STAT 495 (Advanced Data Analysis) mini-project: survey report

Group B - Lucas Bernstein, Ephrata Getachew and Shreya Mathew

2024-09-15

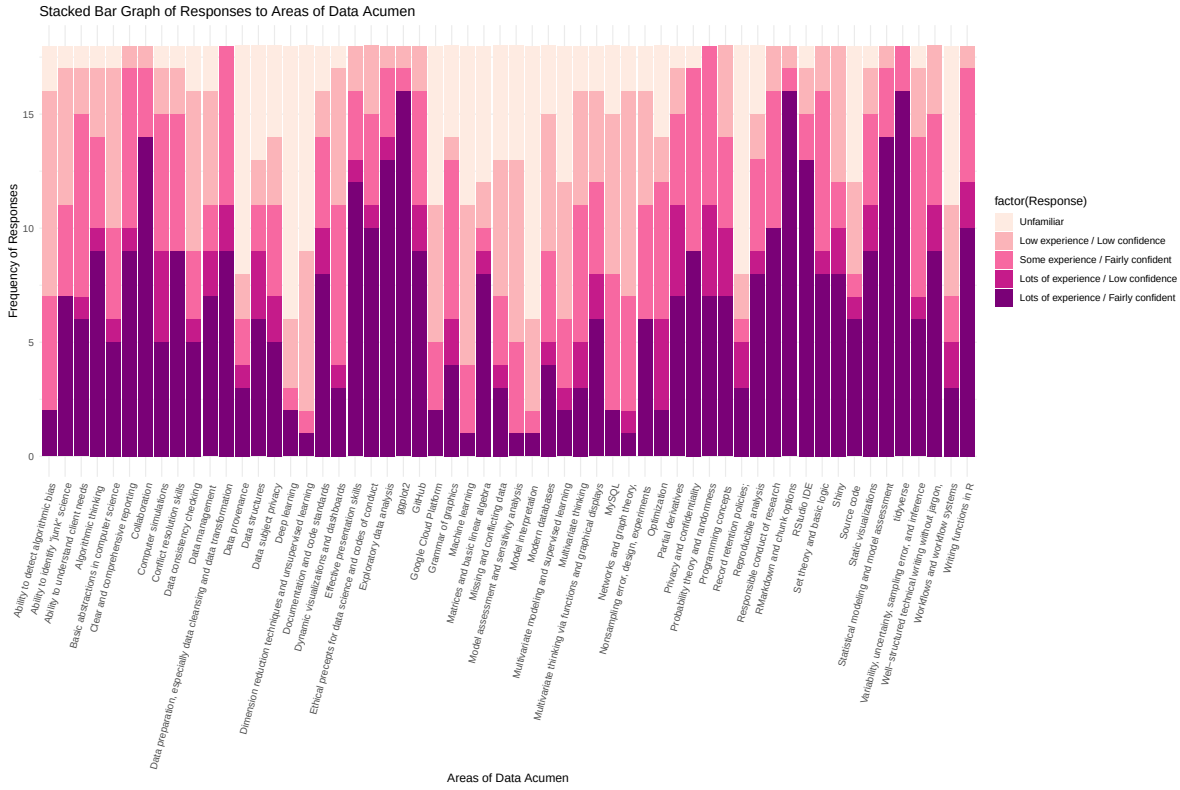
Table of contents

Introduction

Project Data Acumen was conducted in the fall of 2024 as part of STAT 495 to assess the Statistics proficiency of students in the class. The project seeks to evaluate students' experience and confidence across 58 distinct measures of data acumen, spanning critical areas such as mathematical skills, statistical knowledge, computational expertise, and data visualization techniques, among other key dimensions of data competency. This survey-based study collected self-reported levels of experience and confidence from students, enabling a comprehensive analysis of their strengths and gaps in these areas. The collected data was then wrangled and subjected to various analyses to identify trends and patterns in students' reporting. The project aims to answer three central questions: Cohen's Kappa Correlation: What is the correlation between students' agreement on various topics related to data acumen? Experience and Confidence Gaps: Which areas of data acumen do students feel least confident in, and where are the largest gaps in expertise? Common Goals and Interests: What are the most common goals and areas of interest among STAT 495 students in relation to data proficiency?

Data Overview

The dataset was collected from 18 students enrolled in the STAT 495 course, consisting of responses to 60 distinct questions. There were no missing data points. Of these, 58 questions focused on assessing familiarity and confidence across various aspects of data acumen. The response options were categorical, with choices including: "Unfamiliar," "Some experience / Low confidence," "Some experience / Fairly confident," "Lots of experience / Low confidence," and "Lots of experience / Fairly confident." The final two questions were open-ended, addressing students' learning goals and curiosity in the course.



How closely do students' self-reported confidence level align across different topics related data acumen, and what patterns of agreement or disagreement emerge from this comparison?

To address this question we assessed the consistency of students' self-reported confidence across two different topics, comparing how closely their confidence levels align. Cohen's Kappa quantify the level of agreement between two variables, where 1 represents perfect agreement, 0 represents no agreement beyond choice, and negative values indicate disagreement.

Highest Agreement

Table 1: top 3 pairs with highest agreement

Topic_1	Topic_2	Kappa
RMarkdown and chunk options	ggplot2	1.0000000
Writing functions in R	GitHub	0.7339901
RMarkdown and chunk options	tidyverse	0.7272727

RMarkdown and Chunk Options and ggplot2 (Kappa = 1.00) The highest agreement was observed between confidence in using RMarkdown and ggplot2. A perfect Kappa value of 1.00 suggests that students who are confident in one of these tools are equally confident in the other, reflecting the strong connection between effective documentation of code and the visualization of results. These tools are often used together in data analysis, which may explain the perfect alignment in students' confidence. Writing Functions in R & GitHub (Kappa = 0.73): Students who are confident in writing functions in R tend to also feel confident using GitHub for version control. This high Kappa score suggests that those proficient in writing modular, reusable R code are also familiar with managing code with GitHub. These two skills complement each other, as writing functions often requires version control to ensure collaboration. RMarkdown and Chunk Options & tidyverse (Kappa: 0.73): Confidence in using RMarkdown for organizing code chunks and reporting results strongly aligns with confidence in using the tidyverse. Both tools are integral to modern data analysis workflows, and the high agreement suggests that students who excel at organizing their analysis in a clear and reproducible format are also proficient in manipulating and analyzing data using tidyverse packages.

Table 2: Top 3 Pairs with Lowest Agreement

Topic_1	Topic_2	Kappa
Missing and conflicting data	MySQL	-0.2264151
Shiny	Google Cloud Platform	-0.2222222
Probability theory and randomness	Reproducible analysis	-0.2203390

Lowest Agreement

Missing and Conflicting Data & MySQL (Kappa: -0.23): The lowest agreement was found between confidence in handling missing and conflicting data and using MySQL. This suggests that students who are comfortable managing messy data are not necessarily confident in working with structured databases, indicating a divergence between skills in data management and database operations.

Shiny & Google Cloud Platform (Kappa: -0.22): There is a low alignment between students' confidence in using Shiny and the Google Cloud Platform. This suggests that students who excel in building Shiny apps may not feel confident in deploying or managing them in a

cloud-based environment like GCP, highlighting a gap between app development and cloud infrastructure skills.

Probability Theory and Randomness & Reproducible Analysis (Kappa: -0.22): A low Kappa value between confidence in probability theory and reproducible analysis indicates that students who are proficient in foundational statistical theory do not necessarily feel confident in ensuring that their analyses are reproducible. This gap suggests a distinction between theoretical understanding and practical application in ensuring reproducibility in research.

Where do gaps exist between students' experience and confidence across key dimensions of data acumen, and which specific variables show the largest discrepancies?

To address this question, we utilized radar charts as a primary visualization tool to map the self-reported levels of experience and confidence of students across various dimensions of data acumen, including Mathematical, Computational, Statistical, Data Management and Curation, Data Description and Visualization, Data Modeling and Assessment, Workflow and Reproducibility, Communication and Teamwork, Ethics, and Tools and Platforms.

The radar chart is structured to display individual categories of data acumen, with each axis representing a specific skill or concept within that category. The plotted points on each axis range from 0 to 4, where 0 indicates unfamiliarity with the skill, and 4 indicates a high level of experience and confidence. By exploring the overall shape and extent of the radar chart, we can identify patterns that indicate which categories or topics consistently show lower confidence or larger gaps.

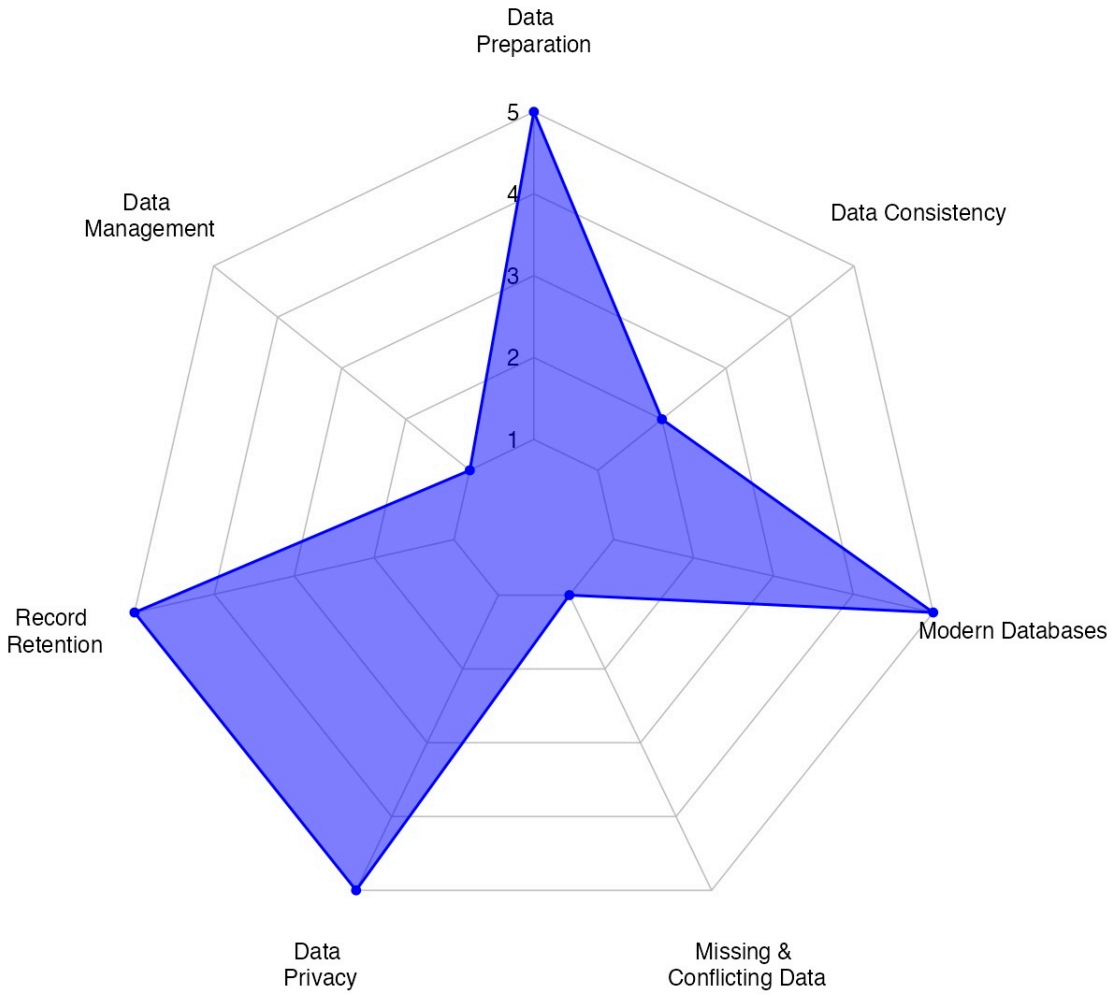


Figure 1: Data management and curation

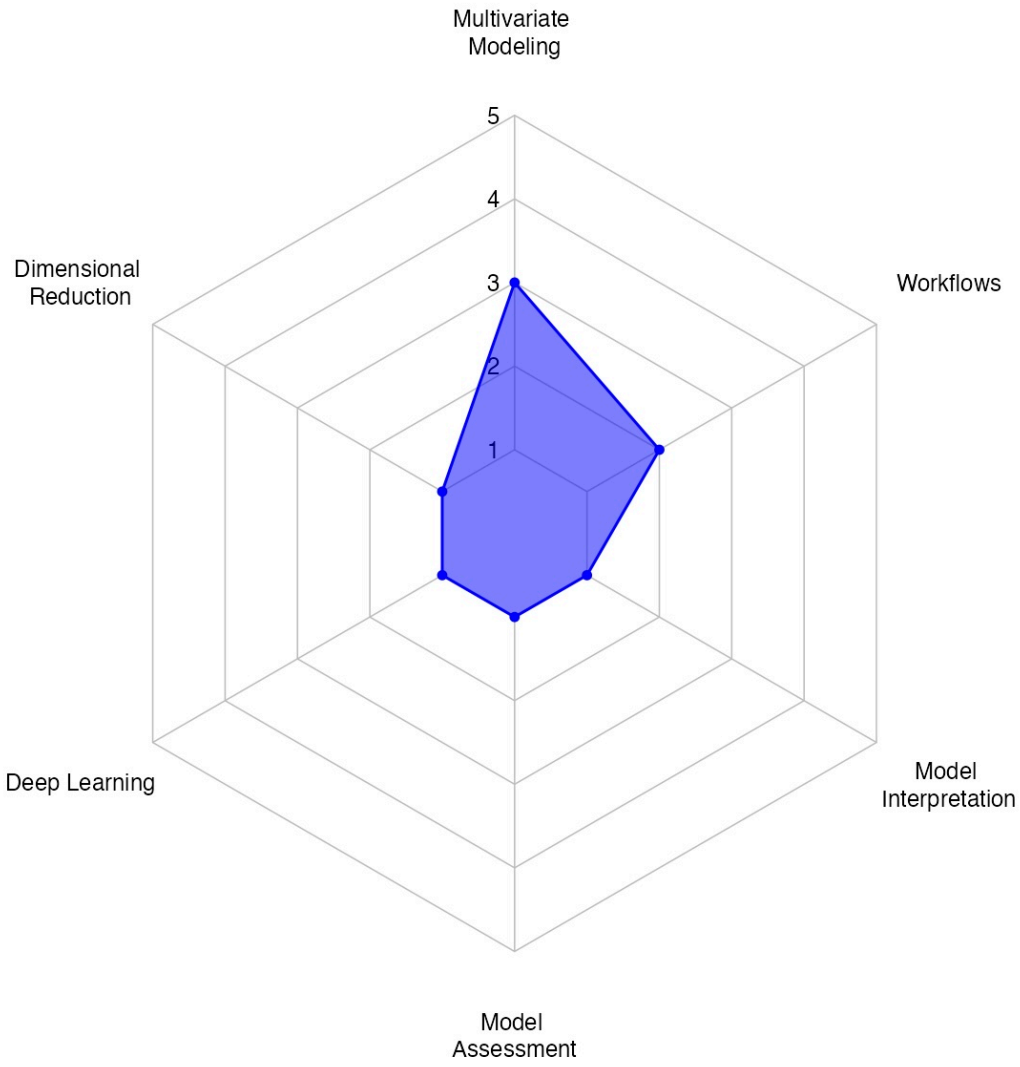
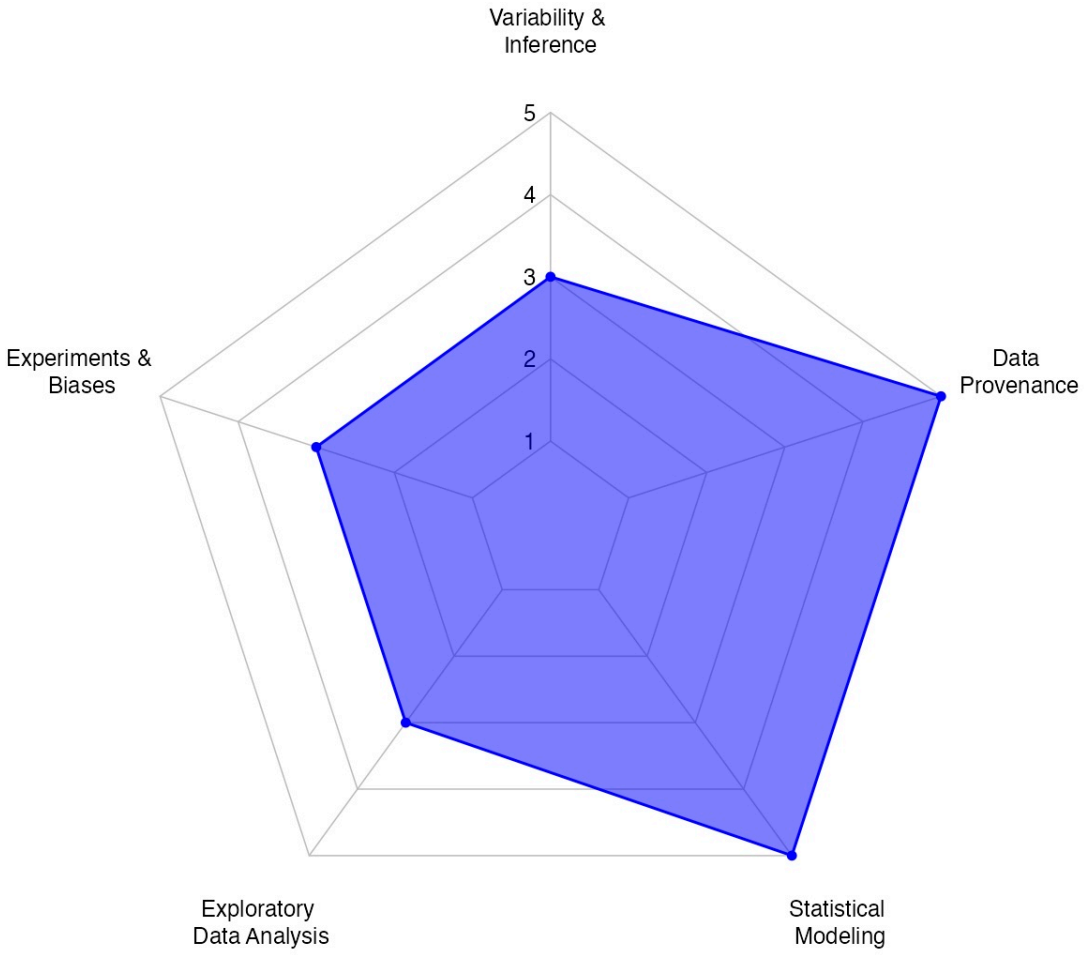
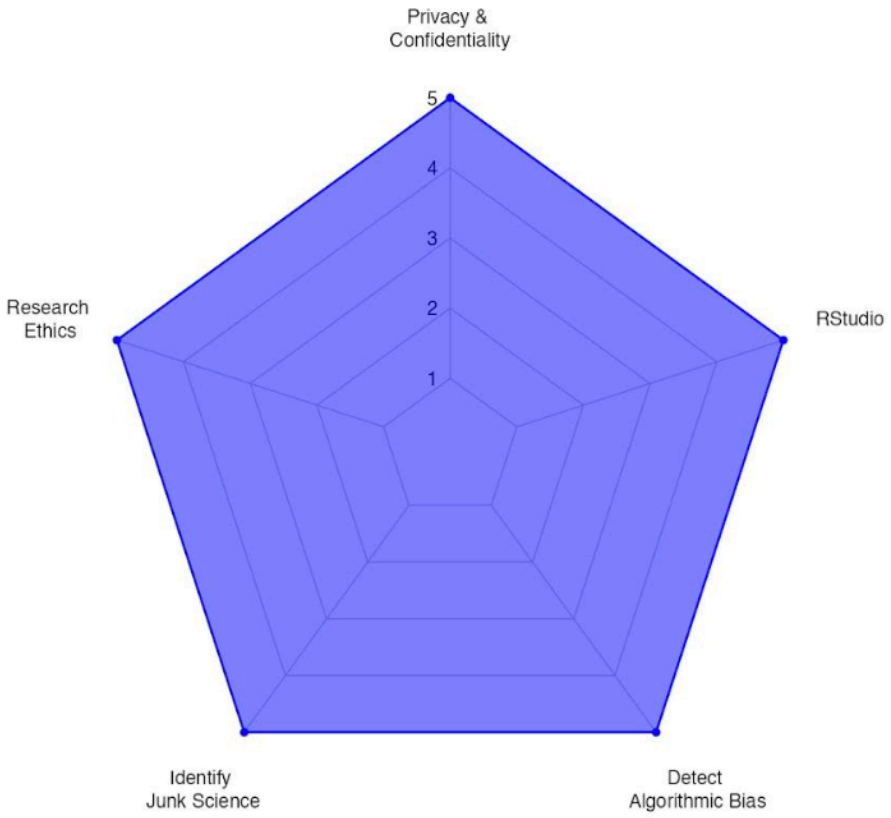


Figure 2: Statistical acumen


Variability &
Inference
5
4
3
2
1
Data
Provenance
Statistical
Modeling
Exploratory
Data Analysis
Experiments &
Biases



In the statistical acumen category, students demonstrate high levels of experience and confidence in both data provenance and statistical modeling. Interestingly, despite reporting only some experience with variability and inference, experiments, and exploratory data analysis (EDA), they still express fairly high confidence in these areas, suggesting a comfort with the application of statistical concepts even with limited hands-on experience. In contrast, the data modeling and assessment dimension shows the most unfamiliarity, with students lacking both experience and confidence in more advanced areas such as deep learning and model interpretation, indicating a need for more exposure to these techniques. The data management and curation category highlights notable variability in student expertise; while students are confident and experienced in data preparation, they are less equipped to handle missing data and more advanced management tasks, pointing to areas where further skill development is necessary. The most surprising finding comes from the ethics dimension, where students show both high experience and confidence across all areas tested, reflecting a strong foundation in ethical considerations related to data.

What are the most common goals and interests among STAT 495 students?

While the Cohen's Kappa correlation analyses and the radar chart category analyses illuminate the skillsets of STAT 495 students more concretely, looking at longer and more freeform responses allows us to more holistically analyze how students would like to engage with STAT 495 going forward.

In order to explore this question, we performed text analysis to find the most common words and bigrams (sets of two words in succession) in the two text based questions at the end of the data acumen survey: "What are your personal goals for this course?" and "Are there specific topics that you are particularly interested in studying?", hereafter referred to as the "goals" question and the "interest" question, respectively. These datasets were then visualized using the `wordcloud` function in R.


learning
prevalent
apps
data
world
rules
experience
limits
love
music
techniques
deep
fields
real
networks
machine
applications
sql
client
supervised
related
expanding
safety
statistics


advanced statistics
identify specific
final project
stat major
final statistics
understand
visualization skills
complex methods
comfortable knowing
theoretical knowledge
real world
grad school
ities strengthen

The word clouds illuminate that there were not a whole lot of common goals among STAT 495 students for the course; most of the common words were general statistics terms. However, in terms of bigrams, "real world" and "applied statistics" both came up twice, indicating at least some shared interest among students Other common themes included getting increased knowledge and confidence with the material (understanding 3x + comfortable 2x + strengthen 2x).

In terms of interests, machine learning and deep machine learning came up more often than any other terms. The same two bigrams appeared twice as those for the goals question, with nothing else occurring more than once. Interests in music, visualization, and new methods also appear in multiple distinct contexts.

Overall, people have fairly disparate goals and interests for this course, or at least conceive of them differently! Although there are at least some shared goals of gaining comfort with certain statistical methods and skills and learning how to apply them in the real world, beyond that STAT 495 students are approaching the course with a wide variety of perspectives and passions. While there are limitations to interpretation due to the numerous ways similar goals and interests can be phrased in a freeform response, examining it through this lens allows us to find some common ground to build an environment that is exciting and enriching for all students.

Conclusion

While analysing the gaps in areas of data acumen, it was revealed that noticeable gaps exists in students' experience and confidence, particularly within Data management and curation dimension. Further, students were most unfamiliar in the data modelling and assessment dimension of data acumen. In general, while students are mostly versed in theoretical concepts, there needs to be more practical exposure for these concepts.

The highest agreement occurs in areas where tools and skills are closely related and frequently used together. For example, RMarkdown, ggplot2, and tidyverse are all part of the same tools in R, so it is not surprising that students proficient in one are also proficient in the others.

On the other hand, the lowest agreement is found between more distinct skills, such as managing missing data and using MySQL, or working with Shiny and Google Cloud Platform. These discrepancies suggest that while students may excel in one area of data science, they may lack confidence in related but more specialized tools or platforms.

To address the discrepancies in confidence, we can offer targeted workshops and real-world projects that require students to use both familiar and unfamiliar tools, helping them build confidence and proficiency. Encouraging peer learning and collaborative projects can further reinforce these skills and support students in developing a well-rounded understanding of the data science pipeline.

When examining students' personal goals, it seems that students are most interested in developing applied statistical knowledge and becoming more comfortable with tools that they have already been exposed to. In regards to their interests surrounding acumen skills, aside from machine learning there is a wide and diverse set of interests, indicating that a balanced and widely applicable curriculum would better serve the class than rigidly focused units. For the most part, students just want to feel set up for their futures beyond Amherst with deeper understandings across the statistical canon.

Overall, our analyses suggest that to create a balanced and effective learning environment, the STAT 495 curriculum should focus on being comprehensive and wide-ranging. Emphasis should be placed on boosting students' confidence with both familiar and unfamiliar tools, while providing more hands-on, "real-world" experiences. By bridging the gaps between theoretical knowledge and practical application, the course can better equip students with the skills and confidence they need to navigate complex data science challenges in future academic and professional settings.