SAGE: Adaptive Recommendation of Spreadsheet Pivot Tables

Whanhee Cho
University of Utah
Salt Lake City, Utah, USA
whanhee.cho@utah.edu

Shamit Fatin
University of Utah
Salt Lake City, Utah, USA
shamit.f@utah.edu

Anna Fariha
University of Utah
Salt Lake City, Utah, USA
afariha@cs.utah.edu

Abstract

Spreadsheet data is ubiquitous, with billions of users relying on spreadsheets for everyday tasks such as financial analysis and decision making. To make sense of spreadsheet datasets and extract data insights, users rely on *data summaries*. *Pivot tables* enable summarization through aggregation—such as `SUM` or `MEAN`—over a few data attributes, grouped by other attributes. Nevertheless, manually searching for interesting pivot tables is tedious—especially for high-dimensional datasets—prompting commercial spreadsheet software to offer pivot table *recommendations*. However, existing approaches often provide poor-quality recommendations of pivot tables that are frequently redundant, semantically uninteresting, and difficult to interpret. Further, they typically lack *adaptiveness* and *customizability*, offering little to no control to the user.

We demonstrate SAGE—a data-semantics-aware system for recommending a *diverse set* of *insightful* and *interpretable* pivot tables. Under the hood, SAGE relies on (1) a novel data-semantics-aware model to quantify the utility of individual pivot tables and the diversity of a set of pivot tables, and (2) a scalable greedy algorithm that efficiently selects a set of high-utility, diverse pivot tables. Beyond improving the overall quality of the recommended pivot table set compared to existing approaches, SAGE offers two additional features: (1) it *adapts* based on user feedback, allowing users to provide positive or negative signals to steer recommendations toward their goals, and (2) it provides enhanced *control*, enabling users to specify preferences such as the number of recommendations, attributes of interest, and the desired degree of diversity.

Demo video link: https://whnhch.github.io/videos/sage.html

1 Introduction

Spreadsheet data is ubiquitous—supporting tasks that range from everyday data analysis to high-stakes decision-making in business, science, and public policy—with tools such as Microsoft Excel [7] and Google Sheets [4] used daily by billions of users [2]. Despite their popularity, effectively extracting *insights* from spreadsheets remains challenging, particularly for users with limited knowledge of the data domain. When users are unfamiliar with the data semantics, manually scanning thousands of tuples and hundreds of attributes to discover insights—such as patterns, trends, or anomalies—becomes tedious. A common and effective approach for extracting insights is *summarization*, such as aggregating data across meaningful groups or *pivots*. For example, rating trends in a restaurant dataset can be observed from average ratings across cities and cuisines (Figure 1). Grouping data by one or more attributes and applying aggregate functions (e.g., `SUM`, `AVG`, `COUNT`) transforms large spreadsheet datasets into concise, consumable summaries called *pivot tables*. However, in an exploratory setting where the user is looking for interesting data trends to discover insights, they must examine a very large number of pivot tables manually.

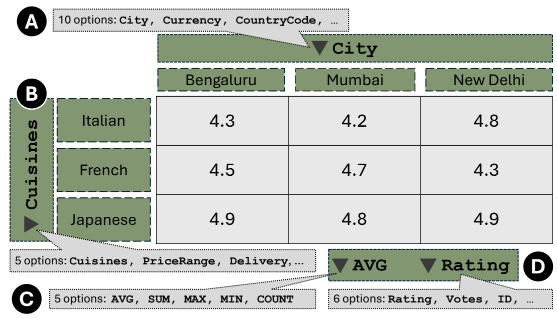

Figure 1: An example pivot table over the Zomato dataset [6] with 21 attributes. Users can specify four parameters: grouping attributes in Ⓐ and Ⓑ, aggregate functions in Ⓒ, and value attributes in Ⓓ.

Example 1.1. *Figure 1 illustrates a pivot table constructed over a restaurant dataset containing 21 attributes, including* `City`, `Currency`, `CountryCode`, `Cuisines`, `Price-Range`, `Delivery`, `Rating`, *etc. In this pivot table, the user selects* Ⓐ `City` *and* Ⓑ `Cuisines` *as the* grouping *attributes,* Ⓓ `Rating` *as the* value *attribute, and* Ⓒ `AVG` *as the aggregate, which is applied to* `Rating`. *Suppose we restrict the choices to* 10 *demographic attributes for* Ⓐ *and* 5 *restaurant-related attributes for* Ⓑ, *and constrain the pivot table to (i) two grouping attributes, (ii) one value attribute (chosen from* 6 *options), and (iii) one aggregate (chosen from* 5 *options). Even under these conservative assumptions, the number of possible pivot tables is* $10 \times 5 \times 6 \times 5 = 1{,}500$. *The combinatorial explosion becomes more severe when multiple grouping and value attributes are allowed. For instance, permitting up to* 4 *grouping attributes and* 2 *value attributes yields a search space of* $\binom{10+5}{4} \times \binom{6}{2} \times 5^2 = 511{,}875$ *possible pivot-table configurations, illustrating how quickly the space of possible pivot tables becomes intractable.*

As manually exploring this combinatorial space to identify useful pivot tables requires substantial effort, commercial spreadsheets provide pivot-table recommendation functionality [4, 7]. However, these systems suffer from several critical limitations: (L_1) they do not exhaustively explore the space of possible pivot tables by materializing and evaluating the candidate pivot tables over the data; instead, they largely rely on surface-level signals such as attribute names and data types; (L_2) they often include semantically invalid—e.g., `SUM(Longitude)`—or uninformative—e.g., `GROUP BY(CountryCode)`—components since they do not account for data semantics; (L_3) they provide top-k recommendations, which frequently exhibit redundancy due to repeated use of similar attribute-aggregation combinations; and (L_4) when relying on LLMs, they may include non-existent attributes due to LLM hallucinations and the lack of validation.

Figure 2 illustrates the limitations of three existing pivot-table recommendation systems: Microsoft Excel, Google Sheets, and ChatGPT. All of them frequently recommend pivot tables involving a

Tool	Recommended Pivot Tables
Google Sheets[1] [4]	(1) COUNT(ID) by City (2) COUNT(Name) by City (3) COUNT(ID) by Country Code
Microsoft Excel[2] [7]	(1) SUM(Longitude) by City (2) SUM(Latitude) by City (3) SUM(Votes), COUNT(ID), SUM(Longitude) by City, Currency (4) SUM(Rating), SUM(Latitude), SUM(Longitude) by Currency, Name (5) SUM(Votes), SUM(Latitude), SUM(Longitude) by Rating Color, Address (6) SUM(Longitude) by Rating Text, Address (7) SUM(Longitude) by Rating Color, Address (8) COUNT(ID), SUM(Longitude), SUM(Country Code) by Rating Text, Address (9) SUM(Country Code), SUM(Cost), SUM(Price Range) by Currency, Name (10) SUM(Cost), SUM(Price Range) by Currency, Name
ChatGPT[3] [8]	(1) AVG(Rating), COUNT(ID) by City (2) AVG(Rating), SUM(Votes) by Cuisine (3) COUNT(ID) by City, Price Range

[1] Google Sheets Results: www.github.com/whnhch/SAGE/blob/main/demo/software/gsheets.png
[2] Microsoft Excel Results: www.github.com/whnhch/SAGE/blob/main/demo/software/excel.pdf
[3] ChatGPT Conversation: www.chatgpt.com/share/69496394-e404-8006-89ca-bb1c7d743350

Figure 2: Existing pivot-table recommendation systems exhibit several limitations. Here, pink denotes semantic invalidity (L_2), purple denotes redundancy (L_3), and blue denotes hallucination (L_4). `Country Code` is cryptic due to containing only numeric identifiers; `SUM(Longitude)` is semantically invalid; and `Cuisine` (instead of `Cuisines`) is invalid, as it is not a data attribute.

repetitive set of attributes (L_3). While Google Sheets involves grouping by cryptic attributes such as `CountryCode`, Microsoft Excel frequently suggests meaningless aggregations such as `SUM(Longitude)` (L_2). LLM-based ChatGPT shows promise in understanding user intent; however, it produces invalid recommendations involving non-existent attributes such as `Cuisine` (L_4). In summary, all these tools suffer from L_1, as they neither perform an exhaustive search nor validate the recommendations against the actual data.

The need for customizability & adaptivity. So far, we have discussed the limitations of existing systems only in terms of their recommendation quality. However, an important requirement for recommendation systems is *user-centeredness*. Except for ChatGPT, current systems are static and non-adaptive, leading to a fifth limitation: (L_5) they cannot incorporate user feedback, do not support user preferences (e.g., the desired number of pivot tables or preferred attributes), and repeatedly recommend the same pivot tables within a session. We illustrate this in Example 1.2.

Example 1.2. *Minho is exploring the Zomato dataset of Example 1.1 and is particularly interested in the attributes* `Rating`, `Cuisines`, *and* `City`. *However, except for ChatGPT, existing tools do not allow him to specify these interests. Moreover, he has no control over the number of recommendations or the degree of diversity in the suggested pivot tables. Google Sheets displays only three recommendations by default, which Minho finds insufficient. In contrast, Microsoft Excel overwhelms him with a long list of redundant and often meaningless recommendations that disregard his preferences, with no way to filter or prioritize the recommendations. After reviewing the suggestions and ignoring irrelevant tables, Minho requests new recommendations, expecting the system to adapt to his feedback. However, both tools return the exact same recommendations, completely ignoring his actions.*

Example 1.2 demonstrates how static pivot-table recommendation systems fail to support user preferences—such as the desired number of recommendations or preferred attributes—resulting in a frustrating user experience. Even with interactive LLM-based systems, users must prompt the system to obtain updated suggestions rather than receiving them automatically. An ideal pivot-table recommendation system should support explicit user preferences, providing *customizability*, while automatically refining its recommendations based on user feedback, ensuring *adaptivity*.

SAGE. To this end, we have developed SAGE [1], a data-semantics-aware system for recommending a k-budgeted set of diverse pivot tables. SAGE overcomes the limitations of prior approaches by satisfying the following properties:

(P_1) SAGE uses an *efficient*, *scalable* greedy approach with optimizations that handle large, high-dimensional data effectively while preserving correctness (addressing L_1 and L_4).
(P_2) Each pivot table recommended by SAGE provides *insightful* and *semantically valid* information in an *interpretable* manner to ensure ease of comprehension by humans (addresses L_2).
(P_3) SAGE recommends a *diverse*, minimally redundant set of pivot tables covering various data aspects (addresses L_3).
(P_4) SAGE provides *customizability*—letting users set the recommendation size, degree of diversity, and attributes of interest—and *adaptivity*—automatically filtering tables deemed uninteresting based on prior interactions (addresses L_5).

Related Work. Insight recommendations [3, 10] usually recommend interesting insights such as deviation or correlation scores. However, they lack the ability to reveal semantically valid insights. While QuickInsights [3] considered semantic insightfulness, it only considers distributional importance rather than semantic surprisingness in the real-world context. OLAP exploration systems [5, 9] help users explore the data cubes interactively to find surprising patterns. However, these systems focus on recommending top-ranked patterns rather than producing a diverse set of patterns. Deep learning-based recommendations [11] leverage useful possible queries to suggest relevant data summarizations. However, these works emphasize recommending top-ranked interesting insights without considering data semantics or diversity. SAGE addresses the limitations of prior work by providing semantically diverse pivot table recommendations.

Demonstration. In our demonstration, participants will observe SAGE's capability to generate diverse, semantically insightful, and interpretable recommendations, with particular emphasis on *adaptivity* and *customizability*. Using real-world datasets, we will showcase not only the high-quality recommendations provided by SAGE, but also how it offers users significantly greater control and customization opportunities compared to existing tools. Participants will experience firsthand how SAGE intelligently adapts to user feedback, continuously refining its recommendations. We provide an overview of SAGE in Section 2 and a detailed walkthrough of our demonstration scenario in Section 3.

2 System Overview

Given a dataset D, a budget k, attributes of interest $\mathcal{A}$, and a diversity threshold θ, SAGE aims to select a set $\mathbf{T}^*$ of k pivot tables from the candidate space $\mathcal{T}_{\mathbf{A}}$ such that (i) each selected pivot table $T \in \mathbf{T}^*$ exhibits high utility, and (ii) the set $\mathbf{T}^*$ satisfies the diversity requirement θ. Naïvely enumerating all k-sized subsets $\mathbf{T} \subset \mathcal{T}_{\mathbf{A}}$ to find the best solution is computationally infeasible, rendering the problem NP-hard. Moreover, achieving this objective requires *data-semantics-aware* measures for both the utility of individual pivot tables and the diversity of a set of pivot tables. These requirements give rise to three key challenges: (1) how to model the *utility* of a single pivot table; (2) how to quantify similarity between pairs of pivot tables to enforce *diversity*; and (3) how to design an *efficient* and *scalable* recommendation algorithm without an exhaustive search. We describe how SAGE addresses these challenges next.

Utility. A pivot table has high utility if it offers *insights* while being easily *interpretable* by humans. To compute the overall utility score of a pivot table T, we linearly combine these two components as follows, where $\alpha \in [0, 1]$ can be adjusted based on user preferences:

$$\textit{Utility(T)} = \alpha \cdot \textit{Insightfulness(T)} + (1 - \alpha) \cdot \textit{Interpretability(T)}$$

Insightfulness. To measure insightfulness of T, SAGE relies on the following four components:[1]

— Interestingness of the attributes involved in a pivot table, $S_{\text{int}}(T)$: To identify interesting attributes, SAGE leverages an LLM to obtain semantic knowledge about the dataset. For example, `City` is a more interesting attribute than `Address` to form groups.

— Informativeness of the table values, $S_{\text{inf}}(T)$: SAGE measure how informative a pivot table is by how much its values differ across groups: larger differences indicate higher informativeness.

— Trends, $S_{\text{trend}}(T)$: Trends observed in a pivot table provide insights. However, the degree of insightfulness hinges on two key factors: the magnitude of the trend metric and how atypical or rare it is. For instance, a positive correlation between income and years of service is generally expected, while a trend showing that new hires earn more on average than long-serving employees contradicts this expectation and thus is particularly insightful. SAGE uses two metrics to quantify the magnitude of a trend: correlation and ratio. Furthermore, to assess the degree of a trend's rarity, it queries an LLM, which provides a broader semantic context.

— Surprisingness, $S_{\text{sur}}(T)$: While surprising values or outliers potentially indicate insights, they need to be validated in a real-world context. For example, high price ranges in touristy cities are expected and therefore not particularly surprising. To this end, SAGE leverages semantic knowledge, obtained from an LLM, to validate genuinely surprising insights.

Intuitively, a pivot table is insightful if it exhibits *any* of the characteristics: informativeness, trend, or surprise, over interesting attributes. Thus,

$$\textit{Insightfulness(T)} = S_{\text{int}}(T) \cdot \max\left(S_{\text{inf}}(T),\ S_{\text{trend}}(T),\ S_{\text{sur}}(T)\right)$$

Interpretability. Interpretability models common-sense reasoning—such as avoiding aggregations like `SUM(Longitude)`—while taking into account the cognitive constraints of humans. SAGE computes interpretability of a pivot table T based on three components:[1]

— Density, S_{den} (T): Excessive nulls in a pivot table hinder interpretability, as humans struggle to draw insights from sparse tables. This motivates high density—i.e., small proportion of null values within a pivot table—as a key interpretability criteria.

— Semantic validity of headers and aggregates, S_{sem} (T): Row and column headers in a pivot table must be semantically meaningful to ensure interpretability. For example, `Country` with values {India, China, Sri Lanka} is more interpretable than `CountryCode` with values {1, 2, 3}. Furthermore, certain aggregate-attribute combinations are nonsensical, such as `SUM(Longitude)`. SAGE consults an LLM to validate whether an aggregate function "makes sense" for the corresponding value attribute to form a meaningful aggregate.

— Conciseness, S_{con} (T): Intuitively, too many cells in a large pivot table reduce comprehensibility. Thus, the third criteria is conciseness, modeled by the size of a pivot table in number of cells.

While insightfulness relies on a strong signal from *any* type of insight, interpretability demands that *all* criteria be met. Thus, interpretability of T is the average of the three components:

$$\textit{Interpretability(T)} = \left(S_{\text{den}}(T) + S_{\text{sem}}(T) + S_{\text{con}}(T)\right)/3$$

Diversity. While utility measures the quality of an individual pivot table, *diversity* captures how effectively a *set* of pivot tables, *collectively*, provides complementary and non-redundant data perspectives. SAGE models two complementary notions of diversity: (i) *query diversity*, which favors semantically distinct combinations of grouping and aggregation attributes, and (ii) *content diversity*, which captures variation in statistical patterns and values across pivot tables. To jointly encode these notions, SAGE represents each pivot table by concatenating its query and content embeddings into a single semantics-preserving vector. We obtain query embeddings using T5, a natural-language encoder fine-tuned on Spider, a text-to-SQL benchmark dataset; while we compute content embeddings using TAPEX, a pre-trained encoder trained on sentence-table pairs.[1]

SAGE controls the desired level of diversity by a user-specified threshold θ, which enforces a minimum allowable distance between any pair of pivot tables in the recommended set. Formally,

$$\theta \leq \text{Diversity}(\mathbf{T}) = \min_{T_i, T_j \in \mathbf{T},\ i<j} \text{dist}(T_i, T_j).$$

Greedy Algorithm. SAGE employs a greedy algorithm with several optimizations that iteratively selects pivot tables in order of decreasing utility scores while ensuring that the diversity constraint is satisfied. In our experiments over four real-world datasets, this approach achieved over 99% reduction in runtime with less than 3% loss in utility [1]. We briefly discuss the key optimizations below:

— *Pruning*. SAGE leverages $S_{\text{int}}(T)$, S_{sem} (T), and S_{con} (T) for early pruning of low-utility candidates, as these scores do not require materialization.

— *LLM Proxy.* Since querying LLMs is time-consuming, we train a lightweight decision-tree—trained on 10K LLM prompt-response pairs—to approximate LLM behavior. The model achieves average accuracies of 90%, 89%, and 64% for correlation, ratio, and surprise scores, respectively, while substantially reducing runtime [1].

— *Sampling and parallelization.* To further enhance practical runtime, SAGE leverages sampling and approximation techniques for score estimation, along with parallel computing.

Customizability and Adaptivity. SAGE allows users to specify the desired number of pivot tables (k), a diversity constraint defined as the minimum allowable distance between pivot tables (θ), and a set of attributes of interest ($\mathcal{A}$). Furthermore, SAGE allows users to express feedback by "liking" or "disliking" the recommended pivot tables. In subsequent iterations, SAGE *adapts* in response to this feedback by favoring pivot tables similar to those the user liked while avoiding those similar to the ones the user disliked.

[1]Due to space constraints, we omit the detailed computations in this demo paper; these can be found in our full paper [1] (accepted to SIGMOD 2026).

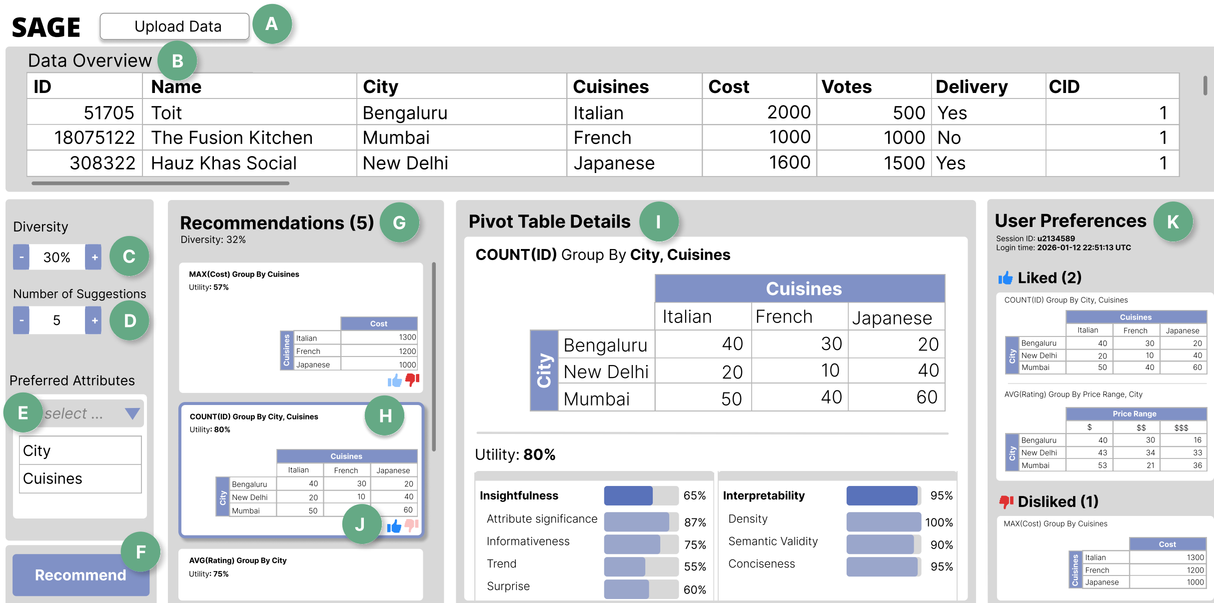

Figure 3: The SAGE interface: Ⓐ upload data, Ⓑ preview data, Ⓒ set up diversity threshold, Ⓓ set up the number of pivot tables, Ⓔ select preferred attributes, Ⓕ request recommendations, Ⓖ SAGE returns recommendations, Ⓗ select a pivot table, Ⓘ preview details of a pivot table, Ⓙ user likes the pivot table. Ⓚ SAGE saves user preferences.

3 Demonstration Scenario

We will demonstrate SAGE using a real-world restaurant dataset Zomato [6] containing 21 attributes including `Cuisines`, `City`, `Cost`, etc. We will guide the users through eleven steps (annotated in Figure 3) impersonating Minho, who is interested in identifying promising cities and cuisines for starting a new restaurant business.

Steps Ⓐ & Ⓑ (Uploading and previewing data) In Ⓐ, the user uploads the data they want to investigate. In this scenario, it is a restaurant dataset. In Ⓑ, the user previews the dataset and can scroll horizontally and vertically to explore it.

Steps Ⓒ, Ⓓ, & Ⓔ (Specifying user preferences) The user sets the minimum distance between pivot tables in the recommendations in Ⓒ. Here, they set it to 30%. Then in Ⓓ, they select the number of desired recommendations using the stepper buttons, setting it to 5. In Ⓔ, they select `City` and `Cuisines` as preferred attributes, hinting SAGE to prioritize these attributes.

Steps Ⓕ & Ⓖ (Requesting recommendations) The user clicks "Recommend" in Ⓕ and SAGE recommends 5 pivot tables in Ⓖ, based on their preferences. The first three tables are `MAX(Cost) Group By Cuisines`; `COUNT(ID) Group By City, Cuisines`; and `AVG(Rating) Group By City`, each with a utility score. Unlike existing tools in Figure 2, SAGE produces results aligned with the user's preferences and includes diverse aggregate functions (`MAX`, `COUNT`, & `AVG`) and attributes (`Cost`, `ID`, & `Rating`).

Steps Ⓗ & Ⓘ (Viewing a pivot table) In Ⓗ, the user selects the second recommendation to see more details in Ⓘ. This table has a utility score of 80% and includes their preferred attributes, `City` and `Cuisines`. The user can also observe the detailed breakdown of *Insightfulness* (65%) and *Interpretability* (95%) components here.

Step Ⓙ (Liking a pivot Table) After reviewing the details, the user indicates that they accept this recommendation by clicking.

Step Ⓚ (Adapting to User Feedback) In Ⓚ, the user reviews their feedback history, consisting of two liked and one disliked pivot tables. SAGE adapts its future recommendations by prioritizing pivot tables similar to those the user liked and avoiding ones resembling the disliked ones. We will demonstrate this adaptive behavior through multiple recommendation iterations.

Beyond the restaurant dataset, we will demonstrate SAGE on real-world datasets—spanning diverse domains such as marketing, real estate, and environment—showcasing its capability to recommend a diverse set of insightful and interpretable pivot tables. The key takeaway from our demo is that users can customize their preferences and provide feedback on the recommendations, while SAGE intelligently adapts to incorporate their feedback.

References

[1] Whanhee Cho and Anna Fariha. 2026. Data-Semantics-Aware Recommendation of Diverse Pivot Tables. *Proc. ACM Manag. Data* 4, 1 (2026).
[2] Clay Team. 2021. *The Many Lives of Spreadsheets*. Clay. www.clay.com/blog/the-many-lives-of-spreadsheets Accessed: 2026-01-08.
[3] Rui Ding, Shi Han, Yong Xu, Haidong Zhang, and Dongmei Zhang. 2019. QuickInsights: Quick and Automatic Discovery of Insights from Multi-Dimensional Data. In *SIGMOD*.
[4] Google Sheets. 2025. Online Spreadsheet Editor. sheets.google.com/
[5] Manas Joglekar, Hector Garcia-Molina, and Aditya G. Parameswaran. 2019. Interactive Data Exploration with Smart Drill-Down. *IEEE TKDE* 31, 1 (2019).
[6] Shruti Mehta. 2025. Zomato Restaurants Data. www.kaggle.com/datasets/shrutimehta/zomato-restaurants-data.
[7] Microsoft Excel. 2025. www.microsoft.com/en-us/microsoft-365/excel
[8] OpenAI. 2025. ChatGPT. chat.openai.com/
[9] Sunita Sarawagi. 2000. User-Adaptive Exploration of Multidimensional Data. In *VLDB*.
[10] Manasi Vartak, Sajjadur Rahman, Samuel Madden, Aditya G. Parameswaran, and Neoklis Polyzotis. 2015. SEEDB: Efficient Data-Driven Visualization Recommendations to Support Visual Analytics. *PVLDB* 8, 13 (2015).
[11] Junjie Xing, Xinyu Wang, and H. V. Jagadish. 2024. Data-Driven Insight Synthesis for Multi-Dimensional Data. *PVLDB* 17, 5 (2024).