Outlier Detection on Sale Transactions Using Box Plot

Wint Wah Hlaing, Tar Tar Khin
Computer University (Pyay)
wintwahhlaing09@gmail.com,cupyay2009@gmail.com

Abstract

Outlier detection is an important task in data mining activities and has much attention in both research and application. A value that lies outside which is much smaller or larger than most of the other value in a set of data, this value is called outlier. Most databases include a certain amount of exceptional values, generally termed as outlier. Extremes values tend to be encountered whenever researchers attempt to measure and characterize real world phenomena. Therefore, researchers in all fields are faced with the problem of extreme observations. An observation that is usually large or small relative to the data values is called univariate outlier. In this paper, we present an approach to automating the process of detection univariate outliers. The process is based on graphical display method of construction box plot. In this paper we used outlier labeling method of box plot to detect outlier in electronic items sale database.

Keywords: Outliers analysis, univariate outlier, data mining, Box plot.

1. Introduction

In many business environments, organizations collect large volume of data from their daily operations .There may be too much data and not enough information .To overcome these data mining tools can be used. Data mining tools can analyze enormous set of data and then to find useful patterns and relationships.

Researcher in marketing frequently works with data, they do encounter outlier. Outlier is observation having value that is extreme or unusual data with respect to the rest of the data. The outlier detection is searching for object in the database which appears to be inconsistent with the remainder of that set of data.

The real process of outlier occurrence is usually unknown to data users or analysts. Sometime, this is an error, resulting from the poor quality of a data set, i.e. a data entry or a data conversion error. Physical measurements, especially when performed with malfunctioning equipment, may produce a certain amount of distorted values. In these cases, no useful information is conveyed by the outlier value. However, it is also possible that an outlier represents correct, though exceptional information [4].

In this paper, we detect univariate outlier by using box plot. The rest of the paper is described as follows: in section 2, the related work is described. Section 3 is discussed outlier labeling method of box plot and section 4 is fulfilled with the process flow and implementation of the system. Section 5 is presented the input and output form of the system. Section 6 is included with the experimental result. The conclusion of this system will be combined at the last section 7.

2. Related Work

In many data analysis tasks a large number of variables are being recorded. One of the first steps is the detection of outlying observation. An exact definition of an outlier often depends on hidden assumptions regarding the data structure and the applied detection method. Some definitions are regarded general enough to cope with various types of data and methods [3].

Outlier may be real or erroneous data. Real outliers are observations whose actual values are in fact and very different than those observed for the rest of the data. Erroneous outliers are observation that is due to data collection process. The extreme observation can be many things both good and bad. There are good outliers that provide useful information that can lead to the discovery of new knowledge or bad outliers that include noisy data points. The isolation of outlier is important both for improving the quality of original data and for reducing the impact of outlying values in the process of knowledge discovery in database [1].

3. Box Plot

One of the most frequently used graphical techniques for analyzing a univariate data set is the box plot. The box plot is constructed from five values: the smallest value, the first quartile, the median, the third quartile, and the largest value. The box plot shows the distribution of the data and is especially useful for comparing distributions graphically. It shows information about the

location, spread, and skewness as well as the tail of the data. They also show how far from most of the data the extreme values are. It is created from a set of five numbers.

The five numbers are calculated as follow:

1. Rank orders the data in ascending order.
2. The smallest value is the minimum and the largest value is the maximum.
3. To find the quartile, first calculate the index for each percentile:
   - Index: i = (P/100) n

Where p is the percentile and n is the number of data values.

- If the index i is not an integer, round up to the nearest integer, e.g. round 3.25 up to 4. i denote the position of the value corresponding to that percentile, e.g. the 4th number.
- If the index i is an integer, average the value in position i with the value in the position (i+1) to get the value for that percentile.
- For Q1, p = 25% and (p/100) = ¼* n.
- ForQ2, (median), p=50% =1/2*n.
- For Q3, p = 75% and (p/100) =3/4*n.
- Interquartile range (IQR) = Q3 - Q1.

The box plot is interpreted as follows: The box itself contains the middle 50% of the data. The right edge of the box indicates the 75th percentile (third quartile) of the data set. The left edge of the box indicates the 25th percentile (first quartile). The different between first quartile and third quartile is known as the interquartile range. The line in the box indicates the median value of the data. If the median line within the box is not equal from the edges, then the data is skewed. Otherwise, the data distribution is symmetric. The ends of the horizontal or vertical lines called "whisker" indicate the minimum and maximum data values [2].

The following figure is an example of a box plot.

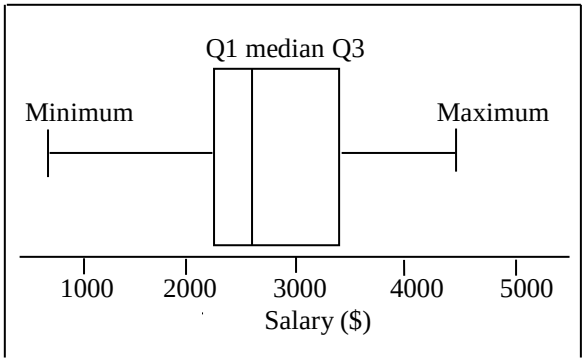

Figure 1: Example of a box plot

The plot may be draw either horizontally or vertically.

3.1 Outlier Analysis

Outlier analysis aims to find a small number of exceptional objects in a database. The data stored in a database may reflect noise, exceptional cases, or incomplete data objects [5].Therefore outlier detection belongs to the most important task in data analysis. The outlier mining problem can be viewed as two sub problems:

- Define what data can be considered as inconsistent in a given data set and
- Find an efficient method to mine the outliers [5].

In general, there are many methods for outlier detection. A quantitative approach to detection of numeric outliers is based on graphical display technique of constructing a box plot. The box plot has become the standard techniques for presenting the 5-number summary which consists of the minimum and maximum range value, the upper and lower quartiles and the median. This method is used to detect outlier in a single variable.

3.2. Univariate outliers

Univariate outliers were identified for each variable within classes. The simplest and the most researched case is the identification of univariate outliers where the distribution of a single variable is examined. Extreme data values are obvious outlier candidates. When the distribution is symmetric, we suspect that candidate's outliers are the extremes of the left or right tail. Correspondingly, the identified outliers are referred to as the lower and upper univariate outliers. In a skewed distribution, the suspect outliers are likely to be the extreme of the longer tail [6].

3.3. Box Plot Outlier Identification

The box plot is a useful graphical method for detecting outlier. The thresholds for lower and upper outliers are defined as follow.

Lower threshold = lower quartile – 1.5(IQR)

Upper threshold = upper quartile + 1.5(IQR)

The value lying outside the lower threshold then the value is lower outliers and the value lying outside the upper threshold then the value is called an upper outlier [6].

4. Implementation of the System

Outliers were searched from electronic items sale transactions data sets. There will be a set of items in the database, and transaction table is updated from the sales of those items. In this system, the database for sale transaction will be built which is the previous transaction of electronic items. We use the outlier labeling method of box plot to detect outlier. The management levels of the user want to know the sale condition of the shop; he or she must enter the user name and password. Then the user can search of his interested month or year. The system finds outlier from the database by using box plot. Later, the user gets the result of sale pattern, which is extreme value in the data set. This system also displays outlier in monthly and yearly.

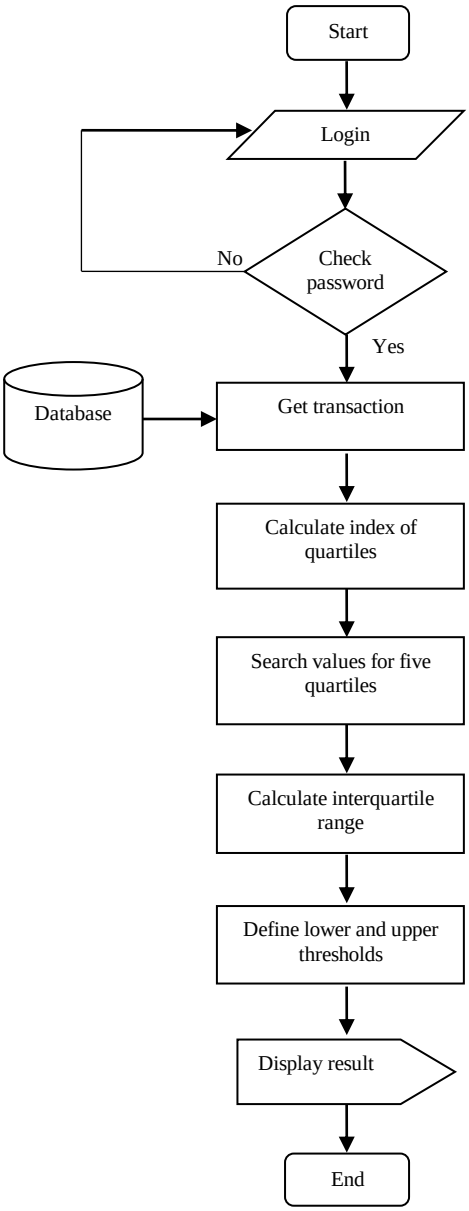

Figure 2: Process flow of the system

5. Input and Output Form of the System

This system analyzes for five years such as year 2005 to year 2009. And displays the highest and lowest seller items in each month and each year and then shows monthly or yearly analysis. The user can choose year and also choose which month of that year.

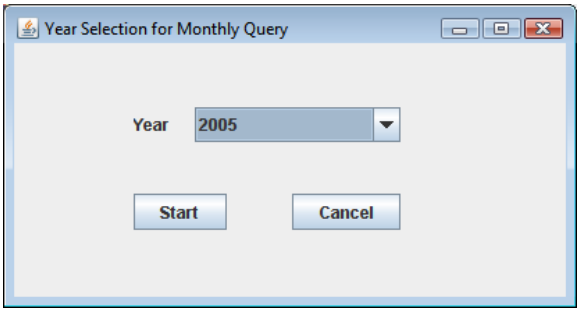

Figure 3: Input form of Year Selection for Monthly Query

The following figure shows the box plot with the result of outlier in year 2005 for month February. In figure: 4, there is one upper outlier and this item name is Refrigerator, item model is NR-A16KN-one door. There is one lower outlier, this item name is Television and item model is 29FS2BL-CTV.

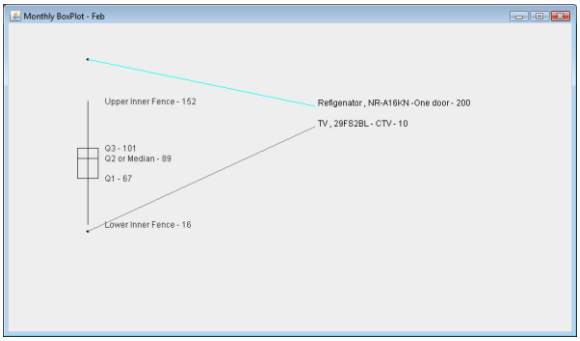

Figure 4: Result of box plot for month February.

Then, the user can search monthly for his interested year. The box plots with the result of outlier in monthly analysis for year 2005 are described in figure: 5, in this figure display 12-months of box plot are comparing.

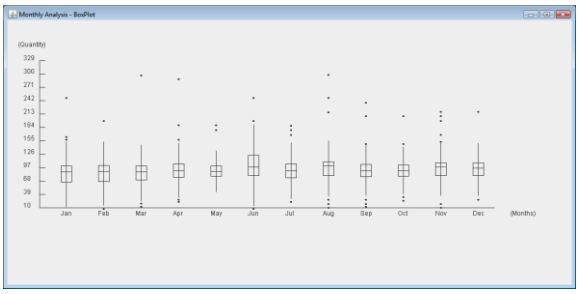

Figure 5: Monthly analysis box plot for year 2005.

The following figure: 6 is showed the box plot with the result of outlier for the whole year in 2008.In this year there is one upper outlier, this item name is Rice cooker and item model is SR42-HP and one lower outlier, this item name is Air conditioner and model is FY10ELNP-Air curtain.

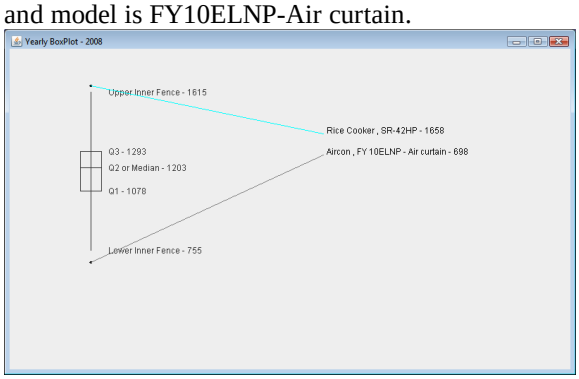

Figure 6: Result of box plot for year 2008.

The next figure is showed the box plot with the result of outlier in yearly analysis for five year (year in 2005, 2006, 2007, 2008 and 2009). In figure: 7 displays five years of box plot is compared.

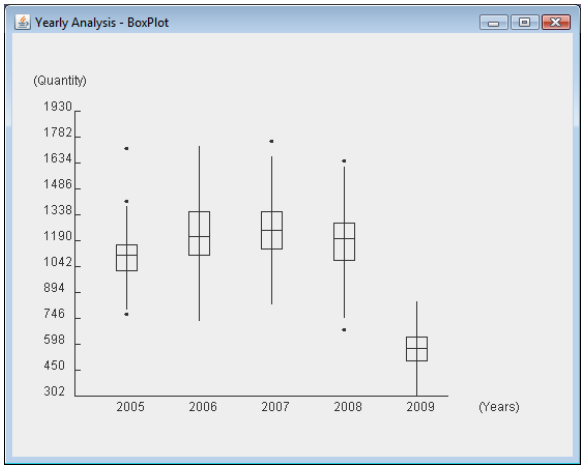

Figure 7: Yearly analysis box plot for five years.

6. Experimental Result

In this system, outlier is detecting by using box plot from the electronic items sale database which is 75 items and total transactions are 4052 during the year 2005 to 2009. This system can detect outlier by each month or year and also monthly or yearly. The plot can be varied with the transaction. If the transaction increase, the length of the plot will vary. The following figure: 8 shows the monthly analysis plot in year 2007 with transactions 2000.

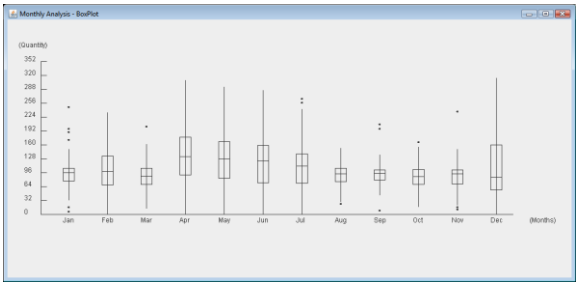

Figure 8: Monthly analysis in year 2007 with transactions 2000.

The next figure: 9 describes monthly analysis plot in year 2007 with transactions 4052. The transactions increase 2000 to 4052 and the length of the plot will vary.

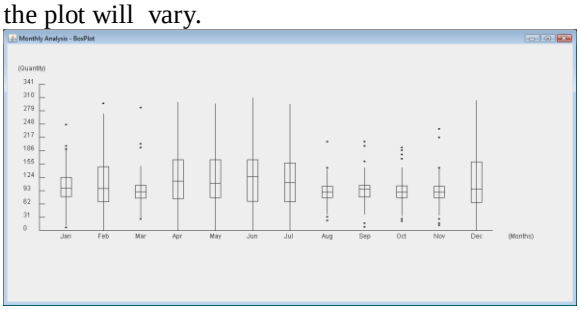

Figure 9: Monthly analysis in year 2007 with transactions 4052.

7. Conclusion

In this paper, we have identified outlier using a graphical display technique of construction a box plot. The detection of outlier can lead to the discovery of truly unexpected knowledge in application area. We intend to find extremeness from a database. In this system we use box plot to detect univariate outliers. This system can only discover outlier depend on the total quantity dimension. The focus of the further work in this system is to determine the impact of an outlier or group of outliers on study results. So there is further analysis using any kinds of association rule. This system may help the analyst or manager to know the sale condition of items.

8. References

[1] Edgar Acuna and Caroline Rodiguez, "On detection of outliers and their effect in supervised classification"

[1] "Box Plot and Five-Number Summaries" Department of Mathematics, Sinclair Community College, Dayton, OH.

[2] Irad Ben-Gal, "Outlier detection"

[3] Mark Last, Abraham Kandel, "Automated Detection of Outliers in Real World- Data"

[4] Jiawei Han, Micheline Kamber, "Data mining concepts and techniques"

[5] Jorma Laurikkala, Martti Juhola and Erna Kentala “Informal identification of outlier in medical data”