ChatGPT- 4 Omni: May vs. August 2024 Chart Testing Report

By Lila Karl & Michael Van Demark, AI Researchers - DiligentIQ

Summary

In August 2024, the DiligentIQ research team conducted an in-depth comparison of the May and August versions of the ChatGPT-4 Omni model within our data chat environment, focusing on chart generation and analytical capabilities. Using a Kaggle xlsx dataset by Matthieu Gimbert, we aimed to identify improvements and persisting challenges in data visualization. The August model revealed enhancements in processing times and chart accuracy, particularly with Stacked Bar Charts, Area Charts, and Line Charts, among others. However, some issues persist in the representation of Pie Charts, Venn Diagrams, Radar Charts, and Ridgeline Plots, indicating areas for further development. Minor issues were also noted with Bubble Charts and Scatter Charts. This report offers stakeholders a clear understanding of the advancements and current limitations in our AI-driven tools, emphasizing our commitment to refining financial data visualization to support the Private Equity Industry.

Successful Charts: Stacked Bar Chart, Area Chart, Correlation Chart, Heat Map, Line Chart, Bar Chart, Histogram
Charts With Minor Issues: Bubble Chart, Scatter Chart, Dendrogram, Box Plot
- Small formatting issues

Unsuccessful Charts: Ridgeline Plot, Pie Chart, Venn Diagram, Radar Chart
- Issues with inaccurate data representation and formatting

Motivation

Chart Accuracy in LLMs

Accurate chart generation is important for industries like Private Equity that rely on precise data visualizations. In the private equity field, significant decisions are based on insights from complex datasets, with visual representations serving to enhance the understanding and communication of that data. The speed at which LLMs can perform financial analysis allows businesses to make decisions quickly.

More specifically, private equity firms largely depend on visual representations, along with their underlying data to improve the evaluation of potential investments, monitor portfolio performance, and assess market trends. If visualizations are not accurately aligned with their underlying data, they can distort growth trajectories or mislead risk assessments. Many other industries such as healthcare and aerospace also demand high standards of data integrity and accuracy, proving the importance of precise AI chart generation.

As LLMs are increasingly used for data analysis and visualization, ensuring they generate and interpret charts correctly is becoming more critical. With reliable LLMs, businesses can make informed decisions efficiently, driving faster growth.

Progress in Model Development

The May and August models of the ChatGPT-4 Omni model were rigorously tested, with the August model demonstrating significant improvements in data visualization and processing. The August model showcased faster response times and more accurate chart generation. These enhancements are particularly advantageous for financial industries, where fast and reliable data interpretation is crucial. The notable progress achieved within just three months demonstrates the rapid evolution of AI technologies and their expanding role as supportive tools in financial decision-making.

Methodology

This evaluation utilized a two-round testing process to analyze chart generation and analysis capabilities of both the May and August Models of ChatGPT-4 Omni:

Charts Tested		
Stacked Bar Chart	Bar Chart	Pie Chart
Area Chart	Histogram	Bubble Chart
Correlation Chart	Box Plot	Radar Chart
Heat Map	Dendrogram	Scatter Chart
Line Chart	Venn Diagram	Ridgeline Plot

In our first round of testing, we evaluated visual appeal, data accuracy, and key performance metrics including Time to First Token (TFT), Response Generation Time (RGT), and total response time using successful prompts from previous chart tests. The second round repeated this process to ensure consistency across various chart types.

Additionally, we revisited prompts that had previously not performed well, slightly modifying select prompts to test for potential improvements. As with the previous two rounds of testing, visual appeal, data accuracy, and key performance indicators were once again analyzed. This approach enabled a comprehensive assessment of the models capabilities while also highlighting areas for improvement.

Test Results

Performance Metrics: Key improvements in the August Model

Speed: The August model demonstrated substantial improvements in processing speed and overall efficiency compared to the May model. Specifically, the Time to First Token (TFT) decreased by an average of 2.45 seconds, from 8.28 seconds in May to 6.11 seconds in August. This indicates a quicker initial response to inputs. The Response Generation Time (RGT) saw a significant reduction of 10.61 seconds, dropping from 39.21 seconds to 28.60 seconds on average. As seen in Figure 1, the speed in the August model decreased for each

metric. This substantial decline highlights enhanced efficiency in the model's ability to process and generate responses. Most notably, the Total Time improved by 14.61 seconds, decreasing on average from 49.47 seconds to 34.86 seconds. These enhancements are particularly valuable for time-sensitive analyses in fields like private equity, where prompt decision-making is critical. The increased efficiency allows users to generate insights and visualizations more rapidly, boosting overall productivity. The complete table of the speed data for each chart generated during this test is displayed in the index (see index section 1)

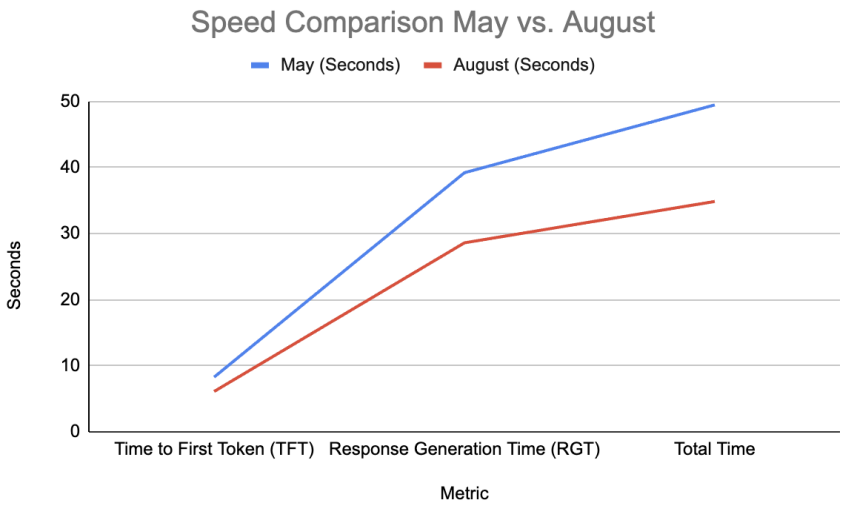

Figure 1. Speed Comparison between May Model and August Model

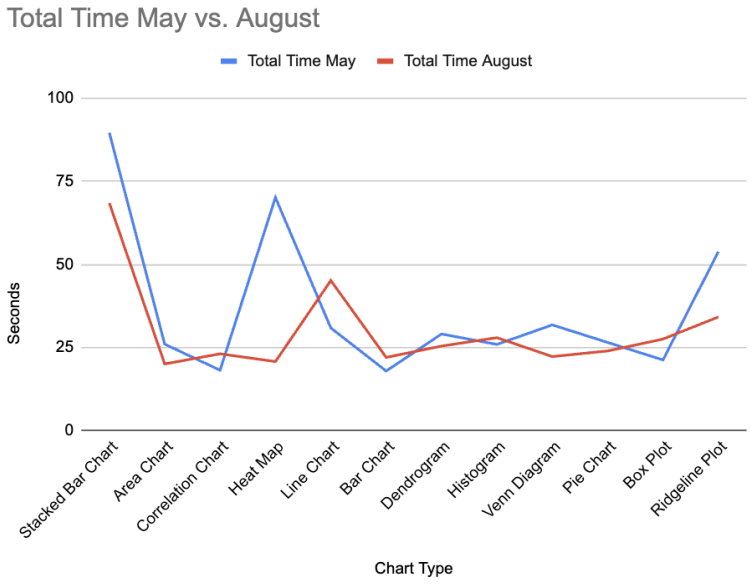

Figure 2. Total Response Time Comparison between May Model and August Model

Successful Charts

The August model demonstrated both improvements and persistent challenges in chart generation accuracy. Several chart types, including the Stacked Bar Chart, Area Chart, Heat Map, Line Chart, Bar Chart, and Histogram, maintained accuracy and clarity in both the May and August models. Figure 3 illustrates an example of a successful chart from both models, showing consistent data representation. The Correlation Charts in Figure 4 shows inconsistencies in the May model but improvement in August with correct and consistent data.

1. **Stacked Bar Chart:** As you can see in Figure 3, the stacked bar charts generated by DiligentIQ in both the May and August models are consistent and accurately represent the dataset.

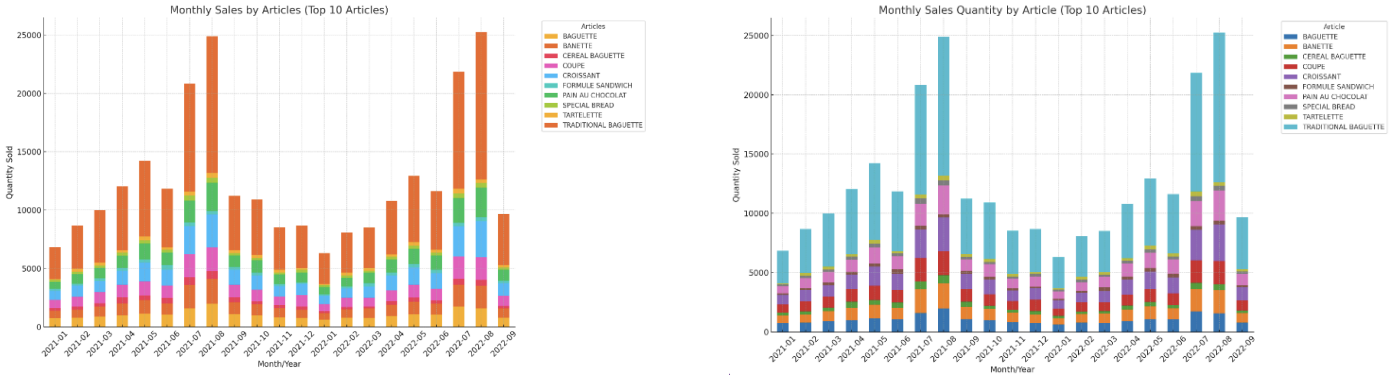

Figure 3. May (left) vs. August (right) Stacked Bar Charts

2. **Correlation Chart:** The May model produced two different Correlation Charts with different data in the two rounds; however, the August model produced consistent Correlation Charts that accurately reflect the data in both rounds. In Figure 4, the improvement can be seen between the May and August models, where the data in the August model is correct and shows enhanced visual clarity. The August model showed improvement in handling Correlation Charts, offering better processing speed and data accuracy. This marks a crucial improvement from the May model, which produced inaccurate and inconsistent charts across rounds.

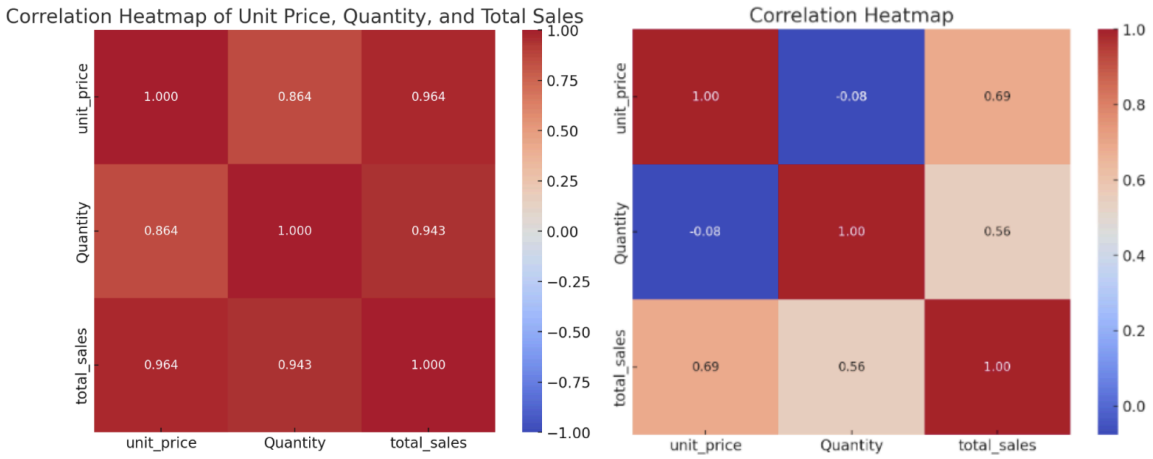

Figure 4. May (left) vs. August (right) Correlation Charts

All charts labeled as successful during testing include those that possessed visual clarity as well as accurate data representation in the August model, with charts from the May model serving as a baseline comparison. The other successful and consistent charts include the Area Chart, Line Chart, Heat Map, Bar Chart, and Histogram (see index section 2).

Charts With Minor Issues

The August model showed some issues in generating Scatter Charts, Bubble Charts, Dendrograms, and Box Plots. It did not follow prompt directions for the desired representation of the data points in the Scatter Chart or Bubble Chart. Additionally, the y-axis titles for the Dendrogram and Box Plot were incorrect. Although the August model generally outperformed the May model, the remaining errors are significant for industries like private equity that rely heavily on precise data for decision-making.These inconsistencies could potentially lead to reduced confidence in data-driven decisions within financial sectors. At DiligentIQ, we are continuously refining our model to deliver reliable charts and data analysis, ensuring that private equity firms can fully rely on our solutions for informed decision-making.

1. **Scatter Chart:** The August model showed progress in generating Scatter Charts, though it still has room for improvement. In the May model, the model only plotted part of the required data given in the prompt directions. In the August model, all of the desired data points were plotted on the chart. However, both models were producing X's as the data points on the chart rather than dots. We used prompt engineering to specify the desired representation of data points on the Scatter Chart by including "Use dots to represent data points on the graph". After refining the prompt, the May model produced dots, but the August model failed to generate a plot. Thus, while the August model initially plotted the correct data points, it was unable to produce the desired format.

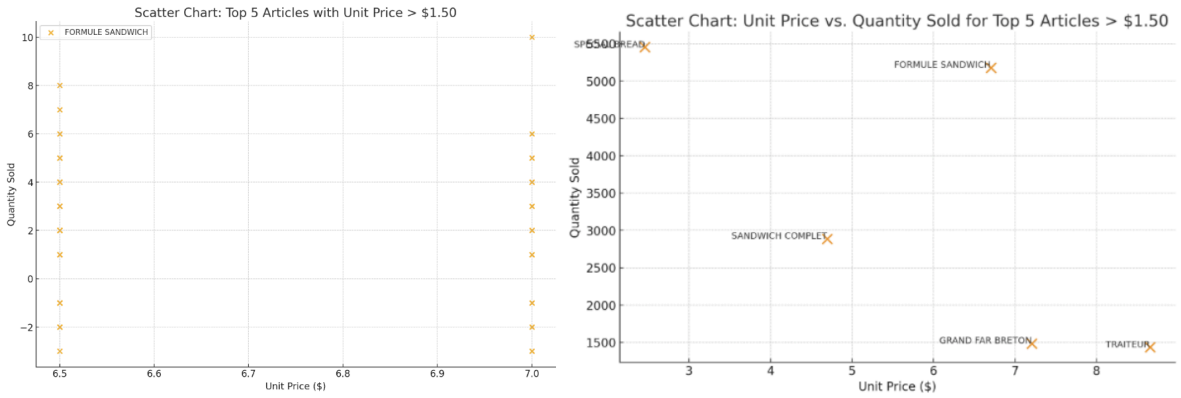

Figure 5. May (left) vs. August (right) Scatter Charts Initial Prompt

2. **Bubble Chart:** Both the May and August models initially faced challenges in properly representing data in bubble charts, using X-shaped points rather than bubbles. We attempted to mitigate this with prompt engineering. We rewrote the prompt to include “circles instead of x's” for the data points; this yielded a successful bubble chart from the May model, while the August model failed to produce the desired format.

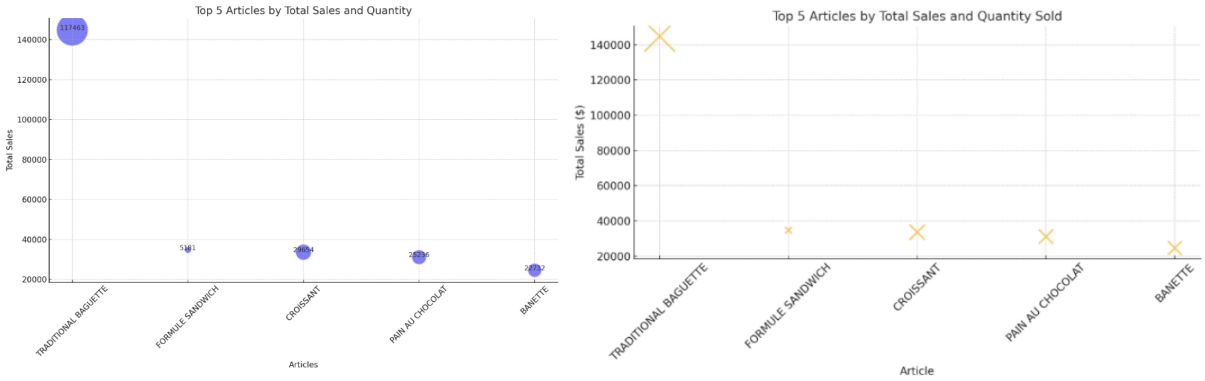

Figure 6. May (left) vs. August (right) Bubble Charts with Refined Prompt

3. **Dendrogram:** Both the May and August models generated the incorrect y-axis title of “Euclidean Distance”, which can be misleading and confusing for a reader trying to interpret this plot. Additionally, the August model produced an inaccurate x-axis titled “Month/Year” as opposed to the correct label: “Article”.

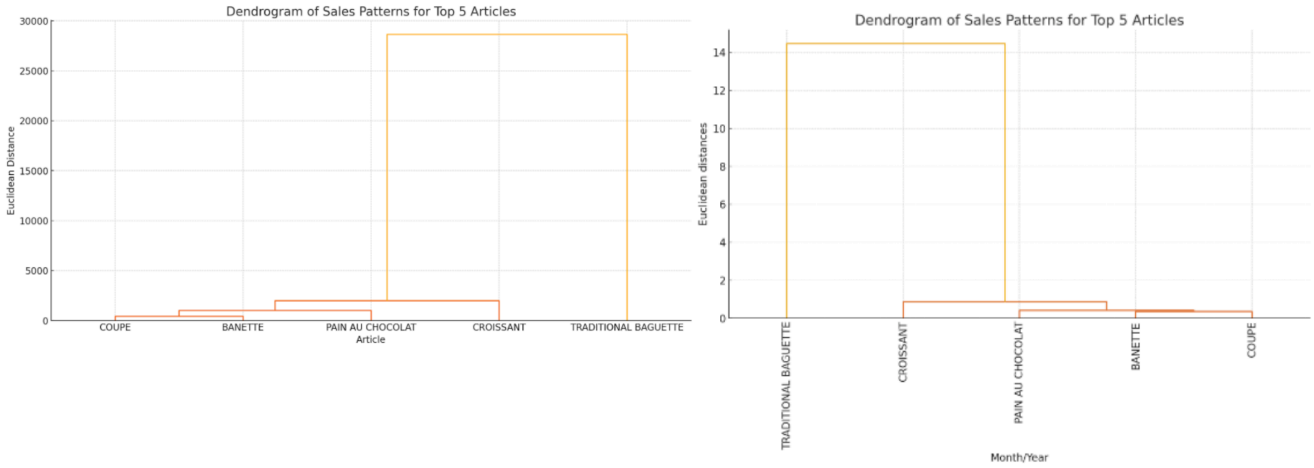

Figure 7. May (left) vs. August (right) Dendrograms

4. **Box Plot:** For the Box Plot, the May model generated a correct y-axis title, while the August model did not. The quantities on the y-axis represent the quantity sold per month, which is not specified in the y-axis title generated with the August model.

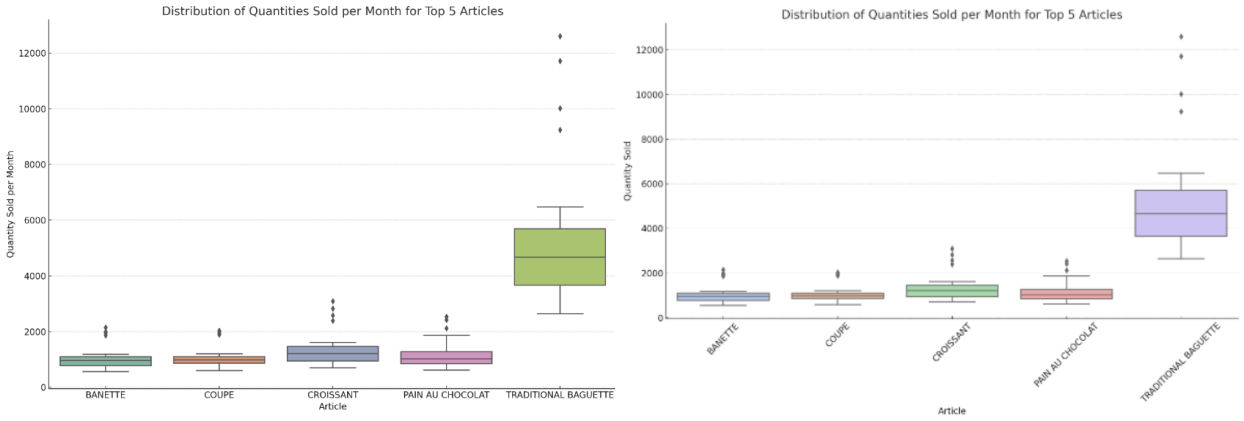

Figure 8. May (left) vs. August (right) Box Plots

Unsuccessful Charts

Despite overall improvements, the August model continued to exhibit persistent issues in data representation, with notable discrepancies observed in the Pie Charts, Venn Diagrams, Radar Charts, and Ridgeline Plots.

1. **Pie Chart:** Despite these charts' visual clarity, the actual sales percentages for items like Formule Sandwich, Traditional Baguette, and others differ significantly from the proportions shown in the pie chart, with some items appearing overrepresented. This

discrepancy highlights both models' difficulty with accurate data representation for pie charts.

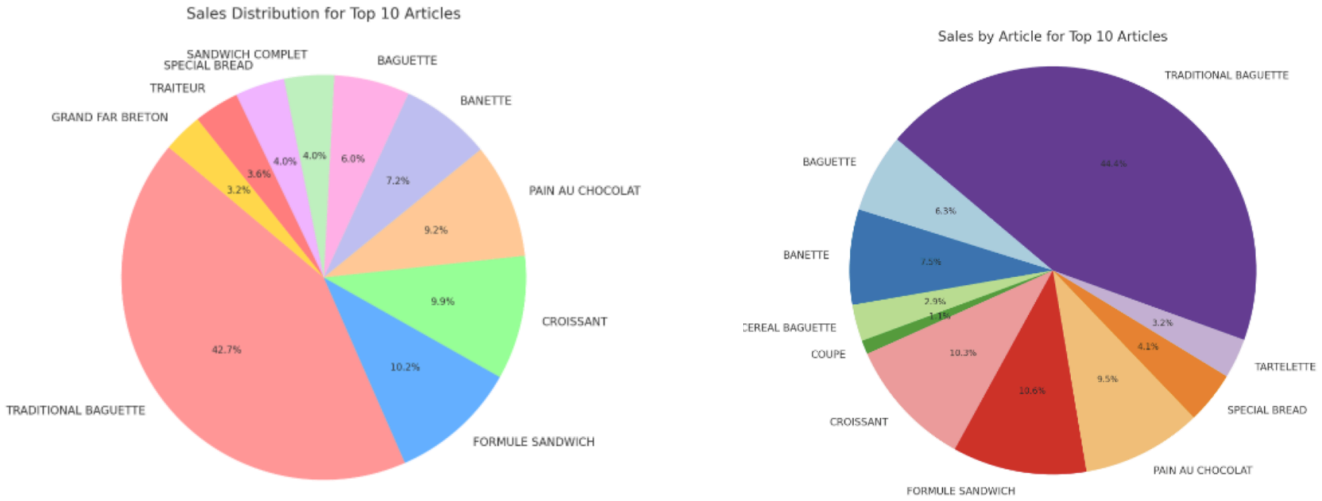

Figure 9. May (left) vs. August (right) Pie Charts

2. **Venn Diagram:** The August model performed worse in this instance, failing to clearly label the top three articles. Both models also presented inaccurate figures, particularly overstating the customer count for traditional baguettes.

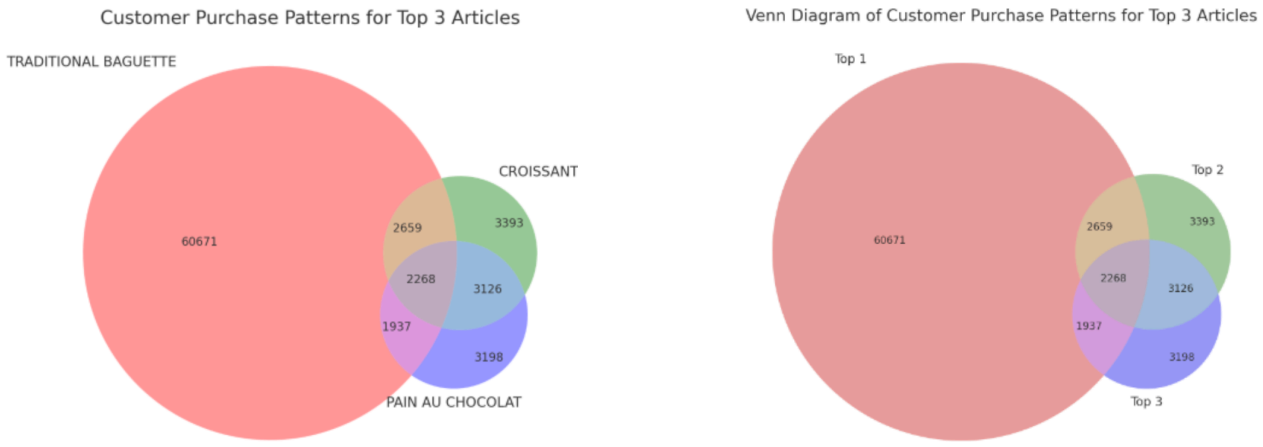

Figure 10. May (left) vs. August (right) Venn Diagrams

3. **Radar Chart:** The May model struggled to generate standard Radar Charts effectively, while the August model showed some improvements but still had significant inconsistencies in formatting and color differentiation, making it difficult to trust the charts for accurate data interpretation.

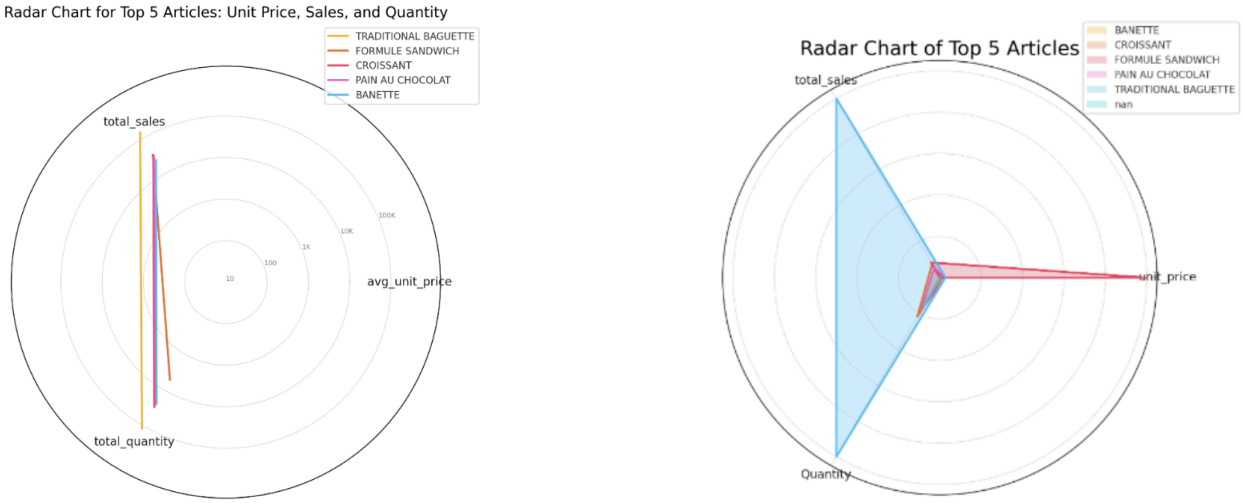

Figure 11. May (left) vs. August (right) Radar Charts

4. **Ridgeline Plot:** The May and August models displayed significant inconsistencies across data representation and formatting, with accuracy varying greatly from chart to chart. These inconsistencies could obscure important data patterns, making it difficult to derive meaningful insights.

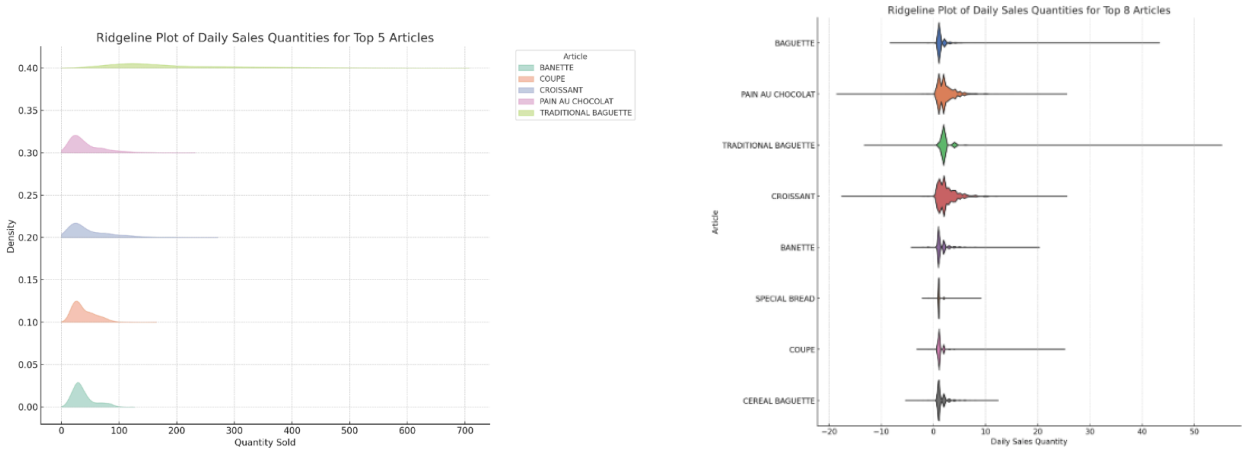

Figure 12. May (left) vs. August (right) Ridgeline Plots

Prompt Engineering

Prompt engineering involves the refining of instructions given to LLMs to improve their output. In our analysis, crafting clearer and more precise prompts lent itself to improved results for bubble charts and radar charts in the May model as shown below, while the August model struggled to yield improved results with prompt engineering.

1. **Bubble Chart:** Specifying “circles instead of x’s” ensures the desired chart formatting is achieved.

    Original Prompt: Create a Bubble Chart based on the data I uploaded. Include top 5 articles, total sales, and quantity.

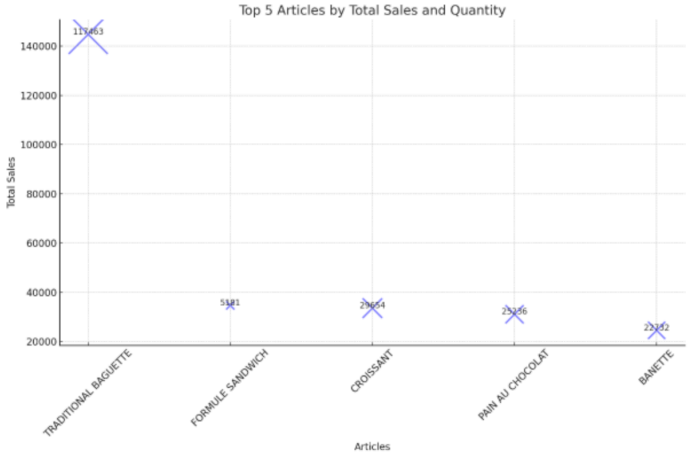

Figure 13. Bubble Chart Before Adjustments (May)

Refined Prompt: Create a Bubble Chart based on the data I uploaded with circles instead of x's. Include top 5 articles, total sales, and quantity.

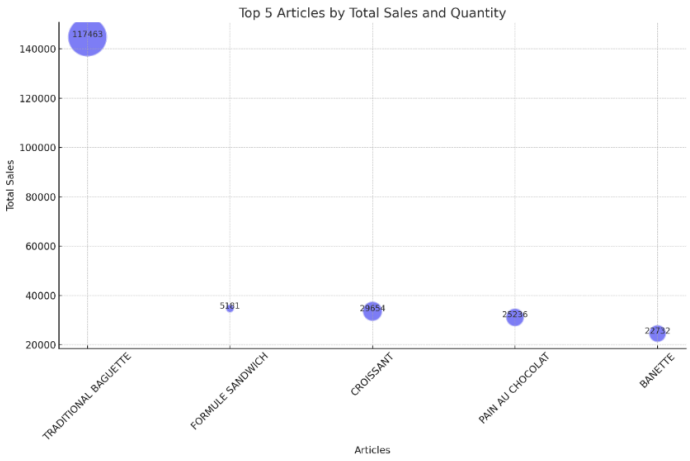

Figure 14. Bubble Chart After Adjustments (May)

2. **Radar Chart:** Specifying both time period and sales metrics outputs a higher quality chart in terms of visual clarity and overall relevance.

Original Prompt: Create a Radar Chart based on the data I uploaded. Use top 5 articles, unit price, sales, and quantity.

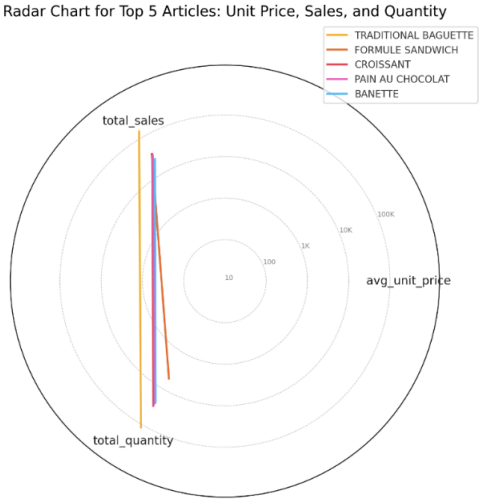

Figure 15. Radar Chart Before Adjustments (May)

Refined Prompt: Create a radar chart showing sales metrics (unit price, quantity sold, and total sales) by time of day (e.g., morning, afternoon, evening) for the top 5 articles.

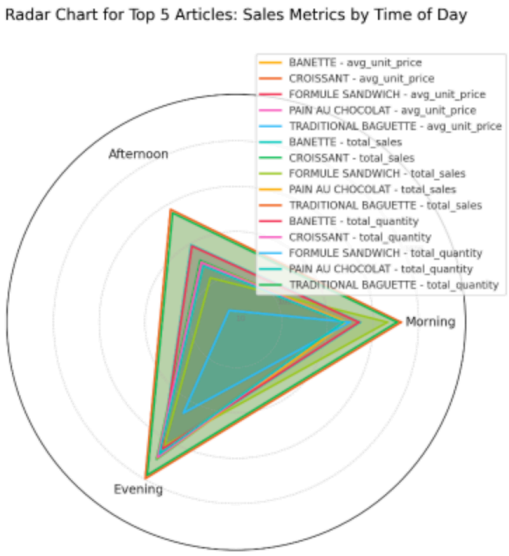

Figure 16. Radar Chart After Adjustments (May)

Industry Takeaways

AI

Our analysis of the ChatGPT-4 Omni model from May to August revealed significant improvements in output processing time, quality, and accuracy. While some issues surrounding data accuracy and formatting persist in chart generation, the progress we observed undoubtedly speaks to the LLMs growing reliability, as well as its ability to evolve. Our challenges in generating correct Bubble and Scatter charts, for example, exemplify areas where refinement is still needed. Nonetheless, with continued improvements in contextual understanding and self-checking, LLMs are becoming even more well-positioned to provide valuable insights that will ultimately drive innovation across numerous industries.

Private Equity

From our analysis, it is clear that the August model's improvements in both chart accuracy and processing speed are particularly exciting for the private equity industry. While PE professionals rely on raw data for their investment decisions, pairing AI-driven visualizations with that data provides a valuable tool that can enhance both the understanding and communication of complex financial data, ultimately allowing for a more accurate and comprehensive analysis.

These recent advancements in the August model represent a significant step in allowing firms to perform more efficient analyses, and supporting firms who manage and work on multiple deals. This improved efficiency allows the "human factor" to allocate more time to bigger-picture analysis, demonstrating how AI technology's evolution will increasingly enable private equity firms to manage complex data more efficiently than ever before. While we are thrilled to report the advancements we are witnessing as models continue to evolve, professionals should continue pairing these visualizations with underlying data to ensure accuracy.

Conclusion:

This report highlights the improvements made in the August version of the ChatGPT-4 Omni model, particularly as it relates to processing speed and data visualization accuracy compared to the May model. Despite this significant progress, formatting issues still persist in the Bubble Chart, Scatter Chart, Dendrogram, and Box Plot; and significant problems with incorrect data representation were observed in the Ridgeline Plot, Pie Chart, Venn Diagram, and Radar Chart. Both models occasionally struggled with highly complex datasets, resulting in slower processing times and sometimes incomplete or incorrect chart outputs. To address this, prompts were refined and charts with errors were retested. The prompt revision methods used include reducing the number of variables, specifying the data point representation, and further defining variables.

Index

Sec. 1 - Speed Data:

Summary Statistics					
TFT	(sec)	**RGT**	(sec)	**Total Time**	(sec)
Average TFT May:	8.28	Average RGT May:	39.21	Average Total Time May:	49.47
Average TFT August:	6.11	Average RGT August:	28.60	Average Total Time August:	34.86
Average Improvement:	2.45	Average Improvement:	10.61	Average Improvement:	14.61

Chart Type	**TFT May**	**TFT August**	**Improvement**	**RGT May**	**RGT August**	**Improvement**	**Total Time May**	**Total Time August**	**Improvement**
Stacked Bar Chart	3.81	2.93	0.88	85.87	65.56	20.31	89.68	68.49	21.19
Area Chart	5.32	1.88	3.44	20.68	18.16	2.52	26	20.04	5.96
Correlation Chart	5.10	3.78	1.32	13.06	19.31	-6.25	18.16	23.09	-4.93
Heat Map	4.48	2.03	2.45	65.69	18.73	46.96	70.17	20.76	49.41
Line Chart	26.18	6.53	19.65	4.70	38.63	-33.93	30.88	45.16	-14.28
Bar Chart	16.31	4.73	11.58	1.60	17.28	-15.68	17.91	22.01	-4.1
Dendrogram	6.68	10.70	-4.02	22.38	14.72	7.66	29.06	25.42	3.64
Histogram	23.93	11.29	12.64	1.98	16.66	-14.68	25.91	27.95	-2.04
Venn Diagram	28.38	6.55	21.83	3.42	15.71	-12.29	31.8	22.26	9.54
Pie Chart	25.24	5.17	20.07	1.27	18.78	-17.51	26.51	23.95	2.56
Box Plot	19.28	9.55	9.73	1.96	17.98	-16.02	21.25	27.53	-6.28
Ridgeline Plot	2.56	7.42	-4.86	51.27	26.77	24.5	53.83	34.19	19.64
Stacked Bar Chart	2.81	4.73	-1.92	95.02	90.55	4.47	98.83	95.28	3.55
Area Chart	4.72	4.31	0.41	29.53	15.14	14.39	34.25	19.45	14.8
Correlation Chart	5.34	4.00	1.34	26.27	23.23	3.04	31.61	27.23	4.38

Heat Map	6.00	3.63	2.37	18.35	18.68	-0.33	24.35	22.31	2.04
Line Chart	6.64	12.82	-6.18	19.15	19.98	-0.83	25.79	32.8	-7.01
Bar Chart	7.36	6.25	1.11	17.58	17.11	0.47	24.94	23.36	1.58
Dendrogram	8.10	5.97	2.13	19.05	19.68	-0.63	27.15	25.65	1.5
Histogram	13.19	6.68	6.51	76.06	15.69	60.37	89.25	22.37	66.88
Venn Diagram	11.04	7.15	3.89	24.43	23.56	0.87	35.47	30.71	4.76
Pie Chart	23.65	9.09	14.56	60.65	66.51	-5.86	84.3	75.6	8.7
Box Plot	10.40	8.32	2.08	35.22	33.10	2.12	45.62	41.42	4.2
Ridgeline Plot	4.11	6.63	-2.52	104.67	25.45	79.22	108.78	32.08	76.7
Bubble Chart	3.88	3.21	0.67	153.3	59.14	94.16	157.18	62.35	94.83
Radar Chart	3.45	11.29	-7.84	11.15	58.61	-47.46	14.6	69.9	-55.3
Radar Chart	5	6.45	-1.45	28.98	41.69	-12.71	33.98	48.14	-14.16
Radar Chart	7.69	5.42	2.27	55.92	42.21	13.71	64.01	47.63	16.38
Scatter Chart	10.4	4.93	5.47	24.82	22.81	2.01	35.22	27.74	7.48
Scatter Chart	26.53	3.85	22.68	2.21	4.9	-2.69	28.74	8.75	19.99
Histogram	3.15	4.18	-1.03	60.68	14.11	46.57	63.83	18.29	45.54
Ridgeline Plot	6.72	7.33	-0.61	107.35	22.65	84.7	114.07	29.98	84.09
Scatter Chart	9.23	9.04	0.19	17.44	43.36	-25.92	26.67	52.4	-25.73
Scatter Chart	9.35	5.43	3.92	73.87	19.96	53.91	83.22	25.39	57.83
Scatter Chart	8.03	4.98	3.05	28.54	20.14	8.4	36.57	25.12	11.45
Scatter Chart	7.22	4.82	2.4	68.67	14.71	53.96	75.89	19.53	56.36
Stacked Bar Chart	6.93	8.52	-1.59	18.13	37.04	-18.91	25.06	45.56	-20.5

Sec. 2 - Other Successful Charts:

Area Chart:

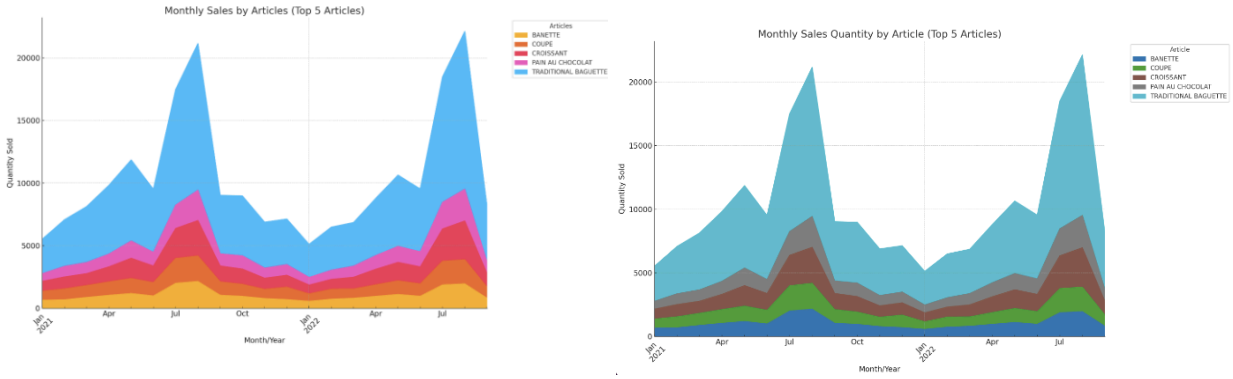

Figure 13. May (left) vs. August (right) Area Charts

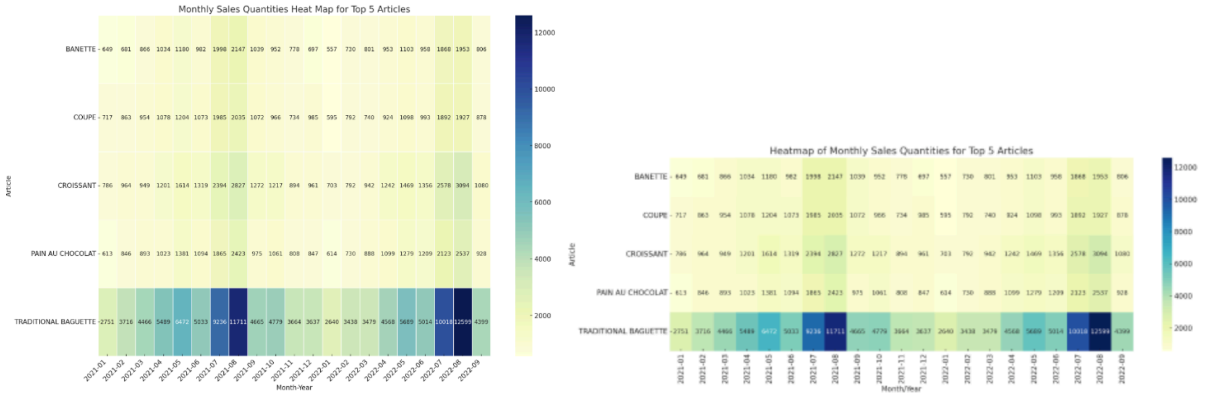

Figure 14. May (left) vs. August (right) Heat Maps

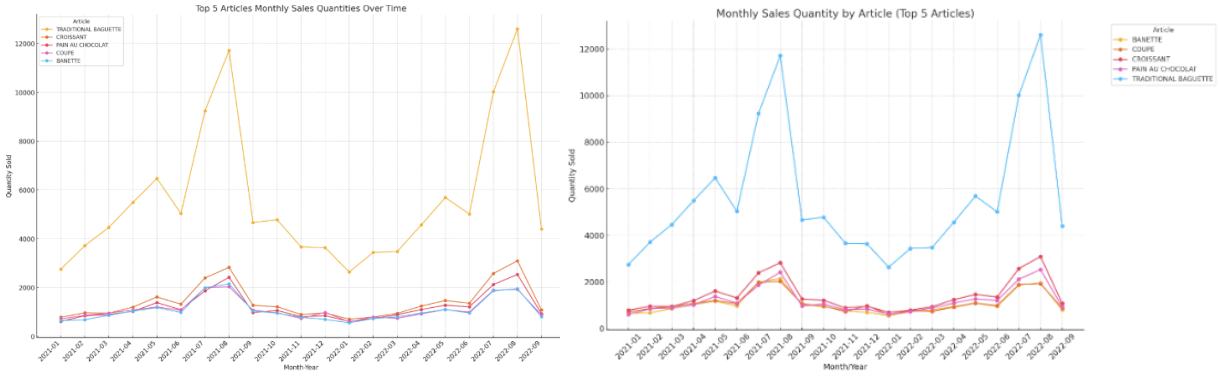

Figure 15. May (left) vs. August (right) Line Charts

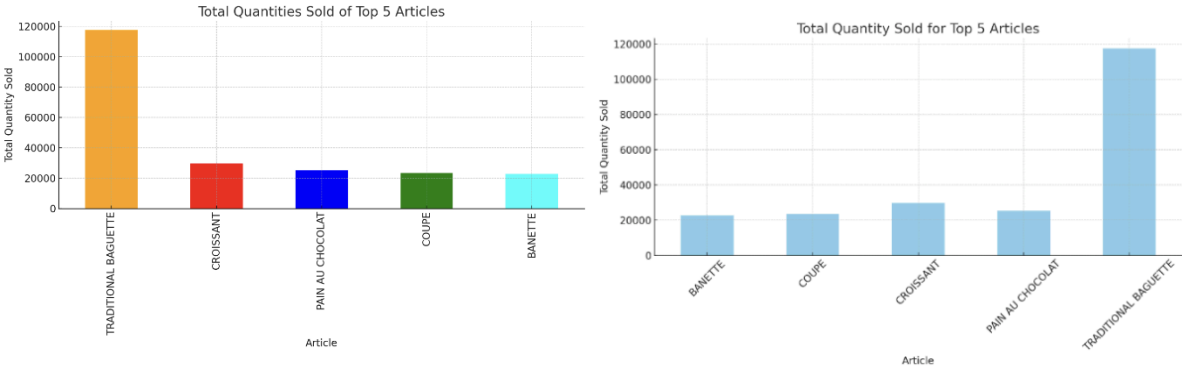

Figure 16. May (left) vs. August (right) Bar Charts

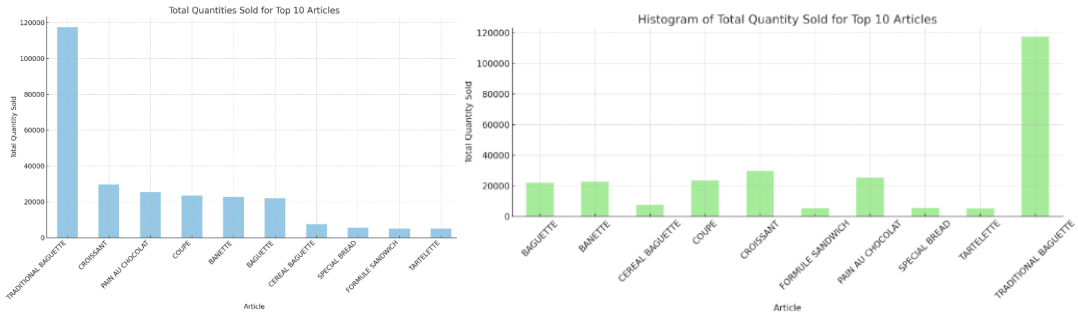

Figure 17. May (left) vs. August (right) Histograms