Data-to-Dashboard: Multi-Agent LLM Framework for Insightful Visualization in Enterprise Analytics

Ran Zhang
Boston University
Boston, MA, USA
ran0925@bu.edu

Mohannad Elhamod
Boston University
Boston, MA, USA
elhamod@bu.edu

arXiv:2505.23695v1 [cs.AI] 29 May 2025

Abstract

The rapid advancement of LLMs has led to the creation of diverse agentic systems in data analysis, utilizing LLMs' capabilities to improve insight generation and visualization. In this paper, we present an agentic system that automates the data-to-dashboard pipeline through modular LLM agents(See Fig. 1) capable of domain detection, concept extraction, multi-perspective analysis generation, and iterative self-reflection. Unlike existing chart QA systems[19, 21, 32, 34], our framework simulates the analytical reasoning process of business analysts by retrieving domain-relevant knowledge and adapting to diverse datasets without relying on closed ontologies or question templates. We evaluate our system on three datasets across different domains. Benchmarked against GPT-4o with a single-prompt baseline, our approach shows improved insightfulness, domain relevance, and analytical depth, as measured by tailored G-Eval metrics[12] and qualitative human assessment.

This work contributes a novel modular pipeline to bridge the path from raw data to visualization, and opens new opportunities for human-in-the-loop validation by domain experts in business analytics. All code can be found here: https://github.com/77luvC/D2D_Data2Dashboard

CCS Concepts

• **Social and professional topics** → **Computing and business**.

Keywords

AI, Data Analytics, Business Insights

ACM Reference Format:
Ran Zhang and Mohannad Elhamod. 2025. Data-to-Dashboard: Multi-Agent LLM Framework for Insightful Visualization in Enterprise Analytics. In *Proceedings of 2nd Workshop on Agentic AI for Enterprise (ACM SIGKDD).* ACM, New York, NY, USA, 7 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

1.1 Contextualization

In enterprise settings, data analytics plays a critical role in generating insights across multiple levels[1], such as: 1. data observations[22, 23] 2. factual data insights (highlighting patterns like trends, distributions, and outliers)[3] 3. hypotheses/question-based insights[9], and 4. domain-knowledge insights, defined as links connecting statistical and/or visual analysis findings with user knowledge[11, 25]. Existing research has explored the first three types of insights, leveraging large language models (LLMs) for data analysis and large vision-language models (LVLMs) for visualization interpretation. These technologies have notably enhanced analyst productivity by automating routine analytical tasks. However, little work has been published on domain-related insight generation by LLMs and LVLMs.

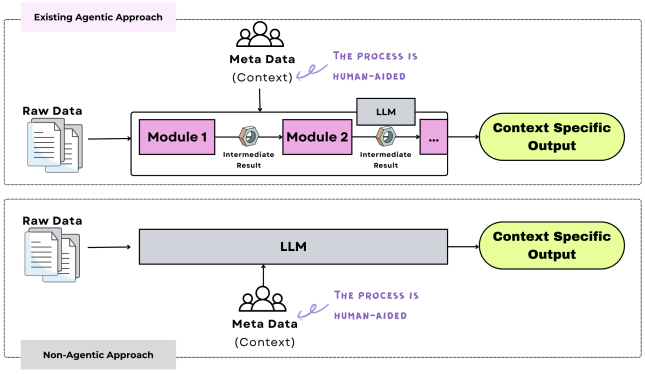

Figure 1: Existing approaches, whether agentic or non-agentic, use language models to obtain context-specific answers and insights, often overlooking the deeper value still embedded in the underlying raw data.

Domain-knowledge hits multiple notes. Stemming from psycholinguistics, it not only supplies factual context, but also shapes how information is structured and transformed into insights[16]. In enterprise settings, domain-knowledge is defined as the company's operational sphere, shaping the critical questions[8]. The assumption here is that domain-knowledge might allow language models to distinguish signals from noise, prioritize metrics synced with business targets, and craft narratives that fit users' mental frameworks[5, 6].

1.2 Motivation

Current studies on data/business analysis using agentic systems with LLMs or LVLMs mainly zero in on factual data identification and hypothesis/question-driven tasks (QA-pairs)[19, 21, 32, 34]. Their methodologies emphasize low-level[28, 33] or high-level analysis[19, 21] aimed at spotting facts or patterns in datasets. This narrow focus leaves a possible gap:

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
ACM SIGKDD, Toronto, ON, Canada
© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

The limited exploration of how LLMs generate broader insights driven by domain-knowledge through agents.

Following this line of investigation, three key questions could be addressed: 1. Will explicit domain identification enhance insights? 2. How do domain-driven insights generated from LLMs differ from those generated by QA tasks? and 3. Can domain-related insights add more meanings to data visualization?

Another challenge we find is how current typical workflows often separate *data-to-chart* [7, 26, 30] and *chart-interpretation* [2, 4, 29] into distinct areas (For further information, refer to Section 2). This separation introduces risks, such as the potential disconnection between chart-derived insights and the original data. If charts are inaccurately generated, any insights based on them are inherently flawed, since they rely on erroneous visual interpretations. Additionally, in many business settings, insights may remain hidden within extensive datasets, indicating that merely interpreting chart insights can overlook the broader insights available from the dataset. This drives us to consider the essential need for developing a system that seamlessly integrates the separated workflows.

1.3 Problem statement

Building on findings in Sections 1.1 and 1.2, the challenge in data analytics with Large Language Model is:

Developing a robust, generalizable agentic system that uses state-of-the-art LLMs to produce domain-related insights, and enhances data visualization with great significance based on these insights.

1.4 Proposed Solution

We propose an end-to-end agentic system that automates the data analysis workflow with a novel pipeline: raw data - domain identification - insights generation - data visualization, a dashboard with charts (See Fig. 2, for detailed information, refer to Section 3).

Leveraging the capabilities of Large Language Models, our framework consists of specialized agents tasked with: (1) detecting the domain and concepts of the dataset, (2) retrieving and applying domain-relevant knowledge to make insightful analysis, and (3) generating visual insights. Each agent is built with role-specific prompting strategies and is capable of memory-based reflection for iterative improvement. This multi-agent architecture not only grounds the analysis in domain semantics but also enables compositional reasoning, making it adaptable to mixed-domain business datasets.

1.5 Contributions

Below, we list the distinct contributions of our work:

- We introduce a novel end-to-end agent-based framework that transforms raw business tables into insightful dashboards through domain-informed reasoning, opening new opportunities for managers in business intelligence, marketing, and finance to validate and refine domain-specific analytics.
- We integrate domain detection and knowledge retrieval mechanisms to support feature selection and multi-perspective analytical reasoning, thereby enhancing the accuracy and robustness of the captured insights.
- We propose a workflow that simulates the cognitive process of business analysts, enabling iterative improvements and reflective interpretation.

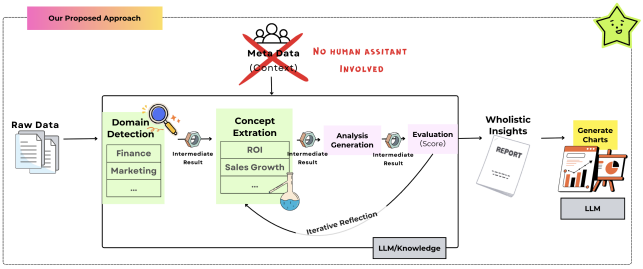

Figure 2: Our end-to-end data-insight-visualization approach provides context-independent domain-aware insights, thus overcoming the limitations of existing systems.

2 Related Work

Prior work on automated data analysis tasks using LLMs and LVLMs (including data insight generation, chart generation and interpretation) generally falls into two categories: (1) **data-to-chart** generation, and (2) **chart-interpretation** summarization. Data-to-chart systems typically target single-table inputs with constrained schema complexity, generating only one or two charts per dataset[32]. Such systems tend to emphasize syntactic chart correctness or alignment with pre-specified templates, rather than insight utility. On the other hand, chart-interpretation systems assume an existing chart and aim to summarize it using LLMs[10, 29]. The resulting insights are usually limited to surface-level features—e.g., maxima, minima, or linear trends—without context-aware reasoning [28].

Even in more advanced settings, such as domain-specific QA over charts (e.g., in finance), the primary focus remains on factual accuracy or verification against known ground-truths[21]. These tasks do not address the cognitive process of deciding *what to visualize* or *what to explore* in the first place. Moreover, very few systems include any notion of domain-knowledge grounding—whether to select relevant metrics, or to contextualize the resulting insights.

Moreover, most existing approaches—whether targeting low-level tasks such as chart factuality check or high-level insight generation—rely heavily on question-answer (QA) pairs[15, 21] to drive the analytical process. While this paradigm is effective in settings where users possess prior domain-knowledge and can articulate precise analytical queries, it inherently limits the generative potential of large language models. By anchoring the analysis to predefined questions, these methods constrain LLMs from autonomously exploring the dataset and surfacing insights that may lie outside the bounds of user expectations.

Recent benchmarks such as *InsightBench*[21] attempt to move toward higher-level insights. However, their approach remains question-driven in structure with enriched context in metadata as input(See Fig. 1). As a result, the opportunity to leverage the

broader reasoning capabilities of LLMs—particularly within agent-based frameworks that simulate open-ended human analytical workflows—remains underexplored.

3 Proposed Approach

Our agentic system operates in two sequential stages: data-to-insight and insight-to-chart (Note that the intermediate "insight" here refers to domain-knowledge insights. This is in contrast to the question-based insights extracted in in the typical workflows discussed in Section 1.2). In the data-to-insight stage, the system orchestrates a suite of specialized agents—collaborate to extract semantic and structural understanding from raw tabular data, identifying the business domain, relevant analytical concepts, and candidate insights. The output of this stage provides structured guidance to the next one. In the insight-to-chart stage, the emphasis is not only on syntactic chart generation but also on the production of insightful visualizations tailored to business reasoning tasks.

3.1 Key Definition

To support flexible reasoning across diverse business tasks, we introduce two foundational abstractions in our system: **domain** and **domain concept**. Rather than imposing a rigid taxonomy, we define these as *relational constructs* that form a hierarchical semantic structure—where a **domain** denotes a broader business context (e.g., finance, operations, incident management), and a **domain concept** refers to more granular elements within that context (e.g., revenue, downtime, churn rate).

This framing intentionally avoids prescriptive definitions or closed ontologies. Inspired by both linguistic theory[16] and the generalization capacity of large language models (LLMs), we posit that strict categorization can constrain discovery. Instead, by prompting the LLM to infer these hierarchical relationships on a case-by-case basis, our agentic system encourages open-ended exploration while maintaining enough structure to guide downstream analytical and visualization tasks.

3.2 Stage 1: Data-to-Insight

In the **data-to-insight** stage, the raw dataset serves as the input. Unlike *BenchInsight*, our approach requires zero supplementary context. The output is: a domain name, domain-specific concepts, and analytical insights. This process involves a series of module agents that perform data profiling, detection of domains and concepts, generation of multi-perspective analyses, evaluation, and iterative self-reflection.

Data Profiler When analysts face large, unfamiliar datasets, a typical approach is to formulate "good questions" that steer a language model toward the desired answers—an idea that underlies much of the question–answer (QA)-driven data-analysis literature [19]. Whether the questions probe high-level trends or low-level facts, the workflow is still anchored in asking before seeing. However, our end-to-end agent takes the opposite tack. We begin with an automated data-profiling stage that constructs a structured, statistical synopsis of the table itself, not just a description of data. Using a tree-of-thought prompting method[35], the agent infers column types, value ranges and units, functional dependencies, and potential keys, thereby reducing the ineffective dimensionality of the raw data before any explicit questions are posed. The resulting profile then serves as a principled scaffold for the subsequent reasoning and insight-generation steps

Domain Detector This module determines the business theme of the dataset by analyzing its structural profile. Rather than applying rigid classification schemes, it generates a flexible domain label accompanied by a concise, one-sentence definition. This label is inferred using external reference knowledge sources, such as Wikipedia, enabling broad transferability across industries without relying on closed taxonomies or predefined ontologies. The generated domain label is unique to each dataset. In future experiments, as more diverse datasets are tested, inconsistencies or imprecisions may emerge. To address this issue, we propose integrating a self-consistency mechanism—as explored in recent work on multi-round verification using LLMs [13]—to enhance stability and balance between generality and accuracy. We do not implement this functionality here as this work is a proof-of-concept.

Concept Extractor Given the domain label and structural metadata, the Concept Extractor agent identifies salient concepts that are likely to drive downstream analysis. These concepts are formulated as natural language phrases (e.g., "monthly active users," "unit cost," "processing delay") that correspond to the domain and data profiling results. The agent ensures that the extracted concepts are not only relevant to the dataset's business theme but also actionable for subsequent analytical tasks performed by the Analysis Generator.

Analysis Generator This agent synthesizes structured insights from the dataset using three lenses: `descriptive` (summarizing distributions or outliers), `predictive` (inferring trends or likely outcomes), and `domain-related`. We pick these lenses because such descriptors offer deeper insight for analysis than non-insight[20] plus domain-knowledge, highlighted in the motivation, is crucial.

The output is a unified JSON object intended to simulate human-like analysis and hypothesis generation. Emphasis is placed on producing novel, non-obvious insights rather than surface-level observations.

Evaluator The Evaluator agent scores the generated outputs across five dimensions: Given that pinpointing the domain and its associated concepts is a crucial cornerstone of our system, the accuracy of domain inference along with the relevance and comprehensiveness of the concepts are key scoring factors. Insightfulness, novelty[14] and depth[31] are vital to enterprise competence and business analytics. These criteria overlap and agree with recent academic benchmarks and are operationalized into a fixed scoring rubric [19]. In addition to numerical scores(ranging between 1 and 4), the Evaluator provides justification for each rating, which gives the next module an enriched textual context to generate better criticism and reflection.

Self-Reflector We adopt the Reflexion framework [24] to enhance reasoning after the evaluation stage. The Self-Reflector receives a composite signal that includes the evaluation scores, contextualized feedback, and memory from prior iterations. The loop runs for up to n iterations or terminates early once all evaluation scores meet a predefined threshold. Notably, we intentionally set a high bar (4 out of 4 across all criteria) to force the LLM to fully utilize its reasoning capabilities, improving insight quality and analytical depth over time.

3.3 Stage 2: Insight to Chart

Our approach implements a Tree-of-Thought (ToT) reasoning framework [35] for transforming analytical insights into domain-appropriate visualizations. The ToT methodology was selected specifically to address the complex decision-making inherent in chart element selection, as it enables structured, multi-step reasoning that simulates deliberative expertise. Unlike single-pass approaches that may prematurely commit to visualization choices, ToT facilitates explicit consideration of alternatives and their domain implications. Through the three-expert consensus mechanism, the system can evaluate competing visualization strategies against domain requirements, debate the effectiveness of different chart types, and scrutinize the selection of visual encodings before committing to a final representation.

This deliberative process culminates in a consensus decision specifying the optimal visualization type, rationale, key insight narrative, and recommended annotations that emphasize domain significance. As a result, the system preserves domain-insight context throughout the visualization pipeline, ensuring that the generated charts serve not merely as data representations but as vehicles for communicating substantive domain knowledge.

It is worth noting that the quality of generated insights was found to inversely impact chart generation accuracy. Particularly, when Stage 1 domain insights are effectively transferred to Stage 2, generating accurate charts, especially legends, becomes more challenging.

4 Evaluation Criteria

To assess our agentic system's ability to generate insightful visualizations from raw business tables, we adopt a multi-faceted evaluation framework.

4.1 Textual Insight Evaluation

G-Eval We evaluate generated insight outputs using *G-Eval*[12], adapting their scoring mechanism to better reflect enterprise needs. Specifically, we tailor the prompts to account for *business reasoning insightfulness* and *alignment with domain concepts*, rather than general QA accuracy. To make the evaluation fair, the outputs are scored across (1) *insightfulness*, (2) *novelty*, and (3) *depth*. [19]

Utilizing InsightBench as a Guide By employing *InsightBench*'s dataset [21], we can assess whether our system successfully determines the proper analytical direction by comparing it with the dataset's ground truth insights. We opted not to adopt *InsightBench* evaluation metric because the workflows of the two systems differ significantly, as illustrated in Fig. 2. While the only input our system takes is the raw data, Sahu et. al.'s [21] approach handholds the data-to-insights process by incorporating more comprehensive context as input. As such, we can only partially leverage their benchmark in our evaluation.

Human Expert Evaluation Our framework provides *domain experts and business analysts* the chance to utilize a unified evaluation rubric for human-in-the-loop validation, thereby aligning the evaluation process with practical implementation. In future work, as our work evolves to the next stage, we aim to include professionals from both industry and academia to carry out statistically significant evaluations.

4.2 Chart Evaluation

Targeting some Kaggle datasets, numerous data analysts have created and shared their visualization projects. We select a subset of highly downloaded datasets and their corresponding visualizations and analyses, and juxtapose them with our own by: 1. reviewing chart characteristics (type, legend, and axis labels), and 2. evaluating the quality of the insights they contain using G-Eval.

5 Experiments

Experiment 1: Data-to-Insight This experiment focuses on exploring three essential research questions: 1. Does explicit domain identification matter? This examines the impact of domain labeling on the relevance and depth of the generated visualization (dataset 1[27]). 2. How does our approach compare to a prompt-only baseline? We assess the performance of our agentic pipeline against a baseline that employs the same model, *GPT-4o*, driven solely by a simple prompt, by evaluating differences in analytical insight, novelty, and domain relevance (dataset 1[27]). 3. How does generating insights through questions vary from deriving insights through domain knowledge? We evaluate our approach againsy the latest question-focused study[21], using one of their datasets.

Experiment 2: Insight-to-Visualization Using a popular finance survey dataset with download number 19.9K[17], we qualitatively examine how our system improves chart generation, analyzing changes in chart type distribution and depth of insights

Dataset Selection For broader coverage, we've chosen datasets encompassing diverse domains, data types, and analytical complexities. These include a comprehensive data source from a leading business school's marketing simulation, InsightBench, a classic Kaggle dataset. In future work, a real-world industrial dataset will be examined using our approach.

6 Results

Result 1: Explicit domain identification matters

As shown in Fig. 3, it is evident that adding a simple domain detection instruction to the prompt,substantially improves the generated insights's coverage, structure, and business relevance. Prompting the model to name the domain implicitly triggers a knowledge frame (customer-centric marketing). That frame guides the analysis toward metrics that marketers actually track (LTV, churn cohorts, channel mix), resulting in both deeper topic granularity and hierarchical structuring. Furthermore, we notice that while customer churn detection does not hinge on adding domain detection, only the domain-aware run links customer churn to strategy by citing "potential issues with retention programmes". Finally, notice that while the baseline run returns broad categories, such as"sectorfocu" and "regionalocus", that are not materially supported by the input raw data, suggesting that the model may be "fishing" for general business axes when no domain context is supplied. This supports our broader hypothesis: *domain grounding is a lightweight yet powerful prompt-engineering lever for any end-to-end insight-generation pipeline.*

Result 2: Our agentic system works better compared to the baseline.

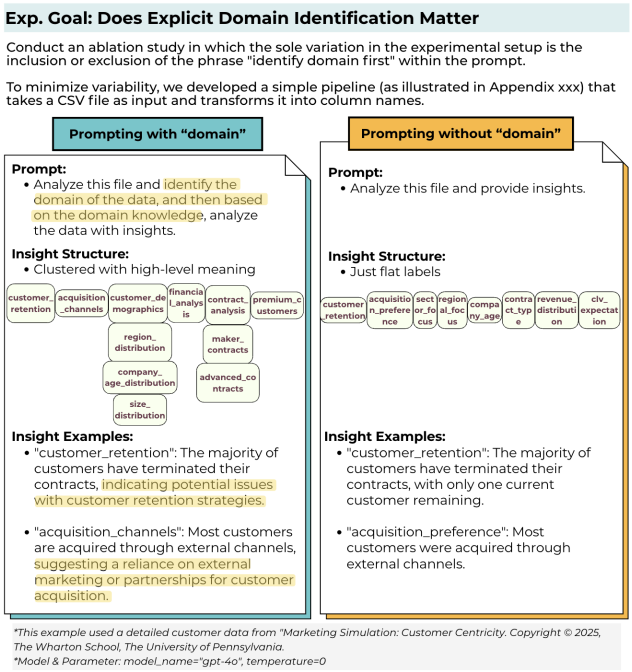

Figure 3: This figure compares the insights obtained with and without domain identification. As can be seen, domain identification grounds the resulting insights, fending it from hallucinations and providing the business analyst with more confidence.

Table 1: Evaluation of our proposed system against non-agentic GPT-4o baseline

G-Eval Metric	GPT-4o	Ours	Relative Lift
Insightful	0.78	0.88	+12 %
Novelty	0.65	0.83	+28 %
Depth	0.75	0.99	+31 %

As shown in Table 1, our approach significantly outperforms a non-agentic GPT-4o baseline with domain awareness in terms of insightfulness, novelty, and depth,

• Insightfulness: Our approach's superior ability to identify nuanced insights, such as the intricate effects of acquisition channels and premium status on customer retention and lifetime value (CLV).

• Novelty: While both our approach and the baseline struggled somewhat in novelty, our system notably surpassed the baseline (0.599 vs. 0.390), successfully generating insights that extend beyond conventional thinking. While these insights may partially correspond to current CRM expertise, since no matter historical or individual novelty are not easy to achieve.

• Depth: Our system outperformed the baseline in terms of the depth of the generated insights (0.942 vs. 0.803), demonstrating that our output exhibits a deeper understanding of CRM complexities, effectively capturing subtle relationships and insights that are implicit in the raw data.

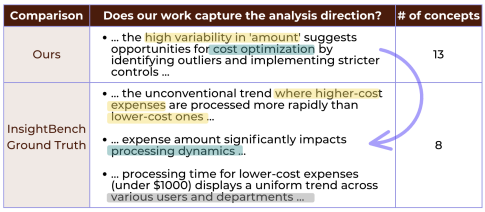

Figure 4: Comparison of our generated insights with InsightBench ground truth. Our system captures the core analytical direction, identifying key themes such as cost variability, processing dynamics, and optimization opportunities, with broader concept coverage. However, it misses one specific ground truth angle—variation across users and departments—highlighting an area for improvement in capturing organizational structure–related insights

Result 3: Insights generated with domain-knowledge can capture the right direction of analysis,

The insights generated with our system successfully aligns with that of *InsightBench*'s ground-truth , capturing the intended direction of analysis by identifying a key pattern:*"The high variability in 'amount' suggests opportunities for cost optimization by identifying outliers and implementing stricter controls."*. Specifically, note that our output contains more diverse and forward-looking concepts, especially concerning automation, compliance, and AI integration, while *InsightBench*'s is tightly focused on operational behavior and expense bracket dynamics.

Moreover, our system highlights several possible analysis directions based on the generated domain-relevant insights, which could be beneficial for broader exploration. As shown in Fig. 4. To maintain academic integrity, and due to the need for domain-expert evaluation, we cannot in this work ascertain the other directions' validity. This will be conduct in future revisions.

Overall, our result suggests an important intuition: domain-knowledge is essential for generating an insightful analysis—a hypothesis that could be further tested through causal experiments in future work.

Table 2: A quantitative evaluation of our proposed Data-to-Dashboard method against a Kaggle user[18]. Notice how our proposed method outperforms the baseline in every single metric.

G-Eval Metric	A Kaggle User	Ours	Relative Lift
Insightful	0.32	0.80	+147%
Novelty	0.34	0.61	+77%
Depth	0.36	0.77	+113%

Result 4: An observation for Human Analyst and Our System The study's goal is to assess the insightfulness of chart creation by juxtaposing our analysis with that of the Kaggle user who received the most "Upvotes" on a personal finance survey dataset[17]. First, we employed G-Eval to assess the insights described in the Kaggle user's work [18], as well as the caption in its generated

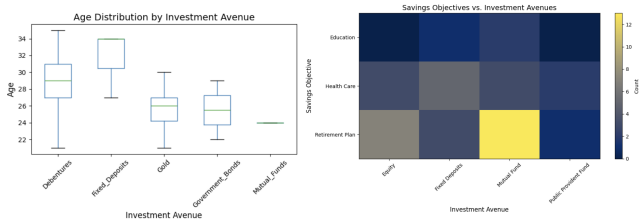

Figure 5: Examples of insightful figures generated by our approach

chart . To evaluate the results in their proper context, we modified the evaluation prompt's criterion from "deep domain expertise" to "Does the analysis demonstrate deep insights?" As shown in Table 2, our system significant outperform on this analysis task.

Secondly, we identify the chart types produced by Kaggle users, which includes 8 bar charts and 1 box plot. In our own work, we created 1 stacked bar chart, 1 heatmap, 1 pie chart, 1 scatter plot, and 1 box plot. Despite the incomplete development of our stage 2 insight-to-chart agentic system due to time constraints, resulting in 2 charts being plotted incorrectly due to coding errors, our work still demonstrates highly insightful chart attributes, shown in Fig. 5.

7 Acknowledgments

We would like to thank Prof. Samuel Engel at Boston University, for generously providing several valuable datasets that supported this project. We would also like to thank Yuan Gao, for the early-stage discussions and paper recommendations that helped shape the initial direction of this project.

References

[1] Leilani Battle and Alvitta Ottley. 2023. What exactly is an insight? a literature review. *2023 IEEE Visualization and Visual Analytics (VIS)* (2023), 91–95.

[2] Alexander Bendeck and John Stasko. 2024. An empirical evaluation of the gpt-4 multimodal language model on visualization literacy tasks. *IEEE Transactions on Visualization and Computer Graphics* (2024).

[3] Eun Kyoung Choe, Bongshin Lee, and m.c. schraefel. 2015. Characterizing Visualization Insights from Quantified Selfers' Personal Data Presentations. *IEEE Computer Graphics and Applications* 35, 4 (2015), 28–37. doi:10.1109/MCG.2015.51

[4] Kiroong Choe, Chaerin Lee, Soohyun Lee, Jiwon Song, Aeri Cho, Nam Wook Kim, and Jinwook Seo. 2024. Enhancing data literacy on-demand: Llms as guides for novices in chart interpretation. *IEEE Transactions on Visualization and Computer Graphics* (2024).

[5] T. Choudhary. 2024. Domain expertise in data analytics: Enhancing insights across industries. *International Journal of Research in Computer Applications and Information Technology (IJRCAIT)* 7, 2 (2024), 69–82.

[6] Nadia Delanoy. 2021. The Importance of Human Domain Knowledge and Business Data Analytics to Support Modern Financial Decisions. *International Journal of Accounting and Finance* 4 (01 2021). doi:10.22158/ijafs.v4n1p1

[7] Victor Dibia. 2023. LIDA: A tool for automatic generation of grammar-agnostic visualizations and infographics using large language models. *arXiv preprint arXiv:2303.02927* (2023).

[8] Eric Evans. 2004. *Domain-driven design: tackling complexity in the heart of software.* Addison-Wesley Professional.

[9] Steven R. Gomez, Hua Guo, Caroline Ziemkiewicz, and David H. Laidlaw. 2014. An insight- and task-based methodology for evaluating spatiotemporal visual analytics. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*. 63–72. doi:10.1109/VAST.2014.7042482

[10] Jiayi Hong, Christian Seto, Arlen Fan, and Ross Maciejewski. 2025. Do LLMs Have Visualization Literacy? An Evaluation on Modified Visualizations to Test Generalization in Data Interpretation. *IEEE Transactions on Visualization and Computer Graphics* (2025).

[11] Eser Kandogan and Ulrich Engelke. 2018. Towards a unified representation of insight in human-in-the-loop analytics: A user study. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics*. 1–7.

[12] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. arXiv:2303.16634 [cs.CL] https://arxiv.org/abs/2303.16634

[13] Zijun Liu, Peiyi Wang, Runxin Xu, Shirong Ma, Chong Ruan, Peng Li, Yang Liu, and Yu Wu. 2025. Inference-time scaling for generalist reward modeling. *arXiv preprint arXiv:2504.02495* (2025).

[14] Michael Lounsbury, Joep Cornelissen, Nina Granqvist, and Stine Grodal. 2021. Culture, innovation and entrepreneurship. In *Culture, Innovation and Entrepreneurship*. Routledge, 1–12.

[15] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244* (2022).

[16] Deborah McCutchen. 1986. Domain knowledge and linguistic knowledge in the development of writing ability. *Journal of memory and language* 25, 4 (1986), 431–444.

[17] Nitindatta. n.d.. *Finance Data.* https://www.kaggle.com/datasets/nitindatta/finance-data Kaggle.

[18] Nitindatta. n.d.. *Finance Data Analysis.* https://www.kaggle.com/code/nitindatta/finance-data-analysis Kaggle Notebook.

[19] Alberto Sánchez Pérez, Alaa Boukhary, Paolo Papotti, Luis Castejón Lozano, and Adam Elwood. 2025. An LLM-Based Approach for Insight Generation in Data Analysis. *arXiv preprint arXiv:2503.11664* (2025).

[20] Md Main Uddin Rony, Fan Du, Ryan Rossi, Jane Hoffswell, Niyati Chhaya, Iftikhar Burhanuddin, and Eunyee Koh. 2023. Augmenting Visualizations with Predictive and Investigative Insights to Facilitate Decision Making. In *Companion Proceedings of the ACM Web Conference 2023*. 77–81.

[21] Gaurav Sahu, Abhay Puri, Juan Rodriguez, Amirhossein Abaskohi, Mohammad Chegini, Alexandre Drouin, Perouz Taslakian, Valentina Zantedeschi, Alexandre Lacoste, David Vazquez, et al. 2024. Insightbench: Evaluating business analytics agents through multi-step insight generation. *arXiv preprint arXiv:2407.06423* (2024).

[22] P. Saraiya, C. North, and K. Duca. 2004. An Evaluation of Microarray Visualization Tools for Biological Insight. In *IEEE Symposium on Information Visualization*. 1–8. doi:10.1109/INFVIS.2004.5

[23] P. Saraiya, C. North, and K. Duca. 2005. An insight-based methodology for evaluating bioinformatics visualizations. *IEEE Transactions on Visualization and Computer Graphics* 11, 4 (2005), 443–456. doi:10.1109/TVCG.2005.53

[24] Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language Agents with Verbal Reinforcement Learning. arXiv:2303.11366 [cs.AI] https://arxiv.org/abs/2303.11366

[25] Arjun Srinivasan, Steven M. Drucker, Alex Endert, and John Stasko. 2019. Augmenting Visualizations with Interactive Data Facts to Facilitate Interpretation and Communication. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2019), 672–681. doi:10.1109/TVCG.2018.2865145

[26] Mara Ströbel, Kai Eckert, and Till Nagel. 2024. Hey ChatGPT, can you visualize my data?–A Multi-Dimensional Study on using an LLM for Constructing Data Visualizations. (2024).

[27] The Wharton School. 2025. *Marketing Simulation: Customer Centricity.* https://interactive.wharton.upenn.edu/academic/customer-centricity-simulation/ © 2025, The Wharton School, The University of Pennsylvania.

[28] Fen Wang, Bomiao Wang, Xueli Shu, Zhen Liu, Zekai Shao, Chao Liu, and Siming Chen. 2025. ChartInsighter: An Approach for Mitigating Hallucination in Time-series Chart Summary Generation with A Benchmark Dataset. *arXiv preprint arXiv:2501.09349* (2025).

[29] Huichen Will Wang, Jane Hoffswell, Victor S Bursztyn, Cindy Xiong Bearfield, et al. 2024. How Aligned are Human Chart Takeaways and LLM Predictions? A Case Study on Bar Charts with Varying Layouts. *IEEE Transactions on Visualization and Computer Graphics* (2024).

[30] Lei Wang, Songheng Zhang, Yun Wang, Ee-Peng Lim, and Yong Wang. 2023. Llm4vis: Explainable visualization recommendation using chatgpt. *arXiv preprint arXiv:2310.07652* (2023).

[31] Norman L Webb. 2002. Depth-of-knowledge levels for four content areas. *Language Arts* 28, March (2002), 1–9.

[32] Yang Wu, Yao Wan, Hongyu Zhang, Yulei Sui, Wucai Wei, Wei Zhao, Guandong Xu, and Hai Jin. 2024. Automated data visualization from natural language via large language models: An exploratory study. *Proceedings of the ACM on Management of Data* 2, 3 (2024), 1–28.

[33] Yifan Wu, Lutao Yan, Leixian Shen, Yunhai Wang, Nan Tang, and Yuyu Luo. 2024. Chartinsights: Evaluating multimodal large language models for low-level chart question answering. *arXiv preprint arXiv:2405.07001* (2024).

[34] Zhengzhuo Xu, Sinan Du, Yiyan Qi, Chengjin Xu, Chun Yuan, and Jian Guo. 2023. Chartbench: A benchmark for complex visual reasoning in charts. *arXiv preprint arXiv:2312.15915* (2023).

[35] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. arXiv:2305.10601 [cs.CL] https://arxiv.org/abs/2305.10601