International Journal of Engineering Education Vol. 39, No. 1, pp. 55–73, 2023
Printed in Great Britain

0949-149X/91 $3.00+0.00

First-Year Engineering Students' Ideation in Data Analytics*

RUBEN D. LOPEZ-PARRA
Purdue University-Main Campus, College of Engineering, ARMS 1300, 701 W. Stadium Ave., West Lafayette, IN 47907, USA.
E-mail: rlopezpa@purdue.edu

ARISTIDES P. CARRILLO-FERNANDEZ
Purdue University-Main Campus, College of Engineering, ARMS 1300, 701 W. Stadium Ave., West Lafayette, IN 47907, USA.
E-mail: carril11@purdue.edu

AMANDA C. EMBERLEY
California Polytechnic State University, San Luis Obispo, college of engineering, 1 Grand Ave. San Luis Obispo, CA 93407, USA.
E-mail: acjohnst@calpoly.edu

TAMARA J. MOORE
Purdue University-Main Campus, College of Engineering, ARMS 1300, 701 W. Stadium Ave., West Lafayette, IN 47907, USA.
E-mail: tamara@purdue.edu

SEAN P. BROPHY
Purdue University-Main Campus, College of Engineering, ARMS 1300, 701 W. Stadium Ave., West Lafayette, IN 47907, USA.
E-mail: sbrophy@purdue.edu

Big data analytics has grown as a valuable tool for professionals from different fields to get insights from large volumes of data and make data-driven decisions. Given this, engineering students need access to learning environments that support learning data analytics. These activities should teach students the skills to assess data, design high-quality questions, perform data analysis, and provide recommendations in a manner that is aligned with client needs. To this end, we developed and implemented the data analytics activity called "The Bike-share problem" for a First-Year Engineering (FYE) design and modeling course. To analyze the students' ideation of questions and recommendations when working on the activity, our summarized research question is: *What are the characteristics of FYE students' proposed questions and recommendations for a client as part of their data analytics project?* We analyzed questions and recommendations from teams' final reports using qualitative content analysis. Our findings show that the students' questions ranged from superficial treatments of the data that required simple analyses to deep explorations of the problem that required more complex analyses. For the recommendations, we found that model responses include considerable detail, support with data, and justification based on the client needs. While both the questions and the recommendations were important separately, we also found differences among teams' ability to align their recommendations to the client with the actual questions they were trying to answer. The differences in student responses to the activity can have many explanations as to the cause; however, we have evidence that perhaps scaffolding in the way the activity is posed and team dynamics may have affected how students responded to the activity. Finally, we provide some effective practices that interested readers may implement to design analytics activities that promote students' ideation.

Keywords: first-year curriculum; professional skills; data literacy; creativity; ideation

1. Introduction

Big data analytics has grown as a useful tool for professionals from different fields to get insights from large volumes of data and make data-driven decisions. From improving customers' service to optimizing energy distribution in electric grids, analytics represents a new world of opportunities for engineers. For this study, analytics is seen as an approach to support a data-driven process of discovery that includes activities such as framing an analytic problem, developing a model, and implementing action plans for companies [1]. Considering the relevance of big data analytics, engineering faculty need to design learning environments that allow students to develop the knowledge, skills, and abilities to perform analytics.

Analytics is more than getting a big data set and running statistical models. It requires the statistical and technical knowledge to make sense of the data and the intuition and creativity to identify problems and recognize meaningful insights for a client [2]. Specifically, there are two critical tasks in analytics: design high-quality questions that guide the data analysis and provide recommendations in a manner that is aligned with client needs. Although designing questions and making recommendations are essential tasks for the success of analytics, there is

* Accepted 7 September 2022.

a lack of research about how to teach them or how students learn them. To this end, we developed and tested the data analytics activity called "The Bike-Share Problem" for an FYE design and modeling course. Based on the students reports, we performed this study as a first attempt to better understand the engineering thinking of first-year engineering (FYE) students when performing analytics. Specifically, we used the theoretical approach of ideation from Shah and colleagues [3] to analyze how the engineering students' ideation (quantity, variety, novelty, and quality) manifests when asking questions based on a large data set and when making recommendations for a client based on statistical findings. Elucidating these points contributes to fostering our understanding of learning analytics and ideation, which will help engineering faculty and curriculum developers design better learning environments. The specific research questions guiding this inquiry include:

What are the characteristics (quantity, variety, novelty, and quality) of first-year engineering students' proposed questions for data analytics?
What are the characteristics (quantity, variety, novelty, and quality) of first-year engineering students' recommendations for a client as part of their data analytics project?

2. Background Literature

Big data analytics was developed to improve business by making decisions based on large data sets. From the pioneering work of Davenport and colleagues [4] about analytics in companies, big data analytics has grown to include applications in fields such as finance (e.g., [5]), logistics (e.g., [6]), healthcare (e.g., [7]), online education (e.g., [8]), and engineering (e.g., [9]). At the same time that analytics was introduced in many fields, its definition also took many shapes [10–12]. However, all of them seem to agree that analytics implies a decision-making process, collecting big data sets, and using technology to make sense of data and get meaningful insights [11, 12].

Big data analytical methods are operationalized using the Data Analytics Lifecycle (DAL) as a framework. The DAL typically involves framing a problem, exploring a big data set, developing a model to get insights, and operationalizing the insights into recommendations for the client [1, 13]. Analytics teams utilize the DAL to get meaningful insights from a big data set. When a company acquires a big data set, analytic teams need to explore and understand it to determine the utility of the information. Namely, based on their initial understanding of the data and their previous knowledge, the team frames a problem that will be addressed through analytics. They scope the problem into a question that will determine the analytics process. Guided by this question, the team analyzes the data and generates models to get insights about the problem. Finally, they present these insights to the client and operationalize them through recommendations for the company. The DAL activities do not represent a sequence of steps; instead, they are actions that may happen simultaneously and iteratively throughout the process of analytics [1]. This constant iteration and simultaneousness convey many cognitive complexities for novices to master analytics, which we explain below.

2.1 Students' Learning of Analytics

The chaotic nature of analytics represents several challenges for learners to grasp. Analyzing big data sets requires analytics teams to apply the knowledge and characteristic cognition related to statistics. Namely, they apply descriptive and inferential statistics to analyze the data [9] and the five elements of statistical thinking identified by Wild and Pfannkuch [16]: (1) using data, (2) requiring contextual knowledge about the data, (3) attention to variation, (4) using modeling tools, and (5) opportunities to discover new things about the data when it is represented differently. Previous research has identified how younger and older students struggle to apply this statistical knowledge and develop this statistical thinking [17–19]. They identified how students struggled to define clear questions which relate data with the context of analytics and, overall, a need for more integration of statistics in the K-16 curricula.

Analytics is more than technical statistical knowledge and thinking, it requires engineering design thinking and creativity. Analytics and engineering design share similar thinking processes. For example, Nelson (2018) compared analytics with design thinking and suggested that analytic teams applied convergent and divergent thinking to identify a problem, make sense of the data, and recognize meaningful insights. Furthermore, Rachel Woods (2019) urged for a design thinking mindset when performing data science. She argued that analytics scientists needed to perform tasks like problem framing in a similar way to designers. Consequently, students doing analytics may experience similar challenges to those identified in engineering design thinking. For example, students may jump into solving a problem without deeply framing it [18, 19], struggle to ask valuable questions that lead the analytics process [20, 21], struggle to contextualize their design knowledge [22, 23], or experience fixation on one design solution or one

section of the data set [14, 24]. Additionally, learning both engineering design and data analytics require students to work in teams to make decisions and be able to use varied methods, such as sketching, to communicate their ideas [25]. Students must learn to balance the social and leadership aspects of teamwork, as well as balance the creativity of the team members [29–31].

Analytics teams engage in creative thinking as a cognitive skill to look at the data in novel ways and get meaningful insights for the client. Creativity is a complex phenomenon that depends on people's brain, cognition, personality, social environment, and culture [29]. For this study, we scope creativity at the level of creative cognition, which includes processes such as ideation, divergent thinking, conceptual combination, restructuring, visual synthesis, and visualization [33–37]. Mainly, we are interested in how students' ideation manifests when performing data analytics. According to Reid and Moriarty, "ideation is the formation of ideas and involves the conception of original thoughts" [38, p. 119]. Ideation has been analyzed in different contexts such as mathematical problem solving (see [35]), however, it is still unexplored in data analytics. In analytics, teams generate ideas and original thoughts by looking at the data from novel perspectives, which may lead to innovative recommendations for the client [2, 13].

2.2 Ideation in Analytics

The role of people's ideation in analytics has not been fully explored. Some recent efforts have been made to understand the relationship between available data and the generation of ideas. Chen and colleagues [39] proposed an artificial intelligence model that helps the user to get inspired during product design. Namely, they studied how participants used this model to get semantic networks and images as cues to promote the participants' ideation of a new spoon. These semantic and visual stimuli seemed to enhance the quantity, variety, and novelty of their ideas. In addition, some research has been done about teaching data analytics to promote innovation, which is strongly related to ideation. Dinter et al. [40] organized a *morphological box* (i.e., a tool to organize information) that described the solution space and parameters for teaching data-driven innovation. Instructors may use this box to design analytics' assignments that promote innovation by reflecting on the appropriate teaching method according to the course setting and content.

Coming up with ideas to design good questions and make meaningful recommendations are vital tasks for the success of the analytics process. This study focuses on two essential tasks to perform big data analytics where ideation is especially relevant: (1) designing good questions that determine the analytics process and (2) making recommendations for the client based on the data analysis. These two tasks require the analytics team to use technical knowledge, intuition, and creativity [2, 13]. On the one hand, teams start performing analytics by defining a problem and designing a question that will lead the rest of the process. Designing a good question is fundamental to producing innovative results from analytics [1, 13]. However, this is not an easy thinking process. For example, Eris (2004) identified several complexities related to asking good questions in engineering design. She related the quality of designers' questions with their decision-making, problem-solving, and creativity overall. On the other hand, analytics' teams need to develop ideas about recommendations for the client based on their statistical findings, which requires their creativity and experience to match them with the company goals [36, 37]. We need to prepare the future professionals with the abilities to ask meaningful questions and recommendations to perform data analytics.

3. Theoretical Framework

The theoretical framework that guides this study is based on the Data Analytics Lifecycle (DAL) of Nelson [1] and the Shah and colleagues' [3] theoretical approach to ideation. Nelson [1] recognizes analytics as a comprehensive strategy to support a data-driven process of discovery. He describes this strategy in terms of the Data Analytics Lifecycle (Fig. 1). The overall DAL is split into four practices: *frame a problem*, *understand and explore the data*, *develop a model*, and *interpret, explain and activate the results*. These four practices are further divided into six tasks: *define* the problem, *identify*, *explore*, and *analyze* the data, *present* the results, and *operationalize* the results. For this study, the FYE students were engaged in all these tasks of the DAL. However, instead of operationalizing the results, which is outside of the scope of the students' activity, they made recommendations for the client based on their data analysis.

The second part of our theoretical framework is the ideation model proposed by Shah and colleagues [3]. Shah proposes to measure students' ideation during engineering design activities based on four elements: novelty, variety, quality, and quantity [38, 39]. *Novelty* denotes how much the design space is expanded. *Variety* refers to how good the design space is explored. *Quality* states how feasible the idea is based on pre-established criteria. *Quantity* focuses on how many ideas are generated by a person or a team. In contrast to the initial purpose

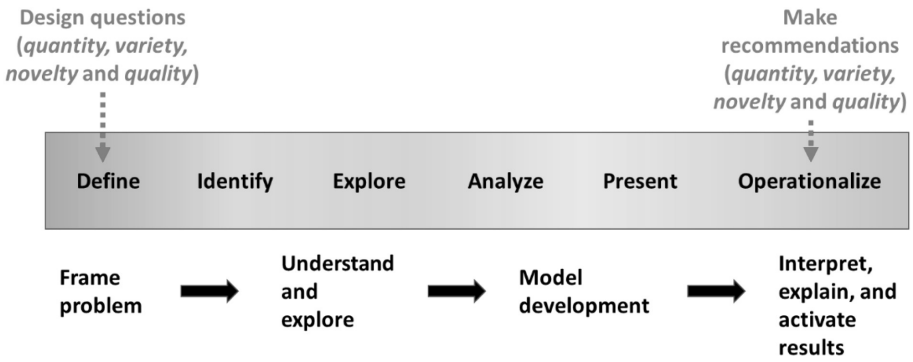

Fig. 1. Data Analytics Lifecycle with the studied elements of ideation (adapted from Nelson (2018)). The tinted font indicates the activities where the students' ideation was explored, and the elements of ideation used to characterize them.

of measuring ideation in engineering design, we will use these four elements to characterize the students' ideation when performing big data analytics. Specifically, we explore the students' ideation of questions that determine the data analysis process and recommendations for a client based on students' statistical analysis. The Data Analysis Plan section will describe how we characterize the four elements of ideation in analytics.

4. Research Design and Methods

4.1 Approach

We used a Qualitative Content Analysis (QCA) approach to characterize the questions students generate when framing an analytics problem and the students' recommendations to improve the client's enterprise. QCA is a "research technique for making replicable and valid inferences from texts (or other meaningful matter) to the context of their use" [44, p. 18]. This technique is characterized by the systematic coding and categorization of large amounts of information to identify patterns and frequencies of the employed words [45–47]. Consequently, QCA allowed us to code and categorize the students' questions and recommendations to recognize their quality, quantity, variability, and novelty. The present research expands on a previously published work about data analytics' learning [14], including additional data and significantly more results and discussion. The following sections describe the study's context, participants, data collection, and analysis. Finally, it presents the applied strategies to promote trustworthiness and the research's limitations.

4.2 Context and Participants

This study was carried out in a large, public, research university located in the Midwest region of the United States. It has a total enrollment of around 31,000 undergraduate students; 43% are female, and 24% are Black, Latinx, and Indigenous domestic students. Of those 31,000, 2,306 students enrolled in FYE; 26% were female, 10% were from traditionally underrepresented minority groups, and 16% were international students. The study's participants were 95 FYE students who were part of a learning community focused on data science in Fall 2019. As a cohort, they took FYE courses and weekly seminars about data science. The students engaged in data analytics, modeling, and engineering design problems as part of their introductory engineering class. The class was guided by an instructor with more than 10 years of experience teaching modeling in engineering and supported by a Ph.D. graduate teaching assistant. The graduate teaching assistant in collaboration with four undergraduate teaching assistants provided feedback during the class time. This study focused on a data analytics problem called "The bike-share problem," which is described below.

4.2.1 The Bike-share Problem

The class instructor and the research team developed the data analytics activity called "The Bike-share problem" for an FYE design and modeling course and implemented it in 19 sections throughout two years. For this study, we used the final version of the activity. This is a data analytics project where students need to analyze user-behavior data from the company CoGo Bike Share to help improve its bicycle-sharing enterprise. These data are freely available through a public internet sharing source. For the course, the 95 students were split into 24 teams of three or four students to work on many aspects of the course, including the bike-share problem. The problem statement was introduced to the students during the first week of the course when they downloaded the data from the company's website (12 Excel files with around 6.3 Mb of information in total). The dataset contained 14 columns of information about users' behavior

that included both numerical (i.e., trip's id and duration; bike id; users' age) and categorical variables (i.e., ride's start and end stations; rides' start and end date and time; user's gender and type).

After downloading the data and during the next seven classes, the teams learned basic statistics (descriptive analysis, data visualization, probability, and regression) and applied it to explore the CoGo Bike Share data using Excel. By the ninth class, the students had to individually propose questions based on their initial data exploration and then meet as a team to negotiate a team question to frame their analytics problem. After that, the students used the large data set to answer their team questions. Based on their analysis, they made recommendations to the company about how this information could be useful for improving the company's enterprise. By working on this project, the students had the opportunity to be engaged in most of the DAL [1].

4.3 Data Collection and Analysis

The students' reports of their analytics process were collected at the end of the course. The 24 reports included the students' descriptions of the problem, their individual and team questions, their statistical analysis, and their recommendations for the client. To better understand the context study, the course slides were also used as an additional data source. Two authors of the paper were the creators of the bike-share problem, as well as the instructor and graduate teaching assistant during the semester of implementation.

We analyzed the 24 teams' reports using QCA [42]. We organized the students' reports into two units of analysis: the individual level and the team level. For the individual information, the coding unit included each one of the 234 questions proposed by the students. For the team information, the coding unit included the 40 teams' questions and the 52 recommendations described in the students' reports. Additionally, the teams were numbered based on the number of individual questions, i.e., Team 1 was the team with the highest number of individual questions and Team 24 was the team with the smallest number of individual questions. Furthermore, each team member was named randomly with the letters A, B, C or D. When we discuss an individual student, we will use their team number and the student letter to refer to them (e.g., Student 4C). In the case of a team, the label will include only the team number (e.g., Team 4). Table 1 includes the details about how the ideation's elements of quantity, novelty, and quality were coded and analyzed for the students' questions and recommendations. The analysis of the variety was more complex; so, it is explained in the next paragraph.

The variety of questions was determined by open coding based on the ideas students wanted to convey. The 274 questions, including individual and team questions, were open-coded by two researchers. First, the two researchers rephrased the first 100 questions to determine the students' main idea and performed open coding simultaneously. Two main categories emerged from this coding: answerability and content. Then, each researcher coded independently the same 50 questions and compared the coding, getting a percentage of agreement of 90%. After that, each researcher coded 50 questions and met to discuss the questions where they had discrepancies. The same procedure was repeated until all questions were coded. The preliminary coding frame was discussed with all the paper's authors to reach a final consensus of the categories for the questions' variety. The coding of all the questions were updated and the totals calculated. The final coding is shown in the Results and Interpretation section.

The recommendations' variety were determined in a similar way as the questions. Two researchers read the 52 recommendations, rephrased, and open-coded them based on the ideas students tried

Table 1. Coding and analysis of the ideation's elements for the questions and recommendations

Ideation's Element	Coding and Analysis of Questions	Coding and Analysis of Recommendations
Quantity	Quantity was determined by frequency: Total number of questions posed by individuals and teams.	Quantity was determined by frequency: Total number of recommendations posed by teams.
Variety	Several rounds of open-coding based on questions' content. See following paragraph for more details.	Several rounds of open-coding based on recommendations' content. See following paragraphs for more details.
Novelty	The least and the most novel questions were based on the counting of the categories for questions' variety.	The least and the most novel recommendations were based on the counting of the categories for recommendations' variety.
Quality	Two researchers open-coded the questions based on the questions' clarity for external readers.	Two researchers code the recommendations based on if they included findings of the data, had some level of elaboration, and justify why the recommendation is relevant for the client.

to convey. The recommendations were grouped into four main initial categories which were further discussed with the rest of the paper's authors. The analysis of the recommendations uses the framework of the Business Model Canvas [43], which includes the categories Key Partners, Key Activities, Key Resources, Value Propositions, Customer Relationships, Customer Segments, and Revenue Streams among others. Through the coding, three main categories for the recommendations emerged: Key Resources, Key Activities, and Customer Relationships. We describe each in the section where they are discussed. The coding of all the recommendations were updated and the totals calculated. The final coding is shown in the Results and Interpretation section.

4.4 Trustworthiness

The reliability and validity are the main criteria to judge the quality or trustworthiness of a QCA [42], [44]. To promote the reliability, two researchers performed the coding and analysis of the questions and recommendations following the procedure described previously to assure consistency. In the case of validity, all the questions and recommendations were included into the coding categories. According to Schreier [47], if most data are in the same category, the coding scheme could have low validity. For this study, the codes for the questions and recommendations were mostly evenly distributed among the categories which supports high validity. Furthermore, three additional researchers with expertise in engineering teaching and learning oversaw and verified the research strategies and implementations employed in the data collection, coding, and analysis to enhance the validity of the study.

4.5 Limitations

We characterized the students' questions and recommendations based only on their written reports; thus, we were limited by how clearly they wrote them. Furthermore, we did not have process data to determine possible teamwork issues, which could have affected the teams' final performances. The findings are based on the results of one implementation of the activity in one class; thus, we invite the reader to consider his or her own context before trying to transfer the findings.

5. Results and Interpretation

The purpose of this study is to investigate the characteristics of students' questions and recommendations for a client when performing the Data Analytics Lifecycle (DAL). This section presents the results and initial interpretation of the quantity, variety, novelty, and quality of the students' questions, and then the recommendations.

5.1 Questions in the Data Analytics Life Cycle

This section describes the results for the quantity, variety, novelty, and quality of both the individual and team questions. We present the results for each of the elements separately and later in the discussion section provide a larger perspective on the integration of the elements.

5.1.1 Questions Quantity

Ideation quantity refers to how many questions a student or team proposes when framing the analytics problem. The students were engaged in two ideation stages: individual ideation and team ideation. During the individual ideation, 84% of the students proposed between one and three questions, and 3% of them showed an abundance of ideas by proposing more than six questions. As Fig. 2 shows, in the teams with more individual questions (teams 1, 2, 3, and 4), most of the team members contributed evenly with questions. For example, each member of Team 1 proposed three, five, seven, and five questions, respectively. Although the ideation of questions was intended to be individual, some teams may have negotiated explicitly or implicitly to come up with a certain number of individual questions.

While most teams (67%) defined one team question to focus on for their client as stated in the instructions, a few teams defined many team questions during the team ideation. The teams that proposed multiple questions proposed between two and six questions (see Fig. 3). For example, the members of Team 20 individually proposed five questions in total, and as a team, they ended up with six questions for the team ideation. When teams ended up with more than one or two questions, the ability of the teams to converge on a specific analytics problem seemed to be diminished. Furthermore, with too many questions, the teams likely needed more time to perform all the statistical analysis and also then more time to develop all of the recommendations that would come from the analyses. Students may need additional support to purposefully select one or two team questions in order to make high-quality recommendations that have detailed analyses rather than surface-level recommendations that cover a wide range of ideas.

5.1.2 Questions Variety

Ideation variety refers to how different the students' questions were when framing the analytics problem. Table 2 summarizes the variety of questions across the pool of students. We identified two main sources of variation in the questions: the questions'

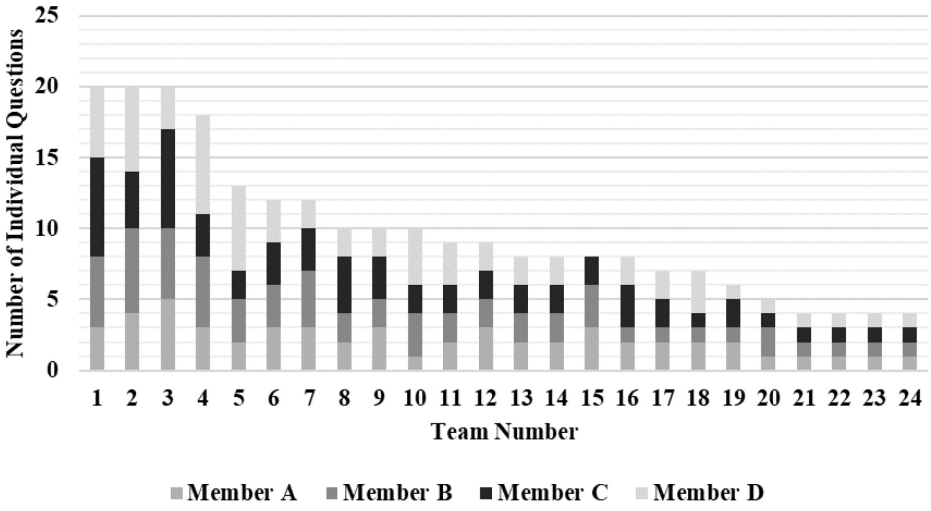

Fig. 2. Number of individual questions proposed by each team member.

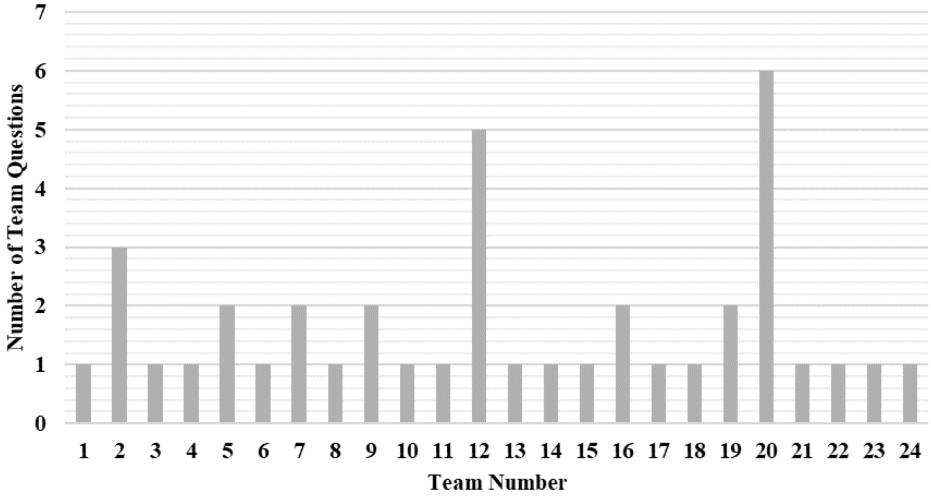

Fig. 3. Number of team questions proposed by each team.

Table 2. Number of students or teams who propose each type of question (We can provide this table in word format upon approval of the manuscript)

Questions' Content		Provided Data		Additional Available Data		No-available Data	
		Individual	Team	Individual	Team	Individual	Team
Descriptive Questions	Frequency	23	2	1	0	0	0
	Maximum - Minimum	30	8	1	0	0	0
	Central Tendency	8	1	0	0	0	0
	Variable Size	1	0	0	0	0	0
	Data Cleaning	1	1	0	1	0	0
Inference Questions	Correlation	28	4	1	0	1	0
	Significant Difference	5	1	0	0	0	0
	Cause-effect	1	0	3	1	0	0
	Investigating patterns	23	1	1	0	1	0
Business Model Questions	Preliminary Recommendations					17	9
	Customer Segments					3	0
	Revenue Streams					2	0
	Key Activities and Resources					9	4

content (Descriptive Questions, Inference Questions, and Business Model Questions), and the questions' answerability (Provided Data, Additional Available Data, and No-available Data). The following paragraphs describe each of the questions' content and their relationship with the answerability.

Descriptive Questions. 68% of the individual students and 54% of the teams proposed questions that aimed to describe data by exploring the frequency distribution of a variable; determining the maximum or minimum of a variable under specific conditions; calculating measures of central tendency; identifying the size of the sample; and trying to clean the data (handle outliers). These descriptive questions were very aligned with the class content which included descriptive statistics and data cleaning. Table 3 shows examples for each category of the Descriptive Questions.

Most of these questions were focused on analyzing the provided large dataset; however, there were three students who asked questions that required additional available data for their solution. For instance, Student 1D asked, "What is the max time allowed for a subscriber ride?" (Maximum-Minimum Question). This question is potentially answerable, but its solution requires the student to look for additional available data by, for example, reading detailed information on the company's website.

Inference Questions. 33% of the individual students and 29% of the teams proposed questions that focused on making inferences from data by calculating if there was a significant difference between variables; determining if there was causation between variables; or investigating why a specific pattern in the data happened. The questions were categorized as correlations when the students or teams used a general term similar to "relationship" without specifying any of the other types of inference questions. While the topics of correlations and data linearization were addressed in the class, performing statistical tests was outside of the course scope which may have limited their ability to continue addressing inference questions. Table 3 shows examples for each category of the Inference Questions.

Compared with descriptive questions, there were more inference questions that required additional available data for their solution. For example, student 5A asked "What are the effects of weather on bike use?" (Cause-effect). Answering this question required the student to look for the forecast and relate this data with a variable that represents the bike use. Additionally, there were a couple of questions that require additional data that do not exist or are not available. For instance, student 4A asked "How exactly does maintenance for the bikes correspond with duration of usage?" (Correlation). Answering this question required additional unavailable data since the students did not have and could not find any information related to the maintenance of the bikes.

Business Model Questions. 33% of the individual students and 54% of teams proposed questions that concentrated on analyzing the company's business model instead of analyzing the provided large dataset. These questions looked for proposing preliminary recommendations for the client; characterizing the company's customer segments; inquiring about the company's profit (revenue streams); and examining the company's key activities and the

Table 3. Examples of students' questions for each category of questions' content

Questions' Content		**Example**
Descriptive Questions	Frequency	"What is the spread of trip duration like across the months?" (Student 21A).
	Maximum – Minimum	"How many stations have very little trips?" (Student 11C).
	Central Tendency	"What is the average of the male and female riding the bikes each month?" (Student 9B).
	Variable Size	"How many subscribers are there in total?" (Student 5D).
	Data Cleaning	"There were many outliers?" (Team 19).
Inference Questions	Correlation	"What is the relationship between stations and trip duration?" (Student 14C).
	Significant Difference	"Difference between customer and subscriber preferences?" (Student 11D).
	Cause-effect	"Does gender/types of users affect the trip duration and satisfaction/review on CoGo Bikeshare?" (Student 6A).
	Investigating patterns	"Why are there several locations that customers seem to frequent more often than subscribers?" (Student 11A).
Business Model Questions	Preliminary Recommendations	"How can changing the time limit for the subscriber encourage more customers to join?" (Team 1).
	Customer Segments	"Are the bikes used more by tourists or locals?" (Student 15D).
	Revenue Streams	"Can we calculate how much money CoGo is making?" (Student 12B).
	Key Activities and Resources	Why are bikes only able to be taken for 30 minutes when the average time used was about an hour?" (Student 4D).

resources required to perform those activities. All these questions required additional unavailable data for their solution since students did not have access to information about the company's business model. Although the focus of the bike-share problem was to learn data analytics, the real context of the project prompted students to think about business concepts and see the problem from a bigger perspective. Table 3 shows examples for each category of the Business Model Questions.

5.1.2.1 Individual and Team Variety of Questions

The students who had a high quantity of questions also showed an interesting variety of questions' ideas. There were 12 students (1A, 1B, 1C, 2B, 2D, 3B, 3C, 3D, 4B, 4D, 5D, 13B) who proposed more than 5 questions. From them, only Student 2B presented all the questions around the same idea, investigating patterns. Most of them included questions with two or three different ideas. For example, Student 5D proposed the following set of question with rich variety:

- "Which events or holidays impact the number of bikes used? How does it differ between subscribers and customers?" Two Inference Questions – Cause-Effect
- "Why are some locations more popular than others for taking and returning bikes?" Inference Question – Investigating Patterns
- "How many subscribers are there in total and on average how many times per month do they use the bikes?" Descriptive Questions – Variable Size and Descriptive Question – Frequency.
- "Are outliers in trip duration more likely to be subscribers or customers?" Descriptive Question – Data Cleaning.

Furthermore, we found that three of the students with high quantity and variety were from team one and the other three from team three. As we mentioned previously, these two teams may have developed an agreement as a team about the activity expectations. Then, three out of four teammates exhibited high quantity and variety in their questions.

For the team ideation stage, only Teams 12 and 20 proposed five or more questions. Although these questions were exploring different aspects of the same problem, they were not as different as the individual students' questions. For instance, Team 12 proposed two big questions: "How does age, starting station, type of user and time of year compare to the total trip duration?" and "Who are the primary users of bike sharing?" Their first question integrated three small questions into one by comparing three variables of the dataset, age, start station, and user type with the same variable trip duration. Their second question is also linked with the variables age and type of user from the first one; these variables can help to characterize the primary users of bike sharing. Generating a number of team questions with less variety is expected for this ideation stage since the teams need to frame the problems that they will address in the rest of the DAL.

5.1.3 Questions Novelty

Novelty refers to the questions' ideas that were less common across the individual and team questions. There were three sources of novelty in the group of questions: questions' ideas that asked for very specific information of the data; questions' ideas that integrated several variables; and questions' ideas that included external information not found in the large dataset. The following paragraphs expand upon these sources.

Some novel questions' ideas were so specific that it was very unlikely for anybody else to consider. This case is shown in the question "What is the max time allowed for a subscriber ride?" (Student 1D). This novel question may help students to identify errors in the dataset but does not provide further information for improving the clients' enterprise. Although these questions' ideas were uncommon in the pool, they may lack quality since they did not clearly include the company's perspective or interests.

The questions' ideas were more novel as the question tried to include more variables of the dataset. Descriptive questions' ideas that aimed to characterize one variable were more common than inference questions' ideas that related two or more variables. For example, only Student 7C asked "How does the age of subscribers affect the location from which the bikes are used?" This question tried to relate the variable user's birth year with either the variable's start location or end location. Furthermore, the inference questions were even more uncommon for the team ideation stage than the individual one. Students may have lacked statistics knowledge to further answer them.

Novel questions were also found when students included external information not found in the large dataset. For instance, student 20A proposed, "How does the availability of bikes correlate with the population densities in different areas in Columbus?" or student 19C proposed, "Why are the reviews so bad for Cogo?" Each of these questions integrated information external to the dataset such as the population density or Cogo's reviews. The students took the initiative to look for that additional information that could provide a competitive advantage in terms of novelty compared with teams based on only the large dataset.

Novel questions ideas were the ones that were proposed by a few students or teams and, when they had quality, had the potential to provide meaningful information for the client. The integration of several variables promotes the combination of variables that can help to discover more complex patterns in the data and the integration of external information can enrich the dataset expanding the problem space. All of them are elements that may promote finding new relevant information for the client.

5.1.4 Questions Quality

Quality questions need to be clear enough to convey the students' ideas. Some students' and teams' questions were unclear due to undefined variables or syntax problems. On the one hand, students employed undefined terms to write the questions. For example, in the question, "How does age relate to the amount traveled?" (Student 12C), the student used the term "amount traveled." The term "amount" could refer to either the distance traveled by the user or the duration of the ride. On the other hand, some team questions tried to include several details that make the sentence less clear for an external reader. For example, Team 11 proposed, "What locations does CoGo Bike Share need to concentrate resources to maximize profit by ensuring bikes are always available at hot locations, during each three-month interval, season?" This question could have been simplified to enhance clarity by saying: What are the company's busiest stations in each three-month interval? Moreover, if the students wanted to include more details, they could have included them in the problem description. They may have thought that all the information related to the problem should be included in the question. Additional clarification about the expectations for the team questions may have helped the students to produce more precise team questions.

5.1.5 Integration of the Ideation Elements When Asking Questions

The student's ideation of questions for data analytics can be analyzed using four elements, quantity, variety, novelty, and quality. Our data showed complex relationships among those four parameters. Quantity represents a baseline for variety. Namely, it is less likely to have different questions (variety) when the number of proposed questions (quantity) was minimum. Furthermore, variety and novelty are also closely related. When students' questions were very different in terms of variety, they also tended to be more novel. For example, Student 5D presented an outstanding quantity and variety with 6 questions with different content and answerability. Moreover, she also proposed the novel question with good quality: "Which events or holidays impact the number of bikes used?" However, the DAL is usually performed in teams which adds a negotiation process that can affect the relationship between the four parameters. For instance, even though team 5 had a student with a remarkable individual ideation of questions; later, during the team ideation, they decided to follow the questions: "Which final destination station is the most popular among all types of users?" and "How does it compare with the destinations with the highest mean trip duration?" which are very common among all the individual and team questions. The team may have compromised to focus an answerable question that each team member could answer quickly for each part of the data. Then, they could come up quicker with a recommendation and finish the project. The following section describes in detail the ideation's elements for the recommendations.

5.2 Recommendations for the Client in the Data Analytics Lifecycle

The recommendations for the client are the final activity of the DAL. Teams came up with recommendations based on their statistical findings for improving the client's enterprise. This section describes the results for the quantity, variety, novelty, and quality of the students' recommendations. We present the results for each of the elements separately and later in the discussion section provide a larger perspective on the integration of the elements.

5.2.1 Quantity of the Recommendations

Ideation quantity for the recommendations refers to how many recommendations the teams proposed. Although teams were asked to provide just one recommendation, we found out that more than 60% of the teams provided more than one. For example, as Fig. 4 shows, Team 15 proposed six recommendations which are related to running incentive programs to get more users, and teams 5 and 8 submitted five and four recommendations respectively, all of them focused on ways to help the client deal with the busiest stations. On the other hand, 12% of the teams did not include any recommendations (teams 13, 14, and 18). A total of 38% of the teams in the class provided one recommendation as requested in the activity. Quantity is related to higher likelihood of getting the solution selected by the client. By providing a higher number of recommendations to the client, chances are that the client identifies a solution that could fit better with their company vision. Thus, students may be encouraged to include several recommendations that have the potential to help the client which

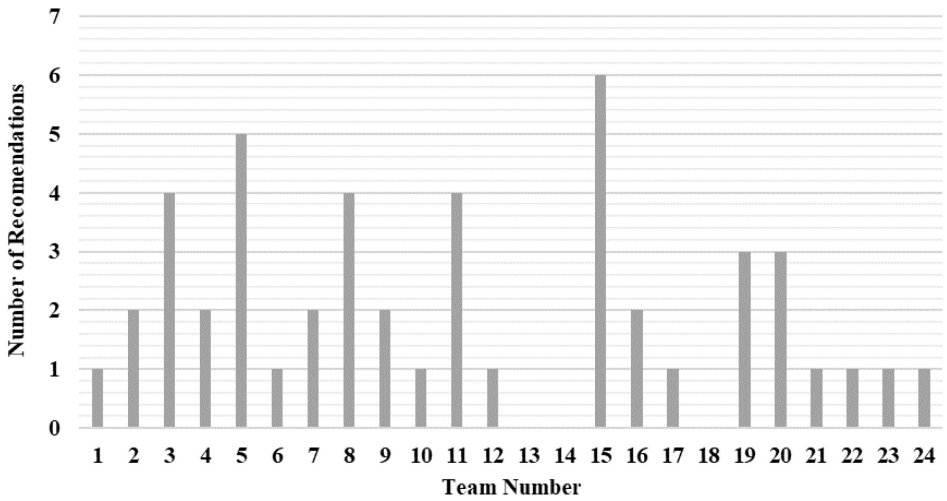

Fig. 4. Number of recommendations proposed by team.

would also help to avoid their fixation on only one recommendation.

5.2.2 Variety of the Recommendations

Ideation variety for the recommendations refers to the differences among the recommendations. Table 4 shows the recommendations' categories and the quantity of teams that included each of them in their report, as well as a team recommendation example of each. Using the terminology of the Business Model Canvas [43], we grouped the recommendations into three main categories, (1) Key Resources, (2) Key Activities, and (3) Customer Relationships.

Key Resources are the most important assets for a successful business model. That is, those resources that enable a business to identify their niche markets, develop a caring contact with customers, and create revenue [43]. Most of the teams (57.1%) proposed a final recommendation in this category. It seems thinking about the company's key resources is the most intuitive approach for the students when generating recommendations. Table 4 shows three subcategories associated with Key Resources: (1) modify available bikes, (2) modify number of docks of stations, and (3) acquire bikes for specific customer segments.

Key Activities are the main things you need to do to have a business model and to provide the solution (i.e., products or services) to clients [43]. A smaller number of teams (37.1%) proposed within this category. Table 4 shows three subcate-

Table 4. Type of recommendations and number of teams who included them with examples

Recommendations' Type		**# Teams**	**Example**
Key Resources	Modify available bikes	7	"For CoGo Bike share to remain most efficient, the company can reduce the number of bikes that are put out on the streets during the winter months" (Team 2).
	Modify number of docks or stations	10	"CoGo should service the hot locations specified in our team finding graphs with enough bikes such that there are always bikes available." (Team 11).
	Acquire bikes for specific customer segments	3	"CoGo Bike Share can host events to encourage more females since there is a disparity between the number of male and female users" (Team 3). "To accommodate for users of this age [older users] . . . we can make using the bike easier" (Team 10).
Key Activities	Change the maintenance program	2	"Keeping high maintenance of these two stations [#45 and #13]" (Team 5).
	Pack Offer subscriptions	5	"Adding a specific marketing strategy aimed at customers who are potential subscribers" (Team 11). "We recommend discount for first time users and special deals for college students" (Team 3).
	Influence revenue streams	6	"CoGo Bike Share restructures their fee system." (Team 6).
Customer Relationships	Modify the app fee renewal availability	1	"The third solution would be to have the ability to renew the rental on the app." (Team 19).
	Be more upfront about pricing	1	"The first solution is easiest for the company to implement is to be more upfront about the pricing" (Team 19).

gories associated with key activities: (1) change the maintenance program, (2) pack offer subscriptions, and (3) influence revenue streams.

Customer Relationships are the final category of recommendations that fell within the Business Model Canvas. The customer relationships depend on the type of business and type of the relationship the company wants to provide to its clients [43]. In this case, a few teams (5.7%) generated a recommendation in this category. Students may be encouraged to think about the relationship between the client and the final user to promote more novel recommendations. Table 4 shows two subcategories associated with customer relationships. Those are: (1) modify the app for the renewal capability, and (2) be more upfront about pricing.

The class had a good recommendations' variety, and there are teams with high and low recommendations' variety. Overall, the class presented a good variety of recommendations that included changing company policies such as the price of renting bikes, changing specific facilities to receive more users, or changing the app to give a better user experience. Inside the teams, there were cases of high and low variety of recommendations. For example, Team 19 proposed three different recommendations around pack offer subscriptions, modify the app fee renewal availability, and be more upfront about pricing. In contrast, Team 11 proposed three recommendations all related to modify the available bikes. Less variety of recommendations per team is expected since the recommendations need to be similar enough to address the same analytics problem. The client would me more interested in having recommendations aligned with the identified problem that random ideas without a support.

5.2.3 Novelty of the Recommendations

Novelty for the recommendations refers to the less common recommendations proposed by the teams. We identified two sources of novelty: (1) recommendations proposed by two or less teams and (2) the recommendations with high elaboration in terms of details. For the first case, Table 4 includes the number of teams that provided each of the recommendations. The most novel recommendations are related to modifying the app for fee renewal capability, being more upfront about pricing, and changing the maintenance program. For example, Team 19 proposed the two novel recommendations in the Customer Relationship category (3): "have the ability to renew the rental on the app," and "be more upfront about the pricing." Moreover, Team 8 proposed a novel recommendation related to changing the maintenance program, "Perform the maintenance between 5–6 am." These examples are representative of the first source of novelty that were proposed by two or less teams.

The second source of novelty was the recommendations' level of elaboration. The recommendations included different levels of elaboration in terms of the amounts of details, which provided additional evidence for considering a recommendation novel. For instance, Team 15 and Team 11 proposed key resources and key activities types of recommendations that were not the least common across the pool of recommendations. However, while Team 15 only mentioned running incentive programs to get more users, Team 11 also included offering a discount for the first-time users to get more users. In this case, Team 11 recommendation would have developed more novelty due to the level of its specificity. Overall, students may be encouraged to include detailed recommendations to promote more novel ideas. Novelty is related to better engineering thinking within the understanding of the problem. From the results we see here, novelty is not only in the newness of the idea but can also be in the new ways of looking at more common ideas by presenting the novelty through a detailed description.

5.2.4 Quality of the Recommendations

Students were directed to include three elements in their recommendations: (1) data to support the recommendation, (2) details of the recommendation, and (3) justification aligned with the needs of the client. The following paragraphs describe the differences in teams' recommendation *quality* due to differences in those three areas.

Teams' recommendations exhibited different levels of quality depending on how much elaboration they had. Table 5 shows teams' recommendations with different levels of quality sorted from lowest to highest quality. For example, team 20 provided a team recommendation with minimum quality since they provided only a justification for the client, without supporting it with data from their analysis nor adding a description to better understand/present their recommendation. In a similar way, Team 7 provided a recommendation supported with low reliable data and without details and justification. Namely, their data mostly came from their own collective previous experiences (i.e., In winter there is a rise in use of ice-melting chemicals that can negatively affect (corrosivity) other metal materials such as those from bikes or cars), and not from the raw data provided to students. In contrast, Team 3 provided a recommendation supporting it with data and added a justification of why that recommendation is relevant for the client. However, they did not elaborate on that recommendation by proposing,

Table 5. Examples of teams' recommendations (order from lowest to highest quality).

Team #	Recommendation	Supporting Data	Details' recommendation	Justification
20	"Placing stations where most people in Columbus live"	None	None	"Offering locals, a way to get around traffic quickly via a convenient and easy to use transportation system is a great way to increase the demand for more stations and grow the CoGo brand"
7	"Reduce or increase the number of bikes in circulation, depending upon the time of the year."	"Considering that bikes can rust easily during the winter, leaving the same number of bikes in circulation all year round will lead to high maintenance costs for the company"	None	None
3	"Discount for first time users and special deals for college students"	"The peak age for bike usage is people who were born in 1989, 1974 and 1967. However, your subscribers could be as young as 16 years of age."	None	"You can entice more subscribers" "To promote services for young people"
2	"Reduce the number of bikes that are put out on the streets during the winter months"	"Looking at the two graphs that were created for overall bike usage and subscriber/ customer usage over the twelve-month interval, it can be said that generally usage decreases over the winter, or colder months"	"During the summer months more bikes can be put out because these are times when more people will be using the bikes for transportation"	"For Cogo-Bike share to remain most efficient"

for example, how much the discount for young users could be. Finally, Team 2 provided an outstanding quality recommendation that was supported with data, included additional details, and was justified based on the client's needs.

High quality recommendations were reached when students based them on reliable data, included details, and aligned with the client's needs. Most of the teams provided these three elements in at least one of the recommendations (e.g., Team 2, Table 5). However, they usually provided additional recommendations with missing information. Although instructors may be interested in reviewing other students' underdeveloped ideas, in a real company context, the clients may want to have full developed ideas to make decisions. In any case, students should be encouraged to propose elaborated recommendations supported on data analysis and that considered the client's interests.

5.2.5 Integration of the Ideation Elements when Making Recommendations

In a similar way to the ideation of questions, the elements of quantity, variety, novelty, and quality are intertwined for the ideation process of recommendations. This relationship generates challenges when assessing the students' generation of recommendations. For example, to evaluate the variety within a team's recommendations, they would need a high quantity of them. In the same way as with the ideation process of questions, a variety of recommendations is necessary to promote the generation of novelty. In contrast to the questions, novel recommendations are not only important for the client because of their uniqueness but also because they need to have a high quality according to the definition presented in the previous section. For example, Team 19 proposed the novel recommendations: "have the ability to renew the rental on the app" from the Customer Relationship category. However, when evaluating the recommendations' quality, the team had minimal elaboration on the idea and lacked a clear justification based on the client's needs. Consequently, although the team had a novel recommendation that no other team proposed, it may not yet be meaningful for the client due to the low quality, i.e., lack of justification and elaboration. Instructors need to consider all those complexities when evaluating the students' ideation process of recommendations. Moreover, to provide students an experience to practice all those four elements, there might be a need for explicitly presenting the definitions of quantity, variety, novelty and quality to the students, as well as their importance when generating questions and making recommendations. Furthermore, perhaps scaffolding with rubrics or other organiza-

tional devices may be useful in helping students make the connections between all the important aspects of developing a solution.

6. General Discussion

Our results section focuses on the questions formulated by the FYE students and their resulting recommendations to the client. We found that the majority of the FYE students were able to translate a large raw dataset into feasible recommendations for a client based on statistical findings and following the DAL. This section presents five themes that emerged throughout the analysis of our data. The first four themes are characteristics of the FYE students' questions and recommendations when performing the DAL. The final theme describes good practices for performing analytics that we observed in the teams with outstanding ideation processes of questions and recommendations.

6.1 Relationship Between the Dataset and the Students Questions' Ideation

The provided large dataset may shape the ideation process of questions. In the case of the COGO dataset, the categorical variables expanded the opportunities to propose questions and subsequent recommendations. For example, many students focused on asking descriptive questions about the distribution of the numerical variable trip duration throughout the year. In contrast, other students used the categorical variable user gender to not only ask for the distribution of trip duration, but also to compare it between males and females. Without that categorical variable, students might have focused on mostly descriptive questions, which may have provided less information for the future recommendations. Furthermore, the variables that were provided in the data set also may have had a limiting effect on the types of questions students asked. For example, we saw that some teams asked questions that were directly answerable from the data given, which was likely limited by the types of variables included in the data set. In contrast, other teams asked questions that went beyond the given data to data they would have to find themselves from other sources or would not be answerable at all without additional data collection.

All these different approaches to asking questions suggest that, at least for some teams, the provided variables and their type limited the types of questions students asked. In today's world, getting large data sets with users' information is not as challenging as getting the personnel who have the ability to glean meaningful information from these data [17, 37]. Companies need to take advantage of any opportunity to promote their personnel's ideation of a variety of questions that can generate value for the organization. Our results suggest that looking for rich datasets with a mix of different types of variables may provide students and potentially professionals with opportunities for coming up with more ideas for analysis that may benefit the company.

6.2 Asking Questions' Role in the DAL

The students' questions showed how they perform the DAL's practices of Frame the Problem and Understand and Explore the data. Asking questions allowed students to explore and expand the problem space which ended up framing their analytics. The questions that students formulated based on the provided data alone represent a more intuitive problem space. In fact, novice data scientists tend to frame their problem solely around what data are available instead of focusing on what would be the most valuable insight for the client [17]. In contrast, the bike-share problem students expanded this intuitive problem space by considering additional information. For example, in the question, "Which events or holidays impact the number of bikes used?" (Student 19A), the student proposed to look for the holidays and contrast that information with the data set. While this example shows that the student was bringing in reliable and easily verifiable data, students may not always choose high-quality data to bring into the problem. For instance, student 6A proposed: "Does gender/types of users affect the trip duration and satisfaction/review on CoGo Bikeshare?" This question required analyzing the publicly available reviews for COGO which may not be a reliable source of information. Allowing students to look for additional data not only may help the client because the students may frame the problem more accurately but also can be a learning opportunity to discuss information quality and trustworthiness.

The questions reflected students had a mix of deep and superficial exploration of the problem space. On the one hand, a greater number of students proposed questions that investigated patterns that showed a deep exploration and understanding of the data. For example, Student 11A proposed, "Why are there several locations that customers seem to frequent more often than subscribers?" a question that showed the student used statistics to determine the most crowded locations and analyzed the results according to the problem context. In a similar way to the engineering design problems, a high integration of the context indicates a deep problem exploration [22, 23]. On the other hand, some students did a superficial exploration of the problem space by proposing recommendations for the client during the ideation stage. For

example, Team 1 asked, "How can changing the time limit for the subscriber encourage more customers to join?" during their team ideation stage. They proposed the recommendation of changing the rides' time limit to the client as part of their question development instead of generating a question to frame an analytics problem. It is likely that in these cases, students are following the pattern of novice designers who tend to skip the problem framing to start solving the problem prematurely [18, 19]. Thus, some students and teams may have needed more support to analyze the problem deeply before moving into a solution.

6.3 Confluence of the Students' Individual Ideation and the Team's ideation

Most of the students did not continue exploring novel or statistically complex questions as a team. The negotiation of team questions from individual questions is a complex process that may impact the final questions variety and novelty. From the individual to the team questions ideation, the students decided to propose mostly descriptive questions and questions that were not answerable. Teams may have found it too difficult to continue comparing variables for their team questions due to the complexity of those questions. In previous studies, we have found that FYE students may be limited by their statistical knowledge when performing data analytics [14]. Answering more complex questions would require statistical tests or correlation analyses that were outside of the course topics. Furthermore, some teams did not choose to explore some of the novel individual questions that the individual team members had identified. For example, as the results showed, Student 5D presented 6 different questions including the novel question: "Which events or holidays impact the number of bikes used?" (Student 5D); however, the team decided to address very common questions. Making decisions as a team is a complex social activity that has been documented in other contexts [29–31]. However, more research needs to be done about how that decision making process impact the individual and team ideation process of questions and recommendations while performing data analytics.

6.4 Recommendations' Development in DAL

Students' recommendations focused on changing the company business models and showed different levels of elaboration in terms of data support, proposal's details, and reference to client needs. The teams used different statistical tools to identify different patterns in the data that would represent problems for the company. Then, they had to decide which pattern or problem was more relevant for the company and, based on that, propose a recommendation for the client. Our findings showed that although all teams' recommendations focus on impacting the company business model in different ways (e.g., Modifying key activities or customer relationships), how in-depth they described their recommendations varied.

For instance, while Team 15 proposed to run incentive programs to get more users, Team 3 proposed to offer a discount for first-time users and special deals for college students to get more users. Both teams focused on getting more users, but Team 3 provided a more elaborated proposal for reaching that goal. Design thinking literature has explored how expert designers presented more detail solutions compared to novice ones [25]. In our context, students seem to follow the same pattern and propose many recommendations but, sometimes, with superficial descriptions. Superficial recommendations may miss company constraints that will end up becoming the recommendation unfeasible. Students need to be encouraged to analyze and describe in depth each recommendation when working in data analytics.

6.5 Promoting Outstanding Questions and Recommendations

Performing the DAL activities of designing questions and making recommendations is challenging and requires experience. Our data showed some common practices across the teams that had an outstanding ideation process of questions and recommendations. These practices can be promoted in the classroom to facilitate the analytics. The following paragraphs describe three students' practices when working on the Bikeshare problem.

6.5.1 Students' Divergent-Convergent Thinking when Performing the DAL

Students' use of a combination of divergent thinking followed by convergent thinking can promote exploration of the problem space and the generation of novel questions. On the one hand, the individual ideation stage of questions requires divergent thinking to propose many questions with a high variety of content and answerability. For example, student 5D demonstrated this thinking by proposing four individual questions including descriptive and inference questions. On the other hand, the team ideation stage required more convergent thinking to frame a specific problem using one or few questions with less variety of ideas. For instance, as the results showed, Team 12 showed this convergent thinking by proposing only two team questions that focused on characterizing the users that have higher or lower trip durations. In contrast, Team 20 proposed only five individual questions (minimum divergent

thinking) and six team questions (minimum convergent thinking). Nelson [1] and R. Woods [20] argued that data analytics teams needed this divergent-convergent thinking sequence to reach innovative solutions for their companies. Furthermore, in other contexts such as engineering design, a positive correlation has been found between divergent thinking and the generation of novel ideas [31]. This divergent-convergent thinking needs to be promoted in the classroom among all students to support the generation of creative data-based questions followed by evidence-based selection of questions that meet the needs of their client.

6.5.2 Alignment between Students' Questions and Recommendations

Students needed to decide which recommendations were aligned and were more relevant to the client and then remove other potential ideas. In general, companies are interested in recommendations that address problems that fit their needs and comply with their constraints [36]. Namely, the recommendations needed to be aligned with the problem and team questions. According to our findings, students still struggled with this alignment. It seems they wanted to include many recommendations without considering if they were or not related to their problem framing. For example, Team 8 proposed four recommendations around ways to incentivize customers to become subscribers. However, they included changing the company's maintenance program as a final recommendation which is not closely related to their question which was related to how to get customers to become subscribers. Students should have done an iteration over the DAL and, after proposing their recommendations, gone back to evaluate if their team questions were aligned with them. Considering that the DAL is a set of practices that overlap and can be done iteratively [1], students should not see their initial framing process as a final problem statement with final team questions but as a preliminary framing that needs to be constantly revised.

This section identified a series of practices that may promote high-quality ideation processes of questions and recommendations. We encouraged students to use a divergent-convergent thinking process when moving from many individual questions to a few team questions that guided their data analysis and started framing the problem. Based on the original team questions, students may run several statistical calculations to identify trends in the data. By looking at those trends, they can go back and reframe their initial problem and come up with a few data-based and elaborated recommendations to help the client. Then, students may evaluate the alignment between their problem, team questions, and recommendations and make the necessary changes to guarantee this alignment. Finally, teams can present the problem, team questions, and recommendations with their statistical support to the client. The next section presents this study's conclusions and implications for teaching and research with future recommendations to continue developing the knowledge about the students' ideation process when performing data analytics.

7. Implications

This study points to several implications and recommendations for teaching data analytics. Our results show the importance of promoting the iterative process of design and DAL by asking students to propose work-in-progress team questions instead of definitive team questions. Students need support to communicate their problem statement in writing and to assure their final proposal is aligned with this problem statement. Additionally, students need scaffolding to consider the problem carefully before jumping to solutions, which is a common struggle for novice designers [19]. To continue developing skills in DAL, students need practice with data analytics throughout their educational careers. Further work is needed to both develop resources to do this and to better understand how students engage in data analytics over the course of their education.

Our results also point to implications about how data sets are used to engage students in DAL. Rich datasets with different types of variables and from real companies provide opportunities for students to practice exploration of data and to engage with real-world data. Our research found that students working on the project in conjunction with learning the course outcomes helped students to reflect and better explore the data. Additionally, students should be encouraged to look for information from outside the provided data sets to expand their problem space and potentially promote novelty. This could be also an opportunity to teach about evaluating the quality of the information that they bring and practice how engineers make decisions based on data. Incorporating meaningful and real-world contexts provides students practice with real-world data analytics and provides opportunities to incorporate many aspects of a multi-faceted problem. However, the context that is provided to the students will affect and guide how the students approach the problem, making a trade-off between promoting creativity and making a solvable problem. Additionally, this study points to misconceptions students have about the relationships between engineering problems and business decisions. For example, some

students proposed recommendations based on incomplete information or misalignment with the problem they investigated that indicate an incomplete understanding of the data and evidence involved in business decisions. Further research is needed to develop principles that guide engineering curriculum developers and instructors in designing high-quality experiences for students as they use data analytics to explore publicly available data sets.

8. Conclusions

In this study, we examined the ways in which students developed questions and recommendations based on a set of big data. Specifically, the students engaged in the DAL to *frame a problem*, *understand and explore the data*, *develop a model*, and *interpret, explain and activate the results.* The data set was taken from the freely available COGO data on bikeshare usage and included thousands of data points across several types of data. We examined the ways in which students engaged in the DAL and included examples of the novelty, quality, quantity, and variety of the students' individual and team questions and recommendations. We found that the data set included variables that guided the students' questions and the different types of variables gave students opportunities to use different types of data. The students asked questions to frame, explore, and understand the analytics problem, but the individual questions were more statistically complex, and novel compared to the teams' questions that they chose to explore. Students applied divergent thinking with their individual questions and then convergent thinking with their team questions to frame their analytics problem. Students identified patterns in the data that represented problems for the company and based on them, made up a variety of recommendations with different levels of elaboration. However, students' final recommendations were not always aligned with the team questions they had posed. Finally, we provided research and teaching implication to promote the students' ideation of questions and recommendations when working on data analytics projects.

Acknowledgements – The authors would like to thank Jill Folkerts, Nastasha Johnson, and Michael Witt for their help in curriculum development/implementation and feedback on the activity used for this work. As this work is part of the Ph.D. dissertation of lead author Ruben D. Lopez-Parra, thank you to his committee members, Professors Allison Godwin and Kevin Solomon, for their feedback on this research paper.

References

1. G. S. Nelson, *The analytics lifecycle toolkit: A practical guide for an effective analytics capability*, Hoboken, NJ: John Wiley and Sons, 2018.
2. F. Provost and T. Fawcett, Data science and its relationship to big data and data-driven decision making, *Big Data*, **1**(1), pp. 51–59, 2013.
3. C. Min, C. Roman and S. Trevor, Big data analytics in financial statements audits, *Account. Horizons*, **19**(2), pp. 423–429, 2015.
4. 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.*, **114**, pp. 343–349, 2018.
5. W. Raghupathi and V. Raghupathi, Big data analytics in healthcare: Promise and potential, *Heal. Inf. Sci. Syst.*, **2**(3), pp. 1–10, 2014.
6. K. A. Douglas, P. Bermel, M. M. Alam and K. Madhavan, Big data characterization of learner behaviour in a highly technical MOOC engineering course, *J. Learn. Anal.*, **3**(3), pp. 170–192, 2016.
7. L. Chiang, B. Lu and I. Castillo, Big data analytics in chemical engineering, *Annu. Rev. Chem. Biomol. Eng.*, **8**, pp. 63–85, 2017.
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, pp. 1–28, 2019.
10. R. Rose, Defining Analytics: A conceptual framework, *ORMS Today*, **43**(3), pp. 34–38, 2016.
11. N. R. Hassan, The origins of business analytics and implications for the information systems field, *J. Bus. Anal.*, **2**(2), pp. 118–133, 2019.
12. I. Y. Song and Y. Zhu, Big data and data science: What should we teach?, *Expert Syst.*, **33**(4), pp. 364–373, 2016.
13. D. Rose, *Data science: Create teams that ask the right questions and deliver real value*, Atlanta, GA: Apress, 2016.
14. R. D. Lopez-Parra, A. Carrillo-Fernandez, A. Johnston, T. J. Moore and S. P. Brophy, Asking questions about data: First-year engineering students' introduction to data analytics, in *2020 ASEE Annual Conference and Exposition*, 2020.
15. 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.*, **7**(1), pp. 68–75, 2017.
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.*, **10**(1), pp. 15–34, 2011.
17. R. Woods, A design thinking mindset for data science, 2019. [Online]. Available: https://towardsdatascience.com/a-design-thinking-mindset-for-data-science-f94f1e27f90. [Accessed: 14-Oct-2021].
18. C. J. Atman, R. S. Adams, M. E. Cardella, J. Turns, S. Mosborg and J. Saleem, Engineering design processes: A comparison of students and expert practitioners, *J. Eng. Educ.*, **96**(4), pp. 359–379, 2007.
19. D. P. Crismond and R. S. Adams, The informed design teaching and learning matrix, *J. Eng. Educ.*, **101**(4), pp. 738–797, 2012.
20. C. Cardoso, P. Badke-Schaub and O. Eris, Inflection moments in design discourse: How questions drive problem framing during idea generation, *Des. Stud.*, **46**, pp. 59–78, 2016.

21. O. Eris, *Effective inquiry for innovative engineering design*, New York, NY: Springer Science+Business Media, 2004.
22. C. J. Atman, K. Yasuhara, R. S. Adams, T. J. Barker, J. Turns and E. Rhone, Breadth in problem scoping: A comparison of freshman and senior engineering students, *Int. J. Eng. Educ.*, **24**(2), pp. 234–245, 2008.
23. D. Kilgore, C. J. Atman, K. Yasuhara, T. J. Barker and A. Morozov, Considering context: A study of first-year engineering students, *J. Eng. Educ.*, **96**(4), pp. 321–334, 2007.
24. D. G. Jansson and S. M. Smith, Design fixation, *Des. Stud.*, **12**(1), pp. 3–11, 1991.
25. D. C. Sevier, K. Jablokow, S. McKilligan, S. R. Daly, I. N. Baker and E. M. Silk, Towards the development of an elaboration metric for concept sketches, in *Proceedings of the ASME 2017 International Design Engineering Technical Conference and Computers in Engineering Conference*, pp. 1–10, 2017.
26. C. Campbell, W. Roth and A. Jornet, Collaborative design decision-making as social process, *Eur. J. Eng. Educ.*, **44**(3), pp. 294–311, 2019.
27. C. A. Toh and S. R. Miller, How engineering teams select design concepts: A view through the lens of creativity, *Des. Stud.*, **38**, pp. 111–138, 2015.
28. M. C. Yang, Consensus and single leader decision-making in teams using structured design methods, *Des. Stud.*, **31**, pp. 345–362, 2010.
29. B. A. Hennessey and T. M. Amabile, Creativity, *Annu. Rev. Psychol.*, **61**, pp. 569–598, 2010.
30. R. A. Finke, T. B. Ward and S. M. Smith, *Creative cognition: Theory, research, and applications*, Cambridge, MA: The MIT Press, 1992.
31. B. Kudrowitz and C. Dippo, Getting to the novel ideas: Exploring the alternative uses test of divergent thinking, in *Proceedings of the ASME 2013 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, pp. 1–6, 2013.
32. M. A. Runco, Divergent thinking, creativity, and ideation, in *The Cambridge handbook of creativity*, J. C. Kaufman and R. J. Sternberg, Eds. Cambridge, England: Cambridge University Press, pp. 413–446, 2010.
33. S. M. Smith, Fixation, incubation, and insight in memory and creative thinking, in *The creative cognition approach*, S. M. Smith, T. B. Ward and R. A. Finke, Eds. Cambridge, MA: The MIT Press, pp. 135–156, 1995.
34. T. B. Ward, What's old about new ideas?, in *The creative cognition approach*, S. M. Smith, T. B. Ward and R. A. Finke, Eds. Cambridge, MA: The MIT Press, pp. 157–178, 1995.
35. K. K. Lim, I. Zaleha and M. Y. Yusof, Fostering mathematical creativity among engineering undergraduates, *Int. J. Eng. Educ.*, **1**(June), pp. 31–40, 2019.
36. H. Mayhew, T. Saleh and S. Williams, Making data analytic work for you instead of the other way around, McKinsey Quarterly, 2016. [Online]. Available: https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/making-data-analytics-work-for-you-instead-of-the-other-way-around [Accessed: 11-Sep-2021].
37. C. Gaur, Top 6 big data challenges and solutions to overcome, 2020 [Online]. Available: https://www.xenonstack.com/insights/big-data-challenges.
38. J. J. Shah, S. M. Smith and N. Vargas-Hernandez, Metrics for measuring ideation effectiveness, *Des. Stud.*, **24**(2), pp. 111–134, 2003.
39. J. J. Shah, N. Vargas-Hernandez, S. M. Smith, D. R. Gerkens and M. Wulan, Empirical studies of design ideation: Alignment of design experiments with lab experiments, in *Proceedings of the ASME 2003 International Conference on Design Theory and Methodology*, 2003.
40. C. Pope, S. Ziebland and N. Mays, Analyzing qualitative data, in *Qualitative research in health care*, 3rd ed., C. Pope and N. Mays, Eds. Oxford, UK: Blackwell Publishing, pp. 63–81, 2006.
41. C. Grbich, *Qualitative data analysis: An introduction*, 1st ed., London, England: SAGE Publications, 2007.
42. M. Schreier, *Qualitative content analysis in practice*, Thousand Oaks, CA: SAGE Publications, 2012.
43. Ashton Mcgill, The business model canvas series: Introducing the business model canvas, 2018. [Online]. Available: https://www.ashtonmcgill.com/business-model-canvas-series-introducing-business-model-canvas/ [Accessed: 02-Apr-2022].
44. K. Krippendorff, *Content analysis: An introduction to its methodology*, 2nd ed., Thousand Oaks, CA: Sage Publications, 2004.

Ruben D. Lopez-Parra (he/his) is a PhD candidate in Engineering Education at Purdue University. He has worked as a K-16 instructor and curriculum designer using various evidence-based active and passive learning strategies. In 2015, Ruben earned an MS in Chemical Engineering at Universidad de los Andes in Colombia where he also received the title of Chemical Engineer in 2012. His research interests are grounded in the learning sciences and include how K-16 students develop engineering thinking and professional skills through diverse learning environments. He aims to apply his research in the design of better educational experiences.

Aristides P. Carrillo Fernandez (he/him) is a PhD candidate in the School of Engineering Education at Purdue University. For over six years, he worked as an export business development manager at a Spanish radio communications company in Madrid, Spain. During that time, he developed new distribution dealer networks in Western Europe and African countries. He earned his MS in Electronics and Systems of Telecommunication at ESIGELEC (École Supérieur D'Ingénieurs en Génie Électrique) at Rouen, France, and his BS in Systems of Telecommunication at the Polytechnic University of Madrid at Madrid, Spain. Aristides' primary research interests are exploring the functional role of empathy in various domains, including engineering team dynamics and teamwork performance, engineering thinking, diversity, and intercultural engineering practice.

Amanda C. Emberley (she/her) is a lecturer in Mechanical Engineering at California Polytechnic State University, San Luis Obispo. Prior to joining Cal Poly, she was a postdoctoral researcher at Purdue University and earned her PhD in engineering education from Purdue in 2020. Her research interests include development of learning environments to

support engineering and STEM learning at the precollege and undergraduate levels and supports for instructors to implement these resources.

Tamara J. Moore, PhD, (she/her) is a Professor in the School of Engineering Education and 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 student learning, instructor implementation, and curriculum development within engineering design-based STEM integration and computational thinking environments.

Sean Brophy, PhD, (he/his) is an Associate Professor in the School of Engineering Education at Purdue. His research in engineering education and learning sciences explores how engineering students think and reason with models as they engage in engineering problem solving activities. This research informs his development of learning environments involving computational analysis and physical computing.