QUIS: Question-guided Insights Generation for Automated Exploratory Data Analysis

IBM Research, India

Abhijit Manatkar
abhijitmanatkar@ibm.com

Ashlesha Akella
ashlesha.akella@ibm.com

Parthivi Gupta
parthivig@kgpian.iitkgp.ac.in

Krishnasuri Narayanam
knaraya3@in.ibm.com

2024 Conference on Empirical Methods in Natural Language Processing Industry Track (EMNLP- 2024)

IBM

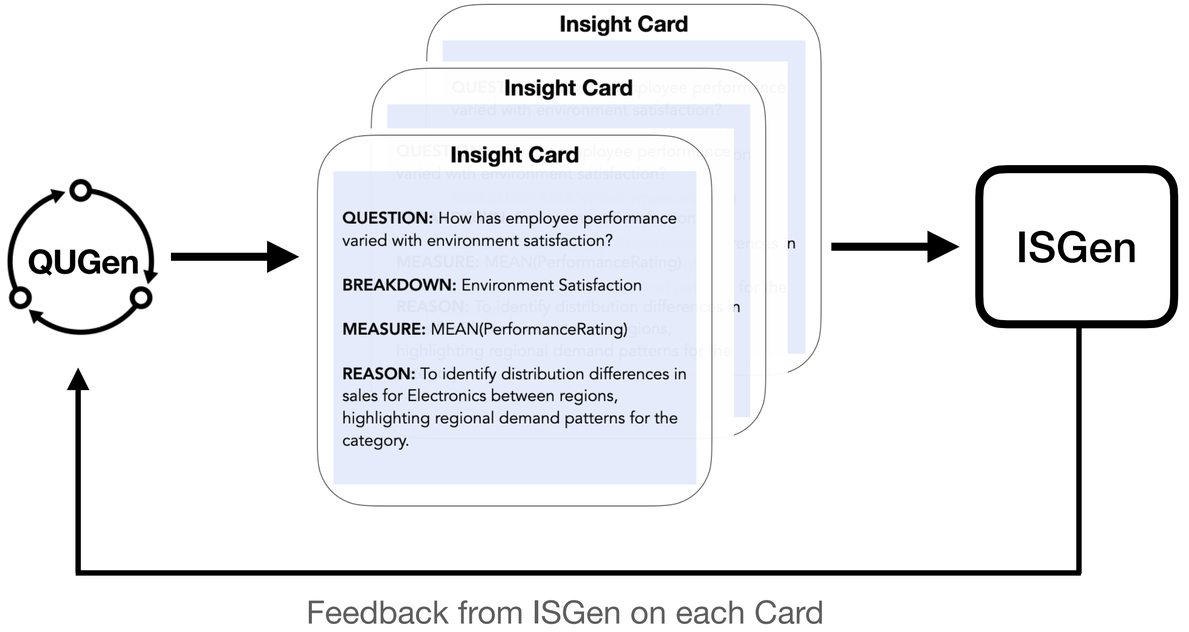
Introduction: Challenges in Automated Data Exploration (ADE)

- **Exploratory Data Analysis (EDA)** is the process of discovering meaningful insights from large datasets.
- **Automated Data Exploration (ADE)** accelerates this process using automation, reducing manual effort.
- **Challenges in existing ADE systems** include the
  - ❖Need for dataset-specific training and the high computational cost of some approaches.
  - ❖Predefined goal-oriented approaches may limit insights to anticipated findings.
  - ❖Using reinforcement learning need dataset-specific training and careful reward shaping.
  - ❖Include statistics-based, interactive, and visualization-driven approaches, but demand significant human resources.
- **Humans typically analyze data** by asking contextually relevant questions based on semantics, guiding the exploration process.

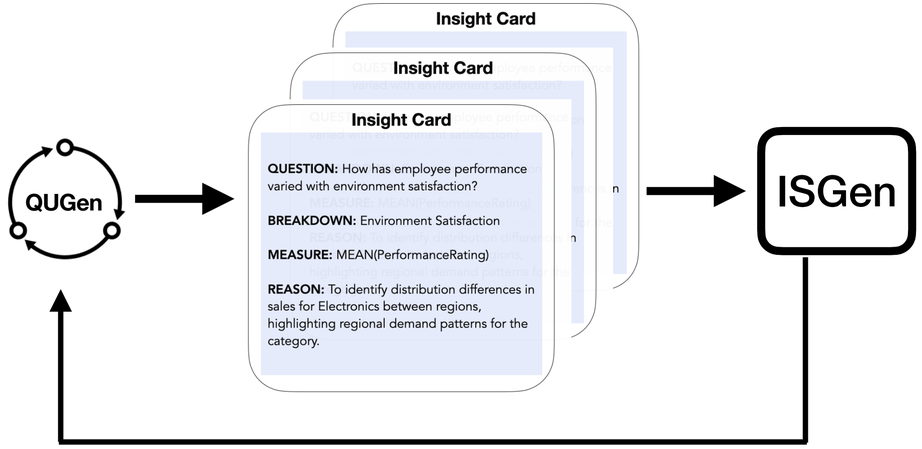
IBM

Introduction: Question-guided Insights Generation for Auto EDA

- **Two-Stage ADE System**: **QUIS** Automates EDA by generating questions from data semantics, then producing insights through statistical analysis.
- **Question Generation Module (QUGen)**: questions are iteratively generated from dataset details (schema).
- **Insight Generation Module (ISGen)**: uses statistical techniques to analyze the data and generate multiple insights for each question.
- **QUIS** eliminates the need for predefined analysis goals
- The system minimizes the need for expert input, accelerates data exploration, and uncovers a broader range of insights.



Question Generation: QUGen



- QUGen generates questions using dataset semantics.
- For each question, QUGen also generates a Reason (R), Breakdown (B), and Measure (M), collectively forming what we refer to as an Insight Card.
- Chain-of-thought responses ensures Breakdown and Measure components are conditioned on the reason and question.

Insight Card

REASON: To analyse whether there are any trends in the average performance of employees over time.

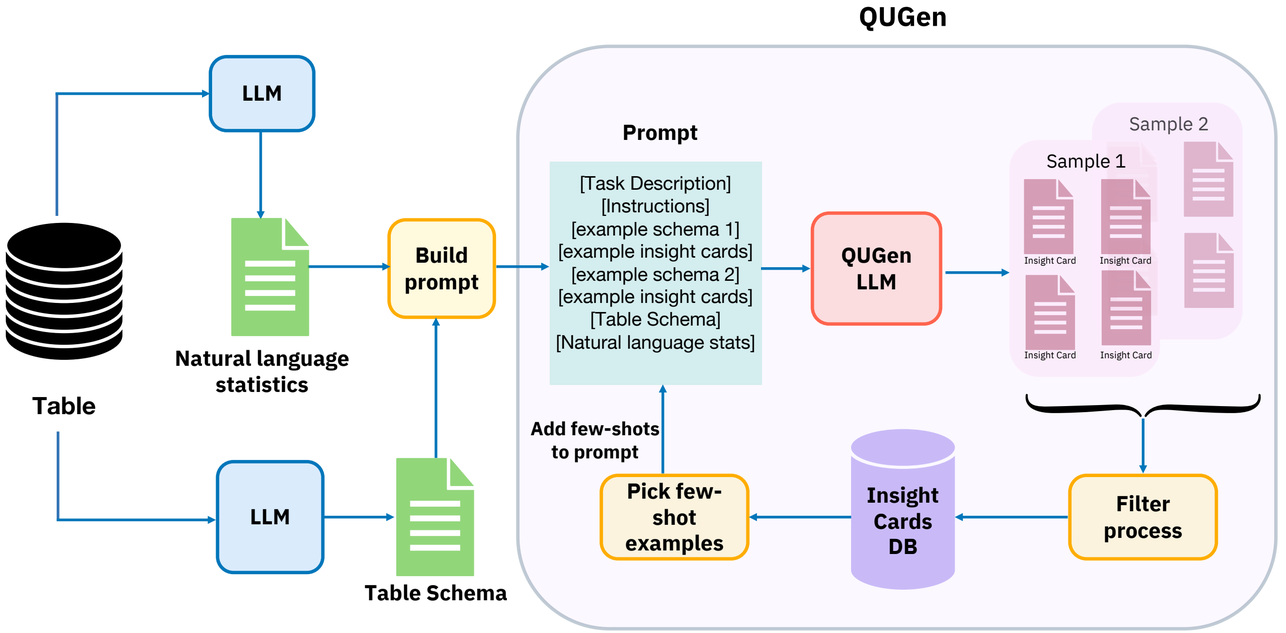
QUESTION: How has employee performance varied over the years?

BREAKDOWN: MEAN(Performance)

MEASURE: Year

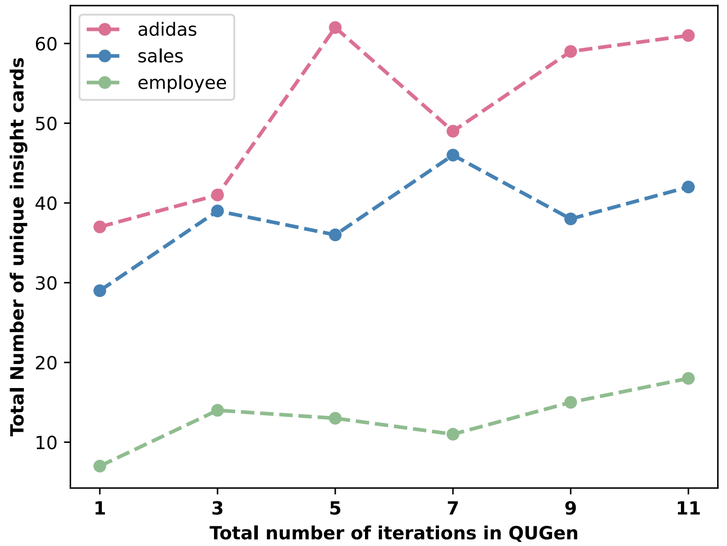
Question Generation: QUGen

- **Insight Card Generation**: The QUGen module prompts the LLM to generate multiple Insight Cards with iterative sampling (***s*** times) and a temperature **t**, resulting in a diverse set of questions for exploration.
- **Filtering**: Insight Cards are filtered for relevance by comparing them to the table schema using semantic similarity, removing irrelevant questions.
- **Duplicate and rudimentary question Filtering**: Duplicate Insight Cards are identified and removed, and simple questions are discarded if their corresponding SQL queries return only one row, ensuring only in-depth questions remain.
- **Iterative Contextual Learning:** Each iteration reuses select Insight Cards from previous iterations as examples.



The Question Generation **QUGen** module of QUIS system generates questions refined over iterations using data semantics

Analysis of Insight Card Diversity in QUIS Across Iterations



Total number of unique insight cards generated by QUIS under non-iterative (1 iteration) and iterative (up to 11 iterations)

Insight Generation: ISGen Terminology

Insight is a four tuple: <B, M, S, P>

- **Perspective**: The perspective is defined by a breakdown attribute (**B**) and a measure (**M**) to identify how specific values of a measure vary across groups defined by the breakdown.
- **Subspace Filtering**: The subspace (**S**), represented by a set of filters, narrows down the dataset to a specific subset, allowing the insight to focus on particular segments of the data.
- **Pattern**: The pattern **P** represents the type of insight observed. QUIS system incorporates the following insight types as candidates
    - ❖ Trend, Outstanding Value, Attribution, Distribution Difference

Employee-Attrition

Age	Attrition	DailyRate	Education	Department	EducationField	YearsAtCompany	Gender	Performance Rating
41	Yes	1102	2	Sales	Life Sciences	6	Female	3
49	No	279	1	Research	Life Sciences	10	Male	4
37	Yes	1373	2	Research	Other	0	Male	3
33	No	1392	4	Sales	Life Sciences	8	Female	3
27	No	591	1	Research	Medical	2	Male	3
32	No	1005	2	Research	Life Sciences	7	Male	3
59	No	1324	3	Sales	Medical	1	Female	4
30	No	1358	1	Sales	Life Sciences	1	Male	4
38	No	216	3	Research	Life Sciences	9	Male	4

Insight Card

REASON: To analyse whether there are any trends in the average performance of employees over time.

QUESTION: How has employee performance varied over the years?

BREAKDOWN: MEAN(Performance)

MEASURE: Year

$$B = \texttt{Year}, M = \texttt{mean}(\texttt{Performance})$$

$$S = \{(\texttt{Department}, \texttt{"Sales"})\}$$

$$P = \texttt{Trend}$$

IBM

Insight Generation: ISGen

- **Insight Identification:** Beam search method and statistical scoring functions to identify interesting insights by evaluating breakdowns, measures, and subspaces for significant patterns in the data
- **Pattern Scoring**: scoring functions are defined to measure the degree to which specific patterns are observed in view **v**
    - **Trend:** we use the Mann-Kendall Trend Test, where $\textsc{ScoreFunc}_{\textsc{Trend}}(v) = 1 - MK(v)$ Is the p-value
    - **Outstanding Value:** The outstanding value pattern occurs when the largest (or most negative) value significantly exceeds the other values (v_{max_1}, v_{max_2}) $\textsc{ScoreFunc}_{\textsc{OV}}(v) =$
    - **Attribution:** when the top value exceeds 50% of the total sum of all values $\textsc{ScoreFunc}_{\textsc{Attr}}(v) = \frac{\max(\{v_1, \ldots v_k\})}{\sum_i v_i}$
    - **Distribution Difference:** when the aggregation measure is COUNT(), using Jensen-Shannon divergence to compare the difference between the initial view v^I and final view v^F $\textsc{ScoreFunc}_{\textsc{DD}}(v^I, v^F) = JSD(\frac{v^I}{\sum_i v_i^I} || \frac{v^F}{\sum_i v_i^F})$

IBM

Insight Generation: ISGen

Initialization: dataset **D**, initial subspace **S_0**, perspective **(B, M)**, and parameters like beam width and maximum depth.

Subspace Expansion: The *expand* function identifies unused columns, assigns weights, weighted sampling of new columns.

Beam Search: expanding and scoring candidate based on the score functions subspaces iteratively for a specified depth.

Scoring and Pruning: Prunes the beam to retain the top **K** subspaces based on their scores.

Final Output: Returns the top K subspaces identified by their scores, highlighting insightful patterns in the data.

Algorithm 1 Insightful Subspace Search

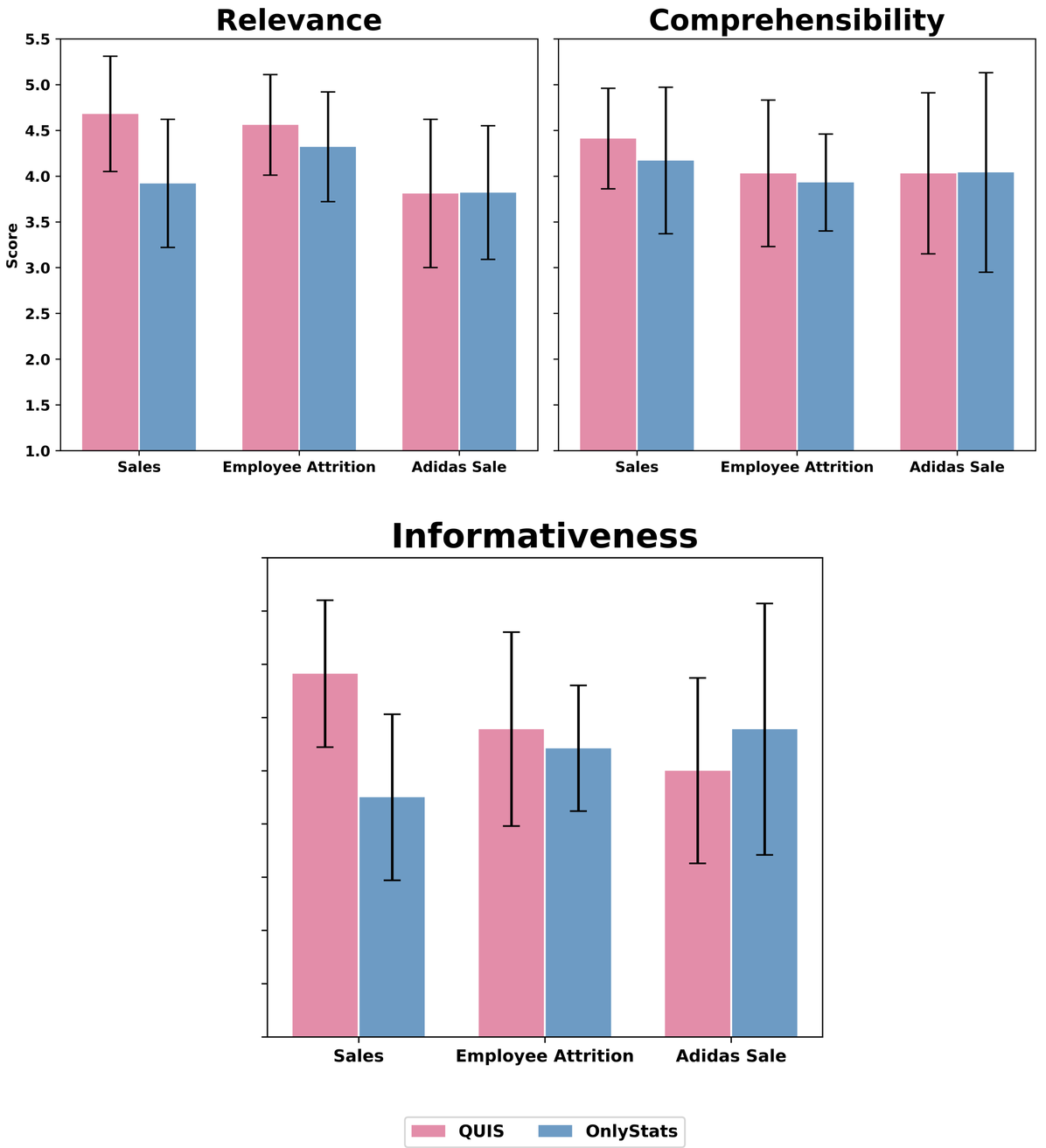
Require: Dataset D, Initial subspace S_0, perspective (B, M), language model LLM, ScoreFunc, beam_width, max_depth, exp_factor

Ensure: Top-K subspaces by score $\{S_1, \ldots S_k\}$

function EXPAND(S)
    avlbl_cols $\leftarrow$ D.cols $-$ S.used_cols
    ▷ S.used_cols are the columns used in the filters so far in S
    w $\leftarrow$ get_weights(avlbl_cols, LLM)
    $X \leftarrow$ sample(avlbl_cols, w)
    $y \leftarrow$ sample($D[X]$)
    return $S + (X, y)$
end function
beam $\leftarrow$ $[(S_0, \text{ScoreFunc}(S_0))]$
for depth $\in \{1, \ldots,$ max_depth$\}$ **do**
    for $(S,$ score$) \in$ beam **do**
        for $i \in \{1, \ldots,$ exp_factor$\}$ **do**
            $S_{new} \leftarrow$ EXPAND(S)
            score $\leftarrow$ ScoreFunc(S_{new})
            beam.add($(S_{new},$ score$)$)
        end for
    end for
    beam $\leftarrow$ top-k(beam, k=beam_width)
end for
return beam

Experiments

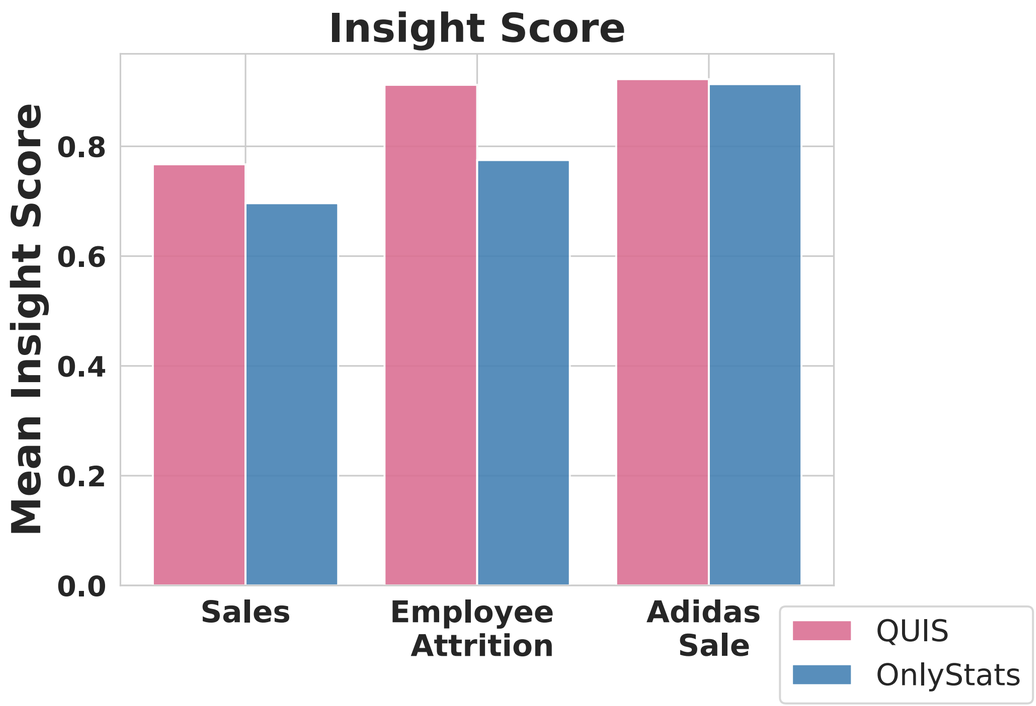
- Evaluated the QUIS's effectiveness using human assessment and insight scores.
- Under 2 conditions:
  1. **OnlyStats**: Replacing QUGEN with a Stats-Based Module to Evaluate ISGEN's Autonomous Performance
  2. **QUIS**: where both QUGEN and ISGEN were involved.
- **Human Evaluation Criteria**
  1. Relevance: To what extent the insight is applicable and useful in a given context?
  2. Comprehensibility: To what extent is this insight understandable and easy to follow?
  3. Informativeness: Does the insight provide substantial information for understanding the data?



Experiments

- Evaluated the QUIS's effectiveness using human assessment and insight scores.
- Under 2 conditions:
  1. **OnlyStats**: Replacing QUGEN with a Stats-Based Module to Evaluate ISGEN's Autonomous Performance
  2. **QUIS**: where both QUGEN and ISGEN were involved.
- **Insight Score:**

❖Comparison of Average Normalized SCOREFUNC Outputs for Insights Across Experimental Conditions



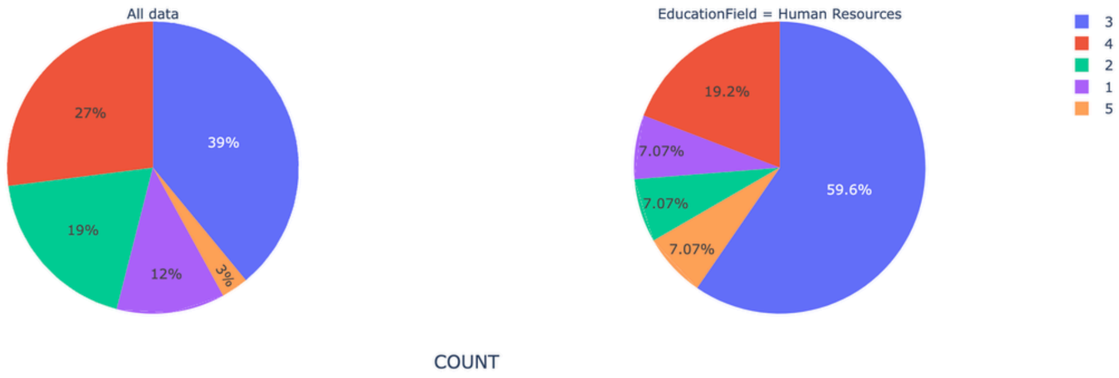
QUIS Example Results

Dataset: Employee Attrition
Question: Do products with higher unit prices result in higher total revenue?

Insight:

Insight Type: Distribution Difference

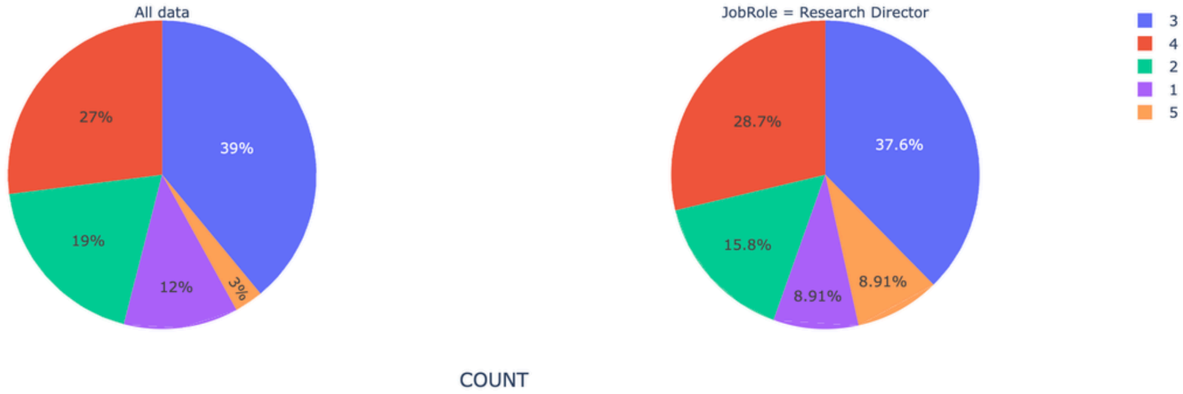
Employees with 3 years of education account for the highest number, with 39% of the total. However, for human resources employees, the highest number is found among those with 3 years of education, with 59% of the total. Employees with 1 year of education account for the lowest number, with 3% of the total.



Insight:

Insight Type: Distribution Difference

In the Research Director role, employees with a high school education still account for the highest number, but with a slightly lower percentage of 38.00%. Conversely, employees with a college degree account for the lowest number, but with a higher percentage of 9.00%.



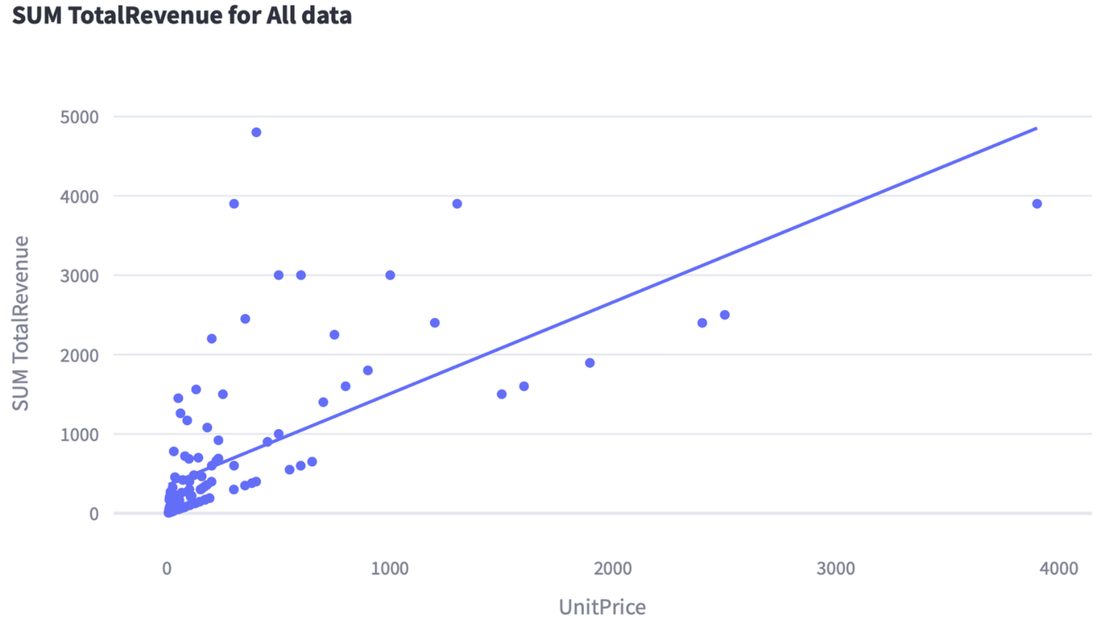
QUIS Example Results

Dataset: Sales
Question: Do products with higher unit prices result in higher total revenue?

Basic Insight:

Insight Type: Trend

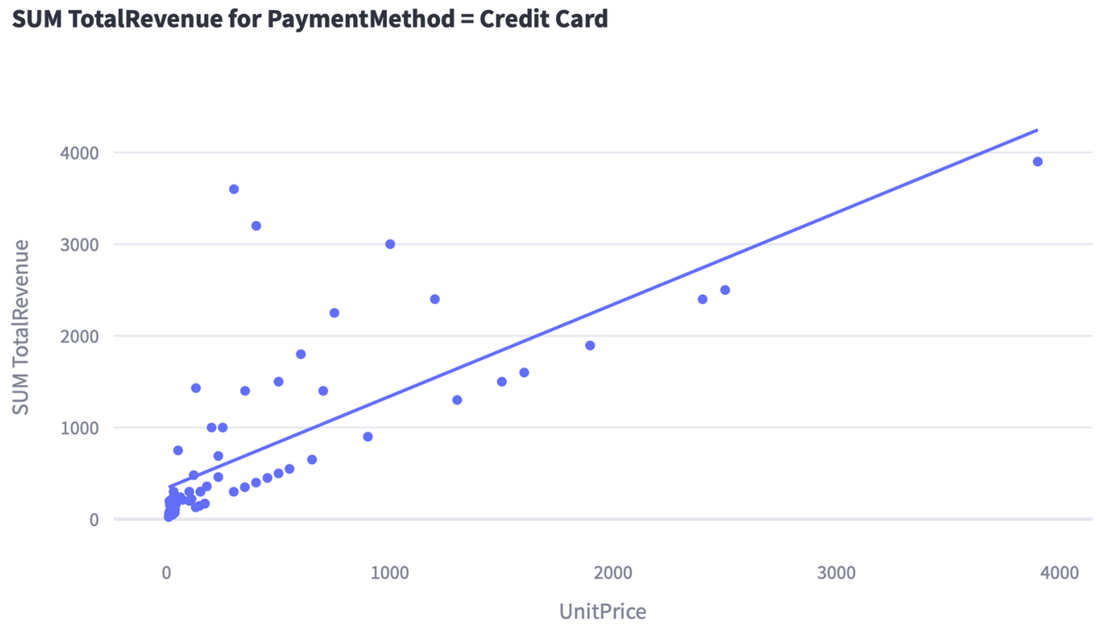
Total Revenue shows an increasing trend with the unit price.



Insight:

Insight Type: Trend

Sales revenue paid via credit card shows an increasing trend with the increase in unit price



IBM

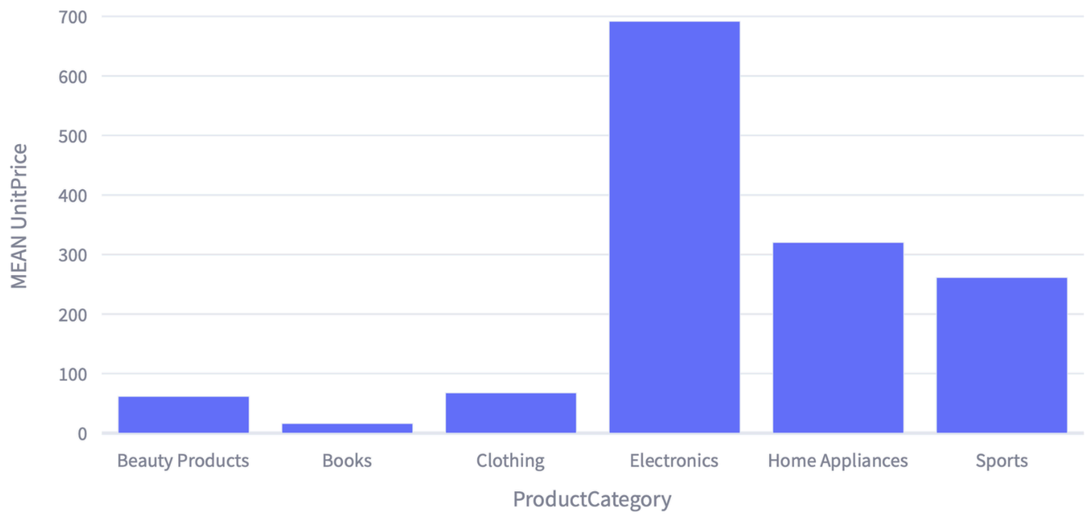
QUIS Example Results

Dataset: Sales
Question: What is the average pricing strategy employed for each product category?

Basic Insight:

Insight Type: Outstanding Value
The mean unit price is significantly higher for Electronics compared to the other product categories.

MEAN UnitPrice for All data

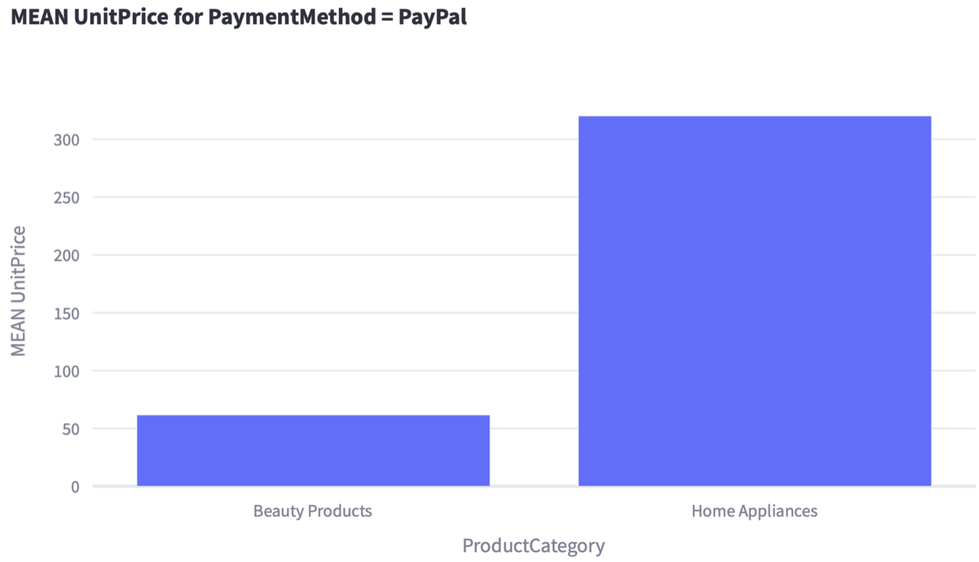


Insight:

Insight Type: Attribution

The average unit price of Home Appliances is significantly higher when paid through PayPal compared to other product categories

MEAN UnitPrice for PaymentMethod = PayPal



Thank you!

Questions please...