Data Carpentry Assessment Report

Analysis of Post-Workshop Survey Results

Kari L. Jordan, PhD

October 2016

TABLE OF CONTENTS

Introduction

Data Carpentry is a fiscally sponsored project of NumFOCUS, a 501(c)3 non-profit that supports world-class, innovative, open source scientific computing projects. The mission of Data Carpentry is to provide researchers high-quality, domain-specific training covering the full lifecycle of data-driven research. This mission is achieved through the development and implementation of workshops taught on the fundamental data skills needed to conduct research.

Data Carpentry is a sibling organization of Software Carpentry, though its focus is on the introductory computational skills needed for data management and analysis in all domains of research. Data Carpentry lessons are domain specific, and range from life and physical sciences to social science. Lessons build upon the existing knowledge of learners to enable them to quickly apply the skills they learn to their own research.

Data Carpentry learners are asked to complete pre- and post-workshop surveys. This report is an analysis of the post-workshop survey responses collected for 40 Data Carpentry workshop sites.

Purpose

The analysis in this report serves the following purposes:
1. To inform the community of the impact Data Carpentry workshops have made on its learners.
2. To provide context for the survey responses as they relate to Data Carpentry learners.
3. To discuss what Data Carpentry is doing well, areas of improvement, and questions we should be asking.

Methodology

The data in this analysis was obtained via online surveys that include demographic questions, likert items, and open-ended responses. Learners were made aware of the purpose of the survey: *to understand the effectiveness of Data Carpentry workshops.*

Responses from the survey were recorded anonymously. There were no risks associated with learners participating in the survey, and they received no compensation. A link to the survey used in this analysis is in the appendix.

Limitations

This analysis has several limitations that may have potentially impacted the results presented here. These limitations are as follows:
- **Sample size -** There have been at least 90 Data Carpentry workshops hosted in various U.S. cities, the United Kingdom, Canada, Australia, Switzerland, and elsewhere. These workshops are designed to support, at most, 30 learners per site. Many of the workshops were filled to capacity, which means we should expect ~2,700 post-workshop survey responses. However, post-workshop survey responses are a major area of improvement. A total of 1,579 survey responses were collected; **492 were post-workshop responses.**
- **Response bias (i.e. self-reported data)** - As the data collected in this analysis is self-reported, it cannot be independently verified. Issues related to learner's selective memory, attribution, and exaggeration may potentially impact the results presented.
- **Cultural differences** - As Data Carpentry workshops are taught globally, cultural differences including attitudes toward completing surveys in general, and attitudes toward specific survey items in particular (ex. questions regarding race/ethnicity) may potentially impact the analysis presented here.

The limitations presented above are characteristic of survey research methodology. Though limitations exist, the results presented here are thought to be generalizable.

Learner Demographics
Data Carpentry post-workshop survey responses were collected using SurveyMonkey. A total of 1,579 survey responses were collected (1,087 pre-workshop responses, 492 post-workshop responses) representing 53 workshop sites.

Of the 53 workshop sites, 40 sites constitute the makeup of post-workshop survey respondents. Of the Data Carpentry learners who completed the post-workshop survey, 49% completed the survey at the event, while 51% completed the survey after leaving the event. Ninety-four percent of the learners were first-time Data Carpentry learners.

Sixty-five percent of Data Carpentry learners attended a workshop in the United States, while 34.6% attended a workshop outside of the United States. Demographic breakdowns of learners by status and race/ethnicity are provided in Tables 1 and 2. Results of learner responses in the "other" category are provided in the text analysis in Figure 1. There were no significant differences in survey responses considering race/ethnicity as a factor.

Table 1: Breakdown of Post-Workshop Survey Respondents by Status

	Frequency	**Percent**	**Cumulative Percent**
Other	51	11.6	11.6
Undergraduate student	18	4.1	15.7
Graduate student	160	36.4	52.1
Post-doc	53	12.0	64.1
Faculty	49	11.1	75.2
Staff	109	24.8	100
Total	**440**		

Figure 1: Text analysis of post-workshop survey respondents *(Status: Other)*

Table 2: Breakdown of Post-Workshop Survey Respondents by Race/Ethnicity

	Frequency	Percent	Cumulative Percent
American Indian/Alaska Native	2	0.8	0.8
Asian/Pacific Islander	44	18.6	19.4
Black or African American	11	4.6	24.1
Hispanic or Latino	12	5.1	29.1
White Caucasian	166	70.0	99.2
Prefer not to say	2	0.8	100.0
Total*	**237**		

*This information was only collected for U.S. survey respondents

Learners represented varying domains of research with ecology, biology (micro, molecular, etc.), genetics, social science, and biochemistry representing the majority (Figure 2).

Figure 2: Text analysis of learner's self-reported domain of research

Analysis of Data Carpentry Workshop Impact
This analysis of learner's self-reported data includes post-workshop attitudes, data management and analysis skills prior to and following Data Carpentry workshops, research computing efficacy, research computing attitudes, feedback on instructors, and open-ended responses. A discussion of these results is included within the analysis.

Post-Workshop Attitudes
Learner attitudes toward Data Carpentry workshops have been favorable. The majority of learners agreed or strongly agreed that they would recommend the workshop to a friend or colleague (Figure 3). Learners felt the material taught in the workshop matched the workshop description (Figure 4). For reference, here is an overview of a Data Carpentry workshop description:

> Data Carpentry's aim is to teach researchers basic concepts, skills, and tools for working with data so that they can get more done in less time, and with less pain. This workshop uses a tabular ecology dataset and teaches data cleaning, management, analysis and visualization. There are no prerequisites, and the materials assume no prior knowledge about the tools. The workshop uses a

4

single tabular data set that contains observations about adorable small mammals over a long period of time in Arizona. See data.md for more information about this data set, including the dowload location. The workshop can be taught using R or python as the base language. Overview of the lessons:

1. Data organization in spreadsheets and data cleaning with OpenRefine
2. Introduction to R or python
3. Data analysis and visualization in R or python
4. SQL for data management

The majority of learners either agreed or strongly agreed that the workshop was worth their time (Figure 5). Lastly, learners felt (agreed or strongly agreed) they could immediately apply what they learned at the workshop (Figure 6).

When asked to self-report their level of involvement in the workshop, learners reported being either very involved or enthusiastically involved in the workshop (Figure 7). Additionally, learners gained either some or a great deal of practical knowledge at the workshop (Figure 8).

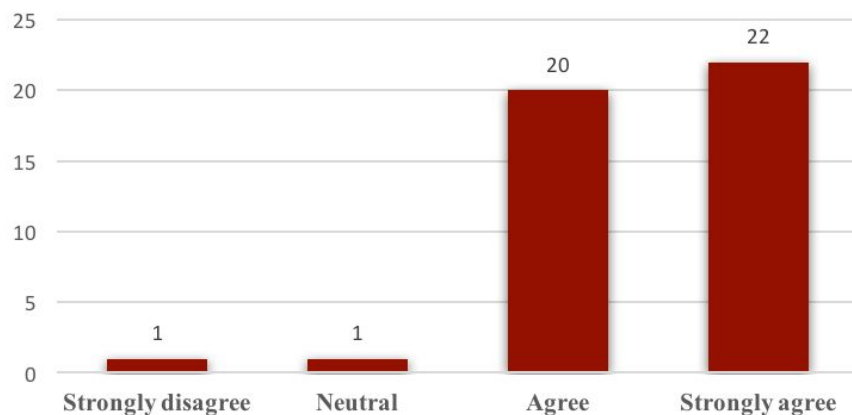**Figure 3:** I would recommend this workshop to a colleague
(n = 44)

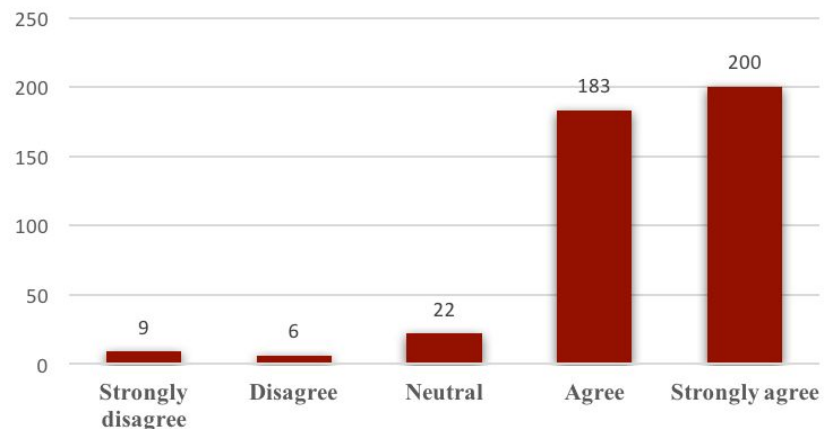**Figure 4:** The material presented matched the workshop description
(n = 420)

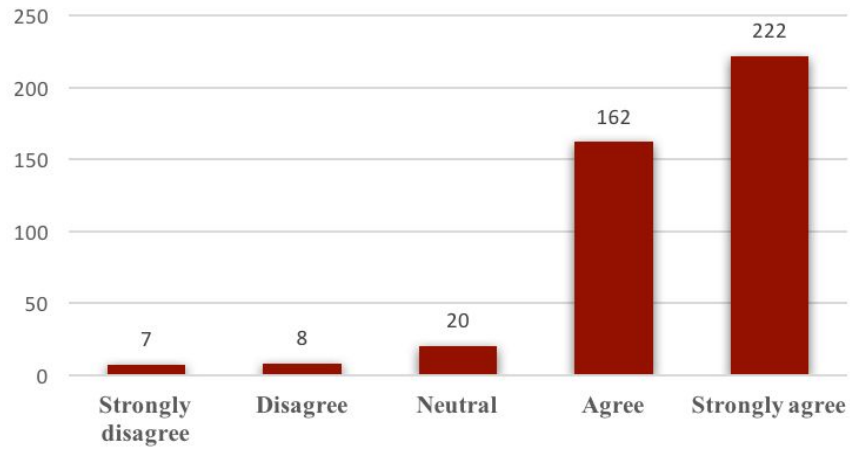**Figure 5:** The workshop was worth my time
(n = 419)

| Strongly disagree | Disagree | Neutral | Agree | Strongly agree |
| 7 | 8 | 20 | 162 | 222 |

Figure 6: I can immediately apply what I learned at the workshop.
(n = 421)

| Strongly disagree | Disagree | Neutral | Agree | Strongly agree |
| 4 | 23 | 80 | 212 | 102 |
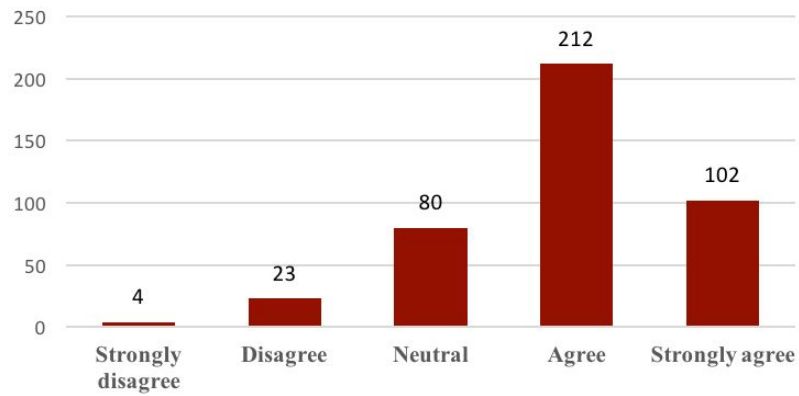
Figure 7: Rate your level of involvement in the activities of this workshop..
(n = 478)

| Somewhat involved | Very involved | Enthusiastically involved |
| 66 | 225 | 151 |

Figure 8: How much practical knowledge have you gained
from this workshop?
(n = 478)
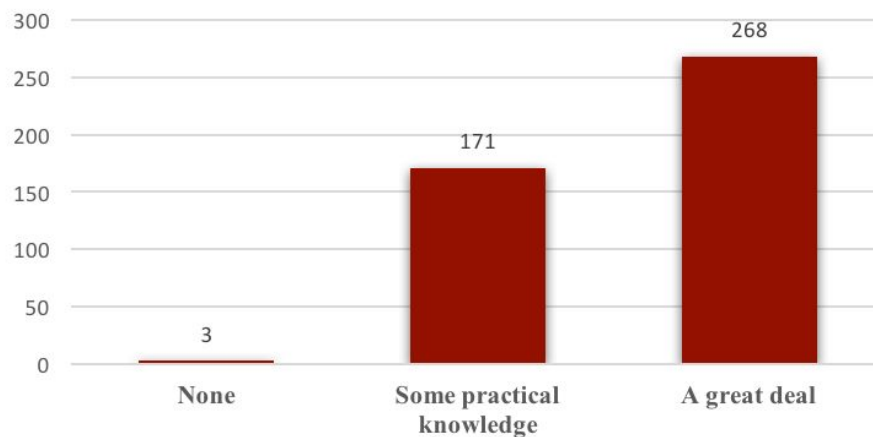
Figures 3-8 tell us that learners are actively participating in Data Carpentry workshops and walking away with practical knowledge they can immediately apply to their research. As a reminder, 94% of learners were attending a Data Carpentry workshop for the first time, so the impact of Data Carpentry workshops has been significant. Figures 3 and 5 are evidence of that, as the majority of Data Carpentry learners would recommend a workshop to a friend or colleague, and feel the workshop was worth their time. Additionally, Data Carpentry is doing well presenting material that matches each workshop description.

Data Management and Analysis Skills
Learners were asked to self-report their level of data management and analysis skills prior to and following the workshop they attended, ranging from very low to very high prior to the workshop, and about the same to much higher following the workshop. Figures 9 and 10 are a display of these responses. Though the majority of Data Carpentry learners were attending a workshop for the first time, we see varying levels of data management and analysis skills prior to the workshop, with nearly 37% responding they had neither high nor low data management and analysis skills prior to the workshop.

Figure 9: Level of data management and analysis skills prior to
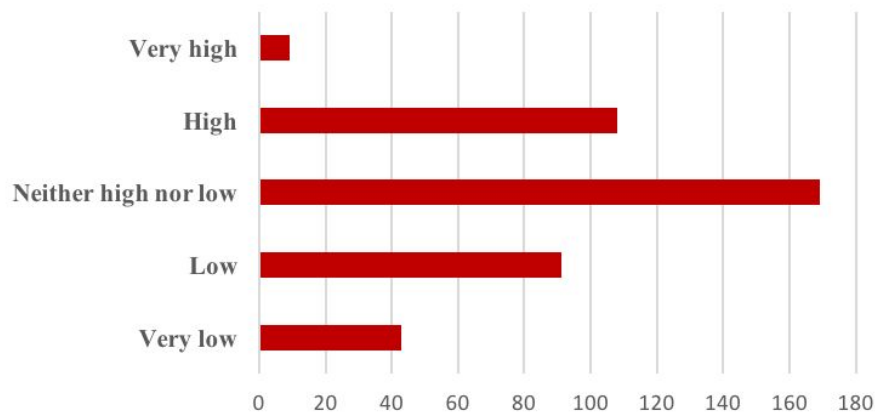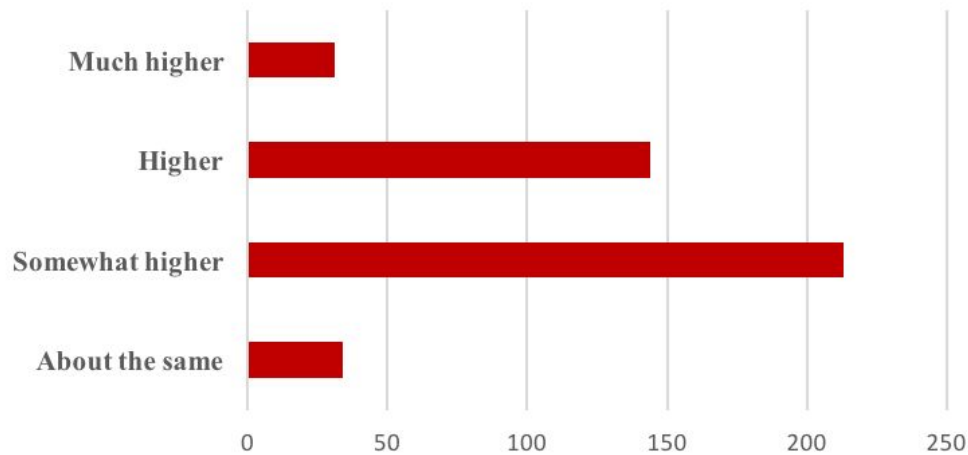the workshop

Figure 10: Level of data management and analysis skills following to the workshop

A comparison of learner data management and analysis skill levels prior to and following the workshop are presented in Tables 3-7. Comparing learners prior to and following Data Carpentry workshops we see that learners having very low, low, and neither high nor low data management and analysis skills prior to the workshop are leaving the workshop with somewhat higher and higher levels of data management and analysis skills (Tables 3-5). Learners having high levels of data management and analysis skills prior to the workshop are leaving with skills somewhat and much higher (Table 6).

Table 3: Comparison of learner levels of data management and analysis skills prior to and following the workshop - *Very Low*

Level of data management and analysis skills prior to the workshop	Level of data management and analysis skills following the workshop	n	%	Cumulative
Very Low	About the same	1	2.33	2.33
Very Low	Somewhat higher	25	58.14	60.47
Very Low	Higher	16	37.21	97.67
Very Low	Much higher	1	2.33	100.00
TOTAL		**43**		

8

Table 4: Comparison of learner levels of data management and analysis skills prior to and following the workshop - *Low*

Level of data management and analysis skills prior to the workshop	Level of data management and analysis skills following the workshop	n	%	Cumulative
Low	About the same	6	6.59	6.59
Low	Somewhat higher	44	48.35	54.95
Low	Higher	37	40.66	95.60
Low	Much higher	4	4.40	100.00
TOTAL		**91**		

Table 5: Comparison of learner levels of data management and analysis skills prior to and following the workshop - *Neither high nor low*

Level of data management and analysis skills prior to the workshop	Level of data management and analysis skills following the workshop	n	%	Cumulative
Neither high nor low	About the same	12	7.10	7.10
Neither high nor low	Somewhat higher	87	51.48	58.58
Neither high nor low	Higher	61	36.09	94.67
Neither high nor low	Much higher	9	5.33	100.00
TOTAL		**169**		

Table 6: Comparison of learner levels of data management and analysis skills prior to and following the workshop - *High*

Level of data management and analysis skills prior to the workshop	Level of data management and analysis skills following the workshop	n	%	Cumulative
High	About the same	14	12.96	12.96
High	Somewhat higher	51	47.22	60.19
High	Higher	29	26.85	87.04
High	Much higher	14	12.96	100.00
TOTAL		**108**		

Table 7: Comparison of learner levels of data management and analysis skills prior to and following the workshop - *Very high*

Level of data management and analysis skills prior to the workshop	Level of data management and analysis skills following the workshop	n	%	Cumulative
Very high	About the same	1	11.11	11.11
Very high	Somewhat higher	5	55.56	66.67
Very high	Higher	1	11.11	77.78
Very high	Much higher	2	22.22	100.00
TOTAL		**9**		

Research Computing Efficacy

A primary goal of Data Carpentry is to increase confidence in a learner's research computing efficacy. Learners were asked to indicate their level of agreement with having a better understanding of how to do the following (compared to before the workshop):
- Effectively organize data in spreadsheets
- Use OpenRefine for data cleaning
- Import a file into Python or R and work with the data
- Do initial visualizations in Python or R
- Construct a SQL query statement

Table 8 shows the mean and mode (value occurring most often) of the responses recorded on a scale of 1 (strongly disagree) to 5 (strongly agree). Mode gives an overall characterization of the distribution. Table 8 tells us Data Carpentry learners are leaving workshops having a better understanding of how to do introductory computational tasks. **Note:** visualizations in Python had a high mean, but lowest mode.

Table 8: Compared to before the workshop I have a better understanding of how to…

Item	n	Mean	Mode
Effectively organize data in spreadsheets	381	4.18	4
Use OpenRefine for data cleaning	311	4.61	5
Import a file into Python and work with the data	143	4.92	4
Import a file into R and work with the data	374	4.49	5
Do initial visualizations in Python	**145**	**4.90**	**3**
Do initial visualizations in R	369	4.47	5
Construct a SQL query statement	273	4.61	5
Use the command line	326	4.47	5

Research Computing Attitudes

A secondary goal of Data Carpentry is related to research computing attitudes. Data Carpentry seeks to shift the perspective of how learners value their skills (i.e. scripting) to improve and promote reproducible research. As such, learners were asked to indicate their level of agreement with the following statements:

- Data organization is a fundamental component of effective and reproducible research.
- Using a scripting language like R or Python can ultimately improve my analysis efficiency.
- Using R or Python makes analyses easier to reproduce.
- A value of using SQL, R or Python is that underlying data cannot accidentally be changed.

Responses were recorded on a scale of 1 (strongly disagree) to 5 (strongly agree). Table 9 presents the mean and mode of the **pre-workshop** responses. Table 10 presents the mean and mode of the **post-workshop** responses. Again, mode is being presented to give an overall characterization of the distribution. Both tables are presented to better understand the impact Data Carpentry workshops have had on learner research computing attitudes.

Table 9: Analysis of Learners' Research Computing Attitudes *(Pre-Workshop)*

Item	n	Mean	Mode
Data organization is a fundamental component of effective and reproducible research.	923	4.47	4
Using a scripting language like R or Python can ultimately improve my analysis efficiency.	920	4.19	4
Using R or Python makes analyses easier to reproduce.	918	3.95	3
A value of using SQL, R or Python is that underlying data cannot accidentally be changed.	912	3.50	3

Table 10: Analysis of Learners' Research Computing Attitudes *(Post-Workshop)*

Item	n	Mean	Mode
Data organization is a fundamental component of effective and reproducible research.	422	4.64	5
Using a scripting language like R or Python can ultimately improve my analysis efficiency.	422	4.43	5
Using R or Python makes analyses easier to reproduce.	420	4.39	5
A value of using SQL, R or Python is that underlying data cannot accidentally be changed.	419	4.06	5

As survey responses were not paired, we cannot assume there are definite relationships between the pre- and post-workshop survey responses; these responses are independent. An independent samples t-test was used to compare means assuming unequal variances, however the results of the analysis produced p-values that were not meaningful. What we do know, however, is that the **mode** for research computing

attitudes of learners who completed the post-workshop survey was high. Ideally, there was a shift in their attitudes compared to before the workshop, as the **mode** for the items under research computing attitudes was lower in the pre-workshop survey compared to that of the post-workshop survey.

Learner Attitudes toward Instructors
The post-workshop survey included questions to assess learner attitudes toward their instructors. These questions were as follows:
1. Could you get clear answers to your questions from the instructors?
2. Were the instructors considerate to you?
3. Were the instructors effective in teaching in the workshop?
4. Were the instructors enthusiastic about the workshop?

Learners responded to these questions selecting either **never, sometimes, usually, most of the time, or always.** Figure 11 provides an analysis of their responses. Of note is learners' response to whether instructors provided clear answers and were considerate. The majority of respondents felt instructors always provided clear answers and were always considerate. Additionally, learners felt their instructors were enthusiastic. Overall, Data Carpentry instructors are leaving a positive impression on learners.

Figure 11: Learner Attitudes Toward Instructors
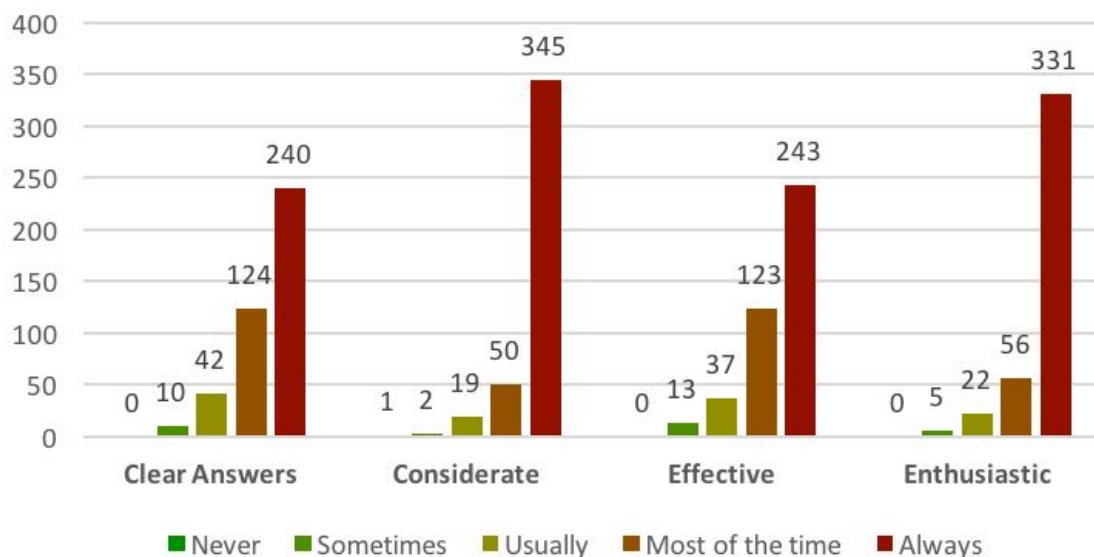
Open-Ended Responses - Workshop Strengths
Learners were asked to open-endedly convey the major strengths of the workshop they attended. A text analysis of the open-ended responses is presented in Figure 12. The open-ended responses tell us that Data Carpentry workshop strengths are directly related to workshop content and instructors, as evidenced by the following open-ended responses:

> "Very practical examples and a wide array of examples are covered to at least set the foundation to use the program immediately and build upon skills outside the workshop. The instructors are very knowledgeable and currently use practices for their work which provides relevance."

"The instructors and helpers are very knowledgeable and the learning material was brought across very clearly. As a whole, the workshop taught me valuable tools which I will indeed apply in current and future research."

"It introduced basic knowledge of data management using robust software and platform. The instructors were well prepared, knowledgeable, very helpful and created an interactive environment to make learning of the skills easier. Overall, they were very enthusiastic in what they were doing."

Figure 12: What are the major strengths of this workshop?

Open-Ended Responses - Workshop Weaknesses
Learners were also asked to open-endedly convey the major weaknesses of the workshop they attended. A text analysis of the open-ended responses is presented in Figure 13. Major concerns were related to time and pace, including the time spent practicing examples, and the time spent on each topic (i.e. SQL vs. R), as evidenced by the following open-ended responses:

"I do think we spent a bit too long on changing directories and listing files. I would have liked to spend more time on loops and making piped commands, since I see that as the best use of command line in it's ability to do multi-step file manipulations. I would have liked some more."

"The workshop can include more "real" genomics data analysis - more advanced [than] what is done in the current curriculum."

"The workshop schedule was somewhat packed."

"Need a bit more hands-on time with own data to apply to newly acquired skills."

Figure 13: What are the major weaknesses of this workshop?

Open-Ended Responses - Comments about Instructors

Learners were given the opportunity to provide additional feedback about their instructors. A text analysis of the responses is presented in Figure 14. The feedback received reiterates one of the major strengths of Data Carpentry workshops: **our instructor**. This is evidenced by the following responses:

"Great Job!!! The enthusiasm for the software and tools really resonates."

"The instructors are highly motivated and clearly understand how difficult the first steps into the world of computer code are."

"They were very helpful, passionate about the topic, and wanted others to learn"

Figure 14: Text analysis of learner comments about instructors

Summary

Data Carpentry workshops have a significant impact on learners. Learners have expressed satisfaction with workshop content and appreciation for the caliber of their instructors. Learners have high means for research computing efficacy, and as Data Carpentry continues to offer more workshops, a shift in the perspective of how researchers view and use computational skills is sure to be realized. Data Carpentry will continue to improve and expand our community of data-driven researchers.

Moving forward, Data Carpentry should develop opportunities for learners to recommend workshops to friends and colleagues. This could be as simple as linking a calendar of upcoming workshops to the post-workshop survey completion page.

Additionally, Data Carpentry should provide a means for learners to immediately apply what they learned in the workshop they attended. Take-home tasks whereby learners use their own data are recommended.

Also, offering periodic skills workshops to learners who attended a Data Carpentry workshop will help learners retain their newfound knowledge. These workshops can be offered in-person or virtually.

To improve our assessment efforts, Data Carpentry should improve its ability to assess paired data. This can be accomplished by having learners create their own unique identifier to include when completing both the pre- and post-workshop survey. Additionally, communicating the importance of these surveys to both the learners, instructors, and helpers could potentially help improve the survey response rate.

Additionally, to assess changes in learner data management and analysis skills and research computing efficacy, the survey questions in both the pre-and post-survey must use identical likert scales.

The link to the post-workshop survey used in this analysis is in the appendix.

Appendix: Data Carpentry Post-Workshop Survey