TALLINN UNIVERSITY OF TECHNOLOGY
School of Information Technology
Department of Software Science

Huseyn Garayev 201438IVSM

A WEB-BASED PLATFORM FOR DETECTING AND HANDLING THE SIMPSON'S PARADOX

Master's Thesis

Supervisor: Rahul Sharma
M.Tech, Early-stage Researcher
Co-Supervisor: Dirk Draheim
Professor

Tallinn 2022

TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond
Tarkvarateaduse instituut

Huseyn Garayev 201438IVSM

VEEBIPÕHINE PLATVORM SIMPSONI PARADOKSI TUVASTAMISEKS JA KÄSITLEMISEKS

Magistritöö

Juhendaja: Rahul Sharma
Uurija
Kaasjuhendaja: Dirk Draheim
Professor

Tallinn 2022

Author's declaration of originality

I hereby certify that I am the sole author of this thesis. All the used materials, references to the literature and the work of others have been referred to. This thesis has not been presented for examination anywhere else.

Author: Huseyn Garayev
Date: 08.05.2022

Abstract

A paradox is a statement or phenomenon which leads to contradictory and unintuitive conclusions. Simpson's paradox is one of the known paradoxes which occurs in statistics and data science and manifests itself as an effect of the reversal of the associations between variables during aggregating and disaggregating data. The Simpson's Paradox can be harmful when making decisions on top of the data that exhibits it; therefore, it is essential to address the paradox, especially in automatic data mining, machine learning, and data science. There are multiple solutions to detect and resolve the Simpson's Paradox; however, there is no unified platform for exploring the datasets with regard to the paradox. This thesis describes the algorithms for detecting the Simpson's Paradox using correlation comparisons, the solution of the Simpson's Paradox using the probability-based adjustments, and the Web-based Platform, which incorporates the aforementioned algorithms and graphical representations of data under the convenient user interface. All the algorithms and methods presented in the thesis are tested on the set of infamous datasets which are studied for the matter of the existence of the Simpson's Paradox and the results are described in detail.

The thesis is written in the English language and contains 51 pages of text, 7 chapters, 21 figures, 22 tables.

Annotatsioon

Paradoks on väide või nähtus, mis viib vastuoluliste ja ebaintuitiivsete järeldusteni. Simpsoni paradoks on näide paradoksist, mis esineb statistikas ja andmeteaduses ning avaldub muutujate vaheliste seoste ümberpööramise tagajärjena kui andmeid koondatakse ja jagatakse. Simpsoni paradoksi olemasolu avaldatavates andmetes võib viia kahjulike otsuste tegemiseni, mistõttu on oluline osata seda käsitleda, eriti automaatse andmekaevandamise, masinõppe ja andmete puhul teadus. Simpsoni paradoksi tuvastamiseks ja lahendamiseks on palju võimalusi, kuid paradoksi osas puudub ühtne platvorm andmekogumite uurimiseks. Selles lõputöös kirjeldatakse Simpsoni paradoksi tuvastamise algoritme korrelatsioonivõrdluste abil, Simpsoni paradoksi lahendust tõenäosuspõhiste kohanduste abil ja veebipõhist platvormi, mis sisaldab ülalnimetatud algoritme ja andmete graafilisi esitusi läbi mugava kasutajaliidese. Kõiki lõputöös esitatud meetodeid testitakse üldtuntud andmekogumite peal, kus uuritakse Simpsoni paradoksi esinemist ja mille tulemusi kirjeldatakse üksikasjalikult.

Lõputöö on kirjutatud inglise keeles ning sisaldab teksti 51 leheküljel, 7 peatükki, 21 joonist, ning 22 tabelit.

List of abbreviations and terms

SP	Simpson's Paradox
IPW	Inverse Propensity Weighting
CRF	Case Fatality Rate
CSV	comma-separated values
API	Application Programming Interface
URL	uniform resource locator
json	JavaScript Object Notation
CLI	Command Line Interface

Table of Contents

List of Figures

List of Tables

1. Introduction

The amount of data generated worldwide is increasing exponentially, and it is harnessed and analyzed automatically by the various artificial intelligence approaches, machine learning algorithms, and manual analysis. But sometimes, data can mislead, and this can have adverse consequences. The existence of bias in data can cause incorrect or unintuitive conclusions during data analysis, and some of these cases can be described as statistical paradoxes.

In this thesis, the author concentrates on the specific case of statistical paradox called Simpson's paradox for artificial intelligence-based applications. The Simpson's paradox is a well-studied phenomenon that was first described by Edward H. Simpson in his paper [1]. Statistical paradoxes can occur in a wide variety of data but require more awareness, particularly in data analysis and artificial intelligence-based applications. For instance, one of the latest occurrences of the Simpson's Paradox has caused a lot of confusion in COVID-19 Case Fatality Rate [2] in China and Italy. This shows how important it is to address the paradoxes in data.

The main objective of my work was to build a web-based platform to automatically detect the existence of Simpson's paradox in the machine learning datasets and resolve the paradox by displaying the adjusted aggregated form of data for unbiased statistical analysis. There have been numerous papers published on the detection of the Simpson's Paradox in recent years [3, 4] via different ways and for different types of data [5]; however, there is no state of the art web-based platform available to check the impact of the Simpson's paradox and adjust the aggregated form of data for unbiased statistical analysis.

In the first part of the thesis, we propose a method to detect the Simpson's paradox and identify the confounding variable by using the preprocessing techniques and Pearson correlation Index to find the association reversal instances in data. This method is applicable to all forms of Simpson's paradox and covers a wide range of data types. In the second part, we develop an algorithm to solve the paradox existing in the form of relative rates[4] by aggregating the data using the probability-based adjustments, namely the Inverse Propensity Weighting method.

The final part consists of the overview of a web-based platform that incorporates the proposed algorithms in one service and provides a visual interface for the users to test or explore their datasets with regard to Simpson's paradox. In each part of the process, experiments are conducted on the set of infamous datasets which have been previously studied for the existence of the Simpson's paradox. For this purpose, we use the Kidney Stone dataset, Berkeley University admission dataset, California DDs Expenditures dataset, Iris Flower dataset, and Penguin dataset.

1.1 Research Goal

1. Developing algorithms for detecting Simpson's paradox and identifying the confounding variables in diverse machine learning datasets (Categorical, Continuous, etc.).

2. Developing an algorithm to resolve the impact of Simpson's paradox.

3. Developing a web-based platform to test and explore machine learning datasets for the existence of the Simpson's paradox.

1.2 Research Questions

1. How to detect the existence of the Simpson's paradox and identify the confounding variables in various machine learning datasets?

2. How to develop an algorithm to resolve the impact of Simpson's paradox on different types of training datasets?

3. How to develop a web-based platform to identify the existence of Simpson's Paradox?

1.3 Development Steps

- **Step 1:** Developing and improving a method for SP detection
- **Step 2:** Developing and improving a method for SP solution
- **Step 3:** Developing, coding and testing an algorithm for SP detection
- **Step 4:** Developing, coding and testing an algorithm for SP solution
- **Step 5:** Building a Web Platform with user interface on top of the developed algorithms

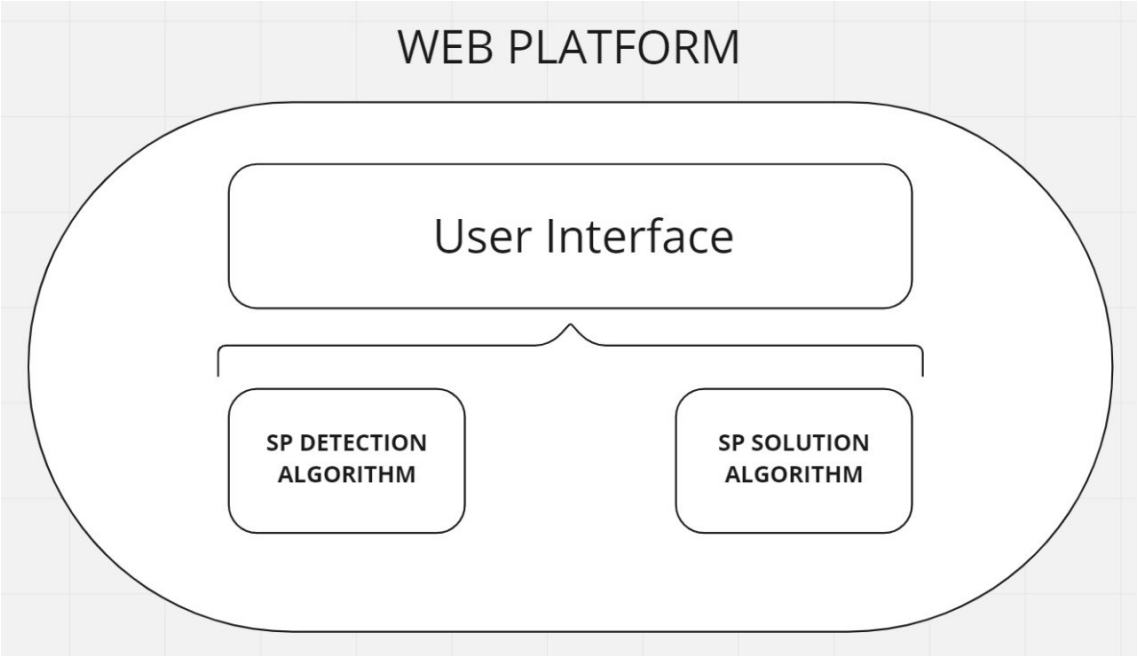

Figure 1. Overview of Development

1.4 Organization of the Thesis

- **Related Work 2:** First part of this chapter contains an overview of the Simpson's Paradox along with a brief history of the phenomenon. The second part describes the work related to the detection of SP. The third part describes the work related to the solution of SP.
- **Detection of Simpson's Paradox and Identification of confounding variable 3:** This chapter demonstrates the algorithm for the identification of the confounding variable and the detection of SP. Next, it gives the results of applying this algorithm to the datasets which contain SP instances.
- **Resolving the Simpson's Paradox 4:** In this chapter, the algorithm for the adjustment of the data aggregation is given along with the overview of IPW, which is used in the algorithm. Next, the results of applying this algorithm to the datasets used in the previous section are demonstrated.
- **Web-based Platform 5:** This chapter describes how the developed algorithms are used in the backend service, contains an explanation of how to use the Web Platform, architecture, choice of technologies, and the description of the Software Development process.
- **Time Evaluations 6:** Table with the time evaluations of the developed algorithms applied to the datasets is given in this chapter.
- **Conclusion 7:** In this chapter, the general overview of the thesis is given along with the discussion of the completeness and limitations of the work done.

2. Related Work

2.1 Simpson's Paradox

Simpson's paradox also known as a reversal paradox is a phenomenon in statistics and data science, where the association between a pair of variables reverses when conditioning by another variable. For aggregated data, the relationship between variable X and Y has one sign, and when disaggregating data into the subpopulations of variable Z the relationship has the opposite sign. Simpson's Paradox is frequently encountered in medical and social sciences. This effect can lead to incorrect or biased conclusions when analyzing the data, therefore it is instrumental to be able to address it. The paradox is getting resolved as we take a closer look at the confounding effect of Z during the statistical analysis or modelling.

Edward Simpson first described the Simpson's paradox in 1951, he reported that the association disappears when disaggregating or aggregating data [1] however the association reversal effect was first noted in Cohen and Nagel's work [6]. In Lindley and Novick's article [7] the Simpson's paradox was explored deeper and important conclusions were made on the decision-making process when encountering the paradox. They demonstrated that we should base our decisions either on the aggregated data or disaggregated data based on the additional information extracted from the context. Today the casualty allows us to take a more detailed look at the Simpson's paradox and we can determine which form of data is applicable to which situation.

Simpson's paradox can appear everywhere, but it requires more attention when it causes confusion in the minds of people and hinders the decision-making process. There has been a huge list of the Simpson's paradox cases in the history of statistical analysis. In this thesis, the author is going to explore some of them in the experiment sections of the paper. A lot of interesting paradox discoveries happened in recent years, for instance, the paper [8] illustrates the paradox occurrence in Quantum Harmonic Oscillator and the Nonlinear Schrodinger Equation. They further speculate on the likelihood of the Simpson's paradox appearing in Quantum Mechanics.

Another popular instance of the Simpson's paradox is studied in [2]. Over the course of recent years, the world has been fighting the deadly pandemic which originates from

the SARS-CoV-2 virus and causes acute respiratory conditions [9]. One of the most important indicators of the pandemic is the CRF or case fatality rate which is the ratio of the fatal cases to the confirmed cases of Covid-19. If we consult the data on CRFs reported by Chinese and Italian institutions in the early stages of the pandemic, we observe the instance of the Simpson's paradox. As given in Fig. 2, the case fatality rate for the entire population is higher in Italy however when we take a look at each of the age groups separately the case fatality rate happens to be higher in China in all of the cases. How can this phenomenon be explained? The most important fact is that CRFs indicate the conditional probabilities of the Covid-19 fatality for the given age group and country therefore the information on the distribution of the cases in each age group remains hidden from us in this context. The distribution of Covid-19 cases over the age groups is vastly different in Italy and China. The Italian population is older than the Chinese population and the majority of Covid-19 cases in Italy were reported for individuals aged 60 and over. Combining this information with the fact that age is positively correlated with the mortality rate for respiratory infections such as Covid-19 explains this unintuitive pattern in data.

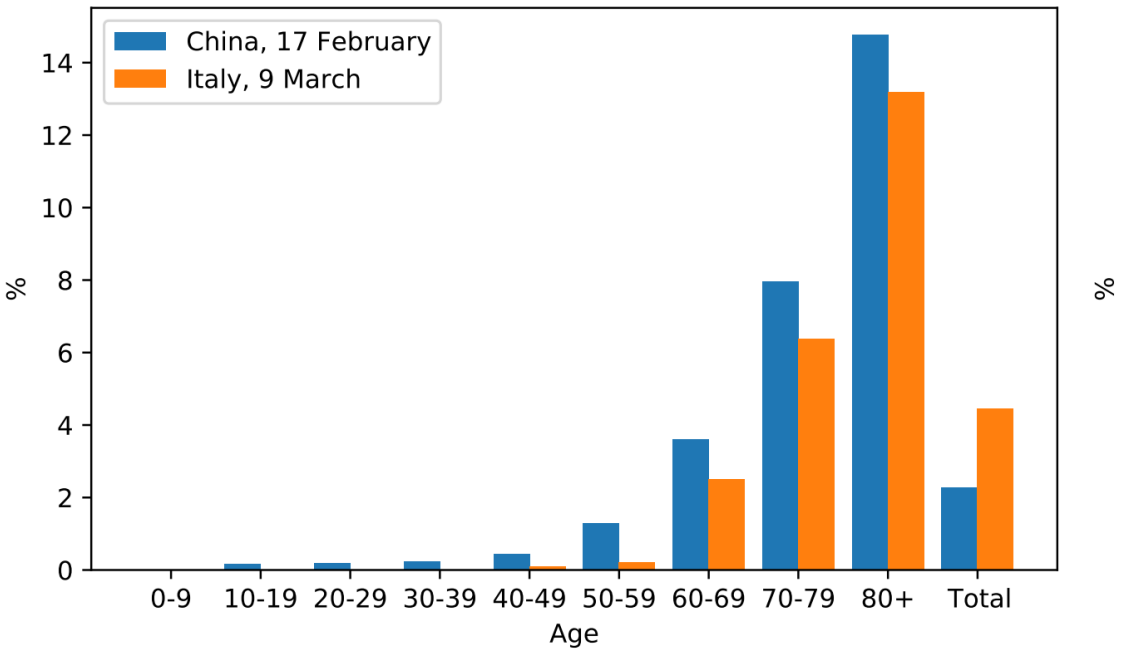

Figure 2. Case fatality rates (CFRs) by age group [2]

2.2 Detection of Simpson's Paradox

Before being able to examine or tackle the Simpson's Paradox we have to detect it first unless it is encountered accidentally. In [10] which is one of the earliest papers on the detection of the Simpson's paradox, the author developed the fundamental algorithm where the attributes are traversed and the dataset is partitioned for each of the attributes, and the paradoxes conditions are checked using the probabilities. This is a very intuitive algorithm

however it contains some limitations and rules enforced on the dataset, attribute form, and composition. In one of the recent papers [3], authors uncover the instances of Simpson's paradox by finding the pairs (dependent variable X and confounding variable Y) which satisfy the rules of the paradox. Linear models are fitted for aggregated and disaggregated data and the reversals are examined with respect to the slope of the linear models. There are also several packages and tools for the detection of the Simpson's paradox. For instance, an R package [11] where the users can pass the independent, dependent variables and the potential confounder to check for the paradox. It works only for continuous data and checks only for one confounder. Another example is Automatic Simpson's Paradox Detector [12] which uses the Regression models and works for a wide range of dataset and attribute types as it contains the preprocessing steps which bring the data to the form applicable for the algorithm.

In paper [13], decision trees are used for the detection of the Simpson's paradox and identification of the confounding variables. Classification and regression trees are used to predict the value of the target variable given the set of input variables. Trees are built by partitioning the dataset into the subsets in a recursive manner, and the partitioning is based on the set of rules that maximize the homogeneity of the subsets. In the method described in this paper, the authors use conditional inference trees [14] where the partitioning is based on the correlation values. The main objective is the structure of the tree and the order of the splits(partitioning), not the prediction of the records. Given the dataset, cause variable X, target variable Y and the set of all potential confounders Z', tree is fitted with the target Y and the predictors comprising X and Z'. In the case of Y being categorical, the classification tree is fitted, in the case of Y being numerical the regression tree is fitted. The presence of X and Z' variables in the splitting sequence, as well as the splitting sequence itself, determines the potential confounders and their confounding behaviour. Trees that exhibit the structure where the first split is done on Z(single confounder) and the next split on X indicate the association between X and Y, confounded by Z and this is the only scenario that might correspond to the instance of the Simpson's paradox, to the instance of the partial paradox (association reversal for some of the subgroups), or even the absence of the paradox where the effect for the subgroups differs only in magnitude.

2.3 Solution of Simpson's Paradox

Any form of bias and paradox in data should be mitigated to avoid issues when working with data or analyzing it [15]. The contradictory impact of the Simpson's Paradox happens because of the unintuitive relationship among the variables and it disappears when the specific questions are asked, and data are represented in a way that allows us to see the effect of the confounding variable. In most cases, the paradox gets resolved when we

understand why the paradoxical effect happens and find the right way to represent data to make the decision-making process unbiased during the data analysis. For that, we have to perform the analysis of the causal graphical models of the system [16]. One of the major pieces of work in a graphical representation of systems is presented in [17] If we consider the relative rates form of the paradox and build the causal model of the system, we can see the full causal explanation of why the paradox occurs.

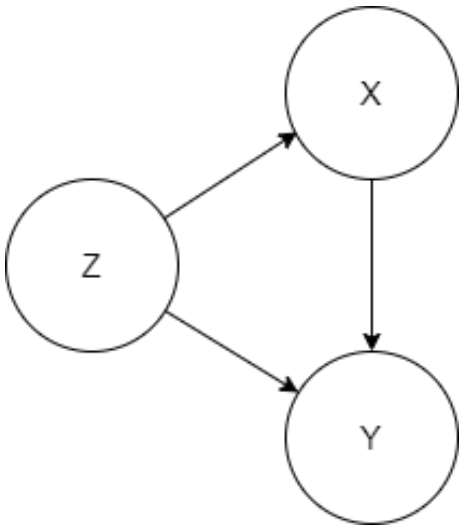

Figure 3. Causal graph with a confounding effect

As we can see from Fig. 3, the confounding variable Z affects both X and Y which corresponds to the case where the occurrence of the Simpson's paradox is possible when we study the effect of X on Y. But how do we avoid the paradox? Is it even possible to avoid the paradox? Yes, we can if we know all the variables that affect the outcome variable [18]. One way to avoid the paradox is to gather the data in a randomized trial manner to make sure that there is an equal distribution of data instances for each subgroup of the confounding variable. But unfortunately, it is not always possible to conduct a randomized trial taking into account all possible scenarios of the statistical bias. It is still possible to eliminate the effect of Simpson's paradox. Judea Pearl's do-calculus solves this problem, and it consists of a set of rules which measures the causal effect among variables and brings a casual system to the probabilistic calculations [19, 20, 16]. Similarly, there are many different methods to adjust for confounding [21]. We will focus on the Inverse Propensity Weighting method in the further sections of the paper.

3. Detection of Simpson's Paradox and Identification of the confounding variable

Simpson's paradox primarily exists in two forms: linear trends and relative rates [4]. Relative rates form of the paradox occurs when variable *X* is categorical and when we examine the causal effect of 2 categories of *X* on variable *Y* the associations reverse during aggregation or disaggregation of data. Linear Trends form of the paradox occur when both *X* and *Y* variables are numerical(continuous) and the linear trend between these variables reverses during aggregation or disaggregation of data. The author explored both of the forms separately as the preprocessing method differs for these forms. The author uses the Pearson correlation index to find out the relationships between variables and identify the confounding variable further on.

Given the input variables *x* and *y*, the Pearson correlation index allows us to measure the strength of the linear association between these variables. The output value lies between -1 and 1, values less than 0 imply a negative association with -1 indicating exact negative association, values greater than 0 imply a positive association with 1 indicating the exact positive association and 0 implies no correlation.

$$r = \frac{\sum(x - m_x)(y - m_y)}{\sqrt{\sum(x - m_x)^2 \sum(y - m_y)^2}}$$

where x and y are input vectors, m_x and m_y are means of the variables respectively. The intuition behind this approach is that if the association reversal happens for certain variables we will be able to get the same reversal effect in the correlation indexes of the variables. The most important indicator is the sign of the correlation coefficient, the magnitude is not important for the identification of the confounder however it might be instrumental in the further explanatory analysis of the paradox effect.

3.1 Relative Rates

Inputs to the algorithm: X - categorical variable by which we condition, *x1* - first category of variable X, *x2* - second category of variable X, Y - continuous or categorical variable (with 2 categories) which is aggregated. Outputs: Z - confounding variable, p - proportion of the subgroups of Z which exhibit reversal of association between X and Y.

The first primary step of the algorithm is to convert the values of the categorical input variables to binary values. The first category of the variable will be substituted with 0 and the second category will be substituted with 1. This preprocessing technique will allow us to apply the Pearson correlation index function and identify the relationship between categorical variables or between categorical and numerical (continuous) variables. All the substitutions are saved in order to be able to return the original values in data tables.

Gender	Result
male	success
female	success
male	failure

Table 1. Example table before preprocessing

Gender	Result
0	1
1	1
0	0

Table 2. Example table after preprocessing

The next part of the algorithm is to calculate the correlation index between X and Y variables with the values of the corresponding columns in the dataset. This way we obtain the information on the sign of the relationship between these variables. Further, we traverse the list of remaining categorical variables, calculate the Pearson index conditioning on each subgroup(category), count the proportion of subgroups where the correlation index reversed relative to the index in aggregated data, and store the value key pairs in an array. We can further get the array element where the value (proportion) is the highest. A value greater than 0 implies the existence of Simpson's paradox and the maximal value of 1 implies a full reversal effect. Therefore the fact of Simpson's paradox occurrence can be decided based on the results of Algorithm 1.

Algorithm 1 Identification of the confounding variable for relative rates form of the Simpson's paradox

Input: dataset *D*, variable *x*, variable *y*
Output: a pair consisting of confounding variable and proportion of reversed association signs

```
d[x] = Preprocess(d[x]) // conversion of categorical column to binary
 d[y] = Preprocess(d[y]) // conversion of categorical column to binary in case it's categorical
 aggreg_index = Pearson(d[x] , d[y]) // calculate corr. index between columns
indexes = [] // initialize index array to store key value pairs where the key is column and value is the number of reversed subgroups
 cols = columns(D) // initialize array of all columns of D
 foreach column ∈ cols do
    if Column Is Not Categorical(column) then
       Continue
    end
    else
       subgroups = Categories(column) // get the categories of a column
       coefficients = [] // initialize empty array to store the correlation indexes foreach subgroup ∈ subgroups do
          disaggreg_index = Pearson( D[x]: where D[column] = subgroup, D[y]: where D[column] = subgroup) calculate corr. index between columns for current subgroup
          Add index of disaggregated to correlation indexes array
       end
    end
    reversed_subgroups = proportionReversedSubgroups(aggreg_index, coefficients ) // calculate proportion of the correlation indexes reversed with respect to the correlation index for the aggregated data
    Add {column, reversed_subgroups} values into indexes
end
Store the max values of indexes pairs into result
 Return result
```

3.1.1 Experiments for Relative Rates

The experiments are performed on the set of infamous datasets where the existence of Simpson's paradox has been studied before. For relative rates, The author of this thesis uses the Kidney stone dataset, Berkeley university admissions dataset, and California DDS dataset. As an initial step, all the datasets are run through the preprocessing functions which

convert the categorical input variables to binary. For each dataset, the author demonstrates the tables with Pearson correlation index between X and Y for the whole data and for each subgroup and explains how the algorithm operates on these tables.

Berkeley university admissions dataset

This is a very popular dataset that contains the results of admission for males and females in different majors. The rate of admission for females is less than for males when data is aggregated, however, when we consider each major separately, female admission rates exceed the rates for males in most of the subgroups. The version of the dataset which is used contains 12764 records with attributes: 'Gender', 'Major', 'Admission'. 'Gender' attribute is set as X variable and 'Admission' attribute as Y variable.

Variable X	Variable Y	Correlation
Gender	Admission	0.0933

Table 3. Correlation coefficient between Gender and Admission

Table 3 demonstrates the Pearson correlation index returned by the algorithm between 'Gender' and 'Admission' variables. Where 'Gender' contains values 0 and 1 (M and F respectively) and 'Admission' contains values 0 and 1 (Accepted and Rejected respectively). The correlation between these variables is 0.0933 which is an indication of a positive association.

Major	Variable X	Variable Y	Correlation
A	Gender	Admission	-0.0630
B	Gender	Admission	-0.0208
C	Gender	Admission	0.0303
D	Gender	Admission	-0.0193
E	Gender	Admission	0.0414
F	Gender	Admission	-0.0288
Other	Gender	Admission	0.0309

Table 4. Correlation coefficients between Gender and Admission for each subgroup of Major

The algorithm traverses the list of all the potential confounders which is just one in this case. It identifies 'Major' as a confounder and the percentage of the subgroups with

association reversal is 57.14%. As we can see from Table 4, correlation index between the 'Gender' and 'Admission' variable is negative for all the subgroups except for 'C', 'E', and 'Other' subgroups. The results correspond to the prior case studies of the sex bias in Berkeley University admissions [22].

Kidney Stone dataset

The Kidney Stone dataset is also the dataset where the existence of the Simpson's paradox has been studied. It represents data for the kidney stone cases and the results of the treatments along with the severity of the illness (size of the stone). The success rate for treatment B happens to be higher than for treatment A in the whole population however when we consider each stone size separately for each sub-population success rate for treatment A is greater than for treatment B. Apparently 'stone_size' is the confounding variable and the algorithm should identify it as a confounder along with the proportion of the reversed subgroups which is 100%. This dataset contains 700 data rows with columns 'treatment', 'success', and 'stone_size'. X is set as 'treatment' and Y as 'success'.

Variable X	Variable Y	Correlation
treatment	success	-0.0574

Table 5. Correlation coefficient between treatment and success

Table 5 demonstrates the Pearson correlation index returned by the algorithm between 'treatment' and 'success' variables. Where 'treatment' contains values 0 and 1 (A and B respectively) and 'success' contains values 0 and 1.

stone_size	Variable X	Variable Y	Correlation
small	treatment	success	0.0400
large	treatment	success	0.0857

Table 6. Correlation coefficient between treatment and success for each subgroup

The algorithm traverses the list of all the potential confounders which is just one in this case. It identifies 'stone_size' as a confounder and the proportion of the subgroups with association reversal is 100%. As we can see from Table 6 correlation index between treatment and success is positive for both small and large stone sizes whereas it is negative for the whole population. The results correspond to the prior case studies on the treatment bias in Kidney Stone [23].

California DDS dataset

The California DDS dataset contains data regarding the allocation of financial resources from the Department of Developmental Services to the individuals in need in California for 2014. After analyzing the data retrieved from the governmental sources the lawsuit was filed against the California DDS claiming that the White Non-Hispanic population was allocated more resources than the Hispanic population (some citation). However, when we disaggregate by the age cohort we encounter a different situation. This is the exemplary case of the Simpson's paradox which demonstrates how frequent and vital it can be. The dataset is individual-oriented and contains 1000 data rows with the columns: 'Id', 'Age Cohort', 'Age', 'Gender', 'Expenditures', and 'Ethnicity'. We set *X* as 'Ethnicity' and *Y* as 'Expenditures'. Additionally as *X* variable contains more than 2 categories we set *x1* as 'White not Hispanic' and *x2* as 'Hispanic'.

Variable *X*	Variable *Y*	Correlation
Ethnicity	Expenditures	-0.3481

Table 7. Correlation coefficient between Ethnicity and Expenditures

Table 7 demonstrates the Pearson correlation index returned by the algorithm between 'Ethnicity' and 'Expenditures' variables. Where 'treatment' contains values 0 and 1 (A and B respectively) and 'success' contains values 0 and 1.

Age Cohort	Variable *X*	Variable *Y*	Correlation
0 to 5	Ethnicity	Expenditures	0.0207
6 to 12	Ethnicity	Expenditures	0.1473
13 to 17	Ethnicity	Expenditures	0.0252
18 to 21	Ethnicity	Expenditures	-0.0290
22 to 50	Ethnicity	Expenditures	0.0514
51+	Ethnicity	Expenditures	0.1876

Table 8. Correlation coefficient between Ethnicity and Expenditures for each subgroup

The algorithm traverses the list of all the potential confounders: 'Age Cohort' and 'Gender'. It identifies 'Age Cohort' as a confounding variable and the proportion of the subgroups with association reversal is 83%. As we can see from Table 8 correlation index between 'Ethnicity' and 'Expenditures' variables is positive for 5 out of 6 age cohorts and negative only for $18-21$ age cohorts whereas it is negative for the whole population. This is

a strong reversal effect and it proves the existence of the Simpson's paradox with the confounding variable 'Age Cohort' [24].

3.2 Linear Trends

Inputs to the algorithm: X - numerical(continuous) variable, Y - numerical(continuous) variable. Outputs: Z - confounding variable, p - proportion of the subgroups of Z which exhibit reversal of association between X and Y.

The algorithm for the detection of Simpson's Paradox in linear trends does not require a preprocessing step. All the steps are identical to the process described above. The first part of the algorithm is to calculate the correlation index between X and Y variables with the values of the corresponding columns in the dataset. This way we obtain the information on the sign of the association between these variables. Further, we traverse the list of remaining categorical variables, calculate the Pearson index conditioning on each subgroup(category), count the proportion of subgroups where the correlation index reversed relative to the index in aggregated data, and store the value key pairs in an array. We can further get the array element where the value (proportion) is the highest. A value greater than 0 implies the existence of Simpson's paradox and the maximal value of 1 implies a full reversal effect. Therefore the fact of Simpson's paradox occurrence can be decided based on the results of Algorithm 2.

Algorithm 2 Identification of the confounding variable for linear trends form of the Simpson's paradox

Input: dataset *D*, variable *x*, variable *y*
Output: a pair consisting of confounding variable and proportion of reversed association signs

```
aggreg_index = Pearson(d[x] , d[y]) // calculate corr. index between columns
indexes = [] // initialize index array to store key value pairs where the key is column and
 value is the number of reversed subgroups
 cols = columns(D) // initialize array of all columns of D
 foreach column ∈ cols do
    if Column Is Not Categorical(column) then
       Continue
    end
    else
       subgroups = Categories(column) // get the categories of a column
       coefficients = [] // initialize empty array to store the correlation indexes foreach
       subgroup ∈ subgroups do
          disaggreg_index = Pearson( D[x]: where D[column] = subgroup, D[y]: where
          D[column] = subgroup) calculate corr. index between columns for current
          subgroup


          Add index of disaggregated to correlation indexes array
       end
    end
    reversed_subgroups = proportionReversedSubgroups(aggreg_index, coefficients ) //
    calculate proportion of the correlation indexes reversed with respect to the correlation
    index for the aggregated data


    Add {column, reversed_subgroups} values into indexes
end
Store the max values of indexes pairs into result
 Return result
```

3.2.1 Experiments for Linear Trends

The experiments are done on the list of datasets that showcase bias in the form of the Simpson's paradox and where the corresponding *X* and *Y* variables are numerical (continuous). For this section Penguin dataset and Iris flower datasets are used. As in relative rates for each dataset, The author demonstrates the tables with Pearson correlation index between *X* and *Y* for the whole data and for each subgroup and explains how the algorithm operates on these tables.

Iris dataset

This is an extremely popular dataset that contains 150 instances of data for the iris flowers [25]. There are 5 attributes: sepal length, sepal width, petal length, petal width, and class. Simpson's paradox is manifested in multiple combinations of variables in this dataset, with the 'class' attribute being the confounder in all of the cases. For the sake of simplicity in the experiment, we will consider the 'sepal length' attribute as X variable and 'sepal width' as Y variable. When we take a look at Table 9 and Table 10 we see why the algorithm identified 'class' as a confounding variable. The correlation index sign reversed for all subgroups when conditioning by the 'class' variable. In Fig. 4 we can observe the visual representation of the reversal effect. Regression lines for subgroups have a positive slope, whereas the trend is negative for the whole data.

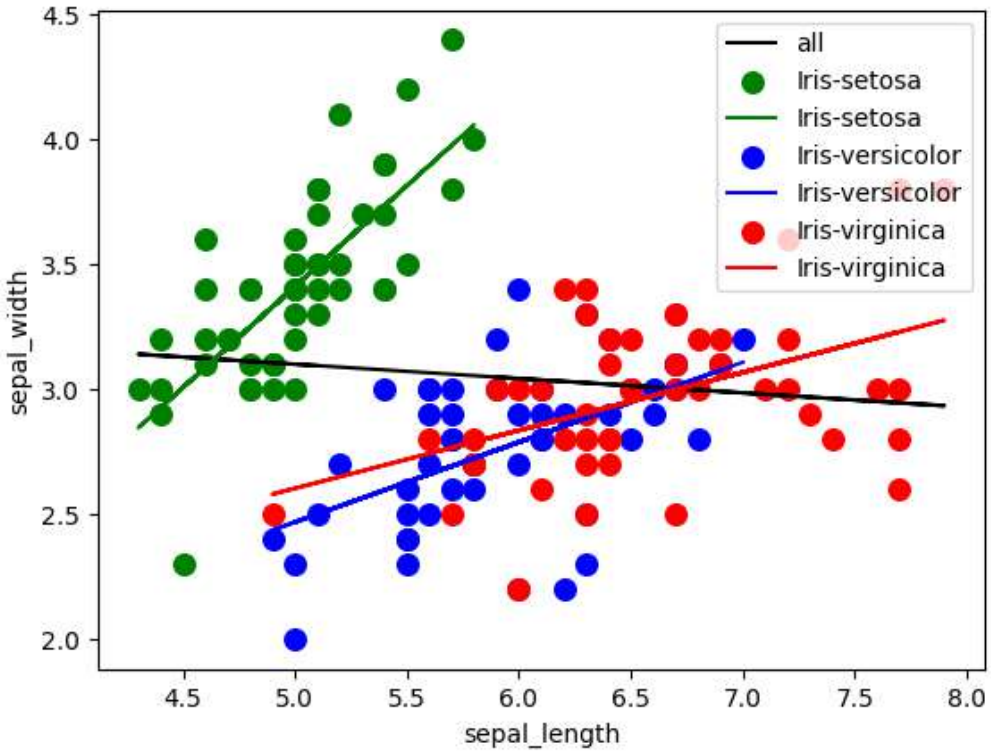

Figure 4. Scatter plot with regression lines for Iris dataset

Variable X	Variable Y	Correlation
sepal length	sepal width	-0.1093

Table 9. Correlation coefficient between sepal length and sepal width

Table 9 demonstrates the Pearson correlation index returned by the algorithm between 'sepal length' and 'sepal width' variables. Where both of the variables are continuous.

class	Variable X	Variable Y	Correlation
Iris-setosa	sepal length	sepal width	0.7467
Iris-versicolor	sepal length	sepal width	0.5259
Iris-virginica	sepal length	sepal width	0.4572

Table 10. Correlation coefficient between sepal length and sepal width for each subgroup

The algorithm traverses the list of all the potential confounders which is just one in this case. It identifies 'class' as a confounder and the proportion of the subgroups with association reversal is 100%. From Table 10 we can see that the correlation index between sepal length and sepal width is positive for all the classes whereas it is negative for the whole population [26].

Penguin dataset

Palmer Archipelago (Antarctica) penguin dataset is also a well-known dataset that was labelled as a replacement for the Iris dataset [27]. It is used for data exploration and visualization for beginners in the data field. The dataset contains the descriptions of 3 species of penguins. There is an instance of the Simpson's paradox in data as the association between the culmen length of the penguin and culmen depth reverses when data is disaggregated by the species. The dataset contains 344 data rows with columns: 'species', 'island' , 'culmen_length_mm', 'culmen_depth_mm', 'flipper_length_mm' , 'body_mass_g' and 'sex'. We set X as 'culmen_length_mm' and Y as 'culmen_depth_mm'. Similarly, here we can see the Simpson's paradox visually in Fig. 5. Regression line slopes are positive for each class and negative for the whole data.

Variable X	Variable Y	Correlation
culmen length	culmen depth	-0.1093

Table 11. Correlation coefficient between culmen length and culmen depth

Table 11 demonstrates the Pearson correlation index returned by the algorithm between "culmen_length_mm' and 'culmen_depth_mm' variables. Where both of the variables are continuous.

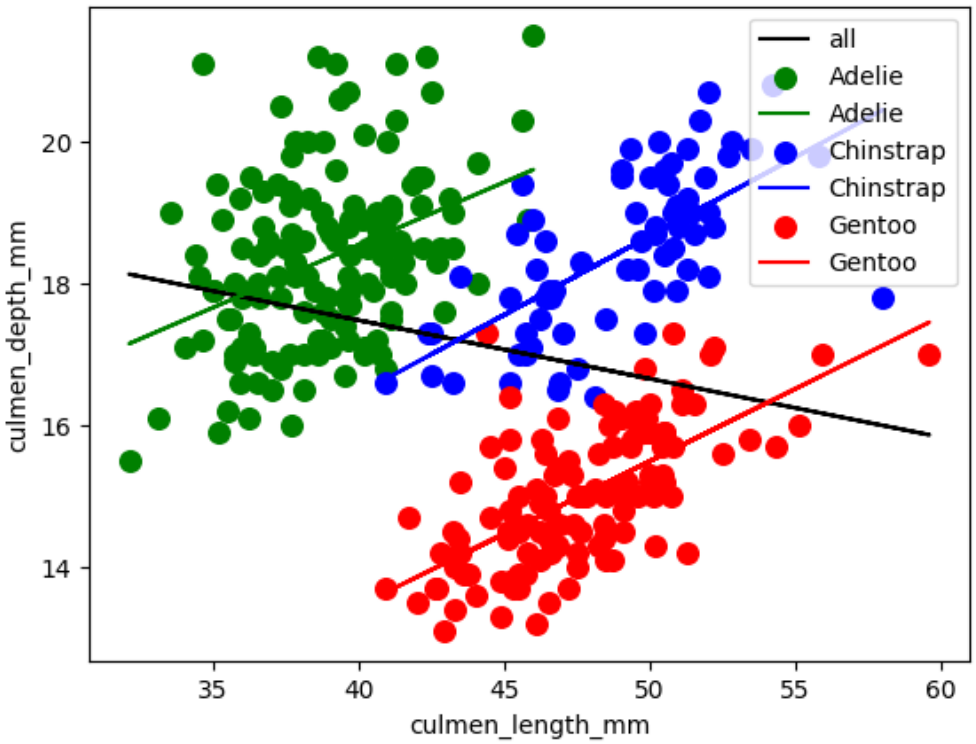

Figure 5. Scatter plot with regression lines for Penguin dataset

class	Variable X	Variable Y	Correlation
Torgersen	culmen length	culmen depth	0.3858
Biscoe	culmen length	culmen depth	0.6535
Dream	culmen length	culmen depth	0.6440

Table 12. Correlation coefficient between culmen length and culmen depth for each subgroup

The algorithm traverses the list of all the potential confounders: 'species', 'island', and 'sex'. It identifies 'species' as a confounder and the proportion of the subgroups with association reversal is 100%. From Table 12 we can see the correlation index between culmen length and culmen depth is positive for all the species whereas it is negative for the whole population [28].

4. Resolving the Simpson's Paradox

The existence of the Simpson's paradox can cause erroneous conclusions during the statistical(data) analysis, therefore it is important to be able to solve it. Solving the paradox can mean different things. In this thesis, we consider the explanation of the paradox and the adjustment of the data for the correct decision-making as a solution. In this section, we concentrate on the Simpson's paradox cases which exist in the relative rates form. The main cause of the paradox is the effect of the confounding variable on both X and Y. The main intuition is that if we eliminate the effect of the confounding variable on X the confusing manifestation of the paradox observed while aggregating the data can be avoided. As we can see from Fig. 6 the paradox occurs because the confounding variable Z has a causal relationship both with X and Y.

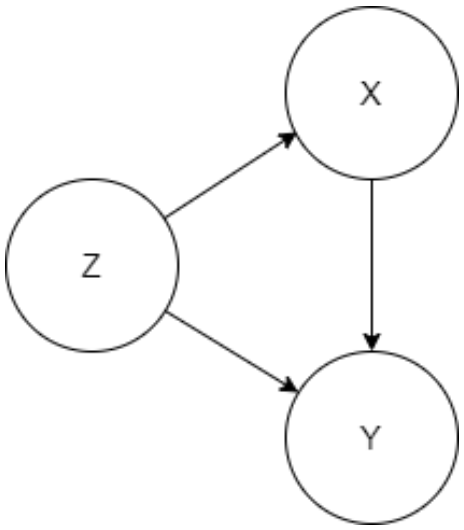

Figure 6. Causal graph with the confounding effect

This confounding effect translates to the uneven distribution of the data instances. So in order to be able to solve the paradox we need to fix the distributions. To make an analogy, as with the Berkeley university admission rates example where 'gender' is X variable, 'admitted' variable and 'major' is the confounding (Z) variable, the distribution of males and females in the subgroups of the 'major' is uneven and is caused by the preferences of the particular gender to apply to the certain majors based on the certain features such as the

competitiveness of the major. So the key to the paradox resolution is an adjustment of the distributions to get rid of the confounding effect and achieve the causal inference system as illustrated in Fig. 7. However, this method does not apply to certain cases of Simpson's paradox. Situations where confounding variables are affected by the X variable and affect Y itself as shown in Fig. 8 require a different approach. Elimination of the confounding effect would also partially eliminate the indirect effect of X on Y. This is impractical as we are interested in a full effect.

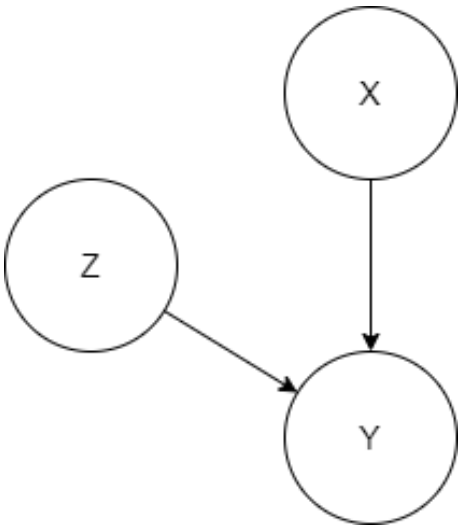

Figure 7. Causal graph without the confounding effect

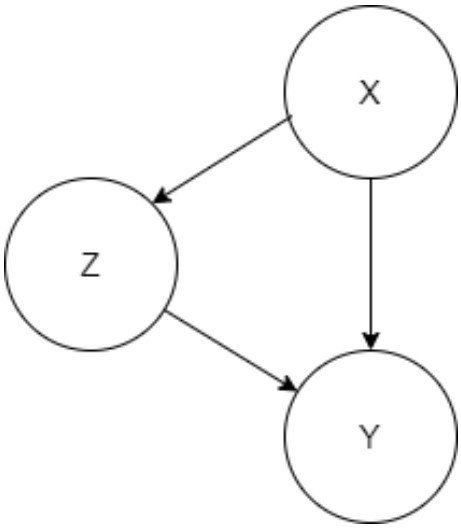

Figure 8. Causal graph with indirect effect of X on Y

4.1 Inverse Probability Weighting

Inverse probability (propensity) weighting is the method to balance the distribution of the data instances in the subgroups [29]. IPW is used in the datasets with the lack of randomness in the observations or with the lack of information caused by some bias. In the case of observations collected for the datasets being performed in a non-randomized manner, the distributions in the subgroups of certain variables will be unbalanced. In most cases this happens due to the confounding effect which can be explained by human factors in the process.

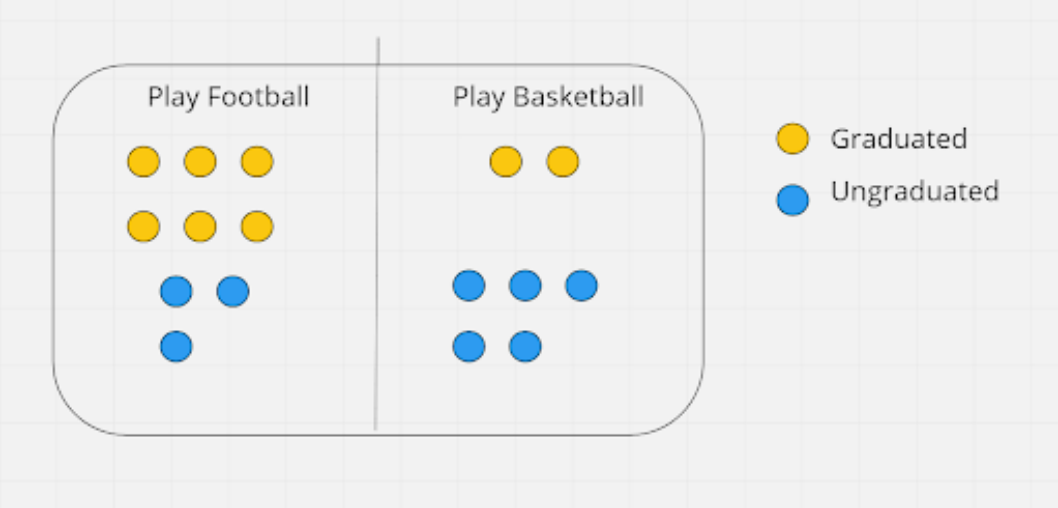

Figure 9. Example of unbalanced subgroups

Fig. 9 is an example case of an unbalanced distribution of data instances in the subgroups. So in order to study the causal effect of the type of sport played on some other variable, we should first fix the distributions with respect to the graduation status, as the graduation status might as well have an effect on that variable. This scenario is illustrated in Fig. 7 and implies the potential existence of Simpson's paradox.

IPW is used to correct the observations during the data analysis. The main intuition behind IPW is that information on probabilities in subpopulations is sufficient to make inferences for the whole population. It works by assigning the inverse propensity as a weight in the aggregation formula. If we denote x as subgroup of X variable and Z_a as all potential confounding variables then the propensity score is $P(x \mid Z_a)$ and the weight is $1/P(x \mid Z_a)$. In the case of Z_a being high dimensional, we have to apply logistic regression or other appropriate machine learning methods to reduce the dimensionality to the level of the probability scalar. As we particularly want to target the individual confounding variable identified in the previous steps of the process we can simplify the propensity score calculation by just considering one confounder. Propensity score in this case will be $P(x \mid Z)$ and the weight $1/P(x \mid Z)$ where Z is a single confounding variable.

$$P(Y|X=x) = \frac{\sum\left(\frac{P(Y,X=x,Z=z)}{P(X=x|Z)}\right)}{\sum\left(\frac{1}{P(X=x|Z)}\right)}$$

The standard formula will be substituted with the formula illustrated above. It will be used in the aggregation step and will be calculated for each value of variable *X*. As we are working with the datasets, the probabilities will be expressed as a fraction of the counts. As we are dealing with a single confounding variable, the propensity scores can be calculated and assigned for each subgroup of that variable as opposed to doing it for each data instance. The detailed steps are described in Algorithm 3. The result of the algorithm is an aggregated data table showing the average rate/amount of *Y* for each category of *X*.

Algorithm 3 Resolving the Simpson's paradox by aggregating the data with the adjustment formula

Input: dataset: *D*, variable: *x*, variable: *y* , confounding variable: *conf*
Output: table
Set *disaggreg_data* with a mean value y of dataset *D* grouping by *{x, conf}*
Set *count_aggreg* with number of values of dataset *D* grouping by *{x}*
Set *count_disaggreg* with number of values of dataset *D* grouping by *{x, conf}*
foreach $(i, row) \in disaggreg\_data$ **do**
    adj = (count_aggreg[y]: where count_aggreg[x]=row[x]) / count_disaggreg[i][y]
    val = row[y] / count_disaggreg[i][y] * (count_aggreg[y]: where count_aggreg[x]=row[x])
    disaggreg_data[i][y] = val
    disaggreg_data[i]['adj'] = adj
end
Set *adjusted_aggreg* with a sum of y values of dataframe *disaggreg_data* grouping by *{x}*
Set *denominator* with a sum of 'adj' column values of dataframe *disaggreg_data* grouping by *{x}*
foreach $(i, row) \in adjusted\_aggreg$ **do**
    val = adjusted_aggreg[i][y] / denominator[i]['adj']
    adjusted_aggreg[i][y] = val
end
Return *adjusted_aggreg*

4.2 Experiments

We use the same dataset we used in the Simpson's paradox detection and confounder identification section for the relative rates form namely: Kidney stone dataset, Berkeley university admissions dataset, and California DDS dataset. Confounding variables had

been identified for each dataset in the previous part of the process. Aggregated form, a disaggregated form of data, and a form aggregated using the adjustment formula are provided for each dataset. Adjusted tables that are considered a resolution to the paradox are based on the pseudo-population generated from the original population by fixing the distributions.

Berkeley University Admissions Dataset

In the previous sections, the algorithm identified the ‘major’ variable as a confounding variable. As described above, the confounding effect is caused by the non-randomness with respect to the different genders applying to different majors. In Fig. 10 we can see that distributions are uneven in the subgroups of ’major’ If we eliminate the effect of ‘major’ on the ‘gender’ variable we can solve the Simpson’s paradox. The average admission rate for each gender calculated in a standard way is illustrated in Table 13, while the same data conditioned by ’major’ is illustrated in Table 14. The average admission rate for each gender calculated with the probabilistically adjusted formula is illustrated in Table 15.

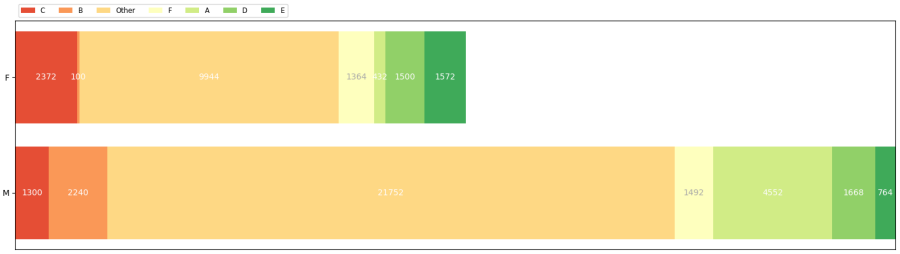

Figure 10. Data Distribution in subgroups of ’major’

Gender	Admission
F	0.3458
M	0.4428

Table 13. Aggregated data with standard formula for Gender and Admission

Gender	Major	Admission
F	A	0.8241
	B	0.6800
	C	0.3390
	D	0.3493
	E	0.2392
	F	0.0733
	Other	0.3769
M	A	0.7250
	B	0.6304
	C	0.3692
	D	0.3309
	E	0.2775
	F	0.0590
	Other	0.4095

Table 14. Disaggregated data conditioned by Major

Gender	Admission
F	0.6271
M	0.3312

Table 15. Aggregated data with adjustment formula for Gender and Admission

Kidney Stone Dataset

In the previous sections, the algorithm identified the 'stone_size' variable as a confounding variable. This implies an imbalance of distributions(Fig. 11) with respect to certain subgroups of 'stone_size' being associated with certain 'treatment' subgroups which are caused by human factors in the process. Indeed doctors tended to allocate patients with large stone sizes to treatment A due to the doctor's understanding that treatment A performs better for large stones and treatment B performs better for small stones. So we aim to eradicate that effect in order to be able to observe the trend for the aggregated data based only on treatment effectiveness. The average success rate for each treatment calculated in a standard way is illustrated in Table 16, while the same data conditioned by 'stone size' is illustrated in Table 17. The average success rate for each treatment calculated with the probabilistically adjusted formula is illustrated in Table 18.

Figure 11. Data Distribution in subgroups of 'stone size'

Treatment	Success
B	0.8257
A	0.78

Table 16. Aggregated data with standard formula for Treatment and Success

Treatment	stone size	Success
B	large	0.6875
	small	0.8666
A	large	0.7300
	small	0.9310

Table 17. Disaggregated data conditioned by Stone Size

Treatment	Success
B	0.7284
A	0.8810

Table 18. Aggregated data with adjustment formula for Treatment and Success

California DDS Dataset

In the previous sections, the algorithm identified the 'age_cohort' variable as a confounding variable. Even though association reversal was not for all the subgroups of 'age_cohort' the proportion of reversed associations was the maximal among the set of variables. The main reason for the paradox is the uneven distribution of individuals with different ethnicities in different age cohorts (Fig. 12). Indeed people of Hispanic ethnicity tend to have a higher density of young population and a lower density of old population. The average expenditures amount for each ethnicity calculated in a standard way is illustrated in Table 19, while the same data conditioned by 'Age Cohort' is illustrated in Table 20.

The average expenditures amount for each ethnicity calculated with the probabilistically adjusted formula is illustrated in Table 21.

Figure 12. Data Distribution in subgroups of 'Age Cohort'

Ethnicity	Expenditures
White not Hispanic	24697.5486
Hispanic	11065.5691

Table 19. Aggregated data with standard formula for Ethnicity and Expenditures

Ethnicity	Age Cohort	Expenditures
White not Hispanic	0 to 5	1366.9
	6 to 12	2052.2608
	13 to 17	3904.3582
	18 to 21	10133.0579
	22 to 50	40187.6240
	51+	52670.4242
Hispanic	0 to 5	1393.2045
	6 to 12	2312.1868
	13 to 17	3955.2815
	18 to 21	9959.8461
	22 to 50	40924.1162
	51+	55585

Table 20. Disaggregated data conditioned by Age Cohort

Ethnicity	Expenditures
White not Hispanic	11453.7385
Hispanic	32131.5521

Table 21. Aggregated data with adjustment formula for Ethnicity and Expenditures

5. Web Platform

The author of this thesis built the website which allows users to explore the Simpson's paradox on their datasets [30] [31]. It combines both detection and solution parts into a single unit aiming to simplify the process of data exploration with the objective of tackling the Simpson's Paradox. Service is written on top of the algorithms described in Part 3 and Part 4 (Algorithms 1, 2, 3) of the thesis. The visual interface requires users to import the dataset on where they intend to perform analysis on. As soon as the file is imported users get the possibility to select the x and y variables from the dropdown list which contains all attributes contained in the dataset file.They can further select *x1* and *x2* variable values from the dropdown list, in case x is categorical and the number of unique values of x is greater than 2. After all the input parameters are ready, users can press the 'Show' button. First, the dataset is preprocessed and passed to the function which detects the presence of Simpson's paradox and identifies the confounding variable. In the next step, the confounding variable along with the preprocessed dataset is passed to the function which builds the analysis tables for the data. The main part of the results is the statement that indicates whether the Simpson's paradox has been detected, the confounding variable identified by the backend functions, and the list of categories of the confounding variable which exhibit association reversals, which is outputted both for linear trends and relative rates form. In the case of the Simpson's paradox existing in the relative rates form, analysis tables as described in the previous section consisting of the data aggregated with the standard formula, disaggregated data conditioned by the confounding variable, data aggregated with the adjusted formula, and the distribution plot are displayed in the interface. In the case of the Simpson's paradox existing in the form of linear trends, the scatter plot which usually explains the reversal effect is displayed in the interface.

5.1 Brief Manual

- **Step 1:** *"Upload the CSV file and wait until the input fields are activated (Fig. 13)."*
- **Step 2:** *"Select the input values for x and y in the dropdown list of the input fields (Fig. 14, 16)"*
- **Step 3:** *"In case x is categorical variables with more than 2 categories, select x1 and x2 from the dropdown list of values of x (Fig. 16)"*
- **Step 4:** *"Press "Show" button and wait until the results are outputted (Fig. 15*

, 17, 18, 19). Error notification is displayed in case of failure."

Figure 13. Index Page

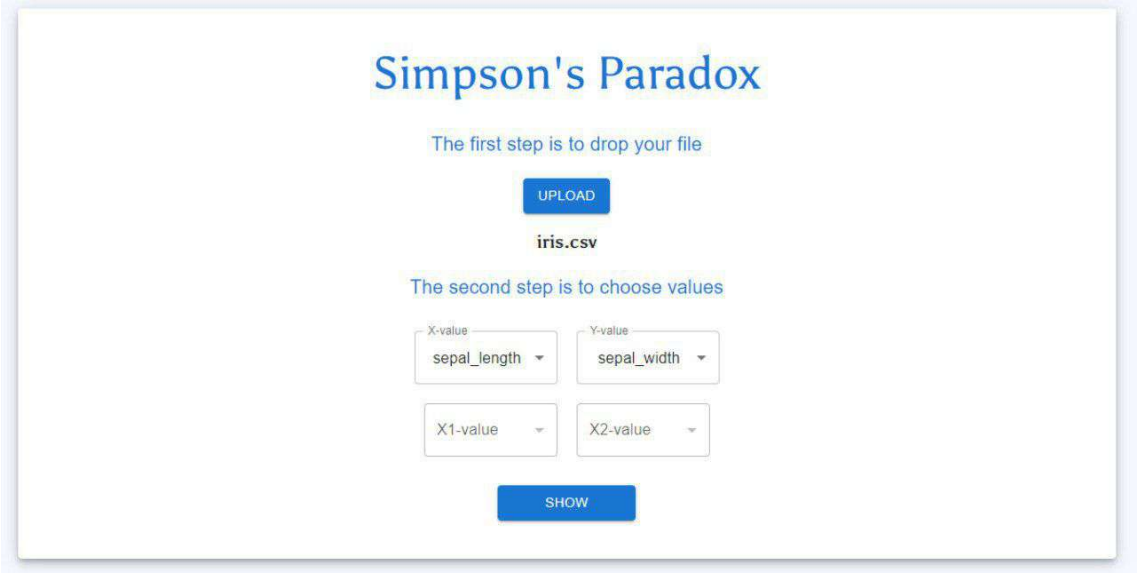

Figure 14. Index Page with selected values for Linear Trends form

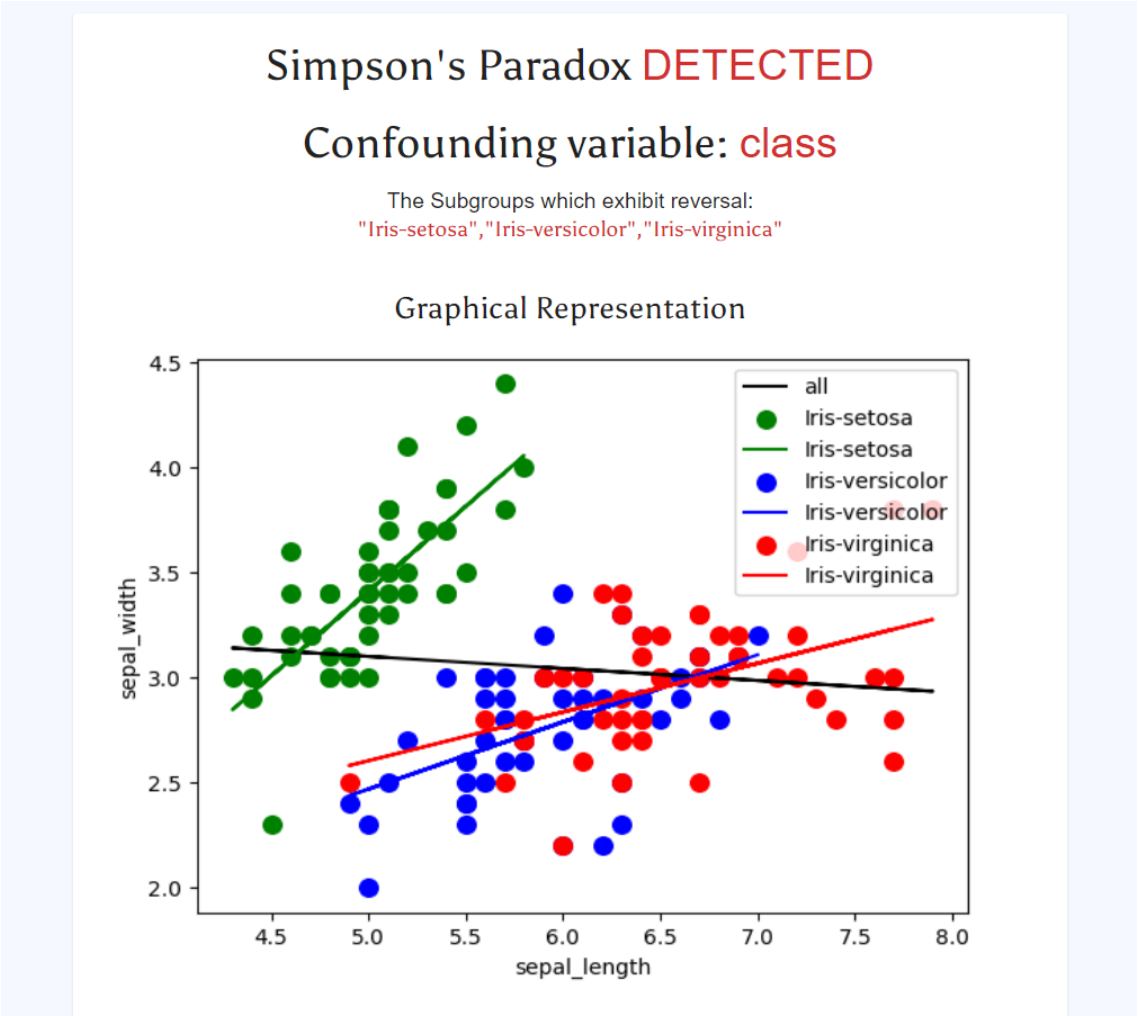

Figure 15. Example output for Linear Trends form

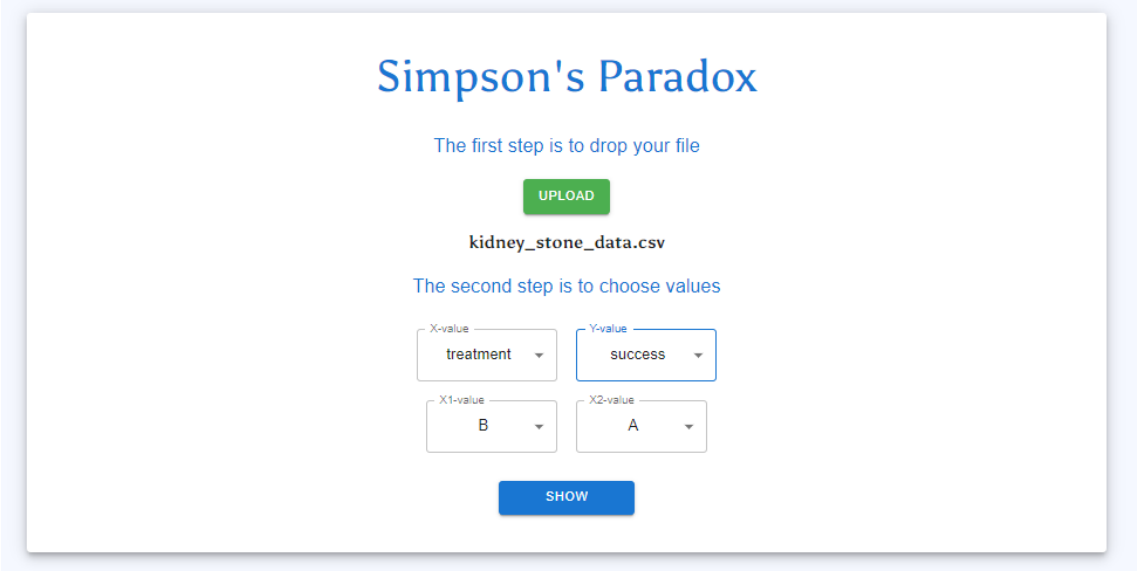

Figure 16. Index Page with selected values for Relative Rates form

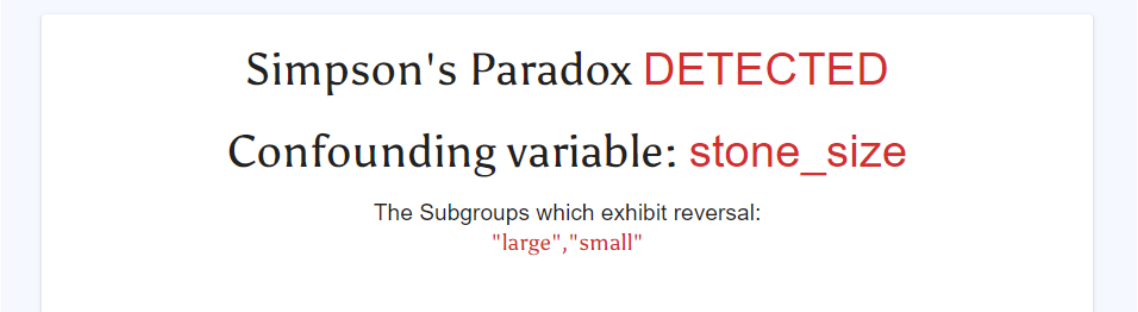

Figure 17. First part of example output for Relative Rates form

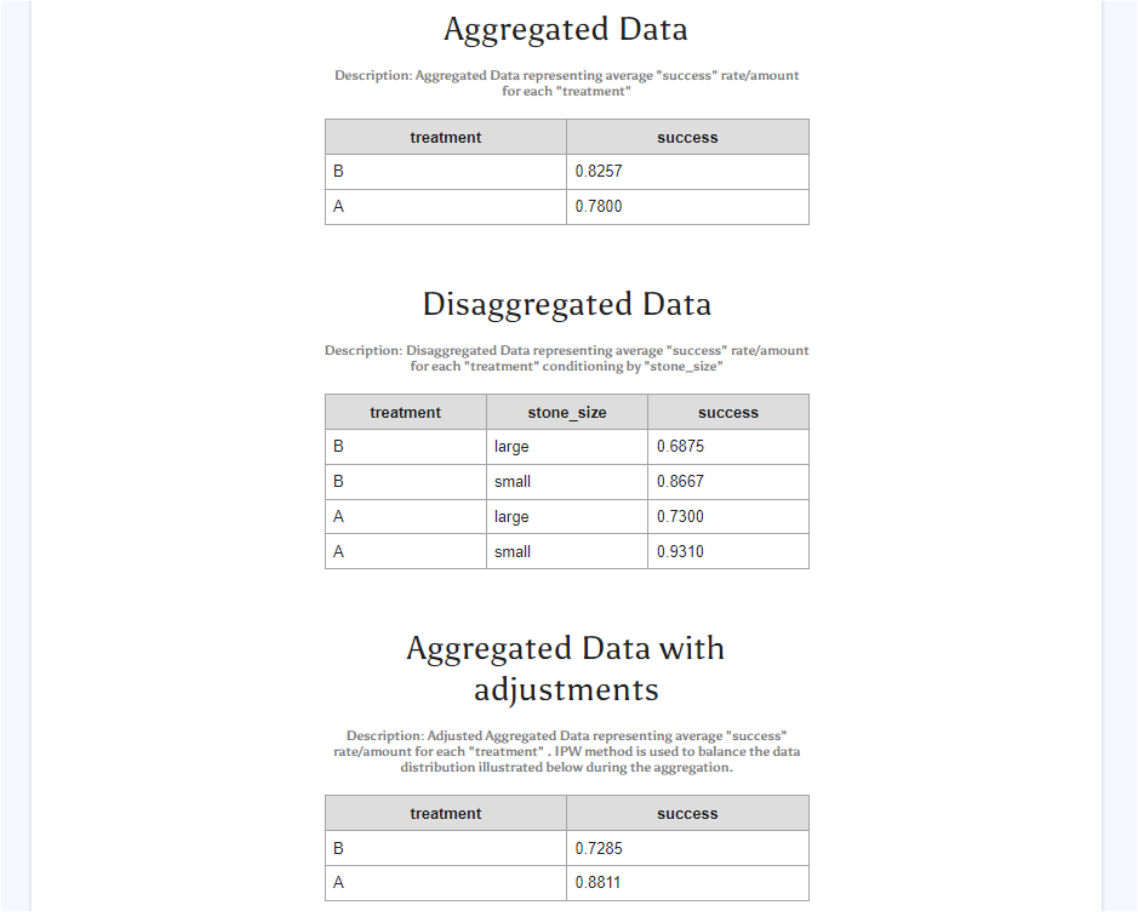

Figure 18. Second part of example output for Relative Rates form

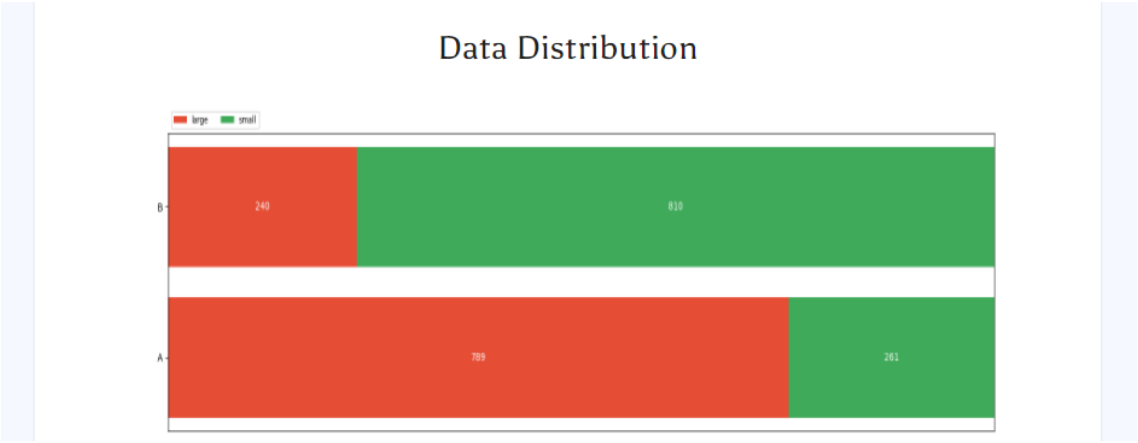

Figure 19. Third part of example output for Relative Rates form

5.2 Architecture

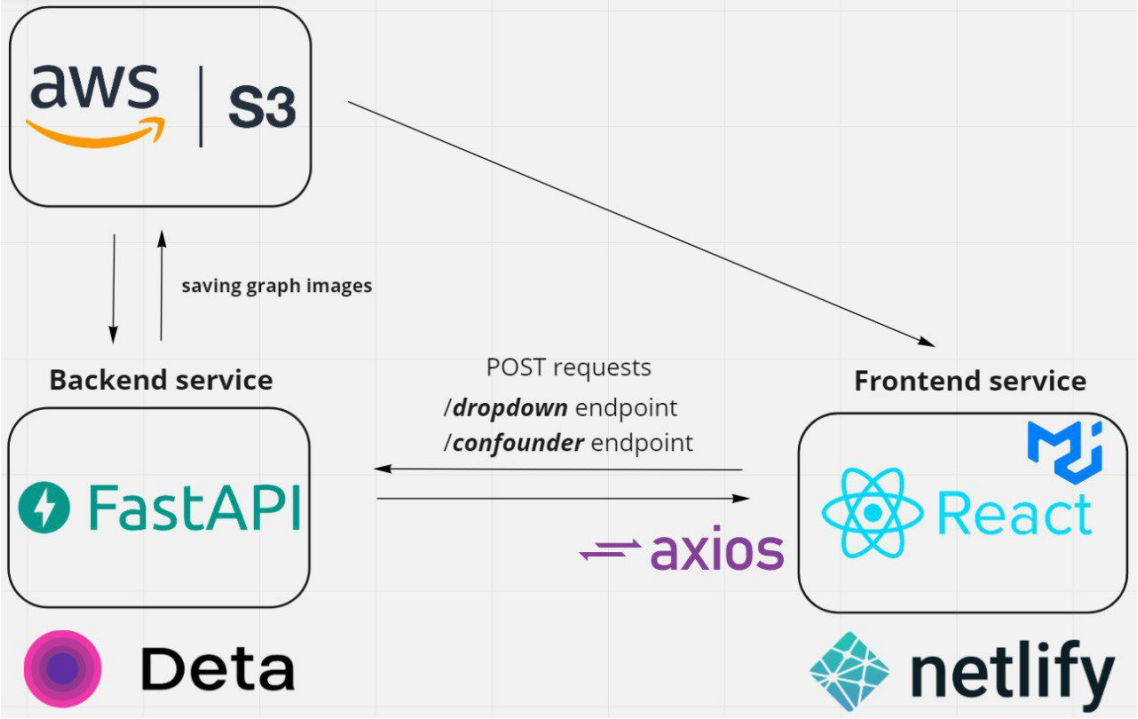

Figure 20. Architecture of the Web Platform

When a user uploads a dataset file to the respective input field in the interface request to /dropdown is made with the uploaded file and the list of attributes and values is returned as a response. The response consists of all attributes which are used as drop-down options in the input fields for *x* and *y*, and a list of values for each categorical attribute that is used as drop-down options for *x1* and *x2* in case attribute selected as x has more than 2 values(categories). When all the values are selected and the "Show" button is pressed, another request is made to /confounder endpoint. At this step, the inputs are fed to Algorithm 1 which takes care of the identification of the confounding variable and detection of SP. Output results are saved in memory and passed to the subsequent functions. In the case of the paradox existing in a relative rates form all the parameters including the confounding variable identified in the previous step are passed to Algorithm 3 which takes care of the building the data table with probabilistic adjustments and the output is once again saved in memory, In parallel, the same parameters are fed to the function which builds the plot demonstrating the distributions of data instances in the subgroups of the confounding variable and the image is uploaded to the AWS S3 bucket through the API Gateway provided by AWS with URL of the image being saved in memory. All the collected output is returned to the front-end side to be parsed. In the case of the paradox existing in the form of a linear trend, only a scatter plot is generated and uploaded to the AWS S3 bucket, in the same way, described previously and all the output is returned to the front-end side.

5.2.1 Backend

The code is written in Python 3.8 programming language. The framework used for the development is FastApi. The main advantage of FastApi is as its name says fast development and high performance. Another upside of this framework is that it is possible to quickly build small APIs and it requires less memory space which can be problematic in the deployment process Application contains 2 endpoints “confounder/” and “dropdown/” where the post requests with the corresponding parameters are made. Parameters are taken from the form submitted by the user in the web interface and posted as a request to the endpoint. The response of the request is the JSON with the required information used in the frontend. Backend service is deployed on the Deta platform. This is a very convenient platform for the deployment of microservices which eliminates the necessity for server configuration and permission management. Deta CLI is used for the deployment and management of API. Downsides are the limitations on the size of an app for the free tier and the possibility to deploy only FastApi and Nodejs apps.

5.2.2 Frontend

Frontend service takes care of the communication with the backend API service and renders the components in the web interface using the data response from the POST request to the "/confounder" and "/dropdown" endpoints. It is written in React Js which enables more functionality in web development than just Javascript and makes the coding easier. Another major advantage of React is the possibility to use a wide range of libraries such as Material-UI which was used in the development of the user interface for this application. This framework provides a huge set of crisp designs to build a visually appealing interface by using just a few lines of code. For the communication with the backend server, we used the Axios library which allows us to conveniently fetch data from external sources. This can be achieved with fetch and AJAX methods but Axios provides more functionality and a level of security. The service was deployed in Github Pages which is free, easy to set up, and allows collaboration with Git and Github. Additionally, it provides a possibility for live updating with Github Workflow which is ideal for continuous development and continuous integration.

5.3 Software Development

For the development of this project, the author used the Agile project management methodology. As the author was working alone, standard Scrum was not applicable so Scrum for One was used. The work was split into 3 sprints and each sprint lasted for about 1-2 weeks

on average. At the beginning of each sprint, The author made the planning and updated the project backlog. Additionally, daily reviews reflecting on work done and the way forward were conducted. 3 user stories were used as the core requirements and each sprint covered all the user stories to some extent to maintain the incremental development process. In Sprint 1 the first version of the application with minimal functionality was developed. In the second sprint, all the parts were extended and enhanced and the functions responsible for the scatter plot generation were written and integration with AWS S3 was made. In the final sprint, everything was revised and improved once again. Distribution plot generation functions were developed and the functionality for drop-down options in the interface was added. Fig. 21 represents the software development plan which was made in the beginning of the process and updated further on.

- **User Story 1:** *"As a user, I want to be able to upload my CSV file and select the corresponding attributes with the drop-down options, so that I submit my inputs conveniently."*
- **User Story 2:** *"As a user, I want to be able to see the structured results for the detection of the Simpson's paradox, so that I can see whether Simpson's Paradox is detected in the uploaded dataset and if detected what is the confounding variable."*
- **User Story 3:** *"As a user, I want to be able to see the structured results for the resolution and explanation of the paradox, so that I can understand why it happens and use the adjusted data for the decision-making."*

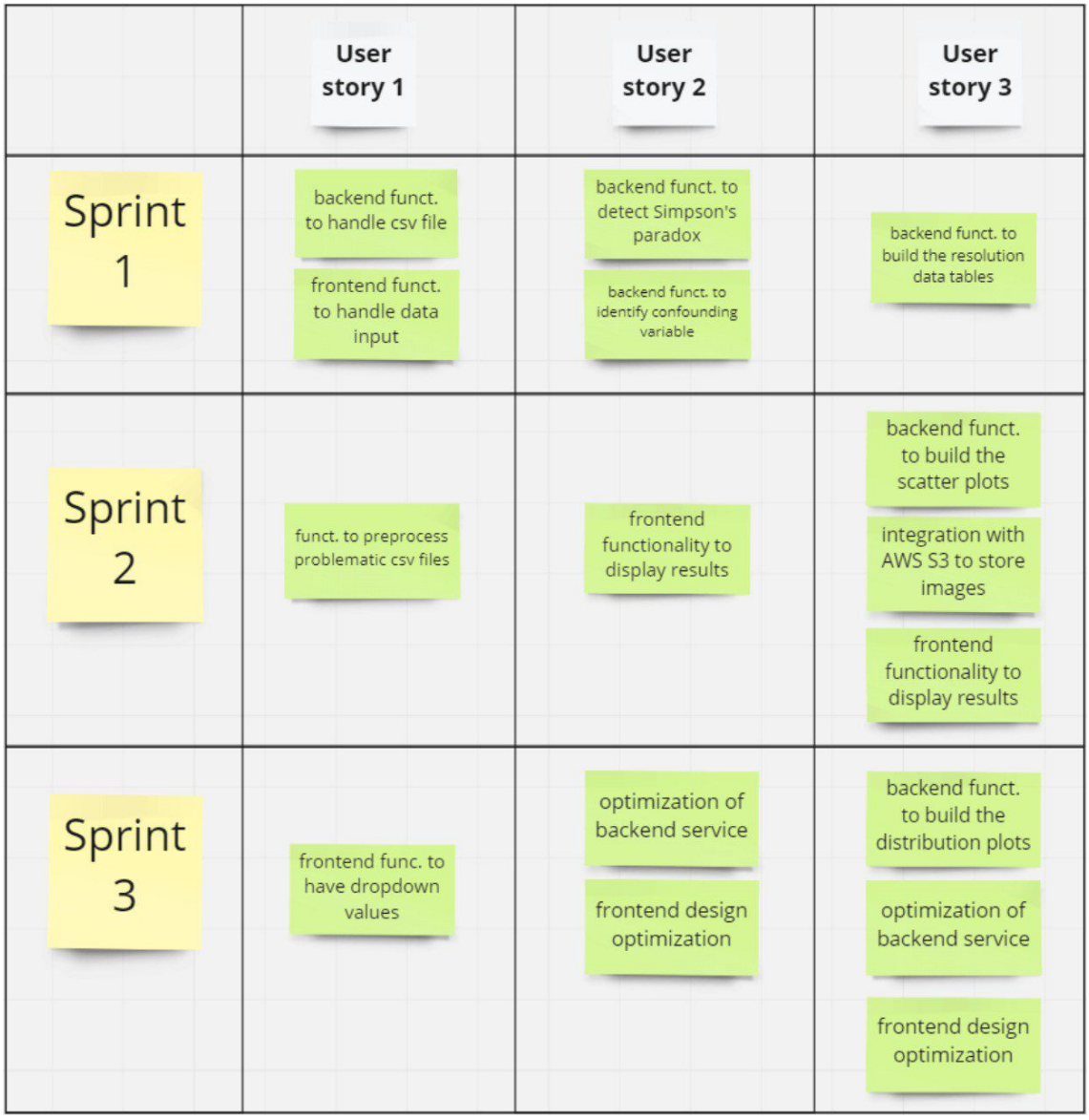

Figure 21. Software development plan

6. Time Evaluations

Time evaluations are performed for each of the algorithms with the same parameters as described in the corresponding Experiments sections. The code is developed in Python 3.8.10 and runs on HP EliteBook 840 G6 with a 1.6GHz Intel Core i5 CPU and 16 GB 1800MHz RAM. All the algorithms replicate the same process described in the experiments sections of the thesis. Algorithms 1 and 2 are for the identification of the confounding variables and Algorithm 3 is for the resolution of the Simpson's Paradox with the probabilistic adjustments. From Table 22 we can see that time strongly correlates with the size of the table. Algorithm 2 is the fastest and Algorithm 3 is the slowest when comparing the time efficiency.

Algorithm	Name	Size	Time
Algorithm 1	Berkeley University Admission	268 kb	0.037 s
	Kidney Stone	7 kb	0.011 s
	California DDS	41 kb	0.018 s
Algorithm 2	Iris	4.5 kb	0.007 s
	Penguin	13 kb	0.02 s
Algorithm 3	Berkeley University Admission	268 kb	0.129 s
	Kidney Stone	7 kb	0.089 s
	California DDS	41 kb	0.109 s

Table 22. Time evaluations

7. Conclusion

The confusion caused by the Simpson's Paradox can turn out to be detrimental in certain cases. The author of this thesis described some historical examples where the understanding of the paradox and its resolution were instrumental. Additionally, the author talked about the methods and tools developed previously that serve the purpose of helping understand or tackle Simpson's paradox. The author found the necessity to improve certain methods and develop a platform that unifies most of the aspects related to the Simpson's Paradox.

In this thesis, the author showed how the Web application for Simpson's paradox exploration in the datasets was built and demonstrated the methods behind the service. First, the author demonstrated the method to identify the confounding variables in the dataset and decide the existence of the Simpson's paradox based on that. This method incorporates preprocessing techniques that allow flexibility in terms of data type variety for datasets, and comparisons of correlation coefficients which uncover the instances of association reversals. It is a very compact algorithm and is applicable to all forms of the Simpson's Paradox. Next, the author explored the algorithm which aggregates the data for X and Y variables using the probabilistic adjustments, which eliminates the confounding effect, therefore solving the paradox. In the final part, the overview and architecture of the developed Web Platform were presented.

All the methods were tested using well-known datasets that exhibit the Simpson's paradox, and the results obtained from the experiments correspond to prior knowledge. Additionally, the processing time of the algorithms was measured for each dataset in correspondence with the conducted experiments and it showed that all the algorithms are relatively efficient. All the experiments and tests conducted for the algorithms were conducted for the whole Web Service as well and the results matched. Service proves to be functional and effective in terms of data exploration, with a focus on the Simpson's paradox, however, there are certain limitations that open a way for further development and research. One potential improvement is that continuous variables can be converted to categorical by dividing the set of values into bins. This would allow us to uncover instances of SP which otherwise wouldn't be possible. Another direction is to alter the SP detection algorithm to omit X variable in the input and search for the pairs (X,Z) instead of just Z. Additionally, the web service components and architecture can be optimized for better performance and speed.

Bibliography

[1] Edward H Simpson. “The interpretation of interaction in contingency tables”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 13.2 (1951), pp. 238–241.

[2] Julius von Kügelgen, Luigi Gresele, and Bernhard Scholkopf. “Simpson’s Paradox in COVID-19 Case Fatality Rates: A Mediation Analysis of Age-Related Causal Effects”. In: *Ieee Transactions on Artificial Intelligence* 2 (2021), pp. 18–27.

[3] Nazanin Alipourfard, Peter G Fennell, and Kristina Lerman. “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*. 2018, pp. 19–27.

[4] Chenguang Xu, Sarah M Brown, and Christan Grant. “Detecting Simpson’s paradox”. In: *The Thirty-First International Flairs Conference*. 2018.

[5] Rahul Sharma et al. “Existence of the Yule-Simpson Effect: An Experiment with Continuous Data”. In: *2022 12th International Conference on Cloud Computing, Data Science Engineering (Confluence)*. 2022, pp. 351–355. DOI: 10.1109/Confluence52989.2022.9734211.

[6] R Cohen Morris, Nage Ernest, et al. “An Introduction to logic and scientific method”. In: *The Journal of Nervous and Mental Disease* 80.4 (1934), pp. 495–496.

[7] David Lindley and Melvin R. Novick. “The Role of Exchangeability in Inference”. In: *Annals of Statistics* 9 (1981), pp. 45–58.

[8] Alessandro Selvitella. “The ubiquity of the Simpson’s Paradox”. In: *Journal of Statistical Distributions and Applications* 4 (2017), pp. 1–16.

[9] Alexander E Gorbalenya et al. “Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. The species severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2”. In: *Nat. Microbiol* 5.4 (2020), pp. 536–544.

[10] Alex A Freitas. “On objective measures of rule surprisingness”. In: *European Symposium on Principles of Data Mining and Knowledge Discovery*. Springer. 1998, pp. 1–9.

[11] *R package. Simpsons: Detecting Simpson's Paradox*. 2018. URL: https://rdrr.io/cran/Simpsons/man/Simpsons.html.

[12] *SimpsonsParadox: Automatic Simpson's Paradox Detector*. 2020. URL: https://github.com/ehart-altair/SimpsonsParadox.

[13] Galit Shmueli and Inbal Yahav. "The forest or the trees? Tackling Simpson's paradox with classification trees". In: *Production and Operations Management* 27.4 (2018), pp. 696–716.

[14] Torsten Hothorn, Kurt Hornik, and Achim Zeileis. "Unbiased recursive partitioning: A conditional inference framework". In: *Journal of Computational and Graphical statistics* 15.3 (2006), pp. 651–674.

[15] Ninareh Mehrabi et al. "A survey on bias and fairness in machine learning". In: *ACM Computing Surveys (CSUR)* 54.6 (2021), pp. 1–35.

[16] Judea Pearl. "Understanding Simpson's Paradox". In: *SSRN Electronic Journal* 68 (Jan. 2013). DOI: 10.2139/ssrn.2343788.

[17] Michael I Jordan. "Graphical models". In: *Statistical science* 19.1 (2004), pp. 140–155.

[18] Norman Fenton, Martin Neil, and Anthony Constantinou. "Simpson's Paradox and the implications for medical trials". In: *arXiv preprint arXiv:1912.01422* (2019).

[19] Judea Pearl et al. "Models, reasoning and inference". In: *Cambridge, UK: CambridgeUniversityPress* 19 (2000), p. 2.

[20] Judea Pearl. "Causal diagrams for empirical research". In: *Biometrika* 82.4 (1995), pp. 669–688.

[21] Edwin Martens. *Methods to adjust for confounding. Propensity scores and instrumental variables*. 2007.

[22] Kevin H Chu et al. "Simpson's paradox: A statistician's case study". In: *Emergency Medicine Australasia* 30.3 (2018), pp. 431–433.

[23] Ned Kock and Leebrian Gaskins. "Simpson's paradox, moderation and the emergence of quadratic relationships in path models: an information systems illustration". In: *International Journal of Applied Nonlinear Science* 2.3 (2016), pp. 200–234.

[24] Stanley A Taylor and Amy E Mickel. "Simpson's paradox: A data set and discrimination case study exercise". In: *Journal of Statistics Education* 22.1 (2014).

[25] Ronald A Fisher. "The use of multiple measurements in taxonomic problems". In: *Annals of eugenics* 7.2 (1936), pp. 179–188.

[26] Dirk Draheim. "Why Not to Trust Big Data: Discussing Statistical Paradoxes". In: ().

[27] Allison Marie Horst, Alison Presmanes Hill, and Kristen B Gorman. “palmerpenguins: Palmer Archipelago (Antarctica) penguin data”. In: *R package version 0.1. 0* (2020).

[28] Jeff Shamp. *Simpson’s Paradox*. 2020. URL: https://rpubs.com/shampjeff/blog_post_2 (visited on 08/31/2020).

[29] Peter C Austin. “An introduction to propensity score methods for reducing the effects of confounding in observational studies”. In: *Multivariate behavioral research* 46.3 (2011), pp. 399–424.

[30] *SimpsonsParadox. Website for SP exploration*. 2022. URL: https://simpsonparadox.netlify.app/.

[31] *SimpsonsParadox. Repository for Website*. 2022. URL: https://github.com/garhus2020/SimpsonP.

Appendices

Appendix 1. Non-exclusive licence for reproduction and publication of a graduation thesis

I Huseyn Garayev

1. grant Tallinn University of Technology free licence (non-exclusive licence) for my thesis "A WEB-BASED PLATFORM FOR DETECTING AND HANDLING THE SIMPSON'S PARADOX", supervised by Rahul Sharma

1.1 to be reproduced for the purposes of preservation and electronic publication of the graduation thesis, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright;

1.2 to be published via the web of Tallinn University of Technology, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright.

2. I am aware that the author also retains the rights specified in clause 1 of the non-exclusive licence.

3. I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights, the rights arising from the Personal Data Protection Act or rights arising from other legislation.

08.05.2022