3.0 INTRODUCTION

Data pre-processing is an often neglected but important step in the data mining process (Zhang et al., 2003). Data gathering methods are often loosely controlled, resulting in out of range values (e.g., Income: -100), impossible data combinations (e.g., Gender: Male, Pregnant: Yes), missing values, etc. Analyzing data that has not been carefully screened for such problems can produce misleading results. Thus, the representation and quality of data is first and foremost before running any analysis. If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training phase is more difficult (Berry and Linoff, 2004). Data preparation and filtering steps can take considerable amount of processing time. Data pre-processing includes cleaning, normalization, transformation, feature extraction and selection, etc. The product of data pre-processing is the final training set.

3.1 DATA PRE-PROCESSING METHODS

Raw data is highly susceptible to noise, missing values and inconsistency. The quality of data affects the data mining results. In order to help improve the quality of the data, and consequently, of the mining results raw data is pre-processed so as to improve the efficiency and ease of the mining process (Fayyad et al., 1996). Data preprocessing is one of the most critical steps in a data mining process which deals with the preparation and transformation of the initial dataset. Data pre-processing methods are divided into following categories:

- Data Cleaning
- Data Integration
- Data Transformation
- Data Reduction
3.1.1 Data Cleaning

Data that is to be analyzed by data mining techniques can have the following defects:

- Incomplete:
  - Lacking attribute values or certain attributes of interest, or containing only aggregate data

- Noisy:
  - Containing errors or outlier values which deviate from the expected.
Chapter 3

Inconsistent:

Containing discrepancies for example department codes used to categorize items.

Incomplete, noisy and inconsistent data are common place properties of large, real-world databases and data warehouses. Incomplete data can occur for a number of reasons (Elisseeti and Isabelle, 2003). Attributes of interest may not always be available, such as customer information for sales transaction data. Other may not be included simply because it was not considered important at the time of entry. Relevant data may be recorded due to a misunderstanding, or because of equipment malfunctions. Data that were inconsistent with other recorded data may have been deleted. Furthermore, the recording of the history or modifications to the data may have been overlooked (Bilgin and Camurcu, 2004). Missing data, particularly for tuples with missing values for some attributes, may need to be inferred. Data can be noisy, having incorrect attribute values, owing to the following. The data collection instruments used may be faulty. There may have been human or computer errors occurring at data entry. Errors in data transmission can also occur. There may be technology limitations, such as limited buffer size for coordinating synchronized data transfer and consumption. Incorrect data may also result from inconsistencies in naming conventions or data codes used. Duplicate tuples also require data cleaning. Therefore, a useful pre-processing step is to run your data through some data cleaning routines (Han and Kamber, 2006).

**Missing Values:**

If it is noted that there are many tuples that have no recorded value for several attributes, then the missing values can be filled in for the attribute by various methods described below (Draper and Smith, 1982):

1. Ignore the tuple
2. Fill in the missing value manually
3. Use a global constant to fill in the missing value
4. Use the attribute mean to fill in the missing value.

5. Use the attribute mean for all samples belonging to the same class as the given tuple.

6. Use the most probable value to fill in the missing value

**Noisy Data:**

Noise is a random error or variance in measured variable. The following data smoothing techniques describes this (Zhang et al., 2003).

- **Binning Methods:**

  This method smoothens a sorted data value by consulting the neighborhood, or values around it. The sorted values are distributed into a number of “buckets” or bins. Because binning methods consult the neighborhood of values, they perform local smoothing values around it (Duda et al., 2001).

- **Clustering:**

  Outliers may be detected by clustering, where similar values are organized into groups or clusters (Frawley et al., 2010).

- **Combined Computer and Human Inspection:**

  Outliers may be identified through a combination of computer and human inspection.

- **Regression:**

  Data can be smoothed by fitting the data to a function, such as with regression (Geof et al., 2013).
Inconsistent data:

There may be inconsistencies in the data recorded for some transactions. Some data inconsistencies may be corrected manually using external references. For example, errors made at data entry may be corrected by performing a paper trace (Green and Silverman, 1994). This may be coupled with routines designed to help correct the inconsistent use of codes. Knowledge engineering tools may also be used to detect the violation of known data constraints (Bilgin and Camurcu, 2004). For example, known functional dependencies between attributes can be used to find values contradicting the functional constraints.

3.1.2 Data Integration

Data integration combines data from multiple sources into a coherent data store, as in data warehousing. These sources may include multiple databases, data cubes or flat files. Databases and data warehouses typically have metadata – that is, data about the data. Such metadata can be used to help avoid errors in schema integration (Berry and Linoff, 2004). Redundancy is another important issue. An attribute may be redundant if it can be derived from another table, such as annual revenue. Inconsistencies in attribute or dimension naming can also cause redundancies in the resulting data set (Lewis and Michael, 1980).

3.1.3 Data Transformation

In data transformation (Berry and Linoff, 2004), the data are transformed or consolidated into forms appropriate for mining. Data transformation can involve the following:

- Normalization, where the attribute data are scaled so as to fall within a small specified range, such as -1.0 to 1.0, or 0 to 1.0.
- Smoothing works to remove the noise from data. Such techniques include binning, clustering and regression.
Aggregation, where summary or aggregation operations are applied to the data. For example, the daily sales data may be aggregated so as to compute monthly and annual total amounts. This step is typically used in construction a data cube for analysis of the data at multiple granularities.

Generalization of the data, where low level or primitive raw data are replaced by higher level concepts through the use of concept hierarchies. For example categorical attributes, like street, can be generalized to higher level concepts, like city or country (Ferreira and Vellasco, 2004). Similarly, values for numeric attributes, like age, may be mapped to higher level concepts, like young, middle-aged and senior.

3.1.4 Data Reduction

Complex data analysis and mining on huge amounts of data may take a very long time, making such analysis impractical or infeasible. Data reduction techniques have been helpful in analyzing reduced representation of the dataset without compromising the integrity of the original data and yet producing the quality knowledge (Zhang et al., 2003). The concept of data is commonly understood as either reducing the volume or reducing the dimensions (number of attributes). There are a number of methods that have facilitated in analyzing a reduced volume or dimension of data and yet yield useful knowledge. Certain partition based methods work on partition of data tuples. That is, mining on the reduced data set should be more efficient yet produce the same analytical results.

Strategies for data reduction include the following.

- **Data Cube Aggregation**, where aggregation operations are applied to the data in the construction of a data cube.

- **Dimension reduction**, where irrelevant, weakly relevant or redundant attributes or dimensions may be detected and removed.
Data compression, where encoding mechanisms are used to reduce the data set size. The methods used for data compression are wavelet transform and Principal component analysis (Lin Y.H et al., 1996).

Numerosity reduction, where the data are replaced or estimated by alternative, smaller data representation such as parametric modes (which need store only the model parameters instead of the actual data e.g. regression and log linear models), or non-parametric methods such as clustering, sampling and the use of histograms.

Discretization and concept hierarchy generation, where raw data values for attributes are replaced by ranges or higher conceptual levels. Concept hierarchies allow the mining of data at multiple levels of abstraction and are a powerful tool for data mining.

3.2 SIGNIFICANCE OF DATA PRE-PROCESSING

Data preparation is an important aspect for both data warehousing and data mining, as real world data tends to be incomplete, noisy and inconsistent. Data preparation includes data cleaning, data integration, data transformation and data reduction. Data cleaning routines can be used to fill in missing values, smooth noisy data, identify outliers and correct data inconsistencies (Zhang et al., 2003). Data integration combines data from multiple sources to form a coherent data store. Metadata, correlation analysis, data conflict detection, and the resolution of semantic heterogeneity contribute towards smooth data integration. Data transformation routines conform the data into appropriate forms for mining. For example, attribute data may be normalized so as to fall between a small range, such as 0 to 1.0. Data reduction techniques such as data cube aggregation, dimension reduction, data compression, numerosity reduction, and discretization can be used to obtain a reduced representation of the data, while minimizing the loss of information content. Concept hierarchies organize the values of attributes or dimensions into gradual levels of abstraction. They are a form of discretization that...
is particularly useful in multilevel mining. Automatic generation of concept hierarchies for categoric data may be based on the number of distinct values of the attributes defining the hierarchy. For numeric data, techniques such as data segmentation by partition rules, histogram analysis, and clustering analysis can be used. Although several methods of data preparation have been developed, data preparation remains an active and important area of research.

3.3 ADOPTED DATA-PREPROCESSING TECHNIQUES

Following data-preprocessing techniques have been used for the collected climatic data of Chennai.

3.3.1 Raw Data Analysis

In raw data analysis, climatic data of the Chennai is overlooked manually how it looks like. Following are the climatic parameter.

- Temperature (High, Avg, Low) in °C
- Dew Point (High, Avg, Low) in °C
- Humidity (High, Avg, Low) in %
- Sea Level Pressure (High, Avg, Low) in (hPa)
- Visibility (High, Avg, Low) in km
- Wind Speed (High, Avg, Low) in (km/hr)
- Precipitation in (mm)

The aim of the research is to study the influence of Sea Level Pressure, Wind Speed, Visibility, Dew Point, Cloud Cover, Humidity, Wind Direction, etc., on Temperature. So Temperature is our key parameter and we did study of Trend Analysis with time period i.e., a scatter plot graph has been constructed between the temperature value and days of the year for the year 2010. Two Graphs has been
constructed one for Max Temperature and another one for Min Temperature per year. As a result for 4 years data, 8 Graphs has been generated and studied.

Fig. 3.2 – Shows Trend Analysis for High Temperature Vs Date (Year 2010)

Fig. 3.3 – Shows Trend Analysis for Low Temperature Vs Date (Year 2010)
3.3.1.1 Results And Discussion

On analyzing all the 8 graphs generated, we inferred that climatic data has to be split into 4 groups, based on the increase and decrease of the temperature based on the time period.

**Table 3.1 Shows the groups in the Time Period based on Trend Analysis**

<table>
<thead>
<tr>
<th>Group No.</th>
<th>Period</th>
<th>Inference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>JAN to MAY</td>
<td>Good Rise in Temperature (Increasing)</td>
</tr>
<tr>
<td>2.</td>
<td>JUN to AUG</td>
<td>Good Drop in Temperature (Decreasing)</td>
</tr>
<tr>
<td>3.</td>
<td>SEP to OCT</td>
<td>Very Gradual Increase in Temperature</td>
</tr>
<tr>
<td>4.</td>
<td>NOV to DEC</td>
<td>Sudden Fall in Temperature</td>
</tr>
</tbody>
</table>

Further to this analysis, the Outlier Analysis and Data mining techniques like Multiple Linear Regression, Correlation Analysis, Classification and Clustering algorithms are separately applied to group 1 (Jan-May) data across 2010 to 2013 and inferences are documented for gaining knowledge.

3.3.2 Normalization

The measurement unit used can affect the data analysis. For example, changing measurement units from meters to inches for height, or from kilograms to pounds for weight, may lead to very different results (Denis Bosq, 2012). In general, expressing an attribute in smaller units will lead to a larger range for that attribute, and thus tend to give such an attribute greater effect or “weight”. To help avoid dependence on the choice of measurement units, the data should be normalized or standardized. This involves transforming the data to fall within a smaller or common range such as [-1,1] or [0.0, 1.0]. The terms standardize and normalize are used interchangeably in data pre-processing, although in statistics, the latter term also has other connotations (Montgomery and Runger, 2014).
Chapter 3

Normalizing the data attempts to give all attributes an equal weight. Normalization is particularly useful for classification algorithms involving neural networks or distance measurements such as nearest-neighbour classification and clustering. For distance-based methods, normalization helps prevent attributes with initially large ranges (e.g., income) from outweighing attributes with initially smaller ranges (e.g., binary attributes). It is also useful when given no prior knowledge of the data (Dash and Liu, 1997).

3.3.2.1 Min-Max Normalization

In data transformation, the data are transformed or consolidated into forms appropriate for mining. We adopt Min-Max Normalization to transform all the collected climatic parameters in the range of 0 to 100, for performing the outlier analysis (Michael Baron, 2013). Min-Max normalization performs a linear transformation on the original data. Suppose that \( \min_A \) and \( \max_A \) are the minimum and maximum values of an attribute, \( A \). Min-max normalization maps a value, \( v \), of \( A \) to \( v' \) in the range \([\text{new}_\min_A, \text{new}_\max_A]\) by computing

\[
v' = \frac{v - \min_A}{\max_A - \min_A} (\text{new}_\max_A - \text{new}_\min_A) + \text{new}_\min_A.
\]

Min-max normalization preserves the relationships among the original data values. It will encounter an “out-of-bounds” error if a future input case for normalization falls outside of the original data range for \( A \).

3.3.2.2 Results And Discussion

Min-Max Normalization data transformation has been applied to the identified group 1 (Jan-May) across 2010 to 2013. Min-Max Normalization has been chosen as part of data pre-processing in order to perform Outlier Analysis and followed by Correlation Analysis followed by Multiple Regression Analysis.
For Example

Maximum in Max Temp : 43
Minimum in Max Temp : 25
Difference : 43 – 25 = 18
Transformation of 29 °C in [0, 100]
is (29 – 25) / 18 = 22.22°C

<table>
<thead>
<tr>
<th>Month</th>
<th>Max Temp (°C)</th>
<th>Min Temp (°C)</th>
<th>Precipitation (mm)</th>
<th>Wind Speed (m/s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan</td>
<td>27</td>
<td>25</td>
<td>100</td>
<td>10</td>
</tr>
<tr>
<td>Feb</td>
<td>29</td>
<td>25</td>
<td>100</td>
<td>10</td>
</tr>
<tr>
<td>Mar</td>
<td>30</td>
<td>25</td>
<td>100</td>
<td>10</td>
</tr>
<tr>
<td>Apr</td>
<td>31</td>
<td>25</td>
<td>100</td>
<td>10</td>
</tr>
<tr>
<td>May</td>
<td>32</td>
<td>25</td>
<td>100</td>
<td>10</td>
</tr>
<tr>
<td>Jun</td>
<td>33</td>
<td>25</td>
<td>100</td>
<td>10</td>
</tr>
<tr>
<td>Jul</td>
<td>34</td>
<td>25</td>
<td>100</td>
<td>10</td>
</tr>
<tr>
<td>Aug</td>
<td>35</td>
<td>25</td>
<td>100</td>
<td>10</td>
</tr>
<tr>
<td>Sep</td>
<td>36</td>
<td>25</td>
<td>100</td>
<td>10</td>
</tr>
<tr>
<td>Oct</td>
<td>37</td>
<td>25</td>
<td>100</td>
<td>10</td>
</tr>
<tr>
<td>Nov</td>
<td>38</td>
<td>25</td>
<td>100</td>
<td>10</td>
</tr>
<tr>
<td>Dec</td>
<td>39</td>
<td>25</td>
<td>100</td>
<td>10</td>
</tr>
</tbody>
</table>

Figure 3.4 – Shows Min-Max Normalization of climatic data of Year 2010 (Screen Shot shows partial data of Yr 2010)

Figure 3.4, clearly shows the transformation of all the climatic parameter by applying the Min-Max Normalization. Above results are used for Outlier Analysis, Correlation Analysis and Multiple Regression Analysis. Chapter 4 broadly elaborates on Correlation Analysis and Multiple Regression Analysis.
3.3.2.3 Z-Score Normalization

A Z-Score (also known as z-value, standard score or normal score) is a measure of the divergence of an individual experimental result from the most probable result, the mean (Pang-Ning et al., 2005). Z is expressed in terms of the number of standard deviations from the mean value. In simple words, Z-Scores tell us whether a particular score is equal to the mean, below the mean or above the mean of a bunch of scores.

If a Z-Score

- Has a value of 0, it is equal to the group mean.
- Is positive, it is above the group mean.
- Is negative, it is below the group mean.
- Is equal +m, it is “m” Standard Deviation above the mean.
- Is equal –m, it is “m” Standard Deviation below the mean.

\[ z = \frac{X - \mu}{\sigma} \]

\( X = \text{Experimental Value} \)
\( \mu = \text{Mean} \)
\( \sigma = \text{Standard Deviation} \)

Z-Scores assuming the sampling distribution of the test statistic (mean in most cases) are normal and transform the sampling distribution into a standard normal distribution. As explained, the standard deviation of a sampling distribution depends on the number of samples. Whenever using Z-Scores it is important to remember a few things:

- Z-Scores normalize the sampling distribution for meaningful comparison.
- Z-Scores require a large amount of data.
- Z-Scores require independent, random data.
3.3.2.4 Results And Discussion

Z-Score Normalization data transformation has been applied to the identified group1 (Jan-May) across 2010 to 2013. Z-Score Normalization has been chosen as part of data pre-processing in order to perform Outlier Analysis and followed by Cluster Analysis.

Fig. 3.5 – Shows Z-Score Transformation for climatic data of Year 2010–2014 (Excel shows partial data)

Fig. 3.5, clearly shows the transformation of all the climatic parameter by applying the Z-Score Normalization. Above results are used for Outlier Analysis, Clustering Analysis. Chapter 6 broadly elaborates on Clustering Analysis.

3.3.2.5 Binning Method – Discretization of Data

Discretization reduces the number of values for a given continuous attribute by dividing the range of the attribute into intervals. Interval labels can then be used to replace actual data values. Some classification algorithms only accept categorical attributes. Binning method is a type of Discretization that has been adopted in this research (Cohen and Jacob, 2003). Binning method first sort data and partition into bins and then one can smooth by bin means, smooth by bin...
median, smooth by bin boundaries, etc. There are two types of Binning method – Equal-Width Partitioning and Equal-depth partitioning. Equal-Width Partitioning method has been adopted for performing the data transformation.

If \( \text{No. of Intervals} = N \)

\[
\text{Width} = \frac{\text{Max Value} - \text{Min Value}}{N}
\]

Intervals are

\(< (\text{Min Value} + \text{Width})\)

Betw. \((\text{Min Value} + \text{Width})\) and \((\text{Min Value} + 2\text{Width})\)

Betw. \((\text{Min Value} + 2\text{Width})\) and \((\text{Min Value} + 3\text{Width})\)

\(> (\text{Min Value} + (n-1)\text{Width})\)

For example for Temperature classification

\(N = 4\)
\(\text{Max Value} = 37\)
\(\text{Min Value} = 23\)
\(\text{Width} = \frac{(37 - 23)}{4} = 3.5\)

Ranges of the temperature and its class are framed as below

<table>
<thead>
<tr>
<th>Temperature in Deg C</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 26.5</td>
<td>T1</td>
</tr>
<tr>
<td>Betw. 26.5 and 29.9</td>
<td>T2</td>
</tr>
<tr>
<td>Betw. 30 and 33.4</td>
<td>T3</td>
</tr>
<tr>
<td>(\geq 33.5)</td>
<td>T4</td>
</tr>
</tbody>
</table>

3.3.2.6 Results And Discussion

Binning Method data transformation has been applied to the identified group1 (Jan-May) across 2010 to 2013. Binning method has been chosen as part of data pre-processing in order to perform Outlier Analysis and followed by
Classification and Clustering techniques. Following table shows the transformation rule used to perform the transformation of continuous climatic data into discrete climatic data.

**Table 3.2 – Shows data Transformation Rules using Binning Method**

<table>
<thead>
<tr>
<th>Visibility in Km</th>
<th>Class</th>
<th>Wind Speed in Km/h</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 3.5</td>
<td>V1</td>
<td>&lt; 3.3</td>
<td>W1</td>
</tr>
<tr>
<td>Betw. 3.5 and 4.9</td>
<td>V2</td>
<td>Betw. 5.3 and 10.5</td>
<td>W2</td>
</tr>
<tr>
<td>Betw. 5 and 6.4</td>
<td>V3</td>
<td>Betw. 10.6 and 15.8</td>
<td>W3</td>
</tr>
<tr>
<td>&gt;= 6.5</td>
<td>V4</td>
<td>&gt;= 15.9</td>
<td>W4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Humidity in %</th>
<th>Class</th>
<th>Temperature in Deg C</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 45</td>
<td>H1</td>
<td>&lt; 26.5</td>
<td>T1</td>
</tr>
<tr>
<td>Betw. 45 and 59</td>
<td>H2</td>
<td>Betw. 26.5 and 29.9</td>
<td>T2</td>
</tr>
<tr>
<td>Betw. 60 and 74</td>
<td>H3</td>
<td>Betw. 30 and 33.4</td>
<td>T3</td>
</tr>
<tr>
<td>&gt;= 75</td>
<td>H4</td>
<td>&gt;= 33.5</td>
<td>T4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sea Lvl Press in hPa</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 1003.8</td>
<td>P1</td>
</tr>
<tr>
<td>Betw. 1003.8 and 1008.5</td>
<td>P2</td>
</tr>
<tr>
<td>Betw. 1008.5 and 1013.3</td>
<td>P3</td>
</tr>
<tr>
<td>&gt;=1013.4</td>
<td>P4</td>
</tr>
</tbody>
</table>

Table 3.2, clearly shows the transformation rule obtained by using Binning Method for all the climatic parameter. Binning method uses four categorization like H1, H2, H3 and H4 for Humidity, etc. The transformed data are used for Outlier Analysis, Clustering Analysis. Chapter 6 broadly elaborates on Clustering Analysis. Figure 3.6 shows the rules applied to the climatic data and its corresponding transformation.
3.4 OUTLIER ANALYSIS

Data examination is a time consuming, but necessary, initial step in any analysis that researchers often overlook. Here we evaluate the impact of missing data, identify outliers, and tests for the assumptions underlying most multivariate techniques (Peter, 2013). The objective of these data examination tasks is as much to reveal what is not apparent as it is to portray the actual data, for the “hidden” effects are easily overlooked.

Outliers are observations with a unique combination of characteristics identifiable as distinctly different from the other observations. Outliers in weather data may be due to data entry problem, faulty data collection instruments, abnormal changes in weather such as tornadoes, hurricane, forest fires, etc. Such non-representative samples can seriously affect the model produced later (Zhao et al., 2003).

There are two strategies for dealing with outliers:

a.) Detect and eventually remove outliers as a part of the preprocessing phase.

b.) Develop robust modeling methods that are insensitive to outliers.
There are two methods for doing Outlier Analysis, one with Graphical method and another without using Graphs. We used both the methods: Graphical and Non-Graphical method (One Class Support Vector Classification).

### 3.4.1 Outlier Analysis – Graphical Method

Graphical techniques are used in identifying the outliers. Graphs are generated portraying the basic characteristics of individual variables and relationships between variables in a simple picture. For example, a simple scatter plot represents in a single picture not only the two basic elements of a correlation coefficient, namely the type of relationship (positive or negative) and the strength of the relationship (Radhika and Shashi, 2009), but also is a simple visual means for assessing linearity that would require a much more detailed analysis if attempted strictly by empirical means (Sun et al., 1995). After applying the min-max normalization to all the climatic parameters, scatter plot graph has been plotted between Temperature and various parameters, through which the outliers are identified and removed for processing it separately.

![Graphs depicting Outliers](image)

**Fig. 3.7.** Shows the Graphs depicting Outliers
Fig. 3.7. Shows the Graphs depicting Outliers

Following figure shows the identified outliers based on Graphical method detection.

Table 3.3. Shows the Identified Outliers

<table>
<thead>
<tr>
<th>IST</th>
<th>Max Temp</th>
<th>Dew Point</th>
<th>Max Humi</th>
<th>Max Sea</th>
<th>Max Vis</th>
<th>Max Wind</th>
<th>Precipitat</th>
<th>CloudCover</th>
<th>WindDirDegrees</th>
</tr>
</thead>
<tbody>
<tr>
<td>8-Jan-2010</td>
<td>16.6666</td>
<td>58.3333</td>
<td>100</td>
<td>72.2222</td>
<td>20</td>
<td>3.960396</td>
<td>75</td>
<td>98.255455</td>
<td>96.255455</td>
</tr>
<tr>
<td>9-Jan-2010</td>
<td>16.6666</td>
<td>58.3333</td>
<td>100</td>
<td>66.6667</td>
<td>0</td>
<td>8.910891</td>
<td>75</td>
<td>97.1590909</td>
<td>97.1590909</td>
</tr>
<tr>
<td>22-Jan-2010</td>
<td>16.6666</td>
<td>58.3333</td>
<td>100</td>
<td>85.3658</td>
<td>80</td>
<td>8.910891</td>
<td>25</td>
<td>89.4883836</td>
<td>89.4883836</td>
</tr>
<tr>
<td>23-Jan-2010</td>
<td>16.6666</td>
<td>25</td>
<td>70.73171</td>
<td>83.3333</td>
<td>60</td>
<td>12.87129</td>
<td>25</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>12-Apr-2010</td>
<td>55.5556</td>
<td>66.6667</td>
<td>17.07317</td>
<td>38.8888</td>
<td>20</td>
<td>12.87129</td>
<td>37.5</td>
<td>53.4090909</td>
<td>53.4090909</td>
</tr>
<tr>
<td>18-May-2010</td>
<td>55.5556</td>
<td>61.6667</td>
<td>85.3658</td>
<td>22.2222</td>
<td>20</td>
<td>12.87129</td>
<td>62.5</td>
<td>88.6363636</td>
<td>88.6363636</td>
</tr>
<tr>
<td>19-May-2010</td>
<td>16.6666</td>
<td>75</td>
<td>100</td>
<td>0</td>
<td>20</td>
<td>23.76283</td>
<td>100</td>
<td>71.87</td>
<td>71.87</td>
</tr>
<tr>
<td>26-May-2010</td>
<td>33.3333</td>
<td>58.3333</td>
<td>73.17073</td>
<td>5.55556</td>
<td>40</td>
<td>21.78218</td>
<td>75</td>
<td>61.4772772</td>
<td>61.4772772</td>
</tr>
<tr>
<td>9-Jul-2010</td>
<td>22.2222</td>
<td>58.3333</td>
<td>100</td>
<td>50</td>
<td>20</td>
<td>12.87129</td>
<td>50</td>
<td>99.71590909</td>
<td>99.71590909</td>
</tr>
<tr>
<td>28-Apr-2011</td>
<td>44.4444</td>
<td>82.3333</td>
<td>100</td>
<td>16.6667</td>
<td>100</td>
<td>12.87129</td>
<td>25</td>
<td>42.8977272</td>
<td>42.8977272</td>
</tr>
<tr>
<td>25-May-2012</td>
<td>88.8888</td>
<td>33.3333</td>
<td>0</td>
<td>5.55556</td>
<td>100</td>
<td>18.81188</td>
<td>37.5</td>
<td>68.75</td>
<td>68.75</td>
</tr>
<tr>
<td>26-May-2012</td>
<td>72.2222</td>
<td>58.3333</td>
<td>7.317073</td>
<td>11.1111</td>
<td>100</td>
<td>16.81188</td>
<td>50</td>
<td>63.6363636</td>
<td>63.6363636</td>
</tr>
<tr>
<td>6-Jun-2013</td>
<td>27.7777</td>
<td>58.3333</td>
<td>100</td>
<td>61.1111</td>
<td>100</td>
<td>25</td>
<td>16.1931812</td>
<td>16.1931812</td>
<td>16.1931812</td>
</tr>
<tr>
<td>13-May-2013</td>
<td>55.5556</td>
<td>83.3333</td>
<td>90.2439</td>
<td>22.2222</td>
<td>100</td>
<td>12.87129</td>
<td>87.5</td>
<td>57.67045455</td>
<td>57.67045455</td>
</tr>
<tr>
<td>14-May-2013</td>
<td>66.6667</td>
<td>91.6667</td>
<td>63.1463</td>
<td>11.1111</td>
<td>100</td>
<td>12.87129</td>
<td>75</td>
<td>53.4090909</td>
<td>53.4090909</td>
</tr>
<tr>
<td>27-May-2013</td>
<td>61.1111</td>
<td>75</td>
<td>26.82927</td>
<td>27.7777</td>
<td>100</td>
<td>15.84158</td>
<td>50</td>
<td>57.67045455</td>
<td>57.67045455</td>
</tr>
<tr>
<td>30-May-2013</td>
<td>66.6667</td>
<td>66.6667</td>
<td>26.82927</td>
<td>22.2222</td>
<td>100</td>
<td>12.87129</td>
<td>37.5</td>
<td>58.23863636</td>
<td>58.23863636</td>
</tr>
</tbody>
</table>
3.4.2 **Outlier Analysis – One Class Support Vector Classification Method**

Outlier detection is implemented as one-class classification, because only one class is represented in the training data (Denis Riordan and Bjarne K Hansen, 2002). An outlier detection model predicts whether a data point is typical for a given distribution or not. An atypical data point can be either an outlier or an example of a previously unseen class (Zhang et al., 2003).

Normally, a classification model must be trained on data that includes both examples and counter-examples for each class so that the model can learn to distinguish between them. For example, a model that predicts side effects of a medication should be trained on data that includes a wide range of responses to the medication (Cherkassy and Ma, 2004; Wang, 2005).

A one-class classifier develops a profile that generally describes a typical case in the training data. Deviation from the profile is identified as an anomaly or outlier (Mohandes et al., 2004; Cortes and Vapnik, 1995). One-class classifiers are sometimes referred to as positive security models, because they seek to identify “good” behaviours and assume that all other behaviours are bad.

3.4.2.1 **Implementation of One Class SVM Method**

Oracle Data Mining uses SVM as the one-class classifier for Outlier detection. When SVM is used for anomaly detection, it has the classification mining function but no target (Wei-Zhen Lu and Wen-Jian Wang, 2005). One-class SVM models, when applied; produce a prediction and a probability for each case in the scoring data. If the prediction is 1, the case is considered typical. If the prediction is 0, the case is considered anomalous. This behaviour reflects the fact that the model is trained with normal data (Jain and Satish, 2009; Mohsen Hayati and Zahra Mohebi, 2007).

As part of implementation, the climatic data were data warehoused into Oracle database and workflow for implementing One-Class classifier on the climatic data has been performed to identify the Outliers. Columns “ANOM_SVM_1_4_PRED” shows the “0” as indication of the Outlier.
Figure 3.8 Shows the Outliers after applying one-class SVM

3.5 CONCLUSIONS

As part of Data pre-processing various techniques like raw data analysis, min-max normalization, Z-Score normalization, Binning Method has been applied successfully on the climatic data for all the groups. Data pre-processing has been considered an essential part of the research, as it is being used as an effective primary step in any data analysis research and it has the direct impact on the results of the research. It also optimizes the data mining techniques by feeding the quality data into it. Chapter 4, Chapter 5 and Chapter 6 makes use of the output of the pre-processed data. Data pre-processing remains an active and important part of this research.