and Dabbish [65] studied graphic designers and the social transparency provided by SNS (Social Network Site) features of a design portfolio website called Dribbble [2]. They found that SNS functionalities like following members and having access to artifacts supported social learning via *legitimate peripheral participation* (LPP) [60] and professional identity development. Holikatti et al. [43] also found heavy use of LPP in Facebook groups for learning how to host living spaces using AirBnB [10], a sharing economy platform for hosting living spaces all over the world. They found that members learned affordances of AirBnB through Facebook by asking questions and interacting with more experienced AirBnB users. In our study, #TidyTuesday on Twitter facilitates these social transparencies via public tweets and links to code, which provides opportunities for skill and professional development via LPP. Kou et al. [57] used CoP as a framework to study the changing practices of user experience (UX) professionals on reddit where they identified social roles in relation to knowledge production and dissemination in the online community of *volatile practice*—rapidly changing occupations. Data scientists face similar challenges in a young field that is ever evolving, and we provide an instance of using an online community of practice for those who are entering the field (more in section 2.3). We extend the literature by studying how the R community use Twitter to improve their data wrangling and visualization skills, learn from each other, and gain a sense of community in an evolving field.

There is also prior work which examines accessibility, motivations and barriers in various online CoPs. For example, Mugar et al. [71] studied the accessibility of participation norms in online communities where participants lack full access to others' work. The authors combined the theory of legitimate peripheral participation with the theory of social translucence to derive practice proxies such as traces of user participation in online environments that act as resources to orient newcomers towards the norms of practice. Similarly, Xu and Bailey [97] uncovered motives for participation and expectations of the critiques within an online community. They provide recommendations for improving the design of systems that support community-based critique of creative artifacts. Our study provides similar insights within the data science context with regards to motivations behind participation in a community-led project like #TidyTuesday, discussing how Twitter can facilitate LPP, but also limit distributed critique.

2.2 Data Scientists

People across a wide range of professions now write code as part of their jobs with the purpose of obtaining insights from data rather than building software. The popular term for this type of work is "data science" and the group of people are often called "data scientists". Data scientists come from various backgrounds like engineering, business, design, and research [28]. They are increasingly prevalent in both industry and academic settings. In industry, data scientists work in numerous sectors like public policy, technology, and healthcare [62]. In academia, data scientists are graduate students, professors and technical staff writing code to make research discoveries [39]. Data scientists are known to be like end-user programmers, writing code as a means to an end—to gain insight into data. Unlike traditional programmers, they are also a group that heavily engages in exploratory programming [51]. In particular, data scientists heavily engage in exploratory data analysis where they continually explore questions about the data and iteratively refine statistical models and visualizations to paint a story. However, data scientists also share similarities with software engineers, writing reusable analysis code to share with others—they engage in what Ko et al. [54] call *user software engineering*.

Despite the growth of data science, there is little understanding of how data scientists are learning and practicing their skills outside of corporate and organizational settings. Prior work has given insight on the activities data scientists engage in at work [39, 50, 51, 53, 79, 81], and how practitioners teach beginners in both industry and academia [59]. However, there are only a few studies examining how data scientists hone several skills such as acquiring, cleaning, wrangling, visualizing, and presenting data [41, 50]. To gain expertise in these skills, data scientists must decide between many different learning paths: attending a university and acquiring a data science major [85, 86], taking MOOCs (massive open online courses), or participating in hands-on workshops [96] and coding bootcamps [13, 21]. Despite data science programs becoming increasingly available, there is still debate around what to include in a data science curriculum [18] and how to prepare students for the industry.

Prior CSCW and HCI research has explored how data scientists work in both organizational and corporate settings, as well as in informal settings. There are several studies on solitary data science practices related to data wrangling tools [40, 83], as well as exploratory analysis and barriers in computational notebooks [24, 52]. In collaborative settings, Passi and Jackson [74, 75] have found that data scientists collaborate in order to resolve tensions around trustworthiness of data and the analysis process. There has also been several studies on how data scientists collectively curate data [72, 101], work together on a project [89, 105], share code in competitions [84], and do data science for social good [104]. For example, Hou and Wang [44] studied collaboration in two civic data hackathons, where data science workers help non-profit organizations and discovered unique broker roles. Tausczik and Wang [84]'s study on Kaggle competitions found that data scientists re-use and share code with one another, which helped individuals practice and hone their skills. We extend these studies by providing insights about how a daily hashtag like #TidyTuesday helps data scientists hone their individual skills, while cultivating a community of practice in the wild and outside of organizational, corporate, and hackathon settings.

2.3 The R Community

The R project [4] was born in 1993 as a free and open source programming language and software environment for statistical computing, bioinformatics, and graphics [48]. The R community is an open source community made up of an R-core, a team of software developers that maintain and evolve the R language, language users and package developers. The R community has also been the subject of extensive research in community evolution [34, 87] and the interplay between different channels [88] for asking questions such as Stack Overflow and mailing lists, where members are active in both channels but noticeable shifts towards the former in recent years. Zagalsky et al. [101]'s study on the R community on Stack Overflow versus R-help mailing lists is especially relevant to our study. They describe how users exchange different types of knowledge on Stack Overflow and mailing lists, including a description of the reasons why members choose one channel over the other: users preferred Stack Overflow for the ability to gain peer recognition and faster turnaround on questions, while others preferred the R-help mailing list for its flexibility on topics and the high activity of experienced users.

There are several key players in the R community that have shaped the modern R community, making significant strides into promoting inclusivity, open source software, and education. RStudio [6], a company behind the popular RStudio IDE (Integrated Development Environment) [5] has been developing programming tools in R, making them more accessible to data scientists. RStudio is also a Certified B Corporation [1] company dedicated to creating and sustaining open source software for data science. Several leaders within the R community work at RStudio like Chief Scientist and Educator, Hadley Wickham and Software Engineer, Jenny Bryan, who have been pushing for more inclusiveness and diversity in the R community. Another key player is the R-Ladies

Global [3], a worldwide organization whose mission is to achieve proportionate representation by encouraging, inspiring, and empowering people of genders currently underrepresented in the R community. They have over 138 chapters in 44 countries and 39000 members, holding meetups and events worldwide in order to introduce minority populations to programming in R. The Carpentries [8] organization have a similar mission to foster diversity and inclusion as well as provide essential data and computational skills for conducting efficient, open, and reproducible research. The Carpentries hold workshops, develop openly available lessons designed using evidence-based teaching practices.

The R community has been shifting towards an online community on Twitter (#rstats) and undergoing a trend towards a new style of R programming called tidy R, which offer a user friendly and consistent way of doing data analysis and visualization. The #rstats community is prevalent on Twitter, and there is even an online textbook called "Twitter for R programmers" [19] to help onboard non-Twitter users. Along with the move to sites like Twitter, there is also a popular programming paradigm in recent years called tidy R, comprised of packages in the "tidyverse" [9] that "share an underlying design philosophy, grammar, and data structures" of data [95]—a framework to tidy data and make it amenable for further analysis. #TidyTuesday encourages the use of this framework and provides yet another online resource to existing online channels like Stack Overflow and R-help mailing lists, but takes place on Twitter, where new types of knowledge might be created. The R community and its dynamics in social media sites like Twitter has not received attention of researchers, and we believe are one of the first to explore how Twitter can be used to share knowledge and learn from each other through #TidyTuesday.

2.4 Hashtag Movements on Twitter

Twitter has been extensively studied to explore how they are used by online communities to organize discussions around a topic, activity, or event. On Twitter, a hashtag (#) character is used for trending "tweet chats" around activism such as challenging engineering stereotypes [61], or exchanging knowledge during breaking events [26] such as natural disasters [76]. There is an emerging body of research that is related to the online communities emerging on Twitter viewed through various analytic lens like social translucence, imagined communities, networks, or linguistic analysis [29, 38, 46, 63, 102, 103]. Gruzd et al. [38]'s Twitter as an "imagined community" comes closest to our study, which found that Twitter is capable of forming a sense of community online for a single individual and their personal network. Our work provides a qualitative approach by interviewing many individuals participating in #TidyTuesday, providing insights into why data scientists participate in the project and how the social interactions lead to a successful CoP on Twitter which also provide facilitation of in-person events like meetups.

Recently, users of Twitter use daily hashtags which repeat on a given day and provide various means to form a community of practice. Daily hashtags have been used for sharing knowledge or organizing discussions for a particular topic or domain. For example, they have been used for academic advising with professional development [73], and for synchronous discussions around healthcare [35]. Within programming communities, #AdventOfCode [90] is an example of a popular daily hashtag where each day of the Advent calendar presents a new programming puzzle to solve. Programmers then post tweets and share their thoughts about their solutions and reflections in solving the puzzles. The data science communities have also recently adopted daily hashtags using a similar structure for practice. #MakeoverMonday [58] is a daily hashtag for data scientists using the Tableau [11] software to produce data visualizations and post their submissions on Twitter every Monday. We extend the research on Twitter-based CoPs by studying #TidyTuesday as an example of a daily hashtag targeted towards data scientists wishing to practice their data science skills and become part of a community to grow in.

3 METHOD

In this section, we present details about #TidyTuesday and how it relates to similar projects in the R community and discuss the research questions we investigated, and the qualitative methods we used to answer them.

3.1 Research Setting: #TidyTuesday

3.1.1 Precursors to #TidyTuesday:. In order to understand the background behind #TidyTuesday, we first interviewed Thomas Mock, the creator of the project. During graduate school, Thomas became involved with an online learning community called R4DS (R for Data Science). Jesse Mostipak started the R4DS project in 2018 as a book club [70] for the R for Data Science textbook [94]. Since then, the project has evolved into a learning space on Slack [7], where programmers of all skill levels can ask questions about R, similar to Q&A sites like Stack Overflow, but designed to foster a friendly, welcoming environment that promotes discussions. We gathered statistics from Jon Harmon who leads R4DS Slack and found that it has 6139 total members, 426 weekly active members, 130 of who have posted. There are also weekly office hours for learners who have more specific questions that mentors can answer.

When Thomas joined the R4DS online learning community, he wanted to start a smaller project designed to provide efficient practice for data scientists using R. The R4DS had experimented with a project called #TidyWeek [12] with a different goal from R4DS: pairing learners and experts with an emphasis on reviewing code. Interested learners would sign up for #TidyWeek, who would be matched with a mentor to get help and feedback on their R project that they were working on. However, the matching process and coordinating between mentors and learners became too difficult to sustain and required a high level of coordination for both learners and mentors.

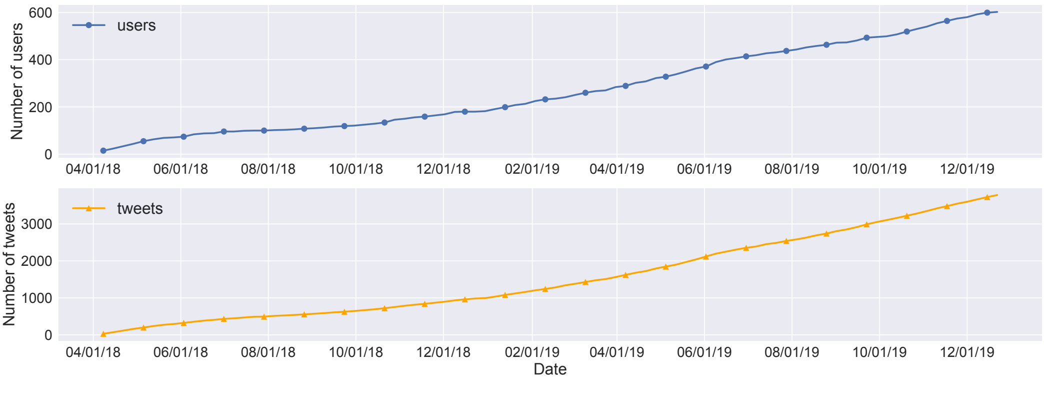
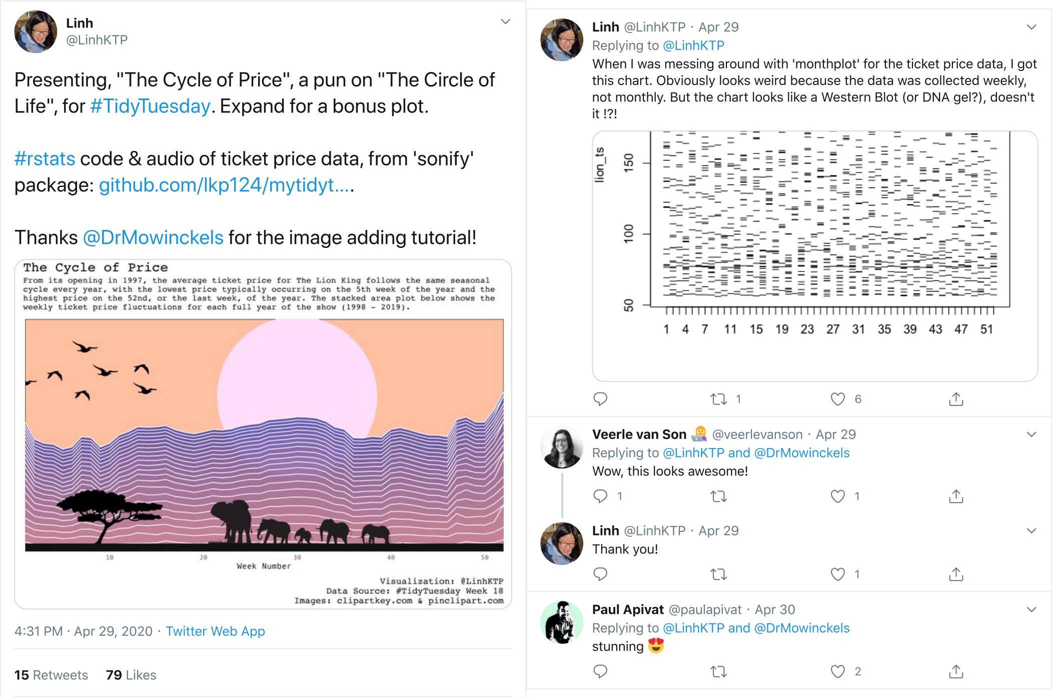


Fig. 1. Cumulative growth of unique #TidyTuesday users and tweets from April, 2018 to Jan, 2020.

3.1.2 Inception and growth of #TidyTuesday:. Thomas then spearheaded #TidyTuesday, which was designed to be a smaller, loosely structured project that could solve the issues of #TidyWeek and his own pain points learning R. After privately experimenting with the project and getting feedback from the R4DS online learning community leaders (C13, I14, I17), Thomas introduced the project to others in the R community, who are prevalent on Twitter (#rstats, #r4ds).[1] His main goal behind #TidyTuesday was to help himself and other data scientists practice their data wrangling and visualization by providing access to weekly released, real-world datasets, and to encourage sharing

[1] https://twitter.com/thomas_mock/status/980921600429252608

of code to facilitate social learning [69]. The #TidyTuesday project has been steadily growing since its inception making it a suitable case study of how an online community of practice for data scientists can grow and sustain itself on Twitter. As shown in Figure 1, #TidyTuesday has become quite popular in the R community with over 3834 unique tweets that contained the hashtag from 607 unique users from April, 2018 (inception) to January, 2020 time period.



Linh
@LinhKTP

Presenting, "The Cycle of Price", a pun on "The Circle of Life", for #TidyTuesday. Expand for a bonus plot.

#rstats code & audio of ticket price data, from 'sonify' package: github.com/lkp124/mytidyt....

Thanks @DrMowinckels for the image adding tutorial!

The Cycle of Price

From its opening in 1997, the average ticket price for The Lion King follows the same seasonal cycle every year, with the lowest price typically occurring on the 5th week of the year and the highest price on the 52nd, or the last week, of the year. The stacked area plot below shows the weekly ticket price fluctuations for each full year of the show (1998 - 2019).

Visualization: @LinhKTP
Data Source: #TidyTuesday Week 18
Images: clipartkey.com & pinclipart.com

4:31 PM · Apr 29, 2020 · Twitter Web App

15 Retweets **79** Likes

Linh @LinhKTP · Apr 29
Replying to @LinhKTP
When I was messing around with 'monthplot' for the ticket price data, I got this chart. Obviously looks weird because the data was collected weekly, not monthly. But the chart looks like a Western Blot (or DNA gel?), doesn't it !?!

Veerle van Son 👩‍💻 @veerlevanson · Apr 29
Replying to @LinhKTP and @DrMowinckels
Wow, this looks awesome!

Linh @LinhKTP · Apr 29
Thank you!

Paul Apivat @paulapivat · Apr 30
Replying to @LinhKTP and @DrMowinckels
stunning 😍

Fig. 2. An example of a #TidyTuesday submission tweet and feedback.

3.1.3 #TidyTuesday tweet anatomy:. An example of a #TidyTuesday submission tweet and feedback are shown in Figure 2. On the left is the submission tweet where the poster provides background about the dataset they analyzed and visualized. To help others reproduce their work, a poster typically includes a link (GitHub) to the code and an attached visualization (stacked area plot) exploring a certain aspect about the dataset. Sometimes, the poster will provide a description of the R packages (`sonify`), functions or tricks which can facilitate disseminating best practices. Moreover, credit is given to individuals (@DrMowinckels_e) who have helped them produce the plot. On the right of Figure 2 is an example of what the feedback typically looks like, where others may compliment them and the poster may provide further elaboration on their submission.

3.2 Research Questions

In order to understand what the #TidyTuesday participants' motivations were, how the project benefited them, and how it formed an online community of practice, we had three main research goals. The first goal was to understand why a data scientist participated in the project. The second goal was to understand the beneficial experiences that data scientists went through as they participated in #TidyTuesday. Finally, the third goal was to identify the various ways in which #TidyTuesday

Table 1. Demographics of interviewees

Social Role *	**ID**	**Gender**	**Degree**	**Field**	**Sector**	**Posts**
Poster	P1	M	Masters	Data Science	Academia	1
Poster	P2	M	Bachelors	Statistics	Academia	6
Poster	P3	F	Doctorate	Environmental Science	Academia	1
Poster	P4	F	Doctorate	Library Science	Academia	1
Poster	P5	M	Doctorate	Data Science	Industry	1
Poster	P6	M	Bachelors	Data Science	Industry	27
Poster	P7	M	Bachelors	Biotechnology	Healthcare	17
Poster	P8	M	Bachelors	Biostatistics	Industry	10
Poster	P9	F	Doctorate	Math Education	Academia	6
Poster	P10	M	Masters	Marine Ecology	Academia	1
Poster	P11	F	Doctorate	Marine Ecology	Academia	11
Poster	P12	F	Doctorate	Ecology Science	Academia	10
Poster	P16	F	Doctorate	Statistics	Academia	15
Poster	P25	M	Doctorate	Radiology	Healthcare	1
Poster	P26	F	Masters	Statistics	Academia	1
Curator	C13	M	Bachelors	Content Science	Industry	50
Curator	C19	M	Masters	Marketing	Industry	15
Curator	C20	M	Doctorate	Data Science	Industry	24
Curator	C21	F	Doctorate	Data Science	Academia	14
Curator	C24	M	Bachelors	Bioengineering	Healthcare	24
Influencer	I14	F	Masters	Data Science	Industry	7
Influencer	I15	M	Doctorate	Data Science	Industry	43
Influencer	I17	M	Doctorate	Data Science	Academia	78
Influencer	I18	F	Doctorate	Data Science	Industry	22
Influencer	I22	M	Doctorate	Ecology Science	Academia	62
Influencer	I23	M	Masters	Psychiatry	Healthcare	72

* Posters focus on posting and sharing their submissions with others. Curators make efforts to organize tweets and learning resources to make it easier for others to participate. Influencers grow and promote the movement.

helped form and sustain an online CoP on Twitter. Thus, we investigated three research questions:

- **Who participates in Tidy Tuesday and what are their motivations and goals?** To understand who typically participates in #TidyTuesday, we asked data scientists about their background, motivations and goals they hoped to accomplish with the project.
- **What do participants gain by participating in Tidy Tuesday?** To understand how #TidyTuesday benefits its participants, we asked questions about their overall experience with the project, including perceived benefits and challenges.
- **How does social activity around Tidy Tuesday cultivate a community of practice?** To investigate whether social activity around #TidyTuesday forms and sustains a CoP, we asked data scientists about their social interactions on Twitter and analyzed them through the lens of the CoP framework.

3.3 Interviews

Demographics and recruitment. We interviewed 26 total data scientists (Table 1). The participants were from 14 different fields and worked in 3 sectors: 13 in academia, 9 in industry, and 4 in healthcare. Participants self-reported age (18-24 = 1, 25-34 = 9, 35-44 = 9, 45-64 = 1, NA = 6), gender (14 men and 12 women), and education level (Bachelor's = 6, Masters = 6, Doctorate = 14). The participants had varying years of programming experience in R ranging from 1 to 3 (7), 3 to 5 (5), and 5 or more years (14). We recruited the participants using a combination of random and snowball sampling. We first used random sampling of the authors of tweets in the April 2018 to Jan 2020 time period and contacted them via email. We ensured that participants were actually making a submission for #TidyTuesday by inspecting their tweets; some were dropped and replaced due to unrelated tweets. For the posters group, we were interested in being able to contrast the experiences between different skill levels, so we tried recruiting between two groups: 7 *one-offs*, who only posted one submission and 8 *persistent*, who posted multiple submissions. To effectively recruit the curators and influencers, who were highly influential for growing #TidyTuesday and the R community at-large, we used snowball recruitment starting from Thomas Mock. We did not offer compensation for participants. To determine who to interview next, we used a constant comparison method [36] to guide our decisions about theoretical saturation.

Social roles: From our preliminary examination of the #TidyTuesday tweets, and tying in past literature on the various skill-sets and roles played by participants within communities of practice, we categorized our participants based on three social roles: poster, curator, and influencer Table 1 presents the full list of participants and their social roles in the project. There were 15 *posters*, who varied in their level of engagement with #TidyTuesday from "one-offs" (P1, P3, P4, P5, P10, P25, P26) who only posted their submission once to those who posted subsequent posts (P2, P6, P7, P8, P9, P11, P12, P16). Liu et al. [61] used a similar category for participants when analyzing tweets related to the #ILookLikeAnEngineer identity hashtag movement; however, we do not focus on "passive" readers for this study since we are only interested in why members participated in the first place. There were 5 participants who served the role of *curators*: in addition to posting their submissions, these participants engaged in organizing, highlighting submissions or contributing new tools to enhance the project. Data curators have been studied before by Zagalsky et al. [101] in Stack Overflow and mailing lists in the R community and by Middleton et al. [68] for baseball analytics within the Sabermetrics community. In this study, curators were interested in improving #TidyTuesday and the larger #rstats community on Twitter by organizing tweets and creating tools to facilitate participation. Finally, 6 were *influencers* who were already immersed in the R community and had a large following on Twitter. They were a mix of individuals who were either highly involved in making contributions with their own #TidyTuesday tweets or spreading the movement and encouraging others to join. This social role is based on Graham and Wright [37]'s study of "superposters" and the role they play in an online forum which found that, despite the potential for negative influence, they had a significant positive effect on others by helping them and being empathetic towards their problems.

Interview Protocol: The first author conducted 30-60 minute semi-structured interviews over Google Hangouts. All the interviews were done remotely using a template[2] which included questions around the following topics:

- Motivation to participate in #TidyTuesday.
- Experience participating in #TidyTuesday and its usage.
- Interactions with the R community on Twitter and elsewhere.

[2] https://go.ncsu.edu/tidytuesday

For each interview, the audio was recorded for transcription and analysis. The first author transcribed all of the interviews. We iteratively developed these questions based on a few pilot interviews to get meaningful responses around motivation, the experiences related to the daily hashtag, and community interaction. Topics like how participants engaged with the R community in contrast to other similar communities did not get much coverage; this is understandable as the R community might be their first one they engaged with. While the structure of the interviews remained constant, we let participants discuss other tangential topics.

3.4 Analysis

We audio-recorded and analyzed the transcripts of the semi-structured interviews. Before analysis, we first segmented the transcripts into different sections reflecting the semi-structured interview questions (Section 3.3). We then began analysis with open coding [23] on each topic looking for similarities and differences across the interviewees' thoughts or actions and assigning short phrases as codes. Some examples of first-level codes include codes like "accountability", "copy-paste code", or "coding alone". We wrote memos and engaged in continual comparison of the codes with one another and we performed focused coding [23], grouping similar codes and analyzing them to identify high level themes like "creating rhythm for practice", "enhancing technical and communication skills", or "community participation on Twitter". Finally, we refined themes in the central concepts of participation in the project, the daily hashtag's impact and cultivation of an online community of practice.

4 RESULTS

In our analysis, we found that participants had various intrinsic and extrinsic motivations, were positively transformed by participating in #TidyTuesday, and their social interactions on Twitter helped build a community of practice for data scientists using R. In the following sections, we discuss themes related to each research question.

4.1 Who participates in Tidy Tuesday and what are their motivations and goals?

Participants had various intrinsic and extrinsic motivates to participate in #TidyTuesday. Posters were mainly concerned with skill development and increasing their public presence in the R community through the project and their tweets. Curators, in addition to posting their submissions, were interested in organizing, highlighting submissions or contributing new tools to enhance the project. Finally, influencers were motivated to improve their skills, but wanted to focus their efforts on growing and promoting the project through various ways. We discuss the themes around motivation below and summarize them in Table 2.

4.1.1 Low barrier of entry:. To participate in #TidyTuesday, all participants were motivated by the low barrier to entry to contribute and get involved with the project. Participants expressed that it was easy to participate in #TidyTuesday because the requirements were minimal:

> *"To me, there's no better on-ramp then download R, download RStudio, install ggplot2 and then make your first plot. From there, when somebody makes their first plot they're like 'holy crap that is way better than me fussing around with Excel chart wizard.'"* (C20)

The lack of a formal onboarding or signup process made it easy for participants to join the project when they felt ready. In fact, many participants (P5, P6, P7, P9, P12, I15) were unsure on participating and passively observed the project on Twitter for a few weeks (3-4), learning about how others tweet and interact with one another before they deciding to get involved. The *"lurking"* (P16) behavior corroborates previous work related to legitimate peripheral participation (LPP), where users are found learning about the community practices and behaviors given access to user

Table 2. High level themes influencing participation

Theme	Representative Example
Low barrier of entry (Section 4.1.1)	"Instead of us having to come up with the idea of the data set Tidy Tuesday does that and it was just perfect." (P12)
Curated datasets for a time-boxed activity (Section 4.1.2)	"I had three hours a day where I was like on the subway essentially and it was the perfect thing to do during that time." (P16)
A weekly rhythm (Section 4.1.3)	"I kind of used it as a practice and a weekly sort of accountability." (P3)
Improve technical and communication skills (Section 4.1.4)	"Tidy Tuesday just seemed perfect because after a few weeks of just seeing a lot of people have done, you kind of pick up some tips about how to do stuff." (P6)
Connecting with the community (Section 4.1.5)	"I wanted to be a part of the R community first, and Tidy Tuesday with the visualizations was like a way to get there." (C20)

profiles and shared artifacts [20, 43, 65]. The choice of Twitter as the #TidyTuesday community public place provided similar affordances enabling LPP—participants can search, save, and observe others' tweets, which include links to the code on open source websites like GitHub, GitLab, or RPubs. Another aspect that attracted all participants to #TidyTuesday was that it is open for individuals of all skill levels, which helps achieve Wenger et al. [92]'s design principle to invite different levels of participation for growth of a CoP. Moreover, because the tweets are "easy access and public" (C20), it also satisfies Jones [49]'s requirement that a virtual settlement should include a common-public-place where members can meet and interact.

However, there were participants who only contributed once or twice which suggests potential barrier of entry issues. P10, P25, P5, and P4 all had similar reasons as to why they haven't posted more submissions on Twitter. P5 and P10 did not post more submissions simply because they did not having enough time. P4 also had time constraints but they also explained that they spent a long time figuring out exactly what they wanted to do with datasets, and would've preferred some ideas to help her get started. P25 differs from the rest because they were not satisfied with the #TidyTuesday visualization that they worked on and therefore decided against sharing it.

Asynchronous submissions: We also found that participants were motivated by the fact that the project does not require any deadlines for a particular #TidyTuesday. All of the posters stated in one form or another that this asynchronous nature of the project offered flexibility. Because #TidyTuesday submissions don't have deadlines, several participants expressed that they felt less pressure finishing it on Tuesday (P4, P6, P11, P16). Sometimes the reluctance to participate on a certain week was because they *"couldn't find anything interesting about [the dataset]"* (I22). If they couldn't practice on a dataset a certain week, participants were assured by the fact that they could always try the next #TidyTuesday. One participant even coined a new hashtag called #TardyTuesday to convey the fact that they don't usually complete their submission on Tuesdays:

> *"I think I coined Tardy Tuesday for a while because I never do them on Tuesdays. I get that if you do it on Tuesday, it's very defined and it's done. But Tuesday's a rough day of*

> *the week. I think on Thursday, I'd be more like able to sit down and do it but on Tuesday I'm either behind on something I need to have done or something else."* (P16)

4.1.2 Curated datasets for a time-limited activity:. Participants expressed that it was important that they had access to datasets every week that were manageable and could be timeboxed—allotting a fixed, maximum unit of time for an activity. As Sutton et al. [83] have noted, large datasets can present multiple problems leading to "death by a thousand wranglings". Several participants (P6, P7, P16, I22, I17, I15) described that the datasets hit a *"sweet spot"* (I15) in terms of the size, making them attractive for a quick analysis. This aligns with all of the participants' treatment of #TidyTuesday as a side activity for extra practice. These manageable datasets allowed participants to limit their practice sessions yet be able to do exploratory analysis and visualization on interesting datasets *"and be done with it in two or three days"* (I22). I15 provided a nice explanation of these characteristics of the #TidyTuesday datasets that attracted them:

> *"There's a sweet spot. You need to have a dataset that answer a variety of questions. It might be between—I would say between—a thousand and a million rows is about the right size for Tidy Tuesday. In fact, it's usually less than a hundred thousand. You could do more but it's still the right area. You could download it quickly and analyze it. And the number of columns between 5 and 20. And if you jump into datasets in the wild, a lot of them won't look like that."* (I15)

4.1.3 A weekly rhythm to stay engaged:. Participants were also interested in the *"rhythm"* (I18) and/or consistency provided by new datasets released every week. P3 described the *"burst"* of activity on Tuesday and throughout the week, which is reflected in Figure 3 where there's a flurry of tweets on Tuesday, with lesser tweets the following days.

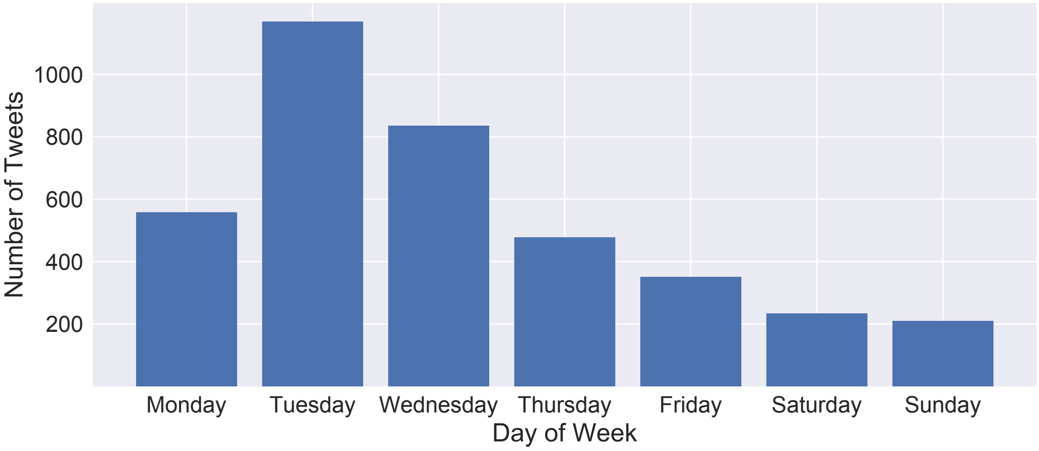


Fig. 3. Number of #TidyTuesday tweets each day of the week from April, 2018 to Jan, 2020.

The #TidyTuesday activity provide "aliveness", a key component to cultivate CoPs identified by Wenger et al. [92] as an indication that the project is active to attract newcomers. Indeed, the *"consistent delivery of a dataset every week"* (I15) attracted participants with either the *"accountability"* (P3, P4, P11) factor or as a source of *"inspiration"* (P3, P16, I17, I22), to help them maintain their engagement with the project and form a practice regimen. Regardless of skill levels, the participants described they needed this routine to stay consistent with their practice.

In addition to rhythm, participants were also motivated by the anticipation of working with new and different types of datasets every week. For example, P3, P12, and C20 described feeling excited about the new dataset for each week, as well as the different types of analysis or visualization that were possible. P3 added that they were inspired by the intricate and diverse visualizations every week. Part of that is thanks to a core set of "super-posters" [37] like I17 and I15, who consistently maintained high levels of posts (47 and 73 posts, respectively) with creative approaches on visualization datasets. Just like Cranshaw and Kittur [25], we found this leadership helped participants (P8, C20, I22, I23) motivate themselves to begin making their own contributions.

4.1.4 Improve technical and communication skills:. All participants were interested in using #TidyTuesday to hone their technical and communication skills, or *repertoire* [91], and build an online presence. In particular, learning was the most common intrinsic motivation between all participants, similar to developers in open source software by Ye and Kishida [100]; participants in our study wanted to engage in "exploratory learning" and "learning by doing". For those who were just getting started with R (P2, P6, P7, P8, C20), traditional resources like classes, MOOCs, online tutorials or books were insufficient in providing adequate practice. They described #TidyTuesday as a push for improving their data science skills and improve their technical abilities to deal with real-world datasets instead of toy datasets built into R datasets.

Those who were already familiar with base R were interested in polishing their skills around the tidyverse style of R (P6, P9, P16, C20, I17, I22, I23) or pushing their data visualization skills. These participants either wanted to get familiarized with tidyverse packages or build on skills they don't normally get to exercise:

> *"I learned R before the tidyverse and then I was like, "Oh I should probably learn the tidyverse, I'm getting a little bit passé." So I went through all the main packages and was teaching myself the differences between base R and whatever like dplyr, purrr—all that stuff. My least favorite thing is actually making plots and ggplot is my nemesis so I was like okay this is perfect because it makes me practice making plots."* (P16)

Participants (P2, P4, P16) were also interested in using #TidyTuesday as a forcing function for improving communication skills through blogging, making screencasts, or building a data visualization portfolio. This motivation has both intrinsic and extrinsic counterparts: participants wanted to improve their own personal skills while attracting the attention of others in the community and potential employers. They were also interested in data visualization techniques that they weren't normally used in their job. Influencers (I17, I22, I23) treated #TidyTuesday as a challenge to continually push the limits of data visualization skills using packages like `ggplot2`.[3]

4.1.5 Connecting and engaging with the R community:. Finally, all of the participants were motivated to participate because they wanted to become part of a community and engage in social learning. Similar to the motivation to improve their communication skills, participants had both intrinsic and extrinsic reasons to connect with the R community. Within McMillan and Chavis [67]'s sense of community (SoC) theory, they were seeking *membership* (feeling a sense of belonging) and *fulfillment of needs* (learning skills from others). For example, P2, P6, P16, and P9 described #TidyTuesday as a good motivator for both fitting into the R community and learning valuable skills from others:

> *"It helps learning a lot easier especially if you're trying to do it on your own. You're not really alone because you've got other people out there on Twitter or wherever sort of learning as well. I think the great thing about it is that it's this place where you sort of learn with other people even if you're not physically in contact with them."* (P6)

[3] https://ggplot2.tidyverse.org

The level of involvement participants wanted with the community was dependent on the participants' skill level and goals. Some of the posters and curators were newcomers to the R community (P8, P12, P16, P3, C20) who were motivated to join #TidyTuesday to absorb best practices and the norms of the R community, and increase their public presence online as part of their professional development. Put another way, #TidyTuesday became a way to audition for the R community on Twitter: *"I wanted to be a part of the R community first, and Tidy Tuesday with the visualizations was like a way to get there"* (C20).

The influencers who were already immersed in the R community were more interested in using #TidyTuesday as a way to *"give back to the community"* (I14, I15, I17, I22, I23) and helping newcomers have a welcoming experience. This fits under the SoC theory definition of *influence*, making a difference to a group. For example, I14, I17 and C13 were interested in fostering the welcoming culture and helping Thomas build the movement by participating themselves. For example, I17 accomplished this by participating heavily in #TidyTuesday to grow the movement and try to welcome beginners who are new to the project:

> *"Self-motivated or self-directed learning is pretty easy for [some people]. But, I know a lot of people where they feel isolated, and it feels challenging for them. I just wanted to help. I saw the community and saw what was there and thought, "Hey this is neat, everyone's helpful and how can I help people learn R?" because if I help others learn R, I'll learn more R. I'll polish my skills and be a helpful guide in the community."* (I17)

4.2 What do participants gain by participating in Tidy Tuesday?

#TidyTuesday transformed all of the participants regardless of their skill level both in regards to skill and professional development as well as getting more involved with the R community in general. We discuss these themes below and summarize them in Table 3.

Table 3. High level themes on the impact of #TidyTuesday

Theme	Representative Examples
Enhancement of skills through LPP (Section 4.2.1)	"I love looking at other people's visualizations and then reading their code and getting ideas on how I would build off of that." (P3)
Hashtags aided information retrieval (Section 4.2.2)	"I think I saw someone like have a package about you can make a gif or something and I was like wow that's a really cool package." (P4)
Expansion of skills outside of occupation (Section 4.2.3)	"Essentially, I just want to try some of the plots I don't get a chance to do in my research work." (P9)
Building an online presence for the job market (Section 4.2.4)	"The way I got that job was because I had all these blog posts and Tidy Tuesday stuff." (P16)

4.2.1 Enhancement of skills through legitimate peripheral participation: All participants commented that reading others' #TidyTuesday code and visualizations helped them enhance their own technical and communication skills. Having access to others' work through their tweets and GitHub links facilitated legitimate peripheral participation for all participants, an important component of situated learning [60] that helps explain how newcomers observed the project initially, then slowly participated by posting their own tweets. For example, P1, P3, P4, P11, and I22 practiced reverse-engineering skills by reusing and modifying others' code while working on their #TidyTuesday submission. Marlow and Dabbish [65] noticed the same social learning mechanism in designers using Dribbble. Influencers like I22, I17, and I23 were pleasantly surprised by how posters *"borrowed and extended"* (I17) their code, giving credence to the effectiveness of sharing code to transform peripheral members into experienced members. Since participants are primarily motivated by learning, this is in contrast to competition-fueled settings like Kaggle, where only a small minority of medium skill-level members shared and re-used code [84].

Through #TidyTuesday tweets, participants also formed impressions about others' skills and their commitment to the project which helped them keep track of those with similar interests or skill-sets. This parallels Dabbish et al. [27] and Marlow et al. [66]'s studies of GitHub, where they also found impression formation mechanisms using visible cues on GitHub like user profile and commits. We see similar visible cues with twitter profiles, tweets, and the visualizations. Among influencers like I17, I23, and I22, this enabled both learning *"creative approaches"* (I22) to visualizations and igniting *"friendly competition"* (I22). #TidyTuesday became a social learning tool for all skill levels to get inspired by others on what is possible with R and anticipate future posts from them:

> *"I love looking at other people's visualizations and then reading their code and getting ideas on how I would build off of that, but I haven't done a lot of them from scratch myself. I learn off of other people's awesome ideas that they share. Yeah that's a big part of it for me: I'm not looking to necessarily practice my skills as much as I am to be inspired and know what I can do based on what other people share."* (P3)

4.2.2 Hashtag as an information retrieval and R package discovery tool:. All participants mentioned that the #TidyTuesday hashtag allowed them to easily search for others' tweets and discover new packages in R. This need for searching others' submissions became such a common task, that a web application tool called Tidy Tuesday Rocks[4] was built to aggregate tweets and make them accessible in a central website. The hashtag made it possible to build the web app which helped participants (P4, P9, C20, I22, I23) search for past submissions that get buried due to high volumes of tweets or are difficult to find via Google searches:

> *"I've also used it to drill into the code, lift that off, and use it in my own. One of the maps I found on Tidy Tuesday rocks and I clicked [their] GitHub code and I was like, "Oh here's how [they] did it!" You see so many pages that are so heavily indexed over so long in Google that it's like really impossible to find what people are doing these days. I know that these things are recent because people did them over the last couple weeks."* (C20)

As C20 points out above, the #TidyTuesday hashtag became useful to find examples of R code that best reflects modern packages and techniques. In this way, the #TidyTuesday hashtag helped aid information retrieval and achieves Wenger et al. [92]'s "design for evolution" recommendation because it allowed aggregation of tweets based on the hashtag. Since the #TidyTuesday tweets were continually updated every week, participants were able to discover modern R packages that they had never used before:

> *"I find it a really good way to practice and try out new packages I've never used before. For example, for first time, I used the gganimate package for animating a plot, or ggalluvial package for doing flow diagram or alluvial plot."* (P6)

[4] tidytuesday.rocks

4.2.3 Expansion of skills outside of an occupation:. #TidyTuesday participants used the project to explore data visualizations that they don't normally get to make in their day-to-day job (P16, C20, I15, I17, I22, I23). Given the diversity of the datasets and the multiple ways of analyzing and visualizing them, #TidyTuesday became a *"choose your own adventure game"* (I17), which allowed participants to *"pursue something really weird"* (P16) beyond traditional visualizations:

> *"With my job, it's often kind of like longer-term projects. We're focused on this one specific thing. So in a given week, I'm not going to be working on mapping and NLP and animation and machine learning and stuff like that. So, it's kind of cool to have a different little thing to play with every week."* (C19)

R veterans like I22 and P9 commented how having access to the visualizations others produce also allowed them to challenge themselves and produce their own unique take on visualizations.

> *"I made a tree map just because it was really interesting. I think I got delayed and then came back like half an hour later and I was like "oh no some people have already used that idea!" But the only difference is you cannot interact with the tree map, and I thought I'm just going to add some additional things to it because it's gonna be boring to see two identical tree maps with different colors."* (P9)

Beyond R programming, the curators were able to use #TidyTuesday as a forcing function to improve other skills which was the same theme in Fiesler et al. [31]'s study of an online fandom community. For example, C20 and C21 were able to use #TidyTuesday to showcase their skills making tools that help others find tweets or understand others' code. C13 learned how to start and maintain a podcast and its respective website, while C24 learned how to use screencasts for covering #TidyTuesday submissions.

Improvement of communication skills with both written and other media. Several participants also used #TidyTuesday to polish their communication skills through blog posts or screencasts, which as a side-effect gave others opportunities for social learning. Often times, participants would expand on their thought process behind the code and visualizations in the form of blog posts (P2, P4, P6, I22, I17, I15, I18). The blog posts helped improve the participants' communication skills, as well as reveal the decisions made behind the code or plot. Influencers like I15 and I18 decided to make screencasts to improve their communication skills and help others learn about how to read, wrangle, analyze and visualize datasets. This helped them improve their own communication skills, while providing additional learning resources for newcomers via LPP [60]:

> *"The barrier to making a screencast is low. We already have quite an amount of written material. So using screencasts as an opportunity of getting another kind of medium out there and show how to use some of these open source packages was a good idea. Tidy Tuesday is a great example of another kind of content, and there are people out there who like watching videos and learn that way."* (I18)

4.2.4 Building an online presence for the job market:. There were many participants who used their #TidyTuesday submissions to build a portfolio for data analysis and visualization (P2, P6, P9, P16, P26, C20, I15, I22, I23). As mentioned earlier, some participants wrote corresponding blog articles on their personal website. We found that tweets, blog posts, and visualization portfolios around #TidyTuesday provided the same transparency and acted as signals for employers as found by Marlow and Dabbish [64] with activity traces and visual cues on candidates' GitHub profiles. By posting their work online either on Twitter, GitHub, YouTube or their personal sites, several participants (P2, P9, P16, I15) were able to attract recruiters and were offered interviews or jobs:

> *"I wrote a post and I think a couple weeks later, there was a consulting firm on the East Coast. They're primarily doing a project in healthcare and they called me and said, "Hey we saw your interactive [plot] on Twitter and we would like to get to know you. We feel like you're gonna be a good fit for one of the senior consultant positions." I was like sure! I was really surprised that everything kind of happened because of that one tweet!"* (P9)

This online presence even benefited experienced data scientist like I15, who made #TidyTuesday screencasts which employers used as convenient indicators of expertise:

> *"It's been awesome from an interviewing standpoint. People say how do you analyze your datasets. I'd say if you want to find examples, they are on YouTube. I did have one hiring manager who watched it all the way through and it definitely went well."* (I15)

4.3 How does social activity around Tidy Tuesday cultivate a community of practice?

We found #TidyTuesday and the social interactions afforded by Twitter helped crowdsource knowledge, disseminate best practices in R, bootstrap offline events, and build an inclusive, welcoming community. We present the high level themes below and summarize them in Table 4.

4.3.1 Promoting modern R practices:. The #TidyTuesday project encouraged the use of the tidyverse packages, which were created to provide R with consistency and ease of use based on the idea of tidy data [95]. Although all of the participants were familiar with this new style of R programming called tidy R, many had not made the transition yet. For example, P11, P6, P16 had significant experience in base R and made the transition to tidy R through #TidyTuesday:

> *"It's funny. Up until about a year ago, I honestly was on the base R side. But tidy R, it's so clean! I am fully committed now. I switched all my classes this last year to teaching tidy R in my undergrad classes too. There's just so much energy and effort that's being put into those packages and I think that's just like how it's moving. I still use [base R] but I moved almost completely away from for loops and things like that which I was completely attached to up until about a year ago. I would say really within the last year, I transitioned to being a more tidy coder."* (P11)

Table 4. High level themes on community building

Theme	**Representative Examples**
Promote best practices (Section 4.3.1)	"Up until about a year ago, I honestly was on the base R side. But tidy R, it's so clean like I am fully committed now." (P11)
Curation to satisfy community needs (Section 4.3.2)	"I just kind of started annotating stuff for myself and then I realized, oh I bet other people would find this useful too." (C19)
Bootstrap offline events (Section 4.3.3)	"Bringing Tidy Tuesday from Twitter and grounding it in real life as hacky hours made sense." (P3)
Promoting an inclusive, welcoming community (Section 4.3.4)	"On Twitter I've kind of had to come out of my shell to post stuff but every time I posted things or interacted with people, they've just been so wonderful and supportive." (C19)

Through #TidyTuesday, participants adopted tidyverse packages like `dplyr` for data wrangling or `ggplot2` for visualization, which can be considered modern coding practices of R due to its rising popularity. Zhu et al. [106]'s study of Wikipedia found that adapting best practices are helpful when the target audience has had some experience already. Thomas Mock and the initial contributors like I17 were already versed in tidy R, and since most participants were at least familiar with the new style of R programming, it proved an effective promotion strategy. #TidyTuesday on Twitter supported "trendspotting" (continued development by keeping up) that Marlow and Dabbish [65] discovered in their study on Dribbble.

4.3.2 Curation as a means to solve rising community needs:. The curators of #TidyTuesday have played a crucial role in enhancing the project and satisfying McMillan and Chavis [67]'s sense of community requirement of "integrating and fulfillment of needs" of the community. It also reflects the tendency of curation by the R community in other channels like Stack Overflow and R-help mailing lists [101], but adding new types of knowledge artifacts like web applications and interactive documents. In response to rising community needs, curators (C13, C19, C20, C21, C24) helped create packages and tools around # TidyTuesday related to organizing tweets, highlighting submissions, and walking through code.

Two packages were made to make dataset retrieval easier for any given week, and a web tool for browsing past submissions in useful ways. The `tidytuesdayr` package makes it easier to access #TidyTuesday datasets without leaving the RStudio IDE (Integrated Development Environment) [5] by requiring a single line of code to load the right dataset from a particular week: `tt_data <- tt_load("2019-01-15")`. This solves the potential challenge for loading data, especially for beginners. Since it is difficult to search for tweets and remembers specific plots, a web application called Tidy Tuesday Rocks[5] allowed the community to browse past submissions by dataset or username. This tool also served as an *"R gallery collection"* (C20), which further inspired several of our participants (P3, P11, I17, I22).

To highlight past submissions and walking through others' code, curators created #TidyTuesday code walk-throughs, and a podcast. The package called `flipbookr` [77] was created as an interactive slide document designed to walk through `ggplot2` code, line-by-line on #TidyTuesday techniques. Similarly, a project called #TidyX [47] was recently created to provide screencasts reviewing past submissions and explaining the code step-by-step, as well as pointing out R packages and techniques for data wrangling and visualization. There are also annotations for #TidyTuesday screencasts which help viewers jump into specific timestamps for information about a particular R function or technique.

4.3.3 Bootstrapping offline events to provide engagement for learners:. Perhaps the most surprising use of #TidyTuesday was bootstrapping in-person events. Some participants (P3, P9, P11, P12) used #TidyTuesday datasets to facilitate in-person events. P9 used #TidyTuesday for organizing a hackathon meetup at their university, which they described as a *"pain-free process"* because #TidyTuesday gave them easy access to the curated datasets. P3 and P12 started a weekly social coding club called "hacky hours" for students at their university holding in-person meetups to alleviate students' fear of programming, and provide them with a welcoming learning space:

> *"I thought it was really important to have an activity or a place where students could just try different things and try learning functions in a social setting that is totally no stress. Or if they fail completely, that's fine and hoping that they would learn something and progress while having fun. So Tidy Tuesday seemed like the perfect thing to kind of center*

[5] http://tidytuesday.rocks

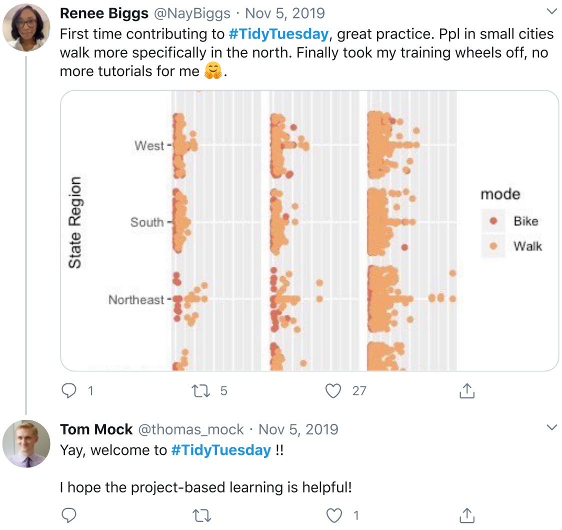
> *that goal and so we started our in-person Tidy Tuesday last spring. We've been sticking with it pretty much every Tuesday since during the academic year!"* (P12)

Similarly, P11 started a #TidyTuesday meetup with students at their university for similar benefits. However, for P11, they had never worked on #TidyTuesday by themselves and started doing them with a group. P3, P11, and P12 all prioritized mentorship of students, encouraging experienced members to help the beginners in the group. This echoes the theme of enabling LPP for learners in Section 4.2, but in an offline setting:

> *"I spend the majority of the time helping others, less so working on my own plots because I wanted to facilitate learning. I'm trying to encourage some of the more senior graduate students now to take that role away from me a little bit more so that they can practice teaching others and starting to feel more comfortable teaching some of the newer graduate students and helping them decode issues."* (P11)

These experiences add a new perspective on Gruzd et al. [38]'s study on Twitter as an imagined community and Wenger et al. [92]'s design components on cultivating CoPs. #TidyTuesday not only propped up an online CoP, but also facilitated in-person meetups, helping reduce efforts behind finding datasets, promoted learning via LPP, and allowed organizers to focus their efforts solely on logistics—setting up a calendar, planning the events, and choosing the locations. In effect, the in-person meetings combined the whole-community gathering taking place online (Twitter) which extends Gruzd et al. [38]'s finding that Twitter can be an imagined community, but complemented an offline small-group gathering (meetups), which adds an additional layer of rhythm and meets [92]'s recommendation of public and private community spaces.

4.3.4 Promoting a welcoming, inclusive community:. The #TidyTuesday community welcomed people of all skill levels, coming from very diverse backgrounds, creating an online space to practice R together. All participants felt they felt welcomed into the larger R community through their participation in #TidyTuesday. Thomas played a large role of ensuring that newcomers felt welcome, and shaping positive behavior for #TidyTuesday through his moderation and example-setting:



Renee Biggs @NayBiggs · Nov 5, 2019

First time contributing to #TidyTuesday, great practice. Ppl in small cities walk more specifically in the north. Finally took my training wheels off, no more tutorials for me 🤗.

Tom Mock @thomas_mock · Nov 5, 2019

Yay, welcome to #TidyTuesday !!

I hope the project-based learning is helpful!

Fig. 4. An example of Thomas welcoming a newcomer to #TidyTuesday.

Thomas expressed that he wasn't as focused on deterring certain behavior, but setting examples of positive behavior with his friendly replies on Twitter. This strategy has worked for #TidyTuesday and is backed by Seering et al. [80]'s finding that users tend to imitate both pro- and anti-social behavior. Participants (P5, P8, P12, C20) described the immediate contrast when asking questions in other online channels Stack Overflow compared to asking on Twitter. The following sentiment resonates with Ford et al. [33]'s study on barriers for participation on Stack Overflow which revealed that askers were hesitant in participating because they feared not receiving an answer back or receiving negative feedback:

> *"With Stack Overflow, everybody is looking to get points so you'll get people who will either not give great answers, but they want to get a point in, or you also get people who are very rude and very standoffish. If you ask how do you do something, instead of telling you an answer or how to get the answer, they'll kind of more or less insult you. We know Twitter to be a very volatile place, but for whatever reason with the R community, it wasn't. People were always very helpful, always very nice and I just enjoyed it a lot more."* (P8)

Despite only receiving passive feedback of likes or retweets, several participants (P6, P16, C19, C20) were surprised by the amount of positive feedback from the community. Within the SoC theory [67], by participating and receiving feedback, participants felt *membership* (feeling like they belong) and *influence* (mattering to its members). For example, P16 was encouraged to continue posting more #TidyTuesday tweets because they weren't used to receiving such feedback in their academic setting:

> *"I think mutual support of just like every time I get a like, I'm like, "oh somebody thinks that I did something cool!" I think that's the big thing in grad school that you also don't get—positive feedback frequently. So I know it's dumb to smile when I get a like on my tweet but I do."* (P16)

Sometimes, there was an *"happy accident"* (Thomas) of interactions between participants and highly renowned data scientists in the R community. For example, P11 recounted how Hadley Wickham liked one of their students' #TidyTuesday post on Twitter, which was very encouraging for the student.

#TidyTuesday also helped several participants increase their online presence. Through their contributions in #TidyTuesday, P3, P11, P12, and C19 dramatically increased their online presence and made an impact by helping others get involved with #TidyTuesday. P3, P11, and P12 commented that #TidyTuesday helped them mentor students and encouraged them to become more active online and reduce their fear of sharing work online, especially for those who were introverted. For C19, they described their transformative experience when they shared their annotations of #TidyTuesday screencasts with the R community on Twitter:

> *"I'm a pretty introverted person so even on Twitter I've kind of had to come out of my shell to post stuff, but I mean every time I posted a #TidyTuesday tweet or interacted with people, they've just been so wonderful and supportive. I've never seen any place on Twitter where people say, "This is so helpful! Thank you! Great job!" It's so amazing that the people are like this on Twitter. So it's been just a wholly positive experience."* (C19)

5 DISCUSSION

In this section, we discuss benefits and challenges behind #TidyTuesday and provide both the R community and similar communities with guidelines on how to effectively use a daily hashtag to build an online community of practice. The guidelines fall under three broad categories: barriers to entry for beginners, technological improvements to facilitate better learning, and social interactions to form and sustain a welcoming, inclusive online community.

5.1 Lowering the barriers to entry

In the following subsections, we discuss some barriers to entry for #TidyTuesday and provide suggestions on how to provide better onboarding.

5.1.1 *Providing better on-boarding for newcomers:.*

Can I participate? There were some data scientists who didn't realize #TidyTuesday was designed for everyone or weren't clear on the skill requirements to participate, even though everyone was welcome. For example, one programmer with minor experience in R asks:

> *"People who participate in #TidyTuesday: how much experience with R/coding in general did you have before doing it the first time? I've tried to do this week's task but I'm finding myself pretty lost."*[6]

This tweet suggests a barrier to entry issue related to getting started in #TidyTuesday, and the implicit skill requirements for a beginner. Steinmacher et al. [82] identified several relevant social barriers that stops students from participating in open source software (OSS), which included barriers like "newcomers need orientation" and "technical hurdles". Potential #TidyTuesday members might have difficulty knowing what skills they need to participate, which has been noted as a problem in Cranshaw and Kittur [25]'s study of Polymath Project, an online collaboration between mathematicians solving open problems. P4 suggested including beginner prompts for datasets about potential actions or questions the analyst can explore which addresses the barrier of "finding a task to start with" [82]. As for "technical hurdles", beginners in #TidyTuesday might face the challenge of the tooling required for #TidyTuesday, such as Git/GitHub for code sharing. To help lower this potential barrier to entry, P12 suggested a small tutorial for learning the bare minimum required for Git/GitHub. Learning resources are linked at the bottom of the Tidy Tuesday GitHub Readme file, but it might make the resources more visible if placed at the beginning, so that a newcomer can better orient themselves to the project.

5.1.2 Reduce fear aversion for novices:. Beginners wishing to participate might also suffer from a case of *fear aversion* after witnessing intricate code and stunning visualizations produced by experts. These submissions could either inspire or hurt learners. To reduce this fear, a possible remedy is for experts to point out they are also learners who sometimes struggle to produce visualizations for #TidyTuesday. As discussed in Section 4.2.1, some of the posters included a blog article or a screencast corresponding to their submission to provide further explanation behind the code and the plots. We believe these *"learning out loud"* (I14) activities done by experts can help beginners by highlighting the gradual, incremental steps taken towards producing those complex, creative plots:

> *"Those [blog posts] are awesome because people think that [experts] easily come up with these polished, awesome work. What the plots doesn't show is that, 'No, we failed and we did like 1500 experiments to get to where we are' or 'there's like a pile of sketches on paper on my desk."'* (I17)

We observed this initiative to blog and record screencasts only among a few of our participants (P4, P16, C24, I22, I15, I17), who were specifically interested in enhancing their communication skills (Section 4.1). We agree with C20 in encouraging experts in the community to learn out loud to help newcomers feel more comfortable participating in online social coding projects like #TidyTuesday.

5.2 Better mechanisms for practice and learning

5.2.1 Provide diverse forms of resources:. #TidyTuesday was effective in sharing knowledge and building an online CoP primarily because participants championed the idea of code-reuse and

[6] https://twitter.com/scottjdavies01/status/1201751167782408192

explanations through tweeting, blogging and making screencasts. Participants provided external, in-depth explanations in the form of blogs (P4, P6, P16, I17, I22) or screencasts (I15, I18). As we discussed in Section 4.2, only the influencers made #TidyTuesday screencasts, and it's important to note that they are experts in R. C20 and I22 suggested encouraging shorter screencasts to nudge the less experienced members of the community to take part creating video formats to showcase tips and tricks in R to further promote social learning:

> *"Stepping through somebody's code doesn't quite get you there. You don't get to hear their design choices or like why they split it up and into these two different things. I would love for the people who are creating these super stellar visualizations to take 10-15 minutes and go back and describe how they got there."* (C20)

Faas et al. [30] have found that live stream coding can support the growth of learning-focused communities that mentor both the streamer and each other during and after streams. Influencers such as I15 and I18 have provided screencasts which can help learners pick up "tricks in R" (I15). Live streams could further enhance the learning experience by allowing questions in-situ, a feature which blog posts and pre-recorded screencasts lack.

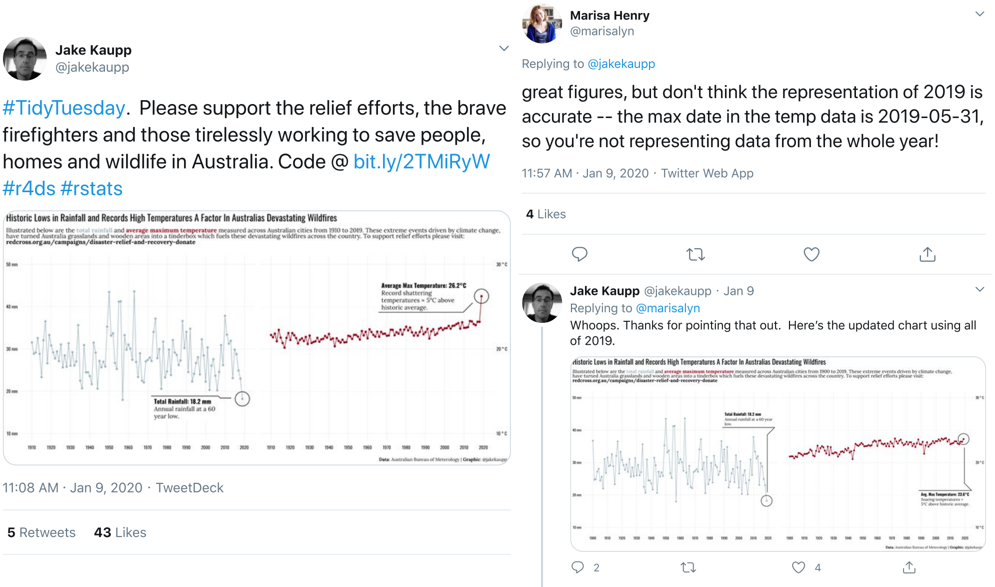
5.2.2 Provide mechanisms for constructive feedback and mentorship:.

Twitter as a platform to promote friendlier learner interactions: The Twitter platform provided a friendly experience for #TidyTuesday to learn and practice R, compared to other channels like Stack Overflow or R-help mailing lists. For example, beyond questions and answers, Twitter replies can initiate discussion threads which are discouraged on Stack Overflow, while still allowing "participatory knowledge creation" [101]. Unlike the aggressive behavior found in R-help mailing lists, #TidyTuesday participants found very little aggression on Twitter, a surprising finding that we will further discuss in Section 5.3.2. All participants expressed that the R community on Twitter (via #TidyTuesday and #rstats) has a welcoming attitude towards beginners, allowing follow up discussions and simple questions without any fear of scorn or negative comments. I14 pointed out that *"this was not always the case"* and the leaders of the R community such as R-Ladies Global and RStudio have helped changed the culture. However, I14 also expressed caution that R-specific community forums—such as RStudio Community[7]—could potentially lead to the "posting is hard" barrier to entry on Stack Overflow [33] by requiring questions to be structured a specific way (for e.g. using a `reprex`[8] for a reproducible example). We suggest the R community and other online communities to consider SNS sites like Twitter to form a community of practice that allows casual dialog, ongoing discussion threads, and friendly interactions.

Limitations of Twitter as a platform for online learning: While #TidyTuesday participants received positive feedback, they did not always provide constructive feedback. Kou and Gray [56] studied *distributed critique*, a set of critique practices whereby geographically distributed creators engage in the critique of design artifacts and processes. We sometimes see evidence of this type of interaction (Figure 5). However, constructive feedback was rare among our participants. There has been an attempt to resolve this with a related hashtag called #RFeedbackFriday, to explicitly ask for feedback, but P6 commented, *"I tried it but did not receive any feedback and it may have fizzled out."* Another online CoP for healthcare has encountered similar difficulties on Twitter [35]. Twitter is a public space and restricts tweets to 280 character limit, so it may stifle meaningful feedback required by learners because of a fear of publicly criticizing others, or not allowing depth. We echo Wenger

[7] https://community.rstudio.com
[8] https://reprex.tidyverse.org

et al. [92]'s suggestion for a "backchannel" for private conversations might help to provide deeper interaction between members of the community, using applications like Slack.



Jake Kaupp
@jakekaupp

#TidyTuesday. Please support the relief efforts, the brave firefighters and those tirelessly working to save people, homes and wildlife in Australia. Code @ bit.ly/2TMiRyW #r4ds #rstats

Historic Lows in Rainfall and Records High Temperatures A Factor In Australias Devastating Wildfires
Illustrated below are the total rainfall and average maximum temperature measured across Australian cities from 1910 to 2019. These extreme events driven by climate change, have turned Australia grasslands and wooden areas into a tinderbox which fuels these devastating wildfires across the country. To support relief efforts please visit: redcross.org.au/campaigns/disaster-relief-and-recovery-donate
50 mm
40 mm
30 mm
20 mm
10 mm
30 °C
20 °C
10 °C
1910 1920 1930 1940 1950 1960 1970 1980 1990 2000 2010 2020
Total Rainfall: 18.2 mm
Annual rainfall at a 60 year low.
Average Max Temperature: 26.2°C
Record shattering temperatures ≈ 5°C above historic average.
Data: Australian Bureau of Meterology | Graphic: @jakekaupp

11:08 AM · Jan 9, 2020 · TweetDeck

5 Retweets 43 Likes

Marisa Henry
@marisalyn

Replying to @jakekaupp

great figures, but don't think the representation of 2019 is accurate -- the max date in the temp data is 2019-05-31, so you're not representing data from the whole year!

11:57 AM · Jan 9, 2020 · Twitter Web App

4 Likes

Jake Kaupp @jakekaupp · Jan 9
Replying to @marisalyn
Whoops. Thanks for pointing that out. Here's the updated chart using all of 2019.

Historic Lows in Rainfall and Records High Temperatures A Factor In Australias Devastating Wildfires
Illustrated below are the total rainfall and average maximum temperature measured across Australian cities from 1900 to 2019. These extreme events driven by climate change, have turned Australia grasslands and wooden areas into a tinderbox which fuels these devastating wildfires across the country. To support relief efforts please visit: redcross.org.au/campaigns/disaster-relief-and-recovery-donate
50 mm
40 mm
30 mm
20 mm
10 mm
30 °C
20 °C
10 °C
1900 1910 1920 1930 1940 1950 1960 1970 1980 1990 2000 2010 2020
Total Rainfall: 18.2 mm
Annual rainfall at a 60 year low.
Avg. Max Temperature: 23.6°C
Soaring temperatures ≈ 5°C above historic average.
Data: Australian Bureau of Meterology | Graphic: @jakekaupp

2 4

Fig. 5. An example of constructive criticism on a #TidyTuesday submission tweet.

Better infrastructure to facilitate feedback and mentorship: To better facilitate deeper learner/mentor interactions, we believe a better technology infrastructure is required that explicitly focuses on receiving feedback and mentorship. C21 engaged in the #MakeOverMonday—a daily hashtag for visualizations using Tableau—and expressed how the project has a central site that is used for critique by experts in a webinar. Such efforts require time, energy and funding which is beyond the scope of a voluntary effort like #TidyTuesday. However, we suggest taking advantage of a central place like the #R4DS online learning community on Slack, which could help support the one-on-one conversations, or Ford et al. [32]'s just-in-time mentorship learners to inform cultural norms. Alternately, one could draw inspiration from Xu et al. [98]'s system Voyant, which allows users to get feedback of their designs from the crowd. A similar system could be re-purposed for data scientists so that they can receive feedback on their plots or code.

5.3 Organically growing an online learning CoP

5.3.1 Choose technologies that support an open structure for growth:. We believe a big part of the success behind #TidyTuesday was Thomas' decision in keeping the core structure of the project loose—a weekly dataset and a few rules to participate—yet allowing others to build on the project by making it open source. In other words, Thomas "designed for evolution" which is recommended by Wenger et al. [92] to organically grow a CoP. #TidyTuesday accomplished this well by choosing the right technologies: Twitter and GitHub.

Hosting #TidyTuesday on GitHub helped Thomas maintain the project for free and make use of crowdsourcing efforts for finding new datasets, fixing problems regarding those datasets, or making improvements to the project. Through GitHub's issues, Thomas was able to get help on fixing uploaded datasets[9], or project-related resources like information in the documentation[10]. However, P16 pointed out that a potential issue for a single moderator is burnout, a state of exhaustion caused by excessive and prolonged stress. This is similar to the burnout identified by Fiesler et al. [31] for experienced coders, which can be an issue for a project maintainer. To combat this, P16 suggested adopting the idea of "RoCur" or Rotating Curator: rotating the spokesperson on a social media account, where every week, a different member of the community manages the Twitter account, sharing their their views on using R, as well as tips and tricks. This is directly inspired by the @WeAreRLadies[11] effort out of the R-Ladies Global. For #TidyTuesday, a RoCur candidate might be a highly motivated individual such as a super-poster [37], a curator or an influencer, to help curate/clean a dataset and interact with and promote others' tweets.

As mentioned in Section 4.2, using Twitter as the sharing platform for #TidyTuesday submissions enabled curations of various forms. Firstly, the hashtag tagging mechanism helped grow the movement by supporting information retrieval (Section 4.2.2), accruing knowledge and forming links to various learning resources. Hashtags became such a useful mechanism for growing #TidyTuesday that it sprung two new daily hashtags from our participants (#TardyTuesday and #Tidydors). Having access to submissions via Twitter also allowed curators to organize the tweets and/or artifacts produced by posters as well as provide further pedagogy on particular submissions for learners. As the project evolved, members started becoming aware of particular challenges of #TidyTuesday and sought to improve the project using their skills outside of R programming alone. Since Thomas welcomed anyone to help fulfill these needs, some members in the community jumped on the opportunity. For example, in Section 4.2, we mentioned the benefits of having access to others' code but passively reading code might not be helpful for learners. Curators stepped in to solve this issue by either highlighting specific packages and techniques (C13), walking through the code line-by-line (C24), or showing the incremental evolution of the code (C21).

To maintain and sustain the growth of an online community, we encourage keeping the core structure loose, and choose technologies that can help promote contributions and promote curations to enhance the project.

5.3.2 *Encourage experts and influencers to engage with newcomers:.*

The success behind #TidyTuesday in fostering an inclusive, welcoming online community of practice on Twitter owes a large part to the involvement of Thomas, influencers and other leaders within the existing R community. At the rstdio::conf 2020 [14], Kate Hertweck delivered a a talk about how R communities are unparalleled in their inclusivity and commitment to learning collectively [42]. She noted several solutions to reduce barriers of entry like managing expectations and creating interest through expectation of activity and continuity. We believe Thomas has accomplished expectations by simply promising a dataset every week, leaving the task of analysis and visualization open-ended. Thomas also used the #TidyTuesday hashtag to create rhythm (Wenger et al. [93]) within the community and helps create what Kate recommended as "FOMO", the fear of missing out.

We also want to highlight the importance of the larger R community and the key players mentioned in Section 2.3 which provided a solid foundation for promoting inclusivity and diversity of members for #TidyTuesday. With the existing #rstats and #R4DS communities on Twitter, which embraces this culture promoted by stakeholders like RStudio, R-Ladies Global, rOpenSci[12], Thomas was able to carry this spirit forward with the #TidyTuesday project by example-setting positive behavior [80]. In a recent tweet, a user asked:

[9] https://GitHub.com/rfordatascience/tidytuesday/issues/186
[10] https://GitHub.com/rfordatascience/tidytuesday/issues/162
[11] https://twitter.com/WeAreRLadies
[12] https://ropensci.org

> *"I'm curious as to how the R community came to be so supportive and welcoming (as opposed to so much of the tech world). Anyone have ideas? #rstats"*[13]

A number of responses followed including Hadley Wickham and Jenny Bryan who provided an insight as to why Twitter is becoming a welcoming and inclusive space for the R community. Hadley commented *"that this wasn't at all the case 10 years ago"* and that perhaps *"each shift to a new space (r-help -> SO -> twitter) can be accompanied a refocusing of shared goals"* or *"it's just dominated by founder effects and shifts tend to be led by younger folks who are more in touch with initial pain of learning."* Indeed, in Section 4.3.4, we mentioned how Hadley liked a #TidyTuesday post by P11's student. Receiving the attention of leaders leads to what McMillan and Chavis [67] calls *membership* (sense of belonging) and *influence* (mattering to the group) within the sense of community theory. Jenny added that the *"growing role of twitter and @RLadiesGlobal creates space for new voices (vs "sorry all spots were filled 10 yrs ago")."* This type of leadership and initiative seems unique to the R community and is embraced within the new space on Twitter, a contrast to other channels like Stack Overflow, which has presented several barriers to entry, especially for women [33].

However, to prevent *"insular"* (P16) data science communities, as suggested by I15 and P16, #TidyTuesday could be opened up for the Python community[14], where users are adopting the tidy data framework [17, 45, 99]. Hence, #TidyTuesday can benefit other similar communities and help them cultivate their own online community of practice around #TidyTuesday, uniting even more data scientists in their efforts to become experts.

To organically evolve a new online CoP over time, we find that influencers and leaders of the community can play a vital role in growing the community by engaging and interacting with members of all skill levels, and instilling the feeling of being part of a community.

6 LIMITATIONS

Our analysis of the #TidyTuesday project represents an initial understanding of the dynamics and nature of participation in daily hashtag, social coding projects. We studied #TidyTuesday for the R community using a qualitative approach through semi-structured interviews. However, there are some limitations to our approach that represent opportunities for future research.

Generalizability: Our findings are drawn from one daily hashtag out of several others targeted at data scientists. For example, we did not study #MakeOverMonday, which is another daily hashtag serving a different need: using the Tableau software to create data visualizations instead programming using R. Hence, the themes we derived around motivations, skill development, professionalization, and community growth only represent the participants' experience at this moment in time for #TidyTuesday. It is important to note that some aspects of our findings might be unique to this project and the R community. Future research should explore other data science communities to improve our understanding of how to use daily hashtags as a tool for growing an online community of practice.

Participation bias: Another potential limitation of our study is a self-selection bias in our interviewee sample. Our sample was selected to contrast the experiences of people who participated in #TidyTuesday but, we do not know the experiences of *readers* [16, 61] who passively engage by viewing, liking, or retweeting tweets. As a result, we may miss out on important issues related to barriers of entry. Participants also knew the study was about a discussion of the #TidyTuesday project, which may have influenced our sample towards those with the strongest feelings about

[13] https://twitter.com/OwenChurches/status/1254634256472485896

[14] https://www.scipy.org

the project. We mitigated this issue by using random sampling for all of the posters and curators and recruiting people of different skill levels and engagement with #TidyTuesday. However, we do not know the feelings or experiences of non-respondents and can only compare their tweets and participation levels.

Qualitative method: The challenges and recommendations we provide for daily hashtags as a way to provide an online learning CoP are based on participants' perceptions and experiences of #TidyTuesday, not quantitative indicators of the hashtag's effect on their behavior. To derive themes, we used qualitative coding to analyze and interpret our data which is limited by theoretical sensitivity and the synthesis conducted by the researchers participating in that process. We followed the guidelines set by Carlson [22] and performed a single-event member check with our results. 22 participants replied and agreed with our results and requested minor changes to their quotations or demographic information. Future studies should examine quantitative aspects of a daily hashtag project such as the dynamics of the tweets, or how the project spread on Twitter. Rosenberg et al. [78], for example, have started an investigation of the posters' code itself, which offers insights on code evolution over time as an indicator for skill development. These quantitative measures can be useful to characterize #TidyTuesday, but we believe our themes provide rich insights and offer new directions for future work to further understand the benefits and challenges associated with the use of daily hashtags.

7 CONCLUSION

In this study, we conducted a qualitative case study on #TidyTuesday—a social coding project for data scientists using R—using the framework of CoP, and extending previous work related to forming and sustaining online CoPs on Twitter. From our analysis of semi-structured interviews with 26 participants, we examined motivations and goals of data scientists participating in #TidyTuesday and how it benefited them. We found that the participants were attracted to the rhythm provided by the project, the opportunity for professional development, and becoming part of the larger R community. Through #TidyTuesday, participants enhanced both technical and communication skills by learning from others, adopting best practices in R, and building an online presence. #TidyTuesday was effective in forming an online CoP by disseminating best practices, providing opportunities for curations to satisfy community needs, bootstrapping offline events and promoting an inclusive, welcoming community. Based on our findings, we discussed several benefits and limitations to using daily hashtags on Twitter to form a community and provide guidelines on cultivating a successful CoP using a daily hashtag such as placing a low barrier of entry for newcomers by providing onboarding, normalizing the sharing of code and artifacts to promote social learning, and making room for evolution for organic growth and sustenance. We believe daily hashtags can be adopted by other data science communities interested in cultivating an online CoP.

ACKNOWLEDGMENTS

This material is based in part upon work supported by the National Science Foundation under Grant No. 2006947.

REFERENCES

[1] [n.d.]. Certified B Corporation. https://bcorporation.net
[2] [n.d.]. Dribbble - Discover the World's Top Designers & Creative Professionals. https://dribbble.com
[3] [n.d.]. R-Ladies Global & R-Ladies is a world-wide organization to promote gender diversity in the R community. https://rladies.org
[4] [n.d.]. R: The R Project for Statistical Computing. https://www.r-project.org
[5] [n.d.]. RStudio - RStudio. https://rstudio.com/products/rstudio/

[6] [n.d.]. RStudio | Open source & professional software for data science teams - RStudio. https://rstudio.com

[7] [n.d.]. Slack. https://slack.com

[8] [n.d.]. The Carpentries. https://carpentries.org

[9] [n.d.]. Tidyverse. https://www.tidyverse.org

[10] [n.d.]. Vacation Rentals, Homes, Experiences & Places - Airbnb. https://www.airbnb.com

[11] [n.d.]. We're changing the way you think about data. https://www.tableau.com/trial/tableau-software

[12] 2018. tidyweek. https://GitHub.com/rfordatascience/tidyweek

[13] 2020. The 50 Most Popular MOOCs of All Time. https://www.onlinecoursereport.com/the-50-most-popular-moocs-of-all-time/

[14] 2020. rstudio::conf. https://rstudio.com/resources/rstudioconf-2020/

[15] Benedict Anderson. 2006. *Imagined communities: Reflections on the origin and spread of nationalism.* Verso books.

[16] Judd Antin and Coye Cheshire. 2010. Readers Are Not Free-Riders: Reading as a Form of Participation on Wikipedia. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work (CSCW '10)*. Association for Computing Machinery, New York, NY, USA, 127–130. https://doi.org/10.1145/1718918.1718942

[17] Tom Augspurger. 2016. Modern Pandas (Part 5): Tidy Data. https://tomaugspurger.github.io/modern-5-tidy.html

[18] Austin Cory Bart, Dennis Kafura, Clifford A Shaffer, and Eli Tilevich. 2018. Reconciling the Promise and Pragmatics of Enhancing Computing Pedagogy with Data Science. In *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*. 1029–1034.

[19] Oscar Baruffa and Veerle van Son. 2020. Twitter for R programmers. https://www.t4rstats.com/index.html

[20] Susan L. Bryant, Andrea Forte, and Amy Bruckman. 2005. Becoming Wikipedian: Transformation of Participation in a Collaborative Online Encyclopedia. In *Proceedings of the 2005 International ACM SIGGROUP Conference on Supporting Group Work (GROUP '05)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/1099203.1099205

[21] Kat Campise. 2020. Guide to Data Science Bootcamps — Complete listing of Bootcamps in the US. https://www.discoverdatascience.org/programs/data-science-bootcamps/

[22] Julie A Carlson. 2010. Avoiding traps in member checking. *Qualitative Report* 15, 5 (2010), 1102–1113.

[23] Kathy Charmaz. 2006. *Constructing grounded theory: A practical guide through qualitative analysis.* sage.

[24] Souti Chattopadhyay, Ishita Prasad, Austin Z Henley, Anita Sarma, and Titus Barik. 2020. What's Wrong with Computational Notebooks? Pain Points, Needs, and Design Opportunities. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[25] Justin Cranshaw and Aniket Kittur. 2011. The polymath project: lessons from a successful online collaboration in mathematics. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1865–1874.

[26] Anqi Cui, Min Zhang, Yiqun Liu, Shaoping Ma, and Kuo Zhang. 2012. Discover breaking events with popular hashtags in twitter. In *Proceedings of the 21st ACM international conference on Information and knowledge management*. 1794–1798.

[27] Laura Dabbish, Colleen Stuart, Jason Tsay, and Jim Herbsleb. 2012. Social coding in GitHub: transparency and collaboration in an open software repository. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*. 1277–1286.

[28] Thomas H Davenport and DJ Patil. 2012. Data scientist. *Harvard business review* 90, 5 (2012), 70–76.

[29] Ingrid Erickson. 2008. The translucence of Twitter. In *Ethnographic praxis in industry conference proceedings*, Vol. 2008. Wiley Online Library, 64–78.

[30] Travis Faas, Lynn Dombrowski, Alyson Young, and Andrew D. Miller. 2018. Watch Me Code: Programming Mentorship Communities on Twitch.Tv. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article Article 50 (Nov. 2018), 18 pages. https://doi.org/10.1145/3274319

[31] Casey Fiesler, Shannon Morrison, R Benjamin Shapiro, and Amy S Bruckman. 2017. Growing their own: Legitimate peripheral participation for computational learning in an online fandom community. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 1375–1386.

[32] Denae Ford, Kristina Lustig, Jeremy Banks, and Chris Parnin. 2018. "We Don't Do That Here": How Collaborative Editing with Mentors Improves Engagement in Social Q&A Communities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, Article Paper 608, 12 pages. https://doi.org/10.1145/3173574.3174182

[33] Denae Ford, Justin Smith, Philip J. Guo, and Chris Parnin. 2016. Paradise Unplugged: Identifying Barriers for Female Participation on Stack Overflow. In *Proceedings of the 2016 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering (FSE 2016)*. Association for Computing Machinery, New York, NY, USA, 846–857. https://doi.org/10.1145/2950290.2950331

[34] Daniel M German, Bram Adams, and Ahmed E Hassan. 2013. The evolution of the R software ecosystem. In *2013 17th European Conference on Software Maintenance and Reengineering*. IEEE, 243–252.

[35] Sarah Gilbert. 2016. Learning in a Twitter-based community of practice: an exploration of knowledge exchange as a motivation for participation in #hcsmca. *Information, Communication & Society* 19, 9 (2016), 1214–1232.

[36] Barney G Glaser and Anselm L Strauss. 1967. Grounded theory: Strategies for qualitative research. *Chicago, lL: Aldine Publishing Company* (1967).

[37] Todd Graham and Scott Wright. 2014. Discursive equality and everyday talk online: The impact of "superparticipants". *Journal of Computer-Mediated Communication* 19, 3 (2014), 625–642.

[38] Anatoliy Gruzd, Barry Wellman, and Yuri Takhteyev. 2011. Imagining Twitter as an imagined community. *American Behavioral Scientist* 55, 10 (2011), 1294–1318.

[39] Philip Jia Guo. 2012. *Software tools to facilitate research programming.* Ph.D. Dissertation. Stanford University Stanford, CA.

[40] Philip J Guo, Sean Kandel, Joseph M Hellerstein, and Jeffrey Heer. 2011. Proactive wrangling: Mixed-initiative end-user programming of data transformation scripts. In *Proceedings of the 24th annual ACM symposium on User interface software and technology.* 65–74.

[41] Johanna Hardin, Roger Hoerl, Nicholas J Horton, Deborah Nolan, Ben Baumer, Olaf Hall-Holt, Paul Murrell, Roger Peng, Paul Roback, D Temple Lang, et al. 2015. Data science in statistics curricula: Preparing students to "think with data". *The American Statistician* 69, 4 (2015), 343–353.

[42] Kate Hertweck. 2020. If you build it, they will come...but then what? Facilitating communities of practice in R. https://rstudio.com/resources/rstudioconf-2020/if-you-build-it-they-will-come-but-then-what-facilitating-communities-of-practice-in-r/

[43] Maya Holikatti, Shagun Jhaver, and Neha Kumar. 2019. Learning to Airbnb by engaging in online communities of practice. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–19.

[44] Youyang Hou and Dakuo Wang. 2017. Hacking with NPOs: collaborative analytics and broker roles in civic data hackathons. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–16.

[45] Jean-Nicholas Hould. [n.d.]. Tidy Data in Python. https://www.jeannicholashould.com/tidy-data-in-python.html

[46] Bernardo A Huberman, Daniel M Romero, and Fang Wu. 2008. Social networks that matter: Twitter under the microscope. *arXiv preprint arXiv:0812.1045* (2008).

[47] Ellis Hughes and Patrick Ward. [n.d.]. TidyX. https://www.youtube.com/channel/UCP8l94xtoemCH_GxByvTuFQ/

[48] Ross Ihaka and Robert Gentleman. 1996. R: a language for data analysis and graphics. *Journal of computational and graphical statistics* 5, 3 (1996), 299–314.

[49] Quentin Jones. 1997. Virtual-communities, virtual settlements & cyber-archaeology: A theoretical outline. *Journal of Computer-Mediated Communication* 3, 3 (1997), JCMC331.

[50] Sean Kandel, Andreas Paepcke, Joseph M Hellerstein, and Jeffrey Heer. 2012. Enterprise data analysis and visualization: An interview study. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (2012), 2917–2926.

[51] Mary Beth Kery, Amber Horvath, and Brad A Myers. 2017. Variolite: Supporting Exploratory Programming by Data Scientists. In *CHI*, Vol. 10. 3025453–3025626.

[52] Mary Beth Kery, Marissa Radensky, Mahima Arya, Bonnie E John, and Brad A Myers. 2018. The story in the notebook: Exploratory data science using a literate programming tool. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems.* 1–11.

[53] Miryung Kim, Thomas Zimmermann, Robert DeLine, and Andrew Begel. 2016. The emerging role of data scientists on software development teams. In *2016 IEEE/ACM 38th International Conference on Software Engineering (ICSE).* IEEE, 96–107.

[54] Amy J Ko, Robin Abraham, Laura Beckwith, Alan Blackwell, Margaret Burnett, Martin Erwig, Chris Scaffidi, Joseph Lawrance, Henry Lieberman, Brad Myers, et al. 2011. The state of the art in end-user software engineering. *ACM Computing Surveys (CSUR)* 43, 3 (2011), 1–44.

[55] Patty Kostkova, Martin Szomszor, and Connie St. Louis. 2014. #swineflu: The use of twitter as an early warning and risk communication tool in the 2009 swine flu pandemic. *ACM Transactions on Management Information Systems (TMIS)* 5, 2 (2014), 1–25.

[56] Yubo Kou and Colin M. Gray. 2017. Supporting Distributed Critique through Interpretation and Sense-Making in an Online Creative Community. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article Article 60 (Dec. 2017), 18 pages. https://doi.org/10.1145/3134695

[57] Yubo Kou, Colin M. Gray, Austin L. Toombs, and Robin S. Adams. 2018. Understanding Social Roles in an Online Community of Volatile Practice: A Study of User Experience Practitioners on Reddit. *Trans. Soc. Comput.* 1, 4, Article Article 17 (Dec. 2018), 22 pages. https://doi.org/10.1145/3283827

[58] Andy Kriebel and Murray Eva. [n.d.]. Makeover Monday. https://www.makeovermonday.co.uk

[59] Sean Kross and Philip J Guo. 2019. Practitioners teaching data science in industry and academia: Expectations, workflows, and challenges. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* 1–14.

[60] Jean Lave, Etienne Wenger, et al. 1991. *Situated learning: Legitimate peripheral participation.* Cambridge university press.

[61] Fannie Liu, Denae Ford, Chris Parnin, and Laura Dabbish. 2017. Selfies as social movements: Influences on participation and perceived impact on stereotypes. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 72.

[62] Steve Lohr. 2017. Where the STEM jobs are (and where they aren't). *New York Times* 1 (2017).

[63] Cecilia Loureiro-Koechlin and Tim Butcher. 2013. The emergence of converging communities via Twitter. *The Journal of Community Informatics* 9, 3 (2013).

[64] Jennifer Marlow and Laura Dabbish. 2013. Activity traces and signals in software developer recruitment and hiring. In *Proceedings of the 2013 conference on Computer supported cooperative work.* 145–156.

[65] Jennifer Marlow and Laura Dabbish. 2014. From Rookie to All-Star: Professional Development in a Graphic Design Social Networking Site. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '14).* Association for Computing Machinery, New York, NY, USA, 922–933. https://doi.org/10.1145/2531602.2531651

[66] Jennifer Marlow, Laura Dabbish, and Jim Herbsleb. 2013. Impression formation in online peer production: activity traces and personal profiles in github. In *Proceedings of the 2013 conference on Computer supported cooperative work.* 117–128.

[67] David W McMillan and David M Chavis. 1986. Sense of community: A definition and theory. *Journal of community psychology* 14, 1 (1986), 6–23.

[68] Justin Middleton, Emerson Murphy-Hill, and Kathryn T Stolee. 2020. Data Analysts and Their Software Practices: A Profile of the Sabermetrics Community and Beyond. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–27.

[69] Thomas Mock. 2018. The Mockup Blog: TidyTuesday. https://themockup.netlify.app/posts/2018-12-11-tidytuesday-a-weekly-social-data-project-in-r/

[70] Jesse Mostipak. [n.d.]. Join the "R for Data Science" online learning community. https://www.jessemaegan.com/post/join-the-r-for-data-science-online-learning-community/

[71] Gabriel Mugar, Carsten Østerlund, Katie DeVries Hassman, Kevin Crowston, and Corey Brian Jackson. 2014. Planet Hunters and Seafloor Explorers: Legitimate Peripheral Participation through Practice Proxies in Online Citizen Science. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '14).* Association for Computing Machinery, New York, NY, USA, 109–119. https://doi.org/10.1145/2531602.2531721

[72] Michael Muller, Ingrid Lange, Dakuo Wang, David Piorkowski, Jason Tsay, Q Vera Liao, Casey Dugan, and Thomas Erickson. 2019. How data science workers work with data: Discovery, capture, curation, design, creation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* 1–15.

[73] Laura A Pasquini and Paul William Eaton. 2019. The# acadv Community: Networked Practices, Professional Development, and Ongoing Knowledge Sharing in Advising. *NACADA Journal* 39, 1 (2019), 101–115.

[74] Samir Passi and Steven Jackson. 2017. Data Vision: Learning to See Through Algorithmic Abstraction. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17).* Association for Computing Machinery, New York, NY, USA, 2436–2447. https://doi.org/10.1145/2998181.2998331

[75] Samir Passi and Steven J. Jackson. 2018. Trust in Data Science: Collaboration, Translation, and Accountability in Corporate Data Science Projects. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 136 (Nov. 2018), 28 pages. https://doi.org/10.1145/3274405

[76] Liza Potts, Joyce Seitzinger, Dave Jones, and Angela Harrison. 2011. Tweeting disaster: hashtag constructions and collisions. In *Proceedings of the 29th ACM international conference on Design of communication.* 235–240.

[77] Gina Reynolds. [n.d.]. Tidy Tuesday Highlights. https://evamaerey.github.io/tidytuesday_walk_through/tidytuesday_highlights.html

[78] Joshua Rosenberg, Anthony Schmidt, Aaron Rosenberg, Jennifer Longnecker, and Michael Mann. 2020. Becoming 'Tidier' Over Time: Studying# tidytuesday as a Social Media-Based Context for Learning to Visualize Data. (2020).

[79] Adam Rule, Aurélien Tabard, and James D Hollan. 2018. Exploration and explanation in computational notebooks. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems.* 1–12.

[80] Joseph Seering, Robert Kraut, and Laura Dabbish. 2017. Shaping pro and anti-social behavior on twitch through moderation and example-setting. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing.* 111–125.

[81] Judith Segal. 2007. Some problems of professional end user developers. In *IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC 2007).* IEEE, 111–118.

[82] Igor Steinmacher, Tayana Conte, Marco Aurélio Gerosa, and David Redmiles. 2015. Social barriers faced by newcomers placing their first contribution in open source software projects. In *Proceedings of the 18th ACM conference on Computer supported cooperative work & social computing.* 1379–1392.

[83] Charles Sutton, Timothy Hobson, James Geddes, and Rich Caruana. 2018. Data diff: Interpretable, executable summaries of changes in distributions for data wrangling. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2279–2288.

[84] Yla Tausczik and Ping Wang. 2017. To Share, or Not to Share? Community-Level Collaboration in Open Innovation Contests. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–23.

[85] Rachel Treisman. 2017. Yale to offer new major in data science. https://yaledailynews.com/blog/2017/03/08/yale-to-offer-new-major-in-data-science/

[86] Alexa Vanhooser. 2018. UC Berkeley announces data science pipeline program for students. https://www.dailycal.org/2018/09/20/uc-berkeley-announces-data-science-pipeline-program-for-students/

[87] BN Vasilescu. 2014. Social aspects of collaboration in online software communities. (2014).

[88] Bogdan Vasilescu, Alexander Serebrenik, Prem Devanbu, and Vladimir Filkov. 2014. How Social Q&A Sites Are Changing Knowledge Sharing in Open Source Software Communities. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '14)*. Association for Computing Machinery, New York, NY, USA, 342–354. https://doi.org/10.1145/2531602.2531659

[89] April Yi Wang, Anant Mittal, Christopher Brooks, and Steve Oney. 2019. How data scientists use computational notebooks for real-time collaboration. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–30.

[90] Eric Wastl. [n.d.]. Advent of Code. https://adventofcode.com

[91] Etienne Wenger. 1999. *Communities of practice: Learning, meaning, and identity.* Cambridge university press.

[92] Etienne Wenger, Richard McDermott, and William M Snyder. 2002. Seven principles for cultivating communities of practice. *Cultivating Communities of Practice: a guide to managing knowledge* 4 (2002).

[93] Etienne Wenger, Richard Arnold McDermott, and William Snyder. 2002. *Cultivating communities of practice: A guide to managing knowledge.* Harvard Business Press.

[94] Hadley Wickham. 2017. *R for Data Science.* O'Reilly.

[95] Hadley Wickham et al. 2014. Tidy data. *Journal of Statistical Software* 59, 10 (2014), 1–23.

[96] Greg Wilson. 2006. Software carpentry: getting scientists to write better code by making them more productive. *Computing in Science & Engineering* 8, 6 (2006), 66–69.

[97] Anbang Xu and Brian Bailey. 2012. What Do You Think? A Case Study of Benefit, Expectation, and Interaction in a Large Online Critique Community. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW '12)*. Association for Computing Machinery, New York, NY, USA, 295–304. https://doi.org/10.1145/2145204.2145252

[98] Anbang Xu, Shih-Wen Huang, and Brian Bailey. 2014. Voyant: Generating Structured Feedback on Visual Designs Using a Crowd of Non-Experts. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '14)*. Association for Computing Machinery, New York, NY, USA, 1433–1444. https://doi.org/10.1145/2531602.2531604

[99] Qiushi Yan. [n.d.]. Tidy Data with Python. https://qiushi.rbind.io/post/python-tidy-data/

[100] Yunwen Ye and Kouichi Kishida. 2003. Toward an understanding of the motivation of open source software developers. In *25th International Conference on Software Engineering, 2003. Proceedings.* IEEE, 419–429.

[101] Alexey Zagalsky, Carlos Gómez Teshima, Daniel M. German, Margaret-Anne Storey, and Germán Poo-Caamaño. 2016. How the R Community Creates and Curates Knowledge: A Comparative Study of Stack Overflow and Mailing Lists. In *Proceedings of the 13th International Conference on Mining Software Repositories (MSR '16)*. Association for Computing Machinery, New York, NY, USA, 441–451. https://doi.org/10.1145/2901739.2901772

[102] Michele Zappavigna. 2011. Ambient affiliation: A linguistic perspective on Twitter. *New media & society* 13, 5 (2011), 788–806.

[103] Michele Zappavigna. 2012. *Discourse of Twitter and social media: How we use language to create affiliation on the web.* Vol. 6. A&C Black.

[104] Ellen Zegura, Carl DiSalvo, and Amanda Meng. 2018. Care and the practice of data science for social good. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*. 1–9.

[105] Amy X Zhang, Michael Muller, and Dakuo Wang. 2020. How do data science workers collaborate? roles, workflows, and tools. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–23.

[106] Haiyi Zhu, Robert E Kraut, and Aniket Kittur. 2016. A contingency view of transferring and adapting best practices within online communities. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 729–743.

Received June 2020; revised October 2020; accepted December 2020