KNOWLEDGE & TECHNOLOGY

Bangladesh Army University of Engineering & Technology

Department of Computer Science and Engineering

A thesis report on

Outlier Detection in Data Mining

A thesis is submitted in partial fulfillment of the requirements of the degree of Bachelor of Science in Computer Science and Engineering.

Submitted by

MD. Touhid Iqbal Sagar

ID No.: 18204023

Supervised by

Md. Muktar Hossain

Lecturer, Department of CSE, BAUET

Department of Computer Science and Engineering

Bangladesh Army University of Engineering & Technology

December, 2022

Bangladesh Army University of Engineering & Technology

Department of Computer Science and Engineering

CERTIFICATE

This is to certify that the thesis entitled **"Outlier Detection in Data Mining"** by **"MD. Touhid Iqbal Sagar"** has been accepted as satisfactory in partial fulfillment of the requirement for the degree of Bachelor of Science in Computer Science and Engineering on December, 2022.

Signature of Supervisor

……

(Md. Muktar Hossain)

(Lecturer)

Department of Computer Science and Engineering (CSE)

Bangladesh Army University of Engineering & Technology (BAUET)

Bangladesh Army University of Engineering & Technology

Department of Computer Science and Engineering

DECLARATION

We hereby declare that thesis titled **"Outlier Detection in Data Mining"** is a debonair thesis. I also ensure that it does not previously submit or published elsewhere for the award of any degree or diploma.

The work has been accepted for the degree of Bachelor of Science in Computer Science and Engineering at Bangladesh Army University of Engineering & Technology (BAUET).

Author

…….

MD. Touhid Iqbal Sagar

ACKNOWLEDGEMENT

At first, I would like to thank The Almighty Allah for giving me the opportunity to complete the thesis work successfully.

I would like to express our honor to Mohammed Golam Sarwar Bhuyan Sir, Associate Professor & Head Dept. of Computer Science and Engineering, BAUET for the love, support and motivation that he has given me continuously throughout my study. We will always be grateful to him for giving us the scope to work at the Mobile Apps and Games Development Lab.

I will forever be indebted to Md. Muktar Hossain for providing all the necessary help throughout the journey. He has not only supervised me but also his advice, guidance, constant encouragement, kind help and the thoughts he has offered have enriched my thesis without which, this work would not have been materialized in the present form. Those sessions lead to new insights, which I could directly apply in my research. During the writing of my thesis, I appreciated the scientific feedback of Md. Muktar Hossain, especially when it was difficult to put things on paper. I am very grateful that he spends a lot of time in reviewing my thesis.

My parents have put me ahead of themselves. Because of their hard work and dedication, I have had opportunities beyond my wildest dreams. My heartfelt thanks to them for giving me all I ever needed to be a successful student and individual.

I also very thankful to lab assistants of Computer Science and Engineering department for their satisfied behavior. Finally, thanks a lot to all my well-wishers.

MD. Touhid Iqbal Sagar
ID: 18204023

ABSTRACT

Outlier detection in data mining becomes a promising field of research. The main purpose of this work is to identify the parameters that are affected by outlier tools from thousands of parameters. This paper represents the research on some existing algorithms on outlier detection. Sometimes this is difficult to predict outliers from a large dataset. The main motive of this work is to gather knowledge without any outlier in the dataset. By this procedure, individual outliers can be detected. The methodology described in the paper can detect any outlier by ensuring the possibly best outlier free dataset. In this paper the main focused algorithms are IQR / Box plot outlier detection method, Scatter Plots outlier detection method and Z score outlier detection method.

Keywords: Outlier, Data Mining, Algorithm, Dataset, Constant Value

List of contents

List of figures

List of tables

Chapter 1

INTRODUCTION

1.1 Introduction

Data mining is the process of finding anomalies, patterns and correlations within large data sets to predict outcomes. Using a broad range of techniques, you can use this information to increase revenues, cut costs, improve customer relationships, reduce risks and more.

The process of digging through data to discover hidden connections and predict future trends has a long history. Sometimes referred to as "knowledge discovery in databases," the term "data mining" wasn’t coined until the 1990s. But its foundation comprises three intertwined scientific disciplines: statistics (the numeric study of data relationships), artificial intelligence (human-like intelligence displayed by software and/or machines) and machine learning (algorithms that can learn from data to make predictions). What was old is new again, as data mining technology keeps evolving to keep pace with the limitless potential of big data and affordable computing power.

Over the last decade, advances in processing power and speed have enabled us to move beyond manual, tedious and time-consuming practices to quick, easy and automated data analysis. The more complex the data sets collected, the more potential there is to uncover relevant insights. Retailers, banks, manufacturers, telecommunications providers and insurers, among others, are using data mining to discover relationships among everything from price optimization, promotions and demographics to how the economy, risk, competition and social media are affecting their business models, revenues, operations and customer relationships. [1]

Outlier is a data object that deviates significantly from the rest of the data objects and behaves in a different manner. An outlier is an object that deviates significantly from the rest of the objects. They can be caused by measurement or execution errors. The analysis of outlier data is referred to as outlier analysis or outlier mining.

An outlier cannot be termed as a noise or error. Instead, they are suspected of not being generated by the same method as the rest of the data objects. [2]

Outliers are of three types, namely –

- ✓ Global (or Point) Outliers
- ✓ Collective Outliers
- ✓ Contextual (or Conditional) Outliers

1.2 Outliers

1.2.1 Global Outliers

They are also known as Point Outliers. These are the simplest form of outliers. If, in a given dataset, a data point strongly deviates from all the rest of the data points, it is known as a global outlier. Mostly, all of the outlier detection methods are aimed at finding global outliers.

For example, In Intrusion Detection System, if a large number of packages are broadcast in a very short span of time, then this may be considered as a global outlier and we can say that that particular system has been potentially hacked.

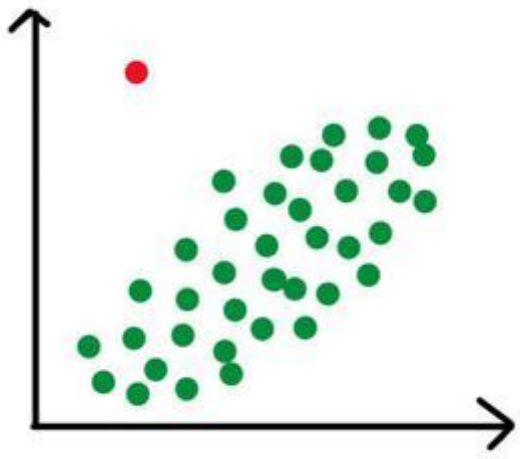


Fig 1.1: The red data point is a global outlier.

1.2.2. Collective Outliers

As the name suggests, if in a given dataset, some of the data points, as a whole, deviate significantly from the rest of the dataset, they may be termed as collective outliers. Here, the individual data objects may not be outliers, but when seen as a whole, they may behave as outliers. To detect these types of outliers, we might need background information about the relationship between those data objects showing the behaviour of outliers.

For example: In an Intrusion Detection System, a DOS (denial-of-service) package from one computer to another may be considered as normal behaviour. However, if this happens with several computers at the same time, then this may be considered as abnormal behaviour and as a whole they can be termed as collective outliers.

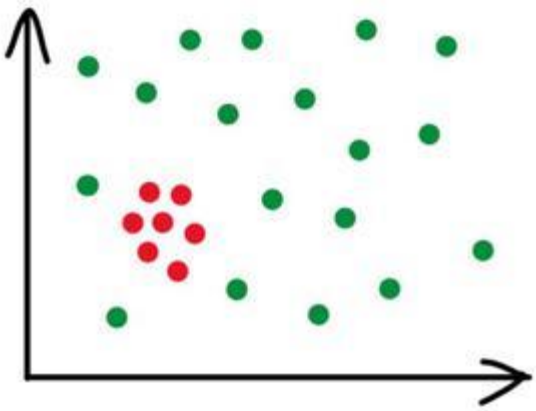


Fig 1.2: The red data points as a whole are collective outliers

1.2.3 Contextual Outliers

They are also known as Conditional Outliers. Here, if in a given dataset, a data object deviates significantly from the other data points based on a specific context or condition only. A data point may be an outlier due to a certain condition and may show normal behaviour under another condition. Therefore, a context has to be specified as part of the problem statement in order to identify contextual outliers. Contextual outlier analysis

provides flexibility for users where one can examine outliers in different contexts, which can be highly desirable in many applications. The attributes of the data point are decided on the basis of both contextual and behavioural attributes.

For example: A temperature reading of 40°C may behave as an outlier in the context of a "winter season" but will behave like a normal data point in the context of a "summer season".

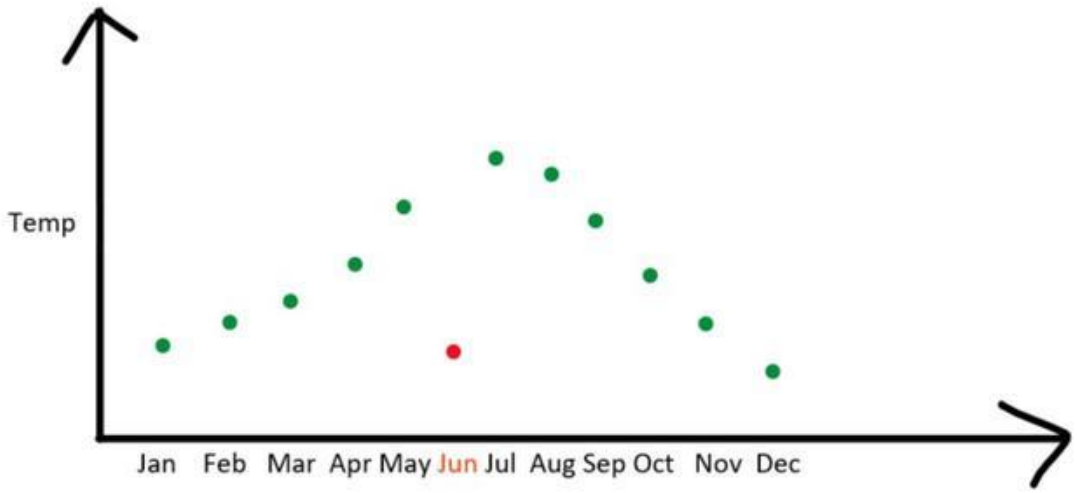


Fig 1.3: A low temperature value in June is a contextual outlier because the same value in December is not an outlier.

1.3 Objectives

Outlier detection is a statistical procedure that aims to find suspicious events or items that are different from the normal form of a dataset. It has drawn considerable interest in the field of data mining and machine learning. Outlier detection is important in many applications, including fraud detection in credit card transactions and network intrusion detection. There are two general types of outlier detection: global and local. Global outliers fall outside the normal range for an entire dataset, whereas local outliers may fall within the normal range for the entire dataset, but outside the normal range for the surrounding data points. [3]

The thesis “Outlier Detection on Data Mining” is aimed to detect the outlier from a dataset. The research is to modify some existing literature on outlier detection like IQR/Box plot outlier detection method, Scatter Plots outlier detection method, Z score outlier detection method. In this research there is also a performance comparison with existing constant values.

1.4 Motivation:

An outlier may indicate bad data. For example, the data may have been coded incorrectly or an experiment may not have been run correctly. If it can be determined that an outlying point is in fact erroneous, then the outlying value should be deleted from the analysis (or corrected if possible).

In some cases, it may not be possible to determine if an outlying point is bad data. Outliers may be due to random variation or may indicate something scientifically interesting. In any event, we typically do not want to simply delete the outlying observation. However, if the data contains significant outliers, we may need to consider the use of robust statistical techniques. [4]

Now a days almost every technology uses database-based service. Dataset is the collection of huge amounts of data. Because it’s a huge number of data it’s very common that there can be any noisy data or outlier. This research is based on that outlier detection. Outlier can give a wrong output that can affect user’s result. It can be difficult to maintain outlier manually. So, this research helps user to gather knowledge without outlier.

1.5 Background:

The definitions of an outlier that are frequently used, the causes of outliers, innovative methods for identifying and detecting outliers, and what to do when an outlier are all covered in this part is discovered. Finally, we present a few fresh applications for outlier detection and offer more resources for follow-up research in these domains.

There have been numerous definitions of what an outlier is since the inception of outlier identification studies. Twelve different interpretations of outliers from the viewpoint of several authors were provided by Ayadi et al. in 2017 [5].

This highlights how difficult it is to define an outlier accurately. Although the term "outlier" is ambiguous and difficult to define, it may generally be summed up as a data point that is significantly different from other data points or a point that does not mimic the expected typical behaviour of the other points [6]. Inliers are data points that are consistent with outliers.

The few existing methods of outlier detection method on data mining are:

- ✓ IQR / Box plot outlier detection method:

  The boxplot method (Exploratory Data Analysis, Addison-Wesley, Reading, MA, 1977) is a graphically-based method of identifying outliers which is appealing not only in its simplicity but also because it does not use the extreme potential outliers in computing a measure of dispersion. The inner and outer fences are defined in terms of the hinges (or fourths), and therefore are not distorted by a few extreme values. Such distortion could lead to failing to detect some outliers, a problem known as "masking". A method for determining the probability associated with any fence or observation is proposed based on the cumulative distribution function of the order statistics. This allows the statistician to easily assess, in a probability sense, the degree to which an observation is dissimilar to the majority of the observations. In addition, an adaptation for approximately normal but somewhat asymmetric distributions is suggested. [7]

- ✓ Scatter Plots outlier detection method:

  We can spot outliers on a scatter plot if it has a regression line. The point or points that are furthest distant from the regression line are considered outliers in a scatter plot. On a scatter plot, there is typically only one outlier, and frequently there is at least one. Keep in mind that outliers in a scatter plot differ significantly from those in a box plot. The length of the line segment that extends perpendicular to the regression line from a point to the regression line determines the distance between the two. A scatter plot has at least one outlier if one of its points is farther away from the regression line than another. [6]

- ✓ Z score outlier detection method:

  The z-score test has long been used to detect outliers in data. In this work, a modified version of the z-score test is proposed which can lead to a reduction of the time-complexity of the traditional z-score test by a value of n each iteration, where n donates the number of spatial points in the data set which have not been discovered as outliers. Using the outlier threshold value to calculate the value of the modified z-score function rather than calculating the z-score value for each observation in the dataset makes this possible.

1.6 Overview of Contents

The rest of the dissertation is organized as follows:

Chapter 2 – Literature Review & Background Study:
The previous study represents my research, motivation for this research, and the previous work on this field. In this chapter there are the reasons why we have chosen outlier detection methods, challenges and previous work.

Chapter 3 – Experimental Analysis of literature review:
The previous study represents literature review and background study of the research. In this chapter we are going to know basic idea about outlier detection and research on the existing outlier detection methods.

Chapter 4 – Methodology
The previous study it represents the details of this research. In this chapter we discuss about our proposed methodology and it's working procedure. We discuss all the steps of our proposed methodology in details.

Chapter 5 – Result Analysis and Discussion:
The previous study represents our methodology. In this chapter there is a discussion about the result in this work and after doing this work we can compare our work with the previous one.

Chapter 6 – Future Work and Conclusion:
The previous study represents the result of our work. In this chapter we discuss about the future work of this work and Conclusion.

References:
All the paper references are given in this section. Referencing is described as giving credit, with citation, to the source of information.

Chapter 2

LITERATURE REVIEW & BACKGROUND STUDY

2.1 Introduction

The appearance of outliers can be caused by a wide range of different problems. Outliers can be caused by a variety of factors, including mechanical problems, shifts in system behavior, dishonest or malevolent behavior, human error, setup or instrument flaws, sample errors, data entry errors, and environmental changes. For instance, human error, such as during data entry and recording during data collection, is frequently the cause of outliers from data errors. How to recognize and handle outliers is the next problem associated with their existence.

How to find outliers is an issue that numerous researchers have attempted to address. Equally crucial issues are the features that must be taken into account and the tests that must be run to find the outliers. There are still ongoing investigations being undertaken to determine the correct answers to these concerns, despite the increased interest in this subject area.

To address these, researchers are coming up with fresh and creative solutions [8], [9]. In machine learning and data mining, the method of outlier identification has gone by various names over the years, including outlier mining, novelty discovery, outlier modeling, anomaly detection, etc. It is crucial to be alert when trying to find and remove outliers. When outliers are removed from accurate data, important hidden information could be lost. Knowing whether a case is multivariate or univariate in terms of the number of features that must be taken into account while trying to identify outliers is also essential. as well, for whether the chosen features can infer the distribution of values for parametric or non-parametric instances using a statistically based method scenario.

2.2 What to do about outliers:

What constitutes an outlier is still a topic of intense debate. The most common practice used by many researchers is to mark a data point as an outlier when it deviates from the mean by three standard deviations or more [10]. To analyse such an argument further, this, however, is a weak supporting premise because it does not apply in all circumstances. This is particularly true nowadays, when we are dealing with massive amounts of dynamic and unstructured data. In order to decide how to handle outliers in the modern world, it is essential to think further about a few key issues. For instance, whether it is wise to leave outliers in the data or accept them as such. Data outliers can occasionally have a harmful effect. This would consequently lead to a longer data training process, fewer accurate models, and ultimately subpar results in outlier detection techniques used in machine learning and deep learning.

New strategies to deal with outliers have been developed as a result of the recent development of new ways to detect outliers. Visual analysis of the data is sometimes preferred [11, 12] to acquire a clear sense of the level of outliers of a data point. Another instance [13] involves the search for data points that exhibit extreme values for a single variable using a method like the univariate methodology.

In most circumstances, one must use their intuition, analytical argument through some experiments, and deliberate thinking before making conclusions in order to address the topic of how to handle outliers. The need of taking into account the context and scenario, as well as the reason for seeking out the outliers, are other crucial queries in the outlier detection process. Understanding the purpose for identifying the outliers and what they ultimately mean is crucial. We'll see in the following sections that distinct methods or application domains necessitate different approaches to dealing with outliers.

2.3 Areas of application for outlier detection

With its increasing attention, outlier identification has a number of applications in many different fields. Outlier detection is used in a wide variety of applications. Due to space restrictions, it is impossible to cover everything completely in a single survey. As a result,

we list and describe current and recent application areas in this study. We shall direct our readers to some earlier studies that thoroughly address the wide range of application sectors where OD approaches are used.

Some existing application areas are:

- ✓ Credit card fraud detection
- ✓ Intrusion detection
- ✓ defect detection from behavioural patterns of industrial machines
- ✓ sensor networks
- ✓ Finding unusual patterns in time-series data
- ✓ Trajectories
- ✓ E-commerce
- ✓ Energy consumption

2.4 IQR / Box plot outlier detection method

It is well known that outliers, observations that are presumed to come from a different distribution than that for the majority of the data set, can have profound influence on the statistical analysis and can often lead to erroneous conclusions. Because outliers and other extreme values can be very influential in most parametric tests, it is not surprising that the detection and accommodation of outliers have received considerable attention in the literature (see for example: Andrews (1974); Andrews and Pregibon, 1978; Atkinson, 1994; Bacon-Shone and Fung, 1987; Barnett, 1978; Barnett and Lewis, 1995; Brant, 1990; Bhandary, 1992; Gnanadesikan and Kettenring, 1972; Hampel, 1985; Hawkins, 1980; Penny, 1996 and Tukey, 1977). The extensive literature on the subject of outliers attests to its relevance as a major concern in the statistical analysis of data.

One simple way commonly employed to identify outliers is based on the concept of the boxplot and involves the use of "inner fences" and "outer fences." This method, suggested by Tukey (1977), has come into common usage, is often included in texts (see for example, Milton, 1999 pp. 55 –58), and has been studied extensively (see for example, Hoaglin et al., 1986; Carling, 2000; Beckman and Cook, 1983; Frigge et al., 1989). This graphically-

based method for identifying outliers is especially appealing not only in its simplicity but, more importantly, because it does not use the extreme potential outliers which can distort the computing of a measure of spread and lessen the sensitivity to outliers. The fences procedure uses the estimated interquartile range often referred to as the H-spread, which is the difference between values of the hinges, i.e., sample third and 3rst quartiles. Speci3cally, the inner fences, f1 and f3, and outer fences, F1 and F3, are usually defined as

$$f1 = q1 - 1{:}5\text{H-spread}; \; f3 = q3 + 1{:}5\text{H-spread};$$

$$F1 = q1 - 3\text{H-spread and } F3 = q3 + 3\text{H-spread}; \qquad (2.4.1)$$

where q1 and q3 are the 3rst and third sample quartiles and H-spread = q3 − q1. Tukey (1977) called observations that fall between the inner and outer fences in each direction "outside" outliners, while those that fall below the outer fence F1 or above the outer fence F3 are "far out" outliers. Inconsistency in de3ning quartiles complicates the construction of fences. Carling (2000) points out that various authors, such as Cleveland (1985), Freund and Perles (1987), Frigge et al. (1989), Hyndman and Fan (1996), Harrell and Davis (1982), Hoaglin et al. (1983) and Hoaglin and Iglewicz (1987) have proposed a variety of definitions for quartiles. Frigge et al. (1989) list eight de3nitions of the sample quartiles. With so many de3nitions it is not surprising that there is variation in determining the criteria for outliers. Carling (2000) compares the outside rate for two of the most common de3nitions, one by Tukey and the "ideal or machine fourth" recommended by Frigge et al. (1989), concluding the "ideal or machine fourth" is an improvement. In this paper, the sample quartiles or hinges are approximated by 3nding the middle point of a set of ordered observations and then 3nding the approximate quartiles q1 and q3 as the middle points of the ordered smaller and larger sets, respectively. See Tukey (1977, pp. 29–38) for more details on the method of 3nding "hinges". Many practitioners might 3nd the outer fences too conservative, causing them to overlook many real outliers. In this paper a new simple more general fences method is suggested which allows flexibility in setting the "outside rate", that is, the probability that an observation from a non-contaminated normal population is outside a speci3ed limit or boundary. While the theoretical development assumes both a normal population.[14]

2.4.1 Example

Data set = (22,25,17,19,33,64,23,17,20,18) [N=10]

Arrange the data ascending format = (17,17,18,19,20,22,23,25,33,64)

Q2 = (20+22)/2 = 21 Q1 = 18 Q3 = 25

Inter-Quartile Range (IQR) = Q3-Q1 = 25-18 = 7

Higher outlier = Q3+(1.5*IQR) = 25+(1.5*7) = 35.5

Lower outlier = Q1- +(1.5*IQR) = 18- +(1.5*7) = 7.5

So, Outlier = 64

Minimum value = 17

Maximum value = 33

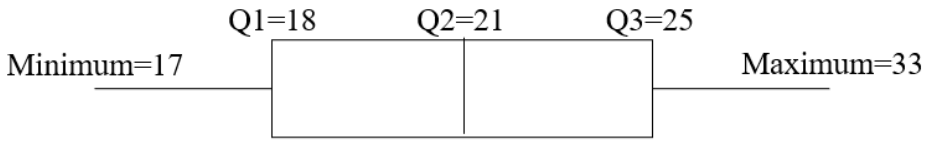


Fig 2.1: Box plot 01

2.5 Scatter Plots outlier detection method

If there is a regression line on a scatter plot, you can identify outliers. An outlier for a scatter plot is the point or points that are farthest from the regression line. There is at least one outlier on a scatter plot in most cases, and there is usually only one outlier. Note that outliers for a scatter plot are very different from outliers for a boxplot.

Distance from a point to the regression line is the length of the line segment that is perpendicular to the regression line and extends from the point to the regression line. If one point of a scatter plot is farther from the regression line than some other point, then the scatter plot has at least one outlier. If a number of points are the same farthest distance from the regression line, then all these points are outliers. If all points of the scatter plot are the same distance from the regression line, then there is no outlier.

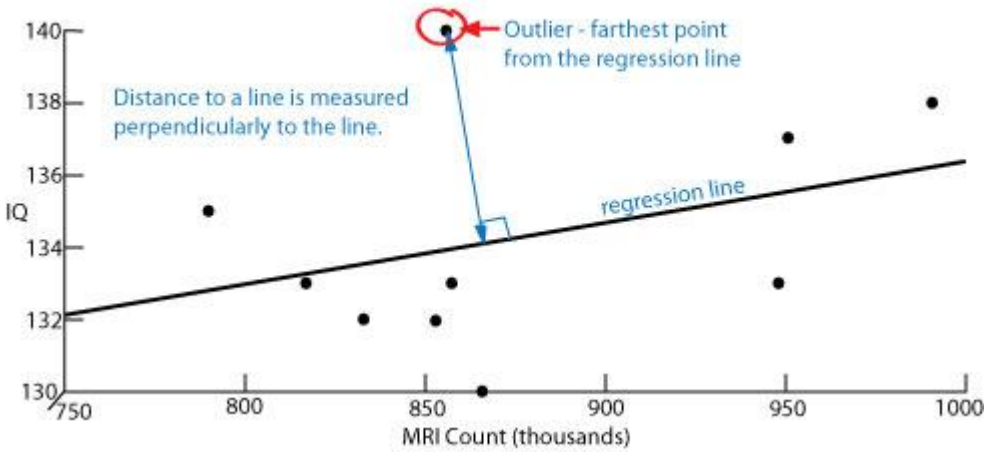


Fig 2.2: Scatter Plots outlier detection 01.

An influential observation (inf. obs.) is a point on a scatter plot that has a large horizontal gap containing no points between it and a vast majority of the other points. As shown in the graph below, there can be more than one influential observation. If there is no large horizontal gap between data points in a scatter plot, there are no influential observations. In many cases, a scatter plot will have no influential observations; but influential observations should be identified if they occur.

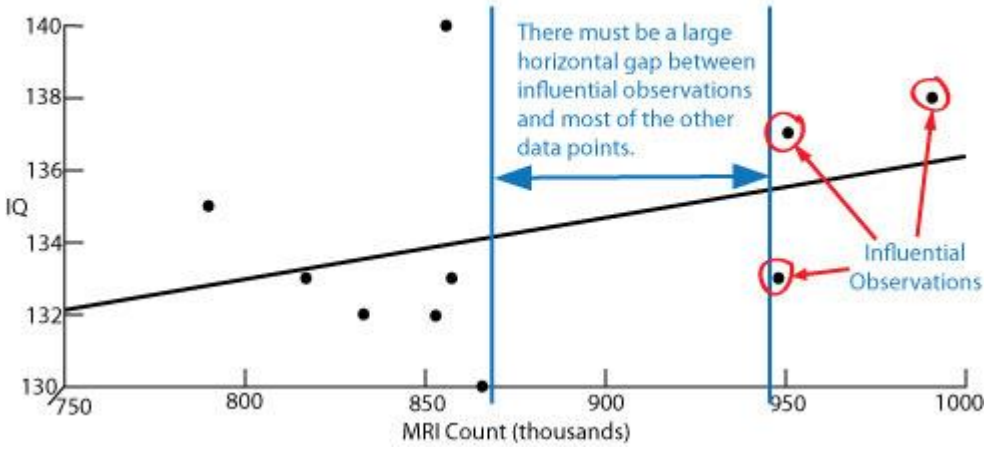


Fig 2.3: Influent observations detection

When an influential observation is moved up or down and the regression line is recomputed, the new line will be much closer to the new location of the influential observation. If a non-influential observation is relocated, the recomputed regression line will be in almost the same position as the original regression line. Thus, the influential observation has "influence" on the location of the regression line. [15,16]

2.5.1 Example:

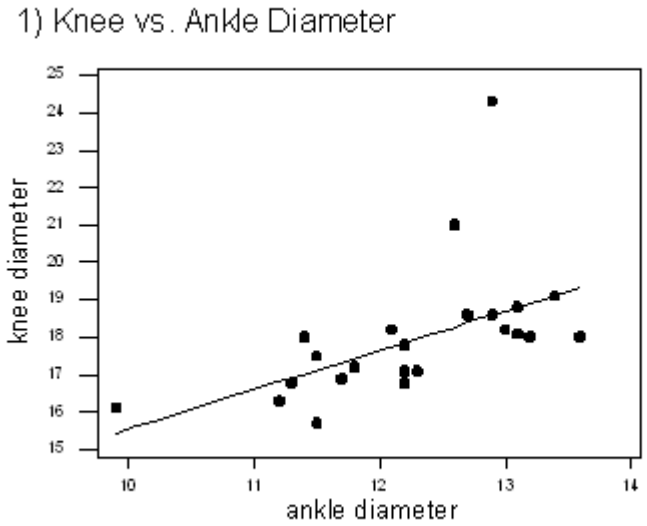


Fig 2.4: Example 01

2.6 Z score outlier detection method

A z-score is defined as a measure of the divergence of a different experimental observation from the most probable result, the mean. Z is expressed in terms of the number of standard deviations from the mean value.

$$Z = (X- \mu)/ \sigma \qquad (2.5.1)$$

X = Experimental Value

μ = Mean

σ = Standard Deviation

It is apparent that the measure of divergence provided by the z-score provides a mechanism to determine the magnitude by which an observation "diverges" from the other observations of the dataset and if found to be large enough, the observation can be deemed an outlier [17]. The traditional z-statistic has also been previously used in spatial outlier detection algorithm which also involves identifying the K nearest neighbours of a data object however involves computing the value of the traditional z-statistic test for each data object and then comparing it to the threshold value (θ) in order to evaluate whether the data object under scrutiny is an outlier or not.

The database consists of a set of spatial points, X= {x1, x2, ..., x n} The function f(xi,xj) is defined as the relationship function mapping from XxX to R (real number set) which is used to determine the relationship between any two spatial points xi and xj the implementer depending upon his/her requirements can choose the relationship function, which may be a Euclidean distance function, Manhattan distance function, etc. The function c(xi) is defined as the comparison function for a spatial point i. The K nearest neighbours of a spatial point xi are denoted as KNBR (xi). The threshold value is denoted as θ and the current comparator value calculated from the threshold value is denoted as current. [18,19]

2.6.1 Z score outlier detection algorithm

Step 1: For the given spatial dataset X= {x1, x2…., xn} find the k nearest neighbour of each spatial point xi ∈ X and sorted their indexes in KNBR (xi).

Step 2: For each point xi ∈, computer c(xi) as $C(xi)=\sum j \in knbr(xi) f(xi,xj)$

Step 3: Compute the mean and standard deviation for the set {c(xi), c(x2) …,c(xn)}, computer the value of CURRENT as

$CURRENT=\theta*\sigma+\mu$

Where θ is the threshold value.

Step 4: For each xi ∈ X, if c(xi)>= CURRENT, then xi is an outlier.

Step 5: Ignoring the discovered outliers find the current K nearest neighbours of each xi ∈ X, and repeat steps 2,3,4 and 5 until no more outliers are found in step4. [19]

2.6.2 Example

Dataset = [11, 10, 12, 14, 12, 15, 14, 13, 15, 102, 12, 14, 17, 19, 107, 10, 13, 12, 14, 12, 108, 12, 11, 14, 13, 15, 10, 15, 12, 10, 14, 13, 15, 10, 109, 123]

For threshold = 2

Outliers = [102, 107, 108, 109, 123]

2.7 Conclusion

In this chapter, there is a discussion about the Literature Review and Background Study of our work. There is also discussed how outliers can be detected and why we detect outlier. I have discussed the challenges and also discuss about previous work related to my work. Many scholars and researcher pay attention in many years to the impact of the outlier detection methods. In their study, they have used different kinds of algorithms and techniques. I have studied some thesis papers and also have shown some algorithms of outlier detection.

Chapter 3

EXPERIMENTAL ANALYSIS OF LITERATURE REVIEW

3.1 Introduction

The previous study represents literature review and background study of the research. In this chapter we are going to implement and research on the existing outlier detection methods. All the experiments of different methods will be present here. The experiments help to improve the method or shows the perfection of the existing methods.

3.2 IQR / Box plot outlier detection method

Experiment 1:

There is a dataset of 10000 data to implement the method. That shows us gender, their height and weight information. This dataset also has some outliers. We have to implement the outlier detection method to detect the outlier.

	Gender	Height	Weight
0	Male	73.847017	241.893563
1	Male	68.781904	162.310473
2	Male	74.110105	212.740856
3	Male	71.730978	220.042470
4	Male	69.881796	206.349801
...	...	...	...
9995	Female	66.172652	136.777454
9996	Female	67.067155	170.867906
9997	Female	63.867992	128.475319
9998	Female	69.034243	163.852461
9999	Female	61.944246	113.649103

10000 rows × 3 columns

Table 3.1: Dataset

By using describe function, we have the calculation that is given below:

	Height	Weight
count	10000.000000	10000.000000
mean	66.367560	161.440357
std	3.847528	32.108439
min	54.263133	64.700127
25%	63.505620	135.818051
50%	66.318070	161.212928
75%	69.174262	187.169525
max	78.998742	269.989699

Table 3.2: Calculation

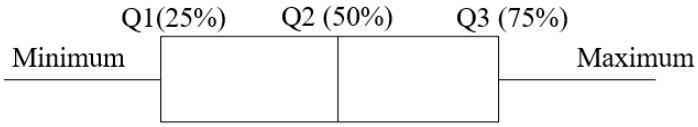


Fig 3.1: Box plot 02

Inter-Quartile Range (IQR) = Q3-Q1 = (69.1742617268347 - 63.505620481218955)

= 5.668641245615746

For constant value 1.5:

Lower limit = Q1 - 1.5*IQR

= 55.00265861279534

Upper limit = Q3 + 1.5*IQR

= 77.67722359525831

Detected outlier is given below:

	Gender	Height	Weight
994	Male	78.095867	255.690835
1317	Male	78.462053	227.342565
2014	Male	78.998742	269.989699
3285	Male	78.528210	253.889004
3757	Male	78.621374	245.733783
6624	Female	54.616858	71.393749
7294	Female	54.873728	78.606670
9285	Female	54.263133	64.700127

Table 3.3: Outlier

Dataset without outlier is given below:

	Gender	Height	Weight
0	Male	73.847017	241.893563
1	Male	68.781904	162.310473
2	Male	74.110105	212.740856
3	Male	71.730978	220.042470
4	Male	69.881796	206.349801
...	...	...	...
9995	Female	66.172652	136.777454
9996	Female	67.067155	170.867906
9997	Female	63.867992	128.475319
9998	Female	69.034243	163.852461
9999	Female	61.944246	113.649103

9992 rows × 3 columns

Table 3.4: Dataset without outliers

For constant value 2:

Lower limit = Q1 - 2*IQR

= 52.168337989987464

Upper limit = Q3 + 2*IQR

= 80.51154421806619

Detected outlier is given below:

Gender	Height	Weight

Table 3.5: No outlier found

Dataset without outlier is given below:

	Gender	Height	Weight
0	Male	73.847017	241.893563
1	Male	68.781904	162.310473
2	Male	74.110105	212.740856
3	Male	71.730978	220.042470
4	Male	69.881796	206.349801
...	...	...	...
9995	Female	66.172652	136.777454
9996	Female	67.067155	170.867906
9997	Female	63.867992	128.475319
9998	Female	69.034243	163.852461
9999	Female	61.944246	113.649103

10000 rows × 3 columns

Table 3.6: Dataset without outliers

For constant value 1:

Lower limit = Q1 - 1*IQR

= 57.83697923560321

Upper limit = Q3 + 1*IQR

= 74.84290297245045

Detected outlier is given below:

	Gender	Height	Weight
23	Male	75.205974	228.761781
190	Male	76.709835	235.035419
197	Male	75.944460	231.924749
202	Male	75.140821	224.124271
232	Male	76.021347	235.926061
...	...	...	...
9761	Female	56.975279	90.341784
9825	Female	55.979198	85.417534
9895	Female	57.740192	93.652957
9904	Female	57.028857	101.202551
9978	Female	57.375759	114.192209

173 rows × 3 columns

Table 3.7: Outlier

Dataset without outlier is given below:

	Gender	Height	Weight
0	Male	73.847017	241.893563
1	Male	68.781904	162.310473
2	Male	74.110105	212.740856
3	Male	71.730978	220.042470
4	Male	69.881796	206.349801
...	...	...	...
9995	Female	66.172652	136.777454
9996	Female	67.067155	170.867906
9997	Female	63.867992	128.475319
9998	Female	69.034243	163.852461
9999	Female	61.944246	113.649103

9827 rows × 3 columns

Table 3 8: Dataset without outliers

Experiment 2:

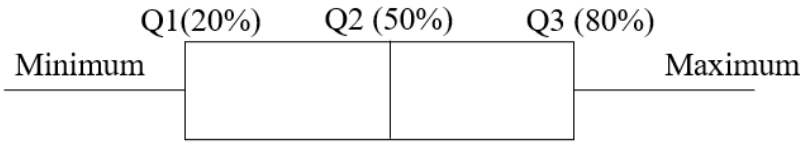


Fig 3.2: Box plot 03

Let’s set the value of Q1 & Q3 as bellow

Inter-Quartile Range (IQR) = Q3-Q1 = (69.81162000527536 - 62.85900650806874)

= 6.95261349720662

Lower limit = Q1 - 1.5*IQR

= 52.43008626225881

Upper limit = Q3 + 1.5*IQR

= 80.24054025108529

Detected outlier is given below:

Gender	Height	Weight

Table 3.9: No outlier found

Dataset without outlier is given below:

	Gender	Height	Weight
0	Male	73.847017	241.893563
1	Male	68.781904	162.310473
2	Male	74.110105	212.740856
3	Male	71.730978	220.042470
4	Male	69.881796	206.349801
...	...	...	...
9995	Female	66.172652	136.777454
9996	Female	67.067155	170.867906
9997	Female	63.867992	128.475319
9998	Female	69.034243	163.852461
9999	Female	61.944246	113.649103

10000 rows × 3 columns

Table 3.10: Dataset without outliers

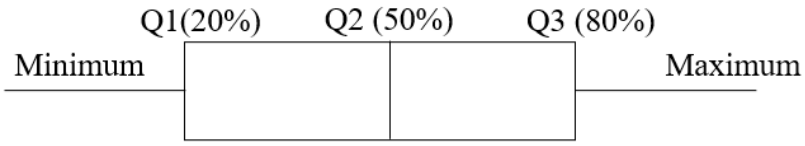


Fig 3.3: Box plot 04

Let’s set the value of Q1 & Q3 as bellow

Inter-Quartile Range (IQR) = Q3-Q1 = (71.47214936325072 - 61.41270131336016)

= 10.059448049890563

Lower limit = Q1 - 1.5*IQR

= 46.323529238524316

Upper limit = Q3 + 1.5*IQR

= 86.56132143808657

Detected outlier is given below:

Gender	Height	Weight

Table 3.11: No outlier found

Dataset without outlier is given below:

	Gender	Height	Weight
0	Male	73.847017	241.893563
1	Male	68.781904	162.310473
2	Male	74.110105	212.740856
3	Male	71.730978	220.042470
4	Male	69.881796	206.349801
...	...	...	...
9995	Female	66.172652	136.777454
9996	Female	67.067155	170.867906
9997	Female	63.867992	128.475319
9998	Female	69.034243	163.852461
9999	Female	61.944246	113.649103

10000 rows × 3 columns

Table 3.12: Dataset without outliers

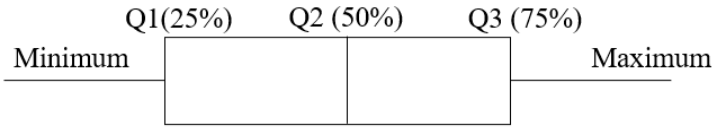


Fig 3.4: Box plot 05

Let’s set the value of Q1 & Q3 as bellow

Inter-Quartile Range (IQR) = Q3-Q1 = (69.1742617268347 - 63.505620481218955)

= 5.668641245615746

Lower limit = Q1 - 1.5*IQR

= 55.00265861279534

Upper limit = Q3 + 1.5*IQR

= 77.67722359525831

Detected outlier is given below:

	Gender	Height	Weight
994	Male	78.095867	255.690835
1317	Male	78.462053	227.342565
2014	Male	78.998742	269.989699
3285	Male	78.528210	253.889004
3757	Male	78.621374	245.733783
6624	Female	54.616858	71.393749
7294	Female	54.873728	78.606670
9285	Female	54.263133	64.700127

Table 3.13: Outlier

Dataset without outlier is given below

	Gender	Height	Weight
0	Male	73.847017	241.893563
1	Male	68.781904	162.310473
2	Male	74.110105	212.740856
3	Male	71.730978	220.042470
4	Male	69.881796	206.349801
...	...	...	...
9995	Female	66.172652	136.777454
9996	Female	67.067155	170.867906
9997	Female	63.867992	128.475319
9998	Female	69.034243	163.852461
9999	Female	61.944246	113.649103

9992 rows × 3 columns

Table 3.14: Dataset without outliers

3.3 Scatter Plots outlier detection method

There is a dataset of trending YouTube videos likes and view ratio. We have almost 200 datasets to implement the model. In this figure the light blue colour shows more disliked videos and dark dots are most liked videos.

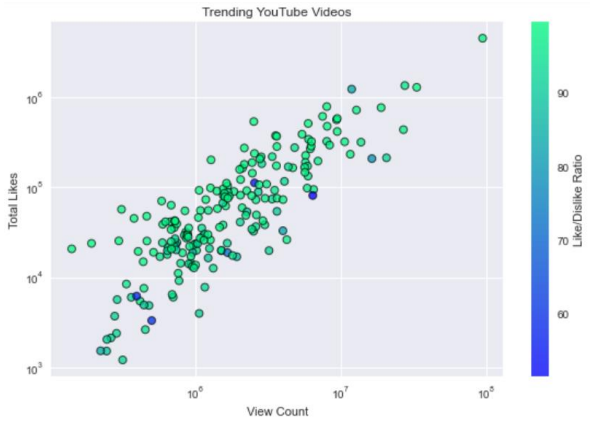


Fig 3.5: Scatter Plots outlier detection

3.4 Z-score outlier detection method

Dataset = [11, 10, 12, 14, 12, 15, 14, 13, 15, 102, 12, 14, 17, 19, 107, 10, 13, 12, 14, 12, 108, 12, 11, 14, 13, 15, 10, 15, 12, 10, 14, 13, 15, 10, 109, 123]

For threshold = 1:

Outliers = [102, 107, 108, 109, 123]

For threshold = 2:

Outliers = [102, 107, 108, 109, 123]

For threshold = 3:

Outliers = []

For threshold = 0/-1/-2/-3…:

Outliers = [11, 10, 12, 14, 12, 15, 14, 13, 15, 102, 12, 14, 17, 19, 107, 10, 13, 12, 14, 12, 108, 12, 11, 14, 13, 15, 10, 15, 12, 10, 14, 13, 15, 10, 109, 123, 102, 107, 108, 109, 123]

3.5 Conclusion

This chapter was based on experiments of the existing literature. Here outlier detection algorithms are implemented. We can compare with many more constant values. They all gives different outputs. We have to compare all the outputs. In Result Analysis and Discussion chapter we have to compare all of these experiments. It helps us to identify all the difference of changing algorithm’s values in the experiment.

Chapter 4

METHODOLOGY

4.1 Introduction

Methodology refers to the systematic, theoretical analysis of the methods applied to a field of study. It comprises the theoretical analysis of the body of methods and principles associated with a branch of knowledge. In the social sciences, methodology generally refers to the theoretical or philosophical analysis of research methods. This includes the assumptions that underpin a particular research approach, the methods used to collect and analyse data, and the ways in which data is interpreted and drawn into conclusions. Methodology is an important aspect of research as it helps to ensure that the results of a study are reliable, valid, and unbiased. It is also an important element of research proposals and academic papers, as it helps to justify the chosen research methods and demonstrate their appropriateness for the research question being studied.

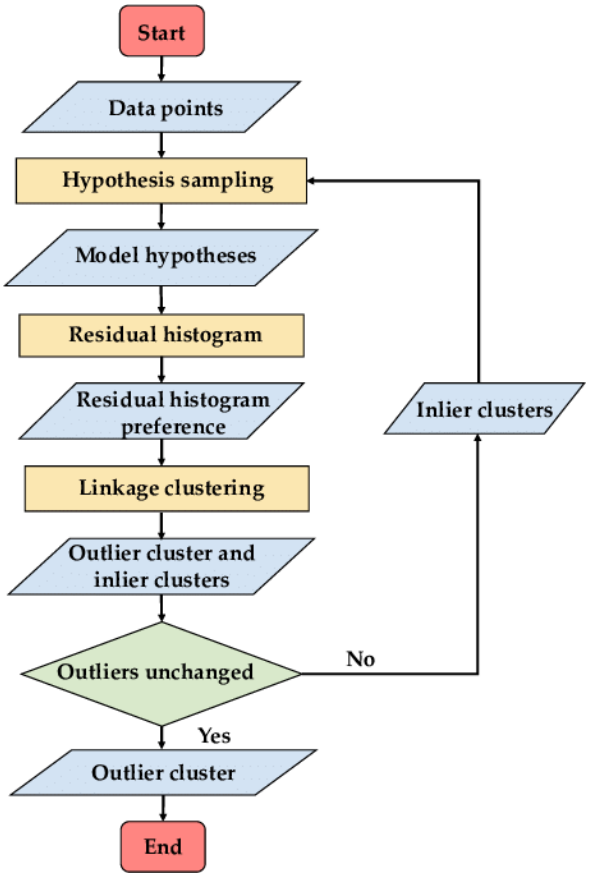


Fig 4.1: Flow chart of outlier detection on data mining

4.2 Methodology description

There are several different methodological approaches that can be used for outlier detection, depending on the characteristics of the data and the research goals. Some common approaches include:

Visual inspection of data: This involves visually examining plots of the data, such as scatterplots or box plots, to identify unusual or extreme observations.

Statistical tests: There are various statistical tests that can be used to identify outliers, including the Grubbs test, the Dixon test, and the Z-score method. These tests calculate the probability that a given observation is an outlier based on the distribution of the data.

Machine learning algorithms: Machine learning algorithms, such as clustering or density-based methods, can be used to identify outliers in large or complex datasets. These

algorithms can identify patterns in the data and flag observations that are significantly different from the rest of the data.

It is important to carefully consider the appropriate method for outlier detection based on the characteristics of the data and the goals of the analysis. It may also be necessary to use a combination of different methods to thoroughly identify and address outliers in the data.

4.3 Proposed methodology:

A research methodology gives research legitimacy and provides scientifically sound findings. It also provides a detailed plan that helps to keep researchers on track, making the process smooth, effective and manageable. The methodology we are proposing consists of some key parts which are:

- ✓ IQR/Box plot
- ✓ Scatter plots
- ✓ Z score

According to these key steps we have constructed our working methodology to achieve the goal. These are the key processes of any methodology related to outlier detection method. We have worked according to this procedure. This works as the blueprint of our whole work which eases the work to achieve the desired outcome.

4.3.1 IQR/Box plot:

The interquartile range (IQR) method, also known as the box plot method, is a simple and effective way to identify outliers in a dataset. The IQR method is based on the concept of the quartiles of a dataset, which divide the data into four equal parts. The first quartile (Q1) is the value at the 25th percentile, the second quartile (Q2) is the value at the 50th percentile (also known as the median), and the third quartile (Q3) is the value at the 75th percentile. The IQR is calculated as the difference between Q3 and Q1.

To use the IQR method for outlier detection, follow these steps:

- ✓ Collect and organize the data.
- ✓ Calculate the first quartile (Q1), second quartile (Q2), and third quartile (Q3) of the data.
- ✓ Calculate the interquartile range (IQR) as the difference between Q3 and Q1.
- ✓ Determine the lower and upper bounds of the data using the following formulas:
- ✓ Lower bound = Q1 - 1.5 * IQR
- ✓ Upper bound = Q3 + 1.5 * IQR

Any observations that fall outside of the lower and upper bounds are considered potential outliers. Determine whether the identified observations are true outliers or if they may be caused by measurement error or other factors. If the identified observations are true outliers, consider whether they should be included or excluded from the analysis. This decision may depend on the research goals and the potential impact of the outliers on the results. If the identified observations are not true outliers, consider whether they should be included or excluded from the analysis. This decision may depend on the potential impact of the observations on the results and the reliability of the data. Analyse the data with or without the identified outliers, as appropriate. It is important to carefully consider the appropriate method for outlier detection and to document the process in order to ensure that the results of the analysis are reliable and valid. [14]

4.3.2 Scatter plots:

Here is a general flow chart for using a scatter plot for outlier detection:

- ✓ Collect and organize the data.
- ✓ Create a scatter plot of the data, with one variable plotted on the x-axis and the other variable plotted on the y-axis.
- ✓ Examine the scatter plot for observations that are significantly different from the rest of the data. These observations may be located far from the main cluster of points or may have unexpected patterns or trends.
- ✓ Determine whether the identified observations are true outliers or if they may be caused by measurement error or other factors.

- ✓ If the identified observations are true outliers, consider whether they should be included or excluded from the analysis. This decision may depend on the research goals and the potential impact of the outliers on the results.
- ✓ If the identified observations are not true outliers, consider whether they should be included or excluded from the analysis. This decision may depend on the potential impact of the observations on the results and the reliability of the data.
- ✓ Analyse the data with or without the identified outliers, as appropriate.

It is important to carefully consider the appropriate method for outlier detection and to document the process in order to ensure that the results of the analysis are reliable and valid. [15,16]

4.3.3 Z score:

The z-score method is a statistical technique that can be used to identify outliers in a dataset. The z-score is a measure of how many standard deviations an observation is from the mean of the data. A z-score of zero indicates that the observation is exactly average, while a z-score of 1.96 or higher indicates that the observation is an outlier (assuming a 95% confidence level).

To use the z-score method for outlier detection, follow these steps:

- ✓ Collect and organize the data.
- ✓ Calculate the mean and standard deviation of the data.
- ✓ Calculate the z-score for each observation in the data using the following formula:
- ✓ z-score = (observation - mean) / standard deviation
- ✓ Identify any observations with a z-score of 1.96 or higher as potential outliers.
- ✓ Determine whether the identified observations are true outliers or if they may be caused by measurement error or other factors.
- ✓ If the identified observations are true outliers, consider whether they should be included or excluded from the analysis. This decision may depend on the research goals and the potential impact of the outliers on the results.

- ✓ If the identified observations are not true outliers, consider whether they should be included or excluded from the analysis. This decision may depend on the potential impact of the observations on the results and the reliability of the data.
- ✓ Analyse the data with or without the identified outliers, as appropriate.

It is important to carefully consider the appropriate method for outlier detection and to document the process in order to ensure that the results of the analysis are reliable and valid. [18,19]

4.4 Conclusion

The choice of methodology for outlier detection will depend on the characteristics of the data and the research goals. Some common approaches include visual inspection of the data, statistical tests, and machine learning algorithms. It may also be necessary to use a combination of different methods in order to thoroughly identify and address outliers in the data.

Regardless of the methodology used, it is important to carefully consider the appropriateness of the method for the specific dataset and research goals. It is also important to document the process of outlier detection in order to ensure the reliability and validity of the results.

In conclusion, the methodology for outlier detection should be carefully chosen and applied in a rigorous and systematic manner in order to accurately identify and address outliers in the data and obtain reliable and valid results.

Chapter 5

RESULT ANALYSIS AND DISCUSSION

5.1 Introduction

Result analysis and discussion refers to the process of evaluating and interpreting the results of a study or experiment. This involves analysing the data, determining the key findings and trends, and discussing the implications and limitations of the results.

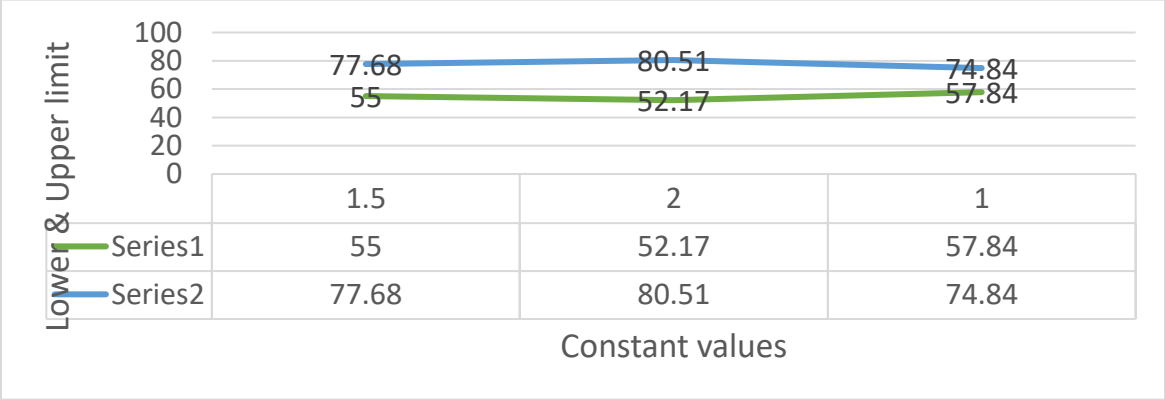
In result analysis, the data is typically organized and summarized in a meaningful way, such as through the use of statistical tests or graphical representations. The results are then interpreted in the context of the research question or hypothesis being tested. Key findings and trends are identified and discussed in relation to the research goals and the existing literature on the topic.

In the discussion section, the results are placed in a broader context and the implications and limitations of the study are considered. The discussion may also address potential applications of the results and suggest directions for future research.

Result analysis and discussion are important components of research as they help to provide insights and understanding of the research question being studied and contribute to the development of knowledge in the field. It is important to carefully and objectively evaluate and interpret the results and to clearly and accurately communicate the findings in the discussion section.

5.2 IQR/Box plot outlier detection method

5.2.1 Experiment 01:

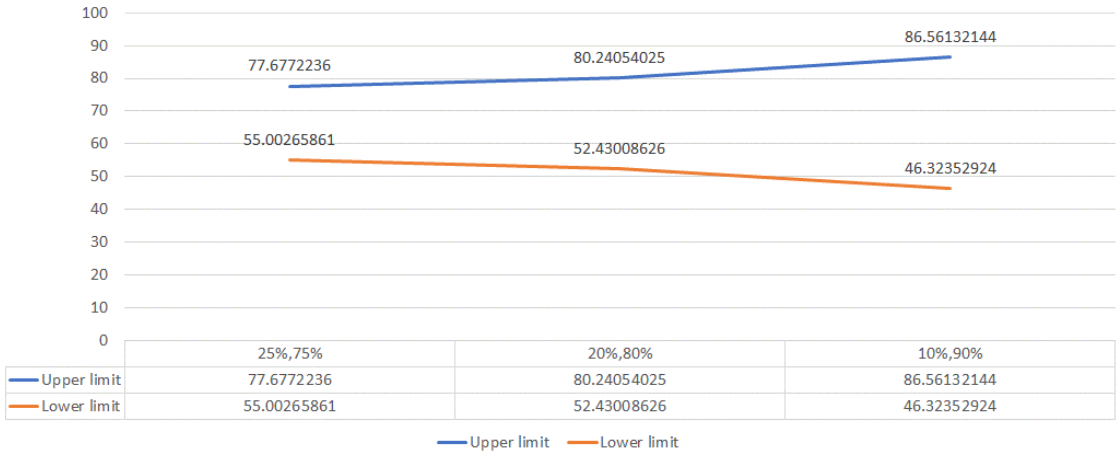


	1.5	2	1
Series1	55	52.17	57.84
Series2	77.68	80.51	74.84

Fig 5.1: Result analysis of IQR Box plot 01

From the figure 5.2.1 we can see different results for different constant values. Here, only 1.5 gives the perfect output and detect the outlier perfectly as required. For other constants like 2 and 1 we can't detect outliers. So, the previous lecturer constant gives the best result.

5.2.2 Experiment 02:



	25%,75%	20%,80%	10%,90%
Upper limit	77.6772236	80.24054025	86.56132144
Lower limit	55.00265861	52.43008626	46.32352924

Fig 5.2: Result analysis of IQR Box plot 02

Here we had experimented on the values of Q1 and Q2. The literature gives the value in 25% and 75%. In my experiment I had used 20%,80% and also 10%-90%. We didn't get better result than the previous result in my experiment.

5.3 Scatter plot outlier detection method

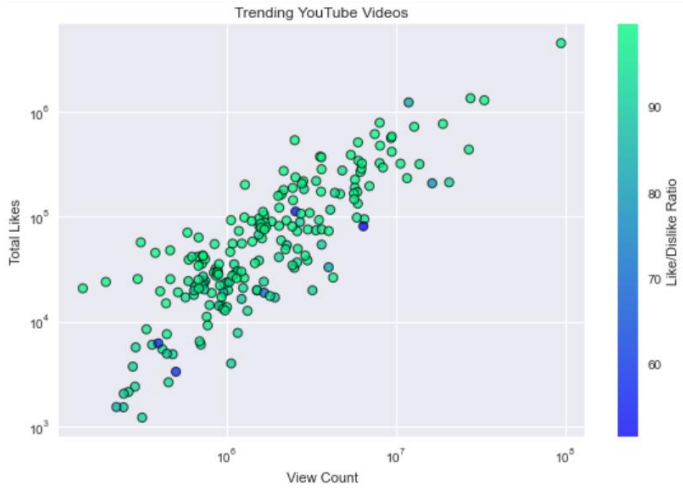


Fig 5.3: Result analysis of Scatter plot

In this we had implemented the algorithm. Here we have a dataset with almost 200 data. There is also outlier exists. The implementation shows that outlier on the top right corner in fig 5.2.2.1.

5.4 Z-score outlier detection method

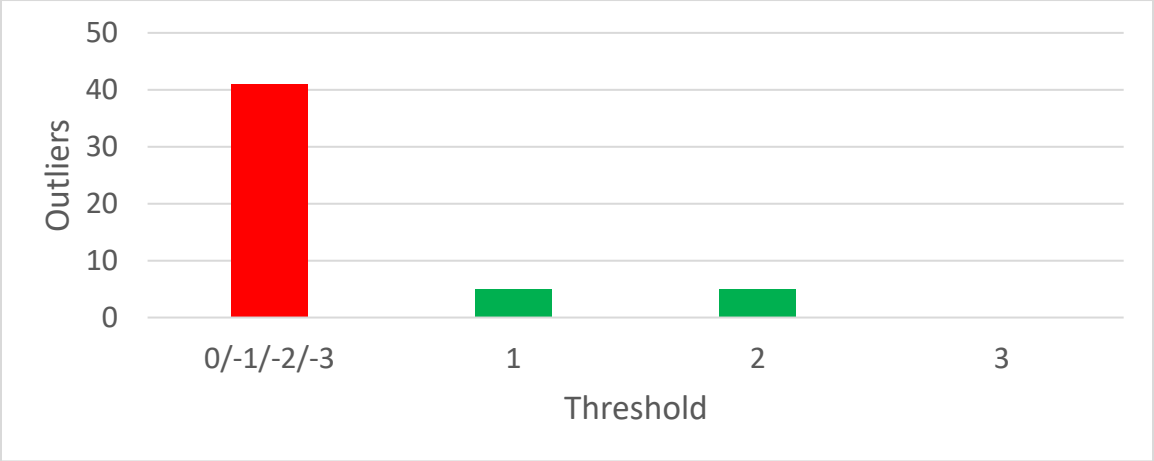


Fig 5.4: Result analysis of Z-score outlier detection method

Fig 5.4.1 shows the outlier detection with changing value of threshold. For Negative or 0 value we can't detect outliers. The program shows all the data as outlier. For threshold value 1&2 we have equal number of outliers. The previous literature shows the outlier value for only threshold value 2 but after my experiment on the algorithm I found that threshold value 1 also gives perfect detection of outlier. For threshold value 3 we didn't get any result. The outlier array was totally empty. So, it also can't be a good outlier detection threshold value.

5.5 Conclusion

The conclusion of a result analysis and discussion in outlier detection should summarize the key findings of the study and discuss their implications. It should also highlight any limitations of the study and suggest directions for future research.

Some key points that could be included in the conclusion of a result analysis and discussion in outlier detection might include:

- ✓ The effectiveness of the outlier detection method used have to successfully identify outliers in the data.
- ✓ The impact of the outliers on the overall analysis.
- ✓ The implications of the results.
- ✓ Limitations of the study.
- ✓ Directions for future research

Overall, the conclusion of a result analysis and discussion in outlier detection should provide a clear and concise summary of the key findings and their implications, as well as outline any limitations and suggest directions for future research.

Chapter 6

FUTURE WORK AND CONCLUSION

6.1 Introduction

The future work and conclusion section of a study on outlier detection in data mining should provide a summary of the key findings and discuss the implications of the results. It should also outline any limitations of the study and suggest directions for future research. In the future work section, you might consider discussing potential ways to improve or extend the outlier detection method used in the study. This could include exploring different algorithms or techniques, or looking at ways to incorporate additional information or context into the analysis.

The conclusion should provide a summary of the key findings of the study and discuss their implications for the field of data mining. This might include discussing the impact of the outliers on the overall analysis, and any patterns or trends that were identified in the data.

It is also important to highlight any limitations of the study, such as constraints on the data or methods used. This can help to contextualize the results and provide a clear understanding of what the study was able to achieve.

Finally, the conclusion should suggest directions for future research, highlighting areas where additional work could be done to better understand outlier detection in data mining. This might include exploring different approaches to outlier detection, or studying the impact of outliers on specific types of data or applications.

6.2 Future work

There are many potential areas of future work in outlier detection in data mining. Some examples might include:

- ✓ Data set implementation in real time
- ✓ More algorithms analysis to detect outlier
- ✓ Incorporating additional context or information: Many outlier detection methods rely on statistical characteristics of the data, but incorporating additional context or information about the data could improve the accuracy of outlier detection. For example, incorporating domain knowledge or prior knowledge about the data could help to identify outliers that may not be apparent using statistical methods alone.
- ✓ Exploring different outlier detection algorithms: There are many different algorithms and techniques that can be used for outlier detection, and there is ongoing research into developing and improving these methods. Future work could focus on comparing different algorithms and identifying the most effective ones for different types of data and applications.
- ✓ Studying the impact of outliers on specific types of data: Different types of data may be more or less sensitive to the presence of outliers. Future work could focus on studying the impact of outliers on specific types of data, such as time series data or text data, to better understand how to detect and handle outliers in these contexts.
- ✓ Incorporating outlier detection into machine learning pipelines: Outlier detection is often used as a pre-processing step in machine learning, but there is ongoing research into integrating outlier detection into machine learning pipelines in a more seamless and automated way.
- ✓ Developing methods for handling outliers: In addition to detecting outliers, it is also important to consider how to handle them once they have been identified. Future work could focus on developing methods for dealing with outliers, such as removing them from the data or replacing them with more appropriate values.

6.3 Conclusion

There is a huge data available in the world. With the needed data there also some unwanted or noisy data. My thesis is to analysis algorithms to detect the outlier or unwanted observations from the dataset and produce only the required data. Outliers have a forward-looking approach to the broad picture and work to connect the dots between diverse information so that the team and organization may stay far ahead. Team culture shouldn't be governed by fear and insecurity since outliers are inherently disruptive and progressive.

The outcomes of statistical modelling and data analysis might be affected by outliers. Some effects of outliers in the data set are listed below: The mean and the standard deviation may be significantly impacted. The normalcy may be reduced if the outliers are not dispersed randomly.

REFERENCES

[1] Data Mining: sas, Oct. 2022. [Online]. Available: https://www.sas.com/en_us/insights/analytics/data-mining.html#:~:text=Data%20mining%20is%20the%20process,relationships%2C%20reduce%20risks%20and%20more

[2] Outlier: geeksforgeeks, Jul. 2021. [Online]. Available: https://www.geeksforgeeks.org/types-of-outliers-in-data-mining/

[3] Outlier Detection: encyclopedia, Oct. 2022. [Online]. Available: https://encyclopedia.pub/entry/7007#:~:text=The%20purpose%20of%20outlier%20detection,data%20points%20in%20a%20dataset

[4] Detection of Outliers: nist, Jun. 2022. [Online]. Available: https://www.itl.nist.gov/div898/handbook/eda/section3/eda35h.htm

[5] A. Ayadi, O. Ghorbel, A. M. Obeid, and M. Abid, ‘‘Outlier detection approaches for wireless sensor networks: A survey,’’ Comput. Netw., vol. 129, pp. 319–333, Dec. 2017.

[6] V. J. Hodge and J. Austin, ‘‘A survey of outlier detection methodologies,’’Artif. Intell. Rev., vol. 22, no. 2, pp. 85–126, 2004.

[7] Schwertman, Neil C., Margaret Ann Owens, and Robiah Adnan. "A simple more general boxplot method for identifying outliers." Computational statistics & data analysis 47.1 (2004): 165-174.

[8] C. C. Aggarwal, Outlier Analysis. 2nd Ed. New York, NY, USA: Springer, 2016.

[9] A. S. Hadi, R. Imon, and M. Werner, Detection of Outliers, vol. 1, no. 1. Hoboken, NJ, USA: Wiley, 2009, pp. 57–70.

[10] V. Barnett and T. Lewis, Outliers in Statistical Data. Hoboken, NJ, USA: Wiley, 1994.

[11] M. F. Jiang, S. S. Tseng, and C. M. Su, ‘‘Two-phase clustering process for outliers’ detection,’’ Pattern Recognit. Lett., vol. 22, pp. 691–700, May 2011.

[12] E. Achtert, H. P. Kriegel, L. Reichert, E. Schubert, R. Wojdanowski, and A. Zimek, ‘‘Visual evaluation of outlier detection models,’’ in Proc. 15th Int. Conf. Database Syst. Adv. Appl. (DASFAA), 2010, pp. 396–399.

[13] Alberto Quesada, Artelnics. Three methods to deal with outliers, Artelnics, Machine Learning Blog. Accessed: Feb. 20, 2018. [Online]. Available: https://www.neuraldesigner.com/blog/3_methods_to_deal_with_outliers

[14] Schwertman, Neil C., Margaret Ann Owens, and Robiah Adnan. "A simple more general boxplot method for identifying outliers." Computational statistics & data analysis 47.1 (2004): 165-174.

[15] N. R. Adam, V. P. Janeja, and V. Atluri. Neighborhood-based detection of anomalies in high-dimensional spatio-temporal sensor datasets. In Proceedings of the 2004 ACM symposium on Applied computing, pages 576–583, Nicosia, Cyprus, 2004

[16] Shao, Lin, et al. "Interactive regression lens for exploring scatter plots." Computer graphics forum. Vol. 36. No. 3. 2017.

[17] Everingham, Mark, et al.” The pascal visual object classes challenge: A retrospective.” International Jour- nal of Computer Vision 111.1 (2015): 98-136.

[18] Long, Jonathan, Evan Shelhamer, and Trevor Darrell.” Fully onvolutional networks for semantic segmen- tation.” Proceedings of the IEEE Conference on Com- puter Vision and Pattern Recognition. 2015.

[19] Aggarwal, Vaibhav, et al. "Detection of spatial outlier by using improved Z-score test." 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI). IEEE, 2019.