Committee Meeting #1

Data Science for the Biomedical Sciences

Daniel Chen

Virginia Tech

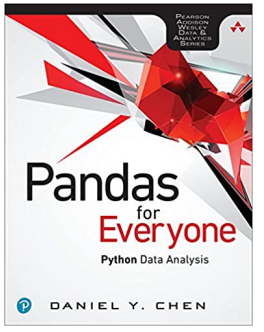
2020/11/17 (updated: 2020-11-17)

Me

- Ph.D, Virginia Tech, 2021, GBCB
- MPH, Columbia, 2014, Epidemiology

- The Carpentries
  - Instructor, 2014
  - Maintainer, 2016 - 2018
  - Instructor Trainer, 2020
  - Maintainer Lead, 2020

- Pearson
  - Book
  - Python + Git classes



- RStudio Intern 2019
  - RStudio Education team
  - `gradethis` package

The committee

Anne Brown

Library

Committee Chair

BBL + DB

Dave Higdon

Statistics

Department Head

Alex Hanlon

Statistics

CBHDS

iTHRIV BERD

Nikki Lewis

Honors College

Computational Research Grant

Background

Data Science Education in the Biomedical Sciences

- "Integrating scientific programming in communities of practice for students in life science"
  - https://dl.acm.org/doi/10.1145/3332186.3333040
- People in the life sciences are afraid to take courses in CS and Stats

Goal

- Create an inviting learning environment
- Start from scratch
  - Positive comment from a learner in post-workshop survey

Dreyfus model of skill acquisition

- Novice
- Competent
- Expert

This isn't new...

- Creating learner personas to focus materials and learning objectives
- Motivating examples with formative assessments
- Surveying learners
- Curriculum is full
  - Figuring out what to take out
    - Plotting and logistic regression

Already existing learning materials

But...

- General purpose, not domain focused
  - Plays a big role in motivation and building mental models

Build a community

At VT

- Help augment the training from Center for Biostatistics and Health Data Science (CBHDS)
  - https://biostat.centers.vt.edu/
- Goal is not to teach statistics but the data literacy side of data science
  - Managing and "cleaning" data
  - Be able to better communicate with analysts and statisticians
- Symbiotic relationship
  - Data Services and Data Bridge at the Library
  - The VT Carpentries community
  - CBHDS

In general

- R/Medicine community
- Biomedical informatics

Sustanining a commmunity of practice

- Open educational materials
- Resources and people to help
- Influx and efflux of people

In-person and online

- The Carpentries Resources for Online Workshops
  - Workshop Logistics and Screen Layouts
- Using OBS for Online Teaching

Overall Dissertation Overview

Phases + research questions

IRB

- **IRB 20-537**: Data Science Workshops for Biomedical and Health Professionals: Persona Identification and Workshop Assessment

3 Phases (i.e., 3 papers + chapters)

Survey based (Adapted from The Carpentries):

1. Pre-workshop student **self-assessment** survey
    - create learner personas
2. **Pre/post workshop** survey
    - assess the workshop materials
    - assess learning objectives
3. **Long-term workshop** survey (6 months out)
    - see if the materials helped with fundamental knowledge to learn more on their own
    - longitudinal study

(Community) Deliverables

In addition to the 3 papers + chapters:

1. Identify and create learner personas for the biomedical community
2. Create a set of CC-0 lessons for the biomedical community
    - Carpentries-inspired
    - Carpenties Incubator: Data Science for Practicing Clinicians
        - https://carpentries-incubator.github.io/Data-Science-for-Docs/
        - Too much emphasis on Medical Doctors

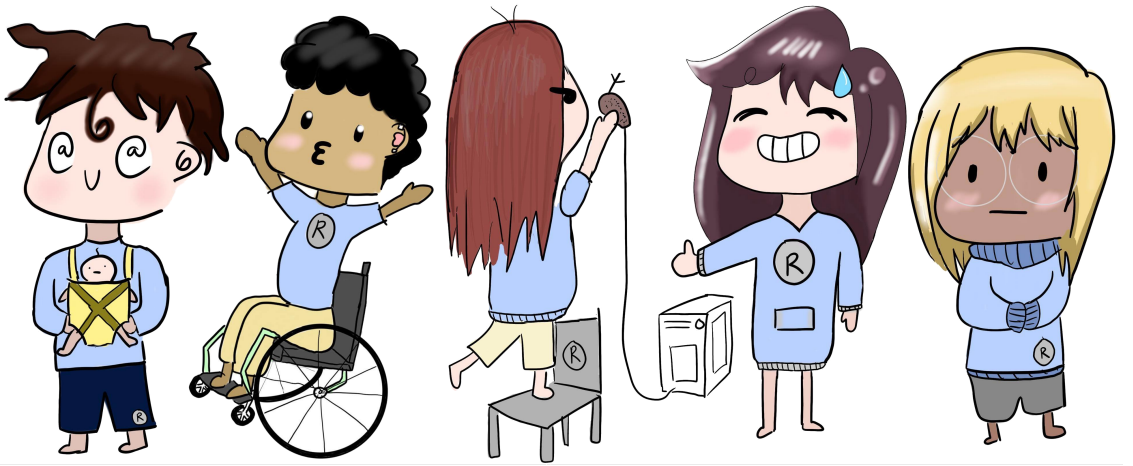
Current progress

1. IRB approved study: IRB 20-537 (Summer 2020)
2. Conducted the pre-workshop learner self-assessment (Summer 2020)
3. Created 4 learner personas (Fall 2020)
4. Created workshop content and learning materials (Fall 2020)
5. Conducted first iteration of workshops (October 20 and 22, 2020)
6. 2 x 1-day 4-hour workshops scheduled in December (8 + 9, 2020)

Planning for

1. More workshops in 2021
    - On the order of once per month
    - Current + new materials
2. Prelims in Feburary 2021

D1. Identify and create learner personas for the biomedical community



4 Personas:

- Alex Academic
- Clare Clinician
- Patricia Programmer
- Samir Student

Persona contains:

1. Background
2. Relevant prior knowledge or experience
3. Perception of needs
4. Special considerations

Creating Personas

- Personas of teachers in a classroom

- Paper: "Through the Eyes of Faculty: Using Personas as a Tool for Learner-Centered Professional Development"
- Methods that combine hierarchical agglomerative cluster analysis with chi-square values or squared Euclidian distance values and complete or average linkage

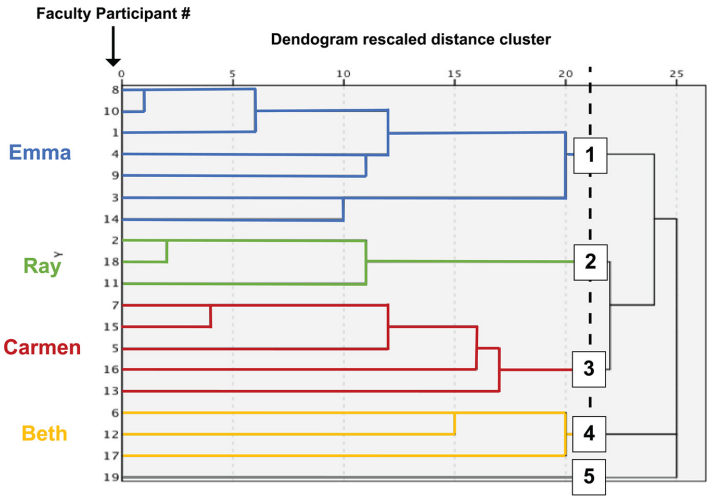


Figure 1: Dendrogram used for persona creation.

https://www.lifescied.org/doi/10.1187/cbe.19-06-0114

Example personas

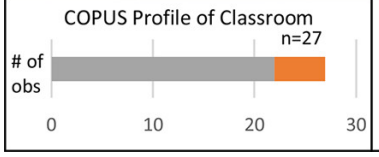
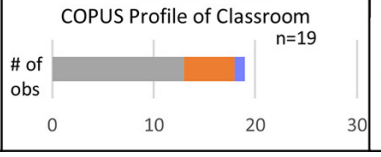
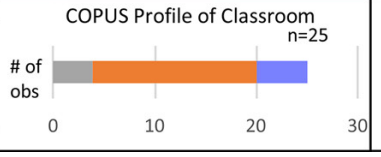
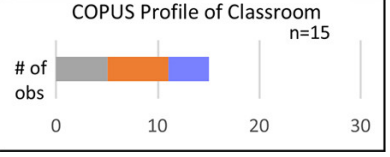
Emma The Expert	**Ray** The Relater	**Carmen** The Coach	**Beth** The Burdened
"I just think that I've put an enormous amount of work into getting good at the way I lecture."	*"They like being active but they don't like you not lecturing to them. That's the problem, right?"*	*"I try to figure out, well, what can I do in class that would help them be able to succeed when they take the exam?"*	*"I tried different ways, you can try activities, problem sets, clicker questions, all kinds of thinking [...] and they still, year after year make the same kinds of mistakes."*
Knowledge of Students • Expects students to do own learning • Knows about student struggles & tendencies & views as deficits	**Knowledge of Students** • Wants to relate to students & bridge gaps • Understands student struggles & perspectives	**Knowledge of Students** • Wants to understand student thinking • Frustrated when students do not engage in class because she believes they learn by doing	**Knowledge of Students** • Wants to see student thinking • Feels defeated when students do not engage in learning opportunities she creates
Teaching Values • Get students enthusiastic about material • Prepare them for upper-level courses • Test synthesis & application on exams	**Teaching Values** • Connect with students • Capture their attention through stories • Invest in their professional development	**Teaching Values** • Engage in problem solving & application • Engage in scientific thinking practices in class	**Teaching Values** • Engage in problem solving & application • Take on responsibility for promoting student engagement • Implement peer interaction
Approaches to Innovations • Feels her expertise & comfort do not lie in active learning pedagogies • Is critical of findings from education literature • Likes learning objectives as organizers for lectures and for providing students some initial scaffold	**Approaches to Innovations** • Uses assessment to inform teaching • Considers how assessments help students • Likes how active learning promotes student engagement • Experiments to find best balance of lecture, storytelling, & activities	**Approaches to Innovations** • Uses assessment to inform teaching • Likes targeted instruction & backwards design • Enjoys implementing active learning	**Approaches to Innovations** • Uses assessment to inform teaching, but feels already knows the misconceptions • Considers how assessments can help inform students • Unsure what to do if students still have struggles after active learning • Enjoys implementing active learning
Perceived Barriers • Feels academic culture prevents a focus on teaching • Feels need to sort/rank student performance for post-college careers	**Perceived Barriers** • Envisioning how to scale up in-class interactions for large classes	**Perceived Barriers** • Trying to fight against the perpetual barriers within academic culture	**Perceived Barriers** • Securing limited resources, such as TAs, to scale up group work in large classes
			

Figure 2: Summaries of personas.

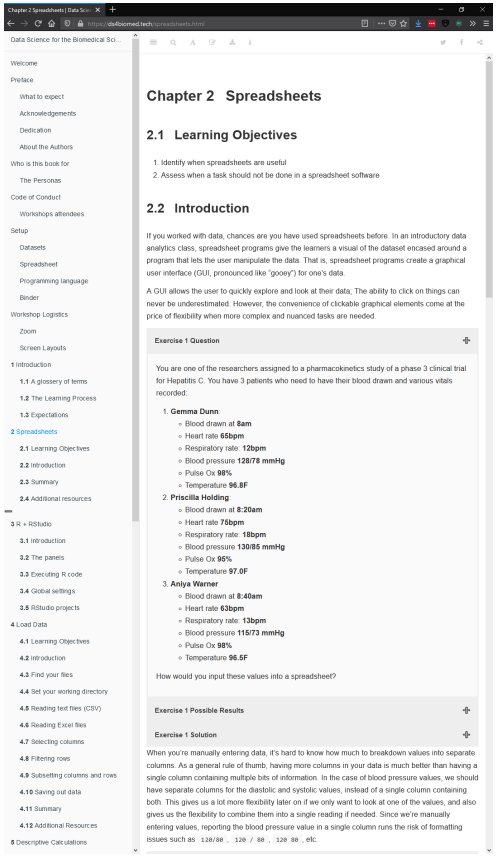
RStudio learner Personas

Extended the RStudio personas to also include Data knowledge

Persona	In Brief	Domain Knowledge	Statistics Knowledge	Programming Knowledge
Anya Academic	A professor who needs training for her research and to pass on to students.	expert	competent	competent
Celine Certified	A certified RStudio instructor.	competent	competent	competent
Exton Excel	A proficient Excel user working in industry who wants to switch to R.	competent	novice	novice
Jacqui Ofalltrades	A data science generalist at a small consulting company.	expert	expert	expert
Katrin Keener	An R enthusiast.	competent	competent	competent
Larry Legacy	A reluctant learner who would really rather just keep using the tools he knows.	expert	expert	novice
M'shelle Manager	An ex-programmer who now leads a team and needs to make decisions about tool adoption and training.	competent	novice	competent
Nang Newbie	An undergraduate student without statistical knowledge, programming skills, and real-world experience.	novice	novice	novice
Toshi Techsupport	A sys admin who has to support data scientists.	expert	novice	expert

https://rstudio-education.github.io/learner-personas/

D2. Create a set of CC-0 lessons for the biomedical community



- Book + workshop materials: https://ds4biomed.tech/
- Slides + papers: bit.ly/ds4biomed-gdrive

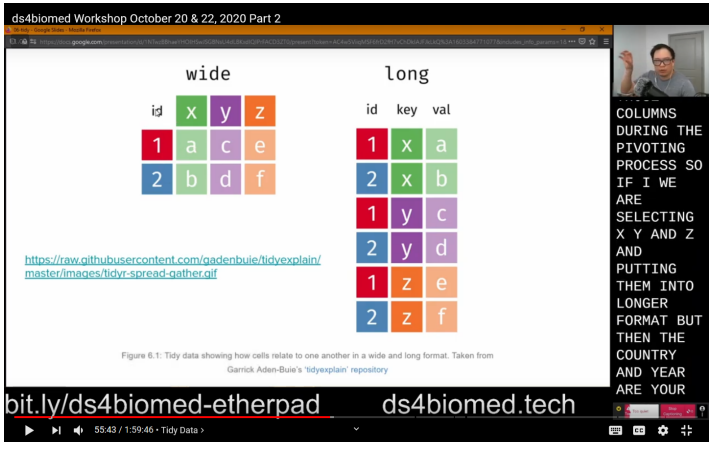
Learning Objectives

Using Bloom's taxonomy:

1. Name the features of a tidy/clean dataset
2. Transform data for analysis
3. Identify when spreadsheets are useful
4. Assess when a task should not be done in a spreadsheet software
5. Break down data processing into smaller individual (and more manageable) steps
6. Construct a plot and table for exploratory data analysis
7. Build a data processing pipeline that can be used in multiple programs
8. Calculate, interpret, and communicate an appropriate statistical analysis of the data

Workshops (i.e., Dissertation Data)

- October 20 + 22, 2020: Pilot workshop over VT listservs
  - Day 1: 20 participants
  - Day 2: 11 participants
- December 8, 2020: Roanoke
- December 9, 2020: Business for Healthcare (Lynchburg)



1. Online + in-person learning
   - Online Workshop Logistics and Screen Layouts
2. Live captions
   - Using OBS for Online Teaching

YouTube Playlist: https://www.youtube.com/watch?v=nQ4lbmKD1no&list=PL4eF1KHNgDfK9rDZLrKc5K2wtQqtJEAnS

Phase 1 / Chapter 1: Persona creation

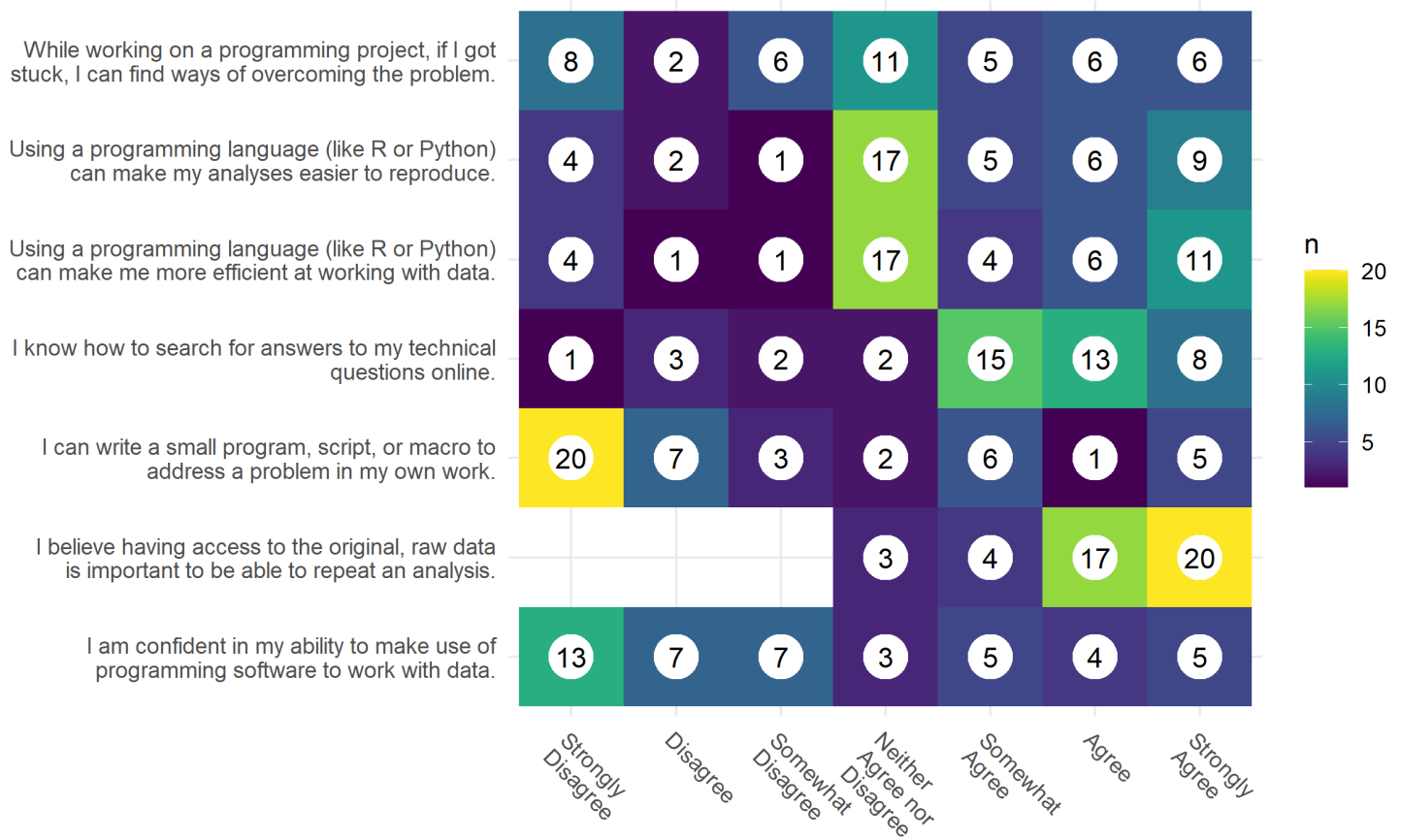
Listservs

1. iTHRIV (Taryn Luoma, MHA)
2. GBCB
3. IGEP (Dennie Munson)
4. VetMed (Andrea Green)
5. MPH (Hannah Menefee, MPH)
6. VT Carpentries workshops (Nathaniel Porter, PhD)
7. VT Roanoke Center (David Conners)

No explicit email sent to the for VCOM listserve

Sample population does **not** contain 8 iTHRIV Scholars

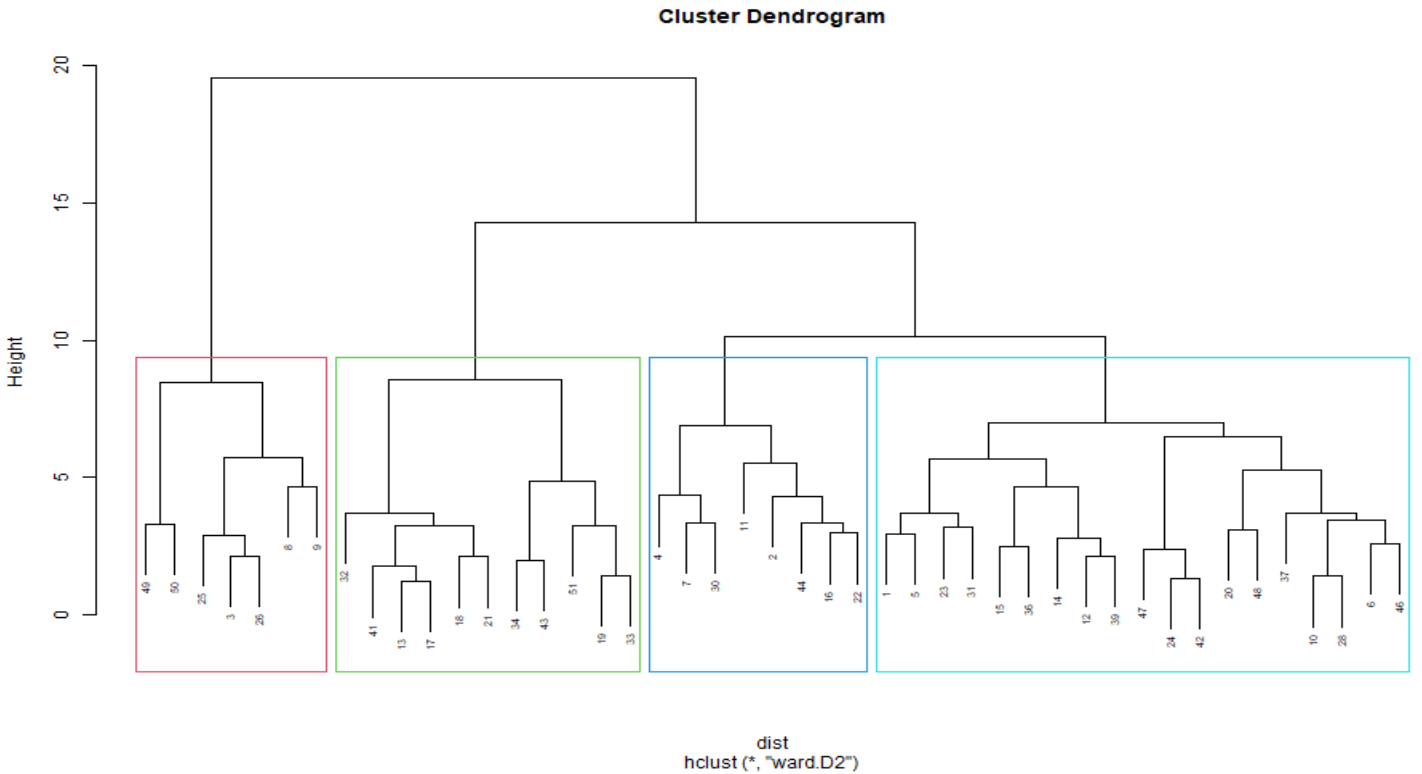
Overall results



Raw data is important, but do not recognize why programming is important.

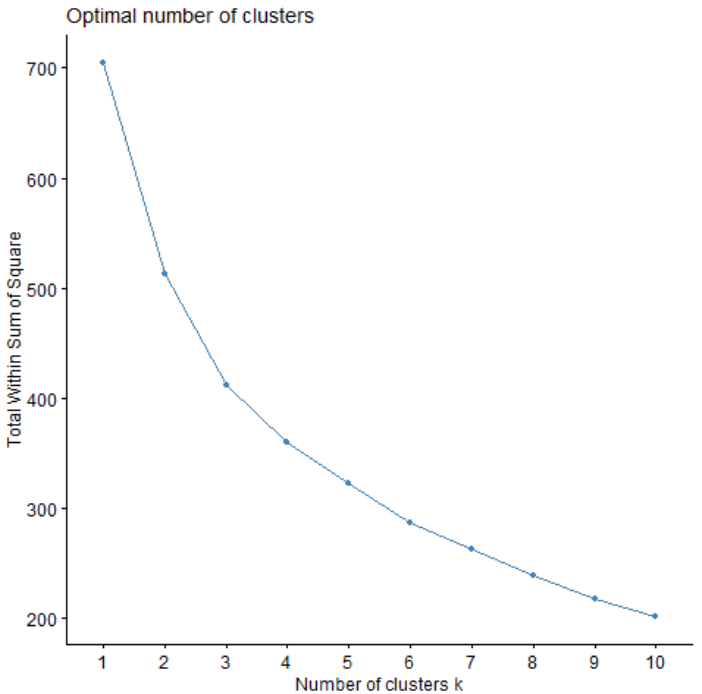
Hierarchical clustering

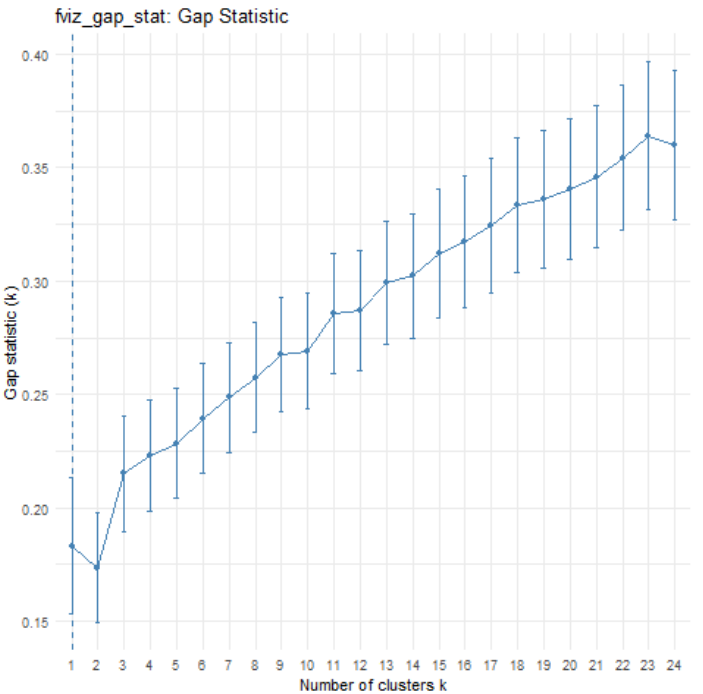
- Centered + Scaled ordinal survey questions
- Euclidean distance
- Ward's Clustering



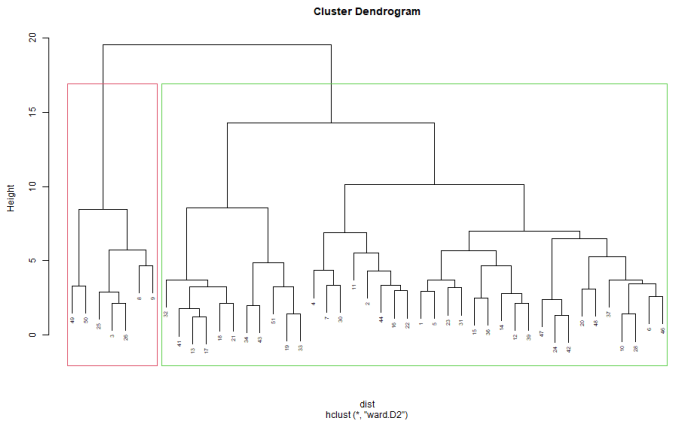
Optimal Clusters

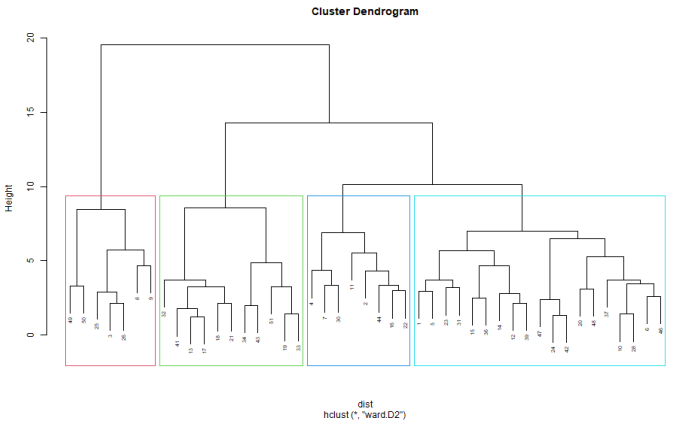
The elbow plot and gap statistic suggests the optimal number of clusters is from 2-4.

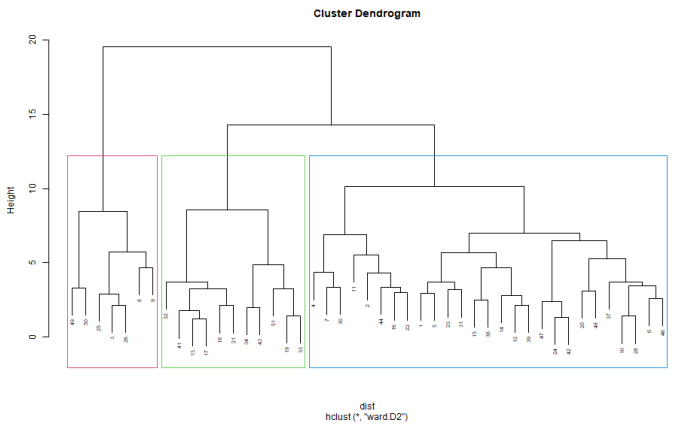




4 Clusters (i.e., personas) for interprebility

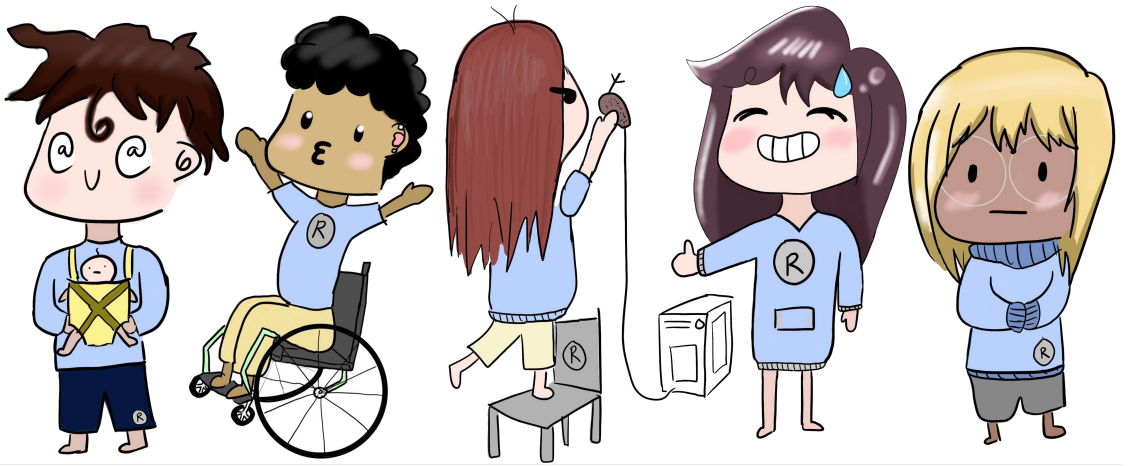






- 2: Experts, Non-Experts
- 3: Experts, Clinicians, Academics
- 4: Experts, Clinicians, Academics (Students, Academics + researchers)

4 Learner Personas



4 Personas:

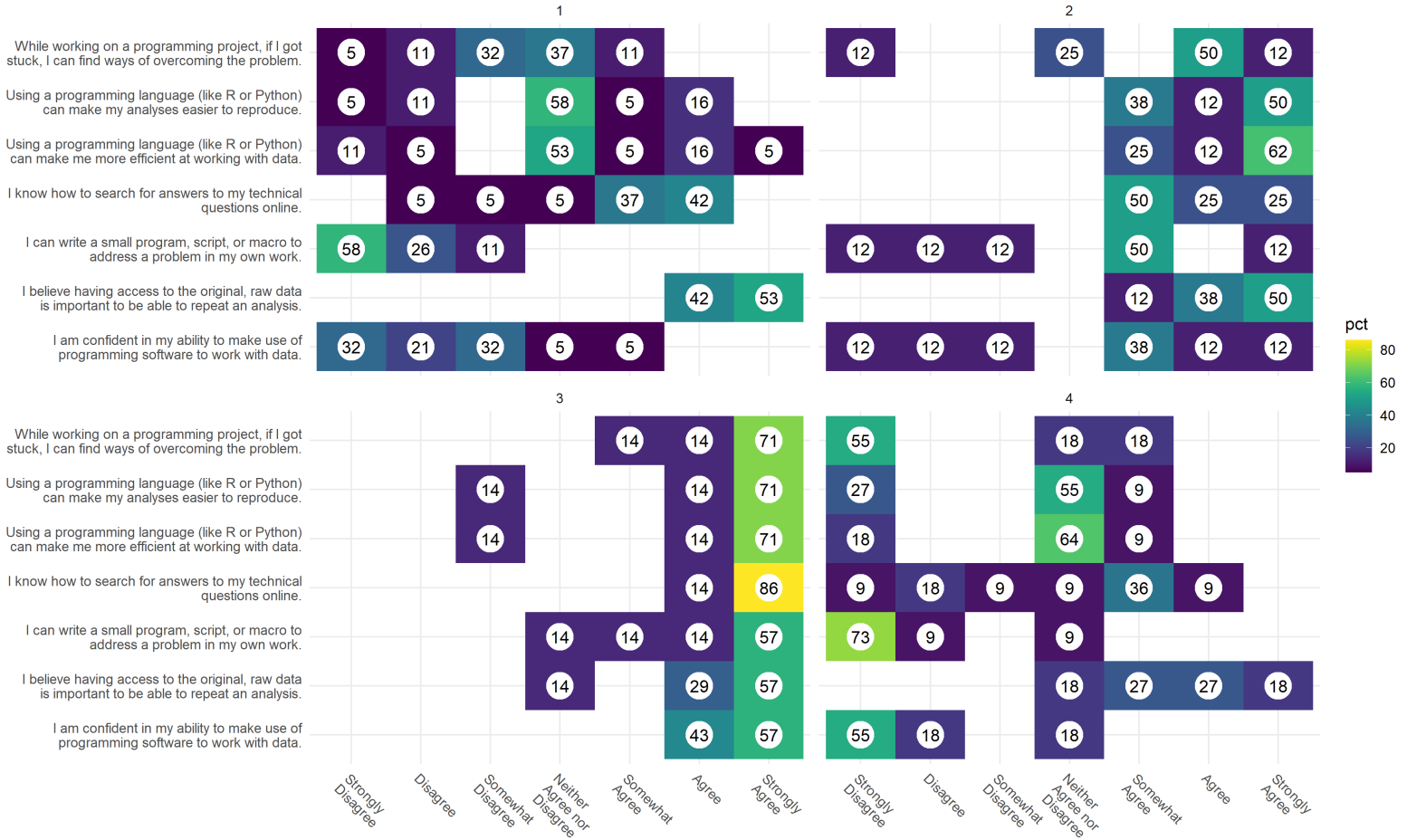
- Alex Academic
- Clare Clinician
- Patricia Programmer
- Samir Student

Persona contains:

1. Background
2. Relevant prior knowledge or experience
3. Perception of needs
4. Special considerations

https://ds4biomed.tech/who-is-this-book-for.html

Overall results by group



(1) Alex Academic, (2) Samir Student, (3) Patricia Programmer, (4) Clare Clinician

PCA

Survey questions were mostly paired, and the PCA results confirm this

```
## Importance of components:
##                            PC1    PC2     PC3     PC4    PC5     PC6
## Standard deviation     2.7051 1.4412 1.22225 1.03560 0.9016 0.82399
## Proportion of Variance 0.4573 0.1298 0.09337 0.06703 0.0508 0.04243
## Cumulative Proportion  0.4573 0.5871 0.68052 0.74755 0.7984 0.84079
##                             PC7     PC8     PC9    PC10    PC11
## Standard deviation     0.76151 0.67018 0.63406 0.62533 0.52667
## Proportion of Variance 0.03624 0.02807 0.02513 0.02444 0.01734
## Cumulative Proportion  0.87703 0.90510 0.93023 0.95467 0.97200
##                            PC12    PC13    PC14    PC15    PC16
## Standard deviation     0.39653 0.34807 0.29827 0.21962 0.17984
## Proportion of Variance 0.00983 0.00757 0.00556 0.00301 0.00202
## Cumulative Proportion  0.98183 0.98940 0.99496 0.99798 1.00000
```

EFA

Picked the highest loaded question in each factor

2 Factors

```
##
## Loadings:
##      Factor1 Factor2
## Q3.1  0.877
## Q3.3  0.972
## Q3.4  1.042  -0.190
## Q3.5  0.957
## Q3.6 -0.135   0.646
## Q3.7 -0.159  -0.192
## Q4.1  0.100   0.428
## Q4.2  0.316   0.417
## Q4.3  0.573
## Q4.4  0.181   0.178
## Q5.1  0.106  -0.372
## Q5.2  0.684   0.153
## Q6.1          0.951
## Q6.2          1.049
## Q6.3          0.925
## Q6.4  0.289   0.290
##
```

3 Factors

```
##
## Loadings:
##      Factor1 Factor2 Factor3
## Q3.1  0.821           0.101
## Q3.3  0.985          -0.123
## Q3.4  0.970  -0.167
## Q3.5  0.984          -0.181
## Q3.6 -0.178   0.483   0.291
## Q3.7 -0.218  -0.292   0.205
## Q4.1          0.196   0.486
## Q4.2  0.248   0.281   0.299
## Q4.3  0.491  -0.115   0.232
## Q4.4         -0.259   0.967
## Q5.1  0.214  -0.143  -0.462
## Q5.2  0.606           0.198
## Q6.1          0.913
## Q6.2          1.021
## Q6.3          0.981  -0.185
## Q6.4  0.115           0.657
##
```

EFA Results

- 3 factors with promax rotation.
    - Each factor coincided with a survey question blocks
- Programming:
    - Q3.3: How familiar are you with interactive programming languages like Python or R?
- Stats:
    - Q6.2: If you were given a dateset containing an individual's smoking status (binary variable) and whether or not they have hypertension (binary variable), would you know how to conduct a statistical analysis to see if smoking has an increased relative risk or odds of hypertension? Any type of model will suffice.
- Data:
    - Q4.4: Do you know what "long" and "wide" data are?

CART

6.2: Logistic regression

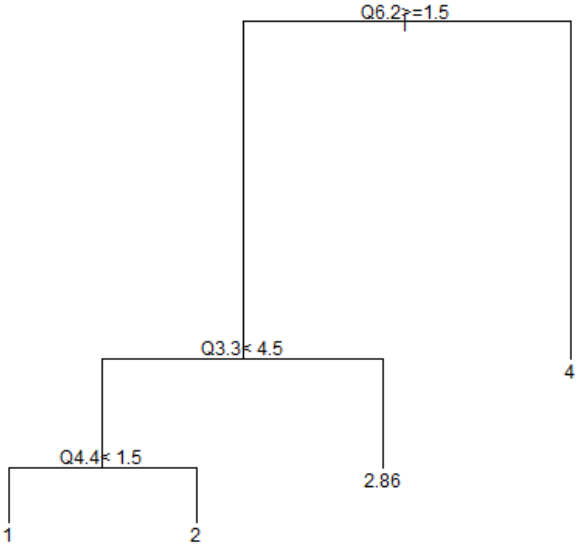
1. "I wouldn't know where to start",
2. "I could struggle through, but not confident I could do it",
3. "I could struggle through by trial and error with a lot of web searches",
4. "I could do it quickly with little or no use of external help"

3.3: Python/R

1. "I do not know what those are",
2. "I have heard of them but have never used them before",
3. "I have installed it, but have only done simple examples with them",
4. "I have written a small program with them before",
5. "I use it to automate certain repetitive tasks",
6. "I have small side projects that I program in it",
7. "I program in them for work"

4.4: Long and wide

1. "I have never heard of the term",
2. "I have heard of it but don't remember what it is.",
3. "I have some idea of what it is, but am not too clear",
4. "I know what it is and could explain what it pertains to"



Phase 2: Workshop results

October 20 + 22, 2020 (T/R)

- 27 People signed up
- 2 Sessions
    - Day 1 12:30 - 15:00 (2.5 Hours)
        - 20 attended
        - Spread sheets, loading data, grouped descriptive statistics
    - Day 2 12:30 - 14:30 (2.0 Hours)
        - 11 attended
        - Tidy data
        - Plotting + logistic regression was blown through

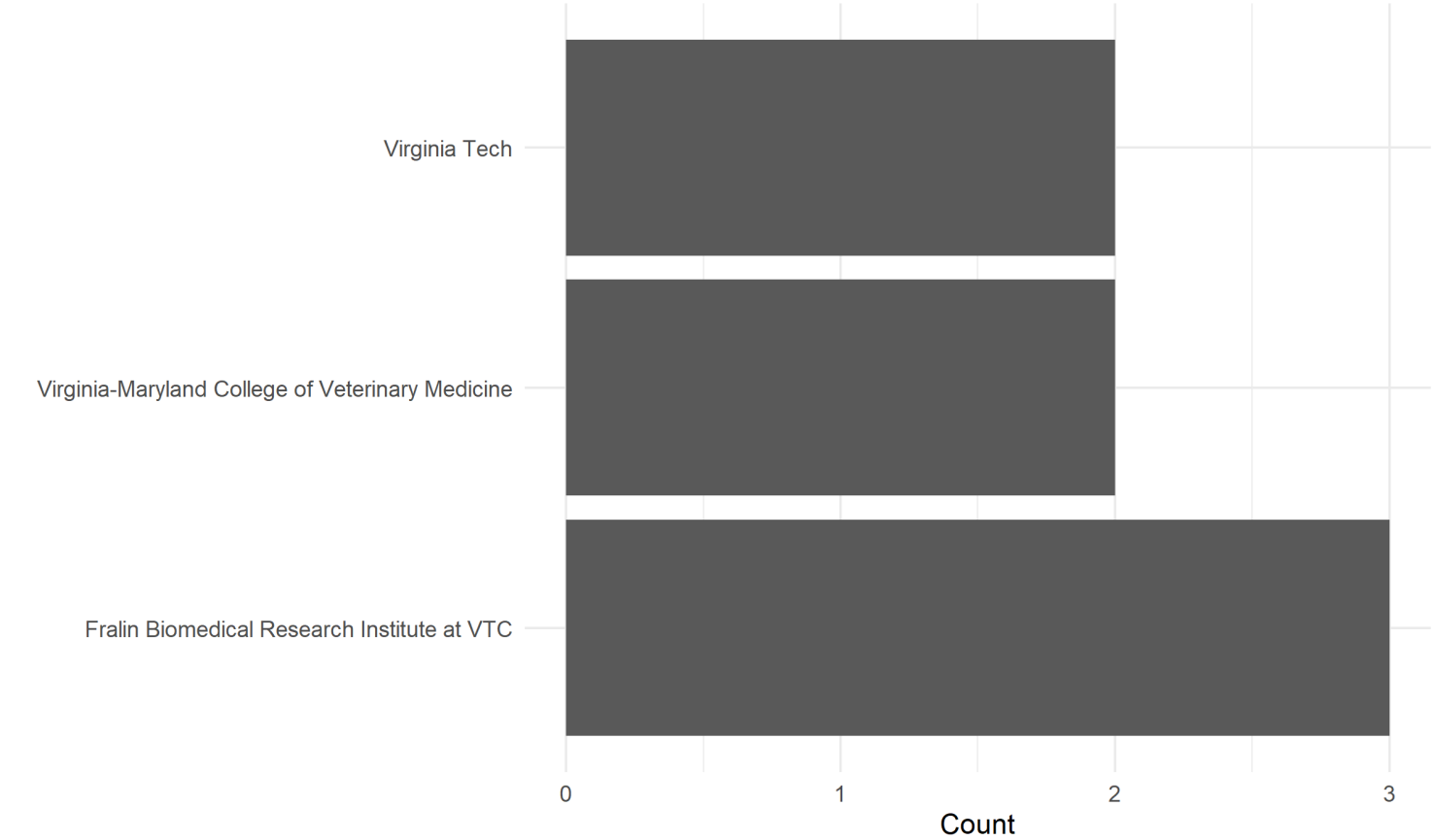
Recordings on YouTube (Bevan & Brown Lab ds4biomed workshop playlist)

Phase 2a: Pre-workshop results

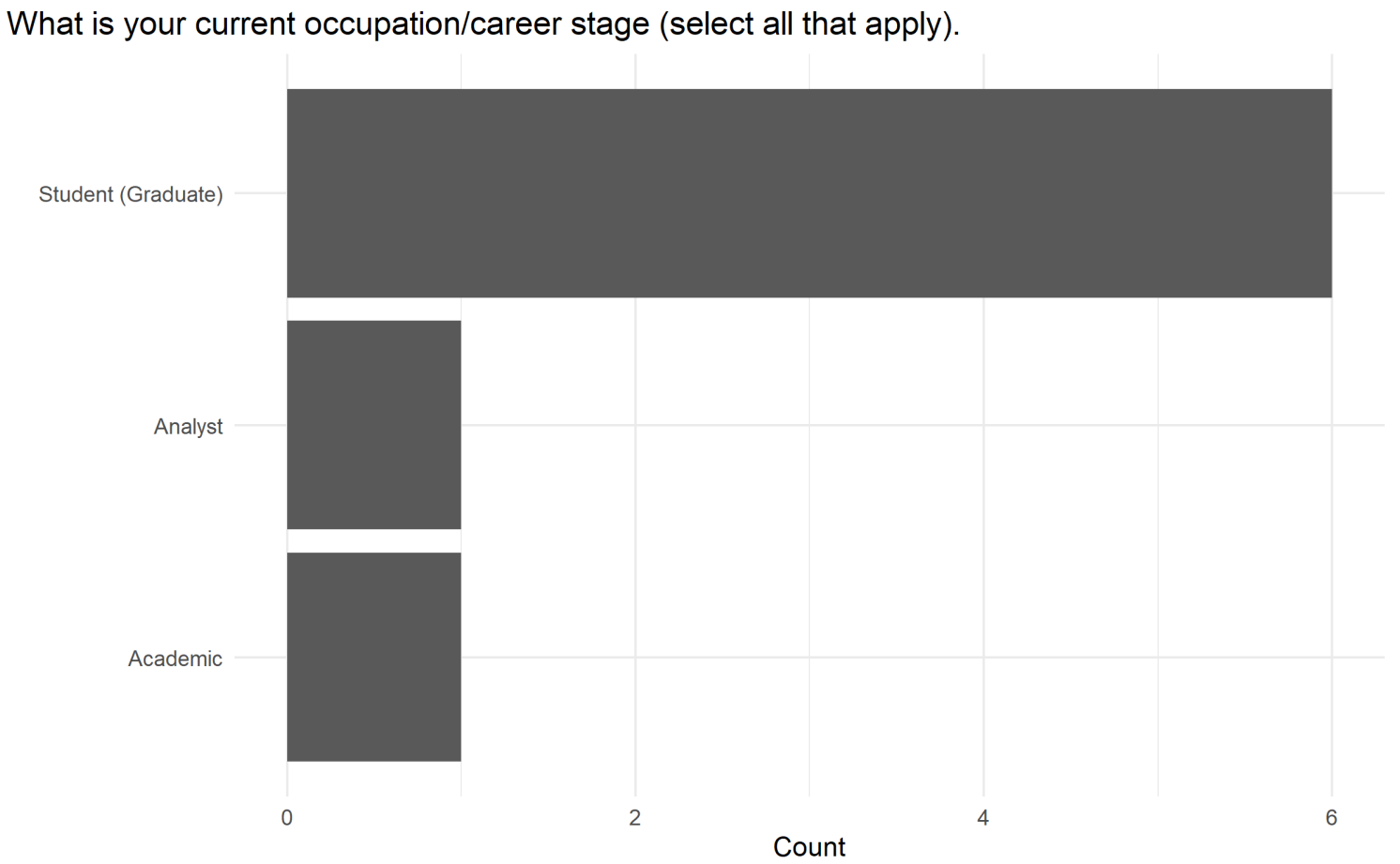
Pre-workshop survey

- n = 8 (7/20 completed survey)

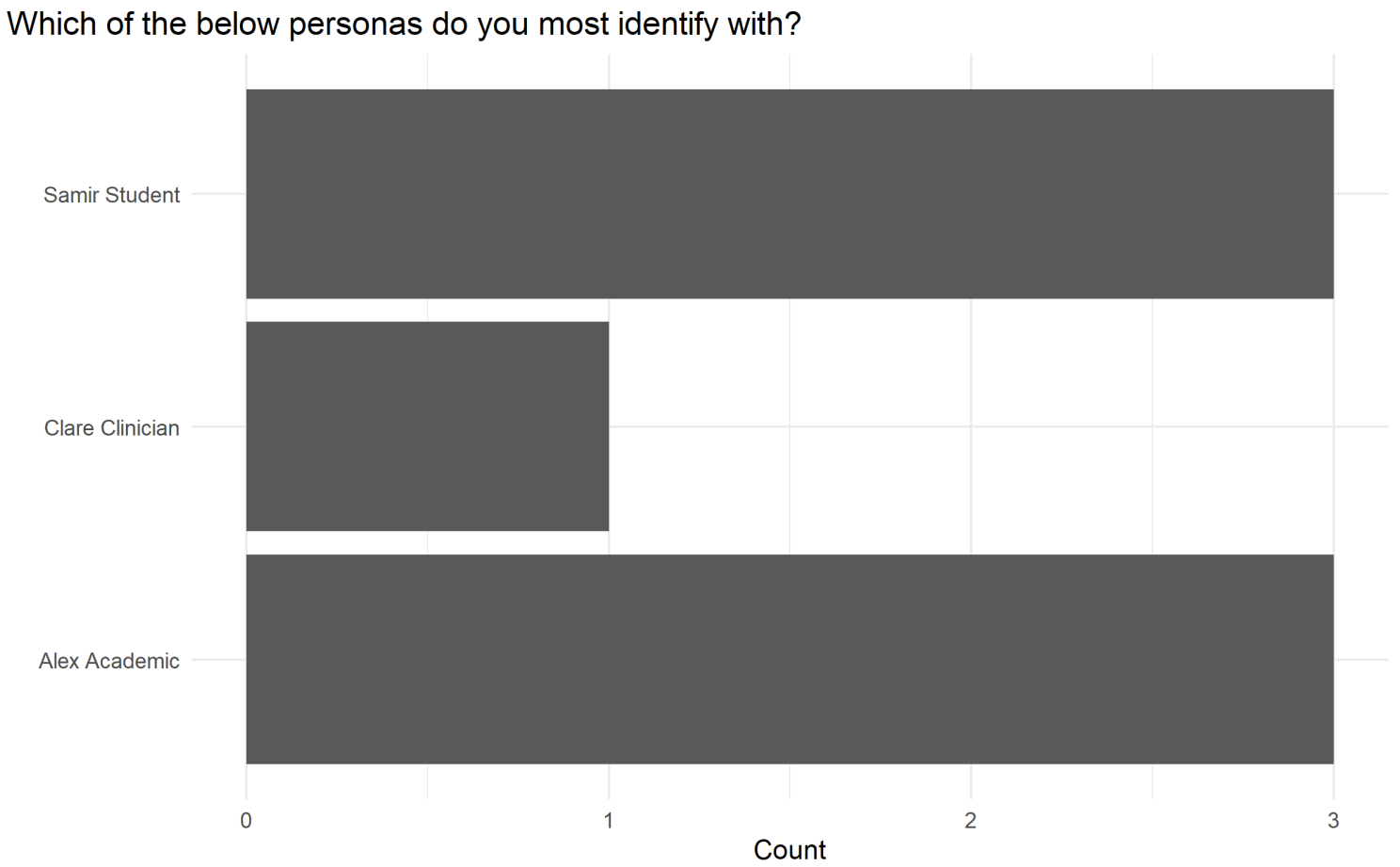
What is your current affiliation?



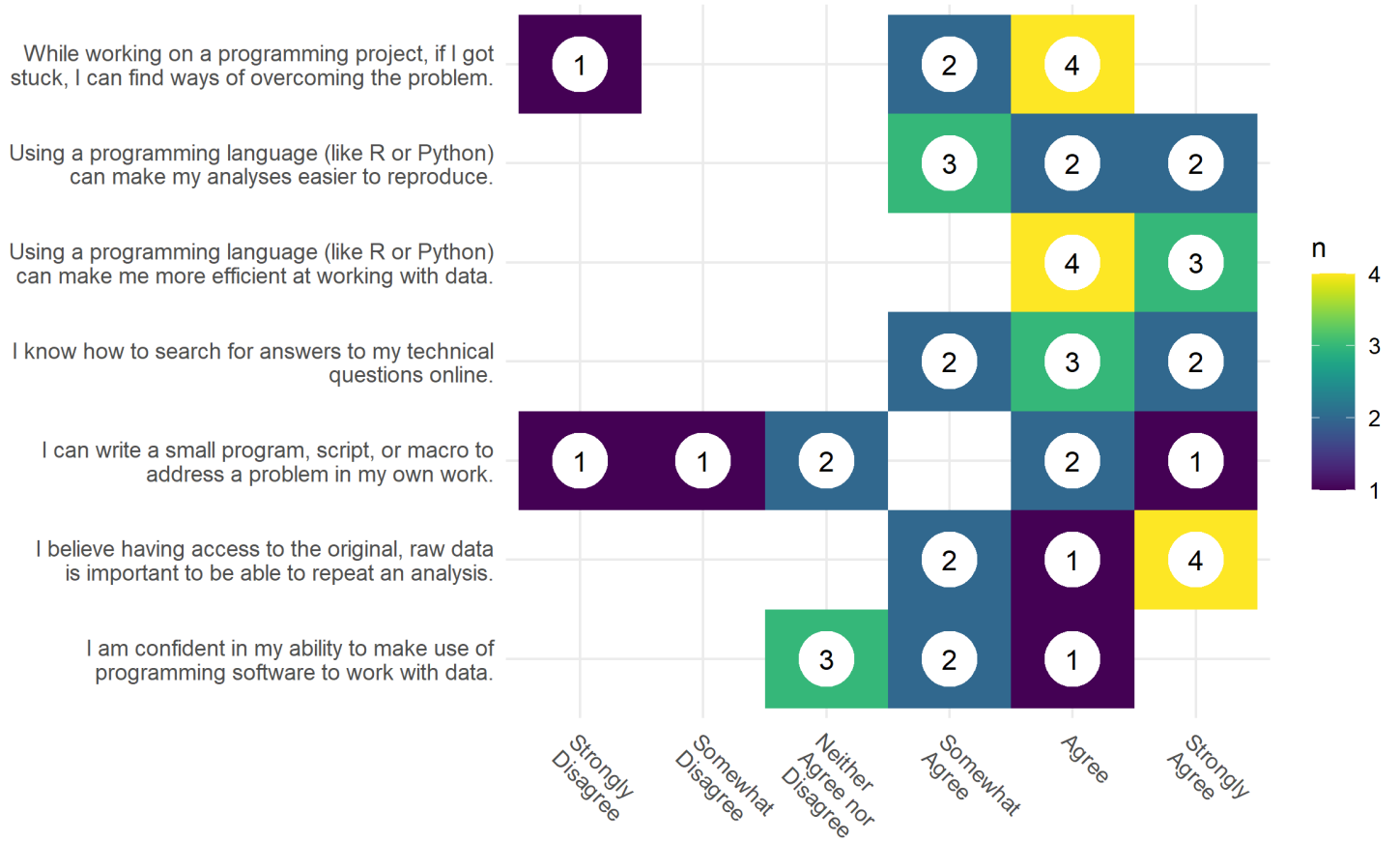
Pre-workshop survey: Occupation



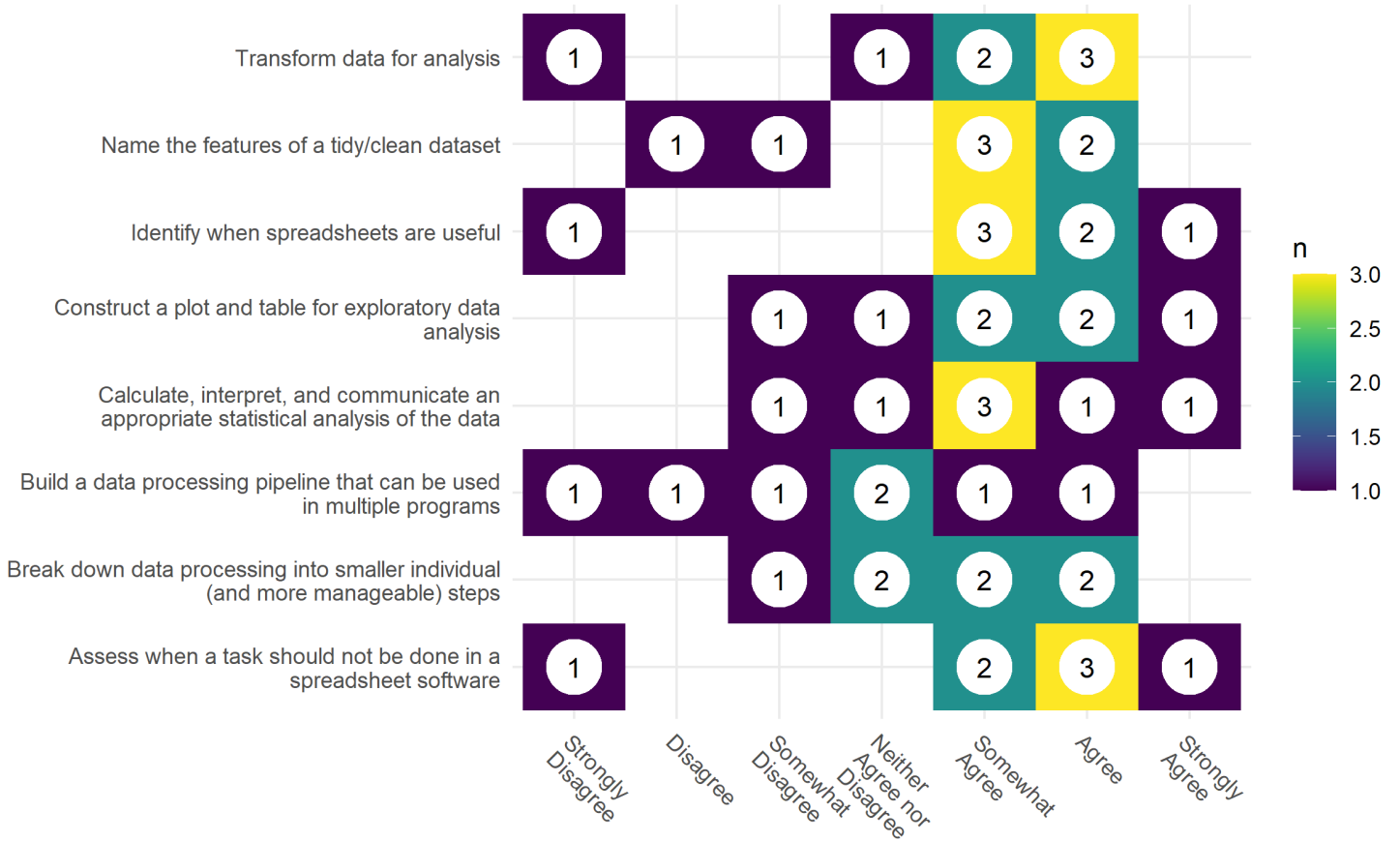
Pre-workshop survey: Self-identification



Pre-workshop survey: Self-assessment

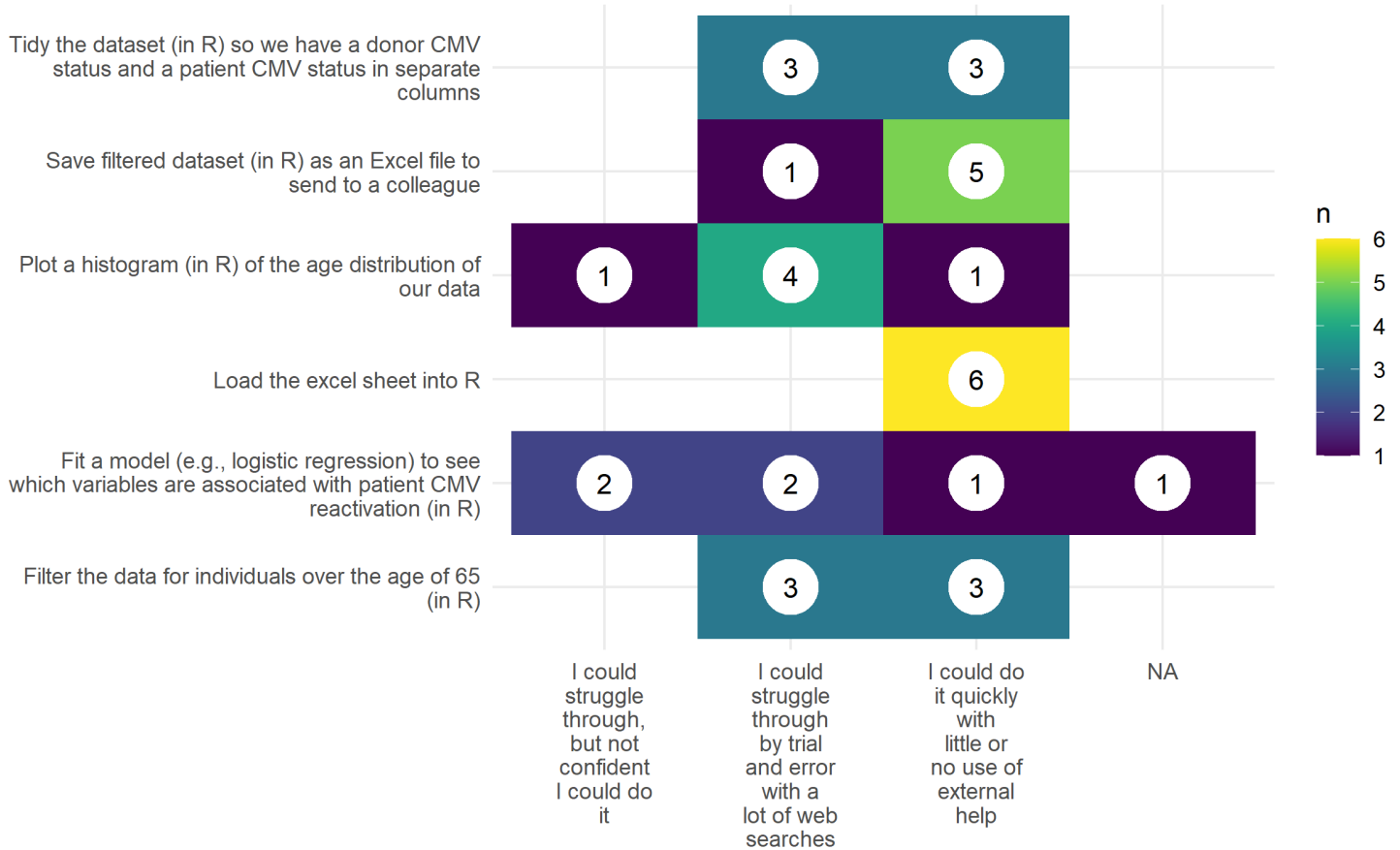


Pre-workshop survey: Learning objectives



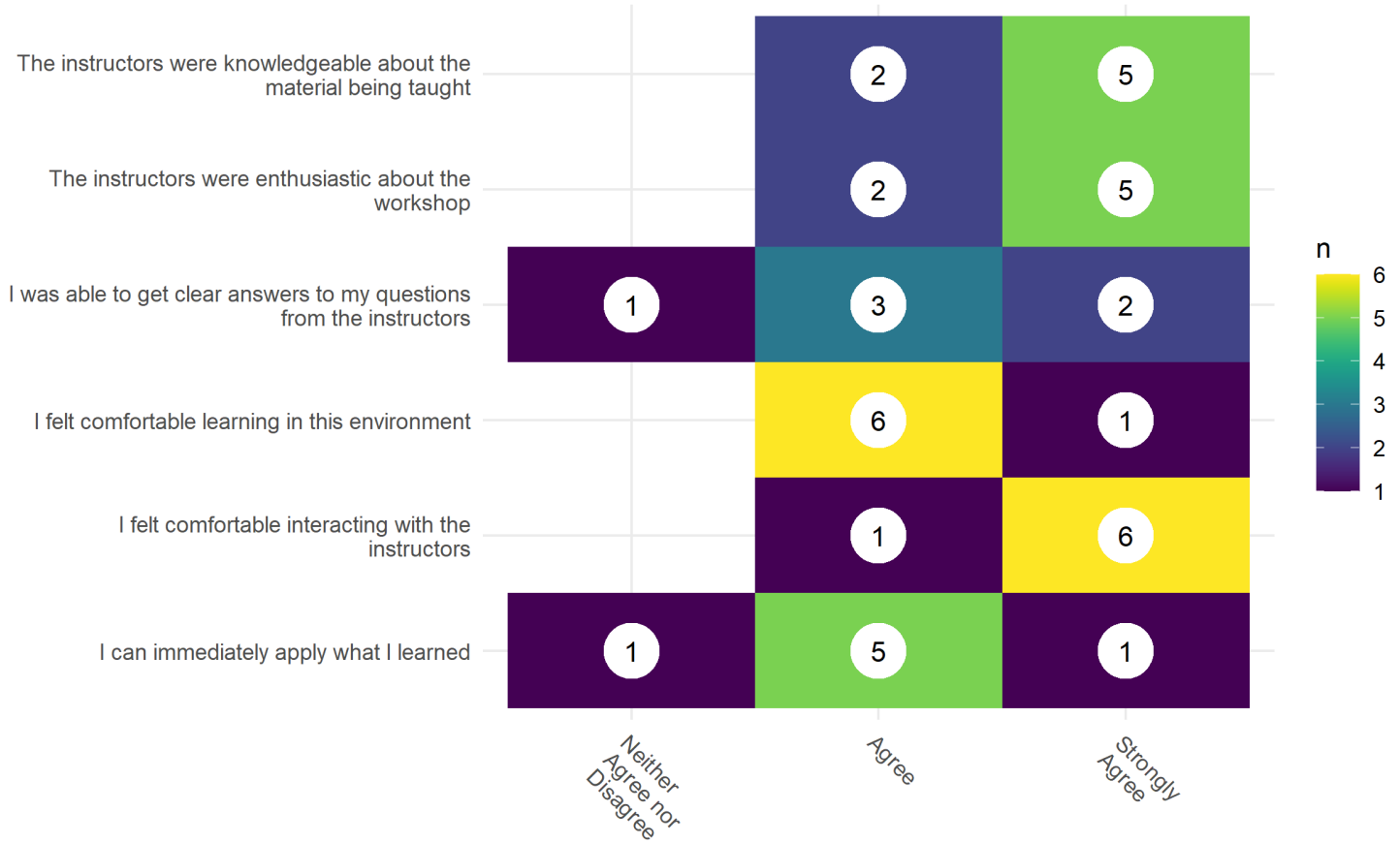
Phase 2b: Post-Workshop results

Post-workshop: Summative assessment



Nobody selected "I wouldn't know where to start"

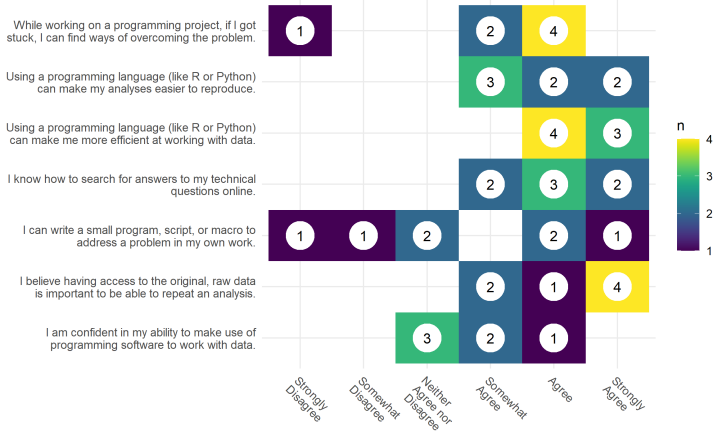
Post-workshop: Environment



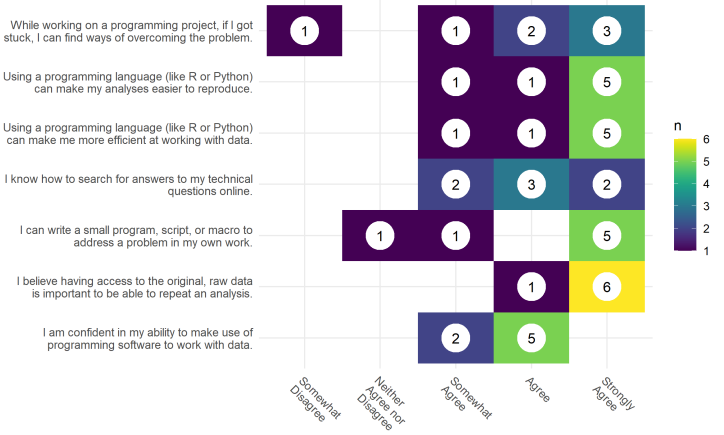
No negative comments about workshop environment

Pre/Post-workshop: Self-assessment

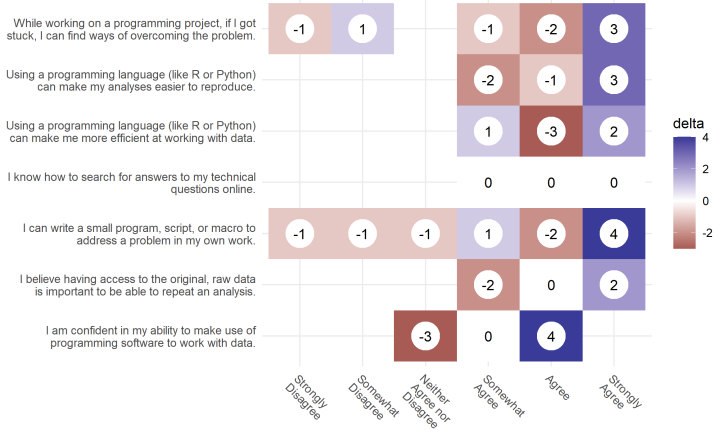
Pre-workshop



Post-workshop

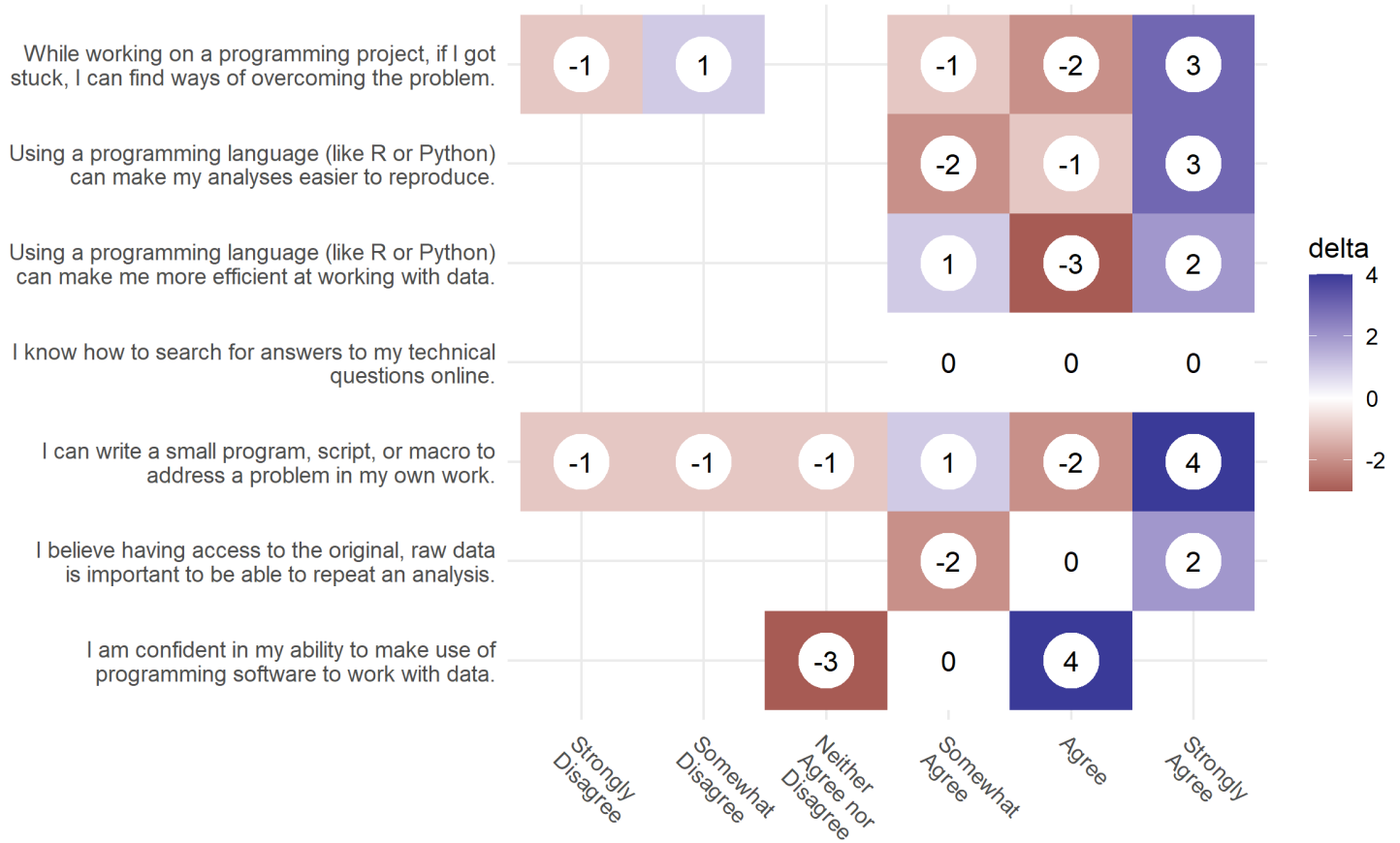


Delta plot



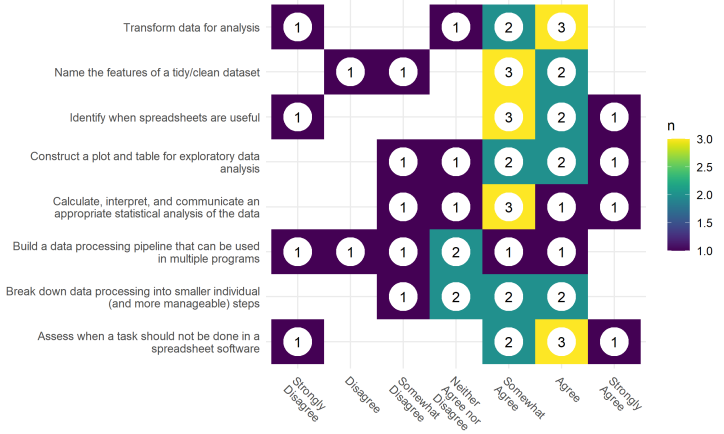
1. ...programming project, if I got stuck, I can find ways of overcoming the problem
2. Using a programming language ... can make my analyses easier to reproduce
3. Using a programming language ... can make me more efficient at working with data
4. I know how to search for answers to my technical questions online
5. I can write a small program, script, or macro ...
6. ... having access to the original, raw data is important...
7. ... confident in my ability to make use of programming software to work with data

Pre/Post-workshop: Self-assessment delta

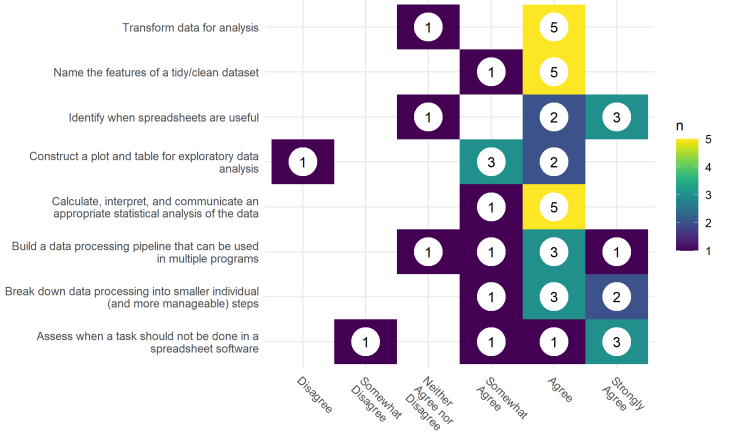


Pre/Post-workshop: Learning objectives

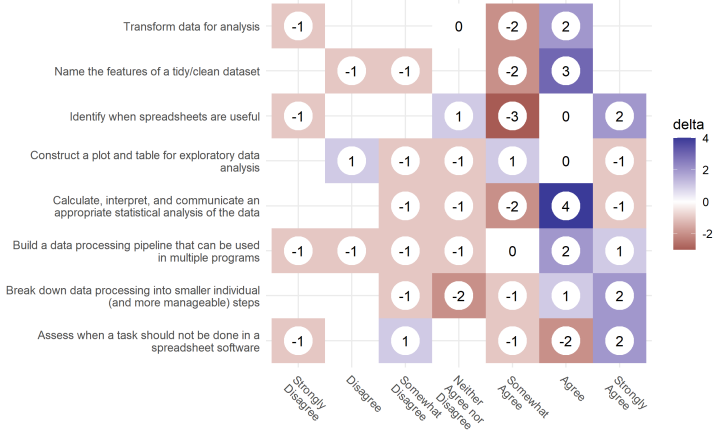
Pre-workshop



Post-workshop

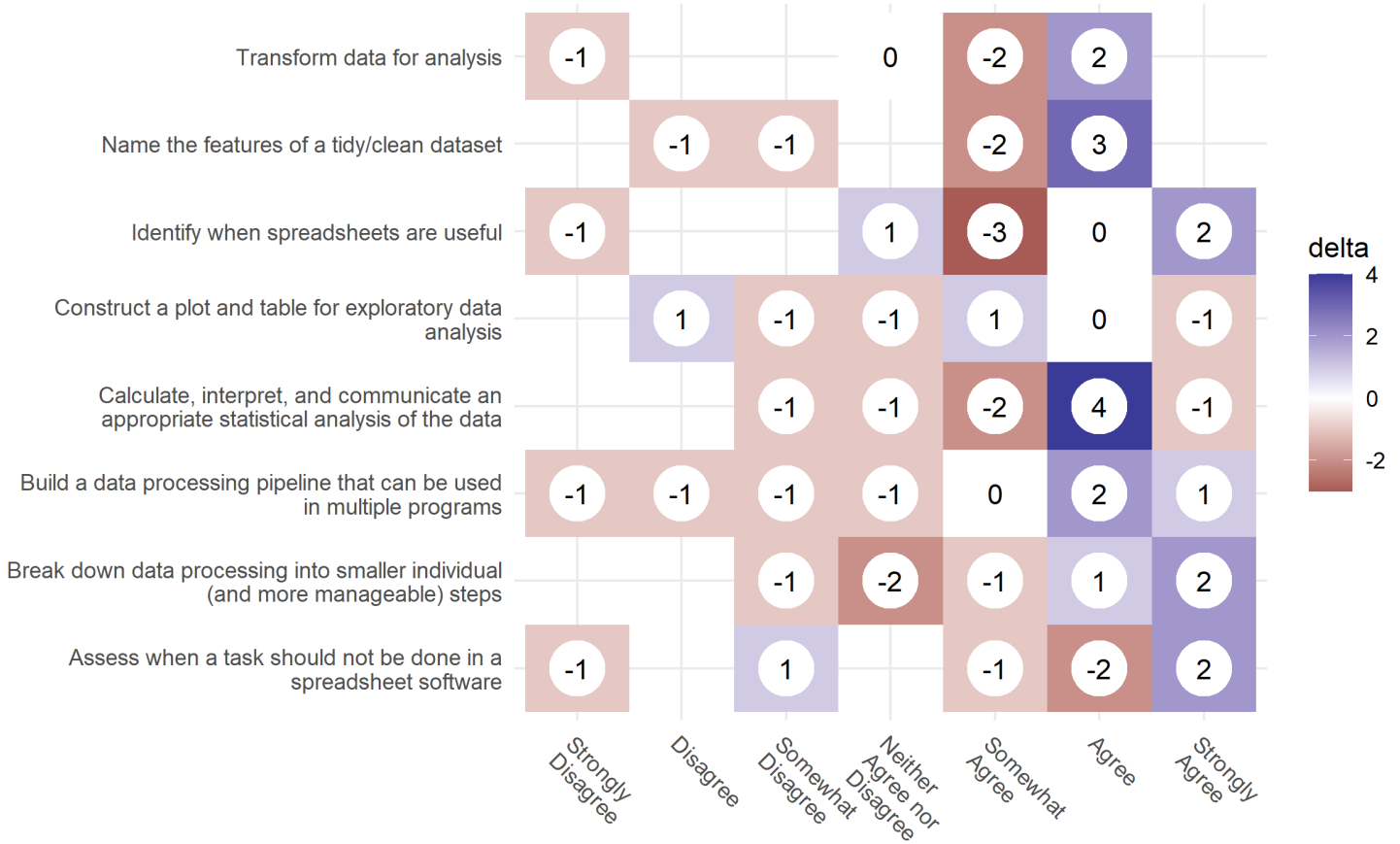


Delta plot



1. Transform data for analysis
2. Name the feature of a tidy/clean dataset
3. Identify when spreadsheets are useful
4. Construct a plot and table for exploratory data analysis
5. Calculate, interpret, and communicate an appropriate statistical analysis of the data
6. Build a data processing pipeline that can b e used in multiple programs
7. Break down data processing into smaller individual (and more manageable) steps
8. Assess when a task should not be done in a spreadsheet software

2 Learning objectives not met



Phase 3: April 2021

Prelims: Feburary 2021

Thanks!

https://github.com/chendaniely/dissertation-presentations