

## International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

# Data Cleaning Using Identify the Missing Data Algorithm (IMDA)

L.Ramesh<sup>1</sup>, N.Marudachalam<sup>2</sup>

M.Phil Research Scholar, P.G & Research, Department of Computer Science, Dr.Ambedkar Govt. Arts College,  
Vyasarpadi, Chennai, India<sup>1</sup>

Assistant Professor, P.G & Research, Department of Computer Science, Dr.Ambedkar Govt. Arts College,  
Vyasarpadi, Chennai, India<sup>2</sup>

**ABSTRACT:** Nowadays, data cleaning solutions are very important for the large amount of data handling users in a company and others. Data cleaning, deals with detecting and removing errors and inconsistencies from data in order to develop the quality of data. There are number of works to handle the noisy data and inconsistencies in the company. But the data cleaning is specially required when integrating volume of data sources and should be addressed together with schema-related data transformations. This paper proposed a work to handle errors in volume of data sources at schema level and this work detecting and removing errors and missing data in a simplified manner and improve the quality of the data in multiple data source of the company having different sources of different locations.

**KEYWORDS:** Data cleaning, Data quality, Missing data, Data Transformations.

## I. INTRODUCTION

Data Cleansing is an activity involving a process of detecting and correcting the errors and inconsistencies in data warehouse. Thus poor quality data i.e.; dirty data present in a data mart can be avoided using various data cleaning strategies, and thus leading to more accurate and hence reliable decision making. The quality data can only be produced by cleaning the data and pre-processing it prior to loading it in the data warehouse. Data quality problems are present in single data collections, such as files and data bases, e.g., due to misspellings during data entry, missing information or other invalid data. This method gives the quality data for the end users or business people for of same kind data sources.

## II. RELATED WORKS

Data cleaning, also called data scrubbing, deals with detecting and removing errors and Inconsistencies from data in order to improve the quality of data. Data quality problems are present in single data collections, such as files and databases, e.g., due to misspellings during data entry, missing information or other invalid data.

### Importance of Data Cleaning

Any organization which uses its database for knowledge discovery and decision making will be required to keep its database maintained and error free. The data warehouse users use the features of data like correctness and accuracy of the data, which reduce with time and regular updates which in turn has an effect on the integrity of the data residing in a data warehouse. These errors thus lead to poor decision making and errors in the trend analysis. Data cleaning is thus a process of maintaining quality data by identifying incorrect or invalid or may be duplicate entries in the information system management. Also the data quality is determined by the quality of the data source.

**The measures of data quality:** Accuracy, Completeness, Consistency, Timelines, Believability, Value-added, Interpretability, Accessibility

# International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

## III. PROCESS OF DATA CLEANING



**1. Original Data :** This data is Retail marketing sale data. It is last and previous sale data to taken in the experiments of this work.

**2. Identify the Error Data:** Which type of error to be occurred in the data set? To analyze the error and then rectify the error in the data. Types of error:

Data entry problem, spelling mistake, Outlier data, incomplete data.

**3. Implement Data Cleaning:** Once the cleansing begins, it should begin to standardize and cleanse the flow of new data as it enters the system by creating scripts or workflows. These can be run in real-time or in batch (daily, weekly, monthly) depending on how much data has been taken for working. These routines can be applied to new data, or to previously keyed-in data.

### 4. Remove the noisy/inconsistent/incomplete using methods

- Data cleaning
- Data methods
- Data Analysis
- Handling noisy data

**5. Maintenance the data:** The database should be backed up continuously. The system should always be prepared for hardware or software failures and data loss. Procedures should be made as simple as possible to ensure that backups are regularly made. As the database involves with time and changes in information technology occur, data collection of data is essential to allow data access of formal data stored in former structure or design.

**6. Control:** We should be controlling your database on a whole. At the end, it is to bring the entire process full circle. Again revisit your plans from the first step and re-valuation. Make changes that outcome of the process and conduct regular reviews to make sure that the data cleaning is running with valuable and accuracy.

# International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

## IV. DATA CLEANING USING R SOFTWARE

Data cleaning May improving the statistical method based on the data. Typical actions like Suggestion or rock formation handling clearly improving the results of statistical analyses. For this reason, data cleaning should be reflecting on something a statistical operation, to be performed in a correctness manner. The R statistical environment provides a good platform for correctness data cleaning since all cleaning actions can be scripted and therefore reproduced.

### The R platform:

R is an integrated suite of software facilities for data manipulation, mathematical calculation and graphical representation. Among other things it has

1. An effective data maintain and storage facility
2. A suite of operators for calculations on arrays, in particular methods
3. A large, coherent, integrated collection of intermediate tools for data analysis, mathematical Graphical facilities for data analysis and display either directly at the monitor or on hardcopy,
4. A well developed, simple and effective programming language (called 'S') which includes conditionals, loops, user defined recursive functions and input and output facilities.
5. R is very much a technology for newly developing methods of interactive data analysis.
6. It has developed rapidly, and has been extended by a large collection of packages. However, most Programs written in R are essentially effective, written for a single piece of data analysis.

## V. PROPOSED WORK

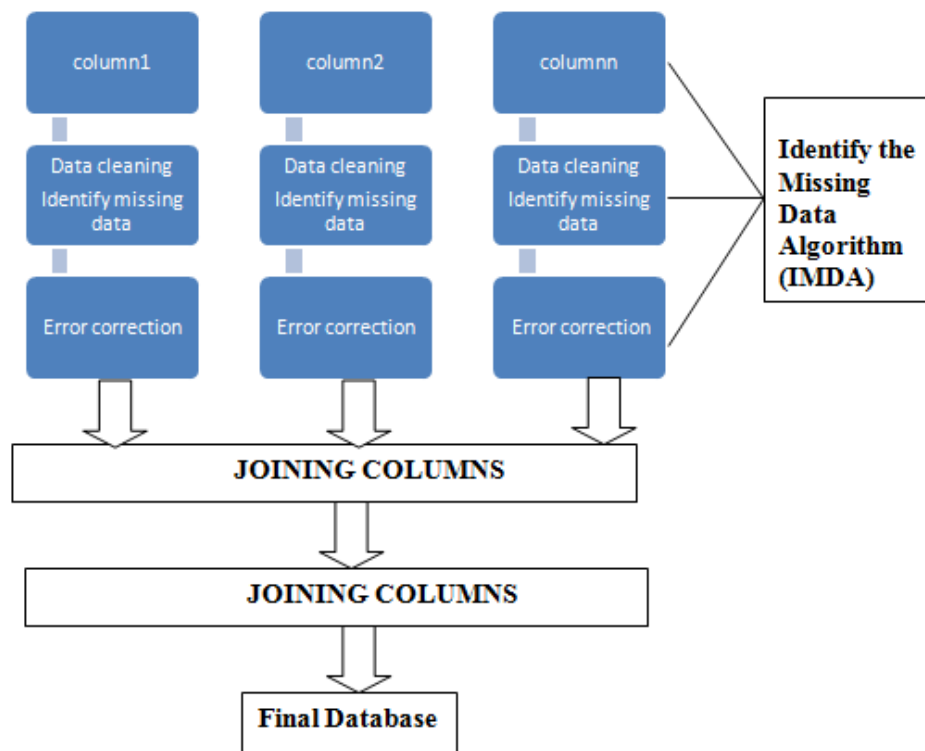


Figure 1: A Framework for Data cleaning in Multiple Data Sources

## International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

The Figure 1: shows that the work implementation of the proposed research work. This step by step process of Data cleaning methodology using R Software explains the user to detect the error and inconsistencies and then clean the noisy data efficiently.

1. Selection of column from multiple data sources
2. using the data cleaning concept for identify the missing data.
3. Correcting the error.
4. Identify Missing data Algorithm (IMDA).
5. Using the R Software.
6. Joining all the columns.
7. Finally get the cleaned data then store in the database.

**1. SELECTION OF COLUMNS FROM MULTIPLE DATA SOURCES:** In this step, data columns have been selected from various sources of single table. Data coming from different field and may have been created different way by different peoples and using different purposes. A company may have information about its customer stored in different tables, because each customer buys different services that are managed by different departments. Selection of table from the multiple data source is the effective research process.

**2. USING THE DATA CLEANING CONCEPT FOR IDENTIFY THE MISSING DATA:** Missing data, or missing values, occur when no data value is stored for the variable in our point of view. Missing data are a common occurrence and can have a meaningful effect on the conclusions that can be written from the data .Missing data can occur because of non responsibility: no information is delivered for several items or no information is provided for a whole unit. Some items are more sensitive for non responsibility.

**3. CORRECTING THE ERROR:** Cleansing ETL (extract, transform, and load) mappings for loading a clean version of the source data into the new corrected database. Compliant rows can be passed through to the clean tables without change. Noncompliant data can be summarized out, reported on, or corrected to be made compliant. More than one common data correction algorithms are built into Oracle Warehouse Builder, or you can implement your own cleansing logics.

#### **4. IDENTIFY THE MISSING DATA ALGORITHM (IMDA)**

**MISSING DATA:** Understanding the reasons why data are missing or inconsistent data can help with analyzing the remaining data. If values are missing at random, the data sample may still be representative of the country population. But if the values are missing systematically, analysis may be harder.

**MISSING COMPLETELY AT RANDOM:** Values in a data set are missing completely at random (MCAR) if the events that lead to any particular data-item being missing are independent both of observable variables and of unobservable limiting factor of interest, and occur entirely at random. When data are missing completely at random, the analyses participated on the data are unused; however, data are rarely MCAR.

**MISSING AT RANDOM:** Missing at random (MAR) is an alternative, and occurs when the missingness is related to a particular variable, but it is not related to the value of the variable that has missing data.

**MISSING NOT AT RANDOM:** Missing not at random (MNAR) is data that is missing for a specific reason (i.e. the value of the variable that's missing is related to the reason it's missing).An example is if men failed to fill in a sadness survey because of their level of depression.

#### **THE IMPORTANCE OF ADDRESSING MISSING DATA:**

1. Fatigue
2. Sensitivity
3. Lack of knowledge

# International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

4. Not applicable
5. Data processing errors
6. Programming errors

## 5. USING THE R SOFTWARE

R software using a data cleaning by identifies the missing data Algorithm (IMDA).

**6. JOINING ALL THE COLUMNS:** After the formation of each column attributes to common table attributes, all the tables would be combined and then the common table format should be saved as final database. After creating unified repository of all the tables, the another column will be created.

**7. FINALLY GET THE CLEANED DATA THEN STORE IN THE DATABASE:** The above step by step process of this work gives the cleaned data of quality for customer use. This quality data will store in to the data base and used for good decision making and conclusion.

## VI. EXPERIMENTAL RESULTS

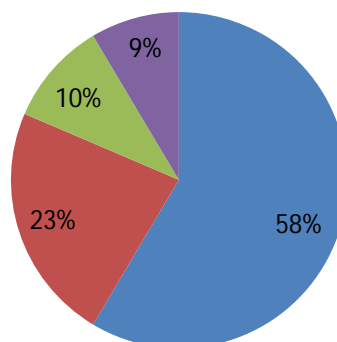
FIGURES SHOWS THE RESULTS

PARTICULARS	ORIGINAL DATA	CORRECT DATA	ERROR DATA	IMDALGORITHM
Site	400	380	120	Clear the error
Sloc	400	400	0	Clear the error
Article description	400	370	30	Clear the error
Mvt	400	200	200	Clear the error
Mat.doc	400	311	89	Clear the error
Item	400	315	85	Clear the error
Posting date	400	276	224	Clear the error
Sales value	400	340	60	Clear the error
Quantity	400	379	21	Clear the error

### 1.1 Data cleaning using Identify the Missing Data Algorithm (IMDA)

#### Experimental result for Pie chart

■ ORIGINAL DATA ■ CORRECT DATA ■ ERROR DATA ■ IMDA ALGORITHM

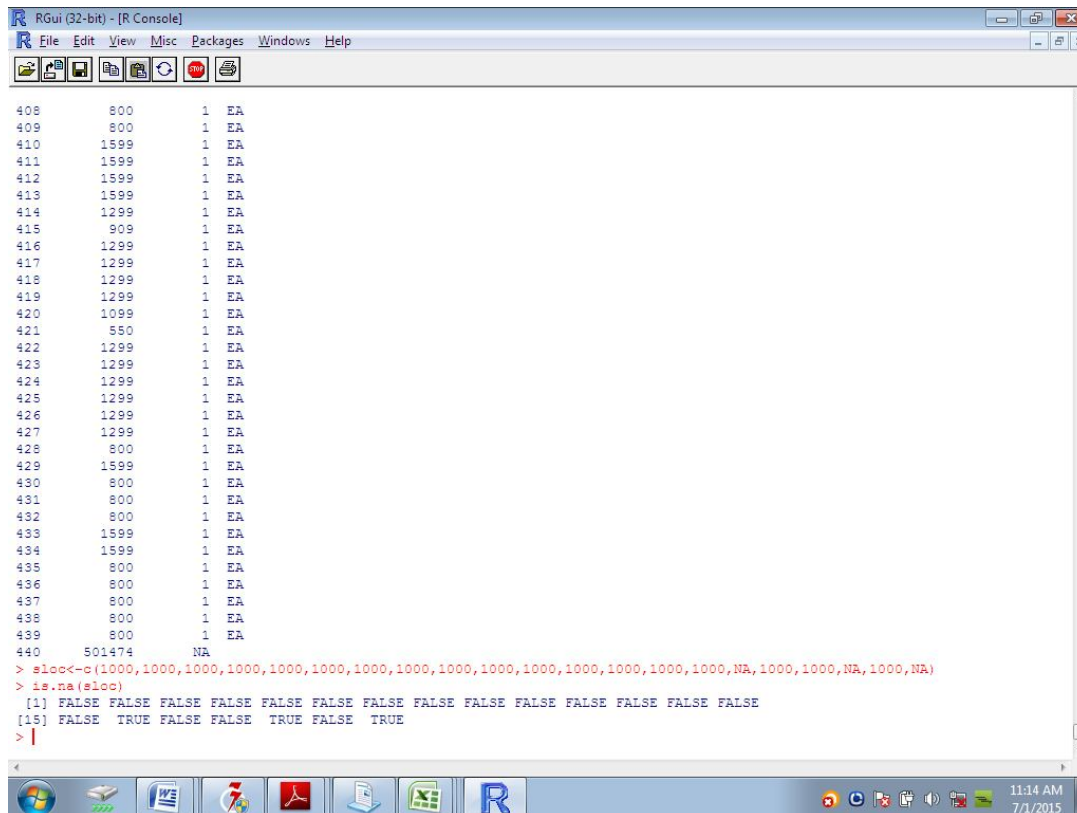


# International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

## 1.2 Experimental result to convert percentage



```

RGui (32-bit) - [R Console]
File Edit View Misc Packages Windows Help
408      800      1 EA
409      800      1 EA
410     1599      1 EA
411     1599      1 EA
412     1599      1 EA
413     1599      1 EA
414     1299      1 EA
415      909      1 EA
416     1299      1 EA
417     1299      1 EA
418     1299      1 EA
419     1299      1 EA
420     1099      1 EA
421      550      1 EA
422     1299      1 EA
423     1299      1 EA
424     1299      1 EA
425     1299      1 EA
426     1299      1 EA
427     1299      1 EA
428      800      1 EA
429     1599      1 EA
430      800      1 EA
431      800      1 EA
432      800      1 EA
433     1599      1 EA
434     1599      1 EA
435      800      1 EA
436      800      1 EA
437      800      1 EA
438      800      1 EA
439      800      1 EA
440    501474     NA
> s1cc<-c(1000,1000,1000,1000,1000,1000,1000,1000,1000,1000,1000,1000,1000,1000,1000,1000,NA,1000,1000,NA,1000,NA)
> is.na(s1cc)
[1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[15] FALSE TRUE FALSE FALSE TRUE FALSE TRUE
> |
  
```

## VII.CONCLUSION

We have implemented an automatic data cleaning for using R Software. This work is analyzed the data cleaning process compare than other approaches previously used. These sequential steps are easy to handling data cleaning and identify the missing data and information retrieval. This new work consists of six steps: Original data, Identify the error, Implement the data cleaning, Remove the noisy/inconsistent/incomplete data, Maintenance the data, Control by using Identify Missing data Algorithm (IMDA). This work will be useful to develop a powerful data cleaning tool by using the existing data cleaning techniques in a sequential order.

## REFERENCES

[1] R.Agusthiyar and K.Narashiman,"A Simplified framework fsor Data cleaning and Information Retrieval in multiple data source problems". August 2014,  
[2]Hernandez, M.A.; Stolfo, S.J.: Real-World Data is Dirty: Data Cleansing and the Merge/Purge Problem. Data Mining and Knowledge Discovery 2(1):9-37, 1998.  
[3] Anjana Sharma and naina Mehta and Iti Sharma .Reasoning with Missing Values in Multi Attribute Datasets.  
[4] Paul Jermyn, Maurice Dixon and Brian J ReadS. Preparing Clean Views of Data for Data Mining  
[5] Hall, B. W., Ward, A. W., & Comer, C. B. (1988). Published educational research: An empirical study of its quality. *The Journal of Educational Research*, 81(3), 182–189.  
[6]Havlicek, L. L., & Peterson, N. L. (1977). Effect of the violation of assumptions upon significance levels of the Pearson r. *Psychological Bulletin*, 84(2), 373–377. doi: 10.1037/0033-2909.84.2.373  
[7]Kassirer, J., & Champion, E. (1994). Peer review: Crude and understudied, but indispensable. *JAMA*, 272(2), 96–97.  
[8]Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B.,Levin, J. R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research*, 68(3), 350–386.

## International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

- [9] Lix, L., Keselman, J., & Keselman, H. (1996). Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance F test. *Review of Educational Research*, 66(4), 579–619.
- Lakshmanan, L.; Sadri, F.; Subramanian, I.N.: *SchemaSQL – A Language for Interoperability in Relational Multi- Database Systems*. Proc. 26th VLDB, 1996.
- [10] Lee, M.L.; Lu, H.; Ling, T.W.; Ko, Y.T.: *Cleansing Data for Mining and Warehousing*. Proc. 10th Intl. Conf. Database and Expert Systems Applications (DEXA), 1999.
- [11] Li, W.S.; Clifton, S.: *SEMINT: A Tool for Identifying Attribute Correspondences in Heterogeneous Databases Using Neural Networks*. In *Data and Knowledge Engineering* 33(1):49-84, 2000.
- [12] Milo, T.; Zohar, S.: *Using Schema Matching to Simplify Heterogeneous Data Translation*. Proc. 24th VLDB 1998.
- [13] Monge, A. E. *Matching Algorithm within a Duplicate Detection System*. IEEE Techn. Bulletin Data Engineering 23 (4), 2000 (this issue).
- [14] Monge, A. E.; Elkan, P.C.: *The Field Matching Problem: Algorithms and Applications*. Proc. 2nd Intl. Conf. Knowledge Discovery and Data Mining (KDD), 1996.