Question-guided Insights Generation for Automated Exploratory Data Analysis

Abhijit Manatkar, Ashlesha Akella, Krishnasuri Narayanam, Sameep Mehta

IBM Research India
{abhijitmanatkar, ashlesha.akella}@ibm.com, {knaraya3, sameepmehta}@in.ibm.com

Abstract

Exploratory Data Analysis (EDA) derives meaningful insights from extensive and complex datasets. This process typically involves a series of analytical operations to identify the patterns within the data. However, the effectiveness of EDA is often limited by the user's domain knowledge and proficiency in data exploration methods. To overcome these challenges, we developed QUIS, a fully automated EDA system that uncovers insights by generating data-related questions and exploring subspaces in the dataset without prior training. QUIS allows users to control key system parameters such as beam width, beam depth, and expansion factor for subspace selection, the interestingness score for filtering valuable insights, and parameters for managing the quality and quantity of generated questions.

Introduction

To fully automate insight extraction from a dataset, we introduce QUIS (Figure 1), an exploratory data analysis (EDA) system, that discovers insights driven by questions generated in accordance with the dataset. QUIS is composed of two modules: QUGEN, to generate semantically relevant questions, which are then used by ISGEN to explore the dataset to uncover statistically significant insights.

Datasets can contain many insights, and multiple exploration methods can extract them. EDA systems aim to ensure that insights are both dominant and found efficiently. QUIS addresses this by assigning an interestingness score to prune insignificant insights and by leveraging data statistics and table semantics to improve exploration efficiency.

The existing EDA literature encompasses a range of approaches, including statistics-based methods (Sellam, Müller, and Kersten 2015; Wang et al. 2020; Ma et al. 2021) and interactive techniques (Milo and Somech 2016, 2018; Agarwal et al. 2023; He et al. 2024), allowing users engage with data through natural language queries or receive suggestions for further analysis. Visualization-based approaches (Vartak et al. 2015; Demiralp et al. 2017; Srinivasan et al. 2018; Wu et al. 2024) offer visual insights and enable additional explorations. However, these methods can become resource-intensive due to extensive user interactions. In contrast, goal-oriented Auto EDA approaches generate insights based on predefined objectives (Tang et al. 2017; Seleznova et al. 2020; Omidvar-Tehrani, Personnaz, and Amer-Yahia 2022; Laradji et al. 2023), which may limit the insights to those aligned with the predetermined goals. EDA using reinforcement learning is studied (Bar El, Milo, and Somech 2020; Personnaz et al. 2021; Garg et al. 2023) to achieve full automation. While these systems minimize user involvement, their demand for dataset-specific training and substantial compute makes the process challenging.

The full technical details of our system demonstration are described in (Manatkar et al. 2024).

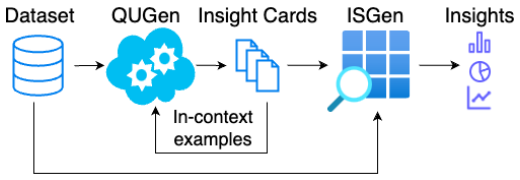

Figure 1: Overview of QUIS workflow

Preliminaries

We define the notion of an insight consistent with previous studies (Ding et al. 2019; Ma et al. 2023). Let $D = \{X_1, X_2, \ldots X_n\}$ be a tabular dataset where each X_i is an attribute (column) of the dataset. An insight, denoted by $Insight(B, M, S, P)$, consists of the following:

1. B represents the *breakdown* column. And M is the *measure*, referring to a quantity of interest from the table. Typically, M is of the form $agg(C)$ where agg is an aggregation function (like `sum()` or `mean()`) and C is the measure column. For each (B, M) pair, an EDA system computes a view $view(D, B, M)$ by grouping the rows of dataset D on B and calculating the measure M for each group. E.g., computing $view(\texttt{CarSales}, \texttt{Year}, \texttt{sum(Sales)})$ on the `CarSales` (Bhatia 2017) dataset is equivalent to applying the SQL query: `SELECT Year, SUM(Sales) FROM Car Sales GROUP BY Year`
2. A subspace $S = \bigcup_i \{(X_i, y) |\ y$ is a value in the domain of column $X_i\}$ is a set of filters that determines a subset D_S consisting of only those rows of D where $D[X_i] = y$

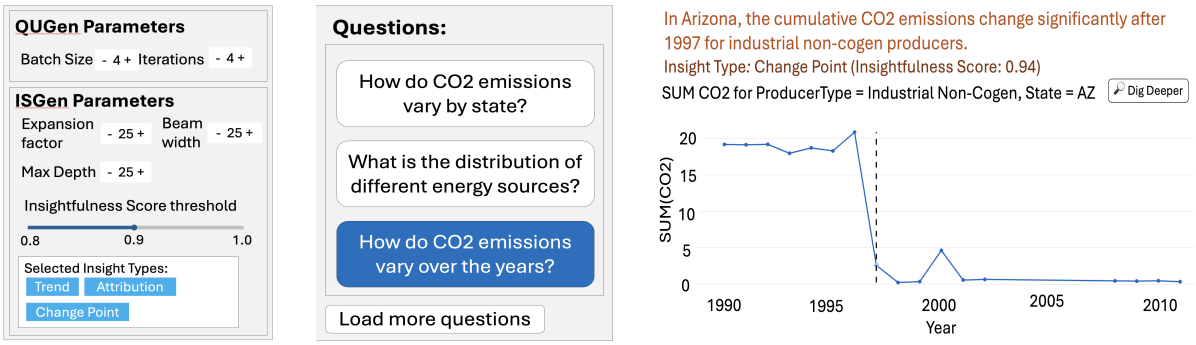

Figure 2: Configuration parameters, a sample of generated questions and an insight

for each $(X_i, y) \in S$. E.g., if `Category` is a filter column, then $\{(\texttt{Category}, \texttt{"Compact"})\}$ is subspace.

3. P represents a type of insight pattern present in D (like trend or outlier). E.g., an insight with pattern *increasing trend* in the sales over the years may be present in the subspace $S = \{(\texttt{Category}, \texttt{"SUV"})\}$

Working of QUIS

QUIS framework begins with QUGEN generating *insight cards* in batches by prompting the LLM `Llama-3-70b-instruct` (AI@Meta 2024). The insight card components *(reason, question, breakdown B, measure M)* are generated in a specific order, with (B, M) pair depending on the reason and question, inspired by the chain-of-thought prompting (Wei et al. 2022). Each new batch of insight cards is iteratively refined by using a subset of previously generated cards and the current batch as in-context examples. Finally, the questions from insight cards will be provided for user's selection. The ISGEN module semantically analyzes the data statistics to extract relevant insights corresponding to each (B, M) pair associated with the user-selected questions.

Data analysis for insights involves checking for each $view(D_S, B, M)$ to find if an insight pattern P can be observed. Since S can be any set that is an element of the power set of D, exhaustively searching all the $2^{|D|}$ views is inefficient. Hence, QUIS adopts a *beam-search* heuristic strategy (Russell and Norvig 2010) to limit the search to subspaces that are semantically meaningful and statistically relevant. The system maintains a beam of subspaces with a specified *Beam Width*. At each iteration, up to a maximum search depth, each subspace in the beam is expanded into a set of k local subspaces by selecting k additional filter columns. The column selection is done by sampling from the unused columns, with the sampling weighted by the semantic relevance of the columns. Each of the expanded subspaces is then scored based on how well it displays an interesting pattern. Finally, the current beam is updated by selecting the highest-scoring subspaces from the set of expanded subspaces, and discarding the rest to maintain the beam width.

To offer an initial data understanding, QUIS explores subspaces upto a maximum depth of 1 (i.e. $|S| \leq 1$) for insight extraction and presents all the significant insights. QUIS enables iterative data exploration by providing a *dig-deeper* feature that allows the user to select subspaces (with $|S| \geq 2$) of interest for in-depth analysis.

Insight types: QUIS supports six types (Chen, Yang, and Ribarsky 2009) of insight patterns: *Trend* - Values showing increasing or decreasing trend. *Outstanding Value* - The largest (or smallest) value in the set is significantly larger (or smaller) than the others. *Attribution* - The highest value accounts for a large proportion ($\geq 50\%$) of the total of all values in the set. *Distribution Difference* - The distribution of values in a set changes notably from one subspace to another. *Change Point* - The mean value/slope of a time series changes significantly in the preceding and succeeding regions of the change point. *Outlier* - A point in a time series shows significant deviation from the ongoing trend.

Knobs for Customized User Experience: QUIS offers below knobs for customized insight extraction. QUGEN *batch size:* chunk size to generate questions QUGEN *iterations:* number of iterations within a batch. *Beam-width:* size of beam maintained during the search. *Expansion factor:* the branching factor that determines the number of expansions on each node in the current beam. *Max search depth:* depth of subspace search for exploration before termination (the maximum number of filter columns). *Interestingness score threshold:* limit to discard insignificant insights.

Dig-Deeper to Explore Subspaces of Interest

Once QUIS presents insights from various subspaces, users can explore further using the "dig-deeper" option. This initiates a beam search within the selected subspace for deeper discovery of insights in iterations. In each iteration, the dataset is filtered by different columns and values, until no more filter columns or interesting insights remain.

Insight type heterogeneity prioritization: As the (B, M) pair remains unchanged during the deeper subspace search, any dominant insight pattern found in a parent subspace is likely to be found in its local subspaces also. E.g., if an insight reveals an increasing trend in Toyota car sales, the deeper search might show the same trend for specific models like Toyota Corolla and Toyota Camry, leading to redundant insights that reduce the overall quality. Hence, QUIS prioritizes identifying insights with diverse types along the subspace exploration path, to enhance quality of the insights.

References

Agarwal, S.; Chan, G. Y.-Y.; Garg, S.; Yu, T.; and Mitra, S. 2023. Fast Natural Language Based Data Exploration with Samples. In *Companion of the 2023 International Conference on Management of Data (SIGMOD)*, 155–158.

AI@Meta. 2024. Llama 3 Model Card.

Bar El, O.; Milo, T.; and Somech, A. 2020. Automatically Generating Data Exploration Sessions Using Deep Reinforcement Learning. In *Proceedings of the 2020 International Conference on Management of Data (SIGMOD)*, 1527–1537.

Bhatia, G. 2017. Car sales. https://www.kaggle.com/datasets/gagandeep16/car-sales. Accessed: 2024-09-22.

Chen, Y.; Yang, J.; and Ribarsky, W. 2009. Toward effective insight management in visual analytics systems. In *2009 IEEE Pacific Visualization Symposium (PacificVis)*, 49–56.

Demiralp, c.; Haas, P. J.; Parthasarathy, S.; and Pedapati, T. 2017. Foresight: recommending visual insights. *Proceedings of the VLDB Endowment (PVLDB)*, 10(12): 1937–1940.

Ding, R.; Han, S.; Xu, Y.; Zhang, H.; and 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 (SIGMOD)*, 317–332.

Garg, S.; Mitra, S.; Yu, T.; Gadhia, Y.; and Kashettiwar, A. 2023. Reinforced Approximate Exploratory Data Analysis. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence (AAAI)*, volume 37, 7660–7669.

He, X.; Zhou, M.; Xu, X.; Ma, X.; Ding, R.; Du, L.; Gao, Y.; Jia, R.; Chen, X.; Han, S.; Yuan, Z.; and Zhang, D. 2024. Text2Analysis: A Benchmark of Table Question Answering with Advanced Data Analysis and Unclear Queries. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, 18206–18215.

Laradji, I. H.; Taslakian, P.; Mudumba, S. R.; Zantedeschi, V.; Lacoste, A.; Chapados, N.; Vazquez, D.; Pal, C.; and Drouin, A. 2023. Capture the Flag: Uncovering Data Insights with Large Language Models. In *Workshop at the Neural Information Processing Systems (NeurIPS)*.

Ma, P.; Ding, R.; Han, S.; and Zhang, D. 2021. MetaInsight: Automatic Discovery of Structured Knowledge for Exploratory Data Analysis. In *Proceedings of the 2021 International Conference on Management of Data (SIGMOD)*, 1262–1274.

Ma, P.; Ding, R.; Wang, S.; Han, S.; and Zhang, D. 2023. InsightPilot: An LLM-Empowered Automated Data Exploration System. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations*, 346–352.

Manatkar, A.; Akella, A.; Gupta, P.; and Narayanam, K. 2024. QUIS: Question-guided Insights Generation for Automated Exploratory Data Analysis. *arXiv preprint arXiv:2410.tbd*.

Milo, T.; and Somech, A. 2016. REACT: Context-Sensitive Recommendations for Data Analysis. In *Proceedings of the 2020 International Conference on Management of Data (SIGMOD)*, 2137–2140.

Milo, T.; and Somech, A. 2018. Next-Step Suggestions for Modern Interactive Data Analysis Platforms. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, 576–585.

Omidvar-Tehrani, B.; Personnaz, A.; and Amer-Yahia, S. 2022. Guided text-based item exploration. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management (CIKM)*, 3410–3420.

Personnaz, A.; Amer-Yahia, S.; Berti-Équille, L.; Fabricius, M.; and Subramanian, S. 2021. DORA THE EXPLORER: Exploring Very Large Data With Interactive Deep Reinforcement Learning. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management (CIKM)*, 4769–4773.

Russell, S. J.; and Norvig, P. 2010. *Artificial intelligence: a modern approach*. Pearson.

Seleznova, M.; Omidvar-Tehrani, B.; Amer-Yahia, S.; and Simon, E. 2020. Guided exploration of user groups. *Proceedings of the VLDB Endowment (PVLDB)*, 13(9): 1469–1482.

Sellam, T.; Müller, E.; and Kersten, M. 2015. Semi-Automated Exploration of Data Warehouses. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (CIKM)*, 1321–1330.

Srinivasan, A.; Drucker, S. M.; Endert, A.; and Stasko, J. 2018. Augmenting Visualizations with Interactive Data Facts to Facilitate Interpretation and Communication. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 25(1): 672–681.

Tang, B.; Han, S.; Yiu, M. L.; Ding, R.; and Zhang, D. 2017. Extracting Top-K Insights from Multi-dimensional Data. In *Proceedings of the 2017 ACM International Conference on Management of Data (SIGMOD)*, 1509–1524.

Vartak, M.; Rahman, S.; Madden, S.; Parameswaran, A.; and Polyzotis, N. 2015. SeeDB: Efficient Data-Driven Visualization Recommendations to Support Visual Analytics. *Proceedings of the VLDB Endowment (PVLDB)*, 8(13): 2182–2193.

Wang, Y.; Sun, Z.; Zhang, H.; Cui, W.; Xu, K.; Ma, X.; and Zhang, D. 2020. DataShot: Automatic Generation of Fact Sheets from Tabular Data. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 26: 895–905.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems (NeurIPS)*, 35: 24824–24837.

Wu, Y.; Wan, Y.; Zhang, H.; Sui, Y.; Wei, W.; Zhao, W.; Xu, G.; and Jin, H. 2024. Automated Data Visualization from Natural Language via Large Language Models: An Exploratory Study. *Proceedings of the ACM on Management of Data (PACMMOD)*, 2(3): 115:1–115:28.