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

Nativ Levy
Technion
nativlevy@campus.technion.ac.il

Michael Cafarella
MIT
michjc@csail.mit.edu

Amir Gilad
Hebrew University
amirg@cs.huji.ac.il

Sudeepa Roy
Duke University
sudeepa@cs.duke.edu

Brit Youngmann
Technion
brity@technion.ac.il

## 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. A companion video for this submission is available at https://tinyurl.com/3hyszb7r.

## 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. *Consider the Stack Overflow developer survey dataset [3], which provides comprehensive information about high-tech employees worldwide. It includes numerous demographic features about individuals, such as gender and age, and role and annual salary. Table 1 shows a few sample tuples with a subset of the attributes. Now consider the following query measuring the average salary by country:*

**Table 1:** A subset of the Stack Overflow dataset.

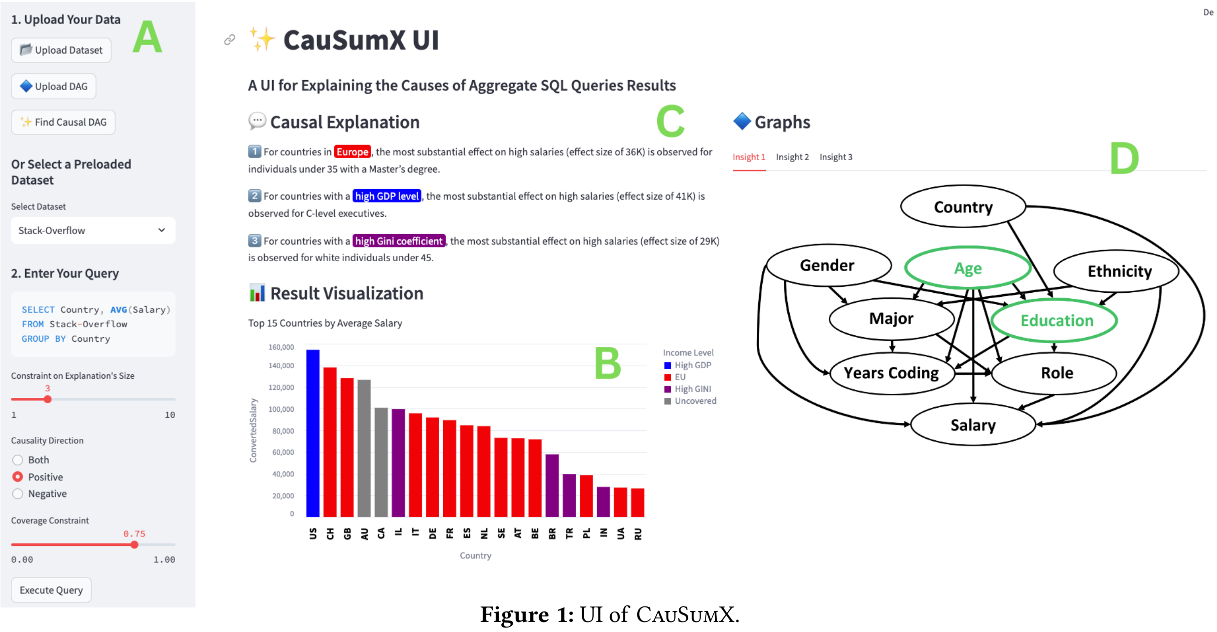
| ID | Count. | Cont. | Gender | Age | Role | Edu. | Salary |
|---|---|---|---|---|---|---|---|
| 1 | US | NA | Male | 26 | Data Scientist | PhD | 180k |
| 2 | US | NA | NB | 32 | QA developer | B.Sc. | 83k |
| 3 | India | AS | Male | 29 | CEO | B.Sc. | 24k |
| 4 | India | AS | Female | 25 | Backend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

```
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 [15, 19]. 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 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.



1. Upload Your Data

Or Select a Preloaded Dataset

Select Dataset: Stack-Overflow

2. Enter Your Query

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

Constraint on Explanation's Size: 3

Causality Direction: ○ Both ● Positive ○ Negative

Coverage Constraint: 0.75

✨ CauSumX UI

A UI for Explaining the Causes of Aggregate SQL Queries Results

💬 Causal Explanation

1 For countries in Europe, the most substantial effect on high salaries (effect size of 36K) is observed for individuals under 35 with a Master's degree.

2 For countries with a high GDP level, the most substantial effect on high salaries (effect size of 41K) is observed for C-level executives.

3 For countries with a high Gini coefficient, the most substantial effect on high salaries (effect size of 29K) is observed for white individuals under 45.

📊 Result Visualization

Top 15 Countries by Average Salary

🔷 Graphs

**Figure 1:** UI of CauSumX.

Example 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. CauSumX explores multiple patterns and evaluates their causal effect on the salary across different countries. Without having to manually explore this large 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 dataset, they have strong* causal effects *as determined by the causal model. Moreover, the analyst knows that these explanations not only hold for one country but hold for several countries that share the same grouping pattern (color).*

This demo paper complements our recent full paper [20] by highlighting the usability of CauSumX, showing its end-to-end implementation and practical scenarios for using the system.

*Related Work.* Explaining SQL query results is an extensively studied problem. Provenance-based solutions show how the output was computed using the input tuples [5]. 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 [10, 15, 19], which are easy to comprehend. While these approaches reveal interesting insights, such as outliers [19], explaining high/low values [10], or comparisons of a set of answers [15], 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 [16, 21]. However, these methods provide a single explanation of the entire view, and do not offer fine-grained explanations for subgroups. The framework presented in [11] 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 [8, 22], however these summaries are not causal.

## 2 TECHNICAL BACKGROUND

We provide an overview of our theoretical foundations of the development of CauSumX. Full details can be found in [20].

### 2.1 Background on Causal Inference

We use Pearl's model for *observational causal analysis* [14]. The broad goal of causal inference is to estimate the effect of a *treatment variable* $T$ (e.g., Education) on an outcome variable $Y$ (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* [14] ($Z$ in Eq. (1)). A sufficient set of confounders can be determined by applying graphical criteria [14], which can be evaluated against a *causal DAG.*

A causal DAG represents direct causal relationships between variables in a given dataset [14]. It can be constructed by a domain expert, or by using existing causal discovery algorithms [18]. 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 [16, 21], unlike non-causal methods, as causal effects estimate the expected difference between two groups.

A *pattern* [15] 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 ($T = 1$ if $\mathcal{P}_t$ evaluates to true for a tuple) and a control group (otherwise).

Intuitively, the grouping pattern $\mathcal{P}_g$ specifies the subpopulation of interest (equivalent to the condition $B$=$b$ for CATE), while the treatment pattern $\mathcal{P}_t$ (equivalent to treatment $T$) 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 3. *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* 36$k$ *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:

(1) **size constraint**: the number of explanation patterns should be at most $k$.
(2) **coverage constraint**: the number of groups in $Q(D)$ explained (i.e., covered) by the patterns must be at least a $\theta$-fraction of all the groups in $Q(D)$.
(3) **redundancy constraint**: an explanation pattern should not explain the same groups explained by another pattern.

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 4. *Fig. 1 depicts an explanation summary for our running example query, where* $k$=3 *and* $\theta$=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:

**Step 1: 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 [4] 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 $Q(D)$.

**Step 2: Mining Treatment Patterns** Our next goal is to identify a treatment pattern $\mathcal{P}_t$ with the highest causal effect on the outcome for each mined 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 [6]. The primary distinction from existing solutions (e.g., [6]) 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 the following optimizations to enhance runtime: **(1)** We remove attributes without a causal path to the outcome, as they do not affect CATE values; **(2)** Only the top 50% treatments with the highest CATE are considered at each lattice level; **(3)** Treatment pattern extraction for each grouping pattern is done in parallel; **(4)** CATE values are estimated using a fixed-size sample of tuples.

**Step 3: Obtaining an Explanation Summary**. Given a collection of the mined explanation patterns with their explainability scores, an integer $k$, and a threshold $\theta$, we formalize the problem of finding an explanation summary via Integer Leaner 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 formulation to LP and use an LP solver to find an explanation summary.

# 3 SYSTEM & DEMONSTRATION

We implemented CauSumX using Python and Streamlit[2] for the UI. To generate explanations in natural language, we used GPT-4 [13]. We used the DoWhy [17] package to compute causal effects, the z3-solver[3] to solve the LP, and the Apyori package [1] implementation of the Apriori algorithm.[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 [7] | 1000 | 20 | 10 |
| Adult [2] | 32.5K | 13 | 13 |
| SO [3] | 38K | 20 | 75 |
| IMPUS-CPS [9] | 1.1M | 10 | 9 |
| Accidents [12] | 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 [18] 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 [20].

- **German** [7]: This dataset contains details of bank account holders, including demographic and financial information.
- **Adult** [2]: This dataset comprises information on individuals including their education, occupation, and annual income.
- **SO** [3]: Described in the Introduction.
- **IMPUS-CPS** [9]: This dataset is derived from the Current Population Survey conducted by the U.S. Census Bureau, which includes demographic details about individuals, e.g., education, occupation, and annual income.
- **Accidents**[12]: This dataset provides comprehensive coverage of car accidents across the USA. It includes numerous environmental stimuli features that describe the conditions surrounding the accidents, such as visibility, precipitation, and traffic signals.

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

Guided tour of CauSumX: The demonstration begins by enabling attendees to 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 of the demo. To illustrate the usability of CauSumX and for the sake of demonstration, 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, and the difference in runtime between the algorithms increases as the size of the dataset increases. Further, we will show users the GPT prompt used to generate Natural Language explanations from the predicates obtained by CauSumX and allow them to select from several other prompts to get different forms of explanations such as the predicates themselves.

# REFERENCES

[1] Apyori python packge. https://pypi.org/project/apyori/, 2019.

[2] Adult census income dataset, 2021. https://tinyurl.com/46zaa4vr.

[3] Stackoverflow developer survey, 2021. https://tinyurl.com/4s298pd2.

[4] R. Agrawal, R. Srikant, et al. Fast algorithms for mining association rules. In *Proc., VLDB*, volume 1215, pages 487–499. Santiago, Chile, 1994.

[5] Y. Amsterdamer, D. Deutch, and V. Tannen. Provenance for aggregate queries. In M. Lenzerini and T. Schwentick, editors, *PODS*, pages 153–164. ACM, 2011.

[6] A. Asudeh, Z. Jin, and H. Jagadish. Assessing and remedying coverage for a given dataset. In *ICDE*, pages 554–565. IEEE, 2019.

[7] A. Asuncion and D. Newman. Uci machine learning repository, 2007.

[8] K. El Gebaly, P. Agrawal, L. Golab, F. Korn, and D. Srivastava. Interpretable and informative explanations of outcomes. *Proc. VLDB*, 8(1):61–72, 2014.

[9] S. Flood, M. King, S. Ruggles, and J. R. Warren. Integrated public use microdata series. *University of Minnesota*, 1, 2015.

[10] C. Li, Z. Miao, Q. Zeng, B. Glavic, and S. Roy. Putting things into context: Rich explanations for query answers using join graphs. In *SIGMOD*, 2021.

[11] P. Ma, R. Ding, S. Wang, S. Han, and D. Zhang. Xinsight: Explainable data analysis through the lens of causality. *Proc. ACM Manag. Data*, jun 2023.

[12] S. Moosavi, M. H. Samavatian, S. Parthasarathy, and R. Ramnath. A countrywide traffic accident dataset. *arXiv preprint arXiv:1906.05409*, 2019.

[13] OpenAI. Gpt-4 technical report, 2023.

[14] J. Pearl. *Causality*. Cambridge university press, 2009.

[15] S. Roy and D. Suciu. A formal approach to finding explanations for database queries. In *SIGMOD*, 2014.

[16] B. Salimi, J. Gehrke, and D. Suciu. Bias in olap queries: Detection, explanation, and removal. In *SIGMOD*, pages 1021–1035, 2018.

[17] A. Sharma and E. Kiciman. Dowhy: An end-to-end library for causal inference. *arXiv preprint arXiv:2011.04216*, 2020.

[18] P. Spirtes et al. *Causation, prediction, and search*. MIT press, 2000.

[19] E. Wu and S. Madden. Scorpion: Explaining away outliers in aggregate queries. *Proc. VLDB*, 2013.

[20] B. Youngmann, M. Cafarella, A. Gilad, and S. Roy. Summarized causal explanations for aggregate views. *Proc. ACM Manag. Data*, 2(1), 2024.

[21] B. Youngmann, M. Cafarella, Y. Moskovitch, and B. Salimi. On explaining confounding bias. *ICDE*, 2023.

[22] C. Yu, L. V. S. Lakshmanan, and S. Amer-Yahia. It takes variety to make a world: diversification in recommender systems. In M. L. Kersten, B. Novikov, J. Teubner, V. Polutin, and S. Manegold, editors, *EDBT*, volume 360, pages 368–378. ACM, 2009.