Pak. J. Statist.
2021 Vol. 37(4), 447-462

A REVIEW AND EMPIRICAL COMPARISON OF UNIVARIATE OUTLIER DETECTION METHODS

Sehar Saleem[§], **Maria Aslam** and **Mah Rukh Shaukat**
Department of Statistics, Lahore College for Women University
Lahore, Pakistan
[§]Email: seharsaleem87@gmail.com

ABSTRACT

Many real-world phenomena generate data sets with outliers i.e., extreme observations that are away from the mainstream of the data. The presence of outliers may cause invalid analysis by violating the conventional assumptions of regression models. Hence identification of outliers holds significant importance in data analysis. This study reviews various outlier labeling methods and shows the comparative detection of outliers by applying these methods on several real data sets with small to large sample sizes and low to high levels of skewness. Some graphical and formal methods of univariate outlier detection are also applied. All labeling methods detected no outlier for symmetric shape except adjusted boxplot. For slightly skewed distribution, Z-score, 3SD method, and 3IQR found resistance for both small and large sample sizes except adjusted boxplot which is resistant in large data only. In the case of mildly skewed and large sample size, the 2Median Absolute Deviation method shown up most sensitive. It is concluded that the Adjusted boxplot, Z-score, 3SD method, and Tukey's 3IQR (interquartile range) method detected fewer outliers among other competing methods. Boxplot and a formal Generalized ESD test identified outlying observations as well as most extreme observations.

KEYWORDS

Labeling methods, Sample size, Skewness measures, Adjusted boxplot, Generalized ESD test, Graphical techniques

1. INTRODUCTION

An outlier is an observation or number of observations that deviate significantly from the mainstream of the other observations, also known as outlying observation or contaminant. Every set of data has its own exact definition of an outlier, based on the hidden assumptions according to the data structure. Outliers might occur from many sources like an error in data compilation, editing, or coding. Outliers arise due to the changes in a system error, fraudulent behavior in financial aspects, human error and instrument error, or simply through natural variations in the samples or populations. The presence of outliers in any data set might produce invalid and spurious results. It usually violates the normality assumption in the simple linear regression technique of an ANOVA test and deals with them improperly making the statistical analysis invalid. Outliers may adversely lead to model misspecification, biased estimates, and inflated standard errors (Ben-Gal, 2005). Hence, outlier detection is a principal step in many data mining

© 2021 Pakistan Journal of Statistics

applications for univariate and multivariate presentations of data (Ben-Gal, 2005). This study is focused on the review and application of univariate outlier detection methods mainly categorized into two types: formal methods and informal methods. Formal methods are test-based methods that usually require some test statistic to test the hypothesis. In this study, Grubb's test, generalized extreme studentized deviate tests, and Dixon test are applied to test the discordancy in several real data sets with different sample sizes and levels of skewness. Informal methods include several outlier labeling methods such as; SD method, Z-score method, modified Z-score method, median Rule, MADe method, Tukey's method, and adjusted boxplot method to depict the outlying observations in real data sets with different sample sizes and levels of skewness.

2. METHODOLOGY

Several outlier detection methods including labeling methods, formal tests and graphical techniques are presented and applied on several real data sets.

2.1 Interval (Labeling) Methods

2.1.1 Z- Score

This method uses two estimators mean and SD to identify the outliers in a data set. Z-score is expressed as:

$$Z_i = \frac{x_i - \bar{x}}{sd} \tag{2.1}$$

where $x_i \sim N(\mu, \delta^2)$ and sd is the standard deviation. The fundamental rule for Z-score is; if x_i follows a normal distribution with mean μ and variance δ^2 i. e. $\{x_i \sim N(\mu, \delta^2)\}$ then Z_i follows the standard normal distribution with mean 0 and variance 1 i.e. $\{Z_i \sim N(0,1)\}$ and the absolute Z-score $|Z|$ value greater than 3 is usually marked as an outlier (Kaliyaperumal, Arumugam, & Kuppusamy, 2015). Shiffler (1988) showed that the maximum value of the Z-score depends on the sample size given as

$$Z_{max} = \frac{(n-1)}{\sqrt{n}} \tag{2.2}$$

This method is not appropriate for small data sets for labeling outliers (Iglewicz & Hoaglin, 1993).

2.1.2 Modified Z-Score

This method uses two estimators for outlier labeling, i.e., median ($\tilde{x}$) and median absolute deviation (MAD) instead of mean and SD to resolve the limitation of Z-Score in which SD can be affected by extreme observation (Seo, 2006). The modified Z-score is denoted by M_i and is calculated as

$$M_i = \frac{0.6745(x_i - \tilde{x})}{MAD} \tag{2.3}$$

where E (MAD) = 0.675 σ for large normal data

$$MAD = median\{|x_i - \tilde{x}|\} \tag{2.4}$$

where $\tilde{x}$ is the sample median and MAD being the sample median absolute deviation. Iglewicz and Hoaglin (1993) proposed that the absolute values of M_i greater than 3.5 i.e. $|M_i| > 3.5$, the observation is considered as an outlier by simulation on pseudo-normal observations for sample sizes of 10, 20, and 40 (Kaliyaperumal, Arumugam, & Kuppusamy, 2015).

2.1.3 Standard Deviation Method

The standard deviation method uses two less robust estimators to identify outliers. These estimators are mean and standard deviation (highly affected by extreme observations) defined as

$$2SD\ Method: \bar{x} \pm 2SD \tag{2.5}$$

$$3SD\ Method: \bar{x} \pm 3SD \tag{2.6}$$

where $\bar{x}$ is the mean and SD is the standard deviation. The observations that do not fall in these intervals are considered outliers (Olewuezi, 2011). This method applies to symmetric data following a normal distribution. According to Chebyshev inequality, if a random variable X exists, it has mean μ and variance δ^2 for any $k > 0$

$$P\ [|X - \mu| \geq k\delta] \leq \frac{1}{K^2} \tag{2.7}$$

$$P\ [|X - \mu| \leq k\delta] \geq 1 - \frac{1}{K^2} \tag{2.8}$$

Inequality $(1 - \frac{1}{K^2})$ tells us the proportion of data that will be within k SD of the mean. Chebyshev inequality has limited application in a way that it gives the smallest proportion of the data. The SD method is quite powerful for large normal data (Seo, 2006).

2.1.4 Tukey's Method (Boxplot)

Tukey (1977) introduced the rules for the construction of a boxplot to screen outliers. This method practices quartiles instead of mean and SD. Tukey's method applies to both symmetric and skewed data. It is less sensitive to extreme observations. This method has the following rules:

$$[Q_1 - 1.5IQR, Q_3 + 1.5IQR] \tag{2.9}$$

$$[Q_1 - 3IQR, Q_3 + 3IQR] \tag{2.10}$$

Q_1 is the first quartile and Q_3 is the third quartile. IQR (interquartile range) is the difference between inner fences and outer fences. The interval with 1.5 IQR called inner fences are situated below Q_1 and above Q_3 at 1.5 IQR distance and the interval with 3 IQR called outer fences are situated below Q_1 and above Q_3 at 3IQR distance. The observations among the inner and outer fences are potential outliers and the observations are considered as a possible outlier that lies outside the outer fences. This method can detect more observations as outliers as the measure of skewness in the data increases (Seo, 2006).

2.1.5 Adjusted Boxplot

Vanderviere and Huber (2004) presented an adjusted boxplot by using medcouple (MC), a robust measure of skewness. If X_n is an independent set from a continuous univariate distribution the MC is defined as

$MC = med\ h(x_i, x_j)$ where h is the Kernel function

$$h(x_i, x_j) = \frac{(x_i - medk) - (medk - x_j)}{x_j - x_i} \tag{2.11}$$

where $medk$ is the median of X_n which satisfy $x_i \leq medk \leq x_j$ and $x_i \neq x_j$. MC ranges between -1 and 1. MC is equal to zero implies data is symmetrical and the adjusted boxplot turns into Tukey's boxplot. If MC is greater than zero, the distribution of the data would be right-skewed. On the other hand, if MC is less than zero, the distribution of the data would be left-skewed. The interval for adjusted boxplot according to MC value is as follows:

$$[L, U] = [Q_1 - 1.5\, exp(-3.5MC)\, IQR, Q_3 + 1.5\, exp(4MC)\, IQR]\ if\ MC \geq 0 \tag{2.12}$$

$$[L, U] = [Q_1 - 1.5\, exp(-4MC)\, IQR, Q_3 + 1.5\, exp(3.5MC)\, IQR]\ if\ MC \leq 0 \tag{2.13}$$

where L and U are lower and upper fences of interval respectively. IQR is the interquartile range and Q_1 is the first quartile and Q_3 is the third quartile of the data. The values that reclined beyond the interval are marked as outliers (Seo, 2006).

2.1.6 Median Absolute Deviation (MAD_e) Method

The method MAD_e is a robust technique that uses median and median absolute deviation (MAD) instead of mean and SD, as they are highly unaffected by extreme observations. This technique is defined as

$$2MAD_e\ Method: Median \pm 2MAD_e \tag{2.14}$$

$$3MAD_e\ Method: Median \pm MAD_e \tag{2.15}$$

where $MAD = (median|x_i - \tilde{x}|\ i = 1, 2, \ldots, n)$. MAD_e is a scaled median absolute deviation if we multiply median absolute deviation (MAD) with 1.483. MAD_e has a 50% breakdown point like the median. The values that lie outside the interval of 2 MAD_e and MAD_e are considered outliers (Seo, 2006).

2.1.7 Median Rule

Median is a measure of central tendency which is also a robust estimator with a 50 % breakdown point. It is located at the center of the data. If $(x_1, x_2, \ldots, x_n)$ is a random variable sorted in some order of magnitude (Olewuezi, 2011). In this method, we use the median instead of quartiles in Tukey's test and a new IQR is established; the scale of IQR is 2.3 (Olewuezi, 2011). It is defined as:

$$[C_1, C_2] = Q_2 \pm 2.3\, IQR \tag{2.16}$$

where Q_2 is the sample median. The median rule is more resistant and the target outlier percentage is less affected. The IQR can be adjusted by the target outlier percentage and Generalized Lambda Distribution (GLD). The observations which do not fall in the interval are marked as outliers. (Seo, 2006).

2.2 Graphical Methods

Histogram, normal P-P plot, and dot plot, SIQR boxplot, and ratio skewed boxplot are utilized to visualize outliers in univariate data.

2.2.1 SIQR Boxplot

Kimber (1990) introduced some adjustments in the fence for skewed data with the use of lower semi-interquartile range and upper interquartile range. The fence rule had defined as:

$$[Q_1 - 3\,SIQRL, Q_3 + 3\,SIQRU] \tag{2.17}$$

where $SIQRL = Q_3 - Q_1$ and $SIQRU = Q_3 - Q_2$. This SIQR boxplot still marked many regular observations as outliers and unable to show a significant difference between regular observation and outlier. Upper whiskers are slightly increased then fewer observations are highlighted as upper outliers and lower whiskers simply marked the smaller observations (Hubert & Vanderveiren, 2008).

2.2.2 Ratio Skewed Boxplot

Ratio skewed boxplot can be applied to univariate, symmetrical, and skewed data sets despite sample size. Kimber (1990) proposed a new rule of fences for boxplots by replacing the IQR with 2SIQR. These fences were more relevant to skewness and worked as old fences for the symmetrical underlying distribution. In the construction of fences use the Bowley's coefficient (sample quartile-based measure of skewness) introduced by Bowley (1920) such as:

$$R_L = \frac{1 - B_c}{1 + B_c} \text{ and } R_U = \frac{1 + B_c}{1 - B_c}$$

Ratio skewed fences are as follows

$$f_L^{RS} = Q_1 - 1.5\text{IQR}R_L \text{ and } f_U^{RS} = Q_3 + 1.5\text{IQR}R_U \tag{2.18}$$

where f_L^{RS} is ratio-skewed lower fence and f_U^{RS} is the ratio-skewed upper fence. By using these fences ratio skewed boxplot can be plotted (Walker, Dovoedo, Chakraborti, & Hilton, 2018).

2.3 Formal (Test-Based) Methods

2.3.1 Grubbs test

Grubbs (1950) introduced a testing procedure to test the hypothesis for a single outlier in a univariate data set. Let a sample of n observations $(x_1, x_2, \ldots, x_n)$ are assumed to be normally distributed and suppose x_n is the largest/extreme observation. It tests the null hypothesis for the presence of a single outlier in a data set against the alternative hypothesis for no outlier present (Grubbs' Outlier Test, 2019).

2.3.2 Dixon Q test

Dixon Q test was introduced by Dixon (1950) used to detect a single outlier in univariate data. It is appropriate for small samples ($n \leq 30$). This test should not be applied more than once in a data set.

2.3.3 Generalized Extreme Studentized Deviate Test

Rosner (1983) proposed a generalized Extreme Studentized Deviate (ESD) test for many outliers. A generalized ESD test can detect many outliers in a univariate data set. It tests the null hypothesis as there are no outliers present in a data with an alternative that there are up to *r* outliers present in a data set. The test statistic is

$$R_i = \frac{\max |x_i - \bar{x}|}{s} \tag{2.19}$$

where $\bar{x}$ is the mean and *s* is the standard deviation. The critical value for the test statistic R_i is

$$\lambda_i = \frac{(n-i)t_{p,\,n-i-1}}{\sqrt{\left(n-i-1+t^2_{p,\,n-i-1}\right)(n-i+1)}} \tag{2.20}$$

where $i = 1,2,\ldots,r$ and $t_{\upsilon,p}$ is the 100 p percentage point from t-distribution with υ degrees of freedom and $p = 1 - \frac{\alpha}{2(n-i+1)}$.

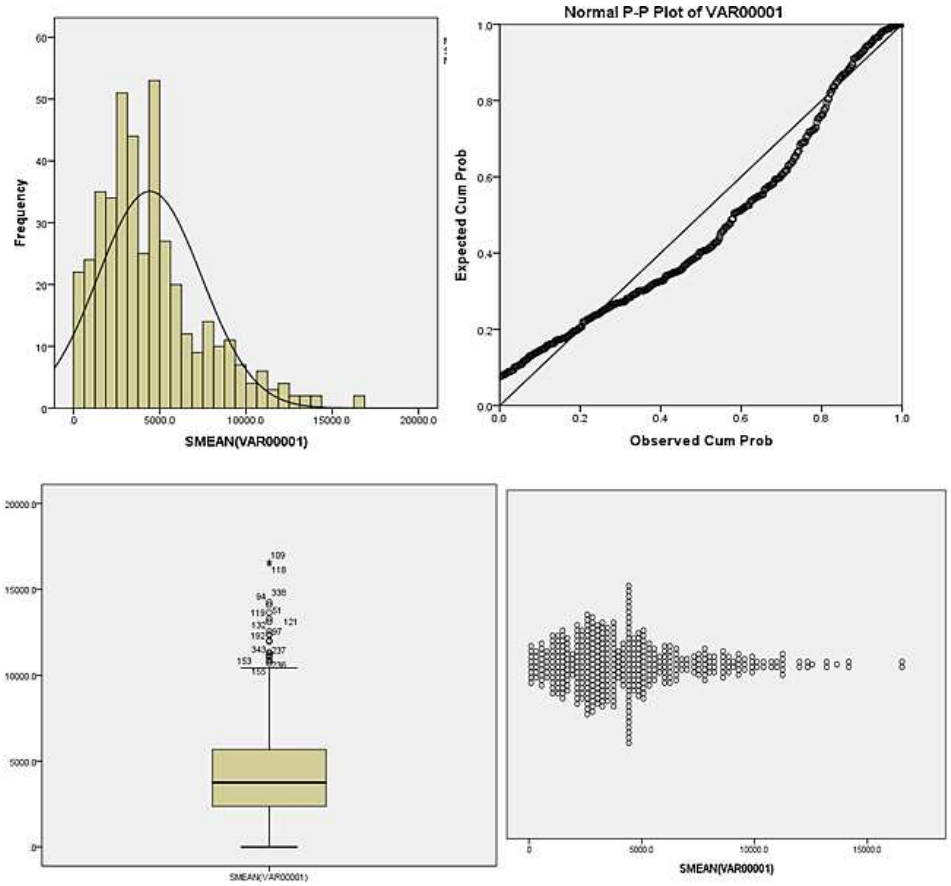
If $R_i > \lambda_i$, then the null hypothesis is rejected. The number of outliers is determined by finding the largest i such that $R_i > \lambda_i$ (Filliben & Heckart, 2013).

3. APPLICATION OF OUTLIER DETECTION METHODS

In this section, some outlier detection methods of univariate data are applied to several real data sets with different measures of skewness and small to large sample sizes. Seven outlier labeling methods and four graphical methods are applied to each data set and then compared the performance by reporting the number of outliers detected by all these methods.

3.1 Graphical Methods

Four graphical representations histogram, normal P-P plot, boxplot, and dot plot are used to highlight the extreme and potential outliers in large and small data sets. These graphs are shown in Figures 3.1, 3.2, and 3.3 for large, small, and modified (omitted two extreme outliers from small data) data sets.

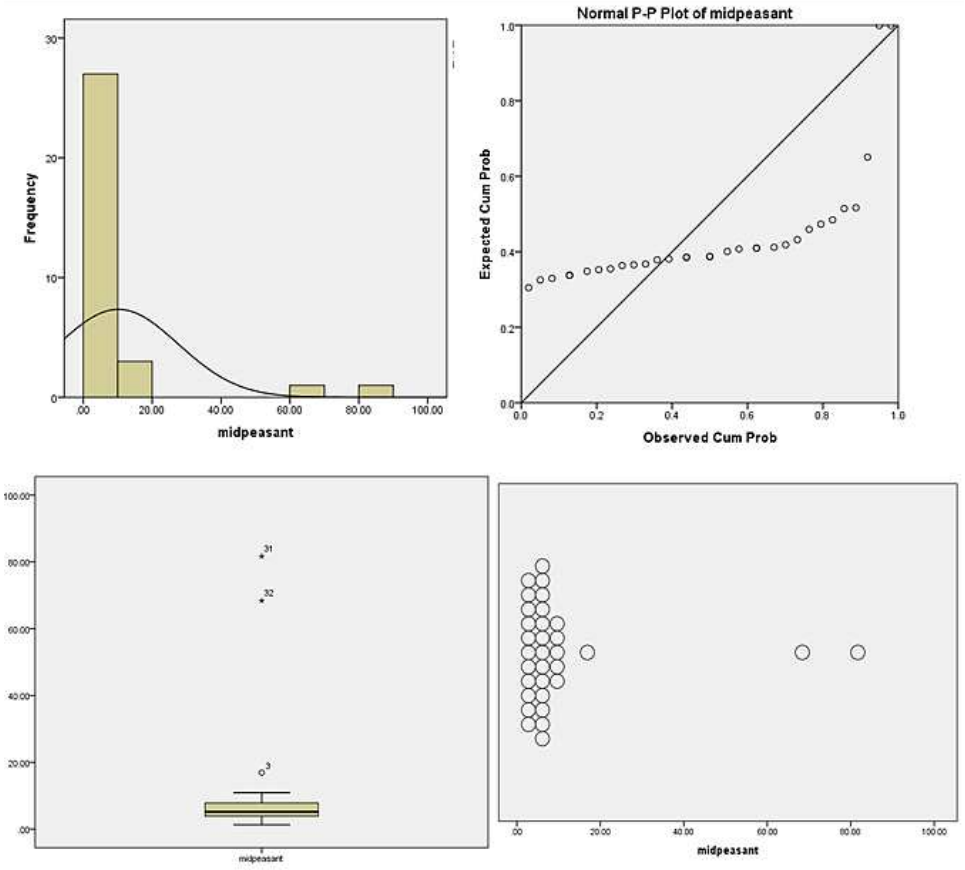


Case 1 (Skewness = 1.166)

Figure 3.1: Histogram, Normal P-P Plot, Boxplot, and Dot Plot for N = 423

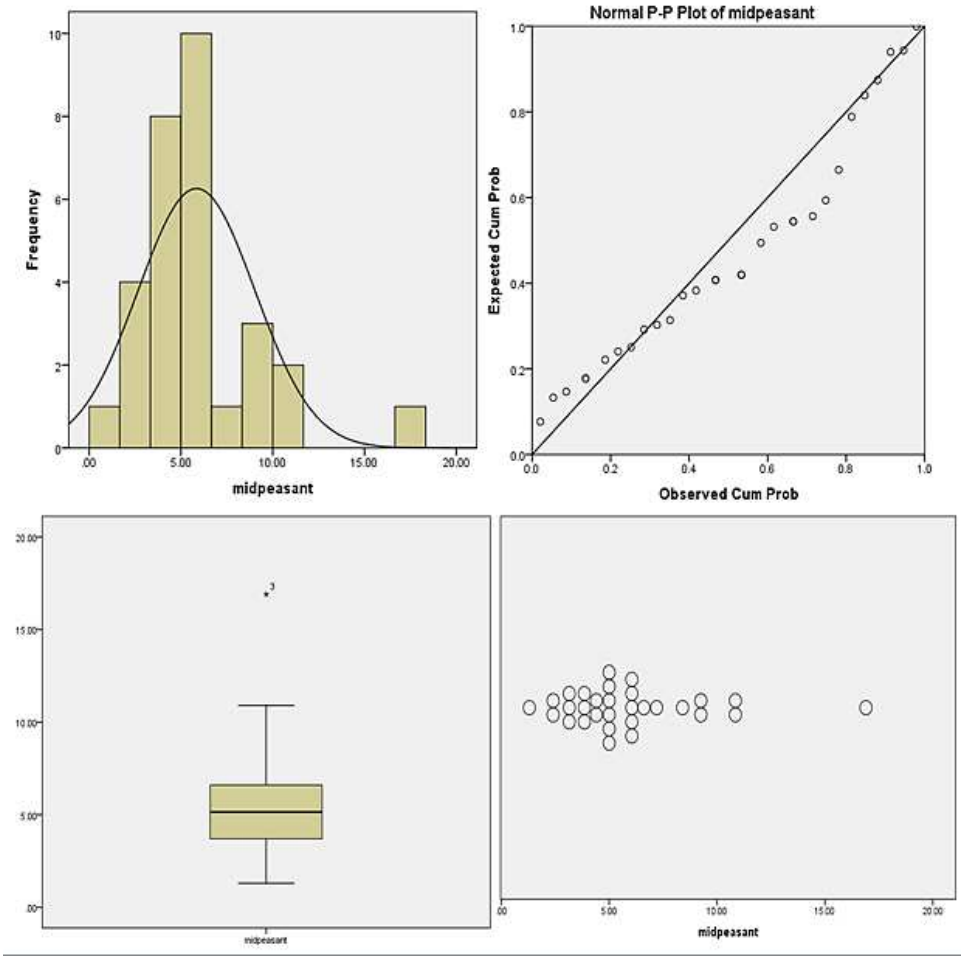
In Figure 3.1, the peak of the histogram is tilting left towards the center, and the graph shows a long right tail. The histogram indicates only one outlier (16609) which is far to the right, an extreme outlier. A normal P-P plot shows that the points have a distinct curvature on the diagonal line which indicates the presence of extreme observations or outliers. In the boxplot, the points which go beyond the upper fence are outliers and the two points that are farther from this upper fence are extreme outliers. Dot plot indicates one point which lies far ahead from the other points of the data as an extreme outlier. In Figure 3.2, the histogram shows two outliers (68.4, 81.7) that are far to the right are extreme outliers. The pattern of the normal P-P plot shows that there do exist outliers in the data.

Three points that go beyond the upper fence are outliers and two points that are farther from this upper fence are considered as extreme outliers. Dot plot shows that the two points (68.4, 81.7) lie far ahead of the other points of the data which are the extreme outliers.



Case 2 (skewness = 3.662)7

Figure 3.2: Histogram, Normal P-P Plot, Boxplot, and Dot Plot for $N = 32$



Case 3 (skewness = 1.662)

Figure 3.3: Histogram, Normal P-P Plot, Boxplot, and Dot Plot for $N = 30$

In Figure 3.3, the histogram exhibits only one outlier (16.9) which is far to the right. On a normal P-P plot, the points show a curve on a diagonal line which indicates that the data has some extreme points or outliers. Boxplot exhibits only a single observation (16.9) that goes beyond the upper fence as an extreme outlier. The dot plot shows only one point (16.9) goes far ahead of the other points of the data which is an extreme outlier.

3.2 Interval (Labeling) Methods

Seven outlier labeling methods are employed on two real data sets. In case 1, the first data set is taken about the sale of sugar in different districts of Punjab in 1995 from the Bureau of Statistics (BOS). The distribution of this data set is mildly skewed to the right. In case 2, the second data set is taken from the 1907 rebellious mid peasants and its distribution is highly skewed to the right because of two highly extreme observations. In case 3, two extreme values are omitted from both data sets, therefore the distribution is

modified to mildly skewed. The purpose of excluding those observations is to show the impact of extreme observations on the outlier labeling methods. Some important descriptive measures of each data set are given below.

Table 1
Descriptive and Other Measures of case 1, 2 and 3

	Case 1	Case 2	Case 3(outlier omitted)
N	423	32	30
Minimum	1.00	1.30	1.30
Maximum	16609	81.70	16.9
Mean	4419.844	10.172	5.847
Median	3756	5.2	5.15
1st quartile	2362	3.8	3.675
3rd quartile	5710	8.1	6.75
Variance	9035152.112	302.018	10.156
Standard deviation	3005.8530	17.379	3.187
Skewness	1.166	3.662	1.622
Kurtosis	1.437	12.873	3.849
Inter Quartile Range (IQR)	3348	4.3	3.075
MADe	2319.41	2.29	2.15
MC (skewness)	0.1495	0.3156	0.00

Case 1 ($n = 423$)

Table 2
Outlier Percentages and Interval/Criteria of Each Outlier Labeling Method

METHODS	Interval/Criteria	Left (%)	Right (%)	Total (%)
2 SD Method	(-1591.86,10431.55)	0 (0)	21 (4.96)	21 (4.96)
3 SD Method	(-4597.72,13437.40)	0 (0)	5 (1.18)	5 (1.18)
Tukey's Method (1.5IQR)	(-2660,10732)	0 (0)	20 (4.73)	20 (4.73)
Tukey's Method (3IQR)	(-7682,15754)	0 (0)	2 (0.47)	2 (0.47)
2 MADe Method	(-882.82,8394.82)	0 (0)	49 (11.6)	49 (11.6)
3 MADe Method	(-3202.23,10714.23)	0 (0)	20 (4.73)	20 (4.73)
Median Rule	(-3944.40,11456.40)	0 (0)	12 (2.84)	12 (2.84)
Adjusted Boxplot	(-613.99,14842.40)	0 (0)	2 (0.47)	2 (0.47)
Z-Score	$\|Zi\| > 3$	0(0)	5(1.18)	5(1.18)
Modified Z – Score	$\|Mi\| > 3.5$	0 (0)	12 (2.84)	12 (2.84)

Remarks: The average results of these outlier labeling methods were similar. The data is a little bit positively skewed. 2SD method, IQR method, and 3MADe method detected almost the same number of outliers (percentage to the right 4.73 approx.) 3IQR and

adjusted boxplot are most resistant because they are least affected by extreme values. The 2MADe method was most sensitive to the observations.

Case 2 ($n = 32$)

Table 3
Outlier Percentages and Interval/Criteria of Each Outlier Labeling Methods

METHODS	Interval/Criteria	Left (%)	Right (%)	Total (%)
2 SD Method	(-24.59,44.93)	0 (0)	2 (6.25)	2 (6.25)
3 SD Method	(-41.97,62.31)	0 (0)	2 (6.25)	2(6.25)
Tukey's Method (1.5IQR)	(-2.65,14.55)	0 (0)	3 (19.38)	3 (9.38)
Tukey's Method (3 IQR)	(-9.10,21.0)	0 (0)	2 (6.25)	2 (6.25)
2 MADe Method	(0.62,9.78)	0 (0)	5 (15.63)	5 (15.63)
3 MADe Method	(-1.62,12.07)	0 (0)	3 (9.38)	3 (9.38)
Median Rule	(-4.69,15.09)	0 (0)	3 (9.38)	3 (9.38)
Adjusted Boxplot	(-1.67,30.98)	0 (0)	2 (6.25)	2 (6.25)
Z – Score	$\|Zi\| > 3$	0 (0)	2 (6.25)	2 (6.25)
Modified Z – Score	$\|Mi\| > 3.5$	0 (0)	3 (9.38)	3 (9.38)

Remarks: The results presented in Table 4 are similar to the results of case 1. All methods detected an approximately equal number of outliers except the 2MADe method which detected the maximum number of outliers as it is sensitive to the extreme observations. The extreme values in the data made the original outliers slip away.

Case 3 ($n = 30$)

Table 4
Outlier Percentages and Interval/Criteria of Each Outlier Labeling Method

METHODS	Interval/Criteria	Left (%)	Right (%)	Total (%)
2 SD Method	(-0.53,12.22)	0 (0)	1 (3.33)	1 (3.33)
3 SD Method	(-3.71,15.41)	0 (0)	1 (3.33)	1 (3.33)
Tukey's Method (1.5 IQR)	(-0.94,11.36)	0 (0)	1 (3.33)	1 (3.33)
Tukey's Method (3 IQR)	(-5.55,15.98)	0 (0)	1 (3.33)	1 (3.33)
2 MADe Method	(0.85,9.45)	0 (0)	4 (13.33)	4 (13.33)
3 MADe Method	(-1.30,11.60)	0 (0)	1 (3.33)	1 (3.33)
Median Rule	(-1.92,12.22)	0 (0)	1 (3.33)	1 (3.33)
Adjusted Boxplot	(-0.94,11.36)	0 (0)	1 (3.33)	1 (3.33)
Z – Score	$\|Zi\| > 3$	0 (0)	1 (3.33)	1 (3.33)
Modified Z- Score	$\|Mi\| > 3.5$	0(0)	1(3.33)	1(3.33)

Remarks: As the extreme values are omitted from the data, skewness, and kurtosis are reduced significantly and detected as only a single outlier (16.9) in the modified data. Also, the 2MADe method detects more outliers than other methods. As the two outliers are quite discrete from the data, so after removing those extreme values the data approached approximately normal distribution.

3.3 Test-Based Methods

Some formal methods also exist in the literature for univariate outlier detection.

3.3.1 Generalized Extreme Studentized Deviate Test

For sample size $n = 423$

H_0: There are no outliers in the data set
H_1: There are up to r outliers in the data set

The level of significance $\alpha = 0.05$ is used and the test statistic is

$$R_i = \frac{\max |x_i - \bar{x}|}{s} \tag{3.1}$$

where $\bar{x}$ is the sample mean and s is the sample SD of a data set.

Output: There are up to 2 outliers tested and the test statistic $R = 4.055$ and the critical value $Pr\ (> |R|) = 0.015$

Remarks: Since p-value = 0.015 which is less than the significance level = 0.05 so, the null hypothesis is rejected and concluded that there are two outliers (16465, 16609) in the data.

For sample size $n = 32$

H_0: There are no outliers in the data set
H_1: There are up to r outliers in the data set

The level of significance $\alpha = 0.05$ is used and the test statistic is

$$R_i = \frac{\max |x_i - \bar{x}|}{s} \tag{3.2}$$

where $\bar{x}$ is the sample mean and s is the sample SD of a data set.

Output: There are up to 3 outliers tested and the test statistic $R = 3.468$ and the critical value $Pr\ (> |R|) = 0.003$

Remarks: **S**ince p-value = 0.003 which is less than the significance level = 0.05 so the null hypothesis is rejected and concluded that there are three outliers (16.9, 68.4, 81.7) in the data.

3.4 Comparison of Outlier Labeling Methods for Different Sample Sizes and Different Skewness

Seven methods are applied to several real data sets with different sample sizes and different skewness. The first data set is taken from daily prices and trading volume of apple stock from July 21st to August 21st in 2016, and the second data set is taken from Barro Growth Data (male secondary education, human capital, investment /GDP) used in Koenker and Machado (1999). The purpose of this comparison is to find the outlier percentages of each outlier labeling method and to identify the robust method according to sample size and skewness type. The average percentages of left, right, and the total outliers for different skewness and sample sizes are given in Table 5.

Table 5
The Average Percentages of Left, Right, and Total Outliers for Different Skewness and Sample Size

Shape of Data	n	SD Method					
		Mean ± 2SD			Mean ± 3SD		
		Left (%)	Right (%)	Total (%)	Left (%)	Right (%)	Total (%)
Approaches to symmetry	66	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
	161	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
Mildly skewed	66	1(1.52)	3(4.55)	4(6.06)	0(0)	1(1.52)	1(1.52)
	161	0(0)	8(4.97)	8(4.97)	0(0)	2(1.24)	2(1.24)
Highly skewed	66	0(0)	4(6.06)	4(6.06)	0(0)	3(4.55)	3(4.55)
	161	0(0)	10(6.21)	10(6.21)	0(0)	4(2.48)	4(2.48)

In Table 5, the outlier percentages computed through the SD method are given. Similarly, the average outlier percentages of other outlier labeling methods are computed. The results of the percentages of seven outlier labeling methods are summarized as follows.

Small sample size:

Symmetric: All methods detect no outlier except adjusted boxplot as it uses robust measure (MC skewness) which is more useful for skewed data.

Mildly Skewed: 2 MADe method detected more outliers than any other method as it is not affected by extreme values. Z-score, 3SD method and Tukey's 3IQR method detected only one observation as an outlier. The median rule identifies the three outliers as it is unaffected by sample size.

Highly Skewed: Adjusted boxplot detected a single outlier because MC skewness is very useful for skewed data. 3 IQR method, 3SD method, and Z-score detected approximately equal number of outliers.

Large Sample Size:

Symmetric: When data approaches symmetry for a large sample size, the outlier percentage in all labeling methods is found to zero. Although Tukey's method applies to large and normal data.

Mildly Skewed: 2SD method and Tukey's 1.5 IQR method detected an equal number of observations as outliers. 3MADe method, Median Rule, and Modified Z-score have almost the same outlier percentages.

Highly Skewed: 2MADe method is most sensitive to extreme observations. Adjusted boxplot, Z-score, and 3SD method detected only four outliers.

4. DISCUSSION

The outlier percentage is higher in the 2MADe method than any other outlier labeling method in skewed data. Senthamarai Kannan et al., (2015) considered the 2MADe as the most robust method for outlier detection. For large normal data, the scale measurements like mean and SD becomes the same as the median and MADe. Left outlier percentages are found the same in Tukey's (1.5 IQR) method and the 3MADe method. On the left side of data most methods like the 3SD method, Tukey's 3IQR method, Median rule, Z-score, Modified Z-score, and Adjusted boxplot have detected no observation as an outlier. For small sample sizes when data approaches symmetry and mildly skewed to the right, the left outlier percentage has increased in the 2MADe method and 2SD method. As the skewness increased in the data the left outlier percentage decreases quickly in these two methods. When the data becomes more skewed, the right outlier percentage is inflated in Tukey's (1.5 IQR) method, MADe method, Median rule, Modified Z-score, and Adjusted boxplot. The SD method irregularly changed in small and large samples. In skewed data, these results are similar to results provided by Seo (2006) in the application of outlier labeling methods on real data sets and simulation studies. According to Seo (2006), three methods such as the MADe, Tukey's method, and the Median rule show similar patterns in skewed data since they employ robust measures to build their intervals. This study found the 3MADe method, Tukey's 1.5 IQR, and the median rule has approximately the same percentages of outlier detected. According to Adil & Irshad (2015), Tukey's boxplot asserts specious outliers on the right side of distribution and exceeds the whiskers on the left side of distribution for skewed distribution. In skewed data for large and small sample sizes, many observations go beyond the upper fences and it is difficult to distinguish the potential and real outliers. Vanderviere and Huber, (2004) described that Tukey's boxplot does not distinguish the potential and real outliers in skewed data. A normal P-P plot indicates the distribution of data and the existence of outliers. A generalized ESD test is performed for large and small sample sizes as this test is used to detect any number of outliers. The results assessed that there are up to two extreme outliers identified in large sample sizes and there are up to three extreme outliers identified in small sample sizes. This test detects largest observation as outliers for large and small sample sizes, According to Kuppusamy & Kaliyaperumal, (2013) Generalized Extreme Studentized Deviate test is significantly better than the Grubbs test and Dixon's test as both tests detected a single observation either maximum or minimum.

5. CONCLUSION

The comparison of the outlier labeling methods in the univariate data set is used to detect the most robust method for outlier detection for small and large sample sizes with different skewness measures, through outlier percentages and their application to real data sets. An adjusted boxplot is sensitive in symmetrical data for small sample size whereas other labeling methods distinguished zero outliers. For large sample size and symmetrical distribution, all outlier labeling methods detected zero outliers. For small sample size Z-score, the 3SD method, and Tukey's (3IQR) method are resistant in slightly skewed data, while adjusted boxplot is the most resistant method is highly skewed distribution and detected only a single outlier. For small and large sample sizes, the 2MADe method is the most sensitive in mildly skewed and highly skewed data as it identified a maximum number

of outliers. For large sample size Z-score, 3SD method, adjusted boxplot, and Tukey's (3IQR) method are resistant in slightly skewed data, these methods are also resistant in highly skewed data except Tukey's (3IQR) method.

REFERENCES

1. Adil, I.H. and Irshad, A.U. (2015). A Modified Approach for Detection of Outliers. *Pakistan Journal of Statistics and Operation Research*, 11(1), 91-102.
2. Andrea, K., Shevlyakov, G. and Smirnov, P.O. (2013). Detection of outliers with Boxplots. *International Conference on Computer Data Analysis and Modeling*, 141-144.
3. Babura, B.I., Adam, M.B., Rahim, A., Samad, A., Fitrianto, A. and Yusif, B. (2018). Analysis and Assessment of Boxplot Characters for Extreme Data. *Journal of Physics*, 1-9.
4. Barbato, G., Barini, E.M., Genta, G. and Levi, R. (2011). Features and performance of some outlier detection methods. *Journal of Applied Statistics*, 38(10), 2133-2149.
5. Ben-Gal, I. (2005). Outlier Detection. In: Maimon O., Rokach L. (eds) *Data Mining and Knowledge Discovery Handbook*. Springer, Boston, MA.
6. Boris, l. and Hoaglin, D.C. (1993). *How to detect and handle outliers. The ASQC Basic References in Quality Control: Statistical Techniques*. American Society for Quality Control, Statistics Division.
7. Carling, K. (1998). Resistant outlier rules and the non-Gaussian case. *Computational Statistics & Data Analysis*, 33(3), 249-258.
8. Carter, N., Schwertman, N. and Kiser, T. (2009). A comparison of two boxplot methods for detecting univariate outliers which adjust for sample size and asymmetry. *Statistical Methodology*, 6(6), 604-621.
9. Cohn, T.A., England, J.F., Berenbrock, C.E., Mason, R.R., Stedinger, J.R. and Lamontagne, J.R. (2013). A generalized Grubbs-Beck test statistic for detecting multiple potentially influential low outliers in flood series. *Water Resources Research,* 49, 5047-5058.
10. Dixon, W.J. (1950). Analysis of Extreme Values. *The Annals of Mathematical Statistics*, 21, 488-506.
11. Dovoedo, Y.H. and Chakraborti, S. (2015). Boxplots- based outlier detection for the location and scale family. *Communication in Statistics- Simulation and Computation*, 44, 1492-1513.
12. Filliben, J.J. and Heckart, A. (2013). Exploratory Data Analysis in Engineering Statistical Handbook. *NIST/SEMATECH*.
13. Grubbs' Outlier Test. (2019). NCSS Statistical Software.
14. Hubert, M. and Vandervieren, E. (2008). An adjusted boxplot for skewed distributions. *Computational statistics & Data Analysis*, 52(12), 5186-5201.
15. Kaliyaperumal, S.K., Arumugam, S. and Kuppusamy, M. (2015). Labeling Methods for Identifying Outliers. *International Journal of Statistics and Systems*, 10(2), 231-238.
16. Kannan, K.S. and Raj, S.S. (2019). Outlier Labeling Methods for Medical Data. In K. Deep, M. Jain and S. Salhi, *Logistics Supply Chain and Financial Predictive Analytics*, 67-75.

17. Kuppusamy, M. and Kaliyaperumal, S.K. (2013). Comparison of methods for detecting outliers. *International Journal of Scientific and Engineering Research*, 4(9), 704-714.
18. Olewuezi, N.P. (2011). Note on the Comparison of Some Outlier Labeling Techniques. Federal University of Technology, Owerri, Nigeria, *Department of Statistics, School of Science,* 2011 Science Publications.
19. Seo, S. (2006). *A Review and Comparison of Methods for Detecting Outliers in Univariate Data Sets*. The University of Pittsburgh.
20. Siraj-Ud-Doulah, M. and Islam, M.H. (2019). An Alternative Robust Measure of Outlier Detection in Univariate Data Sets. *Stm Journals*, 8(1), 1-11.
21. Vanderviere, E. and Huber, M. (2004). An Adjusted Boxplot For a Skewed Distribution. *CompStat*, 1933-1940.
22. Walker, M.L., Dovoedo, Y.H., Chakraborti, S. and Hilton, C. W. (2018). An Improved Boxplot for Univariate Data. *The American Statistician*, 72(4), 348-353.