A Case Study of Human-Authored versus Automatic Dashboard Summaries

Jane Hoffswell
Adobe Research
Seattle, Washington, USA
jhoffs@adobe.com

Victor Soares Bursztyn
Adobe Research
San Jose, California, USA
soaresbu@adobe.com

Shunan Guo
Adobe Research
San Jose, California, USA
sguo@adobe.com

Eunyee Koh
Adobe Research
San Jose, California, USA
eunyee@adobe.com

Abstract

Automatically generated insights can help people interpret key trends in their data; similarly, dashboard summaries can highlight key insights for large and complex analytic dashboards that combine multiple datasets or visualizations. In this work we perform a case study evaluation with five industry professionals to understand how people prioritize insights and author concise summaries; to inform the design of improved automatic techniques, we compare the results to a fully automatic approach. We observed three notable characteristics of human-authored dashboard summaries compared to the automatic method: (1) incorporation of explanations or speculation, (2) improved structural consistency in the text, and (3) careful consideration of the precision for numeric values.

CCS Concepts

• **Human-centered computing** → **Visual analytics**; *Empirical studies in HCI*.

Keywords

Visualization, Dashboards, Automatic Insights, Summarization, Large Language Models.

ACM Reference Format:
Jane Hoffswell, Victor Soares Bursztyn, Shunan Guo, and Eunyee Koh. 2025. A Case Study of Human-Authored versus Automatic Dashboard Summaries. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3706599.3720155

1 Introduction

Analytic dashboards are popular for exploring, monitoring, and analyzing complex data. However, effective interpretation of such dashboards often requires domain expertise, which can make them inaccessible to a larger population of users. Natural language insights can help reduce the interpretation burden by automatically highlighting notable trends for individual charts. However, large dashboards may produce hundreds of insights, which again requires time and expertise to interpret effectively. This challenge leads to our use case: concise summarization of complex dashboards.

In order to inform the design of future automated summarization techniques for complex dashboards, this work aims to understand existing insight prioritization strategies for dashboard analysts. We conduct a case study evaluation with five industry professionals to rank insights and author dashboard summaries, and compare the results to a preliminary LLM-based summarization approach [7]. To structure our discussion, we first analyze the types and prioritization of insights selected by our human participants compared to the automated method (Section 4); we then provide an in-depth discussion of several example summaries (Section 5) and propose next steps for a larger-scale evaluation (Section 6).

2 Related Work

There has been extensive prior work on the automatic generation of natural language insights [5, 6, 8, 10, 21, 24, 25]. To better understand the varying definitions and scope for the term "insight," Battle and Ottley provide a rich review of prior work along with a unified formalism [1]; based on this work, our insights most closely align with the subspace of *data facts*. In particular, we leverage an approach based on Voder [21], which pairs annotated visualizations with automatically generated data facts. Once insights have been generated, a natural next step is to organize them into coherent narratives for larger artifacts [11, 18]. There has also been a variety of related work exploring both automated techniques [4, 19, 20, 26] as well as interactive techniques that can incorporate user preferences or feedback [23, 28, 29]. The importance of storytelling for dashboards has been particularly noted as a key open challenge in the space [15]. Recent advancements in large language models have provided yet another avenue to explore the automatic generation of natural language summaries or narratives [13, 22, 27], which is of particular interest in our work.

3 Evaluation Methodology

This preliminary evaluation explores the research question: *how do people prioritize insights and summarize complex analytic dashboards?* In this section, we introduce the methodology for our human evaluation. As a point of comparison, we also describe an automatic insight selection and LLM-based summarization method [7].

CHI EA '25, Yokohama, Japan

ACM ISBN 979-8-4007-1395-8/25/04
https://doi.org/10.1145/3706599.3720155

3.1 Evaluation with Five Industry Professionals

Task Each task had four basic steps: (1) *review* an interactive dashboard and the automatically-generated insights; (2) *select* between four and fifteen insights as the most important; (3) *explain* the rationale for why these insights were selected; and (4) *write* a natural language summary using the selected insights as appropriate. The full instructions for the task are included in the supplemental material. The dashboards were provided as `.html` files using embedded Vega [17] and Vega-Lite [16] visualizations. Participants were also given a spreadsheet to complete the task deliverables.

Procedure. Participants were assigned to one of six conditions (CHS, CSH, HCS, HSC, SCH, SHC). Each condition had three dashboards with increasing complexity in terms of the dashboard size (small, medium, large). Each dashboard visualized a different dataset (Calls, HR, Sales). The condition abbreviation denotes the dataset order and size; e.g., the CHS condition includes (1) the small Calls dashboard, (2) the medium HR dashboard, and (3) the large Sales dashboard.

Participants. We recruited six industry professionals from Upwork. All participants had a bachelors (5) or masters (1) degree in related fields, e.g., business (2), economics (1), computer science (1), or engineering (2), and self-reported previous experience with data analysis and visualization (mean=8 years, stdev=4.5 years). As part of the screening process, participants reviewed the task description in Upwork and submitted a fixed price bid to participate in the project; selected participants received between USD $100–$200 as compensation for this task according to their project bid. Participants completed the task within one month of accepting the contract; we did not constrain or record the specific task completion time. Due to some difficulty accessing the materials, the participant for condition SHC was not able to complete the study, and is thus excluded from subsequent analysis. The remaining participants thus produced a total of fifteen different summaries, which are included in the supplemental material along with the sample dashboards.

3.2 Automatic Summarization Method

The automatic generation of data-driven summaries is an increasingly important area of research given the rise of large language models [13, 22, 27]. Thus, a secondary goal of this evaluation was to explore differences between human-authored and automatic dashboard summaries. However, a key concern with fully LLM-based techniques is the potential for hallucination; we thus leverage a strategy from Hoffswell et al. that includes a scoring and ranking step to select a subset of insights, which are then passed to an LLM for summarization [7]. This approach provides more explicit guidance in the automation as to what insights are most important, and also mirrors the structure of our human evaluation; the method first identifies a subset of insights to summarize using OpenAI's GPT-3.5 (*gpt-35-turbo-v0613*) [3], with a decoding temperature of 0.5.

We generate one summary for each dataset and size, resulting in nine automatic summaries; for brevity, we leverage a similar naming convention to the human-authored summaries, where the letter indicates the dataset and the position indicates the size (first=small, second=medium, third=large). Hence, CXX refers to the automatic summary for the small Calls dashboard. An example prompt and the summaries are included in the supplemental material.

Our scoring metric uses the same approach as Hoffswell et al. [7], and is defined as follows: $score = 0.3 * layoutScore + 0.7 * valueScore$. The $layoutScore$ is defined as $0.5 * panelScore + 0.5 * tableCol$; the $panelScore$ is then defined as $0.5 * panelRow + 0.5 * panelCol$, and normalized between zero and one. The $panelRow$, $panelCol$, and $tableCol$ values are defined as $value = (max(idx) - idx)/max(idx)$, where idx is the zero-indexed value from the layout and $max(idx)$ is the maximum idx for that value across all charts in the dashboard, which is then normalized between zero and one. The $valueScore$ describes how common the referenced values are in the insights; we thus count the number of times each value is mentioned across all insights, and compute the $valueScore$ as $\frac{1}{n} * \sum c_x$, where n is the number of unique values mentioned in the insight and c_x is the count for each value across all insights; this score is normalized between zero and one. This approach emphasizes both the underlying dashboard layout, as well as the particular insight values commonly referenced across all visualizations; while other prioritization or automation approaches are possible, we focus on this example as once source of comparison for the human-authored summaries.

3.3 Nine Sample Dashboards

We developed nine sample dashboards based on three open source datasets (Calls [2], HR [2], and Sales [9]) and three sizes (small, medium, and large). To make the dashboards more timely, we updated the dates to include or focus on the year 2023. We also manipulated the data in order to generate some more compelling automatic insights; for example, *"The values of 'Sentiment [Very Positive]' are highly skewed towards 'Sundial' (10% in total)"* (`C10SK5`). Images for all nine dashboards are included in the supplemental material. Figure 1 includes all the visualizations across all three datasets and sizes, with the charts arranged in the layout for the large dashboard.

For each dashboard, we automatically generate insights using an approach based on Voder [21]. In particular, we generate twelve different types of automatic insights or data facts: *minimum* (MI), *maximum* (MX), *max extent* (ME), *highest bar* (HB), *skew* (SK), *long tail distribution* (LT), *seasonality* (SE), *trend* (TR), *spike* (SP), *decline* (DE), *anomaly* (AN), and *correlation* (CO). We will use these insight-type abbreviations throughout the paper for brevity. For reference, we assign a unique ID to each insight as follows: the first letter indicates the dashboard (Calls, HR, Sales), the next two characters are the chart number from Figure 1, followed by the two character insight-type abbreviation defined above; for data tables, the final number indicates which column the insight corresponds to.

Calls. Our large *Call Center Overview* dashboard has ten visualizations (Figure 1, *left*) with nine attributes, listed here by the number of visualizations using the attribute: `Calls`, `Sentiment`, `Reason`, `Duration`, `Channel`, `Date`, `Day`, `Client`, `Call Center`. The medium version has five visualizations (C5, C6, C2, C9, C8) with a primary emphasis on `Calls` and `Sentiment`. Finally, the small version includes two visualizations (a donut chart C2 and multi-line chart C8), with an undiluted focus on `Calls` and `Sentiment`. We generate 127, 49, and 28 insights respectively for the three sizes.

HR. Our large *Human Resources Executive Overview* dashboard has nine visualizations (Figure 1, *center*) exploring nine attributes, listed here by the number of visualizations using the attribute: `Headcount`, `Full-Time` or `Termed`, `Department`, `Hire Date`, `Gender`, `Location`,

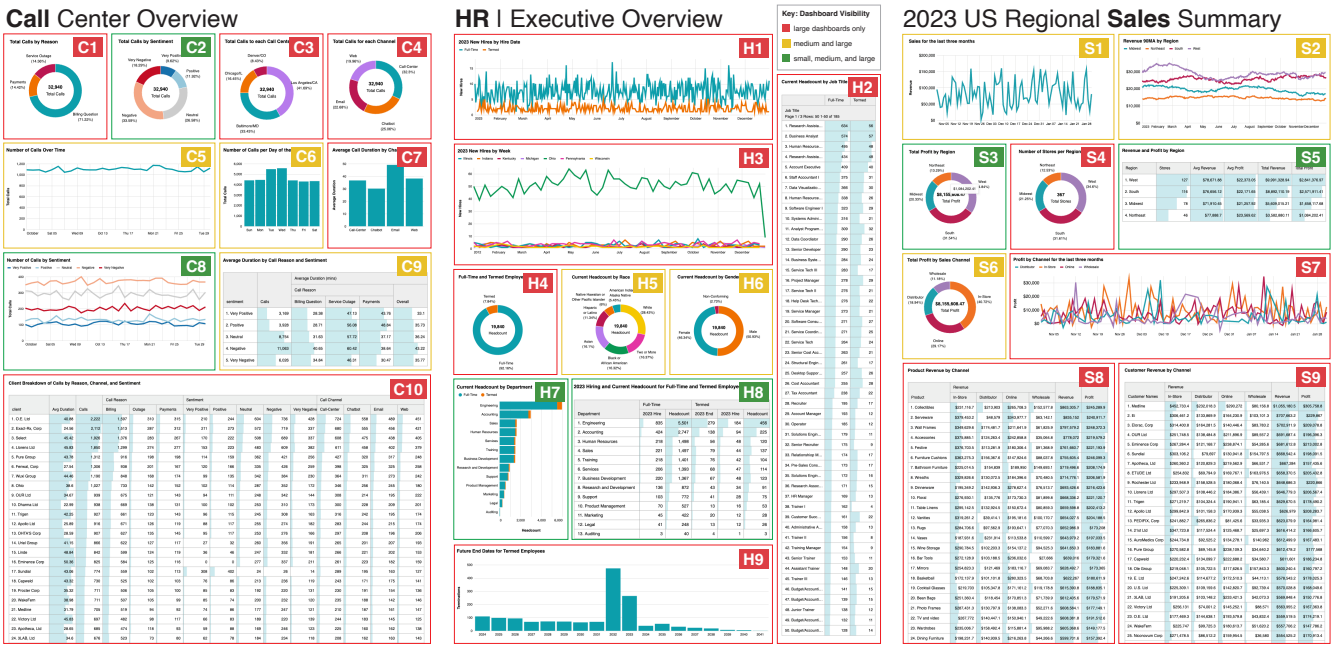

Figure 1: We developed nine dashboards based on three datasets: (left) *Call Center Overview*, (center) *HR Executive Overview*, and (right) *2023 US Regional Sales Summary*. For each dataset, we varied the complexity to create a small, medium, and large dashboard. This figure shows the layout for the large dashboards. The medium and small dashboards use a subset of charts with different layouts (see the supplemental material). The label color indicates the availability of each chart across the three sizes: charts labeled green appear in the small, medium, and large dashboards; yellow in the medium and large; red for the large only.

Race, Job Title, and Term End; this dashboard also uses two filters (2023 Hire and 2023 End) to filter the Headcount based on either the Hire Date or Term End. The medium version has four visualizations (H5, H6, H8, H7) emphasizing Headcount, Department, and Full-Time or Termed. Finally, the small version has two visualizations (a bar chart H7 without the color encoding, and a table H8), which thus has a stronger focus on Headcount and Department. The dashboards have some minor variations in the bar chart H7; for the medium dashboard, we change the orientation; for the small dashboard, we remove the color encoding to reduce the number of insights. We generate 107, 41, and 30 insights for the three sizes.

Sales. The *2023 US Regional Sales Summary* dashboard is based on the "US Regional Sales Data" [2] and dashboard [14]. Our large dashboard has nine visualizations (Figure 1, *right*) which leverage nine different attributes, listed here by the number of visualizations using the attribute: Profit, Revenue, Region, Sales Channel, Date, Stores, Revenue 90MA, Product, and Customer. The medium version has five visualizations (S6, S3, S2, S1, S5) emphasizing Profit and Revenue by Region. Finally, the small version has two visualizations (a donut chart S3 and table S5) emphasizing the Profit by Region. We generate 112, 44, and 16 insights for the three sizes.

4 Evaluation Results: Insight Selection

Across all three datasets and sizes (i.e., 15 summaries), our five participants selected 101 distinct insights out of the 346 possible insights generated (29.2%), whereas our automated method (resulting in 9 summaries) selected 48 insights, 26 of which were also selected by our human participants. Figure 2A shows the number of insights selected overall, broken down by the insight type. The most common insight types selected by our human participants were *maximum* (MX), *minimum* (MI), and *max extent* (ME), whereas the automatic method prioritized *maximum* (MX) and *max extent* (ME).

Per the evaluation methodology (Section 3), participants were instructed to select between four and fifteen of the provided insights for use in their final summary. For each summary, our human participants selected 10.5 insights on average (9.6 for the small dashboard, 10.2 for medium, and 11.6 for large); in contrast, our automatic method selected 7.6 insights for each summary on average (4.3 for the small dashboard, 5.7 for medium, and 12.7 for large). Figure 3 shows the number of insights selected for each summary (i.e., number of rows) as well as the prioritization (i.e., the order).

4.1 Insight Selection Results: Calls

Across the three sizes, our five human participants selected 37 of 127 distinct insights (19 available in the small dashboard, 10 available in the medium, and 8 available in the large). Participants selected an average of 12 insights overall (11.5 for small, 12.5 for medium, and 12* for large); *only one sample is available for the large dashboard.

Figure 2B shows the distribution of selected insights by the insight type. Notably, there were three insights with high agreement across participants (i.e., each insight was selected by four of the five participants, as shown with the darkest purple color, for a selection rate of 0.667). One of these insights, the only *correlation* insight for

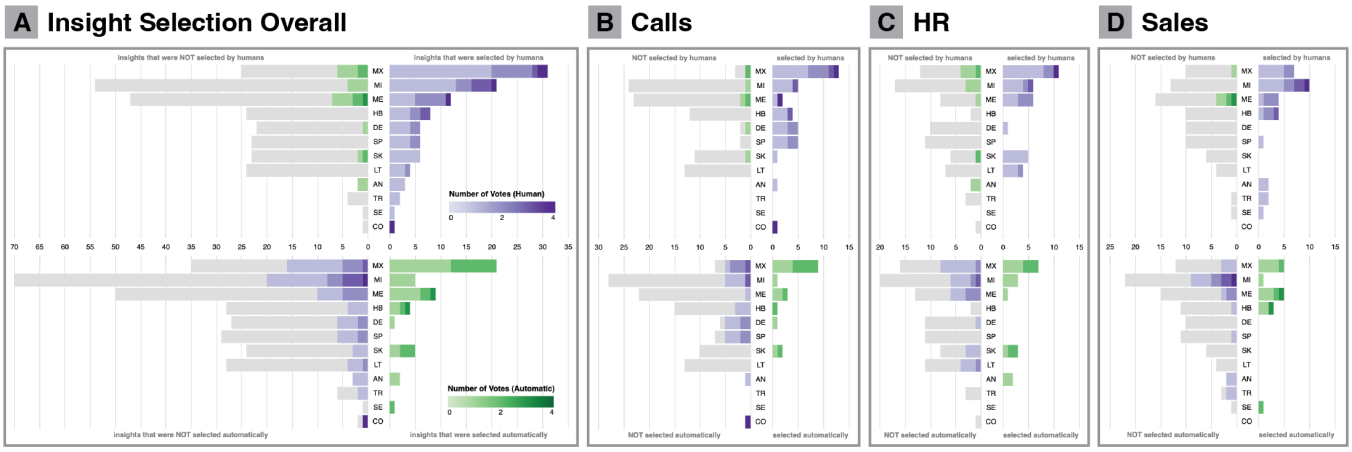

Figure 2: This visualization compares the human-selected (top) and automatically-selected (bottom) insights across the different insight types and datasets: (A) Overall, (B) Calls, (C) HR, (D) Sales. Purple colors correspond to the number of votes from our human participants, whereas green colors correspond to the number of votes from our automatic method (across the three dashboard sizes). The right side of each chart shows the number of insights selected by each method, as well as the number of votes each insight received; darker colors correspond to insights that were more popular. To facilitate comparisons of the two selection approaches, the left side of each chart shows the number of insights that were not selected for each method, with the color corresponding to the number of insights selected by the automatic method but *not* the human participants, and vice versa.

the dashboard (`C08CO`), was *not* selected by any of the automatic summaries, which suggests a possible area of improvement for the automated selection approach. The most popular maximum value insight (`C02MX`), *"'Negative' had the greatest value, with 11,063 in 'Calls' (34% in total),"* was also selected for two of the three automatic summaries (medium and large), whereas the most popular *max extent* insight (`C02ME`) was selected for the small dashboard, which includes similar information: *"The max item, 'Negative', is 26% more than the second highest one, 'Neutral', in 'Calls'."* Of these three insights, `C02MX` had the highest prioritization overall (see Figure 3A), and appeared within the top six for all of the summaries, with an average rank of 3.17 (σ=1.95) compared to 4.60 (σ=3.20) and 5.75 (σ=3.27) for `C02ME` and `C08CO` respectively.

The three insights discussed above were available for all three dashboard sizes; of the insights introduced in the medium dashboard, the *highest bar* insight (`C06HB`) appeared in all five possible summaries (HCS, SCH, HSC, XCX, XXC). As an interesting comparison, the *maximum* insight `C09MX1` appeared in all three possible human summaries, but none of the automatic versions; this insight pulls in new dimensions (`Reason`) and metrics (`Duration`). Notably, the automatic summary for the medium dashboard does not mention the `Duration` or `Reason` at all, whereas the large dashboard selects different insights about the `Duration` (e.g., `C07MX` and `C07SK`) and `Reason` (e.g., `C01MX`), some of which were also selected by our human participant for the large dashboard (HSC).

4.2 Insight Selection Results: HR

Our participants selected 33 of 107 distinct insights (19 from the small dashboard, 5 for medium, and 9 for large). The average number of insights selected for the "HR" dashboard was 9.6 overall (9.5, 9*, and 10 for the three sizes respectively); *only one sample is available for the medium dashboard. Figure 2C shows the distribution of insights by the insight type. The insight `H08MX0` had the highest agreement (i.e., *"'Engineering' had the highest value, with 835 in '2023 Hire [Full-Time]' (30% in total)"*), appearing in four of five human summaries, and two of the automated summaries (small and medium); this insight had an average rank of 5.17 (σ=3.06). In contrast to the Calls dashboard, we saw less agreement in the insights that were selected, hence the lighter colors shown in Figure 3B.

The second most commonly occurring insight was `H05MX`, which first appeared in the medium dashboard, and was used in four of five possible summaries. Participant CSH was the only one not to select this insight, which reported differences in `Headcount` broken down by `Race`. Unlike other participants, all of the insights chosen by CSH focus on one dimension, `Department`, specifically the value "Engineering." Participant CSH explained that *"I chose to focus on the Engineering department within the company. The insights chosen really reinforce the type of jobs that the company is built on."* In contrast, participant SCH preferred to provide broader coverage when producing their summary for the large dashboard: *"I selected insights focusing on the highest values and distribution for each chart."* This difference in approaches showcases a challenge with producing effective summaries: there may not be one true perfect summary.

4.3 Insight Selection Results: Sales

Our participants selected 31 of 112 distinct insights (8 from the small dashboard, 9 for the medium, and 14 for large). The average number of insights selected for the Sales dashboard was 9.8 overall (6*, 8.5, and 13 for the three sizes respectively); *only one sample is available for the small dashboard. Figure 2D shows the distribution of insights by the insight type; while the *maximum* insights were

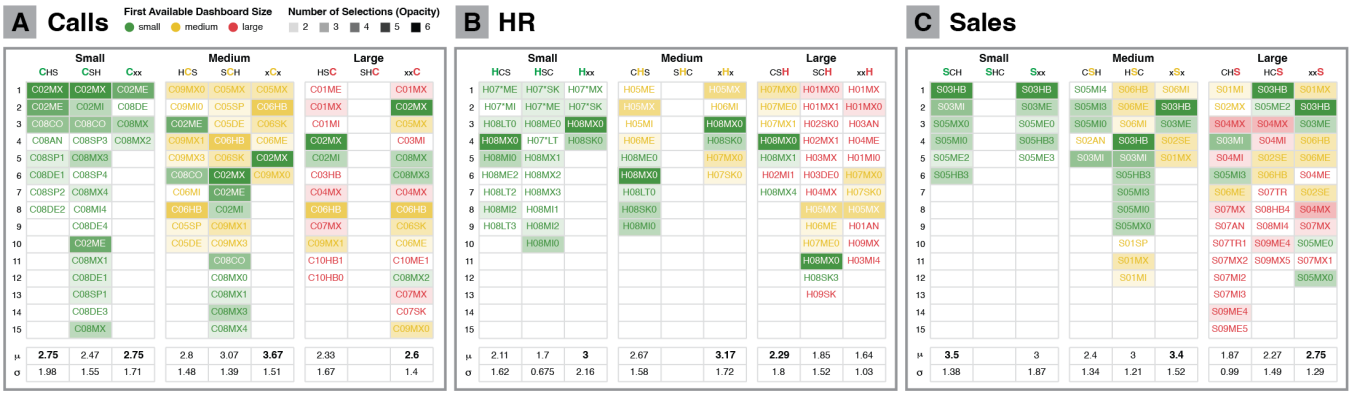

Figure 3: This figure shows the prioritization, popularity, and availability of the selected insights across summaries. Each column corresponds to one human or automatic summary; the opacity shows the popularity of the insights, with darker colors indicating insights that were selected across more summaries for the same dashboard, e.g., the max insight for chart H8 in Figure 1 (`H08MX0`) was selected by four participants (HCS, CHS, CSH, SCH) and two of the automated summaries (HXX, XHX); despite appearing in many summaries, the prioritization (selection order) of the insight varied. For each summary, we also record the mean (μ) and standard deviation (σ) for the number of times the insights were selected across the trials. The color shows the availability of the insight across the dashboard sizes, i.e., green corresponds to insights available for the small, medium, and large dashboard, yellow for the medium and large dashboard, and red for the large dashboard only. The availability can showcase interesting selection patterns; for example, when producing a summary for the medium Sales dashboard, participant CSH selected mostly insights that were available in the small dashboard, and was the only one to select the *anomaly* insight (`S02AN`).

the most popular for the other two dashboards, our human participants selected ten distinct *minimum* insights for this dashboard, whereas our automatic method only selected one (`S06MI`). The automatic method selected 15 distinct insights across the three sizes, five of which were not selected by any of our human participants. Figure 2D shows that these insights correspond to four *max extent* insights and one *maximum*.

Our insight generation approach produces several highly-related insights, i.e., the *maximum* values (MX), the *highest bars* (HB), and the difference between the maximum value and the second highest, a.k.a. the *max extent* (ME). For example, in the `Profit` by `Region` chart (Figure 1 S3), the *highest bars* insight (`S03HB`) states that *"'West' and 'South' contain the highest values, with around 2,706,644.595 in 'Profit' (66% in total),"* and the *max extent* insight (`S03ME`) introduces a comparison of these values by noting that *"The max item, 'West', is 10% more than the second highest one, 'South', in 'Profit'."* For the summary, the automatic LLM-based approach combines these two insights into one sentence: *"The highest bars in terms of profit are in the West and South regions, with the West region having 10% more profit than the South region."* The automatic approach ranks these two insights similarly, thus resulting in both types of insights commonly being selected. In fact, both these insights were selected by the automated method for all three sizes, whereas the *max extent* insight was selected by none of our five human participants, while three selected the *highest bar* insight (`S03HB`), thereby making it the most popular insight for the dashboard.

4.4 Overall Strategies for Insight Prioritization

Across the three datasets, we saw some common selection patterns based on the insight types; for example, while several of the insight types produced similar results, and were thus similarly prioritized by the automated method (see Section 4.3), our human participants tended to select more *maximum* insights than *highest bar* or *max extent* across all of the dashboards (Figure 2A). While this similarity resulted in some strong early agreement for certain insights, there were 67 insights that were only selected for one summary during the evaluation (e.g., insights with a white background in Figure 3), thus highlighting the key differences in how people prioritize insights.

While our automated method followed the same ranking approach for all summaries, our human participants often exhibited different preferences that had a more drastic impact on the insight selection. Hence, automated methods may similarly need to adapt based on the goals of the target audience for the summary. For example, as discussed in Section 4.2, participant CSH and SCH expressed different goals when creating their summary, despite having access to the same dashboard. Interestingly, participant CSH also leveraged differences approaches across dashboards of different sizes; for example, they selected a large number of insights for the small dashboard, and a small number for the medium and large dashboards, whereas other participants often selected a larger number of insights when more were available.

5 Case Study: Human vs. Automatic Summaries

In this section, we discuss some key observations from the final human-authored versus automatic dashboard summaries. The full set of human-authored and automatic summaries is provided in the supplemental material; we also include one comparative example here for the medium "Call Center Overview" dashboard, which we will discuss in more depth in the following section: (1) the summary for participant HCS and (2) the automatically-generated summary

with automatically-selected insights (XCX). Participant HCS selected ten insights, whereas the automatic pipeline selected six (Figure 3A). Only two of the insights were selected by both the human-authored and automated methods: *"'Wednesday' and 'Tuesday' contain the greatest values, with around 5,525.5 in 'Calls' (34% in total)"* (C06HB) and *"'Negative' had the highest value, with 11,063 in 'Calls' (34% in total)"* (C09MX0). The resulting summaries are as follows:

(1) Human-Authored Summary: *"In the month of October, only 22 percent of customers have shown positive sentiment in customer care calls. A high percentage of negative sentiment calls indicates unhappy and sad customers. Calls for billing-related questions demonstrate the most negative sentiment, whereas calls related to payments indicate the highest positive sentiment. Client calls that happen on Tuesday and Wednesday contribute 34% of overall calls compared to the remaining weekdays. The average duration of service-outage related calls is higher compared to other call reasons. Number of calls increased notably by 10 percent at the start of the month (4th to 6th October) and dropped by 10 percent towards the end of the October month."*

(2) Automatic Summary: *"On October 21st, the highest number of calls reached 1,170, which was 7% higher than the average of 1,098. The days with the highest call volume were Wednesday and Tuesday, accounting for 34% of the total calls. The values of calls were highly skewed towards these two days. Wednesday had the highest number of calls, which was 2% higher than Tuesday. The highest value in the negative category for calls was 11,063, accounting for 34% of the total calls."*

5.1 Summary Case Study Results and Discussion

This case study highlights several common trends from the human-authored summaries in our evaluation. Compared to the results from the automatic summaries, participants tended to (1) add explanations (or speculation), (2) improve structural consistency in the narrative flow, and (3) remove approximate or highly precise values.

5.1.1 Explanation or Speculation. Four of the five participants introduced explanations or speculation to the summary (e.g., *"negative sentiment calls indicated unhappy or sad customers"* - HCS), which the automatic approach does not. According to Lundgard and Satyanarayan's four-level model of semantic content [12], these sentences could generally be categorized as L4 insights (i.e., contextual or domain-specific information), whereas our automatic insight-generation approach primarily produces L2 (statistical) or L3 (perceptual) insights, which are then more directly paraphrased by GPT-3.5 for our automatic summaries. To improve the narrative variation in the automatic summaries, future work could explore how to *safely* introduce L4 insights into the results. The key challenge is to introduce these insights without misrepresenting the certainty of the model or hallucinating correlations (or causation) that do not exist. For the human-authored summaries, our participants used words like *"could be"* or *"a possiblity [sic]"* to show their uncertainty with a certain interpretation (e.g., *"Los Angeles/CA and Baltimore/MD report the heighest [sic] number of calls reported, which could be a huge indication of regional-level service issues in those areas"* - HSC), whereas others simply added such context as fact (e.g., *"consumers who are unhappy are more likely to report that vs those who were pleased with the service"* - CSH).

5.1.2 Structural Consistency. Two participants (HCS and SCH) were particularly sensitive to ensuring symmetry in the sentence structure, and often refined the wording or details in the auto-generated insights in order to improve the overall structure and flow of their final summary. In the case study summary, participant HCS took the insights related to the *spike* (C05SP: *"'Calls' increased notably during the period of Oct. 4th to 6th, up by 10% from 1,049 to 1,152"*) and *decline* (C05DE: *"Between Oct. 21st and 26th, the amount of 'Calls' dropped notably, down by 10% from 1,170 to 1,054"*) and combined them using a mirrored structure: *"Number of calls increased notably by 10 percent at the start of the month (4th to 6th October) and dropped by 10 percent towards the end of the October month."*

In order to effectively combine sentences, the participants often had to look up information that was not included in the original, automatically-generated insights. For example, to create a summary for the small Sales dashboard, participant SCH selected insights related to the maximum `Profit` (S03HB: *"'West' and 'South' contain the highest values, with around 2,706,644.595 in 'Profit' (66% in total)"* and `Stores` (S04MX: *"'West' had the greatest value, with 127 in 'Stores' (35% in total)"*), and then extracted the `Stores` data for the "South" `Region` from the provided dashboard in order to write a more structurally consistent summary: *"West has the highest profit at $2.84M, closely followed by South with $2.57M. Combined, West and South account for $5.41M which is 66% of total profit. [...] West and South also have the highest number of stores with 127 and 116, respectively. Combined, West and South account for 243 number of stores which is 66% of all total stores."*

5.1.3 Approximate or Precise Values. All five participants tended to reduce the precision of values reported in the summaries, at least for a subset of the sentences, if not all of them. In the case study example, participant HCS removed the exact values (i.e., *"around 5,525.5 in 'Calls'"*), and only kept the percentages (i.e., *"34% of overall calls"*). On the other hand, two participants (HCS and HSC) also chose to add some precision to the percentages to better match the precision available in the interactive dashboard; for example, participant HSC updated the insight for the maximum `Calls` by `Sentiment` (C09MX0: *"'Negative' had the greatest value, with 11,063 in 'Calls' (34% in total)"*) to add precision and additional details as follows: *"Majority of customers contacting the support are frustated [sic], as almost 52% of the total calls have a negative sentiment (33.59% negative, 18.29% very negative)"* (Figure 1, C2).

6 Limitations and Future Work

We performed a preliminary evaluation and case study with five industry professionals to explore the selection of important insights and summary authoring process compared to an automatic, LLM-based summarization method. While our evaluation results highlighted some examples of insights that were commonly selected across different dashboard sizes, given the small scale of the study, it remains difficult to draw clear conclusions from these results. Furthermore, we also saw different selection and prioritization approaches based on the particular goals or interests of our users, which suggests that there may not be one straightforward set of insights that is best for all scenarios. By extending this study to a larger set of participants in future work, we hope to more clearly identify the types of insight that are more universally important for

summarization as well as the different selection criteria commonly employed. In Section 5, we also reflected on some initial observations regarding the characteristics of human-authored summaries, but we would like to further evaluate the quality of human- and auto-generated summaries directly in future work.

7 Conclusion

In this work we performed a preliminary evaluation with five industry professionals to understand how people prioritize insights and write concise summaries of analytic dashboards; we also compared these results to an automatic LLM-based summarization approach as a case study. We found that participants often introduced explanations (or speculation) to provide more variety or context in the summary. Our participants also tended to focus on the high-level themes and eliminate exact values that were not as pertinent to the main story; however, some participants introduced new details or precision in order to improve the overall consistency (i.e., structural or dashboard-specific) of the summary.

References

[1] Leilani Battle and Alvitta Ottley. 2023. What Do We Mean When We Say "Insight"? A Formal Synthesis of Existing Theory. *IEEE Transactions on Visualization and Computer Graphics* (2023). doi:10.1109/TVCG.2023.3326698

[2] Mark Bradbourne. [n. d.]. *Real World Fake Data.* https://data.world/markbradbourne/rwfd-real-world-fake-data

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[4] Qing Chen, Shixiong Cao, Jiazhe Wang, and Nan Cao. 2023. How Does Automation Shape the Process of Narrative Visualization: A Survey of Tools. *IEEE Transactions on Visualization and Computer Graphics* (2023). doi:10.1109/TVCG.2023.3261320

[5] Rui Ding, Shi Han, Yong Xu, Haidong Zhang, and Dongmei Zhang. 2019. QuickInsights: Quick and Automatic Discovery of Insights from Multi-Dimensional Data. In *ACM International Conference on Management of Data.* doi:10.1145/3299869.3314037

[6] Sneha Gathani, Anamaria Crisan, Vidya Setlur, and Arjun Srinivasan. 2024. Groot: A System for Editing and Configuring Automated Data Insights. In *IEEE Visualization and Visual Analytics.* IEEE. doi:10.1109/VIS55277.2024.00015

[7] Jane Hoffswell, Victor Soares Bursztyn, Shunan Guo, Jesse Martinez, and Eunyee Koh. 2025. Representing Visualization Insights as a Dense Insight Network. *arXiv preprint* (2025). doi:10.48550/arXiv.2501.13309

[8] Shankar Kantharaj, Rixie Tiffany Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. 2022. Chart-to-Text: A Large-Scale Benchmark for Chart Summarization. In *Proceedings of the Association for Computational Linguistics.* doi:10.18653/v1/2022.acl-long.277

[9] Udit kumar Chatterjee. [n. d.]. *US Regional Sales Data.* https://data.world/dataman-udit/us-regional-sales-data

[10] Po-Ming Law, Alex Endert, and John Stasko. 2020. Characterizing Automated Data Insights. In *IEEE Visualization Conference (VIS).* IEEE. doi:10.1109/VIS47514.2020.00041

[11] Bongshin Lee, Nathalie Henry Riche, Petra Isenberg, and Sheelagh Carpendale. 2015. More Than Telling a Story: Transforming Data into Visually Shared Stories. *IEEE Computer Graphics and Applications* (2015). doi:10.1109/MCG.2015.99

[12] Alan Lundgard and Arvind Satyanarayan. 2021. Accessible Visualization via Natural Language Descriptions: A Four-level Model of Semantic Content. *IEEE Transactions on Visualization and Computer Graphics* (2021). doi:10.1109/TVCG.2021.3114770

[13] Pingchuan Ma, Rui Ding, Shuai Wang, Shi Han, and Dongmei Zhang. 2023. InsightPilot: An LLM-Empowered Automated Data Exploration System. In *Conference on Empirical Methods in Natural Language Processing: System Demonstrations.* doi:10.18653/v1/2023.emnlp-demo.31

[14] quantumudit. 2020. *US Regional Sales Report.* https://community.fabric.microsoft.com/t5/Data-Stories-Gallery/US-Regional-Sales-Report/m-p/1249483

[15] Alper Sarikaya, Michael Correll, Lyn Bartram, Melanie Tory, and Danyel Fisher. 2018. What Do We Talk About When We Talk About Dashboards? *IEEE Transactions on Visualization and Computer Graphics* (2018). doi:10.1109/TVCG.2018.2864903

[16] Arvind Satyanarayan, Dominik Moritz, Kanit Wongsuphasawat, and Jeffrey Heer. 2016. Vega-Lite: A Grammar of Interactive Graphics. *IEEE Transactions on Visualization and Computer Graphics* (2016). doi:10.1109/TVCG.2016.2599030

[17] Arvind Satyanarayan, Ryan Russell, Jane Hoffswell, and Jeffrey Heer. 2015. Reactive Vega: A Streaming Dataflow Architecture for Declarative Interactive Visualization. *IEEE Transactions on Visualization and Computer Graphics* (2015). doi:10.1109/TVCG.2015.2467091

[18] Edward Segel and Jeffrey Heer. 2010. Narrative Visualization: Telling Stories with Data. *IEEE Transactions on Visualization and Computer Graphics* (2010). doi:10.1109/TVCG.2010.179

[19] Danqing Shi, Fuling Sun, Xinyue Xu, Xingyu Lan, David Gotz, and Nan Cao. 2021. AutoClips: An Automatic Approach to Video Generation from Data Facts. In *Computer Graphics Forum.* Wiley Online Library. doi:10.1111/cgf.14324

[20] Danqing Shi, Xinyue Xu, Fuling Sun, Yang Shi, and Nan Cao. 2020. Calliope: Automatic Visual Data Story Generation from a Spreadsheet. *IEEE Transactions on Visualization and Computer Graphics* (2020). doi:10.1109/TVCG.2020.3030403

[21] Arjun Srinivasan, Steven M Drucker, Alex Endert, and John Stasko. 2018. Augmenting Visualizations with Interactive Data Facts to Facilitate Interpretation and Communication. *IEEE Transactions on Visualization and Computer Graphics* (2018). doi:10.1109/TVCG.2018.2865145

[22] Nicole Sultanum and Arjun Srinivasan. 2023. DataTales: Investigating the use of Large Language Models for Authoring Data-Driven Articles. In *IEEE Visualization and Visual Analytics.* IEEE. doi:10.1109/VIS54172.2023.00055

[23] Mengdi Sun, Ligan Cai, Weiwei Cui, Yanqiu Wu, Yang Shi, and Nan Cao. 2022. Erato: Cooperative Data Story Editing via Fact Interpolation. *IEEE Transactions on Visualization and Computer Graphics* (2022). doi:10.1109/TVCG.2022.3209428

[24] Benny J Tang, Angie Boggust, and Arvind Satyanarayan. 2017. Extracting Top-K Insights from Multi-dimensional Data. In *ACM International Conference on Management of Data.* doi:10.1145/3035918.3035922

[25] Benny J Tang, Angie Boggust, and Arvind Satyanarayan. 2023. VisText: A Benchmark for Semantically Rich Chart Captioning. In *Proceedings of the Association for Computational Linguistics.* doi:10.18653/v1/2023.acl-long.401

[26] Yun Wang, Zhida Sun, Haidong Zhang, Weiwei Cui, Ke Xu, Xiaojuan Ma, and Dongmei Zhang. 2019. DataShot: Automatic Generation of Fact Sheets from Tabular Data. *IEEE Transactions on Visualization and Computer Graphics* (2019). doi:10.1109/TVCG.2019.2934398

[27] Luoxuan Weng, Xingbo Wang, Junyu Lu, Yingchaojie Feng, Yihan Liu, and Wei Chen. 2024. InsightLens: Discovering and Exploring Insights from Conversational Contexts in Large-Language-Model-Powered Data Analysis. *arXiv preprint* (2024). doi:10.48550/arXiv.2404.01644

[28] Guande Wu, Shunan Guo, Jane Hoffswell, Gromit Yeuk-Yin Chan, Ryan A Rossi, and Eunyee Koh. 2023. Socrates: Data Story Generation via Adaptive Machine-Guided Elicitation of User Feedback. *IEEE Transactions on Visualization and Computer Graphics* (2023). doi:10.1109/TVCG.2023.3327363

[29] Zhuohao Zhang, Sana Malik, Shunan Guo, Jane Hoffswell, Ryan Rossi, Fan Du, and Eunyee Koh. 2022. CODAS: Integrating Business Analytics and Report Authoring. *EuroVA, J. Bernard and M. Angelini, Eds* (2022). doi:10.2312/eurova.20221082