

A Mixed-Method Approach to Investigating Difficulty in Data Science Education

Sydney Shearer^{1^*}, Ellie Strauss^{2^*}; Ethan Hawk^{3^}, Sasha Lioutikova^{4^}, Marius Orehovschi^{5^}, Frankie Vazquez^{3^}

[1] Juniata College, [2] Bates College, [3] Valparaiso University, [4] Yale University, [5] Colby College



Overview

The purpose of this study was to define a methodology to identify disconnect between students and instructors in data science classrooms through analyzing qualitative data. A combined qualitative and quantitative approach was used for analysis of survey data from three institutions. As a whole, the methods used throughout this research process provide direction for researchers in interpretation and analysis of the survey data in an efficient and time-sensitive manner. Although the research was applied to data science classrooms, this method has the potential to be applied into other fields and areas of study when performed with coordination between a field expert and a data scientist.

Dataset

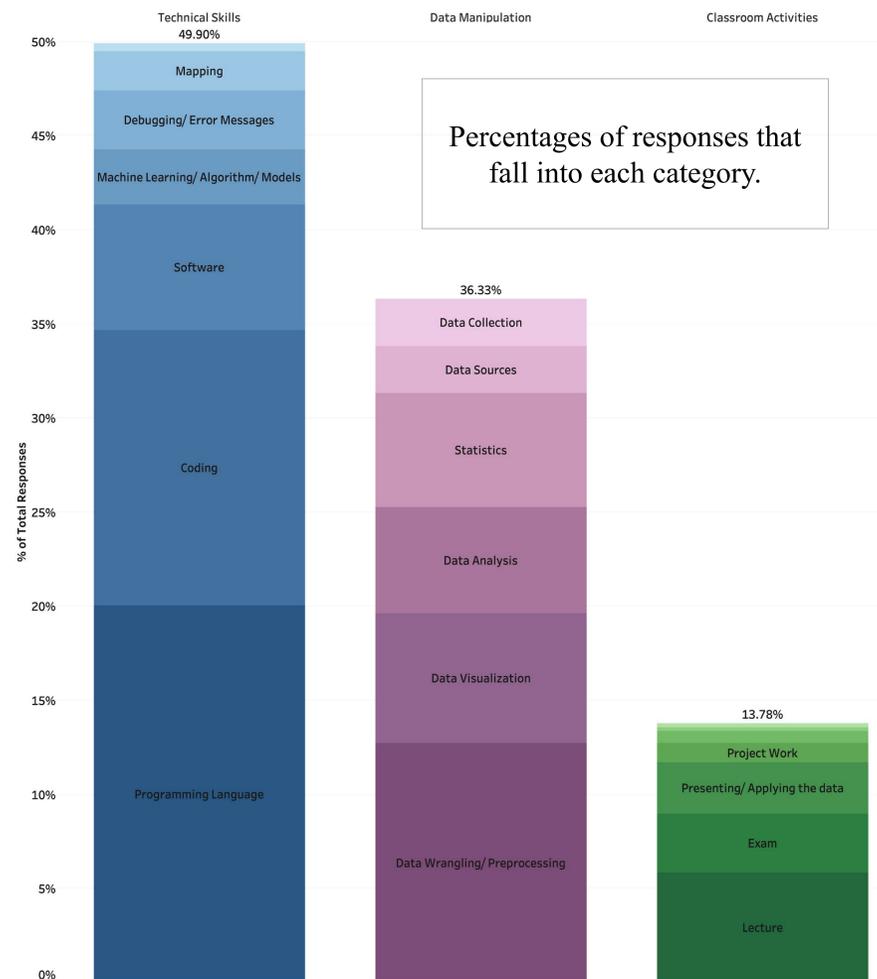
Survey conducted in data science classrooms about aspects of class

- Who was surveyed?**
 - Students (304 responses)
 - Faculty/instructors (112)
 - Teaching assistant (63)
- Where was this conducted?**
 - Brown University
 - Smith College
 - Valparaiso University
- What was asked?**
 - Six Likert Scale Questions
 - Used for overview of dataset
 - Four Open-Ended Questions
 - What topics were covered in class this week?
 - What concepts/activities/ processes did people struggle most with this week?
 - What questions were raised this week?
 - What questions were surprising this week?

Methods

- Natural Language Processing**
 - Removed stop words
 - Lemmatization
 - Tokenization
 - Bags of bigrams
- Quality analysis**
 - Proportion of relevant words in responses
- Manual Content Analysis**
 - Categories created from inductive reasoning, responses categorized
- TF-IDF analysis**
 - Frequency of bigrams
 - Manually labeled into categories
- Corpora**
 - 7 separate corpora used
 - 4 from open-ended questions
 - 3 from respondent roles
 - Responses used as documents
 - Bigrams used as terms

Category Breakdown



TF-IDF Analysis of Bigrams

- TF-IDF: “Term Frequency, Inverse-Document Frequency”
 - Term Frequency (TF): times a term appears in a document
 - Inverse-Document Frequency (IDF): logarithm of the inverse of the number of documents in which the term appears
 - To calculate, multiply TF by IDF to obtain a weight for each term for each document
 - Finds which terms are important within the documents by filtering out words that do not give insight
- | Grouping | “Technical Skills” | “Data Manipulation” | “Classroom Activities” |
|-----------------------------|--------------------|---------------------|------------------------|
| “Questions” (n=13) | 53.85% (n=7) | 30.77% (n=4) | 15.38% (n=2) |
| “Struggles” (n=11) | 36.36% (n=4) | 54.55% (n=6) | 9.09% (n=1) |
| “Surprise Questions” (n=12) | 58.33% (n=7) | 16.67% (n=2) | 25.00% (n=3) |
| “Topics” (n=11) | 9.09% (n=1) | 54.55% (n=6) | 36.36% (n=4) |
| Professor (n=11) | 18.18% (n=2) | 54.55% (n=6) | 27.27% (n=3) |
| Student (n=14) | 14.29% (n=2) | 71.43% (n=10) | 14.29% (n=2) |
| Teaching Assistant (n=12) | 66.67% (n=8) | 16.67% (n=2) | 16.67% (n=2) |

Top 10 TF-IDF terms for each question/role were manually categorized into subcategories

Quality Analysis

Question	Students	Teaching Assistants	Professors
Questions Asked	35.78%	39.64%	42.77%
Surprise Questions	29.12%	20.73%	35.71%
Topics Covered	50.66%	83.33%	65.39%
Topics Struggled	34.71%	39.26%	39.81%

Percentage of non-stop words by respondent role for each survey question.

Subcategory Breakdown

Technical Skills	Data Manipulation	Classroom Activities
Programming Language: 96 (40.17%)	Data Wrangling/ Preprocessing: 61 (35.06%)	Lecture: 28 (42.42%)
Coding: 70 (29.29%)	Data Visualization: 33 (18.97%)	Exam: 15 (22.73%)
Software: 32 (13.39%)	Statistics: 29 (16.67%)	Presenting/ Applying the data: 13 (19.70%)
Debugging/ Error Messages: 15 (6.28%)	Data Analysis: 27 (15.52%)	Project Work: 5 (7.58%)
Machine Learning/ Algorithm/ Models: 14 (5.86%)	Data Sources: 12 (6.90%)	Remote Learning: 3 (4.55%)
Mapping: 10 (4.18%)	Data Collection: 12 (6.90%)	Homework: 1 (1.52%)
Website: 2 (0.84%)		Grading: 1 (1.52%)
GRAND TOTAL: 239 (100.0%)	GRAND TOTAL: 174 (100.0%)	GRAND TOTAL: 66 (100.0%)

Percentages and Counts of responses that fall into each subcategory.

Conclusions/Limitations

- Manual content analysis with an automated algorithmic process gives in-depth understanding to datasets
- Data scientists should collaborate with field experts to gain insight of dataset
 - Without collaboration, an in-depth analysis could not be achieved
- Data was collected during COVID-19 pandemic, resulting in a smaller dataset

Acknowledgements

- Valparaiso University
- EPIC
- TRIPODS+X:EDU NSF Grant
 - DMS #1839257, 1839259, 1839270
- Ruth Wertz, Valparaiso University
- Linda Clark, Brown University
- Katherine M. Kinnaird, Smith College
- Karl R. B. Schmitt, Valparaiso University
- Bjorn Sandstede, Brown University

Interested in learning more?
Read our paper!



^{^*} denotes equal contribution
[^] denotes equal contribution