การประยุกต์ใช้ปัญญาประดิษฐ์เชิงสร้างสรรค์ในการทำความสะอาดและเตรียมข้อมูล: กรณีศึกษาวัสดุคอมโพสิตพอลิโพรพิลีนรีไซเคิลผสมกากชา

Application of Generative Artificial Intelligence in Data Cleaning and Preparation: A Case Study of Recycled Polypropylene Composite Mixed with Tea Residue

Received : May 24, 2024
Revised : June 19, 2024
Accepted : June 21, 2024

อนุชิต คงฤทธิ์, วศ.ม. (Anuchit Khongrit, M.Eng)[1*]
ชีวินทร์ ลิ้มศิริ, ปร.ด.(Cheevin Limsiri, Ph.D.)[2]
สุรีพร มีหอม, วศ.ด. (Sureeporn Meehom, Ph.D.)[3]

บทคัดย่อ

วัตถุประสงค์: การศึกษานี้มีจุดประสงค์เพื่อสำรวจความสามารถของปัญญาประดิษฐ์เชิงสร้างสรรค์ (Generative Artificial Intelligence) ทั้งสิ้น 3 ระบบ Claude.ai ChatGPT 3.5 และ Gemini บนแพลตฟอร์มออนไลน์ในการดำเนินการทำความสะอาดข้อมูล (Data cleaning) และการเตรียมข้อมูล (Data preparation) โดยใช้ข้อมูลของการทดสอบสมบัติทางกลของวัสดุคอมโพสิตพอลิโพรพิลีนรีไซเคิลผสมกากชาเป็นชุดข้อมูล (Data set)

วิธีการวิจัย: สำหรับทำการศึกษาในครั้งนี้ เริ่มต้นจากการสร้างคำถามหรือโจทย์ (Prompt) โต้ตอบ (Conversational prompting) ที่มีลักษณะการพูดคุยแบบสนทนาเพื่อใช้สอบถามปัญญาประดิษฐ์เชิงสร้างสรรค์ในการได้มาซึ่งหลักการ และวิธีการสำหรับการทำความสะอาดข้อมูล และการเตรียมข้อมูล หลังจากนั้นดำเนินการตามขั้นตอน และนำข้อมูลที่ผ่านการทำความสะอาดมาจัดเตรียมเป็นชุดข้อมูล ตลอดจนทำการอภิปรายผลเปรียบเทียบผลลัพธ์ที่ได้จากการใช้งานแต่ละแพลตฟอร์มปัญญาประดิษฐ์เชิงสร้างสรรค์

ผลการวิจัย: ผลการศึกษาพบว่าปัญญาประดิษฐ์เชิงสร้างสรรค์ทั้ง 3 ระบบมีความสามารถในการให้ข้อมูลที่เป็นประโยชน์ในกระบวนการทำความสะอาดข้อมูล และการเตรียมข้อมูลโดยพบว่าข้อมูลที่ได้รับมีความถูกต้องสอดคล้องกับเนื้อหาจากวรรณกรรมที่เกี่ยวข้องกับศาสตร์วิทยาการข้อมูล ทั้งนี้ไม่สามารถใช้แพลตฟอร์มของปัญญาประดิษฐ์เชิงสร้างสรรค์ทั้ง 3 ระบบในการทำความสะอาดข้อมูลได้เนื่องจากข้อจำกัดด้านความสามารถของระบบ จึงใช้ไพธอนไลบรารี่ในการดำเนินการโดยใช้ปัญญาประดิษฐ์เชิงสร้างสรรค์ในแบบกำหนดบุคลิกลักษณะ (Persona-based) สำหรับการสร้างไพธอนสคริปเพื่อการทำความสะอาดข้อมูล และ

จัดเตรียมข้อมูล ชุดข้อมูลหลังการทำความสะอาด และผ่านการเตรียมแล้วพบว่าสามารถนำไปใช้ได้ทั้งโปรแกรมสำหรับการวิเคราะห์ข้อมูล การวิเคราะห์ทางสถิติ ตลอดจนสามารถใช้บนแพลตฟอร์มออนไลน์ได้โดยที่ไม่พบปัญหาหรือข้อผิดพลาดที่เกิดจากความไม่สมบูรณ์หรือความไม่สะอาดของข้อมูล

คำสำคัญ: ปัญญาประดิษฐ์เชิงสร้างสรรค์, การทำความสะอาดข้อมูล, การเตรียมข้อมูล

[1] สาขาการจัดการงานวิศวกรรม คณะวิศวกรรมศาสตร์ มหาวิทยาลัยวงษ์ชวลิตกุล
(Engineering Management, Faculty of Engineering, Vongchavalitkul University)
[2]ผู้ช่วยศาสตราจารย์คณะวิศวกรรมศาสตร์ มหาวิทยาลัยวงษ์ชวลิตกุล
(Assistant Professor, Faculty of Engineering, Vongchavalitkul University)
[3]อาจารย์ สาขาวิชาการจัดการงานวิศวกรรม คณะวิศวกรรมศาสตร์ มหาวิทยาลัยวงษ์ชวลิตกุล
(Lecturer, Engineering Management, Faculty of Engineering, Vongchavalitkul University)
[*]ผู้รับผิดชอบบทความ (Corresponding author)
E-mail: anuchit_kho@vu.ac.th

Abstract

Objectives: This study aims to explore the capabilities of three Generative Artificial Intelligence (AI) systems—Claude.ai, ChatGPT 3.5, and Gemini—on online platforms for data cleaning and data preparation. The dataset used for this study consists of the mechanical property testing data of recycled polypropylene composite mixed with tea waste.

Method: The study begins by creating conversational prompts to interact with the generative AI systems, asking them for principles and methods for data cleaning and data preparation. Following these steps, the cleaned data is prepared, and the results from each AI platform are compared and discussed.

Results: The results indicate that all three Generative AI systems provided useful information for the data cleaning and data preparation processes. The information obtained was accurate and consistent with data science literature. However, the AI platforms could not perform the actual data cleaning due to their system limitations. Consequently, Python libraries were used to carry out the data cleaning, with the generative AI systems serving a persona-based role in generating Python scripts for data cleaning and preparation. The cleaned and prepared dataset was found to be usable in data analysis programs, statistical analysis, and online platforms without any issues or errors stemming from incomplete or unclean data.

Keywords: Generative Artificial Intelligence, Data Cleaning, Data Preparation

1. ความเป็นมาและความสำคัญของปัญหา

การเตรียมข้อมูล (Data preparation) คือการแปลงข้อมูลดิบ (Raw data) จากแหล่งที่มาให้อยู่ในรูปแบบที่เหมาะสมต่อการนำไปใช้ในการวิเคราะห์ตลอดจนสร้างแบบจำลอง ซึ่งกระบวนการนี้ประกอบไปด้วย การรวบรวม (Collecting) ทำความสะอาด (Cleaning) การแปลง (Transforming) และปรับปรุงข้อมูลดิบให้พร้อมสำหรับการวิเคราะห์ โดยเฉพาะอย่างยิ่งการทำความสะอาดข้อมูล (Data cleaning) ถือเป็นขั้นตอนสำคัญที่ช่วยระบุและแก้ไขปัญหาต่างๆ เช่น ความไม่สอดคล้องกัน ข้อผิดพลาด และข้อมูลที่ขาดหายไป การทำความสะอาดข้อมูลช่วยให้นักวิจัยมั่นใจได้ว่า ชุดข้อมูลมีความถูกต้อง สมบูรณ์ และพร้อมสำหรับการวิเคราะห์ ซึ่งนำไปสู่ผลลัพธ์ที่น่าเชื่อถือมากขึ้น (Ye & Wang, 2022) การทำความสะอาดข้อมูล (Data cleaning) มีผลกระทบอย่างมากต่อประสิทธิภาพและประสิทธิผลของกระบวนการวิเคราะห์ข้อมูลในหลากหลายวงการ เช่น วิทยาศาสตร์ วิศวกรรม และเทคโนโลยี เป็นการทำให้มั่นใจว่าข้อมูลที่ใช้ในการวิเคราะห์นั้นมีความถูกต้อง และน่าเชื่อถือ โดยข้อมูลที่ไม่ถูกต้องซึ่งเต็มไปด้วยข้อผิดพลาดต่างๆ เช่น ค่าที่หายไป (Missing value) ข้อมูลซ้ำซ้อน (Duplicate data) และค่าผิดปกติ (Anomalies) อาจนำไปสู่ผลการวิเคราะห์ข้อมูลที่ไม่แน่นอน และไม่สามารถนำไปใช้งานได้ส่งผลกระทบต่อการตัดสินใจและเป็นอุปสรรคต่อการพัฒนาที่ขับเคลื่อนด้วยการวิเคราะห์ข้อมูล (Dasari & Varma, 2022) การพัฒนางานวิจัยทางด้านการวิเคราะห์ข้อมูลหรือการวิเคราะห์ข้อมูลอย่างต่อเนื่องย้ำให้เห็นถึงความสำคัญของวิธีการทำความสะอาดข้อมูลที่มีประสิทธิภาพเพื่อขจัดความไม่ถูกต้อง และเพิ่มคุณภาพของข้อมูลนำเข้าสำหรับกระบวนการวิเคราะห์ข้อมูล การเพิ่มขึ้นของเทคโนโลยีบิ๊กดาต้า (Big data) ทำให้การทำความสะอาดข้อมูลเป็นขั้นตอนสำคัญในการเพิ่มคุณภาพของข้อมูล เนื่องจากข้อมูลที่ไม่ถูกต้อง หรือข้อมูลที่ไม่สะอาด (Dirty data) สามารถลดความแม่นยำ และประสิทธิภาพของกระบวนการตัดสินใจได้อย่างร้ายแรง (Jafarov, 2022; Zou, 2022)

วิธีการทำความสะอาดข้อมูลสามารถแบ่งออกเป็นวิธีการดั้งเดิม (Traditional method) และวิธีการสมัยใหม่ (Modern method) วิธีการและเครื่องมือทำความสะอาดข้อมูลแบบดั้งเดิมที่พบจากงานวิจัยที่เกี่ยวข้อง เช่น การตรวจสอบด้วยสายตา (Manual inspection) เป็นการตรวจสอบข้อมูลทีละแถวหรือทีละคอลัมน์เพื่อค้นหาและแก้ไขข้อผิดพลาด การใช้โปรแกรมไมโครซอฟท์เอกซ์เซล (Microsoft excel) มักถูกใช้ในการทำความสะอาดข้อมูล เนื่องจากมีความสามารถในการจัดการข้อมูลในรูปแบบตาราง สามารถกรอง ค้นหา และแก้ไขข้อมูลได้ง่าย การวิเคราะห์ด้วยภาพ (Visual data profiling) คือการวิเคราะห์ข้อมูลเชิงสถิติเพื่อระบุ และแสดงภาพปัญหาคุณภาพของข้อมูลที่อาจเกิดขึ้น ตลอดจนการใช้กฎ (Rule-based cleaning) การสร้างกฎเพื่อระบุและแก้ไขข้อผิดพลาดในข้อมูล เช่น การลบช่องว่างส่วนเกิน การแปลงตัวพิมพ์เล็ก/ใหญ่ หรือการแทนที่ค่าที่ไม่ถูกต้อง เป็นต้น (Loureiro et al., 2004; Setiyanto & Setiawan, 2022; Shigarov & Mikhailov, 2017; Von Zernichow & Roman, 2017) วิธีการและเครื่องมือทำความสะอาดข้อมูลสมัยใหม่ครอบคลุมหลากหลายแนวทางเพื่อยกระดับคุณภาพและความถูกต้องของข้อมูล เช่น

การใช้เครื่องมือทำความสะอาดข้อมูลแบบหลายกฎที่ผสมผสาน Conditional functional dependencies (CFDs) และ Related dependencies (RDs) เพื่อตรวจจับและแก้ไขข้อผิดพลาด แพลตฟอร์มอัตโนมัติอย่าง Data assist ที่ประยุกต์ใช้ประโยชน์จากการเรียนรู้ของเครื่อง (Machine learning) ในการเตรียมข้อมูล ทำความสะอาด และวิเคราะห์เชิงสำรวจ วิธีการนี้ช่วยประหยัดเวลาในการทำความสะอาดข้อมูลได้อย่างมีนัยสำคัญ และวิธีการที่เป็นนวัตกรรม เช่น DCM (Data Cleaning Model) และเครื่องมือ DCM explorer ที่ช่วยจับภาพและแสดงภาพกระบวนการทำความสะอาดข้อมูลเพื่อให้กระบวนการเข้าใจง่าย นอกจากนี้เครื่องมือวิเคราะห์ข้อมูลขนาดใหญ่ (Big data) ยังมีความสำคัญในการจัดการกับความซับซ้อนของชุดข้อมูลขนาดใหญ่ และขยายตัวอย่างต่อเนื่องโดยมีขีดความสามารถในการทำความสะอาด จัดเก็บ และวิเคราะห์ข้อมูลขนาดใหญ่ได้อย่างมีประสิทธิภาพ โดยเครื่องมือและวิธีการที่ทันสมัยเหล่านี้จะช่วยปรับปรุงกระบวนการทำความสะอาดข้อมูล และคุณภาพโดยรวมของชุดข้อมูลสำหรับการวิเคราะห์และการตัดสินใจได้มีประสิทธิภาพมากขึ้น (Sunne, 2022; Du et al., 2022; Parulian & Ludäscher, 2022; Goyle et al., 2023)

ปัญญาประดิษฐ์เชิงสร้างสรรค์ (Generative Artificial Intelligence; Generative AI) กำลังสร้างการปฏิวัติในการทำความสะอาดข้อมูลด้วยการทำงานแบบอัตโนมัติ ปรับปรุงความถูกต้อง และลดอคติ โมเดลเหล่านี้สามารถเรียนรู้รูปแบบและความสัมพันธ์ภายในข้อมูล ทำให้มีความสามารถในการระบุ และแก้ไขความไม่สอดคล้อง ประมาณค่าที่หายไป และตรวจหาค่าเบี่ยงเบนได้ ซึ่งจะลดเวลาและภาระงานที่จำเป็นสำหรับการทำความสะอาดข้อมูลด้วยตนเอง และนำไปสู่ข้อมูลที่มีคุณภาพสูงสำหรับการวิเคราะห์ (Gill, 2023; Zaamout, 2024) อย่างไรก็ตามประสิทธิภาพของโมเดลเหล่านี้ขึ้นอยู่กับคุณภาพของข้อมูลสำหรับการฝึกอบรม และการเลือกโมเดลที่เหมาะสมสำหรับงานทำความสะอาดเฉพาะนั้นมีความสำคัญ (Sergiienko, 2024) จากการศึกษาเบื้องต้นพบว่ามีการใช้ปัญญาประดิษฐ์เชิงสร้างสรรค์ในการทำความสะอาดข้อมูลโดยวิธีการ Generative Adversarial Networks (GANs) หรือโครงข่ายการเรียนรู้เชิงลึก GANs ช่วยยกระดับกระบวนการทำความสะอาดข้อมูลอย่างมีนัยสำคัญโดยนำเสนอกระบวนการที่เป็นนวัตกรรม เช่น DEGAIN (Deep Embedded Generative Adversarial Imputation Network) เป็นโมเดล ที่ถูกออกแบบมาเพื่อจัดการกับปัญหาข้อมูลที่หายไป (Missing data) โดยเฉพาะ ซึ่งจะเรียนรู้จากข้อมูลที่มีอยู่เพื่อสร้างข้อมูลที่หายไปให้สมบูรณ์ที่สุดเท่าที่จะเป็นไปได้ สามารถสกัดกฎการแก้ไขข้อมูลที่ตีความได้จากข้อมูลที่ไม่สมบูรณ์ ช่วยปรับปรุงคุณภาพและความแม่นยำของข้อมูลโดยไม่ต้องมีการดูแลจากมนุษย์ (Krishnan & Wu, 2019; Lee et al., 2020) โมเดลภาษาขนาดใหญ่ (LLM) ได้รับความสนใจอย่างมากเนื่องด้วยประสิทธิภาพในการทำงานต่างๆ โดย ChatGPT และ GPT-4 จาก OpenAI, Gemini จาก Google, Mixtral จาก Mistral AI และ Claude จาก Anthropic และอื่นๆ เป็นที่นิยมในหมู่ผู้ใช้งานกลุ่มแรกๆ โมเดลเหล่านี้ถูกมองว่าเป็นเทคโนโลยีที่ปฏิวัติวงการต่างๆ สามารถนำไปประยุกต์ใช้ได้หลากหลาย การศึกษาของ Rangapur and Rangapur (2024) ได้ศึกษาเปรียบเทียบโมเดลภาษาขนาดใหญ่ที่มีอยู่ในปัจจุบัน ในการทำงานตอบคำถามสนทนาเพื่อให้เข้าใจถึงจุดแข็ง จุดอ่อน

และประสิทธิภาพโดยรวมของโมเดลแต่ละตัวในการนำไปประยุกต์ใช้ในด้านต่างๆ เช่น การบริการลูกค้า การศึกษา การดูแลสุขภาพ และการเงิน เป็นต้น โดยผลการศึกษาพบว่า 3 ลำดับต้นของโมเดลที่มีประสิทธิภาพในการทำงาน และสามารถตอบคำถามได้ถูกต้องสูงสุดคือ ChatGPT Claude และ Gemini โดย GPT-4 และ Claude มีการพัฒนาอย่างมีนัยสำคัญในการสร้างคำตอบที่เกี่ยวข้องเฉพาะเจาะจง และสอดคล้องกัน ซึ่งทำให้โมเดลทั้งสองเป็นตัวเลือกที่เหมาะสมสำหรับงานตอบคำถามสนทนา ผลการวิจัยนี้มีความสำคัญต่อการพัฒนาผู้ช่วยสเหมือน (Virtual assistant) ที่ต้องอาศัยความสามารถในการประมวลผลและทำความเข้าใจภาษาธรรมชาติ (Rangapur & Rangapur, 2024)

จากที่กล่าวไปข้างต้นนำมาสู่จุดประสงค์ของการศึกษานี้โดยมีวัตถุประสงค์เพื่อสำรวจความสามารถของปัญญาประดิษฐ์เชิงสร้างสรรค์ในการทำความสะอาดข้อมูล และการเตรียมข้อมูล โดยจะทำการทดลองและประเมินประสิทธิภาพของ Generative AI 3 ระบบได้แก่ ChatGPT Claude และ Gemini ในการทำความสะอาดข้อมูลและการเตรียมข้อมูล นอกจากนี้งานวิจัยยังมุ่งเน้นศึกษาแนวทางและวิธีการในการนำ Generative AI มาประยุกต์ใช้ในกระบวนการทำความสะอาดข้อมูลและการเตรียมข้อมูลอย่างมีประสิทธิภาพ เพื่อเป็นแนวทางสำหรับนักวิจัยและผู้ปฏิบัติงานในการนำเทคโนโลยีนี้ไปใช้ในการเพิ่มประสิทธิภาพ และเพิ่มความแม่นยำในการวิเคราะห์ข้อมูลต่อไป โดยข้อมูลดิบที่เลือกใช้ในงานวิจัยนี้เป็นส่วนหนึ่งของการวิจัยที่กำลังดำเนินการอยู่เกี่ยวกับการปรับปรุงสมบัติทางกลของวัสดุคอมโพสิตพอลิโพรพิลีนรีไซเคิลผสมกากชา (อนุชิต คงฤทธิ์ และคณะ, 2565) ซึ่งงานวิจัยที่ผ่านมาได้ทำการศึกษาสมบัติทางกลของวัสดุคอมโพสิต และพบว่ายังมีความแข็งแรงทางกลไม่เหมาะสมกับการนำไปใช้งานจึงได้มีแนวทางการปรับปรุงสมบัติทางกลด้วยสารเติมแต่ง ซึ่งการศึกษาอิทธิพลของสารเติมแต่งที่ส่งผลต่อสมบัติทางกลจะสามารถนำไปวิเคราะห์ และนำไปสู่การปรับปรุงสมบัติทางกลของวัสดุคอมโพสิตได้ซึ่งต้องเริ่มต้นจากการมีชุดข้อมูลที่มีคุณภาพอันนำไปสู่การวิเคราะห์ที่ตรงกับวัตถุประสงค์

2. วัตถุประสงค์การวิจัย

2.1 เพื่อสำรวจความสามารถของปัญญาประดิษฐ์เชิงสร้างสรรค์ในการทำความสะอาดข้อมูล

2.2 เพื่อประเมินประสิทธิภาพของ Generative AI 3 ระบบได้แก่ ChatGPT Claude และ Gemini ในการทำความสะอาดข้อมูล

2.3 เพื่อเสนอแนวทางการนำ Generative AI มาประยุกต์ใช้ในการทำความสะอาดข้อมูลอย่างมีประสิทธิภาพจากผลการวิจัยนี้

3. ขอบเขต และกรอบแนวคิดในการวิจัย

3.1 ขอบเขตในการวิจัย

3.1.1 ชุดข้อมูล (Data set) ที่นำมาใช้เป็นกรณีศึกษาคือไฟล์ .csv ที่บันทึกผลการทดสอบสมบัติการรับแรงดึงวัสดุคอมโพสิตพอลิโพรพิลีนรีไซเคิลผสมกากชาจากเครื่องทดสอบ ด้วยเครื่องทดสอบแรงดึง LLOYD รุ่น LR10K ภาระงาน 10 กิโลนิวตัน (10 kN) มีข้อมูลทั้งสิ้น 7 คอลัมน์ ได้แก่ ID, Type of tea, PP-g-MA, TPE, Yield stress at offset, Yield strain at offset, Percentage of total elongation

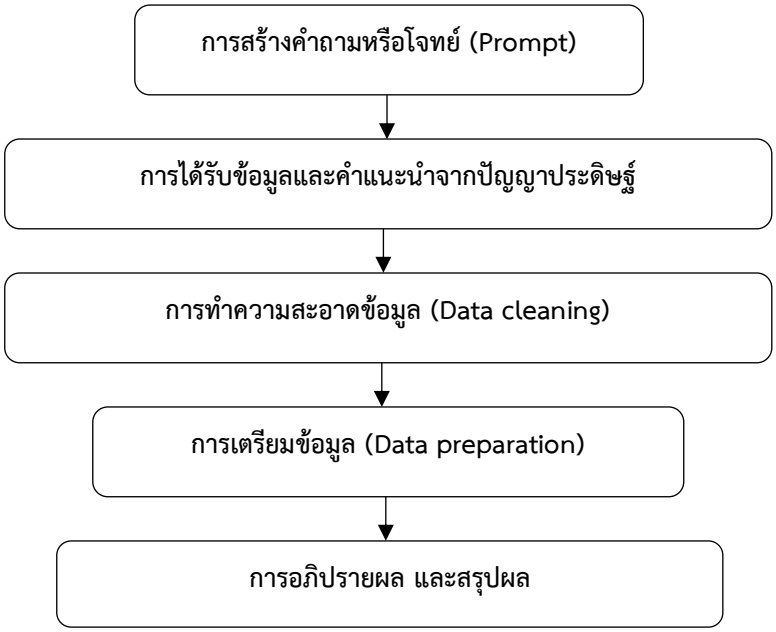
3.1.2 เครื่องมือในการขับเคลื่อนงานวิจัยนี้คือ Generative AI 3 ระบบ (Claude.ai ChatGPT 3.5 และ Gemini) ซึ่งการศึกษานี้เลือกใช้เวอร์ชั่นที่สามารถใช้งานได้โดยไม่เสียค่าบริการ และแพลตฟอร์ม Python Library บนเว็บไซต์ https://colab.research.google.com/ (Google Colab)

3.2 กรอบแนวคิดในการวิจัย

กรอบแนวคิดการวิจัยนี้พัฒนาจากแนวคิดในการปรับปรุงคุณภาพข้อมูลผลการทดสอบสมบัติทางกลของวัสดุคอมโพสิตพอลิโพรพิลีนรีไซเคิลผสมกากชาให้มีคุณภาพเพื่อให้ผลการวิเคราะห์มีความถูกต้อง น่าเชื่อถือ และสามารถนำไปพยากรณ์ผลลัพธ์ได้แม่นยำมากขึ้นโดยมีความต้องการประยุกต์ใช้เทคโนโลยีปัญญาประดิษฐ์เชิงสร้างสรรค์ที่กำลังได้รับความสนใจในเชิงวิชาการและการใช้งานในหลายๆด้าน ณ ปัจจุบันในการให้ข้อมูล หรือหลักการในการทำความสะอาดข้อมูล วิธีการในเชิงปฏิบัติ และการใช้ปัญญาประดิษฐ์เชิงสร้างสรรค์ในการทำความสะอาดข้อมูล และเตรียมข้อมูลสำหรับนำไปวิเคราะห์ด้วยโปรแกรมทางสถิติหรือโปรแกรมที่เกี่ยวข้องกับการวิเคราะห์ข้อมูล หลังจากนั้นจะมีการตรวจสอบความถูกต้องของผลลัพธ์ที่ได้ และสรุปผล

4. วิธีดำเนินการวิจัย

ขั้นตอนการทำความสะอาดข้อมูล และการเตรียมข้อมูลในงานวิจัยนี้ประยุกต์ใช้การทำความสะอาดข้อมูล และเตรียมข้อมูลจากศาสตร์วิทยากรข้อมูลซึ่งใช้เครื่องมือแพลตฟอร์ม Python Library ในการดำเนินงาน โดยขั้นตอนการทำความสะอาดข้อมูลแบบสังเขปมีดังนี้ การนำเข้าไฟล์ข้อมูล (.csv.) การเลือกและจัดระเบียบคอลัมน์ การจัดการค่าที่ขาดหาย การจัดการค่าสุดโต่ง และการสำรวจข้อมูล (Walker, 2020) กระบวนการดำเนินงานวิจัยแสดงดังรูปที่ 1 โดยเริ่มต้นจากการสร้างคำถามหรือโจทย์ (Prompt) เพื่อใช้สอบถามปัญญาประดิษฐ์เชิงสร้างสรรค์ในการได้มาซึ่งหลักการสำหรับการทำความสะอาดข้อมูล (Data cleaning) และการเตรียมข้อมูล (Data preparation) โดยใช้ หลังจากได้รับข้อมูลและคำแนะนำจากปัญญาประดิษฐ์แล้ว ผู้วิจัยดำเนินการทำความสะอาดข้อมูลตามขั้นตอนที่ได้รับมาโดยใช้โปรแกรมหรือแพลตฟอร์มตามที่ปัญญาประดิษฐ์แนะนำ จากนั้นนำข้อมูลที่ผ่านการทำความสะอาดมาจัดเตรียมเป็นชุดข้อมูล (Dataset) เพื่อนำไปวิเคราะห์ด้วยโปรแกรมสำหรับการวิเคราะห์ข้อมูล ได้แก่ Minitab 18, Google Sheet, Rapid miner สุดท้ายจะทำการอภิปรายผลเปรียบเทียบผลลัพธ์ที่ได้จากการใช้งานแต่ละแพลตฟอร์มปัญญาประดิษฐ์เชิงสร้างสรรค์



รูปที่ 1 ขั้นตอนการดำเนินงานของการศึกษานี้

5. ผลการดำเนินการ

5.1 การสร้างคำถาม และผลลัพธ์ที่เกี่ยวข้องกับการทำความสะอาดข้อมูลและจัดเตรียมข้อมูล

การสร้างคำถามหรือโจทย์ (Prompt) เพื่อใช้สอบถามปัญญาประดิษฐ์เชิงสร้างสรรค์ใช้หลักแบบโต้ตอบ (Conversational prompting) ที่มีลักษณะการพูดคุยแบบสนทนาเพื่อให้ AI สามารถเรียนรู้และเข้าใจบริบทของสถานการณ์ได้ลึกซึ้งขึ้นจากการอ้างอิงประเด็นที่ถามหรือตอบไปแล้วในการสนทนา ตลอดจนลดความเสี่ยงของการที่ปัญญาประดิษฐ์เชิงสร้างสรรค์เข้าใจคำสั่งผิดพลาดเนื่องจากมีการโต้ตอบกันไปมา โดยมีการสร้างคำถามแบบตอบโต้ และนำคำตอบกลับมาสร้างเป็นคำถามที่ต่อเนื่องกันเพื่อให้ได้หลักการเพื่อทำความสะอาดข้อมูล (Data cleaning) และการเตรียมข้อมูล (Data preparation) ในเชิงปฏิบัติ โดยผลมีดังนี้

คำถามที่1 : ผมมีไฟล์ data set นามสกุล .csv ที่ได้จากการทดลอง คุณสามารถช่วยทำการ data Cleaning and preparation ให้ได้อย่างไร โดยคำตอบโดยสรุปของปัญญาประดิษฐ์เชิงสร้างสรรค์แสดงดังตารางที่ 1

ตารางที่ 1 คำตอบสำหรับคำถามที่ 1 โดยสรุปจากปัญญาประดิษฐ์เชิงสร้างสรรค์ทั้ง 3 ระบบ

Claude.ai	ChatGPT 3.5	Gemini
มาดูขั้นตอนการทำ data Cleaning และ data preparation สำหรับไฟล์ .csv ของคุณกัน **นำเข้าข้อมูล** - ใช้ library เช่น pandas ในการนำเข้าไฟล์ .csv - ตรวจสอบโครงสร้างของข้อมูลด้วย .head(), .shape, .info() เป็นต้น **ตรวจสอบค่าที่ขาดหายไป (Missing Values)** - ใช้ .isna() หรือ .isnull() เพื่อตรวจหาค่า null/nan - พิจารณาว่าจะลบแถวหรือคอลัมน์ที่มีค่าขาดหายไปมากเกินไป หรือจะทำการเติมค่าด้วยวิธีอื่น เช่น mean, median **แก้ไขค่าผิดพลาด** - ตรวจสอบค่าผิดปกติ (outliers) ด้วย describe(), boxplot เป็นต้น - พิจารณาปรับค่าผิดพลาดหรือลบแถวที่มีค่าผิดปกติออกไป **การแปลงข้อมูล** - ตรวจสอบประเภทข้อมูลของแต่ละคอลัมน์ด้วย .dtypes - ทำการแปลงประเภทข้อมูลให้ถูกต้องตามความเหมาะสม เช่น object -> category, float -> int **Data Normalization หรือ Feature Scaling** - หากต้องใช้อัลกอริทึมบางอย่างที่ต้องการข้อมูลที่ถูกปรับมาตรฐานแล้ว ให้ทำการ normalize หรือ scale ข้อมูลด้วยเทคนิคต่างๆ เช่น StandardScaler, MinMaxScaler	ในการทำ data Cleaning and preparation เพื่อให้ผลลัพธ์ที่ดีที่สุด สิ่งแรกที่ควรทำคือการทำความเข้าใจกับข้อมูลในไฟล์ .csv ของคุณ ซึ่งประกอบไปด้วยข้อมูลในรูปแบบตาราง อาจมีข้อมูลที่สูญหาย ข้อมูลที่ซ้ำซ้อน หรือข้อมูลที่ไม่สมบูรณ์ เราจะตรวจสอบและแก้ไขปัญหาเหล่านี้ตามลำดับดังนี้: **ตรวจสอบข้อมูลที่หายไป** (Missing Data): ตรวจสอบว่ามีข้อมูลที่หายไปหรือไม่ และตรวจสอบว่ามีข้อมูลที่หายไปในแต่ละคอลัมน์มากน้อยเพียงใด หากมีข้อมูลที่หายไป เราสามารถเติมข้อมูลที่สูญหายไปด้วยค่าเฉลี่ย หรือค่ามัธยฐานของคอลัมน์นั้นๆ หรือใช้วิธีการอื่นๆ เช่นการใช้โมเดลเพื่อทำนายค่าที่หายไปได้ **ตรวจสอบข้อมูลที่ซ้ำซ้อน** (Duplicate Data): ตรวจสอบว่ามีข้อมูลที่ซ้ำกันในตารางหรือไม่ หากพบข้อมูลที่ซ้ำซ้อน เราสามารถลบข้อมูลซ้ำทิ้งได้ **ทำความสะอาดข้อมูล** (Data Cleaning): สำหรับข้อมูลที่ไม่ถูกต้องหรือไม่สมบูรณ์ เช่น ข้อมูลที่ไม่เป็นไปตามรูปแบบที่กำหนด หรือมีข้อผิดพลาดทางตรรกะ เราจะทำการแก้ไขให้ถูกต้อง **การเข้ารหัสข้อมูล** (Encoding Data): หากมีข้อมูลที่เป็นข้อความหรือข้อมูลที่ไม่ใช่ตัวเลข เราต้องทำการเข้ารหัสเพื่อแปลงข้อมูลให้เป็นตัวเลข เช่น การใช้ One-Hot Encoding หรือการแปลงข้อความเป็นตัวเลข	สามารถช่วยคุณทำความสะอาดและเตรียมข้อมูล (data Cleaning and preparation) จากไฟล์ .CSV ที่ได้จากการทดลองของคุณได้ ดังนี้ **โหลดข้อมูล** : เริ่มต้นด้วยการโหลดไฟล์ .CSV ของคุณเข้าสู่เครื่องมือวิเคราะห์ข้อมูล เช่น Python หรือ R คุณสามารถใช้ไลบรารีต่างๆ เช่น pandas ใน Python หรือ data.frame ใน R เพื่อโหลดและจัดการข้อมูล .CSV ได้ **ตรวจสอบข้อมูล** : ตรวจสอบข้อมูลของคุณเพื่อหาข้อผิดพลาดหรือความไม่สอดคล้องกันทั่วไป สิ่งนี้อาจรวมถึงค่าที่ขาดหายไป ค่าที่ผิดปกติ หรือรูปแบบข้อมูลที่ไม่ถูกต้อง คุณสามารถใช้ฟังก์ชันการวิเคราะห์ข้อมูลในไลบรารีที่คุณเลือกเพื่อระบุปัญหาเหล่านี้ **ทำความสะอาดข้อมูล** : แก้ไขข้อผิดพลาดและความไม่สอดคล้องกันที่พบในขั้นตอนก่อนหน้า ตัวอย่างเช่น คุณสามารถแทนค่าที่ขาดหายไปด้วยค่าเฉลี่ยหรือค่ามัธยฐาน หรือลบแถวที่มีค่าที่ผิดปกติมาก **เตรียมข้อมูล**: แปลงข้อมูลของคุณให้เป็นรูปแบบที่เหมาะสมสำหรับการวิเคราะห์ สิ่งนี้อาจรวมถึงการแปลงประเภทข้อมูล การจัดการค่าที่หายไป การจัดรูปแบบสตริง และการสร้างตัวแปรใหม่

ตารางที่ 1 คำตอบสำหรับคำถามที่ 1 โดยสรุปจากปัญญาประดิษฐ์เชิงสร้างสรรค์ทั้ง 3 ระบบ (ต่อ)

Claude.ai	ChatGPT 3.5	Gemini
แบ่งข้อมูลเป็น Train/Test Sets - สำหรับการสร้างโมเดลนั้นจำเป็นต้องแบ่งข้อมูลเป็น Training Set และ Test Set - ใช้ train_test_split จาก scikit-learn เพื่อแบ่งข้อมูล	**การเลือกคุณลักษณะ** (Feature Selection): หากมีคุณลักษณะที่ไม่จำเป็น เราสามารถลบคุณลักษณะเหล่านั้นออกจากตารางข้อมูลได้ เพื่อลดความซับซ้อนและเวลาในการประมวลผล **การทำสเกลข้อมูล** (Data Scaling): หากมีคุณลักษณะที่มีการแจกแจงข้อมูลที่แตกต่างกันมาก เราสามารถทำการปรับข้อมูลให้มีการกระจายที่เหมือนกัน เช่น การใช้ Min-Max Scaling หรือ Standard Scaling **การแบ่งข้อมูล** (Data Splitting): หากต้องการใช้ข้อมูลสำหรับการฝึกโมเดลและการทดสอบโมเดล เราจะทำการแบ่งข้อมูลออกเป็นชุดฝึกและชุดทดสอบ	

5.1.1 ความเหมือนของผลลัพธ์

จากปัญญาประดิษฐ์ทั้ง 3 ระบบได้แก่ การโหลดข้อมูลโดยทั้ง 3 ระบบกล่าวถึงการเริ่มต้นด้วยการโหลดข้อมูลจากไฟล์ .csv เข้าสู่เครื่องมือหรือโปรแกรมที่ใช้ในการวิเคราะห์ข้อมูล เช่น Python (โดยใช้ Library pandas) หรือ R. การตรวจสอบข้อมูล โดยใจความสำคัญของการตรวจสอบข้อมูลเพื่อหาข้อผิดพลาด ความไม่สอดคล้อง หรือปัญหาต่าง ๆ ที่อาจเกิดขึ้น เช่น ค่าที่หายไป ค่าผิดปกติ รูปแบบข้อมูลที่ไม่ถูกต้อง การทำความสะอาดข้อมูล ที่อธิบายถึงการแก้ไขปัญหาที่พบในขั้นตอนการตรวจสอบข้อมูล เช่น การจัดการค่าที่หายไป (ลบ หรือแทนที่) การลบข้อมูลซ้ำซ้อน และการแก้ไขค่าผิดพลาด เป็นต้น สุดท้ายเรื่องการเตรียมข้อมูล ที่กล่าวถึงการแปลงข้อมูลให้อยู่ในรูปแบบที่เหมาะสมสำหรับการวิเคราะห์ ซึ่งอาจรวมถึงการแปลงประเภทข้อมูล การจัดการค่าที่หายไป และการปรับขนาดข้อมูล (Scaling) ทั้งนี้แม้คำศัพท์หรือคำที่ใช้อธิบายมีแตกต่างกัน แต่แนวคิดหลักเกี่ยวกับขั้นตอนการทำ Data cleaning and preparation มีความสอดคล้องกัน คือ เริ่มจากการโหลดข้อมูล ตรวจสอบ ทำความสะอาด และเตรียมข้อมูลให้พร้อมสำหรับการวิเคราะห์ต่อไป

5.1.2 ความแตกต่างของผลลัพธ์

พบว่ามีประเด็นที่แตกต่างกันหรือไม่เหมือนกันในด้านเทคนิคที่ใช้ ได้แก่ Claude.ai ให้รายละเอียดและขั้นตอนการดำเนินการรวมถึงการแนะนำเครื่องมือ (pandas) และฟังก์ชันที่ใช้ในแต่ละ

ขั้นตอนอย่างละเอียด (เช่น .head(), .shape, .info(), .isna(), .isnull(), .dtypes, train_test_split) สำหรับ ChatGPT 3.5 เน้นการอธิบายปัญหาที่อาจพบในข้อมูลและเทคนิคที่หลากหลายในการแก้ไขปัญหา เช่น การใช้โมเดลเพื่อเติมค่าที่หายไป การทำ Feature selection การทำ Data scaling คำตอบนี้อธิบายในภาพรวมและแนวทางแก้ปัญหาที่หลากหลาย และ Gemini ให้คำตอบขั้นตอนหลักๆ โดยไม่ได้ลงรายละเอียดเกี่ยวกับเครื่องมือหรือเทคนิคที่ใช้มากนัก แต่มีการแนะนำเครื่องมืออื่นๆ เพิ่มเติมนอกจาก Python เช่น R. เพื่อให้ทราบภาพรวมของกระบวนการ โดยสรุปแล้วทั้ง 3 ระบบอธิบายเกี่ยวกับกระบวนการ Data cleaning and preparation ที่คล้ายกันแต่มีความแตกต่างกันในด้านรายละเอียด การเน้นย้ำในประเด็นต่าง และการใช้คำศัพท์จึงนำมาสู่คำถามที่ 2 เกี่ยวกับวิธีการดำเนินงานส่วนหลักของการวิจัยนี้คือการจัดการจัดค่าสุดโต่ง (Outliers) จัดการกับข้อมูลที่ขาดหาย (Missing value)

5.1.3 ความถูกต้องของผลลัพธ์ เครื่องมือที่ถูกกล่าวอ้างถึงจากปัญญาประดิษฐ์นั้นถูกต้อง ได้แก่ Python (โดยใช้ Library pandas) หรือ R. นอกจากนี้ที่ไม่ถูกกล่าวถึง เช่น Tableau และ SAS เป็นต้น สำหรับหลักการที่สำคัญ เช่น การนำเข้าข้อมูล หรือโหลดไฟล์ข้อมูลในรูปแบบ .csv สำหรับกระบวนการทำความสะอาดข้อมูลที่แสดงดังผลลัพธ์ เช่น ค่าที่หายไป ค่าผิดปกติ รูปแบบข้อมูลที่ไม่ถูกต้อง การจัดการคุณลักษณะข้อมูลนั้นมีความถูกต้องตามหลักการของการทำความสะอาด และจัดเตรียมข้อมูล (Nandi, G., & Sharma, R. K., 2020)

คำถามที่ 2 : อธิบายการดำเนินการทำความสะอาดข้อมูล ได้แก่ การจัดการจัดค่าสุดโต่ง (Outliers) จัดการกับข้อมูลที่ขาดหาย (Missing value handling) โดยคำตอบโดยสรุปของปัญญาประดิษฐ์เชิงสร้างสรรค์แสดงดังตารางที่ 2

ตารางที่ 2 คำตอบสำหรับคำถามที่ 2 โดยสรุปจากปัญญาประดิษฐ์เชิงสร้างสรรค์ทั้ง 3 ระบบ

Claude.ai	ChatGPT 3.5	Gemini
สำหรับการจัดการกับประเด็นต่างๆ ที่คุณสนใจนั้น เรามาดูวิธีการแก้ไขแบบละเอียด **การกำจัดค่า Outliers** - ใช้ boxplot หรือ scatter plot เพื่อระบุค่า outliers - ลบแถวที่มีค่า outliers ออกด้วย dataframe = dataframe[~(dataframe['column'].abs() > threshold)] - หรือใช้ method .clip() หรือ .quantile() เพื่อจำกัดค่าให้อยู่ในช่วงที่เหมาะสม	เป็นไปได้ที่คุณจะทำการกำจัดค่า outliers, แก้ไข missing value, และทดสอบทางสถิติว่าข้อมูลมีการกระจายตัวแบบปกติหรือไม่ เพื่อลดความซับซ้อนและเพิ่มประสิทธิภาพในการวิเคราะห์ข้อมูล ดังนั้น ขั้นตอนที่คุณสามารถทำได้คือดังนี้: **การกำจัดค่า Outliers:** - ใช้วิธีการสถิติ เช่น Z-score หรือ IQR (Interquartile Range) เพื่อตรวจจับและกำจัดค่า outliers - ลบหรือแทนค่า outliers ด้วยค่าที่เหมาะสม เช่น การใช้ค่ามัธยฐานหรือค่าเฉลี่ยของข้อมูลที่ไม่มี outliers	ผมสามารถช่วยคุณจัดการข้อมูลสำหรับการวิเคราะห์ โดยมุ่งเน้นไปที่การกำจัดค่า outliers, แก้ไข missing value, ทดสอบทางสถิติว่ามีการกระจายตัวแบบปกติหรือไม่ดังนี้ **กำจัดค่า outliers:** - วิธี Z-score: คำนวณ Z-score ของแต่ละจุดข้อมูล จุดข้อมูลที่ Z-score มากกว่า 3 หรือน้อยกว่า -3 จะถูกพิจารณาเป็น outliers

ตารางที่ 2 คำตอบสำหรับคำถามที่ 2 โดยสรุปจากปัญญาประดิษฐ์เชิงสร้างสรรค์ทั้ง 3 ระบบ (ต่อ)

Claude.ai	ChatGPT 3.5	Gemini
การแก้ไข Missing Values - ลบแถวหรือคอลัมน์ที่มีค่า null/nan มากเกินไปด้วย dataframe.dropna() - เติมค่าด้วย mean หรือ median: dataframe['column'].fillna (dataframe['column'].mean(), inplace=True) - ใช้ linear regression, kNN หรือ decision tree เพื่อทำนายค่าที่ขาดหายไป	**การแก้ไข Missing Values:** - ใช้วิธีการเติมค่าเพื่อแก้ไข missing values เช่น การใช้ค่าเฉลี่ย หรือค่ามัธยฐานของคอลัมน์นั้นๆ - หากข้อมูลมีความสัมพันธ์กับค่าอื่นๆ สามารถใช้โมเดลทำนายค่าที่หายไปได้	- วิธี IQR: หาค่า Q1 (quartile 1) และ Q3 (quartile 3) ของข้อมูล คำนวณ IQR (interquartile range) = Q3 - Q1 กำผหนดขอบเขตที่ยอมรับได้ = Q1 - 1.5 * IQR และ Q3 + 1.5 * IQR จุดข้อมูลที่อยู่นอกขอบเขตที่ยอมรับได้ จะถูกพิจารณาเป็น outliers - วิธี visual inspection: พล็อตกราฟข้อมูล จุดข้อมูลที่ดูแตกต่างจากจุดอื่นๆ อาจเป็น outliers **แก้ไข missing value:** - การลบข้อมูล: ลบแถวที่มี missing value วิธีนี้ง่าย แต่ทำให้สูญเสียข้อมูล - การแทนที่ค่า: แทนที่ missing value ด้วยค่าเฉลี่ย ค่ามัธยฐาน หรือโหมด วิธีนี้ง่าย แต่ไม่เหมาะกับ missing value จำนวนมาก - การสันนิษฐาน: สันนิษฐาน missing value โดยใช้เทคนิคการ regression หรือ imputation วิธีนี้ซับซ้อน แต่ช่วยรักษาข้อมูลได้ **เครื่องมือและเทคนิค** มีเครื่องมือและเทคนิคมากมายที่คุณสามารถใช้สำหรับการจัดการข้อมูล เครื่องมือที่ได้รับความนิยมบางส่วนได้แก่: - ไลบรารีวิเคราะห์ข้อมูล: Python มีไลบรารี pandas, scikit-learn และ statsmodels ที่ได้รับความนิยมสำหรับการจัดการและวิเคราะห์ข้อมูล R มีแพ็คเกจ data.frame, dplyr และ ggplot2 ที่คล้ายกัน

5.1.4 ความเหมือนของผลลัพธ์ ในการจัดการค่าสุดโต่ง (Outliers) ทั้ง 3 ระบบกล่าวถึงการใช้เทคนิคทางสถิติในการจัดการ Outliers เช่น Z-score และ IQR (Interquartile Range) โดยระบุว่าค่าที่อยู่นอกช่วงที่กำหนดจะเป็น Outliers และสามารถลบหรือแทนที่ด้วยค่าที่เหมาะสม เช่น ค่ามัธยฐาน หรือค่าเฉลี่ย การจัดการค่าที่ขาดหาย (Missing values) ทั้ง 3 ระบบกล่าวถึงวิธีการจัดการหลากหลาย เช่น การลบข้อมูลที่มี Missing Values การแทนที่ด้วยค่าทางสถิติ (เช่น ค่าเฉลี่ย, มัธยฐาน) และการใช้เทคนิคอื่น ๆ เช่น การสร้างโมเดลเพื่อทำนายค่าที่หายไป (Imputation)

5.1.5 ความแตกต่างของผลลัพธ์ สามารถแบ่งเป็น 2 ประเด็นดังนี้ ระดับความละเอียด Claude.ai ให้รายละเอียดและขั้นตอนการดำเนินการรวมถึงตัวอย่างโค้ด Python ที่ใช้ในการลบ Outliers และเติมค่า Missing values ด้วยค่าเฉลี่ย ChatGPT 3.5 อธิบายภาพรวมของวิธีการจัดการ Outliers และ Missing values แต่ไม่ได้ลงรายละเอียดเกี่ยวกับขั้นตอนการดำเนินการ และ Gemini อธิบายหลักการของวิธี Z-score และ IQR ในการจัดการ Outliers ที่ละเอียดและมีขั้นตอนกว่า 2 ระบบแรก แต่ไม่ได้ให้ตัวอย่างโค้ด และกล่าวถึงวิธีการจัดการ Missing values โดยสรุป ประเด็นต่อมาคือเครื่องมือและเทคนิค Claude.ai ไม่ได้กล่าวถึงเครื่องมือหรือเทคนิคอื่นๆ นอกเหนือจากตัวอย่างโค้ด Python ที่ใช้ในการลบ Outliers และเติมค่า Missing values สำหรับ ChatGPT 3.5 กล่าวถึงการใช้สถิติในการจัดการค่า Outlier และ Missing values แต่ไม่ได้ระบุรายละเอียดของเครื่องมือที่ใช้ และ Gemini กล่าวถึงเครื่องมือและเทคนิคที่หลากหลายที่สามารถใช้ในการจัดการข้อมูล เช่น ไลบรารี Pandas, Scikit-learn, Statsmodels ใน Python และแพ็คเกจ Data.frame, Dplyr, Ggplot2 ในโปรแกรม R ที่ชัดเจนกว่า 2 ระบบแรก

5.1.6 ความถูกต้องของผลลัพธ์ ของ Claude.ai กล่าวถึงการกำจัดค่า Outliers ด้วย Boxplot หรือ Scatter plot ตลอดจนจัดการ Missing values และมีตัวอย่างโค้ด Python เป็นการอธิบายที่ถูกต้อง (Finding outliers in a scatter plot / Pandas dataframe?, n.d.; Pandas DataFrame: replace nan values with average of columns, n.d.) สำหรับผลลัพธ์ของ ChatGPT 3.5 ที่อธิบายหลักการของการจัดการค่า Outliers และ Missing Values ด้วยวิธีการทางสถิติมีความถูกต้องเช่นกัน ข้อมูล และผลลัพธ์ของ Gemini ที่แสดงวิธีการด้วยกระบวนการทางสถิติต่างๆ เช่น Z-score, IQR, Regression ตลอดจนเครื่องมือและเทคนิควิเคราะห์ข้อมูลด้วย Python มีไลบรารี Pandas, Scikit-learn และ Statsmodels มีความถูกต้องตามหลักการ (Nandi & Sharma, 2020)

สำหรับคำถามต่อมาจึงเป็นขั้นตอนการทำความสะอาดข้อมูล และสร้างชุดข้อมูลใหม่หลังจากการทำความสะอาดแล้วโดยต้องการทำเป็นไฟล์นามสกุล .csv ไปใช้ในการประมวลผลได้กับโปรแกรมที่หลากหลาย ได้แก่ Google sheet, Minitab, Rapid Miner โดยมีรายละเอียดดังนี้

คำถามที่ 3 : การเตรียมข้อมูล (Data preparation) หลังจากการทำ Cleaning แล้วมีหลักการอย่างไรเพื่อสร้าง data set นามสกุล .csv ไปใช้ในการประมวลผลได้กับโปรแกรมที่หลากหลาย ได้แก่ Google sheet, Minitab, Rapid Miner

ตารางที่ 3 คำตอบสำหรับคำถามที่ 3 โดยสรุปจากปัญญาประดิษฐ์เชิงสร้างสรรค์ทั้ง 3 ระบบ

Claude.ai	ChatGPT 3.5	Gemini
หลังจากทำ data Cleaning แล้ว เพื่อให้ dataset ที่ได้สามารถนำไปใช้งานกับโปรแกรมต่างๆ ได้อย่างกว้างขวาง มีหลักการที่ควรปฏิบัติดังนี้ **จัดรูปแบบข้อมูลให้เป็นมาตรฐาน** - ชื่อคอลัมน์และค่าข้อมูลควรเป็นตัวอักษรพิมพ์เล็กหรือพิมพ์ใหญ่ทั้งหมด - หลีกเลี่ยงการใช้อักขระพิเศษหรือช่องว่างในชื่อคอลัมน์ - ควรแยกวันที่ออกเป็น วัน/เดือน/ปี ในแต่ละคอลัมน์แทนการรวมเป็นคอลัมน์เดียว **เลือกรูปแบบไฟล์ที่เหมาะสม** - สำหรับการแบ่งปันข้อมูลทั่วไป ให้ใช้ไฟล์ .csv - สำหรับการแบ่งปันกับนักวิเคราะห์ข้อมูลหรือนักวิทยาศาสตร์ข้อมูล ให้ใช้ไฟล์ .xlsx หรือ .xls - หากมีข้อมูลซับซ้อนมาก ๆ ให้ใช้ไฟล์ .rdata (R) หรือ .dta (Stata) **กำหนดค่า Delimiters และ Encoding** - ตั้งค่า delimiters (วรรค , ไปป์ | หรือแท็บ) ให้เหมาะสมกับไฟล์และโปรแกรมที่ใช้ ใช้ encoding ที่เป็นมาตรฐาน เช่น UTF-8 สำหรับความเข้ากันได้สูงสุด **เพิ่มเมทาดาทาและคำอธิบาย** - เพิ่มแถวคำอธิบายที่สั้นกะทัดรัดชี้แจงความหมายของคอลัมน์ต่างๆ - ระบุหน่วย, รูปแบบวันที่, และแหล่งที่มาของข้อมูล - หากเป็นไปได้ ให้เพิ่มเมทาดาทาเข้าไปในไฟล์เดียวกันกับข้อมูล	หลักการในการเตรียมข้อมูลหลังจากการทำ Cleaning เพื่อสร้าง data set ที่สามารถใช้งานได้กับโปรแกรมต่างๆ เช่น Excel, Google Sheets, Minitab, หรือ Python ประกอบด้วยขั้นตอนเหล่านี้ **บันทึกข้อมูลเป็นไฟล์ .csv:** หลังจากการทำการ Cleaning เสร็จสิ้นแล้ว คุณจะต้องบันทึกข้อมูลลงในไฟล์ .csv ในการบันทึกไฟล์ .csv คุณสามารถใช้โปรแกรมต่างๆ เช่น Excel, Google Sheets, หรือ Python เพื่อนำข้อมูลไปใช้งานต่อ **การกำหนดชื่อคอลัมน์:** ตรวจสอบและกำหนดชื่อคอลัมน์ที่เหมาะสมให้กับข้อมูลแต่ละคอลัมน์ เพื่อให้ง่ายต่อการอ่านและเข้าใจข้อมูล ระวังให้แน่ใจว่าชื่อคอลัมน์ไม่มีช่องว่างหรืออักขระพิเศษที่ไม่เหมาะสม **การตรวจสอบรูปแบบของข้อมูล:** - ตรวจสอบว่าข้อมูลในแต่ละคอลัมน์มีรูปแบบที่ถูกต้องและเหมาะสมหรือไม่ เช่น ข้อมูลต้องเป็นตัวเลขหรือข้อความตามที่คาดหวัง - สำรวจความถูกต้องของข้อมูลเพื่อป้องกันการเกิดข้อผิดพลาดในการวิเคราะห์หรือการประมวลผลข้อมูลต่อไป **การเข้ารหัสข้อมูล (Encoding Data):** หากมีข้อมูลที่เป็นข้อความหรือข้อมูลที่ไม่ใช่ตัวเลข คุณต้องทำการเข้ารหัสเพื่อแปลงข้อมูลให้เป็นตัวเลข เช่น การใช้ One-Hot Encoding หรือการแปลงข้อความเป็นตัวเลข	หลังจากทำ Data Cleaning เรียบร้อยแล้ว ขั้นตอนต่อไปคือการเตรียมข้อมูลให้พร้อมสำหรับการใช้งานในโปรแกรมต่างๆ เช่น Excel, Google Sheets, Minitab และ Python โดยมีหลักการดังนี้ **ตรวจสอบโครงสร้างข้อมูล:** - ตรวจสอบให้แน่ใจว่าข้อมูลของคุณอยู่ในรูปแบบตารางที่มีแถวและคอลัมน์ที่ชัดเจน - แต่ละคอลัมน์ควรมีชื่อที่สื่อความหมาย - แต่ละแถวควรมีข้อมูลที่เกี่ยวข้องกัน - ตรวจสอบว่าไม่มีค่าที่ขาดหายไป (missing values) หรือค่าที่ผิดพลาด (invalid values) **ตรวจสอบประเภทข้อมูล:** - กำหนดประเภทข้อมูลของแต่ละคอลัมน์ เช่น ตัวเลข ข้อความ วันที่ เวลา ฯลฯ - ตรวจสอบให้แน่ใจว่าประเภทข้อมูลถูกต้องและสอดคล้องกับความเป็นจริง - แปลงประเภทข้อมูลหากจำเป็น **จัดรูปแบบข้อมูล:** - จัดรูปแบบข้อมูลให้สอดคล้องกับโปรแกรมที่ต้องการใช้งาน ตัวอย่างเช่น ใน Excel คุณสามารถใช้เครื่องมือจัดรูปแบบตัวเลข วันที่ เวลา ฯลฯ ใน Python คุณสามารถใช้ไลบรารี Pandas เพื่อจัดรูปแบบข้อมูล - ตรวจสอบให้แน่ใจว่ารูปแบบข้อมูลสอดคล้องกันทั้งไฟล์

ตารางที่ 3 คำตอบสำหรับคำถามที่ 3 โดยสรุปจากปัญญาประดิษฐ์เชิงสร้างสรรค์ทั้ง 3 ระบบ (ต่อ)

Claude.ai	ChatGPT 3.5	Gemini
กำหนดโครงสร้างของข้อมูลให้ชัดเจน - ข้อมูลแบบมิติ (dimensional) ควรถูกจัดในรูปแบบกว้าง (wide) หรือยาว (long) ตามความเหมาะสม - ข้อมูลแบบอนุกรมเวลา (time-series) ควรแยกคอลัมน์วันที่/เดือน/ปี/เวลา ข้อมูลประเภท ID ควรถูกแยกเป็นคอลัมน์เฉพาะ **ตรวจสอบและทดสอบความถูกต้อง** - ตรวจสอบความถูกต้องของทุกส่วนก่อนการแชร์ไฟล์สุดท้าย - ทดสอบการเปิดไฟล์และใช้งานกับโปรแกรมต่างๆ เช่น Excel, Python, R เป็นต้น ตรงสอบกับโปรแกรมย่อยๆ เช่น Google Sheets, SQL เพื่อให้มั่นใจในความเข้ากันได้ การปฏิบัติตามหลักการเหล่านี้ จะทำให้ dataset ของคุณสามารถนำไปใช้งานได้กับหลากหลายโปรแกรม มีความเข้ากันได้สูง และง่ายต่อการแบ่งปันและตีความหมายสำหรับผู้อื่น	**การบันทึกและการนำเข้าข้อมูล:** - บันทึกข้อมูลที่ถูกเตรียมแล้วในรูปแบบไฟล์ .csv โดยใช้โปรแกรมที่เหมาะสม เช่น Excel, Google Sheets, Minitab หรือ Python - ใน Python คุณสามารถใช้ไลบรารี pandas เพื่อนำเข้าข้อมูลและบันทึกเป็นไฟล์ .csv ได้อย่างสะดวก หลังจากที่คุณทำขั้นตอนทั้งหมดนี้เสร็จสิ้น ข้อมูลในรูปแบบไฟล์ .csv นั้น จะพร้อมใช้งานกับโปรแกรมต่างๆ เพื่อการวิเคราะห์หรือประมวลผลข้อมูลต่อไป	**บันทึกข้อมูลเป็นไฟล์ .CSV:** - บันทึกข้อมูลของคุณเป็นไฟล์ .CSV (Comma Separated Values) - ไฟล์ .CSV เป็นไฟล์ข้อความธรรมดาที่สามารถเปิดใช้งานได้กับโปรแกรมต่างๆ - ตรวจสอบให้แน่ใจว่าไฟล์ .CSV มีการเข้ารหัส UTF-8 **ทดสอบการใช้งาน:** - เปิดไฟล์ .CSV ในโปรแกรมที่ต้องการใช้งาน - ตรวจสอบว่าข้อมูลแสดงผลอย่างถูกต้อง - แก้ไขปัญหาใดๆ ที่พบ **เครื่องมือและเทคนิคเพิ่มเติม:** - ไลบรารี Pandas: ไลบรารี Pandas ใน Python เป็นเครื่องมือที่มีประสิทธิภาพสำหรับการจัดการและเตรียมข้อมูล สามารถใช้สำหรับการตรวจสอบโครงสร้างข้อมูล แปลงประเภทข้อมูล จัดรูปแบบข้อมูล และอื่นๆ - เครื่องมือ Data Cleaning: เครื่องมือ Data Cleaning มีหลายประเภทที่สามารถช่วยคุณทำความสะอาดและเตรียมข้อมูล เครื่องมือเหล่านี้สามารถช่วยระบุและแก้ไขข้อผิดพลาดในข้อมูล เช่น ค่าที่ขาดหายไป ค่าที่ผิดพลาด ข้อมูลซ้ำซ้อน ฯลฯ

5.1.7 ความเหมือนของผลลัพธ์ จากปัญญาประดิษฐ์เชิงสร้างสรรค์ทั้ง 3 ระบบ สามารถแบ่งได้เป็น 2 ประเด็นคือการบันทึกข้อมูลเป็นไฟล์ .csv โดยทั้ง 3 ระบบกล่าวถึงการบันทึกข้อมูลที่ผ่านการทำความสะอาดแล้วในรูปแบบไฟล์ .csv เนื่องจากเป็นรูปแบบที่สามารถใช้งานได้กับโปรแกรมต่าง ๆ อย่างกว้างขวาง ประเด็นที่สองคือการตรวจสอบและจัดการข้อมูลโดยทั้ง 3 ระบบกล่าวถึงการจัดรูปแบบโครงสร้างข้อมูลที่สำคัญ เช่น การตรวจสอบโครงสร้างข้อมูล ประเภทข้อมูล และรูปแบบข้อมูล เพื่อให้แน่ใจว่าข้อมูลมีความถูกต้องและสอดคล้องกันก่อนนำไปใช้งาน

5.1.8 ความแตกต่างของผลลัพธ์ Claude.ai ให้รายละเอียดที่ชัดเจนเกี่ยวการจัดรูปแบบข้อมูลให้เป็นมาตรฐาน การเลือกใช้ไฟล์ที่เหมาะสม การกำหนดค่า Delimiters และ Encoding การเพิ่ม Metadata การกำหนดโครงสร้างข้อมูล และการทดสอบความถูกต้องแต่ไม่ได้กล่าวถึงเครื่องมือหรือเทคนิคเฉพาะเจาะจง สำหรับ ChatGPT 3.5 ให้ผลลัพธ์ที่เน้นการบันทึกเป็นไฟล์ .csv การตั้งชื่อคอลัมน์ การตรวจสอบรูปแบบข้อมูล และการเข้ารหัสข้อมูล ตลอดจนกล่าวถึงการใช้โปรแกรม Excel, Google sheets, Python และไลบรารี pandas ในการบันทึกและนำเข้าไฟล์ .csv และ Gemini เน้นการตรวจสอบโครงสร้างข้อมูล ประเภทข้อมูล การจัดรูปแบบข้อมูล และการบันทึกเป็นไฟล์ .csv พร้อมแนะนำเครื่องมือ Pandas ใน Python และเครื่องมือสำหรับการทำความสะอาดข้อมูลทั่วไป

5.1.9 ความถูกต้องของผลลัพธ์ ที่เกี่ยวข้องกับการเตรียมข้อมูล (Data preparation) หลังจากการทำความสะอาดข้อมูลประกอบไปด้วยการตรวจสอบ และรูปแบบของข้อมูล การเข้ารหัสข้อมูล กำหนดโครงสร้างของข้อมูลให้ชัดเจน ตลอดจนการทดสอบความถูกต้อง และการนำไปใช้ ซึ่งจากข้อมูลที่ได้รับปัญญาประดิษฐ์เชิงสร้างสรรค์จึงมีความถูกต้องตามหลักการในการจัดเตรียมข้อมูล (Nandi & Sharma, 2020; Frazer, 2024)

สำหรับคำถามต่อมาหลังจากได้มีการเก็บข้อมูล และหลักการในการทำความสะอาดข้อมูลและเตรียมข้อมูลของปัญญาประดิษฐ์เชิงสร้างสรรค์ทั้ง 3 ระบบแล้ว จึงเป็นที่มาของคำถามที่ 4 ว่าสามารถทำ Data cleaning บนระบบปัญญาประดิษฐ์เชิงสร้างสรรค์ได้หรือไม่โดยมีรายละเอียดดังนี้

คำถามที่ 4 : สามารถทำ Data Cleaning บนระบบปัญญาประดิษฐ์เชิงสร้างสรรค์ได้หรือไม่

ตารางที่ 4 คำตอบสำหรับคำถามที่ 4 โดยสรุปจากปัญญาประดิษฐ์เชิงสร้างสรรค์ทั้ง 3 ระบบ

Claude.ai	ChatGPT 3.5	Gemini
ที่จริงแล้ว Claude ไม่สามารถทำการประมวลผลหรือจัดการไฟล์ข้อมูลโดยตรงได้ครับ เนื่องจากฉันเป็น AI ภายในแชทนี้เท่านั้น ไม่มีการเชื่อมต่อกับระบบไฟล์หรือสามารถเปิดไฟล์ข้อมูลได้	สำหรับการเตรียมข้อมูลหลังจากการทำ Cleaning เช่นการกำจัดค่า outliers, การแก้ไขค่า missing values, การทดสอบทางสถิติว่าข้อมูลมีการกระจายตัวแบบปกติหรือไม่ และการลดขนาดของตัวอย่าง ส่วนใหญ่จะต้องใช้การโปรแกรมมิ่งเพื่อวิเคราะห์และ	โดยทั่วไปแล้ว ใช่ Gemini สามารถช่วยคุณทำสิ่งเหล่านี้ได้จากไฟล์ .CSV ที่ประกอบด้วยข้อมูลเกี่ยวกับผลการทดลองของคุณ คุณต้องการทำความสะอาดและเตรียมข้อมูลนี้ให้พร้อมสำหรับการวิเคราะห์ใน Python

ตารางที่ 4 คำตอบสำหรับคำถามที่ 4 โดยสรุปจากปัญญาประดิษฐ์เชิงสร้างสรรค์ทั้ง 3 ระบบ (ต่อ)

Claude.ai	ChatGPT 3.5	Gemini
สำหรับการทำ data Cleaning, data preparation และการจัดการไฟล์ข้อมูล .csv คุณต้องใช้โปรแกรมหรือสภาพแวดล้อมการประมวลผลข้อมูลจริง เช่น: - Python และ libraries เช่น Pandas, Numpy - R และ packages สำหรับประมวลผลข้อมูล - Microsoft Excel หรือ Google Sheets - เครื่องมือสำเร็จรูปอื่นๆ เช่น SPSS, Minitab, SAS ฉันสามารถให้คำแนะนำและอธิบายขั้นตอนการจัดการข้อมูลในภาพรวมได้ แต่ไม่สามารถดำเนินการกับไฟล์ข้อมูลจริงๆ ได้ครับ คุณจะต้องนำคำแนะนำของฉันไปปฏิบัติในสภาพแวดล้อมการทำงานจริงของคุณ ดังนั้น หากคุณมีไฟล์ข้อมูล .csv ที่ต้องการทำ data Cleaning หรือเตรียมการ โปรดอัพโหลดหรือวางไฟล์นั้นในโปรแกรมที่คุณใช้งาน แล้วปฏิบัติตามคำแนะนำของฉันผ่านทางโปรแกรมนั้นๆ	ประมวลผลข้อมูลในระดับที่ซับซ้อนกว่าที่ ChatGPT 3.5 สามารถดำเนินการได้ในขณะนี้ครับ แต่ ChatGPT 3.5 สามารถช่วยในการให้คำแนะนำเกี่ยวกับขั้นตอนและหลักการในการเตรียมข้อมูลด้วยความรู้ที่มีอยู่ และยังสามารถช่วยในการตอบคำถามเกี่ยวกับการทำงานของขั้นตอนเหล่านี้ได้ด้วยความแม่นยำครับ หากคุณมีคำถามหรือต้องการคำแนะนำเพิ่มเติมเกี่ยวกับการเตรียมข้อมูล กรุณาสอบถามครับ	**การทำความสะอาดข้อมูล**: Gemini สามารถช่วยคุณระบุและแก้ไขข้อผิดพลาดในข้อมูล เช่น ค่าที่ขาดหายไป ค่าที่ผิดพลาด ข้อมูลซ้ำซ้อน ฯลฯ **การเตรียมข้อมูล**: Gemini สามารถช่วยคุณแปลงประเภทข้อมูล จัดรูปแบบข้อมูล และสร้างตัวแปรใหม่ **การวิเคราะห์ข้อมูล**: Gemini สามารถช่วยคุณวิเคราะห์ข้อมูลของคุณ เช่น การคำนวณสถิติ การสร้างกราฟ และการสร้างแบบจำลอง **อย่างไรก็ตาม มีข้อจำกัดบางประการ** - ความสามารถของ Gemini ขึ้นอยู่กับคุณภาพของข้อมูล: หากข้อมูลของคุณมีข้อผิดพลาดมาก Gemini อาจไม่สามารถระบุและแก้ไขข้อผิดพลาดทั้งหมดได้ - Gemini ไม่ได้ออกแบบมาเพื่อแทนที่เครื่องมือวิเคราะห์ข้อมูล: Gemini สามารถช่วยคุณทำความสะอาดและเตรียมข้อมูลของคุณ แต่คุณยังคงต้องใช้เครื่องมือวิเคราะห์ข้อมูลอื่นๆ เช่น Python R หรือ SPSS สำหรับการวิเคราะห์ข้อมูลเชิงลึก - Gemini ยังอยู่ระหว่างการพัฒนา: ความสามารถของ Gemini ยังคงได้รับการพัฒนาอยู่ ฟังก์ชันการทำงานบางอย่างอาจไม่มีให้บริการหรืออาจทำงานได้ไม่สมบูรณ์

5.1.10 ความเหมือนของผลลัพธ์ จากคำตอบดังตารางที่ 4 พบว่า Claude.ai ChatGPT 3.5 และ Gemini ไม่สามารถทำการประมวลผลหรือจัดการไฟล์ข้อมูลโดยตรง ต้องใช้โปรแกรมหรือสภาพแวดล้อมการประมวลผลข้อมูลจริง เช่น Python, R, Microsoft excel, Google sheets เป็นต้น โดยเป็นการใช้ความสามารถทางภาษาเพื่อการเขียนโค้ด หรือฟังก์ชันสำหรับการประมวลผล

5.1.11 ความแตกต่างของผลลัพธ์ แม้ Gemini มีการให้ข้อมูลว่าสามารถช่วยทำสิ่งเหล่านี้ได้จากไฟล์ .csv แต่เมื่อลองนำข้อมูลคัดลอกและวางลงในช่องคำถาม Gemini เป็นข้อความธรรมดา (Plain text) ที่มีลักษณะการเว้นวรรคตามช่องจากไฟล์ .csv และให้คำสั่งสำหรับแปลงข้อความธรรมดาให้เป็นรูปแบบของ .csv Gemini จะสร้างตาราง Google sheets จากข้อมูลที่ป้อนเข้าไป แต่ข้อมูลตอบกลับที่ได้เป็นตารางข้อมูลนั้นขาดความสมบูรณ์อย่างมาก เช่น จำนวนคอลัมน์ไม่ครบตามข้อมูลต้นฉบับ หรือมีแถวข้อมูลที่ถูกข้ามไป เป็นต้น สำหรับ ChatGPT 3.5 ได้ใช้การดำเนินการเดียวกันโดยข้อมูลตอบกลับที่ได้เป็นการเรียงรูปแบบข้อมูลที่มีเครื่องหมาย Comma (,) แต่เนื่องจากชุดข้อมูลนี้มีจำนวนแถวเป็นจำนวนมากจึงทำให้การแปลงข้อมูลใช้เวลานาน และไม่สามารถแปลงข้อมูลเป็นรูปแบบ .csv ได้ครบถ้วนจึงไม่สามารถดำเนินการต่อได้ สำหรับ Claude.ai แนบไฟล์ผลการทดลองนามสกุล .csv ได้ และสามารถอ่านเพื่อให้ข้อมูลเบื้องต้นเกี่ยวกับผลการทดลองได้ สามารถให้ข้อมูลเกี่ยวกับการทดลอง ค่าสมบัติที่ทำการทดสอบ และอิทธิพลเชิงแนวโน้มของปัจจัยที่ศึกษาที่ส่งผลต่อสมบัติการรับแรงดึงของคอมโพสิต ทั้งนี้เมื่อป้อนคำสั่งให้ทำการตรวจสอบและจัดการค่าผิดปกติ (Outliers) และค่าที่ขาดหายไป (Missing values) ซึ่งผลคำตอบที่ได้ Claude.ai สามารถตรวจสอบค่าที่หายไป

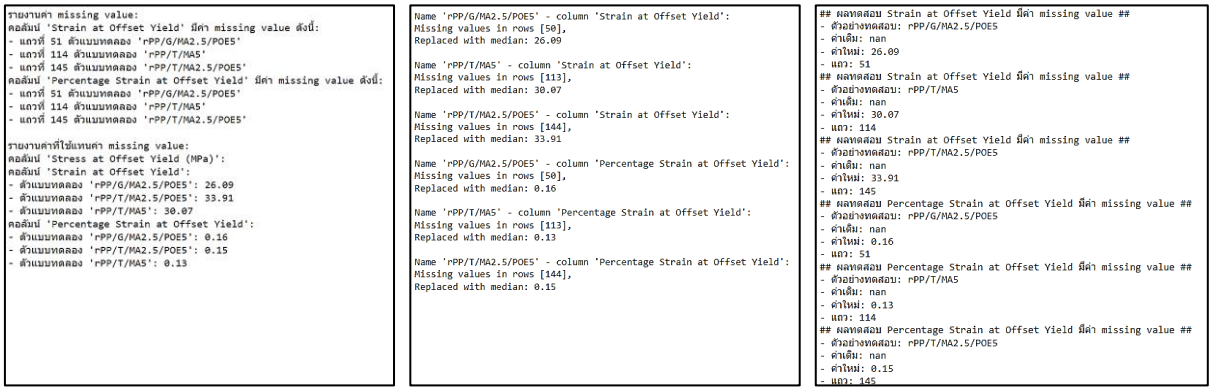
5.1.12 ความถูกต้องของผลลัพธ์ เมื่อลองนำข้อมูลที่ป้อนเข้าไปยังระบบปัญญาประดิษฐ์เชิงสร้างสรรค์ทั้ง 3 ระบบ โดย Claude.ai สามารถแนบไฟล์ผลการทดสอบสมบัติทางกลในรูปแบบ .csv ได้ ChatGPT 3.5 และ Gemini เป็นข้อความธรรมดา (Plain text) ที่มีลักษณะการเว้นวรรคตามช่องจากไฟล์.csv ไปวิเคราะห์หาค่าเฉลี่ยของแต่ละตัวแบบทดลองเพื่อจะนำไปแทนค่าที่หายไป พร้อมกับเปรียบเทียบผลด้วยโปรแกรม Minitab พบว่าผลการคำนวณของปัญญาประดิษฐ์เชิงสร้างสรรค์นั้นไม่ถูกต้อง เช่น ค่าเฉลี่ยที่คำนวณได้จากโปรแกรม Minitab ของตัวแบบทดลองหนึ่งมีค่าเฉลี่ยเท่ากับ 30.19 MPa แต่ Claude.ai คำนวณได้ 30.43 MPa เป็นต้น ดังนั้นผลที่แสดงเกี่ยวกับการคำนวณจึงเชื่อมั่นได้น้อย

สำหรับกรณีศึกษานี้เป็นการดำเนินงานเพื่อเตรียมความพร้อมของชุดข้อมูลที่เกี่ยวข้องกับการวิจัยวัสดุคอมโพสิตพอลิโพรพิลีนรีไซเคิลผสมกากชา โดยมีวัตถุประสงค์หลักในการทำความสะอาดและปรับปรุงคุณภาพของข้อมูลดิบ เพื่อนำไปสู่การวิเคราะห์ในขั้นตอนต่อไป กระบวนการดังกล่าวประกอบด้วยการศึกษาปัจจัยทั้งสิ้น 3 ปัจจัย ซึ่งแบ่งออกเป็นปัจจัยเชิงคุณภาพ 1 ปัจจัย และปัจจัยเชิงปริมาณ 2 ปัจจัย โดยผลการทดสอบสมบัติทางกลจากเครื่องคอมพิวเตอร์เป็นไฟล์นามสกุล .csv ผ่านโปรแกรมบันทึกผลการทดสอบสมบัติ เพื่อดำเนินการประมวลผลข้อมูลดังกล่าวผู้วิจัยได้เลือกใช้ไพธอนสคริป (Python) บนพื้นที่ทำงาน Google colab การเขียนคำสั่ง (Prompt) ในการศึกษานี้ใช้แบบการกำหนดบุคลิกลักษณะ (Persona-based)

กำหนดให้ปัญญาประดิษฐ์เชิงสร้างสรรค์รับบทบาทเป็นนักวิทยาศาสตร์ข้อมูลที่เชี่ยวชาญด้านการเขียน Python code เพื่อการวิเคราะห์ข้อมูล ทำการอธิบายโครงสร้างข้อมูลของ Data set ที่จะทำการวิเคราะห์ว่ามีคอลัมน์ใดบ้าง แต่ละคอลัมน์ทำหน้าที่อะไร บันทึกข้อมูลอะไร ตลอดจนชื่อของ Data set และเส้นทางข้อมูลที่บันทึกไว้บนพื้นที่ทำงาน Google colab จะช่วยให้ปัญญาประดิษฐ์เชิงสร้างสรรค์ตอบสนองได้ตรงกับบทบาทที่กำหนดมากขึ้น ส่งผลให้ผลลัพธ์มีความสอดคล้องกับสิ่งที่คาดหวังในบริบทและบทบาทที่ตั้งไว้เพื่อเขียนโค้ดไพธอนสคริปออกมา จากนั้นจึงนำโค้ดที่ได้จากทั้ง 3 ระบบมาใช้ในการประมวลผลข้อมูลและเปรียบเทียบผลลัพธ์ที่ได้รับในเชิงการอภิปรายโดยมีผลดังนี้

5.2 การจัดการค่าที่ขาดหายไป (Missing values)

การจัดการค่าที่ขาดหายไป (Missing values) จะทำให้ข้อมูลสมบูรณ์มากขึ้นตลอดจนช่วยให้การวิเคราะห์ และจัดการค่าสุดโต่ง (Outlier) มีประสิทธิภาพมากขึ้นเพราะค่าที่หายไปอาจมีผลต่อการคำนวณสถิติเช่น ค่าเฉลี่ย หรือส่วนเบี่ยงเบนมาตรฐานที่ใช้ในการตรวจจับค่าสุดโต่ง การศึกษานี้ทำการกำหนดคำสั่งให้ทำการค้นหาค่าที่หายไปว่าอยู่แถวที่เท่าใด เป็นตัวแบบทดลองใดค่าสมบัติใดที่หายไป จากนั้นให้ทำการเติมค่ามัธยฐาน (Median imputation) ที่คำนวณจากข้อมูลของแต่ละตัวแบบทดลองเนื่องจากค่านี้ไม่อ่อนไหวต่อค่าสุดโต่งทำให้เป็นทางเลือกที่เหมาะสมเมื่อไม่ทราบการกระจายตัวของข้อมูล จากนั้นให้ทำการบันทึกไฟล์ใหม่ในรูปแบบของ .csv เพื่อนำไปดำเนินการในขั้นตอนต่อไป โดยผลลัพธ์จากการจากการวิเคราะห์ข้อมูลจาก Python code จากปัญญาประดิษฐ์ทั้ง 3 ระบบสามารถแสดงได้ดังรูปที่ 2



```
รายงานค่า missing value:
คอลัมน์ 'Strain at Offset Yield' มีค่า missing value ดังนี้:
- แถวที่ 51 ตัวแบบทดลอง 'rPP/G/MA2.5/POE5'
- แถวที่ 114 ตัวแบบทดลอง 'rPP/T/MA5'
- แถวที่ 145 ตัวแบบทดลอง 'rPP/T/MA2.5/POE5'
คอลัมน์ 'Percentage Strain at Offset Yield' มีค่า missing value ดังนี้:
- แถวที่ 51 ตัวแบบทดลอง 'rPP/G/MA2.5/POE5'
- แถวที่ 114 ตัวแบบทดลอง 'rPP/T/MA5'
- แถวที่ 145 ตัวแบบทดลอง 'rPP/T/MA2.5/POE5'

รายงานค่าที่ใช้แทนค่า missing value:
คอลัมน์ 'Stress at Offset Yield (MPa)':
คอลัมน์ 'Strain at Offset Yield':
- ตัวแบบทดลอง 'rPP/G/MA2.5/POE5': 26.09
- ตัวแบบทดลอง 'rPP/T/MA2.5/POE5': 33.91
- ตัวแบบทดลอง 'rPP/T/MA5': 30.07
คอลัมน์ 'Percentage Strain at Offset Yield':
- ตัวแบบทดลอง 'rPP/G/MA2.5/POE5': 0.16
- ตัวแบบทดลอง 'rPP/T/MA2.5/POE5': 0.15
- ตัวแบบทดลอง 'rPP/T/MA5': 0.13
```

```
Name 'rPP/G/MA2.5/POE5' - column 'Strain at Offset Yield':
Missing values in rows [50],
Replaced with median: 26.09

Name 'rPP/T/MA5' - column 'Strain at Offset Yield':
Missing values in rows [113],
Replaced with median: 30.07

Name 'rPP/T/MA2.5/POE5' - column 'Strain at Offset Yield':
Missing values in rows [144],
Replaced with median: 33.91

Name 'rPP/G/MA2.5/POE5' - column 'Percentage Strain at Offset Yield':
Missing values in rows [50],
Replaced with median: 0.16

Name 'rPP/T/MA5' - column 'Percentage Strain at Offset Yield':
Missing values in rows [113],
Replaced with median: 0.13

Name 'rPP/T/MA2.5/POE5' - column 'Percentage Strain at Offset Yield':
Missing values in rows [144],
Replaced with median: 0.15
```

```
## ผลทดสอบ Strain at Offset Yield มีค่า missing value ##
- ตัวอย่างทดสอบ: rPP/G/MA2.5/POE5
- ค่าเดิม: nan
- ค่าใหม่: 26.09
- แถว: 51
## ผลทดสอบ Strain at Offset Yield มีค่า missing value ##
- ตัวอย่างทดสอบ: rPP/T/MA5
- ค่าเดิม: nan
- ค่าใหม่: 30.07
- แถว: 114
## ผลทดสอบ Strain at Offset Yield มีค่า missing value ##
- ตัวอย่างทดสอบ: rPP/T/MA2.5/POE5
- ค่าเดิม: nan
- ค่าใหม่: 33.91
- แถว: 145
## ผลทดสอบ Percentage Strain at Offset Yield มีค่า missing value ##
- ตัวอย่างทดสอบ: rPP/G/MA2.5/POE5
- ค่าเดิม: nan
- ค่าใหม่: 0.16
- แถว: 51
## ผลทดสอบ Percentage Strain at Offset Yield มีค่า missing value ##
- ตัวอย่างทดสอบ: rPP/T/MA5
- ค่าเดิม: nan
- ค่าใหม่: 0.13
- แถว: 114
## ผลทดสอบ Percentage Strain at Offset Yield มีค่า missing value ##
- ตัวอย่างทดสอบ: rPP/T/MA2.5/POE5
- ค่าเดิม: nan
- ค่าใหม่: 0.15
- แถว: 145
```

ก ข ค

รูปที่ 2 ผลลัพธ์โค้ดไพธอนสคริปในการจัดการค่าที่ขาดหายไป (Missing values) จากปัญญาประดิษฐ์เชิงสร้างสรรค์ทั้ง 3 ระบบ ก. Claude.ai ข. ChatGPT 3.5 ค. Gemini

5.2.1 ความเหมือนของผลลัพธ์

ทั้ง 3 ระบบสามารถระบุตัวแบบทดลอง และแถวที่พบค่าที่ขาดหายได้โดยพบที่สมบัติ Strain at offset Yield ในตัวแบบทดลอง rPP/G/MA2.5/POE5rPP/T/MA5rPP/T/MA2.5/POE5 และสมบัติ Percentage Strain at offset Yield ในตัวแบบทดลอง rPP/G/MA2.5/POE5 rPP/T/MA5 และ rPP/T/MA2.5/POE5 ซึ่งค่ามัธยฐานที่ถูกแทนที่คือ 26.09, 33.91, 30.07 สำหรับสมบัติ Strain at offset yield และสมบัติ Percentage Strain at offset yield แทนค่าด้วย 0.16, 0.15 และ 0.13 ตามลำดับตัวแบบทดลองที่เรียงลำดับไปข้างต้น

5.2.2 ความต่างของผลลัพธ์ ส่วนที่แตกต่างอย่างเห็นได้ชัดมี 2 ประเด็นคือการระบุแถวที่พบค่าที่ขาดหายนั้นแตกต่างกัน 1 ช่อง ดังเช่น ChatGPT 3.5 พบค่าที่ขาดหายอยู่ในแถวที่ 50 แต่สำหรับ Claude.ai และ Gemini ระบุว่าอยู่ในแถวที่ 51 ทั้งนี้เนื่องจาก Claude.ai ได้ตีความว่าแถวที่ 1 ซึ่งเป็นแถวที่ใช้ระบุข้อมูลหัวตารางของแต่ละคอลัมน์จึงเริ่มนับแถวที่ 1 ถัดจากหัวตาราง และอีกประเด็นคือรูปแบบการนำเสนอ โดย Claude.ai ได้เสนอว่ามีค่าที่ขาดหายในค่าสมบัติทางกลใดก่อน จากนั้นให้ข้อมูลว่ามีแถวใดบ้างที่มีข้อมูลขาดหาย และเป็นของตัวแบบทดลองใดในย่อหน้าแรกตามลำดับ ย่อหน้าต่อมาจึงเป็นการอธิบายว่าค่ามัธยฐานที่คำนวนได้ของแต่ละตัวแบบทดลองที่จะแทนที่ค่าที่ขาดหายเป็นจำนวนใด สำหรับ ChatGPT 3.5 ได้ให้ข้อมูลว่าตัวแบบทดลองใดจากการทดสอบค่าใดที่มีค่าที่ขาดหายอยู่ในแถวใด และแทนที่ด้วยค่าใดซึ่งแสดงแยกเป็นแต่ละตัวแบบทดลอง และ Gemini มีการแสดงผลที่คล้ายกับ ChatGPT 3.5 แต่มีการเรียงการแสดงผลแยกเป็นบรรทัดเรียงลำดับจากบนลงล่าง ได้แก่ ตัวแบบทดลอง ค่าเดิมที่เป็นค่าที่ขาดหาย ค่าใหม่ที่เป็นค่ามัธยฐาน และแถวที่พบค่าที่ขาดหาย

5.2.3 ความถูกต้องของผลลัพธ์

ทั้ง 3 ระบบสามารถให้คำตอบที่ถูกต้องว่ามีตัวแบบทดลองใด แถวที่เท่าใดในไฟล์ชุดข้อมูลที่มีค่าที่ขาดหายไป โดยค่าที่ขาดหายไปนี้เกิดขึ้นได้จากการทดสอบด้วยเครื่องทดสอบและรายงานผลไปยังคอมพิวเตอร์ควบคุมซึ่งในบางการทดสอบตัวอย่างบางชิ้นอาจมีการจับยึดที่ไม่เหมาะสมส่งผลต่อการอ่านผลที่คลาดเคลื่อนจึงทำให้ไม่สามารถบันทึกข้อมูลได้ สำหรับค่ามัธยฐานจากการประมวลด้วยโปรแกรม Mintab 18 เทียบกับผลการดำเนินงานด้วยปัญญาประดิษฐ์เชิงสร้างสรรค์มีความถูกต้องตรงกันโดยมีค่าตามผลลัพธ์ที่แสดงดังรูปที่ 2

5.3 การจัดการค่าสุดโต่ง (Outlier)

หลังจากการจัดการค่าที่ขาดหายในขั้นตอนที่แล้วได้ทำการบันทึกเป็นไฟล์ใหม่ ขั้นตอนต่อมาคือการจัดการค่าสุดโต่ง (Outlier) รูปที่ 3 แสดงผลลัพธ์จากการใช้โค้ดไพธอนสคริปสำหรับการหาค่าสุดโต่งด้วยวิธี IQR ซึ่งมีเป้าหมายให้ทำการคำนวณหาค่าสุดโต่งของแต่ละตัวแบบทดลอง จากนั้นให้ทำการแสดงผลว่าตัวแบบทดลองใดมีค่าสุดโต่ง มีค่าเป็นเท่าใด อยู่ในแถวที่เท่าใด และแทนที่ด้วยค่าเฉลี่ยของตัวแบบทดลองนั้นๆ เนื่องจากการใช้ค่าเฉลี่ยนั้นคำนวณได้ง่าย เหมาะสมกับข้อมูลที่ไม่มีค่าสุดโต่งเนื่องจากมีการจัดการไปก่อนหน้านี้แล้ว หากตัวแบบทดลองใดไม่มีค่าสุดโต่งจะไม่มีการแสดงผล

```
คอลัมน์ 'Stress at Offset Yield (MPa)' มีค่า outliers ดังนี้:
- ค่า 31.88 จากตัวแบบทดลอง 'rPP/G/POE10' ที่แถวที่ 47
- ค่า 31.7 จากตัวแบบทดลอง 'rPP/G/POE10' ที่แถวที่ 48
- ค่า 32.71 จากตัวแบบทดลอง 'rPP/T/POE10' ที่แถวที่ 132
- ค่า 32.63 จากตัวแบบทดลอง 'rPP/T/POE10' ที่แถวที่ 133
- ค่า 41.57 จากตัวแบบทดลอง 'rPP/T/POE10' ที่แถวที่ 136

ค่าที่ใช้แทนค่า Outlier ในแต่ละตัวแบบทดลอง (ID) สำหรับคอลัมน์ '{col}' ดังนี้:
- ตัวแบบทดลอง 'rPP/G/POE10': 25.413
- ตัวแบบทดลอง 'rPP/T/POE10': 25.771000000000004
คอลัมน์ 'Strain at Offset Yield' มีค่า outliers ดังนี้:
- ค่า 14.19 จากตัวแบบทดลอง 'rPP/G/MA5/POE10' ที่แถวที่ 86
- ค่า 14.91 จากตัวแบบทดลอง 'rPP/T/MA2.5' ที่แถวที่ 107
- ค่า 11.43 จากตัวแบบทดลอง 'rPP/T/POE5' ที่แถวที่ 130
- ค่า 11.42 จากตัวแบบทดลอง 'rPP/T/MA2.5/POE5' ที่แถวที่ 146
- ค่า 40.36 จากตัวแบบทดลอง 'rPP/T/MA2.5/POE5' ที่แถวที่ 148
- ค่า 37.11 จากตัวแบบทดลอง 'rPP/T/MA2.5/POE5' ที่แถวที่ 150

ค่าที่ใช้แทนค่า Outlier ในแต่ละตัวแบบทดลอง (ID) สำหรับคอลัมน์ '{col}' ดังนี้:
- ตัวแบบทดลอง 'rPP/G/MA5/POE10': 23.773999999999997
- ตัวแบบทดลอง 'rPP/T/MA2.5': 29.937000000000005
- ตัวแบบทดลอง 'rPP/T/MA2.5/POE5': 31.053333333333327
- ตัวแบบทดลอง 'rPP/T/POE5': 26.242
คอลัมน์ 'Percentage Strain at Offset Yield' ไม่มีค่า outliers
```

```
------------------------------
Outliers found in column 'Stress at Offset Yield (MPa)':
Number of outliers: 5
Lower bound: 7.699999999999999
Upper bound: 29.98
            Name  Stress at Offset Yield (MPa)
46   rPP/G/POE10                        31.88
47   rPP/G/POE10                        31.70
131  rPP/T/POE10                        32.71
132  rPP/T/POE10                        32.63
135  rPP/T/POE10                        41.57
Replacing outliers for Name 'rPP/G/POE10' with mean value 25.41
Replacing outliers for Name 'rPP/T/POE10' with mean value 25.77
------------------------------
Outliers found in column 'Strain at Offset Yield':
Number of outliers: 5
Lower bound: 16.89625
Upper bound: 37.366249999999994
                 Name  Strain at Offset Yield
85    rPP/G/MA5/POE10                   14.19
106       rPP/T/MA2.5                   14.91
129             rPP/T/POE5              11.43
145  rPP/T/MA2.5/POE5                   11.42
147  rPP/T/MA2.5/POE5                   40.36
Replacing outliers for Name 'rPP/G/MA5/POE10' with mean value 23.77
Replacing outliers for Name 'rPP/T/MA2.5' with mean value 29.94
Replacing outliers for Name 'rPP/T/POE5' with mean value 26.24
Replacing outliers for Name 'rPP/T/MA2.5/POE5' with mean value 31.34
------------------------------
Outliers found in column 'Percentage Strain at Offset Yield':
No outliers found.
```

```
## ผลทดสอบ Stress at Offset Yield (MPa) มีค่า Outliers ##
- ตัวอย่างทดสอบ: rPP/G/POE10
- ค่าเดิม: 31.88
- ค่าใหม่: 25.413
- แถว: 47
- ตัวอย่างทดสอบ: rPP/G/POE10
- ค่าเดิม: 31.7
- ค่าใหม่: 24.7663
- แถว: 48
- ตัวอย่างทดสอบ: rPP/T/POE10
- ค่าเดิม: 32.71
- ค่าใหม่: 25.771000000000004
- แถว: 132
- ตัวอย่างทดสอบ: rPP/T/POE10
- ค่าเดิม: 32.63
- ค่าใหม่: 25.0771
- แถว: 133
- ตัวอย่างทดสอบ: rPP/T/POE10
- ค่าเดิม: 41.57
- ค่าใหม่: 24.321810000000006
- แถว: 136
## ผลทดสอบ Strain at Offset Yield มีค่า Outliers ##
- ตัวอย่างทดสอบ: rPP/G/MA5/POE10
- ค่าเดิม: 14.19
- ค่าใหม่: 23.773999999999997
- แถว: 86
- ตัวอย่างทดสอบ: rPP/T/MA2.5
- ค่าเดิม: 14.91
- ค่าใหม่: 29.937000000000005
- แถว: 107
- ตัวอย่างทดสอบ: rPP/T/POE5
- ค่าเดิม: 11.43
- ค่าใหม่: 26.242
- แถว: 130
- ตัวอย่างทดสอบ: rPP/T/MA2.5/POE5
- ค่าเดิม: 11.42
- ค่าใหม่: 31.339000000000006
- แถว: 146
- ตัวอย่างทดสอบ: rPP/T/MA2.5/POE5
- ค่าเดิม: 40.36
- ค่าใหม่: 33.3309
- แถว: 148
## ผลทดสอบ Percentage Strain at Offset Yield ไม่มีค่า Outliers ##
```

ก ข ค

รูปที่ 3 ผลลัพธ์โค้ดไพธอนสคริปในการจัดการค่าสุดโต่ง (Outlier) จากปัญญาประดิษฐ์เชิงสร้างสรรค์ ทั้ง 3 ระบบ ก. Claude.ai ข. ChatGPT 3.5 ค. Gemini

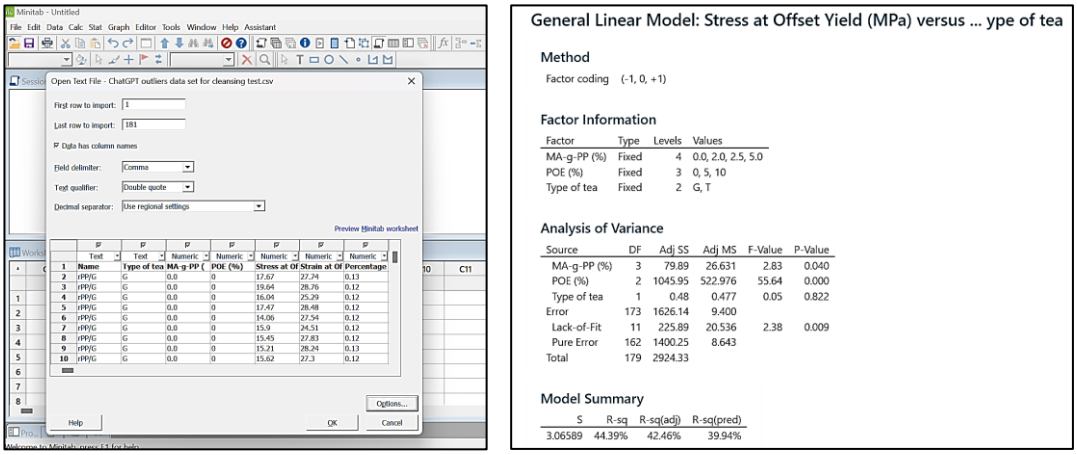
5.3.1 ความเหมือนของผลลัพธ์ ของทั้ง 3 ระบบสามารถแบ่งได้เป็น 3 ประเด็น ได้แก่ ตำแหน่งของ Outliers โดยทั้ง 3 ระบุตำแหน่งของ Outliers ในคอลัมน์ Stress at offset yield (MPa) และ Strain at offset yield ได้ตรงกัน ซึ่งตำแหน่งที่ระบุตรงกันคือ คอลัมน์ค่าสมบัติ Stress at Offset Yield (MPa) แถวที่ 2, 3, 5, 6, 7, 8 และ 9 คอลัมน์ค่าสมบัติ Strain at offset yield แถวที่ 2, 3, 5, 6, 7, 8 และ 9 ประเด็นต่อมาคือ การใช้เทคนิค IQR (Interquartile Range) ในการตรวจจับ Outliers เนื่องจากมีการใช้ค่า Q1, Q3 และ IQR ในการคำนวณ และประเด็นสุดท้ายการจัดการ Outliers ด้วยค่าเฉลี่ยซึ่งทั้ง 3 ระบบเลือกที่จะจัดการ Outliers ด้วยการแทนที่ด้วยค่าเฉลี่ยของแต่ละกลุ่มตัวอย่างทดลอง

5.3.2 ความแตกต่างของผลลัพธ์ พบว่ามีความแตกต่างในด้านรูปแบบการนำเสนอ Claude.ai แสดงผลลัพธ์ในรูปแบบข้อความบรรยายพร้อมทั้งระบุตำแหน่ง (แถวที่) และค่าของ Outliers ก่อนและหลังการแทนที่ด้วยค่าเฉลี่ย ChatGPT 3.5 แสดงผลลัพธ์ในรูปแบบตาราง แสดงค่า Outliers ก่อนการแทนที่ พร้อมระบุจำนวน Outliers ค่าขอบเขตล่างและบน และค่าเฉลี่ยที่ใช้แทนที่ และ Gemini แสดงผลลัพธ์ในรูปแบบข้อความบรรยายแสดงเฉพาะค่า Outliers ก่อนและหลังการแทนที่ นอกจากนี้สิ่งที่แตกต่างคือการระบุแถวที่พบค่าสุดโต่งซึ่งสอดคล้องกับการจัดการค่าที่ขาดหาย

5.3.3 ความถูกต้องของผลลัพธ์ โดยทั้ง 3 ระบบสามารถระบุตำแหน่งของ Outliers ได้อย่างถูกต้องแม้จะมีความต่างในการตีความลำดับแถว นอกจากนี้นเลือกวิธีการจัดการ Outliers ด้วยการแทนที่ด้วยค่าเฉลี่ยซึ่งเป็นวิธีที่เหมาะสม (Chaki et al., 2024) สำหรับค่าเฉลี่ยที่แทนที่ในแต่ละตัวแบบทดลองมีความถูกต้องเมื่อเทียบจากการประมวลด้วยโปรแกรม Mintab18

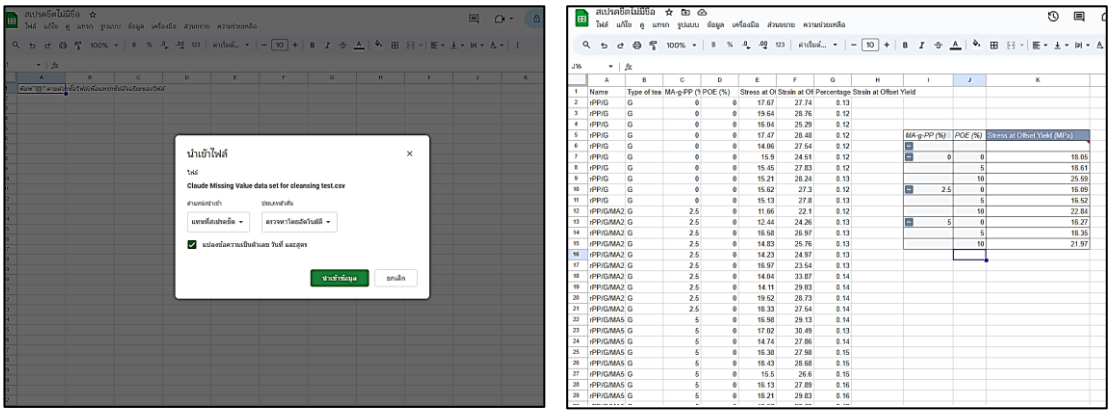
5.4 การตรวจสอบคุณภาพชุดข้อมูลและการนำไปใช้

ภาพที่ 4, 5 และ 6 แสดงตัวอย่างการนำชุดข้อมูลที่ผ่านการทำความสะอาดและจัดเตรียมแล้วมาใช้ พบว่าสามารถนำไปใช้ได้ทั้งโปรแกรม Minitab แพลตฟอร์ม Google sheet โดยที่ไม่พบปัญหาหรือข้อผิดพลาดที่เกิดจากความไม่สมบูรณ์หรือความไม่สะอาดของข้อมูล และในภาพที่ 7 การตรวจสอบคุณภาพข้อมูลของโปรแกรม Rapid Miner ไม่พบความผิดพลาดของข้อมูล (Data error) หรือค่าที่ขาดหาย (Missing value) ดังนั้นจึงสามารถสรุปได้ว่าการประยุกต์ใช้ปัญญาประดิษฐ์เชิงสร้างสรรค์ (Generative Artificial Intelligence) เพื่อช่วยในการดำเนินการทำความสะอาดข้อมูล (Data cleaning) และการเตรียมข้อมูล (Data preparation) นั้นบรรลุตามวัตถุประสงค์



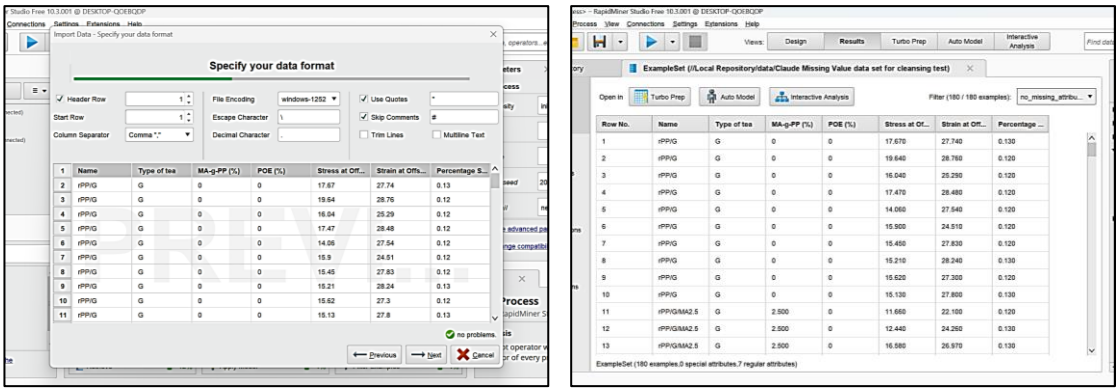
ก ข

รูปที่ 4 (ก) การนำเข้าชุดข้อมูลในโปรแกรม Minitab (ข) การประมวลผลทางสถิติด้วยโปรแกรม Minitab



ก ข

รูปที่ 5 (ก) การนำเข้าชุดข้อมูลในแพลตฟอร์ม Google sheet (ข) การประมวลผลทางสถิติด้วยในแพลตฟอร์ม Google Sheets



ก ข

รูปที่ 6 (ก) การนำเข้าชุดข้อมูลในโปรแกรม Rapid Miner (ข) การตรวจสอบคุณภาพข้อมูลก่อนการวิเคราะห์ด้วยโปรแกรม Rapid Miner

6. อภิปรายผลการวิจัย

จากการสร้างคำถามหรือโจทย์ (Prompt) เพื่อให้ได้วิธีการประยุกต์ใช้ปัญญาประดิษฐ์สร้างสรรค์เพื่อการทำความสะอาดข้อมูล และการเตรียมข้อมูลจนมาสู่การใช้แพลตฟอร์ม Google colab เพื่อการทำความสะอาดข้อมูลในงานวิจัยนี้ ไพธอนสคริปที่ได้จากการวิจัยนี้มิได้เกิดจากการสร้างเพียงครั้งเดียว โดยจากการเริ่มต้นคำถามและคำสั่งด้วยวิธีการกำหนดบทบาทพร้อมทั้งให้คำสั่งและบริบท เมื่อปัญญาประดิษฐ์เชิงสร้างสรรค์ทั้ง 3 ระบบให้คำตอบออกมาและนำไปใช้บนแพลตฟอร์ม Google colab เมื่อดำเนินการกับไพธอนสคริปที่ได้รับพบว่าผลลัพธ์ที่ได้ยังมีข้อผิดพลาด (Error) โดยเมื่อนำข้อความแจ้งเตือนความผิดพลาดกลับไปตั้งเป็นคำถามให้กับปัญญาประดิษฐ์ก็สามารถแก้ไขไพธอนสคริป (Python code) จนสามารถดำเนินการได้ แต่ทั้งนี้ต้องอาศัยการให้ข้อมูลเพิ่มเติม และตั้งคำถามเชิงตอบโต้เพื่อให้ปัญญาประดิษฐ์เชิงสร้างสรรค์เข้าใจสิ่งที่เราต้องการดำเนินการมากขึ้น ทั้งนี้นอกเหนือจากไพธอนสคริปที่ได้รับ ยังมีคำอธิบายการทำงานของไพธอนสคริปที่สร้างขึ้นเป็นขั้นตอน โดยทั้ง 3 ระบบมีการให้ข้อมูลแตกต่างกันไปสอดคล้องกับไพธอนสคริปที่สร้างขึ้นโดยผู้ใช้สามารถนำไปใช้พร้อมกับทำความเข้าใจหลักการทำงานของไพธอนสคริปที่ได้รับ ทั้งนี้สิ่งที่ช่วยให้การสร้างไพธอนสคริปของปัญญาประดิษฐ์สามารถนำไปวิเคราะห์ข้อมูลได้ตามจุดประสงค์คือการอธิบายโครงสร้างข้อมูลที่ถูกต้อง เช่น จำนวนคอลัมน์ ชื่อของคอมลัมน์ บทบาทของแต่ละคอลัมน์ เป็นต้น และแม้จะทำการกำหนดบทบาทให้ปัญญาประดิษฐ์สวมบทบาทเป็นนักวิทยาศาสตร์ข้อมูลเพื่อดึงศักยภาพของปัญญาประดิษฐ์มาใช้อย่างเต็มที่ แต่ยังต้องมีการให้ข้อมูล และเพิ่มถามคำถามในลักษณะโต้ตอบเพื่อให้มีข้อมูลเพียงพอในการสร้างไพธอนสคริปสำหรับการนำไปทำความสะอาดข้อมูลแต่ละขั้นตอน ทั้งนี้เมื่อมีการกำหนดคำสั่งหรือถามคำถามมากขึ้นปัญญาประดิษฐ์จะมีการตอบโต้จากการเรียนรู้มากขึ้นเช่นกัน ตลอดจนการเขียนไพธอนสคริปของการทำความสะอาดข้อมูลแต่ละขั้นตอนปัญญาประดิษฐ์จะใช้รูปแบบที่เคยเขียนออกมาก่อนหน้านี้ ดังนั้นจึงทำให้การแสดงผลของแต่ละขั้นตอนมีลำดับ และรูปแบบที่คล้ายกันในแต่ละระบบของปัญญาประดิษฐ์

จากงานวิจัยที่เกี่ยวข้องพบว่าการทำความสะอาดข้อมูลด้วยภาษาไพธอนเป็นกระบวนการสำคัญในการจัดการกับค่าที่หายไปและค่าผิดปกติในชุดข้อมูล ทำให้ข้อมูลมีความถูกต้องและน่าเชื่อถือสำหรับการวิเคราะห์ ไลบรารี่ไพธอนมีเทคนิคและเครื่องมือที่มีประสิทธิภาพสำหรับการทำความสะอาดข้อมูล ซึ่งเหนือกว่าวิธีการแบบดั้งเดิมเช่น Excel ในด้านความเร็วและความแม่นยำ เทคนิคต่างๆ ที่ถูกนำมาใช้มีอยู่ด้วยกันหลายเทคนิค เช่น วิธีการประมาณค่า การประมวลผลการแจกแจงปกติ และอัลกอริทึม K-Nearest Neighbor (KNN) เป็นต้น การศึกษาของ Jin (2022) ได้เสนอเทคนิคเพื่อการทำความสะอาดข้อมูลด้วยกราฟ (Graphical data cleaning) ได้แก่ ฮีตโตแกรม กราฟเส้น กราฟแท่ง แผนภาพการกระจาย แผนภาพควอไทล์ โดยผลการศึกษาพบว่ากราฟแต่ละชนิดมีความเหมาะสมกับประเภทข้อมูลที่ต่างกัน ไม่สามารถใช้ชนิดใดชนิดหนึ่งได้สำหรับทุกๆ ประเภทข้อมูล เนื่องจากกราฟแต่ละประเภทมีข้อจำกัด เช่น กราฟแท่ง มีความเหมาะสมสำหรับชุดข้อมูลขนาดเล็กที่ไม่เกิน 1,000 ข้อมูล หรือหากข้อมูลไม่ได้มีการแจกแจงแบบปกติ การใช้ฮีสโตแกรมอาจไม่สามารถตรวจจับค่าสุดโต่งได้ชัดเจน เป็นต้น ดังนั้นจึงจำเป็นต้องใช้วิธีการทางคณิตศาสตร์ขั้นสูงในการดำเนินการทำความสะอาดข้อมูล ซึ่งสอดคล้องกับการศึกษาของ Dasari and Varma (2022) ได้ทำการศึกษาการใช้ไพธอนสำหรับการทำความสะอาดข้อมูลด้วยเทคนิคต่างๆ ผ่านกรณีศึกษาเพื่อเปรียบเทียบกับการใช้โปรแกรมไมโครซอฟเอ็กเซล (Microsoft excel) สำหรับการจัดการค่าที่ขาดหาย ค่าที่ซ้ำซ้อน ตลอดจนการสะกดคำผิด ซึ่งปัญหาเหล่านี้นำมาซึ่งคุณภาพข้อมูลที่ต่ำและทำให้การนำไปวิเคราะห์ขาดคุณภาพ โดยผลการศึกษาพบว่าการใช้ Excel สามารถทำความสะอาดข้อมูลได้ตามวัตถุประสงค์ แต่มีข้อด้อยคือเวลาที่ใช้ในการดำเนินการที่นาน และซับซ้อน อีกทั้งต้องการองค์ความรู้ขั้นสูงในการประยุกต์ใช้สูตรการคำนวณ แต่การใช้ไพธอนนั้นสามารถดำเนินการได้อย่างรวดเร็วและแม่นยำกว่า แต่ผู้ใช้ยังจำเป็นต้องมีความรู้ทางด้านการโปรแกรมภาษาไพธอนเพื่อสร้างสคริปในการประมวลผล การทำความสะอาดข้อมูลขั้นที่สูงขึ้นโดยใช้อัลกอริทึมของปัญญาประดิษฐ์ ได้ทำการศึกษาโดย Martinez-Luengo et al. (2019) โดยศึกษาการใช้ K-Nearest Neighbor (KNN) และ Artificial Neural Network (ANN) เพื่อการแทนค่าที่ขาดหาย และค่าความคลาดเคลื่อนต่างๆ (Noise) สำหรับการจัดการข้อมูลของระบบการตรวจสภาพโครงสร้างของกังหันลมนอกชายฝั่ง โดยมุ่งเน้นไปที่การเพิ่มความแม่นยำในการประเมินความล้าตัวของโครงสร้างโดยใช้ข้อมูลต่อเนื่องแบบด้วยเวลาจริง (Real time monitoring) ซึ่งอัลกอริทึมจะรับข้อมูลแบบเวลาจริงเข้าในระบบ และเมื่อตรวจพบข้อมูลที่ขาดหายจะนำเข้าสู่อัลกอริทึมเพื่อประมวลผลและทำนายค่าที่เป็นไปได้มากที่สุดและแทนที่ลงในชุดข้อมูลโดยสรุปผลพบว่ามีความผิดพลาดเพียง 2.1% ส่งผลให้การตรวจสอบคุณภาพและประเมินอายุการใช้งานของโครงสร้างกังหันลมนอกชายฝั่งได้แม่นยำมากขึ้น (Dasari & Varma, 2022; Jin, 2022; Torkey et al., 2021; Mertz, 2021; Martinez-Luengo et al., 2019) แต่วิธีการนี้จำเป็นต้องมีการสร้างเป็นโปรแกรมเฉพาะทางและต้องเป็นผู้มีความรู้เฉพาะด้านในการใช้งานจึงอาจทำให้ผู้ใช้งานทั่วไปไม่สามารถใช้ได้ แต่การศึกษานี้แสดงให้เห็นว่าการประยุกต์ใช้ปัญญาประดิษฐ์เชิงสร้างสรรค์สามารถช่วยให้ผู้มีความรู้พื้นฐานเกี่ยวกับการตั้งโจทย์ (Prompt) และ

การวิเคราะห์ข้อมูลสามารถใช้โปรแกรมไพธอนที่มีศักยภาพในการวิเคราะห์และจัดการข้อมูลได้รวดเร็ว และมีประสิทธิภาพเพราะนอกจากการสร้างไพธอนสคริปแล้ว ปัญญาประดิษฐ์เชิงสร้างสรรค์สามารถเป็นผู้ให้คำแนะนำในด้านการวิเคราะห์ข้อมูล ตลอดจนการแก้ปัญหาได้รวดเร็ว แต่ทั้งนี้ผู้ใช้ควรมีการศึกษาด้านการตั้งโจทย์ และความรู้ด้านการวิเคราะห์ข้อมูลควบคู่กันเนื่องจากสิ่งที่ปัญญาประดิษฐ์ให้คำตอบมานั้นอาจไม่เป็นจริงเสมอไป (Hallucination) เนื่องจากหลักการทำงานของปัญญาประดิษฐ์เชิงสร้างสรรค์จะใช้ความน่าจะเป็นจากข้อมูลที่ผ่านการเรียนรู้มาเพื่อตอบคำถาม

7. สรุปผลการวิจัย

จากจุดประสงค์ของการศึกษานี้สามารถสรุปการศึกษาได้ดังนี้

7.1 สำรวจความสามารถของปัญญาประดิษฐ์เชิงสร้างสรรค์ในการทำความสะอาดข้อมูล พบว่าความสามารถของปัญญาประดิษฐ์เชิงสร้างสรรค์ทั้ง 3 ระบบมีความสามารถในการให้ข้อมูล และหลักการที่จำเป็นสำหรับการทำความสะอาดข้อมูลและจัดเตรียมข้อมูลได้ถูกต้อง (ดังรายละเอียดในหัวข้อ 5.1 การสร้างคำถาม และผลลัพธ์ที่เกี่ยวข้องกับการทำความสะอาดข้อมูลและจัดเตรียมข้อมูล) ผ่านการใช้คำสั่งแบบสนทนาโต้ตอบ (Conversational prompting) ที่มีลักษณะการพูดคุยแบบสนทนาเพื่อให้ปัญญาประดิษฐ์สามารถเรียนรู้และเข้าใจบริบทของสถานการณ์ได้ลึกซึ้งขึ้น นอกจากนี้ยังสามารถช่วยเขียนไพธอนสคริป (Python scrip) สำหรับการทำความสะอาดข้อมูล และจัดเตรียมข้อมูลได้ (ดังรายละเอียดในหัวข้อ 5.2 การจัดการค่าที่ขาดหายไป (Missing values) และ 5.3 การจัดการค่าสุดโต่ง (Outlier)) นอกจากนี้หากมีความผิดพลาดในการประมวลผลไพธอนสคริป สามารถนำข้อความที่แสดงมาใช้เป็นคำสั่งควบคู่กับเพิ่มการให้ข้อมูลส่วนที่จำเป็นเพื่อให้ปัญญาประดิษฐ์ทำการแก้ไขได้อย่างถูกต้อง และให้ผลลัพธ์ตามที่ต้องการ

7.2 ประเมินประสิทธิภาพของ Generative AI 3 ระบบได้แก่ Claude.ai ChatGPT 3.5 และ Gemini ในการทำความสะอาดข้อมูล และเตรียมข้อมูลในการศึกษานี้สามารถพิจารณาโดยดูจากผลสัมฤทธิ์ของการดำเนินงานแต่ละขั้นตอนดังตารางที่ 5 พบว่ามีเพียงขั้นตอนการดำเดินการบนแพลตฟอร์มที่ไม่สามารถทำได้โดยรายละเอียดแสดงในคำถามที่ 4 สามารถทำ Data cleaning บนระบบปัญญาประดิษฐ์เชิงสร้างสรรค์ได้หรือไม่ นอกเหนือจากนั้นปัญญาประดิษฐ์เชิงสร้างสรรค์ทั้ง 3 ระบบสามารถดำเนินการได้ตามวัตถุประสงค์และมีความถูกต้องทั้งข้อมูลที่ให้ ตลอดจนผลลัพธ์จากการทำความสะอาดข้อมูล

ตารางที่ 5 เปรียบเทียบประสิทธิภาพ Generative AI 3 ระบบ ในงานวิจัยนี้

หัวข้อ	Claude.ai	ChatGPT 3.5	Gemini	หมายเหตุ
อธิบายหลักการทำความสะอาดข้อมูลและการเตรียมข้อมูล	สามารถให้ข้อมูลได้ถูกต้อง	สามารถให้ข้อมูลได้ถูกต้อง	สามารถให้ข้อมูลได้ถูกต้อง	
อธิบายวิธีการทำความสะอาดข้อมูล (ค่าที่ขาดหาย และค่าสุดโต่ง)	สามารถให้ข้อมูลได้ถูกต้อง	สามารถให้ข้อมูลได้ถูกต้อง	สามารถให้ข้อมูลได้ถูกต้อง	
อธิบายการเตรียมข้อมูลหลังจากการทำความสะอาด	สามารถให้ข้อมูลได้ถูกต้อง	สามารถให้ข้อมูลได้ถูกต้อง	สามารถให้ข้อมูลได้ถูกต้อง	
ความสามารถในการดำเนินการบนแพลตฟอร์ม	สามารถทำได้ แต่ผลลัพธ์ไม่ถูกต้อง	สามารถทำได้ แต่ผลลัพธ์ไม่ถูกต้อง	สามารถทำได้ แต่ผลลัพธ์ไม่ถูกต้อง	
การจัดการค่าที่ขาดหายด้วย python บน google colab	สามารถทำได้ ผลลัพธ์ถูกต้อง	สามารถทำได้ ผลลัพธ์ถูกต้อง	สามารถทำได้ ผลลัพธ์ถูกต้อง	
การจัดการค่าที่สุดโต่งด้วย python บน google colab	สามารถทำได้ ผลลัพธ์ถูกต้อง	สามารถทำได้ ผลลัพธ์ถูกต้อง	สามารถทำได้ ผลลัพธ์ถูกต้อง	
การบันทึกข้อมูล และการทดสอบการใช้ข้อมูล	สามารถทำได้ ผลลัพธ์ถูกต้อง	สามารถทำได้ ผลลัพธ์ถูกต้อง	สามารถทำได้ ผลลัพธ์ถูกต้อง	

ทั้งนี้การให้ข้อมูลในเชิงปฏิบัติยังนั้นยังมีความผิดพลาดอยู่บ้าง เช่น ในการเขียนสคริปของภาษาไพธอนในกระบวนการทำความสะอาดข้อมูลตามที่กำหนดแล้วต้องมีการแสดงผลเป็นรายงานด้วย แต่หลายๆ ครั้ง ปัญญาประดิษฐ์เชิงสร้างสรรค์ไม่ได้เขียนสคริปในการแสดงผลตามที่ได้กำหนดในคำสั่งจึงต้องมีการให้คำสั่งเพิ่มเติมเพื่อการนำเสนอผลการดำเนินการด้วย เป็นต้น สำหรับ Claude.ai และ ChatGPT 3.5 ผู้วิจัยพบว่าทั้ง 2 ระบบมีความสามารถในการแปลความต้องการและข้อมูลที่ให้ไปเพื่อการสร้างสคริปภาษาไพธอนในการทำความสะอาดข้อมูลและจัดเตรียมชุดข้อมูล สามารถแก้ไขความผิดพลาด (Error) ของสคริปได้จากการให้ข้อความแจ้งเตือนความผิดพลาดที่เกิดขึ้น แต่สำหรับ Gemini พบว่าต้องมีการปรับเปลี่ยนการให้คำสั่งพร้อมอธิบายเพิ่มเติมมากกว่า Claude.ai และ ChatGPT 3.5 และการแก้ไขความผิดพลาดของสคริปจะเป็นลักษณะของการให้ข้อมูลเพิ่มเติมมากกว่าการแก้ไขสคริป อย่างไรก็ตามยังมีผลการศึกษาไม่เพียงพอที่จะสามารถบอกได้อย่างชัดเจนว่าระบบใดมีความสามารถสูงที่สุด แต่ผลการดำเนินการนั้นบรรลุตามวัตถุประสงค์ และผู้วิจัยเห็นว่าปัญญาประดิษฐ์เชิงสร้างสรรค์นั้นสามารถช่วยให้ผู้ที่ต้องการเตรียมชุดข้อมูลให้มีคุณภาพเพื่อการวิเคราะห์ข้อมูล หรือการนำไปวิเคราะห์ทางสถิติ

7.3 เสนอแนวทางการนำ Generative AI มาประยุกต์ใช้ในการทำความสะอาดข้อมูลอย่างมีประสิทธิภาพจากผลการวิจัยนี้มีดังนี้

7.3.1 ผู้ใช้ควรมีการศึกษาเรื่องการสร้างคำถามหรือคำสั่ง (Prompt engineer) เพื่อให้สามารถดึงศักยภาพของปัญญาประดิษฐ์เชิงสร้างสรรค์มาใช้ในการให้ข้อมูล ตลอดจนวิธีการทำความสะอาดข้อมูล และการจัดเตรียมข้อมูลได้เหมาะสมกับประเภทของชุดข้อมูลที่ต้องการวิเคราะห์ ทั้งนี้ข้อจำกัดของการใช้ปัญญาประดิษฐ์เชิงสร้างสรรค์นั้นมีความไม่แน่นอนในการแสดงผลลัพธ์ แม้จะใช้คำสั่งเดียวกันก็ตาม

7.3.2 ผู้ใช้มีความจำเป็นต้องศึกษาเรื่องการวิเคราะห์ข้อมูล หรือวิทยาการข้อมูลเบื้องต้นเพื่อให้สามารถสร้างคำสั่งให้ปัญญาประดิษฐ์เชิงสร้างสรรค์ทำการสร้าง Python scrip ที่นำไปใช้ทำความสะอาดข้อมูลได้อย่างเหมาะสม

8. เอกสารอ้างอิง

1. อนุชิต คงฤทธิ์, สงวน วงษ์ชวลิตกุล และมารุต โคตรพันธ์. (2565). การศึกษาสมบัติเบื้องต้นของวัสดุเชิงประกอบจากพอลิโพรพิลีนเกรดรีไซเคิลผสมผงกากชา. *วารสารวิศวกรรมและเทคโนโลยี มหาวิทยาลัยรังสิต*, *25*(2), หน้า 43-56.
2. Chaki, R., Chaki, N., Cortesi, A., & Saeed, K. (2024). Applied computing for software and smart systems. *Proceedings of ACSS 2023* (p. 71). Springer Nature.
3. Dasari, D., & Varma, P. (2022). Employing various data cleaning techniques to achieve better data quality using Python. *2022 6th International Conference on Electronics, Communication and Aerospace Technology*. IEEE. htpst://doi.org/10.1109/iceca55336.2022.10009079
4. Du, S., Shi, W., Li, S., & Zhao, G. (2022). Research on data cleaning technology based on RD-CFD method. *7th International Symposium on Advances in Electrical, Electronics, and Computer Engineering*. IEEE. https://doi.org/10.1117/12.2639864
Finding outliers in a scatter plot / pandas dataframe? (n.d.). Stack Overflow. https://stackoverflow.com/questions/75993462/finding-outliers-in-a-scatter-plot-pandas-dataframe
5. Frazer, G. (2024). *Data Cleaning with Power BI: The definitive guide to transforming dirty data into actionable insights*. Packt Publishing Ltd.
6. Gill, J. K. (2023, November 24). *Generative AI for data analytics And management*. XENONSTACK. https://www.xenonstack.com/blog/generative-ai-data-analytics

7. Goyle, K., Xie, Q., & Goyle, V. (2023). DataAssist: A machine learning approach to data cleaning and preparation. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2307.07119

8. Jafarov, E. J. E. (2022). Technologies of data processing and cleaning, noise identification and removal at time series. *Proceedings of Azerbaijan High Technical Educational Institution*, *23*(12), 73–82. https://doi.org/10.36962/pahtei23122022-73

9. Jin, Z. (2022). Principle, methodology and application for data cleaning techniques, *BCP Business & Management*, *26*, 724–732. https://doi.org/10.54691/bcpbm.v26i.2032

10. Krishnan, S., & Wu, E. (2019). AlphaClean: Automatic generation of data cleaning pipelines. arXiv Preprint *arXiv:1904.11827*. https://arxiv.org/abs/1904.11827

11. Lee, J., Kim, D., Hong, H., & Kuk, J. (2020). Apparatus and method of generating map data of cleaning space. *(Patent)*. https://patents.google.com/patent/US20200218274A1/en

12. Loureiro, A., Torgo, L., & Soares, C. (2004). Outlier detection using clustering methods: A data cleaning application. *In Proceedings of KDNet Symposium on Knowledge-based Systems for the Public Sector*. Bonn, Germany.

13. Martinez-Luengo, M., Shafiee, M., & Kolios, A. (2019). Data management for structural integrity assessment of offshore wind turbine support structures: Data Cleaning and missing data imputation, *Ocean Engineering*, *173*, 867–883. https://doi.org/10.1016/j.oceaneng.2019.01.003

14. Mertz, D. (2021). Cleaning data for effective data science: Doing the other 80% of the work with Python. *R, and command-line tools*. Packt Publishing Ltd.

15. Nandi, G., & Sharma, R. K. (2020). *Data Science Fundamentals and Practical Approaches: Understand Why Data Science Is the Next*. BPB Publications. *pandas DataFrame: replace nan values with average of columns*. (n.d.). Stack Overflow. https://stackoverflow.com/questions/18689823/pandas-dataframe-replace-nan-values-with-average-of-columns

16. Parulian, N. N., & Ludäscher, B. (2022). DCM explorer: A tool to support transparent data cleaning through provenance exploration. In Proceedings of the *14th International Workshop on the Theory and Practice of Provenance (TaPP '22)* (Article 10, pp. 1–6). Association for Computing Machinery. https://doi.org/10.1145/3530800.3534539

17. Rangapur, A., & Rangapur, A. (2024). The Battle of LLMs: A Comparative Study in Conversational QA Tasks. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2405.18344

18. Sergiienko, B. (2024, April 16). How to Use Generative AI in Data Analytics for Enhanced Decision-Making and Strategic Growth. *Master of Code Global.* https://masterofcode.com/blog/generative-ai-for-data-analytics

19. Setiyanto, S., & Setiawan, I. (2022). Data science with Excel. *International Journal of Computer and Information System*, *3*(3), 104–110. https://doi.org/10.29040/ijcis.v3i3.79

20. Shigarov, A. O., & Mikhailov, A. A. (2017). Rule-based spreadsheet data transformation from arbitrary to relational tables. *Information Systems*, *71*, 123–136. https://doi.org/10.1016/j.is.2017.08.004

21. Sunne, S. (2022). Cleaning data. *In Routledge eBooks* (pp. 71–85). https://doi.org/10.4324/9781003273301-5

22. Torkey, H., Ibrahim, E., Hemdan, E. E., El-Sayed, A., & Shouman, M. A. (2021). Diabetes classification application with efficient missing and outliers data handling algorithms. *Complex & Intelligent Systems*, *8*(1), 237–253. https://doi.org/10.1007/s40747-021-00349-2

23. Von Zernichow, B. M., & Roman, D. (2017). Usability of visual data profiling in data cleaning and transformation. *In Lecture Notes in Computer Science* (pp. 480–496). https://doi.org/10.1007/978-3-319-69459-7_32

24. Walker, M. (2020). *Python Data Cleaning Cookbook: Modern Techniques and Python Tools to Detect and Remove Dirty Data and Extract Key Insights.* Packt Publishing Ltd.

25. Ye, A., & Wang, Z. (2022). Data preparation and engineering. In *Apress eBooks* (pp. 95–179). https://doi.org/10.1007/978-1-4842-8692-0_2

26. Zaamout, K. (2024, April 30). *AI-Driven Operational Efficiency and Data Management*. Caylent. https://caylent.com/blog/ai-driven-operational-efficiency-and-data-management

27. Zou, F. (2022). Research on data cleaning in big data environment. 2022 International Conference on Cloud Computing, *Big Data and Internet of Things (3CBIT)*, 145–148. https://doi.org/10.1109/3CBIT57391.2022.00037