Asking Questions about Data: First-year Engineering Students' Introduction to Data Analytics

Mr. Ruben D. Lopez-Parra P.E., Purdue University-Main Campus, West Lafayette (College of Engineering)

Ruben D. Lopez-Parra is a graduate research assistant at Purdue University pursuing a Ph.D. in Engineering Education. Previously, he worked as a Natural Science teacher in High School where he, as a scholarly teacher, constantly assessed his performance to design better learning environments that promote students' conceptual understanding. In 2015, Ruben earned the M.S in Chemical Engineering at Universidad de los Andes in Colombia where he also received the title of Chemical Engineer in 2012. His research interests include cognition and metacognition in the engineering curriculum.

Mr. Aristides Carrillo-Fernandez, Purdue University

Arístides Carrillo Fernández is a Ph.D student in School of Engineering Education at Purdue University. He was previously an export business development manager at a Spanish radio communications company in Madrid, Spain. For over six years., he was developing new distribution dealer networks in South Europe and West Africa countries. He earned his M.S. in Electronics and Systems of Telecommunication at ESIGELEC (École Supérieur D'Ingénieurs en Génie Électrique) at Rouen, France in 2009, and his B.S. in Systems of Telecommunication at Polytechnic University of Madrid at Madrid, Spain in 2006. Arístides' research interests include the role of empathy and reflection in learning in engineering education and practice contexts, and professional development in global environments.

Amanda Johnston, Purdue University-Main Campus, West Lafayette (College of Engineering)

Amanda Johnston is a PhD candidate in engineering education at Purdue University.

Prof. Tamara J Moore, Purdue University-Main Campus, West Lafayette (College of Engineering)

Tamara J. Moore, Ph.D., is a Professor in the School of Engineering Education and Interim Executive Director of the INSPIRE Institute at Purdue University. Dr. Moore's research is centered on the integration of STEM concepts in K-12 and postsecondary classrooms in order to help students make connections among the STEM disciplines and achieve deep understanding. Her work focuses on defining STEM integration and investigating its power for student learning. Tamara Moore received an NSF Early CAREER award in 2010 and a Presidential Early Career Award for Scientists and Engineers (PECASE) in 2012.

Dr. Sean P Brophy, Purdue University at West Lafayette

Dr. Sean Brophy is the Co-Leader of the Educational, Outreach and Training them for the George E. Brown Network for Earthquake Engineering Simulation (NEES). His research in engineering education and learning sciences explores how children learn through interactions with technologies ranging from manual manipulative like structures students design build and test with shake tables to digital manipulative with mobile devices. He continues to explore new methods to enhance informal and formal learning experiences.

Asking Questions about Data: First-year Engineering Students' Introduction to Data Analytics

Abstract

This complete research paper aims to understand the question design's process of first-year engineering students when performing data analytics. Specifically, we aim to answer the research question: *How do first-year engineering students use a large data set to ask questions focused on the client's needs?* While most research in the area of analytics has focused on how to perform the data analytics cycle successfully, the learning process behind the practice of data analytics is still not totally understood. This study was conducted with 53 first-year engineering students who worked in 14 teams to solve an open-ended problem of data analytics called: The Bike-share Problem. Students were tasked to download the freely available data from Capital Bikeshare company (~3 million data points) and do a preliminary analysis to understand the data set and the company itself. Students proposed questions individually to explore the data and, as a team, design a team question focused on the client's needs. We used content analysis to develop a codebook and analyze the 92 individual and 13 team questions. Our results showed students were able to take the data and frame a problem until designing a question that considered the client's perspective. However, some students ended up writing questions that were not clear due to ambiguous terms, so simple as not to be useful to the client, or too complex to be answerable with the date and time they had. Additionally, students' questions often focused on a single, simple variable of the data and not utilizing the breadth of the data available to them. Finally, the student's previous knowledge of statistics could have mediated their question design practice and limited their ability to answer their questions. We presented some suggestions for researchers and professors who want to study and teach analytics. The Bike-share problem is an example of how analytics can be successfully integrated early in engineering curricula, and we animate professors to implement similar activities in their courses.

Introduction

Gathering and analyzing large data sets from customers' behaviors have allowed companies to propose novel strategies to improve their business by making data-driven decisions. By applying novel tools such as machine learning, computer scientists have designed algorithms that enable organizations to gather, store, and analyze large data sets. These data sets contain, for example, information about what products a person bought the last month or what stores that person visited. However, this information remains unusable until somebody analyzes it by looking for trends. The companies can then use the analysis to improve their business [1], [2]. Traditionally, computer scientists were in charge of this labor, but nowadays, analytics has become so popular that engineers and people from different disciplines are also participating in this practice [3]. Thus, we need to prepare engineering students for this new demand. While most research in this field has focused on how to apply analytics strategies to address problems in different areas [4]–[6], the students' learning and inquiry process behind the practice of data analytics is still not totally understood. If we want to prepare our future engineering workforce for data analytics demands, we need to understand better how engineering students learn data analysis skills, such as identifying useful information from large data sets and translating this information in real

proposals for companies. Consequently, for this work, we aim to answer the research question: *How do first-year engineering students use a large data set to ask questions focused on the client's needs?*

Literature Review

Data analytics (or just "analytics") has gained popularity among companies, but it still lacks a standard definition among the data related disciplines. Since the pioneering report "Competing on Analytics" [1], Google searches, and usage of the term "analytics" have grown intensely [7]. This spread of analytics has generated an extended debate on the definition and characteristics of analytics. While Davenport and Harris [8] define analytics as a set of practices to drive decisions and actions, Rose [7] sees it as a term that groups data science and operation research and Keenan, Owen, and Schumacher [9] consider it as a process by which companies make better decisions through analyzing data. In this sense, Hassan [10] provides a rich discussion about the definition of analytics and the relationship of analytics with other disciplines. Despite the differences, most authors seem to agree about the importance of decision-making, technology, and data as critical elements of analytics [11]; thus, we would want our first-year engineering students to grasp the bases of these elements to start to understand analytics.

As in engineering, questions are fundamental for data science teams, because they lead the team to new findings [3]. Thus, the type of questions that students aim to answer with the given data impacts the quality of their proposed solutions. The questions that engineers ask when working on complex problems frame their work on that problem. In the same way, the quality and complexity of the questions that students generated around data may impact their forward analysis positively or negatively. For example, higher-level, more complex questions are positively correlated to a team's design success [12]. When engineers generate good guiding questions to frame a problem, to understand data, or to gather information for a client, their questions must be framed to give the most useful information to make decisions for the given situation [3].

The questioning process gets more complicated when working in teams because the team needs to coordinate and communicate effectively throughout the question generation process and balance multiple viewpoints about different aspects of the situation. Research shows that the teams' ideas about what they need to learn (their knowledge needs) are a factor that could affect their question generation process. For instance, "when novice designers are unaware of their knowledge needs, they are subsequently unable to ask questions or to employ a clear design strategy that is capable of learning to the pertinent information sharing" [13, p. 2]. In other words, students should reflect on what they need to know about the task (i.e., data analytics to provide beneficial solutions to the client) before "jumping into" questions generation. This reflective process could help novice engineers to become aware of what they need to know about the client's needs to formulate better-designed questions to provide more valuable outcomes from their data analysis.

The activity the students engaged within this study included all five elements of statistical thinking identified by Wild and Pfannkuch [14]: using data, requiring contextual knowledge

about the data, attention to variation, using modeling tools, and opportunities to discover new things about the data when it is represented differently, what they call *transnumeration*. There are many benefits of working with real data, as the students did in this study. These benefits include learning to deal with different types of data, defining categories of classification, and applying ideas such as sampling and distributions of variables [15], [16]. Moreover, as McKinney Jr. and Niese [17] recommended, this data allows students to reflect on the limitations and assumptions about the process of analytics. The assignment examined in this study incorporated ideas suggested by Shaughnessy [18] with a structure designed to help students overcome some of the major conceptual challenges students have with statistics. These include, for example, building on the intuitive understandings that students have about center and variability, transforming their knowledge to incorporate the different types of averages and variabilities, and letting students engage directly in obtaining the data sample to see relationships between the different samples.

Theoretical Framework

The theoretical framework that guided our analysis is based on the data analytics cycle described by Nelson [19], specifically the Analytics Lifecycle, which Figure 1 displays. The process of data analytics "helps us to know the truth and understand the context" [18, p. 131] of the problem we are trying to understand. Analytics can be used to "solve a problem...support a narrative...understand a phenomenon…[and/or] discover something new" [18, p. 133]. Nelson [19] uses the Analytics Lifecycle to break down the activities that generally occur in this process, activities that we hoped the students in this study would engage. Those activities or practices include: problem framing, understanding and exploring or data sensemaking, developing an analytic model, and interpreting, explaining, and activating results. Within this process, the activities are further broken down into the tasks of *define*, *identify*, *explore*, *analyze*, *present*, and *operationalize*. In the class project, students engaged in all these tasks except operationalize, which would require managing the implementation of the solution over some time. For the present study, we focus our analysis on the process of question design from the task *define*.

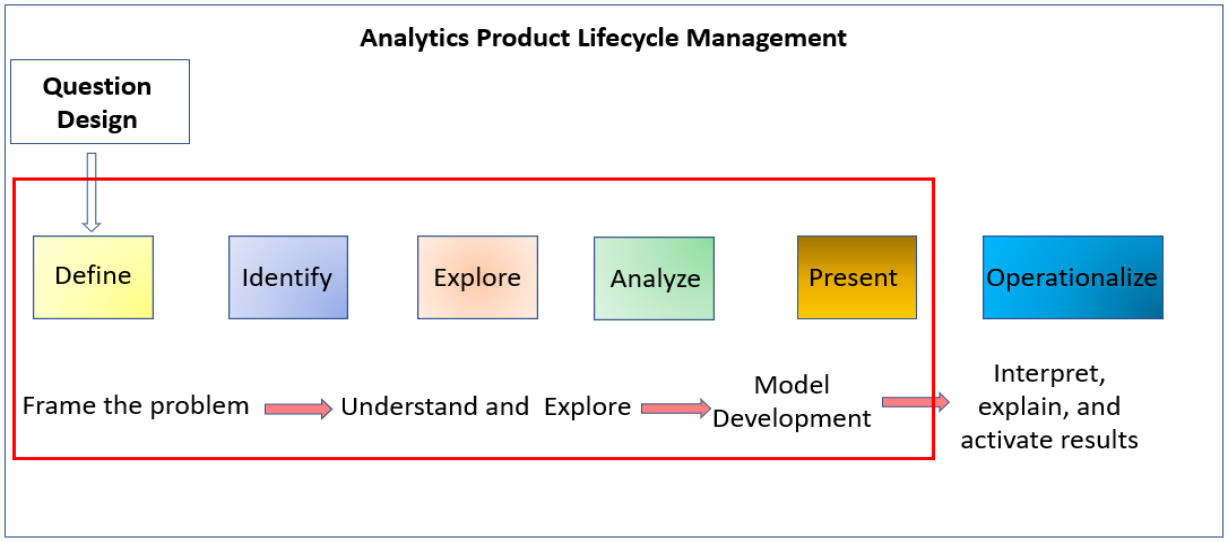

Figure 1. Data Analytics Lifecycle (adapted from [19])

Methods

This study focused on how first-year engineering students asked questions as part of their data analytics process using a large data set. We aimed to answer the research question: How do first-year engineering students use a large data set to ask questions focused on the client's needs?

Participants and Context

This study was conducted in a large public research university in the Midwest region of the United States with a total enrollment of approximately 31,000 undergraduate students, a campus size of around 2,500 acres, and an acceptance rate of 57%. The university has a distribution of 43% female students, 24% minority domestic students, and the most pursued majors are engineering and business-related majors.

The participants of this study were 53 first-year engineering students who were part of a data science-focused learning community. The students self-selected into the learning community. This learning community was created in partnership with engineering, English, and library science professors. The students took introductory engineering, English, and weekly seminars together, in addition to their other courses. During the semester, they engaged in data analytics, modeling, and engineering design problems and projects. This study focused on one of those problems, "the bike-share problem," which is described in the next section.

The Bike-share Problem

The instructor, in collaboration with the research team, wrote the bike-share problem. In fall 2018, the 53 students worked in 14 teams of three or four to complete the assignment. Before this assignment, they had worked with these same teams on several other problems focused around learning data analysis and statistics; however, all of their previous assignments had been much more structured and used small data sets. In previous assignments, the students had practiced using Excel to perform descriptive statistics, use several types of graphs for data visualization, and perform simple statistical tests.

For the bike share problem, the instructor asked students to help the company Capital Bikeshare to understand the usage of the shared bicycles and, based on their analysis of the data downloaded from Capital Bikeshare website, propose an improvement in the bike-share service for their enterprise. The instructor provided them with written instructions as a scaffold of the problem-solving process. Namely, the students had to follow the following steps:

1. First, the students were asked to explore bike share programs, including considering the benefits and drawbacks and the cost structure.
2. Next, the students downloaded the data set as an Excel spreadsheet directly from the company's website (this data set is freely and readily available). In each team, one student downloaded one quarter of the year. Each data set contained between 500,000 and 1.3 million data points, for a total of around three million data points per team. Each spreadsheet contained nine columns of information about the behavior of Capital

Bikeshare users (See Figure 2). For each ride, there was data for: The duration of the ride in minutes (duration), the start day and hour (start date and start time), the end day and hour (end date and end time), the code for the start station, the name of the start station (start station), the code for the end station, the name of the end station (end station), the reference number of the used bike (bike number), and casual or member user (type of user).

3. Next, the students were asked to individually write potential questions to explore the data and play with the data.
4. Then, the students met with their team, discussed their initial findings, and identified a team research question to help Capital Bikeshare improve its bicycle-sharing enterprise.
5. The students then worked as a team to answer the team research question by analyzing the data with Excel. They were explicitly asked to utilize pivot tables in their analysis, which was a new topic in the class.
6. Finally, as a team, they were asked to state a proposal, based on the analysis of the results, to help Capital Bikeshare improve their business.

	A	B	C	D	E	F	G	H	I
1	Duration	Start date	End date	Start station number	Start station	End station number	End station	Bike number	Member type
2	221	1/1/2017 0:00	1/1/2017 0:04	31634	3rd & Tingey St SE	31208	M St & Ne	W00869	Member
3	1676	1/1/2017 0:06	1/1/2017 0:34	31258	Lincoln Memorial	31270	8th & D St	W00894	Casual
4	1356	1/1/2017 0:07	1/1/2017 0:29	31289	Henry Bacon Dr & Lincoln I	31222	New York	W21945	Casual
5	1327	1/1/2017 0:07	1/1/2017 0:29	31289	Henry Bacon Dr & Lincoln I	31222	New York	W20012	Casual
6	1636	1/1/2017 0:07	1/1/2017 0:34	31258	Lincoln Memorial	31270	8th & D St	W22786	Casual

Figure 2. Screenshot of the Capital Bikeshare data from one student's Excel file.

Data Collection and Analysis

We collected all the students' artifacts related to the Bike-share problem, which consisted of students' description of the problem, students' individual and team questions, students' results of their team questions (including data summaries and graphs), and students' written proposals for the client based on all the information. Initially, we explored the artifacts to understand better how each team performed the Data Analytics Lifecycle, and then we focused on the problem framing practice. Particularly, we studied the students' design questions task in order to address our research question. The following paragraphs described how we analyzed the students' questions and the codebook used to categorize them.

The data analysis was conducted in several steps. First, we explored all the students' reports and Excel files individually. Then, we organized the data into a single spreadsheet that included each of the individual and team questions, along with shortened descriptions of each team's proposal. Later, we worked individually and met several times as a whole group to develop the codebook represented in figure 3. Following this codebook, we classified the individual and team questions based on their answerability. Answerable questions are the ones that students can answer with the provided data from Capital Bikeshare or with additional data that they could obtain quickly. Not answerable questions are questions that students could not answer because they are

ambiguous, open, or because they would require additional studies to collect more information. In order to code the answerable questions, we started by classifying the nine columns of the spreadsheet (figure 2) into five variables according to the similarities in the information (the students split the column start date and end date to get the information of start time and end time respectively):

1. User: Type of user (member or casual user)
2. Time: Duration, start time, end time.
3. Date: Start date, end date.
4. Station: Start station, end station, start station number, end station number.
5. Bike number: Bike number.

Then, we categorized the answerable questions in terms of the number of variables needed into simple (one data variable), middle (2 data variables), and complex (3 or more data variables and/or data from outside the data set). Figure three represents the codebook with examples from our data for each category of questions.

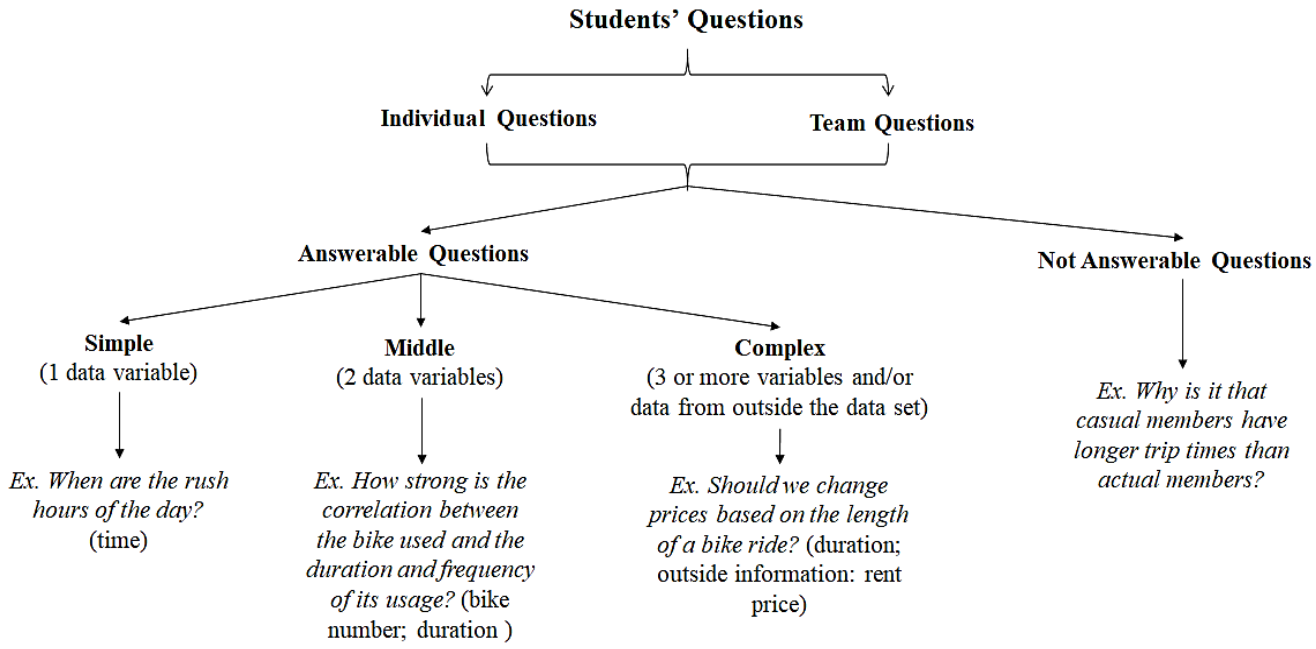

Figure 3. The codebook used to categorize the students' questions when doing analytics.

Limitations

This study focused on student work from a single class of 53 students during their first semester as undergraduate engineering students. Therefore, these results are limited by the small sample size. Although the students had the opportunity of performing previous assignments related to big data analytics, we only used data from the Bike-share Problem. This is not a large limitation

because, by addressing this problem, the students worked through almost the entire Data Analytics Lifecycle. Consequently, we could focus on their questioning process without losing the big picture of the entire cycle. Finally, we are limited by the information of the students' artifacts. Additional factors, such as their teamwork skills, could have influenced their design questions practice.

Results

The purpose of this study was to investigate how first-year undergraduate engineering students developed questions using a large data set. The students performed the problem framing and data sensemaking practices of the Analytics Lifecycle almost simultaneously. In order to frame the problem, the first-year engineering students tried to understand better the Capital Bikeshare enterprise by looking into the company's website and magazine articles. Once they had a general idea about the company business, they started the data sensemaking by downloading and exploring the large data set individually. After that, the students came up with individual and team questions by integrating their initial understanding of the problem and data exploration. Aiming to characterize the students' usage of the Capital Bikeshare large data set, the following paragraphs describe their initial approach to exploring the data, their individual data-exploratory questions, and their team client-needs questions.

Preliminary Data Exploration

The students followed different strategies to explore the data and reach sensemaking. The assignment prompt asked students to explore the data individually by using Excel and then to propose a question to be discussed with their teammates. However, only Teams G and K followed the exact instructions (i.e., explore the data individually and then formulate one or more questions); the rest of the teams explored the data in three alternative ways. First, Team A did not present evidence of their data exploration; instead, they enlisted their individual questions and described possible problems of the company according to their personal experience. Second, Teams F, H, L, M, and P; in addition to exploring the data and formulating questions, they also provided general findings from their individual questions. Finally, the rest of the teams did the same as the previously described teams, but they also answered each question carefully and even added some personal opinions about their findings. Many of the students jumped into proposing solutions without spending the expected time on the problem framing. They may not have recognized the importance of discussing the possible problems of the company through the questions.

Table 1. The first strategies followed by the teams to generate individual questions.

Team strategy	**Teams**
Explore the data and post individual questions (as proposed).	G, K
Not show the exploration of the data and post the individual questions.	A

Explore the data, post the individual questions, and provide general findings from all the questions.	F, H, L, M, P
Explore the data, post the individual questions, and fully answer and analyze each of the questions.	B, C, D, J, N, E,

Individual Data-exploratory Questions

The students proposed 92 individual questions that we analyzed to understand how they used the variables of the data set. As we previously explained in the methods section, the large data set of Capital Bikeshare had seven variables (start station, end station, user type, date, time, and bike number) that students could explore to identify a client's need. The students used these variables in many forms, as the paragraphs below describe.

Answerable Questions

The students asked questions of varying complexity in terms of the number of data types needed. These various levels of questions approached different levels of consideration of the client's needs and different levels of incorporation of the different types of data. Students who considered multiple types of data were better able to develop questions that considered a broader scope of the problem. For example, one member of Team D asked a simple question that only needed one variable, *time,* and asked, "What is the average time riding a bike?" and a member of Team G asked the question, "Do members or casuals use the bikes more often?", which only required the use of one variable, *member type.* Most of the students (≈59%) proposed simple questions like these that used only one variable from the data set. In the cases of these simple questions, the students may have restricted their problem framing process by analyzing only the specific variable. In contrast, a more complex question was asked by one student of Team B, which needed both of the variables *bike number* and *time*, "How strong is the correlation between the bike used and the duration and frequency of its usage?" By asking this complex question, the student used more variables, which allowed him to have a richer picture of the Capital Bikeshare enterprise and their needs. However, most of the students missed the learning opportunity to develop complex questions using multiple data types.

The students' individual questions used some data types for their variables more than others. Most of the questions used the data that was coded into the *time* category (42%) or *station* category (31%). For example, a student from Team N asked a question focusing on data from the *time* category, "At which point of day do the users of Capital Bikeshare use the share bike least frequently?" and a student from Team P asked a question focusing on data from the *station* category and user *type* category, "Which member types are common at each starting and ending stations?" In contrast, the students scarcely used the variable bike number (5%). For instance, a student from Team D proposed one of the few questions with this variable: "[Which are the] most used bikes?" To answer this question, the student would need to count how many rides each bike had based on its identification number. By overlooking some variables, the students may have missed information from the data set that could potentially help or reveal potential problems of the Capital Bikeshare enterprise.

Not Answerable Questions

Some students looked for another kind of information from outside sources (i.e., Capital Bikeshare enterprise website, other companies that provide the same services in other cities, countries, etc.) when exploring the data to understand the problem. This initial step focused on asking questions individually to explore the data; however, some of the questions asked for information outside of the data. 38% of the questions asked for information that was not in the data set. For example, one team member from Team D asked the question: "Should we change prices based on length of bike ride? Another one, from Team F, said: "How does the revenue from casual riders compare to the revenue from members?" Although some of these questions asked for information that students could obtain, such as the ride price, most of them were not answerable with the provided data since they required information that was not available within the timeframe of the assignment. For instance, one member of Team G asked the question: "Do certain streets have a higher probability of attracting potential renters?" This question is connected with the potential client's needs. However, it is not answerable because students would need to perform additional studies to determine how to attract the new renters, and that is outside of the project's scope. In summary, students seem to practice perspective-taking in order to conceive ways that might help improve the Capital Bikeshare enterprise. However, this percentage of students lost focus on what the project was asking them to do and ended up asking not answerable questions.

Communication and Additional Issues

We observed that students formulated questions with ambiguous words. Many of their questions used vague terms such as "bike usage" and "heaviest traffic flow" that could have referred to a variety of data types, indicating that students struggled to clearly articulate their ideas or understand the importance of precise questions. Additionally, there were some differences in terms of the number and type of questions each student asked. While some students provided one question, others proposed more than three questions from the data. For example, one student proposes seven questions, but almost all of them were either ambiguous or not answerable. Although this student was creative by proposing many questions, he did not take into account the goal of the project, which was to help the company by using the provided data.

Team Client-needs Questions

The assignment prompt asked the members of each group to discuss their individual questions and coming up with a team question that could help Capital Bikeshare improve its enterprise. The students integrated their initial individual reasoning in different ways from starting with totally new questions to just choosing one of their previous individual questions. Although there were 14 teams, we analyzed 13 team questions since team A did not propose a question explicitly. The following paragraphs characterize the students' reasoning to formulate the team questions, the questions' structure, and the similarities and differences with the individual questions.

Most of the students' teams designed questions based on problems they identified or opportunities to improve the Capital Bikeshare enterprise. The teams identified problems such as there are not enough bikes available at the busiest stations (Team B), there are not popular enough stations (Team J, and K), and the reduction of the demand for bikes during cold months (Team C). Some teams aimed to improve the enterprise by gathering more users and profit (Team N), optimizing the company resources during the peak hours or at the busiest stations (Teams H, G, M, and P), taking advantage of how the demand changed during the day (Team F). The teams D, E, and L did not state explicitly in their reports why they proposed their client-needs questions. In general, most of the teams considered the client-needs to propose the questions during this problem framing and data sensemaking. As a result, the students came up with the questions listed in Table 2; some of them were extensions of the individual questions (Team D, E, N, F), and the others were new questions.

Table 2. Team client-needs questions proposed by the teams to improve the Capital Bikeshare enterprise.

Team	Question
A	Did not propose a team question.
B	Which start station has the highest traffic based on the count function in Excel?
C	How can Capital Bikeshare promote the use of their bikes in colder seasons, and how can they accommodate for demand in warmer seasons?
D	Where are the most popular destinations that people arrive at using city bikes?
E	What is the most popular station? Why that station is so popular?
F	Will changing the pricing plan for Capital Bikeshare for casual riders to a demand-based price per minute model increase their revenue compared to their current model?
G	What the busiest stations are, when are they the busiest, and who uses them the most?
H	What hub is the most common station in Washington D.C. and how can the company optimize that location?
J	Which bicycle stations are most popular throughout the course of 2017? In other words, which stations possess the most rentals, and which stations possess the most returns?
K	When, where, and at what time do their employees have to be the most active with supplying bikes to the docking station?
L	How can we encourage casual riders to in the winter months and member riders in the summer months?

M	Which locations are the most popular for Capital Bikeshare? What bike stations receive the most traffic (users) according to location? Do the popular locations change according to the time of year (i.e. summer vs. winter)?
N	How does the usage of bikes differ between members and casual users?
P	A positive skew is evident when making a histogram of the frequency of number of bikes and the average duration. Can we add more stations to stations at the far end of the spectrum (3-4 standard deviations above the mean) to minimize traffic and allow for more use of those highly used areas?

The team questions usually were more complex that individual questions in terms of the number of employed variables, but both types of questions tended to ask for descriptive information from the data. While many students used one or two variables in the individual question, many team questions used two or more variables (Teams G, K, L, and M). For example, Team G proposed a team question that groups three different individual questions – What are the busiest stations? When is the station busiest? and Who does use the stations the most? – in one question that uses three variables: User, Date, and Time. Despite these differences, most of the team questions were similar to the individual questions, as both types asked for descriptive information from the data set. For instance, Team E was still asking for what the most popular station is, in the same way, that most students did individually. It may indicate that students tried to balance proposing meaningful questions related to the client's needs by proposing questions that they can answer using pivot tables and their statistics knowledge.

Some teams tried to propose questions that asked for no descriptive information, but they ended up being not answerable with the available data. Many of the team questions posed ideas that were not answerable because they require information that the students did not have. For example, Team E asked why the most popular station is so popular, or Team C asked how Capital Bikeshare can promote the use of their bikes in colder seasons. To address these questions, students would need additional qualitative information about what was outside of the scope of the project. In contrast, the Team F question could be answerable; however, the students would need to generate mathematical models and run simulations in order to identify the effect of changing the pricing plan. As they are first-year students, they would not have the skills and knowledge to do these processes. Although students wanted to propose different types of questions, it seems their skills and previous knowledge limit them.

Even though they spent time explicitly brainstorming questions first individually and then as a team, most of the teams ended up asking the same question, although they phrased it differently. The question was: What are the most popular stations? This question was formulated in different ways by the Teams B, D, E, G, H, J, K, and M. In contrast, the Teams N, L, P, and C proposed different questions, but they were ambiguous or not answerable with the available data. The most different question in terms of content was proposed by Team F, who asked about changing the pricing system of Capital Bikeshare (see Table 2). Although they tried to answer this question throughout the project, its solution would require statistical tests that were outside of the students' expertise.

Discussion

The purpose of this study was to investigate the research question, how do first-year engineering students use a large data set to develop questions focused on the client's needs? Overall, although they struggled with some practices of the Data Analytics Lifecycle [19], the students in this sample were able to take an extensive data set (~3 million data points) and frame a problem until designing a question that considers the client's perspective. Moreover, as they were designing the questions, they performed tasks not only from the problem framing practice but also from the data sensemaking practice. For example, some teams thoroughly explored the data individually before proposing their team research question (see Table 1). As Nelson [19] describes, analytics is not a rigid, sequential set of steps; instead, as we evidenced with these students, it is a fluidly transition back and forth throughout the practices of the Data Analytics Lifecycle. By analyzing our data, three themes emerged as areas that characterized the students' usage of the data set to develop questions: *Students' Inquiry in Data Analytics, Students' Data Fixation,* and Students' *Statistics Knowledge to Perform Analytics.*

Students' Inquiry in Data Analytics

The students' questions showed the many difficulties associated with the inquiry process during the data-driven analysis. Although all the teams were able to perform the analysis to some extent, their questions varied notably. Some students proposed answerable simple and complex questions that integrated variables from the data and considered the clients' perspective. Examples of these are presented in the Answerable Question section in the results. However, many students designed unanswerable questions that required not provided information to be answered or used vague terms. The following paragraphs will discuss these two aspects of the unanswerable questions.

Questions that Required more Information

Although the questions that required additional information to be answered could not be used in the Bikeshare problem, those were often open questions that could have helped students to understand the problem better. Rose [3] suggests that analytics teams should have a balance between open and closed questions to keep the team focus on both the big picture and details of the problem. Moreover, she explains that open questions promote discussions inside analytics teams, which could contribute to getting more exciting insights about the problem. In our study, even though students could not use the individual open questions to directly analyze the data -which was the initial objective of the activity-, these questions could have aided students' inquiry reasoning. By formulating those individual open questions, students may have ended up proposing the more closed and answerable team questions that frame the problem. Further research needs to be done on how open questions could aid the students' thinking when performing data analytics.

Questions that Used Vague Terms

Many individuals and teams struggled to communicate their questions clearly and to connect their questions to the problems faced by the company -their client-, resulting in ambiguous

questions. Examples of these are presented in the unanswerable questions section in the results. Students often developed questions that were unclear, potentially to both themselves, their teammates, and outside observers. Moreover, they struggled to formulate questions that can be directly addressed with the given data. Glancy, Moore, Guzey, and Smith [20] found among younger students (5th-grade students) there was "a mismatch between the type of data the students collected and the data they used to draw conclusions or make decisions." [19, p. 74]. Although the conclusions drawn by the students in this study were based on the data they had, their questions, at the beginning of the data analytics cycle, were often not based on that data. This mismatch suggests that although the first-year undergraduate students were able to use their data in their decisions eventually, unlike the younger students, many of them poorly align the data with their initial questions. Therefore, students were not able to evaluate if their question was answerable with the provided data, as well as assess if the data supported the obtained answers. In general, the students' inquiry could be enhanced by more direct instruction about how to formulate questions before engaging in the next stages of the Data Analytics Lifecycle.

Students' Data Fixation

The students showed fixation on a small amount of the data when designing the questions. Most of the students' questions focused on only certain types of data, and they did not use the entire dataset. Nelson characterizes the problem framing practice of Data Analytics Lifecycle as a design activity where analytics teams use divergent and convergent thinking to explore the problem space [19]. In the Bike-share Problem, the project aimed to promote the students' divergent thinking by asking them to propose several individual questions and their convergent thinking by asking them to negotiate and define one team question. We would expect this activity to facilitate creativity and ideation among the students; however, as described in the Results section, many of the team questions focused around the same variable and idea: counting the most popular stations. This analysis, with only one type of data, did not provide the students with meaningful information from their analysis to give to the clients, which is what they were tasked to do. Future studies could identify teaching practices that help students overcome the fixation on some variables of the data.

Students' Statistics Knowledge to Perform Analytics

During the data analysis process, students were limited by their specific skills, particularly in their limited knowledge of statistics. Additionally, students were required to utilize certain functions in Excel, namely pivot tables, that may have limited the scope of their analyses. Additional knowledge of statistics would allow them to formulate and address more complex types of questions. For example, additional knowledge of inferential tests or correlations could have prompted students to think about different questions. However, even with their limited skill set, they were able to engage with the data analytics cycle and may be able to transfer the knowledge of that process to future problems as they improve their data analysis skillset. Hjalmarson, Moore, and Delmas [16] suggested that tasks for first-year engineering students could be focused on designing methods to answer particular questions that would "provide additional practice with applying concepts" [16, p. 30] in statistics. This study provided a different context to show that students were indeed able to apply statistics concepts when answering questions about data.

Conclusions and Implications

In this study, we examined how first-year engineering students used a large data set to ask questions focused on a client's needs. We proposed a codebook to analyze the students' questions and, overall, we found that the students were able to work through the Data Analytics Lifecycle [19]. Many students formulated a variety of questions that were based on the data and considered the client's needs. However, some students ended up writing questions that were not clear due to ambiguous terms, so simple as not to be useful to the client, or too complex to be answerable with the date and time they had. Nevertheless, some of these questions could have aided their thinking to develop the more meaningful questions. Learners need to practice writing questions and considering the constraints of the data and how that affects their data analytics processes. Instructors can support the students' question writing process by introducing case studies where other analytics projects use questions successfully to identify and answer the client's needs.

The students' questions often focused on a single, simple variable of the data and not utilizing the breadth of the data available to them. They should spend more time understanding the data before jumping into its analysis. Particularly, they could be encouraged to think about the opportunities that each variable provides as a source of information for the client. More research needs to be done about what teaching practices help students overcome the fixation on some variables of the data.

The students' previous knowledge of statistics could have mediated their question design practice and limited their ability to answer their questions. Although the students were able to apply some of the statistical skills they had learned in the course, they did not have the skills to address some of the questions they were interested in. Similar activities, where students learn and practice statistics, could be integrated into the curriculum at the pre-college and college level. The Bike-share problem is an example of how analytics and statistics can be integrated early in engineering education successfully. In the future, we plan to implement the suggestions described in the previous paragraphs and continue to explore ways in which first-year engineering students approach questioning and data analytics.

References

[1] T. H. Davenport, D. Cohen, and A. Jacobson, "Competing on analytics," Babson Park, MA, 2005.

[2] McKinsey Global Institute, *The age of analytics: Competing in a data-driven world*. McKinsey & Company, 2016.

[3] D. Rose, *Data science: Create teams that ask the right questions and deliver real value*. Atlanta, GA: Apress, 2016.

[4] C. Min, C. Roman, and S. Trevor, "Big data analytics in financial statements audits," *Account. Horizons*, vol. 19, no. 2, pp. 423–429, 2015.

[5] K. Govindan, T. C. E. Cheng, N. Mishra, and N. Shukla, "Big data analytics and applications for logistics and supply chain management," *Transp. Res. Part E Logist. Transp. Rev.*, vol. 114, pp. 343–349, 2018.

[6] W. Raghupathi and V. Raghupathi, “Big data analytics in healthcare: Promise and potential,” *Heal. Inf. Sci. Syst.*, vol. 2, no. 3, pp. 1–10, 2014.
[7] R. Rose, “Defining Analytics: A conceptual framework,” *ORMS Today*, vol. 43, no. 3, pp. 34–38, 2016.
[8] T. H. Davenport and J. G. Harris, *Competing on analytics: The new science of winning*. Boston, MA: Harvard Business Review Press, 2007.
[9] P. T. Keenan, J. H. Owen, and K. Schumacher, “Introduction to analytics,” in *Informs analytics body of knowledge*, J. J. Cochran, Ed. Hoboken, NJ: John Wiley and Sons, 2019, pp. 1–28.
[10] N. R. Hassan, “The origins of business analytics and implications for the information systems field,” *J. Bus. Anal.*, vol. 2, no. 2, pp. 118–133, 2019.
[11] I.-Y. Song and Y. Zhu, “Big data and data science: what should we teach?,” *Expert Syst.*, vol. 33, no. 4, pp. 364–373, 2016.
[12] O. Eris, *Effective inquiry for innovative engineering design*. New York, NY: Springer Science+Business Media, 2004.
[13] E. S. Fleming and A. E. Coso, “Viewing an interdisciplinary human-centered design course as a multiteam system: Perspectives and information sharing,” in *Design Thinking Research Symposium 2014*, 2014, pp. 1–18.
[14] C. J. Wild and M. Pfannkuch, “Statistical thinking in empirical enquiry,” *Int. Stat. Rev.*, vol. 67, no. 3, pp. 223–265, 1999.
[15] J. Hall, “Engaging teachers and students with real data: Benefits and challenges,” in *Teaching statistics in school mathematics-challenges for teaching and teacher education A joint ICMI/IASE study: The 18th ICMI study*, C. Batanero, G. Burrill, and C. Reading, Eds. New York, NY: Springer Science+Business Media, 2011, pp. 335–346.
[16] M. A. Hjalmarson, T. J. Moore, and R. Delmas, “Statistical analysis when the data is an image: Eliciting student thinking about sampling and variability,” *Stat. Educ. Res. J.*, vol. 10, no. 1, pp. 15–34, 2011.
[17] E. H. McKinney Jr. and B. D. Niese, “Big data critical thinking skills for analysts—Learning to ask the right questions,” in *Twenty-second Americas Conference on Information Systems*, 2016, pp. 1–8.
[18] M. J. Shaughnessy, “Research on students’ understanding of some big concepts in statistics,” in *Thinking and reasoning with data and chance*, G. Burrill and P. C. Elliott, Eds. Reston, VA: National Council of Teachers of Mathematics, 2006.
[19] G. S. Nelson, *The analytics lifecycle toolkit: A practical guide for an effective analytics capability*. Hoboken, NJ: John Wiley and Sons, 2018.
[20] A. W. Glancy, T. J. Moore, S. Guzey, and K. A. Smith, “Students’ successes and challenges applying data analysis and measurement skills in a fifth-grade integrated STEM unit,” *J. Pre-College Eng. Educ. Res.*, vol. 7, no. 1, pp. 68–75, 2017.