Why Not to Trust Big Data: Discussing Statistical Paradoxes*

Rahul Sharma[1][0000−0002−9024−8768], Minakshi Kaushik[1][0000−0002−6658−1712], Sijo Arakkal Peious[1][0000−0002−7858−9463], Mahtab Shahin[1][0000−0003−0034−018x], Ankit Vidyarthi[2][0000−0002−8026−4246], Prayag Tiwari[3][0000−0002−2851−4260], and Dirk Draheim[1][0000−0003−3376−7489]

[1] Information Systems Group,
Tallinn University of Technology,
Akadeemia tee 15a, 12618, Tallinn, Estonia
{rahul.sharma,minakshi.kaushik,sijo.arakkal,
mahtab.shahin,dirk.draheim}@taltech.ee
[2] Jaypee Institute of Information Technology, Noida, India
dr.ankit.vidyarthi@gmail.com
[3] Department of Computer Science, Aalto University, Finland
prayag.tiwari@aalto.fi

Abstract. Big data is driving the growth of businesses, data is the money, big data is the fuel of the twenty-first century, and there are many other claims over Big Data. Can we, however, rely on big data blindly? What happens if the training data set of a machine learning module is incorrect and contains a statistical paradox? Data, like fossil fuels, is valuable, but it must be refined carefully for the best results. Statistical paradoxes are difficult to observe in datasets, but they are significant to analyse in every small or big dataset. In this paper, we discuss the role of statistical paradoxes on Big data. Mainly we discuss the impact of Berkson's paradox and Simpson's paradox on different types of data and demonstrate how they affect big data. We provide that statistical paradoxes are more common in a variety of data and they lead to wrong conclusions potentially with harmful consequences. Experiments on two real-world datasets and a case study indicate that statistical paradoxes are severely harmful to big data and automatic data analysis techniques.

Keywords: Big Data · Artificial Intelligence · Machine Learning · Data Science· Simpson's Paradox · Explainable AI

1 Introduction

Data has always been critical in making decisions. Earlier, statistics and mathematics have been used to draw insights from data. However, in the last two

* This work has been partially conducted in the project "ICT programme" which was supported by the European Union through the European Social Fund.

decades, with the emergence of social media and big data technologies, data science, artificial intelligence (AI) and machine learning (ML) techniques have gained massive ground in practice and theory. These decision support techniques are being used widely to develop intelligent applications and acquire deeper insights from structured and unstructured data. In most AI use cases, ML based trained artificial systems provide fast and accurate outcomes; however, it does not guarantee accurate results for every use case in real life. Moreover, like statistics, understanding causal relationships and evaluating the existence of statistical paradoxes in the training dataset is not in the mainstream data science, AI and ML application scenarios. AI, machine learning, and big data are now widely used in medical sciences, social sciences, and politics, and they have a direct or indirect impact on human life and decisions. Therefore, understanding causal relationships and evaluating the existence of statistical paradoxes is essential for fair decision making [13, 14].

Statistical reasoning and probability theory are the foundation of many AI, big data and data science techniques, e.g., random forest [7], support vector machines [11], etc. Therefore, it is usual to have causal relationships and statistical paradoxes in these decision support techniques. A paradox can be a statement that leads to an apparent self-contradictory conclusion. Even the most well-known and documented paradoxes frequently confound domain specialists because they fundamentally violate common sense.

There are many statistical paradoxes (e.g., Simpson's Paradox, Berkson's Paradox, Latent Variables, Law of Unintended Consequences, Tea Leaf Paradox, etc.). Statistical paradoxes are not new to be discussed in statistics and mathematics; expert mathematicians and statisticians adequately addressed the severe impact of paradoxes. However, in modern decision support techniques, specifically AI and data science, causal relationships, data fallacies and statistical paradoxes are not appropriately addressed. In this article, we discuss the impact of Berkson's Paradox, Yule-Simpson paradox and causal inference on big data. We highlight several hidden problems in data that are not yet discussed in big data mining. We use two benchmark datasets for machine learning and a case study to demonstrate the existence of Simpson's paradox in different types of data.

The paper is organised as follows. In Sect. 2, we discuss why not trust data science, AI, ML and big data. In Sec. 3, we discuss two statistical paradoxes and discuss their impacts on big data mining. In Sect. 4, we use two benchmark datasets for machine learning to demonstrate the effects of Simpson's paradox. In Sect. 5 we provide a case study to analyse the impact of Simpson's paradox in real life. Finally, a discussion and conclusion is provided in Sect. 6 and Sect. 7.

2 Why not to trust on data science, AI, ML and big data

In AI, ML and Data Science, observing trends, mean and correlation between two variables for making decisions is not always correct. E.g., suppose in a city, the Covid-19 infection rate of smokers is less than the infection rate of the non-

smokers. Can we claim that smoking prevent Covid-19? It is a perfect case of poor data science where all the variables and features in the dataset are not appropriately observed. In today's world, data literacy may not seem exciting when compared to machine learning algorithms or big data mining, but it should be the foundation for all data mining processes.

Datasets, irrespective of their size and type, are not self-explanatory. It's all numbers and statistics responsible for creating stories out of datasets. Therefore, it's essential to validate a dataset statistically and evaluate the existence of any statistical paradoxes. AI, ML, big data and data science based techniques generate knowledge from data. Therefore, decision support techniques are easily prone to statistical paradoxes and can not be trusted.

3 Statistical Paradoxes

Statistical paradoxes aren't something that hasn't been discussed before. These terms are widely used in statistics and have been around for over a century. Statistical paradoxes are fundamentally related with various statistical challenges and mathematical logic including causal inference [27, 28], the ecological fallacy [24,31], Lord's paradox [36], propensity score matching [32], suppressor variables [10], conditional independence [12], partial correlations [16], p-technique [8] and mediator variables [26]. The instances of statistical paradoxes specifically Simpson's paradox have been discussed in various data mining techniques [17], e.g., association rule mining [2] and numerical association rule mining [20,21,34]

More recently, Kügelgen et al. [37] pointed out the importance of statistical analysis of real data and demonstrated instances of Simpson's paradox in Covid-19 data analysis. They provide that the overall case fatality rate (CFR) was higher in Italy than in China. However, in every age group, the fatality rate was higher in China than in Italy. These observations raise many questions on the accuracy of data and its analysis. Heather et al. [25] have addressed the existence of Simpson's paradox. In psychological science, Kievit et al. [22] examined the instances of Simpson's paradox. Alipourfard et al. [3] have discovered the existence of Simpson's paradox in social data and behavioural data [4]. Therefore, understanding data, especially big data, is more critical than its processing. In the following two sections, we discuss Berkson's paradox and Yule-Simpson's Paradox to demonstrate their vast impact on big datasets.

3.1 Berkson Paradox

Berkson's paradox can make it appear as if there is a relationship between two independent variables when there is no relationship between the variables. In 1946, despite diabetes being a risk factor for cholecystitis, Berkson [5] observed a negative correlation between cholecystitis and diabetes in hospital patients. Berkson state that If at least one of two independent events occurs, they become conditionally dependent. In other words, two independent events become conditionally dependent, given that at least one of them occurs. Statistically,

Berkson's paradox and Simpson's paradox are very close to each other. Berkson's paradox is a type of selection bias caused by systematically observing some events more than others.

$$if\ 0 < P(A) < 1,\ 0 < P(B) < 1\ and \tag{1}$$

$$P(A|B) = P(A)\ then \tag{2}$$

$$P(A|B, A \cup B) = P(A)\ Hence \tag{3}$$

$$P(A|B, A \cup B) > P(A) \tag{4}$$

As given in Eq. 1 to Eq. 4, $P(A|B)$, a conditional probability, is the probability of observing event A given that B is true. The probability of A given both B and (A *or* B) is smaller than the probability of A given (A *or* B).

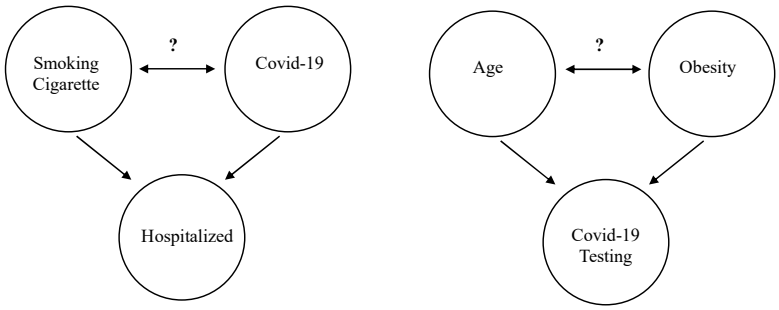

Fig. 1. Berkson's paradox: two noticeable example of Covid-19 which introduce a collider.

As we all know, smoking cigarettes is a well-known risk factor for respiratory diseases. However, recently Wenzel T. [9] observed a negative co-relation between Covid-19 severity and smoking cigarettes. In another observation, Griffith et al. [18] describe it as a Collider Bias or Berkson's paradox. In Fig. 1, we demonstrate an example of collider. Here Smoking cigarettes, Covid-19 are two independent variables, but they collide with another random variable, hospitalised. Here, the variable hospitalised is collider for both smoking cigarettes and Covid-19.

3.2 Yule-Simpson's Paradox

In the year 1899, Karl Pearson et al. [29] demonstrated a statistical paradox in marginal and partial associations between continuous variables. Later in 1903, Udny Yule [38] explained "the theory of association of attributes in statistics"

and revealed the existence of an association paradox with categorical variables. In a technical paper published in 1951 [33], Edward H. Simpson described the phenomenon of reversing results. However, in 1972, Colin R. Blyth coined the term "Simpsons Paradox" [6]. Therefore, this paradox is known with different names and it is popular as the Yule–Simpson effect, amalgamation paradox, or reversal paradox [30].

We start the discussion on the paradox by using the real-world dataset from Simpson's article [33]. In this example, analysis for medical treatment is demonstrated. Table 1 summarises the effect of the medical treatment for the entire population ($N = 52$) as well as for men and women separately in subgroups. The treatment appears effective for both male and female subgroups; however, the treatment seems ineffective at the whole population level.

Table 1. 2×2 contingency table with sub population groups D1 and D2.

	Population $D = D_1 + D_2$		**Sub-population** D_1		**Sub-population** D_2	
	Success (S)	Failure ($\neg S$)	Success (S)	Failure ($\neg S$)	Success (S)	Failure ($\neg S$)
Treatment (T)	$a_1 + a_2$	$b_1 + b_2$	a_1	b_1	a_2	b_2
No Treatment ($\neg T$)	$c_1 + c_2$	$d_1 + d_2$	c_1	d_1	c_2	d_2

Definition 1. *Consider D groups of data such that group D_1 has A_i trials and $0 \leq a_i \leq A_i$ "successes". Similarly, consider an analogous D groups of data such that group D_2 has B_i trials and $0 \leq b_i \leq B_i$ "successes" Then, Simpson's paradox occurs if*

$$\frac{a_1}{A_1} \geq \frac{b_1}{B_1} and \frac{a_2}{A_i} \geq \frac{b_2}{B_2} \text{ for all } i = 1, 2, \ldots, n \text{ but } \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} A_i} \leq \frac{\sum_{i=1}^{n} b_i}{\sum_{i=1}^{n} B_i} \tag{5}$$

we use the following example to show how this equation works.

$$\frac{10}{20} > \frac{30}{70} and \frac{10}{50} > \frac{10}{60} \text{ but } \frac{10+10}{20+50} < \frac{30+10}{70+60}, \tag{6}$$

We could also flip the inequalities and still have the paradox since A and B are chosen arbitrarily.

Classically the paradox is expressed via contingency tables. Let a 2×2 contingency table for treatment (T) and sucess (S) in the i^th sub-population is represented by a four-dimensional vector of real numbers $D = (a_1, b_1, a_2, b_2)$. Then

$$D = \sum_{i=1}^{N} D_i = \left(\sum a_i, \sum b_i, \sum c_i, \sum d_i\right) \tag{7}$$

is the aggregate dataset over N sub populations. This can be read as given in Table 1.

We can also demonstrate the Simpson's paradox scenario via probability theory and conditional probabilities. Let $T = treatment$, $S = successful$, $M = Male$, and $F = Female$ then,

$$\mathsf{P}(S \mid T) = \mathsf{P}(S \mid \neg T) \tag{8}$$

$$\mathsf{P}(S \mid T, M) > \mathsf{P}(S \mid \neg T, M) \tag{9}$$

$$\mathsf{P}(S \mid T, \neg M) > \mathsf{P}(S \mid \neg T, \neg M) \tag{10}$$

Based on Eq. 8, 9 and 10, one should use the treatment or not? As per the success rate for the male and female population, the treatment is a success, but overall, the treatment is a failure. This reversal of results between groups population and the total population has been referred to as Simpson's Paradox. In statistics, this concept has been discussed widely and named differently by several authors [29, 38].

4 Existence of Simpson's Paradox in Big Data

Simpson's paradox can exist in any dataset irrespective of its size and type [23]. The paradox demonstrates the importance of having human experts in the loop to examine and query Big Data results. In this section, we present datasets to analyse the presence and implications of Simpson's paradox on big data.

To identify an instance of the Simpson paradox in a continuous dataset with n continuous variable and m discrete variables, we can compute a correlation matrix $(n \times n)$ for all the data. Then for m discrete variable with k_m levels, an additional $(n \times n)$ matrix needs to be calculated for each level of variables as follows. Therefore, we need to calculate the $1 + \sum_i^m = k_i$ correlation matrices of size $(n \times n)$ and compare it with the lower half of $\sum_i^m = k_i$ for subgroup levels. We have also discussed the measures to find the impact of one numerical variable to another numerical variable [19].

4.1 Datasets

We use the iris dataset and miles per gallon (mpg) dataset, the two benchmark datasets for machine learning to demonstrate the presence of Simpson's paradox in data.

Iris Dataset: Ronald Fisher introduced the iris dataset in a research paper [15]. It consists three types of iris species (Setosa, Versicolor, Virginicare), each with 50 data samples. The species names are categorical attributes, length and width are continuous attributes.

In order to identify the existence of Simpson's paradox in the iris datasets, we first visualise the relationship between the length and width of each pair of

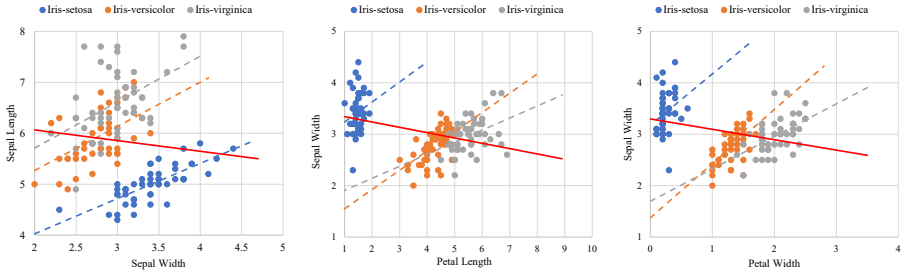

Fig. 2. Simpson's paradox in Iris dataset: there is a positive correlation between the three pairs of sepal length and petal width for the Iris-setosa, Iris-versicolor and Iris-virginicare (dashed lines). However, the overall trend for the length and width for the entire population is negative (solid red line) in all three combinations.

candidate attributes. As shown in Fig. 2, in the iris dataset, we identify the existence of Simpson's paradox for three pairs of measurements. 1. sepal length and width, 2. sepal width and petal length, and 3. sepal width and petal width.

In Fig. 2, the correlation between sepal width and sepal length is positive (dashed line) for each species. However, the correlation between sepal width and sepal length for the entire population is negative (solid red trend line). Similarly, the pair of petal length, width, and the pair of petal width and sepal width have positive trends for each species; however, the overall trend for the length and width for the entire population is negative in both cases. Therefore, this is a clear case of Simpson's paradox in the iris dataset.

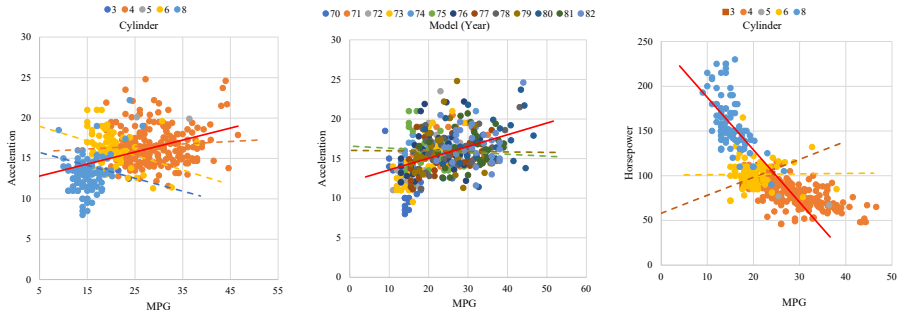

Fig. 3. Simpson's paradox in auto MPG dataset: There is a negative correlation between MPG and acceleration for three cylinders engines and six cylinders engines; however, the overall trend between MPG and acceleration is positive (solid red line). Similarly, the overall trend is positive for MPG and acceleration with respect to the model year. However, the overall trend between MPG and horsepower according to the engine cylinders is negative.

The MPG dataset: Ross Quinlan used the Auto MPG dataset in 1993 [30]. The dataset contains 398 automobile records from 1970 to 1982, including the vehicle's name, MPG, number of cylinders, horsepower, and weight. The dataset includes three multi-valued discrete attributes and five continuous attributes.

In order to identify the existence of Simpson's paradox in the MPG datasets, we visualise the relationship between MPG, acceleration and horsepower for two categorical attributes (number of cylinders and model year). The goal of analysing the dataset is to know the factors that influence each car's overall fuel consumption. The dataset consists of fuel consumption in mpg, horsepower, number of cylinders, displacement, weight, and acceleration.

In the MPG dataset, we identify the existence of Simpson's paradox in three pairs of measurements. 1. MPG with acceleration according to the engine cylinders, 2. MPG with acceleration with respect to their model year, and 3. MPG with horsepower according to the engine cylinders. In the figure. 3, it is visualised that there is a negative correlation between MPG and acceleration for three cylinders engines and six cylinders engines; however, the overall trend between MPG and acceleration is positive (solid red line). Similarly, the overall trend is opposite for MPG with acceleration with respect to the model year and MPG with horsepower according to the engine cylinders.

5 Analysis Simpson's Paradox in Real Life: A Case Study

The case study is from the California Department of developmental services (CDDS), United States of America [35]. As per the annual reports published by the department, the average annual expenditures on Hispanic residents were approximately one-third (1/3) of the average expenditures on White non-Hispanic residents. According to the marginal analysis, it was a solid gender discrimination case. However, a conditional analysis of ethnicity and age found no evidence of ethnic discrimination. Furthermore, except for one age group, the trends were completely opposite. The average annual expenditures on White non-Hispanic residents were less than the expenditures on Hispanic residents. Therefore, it is a perfect case of Simpson's paradox in real life.

Table 2. Number of residents by ethnicity and percentage of expenditures.

Ethnicity	Sum of Expend. ($)	% of Expend.	# of Residents	% of Residents
American Indian	145753	0.81	4	0.4
Asian	2372616	13.13	129	12.9
Black	1232191	6.82	59	5.9
Hispanic	4160654	23.03	376	37.6
Multi Race	115875	0.64	26	2.6
Native Hawaiian	128347	0.71	3	0.3
Other	6633	0.04	2	0.2
White not Hispanic	9903717	54.82	401	40.1
Total	18065786	100%	1000	100%

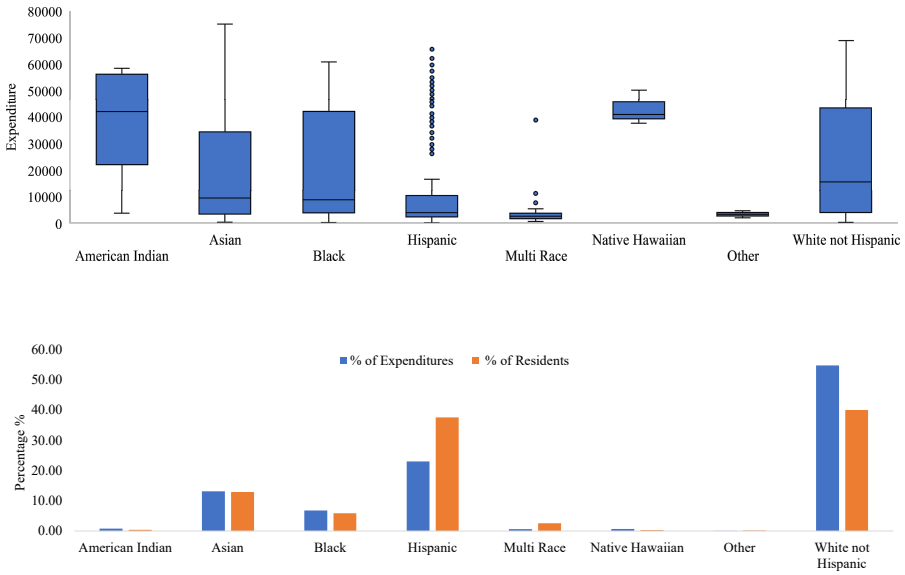

Fig. 4. Distribution of expenditure as per the ethnic groups.

5.1 The Dataset

We use the same dataset to analyse the original claims. The dataset is publicly available at [1]. The dataset mainly consists various information of one thousand disabled residents (DRs) under six important variables (ID, age age group, gender, expenditures, ethnicity). Each DR has a unique identification, i.e., "ID". The state department uses AGE to decide the financial needs and other essential needs of DRs. The age groups of the residents are divided into six age groups. (0-5 years old, 6-12 years old, 13-17 years old, 18-21 years old, 22-50 years old, and 51 years old). These groups are based on the amount of financial assistance required at each stage of age. E.g., The 0-5 age group (preschool age) has the fewest needs and thus requires the least funding.

The"Expenditures" variable represents the annual expenditures made by the state to support each resident and their family. Information about the expenditures, the number of residents and their percentage as per ethnicity is given in Table 2. The expenditures include all the expenses, including psychological services, medical fees, transportation and housing costs such as rent (especially for adult residents). As far as the case is concerned, "ethnicity" is the most important demographic variable in the dataset. The dataset includes eight ethnic groups.

As demonstrated in Fig. 4, the population difference between the Hispanic and the White non-Hispanic people is significantly less. However, there is a big difference between the distribution of assistance to the Hispanic and the White

non-Hispanic group. Therefore, these two populations are selected for the case study for further investigation.

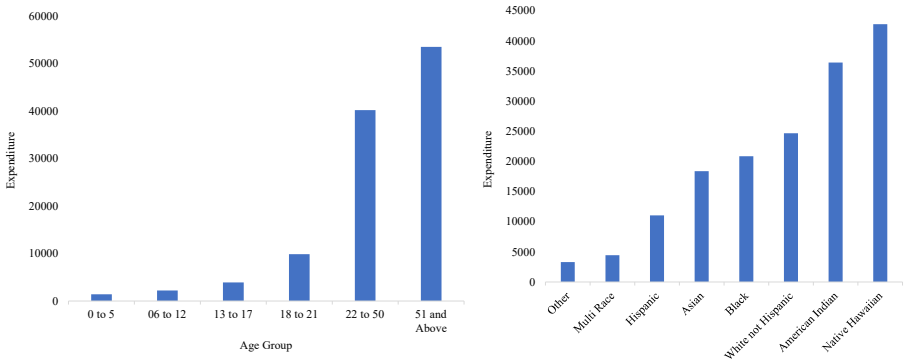

Fig. 5. 1. Average expenditure by age group, 2. average expenditure by ethnicity.

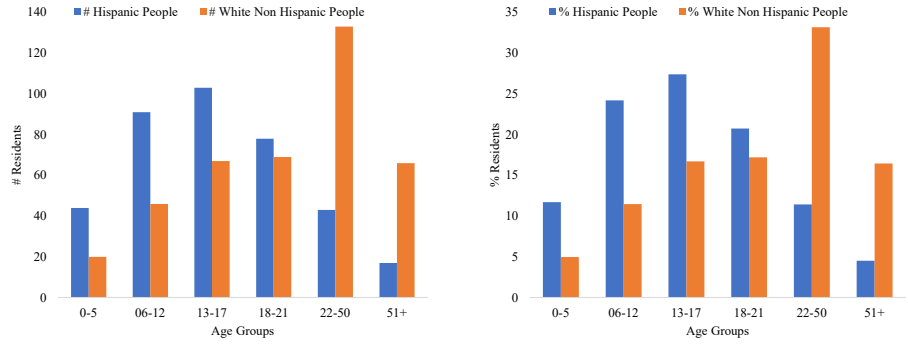

Fig. 6. 1.Hispanic and white non Hispanic residents with their age groups, 2. Percentage of Hispanic and white non Hispanic residents according the age groups.

5.2 Data Analysis

We begin the data analysis by comparing the total amount of expenditure in relation to different ethnic groups. As per the bar chart given in Fig. 5, It is clear that the average expenditure on Hispanic residents is significantly lower than the

White non-Hispanic residents. Moreover, the analysis of average expenditure by the age groups shows that the average expenditure was very high for the older age groups. As per Fig. 5, it is also a clear case of age discrimination. However, age is not considered a factor for the discrimination because older people are eligible to get higher expenditures.

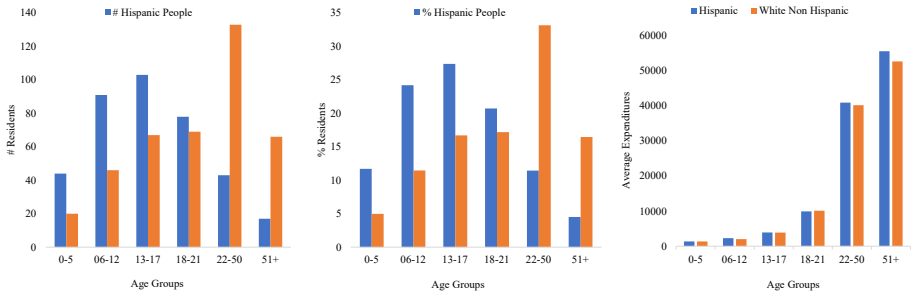

Fig. 7. Average expenditures by ethnicity and age groups.

The overall Hispanic population receiving assistance is younger than the white non-Hispanic population receiving assistance. As the age is showing discriminatory behaviour, therefore, we compare the average amount of funds received by the two observed ethnic groups as per their age groups in Fig. 6. It is clear that the number of beneficiaries from the Hispanic group is higher in the lower age groups, while the number of beneficiaries from the white non-Hispanic group is higher in the older groups. As white non-Hispanic are older people, therefore, they are receiving more support.

Now we see an opposite picture of the case, in Fig. 6. The aggregated data shows that white non-Hispanic people have more support from the department; however, for most of the age groups except one age group, the average expenditure for the Hispanics was higher. So, we are witnessing Simpson's paradox!. The age group variable proved to be lurking in this case, without which we can not show any results in marginal data.

6 Discussion

The existence of statistical paradoxes in benchmark datasets and in real-life case studies provides a direction to understand the causality in decision making. We noticed that most machine learning and deep learning algorithms focus only on identifying correlations rather than identifying the real or causal relationships between data items. Therefore, understanding and evaluating causality is an important term to be discussed in big data, Data Science, AI and ML.

7 Conclusion

Handling statistical paradoxes is a complex challenge in AI, ML and Big Data. Different paradoxes state the possibilities of errors in the outcomes of automatic data analysis conducted for AI, Ml and big data based applications. In this paper, we discussed the existence of Berkson's paradox and demonstrate the existence of Simpson's paradox and in two real datasets. Statistical paradoxes in data reflect the importance of probabilities and causal inference and seek a manual inspection of datasets. We argue that if confounding effects are not properly addressed in datasets, outcomes of an data analysis can be completely opposite. However, with the right tools and data analysis, a good analyst or data scientist can handle it in a better way. The statistical paradoxes confirm essential statistical evaluation for datasets and demonstrate the importance of human experts in the loop to examine and query Big datasets.

References

1. California Department of Developmental Services CDDS expenditures, `https://kaggle.com/wduckett/californiaddsexpenditures`
2. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: Proceedings of VLDB'1994 – the 20th International Conference on Very Large Data Bases. p. 487–499. Morgan Kaufmann (1994)
3. Alipourfard, N., Fennell, P.G., Lerman, K.: Can you trust the trend? discovering simpson's paradoxes in social data. In: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. p. 19–27. WSDM '18, Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3159652.3159684
4. Alipourfard, N., Fennell, P.G., Lerman, K.: Using simpson's paradox to discover interesting patterns in behavioral data. In: Proceedings of the Twelfth International AAAI Conference on Web and Social Media. AAAI Publications (2018)
5. Berkson, J.: Limitations of the application of fourfold table analysis to hospital data. Biometrics Bulletin **2**(3), 47–53 (1946), `http://www.jstor.org/stable/3002000`
6. Blyth, C.R.: On Simpson's paradox and the sure-thing principle. Journal of the American Statistical Association **67**(338), 364–366 (Jun 1972)
7. Breiman, L.: Random forests. Machine learning **45**(1), 5–32 (2001)
8. Cattell, R.B.: P-technique factorization and the determination of individual dynamic structure. Journal of Clinical Psychology (1952)
9. Commission, E., Centre, J.R., Wenzl, T.: Smoking and COVID-19 : a review of studies suggesting a protective effect of smoking against COVID-19. Publications Office (2020). https://doi.org/doi/10.2760/564217
10. Conger, A.J.: A revised definition for suppressor variables: A guide to their identification and interpretation. Educational and psychological measurement **34**(1), 35–46 (1974)
11. Cortes, C., Vapnik, V.: Support-vector networks. Machine learning **20**(3), 273–297 (1995)
12. Dawid, A.P.: Conditional independence in statistical theory. Journal of the Royal Statistical Society: Series B (Methodological) **41**(1), 1–15 (1979). https://doi.org/https://doi.org/10.1111/j.2517-6161.1979.tb01052.x

13. Draheim, D.: DEXA'2019 keynote presentation: Future perspectives of association rule mining based on partial conditionalization., Linz, Austria, 28th August 2019. https://doi.org/10.13140/RG.2.2.17763.48163
14. Draheim, D.: Future perspectives of association rule mining based on partial conditionalization. In: Hartmann, S., Küng, J., Chakravarthy, S., Anderst-Kotsis, G., A Min Tjoa, Khalil, I. (eds.) Proceedings of DEXA'2019 - the 30th International Conference on Database and Expert Systems Applications. LNCS, vol. 11706, p. xvi. Springer, Heidelberg New York Berlin (2019)
15. Fisher, R.A.: The use of multiple measurement in taxonomic problems. Annals of Eugenics **7**(2), 179–188 (Sep 1936). https://doi.org/10.1111/j.1469-1809.1936.tb02137.x
16. Fisher, R.A.: Iii. the influence of rainfall on the yield of wheat at rothamsted. Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character **213**(402-410), 89–142 (1925)
17. Freitas, A.A., McGarry, K.J., Correa, E.S.: Integrating bayesian networks and simpson's paradox in data mining. In: Texts in Philosophy. College Publications (2007)
18. Griffith, G.J., Morris, T.T., Tudball, M.J., Herbert, A., Mancano, G., Pike, L., Sharp, G.C., Sterne, J., Palmer, T.M., Davey Smith, G., Tilling, K., Zuccolo, L., Davies, N.M., Hemani, G.: Collider bias undermines our understanding of COVID-19 disease risk and severity. Nature Communications **11**(1), 5749 (2020). https://doi.org/10.1038/s41467-020-19478-2, https://doi.org/10.1038/s41467-020-19478-2
19. Kaushik, M., Sharma, R., Peious, S.A., Draheim, D.: Impact-driven discretization of numerical factors: Case of two- and three-partitioning. In: Big Data Analytics. pp. 244–260. Springer International Publishing, Cham (2021)
20. Kaushik, M., Sharma, R., Peious, S.A., Shahin, M., Yahia, S.B., Draheim, D.: On the potential of numerical association rule mining. In: International Conference on Future Data and Security Engineering. pp. 3–20. Springer (2020)
21. Kaushik, M., Sharma, R., Peious, S.A., Shahin, M., Yahia, S.B., Draheim, D.: A systematic assessment of numerical association rule mining methods. SN Computer Science **2**(5), 348 (2021). https://doi.org/10.1007/s42979-021-00725-2
22. Kievit, R., Frankenhuis, W., Waldorp, L., Borsboom, D.: Simpson's paradox in psychological science: a practical guide. Frontiers in Psychology **4**, 513 (2013). https://doi.org/10.3389/fpsyg.2013.00513
23. Kim, Y.: The 9 pitfalls of data science. The American Statistician **74**(3), 307–307 (2020). https://doi.org/10.1080/00031305.2020.1790216
24. King, G., Roberts, M.: Ei: a (n r) program for ecological inference. Harvard University (2012)
25. Ma, H.Y., Lin, D.K.J.: Effect of simpson's paradox on market basket analysis. Journal of Chinese Statistical Association **42**(2), 209–221 (Jun 2004). https://doi.org/10.29973/JCSA.200406.0007
26. MacKinnon, D.P., Fairchild, A.J., Fritz, M.S.: Mediation analysis. Annual Review of Psychology **58**(1), 593–614 (2007). https://doi.org/10.1146/annurev.psych.58.110405.085542, pMID: 16968208
27. Pearl, J.: Causal inference without counterfactuals: Comment. Journal of the American Statistical Association **95**(450), 428–431 (2000)
28. Pearl, J.: Understanding simpson's paradox. SSRN Electronic Journal **68** (01 2013). https://doi.org/10.2139/ssrn.2343788

29. Pearson Karl, L.A., Leslie, B.M.: Genetic (reproductive) selection: Inheritance of fertility in man, and of fecundity in thoroughbred racehorses. Philosophical Transactions of the Royal Society of London: Series A **192**, 257–330 (Dec 1899)
30. Quinlan, J.: Combining instance-based and model-based learning. In: Machine Learning Proceedings 1993, pp. 236–243. Elsevier (1993). https://doi.org/10.1016/B978-1-55860-307-3.50037-X
31. Robinson, W.S.: Ecological correlations and the behavior of individuals. American Sociological Review **15**(3), 351–357 (1950)
32. Rosenbaum, P.R., Rubin, D.B.: The central role of the propensity score in observational studies for causal effects. Biometrika **70**(1), 41–55 (1983)
33. Simpson, E.H.: The interpretation of interaction in contingency tables. Journal of the Royal Statistical Society: Series B (Methodological) **13**(2), 238–241 (1951)
34. Srikant, R., Agrawal, R.: Mining quantitative association rules in large relational tables. In: Proceedings of the 1996 ACM SIGMOD international conference on Management of data. pp. 1–12 (1996)
35. Taylor, S.A., Mickel, A.E.: Simpson's paradox: A data set and discrimination case study exercise. Journal of Statistics Education **22**(1), 8 (Mar 2014). https://doi.org/10.1080/10691898.2014.11889697
36. Tu, Y.K., Gunnell, D., Gilthorpe, M.S.: Simpson's paradox, lord's paradox, and suppression effects are the same phenomenon–the reversal paradox. Emerging themes in epidemiology **5**(1), 1–9 (2008)
37. Von Kugelgen, J., Gresele, L., Scholkopf, B.: Simpson's paradox in COVID-19 case fatality rates: A mediation analysis of age-related causal effects. IEEE Transactions on Artificial Intelligence **2**(1), 18–27 (Feb 2021). https://doi.org/10.1109/tai.2021.3073088
38. Yule, G.U.: Notes on the theory of association of attributes in statistics. Biometrika **2**(2), 121–134 (02 1903)