J Vis (2024) 27:503–520
https://doi.org/10.1007/s12650-024-00975-1

REGULAR PAPER

Qi Jiang · Guodao Sun · Tong Li · Jingwei Tang · Wang Xia · Sujia Zhu · Ronghua Liang

Qutaber: task-based exploratory data analysis with enriched context awareness

Received: 11 November 2023 / Accepted: 31 January 2024 / Published online: 11 March 2024

Abstract Exploratory data analysis (EDA) has emerged as a critical tool for users to gain deep insights into data and unearth hidden patterns. The integration of recommendation algorithms has enhanced its capabilities and further popularized its utilization. Most recommendation-based EDA methods concentrate on the extraction of pivotal insights from datasets, and the taxonomy of these insights is well-established. However, the support for further analytical endeavors to expand these initial findings remains constrained, as evidenced by the restricted scope of analytical intents that are tailored to specific scenarios. Moreover, these systems often lack sufficient context-awareness capabilities, failing to equip users with the necessary tools for a thorough exploration of extensive recommendations. To address these limitations, we introduce Qutaber, a task-based EDA system with enriched context-awareness. We first summarize six core analytical tasks tailored for EDA scenarios through literature reviews and expert interviews. Then, Qutaber integrates the use of small multiples, enhanced with a multi-metric re-ranking function, to enable a thorough and efficient examination of expanded charts pertaining to various analytical tasks. Furthermore, a machine learning method is leveraged to characterize the semantic features of these charts for a holistic landscape of recommended charts. Finally, a case study using a real-world dataset demonstrates Qutaber's practical application, followed by a user study to further evaluate the usability of the proposed techniques. Our findings illustrate that Qutaber facilitates an effective and context-rich EDA experience for users.

Keywords Exploratory data analysis · Task-based · Mixed-initiative interaction · Context-awareness

Q. Jiang · G. Sun (✉) · T. Li · J. Tang · W. Xia · S. Zhu · R. Liang
College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou, China
E-mail: guodao@zjut.edu.cn

Q. Jiang
E-mail: jiangqi@zjut.edu.cn

T. Li
E-mail: litong@zjut.edu.cn

J. Tang
E-mail: jwtang@zjut.edu.cn

W. Xia
E-mail: xiawang@zjut.edu.cn

S. Zhu
E-mail: zhusujiashiny@gmail.com

R. Liang
E-mail: rhliang@zjut.edu.cn

1 Introduction

Exploratory data analysis (EDA) is a powerful and widely used method that helps users deepen their understanding of data and uncover hidden patterns (Li et al. 2023). Traditionally, EDA is characterized by an iterative process of exploration, discovery, and re-exploration, which allows users to flexibly adjust their analytical strategies based on continuously emerging insights. Despite its benefits, prior EDA tools often required users to possess ample domain expertise for the manual selection of data dimensions, aggregation operations, and visual encoding (Battle and Heer 2019; Wongsuphasawat et al. 2017). However, when dealing with large and complex datasets, finding meaningful patterns can be a tedious and time-intensive task.

Recently, researchers have attempted to address these limitations by incorporating recommendation systems into EDA (Mackinlay et al. 2007; Dibia and Demiralp 2019; Hu et al. 2019; Luo et al. 2018; Wongsuphasawat et al. 2015, 2017; Zhou et al. 2023). These systems streamline the exploratory process for users by providing a set of suitable suggestions for effective visual encodings or potentially insightful visualizations. According to the analytical workflow, recommendation-based EDA systems are broadly categorized into two categories. The first category focuses on extracting insights from a given dataset (e.g., identifying a strong correlation between two variables in a scatterplot), with the goal of swiftly familiarizing users with the preliminary facets of their data and emphasizing notable patterns (Dibia and Demiralp 2019; Shen et al. 2022; Ding et al. 2019; Shen et al. 2021). The second category seeks to enhance users' comprehension of these initial insights by suggesting avenues for further analysis (Wongsuphasawat et al. 2017; Pandey et al. 2023). Most existing work in visualization recommendation concentrates on the first category, with extensive taxonomic studies undertaken to classify these insight tasks (Amar et al. 2005; Shen et al. 2021). However, guidance for post-insight analysis is less represented in the literature. Studies addressing this area often display a limited scope, typically focusing on a restricted array of analytical intents (Wongsuphasawat et al. 2017; Hu et al. 2018; Cui et al. 2019), or are customized for particular use cases (Pandey et al. 2023; Deng et al. 2022). Thus, the first challenge is *how to provide nuanced and insightful analytical task suggestions within the scope of EDA to effectively guide users toward more profound insights*.

Most existing recommendation-based systems opt to present a limited selection of key suggestions, aiming to streamline the decision-making process and reduce cognitive burden (Tang et al. 2017; Cui et al. 2019). While this method has its merits, it also has its inherent limitations (Roy and Dutta 2022). An overly concise selection of recommendations might inadvertently lead users to overlook other potentially valuable suggestions, thereby restricting their ability to make fully informed choices. Moreover, a narrower range of choices can exacerbate the effects of algorithmic bias, potentially skewing users' decisions away from their genuine interests and needs (Li et al. 2021). Nonetheless, most of the existing systems lack the necessary tools for users to delve into an extensive spectrum of recommendations. This leads to a secondary challenge regarding *how to help users gain a productive and comprehensive insight into available analysis suggestions*.

To address these challenges, we introduce a task-based EDA with enriched context-awareness. Previous research has highlighted the critical role of user-driven analytical tasks in determining the effectiveness of visualizations (Saket et al. 2018; Shen et al. 2021). To this end, we first identified six core analysis tasks specific to EDA (including their analysis intents and operations), which was accomplished through an extensive literature review and interviews with domain experts. To provide a detailed view of the extensive analytical suggestions, we opted for the utilization of small multiples to display potential charts, organizing them according to their associated tasks. While introducing a broader spectrum of recommendations has the potential to increase cognitive load, this effect can be mitigated since the suggested tasks are closely related to the visualization currently engaging the user. This situation narrows the search space and confines the scale of recommended results to a manageable scope. Furthermore, our system also incorporates a multi-metric re-ranking mechanism within the small multiples, assisting users in efficiently navigating and prioritizing visualizations. To further enhance the holistic perception of provided analytical recommendations, a machine learning approach was utilized to capture the semantic features of these task-related visualizations. The generated semantic vectors are then projected onto a two-dimensional scatterplot for a holistic landscape.

These ideas mentioned are implemented in a task-based EDA system, named Qutaber. The system adopts a top-down exploration model, which begins with providing users with an initial comprehension of their dataset by showcasing its overarching characteristics. Next, the system recommends charts that illustrate significant patterns within the data to ease users into the analytical process. When users seek to delve into particular findings of interest, they naturally engage in a task-based EDA process. This process is iterative, involving a cycle of discovery and re-exploration of emerging patterns. Finally, we conducted a case study and a user study using real-world datasets to evaluate the usability of Qutaber and verify its effectiveness in promoting data analysis. Reviewing the building and evaluation through Qutaber, we highlight the following contributions.

- A design space tailored for user-centered analysis tasks was identified through expert interviews and a review of relevant literature to facilitate the expanded analysis of focused visualizations.
- A suite of visual analytics tools to enhance contextual awareness in data exploration, including small multiples enhanced by multi-metric re-ranking feature, and a machine learning-based chart embedding approach to provide a holistic picture of all task-related visualizations.
- A runnable EDA system, named Qutaber, is engineered to empower users in conducting thorough and effective data exploration. The usability and efficacy of Qutaber have been empirically validated through a case study and a user study employing real-world datasets.

2 Related work

In this section, we elaborate on the recent studies that are most relevant to this work, including visual analysis tasks and exploratory data analysis.

2.1 Visual analysis tasks

Visual analysis tasks are essential for understanding and describing the diverse activities undertaken by users as they explore and analyze complex data. The academic community has developed well-established taxonomies to define and categorize these analytical tasks. One of the earliest and most influential task taxonomies was proposed by Wehrend and Lewis (1990), who divided visualization tasks into eleven categories. Later, Amar et al. (2005) refined this taxonomy by using affinity diagramming and proposed ten low-level analytical tasks (including *retrieve value, filter, compute derived value, find extremum, sort, determine range, characterize distribution, find anomalies, cluster, and correlate*). Moreover, there are other studies that synthesized these tasks into two broader groupings: value tasks and summary tasks (Jiang et al. 2021; Moritz et al. 2018). Beyond defining and categorizing these tasks, empirical research has also been conducted to assess the efficiency of different charts and visual representations in facilitating these tasks (Saket et al. 2018; Kim and Heer 2018).

The analytical tasks mentioned above aim to assist users in deriving insights from datasets, such as detecting outliers, identifying growth trends, and uncovering strong correlations. Furthermore, further investigations by researchers are delving into subsequent analytical tasks that build upon and deepen these preliminary findings. For instance, FlowSense (Yu and Silva 2019) summarized six common analytical tasks in the visual analytics workflow, including *visualizing, visual encoding, filtering and finding extremum, subset manipulation, highlighting, and linking*. Pandey et al. (Pandey et al. 2023) facilitated users to create a dashboard by summarizing four categories of analytic intent, along with their ten corresponding tasks. Moreover, research efforts have also been devoted to investigating particular tasks within EDA, e.g., drilling down (Lee et al. 2019; Joglekar et al. 2015) and comparing (Law et al. 2018; Wang et al. 2022). Additionally, the effectiveness of these analytical tasks has been evaluated by empirical studies. Lee's work (Lee et al. 2021) is one of these, in which they probed into the practicality and impact of ten task categories in the visual analysis process.

In this work, we concentrate on the analytical tasks that become crucial after initial insights are acquired. While our work is grounded in the analytical tasks identified within the existing literature, we enhance and expand these frameworks with valuable insights obtained from interviews with domain experts. Specifically, we draw upon the task classifications proposed by Lee et al. (2021) and Pandey et al. (2023), alongside a synthesis of task classifications from a diverse set of studies (Amar et al. 2005; Shen et al. 2021; Wongsuphasawat et al. 2017; Yu and Silva 2019; Shen et al. 2022). Following this, we refine the collected task categories through expert interviews, extending their scope to incorporate additional analytical scenarios and intentions.

2.2 Exploratory data analysis

EDA stands as a critical phase within the data science workflow, and it is often enhanced by diverse visualization techniques (Sun et al. 2013; Zhao et al. 2022; Zhu et al. 2023). The iterative nature makes it allow analysts to incrementally discover insights and adjust their analysis strategies in real-time. Early EDA tools (e.g., Excel) relied heavily on the expertise of the analyst, necessitating their involvement in nearly all aspects of data analysis, from the selection of data fields and corresponding aggregation functions to the crafting of visual representations. While these tools provided certain intuitive views for data examination, their reliance on expertise still makes them both time-consuming and laborious.

Recognizing these limitations, the academic community has introduced visualization recommendation systems to streamline the exploratory process. Initial work mostly used rule-based algorithms to offer visual suggestions by assessing the characteristics and distributions of data. For instance, Show Me (Mackinlay et al. 2007) implemented automatic analysis of the attributes selected by users to suggest suitable visual designs. Following this, Voyager (Wongsuphasawat et al. 2015) and Voyager2 (Wongsuphasawat et al. 2017) advanced the field by using heuristics to recommend alternative visualizations, thereby improving user efficiency in data exploration. Similarly, Foresight (Demiralp et al. 2017) automated insight extraction and allowed further inquiries into related insights. Additionally, Draco (Moritz et al. 2018) introduced a knowledge-based framework for visual design, employing constraint-solving techniques to refine visual encodings and recommendations. In contrast to rule-based approaches, another research branch has explored data-driven models for generating visualization suggestions. Representative work such as Data2Vis (Dibia and Demiralp 2019), DeepEye (Luo et al. 2018), and VisML (Hu et al. 2019) have employed deep learning to identify patterns and designs within extensive datasets of user-generated charts. Despite the advancements in data-driven approaches, our research opts for a rule-based methodology to generate suggestions for subsequent analysis, prioritizing its proven reliability and the transparency of its results.

Previous systems have predominantly concentrated on the precision of recommendations, often neglecting to provide a broader context for these suggestions. Concurrently, the visualization community has seen a growth in interest in chart similarity computation and chart embedding research. For instance, Chart Constellations (Xu et al. 2018) employed a hybrid approach to characterize charts from various perspectives, facilitating a collective analytical process. Similarly, ChartSeer (Zhao et al. 2020) used a deep learning method to interpret analyst-generated charts to generate visual summaries. Inspired by these studies, we also aim to enhance users' contextual perception of recommended charts by providing a global view. Nonetheless, these efforts may grapple with challenges stemming from their dependence on heuristic rules or high-quality training datasets, potentially restricting their broader application and practical deployment. Diverging from existing approaches, our method utilizes a pretrained textual model to implement chart embedding. This process begins with the transformation of chart specifications into detailed textual narrative descriptions, which are then used by the pretrained model to generate semantic vectors. Furthermore, we integrate the use of small multiples with a multi-metric re-ranking mechanism to improve the context of exploration and the efficiency of browsing.

3 Preliminary study

In order to design a more efficient, scalable, and practical Qutaber, it is essential to gain a deep understanding of users' core tasks and requirements as they engage in data analysis exploration. Therefore, we conducted this preliminary study with the aim of providing powerful support for succeeding system design and implementation.

The preliminary study is comprised of two principal components. First, we conducted a literature review of previous studies on EDA (Mackinlay et al. 2007; Wongsuphasawat et al. 2015, 2017; Shen et al. 2022; Wang et al. 2022; Li et al. 2023; Joglekar et al. 2015; Lee et al. 2019; Cui et al. 2019) and visualization recommendation (Zhu et al. 2020; Demiralp et al. 2017; Lee et al. 2021; Shen et al. 2021) to collect the analytical tasks commonly used by users during their detailed exploration of data. Second, we conducted semi-structured interviews with two professional data scientists (referred to as E1 and E2 below), aiming to discuss the analysis tasks identified from the literature review and to gain insights into their specific requirements for a task-based EDA system. Both experts hold PhDs in computer science, are employed as research faculty at universities, and have over five years of experience in the domain of data visual analysis.

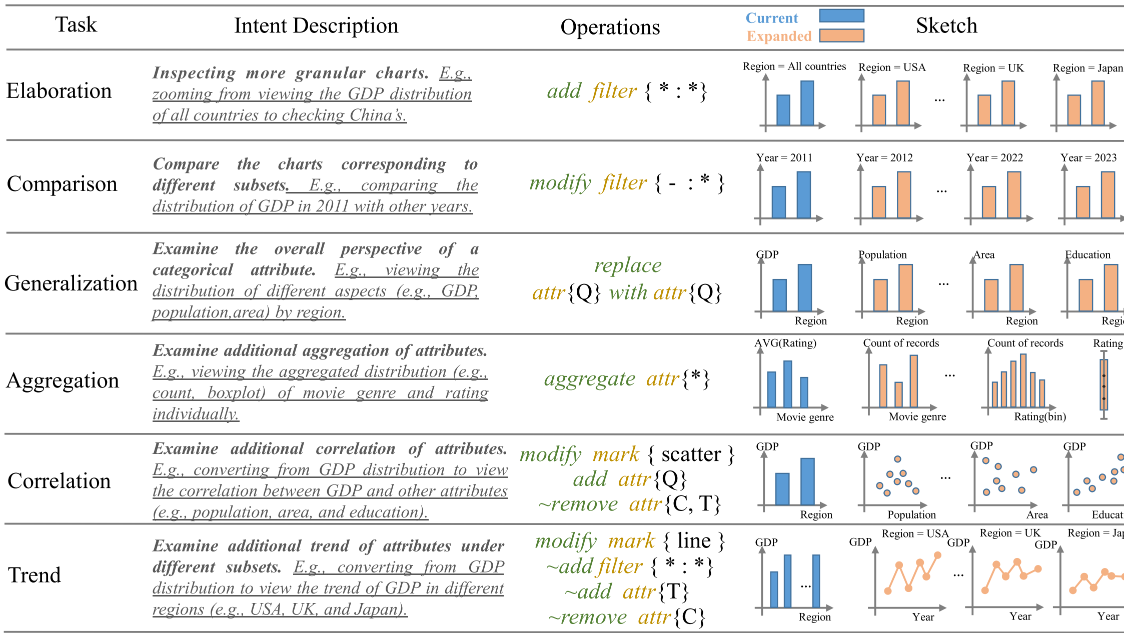

Fig. 1 Design space for expanded analysis tasks. Each row within the design space represents an expanded analysis task. The *Intent Description* column elucidates the analytical objectives underlying these tasks, providing insight into the expected outcomes of each analysis. In *Operation* column, actions and objects of operations are color-coded for intuitive understanding: actions to be undertaken are highlighted in green, the objects of these actions are marked in yellow, and parameters governing the operations are enclosed in curly brackets { }. Moreover, attributes are categorized by type—**Q**: quantitative, **C**: categorical, and **T**: temporal. Actions that are not marked with a tilde (~) indicate that they will definitely be executed, while those marked with this symbol indicate that they may be executed. For *filter*, the part preceding (**:**) denotes the attribute used for filtering, while the part following (**:**) denotes the value being filtered, an asterisk (*) indicates all possible options, while a hyphen (**-**) signifies that the original options should remain unchanged. In *Sketch* column, the blue chart denotes the current chart, while the orange charts represent the expanded charts

During the interview process, we first presented analysis tasks identified during the literature review to experts, seeking their insights and recommendations. Through collaborative dialogue, we structured and defined the fundamental aspects of these tasks, including the task type, its analytic purpose, and the specific operations required for implementation. Furthermore, we also broadened the scope to include additional scenarios and their analysis tasks where necessary. *Correlation* and *Trend* are examples of these as their practical significance and their capacity to produce meaningful insights in real-world analysis contexts. Overall, we excluded tasks that are easily accomplished via basic chart editing tools, such as alternative encoding methods (Wongsuphasawat et al. 2017). Besides, complex statistical analyses like linear regression have not been explored in this context (Hu et al. 2018). Currently, our emphasis is on analytical tasks that offer a broader spectrum of outcomes and richer informational depth. To manage complexity and ensure feasibility, we ultimately identified six key analysis tasks, namely *Elaboration, Comparison, Generalization, Aggregation, Correlation, and Trend*. These tasks have been applied to three essential chart types: *bar, line, and scatterplot*, as illustrated in our analysis task space (Fig. 1). For enhanced understanding, we also augmented the task space with representative sketches, specifically using bar charts. Notably, our focus in this work has been on defining these core analysis tasks and validating their effectiveness within the fundamental chart types, with more chart scenarios (e.g., pie charts and maps) to be implemented in the future.

Another branch of the expert interviews involves identifying the challenges and requirements they face when using task-based EDA systems. For this purpose, we formulated a series of targeted questions to uncover the practical hurdles that experts face in real-world application scenarios. These pivotal questions encompassed: *"What difficulties and challenges have you encountered in using existing EDA tools?,"* *"What are the limitations you perceive in recommendation-based EDA systems?,"* *"What features and capabilities do you envision for an ideal task-based EDA system?,"* *"How does the level of context-awareness in existing EDA systems impact your analytical processes?,"* and *"How do you think context-awareness can be enhanced in EDA systems?."* The interview process lasted about 90 min, and at the end, we collected comments and requirements from experts, which are also described in the follow-up requirement analysis (Sect. 4). Moreover, an extended discussion of specific techniques of visual design is involved, which is presented in Sect. 5.3.

4 The design of Qutaber

In this section, we delve into the analytical requirements of our target users and delineate the design goals for Qutaber. Moreover, we detail the system's workflow, offering a comprehensive overview of how Qutaber operates.

4.1 Requirement analysis

Qutaber is designed as an EDA system tailored for tabular data, which aims to augment the capabilities of professional data analysts by providing a task-oriented approach coupled with enriched context-awareness during data exploration. Through interviews with domain experts and a comprehensive review of existing EDA literature, we have summarized four requirements that Qutaber is engineered to address.

R1. Understand the general characteristics of the data. Most existing EDA tools commence with presenting a data overview, which enables users to gain an initial understanding of the dataset (Ghosh et al. 2018; Sun et al. 2021). This point is also indicated in our interviews with experts, both E1 and E2 concurred that an introductory overview of the dataset is essential for EDA tools. E2 commented, "*Providing users with a statistical-level data overview enables a rapid acquisition of the data's fundamental characteristics, fostering further detailed exploration*." Therefore, users first expect our system to provide an overview of data characteristics at the statistical level.

R2. Steer proactive and task-based data exploration. Following the discovery of interesting patterns, users often embark on an in-depth exploration of these findings. Accordingly, there is an expectation for the system to offer efficient and convenient support as users widen their investigative scope. This necessitates the system's ability to proactively provide relevant suggestions for further exploration that align with the users' analytical goals or intents (Srinivasan et al. 2018; Shen et al. 2022). E1 underscored this by stating, "*The EDA system should automate low-level and labor-intensive operations and enable a task-oriented guidance for subsequent analysis, thus reducing the exploratory workload for users*."

R3. Provide enriched context-awareness. As users navigate through the iterative process of data exploration, they also demand that our system be able to provide an enriched contextual backdrop (Wongsuphasawat et al. 2019; Demiralp et al. 2017). This requirement has been echoed by both experts who pointed out that existing recommendation-based EDA tools often fail to meet this need, as they focus on providing more precise recommendations. Furthermore, E1 added, "*Effectively capturing and presenting enriched contexts is crucial in guiding users to discover more potential insights*." Therefore, users expect our system to incorporate comprehensive contextual information throughout the exploration process, enabling users to identify additional data patterns and make well-informed decisions.

R4. Inspect individual visualization charts. The inspection and modification of visualization charts are frequent analytical activities in EDA. It is not solely about information gathering, also encompasses the need for users to swiftly adjust visualizations to validate hypotheses. Such an iterative endeavor necessitates an interface designed for seamless reconfiguration of visual displays (Wongsuphasawat et al. 2015, 2017). Both E1 and E2 further underscored the necessity of this feature, advocating for a system that provides an accessible and intuitive view for users to inspect and modify visualization charts. Therefore, there is a clear expectation among users for a feature like a parameter panel, which would enable them to easily interact with and edit visualizations.

4.2 Design goals

With the aim of better designing and implementing Qutaber, we have delineated the functionalities that the system should support, deriving from the previously detailed requirements. The specific design goals are as follows.

G1. Present a data overview and avoid cold starts. At the outset of EDA, users face challenges in determining the appropriate initial steps due to unfamiliarity with the dataset. Therefore, the system should provide a comprehensive overview of the data, elucidating key aspects such as the number of dimensions, their respective categories, and distribution patterns (**R1**). This foundational overview is crucial for enabling users to swiftly grasp the intrinsic characteristics of the dataset. Furthermore, to avoid cold starts, the system should provide preliminary insights that highlight significant data points and emerging patterns, thus equipping users with a starting point for deeper and more targeted exploration (**R2**).

G2. Support task-based expansion of user focus. The nature of EDA entails an iterative process of exploration, discovery, and re-exploration. With this in mind, the system should support a task-oriented expansion of the user's focus by providing a selection of automatically generated analytical recommendations. These suggestions are not arbitrary but aligned with the user's ongoing analysis intentions. To facilitate the effective expansion of exploration, it is also essential to comprehend the analytical tasks users undertake during in-depth exploration and to empower them with the flexibility to select from task-related recommendations that align with their specific analytical needs(**R2**).

G3. Provide an enriched context for iterative exploration. The enhancement of context within EDA plays a pivotal role in uncovering more profound patterns and facilitating a more informed decision-making process. To this end, it is imperative that the system offers enriched contextual backing throughout each iteration of the user's exploratory journey. This entails ensuring that users have access to a wider array of insights and the ability to prioritize this information according to various criteria. Additionally, the system should also provide a visual summary of all recommendations by characterizing their semantic features, thus further elucidating their relevance and background (**R2 and R3**).

G4. Enable convenient user interaction. Intuitive user interfaces and convenient user interactions are crucial for an EDA system, as these ensure that users are able to quickly become proficient with the tools. In this regard, the system is designed to offer an intuitive interface, thereby reducing the learning curve for novice users Furthermore, our system should also incorporate real-time and accessible interactive functions to meet diverse analytical requirements. These features support a range of functionalities, including the selection of a focused view, real-time modifications to charts, and the ability to engage in deeper analysis directly from the focused view (**R1 to R4**).

4.3 System workflow

Qutaber, designed as a task-based EDA system, adopts a top-down exploration approach to conduct the analytical process. This system empowers users to concentrate on advanced analytical endeavors, minimizing distractions from menial and repetitive operations, thereby improving overall efficiency and effectiveness in data exploration. Furthermore, Qutaber enhances the discovery of potential patterns and analytical decision-making by providing enriched contextual information. As shown in Fig. 2, the system workflow can be divided into two stages: S1 data overview and initial recommendations; S2 task-based EDA with enriched context-awareness.

S1. Data overview and initial recommendations. In this phase, the system presents key characteristics of the uploaded dataset to provide users with an initial insight into its structure. To avoid cold starts, Qutaber executes an automated inspection of data patterns, offering preliminary recommendations. These insights aim to spark user curiosity and pave the way for further EDA processes.

S2. Task-based EDA with enriched context-awareness. When focusing on a specific chart, users progress to more in-depth task-based EDA. In this context, the system dynamically generates expanded charts tailored to various analytical tasks identified in Sect. 3, which are derived from the chart currently under review. These expanded charts are organized into small multiples, categorized by task type to which they pertain, thereby facilitating comparative analysis and pattern recognition. A re-ranking feature is also integrated to allow users to customize the arrangement of these charts according to their individual preferences and requirements. Furthermore, Qutaber leverages machine learning techniques to present a semantic overview of these expanded charts, enriching the user's contextual awareness.

5 The Qutaber system

In this section, we introduce the Qutaber system, delineating its visual design, the array of techniques it utilizes, and the methods by which users leverage these for exploratory analysis.

The user interface of Qutaber, illustrated in Fig. 3, offers an intuitive data analysis platform for users. Next, we delve into the system's core views which include data summary view (Fig. 3B), initial recommendation view (Fig. 3C), expansion detail view (Fig. 3E), and expanded charts overview view (Fig. 3F1).

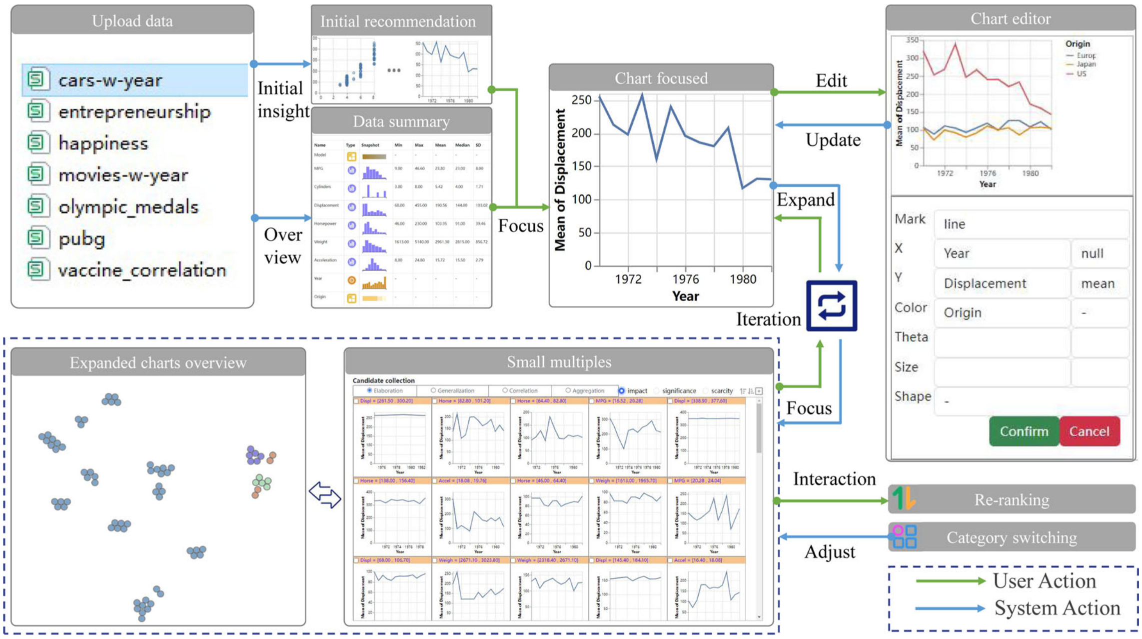

Fig. 2 The workflow of Qutaber. First, the system automatically generates a data overview and provides certain initial insights based on the uploaded data. From these, users then focus on a specific chart and perform an in-depth and iterative exploration. Note that, the green line indicates system action, and the blue line indicates user action

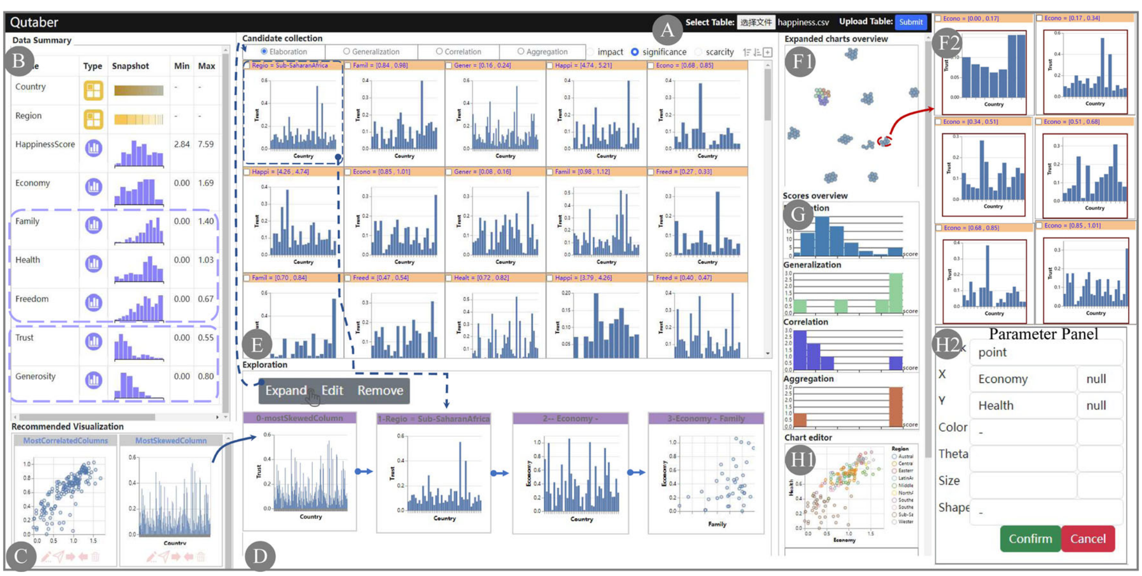

Fig. 3 The user interface of Qutaber. When a dataset is uploaded (**A**), the statistics of the dataset are presented in the data summary view (**B**), and certain initial recommendations are presented in (**C**). Pick a specific chart into the exploration area (**D**), where users can expand the chart for follow-up analysis. The expanded charts are presented by category in (**E**) and their corresponding score distributions are displayed in (**G**), and these charts are allowed to be re-ranked by different metrics. The overall overview of all expanded charts is shown in (**F1**), and (**F2**) are the charts for a cluster in (**F1**). **H1** and **H2** are the editing areas of the selected chart

5.1 Data summary view

The data summary view is engineered to present a comprehensive overview of key characteristics pertaining to the uploaded dataset. As depicted in Fig. 3B, Qutaber enumerates the dataset's attributes along with their respective categories (e.g., temporal, quantitative, or categorical). Additionally, the system offers essential statistical indicators commonly employed in quantitative data analysis (i.e., minimum, maximum, mean, median, and standard deviation). Data snapshots are also available in the form of histograms for a detailed view of the data distribution.

5.2 Initial recommendation view

Within this view (Fig. 3C), the system automatically extracts significant patterns within the data to help users swiftly identify key information. Currently, our system leverages a rule-based approach to generate these insights rather than employing a machine-learning approach. The decision to do this is motivated by the former’s ability to yield more controlled, predictable, and interpretable outcomes. Ultimately, we selected three categories of insights for initializing recommendations—correlation, trend, and distribution. The primary rationale for this choice is twofold. First, these insights are not only commonly utilized but also critical in shaping the foundational comprehension of the dataset. Second, our strategy is to begin with the most accessible and universally employed insights, thereby ensuring that users are presented with a concise and digestible entry point. From there, as users become more comfortable and skilled in data analysis, they can progressively engage with more complex analytical concepts.

I1. Correlation is understood as the degree to which variables are related or dependent on each other. For this type, the system recommends a scatterplot that shows the most relevant attribute pair.

I2. Trend analysis examines the movement or pattern of data points over a period. Accordingly, the system recommends line charts that reveal the attributes with the most significant increases or decreases. However, if the dataset does not include time-related attributes, no recommendations of this kind are offered.

I3. Distribution is a representation of the frequency or occurrence of values, illustrating how the data is dispersed or concentrated across different categories or ranges. For these insights, bar charts are recommended to demonstrate attributes with the most notable or subtle variations in distribution.

5.3 Expansion detail view

When users focus on a specific view, they have the option to perform an “expand” action to obtain further analysis views based on different tasks (summarized in Fig. 1). During this process, the system automatically proceeds in sequence with the following steps: task expansion, chart rectification, and visual presentation. The details of these steps are described as follows.

Task expansion. For this stage, the input is the chart currently focused, and the system outputs all available expanded charts accordingly. More specifically, the system identifies which of the six predefined analysis tasks (Fig. 1) are executable based on the properties of the chart currently focused. This decision-making process is informed by the conditions necessary for each task, e.g., *Trend* requires temporal attributes in the dataset, while *Comparison* needs at least one filter to be present on the focused chart. Once the executable tasks are determined, the system automatically performs the corresponding analysis actions. An example is provided where the *Elaboration* task is applied to a specific chart titled “*0-mostSkewedColumn*” (Fig. 3). In this case, the system generates expanded charts by applying various filters.

Chart rectification. During this stage, a rule-based method is employed to identify and correct any conflicts or errors that arise during expansion. This approach is chosen for its reliability and is grounded in established practices from previous studies (Moritz et al. 2018; Chen et al. 2021). The detection rules and rectification actions in our current implementation are drawn from prior literature, in particular by VizLinter (Chen et al. 2021). Specifically, we examine the generated charts by iterating through their corresponding Vega-Lite specifications. If a specific error is identified, a pre-defined resolution is executed. For instance, consider a line chart with the X-axis representing years. When performing the expansion for the *Elaboration* task by adding an additional filter (e.g., Year = 2023), the expanded chart is in violation (the X-axis being Year and the filter “Year = 2023” are in conflict for a line chart). Accordingly, the solution to the conflict is to remove the X-axis channel.

Visual presentation. Next, we delineate the display techniques for charts derived from task-based expanded chart extractions. Our preliminary study involved interviewing with experts on three visual designs for presenting multiple charts from previous research: (1) small multiples, which displays numerous chart instances in a single view (Mackinlay et al. 2007; Lekschas et al. 2020; Brehmer et al. 2019); (2) top-k, which just presents the salient k items a broader dataset (Wongsuphasawat et al. 2015); and (3) pagination, which entails dividing the collection of charts into separate pages (Demiralp et al. 2017). The consensus among the experts was that the small multiples outperformed the other options in providing comprehensive overview and facilitating inter-chart comparisons. Another reason supporting this choice is empirical evidence from our tests, which suggests that the charts resulting from expansion are generally moderate in size. It is worth noting that small multiples here is a simplified version that does not necessitate uniform axes or consistent measurement scales.

Further, arranging these charts in a well-organized manner, e.g., by placing charts of high relevance adjacent to each other, helps to reveal patterns and improve navigation efficiency. With this in mind, we introduce an interactive re-ranking feature, which empowers users to personalize chart arrangements based on specific metrics. After investigating the related work on chart assessment, we summarized the following three primary metrics: impact, significance, and scarcity. These metrics were chosen due to their ability to encapsulate the essential attributes of charts. Additionally, they provide a multifaceted yet precise approach to assess the informational value that each chart conveys.

- *Impact*. This impact metric is utilized to quantify the data density of charts (Ding et al. 2019; Shi et al. 2020; Tang et al. 2017). Charts with a greater number of data points are assigned a higher impact score. Accordingly, we determine the impact score by calculating what percentage of the data items are accounted for in each chart. Users can employ this metric to decide whether to prioritize more information-rich charts with a higher number of data points or simpler charts with fewer data points.
- *Significance*. This metric measures the statistical characteristics of a chart, i.e., charts with more distinct statistical features (e.g., strongly correlated scatter charts, fast-growing line charts) tend to receive higher significance scores (Ding et al. 2019; Tang et al. 2017; Cao et al. 2022). Various statistical methods, such as standard deviation, series analysis, and Pearson's correlation coefficient, are used to calculate the scores for this indicator.
- *Scarcity*. The scarcity metric is used to evaluate the rarity of data items presented in a chart within the broader collection of charts (Shi et al. 2020). It is calculated by determining the occurrence frequency of these data items. Users are allowed to prioritize the display of charts with more unique data items or those with more commonly occurring data.

In the end, we calculate an overall score for each chart by performing a weighted sum of the scores for impact, significance, and scarcity. Based on experiments, the default weights for these metrics are 0.5, 0.25, and 0.25, respectively. The distribution of these scores for the expanded charts is depicted in Fig. 3G. This allows users to easily notice variations in chart importance, both across different tasks and within each individual task, thereby facilitating the identification of key information locations.

5.4 Expanded charts overview view

In the expansion detail view, the expanded charts are presented using small multiples categorized by task. Further, we enhance this overview by characterizing the semantic features of these charts, thereby enabling users to understand the interrelations among various charts across different task categories. To achieve this, we employ the Universal Sentence Encoder (USE) (Cer et al. 2018) to vectorize the charts. The resultant vectors are dimensionally reduced to a two-dimensional space using the t-SNE (van der Maaten and Hinton 2008), and then projected onto a scatterplot for a holistic picture. More specifically, the entire process is depicted in Fig. 4, which consists of three distinct phases.

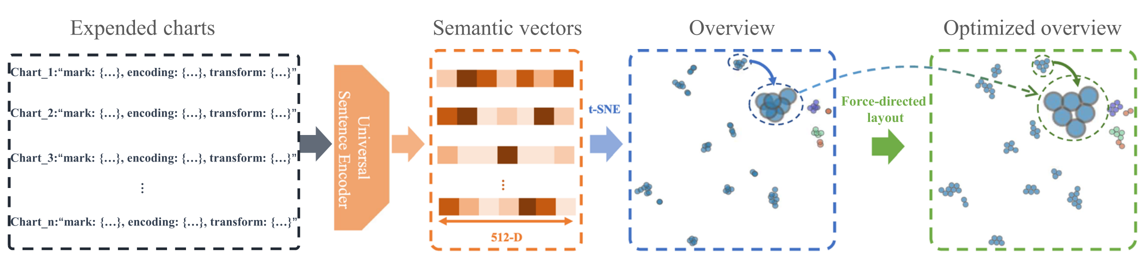

Fig. 4 The illustration of chart embedding, dimensionality reduction, and layout optimization. Firstly, the charts in Vega-Lite format are converted into string format, and after eliminating the uninformative fields, they are input to the Universal Sentence Encoder (USE) model to get the embedded vectors, these vectors are dimensionality reduced by t-SNE and projected to a 2-dimensional scatter plot, and finally, the scatter overlap is optimized by a force-guided algorithm

Semantic embedding Our approach begins by transforming charts presented in Vega-Lite format into semantic vectors. This process involves several potential methods for data encoding, including hard coding, auto-encoders, and unsupervised pretrained models. Among these, unsupervised pretrained models are more suitable for our task, due to their superior semantic information capturing ability and broader applicability compared to the more rigid hard coding and the resource-intensive auto-encoders. Ultimately, we opted for USE (Cer et al. 2018) over other pretrained models like Bert (Devlin et al. 2018) due to its superior real-time performance and out-of-the-box usability. Specifically, we begin by converting charts with the Vega-Lite specification into textual string format, removing fields with unnecessary information other than *mark*, *encoding*, and *transform*. These strings are then fed into the USE model, resulting in the generation of a 512-dimensional semantic vector.

Dimension reduction and visual projection Next, our system applies dimensionality reduction (DR) on the semantic vectors generated in the preceding stage, and then maps them onto a scatterplot for visual analysis. To determine the most suitable DR method for our application, we undertook an empirical evaluation using a dataset pertaining to *cars*.[1] This evaluation aimed to compare the effectiveness of several DR methods, including t-SNE (van der Maaten and Hinton 2008), UMAP (McInnes et al. 2018), MDS (Borg and Groenen 2006), and PCA (Abdi and Williams 2010). The results (Fig. 5) indicated that nonlinear DR techniques (e.g., t-SNE, UMAP) are more effective than linear methods (e.g., PCA) in capturing the local features of semantic vectors. Ultimately, we selected t-SNE as the default DR method, as it demonstrated superior performance in preserving local structural features in our iterative tests. Subsequently, these vectors are mapped to a scatterplot, where each point symbolizes a specific chart, the color of each point denotes different task types, and the details of the chart are presented through interactive tooltips.

Layout optimization Moreover, we identified the presence of overlapping points in the previously mentioned scatterplot, which hindered users from scrutinizing local structural details. Regarding this, our goal was to reposition these points to minimize overlap while maintaining their original distribution characteristics. Initially, we modeled this as an optimization problem focusing on minimal distribution changes and preventing overlaps. However, this approach was computationally intensive, impacting its practicality. To identify a more efficient optimization method, we conducted an analytical investigation on the scatterplot distributions in task-based EDA scenarios. Our findings indicate that in the vast majority of cases, the overview of the expanded charts exhibits dense intra-cluster distribution with distinct inter-cluster boundaries (e.g., the t-SNE result in Fig. 5). This distributional feature is particularly suitable for the application of force-directed algorithms, and such algorithms also meet our needs for efficiency and ease of implementation. Specifically, to maintain the scatterplot's original distribution, we applied horizontal (forceX) and vertical (forceY) forces, with the corresponding parameters being the x and y coordinates of the point, respectively. Meanwhile, we applied a collision force (forceCollide) to each point to avoid overlaps. As a rule of thumb, the most suitable parameter for forceCollide was the radius value of the data point. In Qutaber, the corresponding parameter value is 3.5.

6 Case study

In this section, we present a case study utilizing real-world data to illustrate the practical application and effectiveness of Qutaber. Our case study involves an experienced data analyst, Tom, who has an extensive background in data analysis, with over five years of experience. He is skillful in multiple data analysis tools (e.g., Plotly, Tableau, and Power BI). To initiate the study, we introduced Tom to Qutaber, focusing on its user interface and functions. We then provided Tom with a dataset on *Happiness*, comprising 158 data items across 9 attributes: *Country*, *Region, Happiness Score*, *Economy*, *Family*, *Health*, *Freedom*, *Trust*, and *Generosity*. Tom was encouraged to freely explore this dataset using Qutaber, without specific guidance or constraints. This method allowed us to observe Tom's natural interaction with Qutaber and gain insights into its usability and effectiveness in real-world analysis scenarios. Throughout the process, Tom's findings and comments were recorded as evaluation material. The following is a detailed description of Tom's complete exploration process.

Early Exploration. Once Tom uploaded the dataset into Qutaber (Fig. 3A), the system provided a statistical overview of the data attributes (Fig. 3B) (**G1**). From the data summary view, Tom found that there are only categorical and quantitative attributes. In addition, He noted a peculiar observation, "*except for the*

[1] https://github.com/nl4dv/nl4dv/blob/master/examples/assets/data/cars-w-year.csv.

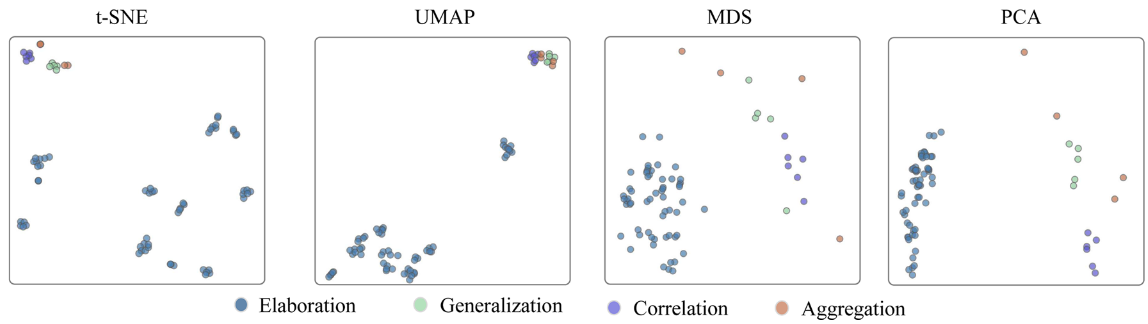

Fig. 5 The comparison of the results of different DR techniques on a certain expansion task of a car dataset

Happiness Score, all quantitative attributes have a minimum value of zero, suggesting potential issues with missing or unclean data." Further examination of attribute snapshots (Fig. 3B), he noticed that the distribution of *Family*, *Health*, *and Freedom* was concentrated in the higher range, while the distribution of *Trust* and *Generosity* tended to fall in the lower range. Upon reviewing the initial recommendation view (Fig. 3C), Tom observed that the absence of a temporal attribute resulted in the exclusion of the line chart in the recommendations (**G1**). Instead, the recommended chart contains a scatter chart presenting the most relevant attributes and two bar charts showcasing the attributes with the most and least variance in their distributions.

Task-based Exploratory Data Analysis. Among the recommended charts, a bar chart depicting the attribute "*Trust*" caught his attention due to its pronounced distributional fluctuations. Tom selected this chart into the exploration area and performed the first round of expansion (Fig. 3C). Accordingly, the system proceeded to generate an expanded set of charts in the backend. These charts were categorized by tasks and displayed in Fig. 3E, along with their corresponding score distributions in Fig. 3G. A noteworthy observation from Tom was the distinctive scoring pattern among these tasks. Charts categorized under *Elaboration* and *Correlation* were predominantly distributed within the lower scoring range, while those under *Generalization* and *Aggregation* exhibited a distribution in the higher scoring range. Furthermore, an overall overview of these charts across task categories is shown in Fig. 3F1 (**G2 and G3**). From this, Tom noticed "*The distribution of the charts exhibits a distinct clustering phenomenon, suggesting that charts with similar characteristics tend to group together*." Delving into a specific cluster, his investigation revealed that charts within this cluster differed only in the filtered interval, while maintaining consistency in other elements (shown in Fig. 3F2).

Tom continued his exploration of the expanded charts, examining them according to task categories and re-ranking them by various metrics (**G3 and G4**). Specifically, within the context of the Elaboration task, Tom noted a diversity in the behavioral patterns of the expanded charts. Contrary to expectations of uniform sharp fluctuations across the board, he observed that several charts displayed surprisingly stable, flat distributions. This deviation prompted a more granular analysis. By applying a significance-based descending order, Tom identified that the bar chart with the added filter (*Region* = *Sub-SaharanAfrica*) demonstrated the most pronounced distributional fluctuation. From these observations, Tom hypothesized, "*Countries in the Sub-SaharanAfrica region could potentially be the primary factor to the observed sharp fluctuations in the Trust attribute's distribution*."

Next, Tom incorporated the chart titled "*Region* = *Sub-SaharanAfrica*" into the exploration area and initiated a second round of expansion centered on it (**G2**). Correspondingly, the system replicated actions akin to the first expansion. In this round of expanded charts, Tom's attention was drawn to the distribution patterns of various attributes. He observed that, alongside *Trust*, the *Economy* attribute also exhibited significant distributional fluctuations. Consequently, he added this chart to the exploration area and started the next round of expansion. In the third round of exploration, Tom's focus shifted to investigating the interplay between the *Economy* attribute and other variables, and he identified the strongest correlation to be between *Economy* and *Family*. From this insight, Tom inferred, "*In Sub-SaharanAfrica, countries with strong Economy exhibit higher family values. However, Family is not solely dependent on the Economy and could be influenced by other factors as well*."

Overall, Tom's engagement with Qutaber demonstrated its capabilities as a potent and efficient tool for data analysis. He found that employing task-based EDA through Qutaber was not only practical but also significantly streamlined his analytical processes. Furthermore, Qutaber's ability to provide enriched context awareness greatly enhances Tom's exploratory activities, giving him a deeper and more nuanced understanding of the data.

7 User study

To further evaluate the effectiveness and usability of the Qutaber system, we conducted a controlled user study. This study was designed with three primary objectives: 1) determine the extent to which the system's semantic embedding of charts aligns with user perceptions; 2) evaluate the efficiency and applicability of task-based EDA; 3) gauge users' attitudes toward the proposed techniques and features of our system.

7.1 Participants, dataset, and baseline system

In our user study, we recruited a total of 12 participants, consisting of 7 males and 5 females, aged 21–29 years (Mean ≈ 24 years). The participants included 4 Ph.D. students (marked as P1 to P4) and 8 Master's students (marked as P5 to P12). All participants were experienced in data analysis and visualization, possessing 1.5–5.5 years (Mean ≈ 3.5 years). The study employed four diverse real-world datasets[2]: *cars* (303 records and 9 fields), *movies* (709 records and 10 fields), *Olympic medals* (8668 records and 10 fields), *colleges* (1215 records and 14 fields). These selected datasets encompass a wide range of domains, exhibit diverse scales of data items, and include varied distributions in terms of field types and quantities. This diversity ensures a thorough and comprehensive evaluation of Qutaber's analytical capabilities. Voyager 2[3] was selected as the baseline system for comparative analysis. The reason for this choice was that Voyager 2 aligns closely with Qutaber in terms of its feature set and analytical approach, notably starting with initial insights and supporting progressive analytical depth. Additionally, both systems offer additional views and insights, enhancing the data exploration process.

7.2 Design, tasks, and procedure

This user study was designed as a within-subjects experiment and tasked participants with an open-ended data exploration using both Qutaber and the baseline system. We chose this type of experimental task to better observe users' interactions and responses to differing system features. Emphasizing on a comprehensive understanding followed by bold exploration, we aimed to foster an environment conducive to insightful feedback.

First, we spent approximately 15 min to train the participants. The training program consisted of a brief tutorial on both systems to familiarize them with the interfaces and functionalities, followed by detailed explanations of various user-related concerns. Next, we asked participants to individually explore the dataset of interest using both Qutaber and the baseline system. During the process, participants were encouraged to think aloud and share their thoughts on system usability, effectiveness, and any challenges they faced. Following the exploration, we conducted a brief questionnaire based on a 7-point Likert Scale to collect subjective feedback on the participants' experiences with both systems.

Second, our focus extended to evaluating specific features within Qutaber, including expansion tasks, chart scoring metrics, re-ranking feature, and chart embedding. To facilitate this, we designed four distinct expansion scenarios, each tailored to a different dataset. These scenarios included a randomly selected chart for expansion analysis, along with its corresponding expanded results. Participants then were tasked with assessing these scenarios on three critical aspects: 1) *how helpful the expansion tasks were in their exploratory analysis*; 2) *whether the chart scoring aligned with their cognition and their satisfaction with the re-ranking feature*; and 3) *whether the chart embedding results align with their cognitive expectations*. For this evaluation, we utilized a 5-point Likert Scale to collect participant feedback. Note that the 7-point Likert Scales were previously employed to obtain a nuanced comparison of the functionalities between Qutaber and the baseline system, whereas now the switch to the 5-point Likert Scale was intended to reduce the cognitive load on participants and make them more willing to make definitive choices.

[2] https://github.com/nl4dv/nl4dv/tree/master/examples/assets/data.

[3] http://vega.github.io/voyager/.

7.3 Study results

Next, we proceeded to analyze and discuss the results collected from our study.

Rating of techniques, Fig. 6 illustrates the ratings provided by participants regarding the key techniques in Qutaber, ranging from 1 (very poor) to 5 (very great). Overall, participants expressed satisfaction with these techniques, as reflected in ratings above 3.5 with a standard deviation within 0.95 for all evaluated aspects. To determine the impact of dataset variability on these techniques, we employed an ANOVA test. Our findings indicate a dataset sensitivity for chart embedding (p = 0.044), whereas expansion tasks (p = 0.31), and chart scoring metrics along with re-ranking (p = 0.29) showed relative robustness against different datasets. To further analyze the differences in chart embedding across different datasets, a t-test was conducted. This test revealed significant differences, particularly when comparing the *college* dataset against others The *college* dataset's larger size and greater complexity likely contribute to these differences. Specifically, the task expansion process in the context of this dataset generates a more extensive and complex array of charts. This increase in number and intricacy could potentially diminish the clarity of cluster edges and increase the risk of overlapping, impacting the overall effectiveness of chart embedding in Qutaber.

Rating of system, Fig. 7 presents the user ratings on various aspects of Qutaber and Voyager 2. A t-test was applied to assess the statistical significance of these ratings. In terms of ease of learning, no significant difference was observed between the two tools (Qutaber: $\mu = 5.83$, $\sigma = 0.88$; Voyager 2: $\mu = 5.42$, $\sigma = 0.81$; $p = 0.279$). Similarly, the ease to use ratings for Qutaber ($\mu = 5.42$, $\sigma = 1.36$) and Voyager 2 ($\mu = 5.67$, $\sigma = 0.97$) did not show a significant difference ($p = 0.576$). However, in the context-aware, Qutaber demonstrated a significant advantage over Voyager 2 (Qutaber: $\mu = 5.42$, $\sigma = 0.99$; Voyager 2: $\mu = 4$, $\sigma = 1.27$; $p = 0.004$), attributed to its enriched and comprehensive approach in presenting expanded charts using small multiples and chart embedding techniques. Moreover, participants expressed a higher satisfaction with the EDA process in Qutaber Moreover, most of the participants were more satisfied with the EDA exploration process in Qutaber compared to Voyager 2 (Qutaber: $\mu = 5.5$, $\sigma = 1.55$; Voyager 2: $\mu = 4.08$, $\sigma = 1.17$; $p = 0.007$), as it provided expansion tasks that were more in line with their analysis needs. Lastly, in terms of system interface and functions, Qutaber ($\mu = 5.17$, $\sigma = 1.24$) showed a marginally significant superiority over Voyager 2 ($\mu = 4.25$, $\sigma = 1.48$), $p = 0.067$.

7.4 Feedback

The entire evaluation process lasts approximately 60 min, followed by concise interviews to gather participants' feedback and detailed insights on the system's specific features. The consensus among participants

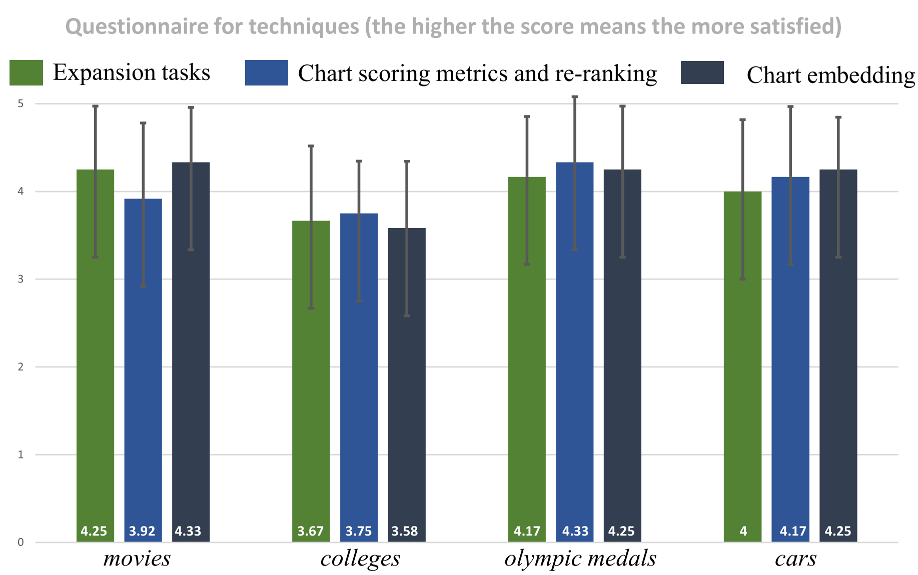

Fig. 6 Results on participants' ratings on expansion tasks, chart scoring metrics and re-ranking, and chart embedding in four specific scenarios

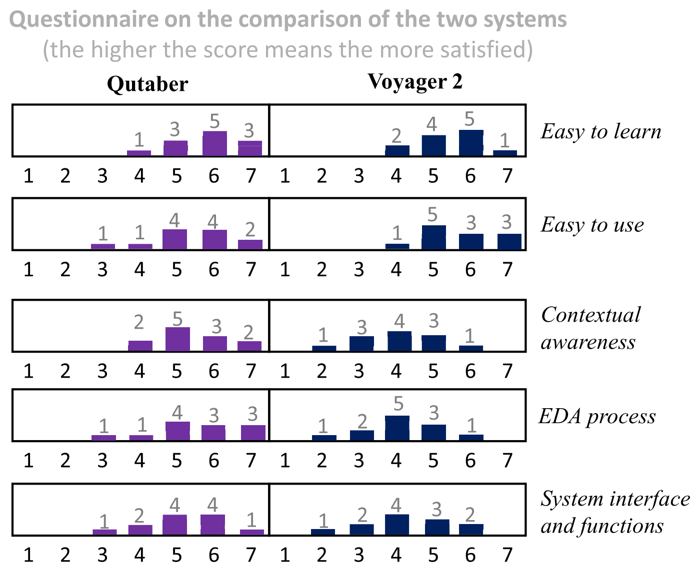

Fig. 7 Participants' ratings of the two systems from various aspects

was that Qutaber significantly enhanced their exploratory capabilities. P2 commented, *"Qutaber automates numerous analysis operations that traditionally require manual execution, thereby enabling users to devote more effort to scrutinizing results and uncovering valuable insights.*" Furthermore, P9 commended the expansion tasks for their thoughtful design and practicality, which facilitated a deeper exploration of insights related to the focused chart. Regarding the chart embedding, a majority of participants reported the ease of identifying distribution characteristics in expanded charts through this feature. Despite the positive feedback, participants identified areas for improvement. More than two-thirds of participants suggested that Qutaber could benefit from more comprehensive initialization recommendations. Furthermore, P7 suggested, "*Currently the chart embedding only uses information from Vega-Lite, incorporating additional chart features (e.g., statistical features, text descriptions) would capture more accurate chart semantics*."

8 Limitations and future work

In our case study and user evaluations, Qutaber has proven effective in aiding data exploration and pattern discovery. Despite its strengths, like many empirical studies, it has its limitations, which we have identified through reviewing our system implementation and user feedback.

Visual presentation. The current Qutaber implementation uses small multiples for a comprehensive overview. However, with larger datasets, these small multiples can overwhelm users. A potential improvement is the introduction of hierarchical small multiples, offering varying detail levels tailored to user needs. Additionally, the system's direct display of expanded charts complicates smooth transitions between charts. P5 recommended animated transitions for each expanded chart, explaining that *"Animated transitions would provide users with visual cues to better understand the connections between the current chart and the target chart*." Moreover, the current charts lack textual descriptions, which may further increase the cognitive load for users. Future enhancements could involve employing large language models, like ChatGPT (Brown et al. 2020; Radford et al. 2019), for the automated generation of textual descriptions to elucidate key patterns. P4 highlighted the benefits of this integration, *"Incorporating natural language generation can enrich user experience and offer an accessible approach to understanding complex data patterns"*.

Exploration guidance. The current system also has several limitations in supporting user exploration. First, Qutaber's initial recommendations are confined to a narrow range of chart types. For this, P1 suggested, "*Presenting richer initial insights through data mining and machine learning techniques not only enhances the user's understanding of the data but also drives them faster into the in-depth exploration*." Second, the system's task-based expansion process limits users to expanding only one chart at a time, preventing simultaneous exploration of multiple charts. Moreover, when a new expansion is performed, prior expansion results are not retained. Addressing this, P8 suggested enhancements, *"Qutaber should enable simultaneous exploration of multiple charts and incorporate a history interface for easy review of past expansions*."

User Interaction. The current iteration of Qutaber predominantly utilizes traditional keyboard and mouse inputs, requiring users to interact through components like the parameter editing panel for chart manipulation. This setup potentially steepens the learning curve, particularly for less experienced users. A potential improvement is to introduce natural language interfaces that enable more intuitive and efficient user-system interactions. P3 highlighted the advantages of this approach, *"Natural language interfaces can enhance system accessibility and user-friendliness, thereby lowering the barriers to entry for non-expert users*." Correspondingly, we plan to integrate natural language interfaces across the user exploration and analysis journey. This would empower users to perform tasks like data manipulation (e.g., selecting dimensions, and filtering data) and adjusting visual styles through the use of natural language commands.

9 Conclusion

In this work, we introduce Qutaber, a runnable system designed for the analysis of tabular data. We first identified six critical expansion analysis tasks specific to EDA scenarios through a preliminary study that included expert interviews and a review of relevant literature. Leveraging these defined tasks, Qutaber empowers users to engage in an expanded exploration from the chart of current interest. To augment contextual awareness during the analysis process, we used small multiples integrated with a multi-metric re-ranking feature to present an array of recommended charts more comprehensively. Furthermore, we adopted a machine learning methodology to characterize the semantic vectors of these charts. The outcomes are then visualized on a 2D scatterplot to help users get a holistic understanding of the findings from their expanded exploration. To validate the effectiveness of Qutaber, we conducted a case study and a user study utilizing real-world datasets. The results from these studies highlight Qutaber's efficacy in aiding users to explore and discover hidden data patterns. Finally, we discussed the limitations of the current system and proposed potential avenues for further enhancement, which would be implemented in future work.

Acknowledgements The work is partly supported by the National Key Research and Development Program of China (2020YFB1707700), Zhejiang Provincial Natural Science Foundation of China (LR23F020003 and LTGG23F020005), Fundamental Research Funds for the Provincial Universities of Zhejiang (RF-B2023006), and National Natural Science Foundation of China (62372411, 62036009). The authors thank all other participants for their joining the discussion, suggestion, and evaluation.

References

Abdi H, Williams LJ (2010) Principal component analysis. Wiley Interdiscip Revi: Comput Stat 2(4):433–459

Amar R, Eagan J, Stasko J (2005) Low-level components of analytic activity in information visualization. In: IEEE symposium on information visualization, 2005. INFOVIS 2005. IEEE, pp 111–117

Battle L, Heer J (2019) Characterizing exploratory visual analysis: a literature review and evaluation of analytic provenance in tableau. Comput Graph Forum 38(3):145–159

Borg I, Groenen P (2006) Modern multidimensional scaling: theory and applications. J Educ Meas 40:277–280

Brehmer M, Lee B, Isenberg P, Choe EK (2019) A comparative evaluation of animation and small multiples for trend visualization on mobile phones. IEEE Trans Vis Comput Graphics 26(1):364–374

Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A et al (2020) Language models are few-shot learners. Adv Neural Inf Process Syst 33:1877–1901

Cao Y-R, Li X-H, Pan J-Y, Lin W-C (2022) Visguide: user-oriented recommendations for data event extraction. In: Proceedings of the 2022 CHI conference on human factors in computing systems, CHI '22, pp 1–13

Cer D, Yang Y, Kong S-y, Hua N, Limtiaco N, John RS, Constant N, Guajardo-Cespedes M, Yuan S, Tar C, et al (2018) Universal sentence encoder. arXiv:1803.11175

Chen Q, Sun F, Xu X, Chen Z, Wang J, Cao N (2021) VizLinter: a linter and fixer framework for data visualization. IEEE Trans Vis Comput Graphics 28(1):206–216

Cui Z, Badam SK, Yalçin MA, Elmqvist N (2019) Datasite: proactive visual data exploration with computation of insight-based recommendations. Inf Vis 18(2):251–267

Demiralp c, Haas PJ, Parthasarathy S, Pedapati T (2017) Foresight: recommending visual insights. In: Proceedings of the VLDB Endowment, vol 10, pp 1937–1940

Deng D, Wu A, Qu H, Wu Y (2022) Dashbot: insight-driven dashboard generation based on deep reinforcement learning. IEEE Trans Vis Comput Graphics 29(1):690–700

Devlin J, Chang M-W, Lee K, Toutanova K (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805

Dibia V, Demiralp Ç (2019) Data2vis: automatic generation of data visualizations using sequence-to-sequence recurrent neural networks. IEEE Comput Graphics Appl 39(5):33–46
Ding R, Han S, Xu Y, Zhang H, Zhang D (2019) Quickinsights: quick and automatic discovery of insights from multi-dimensional data. In: Proceedings of the 2019 international conference on management of data, pp 317–332
Ghosh A, Nashaat M, Miller J, Quader S, Marston C (2018) A comprehensive review of tools for exploratory analysis of tabular industrial datasets. Vis Inform 2(4):235–253
Hu K, Bakker MA, Li S, Kraska T, Hidalgo C (2019) VizML: a machine learning approach to visualization recommendation. In: Proceedings of the 2019 CHI conference on human factors in computing systems, CHI '19
Hu K, Orghian D, Hidalgo C (2018) DIVE: a mixed-initiative system supporting integrated data exploration workflows. In: Proceedings of the workshop on human-in-the-loop data analytics, pp 1–7
Jiang Q, Sun G, Dong Y, Liang R (2021) DT2VIS: a focus+ context answer generation system to facilitate visual exploration of tabular data. IEEE Comput Graphics Appl 41(5):45–56
Joglekar M, Garcia-Molina H, Parameswaran A (2015) Smart drill-down: a new data exploration operator. In: Proceedings of the VLDB endowment, vol 8, no 12, pp 1928–1931
Kim Y, Heer J (2018) Assessing effects of task and data distribution on the effectiveness of visual encodings. Comput Graphics Forum 37(3):157–167
Law P-M, Basole RC, Wu Y (2018) Duet: helping data analysis novices conduct pairwise comparisons by minimal specification. IEEE Trans Vis Comput Graphics 25(1):427–437
Lee DJ-L, Dev H, Hu H, Elmeleegy H, Parameswaran A (2019) Avoiding drill-down fallacies with vispilot: assisted exploration of data subsets. In: Proceedings of the 24th international conference on intelligent user interfaces, IUI '19, New York, NY, USA, pp 186–196
Lee DJ-L, Setlur V, Tory M, Karahalios K, Parameswaran A (2021) Deconstructing categorization in visualization recommendation: a taxonomy and comparative study. IEEE Trans Vis Comput Graphics 28(12):4225–4239
Lekschas F, Zhou X, Chen W, Gehlenborg N, Bach B, Pfister H (2020) A generic framework and library for exploration of small multiples through interactive piling. IEEE Trans Vis Comput Graphics 27(2):358–368
Li Q, Lin H, Tang CF, Wei X, Peng Z, Ma X, Chen T (2021) Exploring the "double-edged sword" effect of auto-insight recommendation in exploratory data analysis. In: IUI workshops
Li X, Zhang Y, Leung J, Sun C, Zhao J (2023) EDAssistant: supporting exploratory data analysis in computational notebooks with in situ code search and recommendation. ACM Trans Interact Intell Syst 13(1):1–27
Luo Y, Qin X, Tang N, Li G (2018) DeepEye: towards automatic data visualization. In: 2018 IEEE 34th international conference on data engineering (ICDE). IEEE, pp 101–112
Mackinlay J, Hanrahan P, Stolte C (2007) Show me: automatic presentation for visual analysis. IEEE Trans Vis Comput Graphics 13(6):1137–1144
McInnes L, Healy J, Melville J (2018) UMAP: uniform manifold approximation and projection for dimension reduction. arXiv:1802.03426
Moritz D, Wang C, Nelson GL, Lin H, Smith AM, Howe B, Heer J (2018) Formalizing visualization design knowledge as constraints: actionable and extensible models in draco. IEEE Trans Vis Comput Graphics 25(1):438–448
Pandey A, Srinivasan A, Setlur V (2023) MEDLEY: intent-based recommendations to support dashboard composition. IEEE Trans Vis Comput Graphics 29(01):1135–1145
Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I et al (2019) Language models are unsupervised multitask learners. OpenAI blog 1(8):9
Roy D, Dutta M (2022) A systematic review and research perspective on recommender systems. J Big Data 9(1):59
Saket B, Endert A, Demiralp Ç (2018) Task-based effectiveness of basic visualizations. IEEE Trans Vis Comput Graphics 25(7):2505–2512
Shen L, Shen E, Tai Z, Song Y, Wang J (2021) TaskVis: task-oriented visualization recommendation. In: Proceedings of EuroVis, vol 21, pp 1–5
Shen L, Shen E, Tai Z, Xu Y, Dong J, Wang J (2022) Visual data analysis with task-based recommendations. Data Sci Eng 7(4):354–369
Shi D, Xu X, Sun F, Shi Y, Cao N (2020) Calliope: automatic visual data story generation from a spreadsheet. IEEE Trans Vis Comput Graphics 27(2):453–463
Srinivasan A, Drucker SM, Endert A, Stasko J (2018) Augmenting visualizations with interactive data facts to facilitate interpretation and communication. IEEE Trans Vis Comput Graphics 25(1):672–681
Sun G, Zhu S, Jiang Q, Xia W, Liang R (2021) EvoSets: tracking the sensitivity of dimensionality reduction results across subspaces. IEEE Trans Big Data 8(6):1566–1579
Sun G-D, Wu Y-C, Liang R-H, Liu S-X (2013) A survey of visual analytics techniques and applications: state-of-the-art research and future challenges. J Comput Sci Technol 28:852–867
Tang B, Han S, Yiu ML, Ding R, Zhang D (2017) Extracting top-k insights from multi-dimensional data. In: Proceedings of the 2017 ACM international conference on management of data, pp 1509–1524
van der Maaten L, Hinton G (2008) Viualizing data using t-SNE. J Mach Learn Res 9:2579–2605
Wang AY, Epperson W, DeLine RA, Drucker SM (2022) Diff in the loop: supporting data comparison in exploratory data analysis. In: Proceedings of the 2022 CHI conference on human factors in computing systems, CHI '22, New York, NY, USA
Wehrend S, Lewis C (1990) A problem-oriented classification of visualization techniques. In: Proceedings of the first IEEE conference on visualization. IEEE, pp 139–143
Wongsuphasawat K, Liu Y, Heer J (2019) Goals, process, and challenges of exploratory data analysis: an interview study. arXiv:1911.00568
Wongsuphasawat K, Moritz D, Anand A, Mackinlay J, Howe B, Heer J (2015) Voyager: exploratory analysis via faceted browsing of visualization recommendations. IEEE Trans Vis Comput Graphics 22(1):649–658

Wongsuphasawat K, Qu Z, Moritz D, Chang R, Ouk F, Anand A, Mackinlay J, Howe B, Heer J (2017) Voyager 2: augmenting visual analysis with partial view specifications. In: Proceedings of the 2017 CHI conference on human factors in computing systems, CHI '17, New York, NY, USA, pp 2648–2659

Xu S, Bryan C, Li JK, Zhao J, Ma K-L (2018) Chart constellations: effective chart summarization for collaborative and multi-user analyses. Comput Graphics Forum 37(3):75–86

Yu B, Silva CT (2019) FlowSense: a natural language interface for visual data exploration within a dataflow system. IEEE Trans Vis Comput Graphics 26(1):1–11

Zhao J, Fan M, Feng M (2020) ChartSeer: interactive steering exploratory visual analysis with machine intelligence. IEEE Trans Vis Comput Graphics 28(3):1500–1513

Zhao Y, Ge L, Xie H, Bai G, Zhang Z, Wei Q, Lin Y, Liu Y, Zhou F (2022) ASTF: visual abstractions of time-varying patterns in radio signals. IEEE Trans Vis Comput Graphics 29(1):214–224

Zhou Y, Meng X, Wu Y, Tang T, Wang Y, Wu Y (2023) An intelligent approach to automatically discovering visual insights. J Vis 26(3):705–722

Zhu S, Sun G, Jiang Q, Zha M, Liang R (2020) A survey on automatic infographics and visualization recommendations. Vis Inform 4(3):24–40

Zhu Z, Shen Y, Zhu S, Zhang G, Liang R, Sun G (2023) Towards better pattern enhancement in temporal evolving set visualization. J Vis 26(3):611–629