FEDEX: An Explainability Framework for Data Exploration Steps

Daniel Deutch
Tel Aviv University
danielde@post.tau.ac.il

Amir Gilad
Duke University
agilad@cs.duke.edu

Tova Milo
Tel Aviv University
milo@cs.tau.ac.il

Amit Mualem
Tel Aviv University
amitmualem@mail.tau.ac.il

Amit Somech
Bar-Ilan University
somecha@cs.biu.ac.il

ABSTRACT

When exploring a new dataset, Data Scientists often apply analysis queries, look for insights in the resulting dataframe, and repeat to apply further queries. We propose in this paper a novel solution that assists data scientists in this laborious process. In a nutshell, our solution pinpoints the most interesting (sets of) rows in each obtained dataframe. Uniquely, our definition of interest is based on the contribution of each row to the interestingness of different columns of the entire dataframe, which, in turn, is defined using standard measures such as diversity and exceptionality. Intuitively, interesting rows are ones that explain why (some column of) the analysis query result is interesting as a whole. Rows are correlated in their contribution and so the interesting score for a set of rows may not be directly computed based on that of individual rows. We address the resulting computational challenge by restricting attention to semantically-related sets, based on multiple notions of semantic relatedness; these sets serve as more informative explanations. Our experimental study across multiple real-world datasets shows the usefulness of our system in various scenarios.

PVLDB Reference Format:
Daniel Deutch, Amir Gilad, Tova Milo, Amit Mualem, and Amit Somech. FEDEX: An Explainability Framework for Data Exploration Steps. PVLDB, 15(13): 3854 - 3868, 2022.
doi:10.14778/3565838.3565841

PVLDB Artifact Availability:
The source code, data, and/or other artifacts have been made available at https://github.com/TAU-DB/FEDEx.

1 INTRODUCTION

Exploratory Data Analysis (EDA) is an essential process performed by data scientists and analysts in order to up-close examine a new dataset, better understand its nature and characteristics, and extract preliminary insights from it. EDA is typically performed in *scientific notebooks* [38] using a dataframe library [53] which allows users to interactively transform and analyze datasets via programmatic query operations (such as filter, aggregation, join, pivot, etc.). In each exploratory step, the analyst runs a query over the previously obtained dataframe, examines the resulting dataframe to derive intermediate insights, and decides on the next exploratory step.

EDA is a challenging process, where arguably the most significant challenge lies in scanning the (possibly many) query results obtained in each step to identify interesting patterns and trends. These discoveries are crucial both for gathering insights and conceiving the next exploratory steps. To illustrate the difficulty faced by analysts in interpreting the results of exploratory steps, consider the following example EDA scenario.

Example 1.1. Data scientist Clarice works on a dataset of song information, published by Spotify[1].

The dataset (see samples in Figure 1) contains a *popularity score* for each song, alongside 37 descriptive features (e.g., artist name, year, decade) as well as audio-analysis features such as *loudness*, *danceability*, and others.

Clarice is first interested in the question "what makes songs popular?". She, therefore, composes a filter operation showing songs that have a *high* popularity score (> 65). Applying this operation yields a dataframe with 7000 rows and 37 columns (see Figure 1a for a small sample). Clarice now needs to examine these results in order to understand what is interesting about them. One such notion of interestingness may be captured by the question "In what way are these popular songs different than the rest of the songs in the dataset?" Clarice therefore needs to sift through the rows, apply additional data transformations, and data visualizations.

Clarice then decides to focus her analysis only on recent songs (released after 1990) and investigate their characteristics. To do so, she first filters the original dataframe to output songs released after 1990, and then performs a group-by operation to view the mean *loudness* and *danceability* values for each year.

A sample of the result of this group-by step appears in Figure 1b. This step results in a much smaller dataframe than before, including about 30 rows and 3 columns. However, interesting patterns and trends are still not clearly visible – are there any particular years with more 'quiet' songs than the other? is there a trend where newer songs are more 'loud' and 'danceable'? Answering these questions still requires a substantial effort.

In this paper, we present FEDEX (Ef**fici**e**n**t **D**ata Exploration **Ex**planations), an EDA explanation framework that assists users in analyzing and understanding the results of their exploration steps. FEDEX explains exploratory steps using a twofold process: (1) as an analyst would do, FEDEX inspects the resulted dataframe and discover interesting aspects of it. For example, does it show outliers? Or perhaps a highly diverse set of values in one of the columns?;

[1] https://www.kaggle.com/c/bfh-spotify-challenge/overview

Proceedings of the VLDB Endowment, Vol. 15, No. 13 ISSN 2150-8097.
doi:10.14778/3565838.3565841

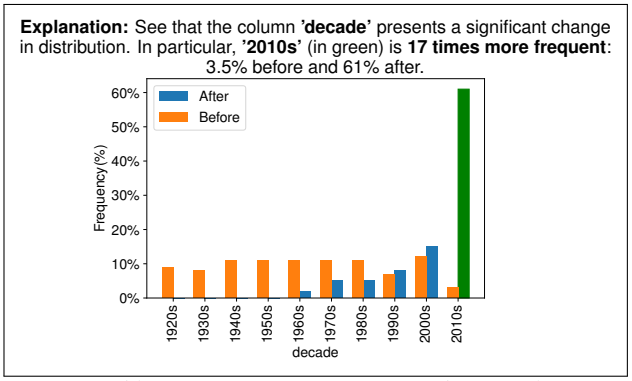
name	main_artist	year	decade	...	danceability	loudness	popularity
Maybe I'm Amazed...	Paul McCartney	1970	1970	...	0.471	-10.407	66
Life on Mars?...	David Bowie	1971	1970	...	0.442	-14.635	71
Time in a bottle	Jim Croce	1972	1970	...	0.544	-11.952	67
Desperado - 2013...	Eagles	1973	1970	...	0.228	-12.749	67
...	...	...	...	...	...	...	...

(a) Filter results

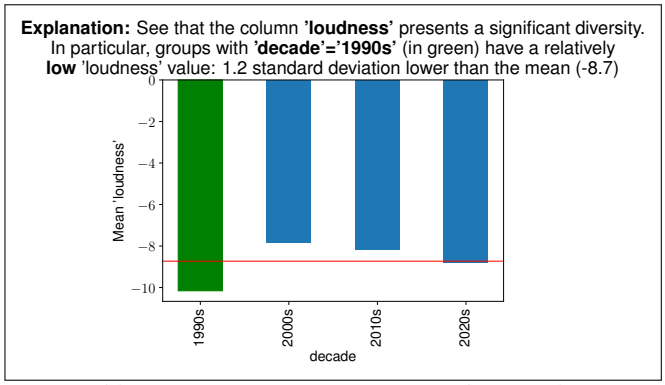
year	loudness	danceability
1991	-11.076072	0.555551
2014	-7.826855	0.586153
1992	-10.694651	0.555135
2013	-8.238654	0.593881
...	...	...

(b) Group-by results

Figure 1: Result samples of filter and group-by operations over the Spotify "Song Popularity Dataset". Figure 1a shows partial results of a filter operation that leaves in only songs with 'popularity' scores above 65. Figure 1b depicts partial results of a group-by operation, showing for each year – since 1990 – the mean 'loudness' and 'danceability' values.



(a) Explanation for the filter step (Figure 1a)



(b) Explanation for the group-by step (Figure 1b)

Figure 2: Multi-modal explanations generated by FEDEX. Figure 2a explains the filter results in Figure 1a. Our framework has identified that the highest contribution to the interestingness score is due to songs made in the 2010s – they form 61% of the popular songs, compared to only 3.5% in the entire dataset, leading to the insight that *new songs tend to be more popular than older songs*. Figure 2b explains group-by results in Figure 1b. Our explanation points out that *songs made in the 90s tend to be less loud than later songs*.

(2) since naturally, not all underlying tuples equally contribute to the interesting pattern discovered, FEDEX detects which data subsets *cause* the resulted dataframe to be interesting. Therefore, a good *explanation* for an exploratory step is a set-of-rows from the dataframe which *significantly contributes* to a *highly interesting* aspect of the results dataframe. Importantly, FEDEX tailors the explanation to the *context* of the generated dataframe, i.e., the EDA operation or query and the source dataframe. FEDEX then presents this set-of-rows in a hybrid format via the notebook interface as coherent, easy-to-read *captioned visualizations*. This allows users to quickly understand and derive immediate insights from each exploratory step they make.

Example 1.2 (FEDEX Explanations). Figure 2 shows the explanations produced by FEDEX, for the two exploratory steps on the Spotify dataset in Example 1.1 (see Figure 1). For the filter step, which focuses only on songs with a popularity score greater than 65, FEDEX detects that in the 'decade' column there exists a high, interesting *deviation* from the input dataframe (namely, the dataframe Clarice applied the filter on). It then detects that most of the deviation is due to *songs from the 2010s*. Using a side-by-side bar plot, FEDEX highlights the difference in the number of songs from the 2010s decade before and after the filter, showing that *newer songs from the last decade tend to be more popular than older songs*. As for the group-by step, showing the mean 'loudness' and 'danceability' values for each release year, FEDEX identifies an interesting pattern in the column 'loudness', whose values greatly differ from one another. It then detects that the large diversity in values occurs mainly due to years in '1990s', highlighting that *the songs released in the 1990s are significantly less loud than songs made in later years*.

Multiple approaches exist for assisting data exploration, such as recommending the next exploratory steps [50, 65], query autocompletion [39, 40], visualization [41, 76, 78] and even insights [3, 72] recommendations. These works reduce the manual effort of composing queries, yet to the best of our knowledge, FEDEX is the first system to automatically generate explanations for output dataframes *in the context of exploratory steps*.

Our main contributions can be summarized as follows.

- We devise FEDEX, a novel explainability framework for data exploration, using a hybrid, twofold approach: (1) FEDEX uses measures of interestingness [31, 63] to detect interesting columns in the resulted output dataframe. (2) FEDEX then applies a notion of *contribution of a set of rows*, inspired by previous work [79], to identify meaningful subsets of the input dataframe that significantly contribute to the interestingness score of one of the columns in the output dataframes. We define an explanation candidate as a tuple (R, A), where R is a set of rows in the input dataframe that contribute to the interestingness of a column A in the step's output dataframe.
- A naive computation of explanation candidates would imply exponential complexity as we would have to traverse all possible pairs of columns and row subsets. Therefore, FEDEX employs the following efficient process for finding the most promising

explanations in terms of both interestingness and contribution, which are also semantically meaningful: First, FEDEX uses a two-step greedy approach to calculate the most promising explanation candidates: (1) it identifies the interesting columns, utilizing a sampling-based calculation of interestingness, which significantly accelerates the computation while maintaining the quality of the results. On the resulted interesting columns, it then (2) computes the contribution of only *selected* sets of rows (rather than all of them) by partitioning the rows into *semantically meaningful sets-of-rows*. This is done by several automatic partitioning methods which dramatically reduce the number of candidate and keep only the meaningful ones. Finally, FEDEX uses the skyline operator [13] over the compact set of semantically-meaningful explanation candidates, to retrieve only the *dominating* ones, in terms of contribution and interestingness scores. Each skyline explanation is presented to the user via a dedicated captioned visualization.

- We have implemented a prototype version of FEDEX (available in [71]), performed comprehensive user studies, and simulated experiments. The results indicate that our approach is 1.7 times more helpful than commonly used baselines on average, allows users to get approximately 4 more insights on average than unassisted EDA, and runs at interactive speed for large data.

2 RELATED WORK

As mentioned above, multiple lines of work have studied solutions to assist data scientists in data exploration, focusing mainly on the aspect of formulating and composing exploratory operations. For example, systems for query suggestions [50, 65], simplified EDA interfaces for non-programmers [4, 10, 69, 70], query formulation assistance [39, 40], and EDA guidance tools [36, 62]. We focus here on the challenge of *explaining the results of the exploratory steps and assisting users in quickly understanding them*. In this context, we survey three lines of related work: (1) interestingness assessment in data exploration, (2) query explanation frameworks, and (3) automatic visualizations and insights discovery.

Modeling and predicting interest in exploratory sessions. Modeling interestingness in EDA, i.e., assessing whether a resulted view of an EDA operation is interesting or not, is a known challenge: previous work showed that interestingness in EDA is multi-faceted [28, 46], often subjective [21] and dynamically changing, even in the same exploratory session [49].

Consequently, several works attempt to model or predict user interests in EDA sessions: works such as [45, 49] propose predicting the most suitable notion of interestingness by mining logs from previous EDA sessions. The system described in [49] is used to dynamically select, before the user employs another EDA operation, the most suitable among a set of *existing* interestingness measures, whereas in [45] a *learning-to-rank* model is used to construct an ad hoc interestingness notion. Another line of work proposes modeling user interests by collecting live feedback: systems such as [24, 56] ask the user, e.g., to annotate presented tuples as "interesting" or "non-interesting", while *explore-by-example* systems [23, 66] ask the users to provide examples for interesting tuples in advance.

While FEDEX employs measures of interestingness, the end goal is fundamentally different. Rather than using these measures to recommend interesting operations or views as in previous work, FEDEX uses the measures to evaluate the users' operations and generates explanations based on sets-of-rows that impact the interestingness. This allows FEDEX to identify *why* a user's operation is interesting. Furthermore, in Section 4.2 we show that explanations generated by directly applying interestingness measures are generally less useful compared to those generated by FEDEX.

Explaining query results. Explaining the results of SQL queries for debugging purposes has been extensively researched within the database community. Prominently, such research works utilize data provenance, as well causality-stemmed notions such as intervention and influence, in order to identify tuples whose existence or absence affects the result of the inspected query. For example, [15, 29] explain the *existence* of particular tuples in the query result, whereas works e.g. [11, 12, 17, 73] explain the *absence* of particular tuples. Explaining aggregation results has been suggested in [2], and outliers in [61, 67, 79]. In addition, works e.g., [22, 42, 48] propose augmenting provenance with additional sources of information such as counter-balancing patterns [48], related tables [42], and natural language questions [22]. Last, data debugging tools also utilize notions of causality and intervention, e.g., to detect erroneous tuples that significantly impact an aggregation result [6], and to detect faults in predictive models that may be caused by non-conforming data subsets [27].

The main difference from FEDEX is that these works explain which input tuples affect the *direct output* of a query (i.e., why certain tuples appear/do not appear in the results), whereas the goal of our work is focused on explaining *why* an EDA operation is *interesting*, by detecting sets-of-rows that positively impact the interestingness of users' exploratory steps.

Automatic visualization generation & insights discovery. In recent years, a plethora of visualization generation systems have been devised [8, 9, 45, 57, 76, 78, 81], taking a dataset as input and generating useful data visualizations (e.g., histograms, line plots, heatmaps). Many such systems utilize a notion of *utility* or *importance*, and scan the input dataset to show the top-k visualizations obtaining the highest score [57, 76, 78].

Using roughly similar methods, previous works propose solutions for facts and insights discovery. Given an input dataset, these solutions can automatically detect several types of useful insights, such as *outliers* [34, 72], *rare categories* [32], *conditional correlations* [51, 52], and *trend-lines* [3, 25, 72]. Other solutions propose insights discovery for *specific* use-cases such as *user rating data* [74], *smart meters data* [44], *OLAP* [1] and *data fusion* [26].

A step further, systems like [18, 77] generate random interesting facts (in a similar manner to, e.g., [25, 72]), then organize them in a comprehensive medium such as fact sheets [68, 77] and infographics [18] or use them as guidance in interactive exploration [70].

Similarly, FEDEX also assists users in discovering insights from the data. However, rather than generating random, arbitrary facts from the dataset (as e.g., [57, 72, 76–78]), FEDEX generates insights that explain each exploratory step made by the user. The methods used in FEDEX are also different than previous work, and ours is the only work, to our knowledge, that derives insights by interweaving interestingness and causality assessment. In our experimental

evaluation, we compare the explanations of FEDEX to visualizations and insights generated by [76] and [77], and show that FEDEX is significantly more effective in explaining exploratory steps.

3 EXPLANATION FRAMEWORK

We begin with a brief overview of our data model for notebook-based exploration, then detail each of the components in FEDEX.

3.1 Model for Notebook-based EDA

As mentioned, the notebook interface allows users to explore datasets interactively using a dataframe module [53]. The user typically loads a dataset as a dataframe from an external source (e.g., a CSV file, spreadsheet, or an SQL query result from a database server), then interacts with them by employing programmatic exploratory operations, such as filter, group-by, aggregate, etc. The results of EDA operations are also dataframes; these dataframes may be viewed and/or used as input for subsequent exploration steps.

Dataframe. Formally, a dataframe d is equivalent to a single relational table or view. It comprises of a multiset of rows over a schema $\mathcal{A}(d)$. Each row $r \in d$ represents either a tuple or a group (in case d is resulted from a group-by operation). Correspondingly, an attribute $A \in \mathcal{A}(d)$ specifies either a table attribute, or an aggregated column. Furthermore, we refer to a dataframe *column* by $d[A]$, i.e., the multiset of all rows' values associated with the attribute A in dataframe d.

EDA operations. We consider popular, commonly used (see [80]) of filter, join, group-by, and union. Additional, advanced EDA and OLAP operations such as *pivot*, *diff*, and *roll-up* can be supported by a simple extension of our model. We further denote a full exploratory step by $Q = (D_{in}, q, d_{out})$, i.e., applying operation q on an input dataframe(s) $d_{in}(D_{in})$, resulting in an output dataframe d_{out}. Multiple input dataframes ($|D_{in}| > 1$) may be used when the operation is a join or union.

The following example demonstrates the workflow of a notebook-based EDA process.

Example 3.1. Reconsider the Spotify songs dataset from Example 1.1. After a user loads it into a dataframe d_0, she performs the first exploratory step by employing a *filter* operation q_1 equivalent to the SQL query *'select * from d_0 where popularity>65'* This results in a new dataframe $d_{out} = q_1(d_0)$, shown in Figure 1a, that includes the subset of rows that comply to the criterion of "popularity>65".

The group-by operation from our running example, specified by q_2, is equivalent to the query *'select AVG(loudness), AVG(danceability) from d_0 where year>=1990 group by year'*.

The resulting dataframe (depicted in Figure 1b) shows the mean values for *loudness* and *danceability* per year, starting from 1990.

3.2 Interestingness of Exploration Steps

Measuring interestingness of data analysis operations has been intensively discussed in previous work (see [28, 31] for surveys). These works define numerous abstract facets of interestingness, such as novelty, surprisingness, exceptionality, and diversity. Often such notions are implemented differently, depending on the data and task [57, 76, 78]. FEDEX can take an existing or a custom, user-defined notion of interestingness (See Section 3.8). As default measures, FEDEX utilizes two interestingness functions corresponding to the notion of *exceptionality* and *diversity*. Similar implementations of such functions were proven useful in the context of EDA tasks [5, 49, 75, 76]. These two measures are applicable, as described below, for all the EDA operations supported by FEDEX (filter, join, group-by, and union).

Since applying q does not equally affect all columns in the output dataframe d_{out}, following [49, 50], we assess the interestingness of the step $Q = (D_{in}, q, d_{out})$ separately for each column in the output dataframe d_{out}. We denote by $I_A(Q)$ the *interestingness score* of step Q, by an interestingness function I, w.r.t. attribute A.

Exceptionality (Filter/Join/Union). Inspired by [63, 75, 76], we introduce an exceptionally measure that is particularly suitable for filter and join steps, as follows. For the case of a filter operation q_f, this measure intuitively deems the filter result interesting if it produces an output dataframe which significantly *deviates* from the input dataframe. Namely, the filter operation caused a significant change in the distribution of the values of a given column. We, therefore, quantify the deviation of the step $Q = (d_{in}, q_f, d_{out})$ w.r.t Attribute $A \in \mathcal{A}(d_{out})$ by measuring the differences between the values of $d_{in}[A]$ and $d_{out}[A]$, which are the columns associated with attribute A in dataframes d_{in} and d_{out} (resp). Concretely, we use the two-sample Kolmogorov–Smirnov (KS) test [60], a well-known statistical test used to assess whether two samples originate from the same probability distribution. First, we define the column probability distribution $Pr(d[A])$ based on the relative frequency of its values (i.e., for each value $v \in d[A]$, $Pr(v)$ is the probability to choose v uniformly at random). We then calculate the exceptionality-based interestingness score:

$$I_A(d_{in}, q_f, d_{out}) := KS(Pr(d_{in}[A]), Pr(d_{out}[A])) \quad (1)$$

The same measure is also suitable for quantifying the interestingness of the join and union operations since, oftentimes, the results of a join (or union) contain a different number of tuples than in the original input dataframes (e.g., if not all tuples in the input dataframes match on the join condition).

For a join operation q_j and a column A in the output dataframe d_{out}, the interestingness is measured by $I_A(d'_{in}, q_j, d_{out})$, where d'_{in} is the corresponding input dataframe in D_{in}, s.t. $A \in \mathcal{A}(d'_{in})$ and I_A is calculated as in Equation 1.

For a union operation q_U, the interestingness of a column A is defined by $\max_{d_{in} \in D_{in}} I_A(d_{in}, q_U, d_{out})$, namely the maximal KS difference of the output column $d_{out}[A]$ and each of the input dataframes to be unionized.

Diversity (group-by). We have also implemented an interestingness function that captures *diversity* [7, 31]. It is particularly suitable for group-by operations, since, intuitively, a group-by step that yields a dataframe with a highly diverse set of aggregated values, implies a large difference between the groups. Therefore, our interestingness function quantifies the diversity of a column in the output dataframe of a group-by operation q_g by calculating the coefficient of variation (CV) [7] of the aggregated values:

$$I_A(d_{in}, q_g, d_{out}) := CV(d_{out}[A]) = \frac{1}{\bar{a}} \cdot \sqrt{\frac{\sum(a_i - \bar{a})^2}{n-1}} \quad (2)$$

Table 1: Notations Summary

Notation	Description
d, $\mathcal{A}[d]$	Dataframe, Dataframe Schema
$d[A]$, $A \in \mathcal{A}[d]$	Dataframe column
q	EDA operation specifications
$d_{in}(D_{in}), d_{out}$	Input Dataframe(s), Output Dataframe
$Q = (D_{in}, q, d_{out})$	Exploratory step
$I_A(Q)$	Interestingness of step Q w.r.t attribute A
$R \subseteq d$	Set-of-Rows in a dataframe d
$C(R, A, Q)$	Contribution of a set-of-rows R to interestingness of a column A in step Q
$\mathcal{R} = \{R_1, R_2, \ldots, \hat{R}\}$	Row partition of a dataframe to sets-of-rows
$\bar{C}(R, A)$	Standardized contribution of a set-of-rows $R \in \mathcal{R}$, compared to the rest of the row-sets in $\mathcal{R}$
$(R, A) \in EC$	The set of explanation-candidates, each is a pair of set-of-rows and a single column in d_{out}
$E \in EX$	A dominating explanation in terms of contribution and interestingness

Where the sum in the numerator is over $a_1, a_2, \ldots, a_n$ which are the aggregated values in column A after employing q_g, $\bar{a}$ is mean value, and n is the number of groups. The following example demonstrates the usage of the exceptionality and diversity measures as defined above, to evaluate the interestingness of the filter and group-by operations shown in Figure 1.

Example 3.2. Consider again the filter and group-by results depicted in Figure 1. For the group-by results (Figure 1b), it is clearly visible that the values in column 'loudness' are more diverse than that of 'danceability', which shows value distribution tightly around 0.55. Indeed, the diversity-based score of the column 'loudness' is 0.13, whereas the score of the column 'danceability' is 0.04. Hence, our framework will focus on the former column to explain the interestingness of the view (See Figure 2b). As for the filter results, while this is not visible in the sample in Figure 1a, the highest deviation was measured for the column 'decade', obtaining an interestingness score of 0.56, followed by the columns 'year' and 'loudness', for which the scores are 0.54 and 0.41, respectively.

3.3 Contribution of Sets of Rows

Given an exploratory step $Q = (D_{in}, q, d_{out})$, we further aim to focus the explanation by quantifying the *contribution* of a set-of-rows $R \subset D_{in}$ to the interestingness of a column A, $I_A(Q)$. Hence, we define a contribution function, as follows. Our definition of contribution draws on the notion of intervention, first defined in the context of causality [54] to measure the change in the outcome when the input changes. This notion was adopted by the database community for measuring the change in query results when some of the rows are absent [47, 61, 79]. In particular, our notion of contribution function is inspired by [79] and defined as follows.

Definition 3.3 (Contribution of a set-of-rows). For a given step $Q = (D_{in}, q, d_{out})$, the amount that a set-of-rows $R \subset D_{in}$ *contributes* to the interestingness of column $A \in \mathcal{A}(d_{out})$ is calculated by:

$$C(R, A, Q) = I_A(D_{in}, q, d_{out}) - I_A(D_{in} - R, q, d'_{out})$$

Namely, we remove the rows set R from the input dataframes D_{in}, employ again the operation q which now results in a new output dataframe d'_{out}, and recalculate the interestingness score I_A. Intuitively, the higher the decrease in the interestingness score caused by removing R, the higher its contribution is to the interestingness of the column $d_{out}[A]$.

Example 3.4. Recall that Example 3.2 showed that the most interesting column in the filter step (Figure 1a) is 'decade', obtaining a score of 0.56. Calculating, for example, the contribution of the 'decade'="2010s", we first omit from d_{in} all rows where 'decade'="2010s". We then employ the operation q, i.e., *"filter by 'popularity' greater than 65"*) and recalculate the interestingness. The resulted score is now 0.47, which means that the contribution score of "2010s" is $C(R_{\text{"2010s"}}, A_{\text{decade}}, Q) = 0.086$. This is a relatively high contribution score, as the interestingness of the column 'decade' decreased by 16% when removing songs made in the 2010s.

Note that generally, the contribution of a set-of-rows may be negative, if removing this set-of-rows reduces the value of the interestingness function. For example, for group-by queries with diversity as an interestingness function, the contribution function may either be negative or positive. To see this, consider the dataframe $d_{in} = \{(x, 1), (x, 2), (y, 3)\}$ and the group-by query that groups the first attribute and sums the second. Here $d_{out} = \{(x, 3), (y, 3)\}$ (diversity $= 0$), whereas if we remove $(x, 2)$ from d_{in} and perform the same query, we will get $d_{out} = \{(x, 1), (y, 3)\}$ (diversity > 0). Thus, the contribution of $(x, 2)$ is negative. However, if we consider the dataframe $d_{in} = \{(x, 1), (x, 1), (y, 1)\}$ with the same query, we will get $d_{out} = \{(x, 2), (y, 1)\}$ (diversity > 0). If we remove one of the $(x, 1)$ tuples and evaluate the same query, we get $d_{out} = \{(x, 1), (y, 1)\}$ (diversity $= 0$). Thus, the contribution of $(x, 1)$ is positive. As we describe below, we are interested in sets-of-rows that obtain a significantly high, positive contribution score, compared to other sets-of-rows. If there are no sets-of-rows with a positive contribution, no explanation will be generated for the exploration step (see Section 3.7).

3.4 Explanation Candidates

As mentioned above, FEDEX *explains* a given exploratory step by identifying sets-of-rows in its *input* dataframes that contribute to the interestingness score $I_A(Q)$, on column A in the *output dataframe*. An explanation-candidate is defined as follows:

Definition 3.5 (Explanation Candidate). For exploratory step $Q = (D_{in}, q, d_{out})$, an explanation-candidate E is defined by $E := (R, A)$ where $R \subset D_{in}$ and $A \in \mathcal{A}(d_{out})$

Example 3.6. Consider again the filter step depicted in Figure 1a resulting in songs with popularity over 65. An example explanation-candidate is $(R_{'decade'=2010s'}, A_{'decade'})$, namely the rows from d_{in} (i.e., the full songs dataframe) depicting songs released in the 2010s decade, and the 'decade' column in the output dataframe. To understand how $(R_{'decade'=2010s'}, A_{'decade'})$ explains the filter step, recall that the interestingness score here measures the deviation from d_{in} to d_{out}. As calculated in Example 3.4, the contribution $C(R_{'decade'=2010s'}, A_{'decade'}, Q)$ is high, which means that songs from 2010s *explain* this deviation. Indeed, as also illustrated in the final explanation (figure 2a), *songs (rows) from 2010s are highly more frequent after the filter, than in the entire dataset.*

Naturally, not all explanation-candidates are useful. We next explain how we restrict the set-of-rows to include only the ones that are semantically related, then describe the quality metric of the explanation-candidates and how FEDEX only selects the best ones and return them as coherent captioned visualizations.

3.5 Partitioning the Input Dataframe

While, in theory, one could calculate the contribution of individual rows, or, alternatively, do so for *all* sets-of-rows, FEDEX focuses on the contribution of *semantically related sets-of-rows*. This both allows for a faster computation (as is shown in the sequel) and yields explanations that depict the "bigger picture", allowing users to obtain meaningful, high-level insights from each exploratory step. The following example illustrates the latter point.

Example 3.7. Consider again the group-by results in Figure 1b. Recall from Example 3.2 that the column 'loudness' obtains high interestingness score, as it contains a diverse set of values. Had we computed the contribution of, e.g., each single row in d_{in}, we would have seen that the top-2 contribution values are of '1991' and '2007', since they have the most extreme values (-11.07 and -7.49, resp). Yet instead, if we group together years by their corresponding decade, we now obtain that the highest contribution is made by the decade '1990s'. This now yields a higher-level interesting pattern indicating that *songs released in the 1990s tend to be less loud than in later decades*. As detailed in the sequel, this observation will be highlighted by the output explanation of FEDEX (See Figure 2b).

We next define the row-partition scheme and detail the specific methods currently implemented in FEDEX.

Definition 3.8 (Row Partition). Given an input dataframe $d_{in} \in D_{in}$, a row partition divides d_{in} into $n+1$ disjointed sets-of-rows: $\mathcal{R} = \{R_1, R_2, \ldots, R_n, \hat{R}\}$ such that:

$$\forall R_i, R_j \in \mathcal{R},\ R_i \subset d_{in} \wedge R_j \subset d_{in} \wedge \bigcup_{R_i \in \mathcal{R}} R_i \cup \hat{R} = d_{in} \wedge R_i \cap R_j = \emptyset$$

The special set-of-rows $\hat{R} \in \mathcal{R}$ (can be empty) is called an *ignore-set*, as it cannot become an explanation-candidate, as detailed below.

FEDEX supports three types of row partition methods and utilizes them when generating explanation (as described in Section 3.7). Other partition methods are supported, as long as they comply with definition 3.8. The partition methods in FEDEX are as follows:

Frequency-based partition. Given either a numeric or categorical attribute A we divide d_{in} to n sets-of-rows, corresponding to the n most prevalent values in the column $d_{in}[A]$. We then assign the rest of rows in the ignore set $\hat{R}$.

Numeric-based partition. Given a numeric attribute A, we divide the rows in d_{in} according to their corresponding values in $d_{in}[A]$ to n sets-of-rows, using *equal-frequency binning*. Namely, each set-of-rows in this partition correspond to an interval of values of $d_{in}[A]$, s.t. the number of values in each interval is equal. The ignore-set $\hat{R}$ is empty in this case.

Many-to-one partition. This method partitions the rows by mining many-to-one relationships between columns in d_{in}. Given an attribute A, we look for a different attribute B s.t. each value in $d_{in}[A]$ is mapped to a single value in $d_{in}[B]$, yet there exist at least two values in $d_{in}[B]$ mapped to different values of $d_{in}[A]$. Formally, for an attribute A in d_{in} we search for all columns B such that both of the following conditions hold:

(1) $\forall r_i, r_j \in d_{in}, (r_i[A] = r_j[A]) \rightarrow (r_i[B] = r_j[B])$
(2) $\exists r_i, r_j \in d_{in}, (r_i[B] = r_j[B]) \wedge (r_i[A] \neq r_j[A])$

Intuitively, these conditions ensure that the partition of values from A according to column B are consistent, i.e., every pair of equal values is placed in the same set (Condition 1), and that the partition according to B will be strictly coarser compared to the partition according to A (Condition 2).

After mapping the values from A to B we split the set-of-rows using the frequency-based partition (as defined above), over the column B. This partition is particularly useful for group-by dataframes.

Example 3.9. Reconsider the group-by output dataframe in Figure 1b, where each row represents a single year. Partitioning the rows via the 'year' column could be done using the frequency-based method, or many-to-one (as the column is categorical, numeric binning is not applicable). Using the frequency-based method, we can partition the rows in the input dataframe (i.e., before applying the group-by operation) according to the n most frequent 'year' values and placing the rest of the rows in the ignore-set $\hat{R}$. In the many-to-one partition, the column 'decade' has a many-to-one relationship with 'year'. This method has yielded a preferable explanation, particularly for the set-of-rows associated with *'decade'='1990s'*.

3.6 Quality of Explanation Candidates

Intuitively, an explanation-candidate $E = (R, A)$ is a good explanation for the exploratory step Q if the contribution $C(R, A, Q)$ to the interestingness of A is significant, and the interestingness score $I_A(Q)$ is itself high. We next explain how the significance of contribution is calculated, and how FEDEX balances between interestingness and contribution.

We evaluate the significance of contribution for a given set-of-rows $R \in \mathcal{R}$ as follows. Rather than considering the raw contribution score of a set-of-rows R, we compare the contribution of R to the contributions of the other sets-of-rows in $\mathcal{R}$. We then define the *standardized contribution* of a set-of-rows $R \in \mathcal{R}$ by:

$$\bar{C}(R, A) = \frac{C(R, A, Q) - \mu_{\mathcal{R}}}{s_{\mathcal{R}}}$$

Where $\mu_{\mathcal{R}}$ and $s_{\mathcal{R}}$ are the mean and standard deviation of the contribution scores of all sets-of-rows in the partition $\mathcal{R}$.

This quantifies the significance of the contribution of R, since the higher the *standardized* contribution, the farther the contribution of R from the mean contribution of its fellow row sets.

Correspondingly, the *quality* of an explanation (R, A) is measured by using two metrics: (1) the standardized contribution $\bar{C}(R, A)$ and (2) the interestingness of the column A, $I_A(Q)$.

Skyline of contribution & interestingness. We next define the set of desired explanations based on contribution and interestingness. Denote by $EC(Q)$ the set of all explanation-candidates considered in FEDEX for a step Q is, formally:

$$EC(Q) \coloneqq \bigcup_{\mathcal{R}} \bigcup_{R \in \mathcal{R}} \bigcup_{A \in \mathcal{A}(d_{out})} \{(R, A)\}$$

Algorithm 1: FEDEX Explanations Generation

Input: Exploratory step $Q = (D_{in}, q, d_{out})$
Output: Explanations for Q

1 **for** $A \in \mathcal{A}(d_{out})$ **do**
2 | Calculate interestingness score $I_A(Q)$
3 $\mathcal{SR} \leftarrow \emptyset$
4 **foreach** *row-partition* **do**
5 | $\mathcal{R} \leftarrow$ *row-partition*(D_{in})
6 | $\mathcal{SR} \leftarrow \mathcal{SR} \cup \{\mathcal{R}\}$
7 $EC \leftarrow$ Empty Dictionary
8 **foreach** $\mathcal{R} \in \mathcal{SR}, A \in \mathcal{A}(d_{out})$ **do**
9 | **foreach** $R \in \mathcal{R}$ **do**
10 | | Calculate contribution $C(R, A, Q)$
11 | | **if** $C(R, A, Q) > 0$ **then**
12 | | | $EC[(R, A)] \leftarrow \left(\bar{C}(R, A), I_A(Q)\right)$
13 $EX = \underset{(R,A)\in EC}{\text{SKYLINE}} \left(\bar{C}(R, A), I_A(Q)\right)$
14 **foreach** $E \in EX$ **do**
15 | GenerateVisualExplanation(E)

Given the set $EC(Q)$ of all explanation-candidates, we look for ones that obtain *both* a good contribution and a high interestingness score. To balance the two metrics we use a skyline-operator [13] calculation. Namely, we define the set EX of desired explanations as a maximal subset of $EC(Q)$ satisfying the following:

$$\forall (R, A) \in EX.\ \nexists (R', A') \in EC(Q).\ (I_{A'}(Q) > I_A(Q) \wedge \bar{C}(R', A') > \bar{C}(R, A))$$

Example 3.10. Reconsider the group-by step, depicted in Figure 1b. Recall from Example 3.2 the interestingness of column 'loudness' is higher than of 'danceability'. The set-of-rows with the highest contribution is the one where 'decade'="1990s" (when using the many-to-one partition), obtaining raw influence score of 1.12, whereas the sets of rows associated with "2000s", "2010s", "2020s" obtained contribution of −0.04,−0.35, and −0.055 (resp). Therefore, "1990s" obtains the highest *standardized* contribution of 1.69. This explanation-candidate is indeed a *dominating* one, according to the skyline computation, and therefore returned to the user by FEDEX. Note that the skyline outputs another dominating explanation, having a *lower* interestingness score, yet a *higher* standardized contribution. This is 'decade'="2020" w.r.t. interestingness $I_{danceability}$, which has an interestingness score of 0.04 (as is shown in Example 3.2), and a standardized contribution of 1.7. According to this explanation (visualization omitted), *Songs made in the 2020's are relatively more "danceable" than older songs*.

3.7 Explanations Generation Process

The explanations generation process for an exploratory step Q is detailed in Algorithm 1. It takes as input an exploratory step $Q = (D_{in}, q, d_{out})$ and operates as follows.

Pre-processing: interestingness & row partitioning. First, depending on the type of operation q (filter, group-by, join) we calculate the interestingness scores $I_A(Q)$ using the corresponding function, for each attribute A in the output dataframe d_{out} (Lines 1–2 in the algorithm). Next, we use our row partition techniques (Section 3.5) to split the input dataframes into multiple partitions $\mathcal{R}_1, \mathcal{R}_2, \ldots$, and unify all sets-of-rows into the set $\mathcal{SR}$ (Lines 3–6).

Forming explanation-candidates and calculating standardized contribution. We initialize an empty dictionary EC (Line 7), which will contain the quality scores for each explanation candidate. Then, we iterate over all partitions and all output attributes (Line 8). For each pair (partition, attribute), we iterate over every set-of-rows in the partition (Line 9) and compute its contribution to A (Line 10). If the contribution is positive (Line 11), then both the standardized contribution (w.r.t. the partition, as explained in Section 3.6), and the interestingness score of A are stored in $EC[(R, A)]$ (Line 12).

Calculating the interestingness/contribution skyline. Now that the dictionary EC contains for each explanation candidate (R, A) its (standardized) contribution and interestingness score, we employ the skyline operator (see Section 3.6) which outputs only the dominating explanations, that are not inferior to any other candidates in both contribution or interestingness (Line 13).

Note that to further limit the resulted explanations, one can use, e.g., a weighted average between the interestingness score and the standardized contribution. Namely, given user-defined weights W_I and W_C, let the weighted score of an explanation candidate be $SCORE(R, A) = \frac{W_I \cdot I_A(Q) + W_C \cdot \bar{C}(R,A)}{W_I + W_C}$. We can then use this score to rank the explanations generated by the skyline operator (Line 13) and keep the top k ones.

Generating captioned visualizations for each resulted explanation. For each dominating explanation (R, A), s.t. $R \in \mathcal{R}$ we now (Lines 14–15) generate a corresponding captioned visualization (as illustrated in Figure 2). FEDEX produces a different visualization for each type of interestingness measure: (1) for exceptionality-based explanations, we highlight the deviation in column A, caused by the filter/join operation. This is done, as illustrated in Figure 1a, via a side-by-side bar plot, in which the left-hand side depicts the mean $d_{in}[A]$ value, for all sets-of-rows in the partition $\mathcal{R}$, and the right-hand side depicts the mean $d_{out}[A]$. The chosen set-of-rows for the explanation, R, is colored in green. As for the caption, we use a natural language template and plug-in the attribute name A, and the label of R according to the partition method. The label is set based on the partition approach (Section 3.5). If the chosen partition is numeric-based, the end values of the interval are set as the label. If the partition is many-to-one, the value in column B will be the label. Otherwise, the partition will be frequency-based, and the label is the value itself. We then describe the deviation of R from $d_{in}[A]$ to $d_{out}[A]$ in percentages and multiplications, as illustrated in Figure 2a. (2) Diversity-based explanations highlight the extremity in terms of the aggregated A values in $d_{out}[A]$ obtained by the rows in R, compared to the rest of the sets-of-rows in the partition $\mathcal{R}$. As illustrated in Figure 1b, we use a bar chart to depict the mean aggregated values obtain by each set-of-rows in $\mathcal{R}$, where the mean value of the rows in R is again colored in green. To further emphasize the extremity of R, we also depict the mean $d_{out}[A]$ value, using a horizontal red line. The caption is generated in a similar manner to that of the exceptionality-based explanations, but supports the visualization by emphasizing how far is the value of R_j from the mean, in terms of standard deviations.

Sampling optimization. To reduce explanations generation times, we employ a sampling approach to optimize Algorithm 1. Instead of considering all rows in d_{in} (Lines 1–2), we calculate the interestingness scores over a sample of the rows, obtained using uniform

sampling. All other parts of the algorithm stay intact, and, in particular, the contribution is still computed over all rows. In Section 4, we show that with a relatively small sample size (5K), this approach achieves good accuracy while substantially improving interactivity.

3.8 Customization & Extensions of FEDEX

We next detail several extensions for FEDEX.

General interestingness functions. While FEDEX utilizes established interestingness measures shown useful in exploration tasks [5, 49, 75, 76], FEDEX can take any interestingness function as input, with no required properties (e.g., monotonicity or non-negativity). Examples measures are compactness/coverage [16] for group-by operations, and *surprisingness* [43], as well as learned-based measures (inspired by [45]).

User-specified columns. Furthermore, to give expert users more control, they can specify the columns that they are interested in. Then, FEDEX will only compute the skyline explanations for the chosen columns based on their interestingness and the contribution of the different sets-of-rows to these columns. This is simply done by projecting the input and output dataframes over the user-selected set of attributes, before employing Algorithm 1.

For instance, to explain the filter step in our running example (See Figure 1a) the user could potentially restrict FEDEX to, e.g., the 'danceability' and 'loudness' columns (rather than on all columns), hence obtaining explanations regarding these columns only.

Custom partitioning of rows. Last, as mentioned above, our framework can be further extended with a user-defined partitioning scheme, as long as they comply with Definition 3.8. For instance, users can add a custom partition for date/time columns and group the tuples by month or years; partition a geo-location column by city or state, etc. The new partitions are added to the existing ones, and the system uses them all to generate explanations.

4 EXPERIMENTS

We evaluated the quality and performance of FEDEX on three real-world datasets, and compared them to existing baselines. The results show that the hybrid explanations generated by FEDEX and its sampling optimization version, FEDEX-SAMPLING, are clearer and more interesting than the ones generated by our baselines, and that the sampling optimization employed in FEDEX allows to generate explanations in interactive time while not significantly reducing the accuracy of obtained explanations w.r.t. the skyline.

4.1 Setup, Datasets, Queries, and Baselines

Our experimental study includes 3 real-world datasets and 4 baselines that are based on expert knowledge and previous work.

Implementation. FEDEX is implemented in Python 3.8. It uses Pandas [53] to store and manipulate the database and uses NumPy [30] to compute the explanations and Matplotlib [33] to generate the visualizations. We have made the source code available [71]. The experiments were run on Windows 10 laptop with 16GB RAM and 1.9 GHz Quad-Core Intel Core i7 processor.

Datasets. We have used the following 3 datasets:

1. Spotify [20]: containing information about tracks (e.g., duration, popularity, danceability) and artists (e.g., genres and popularity). It comprises a single table of 174,389 rows and 20 columns.

2. Credit Card Customers [19]: containing customer details such as age, gender, education level, and credit card category. It has a single table with 10,127 rows and 21 columns.

3. Products and Sales [55]: consisting of two tables: a Products table (9,977 rows, 16 columns) with information about beverage products (e.g., name, vendor, price), and a Sales table (3,049,913 rows and 17 columns) that contains a record of all the sales of products in a given chain (e.g., store id, quantity, date). The join view of the two tables has 3,049,913 rows. For our scalability experiments (Section 4.3) we uniformly sampled additional 6,950,087 rows (i.e., duplicates) to get a view with exactly 10M rows.

Further note that the datasets all contain skewed columns. In total, over 31 of the columns are highly skewed, and over 41 of the columns are moderately skewed. In particular, Fisher-Pearson standardized moment coefficient [14] was high for multiple columns in each dataset. For example, the top-1 column had a measurement of 10.16, 2.06, and 205.89 for the Spotify, Credit Card Customers, and Products and Sales datasets, respectively.

Queries. We have composed 5 filter/join queries for each dataset, and 5 group-by queries for each dataset. The tables containing the join, filter, and group-by queries, along with their reference numbers, can be found in Appendix A .

Baselines and optimized version. Our experiments included a comparison of FEDEX to the following alternatives:

1. SEEDB [76]: The solution of Vartak et. al. (implemented in [58]) automatically generates visualizations that are meant to emphasize trends in a given view stemming from a query over a source view.

2. RATH [72]: RATH (implemented in implementation is taken from [59]) automatically generates the top-k insightful visualizations using a single score function for all operations that is both applicable to different types of insights and fair across different types of insights (for a detailed comparison see Section 2).

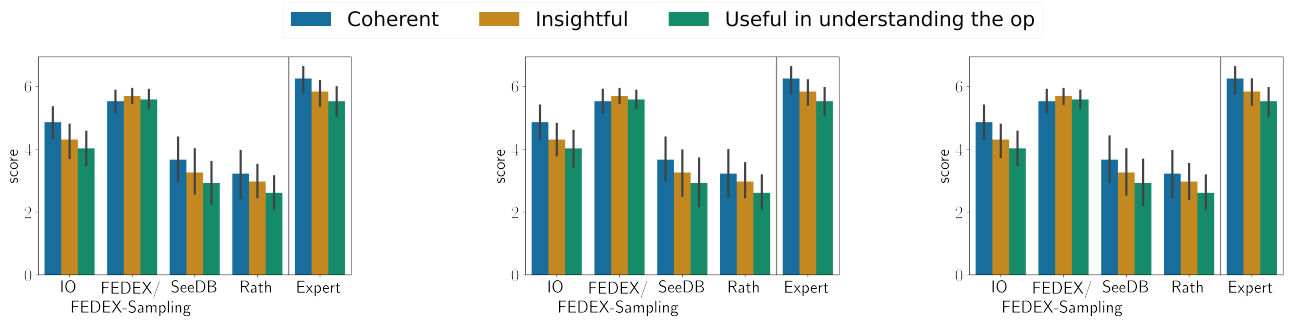
3. IO: The **I**nterestingness **O**nly baseline is based on previous work [79] where the influence of an attribute is measured by checking the difference in interestingness of an attribute in d_{out} w.r.t. D_{in}.

4. EXPERT: As part of our user study (Section 4.2), we have asked three experts to manually formulate their own explanations for each of the three datasets, each represented in its own notebook. These explanations consisted of a detailed textual description. The experts had access to the query, the dataset, and the resulting dataframe.

5. FEDEX-SAMPLING: The optimized version of FEDEX, where uniform sampling of the rows is used to compute the interestingness. After performing experiments to choose a reasonable sample size (see the paragraph "Accuracy of FEDEX-SAMPLING" in Section 4.3), we have concluded that a sample of 5K rows delivers good accuracy.

4.2 User Studies

We detail two user studies that were performed to evaluate the quality of the explanations generated by FEDEX and FEDEX-SAMPLING. For our user studies, we have used separate notebooks for each dataset and the appropriate queries. The maximal number of explanations presented to users (the size of the skyline set) over all queries in both studies was 2.



(a) Results for the Credit Card dataset (b) Results for the Spotify dataset (c) Results for the Products dataset

Figure 3: User study results on the three datasets

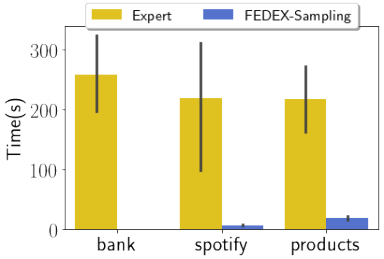


Figure 4: Explanation generation time for the user study in Figure 3

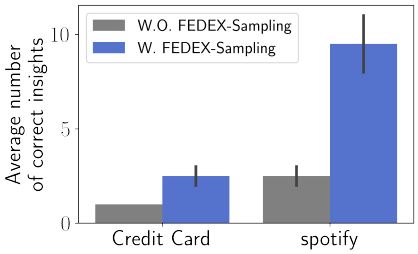


Figure 5: Interactive user study

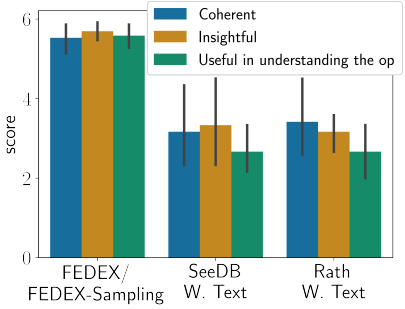
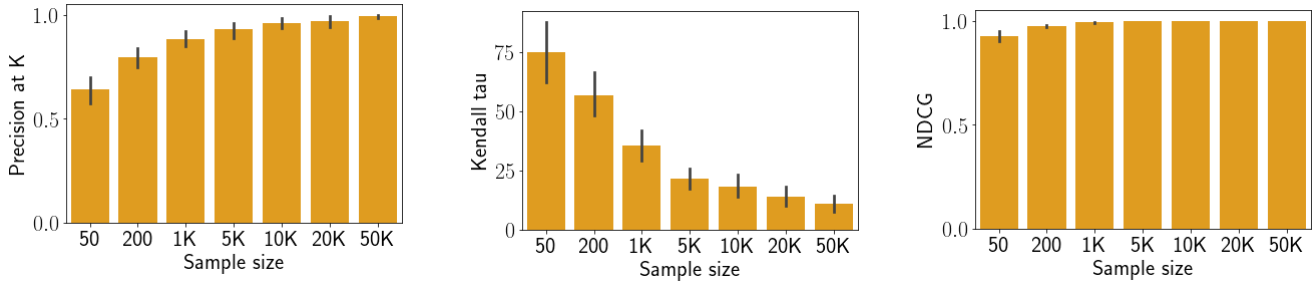


Figure 6: User study results for baselines augmented with NL explanations

Comparison to existing baselines. We have performed a user study with the goal of evaluating the quality of the explanations generated by FEDEX compared to existing baselines and expert-generated explanations. that included 25 participants. The users had different backgrounds; 17 of the users were non-experts with minimal SQL knowledge and 8 were CS graduate students. The study was based on three notebooks, one for each of the three datasets. The notebooks also included details regarding the given dataset and the goal of the exploration session (the notebooks can be found in the FEDEX repository [71]). The Spotify notebook contained the filter/join queries 6, 7 and the group-by queries 21, 22 (see Appendix A); the Credit Card Customers notebook contained the filter/join queries 11, 12, 13 and group-by query 27 from Appendix A; and the Product and Sales notebook contained the filter/join queries 1, 5 and group-by queries 16, 17, 18. The Spotify and Credit Card notebooks were each shown to 9 users, and the notebook with the Products and Sales session was shown to 10 users. Each notebook also included the explanations generated by SeeDB, Rath, IO, Expert, and FEDEX (the explanations computed by FEDEX-Sampling were identical to those computed by FEDEX, i.e. the skyline set was identical). For the Expert baseline, we have asked 3 experts to analyze the notebooks and generate an explanation for each operation manually. Figure 4 shows the time it took the experts to generate the explanations compared to the generation time of FEDEX. Naturally, the explanation generation time of the experts was substantially larger than for FEDEX.

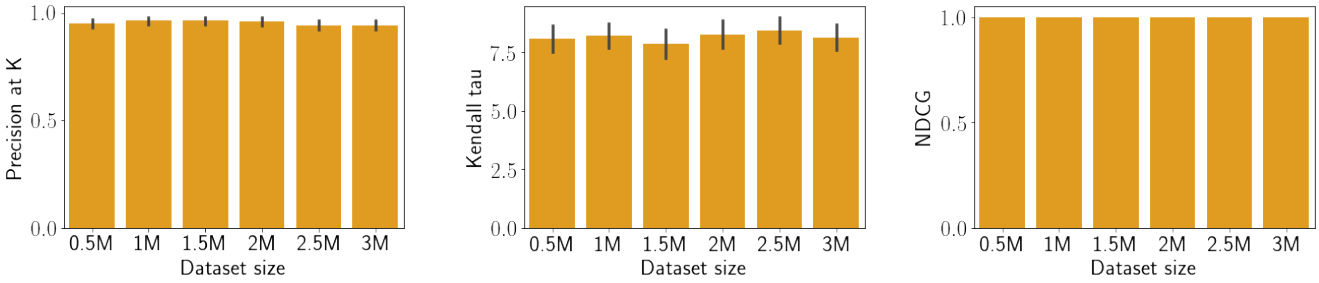
For each user, we presented the query, the original input dataset and the query output. Then, we showed the user up to five explanations (the SeeDB baseline could not generate explanations for group-by queries as it compares d_{in} and d_{out}, but in group-by operations the input and output columns are different). Users were asked to grade the explanations on a scale of 1–7 w.r.t. *coherency* (is the explanation easy to understand?), *insight level* (does the explanation provide an interesting insight?), and *usefulness* in understanding the operation (does the explanation assist in understanding the EDA operation results?).

For all three datasests, we observed that users have generally preferred the Expert explanations. The average score for these explanations for coherency, interestingness level, and relevance was 6.33, 5.5, 5.33 respectively across all three datasets. From the automatically generated explanations, those generated by FEDEX were clearly preferred. The average score of the explanations generated for the Spotify dataset across the different measures is 5.1, whereas the average score of IO, SeeDB, and Rath for this dataset is 3.8, 3, 2.8 respectively. For the Credit Card Customers dataset, the average score for FEDEX across the different measures is 5.6, compared to an average score of 4.4, 3.3, 2.9 for IO, SeeDB, and



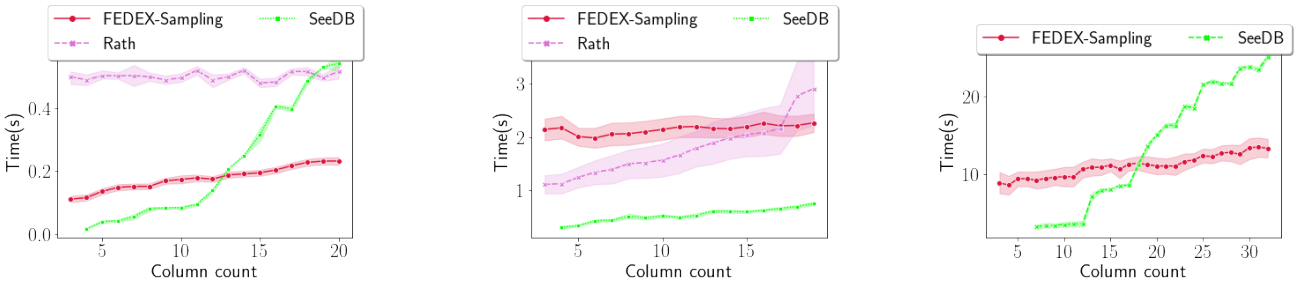
(a) Precision@k for FEDEX-SAMPLING (b) Kendall-Tau distance for FEDEX-SAMPLING (c) nDCG for FEDEX-SAMPLING

Figure 7: Accuracy results for FEDEX-SAMPLING averaged over the Spotify and Products datasets with all relevant filter and join queries in Appendix A



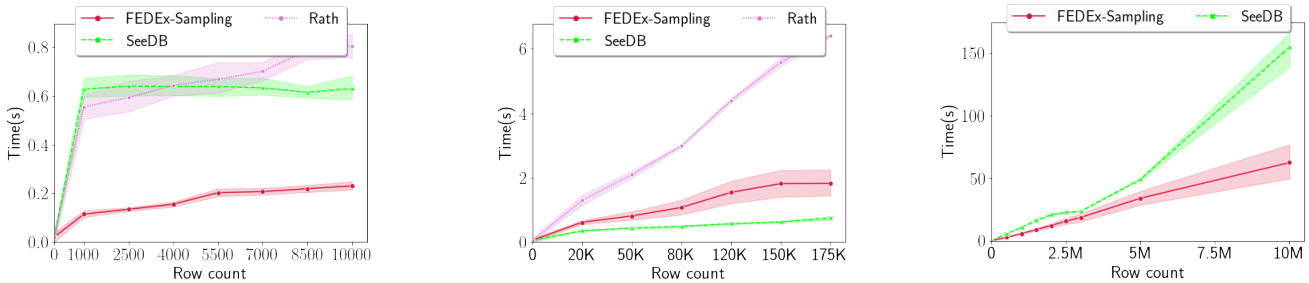
(a) Precision@k for FEDEX-SAMPLING (b) Kendall-Tau distance for FEDEX-SAMPLING (c) nDCG for FEDEX-SAMPLING

Figure 8: Accuracy results for FEDEX-SAMPLING for the Products datasets with all relevant filter and join queries in Appendix A



(a) Credit Card Customers dataset (b) Spotify dataset (c) Products and Sales dataset

Figure 9: Runtime as a function of column number for all three datasets for FEDEX-SAMPLING and the baselines averaged over the filter and join queries from Appendix A



(a) Credit Card Customers dataset (b) Spotify dataset (c) Products and Sales dataset

Figure 10: Runtime as a function of row number for all three datasets for FEDEX and FEDEX-SAMPLING averaged over the filter and join queries from Appendix A

Rath, respectively. Finally, for the Products and Sales dataset, the average score of the explanations generated by fedex across the different measures is 5.3, whereas the average score of IO and SeeDB explanations for this dataset are 3.2, 3.8, respectively. Rath timed out for every query over this dataset (after more than 3 hours of computation), and thus its explanations were excluded. It should be noted that fedex generates hybrid explanations that include both visualization and text, making them easily understandable. This may explain the users' preference for these explanations over the explanations generated by some of the baselines. Interestingly, the scores obtained by fedex were very close to these of Expert on the Products and Sales Notebook. fedex showed such high scores for this notebook due to the join operation; Expert did not explain this join while fedex noticed a change in the distribution and pointed that out to the user. fedex generated more insightful explanations (an average score of 5.4 compared to an average of 5.1 for Expert) and slightly more relevant to the task (average of 5.1 compared to 5 for Expert); in contrast, the Expert baseline got higher coherency scores with an average of 6.3 compared to an average of 5.4 for fedex.

Comparison to unassisted EDA. We have compared standard unassisted EDA with EDA assisted by fedex/fedex-Sampling, using the Spotify and Credit Card datasets. The task for the Spotify dataset was to find which songs are more popular and why; the task for the Credit Card Customers dataset was to find out why people leave the service and how can we anticipate it. The study included 8 participants who were asked to identify as many insights as possible that are related to the task, when presented with an empty notebook. Then, an expert was asked to denote for each user-generated insight, whether this insight is correct and directly related to the task or not. For example, for the Spotify dataset, "acoustic songs (with acousticness > 0.5) are usually less popular" is a correct insight while "songs from the 90s are louder than other decades" is an unrelated insight. We counted the number of insights gathered by the participants over 10 minutes. The average number of insights for the two datasets is shown in Figure 5. The average numbers were 2.5 (9.5) with fedex-Sampling and 1 (2.5) without it for the Credit Card (Spotify) dataset. The results clearly show the benefit of using fedex and fedex-Sampling and indicate that fedex is able to assist users in gaining insights about the EDA task.

Comparison to augmented baselines. We have further added textual explanations to the SeeDB and Rath baselines (in addition to their 'organic' visualizations) and performed an additional user study in which the output of all baselines includes both a visualization and a caption. In the study, four participants were asked to analyze the Credit Card Customers dataset and its notebook from the first user study (including the five relevant queries that can be found in Appendix A). Since the quality of automatic captioning methods may vary, we have asked an expert to manually devise a textual description for the visualizations generated by SeeDB and Rath. The results in Figure 6 indicate that even with experts-generated textual explanations for the baselines, fedex is able to generate explanations that are significantly more coherent, insightful, and useful. In particular, the scores were 5.52 for fedex, 3.17 for SeeDB augmented with textual explanations, and 3.42 for Rath augmented with textual explanations.

4.3 Simulated Experiments

We first measure the accuracy of the explanations generated by fedex-Sampling. After establishing that fedex-Sampling is indistinguishable from fedex in terms of accuracy, we measure its runtimes compared to the baselines when varying the parameters of the database and further measure the trade-off between the number of sets-of-rows and accuracy. In the following experiments, the number of sets-of-rows was set to either 5 or 10 (fedex tries to divide the values into both 5 and 10 sets-of-rows and computes the skylines for all options). For each measured point, we run fedex and/or fedex-Sampling (depending on the experiment) three times for each one of the relevant queries.

Accuracy of fedex-Sampling. In this set of experiments, we have averaged over the join/filter queries 1-10 and group-by queries 16-25 from Appendix A, for the Products and Spotify datasets. Figure 7a depicts the precision@k [64] of fedex-Sampling w.r.t. the output of fedex (k was set to 3 since in most cases the number of explanations in the skyline set was $\leq$ 3), used as the ground truth. The measurements were performed over the Spotify and Products datasets, as the Credit Card dataset is too small. The average precision over the different queries is very high, reaching over 93%, 96%, 97%, and 99% for sample sizes of 5K, 10K, 20K, and 50K, respectively. This suggests that a relatively small sample can accurately predict the set of skyline explanations. Figure 7b shows the Kandall-Tau distance [37] between the ground truth results and the results obtained by fedex-Sampling, when considering all explanations in the skyline. The distance gradually decreases and falls from 74.8 at sample size 50 to 10.8 for sample size 50K. In particular, for a 5K sample the distance is 21.6 – lower than the mean value of 33.1 among all sample sizes. Figure 7c shows the nDCG score [35] of fedex-Sampling. The starting point of the score is already high (92.6%), and it further increases gradually with the sample size where the most significant increase is observable as we move from sample sizes of 50 to 200 (92% to 97%, respectively). In particular, for a sample size of 5K, the nDCG score is 99.8%.

We have further measured the change in accuracy of fedex-Sampling with a sample size of 5K for a varying number of rows for the Products and Sales dataset, which is the largest dataset out of the three. Our results are shown in Figure 8. All three subfigures show that the accuracy remains high for all sets of rows w.r.t. all three metrics. In particular, for 3M rows, the precision@k was 0.942, the Kendall-Tau distance was 8.1, and the nDCG value was 0.9985. Kendall-Tau has minor fluctuations since the different columns have very close interestingness scores, and Kendall-Tau is sensitive to this, as a slight change in interestingness can lead to a change in the ranking of the explanations. *In light of these results, in our scalability experiments, we have examined fedex-Sampling with a fixed sample size of 5K rows.*

Last, while an inaccurate interestingness calculation by fedex-Sampling may result in sub-optimal explanations, recall from Section 4.2 that fedex-Sampling produced the exact same output as fedex. In general, there are no guarantees that the skyline will be the same for fedex and fedex-Sampling. To explain the reason for the lack of change in the skyline in Section 4.2, recall that the skyline set of explanations is determined via the following two

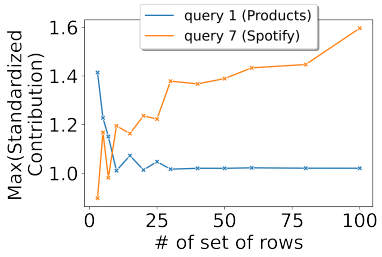


Figure 11: Contribution score (see Section 3.5) for varying number of sets-of-rows for query 1 over Products and Sales, and query 7 over Spotify (see Appendix A for the queries)

steps: (1) computing interestingness and for the top-k most interesting columns (2) computing the contribution for all sets of rows. Note that Figure 7 measures only the accuracy for step (1), which is where we see a small loss by the sampling. This loss was not significant enough to affect the final skyline set in Section 4.2.

Runtime analysis for varying column number. In this set of experiments, we have examined the runtime as a function of the number of columns in the dataset. The experiment was run over all three datasets, where we averaged over the join/filter queries the 1-15 in Appendix A. We have set the number of rows to its maximum size for each dataset and have gradually increased the number of columns. We always included the attribute that the query needs (e.g., X) and the attribute with the highest interestingness score (e.g., Y). Then, we perform a random permutation on the attributes and add the columns in a fixed order (if the permutation is $A_1, A_2, \ldots, A_m$, we first check X, Y, A_1, and then X, Y, A_1, A_2 and so on). Figure 9 depicts the execution time of FEDEX-SAMPLING compared to the baselines SEEDB and RATH. Figure 9a shows the runtime for an increasing number of columns for the Credit Card dataset. For 20 columns, the runtime of FEDEX-SAMPLING was 0.23s, whereas for SEEDB and RATH it was 0.54s and 0.52s, respectively. In Figure 9b, at the maximum number of columns for the Spotify dataset (20 columns), the runtime of FEDEX-SAMPLING was 2.27s, while the runtime for SEEDB and RATH was 0.75s, 2.9s, respectively. SEEDB performs better for this specific dataset because it contains mostly numeric attributes, and SEEDB counts on both categorical values for grouping and numeric attributes for aggregations. The lack of categorical attributes reduces the number of possible views and reduces the runtime. For the Products dataset (Figure 9c) with 33 columns, the runtime was 13.3s for FEDEX-SAMPLING and 25.1 for SEEDB. RATH was not able to run on 3M rows due to high memory usage (taking more than 17GB, resulting in an 'out of memory' error) and long processing times (average of 50% CPU usage for more than 10 minutes), therefore omitted from the graph. These results suggest that FEDEX-SAMPLING with a sample size of 5K performs better than SEEDB and RATH on datasets with a moderate to large schema size when interactive performance is desired.

Execution time analysis for varying row number. In this set of experiments, we have computed average execution times separately for the group-by queries and filter/join queries. Again, the sample size of FEDEX-SAMPLING was set to 5K. Here, for Figure 10c, we sampled additional rows to increase the size of the view to a maximum of 10M.

Figure 10a shows that FEDEX-SAMPLING performs better than SEEDB and RATH in most cases as the data size increases with a relatively small increase in execution time as the number of rows grows. For 10K rows, the execution time is 0.23s, as opposed to an execution time of 0.63s, 0.81s for SEEDB and RATH, respectively. In Figure 10b, for 174, 389 rows, the execution time of FEDEX, SEEDB, and RATH were 1.81s, 0.7s, 6.4s, respectively. SEEDB slightly outperforms FEDEX-SAMPLING for this dataset for the same reason detailed in the previous paragraph regarding Figure 9b. In Figure 10c, for 10M rows, the execution times of FEDEX-SAMPLING and SEEDB were 62.4s, 154.9s, respectively. RATH was not able to run on these data sizes (as in the explanation for Figure 9, due to high memory usage and large processing times), so it does not appear in the figure. This experiment shows that FEDEX-SAMPLING performs better or in a comparative manner with the examined baselines in terms of runtime for large data sizes.

Accuracy for varying sets-of-rows sizes. Figure 11 shows the contribution score w.r.t. the numbers of sets-of-rows (see Section 3.7) for queries 3 and 7 shown in Appendix A for the Products and Sales and Spotify datasets, respectively. We have chosen two specific queries to ensure that the column will remain constant and that only the number of sets-of-rows will vary. There is no clear trend in the results: the optimal number of sets-of-rows w.r.t. contribution depends on the query and the values of the chosen attribute. We note that a partition to large number of sets-of-rows resulted in an overloaded and less clear visualization that includes many ticks on the X-axis. Thus, for understandable explanations, choosing a lower number of sets-of-rows appears to be a better strategy (in the user study we set it to 5 or 10).

5 CONCLUSIONS AND FUTURE WORK

We have presented FEDEX, a system that provides coherent explanations for data exploration steps. The resulted explanations assist users in understanding what is interesting in the results of their exploration steps and gather actionable insights. Our explanations are based on a novel approach, that holistically examines an exploration step and calculates the contribution of sets-of-rows from the *input dataframe* to the interestingness score of a column in the *output dataframe*. Those sets-of-rows that significantly contribute to highly interesting attributes are transformed to coherent explanations in the form of visualization with a natural language caption. We have performed an extensive evaluation of our approach that shows its usefulness over real-world datasets and tasks. For future work, we intend to study different ways of measuring the contribution of row sets, such as causal responsibility. Another intriguing direction is to develop new measures that capture additional interestingness facets, as well as to study dedicated optimization techniques for interestingness calculations.

ACKNOWLEDGMENTS

This research has been partially funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 804302), the Israel Science Foundation, BSF - the Binational US-Israel Science foundation and the Tel Aviv University Data Science center, and the NSF awards IIS-1703431, IIS-1552538, IIS-2008107.

REFERENCES

[1] Deepak Agarwal, Dhiman Barman, Dimitrios Gunopulos, Neal E Young, Flip Korn, and Divesh Srivastava. 2007. Efficient and effective explanation of change in hierarchical summaries. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. 6–15.

[2] Yael Amsterdamer, Daniel Deutch, and Val Tannen. 2011. Provenance for aggregate queries. In *Proceedings of the thirtieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. 153–164.

[3] Abolfazl Asudeh, Hosagrahar Visvesvaraya Jagadish, You Wu, and Cong Yu. 2020. On detecting cherry-picked trendlines. *Proceedings of the VLDB Endowment* 13, 6 (2020), 939–952.

[4] Zhifeng Bao, Yong Zeng, HV Jagadish, and Tok Wang Ling. 2015. Exploratory keyword search with interactive input. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. 871–876.

[5] Ori Bar El, Tova Milo, and Amit Somech. 2020. Automatically generating data exploration sessions using deep reinforcement learning. In *SIGMOD*. 1527–1537.

[6] Daniel W Barowy, Dimitar Gochev, and Emery D Berger. 2014. Checkcell: Data debugging for spreadsheets. *ACM SIGPLAN Notices* 49, 10 (2014), 507–523.

[7] Arthur G Bedeian and Kevin W Mossholder. 2000. On the use of the coefficient of variation as a measure of diversity. *Organizational Research Methods* 3, 3 (2000), 285–297.

[8] Rachel Behar and Sara Cohen. 2020. Optimal End-Biased Histograms for Hierarchical Data. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 3261–3264.

[9] Rachel Behar and Sara Cohen. 2020. Optimal Histograms with Outliers.. In *Extending database technology (EDBT)*. 181–192.

[10] Ramon Bespinyowong, Wei Chen, HV Jagadish, and Yuxin Ma. 2016. ExRank: An exploratory ranking interface. *PVLBD* 9, 13 (2016), 1529–1532.

[11] Nicole Bidoit, Melanie Herschel, and Aikaterini Tzompanaki. 2015. Efficient computation of polynomial explanations of why-not questions. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. 713–722.

[12] Nicole Bidoit, Melanie Herschel, and Katerina Tzompanaki. 2014. Query-based why-not provenance with nedexplain. In *Extending database technology (EDBT)*.

[13] Stephan Borzsony, Donald Kossmann, and Konrad Stocker. 2001. The skyline operator. In *Proceedings 17th international conference on data engineering*. IEEE, 421–430.

[14] Stan Brown. 2011. Measures of shape: Skewness and kurtosis.

[15] P. Buneman, S. Khanna, and W.C. Tan. 2001. Why and Where: A Characterization of Data Provenance. In *ICDT*. 316–330.

[16] Varun Chandola and Vipin Kumar. 2007. Summarization–compressing data into an informative representation. *Knowledge and Information Systems* 12, 3 (2007), 355–378.

[17] Adriane Chapman and HV Jagadish. 2009. Why not?. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*. 523–534.

[18] Weiwei Cui, Xiaoyu Zhang, Yun Wang, He Huang, Bei Chen, Lei Fang, Haidong Zhang, Jian-Guan Lou, and Dongmei Zhang. 2019. Text-to-viz: Automatic generation of infographics from proportion-related natural language statements. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 906–916.

[19] Credit Card Customers Dataset. 2021. https://www.kaggle.com/sakshigoyal7/credit-card-customers/tasks?taskId=2729.

[20] Spotify Dataset. 2021. https://www.kaggle.com/mrmorj/dataset-of-songs-in-spotify.

[21] Tijl De Bie. 2013. Subjective interestingness in exploratory data mining. In *Advances in Intelligent Data Analysis XII*. Springer, 19–31.

[22] Daniel Deutch, Nave Frost, and Amir Gilad. 2017. Provenance for Natural Language Queries. *PVLDB* 10, 5 (2017), 577–588.

[23] Daniel Deutch and Amir Gilad. 2016. QPlain: Query by explanation. In *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*. IEEE, 1358–1361.

[24] Kyriaki Dimitriadou, Olga Papaemmanouil, and Yanlei Diao. 2016. AIDE: An Active Learning-based Approach for Interactive Data Exploration. *TKDE* (2016).

[25] Rui Ding, Shi Han, Yong Xu, Haidong Zhang, and Dongmei Zhang. 2019. Quicklnsights: Quick and automatic discovery of insights from multi-dimensional data. In *Proceedings of the 2019 International Conference on Management of Data*. 317–332.

[26] Xin Luna Dong and Divesh Srivastava. 2013. Compact explanation of data fusion decisions. In *Proceedings of the 22nd international conference on World Wide Web*. 379–390.

[27] Anna Fariha, Ashish Tiwari, Arjun Radhakrishna, and Sumit Gulwani. 2020. Extune: Explaining tuple non-conformance. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*. 2741–2744.

[28] Liqiang Geng and Howard J Hamilton. 2006. Interestingness measures for data mining: A survey. *ACM Computing Surveys (CSUR)* 38, 3 (2006), 9–es.

[29] T.J. Green, G. Karvounarakis, and V. Tannen. 2007. Provenance semirings. In *PODS*. 31–40.

[30] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. 2020. Array programming with NumPy. *Nature* 585, 7825 (2020), 357–362.

[31] Robert J Hilderman and Howard J Hamilton. 2013. *Knowledge discovery and measures of interest*. Vol. 638. Springer Science & Business Media.

[32] Hao Huang, Qian Yan, Wei Lu, Huaizhong Lin, Yunjun Gao, and Lei Chen. 2019. LERI: Local Exploration for Rare-Category Identification. *IEEE Transactions on Knowledge and Data Engineering* 32, 9 (2019), 1761–1772.

[33] J. D. Hunter. 2007. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering* 9, 3 (2007), 90–95.

[34] Ihab F Ilyas, Volker Markl, Peter Haas, Paul Brown, and Ashraf Aboulnaga. 2004. CORDS: Automatic discovery of correlations and soft functional dependencies. In *Proceedings of the 2004 ACM SIGMOD international conference on Management of data*. 647–658.

[35] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.* 20, 4 (2002), 422–446.

[36] Manas Joglekar, Hector Garcia-Molina, and Aditya Parameswaran. 2017. Interactive data exploration with smart drill-down. *IEEE Transactions on Knowledge and Data Engineering* 31, 1 (2017), 46–60.

[37] Maurice George Kendall. 1948. Rank correlation methods. (1948).

[38] Mary Beth Kery, Marissa Radensky, Mahima Arya, Bonnie E John, and Brad A Myers. 2018. The story in the notebook: Exploratory data science using a literate programming tool. In *CHI*.

[39] Nodira Khoussainova, YongChul Kwon, Magdalena Balazinska, and Dan Suciu. 2010. SnipSuggest: Context-Aware Autocompletion for SQL. *Proc. VLDB Endow.* 4, 1 (2010), 22–33.

[40] Marie Le Guilly, Jean-Marc Petit, Vasile-Marian Scuturici, and Ihab F Ilyas. 2019. ExplIQuE: Interactive Databases Exploration with SQL. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 2877–2880.

[41] Doris Jung-Lin Lee, Dixin Tang, Kunal Agarwal, Thyne Boonmark, Caitlyn Chen, Jake Kang, Ujjaini Mukhopadhyay, Jerry Song, Micah Yong, Marti A Hearst, et al. 2021. Lux: Always-on Visualization Recommendations for Exploratory Data Science. *arXiv preprint arXiv:2105.00121* (2021).

[42] 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 *Proceedings of the 2021 International Conference on Management of Data*. 1051–1063.

[43] Bing Liu, Wynne Hsu, Lai-Fun Mun, and Hing-Yan Lee. 1999. Finding interesting patterns using user expectations. *IEEE Transactions on Knowledge and Data Engineering* 11, 6 (1999), 817–832.

[44] Xiufeng Liu, Lukasz Golab, and Ihab F Ilyas. 2015. SMAS: A smart meter data analytics system. In *2015 IEEE 31st International Conference on Data Engineering*. IEEE, 1476–1479.

[45] Yuyu Luo, Xuedi Qin, Nan Tang, and Guoliang Li. 2018. DeepEye: Towards Automatic Data Visualization. ICDE.

[46] Ken McGarry. 2005. A survey of interestingness measures for knowledge discovery. *The knowledge engineering review* 20, 1 (2005), 39–61.

[47] Alexandra Meliou, Wolfgang Gatterbauer, Joseph Y Halpern, Christoph Koch, Katherine F Moore, and Dan Suciu. 2010. Causality in databases. *IEEE Data Engineering Bulletin* 33 (2010), 59–67.

[48] Zhengjie Miao, Qitian Zeng, Boris Glavic, and Sudeepa Roy. 2019. Going beyond provenance: Explaining query answers with pattern-based counterbalances. In *Proceedings of the 2019 International Conference on Management of Data*. 485–502.

[49] Tova Milo, Chai Ozeri, and Amit Somech. 2019. Predicting "What is Interesting" by Mining Interactive-Data-Analysis Session Logs. In *EDBT*. 456–467.

[50] Tova Milo and Amit Somech. 2018. Next-Step Suggestions for Modern Interactive Data Analysis Platforms. In *KDD*. ACM, 576–585.

[51] Katsiaryna Mirylenka, Graham Cormode, Themis Palpanas, and Divesh Srivastava. 2015. Conditional heavy hitters: detecting interesting correlations in data streams. *The VLDB Journal* 24, 3 (2015), 395–414.

[52] Katsiaryna Mirylenka, Themis Palpanas, Graham Cormode, and Divesh Srivastava. 2013. Finding interesting correlations with conditional heavy hitters. In *2013 IEEE 29th International Conference on Data Engineering (ICDE)*. IEEE, 1069–1080.

[53] The pandas development team. 2020. *pandas-dev/pandas: Pandas*. https://doi.org/10.5281/zenodo.3509134

[54] Judea Pearl et al. 2009. Causal inference in statistics: An overview. *Statistics surveys* 3 (2009), 96–146.

[55] Products and Sales dataset. 2018. https://data.world/classrooms/guide-to-data-analysis-with-sql-and-datadotworld.

[56] Li Qian, Jinyang Gao, and HV Jagadish. 2015. Learning user preferences by adaptive pairwise comparison. *Proceedings of the VLDB Endowment* 8, 11 (2015), 1322–1333.

[57] Xuedi Qin, Yuyu Luo, Nan Tang, and Guoliang Li. 2020. Making data visualization more efficient and effective: a survey. *The VLDB Journal* 29, 1 (2020), 93–117.

[58] Rath repository. 2018. https://github.com/snknitin/-SeeDB.

[59] Rath repository. 2022. https://github.com/Kanaries/Rath.

[60] Sheldon M Ross. 2004. *Introduction to probability and statistics for engineers and scientists*. Elsevier.

[61] Sudeepa Roy and Dan Suciu. 2014. A formal approach to finding explanations for database queries. In *SIGMOD*, Curtis E. Dyreson, Feifei Li, and M. Tamer Özsu (Eds.). 1579–1590.

[62] Sunita Sarawagi. 2001. User-cognizant multidimensional analysis. *The VLDB Journal* 10, 2 (2001), 224–239.

[63] Sunita Sarawagi, Rakesh Agrawal, and Nimrod Megiddo. 1998. Discovery-driven exploration of OLAP data cubes. In *EDBT*.

[64] Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2008. *Introduction to information retrieval*. Vol. 39. Cambridge University Press Cambridge.

[65] Mariia Seleznova, Behrooz Omidvar-Tehrani, Sihem Amer-Yahia, and Eric Simon. 2020. Guided exploration of user groups. *Proceedings of the VLDB Endowment (PVLDB)* 13, 9 (2020), 1469–1482.

[66] Thibault Sellam and Martin Kersten. 2016. Cluster-driven navigation of the query space. *IEEE Transactions on Knowledge and Data Engineering* 28, 5 (2016), 1118–1131.

[67] Masoumeh Shafieinejad, Florian Kerschbaum, and Ihab F Ilyas. 2021. PCOR: Private Contextual Outlier Release via Differentially Private Search. In *Proceedings of the 2021 International Conference on Management of Data*. 1571–1583.

[68] 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* 27, 2 (2020), 453–463.

[69] Manish Singh, Michael J Cafarella, and HV Jagadish. 2016. DBExplorer: Exploratory Search in Databases. *EDBT* (2016).

[70] 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* 25, 1 (2018), 672–681.

[71] FEDEX Repository. 2022. https://github.com/TAU-DB/FEDEx.

[72] Bo Tang, Shi Han, Man Lung Yiu, Rui Ding, and Dongmei Zhang. 2017. Extracting top-k insights from multi-dimensional data. In *Proceedings of the 2017 ACM International Conference on Management of Data*. 1509–1524.

[73] Balder ten Cate, Cristina Civili, Evgeny Sherkhonov, and Wang-Chiew Tan. 2015. High-level why-not explanations using ontologies. In *Proceedings of the 34th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*. 31–43.

[74] Saravanan Thirumuruganathan, Mahashweta Das, Shrikant Desai, Sihem Amer-Yahia, Gautam Das, and Cong Yu. 2012. Maprat: Meaningful explanation, interactive exploration and geo-visualization of collaborative ratings. *Proceedings of the VLDB Endowment (PVLDB)* 5, 12 (2012), 1986–1989.

[75] Matthijs van Leeuwen. 2010. Maximal exceptions with minimal descriptions. *Data Mining and Knowledge Discovery* 21, 2 (2010), 259–276.

[76] Manasi Vartak, Sajjadur Rahman, Samuel Madden, Aditya G. Parameswaran, and Neoklis Polyzotis. 2015. SEEDB: Efficient Data-Driven Visualization Recommendations to Support Visual Analytics. *Proc. VLDB Endow.* 8, 13 (2015), 2182–2193.

[77] 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* 26, 1 (2019), 895–905.

[78] Kanit Wongsuphasawat, Dominik Moritz, Anushka Anand, Jock Mackinlay, Bill Howe, and Jeffrey Heer. 2016. Voyager: Exploratory analysis via faceted browsing of visualization recommendations. *TVCG* (2016).

[79] Eugene Wu and Samuel Madden. 2013. Scorpion: Explaining Away Outliers in Aggregate Queries. *Proc. VLDB Endow.* 6, 8 (2013), 553–564.

[80] Cong Yan and Yeye He. 2020. Auto-Suggest: Learning-to-Recommend Data Preparation Steps Using Data Science Notebooks. In *SIGMOD*. 1539–1554.

[81] Fabian Colque Zegarra, Juan C Carbajal Ipenza, Behrooz Omidvar-Tehrani, Viviane P Moreira, Sihem Amer-Yahia, and João LD Comba. 2020. Visual exploration of rating datasets and user groups. *Future Generation Computer Systems* 105 (2020), 547–561.

A QUERIES FOR THE EXPERIMENTS

Table 2: Join and filter queries used in the experiments (used with the exceptionality measure from Section 3.2)

Num.	Dataset and Type	Query
1	Products (J)	SELECT * FROM products INNER JOIN sales ON products.item=sales.item;
2	Products (J)	SELECT * FROM counties INNER JOIN sales ON counties.county=sales.county;
3	Products (J)	SELECT * FROM counties INNER SELECT * FROM stores INNER JOIN sales ON stores.store=sales.store;
4	Products (F)	SELECT * FROM products_sales WHERE sales_liter_size ≤ 500;
5	Products (F)	SELECT * FROM products_sales WHERE sales_pack == 12;
6	Spotify (F)	SELECT * FROM spotify WHERE popularity > 65;
7	Spotify (F)	SELECT * FROM spotify WHERE year > 1990;
8	Spotify (F)	SELECT * FROM spotify WHERE loudness > -12;
9	Spotify (F)	SELECT * FROM spotify WHERE duration_minutes < 3;
10	Spotify (F)	SELECT * FROM spotify WHERE tempo > 100;
11	Bank (F)	SELECT * FROM Bank WHERE Attrition_Flag != "Existing Customer";
12	Bank (F)	SELECT * FROM [SELECT * FROM Bank WHERE Attrition_Flag != 'Existing Customer'] WHERE Total_Count_Change_Q4_vs_Q1 > 0.75;
13	Bank (F)	SELECT * FROM Bank WHERE Months_Inactive_Count_Last_Year > 2;
14	Bank (F)	SELECT * FROM Bank WHERE Customer_Age < 30;
15	Bank (F)	SELECT * FROM Bank WHERE Income_Category == “Less than $40K”;

Table 3: Group-by queries used in the experiments (used with the diversity measure from Section 3.2)

Num.	Dataset and Type	Query
16	Products (GB)	SELECT count(item) FROM products_sales GROUP BY sales_vendor;
17	Products (GB)	SELECT count(item) FROM products_sales GROUP BY sales_county, sales_category_name;
18	Products (GB)	SELECT count(item) FROM products_sales GROUP BY products_sales_pack;
19	Products (GB)	SELECT mean(sales_total), mean(sales_pack) FROM products_sales GROUP BY sales_bottle_quantity;
20	Products (GB)	SELECT mean(products_bottle_size) FROM products_sales GROUP BY products_pack, products_inner_pack;
21	Spotify (GB)	SELECT mean(popularity), max(popularity), min(popularity) FROM spotify GROUP BY year;
22	Spotify (GB)	SELECT mean(danceability), max(danceability), mean(instrumentalness), max(instrumentalness), mean(liveness) FROM spotify GROUP BY year;
23	Spotify (GB)	SELECT mean(danceability), mean(popularity) FROM spotify GROUP BY key;
24	Spotify (GB)	SELECT max(duration_minutes), mean(duration_minutes) FROM spotify GROUP BY decade;
25	Spotify (GB)	SELECT mean(loudness), mean(liveness), mean(tempo) FROM spotify GROUP BY mode, key;
26	Bank (GB)	SELECT mean(Credit_Used), mean(Total_Transitions_Amount) FROM Bank GROUP BY Marital_Status, Income_Category;
27	Bank (GB)	SELECT count FROM Bank GROUP BY Marital_Status, Gender, Education_Level;
28	Bank (GB)	SELECT mean(Credit_Used), mean(Total_Transitions_Amount) FROM Bank GROUP BY Marital_Status;
29	Bank (GB)	SELECT mean(Customer_Age) FROM Bank GROUP BY Gender, Income_Category;
30	Bank (GB)	SELECT count FROM Bank GROUP BY Registered_Products_Count, Attrition_Flag;