

Teaching Case - Student Materials

Evaluating Generative AI for Data Cleaning: A Case in Quality Assurance and Critical Thinking

Benjamin Larson
blarson@aum.edu

Pawan Chandu Veldi
pveldi@aum.edu

Prathyusha Vemula
pvemula1@aum.edu

Jeffrey Bohler
jbohler1@aum.edu

Auburn University at Montgomery
Montgomery, AL, 36124, USA

Hook

Generative AI can clean a dataset in seconds, but can you trust every decision it makes? This case challenges students to critically evaluate AI-assisted data preparation using the CRISP-DM methodology and to apply quality assurance principles to determine whether the resulting data are truly ready for analysis.

Abstract

Data preparation consumes much of the effort in an analytics project, which makes it an attractive place to apply generative artificial intelligence (AI). It is also a place where uncritical use is expensive, because a model that cleans a dataset confidently can produce a dataset that looks finished but is wrong. This teaching case asks students to clean the same retail transaction dataset twice. Students first upload the raw file and ask an AI system to clean it with no additional context. They then re-prompt with a data dictionary and documented business rules, compare the two outputs, and defend which they would trust. A final section requires an independent quality assessment against the Cross-Industry Standard Process for Data Mining (CRISP-DM) Data Understanding and Data Preparation deliverables, concluding with a judgment on whether the data will support a time-series forecast. The case has been administered in two sections of graduate students (25 and 13 students, respectively); the design reported here reflects revisions made after students in the initial offering accepted invalid AI reasoning despite being warned to verify it, a pattern consistent with the automation bias literature. The case requires no prior AI knowledge, uses only descriptive statistics as a prerequisite, and provides a dataset, a data dictionary, a notes handout, an assignment, and a grading rubric.

Keywords: generative AI, data preparation, CRISP-DM, data quality, automation bias, critical thinking

Evaluating Generative AI for Data Cleansing: A Case in Quality Assurance and Critical Thinking

Benjamin Larson, Pawan Chandu Veldi, Prathyusha Vemula and Jeffrey Bohler

1. INTRODUCTION

Data mining traditionally requires the largest share of project time for data acquisition and preparation (Munson, 2012). This may lead teams to turn to large language models (LLMs) or agentic AI to speed up the process, but they require a systematic approach. Without understanding the problem's context and adequate descriptions, LLMs or agents may not clean or prepare the data appropriately. To support this effort, humans need to apply a systematic approach to understanding context and needs and use critical thinking to evaluate the results. The systematic approach we use in this case is the Cross-Industry Standard Process for Data Mining (CRISP-DM). However, other similar structures, such as Cognitive Project Management for AI (CPMAI), may be used as appropriate, depending on the project type. In most quantitative methods courses, students are given tidy, pre-cleaned spreadsheets or datasets that are already suitable for analysis. By contrast, in real business settings, analysts usually receive raw, messy data that must be cleaned before proper analysis (Snyder, 2019). Some teaching cases focus on data cleaning (Collier, 2024; Yue, 2012); however, this case differs by including the CRISP-DM framework and engaging an LLM as a partner in data cleaning. While some evidence shows LLMs can perform data preparation using CRISP-DM, human oversight is still needed (Musazade et al., 2024).

Quality assurance and critical thinking matter across all six CRISP-DM phases, but they bind most tightly in Data Understanding and Data Preparation, where this case sits. The decisions made there about what to keep, what to repair, and what to discard will propagate into every model built downstream, and they are effectively invisible by the time anyone looks at a forecast.

This case represents the data-cleaning portion of that process, performed in the Data Preparation phase. However, it is also related to the other phases of CRISP-DM. The assignment engages students in cleaning and evaluating a retail dataset, helping them become familiar with the business problem as they learn to apply specific steps in the process.

2. CASE

2.1 Case Description

Merger Grocery is a grocery store chain formed by combining several locations across the southeastern United States. The company sells products across several departments, including Bakery, Meat, Dairy, Produce, Beverages, and Household Goods. The company needs to improve inventory planning and reduce food waste by implementing time-series forecasts to predict future product demand. We used a synthetically produced grocery store dataset with seeded quality issues because students are familiar with its concrete consequences.

The data management team has extracted several years of historical point-of-sale transactions from all retail management systems across the stores. Each transaction includes store location, customer, product category, item purchased, price, quantity sold, total amount spent, payment method, sale date, and coupon usage. Despite the team's best efforts, you are left with the data following(a dataset, whose sample appears in Appendix A) that appears to have issues that need to be addressed before it can be used. The case tasks the individual or team with preparing and evaluating data quality for performing a time-series forecast.

2.2 Target Audience and Prerequisites

This case is designed for undergraduate or graduate students in a data mining, data wrangling, machine learning, or data analytics course. Students should have some knowledge of analytics and statistics,

including a basic understanding of descriptive statistics. Students should also have some basic knowledge of data types and a deeper understanding of the first three phases of the CRISP-DM process. This knowledge may have been acquired in another course or in the current course. No specific knowledge related to AI is required for the assignment. We recommend two to four one-hour lectures on these topics and provide slides to support our instructional materials.

2.3 Learning Objectives

Upon completion of this case, students will be able to:

1. Apply the Data Understanding and Data Preparation deliverables of CRISP-DM to a transactional dataset.
2. Construct and improve a data-cleaning prompt that supplies business context, a data dictionary, and explicit cleaning rules.
3. Evaluate AI-generated cleaning decisions against documented business rules and descriptive statistics, identifying both errors of omission and errors of commission.
4. Compare AI outputs produced under different prompt conditions and justify which is more trustworthy using evidence drawn from the data.
5. Judge whether a prepared dataset is fit for time-series forecasting and defend that judgment.

2.4 Supplied Instructional Materials

Instructional materials are available on request:

1. Synthetically generated dataset (an example is presented in Appendix A)
2. Slides for CRISP-DM Phases 1–3
3. Data dictionary of the dataset (Appendix B)
4. Notes handout for data preparation (Appendix C)
5. Assignment (Appendix D)

2.5 Implementation

This assignment may be implemented in several different ways. The assignment can be completed individually or in a group. If the assignment is given as a group, each student should prompt the AI system individually; but group members may then discuss the results within their group and turn in the assignment as a group. The assignment may be worked on in a flipped classroom or given as a take-home assignment. Table 1 provides an estimate of the time needed for a traditional classroom.

Assignment Part / Lecture	Description	Time
Assignment Part 1: Prompt	Prompt the AI to clean the data.	5 minutes
Assignment Part 2: Summary Statistics	Run descriptive statistics.	20 minutes
Assignment Part 2: Discussion	Discuss in class.	25 minutes
Data Preparation Lecture	Fill out the notes provided in Appendix C.	50 minutes
Assignment Part 3: Improve the Prompt	Allow time to integrate lecture notes into the prompt (if in class).	30 minutes
Assignment Part 4: Compare the Results	Run descriptive statistics.	10 minutes
Assignment Part 4: Discussion	Discuss differences in prompt results.	40 minutes
Assignment Part 5: Independent Data Quality Assessment	Discuss and report on quality issues.	40 minutes

Table 1. Time Estimates for a Traditional Classroom

3. CRISP-DM

The Cross-Industry Standard Process for Data Mining (CRISP-DM) is one of the most widely adopted methodologies for conducting data mining and analytics projects because it provides a structured, domain-independent framework for transforming business problems into data-driven solutions (Chapman et al., 2000; Wirth & Hipp, 2000). The methodology consists of six iterative phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment, which

collectively guide the analytics lifecycle. Unlike traditional sequential development approaches, CRISP-DM recognizes that data mining is an iterative process in which discoveries made in later phases frequently require analysts to revisit earlier phases to improve data quality, redefine business objectives, or refine analytical methods (Bošnjak et al., 2009; Shearer, 2000).

Each phase in CRISP-DM has relationships with both prior and subsequent phases. For example, Business Understanding defines the criteria for success, objectives, resources, and project planning. It is grounded in business understanding; the data collected, described, and explored in Data Understanding leads to Data Preparation and Modeling. Figure 1 shows that if an issue or new insight arises during Data Preparation or Modeling, any prior phase may be revisited to address the concern or capitalize on an opportunity.

CRISP-DM remains widely used more than two decades after its release, though not uncritically. Martínez-Plumed et al. (2021) document its continued adoption, argue that its linear phase structure poorly describes contemporary data science work, and propose recasting it as non-linear data science trajectories. We adopt CRISP-DM here for a narrower reason than fidelity to expert practice: its explicit phase boundaries give novices a vocabulary for stating where a problem originated and where it has to be fixed. That is a pedagogical justification, not a claim that the six phases describe how experienced teams actually work.

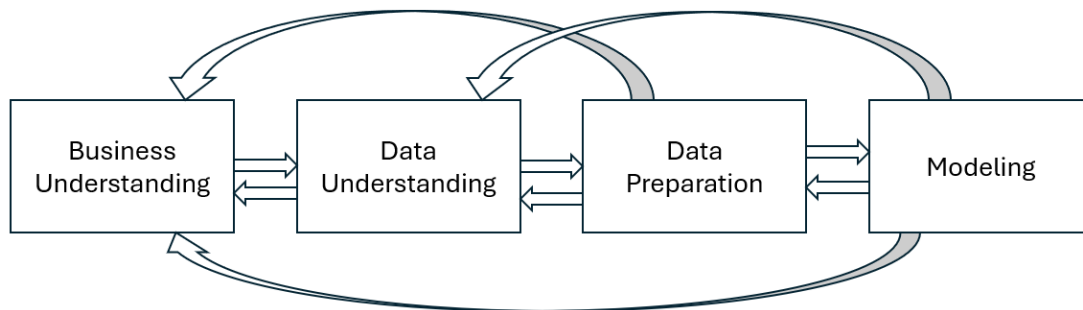


Figure 1. CRISP-DM Phases 1–4 Feedback

Among these phases, Data Preparation is critical because it transforms raw, heterogeneous, and often incomplete data into datasets suitable for analysis. Bošnjak et al. (2009) and Purbasari et al. (2021) report that data quality directly influences model performance, decision-making, and overall project success. Zhang et al. (2003) describe Data Preparation as the bridge between raw data and knowledge discovery. Martínez-Plumed et al. (2021) report that data preparation remains the phase where practitioners spend the most effort and iterate most often, even as they question whether CRISP-DM's phase model is the right way to describe that work. Kannengiesser and Gero (2023) make a similar point from a design-process perspective: the effectiveness of every subsequent phase depends on the quality, completeness, and usability of the prepared dataset.

The Data Preparation phase consists of multiple interrelated activities that convert raw operational data into datasets suitable for analytical modeling. The original CRISP-DM framework identifies five primary activities: data selection, data cleaning, data construction, data integration, and data formatting (Chapman et al., 2000). Although these activities are presented separately, Bošnjak et al. (2009) and Martínez-Plumed et al. (2021) describe them as iterative tasks refined throughout the project lifecycle rather than completed in a single sequence.

Data selection involves identifying relevant variables and records that support the project's business objectives while eliminating unnecessary information that may increase analytical complexity. Once relevant data have been identified, data cleaning addresses missing values, duplicate records, inconsistent formats, inaccurate entries, noise, and outliers that could negatively influence subsequent analyses (Zhang et al., 2003). Initial identification relies heavily on having the data collected, stored, evaluated, and defined during the Data Understanding phase, as tools such as data dictionaries inform preparation about data accuracy and relevance.

Assessment of the Data Preparation phase extends beyond verifying whether data have been cleaned or transformed successfully. Within the CRISP-DM framework, Data Preparation is evaluated by determining whether the activities performed during this phase produce datasets that are accurate, complete, consistent, and suitable for subsequent analytical tasks, while supporting the project's overall business objectives (Chapman et al., 2000; Wirth & Hipp, 2000). Rather than relying on a single measure, Zhang et al. (2003) and Purbasari et al. (2021) assess Data Preparation using established data quality dimensions, including accuracy, completeness, consistency, validity, uniqueness, and timeliness. These dimensions provide a structured approach to determining whether preparation activities have corrected errors, addressed missing values, standardized formats, and improved overall dataset quality.

Machine-learning and data-mining projects can be represented through different process frameworks, each emphasizing aspects of the analytics lifecycle. Knowledge Discovery in Databases (KDD) and Sample, Explore, Modify, Model, and Assess (SEMMA) are two alternative approaches. KDD emphasizes transforming raw data into useful knowledge, whereas SEMMA emphasizes the sequence of sampling, exploration, modification, modeling, and assessment. CRISP-DM extends the process beyond the technical modeling activities by explicitly incorporating business understanding, data understanding, data preparation, modeling, evaluation, and deployment (Azevedo & Santos, 2008).

The distinction is particularly relevant from a systems-thinking perspective. In a machine-learning project, the model is only one component of a larger system. The definition of the analytical problem influences the data collected and examined; data preparation influences the information available to the model; model outputs influence evaluation; and evaluation influences decisions about whether and how to use the analytical solution. CRISP-DM represents these relationships through connected phases rather than focusing exclusively on the modeling component.

Compared with KDD, CRISP-DM makes the relationship between the analytical process and the organizational or business problem more explicit. Compared with SEMMA, CRISP-DM provides a broader lifecycle by including activities before the modeling process and the deployment of analytical results. These differences make CRISP-DM particularly useful when the objective is to examine an analytics project as a connected system in which decisions at one stage can create consequences at later stages.

Martínez-Plumed et al. (2021) explain that contemporary data-science work has evolved beyond the traditional process representation and may involve flexible trajectories and repeated movement between activities. Therefore, the value of CRISP-DM lies not in requiring every project to follow the same order, but in providing a recognizable representation of the major components and relationships within an analytics process.

Accordingly, CRISP-DM can be distinguished from alternative frameworks by the breadth of its lifecycle perspective. KDD primarily emphasizes knowledge discovery, SEMMA concentrates on the data-mining workflow, and CRISP-DM explicitly links problem context, data, preparation, modeling, evaluation, and deployment. This comparison provides a systems-oriented basis for positioning CRISP-DM within the broader landscape of machine-learning and data-mining process frameworks. While we focus on the CRISP-DM framework, it is important to note that, whether using CRISP-DM or other processes, obtaining, selecting, and cleaning data are integral parts of analytics and must be assessed critically.

4. CRITICAL THINKING

Critical thinking is a purposeful and self-regulatory process of judgment that involves interpreting, analyzing, evaluating, and drawing inferences from available evidence to make reasoned decisions. According to Facione (1990), critical thinking encompasses six core cognitive skills: interpretation, analysis, evaluation, inference, explanation, and self-regulation. These skills help people solve complex problems, assess evidence objectively, and justify conclusions with logical reasoning. Rather than accepting information at face value, critical thinkers question assumptions, assess evidence credibility, compare alternative solutions, and justify their decisions with reasoned judgment.

Developing information quality professionals requires integrating theoretical knowledge with practical problem-solving skills (Lee et al., 2007). Whether the ability to interpret AI-generated output is sufficient to produce warranted trust in it is less settled than it may appear. In their study of AI adoption in higher

education, Sabri et al. (2026) found that data quality predicted the interpretability of AI-generated knowledge. However, interpretability did not significantly influence ethical trust, and trust did not directly predict adoption success, a pattern they term the trust-interpretability paradox. Their result matters for how we teach this. If showing the model's reasoning alone does not calibrate a user's trust, then it is not enough to ask students to read an AI's explanation and judge it. They need a procedure for testing that explanation against the data. This case provides that procedure.

In using generative AI for data preparation, critical thinking is essential because AI-generated recommendations may contain inaccuracies, hallucinations, or bias. Users must evaluate the accuracy, relevance, and appropriateness of AI-assisted logic before applying it to data-preparation tasks. For trainers or instructors who allow learners to use generative AI as a tool or scaffolding for data preparation, it is important to incorporate reflection and to guide the prompting process to foster ethical use (Addiucci & Temperini, 2026; Aldulaijan & Almalki, 2025). Gimpel et al. (2025) argue that ChatGPT can help achieve learning goals and train students to question outputs, check claims, and iteratively refine prompts.

In our case, we want to promote critical thinking to systematically interpret, analyze, evaluate, and validate AI-generated recommendations, supporting accurate, transparent, evidence-based data-cleansing decisions. For educators, this may translate into a rubric with five key components: evaluating evidence that may support the decision, analyzing the evidence's meaning, synthesizing the evidence with a problem, forming the argument in a structured manner, and making a valid argument in the discipline context (Reynders et al., 2020). Learners should evaluate whether the AI has identified erroneous data, interpreted the issue correctly so it can be fixed, and explained the process used in a manner consistent with the problem's context.

5. AUTOMATION BIAS

The tendency to accept an automated system's recommendation without adequate verification is well documented and predates generative AI by decades. Parasuraman and Riley (1997) distinguish misuse of automation (overreliance on an imperfect system) from disuse the rejection of a system that would help) and locate both in a mismatch between what operators expect of automation and what it actually delivers. Skitka et al. (1999) demonstrated the cost directly. In a simulated flight task, participants working without an automated aid outperformed participants given a highly but imperfectly reliable one on a monitoring task. The aid did not simply fail to help. It degraded the human's own vigilance.

This is the failure mode that a data preparation task with generative AI invites. An AI system that cleans a dataset produces output with two properties that suppress scrutiny: it is fluent and complete. No blank cells remain to signal that something needs attention. A student looking at a cleaned file has no visual cue to distinguish a well-reasoned imputation from a confident error, and the model's explanation of its reasoning does not reliably provide one, as Sabri et al.'s (2026) finding on interpretability suggests.

Warning students to be critical does not appear to be sufficient. Buçinca et al. (2021) found that cognitive forcing functions, design interventions that interrupt the user and require deliberate engagement before an AI recommendation can be accepted, reduced overreliance on AI relative to standard explainable-AI interfaces, which presented reasoning but left the decision to accept it frictionless. The distinction is between showing someone the reasoning and requiring them to act on it.

That distinction shaped this case. Parts 3 through 5 of the assignment are cognitive forcing functions in the sense described by Buçinca et al. (2021). Part 3 requires the student to provide the business rules before re-prompting, meaning that articulating the constraints of the correct answer before seeing the model's answer. Part 4 requires an explicit comparison, so you cannot overlook any difference between the two outputs. Part 5 requires an independent audit conducted on the assumption that the AI is wrong. None of these can be completed by reading and nodding.

6. DATA QUALITY ASSURANCE

Quality assurance refers to the systematic activities implemented to ensure that products, services, or systems consistently meet specified quality requirements. Unlike traditional software, AI models may generate different outputs for similar inputs, making continuous evaluation and monitoring essential to

ensure trustworthy, reliable, and responsible AI systems. Quality assurance is critical because AI systems are probabilistic rather than deterministic.

Within our case, the key aspect we need to evaluate is data quality. Wang et al. (2024) identify five dimensions to focus on: completeness, accuracy, consistency, timeliness, and relevance. These data quality concepts must be defined and used throughout the CRISP-DM process. In the Business Understanding phase, set objectives to inform the need for timely data, its relevance, and the data required to form a complete picture. Relevance may change as variables are found to be insignificant, and model evaluation may identify gaps in completeness. Knowledge of accuracy and consistency should be derived from a combination of the Business Understanding and Data Understanding phases and should be set to a reasonable level through critical assessment.

7. PROMPT ENGINEERING

The usefulness of an AI-generated result depends heavily on how the task was specified. In a systematic review of prompting in education, Qian (2025) separates prompting strategies into two families: technique-based approaches, which target a specific output goal, and process-based approaches, which support cognitive engagement and collaborative thinking with the model. This case works in the second family. Rather than treating prompt engineering as a search for the one prompt that produces the right answer, we treat it as a loop in which the student supplies context, inspects the model's stated reasoning, and revises, with the student, not the model, holding the decision. Once that logic stabilizes, it can be documented and reused on similar datasets. Figure 2 illustrates the loop.

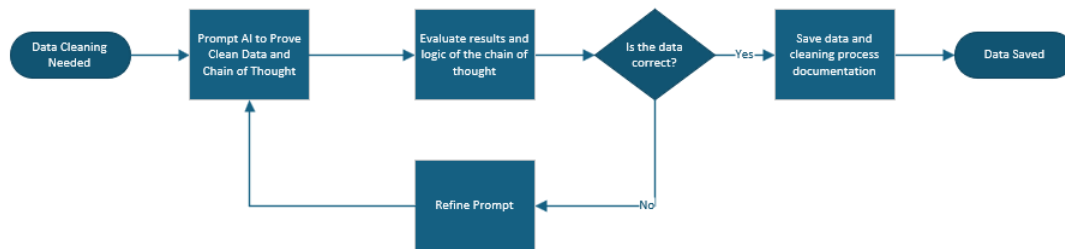


Figure 2. Cleaning Data Using Process-Based Prompt Engineering

Asking the AI to report its chain of thought and the steps it used, and then adjusting the corresponding steps, can make the data-cleaning process a collaborative chain-of-thought process (Yoo, 2025). Table 2 can be used as a template for documenting the steps taken in the data-cleaning process, and the number of instances should also be recorded. By asking the AI to provide such a table in its chain of thought, the human evaluator can assess whether the process steps make sense in the context of the problem and use the table to refine the original prompt.

Attribute	Situation	Algorithm
Total Spent	Total Spent value is null or 0, and Price and Quantity > 0	Total Spent = Price * Quantity
Total Spent	Total Spent value is null or 0, and Price or Quantity is null or 0	Remove record
Price	Price value is null or 0, and Total Spent and Quantity > 0	Price = Total Spent / Quantity
Price	Price value is null or 0, and Total Spent or Quantity is null or 0	Remove record

Table 2. Data Cleaning Chain of Thought Documentation

8. ASSIGNMENT

8.1 Assignment Development

The assignment (Appendix D) is broken down into five parts:

Part 1 is for students to upload data from an Excel file and prompt generative AI to clean it, providing the rationale for each action.

Part 2 is for students to run descriptive statistics and discuss the accuracy and logic of the AI's decisions.

Part 3 asks students to modify the prompt by providing context, such as the data dictionary, and apply the data-cleaning steps from the data preparation lecture.

Part 4 asks students to run descriptive statistics and discuss differences between the two results.

Part 5 requires students to provide a summary of the remaining quality issues.

Table 3 maps each learning objective to the assignment part and rubric line that assesses it.

Assignment Learning Objectives:

1. Apply the Data Understanding and Data Preparation deliverables of CRISP-DM to a transactional dataset.
2. Construct and improve a data-cleaning prompt that supplies business context, a data dictionary, and explicit cleaning rules.
3. Evaluate AI-generated cleaning decisions against documented business rules and descriptive statistics, identifying both errors of omission and errors of commission.
4. Compare AI outputs produced under different prompt conditions and justify which is more trustworthy using evidence drawn from the data.
5. Judge whether a prepared dataset is fit for time-series forecasting and defend that judgment.

Assignment	Learning Objective Assessed	Rubric Line
Part 1	2. Construct and improve a data-cleaning prompt	Created initial prompt, AI output, and conversation transcript
Part 2	3. Evaluate AI-generated cleaning decisions	Evaluation of the first AI response
Part 3	2. Construct and improve a data-cleaning prompt	Improved prompt with data dictionary and business rules
Part 4	4. Compare AI outputs	Comparison of AI outputs
Part 5	1. Apply Data Understanding and Data Preparation	Independent data quality assessment
Part 5	5. Judge and defend dataset fitness	Independent data quality assessment

Table 3. Learning Objectives Assessment Mapping

Note that LO1 cannot be fully assessed until Part 5 is completed.

8.2 Tool Dependence and Reproducibility

The pilot sections used ChatGPT (GPT-4o, accessed September–November 2025). The tool used matters because the contrast between Part 1 and Part 3 depends on the model making mistakes when given no context, and model capability on tabular tasks is advancing quickly. A model without code execution and a model with a Python sandbox behave differently on this dataset. Narayan et al. (2022) found that foundation models achieve state-of-the-art performance on several data cleaning and integration tasks despite not being trained for them, and capabilities have advanced since then.

Instructors should expect variation on three axes: between models (the assignment permits ChatGPT, Copilot, Claude, or Gemini), between runs of the same model, and across model versions over time. We treat this as a feature rather than a defect. The Part 2 discussion question asking whether the results are the same within a group makes the variance visible, and non-determinism in a tool students are being asked to trust is itself a quality assurance finding worth ten minutes of class discussion.

If the model cleans the dataset well on the naive prompt, the case still works, but the emphasis shifts. Parts 1 through 4 become a comparison of reasoning rather than of outcomes: students examine whether the model's stated rationale for a correct repair is actually sound, or whether it arrived at a

defensible answer by an indefensible route, for example, by applying $\text{Total Spent} = \text{Price} * \text{Quantity}$ to a record whose Price is the anomalous field. Part 5 is unaffected. An independent audit of a clean dataset is the same professional task as an audit of a dirty one, and concluding that the data is fit for purpose with evidence is a legitimate result.

Instructors adopting this case should run the Part 1 prompt against their intended model before assigning it and consult the instructor materials for the errors the model missed in our sections.

8.3 Discussion Questions

After Part 2, students evaluate the first AI response. Suggested questions include:

- Are the results the same within your group?
- Which data quality issues were correctly identified?
- Which issues were overlooked?
- Did the AI make unsupported assumptions?
- Were any values incorrectly changed?
- Did the AI introduce new errors?
- Were missing values handled appropriately?
- Were calculated fields validated?
- Could you trust this cleaned dataset for forecasting?

After Part 4, students compare the two AI outputs. Suggested questions include:

- Which prompt produced better results?
- What additional issues were identified?
- Which recommendations have changed?
- How did additional business knowledge improve the output?
- Which version would you trust more?

9. CONCLUSION

Results from the initial offering were instructive and mixed. Students were explicitly warned to evaluate the AI's output, and many nonetheless accepted at least one invalid repair, most commonly allowing Total Spent to be computed as zero because the Price field was blank or zero. The warning was not the problem; the absence of a task that made accepting the output effortful was. This is consistent with Skitka et al. (1999) and with Bućinca et al.'s (2021) finding that explanation without friction does not reduce overreliance.

That observation drove the current design. Parts 3 through 5 were added so that students would have to state the business rules before seeing the model's answer, compare two outputs rather than judge one in isolation, and audit the result independently. Students in the revised offering reported that the comparison in Part 4 was where the AI's errors became visible to them. We report these as instructor observations, not as measured outcomes; a controlled assessment comparing the two designs is the obvious next study and one we intend to run.

The broader point is one that the case makes better than a lecture can. Data preparation with generative AI is not a delegation problem. It is a supervision problem, and supervision requires knowing what the right answer looks like before you ask.

10. REFERENCES

- Addiucci, L., & Temperini, M. (2026). Use of prompt engineering in information teaching and learning in secondary education: A literature review and research perspective. *Journal of Information Technology Education: Research*, 25, Article 18. <https://doi.org/10.28945/5773>
- Aldulaijan, A. T., & Almalky, S. M. (2025). The impact of generative AI tools on postgraduate students' learning experiences: New insights into usage patterns. *Journal of Information Technology Education: Research*, 24, Article 3. <https://doi.org/10.28945/5428>

- Azevedo, A., & Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: A parallel overview. In *Proceedings of the IADIS European Conference on Data Mining, 2008* (pp. 182–185). IADIS.
- Bošnjak, Z., Grljević, O., & Bošnjak, S. (2009). CRISP-DM as a framework for discovering knowledge in small- and medium-sized enterprises' data. In *Proceedings of the 5th International Symposium on Applied Computational Intelligence and Informatics* (pp. 509–514). IEEE. <https://doi.org/10.1109/SACI.2009.5136302>
- Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), Article 188. <https://doi.org/10.1145/3449287>
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. CRISP-DM Consortium.
- Collier, C. A. (2024). Teaching Case: Cleaning House: A Case Teaching Data Cleaning Using Real-World Zillow Real Estate Data. *Journal of Information Systems Education*, 35(4), 456–460. <https://doi.org/10.62273/OQXC8468>
- Facione, P. A. (1990). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction* (The Delphi Report). American Philosophical Association.
- Gimpel, H., Hall, K., Decker, S., Eymann, T., Gutheil, N., Lämmermann, L., Braig, N., Maedche, A., Röglinger, M., Ruiner, C., Schoch, M., Schoop, M., Urbach, N., & Vandirk, S. (2025). Using generative AI in higher education: A guide for instructors. *Journal of Information Systems Education*, 36(3), 237–256. <https://doi.org/10.62273/QLLG7172>
- Kannengiesser, U., & Gero, J. S. (2023). Modelling the design of models: An example using CRISP-DM. *Proceedings of the Design Society*, 3, 2705–2714. <https://doi.org/10.1017/pds.2023.271>
- Lee, Y. W., Pierce, E., Talburt, J., Wang, R. Y., & Zhu, H. (2007). A curriculum for a Master of Science in Information Quality. *Journal of Information Systems Education*, 18(2), 233–242. <https://jise.org/Volume18/n2/JISEv18n2p233.html>
- Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernández-Orallo, J., Kull, M., Lachiche, N., Ramírez-Quintana, M. J., & Flach, P. (2021). CRISP-DM twenty years later: From data mining processes to data science trajectories. *IEEE Transactions on Knowledge and Data Engineering*, 33(8), 3048–3061. <https://doi.org/10.1109/TKDE.2019.2962680>
- Munson, M. A. (2012). A study on the importance of and time spent on different modeling steps. *ACM SIGKDD Explorations Newsletter*, 13(2), 65–71. <https://doi.org/10.1145/2207243.2207253>
- Musazade, N., Mezei, J., & Wang, X. (2024). Exploring the Performance of Large Language Models for Data Analysis Tasks Through the CRISP-DM Framework. In Á. Rocha, H., Adeli, G., Dzemyda, F., Moreira, & A. Poniszewska-Marańda (Eds.), *Good Practices and New Perspectives in Information Systems and technologies* (Lecture Notes in Networks and Systems, Vol. 989, pp. 56–65). Springer. https://doi.org/10.1007/978-3-031-60227-6_5
- Narayan, A., Chami, I., Orr, L., & Ré, C. (2022). Can foundation models wrangle your data? *Proceedings of the VLDB Endowment*, 16(4), 738–746. <https://doi.org/10.14778/3574245.3574258>
- Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2), 230–253. <https://doi.org/10.1518/001872097778543886>
- Purbasari, A., Rinawan, F. R., Zulianto, A., Susanti, A. I., & Komara, H. (2021). CRISP-DM for data quality improvement to support machine learning of stunting prediction in infants and toddlers. In *Proceedings of the 8th International Conference on Advanced Informatics: Concepts, Theory and Applications* (pp. 1–6). IEEE. <https://doi.org/10.1109/ICAICTA53211.2021.9640294>

- Qian, Y. (2025). Prompt engineering in education: A systematic review of approaches and educational applications. *Journal of Educational Computing Research*, 63(7-8), 1782-1818. <https://doi.org/10.1177/07356331251365189>
- Reynders, G., Lantz, J., Ruder, S. M., Stanford, C. L., & Cole, R. S. (2020). Rubrics to assess critical thinking and information processing in undergraduate STEM courses. *International Journal of STEM Education*, 7, Article 9. <https://doi.org/10.1186/s40594-020-00208-5>
- Sabri, O., Al-Shargabi, B., & Attia, O. N. (2026). Data quality and the trust-interpretability paradox: Toward a mid-range framework for AI adoption in educational information systems. *Journal of Information Technology Education: Research*, 25, Article 17. <https://doi.org/10.28945/5775>
- Shearer, C. (2000). The CRISP-DM model: The new blueprint for data mining. *Journal of Data Warehousing*, 5(4), 13-22.
- Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making? *International Journal of Human-Computer Studies*, 51(5), 991-1006. <https://doi.org/10.1006/ijhc.1999.0252>
- Snyder, J. (2019). Data cleansing: An omission from data analytics coursework. *Information Systems Education Journal*, 17(6), 22-29. <https://isedj.org/2019-17/n6/ISEDJv17n6p22.html>
- Wang, J., Liu, Y., Li, P., Lin, Z., Sindakis, S., & Aggarwal, S. (2024). Overview of data quality: Examining the dimensions, antecedents, and impacts of data quality. *Journal of the Knowledge Economy*, 15(1), 1159-1178. <https://doi.org/10.1007/s13132-022-01096-6>
- Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining* (pp. 29-39).
- Yoo, S. (2025). *Co-CoT: A prompt-based framework for collaborative chain-of-thought reasoning*. arXiv. <https://doi.org/10.48550/arXiv.2504.17091>
- Yue, K. -B. (2012). A Realistic Data Cleansing and Preparation Project. *Journal of Information Systems Education*, 23(2), 205-216.
- Zhang, S., Zhang, C., & Yang, Q. (2003). Data preparation for data mining. *Applied Artificial Intelligence*, 17(5-6), 375-381. <https://doi.org/10.1080/713827180>

APPENDIX A

Example Data (Synthetically Generated Sample)

Location	Customer ID	Category	Item	Price	Quantity	Total Spent	Payment Method	Sale Date	Coupon
5	1725	Bakery	1	4.5	1	4.5	Cash	6/17/2025	
3	345	Bakery	1	0	1	0	Credit Card	10/25/2025	TRUE
5	1587	Bakery	1	4.5	1	4.5		10/31/2025	
5	1725	Bakery	1	4.5	1	4.5	Cash	4/7/2026	TRUE
1	1242			4.5	1	4.5	Credit Card	8/18/2026	TRUE
1	345			4.5	1	4.5	Cash	9/6/2026	FALSE
4	414	Bakery	1	4.5	2	9	Cash	10/22/2024	TRUE
4	1311	Bakery	3	4.5	2	9	Credit Card	12/25/2024	FALSE
6	552	Bakery	1	4.5	2	9	Cash	2/9/2025	FALSE
6	414	Bakery	1	3.5	2	0	Cash	8/30/2026	
2	69	Bakery	1	4.5	2	9	Cash	11/10/2026	TRUE
3	1242	Meat	5	4.5	3	13.5	Cash	5/23/2024	
5	276	Bakery	1	4.5	3	13.5	Credit Card	6/30/2024	

APPENDIX B

Data Dictionary

Field Name	Data Type	Required	Constraints	Description / Notes
Location	NUM(9,0)	Yes	> 0	Store where the product was sold
Customer ID	NUM(9,0)	No	> 0	Customer number
Category	VARCHAR2(35)	Yes		Category of the item
Item	VARCHAR2(35)	Yes		Item identifier
Price	NUM(6,2)	Yes	> 0 and < 50	Price of item
Quantity	NUM(2,0)	Yes	> 0 and < 50	Quantity sold
Total Spent	NUM(6,2)	Yes	> 0 and < 2500	Total amount of line item
Payment Method	VARCHAR2(35)	Yes		How the customer paid
Sale Date	DATE	Yes	> 1/1/2022 and <= 6/30/2026	Date of the sale. The extract was pulled on 6/30/2026; no valid sale can postdate the extract.
Coupon	CHAR(5)	No	TRUE or FALSE	Whether a coupon was applied. Blank is treated as FALSE.

APPENDIX C

Data Preparation Notes Handout

An algorithm is a set of instructions, carried out in order, that gets you from a starting point to a result.

Filtering / Sampling Data

- When should we filter?
- When should we sample?

Duplicates

- When should ~~you~~we remove duplicates?
- When might ~~you~~we add duplicates?

Missing Data

- How can we identify missing data?
- What can we do to handle missing records?
- When should we remove it?
- When should we use a default?
- When should we use a formula?
- When should we impute?

Missing Field / Attribute	Situation	Algorithm

Feature Engineering

Create features or values to reflect relationships, simplify data, or be model-appropriate.

- Prepare Coupon to be a Boolean.
- Prepare the data for a time-series plot of the entire company.
- Prepare the data for a time-series plot by location.

Field / Attribute	Situation	Algorithm

APPENDIX D

Assignment: Prompt Engineering for AI-Assisted Data Cleaning

Overview

Generative AI tools are becoming increasingly common in data analytics. While these tools can automate parts of the data-cleaning process, the quality of the results depends heavily on the prompts and the business context provided by the user.

In this assignment, you will investigate how prompt engineering affects AI-assisted data cleaning. You will first ask an AI system to clean a retail grocery sales dataset using only the raw data. You will then improve your prompt by adding business knowledge, including a complete data dictionary and established data-cleaning procedures. Finally, you will critically evaluate both AI outputs and determine whether the resulting dataset is suitable for time-series forecasting.

The purpose of this assignment is to demonstrate that successful AI-assisted analytics requires both technical knowledge and critical human evaluation.

Learning Objectives

Upon completion of this assignment, you will be able to:

1. Apply the Data Understanding and Data Preparation deliverables of CRISP-DM to a transactional dataset.
2. Construct and improve a data-cleaning prompt that supplies business context, a data dictionary, and explicit cleaning rules.
3. Evaluate AI-generated cleaning decisions against documented business rules and descriptive statistics, identifying both errors of omission and errors of commission.
4. Compare AI outputs produced under different prompt conditions and justify which is more trustworthy using evidence drawn from the data.
5. Judge whether a prepared dataset is fit for time-series forecasting and defend that judgment.

Part 1 — Initial AI Prompt

Upload the provided grocery store sales dataset into a generative AI system (such as ChatGPT, Microsoft Copilot, Claude, or Gemini). Without providing any additional information, ask the AI to clean the dataset. Your prompt should request that the AI:

- Identify all data quality problems.
- Clean the dataset.
- Explain the assumptions and reasoning behind each cleaning decision.
- Describe any uncertainties or alternative approaches.
- Explain how the cleaned dataset would support time-series forecasting.

Deliverables include:

- Your complete prompt.
- Screenshots or exported conversation.
- The AI-generated cleaned dataset.
- The AI's explanation of its cleaning decisions.

Part 2 — Evaluate the AI Output

Before you can evaluate the AI's work, you need the evidence to evaluate it against.

Run descriptive statistics on the raw dataset and on the AI-cleaned dataset, and place them side by side:

- Record count.
- For Price, Quantity, and Total Spent: count, mean, median, minimum, maximum, and the number of nulls and zeros.
- Frequency tables for Location, Category, Payment Method, and Coupon.
- Count of records where Total Spent does not equal Price times Quantity.
- Earliest and latest Sale Date.

Every difference between the two columns is something the AI did. Some of those changes are repairs. Some do damage. Telling them apart is your job here.

Critically evaluate the AI's work. Consider questions such as:

- Which data quality issues were correctly identified?
- Which issues were overlooked?
- Did the AI make unsupported assumptions?
- Were any values incorrectly changed?
- Did the AI introduce new errors?
- Were missing values handled appropriately?
- Were calculated fields validated?
- Could you trust this cleaned dataset for forecasting?

Provide specific examples from the dataset.

Part 3 == Improve the Prompt

Rewrite your prompt to include the business context the AI did not have. Provide the complete data dictionary (Appendix B) and the data-cleaning rules developed in the data preparation lecture (Appendix C). Ask the AI to document its chain of thought using the structure in the table below, listing the known steps it should take and any additional steps it takes on its own.

Attribute	Situation	Algorithm	Instances

Run the improved prompt and save the results.

Part 4 == Compare the Results

Run the same descriptive statistics on the dataset produced by your improved prompt. You should now have three sets: raw, post-cleaning 1, and post-cleaning 2. Compare all three, then compare the outputs from Parts 1 and 3. Discuss:

- Which prompt produced better results?
- What additional issues were identified?
- Which recommendations have changed?
- How did additional business knowledge improve the output?
- Which version would you trust more?

Support your conclusions with examples.

Part 5 == Independent Data Quality Assessment

Finally, perform your own professional evaluation of the cleaned dataset. Do not assume the AI is correct. Evaluate the dataset for:

- Missing values

- Invalid values
- Incorrect calculations
- Duplicate records
- Outliers
- Inconsistent formatting
- Logical inconsistencies
- Data integrity
- Time-series readiness

For every remaining issue:

- Describe the problem.
- Explain why it matters.
- Recommend how it should be corrected.
- State whether the issue would affect forecasting accuracy.

Conclude with an overall assessment of whether the dataset is suitable for developing a reliable time-series forecasting model.

Submission Requirements

Submit a single report containing:

- Initial AI prompt.
- AI response from Part 1.
- Evaluation of the initial response.
- Improved prompt.
- AI response from Part 3.
- Comparison of both AI outputs.
- Final independent data quality assessment.
- Summary statistics and frequency tables prior to cleaning, post-cleaning 1, and post-cleaning 2.
- An appendix containing the cleaned dataset and the AI conversation transcript or screenshots.

Grading Rubric (100 Points)

Criteria	Points
Part 1: initial prompt, AI output, and conversation transcript	5
Evaluation of the first AI response	20
Improved prompt with data dictionary and business rules	20
Comparison of AI outputs	15
Independent data quality assessment	20
Reflection questions	10
Professional writing, organization, and supporting evidence	10
Total	100

Reflection Questions

Answer the following questions at the end of your report:

1. What data quality issues did the initial AI miss?
2. Which prompt engineering techniques most improved the AI's performance?
3. Why is human oversight still necessary when using AI for data preparation?
4. What additional information would further improve the AI's ability to clean the dataset?
5. If this dataset were used to train a forecasting model without cleaning, what business risks might result?