Shih, JL. et al. (Eds.) (2023). Proceedings of the 31st International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

# Unveiling University Students' Data Literacy: A Case Study on Modeling Reasoning in Data Mining Projects

*Tianqi ZHANG*
*East China Normal University, China*
51214108044@stu.ecnu.edu.cn

**Abstract:** In the era of big data, cultivating students' data literacy is of paramount importance. Data literacy encompasses the abilities to collect, manage, analyze, and apply data. As data science is a superset composed of mathematics and statistics, computer science, and specific application fields, it is crucial to investigate data literacy development through the lens of statistical education. An essential component of statistical reasoning is modeling reasoning, which is also fundamental to data literacy. However, limited research exists on the manifestation of data literacy among university students in real-world tasks. Therefore, this study explores the forms of modeling reasoning exhibited by university students in a data mining project. The findings reveal that, in a real teaching setting, university students' modeling reasoning ability follows a spiral progression. Considering these insights, modeling reasoning, as the core of data mining activities, plays a pivotal role in fostering students' data literacy. To cultivate data literacy among university students in the age of big data, we recommend implementing project-based learning, incorporating ill-structured problems and real, complex, massive datasets as project backgrounds.

**Keywords:** Data literacy, modeling reasoning, statistical reasoning, project-based learning

## 1. Introduction

In the era of big data, the rapid advancement of intelligent technology and its increasing integration into everyday life have transformed data into a valuable resource in the 21st century. Both society and the labor market now demand graduates who possess the knowledge and skills to effectively utilize statistics, data management, and computer science to make informed decisions. Consequently, the cultivation of data literacy, which encompasses the abilities to collect, manage, analyze, and apply data, has become increasingly crucial. In fact, data literacy has emerged as an upgraded version of information literacy within the realm of 21st-century skills. To address the growing demand for data experts in the age of big data, universities have established data science majors. Researchers argue that data science should encompass mathematics and statistics, computer science and programming, as well as specific application fields (Rosenberg, Lawson, Anderson, Jones & Rutherford, 2019). As data availability improves and technological advancements introduce user-friendly data processing software, researchers in statistical education are actively exploring methods to foster students' data literacy in the era of big data (Ben-Zvi, Makar & Garfield, 2018).

Within statistical education, modeling reasoning is a fundamental underlying ability in both statistical reasoning and data literacy (Bakker & Hoffmann, 2005; Wild & Pfannkuch, 1999). Rubin (2019) synthesized key aspects derived from a series of works related to data processing in the era of big data, including context, variation, aggregation, visualization, and inference. In particular, inference, specifically Informal Statistical Inference (ISI), is closely associated with modeling inference. In the realm of statistical education, modeling reasoning serves as the foundation for data analysis and result interpretation. It involves simplifying

phenomena based on data and theories, generating models that can explain, predict, or define these phenomena (Aridor & Ben-Zvi, 2017).

Modeling reasoning not only simplifies the process of defining, explaining, and predicting phenomena based on data and theories (Aridor & Ben-Zvi, 2017), but it also constitutes a key element of data literacy and statistical reasoning (Wild & Pfannkuch, 1999). Generally, models are regarded as forms of interpretation, while modeling is a process that simplifies, evaluates, and enhances our understanding of a phenomenon using key theories and data from a specific discipline, either by incorporating it into existing theoretical frameworks or by generating new discoveries (Lehrer & Schauble, 2010). Models can be classified as abstract models (conceptual models) or concrete models (e.g., figures and tables). Abstract models represent real-world systems and conjecture about their behavior to describe, explain, predict, and elaborate on their behavior (Wild & Pfannkuch, 1999). Concrete models, on the other hand, serve as tools to represent processes such as identifying key components or properties of a population, making predictions or samples, and drawing inferences about the representativeness of a random sample (Aridor & Ben-Zvi, 2017). Modeling is a multifaceted process in which the role of the model evolves alongside changes in thinking (Gravemeijer, 1999). Initially, a model takes the form of an "informal reasoning model." As a new concept emerges, the model's role changes, and it transforms into a "formal reasoning model." These two processes are accompanied by a third process that shapes the model into a set of symbols guiding previous reasoning processes (Gravemeijer, 1999).

Practical activities in statistical education are often seen as a form of modeling as they involve understanding variation and uncertainty, processing data, and constructing models (Lehrer & English, 2018). The modeling process entails evaluating and optimizing models, including generating new theoretical ideas or discoveries based on data (Dvir & Ben-Zvi, 2018). Modeling reasoning can be viewed as an analogical process that simplifies real phenomena, elucidates connections and relationships among their components, and deals with inherent uncertainties (Wild & Pfannkuch, 1999). Familiarity with the modeling reasoning process greatly aids the cultivation of data literacy. In an effort to enhance the quality of statistical education in primary schools, Biehler et al. interviewed four pre-service primary school teachers who participated in a statistical modeling reasoning training course. They evaluated and analyzed the teachers' reasoning processes, scrutinized their utilization of Tinkerplots to model statistical situations, and assessed their ability to evaluate models in accordance with the given statistical situations, providing recommendations accordingly (Biehler, Frischemeier & Podworny, 2017). The learning environment associated with modeling reasoning can also facilitate the development of students' data literacy. Conway et al. conducted quasi-experimental research on primary school students and found that adhering to the principles of a statistical reasoning learning environment in both beliefs and practices had a more positive impact on students' statistical reasoning ability compared to traditional classroom teaching (Conway, Martin, Strutchens, Kraska & Huang, 2019). This indicates that training in modeling reasoning is pivotal in statistical and data science teaching activities, and that modeling reasoning ability plays a vital role in students' comprehension, analysis, and application of data.

Despite the considerable research on modeling reasoning in primary and middle school students, few studies have focused on the data literacy of university students (Setiawan & Sukoco, 2021). Currently, there is a dearth of empirical research examining the modeling reasoning ability of university students. Thus, a better understanding of how university students manifest modeling reasoning ability is crucial for the development of data science programs. Additionally, as there is limited guidance on the theoretical exploration and teaching practices for specific data literacy training, further theoretical and empirical research is necessary to guide the development and implementation of data science education.

Therefore, this study aims to explore the manifestation of data literacy among university students during real-world tasks, specifically by examining the forms of modeling reasoning employed by university students in a data mining project. Due to the limited knowledge regarding modeling reasoning among university students, a qualitative research

approach, specifically a case study, was deemed appropriate for exploring this phenomenon. The study utilized a bottom-up grounded theory method to provide a detailed analysis of the manifestation of modeling reasoning among university students.

## 2. Method

### *2.1 Participants*

The participants in this study were three university students who enrolled in an educational data mining course at a university in Shanghai, China. These students worked collaboratively on a data mining project as part of the course. The project involved analyzing real and complex education data to predict students' answering efficiency in a specific scenario. The scenario was as follows: Some students took an online math test consisting of two sets of exams (exam A and exam B) with the same type and number of questions. Each test lasted for 30 minutes, and once the time expired, students were automatically terminated and unable to continue. The project provided students with four data documents: (1) a sequence of students' 30-minute behavior on exam A for Group 1 students, (2) the efficiency of Group 1 students on exam B, (3) the performance of Group 2 students on exam A, and (4) the prediction results of the answering efficiency of Group 2 students on exam B. The objective of the project was for students to build a model based on the performance of Group 1 students on exam A and their efficiency in completing exam B, and then use this model to predict the efficiency of Group 2 students in completing exam B based on their performance on exam A.

The dependent variable in this project was the answering efficiency of exam B, which was defined as a binary variable: “Yes” for efficient and “No” for inefficient. Efficiency was determined based on two criteria: (1) whether all questions on exam B were answered and (2) whether the time allocation for each question was reasonable. Reasonableness was determined by comparing the time distribution of all students for each question and setting the threshold as the shortest possible time for each question (specifically, the first 5% of the time distribution).

The independent variables included student ID, test type, a unique question number for each question, question type, specific behaviors of students during the answering process, additional information about their behaviors, and the time stamp of each answering behavior.

### *2.2 Procedure*

Throughout the 9-week course, the students utilized the Zoom meeting tool to engage in discussions twice a week, with each session lasting 50-60 minutes. The entire discussion process was recorded using the Zoom meeting feature. These recorded videos were manually transcribed to obtain a written record of each discussion.

### *2.3 Data Analysis*

The transcribed discussions resulted in a conversation document comprising a total of 82,003 words and 3,209 statements. A statement was defined as a complete conversation in which a student participated and contained at least one sentence.

The data analysis employed a bottom-up grounded theory method to derive a comprehensive understanding of the data. Initially, the conversation document was analyzed sentence by sentence using open coding to identify and extract several categories. Subsequently, specific categories were grouped together through axial coding to generate overarching themes. Finally, each category was further refined using selective coding to extract core concepts and capture the key content of the conversations.

To ensure the reliability of the coding and analysis, two researchers coded respectively under the guidance of professionals and got good consistency (Kappa=0.81). Furthermore,

data from different sources, such as the discussion videos and the team's final project report, were also integrated into the category analysis. This additional data served as a cross-reference and validation for the coding process.

## 3. Findings

### *3.1 Overall Status*

To depict the manifestation of university students, the coding results are shown in Table 1. By coding the conversation document sentence by sentence, and with repeated reading, comparison and modification, the whole conversation document was finally divided into 41 activity segments, each representing a complete activity discussion among the students in the group, such as understanding the task requirements and dealing with missing values. After removing repeated activities, 19 first-level codes were identified through open coding (see Table 1). Axial coding was then conducted to analyze the categories based on the 19 first-level codes and establish relationships between them. The first-level codes were organized chronologically, and their relationships in terms of causality, situation, similarity, difference, function, and process were considered, resulting in the formation of 8 second-level codes (see Table 1). Finally, selective coding was performed, and all categories were classified into corresponding core categories that encompassed all the activities in the session document. Five core categories, referred to as third-level codes, were extracted: understanding tasks, organizing data, selecting variables, selecting sampling methods, and selecting modeling methods (see Table 1).

Table 1. *Coding Results of Bottom-Up Grounded Theory*

| First-level Codes | Second-level Codes | Third-level Codes |
|---|---|---|
| Understanding the task | Understanding the task | Understanding the task |
| Assigning phased task | | |
| Preparing software environment | Preparing software environment | |
| Dealing with missing values | Preprocessing data | Organizing data |
| Exploring preprocessing data | | |
| Preprocessing data | | |
| Reflecting on preprocessing data | | |
| Exploring understanding data | Understanding data | Selecting variables |
| Reflecting on understanding data | | |
| Exploring independent variables selection | Selecting independent variables | |
| Reflecting on independent variables selection | | |
| Exploring sampling methods | Selecting sampling method | Sampling |
| Understanding sampling | | |
| Reflecting on sampling methods | | |
| Exploring modeling process | Understanding modeling process | Modeling |
| Reflecting on modeling process | | |
| Selecting model | Selecting model | |
| Evaluating model | | |
| Optimizing model | | |

By summarizing the second-level and third-level codes, it was observed that the modeling reasoning process of university students in the educational data mining project followed a spiral pattern. This process involved the steps of defining the problem, generating

and selecting variables and attributes, selecting modeling and sampling methods, organizing and structuring data, exploring and analyzing data, interpreting data, and presenting the results (see Figure 1). The process went through several iterations, with the students continuously optimizing and improving their understanding of specific concepts, selection of independent variables, data processing methods, sampling techniques, and modeling methods over time, ultimately leading to a progressive spiral form.

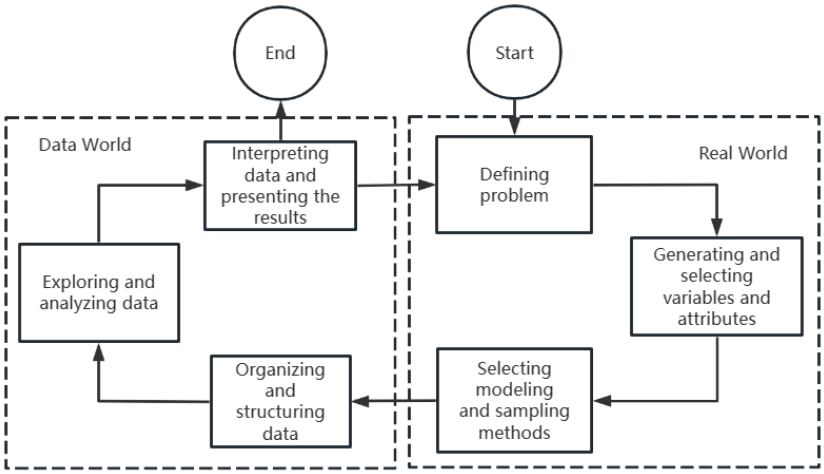


*Figure 1.* The spiral progress of modeling reasoning

### *3.2 Detailed Manifestation*

Since the goal of the data mining project was to build a prediction model for students' answer efficiency, the entire process of completing the project represented a comprehensive modeling reasoning process. Additionally, since the students engaged in discussions twice a week, it was evident from the analysis that they would review the task flow at the beginning of almost every weekly discussion and allocate tasks accordingly. This iterative nature of their discussions contributed to the spiral development of their modeling reasoning process, wherein each week built upon the progress made in the previous week. Therefore, to demonstrate the spiraling process of modeling reasoning, relevant conversation fragments that were important and non-repetitive were selected based on the chronological order of task completion.

#### *3.2.1 Understanding the meaning of the task*

After reading the task description document, the students (referred to as student X, student Y, and student Z) engaged in a discussion to accurately comprehend the task's purpose, define its objectives, and determine the analytical process and steps [2-104] (Numbers in square brackets denote the statement numbers of the conversation document). Since the process of understanding the meaning of the task was relatively fragmented, and their understanding of the task's purpose did not undergo significant changes thereafter, the project report is used as a demonstration. The students clarified the task's objectives as follows: "How to predict the answering efficiency of (another) group of students in exam B by analyzing the data of the existing group of students' answering behavior in exam A and the answering efficiency in exam B, with the ultimate goal of maximizing the accuracy of the prediction" [Project report, p1] (means the source of information is from page1 of the team project report). They also depicted the analysis process in Figure 2 [Project report, p2]. Although the students had a clear understanding of the general process, they encountered some difficulties and uncertainties in the execution and comprehension of specific steps. For instance, student Z expressed hesitation regarding the final result, questioning whether it is a probability value for a binary variable: "Yes, then what we finally get should not be, is a classification of 0 and 1, right? The probability of 0 is probably 40%, and the probability of 1 is 60%, right? That's probably the case, right?" [54]. However, the students demonstrated a

clear understanding of the general modeling process, enabling them to confirm and clarify related concepts, explore and select better methods guided by a clear analytical process, and ultimately develop a spiral modeling reasoning process.

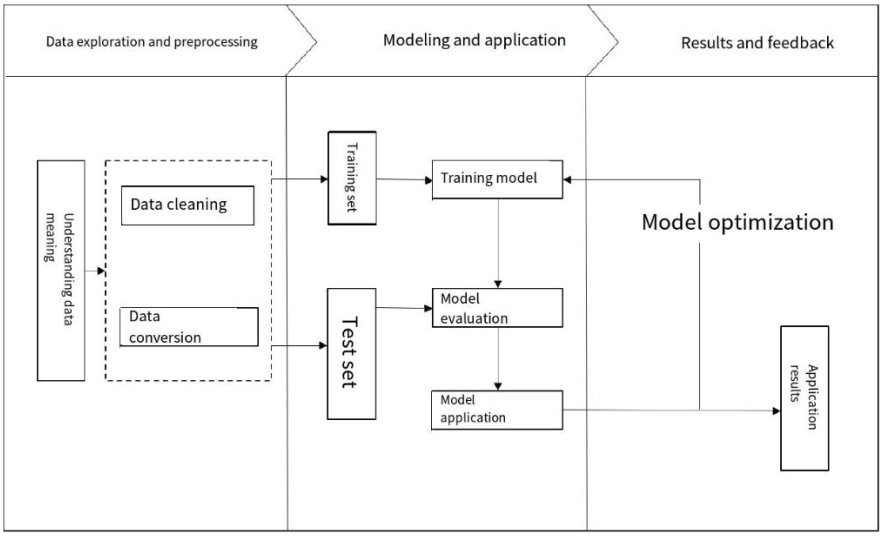


*Figure 2.* The analysis process made by the students

*3.2.2 Exploration of model selection*

During the initial exploration, the students contemplated the types of predictive models that should be built. Student Y suggested “examining numerous existing models” [111]. Student X proposed the idea of “an integrated model, considering the limitations of a single model's predictive performance” [114]. Student X exhibited a thorough understanding of integrated models, explaining the concept of re-modeling a single model and providing examples to elucidate the voting and stacking methods to other students [125,127]. This initial assumption served as a guiding principle for the subsequent modeling process.

*3.2.3 Understanding the data*

Comprehending the meaning of the data served as a prerequisite for variable selection and data processing. The students referred to the task description document to understand the variables in the dataset, their respective representations, and the type of data [201,204,211]. Student X raised concerns about missing values, although the discussion on how to handle them remained superficial, yet still acknowledged the importance of addressing this issue [201,203]. However, student Z displayed limited ideas regarding missing value treatment, suggesting their deletion without further elaboration [202].

*3.2.4 Data preprocessing*

The students encountered difficulties in selecting independent variables from the real dataset, even after understanding the data's meaning. Extracting information from the dataset proved challenging due to its complex presentation. The students actively engaged in discussions on data preprocessing, particularly regarding the conversion between long and wide formats. Student X demonstrated understanding of long data, explaining that each line represents a record of behaviors for each student [295]. Similarly, student Z expressed comprehension of wide data, noting the need to combine all the students to analyze "student" as a whole [297]. The mutual conversion between long and wide format data is a crucial step in data preprocessing for data mining projects, as it facilitates subsequent analysis and modeling by providing "tidy data" (Wickham, 2014).

### *3.2.5 Selection of independent variables*

The selection of independent variables was a crucial step in the modeling reasoning process as it bridged the understanding of data meaning and the establishment of the model, providing clues to connect the real world with the data world (Lehrer & Schauble, 2004). Through discussion, the students identified two main independent variables for predicting answering efficiency: "the answering time of each question" and "the number of questions answered within the specified time" [395]. Additionally, the students decided to initially include all variables and later choose the most influential ones through variable analysis [396]. This approach demonstrated the generation and selection of variables and attributes in the modeling reasoning process, incorporating problem definition, data understanding, and exploratory data analysis (EDA) to inform important modeling decisions.

### *3.2.6 Selection of sampling method*

The selection of a sampling method was an indispensable step in the modeling reasoning process, as it determined the use of samples in model construction. The students not only employed "principal component analysis" and "factor analysis" as part of EDA to select independent variables [778], but also demonstrated understanding and discussion of sampling methods. Student X initially perceived sampling as "data segmentation" and described the ten-fold cross-validation method as dividing the data into ten parts, repeatedly selecting one part as the test set while using the others as the training set [780]. After extensive discussions and trials, the students ultimately chose the ten-fold cross-validation sampling method with three repetitions based on the results of model evaluation and optimization [Project report, p4].

### *3.2.7 Selection of modeling method*

The selection of a modeling method directly influenced the modeling results. However, in educational data mining projects, specific modeling methods were often chosen based on the results of evaluation models. The students utilized data analysis software RStudio to convert complex probability formulas into specific codes, facilitating the implementation and analysis of different modeling techniques. The selection of modeling methods followed the principles of EDA, saving time and allowing the students to focus more on the data itself rather than the application of intricate statistical probability formulas. The students used RStudio and focused on understanding how the code was implemented when choosing the modeling method, considering the previously mentioned integrated modeling approach, specifically the voting and stacking methods [1145, 1151]. The choice of modeling method primarily considered the kappa value, which aligned with the EDA approach of focusing on the results and outcomes rather than becoming overly fixated on the modeling method itself [1149].

### *3.2.8 Model evaluation*

Model evaluation closely followed the model selection process and significantly impacted all previous decisions, including the choice of independent variables, sampling method, and modeling method. The evaluation of the model focused on two key metrics: kappa and AUC [Project report, p4]. During the evaluation, the students primarily paid attention to the kappa value, while AUC was not considered at this stage. Different modeling methods produced varying kappa values, allowing for a direct comparison of their strengths and weaknesses. For instance, "regression is 0.15, decision tree is 0.17, and random forest is also 0.17" [1724]. This comparison revealed that both random forest and decision tree methods performed similarly and better than the regression method.

### *3.2.9 Model Optimization*

Model optimization involved adjusting the input model data and parameters based on the evaluation results. The evaluation outcomes influenced the selection of independent variables and data preprocessing. The students attempted to optimize the model by adjusting its parameters, referred to as parameter adjustment. The presence of underfitting or overfitting was an important consideration during model optimization and closely tied to the model's parameters. Underfitting occurred when the model failed to predict results accurately, while overfitting referred to the model performing exceptionally well on sampled data but lacking generalization ability to apply to the entire population. The AUC value partly reflected the model's generalization ability. The students demonstrated an understanding of underfitting, overfitting, and generalization ability, engaging in discussions about these concepts: "It's under-fitting, so it lacks an ability to identify different independent variables, so he finally gets the same probability" [3076]; "Does underfitting mean no generalization ability? " [3080]; "Sure, it's too precise " [3081].

## 4. Discussion

Data literacy has emerged as a crucial component of 21st-century skills in the era of big data, encompassing various aspects such as modeling reasoning, data aggregation, context understanding, and data visualization (Rubin, 2019). Among these, modeling reasoning plays a significant role in establishing connections between the real world and the data world, enabling individuals to define, explain, and predict phenomena based on data and theories (Aridor & Ben-Zvi, 2017). However, there is limited understanding of how university students manifest modeling reasoning, which can impede the development of data science subjects in universities and hinder the cultivation of students' data literacy. Therefore, this study aimed to explore university students' manifestation of modeling reasoning through a case study approach.

The findings of this study reveal that the process of modeling reasoning among university students follows a spiral development form in educational data mining projects. This spiral form is observed due to the iterative and continuous optimization nature of modeling reasoning itself, as well as its role as the framework for data mining activities. The analysis process made by the students (see Figure 2) is divided into three parts: data exploration and preprocessing, modeling and application, results and feedback, which demonstrates their clear understanding of the general modeling process that enables the students to develop a spiral progress of modeling reasoning (see Figure 1). This spiral development form shares similarities with the integrated modeling approach (Bakker & Hoffmann, 2005), which also exhibits a spiral rise in its entirety. However, in the current study, university students utilized the idea of integration to link the real world and the data world at the initial stage of the modeling reasoning process, specifically during the task understanding phase. Consequently, the spiral progressive form was manifested as circular steps of modeling reasoning, accompanied by a deepening understanding of relevant concepts. Establishing the connection between data and probability is crucial in data mining activities (Konold & Kazak, 2008), which is also evident in this study through repeated understanding of sampling methods and repeated selection of modeling methods.

Teaching data science or big data remains an area with limited research (Saltz & Heckman, 2015). This study not only sheds light on the manifestation of university students' modeling reasoning but also provides insights into cultivating data literacy in universities. Project-based learning has been recognized as an effective teaching method in statistical education, promoting the development of statistical reasoning and data literacy (Ben-Zvi et al., 2018). Additionally, in the era of big data, exposing students to complex and real scientific datasets facilitates the development of statistical thinking and data literacy (Saltz & Heckman, 2015). The present study aligns with these findings by employing a project-based learning approach, providing students with real, unprocessed data to tackle authentic problems. This approach significantly enhances and develops students' modeling reasoning abilities (Konold & Kazak, 2008). Particularly in the establishment of data science programs in universities, the training of students' data literacy holds significant importance. While well-

structured problems and small dataset cases have proven effective in training students' data literacy in the past (Ben-Zvi et al., 2018; Koparan & Güven, 2015), the era of big data necessitates consideration of ill-structured problems and real, complex, and massive datasets within a project context when implementing project-based learning to cultivate university students' data literacy.

In this study, the utilization of RStudio, a data analysis software, allows students to explore the data analysis process through exploratory data analysis (EDA), transcending the limitations of understanding statistical formulas and probability models alone (Moore, 1997). While this approach enables students to focus on the analysis process in real data mining projects, it is important to investigate whether the excessive focus on result-oriented EDA may have adverse effects on the cultivation of students' data literacy. Furthermore, computational thinking holds significant importance in modeling reasoning activities. In the educational data mining project examined in this study, the field of data science encompasses not only knowledge of mathematical statistics and specific domain knowledge (e.g., education-related knowledge), but also knowledge of computer programming (Rosenberg et al., 2019). Given the reliance on data analysis software RStudio, understanding and proficiently utilizing coding is crucial during the programming process. The students encountered difficulties and had to halt the modeling reasoning process due to unfamiliar code, emphasizing the need for computational thinking. This study highlights the importance of preparing the software environment before initiating the modeling reasoning process among university students. Consequently, further research is warranted to explore the relationship between computational thinking and modeling reasoning.

## 5. Conclusion

In conclusion, this study contributes to the understanding of university students' manifestation of modeling reasoning, shedding light on the cultivation of data literacy in higher education. The findings reveal that the process of modeling reasoning among university students follows a spiral development form in educational data mining projects. Besides, it is suggested to apply project-based learning and the utilization of real, complex, and massive datasets in developing students' statistical thinking and data literacy. However, the potential impacts of result-oriented EDA on students' data literacy and the significance of computational thinking in modeling reasoning warrant further investigation. By addressing these areas, educators can enhance their pedagogical approaches to effectively cultivate students' data literacy in the era of big data.

## Acknowledgements

This essay was funded by the Special Fund for Postgraduates of East China Normal University to Participate in International Conferences. I would like to express my sincere gratitude to Associate Professor Bian Wu for his meticulous guidance and support throughout this research project.

## References

Aridor, K., & Ben-Zvi, D. (2017). The Co-Emergence of Aggregate and Modelling Reasoning. *Statistics Education Research Journal, 16*, 38-63.

Bakker, A., & Hoffmann, M.H. (2005). Diagrammatic Reasoning as the Basis for Developing Concepts: A Semiotic Analysis of Students' Learning about Statistical Distribution. *Educational Studies in Mathematics, 60*, 333-358.

Ben-Zvi, D., Makar, K., & Garfield, J. (2018). *International handbook of research in statistics education*. Cham: Springer.

Biehler, R., Frischemeier, D., & Podworny, S. (2017). Elementary Preservice Teachers' Reasoning about Modeling a "Family Factory" with TinkerPlots--A Pilot Study. *Statistics Education Research Journal, 16*, 244-286.
Conway, B., Martin, W. G., Strutchens, M.E., Kraska, M.F., & Huang, H. (2019). The Statistical Reasoning Learning Environment: A Comparison of Students' Statistical Reasoning Ability. *Journal of Statistics Education, 27*, 171 - 187.
Dvir, M., & Ben-Zvi, D. (2018). The role of model comparison in young learners' reasoning with statistical models and modeling. *ZDM, 50*, 1183-1196.
Gould, R.G. (2017). Data Literacy is Statistical Literacy. *Statistics Education Research Journal, 16*, 22-25.
Gravemeijer, K.K. (1999). How Emergent Models May Foster the Constitution of Formal Mathematics. *Mathematical Thinking and Learning, 1*, 155-177.
Konold, C.E., & Kazak, S. (2008). Reconnecting Data and Chance. *Technology Innovations in Statistics Education, 2*.
Koparan, T., & Güven, B. (2015). The effect of project-based learning on students' statistical literacy levels for data representation. *International Journal of Mathematical Education in Science and Technology, 46*, 658 - 686.
Lehrer, R., & English, L. (2018). Introducing children to modeling variability. In D. Ben-Zvi, J. Garfield, & K. Maka (Eds.), International handbook of research in statistics education (pp. 229-260). Cham: Springer.
Lehrer, R., & Schauble, L. (2004). Modelling natural variation through distribution. *American Educational Research Journal, 41*(3), 635-679.
Lehrer, R., & Schauble, L. (2010). What Kind of Explanation is a Model. In M. K. Stein (Eds.), Instructional explanations in the disciplines (pp. 9-22). New York, US: Springer.
Moore, D.S. (1997). New Pedagogy and New Content: The Case of Statistics. *International Statistical Review, 65*.
Rosenberg, J.M., Lawson, M.A., Anderson, D., Jones, R.S., & Rutherford, T. (2019). Making Data Science Count In and For Education. *Research Methods in Learning Design and Technology*.
Rubin, A. (2019). Learning to Reason with Data: How Did We Get Here and What Do We Know? *Journal of the Learning Sciences, 29*, 154-164.
Saltz, J., & Heckman, R. (2015). Big Data science education: A case study of a project-focused introductory course[J]. *Themes in Science & Technology Education, 8*(2), 85-94.
Setiawan, E. P., & Sukoco, H. (2021). Exploring first year university students' statistical literacy: A case on describing and visualizing data. *IndoMS-Journal on Mathematics Education, 12*(3), 427-448.
Wickham, H. (2014). Tidy data. *Journal of statistical software, 59*(10),1-23.
Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry (with discussion). *International Statistical Review, 67*(3), 223-248.