Volume 9, Issue 08, August 2024 ISSN (online): 2456-0006

International Journal of Science Technology Management and Research

Available online at: www.ijstmr.com

Analysis and Identify Correct Outlier and Providing Statistical Treatment to Avoid Information Loss

Kapish bhaydia
M Tech CSE 4th sem
LNCT (Bhopal) Indore Campus
Indore M.P. India

Dikshika Maliwad
Assistant Professor CSE department
LNCT (Bhopal) Indore Campus
Indore M.P. India

Abstract: ***An outlier is an individual point of data that is distant from other points in the dataset. It is an anomaly in the dataset that may be caused by a range of errors in capturing, processing or manipulating data. Outliers can skew overall data trends, so outlier detection methods are an important part of statistics. In this paper we used boxplot approach inter quartile range to find out outliers. Our main contributions are, we used inter quartile range(IQR) approach in which IQR also called the mid spread or middle, or technically H-spread, is a measure of statistical dispersion, being equal to the difference between 75th and 25th percentiles, or between upper and lower quartiles, IQR = Q3 − Q1. It is a trimmed estimator, defined as the 25% trimmed range, and is a commonly used robust measure of scale. The IQR is a measure of variability, based on dividing a data set into quartiles. Quartiles divide a rank-ordered data set into four equal parts. The values that separate parts are called the first, second, and third quartiles; and they are denoted by Q1, Q2, and Q3, respectively. Quartiles are calculated recursively, by using median. If the number of entries is an even number 2n, then the first quartile Q1 is defined as First quartile Q1 = median of the n smallest entries Third quartile Q3 = median of the n largest entries If the number of entries is an odd number 2n+1, then the first quartile Q1 is defined as First quartile Q1 = median of the n smallest entries The third quartile Q3 = median of the n largest entries The second quartile Q2 is the same as the ordinary median Set lower and upper limit as decide the outlier. We used Python language to implement the proposed work;***

Keywords: Outlier, Quartiles, Boxplot, Mean, Median, Dispersion

I. INTRODUCTION

The important role of data analytics is to involve in long process, long before the actual analysis phase begins. In fact, up to two-thirds of the time taken in the data analytics process is spent cleaning what's known as "dirty" data, data that needs to be edited, worked on, or otherwise manipulated before it's suitable for analysis. During the cleaning phase, a data analyst may find outliers in the "dirty" data, which leads to either removing them from the dataset entirely, or handling them in another way. In data analytics, outliers are values within a dataset that vary greatly from the others they're either much larger, or significantly smaller. Outliers may indicate variability in a measurement, experimental errors, or a novelty. In a real-world example, the average height of a giraffe is about 16 feet tall. However, there have been recent discoveries of two giraffes that stand at 9 feet and 8.5 feet, respectively. These two giraffes would be considered outliers in comparison to the general giraffe population[8,9].

When going through the process of data analysis, outliers can cause anomalies in the results obtained. This means that they require some special attention and, in some cases, will need to be removed in order to analyze data effectively.

There are two main reasons why giving outliers' special attention is a necessary aspect of the data analytics process:

1. Outliers may have a negative effect on the result of an analysis
2. Outliers or their behavior may be the information that a data analyst requires from the analysis

Outlier detection is a key consideration within the development and deployment of machine learning algorithms. Models are often developed and leveraged to perform outlier detection for different organizations that rely on large datasets to function. Economic modeling, financial forecasting, scientific research, and ecommerce campaigns are some of the varied areas that machine learning-driven outlier detection is used. Identifying and dealing with outliers is an integral part of working with data, and machine learning is no different. Algorithm development usually relies on huge arrays of training data to achieve a high level of accuracy. Once deployed, models will process huge amounts of data, providing insights into trends and patterns. In this data-rich environment, organizations can expect to have to deal with outlier data. Outliers can skew trends and have a serious impact on the accuracy of models. The presence of outliers can be a sign of concept drift, so ongoing outlier analysis in machine learning is needed. Machine learning models learn from data to understand the trends and relationship between data points. Outliers can skew results, and anomalies in training data can impact overall model effectiveness. Outlier detection is a key tool in safeguarding data quality, as anomalous data and errors can be removed and analyzed once identified.

Outlier detection is an important part of each stage of the machine learning process. Accurate data is integral during the development and training of algorithms, and outlier detection is performed after deployment to maintain the effectiveness of models. This guide explores the basics of outlier detection techniques in machine learning, and how they can be applied to identify different types of outlier.

II. IDENTIFY OUTLIERS USING VISUALIZATIONS

Following are common methods are used for identifying outliers [9,10,11]

A. *Identify outliers using visualizations*

We must know that how each type of outlier can be categorized. We must also know that how we can detect and handle outliers. With small datasets, it can be easy to spot outliers manually (for example, with a set of data being 28, 26, 21, 24, 78, you can see that 78 is the outlier) but when it comes to large datasets or big data, other tools are required. There are some methods commonly used to identify outliers with visualizations or statistical methods, but there are many others available for implementation into data analytics process. The method that we end up using will depend on the type of dataset working with.

B. *Identify outliers using visualizations*

In data analytics, analysts create data visualizations to present data graphically in a meaningful and impactful way, in order to present their findings to relevant stakeholders. These visualizations can easily show trends, patterns, and outliers from a large set of data in the form of maps, graphs and charts. We can read more about the different types of data visualizations, but here are two that a data analyst could use in order to easily find outliers.

C. *Identifying outliers with box plots*

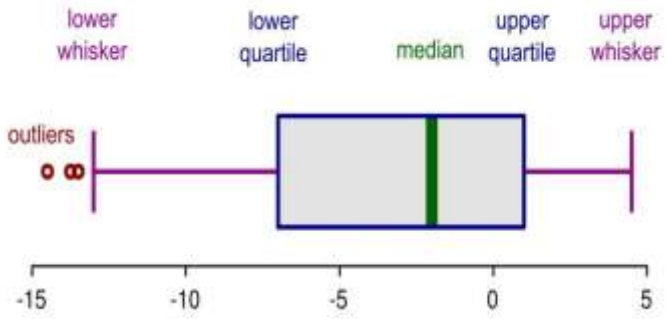

Fig 1 box plot approach for outlier detection

Visualizing data as a box plot makes it very easy to spot outliers. A box plot will show the "box" which indicates the interquartile range (from the lower quartile to the upper quartile, with the middle indicating the median data value) and any outliers will be shown outside of the "whiskers" of the plot, each side representing the minimum and maximum values of the dataset, respectively. If the box skews closer to the maximum whisker, the prominent outlier would be the minimum value. If the box skews closer to the minimum-valued whisker, the prominent outlier would then be the maximum value. Box plots can be produced easily using Excel or in Python, using a module such as Plotly.

D. *Identifying outliers with scatter plots*

As the name suggests, scatter plots show the values of a dataset "scattered" on an axis for two variables. The visualization of the scatter will show outliers easily these will be the data points shown furthest away from the regression line (a single line that best fits the data). As with box plots, these types of visualizations are also easily produced using Excel or in Python.

A.

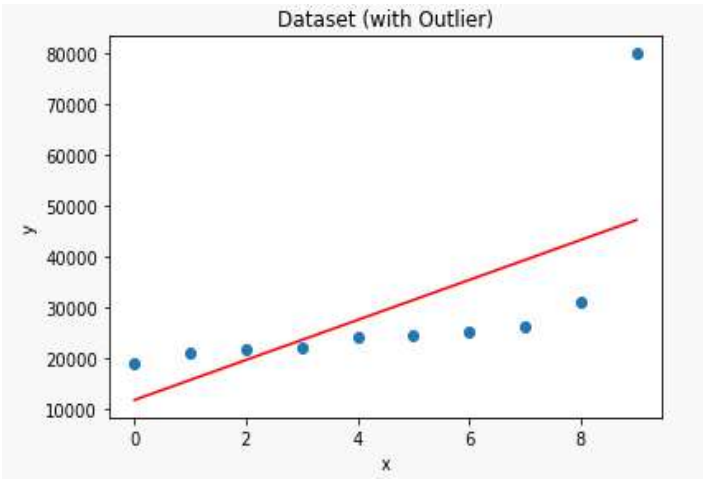

Fig 2 Scatter plot approach for outlier detection

E. *Identify outliers using statistical methods*

There are some commonly-used statistical methods are available for finding outliers. A data analyst may use a statistical method to assist with machine learning modeling, which can be improved by identifying, understanding, and in some cases removing outliers.

F. Identifying outliers with DBSCAN

DBSCAN (Density Based Spatial Clustering of Applications with Noise) is a clustering method that's used in machine learning and data analytics applications. Relationships between trends, features, and populations in a dataset are graphically represented by DBSCAN, which can also be applied to detect outliers.DBSCAN is a density-based clustering non-parametric algorithm, focused on finding and grouping together neighbors that are closely packed together. Outliers are marked as points that lie alone in low-density regions, far away from other neighbors.

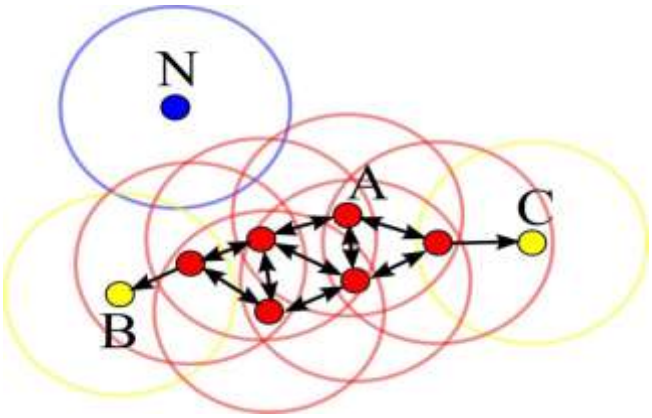

Fig 3 DBSCAN approach for outlier detection

Illustration of a DBSCAN clusters analysis. Points around A are core points. Points B and C are not core points, but are density-connected via the cluster of A (and thus belong to this cluster). Point N is Noise, since it is neither a core point nor reachable from a core point.

APPLICATIONS OF OUTLIER DETECTION

Outlier's detection can be applied on lot of data sets for various purposes. Some of which are discussed below:

- **Fraud detection** - Detecting fraudulent applications for credit cards, state benefits or detecting fraudulent usage of credit cards or mobile phones. Fraud refers to criminal activities occurring in commercial organizations such as banks, credit card companies, insurance agencies, cell phone companies, stock market, etc. Malicious users could be actual customers of the organization or resorting to identity theft (posing as customers). The detection activity aims at detection of unauthorized consumption of resources provided by the organization to prevent economic losses.
- **Fraudulent Usage of Credit Card:** Associated with credit card thefts. The data records are defined over several dimensions such as the user ID, spent amount, time between consecutive card usage, etc. The frauds are typically reflected in transactional records (point outliers) and correspond to high payments; high rate of purchase, purchase of items never purchased by 27 the user before, etc. Availability of labeled records is no problem since credit companies have complete data available. Moreover, the data falls into distinct profiles based on the credit card user. Hence profiling and clustering based techniques are typically used in this domain.
- **Intrusion detection** - Detecting unauthorized access in computer networks. Intrusion detection refers to detection of malicious activity (break-ins, penetrations, and other forms of computer abuse) in a computer related system interesting from a computer security perspective. Being different from normal system behavior, intrusion detection is a perfect candidate for applying outlier detection techniques.

III. LITERATURE REVIEW

In 2017 Rasim M. Alguliyev et al "An Anomaly Detection Based on Optimization". At present, anomaly detection is one of the important problems in many fields. The rapid growth of data volumes requires the availability of a tool for data processing and analysis of a wide variety of data types. The methods for anomaly detection are designed to detect object's deviations from normal behavior. It is difficult to select one tool for all types of anomalies due to the increasing computational complexity and the nature of the data. They proposed an improved optimization approach for a previously known number of clusters, where a weight is assigned to each data point, is proposed. Their aim is to show that weighting of each data point improves the clustering solution. The experimental results on three datasets show that the proposed algorithm detects anomalies more accurately. They compared to the k-means algorithm. The quality of the clustering result was estimated using clustering evaluation metrics. They proposed a new approach to detect abnormal values in Big data. The aim of the algorithm presented in the paper is to improve the anomaly detection. This research, the weights were assigned to each point (instance) and determine the relative position of this point in the entire data set [1].

In 2018 Victoria J. Hodge et al proposed "An Evaluation of Classification and Outlier Detection Algorithms". They evaluate algorithms for classification and outlier detection accuracies in temporal data. They focus on algorithms that train and classify rapidly and can be used for systems that need to incorporate new data regularly. They compare the accuracy of six fast algorithms using a range of well-known time-series datasets. The analyses demonstrate that the choice of algorithm is task and data specific but that can derive heuristics for choosing. Gradient Boosting Machines are generally best for classification but there is no single winner for outlier detection though Gradient Boosting Machines (again) and Random Forest is better. Hence, we recommend running evaluations of a number of algorithms using our heuristics." They evaluated algorithms for both classification and outlier detection for an on-line system that assimilates new data regularly. They aimed to derive heuristics for the best algorithms. For a dataset with complex structure (non-linearity) where an

ensemble method will fail and a very specific approach is needed, then k-NN or C4.5 are better. It is not possible to derive a simple heuristic for outlier detection indicating that a number of algorithms need to be evaluated [2].

In 2018 Aurore Archimbaud et al proposed "ICSOutlier: Unsupervised Outlier Detection for Low-Dimensional Contamination Structure" Detecting outliers in a multivariate and unsupervised context is an important and ongoing problem notably for quality control. In this particular context, the Invariant Coordinate Selection (ICS) method shows remarkable properties for identifying outliers that lie on a low-dimensional subspace in its first invariant components. It is implemented in the ICS Outlier package. The main function of the package, ics. Outlier offers the possibility of labelling potential outliers in a completely automated way. Four examples, including two real examples in quality control, illustrate the use of the function. Comparing with several other approaches, it appears that ICS is generally as efficient as its competitors and shows an advantage in the context of a small proportion of outliers lying in a low-dimensional subspace. The cut-off for outlier identification can be adjusted by looking at the ICS distances plot. The ability of the method to check for the absence of outliers is an advantage compared with methods such as the Mahalanobis distance. The real HTP data set included in the package shows that the ICS method is useful for outlier detection in the context of quality control, among possible applications. Finally the ICS method competes with common approaches based on distances (the Mahalanobis distance, its robust version, the PCA methods, its variants and improvements) as well as other methods based on the density (LOF) or on angles (ABOD)[3].

In 2018 Charmgil Hong and Milos Hauskrecht proposed "Multivariate Conditional Outlier Detection: Identifying Unusual Input-Output Associations in Data". They studied multivariate conditional outlier detection, a special type of the conditional outlier detection problem, where data instances consist of continuous input (context) and binary output (responses) vectors. They presented a novel outlier detection framework that identifies abnormal input-output associations in data using a decomposable conditional probabilistic model. Since the components of this model can vary in their quality, they combined them with the help of weights reflecting their reliability in assessment of outliers. They proposed two ways of calculating the component weights: global that relies on all data and local that relies only on the instances similar to the target instance. They also presented a probabilistic framework for the multivariate conditional outlier detection (MCOD) problem that relies on a decomposable model of conditional joint probability, where data instances that are assigned a low probability by the model are considered to be outliers [4].

In 2019 Yue Zhao et al proposed PyOD: A Python Toolbox for Scalable Outlier Detection PyOD: A Python Toolbox for Scalable Outlier Detection. PyOD is an open-source Python toolbox for performing scalable outlier detection on multivariate data. Uniquely, it provides access to a wide range of outlier detection algorithms, including established outlier ensembles and more recent neural network-based approaches, under a single, well-documented API designed for use by both practitioners and researcher. With robustness and scalability in mind, best practices such as unit testing, continuous integration, code coverage, maintainability checks, interactive examples and parallelization are emphasized as core components in the toolbox's development. PyOD is compatible with both Python 2 and 3 and can be installed through Python Package Index (PyPI). They presented PyOD, a comprehensive toolbox built in Python for scalable outlier detection. It includes more than 20 classical and emerging detection algorithms and is being used in both academic and commercial projects. As avenues for future work, we plan to enhance the toolbox by implementing models that work well with time series and geospatial data, improving computational efficiency through distributed computing and addressing engineering challenges such as handling sparse matrices or memory limitations [5].

In 2019 Agnieszka Duraj et al proposed "Detection of outliers in data streams using grouping methods" Efficient processing of data streams usually requires their initial processing, including on the removal of anomalies caused by, for example, measuring errors. Such errors may result in misinterpretation of the phenomena being analyzed .The literature describes several methods for detecting exceptions in data streams. Each of them requires proper selection of operating parameters. In addition, the effectiveness of methods may vary depending on the data set being analyzed. The article describes current methods for detecting exceptions in data streams and analyzed their operation on gas consumption data[6].

In 2020 Atiqur Rehman proposed "Unsupervised outlier detection in multidimensional data" Detection and removal of outliers in a dataset is a fundamental preprocessing task without which the analysis of the data can be misleading. In order to detect the anomalies in a dataset in an unsupervised manner, some novel statistical techniques are proposed. The proposed techniques are based on statistical methods considering data compactness and other properties. They proposed ideas are found efficient in terms of performance, ease of implementation, and computational complexity. Furthermore, two proposed techniques presented in this paper use transformation of data to a uni dimensional distance space to detect the outliers, so irrespective of the data's high dimensions, the techniques remain computationally inexpensive and feasible. Comprehensive performance analysis of the proposed anomaly detection schemes is presented in the paper, and the newly proposed schemes are found better than the state-of-the-art methods when tested on several benchmark datasets[7].

IV. PROBLEM STATEMENT

An outlier can cause serious problems in statistical analyses. Outlier points can therefore indicate faulty data, erroneous procedures, or areas where a certain theory might not be valid Outliers are unusual values in dataset, and they can distort statistical analyses and violate their assumptions. As lot of outlier detection algorithms exists for detecting outliers and the usage of all these vary according to the type. Some of the problem that needs to be considers when we have outlier in the data set these are

1. First we need to decide the method or tools for founding outliers in the data set.
2. How the outlier affect the mean, variance and standard deviation of the given dataset.
3. If the outliers are non-randomly distributed, they can decrease normality of data set.
4. Outlier are reduces the power of statistical tests.

5. In a data distribution, with extreme outliers, the distribution is skewed in the direction of the outliers which makes it difficult to analyze the data.
6. Should we remove outliers? Removing outliers is legitimate only for specific reasons. Outliers can be very informative about the subject-area and data collection process

PROPOSED APPROACH

The proposed is based on inter quartile range(IQR).In descriptive statistics, the inter quartile range (IQR), also called the mid spread or middle, or technically H-spread, is a measure of statistical dispersion, being equal to the difference between 75th and 25th percentiles, or between upper and lower quartiles, IQR = Q3 − Q1. In other words, the IQR is the first quartile subtracted from the third quartile; these quartiles can be clearly seen on a box plot on the data. It is a trimmed estimator, defined as the 25% trimmed range, and is a commonly used robust measure of scale. The IQR is a measure of variability, based on dividing a data set into quartiles. Quartiles divide a rank-ordered data set into four equal parts. The values that separate parts are called the first, second, and third quartiles; and they are denoted by Q1, Q2, and Q3, respectively.
Quartiles are calculated recursively, by using median.
If the number of entries is an even number 2n, then the first quartile Q_1 is defined as
First quartile Q_1 = median of the n smallest entries
Third quartile Q_3 = median of the n largest entries
If the number of entries is an odd number 2n+1, then the first quartile Q_1 is defined as
First quartile Q_1 = median of the n smallest entries
The third quartile Q_3 = median of the n largest entries
The second quartile Q_2 is the same as the ordinary median

V. RESULT AND ANALYSIS

Boxplot for total_bill with Q1,Q3, IQR and Lowe bound, Upper bound
We calculate first and thirds quartile and then we found interquartile range after that we calculate the upper and lower bond value to found the outliers for total_bill here we can see that the value of first quartile is 13.34 , value of third quartile is 24.127 and the value of IQR is 10.7799. The value of Lowe bound is -2.822499 and Upper bound is 40.297499

TABLE 1
Value of Q1, Q3, IQR, Lowe bound, Upper bound for total_bill

Q1	Q3	IQR	Lowe bound	Upper bound
13.3475	24.12749	10.7799	-2.822499	40.297499

print(sns.boxplot(data1["total_bill"]))

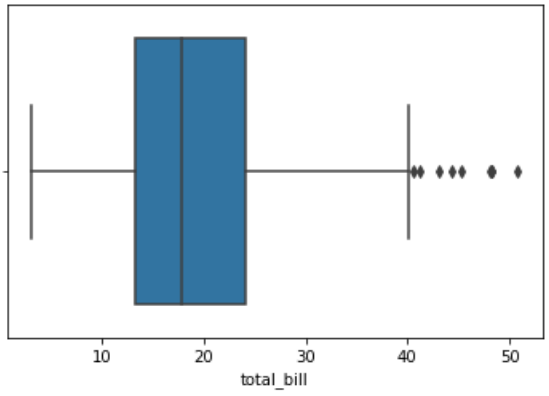

Fig 4 Boxplot for total_bill

5.2 Boxplot for tips with Q1,Q2, IQR and Lowe bound, Upper bound
We calculate first and thirds quartile and then we found interquartile range after that we calculate the upper and lower bond value to found the outliers for total_bill here we can see that the value of first quartile is 2.0, value of third quartile is 3.56249 and the value of IQR is 1.5625. The value of Lower bound is -0.34375and Upper bound is 5.90625

TABLE 2
Value of Q1, Q3, IQR, Lowe bound, Upper bound for tips

Q1	Q2	IQR	Lower bound	Upper bound
2.0	3.56249	1.5625	-0.34375	5.90625

print(sns.boxplot(data1["total_bill"]))

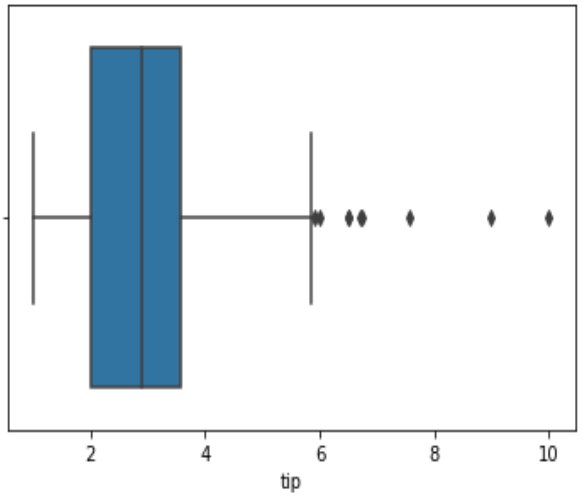

Fig 5 Boxplot for tips

5.3 Boxplot for size with Q1,Q2, IQR and Lowe bound, Upper bound

We calculate first and thirds quartile and then we found interquartile range after that we calculate the upper and lower bond value to fou nd the outliers for total_bill here we can see that the value of first quartile is 2.0, value of third quartile is 3.0and the value of IQR is 1.0. The value of Lower bound is 0.5 and Upper bound is 4.5

TABLE 3
Value of Q1, Q3, IQR, Lowe bound, Upper bound for size

Q1	Q2	IQR	Lower bound	Upper bound
2.0	3.0	1.0	0.5	4.5

print(sns.boxplot(data1["size"]))

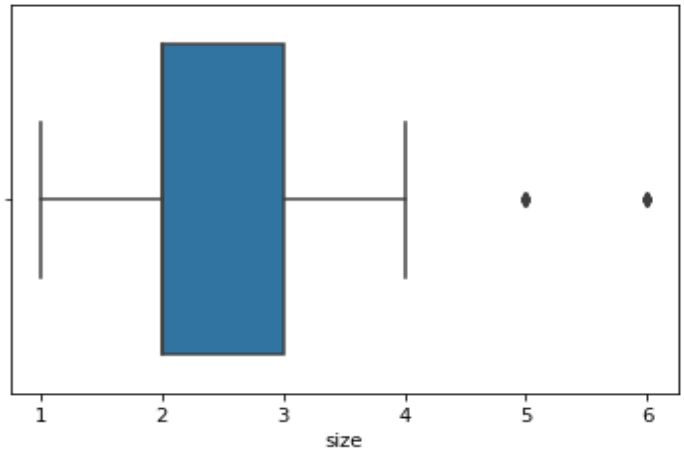

Fig 6 Boxplot for size

CONCLUSION

In this paper we compares performance proposed approach with distance based approach for outlier detections. The proposed is based on inter quartile range(IQR).In descriptive statistics, the inter quartile range (IQR), also called the mid spread or middle, or technically H-spread, is a measure of statistical dispersion, being equal to the difference between 75th and 25th percentiles, or between upper and lower quartiles, IQR = Q3 − Q1. In other words, the IQR is the first quartile subtracted from the third quartile; these quartiles can be clearly seen on a box plot on the data. It is a trimmed estimator, defined as the 25% trimmed range, and is a commonly used robust measure of scale. The IQR is a measure of variability, based on dividing a data set into quartiles. Quartiles divide a rank-ordered data set into four equal parts. The values that separate parts are called the first, second, and third quartiles; and they are denoted by Q1, Q2, and Q3, respectively. In this paper we not only detect the outlier efficiently but we also give proposer treatment to them. We replace the outliers using either upper bond or any other statistical method.

REFERENCE

1. Rasim M. Alguliyev, Ramiz M. Aliguliyev, Yadigar N. Imamverdiyev An Anomaly Detection Based on Optimization I.J. Intelligent Systems and Applications, 2017, 12, 87-96 Published Online December 2017 in MECS (http://www.mecs-press.org/) DOI: 10.5815/ijisa.2017.12.08
2. Victoria J. Hodge and Jim Austin An Evaluation of Classification and Outlier Detection Algorithms Digital Creativity Labs, Department of Computer Science, University of York, UK {victoria.hodge, jim.austin}@york.ac.uk arXiv:1805.00811v1 [stat.ML] 2 May 2018
3. Aurore Archimbaud, Klaus Nordhausen, and Anne Ruiz-Gazen ICSOutlier: Unsupervised Outlier Detection for Low-Dimensional Contamination Structure The R Journal Vol. 10/1, July 2018 ISSN 2073-485
4. Charmgil Hong and Milos Hauskrecht Multivariate Conditional Outlier Detection: Identifying Unusual Input-Output Associations in Data Department of Computer Science University of Pittsburgh Pittsburgh, PA 15260 Copyright c 2018, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.
5. Yue Zhao Zain Nasrullah PyOD: A Python Toolbox for Scalable Outlier Detectio Journal of Machine Learning Research 20 (2019) 1-7 Submitted 1/19; Revised 4/19; Published 5/19 arXiv:1901.01588v2 [cs.LG] 10 Jun 2019
6. Agnieszka Duraj , Łukasz Chomątek Politechnika Łódzka, Instytut Informatyki, Wydział Fizyki Detection of outliers in data streams using grouping methods Przegląd Elektrotechniczny, ISSN 0033-2097, R. 95 NR 2/2019
7. Atiqur Rehman and Samir Brahim Belhaouari Unsupervised outlier detection in multidimensional data ur Rehman and Belhaouari J Big Data (2021) 8:80 https://doi.org/10.1186/s40537-021-00469-z atiqjadoon@gmail.com ICT Division, College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qata
8. Shahrooz Abghari Data Mining Approaches for Outlier Detection Analysis Blekinge Institute of Technology Doctoral Dissertation Series No 2020:09 2020 Department of Computer Science Publisher: Blekinge Institute of Technology SE-371 79 Karlskrona, Sweden Printed by Exakta Group, Sweden, 2020 ISBN: 978-91-7295-409-0 ISSN: 1653-2090 urn:nbn:se:bth-20454
9. Afrah Yahya AL Rezami Effect of outliers on the coefficient of determination in multiple regression analysis with the application on the GPA for student International Journal of Advanced and Applied Sciences, 7(10) 2020, Pages: 30-37 Contents lists available at Science-Gate International Journal of Advanced and Applied Sciences Journal homepage: http://www.science-gate.com/IJAAS.html
10. Saima Afzal , Ayesha Afzal , Muhammad Amin , Sehar Saleem , 4 Nouman Ali , 5 and Muhammad Sajid A Novel Approach for Outlier Detection in Multivariate Data Hindawi Mathematical Problems in Engineering Volume 2021, Article ID 1899225, 12 pages https://doi.org/10.1155/2021/1899225
11. Ishani Chatterjee , Mengchu Zhou , Abdullah Abusorrah Statistics-Based Outlier Detection and Correction Method for Amazon Customer Reviews Entropy 2021, 23, 1645. https://doi.org/10.3390/e23121645 Academic Editor: Ernestina Menasalvas Received: 23 October 2021 Accepted: 30 November 2021 Published: 7 December 2021