FOX: Visual Exploration of Data Fact Outliers

Yikai Li and Yong Wang

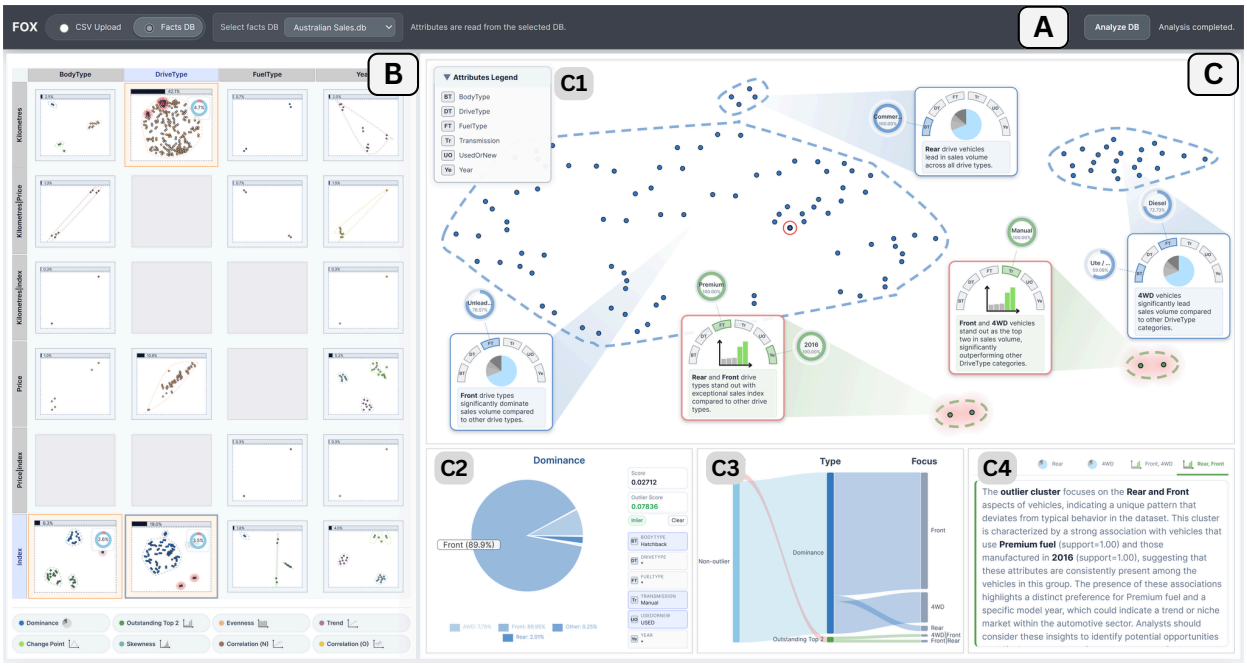


Figure 1: The user interface of *FOX*, which is organized into two parts: preparation and exploration. The first part is supported by the Upload Panel (A), which enables dataset selection, fact extraction, and preprocessing (including grouping, similarity computation, and outlier scoring). The second part consists of the Overview Panel (B), which presents grouped facts as an interactive matrix, and the Main Panel (C), which supports detailed analysis through four views: Cluster Map (C1), Fact Detail (C2), Fact Flow (C3), and Cluster Insights (C4).

Abstract—Exploratory Data Analysis (EDA) systems extract and present data facts to summarize meaningful patterns such as trends and correlations for efficient dataset exploration. However, existing approaches rarely consider outlier detection at the level of data facts, and heterogeneous facts from different analytical scopes are often aggregated in a single view, making it difficult to define meaningful metrics and effectively analyze data fact outliers. To fill this gap, we present *FOX*, a novel visual analytics system for interactive data **F**act **O**utlier e**X**ploration. *FOX* organizes data facts into groups with consistent analytical scopes and computes a unified outlier score that combines distribution-based and pattern-based components. Its interface comprises an *Upload Panel* for data preparation and two coordinated exploration panels: the *Overview Panel* employs a matrix-based visualization to enable an intuitive overview of all data facts, and the *Main Panel* provides four linked views for cluster-level and fact-level analysis. We evaluated the usability and effectiveness of the system through two usage scenarios on public datasets and in-depth interviews with 12 participants. The results show that FOX enables meaningful detection, analysis, and explanation of data fact outliers.

Index Terms—Data fact outlier, visual analytics, exploratory data analysis

1 Introduction

Exploratory Data Analysis (EDA) aims to uncover patterns and derive insights from data [2, 24]. To facilitate effective EDA, recent research has developed automated data-mining approaches [16, 44] and interactive visual analytics systems [27, 50] that extract and present *data facts* [16, 27], which are interesting patterns in a dataset, such as trends and correlations, and are often represented in a tuple-based format [16].

Outlier detection plays an important role in data analysis by helping analysts identify unexpected patterns that may reveal valuable insights [7]. However, existing approaches rarely examine outliers at the level of data facts. We define a *data fact outlier* as a data fact whose pattern deviates substantially from other facts within the same analytical scope. For example, in a retail sales dataset, most facts concerning the relationship between discounts and sales may exhibit a positive correlation, whereas a fact restricted to technology products shipped via second-class delivery may exhibit a negative correlation. Such a context-specific deviation can reveal potentially valuable behavior that would be difficult to identify through conventional analysis. Nevertheless, methods for detecting and interactively exploring these data fact outliers remain underexplored. Although rare-pattern mining and outlier-detection methods have been extensively studied [4, 7, 23], they generally focus on raw data instances, transactions, or individual patterns rather than deviations among a collection of comparable data

• *Yikai Li is with Nanyang Technological University. E-mail: liyi0084@e.ntu.edu.sg.*
• *Yong Wang is with Nanyang Technological University. E-mail: yong-wang@ntu.edu.sg.*

Manuscript received xx xxx. 202x; accepted xx xxx. 202x. Date of Publication xx xxx. 202x; date of current version xx xxx. 202x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.202x.xxxxxxx

facts. Similarly, existing visual analytics systems primarily present data facts to support exploration or storytelling [28, 42], without explicitly identifying and supporting the investigation of data fact outliers.

Supporting such investigation presents two related challenges: ensuring meaningful comparisons among data facts and maintaining a clear visual representation. Existing visualization approaches often aggregate heterogeneous data facts into a single view to reveal global patterns and distributions [19, 50]. Although this strategy is useful for obtaining an overall summary, it is problematic for outlier analysis because data facts may belong to different analytical scopes. For example, facts describing sales across countries and those describing sales across product types use different breakdown attributes and therefore cannot be meaningfully compared when computing outlier scores. Reliable fact-level outlier analysis consequently requires data facts to be organized into consistent analytical scopes, within which tailored comparison and scoring metrics can be defined. Moreover, displaying a large number of heterogeneous facts in one view can produce substantial visual clutter, making unusual patterns difficult to identify and inspect.

To address these challenges, we propose *FOX*, a novel visual analytics system for data **F**act **O**utlier e**X**ploration that enables users to conduct data fact outlier analysis within each group through coordinated views, as shown in Fig. 1. The system automatically extracts data facts from datasets, organizes them into groups with consistent analytical scopes, computes numerical and categorical similarities, and derives outlier scores by combining distribution-based and pattern-based scores. The Upload Panel allows users to upload a target dataset and configure the settings for fact extraction. The Overview Panel presents the global distribution of facts using a matrix-based visualization and integrates interactive exploration features that allow users to select cells of interest for deeper analysis. For a selected cell, analysts can further explore the data in the Main Panel through coordinated views, where the Cluster Map view visualizes facts as a scatterplot with tailored labels and contextual descriptions, the Fact Detail view enables inspection of selected facts, the Fact Flow view provides an overview of fact distributions within the group, and the Cluster Insights view generates descriptions to summarize clusters and highlight potentially valuable insights. In summary, the main contributions of this paper are as follows:

- We present *FOX*, a novel visual analytics system that extracts data facts and organizes them into groups to facilitate outlier analysis. We evaluate the system through two usage scenarios and in-depth interviews with 12 participants, demonstrating its effectiveness and usability. The source code is publicly available online[1].
- We propose a unified outlier scoring framework with tailored metrics for data fact outliers, enabling the quantification of outlier scores by combining distribution-based and pattern-based characteristics.

2 Related Work

In this section, we review existing work on automatic fact mining tools, as well as methods for outlier detection and rare pattern mining.

2.1 Automatic Fact Mining Tools

Over the past decade, prior EDA research has developed structured frameworks and explicit definitions for extracting data facts and mining interesting patterns from datasets. Demiralp et al. [13] supported rapid data exploration through guideposts and statistical descriptors, serving as an early precursor to the data fact concept. Tang et al. [44] formally introduced the concept of data facts, capturing interesting observations derived from aggregation results to reduce analysts' manual effort during exploration. Ding et al. [16] established a more comprehensive and unified formulation of data facts using a 5-tuple representation, together with a systematic mining framework based on this formulation. Ma et al. [31] further introduced the notion of homogeneous data patterns for grouping data facts and developed a scoring framework.

Apart from theoretical design, a variety of visual fact mining and analytics systems have been proposed to leverage visual representations for improving human understanding and supporting insight discovery beyond traditional analysis methods [27]. Prior work can be broadly grouped into data storytelling systems, visualization recommendation approaches, and LLM-assisted tools that enable mixed-initiative data analysis. Data storytelling systems aim to extract and structure insights into coherent narratives, such as Calliope [42], which explores tabular data to construct data-driven stories, CoInsight [26], which derives and relates insights from hierarchical tables, FactExplorer [19], which exposes the space of possible facts to support exploration, and DataShot [46], which automates the generation of concise analytical summaries. Visualization recommendation approaches focus on mapping data to effective visual representations, including Data2Vis [15], which generates visualizations from high-level specifications, DashBot [14], which learns to construct analytical dashboards, and DeepEye [30], which automatically selects data transformations and visualization types. More recently, LLM-assisted tools have supported interactive and mixed-initiative analysis [28, 32, 48]. For example, InkSight [28] documents analytical findings, InsightLens [48] helps users navigate complex conversational contexts, and InsightPilot [32] guides analysis workflows and insight generation. Our system organizes data facts into groups and focuses on detecting and analyzing data fact outliers, with LLM-generated summaries to aid interpretation.

2.2 Outlier Detection and Rare Pattern Mining

Pattern mining was introduced by Agrawal et al. [1] to discover frequent patterns in transactional data and was later extended to identify rare but meaningful patterns. Rare pattern mining methods are generally Apriori- or FP-Growth–based. MS-Apriori [4, 29] introduced multiple minimum support thresholds, followed by extensions such as ARIMA [43] and HURI [39]. However, Apriori-based methods scale poorly because retaining rare items causes rapid candidate growth [4, 23]. FP-Growth–based methods, including MCCFP [22], MSFP [17], and MCRP [3], avoid explicit candidate generation. Nevertheless, most studies focus on rare association rules in static data [23], while sequential rare pattern mining remains less explored due to temporal dependencies and exponential candidate growth [4, 23].

Outlier detection identifies rare instances that deviate substantially from the majority [7]. Depending on label availability, methods are classified as supervised, semi-supervised, or unsupervised, with unsupervised approaches most widely studied because labeled anomalies are scarce. Existing techniques are broadly proximity- or projection-based. Proximity-based methods use neighborhood relationships: LOF [8] measures local-density deviation, while COF [45] and LOCI [38] consider connectivity and multi-scale density, respectively. However, their reliance on nearest-neighbor searches can be computationally costly. Projection-based methods improve scalability through dimensionality reduction; for example, PINN [12] uses random projections to approximate k-nearest neighbors efficiently.

To the best of our knowledge, our system is among the first systems to explicitly support outlier analysis at the level of data facts and proposes a new metric to identify such outliers by combining distribution-based and pattern-based features, leveraging both LOF and information-theoretic measures.

3 System Design

In this section, we discuss the design requirements derived from prior work and present the framework of our system.

3.1 Design Requirements

Our system aims to support effective dataset exploration by enabling users to conduct structured outlier analysis in an efficient and intuitive manner. Based on this goal, we derive six key design requirements from prior work that guide the development of our system, following a workflow from data processing to interactive exploration.

R1: Automatic Fact Extraction. To reduce manual effort and enable systematic data exploration, automatic data fact extraction has been widely recognized as an established practice in visual analytics research [16, 42, 44]. However, prior systems often rely on prioritization strategies such as best-first search [16] to identify top-ranked facts, which may overlook facts that are less prominent under the criteria

[1] https://github.com/lyk6666/FOX

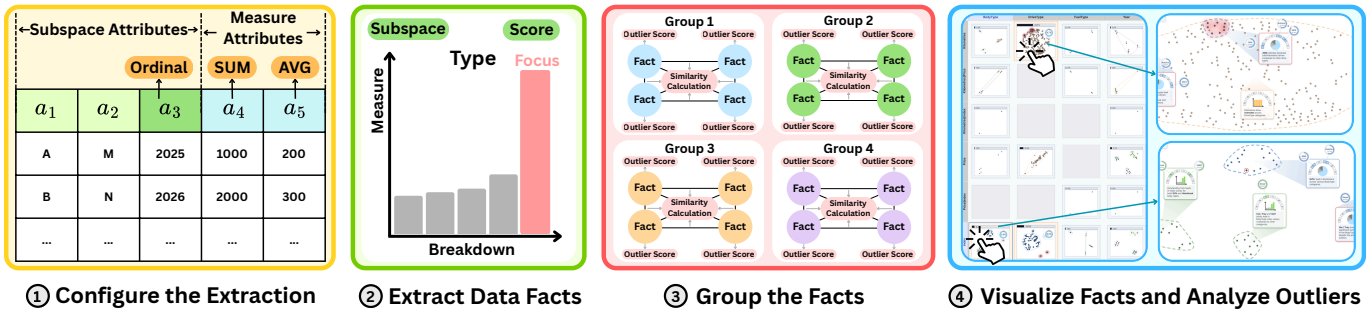


Figure 2: The pipeline consists of four stages: (1) dataset upload and configuration for subspace and measure attributes, (2) automatic fact extraction, (3) fact grouping with similarity computation and outlier scoring, and (4) an interactive interface for visualizing data facts and analyzing fact outliers through cluster exploration, fact inspection, and descriptive summaries.

but still valuable as outliers relative to other facts. Therefore, *FOX* should adopt an extraction strategy that more comprehensively covers diverse subspaces, providing a broader fact population for downstream grouping and outlier scoring.

R2: Structured Fact Grouping for Consistent Analysis. Presenting all data facts within a single unified view provides an intuitive overview of their global distribution and patterns [19, 50]. However, effective outlier analysis requires facts to be grouped according to consistent breakdown and measure attributes [16]. Such structured grouping enables the design of outlier metrics tailored to specific analytical scopes. By ensuring that all facts within a group share the same breakdown–measure scope, pattern interpretation and comparison remain coherent and meaningful.

R3: Tailored Metrics for Fact-Level Outlier Scoring. Extensive research has been conducted on outlier and anomaly detection [7, 9]. However, existing metrics are not directly applicable to the analysis of data facts. Prior work such as MetaInsight [31] introduced the idea of grouping data facts into commonness and exception categories at a global level, but did not propose dedicated metrics to quantify fact-level outlierness, instead relying on simple overlap-based similarity measures. Therefore, there is a need to design system-specific and context-aware metrics that can effectively quantify outlier scores for individual data facts and support subsequent analysis.

R4: Global Overview and Group Identification. Analysts need to efficiently monitor the global landscape of data facts spanning multiple heterogeneous breakdown and measure attributes [26]. The system should support a rapid visual scan across distinct fact groups, enabling users to immediately identify which analytical scopes contain anomalous behaviors or high proportions of outliers without initial exposure to the complexity of individual data facts.

R5: Intra-Group Exploration with Drill-Down. Once an interesting analytical group is isolated, users require capabilities to progressively explore its local structure [19, 50]. Analysts need to examine detailed fact distributions, inspect individual subspaces, and interactively trace how data patterns correlate across different dimensions such as outlier status, pattern types, and focuses.

R6: Cognitive Load Reduction. Evaluating large clusters of data facts can be time-consuming and cognitively overwhelming, especially for users without deep domain expertise. Instead of forcing analysts to interpret abstract patterns manually, the system should automatically turn complex cluster characteristics into concise, natural-language summaries that highlight key analytical takeaways [32, 48].

3.2 Framework

As illustrated in Fig. 2, the framework consists of four main stages. First, the analyst uploads a dataset and configures the analysis by specifying subspace attributes, which also serve as breakdown attributes, and measure attributes. Ordinal attributes can be selected within the subspace attributes, and an aggregation method is assigned to each measure attribute to prepare the system for fact extraction. The system then automatically extracts data facts that satisfy the predefined configurations, generating a file containing all extracted facts (**R1**).

Next, the analyst selects the extracted fact file for further processing. The system organizes the facts into groups (**R2**), computes pairwise similarity within each group, and calculates an outlier score for each fact using our proposed metrics (**R3**), thereby identifying outlier facts and preparing the data for visualization and exploration. After processing, the analyst can interactively explore the facts and analyze fact outliers through the system interface (**R4**). By selecting groups of interest (**R5**), users can explore detailed information through coordinated views. These views support the inspection of data facts and their relationships across contexts, while LLM-assisted summaries and labels are generated through a designed prompting strategy (**R6**).

4 Data Fact Processing

This section describes the data fact processing pipeline, which entails extracting insightful data facts from the dataset, grouping them according to predefined rules, and calculating their outlier scores.

4.1 Data Facts

Data facts capture meaningful patterns hidden in data and serve as fundamental building blocks for data exploration. Following prior work [16, 31, 44], we adopt a formalized representation of data facts to systematically extract valuable insights that satisfy predefined requirements. In our formulation, each data fact is defined by six core elements and represented as a 6-tuple:

$$\{subspace, breakdown, measure, type, score, focus\}.$$

Consider a hypothetical global retail dataset comprising annual sales records from different countries and three attributes: *Country*, *Year*, and *Sales*. A data fact derived from this dataset may characterize the distribution of total sales across countries in a given year, identifying the dominant country and quantifying the interestingness of the observed pattern. For example, such a fact can be represented as $\{\{Country = *, Year = 2025\}, Country, Sales, Dominance, 0.5, Country\,A\}$.

Subspace defines the analytical scope of a data fact through attribute filters over nominal and ordinal attributes, which we collectively refer to as subspace attributes. It is represented as a fixed-length tuple of size n, where the k-th attribute a_k takes value $v_k \in V_k \cup \{*\}$, and V_k denotes the set of all distinct values of a_k. The wildcard $*$ indicates that no filter is applied to the corresponding attribute, and the number of specified attribute values determines the subspace degree. Formally, a subspace is defined as:

$$\{\{a_1 = v_1\}, \{a_2 = v_2\}, \ldots, \{a_n = v_n\}\}, \quad v_k \in V_k \cup \{*\}.$$

In our example, the subspace $\{Country = *, Year = 2025\}$ defines the fact's analytical context as records in the year 2025. **Breakdown** performs a group-by operation on a subspace using an attribute without an applied filter, partitioning the data for aggregation. **Measure** refers to one or two numerical attributes aggregated using functions such as sum, average, minimum, or maximum. **Type** describes the pattern category of a data fact. Following prior work [10, 16, 44], we adopt

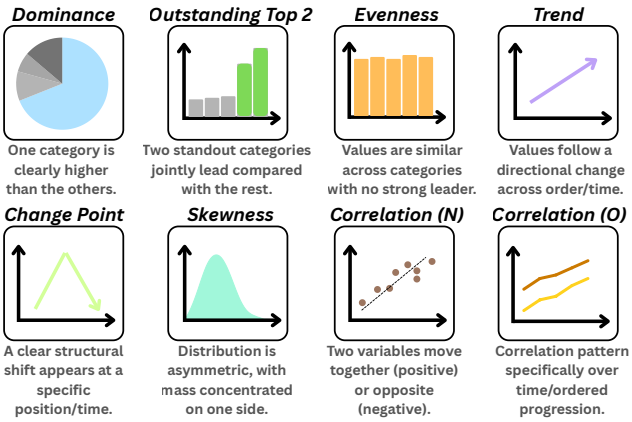


Figure 3: Our system supports eight types of data facts: *Dominance*, *Outstanding Top 2*, *Evenness*, *Trend*, *Change Point*, *Skewness*, *Correlation (N)*, and *Correlation (O)*, each accompanied by a brief description. Here, *Correlation (N)* refers to correlations over nominal attributes, while *Correlation (O)* refers to correlations over ordinal attributes.

eight types in our system, as illustrated in Fig. 3, where each type is encoded with a distinctive color consistent with the system design. Selecting *Country* as the breakdown and *Sales* as the measure implies that the fact evaluates the distribution of sales volumes across countries within the subspace. **Score** quantifies the interestingness of a fact using impact and significance. Impact is computed using a count aggregation that satisfies the anti-monotonicity property [16, 36], while significance follows prior formulations [16, 44]. The final score is defined as:

$$score_i = impact_i \times significance_i. \tag{1}$$

Focus highlights the key element of a fact that requires attention, such as a *Dominance* type with a focus on *Country A*, which indicates that *Country A* leads sales within this specific context with a score of 0.5.

Based on the above representation, *FOX* automatically extracts data facts by enumerating candidate subspaces within the user-configured attribute space. Unlike prior systems that prioritize fact extraction tasks using impact-based best-first search [16], *FOX* employs depth-first search to systematically traverse the attribute space and retain all facts that satisfy the predefined extraction criteria (**R1**). To improve computational efficiency, the system applies impact-based pruning to terminate branches that cannot produce eligible facts, while the fact types and their associated scores are computed using established methods from prior work [16, 44].

4.2 Fact Grouping

Since data facts are generated under distinct breakdown–measure combinations, analyzing them within a single global scope is inappropriate. For example, comparing facts about sales across years with facts about discounts across countries is conceptually inconsistent because they are defined over different analytical scopes, meaning that they involve different breakdown or measure attributes. Performing outlier detection or similarity measurements across different scopes yields uninterpretable results because cross-scope comparisons lack explainability.

While local outlier analysis within global clusters is a potential alternative [7, 8], further subdividing facts can reduce the system's sensitivity to meaningful deviations. To ensure valid and effective comparisons, we restrict analysis to facts sharing identical breakdown and measure attributes. Accordingly, we introduce a grouping mechanism that partitions data facts by identical breakdown and measure attributes, restricting each analysis to a specific scope (**R2**).

4.3 Similarity

After grouping data facts by breakdown and measure, we compare facts within each group to support outlier detection. Since grouped facts share the same analytical scope, they can be represented consistently and compared meaningfully. Our similarity measure combines numerical similarity between aggregated value distributions with categorical similarity based on fact type and focus, as illustrated in Fig. 4.

Each data fact aggregates records according to the breakdown attribute and computes aggregated values based on the measure attribute, producing a set of aggregated results. These results can be represented as a vector in which the keys correspond to breakdown attribute values and the entries represent the aggregated measure values. Since facts within the same group share the same breakdown attribute, their vectors have identical dimensions and key sets. Therefore, our grouping mechanism provides a consistent representation that enables meaningful similarity comparison.

Although different subspaces may contain varying numbers of records, resulting in different aggregated magnitudes, our analysis focuses on the distributions of values rather than their absolute scales. For this reason, we adopt cosine similarity to measure the similarity between data facts [34, 40]. Let the measure attribute be m, let v_k^i denote the i-th distinct value in V_k, and let m^i denote the aggregated measure value corresponding to v_k^i. The vector representation of a fact can therefore be written as:

$$\{\{v_k^1 : m^1\}, \{v_k^2 : m^2\}, \ldots, \{v_k^{|V_k|} : m^{|V_k|}\}\}.$$

Given two facts $fact_a$ and $fact_b$ in the same group, their numerical similarity is measured using cosine similarity [34, 40]:

$$sim_{cos}(fact_a, fact_b) = \frac{\sum_{i=1}^{|V_k|} m_a^i m_b^i}{\sqrt{\sum_{i=1}^{|V_k|} (m_a^i)^2}\sqrt{\sum_{i=1}^{|V_k|} (m_b^i)^2}}. \tag{2}$$

This metric captures the similarity between the value distributions of the two facts while being invariant to their absolute magnitudes.

However, relying solely on cosine similarity is insufficient in the context of data facts. In practice, we observe several limitations. For example, when the number of breakdown categories $|V_k|$ is large, cosine similarity may dilute the influence of the main characteristic of a fact, reducing its discriminative power. In addition, small differences in value distributions may lead to substantially different interpretations. For instance, a proportion of 49% versus 51% may produce similar cosine similarity values, yet the latter indicates a *Dominance* type.

To better capture the categorical characteristics of data facts, we additionally incorporate the type and focus into the similarity calculation. They explicitly describe the type category and the key element of a fact, providing complementary information beyond numerical distributions. We therefore adopt a simple overlap metric [6] for these categorical attributes, where the similarity equals 1 if two values are identical and 0 otherwise, denoted as sim_{type} and sim_{focus}, respectively. The final similarity between two data facts is defined as a weighted combination of numerical and categorical similarities:

$$sim(fact_a, fact_b) = \alpha_1 \cdot sim_{cos} + \alpha_2 \cdot sim_{type} + \alpha_3 \cdot sim_{focus}, \tag{3}$$

where $\alpha_1 + \alpha_2 + \alpha_3 = 1$, and the weights control the relative importance of distribution similarity and categorical characteristics.

4.4 Outlier Score

After computing similarity, we estimate the outlier score of each data fact within its group. A straightforward approach is to apply LOF [7, 8] to the similarity-based distance matrix to detect distribution anomalies. However, LOF alone is not sufficient in practice. Different groups may exhibit distinct distribution characteristics, and a single parameter setting cannot reliably capture all meaningful deviations. As a result, facts that are statistically close to others but exhibit distinctive pattern-level characteristics may remain undetected.

Previous studies note that deviations among data facts often stem from uncommon characteristics [31]. This typically manifests in two ways: either certain fact types occur rarely within a group dominated by others, or specific values appear infrequently within the same fact type. For example, while most facts about sales across countries show an

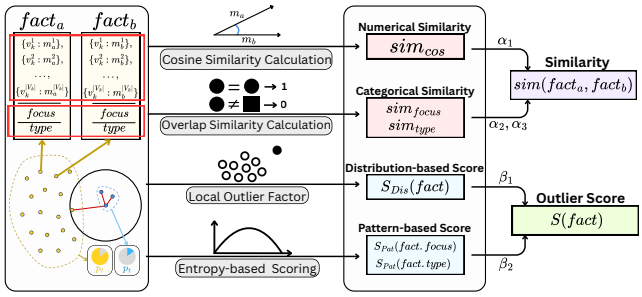


Figure 4: Computation of similarity and outlier scores. Numerical and categorical similarities are combined to compute fact similarity, while the final outlier score integrates a distribution-based score using LOF and a pattern-based score derived from entropy.

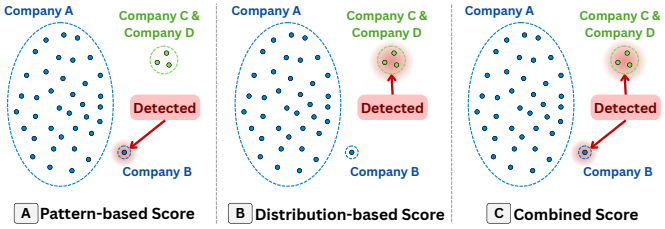


Figure 5: An example illustrating the complementary roles of the two outlier scoring components. Three clusters are constructed: a majority cluster (*Dominance*, *Company A*), a cluster with one fact (*Dominance*, *Company B*), and a cluster with three facts (*Outstanding Top 2*, *Company C & Company D*). (A) The pattern-based score detects the one-fact cluster but misses the three-fact cluster. (B) The distribution-based score detects the three-fact cluster but misses the one-fact cluster. (C) The combined score detects both.

even distribution, only a few may reveal the dominance of a single country. Similarly, within that *Dominance* type, a specific country might dominate rarely compared to others. While these rare characteristics often harbor valuable insights, existing simple overlap-based metrics [31] fail to adequately capture their underlying statistical properties.

To better quantify such deviations, we introduce a pattern-based outlier metric inspired by information theory [18, 41]. The metric evaluates the rarity of pattern elements by considering two distributions: the distribution of fact types within a group and the distribution of fact focuses within each fact type. Let p_t denote the probability of observing a pattern element t, which corresponds to either the probability of a fact type within a group or the probability of a fact focus within a specific fact type. The pattern-based outlier score is defined as:

$$S_{Pat}(t) = (1-p_t)(1-E), \quad E = \frac{-\sum_i p_i \log p_i}{\log K}. \tag{4}$$

Here, K denotes the number of possible element values, corresponding to either fact types or focuses. The entropy term captures the overall skewness of the distribution, while $(1-p_t)$ reflects the outlierness of a specific element. Together, these two factors characterize the degree to which an element deviates from the prevailing patterns in the distribution. Consequently, elements that occur infrequently within a highly skewed distribution receive higher scores, indicating stronger outlierness in terms of pattern characteristics.

To obtain the final outlier score for each data fact, we combine a distribution-based score with a pattern-based score. The distribution-based score is computed via LOF using the combined similarity-based distance matrix prior to dimensionality reduction. It captures local deviations in neighborhood structure based on numerical distributions, fact types, and focuses. The pattern-based score separately quantifies the group-level rarity of fact types and focuses. They capture different aspects of outlierness: the distribution-based score evaluates pairwise neighborhood relationships, whereas the pattern-based score characterizes categorical prevalence within the group. To integrate them, we normalize the LOF score and denote it as S_{Dis} [25]. Since outlierness may arise from either fact type or focus, we take the maximum of the two pattern-based scores as the pattern signal. The final outlier score of a data fact f is defined as:

$$S(f) = \beta_1 \cdot S_{Dis}(f) + \beta_2 \cdot \max\left(S_{Pat}(f.type), S_{Pat}(f.focus)\right), \tag{5}$$

where β_1 and β_2 control the relative contributions of the distributional outlier score and the pattern-based outlier score, with $\beta_1 + \beta_2 = 1$. The resulting value represents the overall outlier score of a data fact. A fact is identified as an outlier if its score exceeds a predefined threshold. Fig. 4 provides an intuitive illustration of how similarity and outlier scores are computed (**R3**).

Since the proposed outlier score targets fact-level anomalies, a concept not explicitly addressed by existing metrics, there is no direct baseline available for quantitative comparison. Therefore, we illustrate the behavior and practical utility of the proposed metric through a component analysis and two usage scenarios. These evaluations demonstrate how its components contribute to detecting different types of data fact outliers. To illustrate the complementary roles of the two components, we construct a synthetic example consisting of three clusters of data facts. The first cluster corresponds to facts with type *Dominance* and focus *Company A*. These facts form the majority of data facts in the group. The second cluster contains a single fact with type *Dominance* and focus *Company B*, representing a rare focus within the same fact type. The third cluster contains three facts with type *Outstanding Top 2* and focus *Company C & Company D*, representing a different pattern.

As shown in Fig. 5A, when only the pattern-based outlier score is applied, the fact associated with *Company B* is detected as an outlier because its focus appears only once among facts of the same type. According to the pattern-based scoring formula in Eq. 4, this rare occurrence results in a very small p_t, producing a high pattern-based outlier score. In contrast, the cluster involving *Company C & Company D* is not identified because this pattern occurs multiple times within the group, leading to a larger p_t and therefore a lower pattern-based score that does not exceed the detection threshold. As shown in Fig. 5B, when only the distribution-based outlier score is applied, the cluster corresponding to *Company C & Company D* is detected because these facts are positioned farther from the majority cluster in the similarity space, resulting in a high LOF score. In contrast, the fact associated with *Company B* is not detected because it lies close to the majority cluster centered around *Company A* and therefore does not exhibit strong distributional deviation. When the combined score is applied (Fig. 5C), both the *Company B* fact and the *Company C & Company D* cluster are successfully detected. This result demonstrates that the pattern-based and distribution-based components capture complementary aspects of data fact outliers.

In addition, when applied to the datasets used in Sec. 6, the proposed outlier score produced results broadly consistent with our qualitative observations. These examples provide preliminary evidence that the metric can help analysts identify potentially meaningful data fact outliers during exploration.

5 Interface and Visual Design

Our system interface is organized into two parts: preparation and exploration. The first part is supported by the Upload Panel, while the second part consists of the Overview Panel and the Main Panel.

Upload Panel. The Upload Panel consists of two stages. In the first stage, users upload a dataset and configure the subspace and measure attributes. Ordinal subspace attributes can represent sequential variables, such as rankings or years, while an aggregation method can be assigned to each measure attribute. Once the configuration is confirmed, the system automatically extracts the eligible data facts using the procedure described in Sec. 4.1 (**R1**). In the second stage, the system groups the extracted facts, constructs intra-group similarity matrices, and computes their outlier scores (**R2**, **R3**), thereby preparing them for visualization and exploration.

Overview Panel. After fact processing is completed, the Overview Panel presents a matrix-based overview of all fact groups, as shown

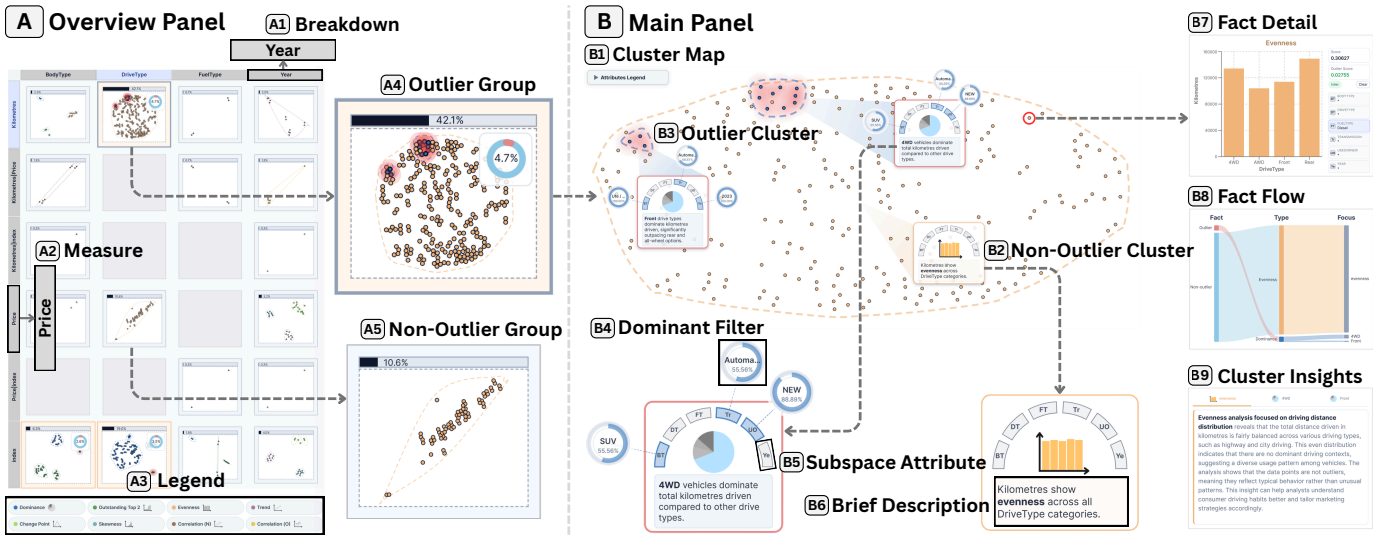


Figure 6: The user interface of FOX and its visual design. This figure presents the Overview Panel (A) and the Main Panel (B) of the system interface, illustrating the key design components and visual encodings used in the system.

in Fig. 6A (**R4**). Because each group is defined by a unique combination of breakdown and measure attributes, we adopt a matrix layout in which the top row encodes breakdown attributes (Fig. 6A1) and the left column encodes measure attributes (Fig. 6A2). Each matrix cell therefore corresponds to a distinct fact group defined by a particular combination of breakdown and measure attributes. Different fact types are represented using distinct colors, with the legend displayed at the bottom of the panel (Fig. 6A3). Within each cell, the top progress bar indicates the proportion of facts in that group relative to the full set of extracted facts (Fig. 6A4). The lower portion of each cell displays a scatterplot preview of the contained facts, positioned according to pairwise similarity through dimensionality reduction. Different methods emphasize different structural properties: t-SNE and UMAP preserve local neighborhood relationships [37], PCA prioritizes variance [33], and MDS preserves pairwise dissimilarities [5]. The method can be selected according to the analytical objective; in our implementation, we use MDS to better preserve similarities among facts. Facts are further organized by their focus, such that facts sharing the same focus are grouped into a single cluster. To visually delineate clusters, we render a smoothed convex hull around each cluster: facts belonging to the same cluster are enclosed by an outer polygon, which is padded, clipped to the canvas, and converted into a smooth closed path drawn as a dashed outline in the cluster color. Cells encode outlier information using distinct visual styles. Outlier groups, as illustrated in Fig. 6A4, are highlighted with a light red background and displayed with a slightly larger cell size. In addition, a donut chart appears in the top-right corner to indicate the proportion of outlier facts within that group. Individual outlier facts within the scatterplot are emphasized using a red halo. Non-outlier groups, shown in Fig. 6A5, are displayed with a light blue background and a smaller cell, and the outlier proportion indicator is omitted for brevity. Each cell is interactive, and selecting a cell reveals its detailed information in the Main Panel.

Main Panel. The Main Panel is organized into four views: (1) Cluster Map, (2) Fact Detail, (3) Fact Flow, and (4) Cluster Insights, as shown in Fig. 6B.

1) Cluster Map. This view is an enlarged version of the selected cell from the Overview Panel and enables a detailed exploration of data fact clusters organized by focus, as shown in Fig. 6B1 (**R5**). Within this view, facts belonging to the same focus are grouped into clusters and enclosed by smooth hull boundaries to visually distinguish cluster structures. Clusters containing outlier facts are highlighted differently from those containing only non-outlier facts, allowing analysts to quickly identify unusual patterns. For example, Fig. 6B2 illustrates a typical non-outlier cluster, while Fig. 6B3 shows an outlier cluster that contains facts with high outlier scores, each highlighted with a red halo. A key addition in this view is the cluster label, a rectangular box surrounded by donut charts. It summarizes important information about each cluster. The label consists of three components. The first component is an icon that conveys the dominant characteristics of the cluster, determined by fact type and focus. The second component is a brief natural-language description generated with LLM assistance, based on the group's breakdown and measure as well as the cluster's type and focus, as illustrated in Fig. 6B6 (**R6**). The third component summarizes the cluster's contextual subspace analysis. In this design, subspace attributes are arranged along a radial attribute band (Fig. 6B5), where each band segment corresponds to one attribute and is displayed in abbreviated form, with the attribute legend shown in the upper-left corner of the view (Fig. 1C1). Because each subspace consists of a set of attribute–value filters, we compute each filter's support within a cluster, namely the proportion of facts containing that filter, to identify frequently occurring subspace filters. When the support of a filter exceeds 50%, it is visualized as a donut chart connected to the corresponding attribute segment, where the slice encodes the dominant attribute value and its frequency (Fig. 6B4). This design enables analysts to quickly identify salient contextual subspace filters without manually inspecting individual facts. These dominant filters reveal the common contexts in which facts of particular types or focuses tend to occur, thereby providing additional analytical insight. By default, labels are placed outside the cluster hull. We sample candidate positions along the hull perimeter, select a nonoverlapping position greedily, and then apply a force-directed adjustment to reduce remaining collisions. In addition, users can manually reposition labels via direct manipulation to suit their preferences. To mitigate occlusion, labels become semi-transparent on hover, allowing users to inspect the underlying facts without obstruction.

2) Fact Detail. When a fact is selected, this view presents its chart, score, subspace, and outlier score in detail. The left side shows a visual chart of the selected data fact, with the chart format determined by the fact type shown in the title. The right side displays the associated metadata, including the score, outlier score, and subspace. Subspace attributes are represented as cards, and non-wildcard attributes are highlighted to indicate the active contextual constraints. An example of this encoding is shown in Fig. 6B7, where the highlighted pills represent the attribute–value filters that define the subspace. In this example, the subspace is constrained only by the filter $\{FuelType = Diesel\}$.

3) Fact Flow. This view provides a visual overview of how data facts in the selected group are distributed across several analytical dimensions using a Sankey-style diagram [11]. The flow begins with

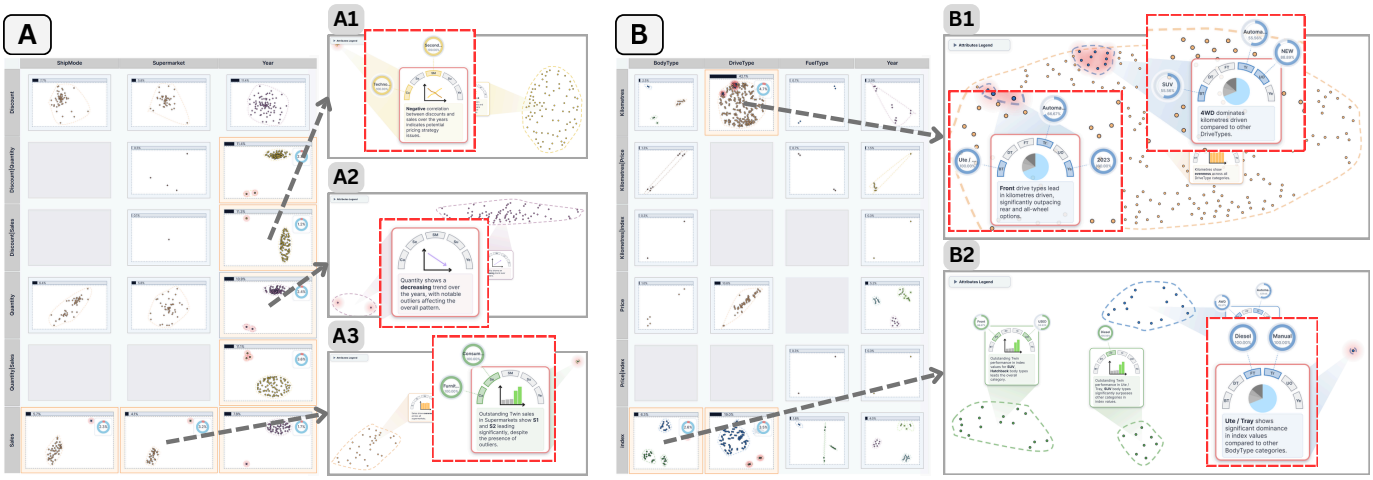


Figure 7: Usage Scenarios. This figure demonstrates the Overview Panel and the corresponding Cluster Map view in the Main Panel after selecting a cell. Panel (A) corresponds to Usage Scenario 1, while panel (B) corresponds to Usage Scenario 2. In each expanded Cluster Map view, the labels of outlier clusters are highlighted with red rectangles.

the outlier status in the first column, distinguishing between outlier and non-outlier facts. These facts then flow into the second column, which represents the fact type, and finally into the third column, which represents the focus of each fact. The width of each flow encodes the number of facts belonging to that category, allowing analysts to quickly understand how facts are distributed and connected across outlier status, type, and focus. Additionally, hovering over a flow segment reveals the exact number of facts represented. By visually tracing these flows, analysts can easily observe the composition and relationships among different categories in the selected group, as illustrated in Fig. 6B8.

4) Cluster Insights. This view delivers a coherent, natural-language explanation of each cluster, allowing users to select clusters of interest from an interactive bar and view details below, thereby reducing cognitive load and supporting faster understanding (**R6**). Drawing on established prompt-engineering practices [35, 49], we design a staged, role-based prompting pipeline implemented via the OpenAI API using `gpt-5.4-mini` with the temperature set to 0.2. Inspired by recent advances in task decomposition and structured prompt engineering [21, 47, 51], the pipeline decomposes the summarization task into three sequential stages rather than relying on a single monolithic prompt, helping mitigate hallucinations and reduce cascading reasoning errors. First, a *meaning analyzer* translates abstract schema attributes (i.e., breakdowns, measures, types, and focuses) into plain-language terminology. Second, a *subspace analyzer* contextualizes the cluster by evaluating its frequently occurring subspace filters and the subspaces of data fact outliers. Third, a *group summarizer* ingests these intermediate analytical outputs to synthesize a cohesive, cluster-level narrative. An example of the generated summary is shown in Fig. 6B9.

To facilitate initial use, we incorporate an onboarding feature that automatically guides analysts through the system's concepts, functionalities, and visual encodings whenever they enter a new interface state.

6 Usage Scenarios

We present two usage scenarios using public datasets to illustrate how analysts can explore data and analyze data fact outliers with the system.

6.1 Scenario 1: Supermarket Sales

In the first scenario, we use a supermarket sales dataset[2] to demonstrate how *FOX* assists users in analyzing data fact outliers during exploratory analysis. Before using the system, the analyst has already conducted a rough inspection of the dataset but has not identified any particularly surprising findings. After loading the dataset, extracting data facts, and performing the subsequent analysis, the system identifies 14 cells containing facts, among which seven groups include outlier facts that deserve further inspection, as shown in Fig. 7A.

The analyst first explores the relationship between discounts and sales over time. After opening the corresponding cell, as shown in Fig. 7A1, the analyst finds two clusters of facts. The larger cluster indicates a positive correlation between discounts and sales, which is consistent with common expectations that higher discounts tend to stimulate purchases. However, the analyst is surprised to discover another cluster showing a negative correlation between discounts and sales. By examining the associated subspaces, the analyst finds that this pattern occurs for technology products shipped by second-class delivery. This suggests that higher discounts do not always increase sales and may even reduce them in some contexts.

Next, the analyst investigates which supermarkets dominate sales. After selecting the relevant cell, as shown in Fig. 7A2, the analyst finds that, in most cases, sales are distributed relatively evenly across the six supermarkets. However, one outlier cluster reveals a different pattern: for furniture products purchased by consumer customers, supermarkets S1 and S2 together account for a much larger share of sales than the other four supermarkets. This finding suggests that S1 and S2 may be particularly competitive in the consumer furniture market, possibly because they place greater emphasis on this category or employ more effective sales strategies.

The analyst then turns to quantity trends over the years. Most facts in this group show an increasing trend, while only two facts exhibit a decreasing trend, as shown in Fig. 7A3. Since the automatically generated labels do not reveal any obvious common context for these decreasing-trend facts, the analyst inspects them individually. This closer examination shows that one decreasing trend appears among home office orders delivered by first-class shipping, while another appears for corporate customers purchasing office supplies. These findings help the analyst identify specific contexts in which quantity changes differ from the overall upward trend.

6.2 Scenario 2: Australian Car Sales

In the second scenario, the analyst uses our system to explore an Australian car sales dataset[3]. Compared with the first scenario, this dataset is more complex, making it more difficult to identify data fact outliers through straightforward inspection. After automatic fact extraction and subsequent processing, the system identifies 18 groups of data facts, among which three groups contain fact outliers that merit closer examination, as shown in Fig. 7B.

[2] https://www.kaggle.com/datasets/yikai6/supermarket-resale

[3] https://www.kaggle.com/datasets/nelgiriyewithana/australian-vehicle-prices

The analyst first explores the group describing the relationship between kilometers driven and drive type. After opening this group, as shown in Fig. 7B1, the distribution of data facts appears relatively complex at first glance. To gain an initial understanding, the analyst refers to the Fact Flow view, which provides a clear visual overview of how data facts are distributed across categories. From this overview, the analyst observes that, in most contexts, kilometers driven are relatively evenly distributed across different drive types. However, two outlier clusters exhibit distinctly different patterns. In one cluster, the driving distance is dominated by front-wheel-drive vehicles, whereas in the other it is dominated by four-wheel-drive vehicles. By inspecting the labels and accompanying summaries, the analyst finds that one cluster is primarily associated with new vehicles, automatic transmission, and the SUV category. Within this cluster, four-wheel-drive vehicles account for a large share of kilometers driven, suggesting a strong association between vehicle type, drivetrain, and usage patterns. In the other cluster, front-wheel-drive vehicles dominate the driving distance and are frequently associated with utility vehicles from 2023 with automatic transmission. The summary panel further helps the analyst confirm and interpret these patterns.

The analyst then examines another group of interest: the relationship between the number of vehicles sold and body type. Using count aggregation, this group can be interpreted as reflecting sales volume. After selecting the corresponding cell, as shown in Fig. 7B2, the analyst inspects the Fact Flow view to understand the distribution of data facts. The visualization reveals two types of data facts and four distinct focuses, each corresponding to a cluster. Compared with the previous group, this one exhibits a more balanced distribution of *Dominance* and *Outstanding Top 2* facts. A closer inspection of the *Dominance* facts shows that most are associated with SUV sales dominance, while only a single fact indicates dominance by utility vehicles. This observation motivates further exploration of the clusters. The analyst finds that many SUV-dominant facts are associated with contexts involving all-wheel-drive vehicles and automatic transmission, suggesting a strong co-occurrence between these attributes and higher SUV sales. In contrast, the utility vehicle dominance appears in a more specific context, namely diesel vehicles with manual transmission, and is accompanied by a relatively high outlier score of 0.67. Overall, these findings suggest that SUV sales are frequently associated with a range of common configurations, whereas utility vehicles tend to become dominant only under more specific conditions. By combining these observations with information from the Cluster Insights view, the analyst gains a comprehensive understanding of the characteristics, underlying patterns, and contextual associations within each cluster. These findings clarify how vehicle characteristics relate to driving patterns and sales outcomes.

7 User Interview

To evaluate the usability and effectiveness of *FOX*, we conducted in-depth interviews with 12 participants. In this section, we describe the participants, study procedure, and results derived from interviews.

7.1 Participants

We recruited 12 participants (P1–P12) from a university population (3 female, 9 male; $age_{mean} = 23.92$, $age_{sd} = 2.78$). All reported normal or corrected-to-normal vision, no color-vision deficiencies, relevant academic backgrounds, and at least one year of experience with data analysis and visualization tools. The study was approved by the Nanyang Technological University Institutional Review Board. Written informed consent was obtained from all participants.

7.2 Procedure

The interview consisted of four stages and lasted approximately one hour. In the first stage, we provided a brief introduction to the system, including its background and key concepts (e.g., data facts and outlier scores), to familiarize participants with its purpose. In the second stage, we used the dataset described in Sec. 6.1 to demonstrate the system's functionality and train participants on how to use and explore it. Participants were encouraged to ask questions during this stage. In the third stage, participants conducted a task-based, think-aloud exploration using the dataset from Sec. 6.2, where they were asked to identify an outlier group using the Overview Panel, select a data fact outlier within it, interpret the meaning of a specific cluster, and use the Main Panel views to further explore data fact outliers and enhance their understanding. Afterward, they were invited to freely explore the system and provide feedback, suggestions, and insights. In the final stage, participants completed a post-study questionnaire consisting of 11 questions. Nine closed-ended questions used a 5-point Likert scale [20] (from Strongly Disagree to Strongly Agree) to evaluate different aspects of the system, including the onboarding feature (Q1), the Overview Panel (Q2–Q3), the Main Panel (Q4–Q7), and overall usability (Q8–Q9). Two open-ended questions (Q10–Q11) were included to collect additional feedback and suggestions for improvement. All participants provided informed consent prior to the study and received compensation of SGD 15 upon completion.

7.3 Results

Fig. 8 presents participants' responses to the closed-ended questions (Q1–Q9). Overall, the system received positive feedback from participants. In the following, we provide a detailed analysis of user feedback based on these results.

Onboarding Feature. Overall, the onboarding feature received a high rating (Q1: $M = 4.67$). Participants P1 and P5 noted that the onboarding process was particularly helpful for first-time use. However, P2 found the onboarding tour helpful but suggested that certain aspects could be explained in greater detail, as understanding several key concepts is essential for effectively using the system. Consequently, P2 assigned a rating of 3 for this question.

Overview Panel. Participants generally assigned high ratings to the Overview Panel. The matrix representation was considered effective in providing an overview of the distribution of data facts (Q2: $M = 4.58$) (**R4**), while participants found the visual encoding helpful for identifying groups to explore (Q3: $M = 4.42$) (**R4**). During task-based exploration, most participants first located cells with a red background and then examined the corresponding distributions in detail. In contrast, P4 and P5 relied more on percentage values to select cells, tending to prioritize those containing more data facts; they described the encoding as both concise and comprehensive. However, some participants identified limitations. P11 noted that while the matrix representation offers a clear overview of grouped data, it lacks a global view of all data facts within a single unified visualization. Additionally, P6 found the visual encoding of cells insufficiently salient, initially focusing only on the distribution plots below. P6 suggested that key indicators, such as size and color used to represent outlier groups, should be more visually prominent and easier to compare across cells.

Main Panel. Overall, participants provided positive feedback across these components: the Cluster Map view effectively conveys the distribution of data facts within clusters (Q4: $M = 4.50$), the Fact Detail view supports clear inspection of individual facts (Q5: $M = 4.42$), and the Fact Flow view offers an intuitive visual summary (Q6: $M = 4.25$) (**R5**). The Cluster Insights view provides valuable textual explanations (Q7: $M = 4.67$) (**R6**). Despite the generally positive feedback, participants identified several areas for improvement. For the Cluster Map view, P7 suggested removing the breakdown attribute from the label ring bands, as it is not filterable under the current constraints and therefore unnecessary to include. P10 commented that the label text is too small and expressed less interest in distributional views, preferring concise, high-level insights. Regarding the Fact Detail view, P9 noted that although the feature is useful, it may be redundant in practice, as facts within the same cluster often exhibit similar patterns; thus, the Cluster Map view alone may suffice for analysis. For the Fact Flow view, P2 found it helpful but suggested that it may be unnecessary when the distribution within a group is relatively simple. Similarly, P11 observed that it serves a role similar to the Cluster Map view and suggested aligning their layouts to reflect their comparable overview functions. For the Cluster Insights view, we observed differing user preferences. Participants P3, P4, and P8 found the LLM-assisted summaries highly beneficial, as they reduce cognitive effort and provide intuitive, text-based insights; P8 further suggested elevating this com-

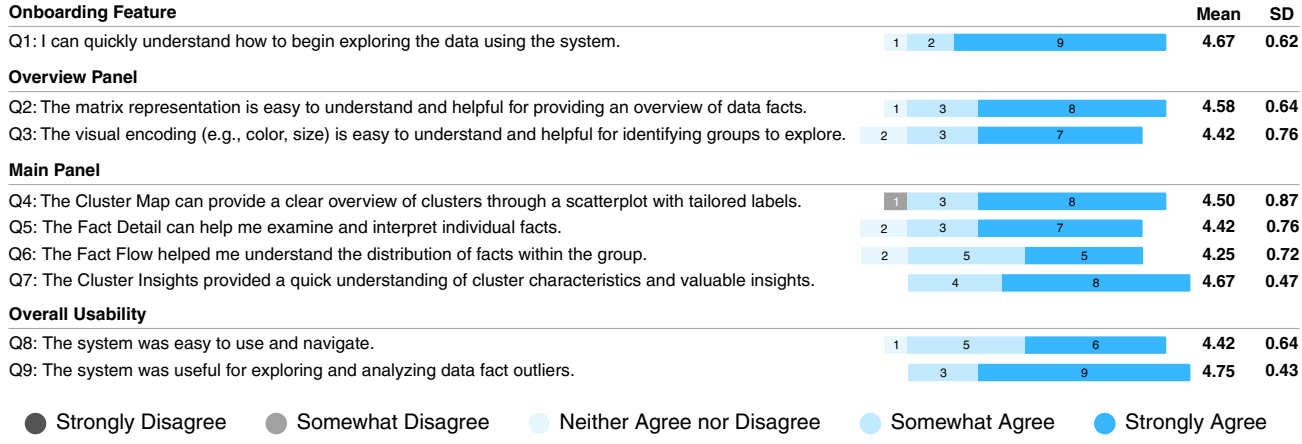


Figure 8: User interview questionnaire results. The results are organized into four parts: Onboarding (Q1), Overview Panel (Q2–Q3), Main Panel (Q4–Q7), and Overall Usability (Q8–Q9). The figure presents the closed-ended questions, their corresponding components, and the mean and standard deviation for each question.

ponent to a primary view to better support exploration. In contrast, P9 preferred visual representations over textual descriptions and reported a reluctance to read longer textual summaries during analysis.

Overall Usability. Overall, participants found the system easy to use and navigate (Q8: $M = 4.42$) and effective for exploring and analyzing data fact outliers (Q9: $M = 4.75$). P1 noted that the concept of data fact outliers was novel and enabled exploration across different groups in ways she had not previously considered. P11, who works as an analyst, expressed interest in gaining a deeper understanding of the system and potentially applying it in future practice. P9, who has a strong background in visualization, noted that the Overview and Main Panels support both global overview and drill-down analysis, offering a clearer design than combining all information within a single view. However, she also emphasized that a solid understanding of the fundamental concepts is important for effectively using the system in analysis.

Open-ended Questions. In this section, we evaluated participant preferences regarding component utility (Q10) and gathered suggestions for improvement (Q11). Regarding utility, participants expressed two main preferences: P1, P2, P6, P7, P9, and P11 favored the Cluster Map view for its informative labels and effective distribution overview, while the remaining participants preferred the Cluster Insights view for providing intuitive insights that minimize cognitive load and learning effort. Participants also provided several suggestions for improvement. For the interface, P5 proposed a color configuration panel to customize data fact encodings, while P11 and P12 suggested increasing the prominence of the Fact Flow via layout adjustments or richer visual encodings. Regarding the LLM features, P8 suggested expanding the LLM-based summary to a global view with natural-language exploration, whereas P9 advised reducing summary verbosity in favor of more intuitive visualization.

8 Discussion and Future Work

In this section, we discuss limitations of our system and outline possible directions for future work.

Extending Supported Data Fact Types. Currently, our system supports eight predefined data fact types, which may not capture all analytical patterns of interest, such as seasonality [16]. To address this, future work could explore mechanisms for user-defined pattern specification, allowing analysts to describe new fact types that can be integrated into the extraction algorithms for more flexible analysis.

Supporting Diverse User Preferences. Our interviews revealed that users exhibit differing preferences in how they interact with the system. For example, P8 preferred the LLM-assisted summary and suggested expanding its role, whereas P9 favored visualization-based exploration and preferred minimizing textual descriptions. To accommodate such differences, future work could introduce adaptive or customizable interfaces, such as a toggle mechanism that allows users to switch between text-driven and visualization-driven exploration modes.

Supporting Natural-Language Interaction. While existing LLM-based visual analytics tools are effective at generating general explanations [27, 32], *FOX* focuses on targeted data fact outlier analysis by using deterministic scoring to identify data fact outliers. However, the current LLM pipeline only generates static summaries for detected clusters and does not allow users to further explore the results. Future work could integrate an LLM-based conversational interface with natural-language query capabilities, enabling users to ask follow-up questions, request clarification, and investigate specific patterns or data fact outliers in greater depth.

Enriching Outlier Semantics and Relationships. Currently, our system uses a binary classification of data fact outliers and focuses on intra-group exploration, overlooking the distinct semantic roles that outliers may play. For example, an increasing trend among predominantly decreasing trends contradicts the prevailing pattern, whereas a *Dominance* fact focused on *Country B* among facts predominantly focused on *Country A* represents a distinctive alternative focus. Moreover, the relationships between group-level outliers and global dataset patterns remain unexplored, and the system does not identify which breakdown or measure attributes are associated with the highest proportions of outliers. Future work will address these limitations by enriching data fact outlier semantics and linking intra-group findings to the global matrix to support broader analysis.

Improving Scalability. When applied to large datasets or high volumes of data facts, the current extraction algorithm and Overview Panel face performance and visualization scalability limits. Future work will focus on optimizing large-scale fact extraction and introducing hierarchical or sampling-based techniques to manage complexity.

9 Conclusion

We present *FOX*, a visual analytics system for exploring data fact outliers. The system groups data facts into consistent analytical scopes and introduces a unified scoring framework that combines distribution-based and pattern-based scores. We design an interactive interface with coordinated visualizations and LLM-assisted summaries to help analysts examine outlier clusters and interpret individual facts. Through two usage scenarios and interviews with 12 participants, we demonstrate that *FOX* enables effective exploration of data fact outliers. Future work will support user-defined fact types, customizable interfaces, natural-language interaction, richer semantics, and greater scalability.

Acknowledgments

This project is supported by the Ministry of Education, Singapore, under its Academic Research Fund Tier 1 (NTU Tier 1, RG104/25) and the NTU Start Up Grant awarded to Yong Wang. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not reflect the views of the Ministry of Education, Singapore.

References

[1] R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. In *Proc. SIGMOD*, pp. 207–216. ACM, New York, 1993. doi: 10.1145/170035.170072 2

[2] J. T. Behrens. Principles and procedures of exploratory data analysis. *Psychol. Methods*, 2(2):131–160, 1997. doi: 10.1037/1082-989X.2.2.131 1

[3] U. Bhatt and P. Patel. A novel approach for finding rare items based on multiple minimum support framework. *Procedia Comput. Sci.*, 57:1088–1095, 2015. doi: 10.1016/j.procs.2015.07.391 2

[4] A. Borah and B. Nath. Rare pattern mining: Challenges and future perspectives. *Complex Intell. Syst.*, 5(1):1–23, 2019. doi: 10.1007/s40747-018-0085-9 1, 2

[5] I. Borg and P. J. F. Groenen. *Modern Multidimensional Scaling: Theory and Applications*. Springer Ser. Stat. Springer, New York, 2nd ed., 2005. doi: 10.1007/0-387-28981-X 6

[6] S. Boriah, V. Chandola, and V. Kumar. Similarity measures for categorical data: A comparative evaluation. In *Proc. SDM*, pp. 243–254. SIAM, Philadelphia, 2008. doi: 10.1137/1.9781611972788.22 4

[7] A. Boukerche, L. Zheng, and O. Alfandi. Outlier detection: Methods, models, and classification. *ACM Comput. Surv.*, 53(3):55:1–55:37, 2020. doi: 10.1145/3381028 1, 2, 3, 4

[8] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. LOF: Identifying density-based local outliers. In *Proc. SIGMOD*, pp. 93–104. ACM, New York, 2000. doi: 10.1145/335191.335388 2, 4

[9] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3):15:1–15:58, 2009. doi: 10.1145/1541880.1541882 3

[10] Q. Chen, N. Chen, W. Shuai, G. Wu, Z. Xu, H. Tong et al. CalliopeNet: Automatic generation of graph data facts via annotated node-link diagrams. *IEEE Trans. Vis. Comput. Graph.*, 30(1):562–572, 2024. doi: 10.1109/TVCG.2023.3326925 3

[11] D. Daniel and K. West-Mitchell. The Sankey diagram: An exploratory application of a data visualization tool. *Transfusion*, 64(6):967–968, 2024. doi: 10.1111/trf.17803 6

[12] T. de Vries, S. Chawla, and M. E. Houle. Finding local anomalies in very high dimensional space. In *Proc. ICDM*, pp. 128–137. IEEE, Piscataway, 2010. doi: 10.1109/ICDM.2010.151 2

[13] Ç. Demiralp, P. J. Haas, S. Parthasarathy, and T. Pedapati. Foresight: Rapid data exploration through guideposts. *CoRR*, abs/1709.10513, 2017. doi: 10.48550/arXiv.1709.10513 2

[14] D. Deng, A. Wu, H. Qu, and Y. Wu. DashBot: Insight-driven dashboard generation based on deep reinforcement learning. *IEEE Trans. Vis. Comput. Graph.*, 29(1):690–700, 2023. doi: 10.1109/TVCG.2022.3209468 2

[15] V. Dibia and Ç. Demiralp. Data2Vis: Automatic generation of data visualizations using sequence-to-sequence recurrent neural networks. *IEEE Comput. Graph. Appl.*, 39(5):33–46, 2019. doi: 10.1109/MCG.2019.2924636 2

[16] R. Ding, S. Han, Y. Xu, H. Zhang, and D. Zhang. QuickInsights: Quick and automatic discovery of insights from multi-dimensional data. In *Proc. SIGMOD*, pp. 317–332. ACM, New York, 2019. doi: 10.1145/3299869.3314037 1, 2, 3, 4, 9

[17] E. M. Elgaml, D. M. Ibrahim, and E. A. Sallam. Improved FP-growth algorithm with multiple minimum supports using maximum constraints. *Int. J. Comput. Electr. Autom. Control Inf. Eng.*, 9(5):1087–1094, 2015. doi: 10.5281/zenodo.1100847 2

[18] R. M. Gray. *Entropy and Information Theory*. Springer, New York, 2nd ed., 2011. doi: 10.1007/978-1-4419-7970-4 5

[19] Q. Jiang, G. Sun, Y. Dong, L. Pan, B. Chang, L. Jiang et al. FactExplorer: Fact embedding-based exploratory data analysis for tabular data. *Int. J. Hum.-Comput. Interact.*, 41(19):12361–12379, 2025. doi: 10.1080/10447318.2025.2459172 2, 3

[20] A. Joshi, S. Kale, S. Chandel, and D. K. Pal. Likert scale: Explored and explained. *Br. J. Appl. Sci. Technol.*, 7(4):396–403, 2015. doi: 10.9734/BJAST/2015/14975 8

[21] T. Khot, H. Trivedi, M. Finlayson, Y. Fu, K. Richardson, P. Clark et al. Decomposed prompting: A modular approach for solving complex tasks. In *Proc. ICLR*. OpenReview.net, 2023. doi: 10.48550/arXiv.2210.02406 7

[22] R. U. Kiran and P. Krishna Reddy. An efficient approach to mine rare association rules using maximum items' support constraints. In *Proc. BNCOD*, vol. 6121 of *Lect. Notes Comput. Sci.*, pp. 84–95. Springer, Berlin, 2012. doi: 10.1007/978-3-642-25704-9_9 2

[23] Y. S. Koh and S. D. Ravana. Unsupervised rare pattern mining: A survey. *ACM Trans. Knowl. Discov. Data*, 10(4):45:1–45:29, 2016. doi: 10.1145/2898359 1, 2

[24] M. Komorowski, D. C. Marshall, J. D. Salciccioli, and Y. Crutain. Exploratory data analysis. In *Secondary Analysis of Electronic Health Records*, chap. 15, pp. 185–203. Springer, Cham, 2016. doi: 10.1007/978-3-319-43742-2_15 1

[25] H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek. Interpreting and unifying outlier scores. In *Proc. SDM*, pp. 13–24. SIAM, Philadelphia, 2011. doi: 10.1137/1.9781611972818.2 5

[26] G. Li, R. Li, Y. Feng, Y. Zhang, Y. Luo, and C. H. Liu. CoInsight: Visual storytelling for hierarchical tables with connected insights. *IEEE Trans. Vis. Comput. Graph.*, 30(6):3049–3061, 2024. doi: 10.1109/TVCG.2024.3388553 2, 3

[27] Y. Lian, J. Hao, W. Zeng, and Q. Luo. A survey of visual insight mining: Connecting data and insights via visualization. *Vis. Inform.*, 9(4):100271, 2025. doi: 10.1016/j.visinf.2025.100271 1, 2, 9

[28] Y. Lin, H. Li, L. Yang, A. Wu, and H. Qu. InkSight: Leveraging sketch interaction for documenting chart findings in computational notebooks. *IEEE Trans. Vis. Comput. Graph.*, 30(1):944–954, 2024. doi: 10.1109/TVCG.2023.3327170 2

[29] B. Liu, W. Hsu, and Y. Ma. Mining association rules with multiple minimum supports. In *Proc. KDD*, pp. 337–341. ACM, New York, 1999. doi: 10.1145/312129.312274 2

[30] Y. Luo, X. Qin, N. Tang, and G. Li. DeepEye: Towards automatic data visualization. In *Proc. ICDE*, pp. 101–112. IEEE, Piscataway, 2018. doi: 10.1109/ICDE.2018.00019 2

[31] P. Ma, R. Ding, S. Han, and D. Zhang. MetaInsight: Automatic discovery of structured knowledge for exploratory data analysis. In *Proc. SIGMOD*, pp. 1262–1274. ACM, New York, 2021. doi: 10.1145/3448016.3457267 2, 3, 4, 5

[32] P. Ma, R. Ding, S. Wang, S. Han, and D. Zhang. InsightPilot: An LLM-empowered automated data exploration system. In *Proc. EMNLP Syst. Demonstrations*, pp. 346–352. ACL, Stroudsburg, 2023. doi: 10.18653/v1/2023.emnlp-demo.31 2, 3, 9

[33] Y. Ma and L. Zhu. A review on dimension reduction. *Int. Stat. Rev.*, 81(1):134–150, 2013. doi: 10.1111/j.1751-5823.2012.00182.x 6

[34] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, 2008. doi: 10.1017/CBO9780511809071 4

[35] G. Marvin, N. Hellen, D. Jjingo, and J. Nakatumba-Nabende. Prompt engineering in large language models. In *Proc. ICDICI*, Algorithms Intell. Syst., pp. 387–402. Springer, Singapore, 2024. doi: 10.1007/978-981-99-7962-2_30 7

[36] R. T. Ng, L. V. S. Lakshmanan, J. Han, and A. Pang. Exploratory mining and pruning optimizations of constrained association rules. In *Proc. SIGMOD*, pp. 13–24. ACM, New York, 1998. doi: 10.1145/276304.2763074

[37] L. G. Nonato and M. Aupetit. Multidimensional projection for visual analytics: Linking techniques with distortions, tasks, and layout enrichment. *IEEE Trans. Vis. Comput. Graph.*, 25(8):2650–2673, 2019. doi: 10.1109/TVCG.2018.2846735 6

[38] S. Papadimitriou, H. Kitagawa, P. B. Gibbons, and C. Faloutsos. LOCI: Fast outlier detection using the local correlation integral. In *Proc. ICDE*, pp. 315–326. IEEE, Piscataway, 2003. doi: 10.1109/ICDE.2003.1260802 2

[39] J. Pillai and O. P. Vyas. High utility rare itemset mining (HURI): An approach for extracting high-utility rare item sets. *i-Manag. J. Future Eng. Technol.*, 7(1):25–33, 2011. doi: 10.26634/jfet.7.1.1677 2

[40] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manag.*, 24(5):513–523, 1988. doi: 10.1016/0306-4573(88)90021-0 4

[41] C. E. Shannon. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27(3):379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x 5

[42] D. Shi, X. Xu, F. Sun, Y. Shi, and N. Cao. Calliope: Automatic visual data story generation from a spreadsheet. *IEEE Trans. Vis. Comput. Graph.*, 27(2):453–463, 2021. doi: 10.1109/TVCG.2020.3030403 2

[43] L. Szathmáry, A. Napoli, and P. Valtchev. Towards rare itemset mining. In *Proc. ICTAI*, pp. 305–312. IEEE, Piscataway, 2007. doi: 10.1109/ICTAI.2007.30 2

[44] B. Tang, S. Han, M. L. Yiu, R. Ding, and D. Zhang. Extracting top-k insights from multi-dimensional data. In *Proc. SIGMOD*, pp. 1509–1524. ACM, New York, 2017. doi: 10.1145/3035918.3035922 1, 2, 3, 4

[45] J. Tang, Z. Chen, A. W.-C. Fu, and D. W. Cheung. Enhancing effectiveness of outlier detections for low density patterns. In *Advances in Knowledge Discovery and Data Mining*, vol. 2336 of *Lect. Notes Comput. Sci.*, pp. 535–548. Springer, Berlin, Heidelberg, 2002. doi: 10.1007/3-540-47887-6_53 2

[46] Y. Wang, Z. Sun, H. Zhang, W. Cui, K. Xu, X. Ma et al. DataShot: Automatic generation of fact sheets from tabular data. *IEEE Trans. Vis. Comput. Graph.*, 26(1):895–905, 2020. doi: 10.1109/TVCG.2019.2934398 2

[47] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia et al. Chain-of-thought prompting elicits reasoning in large language models. In *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 24824–24837. Curran Associates, Red Hook, 2022. doi: 10.52202/068431-1800 7

[48] L. Weng, X. Wang, J. Lu, Y. Feng, Y. Liu, H. Feng et al. InsightLens: Augmenting LLM-powered data analysis with interactive insight management and navigation. *IEEE Trans. Vis. Comput. Graph.*, 31(6):3719–3732, 2025. doi: 10.1109/TVCG.2025.3567131 2, 3

[49] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert et al. A prompt pattern catalog to enhance prompt engineering with ChatGPT. In *Proc. PLoP*, pp. 1–31. The Hillside Group, Monticello, 2023. doi: 10.64346/PLoP2023p05 7

[50] S. Wu, W. Luan, Y. Wang, D. Zeng, Q. Shen, and B. Tang. Data insights as data: Quick overview and exploration of automated data insights. In *Proc. CHI EA*, pp. 167:1–167:7. ACM, New York, 2025. doi: 10.1145/3706599.3719702 1, 2, 3

[51] D. Zhou, N. Schärli, L. Hou, J. Wei, N. Scales, X. Wang et al. Least-to-most prompting enables complex reasoning in large language models. In *Proc. ICLR*. OpenReview.net, 2023. doi: 10.48550/arXiv.2205.10625 7