On Demand Detections and Explanations of Correlation Pattern Bias in EDA Workflows

Itay Elyashiv and Shira Turgeman

April 2024

1 Abstract

This article introduces a novel approach to enhancing Exploratory Data Analysis (EDA) and feature selection processes within the Data Science pipeline. Focusing on the pervasive issue of correlation bias(Simpson's Paradox), our solution provides automated detection and explanation of such occurrences within datasets, on user demand and based on user needs. Through experimental evaluation and user studies across diverse real-world datasets, we demonstrate the effectiveness and necessity of our feature in uncovering hidden patterns and mitigating erroneous conclusions.

2 Problem Description

This project focuses on enhancing the Exploratory Data Analysis (EDA) phase and feature selection within the Data Science pipeline process.

Exploratory Data Analysis (EDA) stands as a crucial initial step undertaken by data scientists and analysts to delve deeply into a new dataset. Its primary objectives include gaining a comprehensive understanding of the dataset's characteristics, identifying patterns, and extracting preliminary insights. However, novice analysts may overlook significant phenomena or fall into "statistical traps" that could skew their conclusions.

One such trap is what we refer to as "correlation bias", with the specific instance of Simpson's paradox, a phenomenon well-known among experienced analysts and researchers. Simpson's paradox occurs when a trend evident in different groups reverses or disappears when these groups are combined. This paradox can elude detection, especially by those less experienced in data analysis, leading to erroneous conclusions.

The consequences of overlooking such occurrences include flawed feature selection, false pattern mining and even biased models.

Furthermore, we noticed that the paradox is often made apparent by some subgroup that causes bias in the overall correlation between 2 attributes. Our goal is to identify such occurrences, find the responsible subgroups and inform the analyst.

Example 2.1 *Even though each group in the Iris dataset* [1] *shows a positive trend, when combined, the overall trend is negative, showcasing Simpson's paradox where the aggregated data leads to a different conclusion than the individual groups. See Figure 1.*

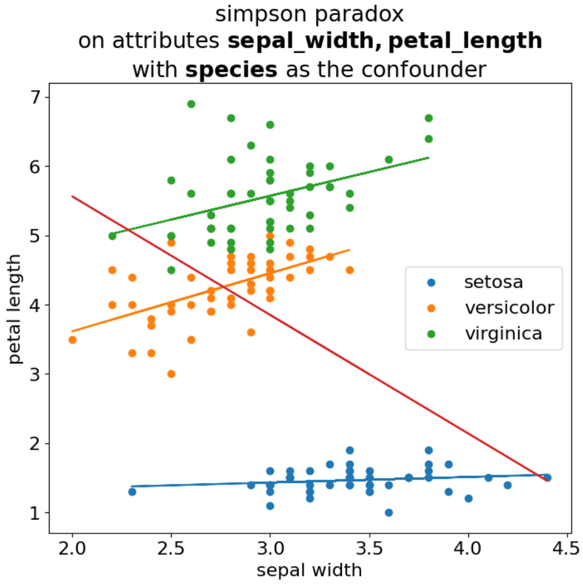


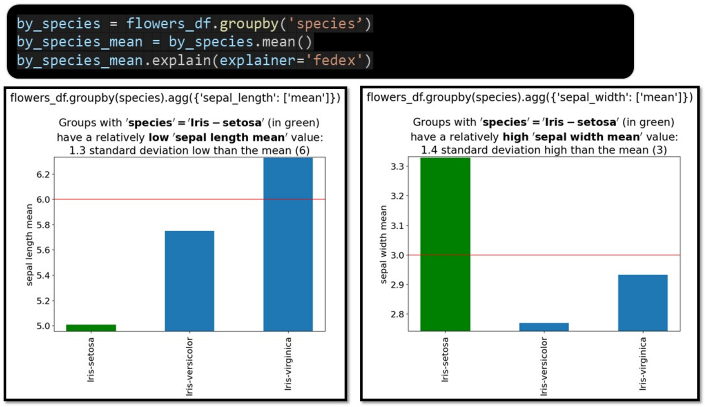
Figure 1: Simpson's paradox in Iris dataset

3 Solution overview

Our solution detects correlation bias occurrences within the data on demand, notifies the user about their existence and pinpoints the subset (if exists) that's responsible for the paradox. We aim to integrate this solution as a feature within data analysis frameworks that help discovering insights in data and explanations for exploration steps. As a first step, the solution is integratable in the pd-explain library which is an implementation of FEDEX([2]).

FEDEX (Efficient Data Exploration Explanations), is an EDA explanation framework that assists users in analyzing and understanding the results of their exploration steps. Its implementation is called pd-explain. FEDEX explains exploratory steps using a twofold process: (1) as an analyst would do, FEDEX inspects the resulted dataframe and discovers interesting aspects of it. For example, does it show outliers? Or perhaps a highly diverse set of values in one of the columns? (2) since naturally, not all underlying tuples equally contribute to the interesting pattern discovered, FEDEX detects which data subsets cause the resulted dataframe to be interesting. FEDEX tailors the explanation to the context of the generated dataframe, i.e., the EDA operation or query and the source dataframe, and then presents a coherent, easy-to-read captioned visualizations. This allows users to quickly understand and derive immediate insights from each exploratory step they make.

[1] Link to Iris data



```
by_species = flowers_df.groupby('species’)
by_species_mean = by_species.mean()
by_species_mean.explain(explainer='fedex')
```

Figure 2: Fedex results

In our case, users will be able to receive a correlation pattern bias explanation(including the problematic subset- group), given a grouped dataframe or a grouping attribute.

Example 3.1 *Data scientist Bob works on the Iris flowers dataset. The dataset consists of 150 samples of iris flowers, defined by four features: sepal length, sepal width, petal length, and petal width. Those features are used to classify the iris flowers into three species: Setosa, Versicolor, and Virginica. Each sample is labeled with the species it belongs to.*
Bob wishes to delve into the dataset. He performs a group-by operation based on the flower species, and utilizes the ‘pd-explain‘ library to generate the graphs depicted in Figure 2.
He received two explanations:
(1) The species ’setosa’ has a relatively low sepal length mean value, 1.3 standard deviations lower than the mean.
(2) The species ’setosa’ has a relatively high sepal width mean value, 1.4 standard deviations higher than the mean.
Though informative, those explanations depend on an aggregation function. If the division unfolds an occurrence of the Simpson’s paradox, the existing system isn’t able to detect it.
Bob calls our solution on dataframe ’by_species’, and receives an informative multimodal explanation about a Simpson’s paradox occurrence between the sepal width and sepal length features in the different groups. In addition, the explanation states that the group of Iris setosa is causing this effect. See figure 3.

Implementation of the solution:

The process of detecting Simpson’s paradox and generating explanations for a group-by operation:

1. Initially, the user performs a groupby operation on the data, utilizing a feature already available within the system. This operation organizes the data based on categorical values from column (A) and applies an aggregation function to another specified column (B), as exemplified by the syntax: ‘$df.groupby('A')$‘.

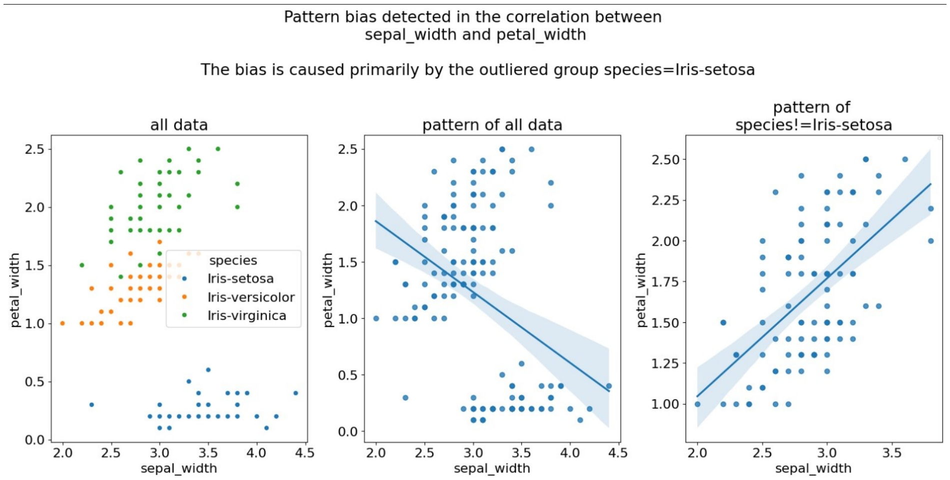


Figure 3: Explanation for Simpson's paradox in Iris dataset

2. Recognizing the user's interest in attribute A and desire to find patterns w.r.t this attribute, the system runs algorithm 1.

3. The algorithm traverses attribute pairs (B, C), to find pairs where their overall correlation ($corr(B,C)$) is different from that within each group of A:

   $\{corr_g(B,C) | g \in dom(A)\}$

4. To choose the most significant effect, the system calculates each detected appearance's interestingness:

   $I_A(d,B,C) = \frac{1}{|dom(A)|} \sum_{g \in dom(A)} |(corr_g(B,C) - corr(B,C))|$

5. To build a good explanation, it's crucial to identify the group that contributes the most to the Interestingness score. I.e, the group that causes the bias. To achieve this, we calculate each group's contribution to the effect- the difference in interestingness with and without the group subset:

   $$C(g,d,A,B,C) = |I_A(d,B,C) - I_A(d[A \neq g],B,C)|$$

   Then, we proceed to normalize the contributions in the following manner:

   $$\bar{C}(g,d,A,B,C) = \frac{C(g,d,A,B,C)}{\sum_{g' \in dom(A)} C(g',d,A,B,C)}$$

6. Given the set $EC_A = \{(B,C,g)\}$ of all explanation-candidates, we look for ones that obtain both a good contribution($\bar{C}(g,d,A,B,C) > \alpha$)and a high interestingness ($I_A(d,B,C) > \beta$) score, with predefined thresholds. To balance the two metrics we use a skyline-operator calculation. Namely, we define the set EX_A of desired explanations as a maximal subset of EC_A satisfying the following:

   $\forall (B,C,g) \in EX_A . \nexists (B',C',g') \in EC_A . (I_A(B',C',g') > I_A(B,C,g) \wedge \bar{C}(g',d,A,B',C) > \bar{C}(g,d,A,B,C))$

7. Finally, the solution presents the EX_A set in a user-friendly and accessible format, ensuring a seamless and intuitive experience.

Algorithm 1 A Modified FedEx Explanations Generation

```
Data: DataFrame $d$, categorical column $A$
Result: Explanation Candidates $\{ec_1, ..., ec_k\}$
$SPC \leftarrow empty\_group$
 for $column\ pairs\ (B, C) \subseteq d.columns\_numerical$ do
    if $Calculate\ interestingness\ score\ Ic(d, A, B, C) > \beta$ then
        $SP\_candidates \leftarrow (d, A, B, C)$;
foreach $candidate \in SP\_candidates$ do
    for $group\ g \in dom(A)$ do
        if $Calculate\ contribution\ score\ \bar{C}g(d,\ A,\ B,\ C) > \alpha$ then
            $EC \leftarrow (\bar{C}g(d, A, B, C, g), Ic(d, A, B, C))$
$EX = \underset{(B,C,g)\in EC}{\text{SKYLINE}}(I, \bar{C}g)$
 foreach $E \in EX$ do
    GenerateVisualExplanation($E$);
return $EX$
```

4 Experimental evaluation

We evaluated the quality and performance of our feature on three real world datasets. The results show that our explanations are necessary.

4.1 Datasets

We have used the following 3 datasets:

1. Iris species - The dataset contains a set of 150 records under 5 attributes - Petal Length, Petal Width, Sepal Length, Sepal width and Class(Species).

2. Penguins lter - The dataset contains observations of penguin species including measurements like bill length, bill depth, flipper length, and body mass, aiding in ecological research and species identification.

3. auto-mpg - The dataset contains information about various car models, including attributes like miles per gallon (mpg), cylinders, displacement, horsepower, weight, acceleration, model year, and origin.

4.2 Performance Evaluation

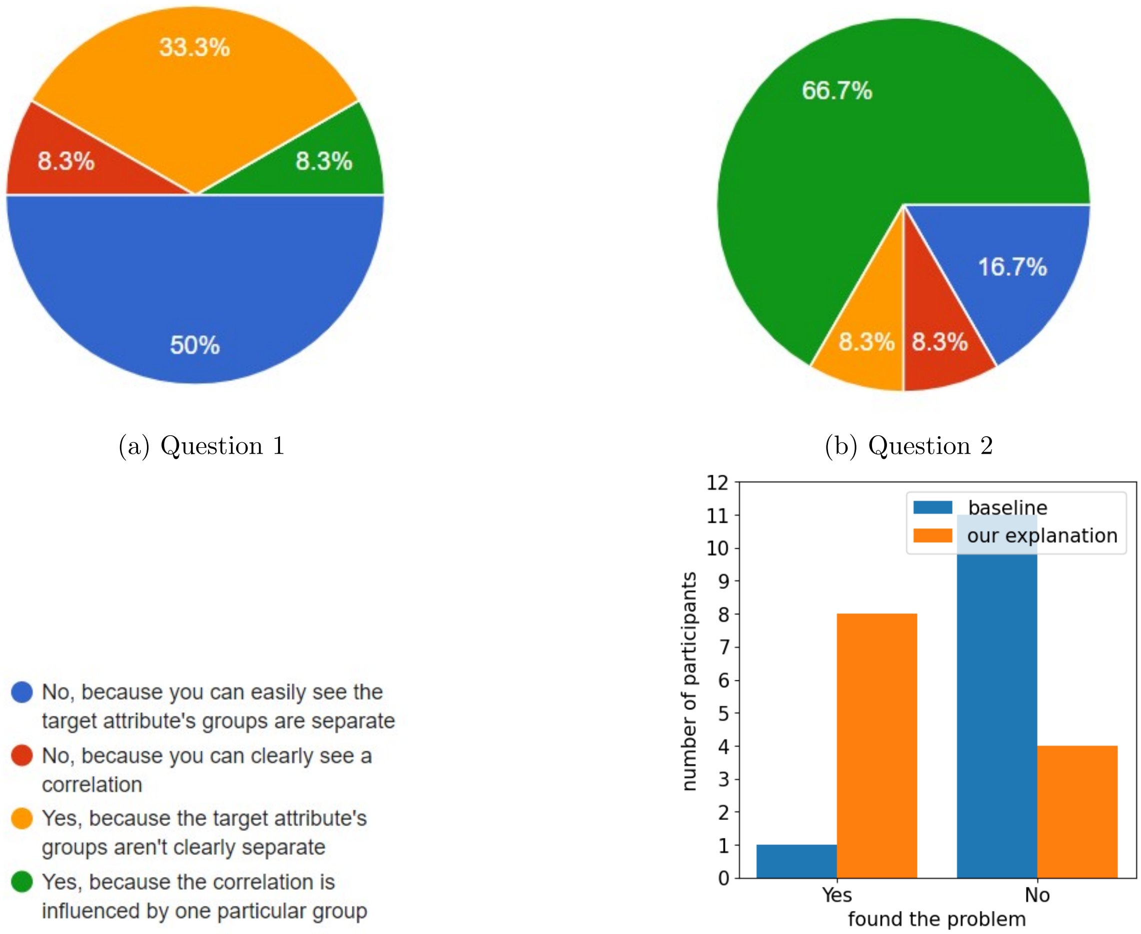
We conducted experiments using above three datasets known to exhibit Simpson's paradox. We tested our feature to ensure that it accurately detects the paradox within the data and provides an explanation for its occurrence. Results showed that our system successfully identified the paradox in all three datasets and generated clear explanations for each case. You can see the result on the Iris dataset in figure 3 and the full test and the results in the notebooks attached to GitHub([1]).

4.3 User Studies

We detail user studies that was performed to evaluate the quality and necessity of the explanation generated by our feature on Iris dataset.
The experiment involved 12 participants with different backgrounds: 5 participants held bachelor's degrees in Computer Science, while 7 were graduate students in Computer Science. They were asked to find a problem in the Iris data twice- first after looking at the baseline(simple grouping visualization) and then after looking at our explanation. They answered a Google Forms link with 3 questions:

1. Baseline: They saw the left graph in Figure 3 and were asked the following question: Is there a problem with the Iris data, visible in the graph?
   (a) No, because you can easily see the target attribute's groups are separate.
   (b) No, because you can clearly see a correlation.
   (c) Yes, because the target attribute's groups aren't clearly separate.
   (d) Yes, because the correlation is influenced by one particular group.
2. Our System: They saw the entire explanation in Figure 3 (3 graphs with explanation) and were asked the same question again.
3. They were asked if they were familiar with Simpson's paradox.

Results: The findings are clear: most participants didn't notice the Simpson's paradox in the data until it was explained to them. Before the explanation, over half didn't see any issue, and only one (8.3%) recognized the correlation bias. After our explanation though, the recognizers portion skyrocketed to 8(66.7%), see figure 4. When the null hypothesis is that our explanation didn't improve the problem detection within the participants, the p-value is low, hence our improvement is significant:
$\chi^2(1, N = 12) = \frac{(1-8)^2}{8} + \frac{(11-4)^2}{4} = 18.375, p = .000018$
In addition, 83% of users(figure 5), even though being computer science graduate students, weren't familiar with the Simpson's paradox, a finding that supports our claim that many researchers aren't aware of it and often neglect issues it rises.



(a) Question 1

(b) Question 2

(c) legend

(d) comparison between the baseline and our solution

Figure 4: Results from user studies

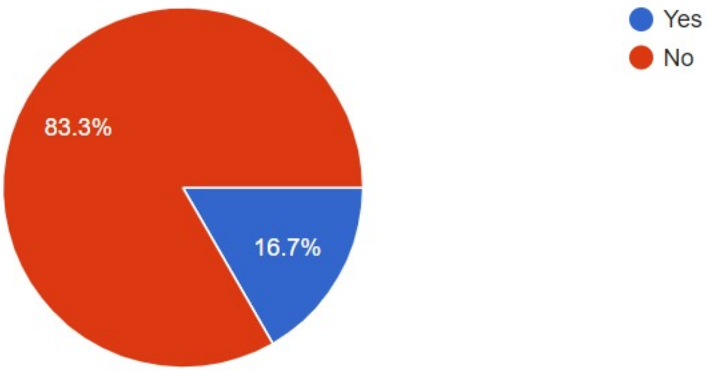


Figure 5: Question 3

5 Related work

Detecting Simpson's Paradox. previous work([3],[5]) has suggested ways to find occurrences of the phenomenon in datasets. Nevertheless, they didn't address the broader problem of correlation bias, attempt to find the cause of the occurrence, or integrate the solution into a programmatic EDA workflow.

On-Demand Exploratory Explanations/Insights. Our solution is interactive data analysis, which requires an on-demand framework. There are such frameworks for explaining exploration steps([2]) and providing insights on certain attributes in dataframes([4]).

6 Conclusion

In conclusion, this article presents a novel approach to enhancing the EDA phase and feature selection within the Data Science pipeline process by addressing the pervasive issue of correlation bias and Simpson's paradox. By integrating a solution that detects and explains such occurrences within datasets, our framework provides data scientists and analysts with valuable insights to avoid erroneous conclusions and biased models. Through experimental evaluation and user studies, we have demonstrated the effectiveness and necessity of our feature across diverse real-world datasets. Moreover, our findings underscore the importance of raising awareness about correlation bias and Simpson's paradox within the data science community and highlights the significance of on-demand exploratory explanations for facilitating informed decision-making via data analysis. Overall, our solution represents a crucial step forward in empowering data scientists with the tools and knowledge needed to extract meaningful insights and uncover hidden patterns within complex datasets.

References

[1] Tds-research repository. https://github.com/ItayELY/TDS-Research, 2024.

[2] Daniel Deutch, Amir Gilad, Tova Milo, Amit Mualem, and Amit Somech. Fedex: An explainability framework for data exploration steps. *arXiv preprint arXiv:2209.06260*, 2022.

[3] Yue Guo, Carsten Binnig, and Tim Kraska. What you see is not what you get! detecting simpson's paradoxes during data exploration. In *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics*, pages 1–5, 2017.

[4] Doris Jung-Lin Lee, Dixin Tang, Agarwal, et al. Lux: always-on visualization recommendations for exploratory dataframe workflows. *PVLDB*, 15(3):727–738, 2021.

[5] Rahul Sharma, Minakshi Kaushik, Sijo Arakkal Peious, Markus Bertl, Ankit Vidyarthi, Ashwani Kumar, and Dirk Draheim. Detecting simpson's paradox: A step towards fairness in machine learning. In *European Conference on Advances in Databases and Information Systems*, pages 67–76. Springer, 2022.