SHORT-PAPER

CauSumX: Summarized Causal Explanations For Group-By-Average Queries

NATIV LEVY, Technion - Israel Institute of Technology, Haifa, Israel

MICHAEL JOHN CAFARELLA, Massachusetts Institute of Technology, Cambridge, MA, United States

AMIR GILAD, Hebrew University of Jerusalem, Jerusalem, Israel

SUDEEPA ROY, Duke University, Durham, NC, United States

BRIT YOUNGMANN, Technion - Israel Institute of Technology, Haifa, Israel

Open Access Support provided by:

Hebrew University of Jerusalem

Technion - Israel Institute of Technology

Duke University

Massachusetts Institute of Technology

Published: 22 June 2025

Citation in BibTeX format

SIGMOD/PODS '25: International Conference on Management of Data
June 22 - 27, 2025
Berlin, Germany

Conference Sponsors:
SIGMOD

https://doi.org/10.1145/3722212.3725088
ISBN: 9798400715648

CauSumX: Summarized Causal Explanations For Group-By-Average Queries

Nativ Levy
nativlevy@campus.technion.ac.il
Technion
Haifa, Israel

Michael Cafarella
michjc@csail.mit.edu
CSAIL, MIT
Boston, USA

Amir Gilad
amirg@cs.huji.ac.il
Hebrew University
Jerusalem, Israel

Sudeepa Roy
sudeepa@cs.duke.edu
Duke University
Durham, USA

Brit Youngmann
brity@technion.ac.il
Technion
Haifa, Israel

Abstract

Group-by-average SQL queries are a cornerstone of data analysis, often employed to uncover patterns and trends within datasets. However, interpreting the results of these queries can be challenging and time-intensive, particularly when working with large, high-dimensional datasets. Automating the generation of explanations for such queries can greatly enhance analysts' ability to derive meaningful insights while reducing human effort. Effective explanations must balance succinctness and depth, offering insights into different patterns across aggregate results, while crucially reflecting cause-effect relationships rather than mere correlations. This ensures that users can make informed, data-driven decisions grounded in reality. In this demonstration, we present CauSumX, a system that produces concise and causal explanations for group-by-average queries. Leveraging background causal knowledge, CauSumX identifies the key causal factors driving variations in the outcome variable across different groups. The system employs an efficient algorithm based on a recently published paper. We will demonstrate the utility of CauSumX for generating useful summarized causal explanations by interacting with the SIGMOD'25 participants, who will act as data analysts aiming to explain their query results.

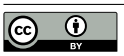
CCS Concepts

• **Information systems** → *Data management systems*; *Relational database model*; *Data mining*.

Keywords

Causal Inference, SQL, Query Result Explanation

ACM Reference Format:
Nativ Levy, Michael Cafarella, Amir Gilad, Sudeepa Roy, and Brit Youngmann. 2025. CauSumX: Summarized Causal Explanations For Group-By-Average Queries. In *Companion of the 2025 International Conference on Managementof Data (SIGMOD-Companion '25), June 22–27, 2025, Berlin, Germany.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3722212.3725088



SIGMOD-Companion '25, Berlin, Germany

ACM ISBN 979-8-4007-1564-8/2025/06
https://doi.org/10.1145/3722212.3725088

1 Introduction

Group-by-average SQL queries are frequently used in data analysis applications and are often represented as bar charts. These queries provide an *aggregate view* on the data, such as the average salary by occupation or gender or determining the average product satisfaction rate. However, interpreting the results of such can be both complex and time-consuming, particularly when dealing with large or high-dimensional datasets. Moreover, understanding the *causal factors* driving variations in averages across groups is crucial for making informed decisions. For example, understanding the causal reasons behind the lower average severity rate of car accidents in certain regions of the United States can guide policymakers in implementing effective corrective measures—insights that non-causal associations alone cannot provide. To illustrate, consider the following example.

Example 1.1. Consider the Stack Overflow developer survey [2], which provides comprehensive information about high-tech employees worldwide. It includes numerous demographic features about individuals, such as gender and age, and annual salary. Table 1 shows a subset of the dataset. Consider the following query measuring the average salary by country:

```
SELECT Country, AVG(Salary)
FROM Stack-Overflow
GROUP BY Country
```

In the lower part of Fig. 1, the results are displayed as a bar chart marked as part B (the colors will be explained later). There is a significant disparity in average salary across different countries. A useful explanation should highlight both the primary factors contributing to this variation, and within each country, the factors influencing the salary. However, the dataset is too big regarding the number of tuples (over 380k tuples) and attributes (20 attributes) to look for a succinct and informative explanation by manual inspection. While tools like Tableau give highly sophisticated visualizations by slicing and dicing the data across several dimensions, they return the aggregates for these dimensions and do not differentiate between causal and non-causal reasons.

Explaining the outcomes of group-by SQL queries has been thoroughly explored in the literature [14, 18]. While many approaches provide explanations in the form of *predicates*, which are easily understandable, they fall short of offering a concise explanation for the entire aggregate view. Moreover, although these explanations

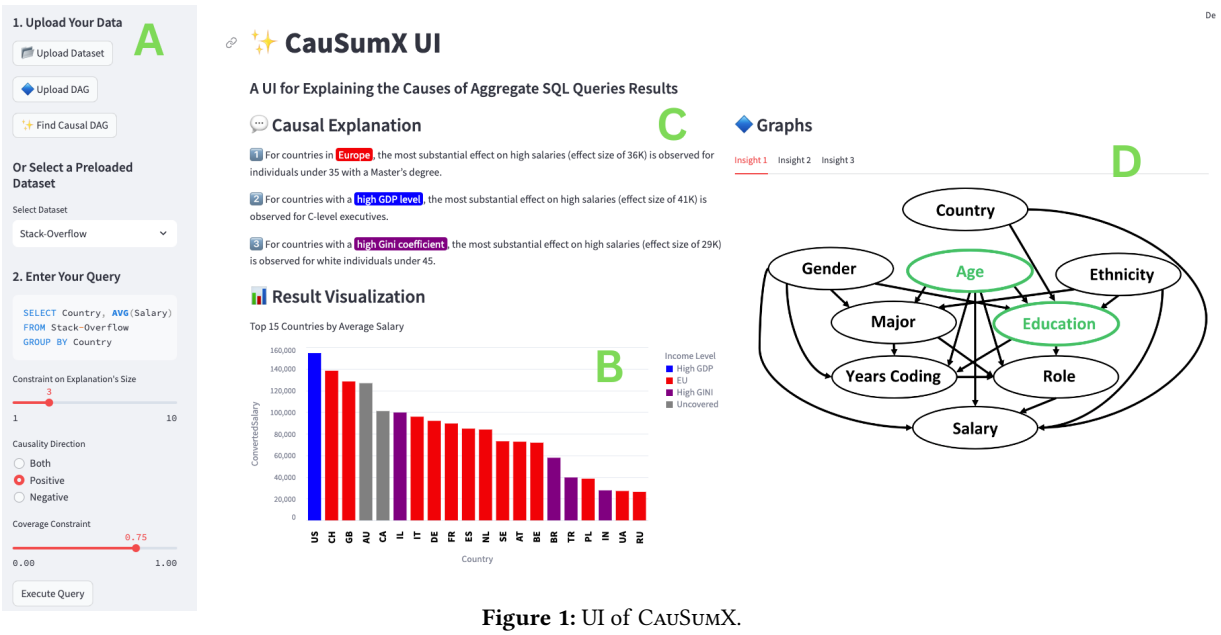


Figure 1: UI of CauSumX.

Table 1: A subset of the Stack Overflow dataset.

ID	Count.	Cont.	Gender	Age	Role	Edu.	Salary
1	US	NA	Male	26	Data Scientist	PhD	180k
2	US	NA	NB	32	QA developer	B.Sc.	83k
3	India	AS	Male	29	CEO	B.Sc.	24k
4	India	AS	Female	25	Backend dev.	M.S.	7.5k
5	China	AS	Male	21	Backend dev.	B.Sc.	19k

unveil various interesting insights, they are not causal. *Causal inference* has been studied in AI and Statistics extensively. It enables analysts to assess the impact of a *treatment* variable on an *outcome* and has been used in decision-making in numerous fields including social science, economics, biology, and medicine.

In this work, we demonstrate CauSumX, a system that generates *a summarized causal explanation for the results of a group-by-average query*. The explanation summary consists of a set of *explanation patterns* (predicates). Each explanation pattern is defined by a *grouping pattern* capturing a subset of the results of the query answer, and a *treatment pattern*, representing the factor with the most substantial causal effect on the outcome (average attribute). CauSumX identifies a set of explanation patterns that collectively capture the entire aggregate view. CauSumX further enables analysts to balance between the summary size and the coverage of the aggregate view.

Example 1.2. Continuing our example, the analyst uses CauSumX to explain her query results while limiting the number of insights to three. The explanation summary is displayed in Fig. 1. The mapping between countries and insights is visualized using the bars' color. Without having to manually explore the dataset, the analyst learns the main reasons for high and low salaries in different countries. These reasons are not just predicates summarizing tuples in the data, they have strong *causal effects*. Moreover, the analyst knows that these explanations not only hold for one country but also hold for several countries that share the same grouping pattern (color).

Related Work. Explaining SQL query results is an extensively studied problem. Provenance-based solutions show how the output was computed using the input tuples [4]. However, an aggregate answer over a large dataset uses many input tuples, hence several approaches have focused on providing high-level explanations as *predicates* on input tuples [9, 14, 18], which are easy to comprehend. While these approaches reveal interesting insights, such as outliers [18], explaining high/low values [9], or comparisons of a set of answers [14], they do not provide a summarized explanation for the entire view, and they are not causal.

Recent works have provided explanations for group-by-avg queries using causal analysis, focusing on revealing unobserved factors influencing the results [15, 20]. However, these methods provide a single explanation of the entire view, and do not offer fine-grained explanations for subgroups. The framework presented in [10] identifies causal predicates explaining differences in two average outcomes. However, they do not search for treatments affecting the outcome and do not give a summarized explanation of the entire view. In contrast, *data summarization* techniques offer methods to summarize large datasets considering factors such as diversity and coverage [7, 21], however these summaries are not causal.

2 Technical Background

2.1 Background on Causal Inference

We use Pearl's model for *observational causal analysis* [13]. The broad goal of causal inference is to estimate the effect of a *treatment variable* T (e.g., `Education`) on an outcome Y (e.g., `Salary`). One popular measure of causal estimate is *Average Treatment Effect* (ATE), defined as the difference in the average outcomes of the treated and control groups:

$$ATE(T, Y) = \mathbb{E}_Z \left[\mathbb{E}[Y \mid T = 1, Z = z] - \mathbb{E}[Y \mid T = 0, Z = z] \right] \quad (1)$$

Since ATE is computed over observational data, the treatment and control groups may not be assigned randomly. Therefore, to mitigate the effect of *confounding factors* (i.e., attributes that can affect the treatment assignment and outcome), we must control for *confounding variables* [13] (Z in Eq. (1)). A sufficient set of confounders can be determined by applying graphical criteria [13], which can be evaluated against a *causal DAG*.

A causal DAG represents direct causal relationships between variables in a given dataset [13]. It can be constructed by a domain

expert, or by using existing causal discovery algorithms [17]. For example, Fig. 1 shows a partial causal DAG for the attributes of the SO dataset. It represents the fact that the `Role` of a coder depends on the values of her `Education` and `Age` attributes.

In CauSumX, where the treatment with maximum effect may vary among different groups in the query answer, we are interested in computing the *Conditional Average Treatment Effect* (CATE), which measures the effect of a treatment on an outcome within *a subpopulation of interest*. Given a subpopulation defined by a predicate $B = b$, we compute $CATE(T, Y \mid B = b)$ by adding this predicate to the conditioning sets in Eq. (1).

2.2 Problem Formulation

We consider a database associated with a causal DAG. Given a group-by-avg query Q[1] and a database D, the result of evaluating Q over D is denoted by $Q(D)$. Restricting to AVG is fundamental for causal explanations [15, 20], unlike non-causal methods, as causal effects estimate the expected difference between two groups.

A *pattern* [14] is a conjunction of predicates (attribute-value assignments). An **explanation pattern** consist of pairs of patterns: (i) A **grouping pattern** $\mathcal{P}_g$ captures a subset of groups in $Q(D)$, e.g., a subset of counties in the same continent. A group in $Q(D)$) is either *covered* or not by $\mathcal{P}_g$. Thus, $\mathcal{P}_g$ only contains attributes having a functional dependency with the grouping attribute(s) of Q. (ii) A **treatment pattern**, $\mathcal{P}_t$, is defined over D and partitions the input tuples into treated (if $\mathcal{P}_t$ evaluates to true) and a control group (otherwise). Intuitively, the grouping pattern specifies the subpopulation of interest (equivalent to the condition B=b for CATE), while the treatment pattern explains the outcome Y within that subpopulation as per the CATE value.

To evaluate the effectiveness of an explanation pattern $(\mathcal{P}_g, \mathcal{P}_t)$, we define its *explainability* by estimating the causal effect (as per CATE) of $\mathcal{P}_t$ on the outcome variable of Q, within the subpopulation defined by $\mathcal{P}_g$. Since CATE can be either positive or negative, the explainability is defined as the absolute CATE value.

We emphasize that each selected explanation pattern represents an explanation that is statistically significant. Specifically, based on causal analysis, the expected explainability reflects the anticipated average increase in the outcome for the specific subpopulation to which the explanation applies.

Example 2.1. With $\mathcal{P}_g$: (`Cont.=EU`) and $\mathcal{P}_t$: (`Edu. = M.S`), the treatment group comprises individuals with a Master's degree from European countries, while the control group consists of individuals without a Master's degree from European countries. The explainability of $(\mathcal{P}_g, \mathcal{P}_t)$=36$K$, indicating that, on average, individuals from European countries with a Master's degree earn 36k more than those without a Master's degree from European countries.

Given an outcome variable, a *positive explanation pattern* explains what increases the outcome, while a *negative explanation pattern* explains what decreases it. CauSumX's UI lets analysts choose to view either one or both types of explanations.

Problem Definition: We aim to obtain a succinct yet comprehensive set of explanation patterns for the groups in $Q(D)$ from the huge search space of possible explanation patterns. To achieve this, we frame a constrained optimization problem: **size constraint**: the number of explanations should be at most k; **coverage constraint**: the number of groups in $Q(D)$ explained (i.e., covered) by the patterns must be at least a θ-fraction of all the groups in $Q(D)$; **redundancy constraint**: an explanation pattern should not explain the same groups explained by another pattern; **objective**: Our objective is to find the set of explanation patterns (referred to as the explanation summary) that abide by these constraints and whose overall explainability is maximized.

Example 2.2. Fig. 1 depicts an explanation summary for our running example query, where k=3 and θ=0.75, i.e., we aim to find a set of at most 3 explanation patterns that reveal what affects salary for at least 75% of the countries in the SO dataset.

2.3 Algorithms

A naive approach considers all grouping and treatment patterns and results in long runtimes. Instead, CauSumX employs an efficient three steps algorithm, described as follows:

Mining Grouping Patterns Considering every possible grouping pattern is infeasible as their number is exponential in the dataset size. Instead, CauSumX utilizes the Apriori algorithm [3] to generate frequent candidate grouping patterns. We use Apriori to extract patterns based on attributes with an FD to the grouping attribute in Q. A hash table is then employed to consider only grouping patterns defining distinct groups in $Q(D)$.

Mining Treatment Patterns Our next goal is to identify a treatment pattern $\mathcal{P}_t$ with a high causal effect on the outcome for each grouping pattern $\mathcal{P}_g$. Since the number of potential treatment patterns for $\mathcal{P}_g$ is large, CauSumX employs a greedy approach to assess the CATE only for *promising* treatment patterns, guided by lattice traversal [5]. The primary distinction from existing solutions (e.g., [5]) lies in the non-monotonic nature of CATE, where adding a predicate can change its direction. All treatment patterns form a lattice, with nodes representing patterns and edges indicating derivations by adding a single predicate. W.L.O.G., assume we are searching for a positive treatment. We traverse this lattice top-down. Nodes are materialized only if all their parents have positive CATE. This ensures that we account for most treatments with a positive CATE. We apply several optimizations to enhance runtime, including sampling. parallelism and pruning, as detailed in [19].

Obtaining an Explanation Summary. Given a collection of the mined explanation patterns with their explainability scores, an integer k, and a threshold θ, we formalize the problem of finding an explanation summary via Integer Linear Programming (ILP), by extending the ILP for the max-k-cover problem. We then employ the standard randomized rounding algorithm for max-k-cover to relax the ILP to LP and use an LP solver to find a solution.

3 System & Demonstration

We implemented CauSumX using Python and Streamlit[2] for the UI. To generate explanations in natural language, we used GPT-4 [12]. We used the DoWhy [16] package to compute causal effects, and the z3-solver[3] to solve the LP.[4]

[1] possibly with multiple grouping attributes

[2] https://streamlit.io/

[3] https://pypi.org/project/z3-solver/

[4] Our code repository is available at https://github.com/TechnionTDK/causumx.

Table 2: Datasets for the Demonstration.

Dataset	# of tuples	# of attributes	# of grouping patterns
German [6]	1000	20	10
Adult [1]	32.5K	13	13
SO [2]	38K	20	75
IMPUS-CPS [8]	1.1M	10	9
Accidents [11]	2.8M	40	15

System overview: CauSumX features a UI (shown in Fig. 1) and three key components. The analyst inputs a dataset D and specifies a query Q. If no causal DAG is provided, CauSumX employs the PC causal discovery algorithm [17] to obtain one (using the DoWhy implementation). CauSumX initially mines frequent candidate grouping patterns. Then, for each mined pattern, it identifies a treatment pattern with high explainability. It further generates an explanation summary using an LP solver. Finally, the analyst is displayed with an explanation summary for her query results.

3.1 Demo Scenario

We demonstrate the operation of CauSumX over multiple public datasets (statistics are given in Table 2). The causal DAGs were constructed by relying on prior work [19].

The SIGMOD participants will act as data analysts, seeking to explain the results of their group-by-average queries.

Guided tour of CauSumX: Attendees can choose the dataset they wish to explore and load a corresponding causal DAG (marked as A in Fig. 1). We will then explain the function of the different settings for the explanations, shown on the bottom left-hand side in Fig. 1, setting them to initial values in this part. To illustrate the usability of CauSumX, the audience will investigate queries inspired by real-world sources, such as the Stack Overflow annual report, media websites, and academic papers. For each dataset, CauSumX will present an example query. By clicking on the Execute Query button, CauSumX visualizes the query results (marked as B in Fig. 1), and presents the obtained explanation summary. The colors of each bar link groups in the query results (Countries) to their relevant explanation (marked by C in Fig. 1). CauSumX also displays, for each insight (i.e., explanation pattern), the relevant part in the input causal DAG, highlighting variables with negative causal effects in red (omitted from the presentation as we focus on positive explanation only) and those with positive effects in green, i.e., `Age` and `Education` (marked in D in Fig. 1). The audience will then explore how changing the size and coverage constraints affect the balance between detailed explanations and concise summaries.

Manual exploration of CauSumX: The participants would be invited to insert their own queries using the system UI. They could then inspect the generated explanation summaries and investigate them by exploring the relevant parts in the underlying causal DAG. This scenario simulates a common real-life task where a data analyst tries to explain the results of her query. We will further allow users to change the causal DAG by either uploading a pre-prepared DAG designed by experts or using an automatic algorithm to discover the causal DAG. Users could then observe the changes in the DAG and the obtained explanations to examine the effect of the causal DAG on the outputted explanations.

Looking under the hood: Interested participants will be invited to examine how various parameters affect quality and performance. For example, we will run the naïve approach that considers all explanation patterns, and show that while the explanation summary generated by CauSumX is similar to that of the naïve approach, CauSumX is at least one order of magnitude faster than this baseline. Further, we will show users the GPT prompt used to generate Natural Language explanations and allow them to select from several other prompts to get different forms of explanations.

Acknowledgments

This work was funded by the NSF awards IIS-2008107 and IIS-2147061, the ISF under grant 1702/24, the Scharf-Ullman Endowment, and the Alon Scholarship. We also appreciate the support from DARPA ASKEM Award HR00112220042, the ARPA-H Biomedical Data Fabric project, and Liberty Mutual.

References

[1] 2016. Adult Census Income Dataset. https://www.kaggle.com/datasets/uciml/adult-census-income. Accessed: 2024-04-04.
[2] 2021. Stackoverflow Developer Survey. https://tinyurl.com/4s298pd2.
[3] Rakesh Agrawal, Ramakrishnan Srikant, et al. 1994. Fast algorithms for mining association rules. In *Proc., VLDB*, Vol. 1215. Santiago, Chile, 487–499.
[4] Yael Amsterdamer, Daniel Deutch, and Val Tannen. 2011. Provenance for aggregate queries. In *PODS*, Maurizio Lenzerini and Thomas Schwentick (Eds.). ACM, 153–164. doi:10.1145/1989284.1989302
[5] Abolfazl Asudeh, Zhongjun Jin, and HV Jagadish. 2019. Assessing and remedying coverage for a given dataset. In *ICDE*. IEEE, 554–565.
[6] Arthur Asuncion and David Newman. 2007. UCI machine learning repository.
[7] Kareem El Gebaly, Parag Agrawal, Lukasz Golab, Flip Korn, and Divesh Srivastava. 2014. Interpretable and informative explanations of outcomes. *Proceedings of the VLDB Endowment* 8, 1 (2014), 61–72.
[8] Sarah Flood, Miriam King, Steven Ruggles, and J Robert Warren. 2015. Integrated public use microdata series. *University of Minnesota* 1 (2015).
[9] Chenjie Li, Zhengjie Miao, Qitian Zeng, Boris Glavic, and Sudeepa Roy. 2021. Putting Things into Context: Rich Explanations for Query Answers using Join Graphs. In *SIGMOD*.
[10] Pingchuan Ma, Rui Ding, Shuai Wang, Shi Han, and Dongmei Zhang. 2023. XInsight: EXplainable Data Analysis Through The Lens of Causality. *Proc. ACM Manag. Data*, Article 156 (jun 2023), 27 pages.
[11] Sobhan Moosavi, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. 2019. Accident risk prediction based on heterogeneous sparse data: New dataset and insights. In *SIGSPATIAL*. 33–42.
[12] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]
[13] Judea Pearl. 2009. *Causality.* Cambridge university press.
[14] Sudeepa Roy and Dan Suciu. 2014. A formal approach to finding explanations for database queries. In *SIGMOD*.
[15] Babak Salimi, Johannes Gehrke, and Dan Suciu. 2018. Bias in olap queries: Detection, explanation, and removal. In *SIGMOD*. 1021–1035.
[16] Amit Sharma and Emre Kiciman. 2020. DoWhy: An End-to-End Library for Causal Inference. *arXiv preprint arXiv:2011.04216* (2020).
[17] P. Spirtes et al. 2000. *Causation, prediction, and search.* MIT press.
[18] Eugene Wu and Samuel Madden. 2013. Scorpion: Explaining away outliers in aggregate queries. *Proc. VLDB* (2013).
[19] Brit Youngmann, Michael Cafarella, Amir Gilad, and Sudeepa Roy. 2024. Summarized Causal Explanations For Aggregate Views. *Proc. ACM Manag. Data* 2, 1 (2024). doi:10.1145/3639328
[20] Brit Youngmann, Michael Cafarella, Yuval Moskovitch, and Babak Salimi. 2023. On Explaining Confounding Bias. *ICDE* (2023).
[21] Cong Yu, Laks V. S. Lakshmanan, and Sihem Amer-Yahia. 2009. It takes variety to make a world: diversification in recommender systems. In *EDBT*, Martin L. Kersten, Boris Novikov, Jens Teubner, Vladimir Polutin, and Stefan Manegold (Eds.), Vol. 360. ACM, 368–378.