Understanding Different Techniques Of Data Cleaning And Different Operations Involved

[1]Ms. Muskan Agrawal, [2]Mr. Abhinav Mani Tripathi, [3]Mr. Vishal Dudgikar, [4]Ms. Neha Vekhande

[1]B.Tech in Computer Engineering
Ajeenkya DY Patil University
Pune, India
muskan.agrawal@adypu.edu.in

[2]B.Tech in Computer Engineering
Ajeenkya DY Patil University
Pune, India
abhinav.tripathi@adypu.edu.in

[3]B.Tech in Computer Engineering
Ajeenkya DY Patil University
Pune, India
vishal.dudgikar@adypu.edu.in

[4]Assistant Professor
School of Engineering-ADYPU
Pune, India
neha.vekhande@adypu.edu.in

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 10 May 2021

Abstract— Identifying the problems related to inaccurate data and then correcting of detected errors addresses to data cleansing and provides an overview of the main solution and its scope. The improved quality of unreasonable data helps in data redundancy avoiding all sorts of omissions and increasing the consistency of data in data warehouse. The uncleansed data diminishes the importance of Data Mining and Data Warehousing. Since, data is a major asset in many companies and industries, the specifying of problem statement and solving it to enhance the identification of potential errors leads to better understanding of Data and Data Mining assets. The different methods of data cleansing are surveyed to provide a complete overview on benefits of data cleaning and various techniques involved in it. The main purpose of this article is to meet the growing demands of industry by providing the standardized data with the help of research on different algorithms available in the market which proves out to be beneficial in cleaning of data.

Keywords- data cleaning, machine learning algorithms, artificial intelligence, big data, binning techniques, data cleaning cycle, big data, deep learning, data mining.

I. INTRODUCTION

Undoubtedly, data cleaning also known as data scrubbing is a very crucial part to maintain the correct data for business decisions. The heedless of removal of data errors may cause serious predicament which could result in bad quality, incorrect spelling, wrong data entry, missing information, etc. One has to forge ahead in order to deal with all the problems related to data inconsistency by providing data cleansing rules and techniques. Use of techniques such as parsing, correcting, standardizing, matching, consolidating, handling missing and noisy data will meet the requirements needed for solving bigger problems of data cleaning [1]. The industries should hire a team of professional experts to go with the entire flow of data cleansing because the process turns out to be very time consuming and expensive job. The task has to be divided into smaller threads. Right from the isolating of data through developing advanced framework, maintaining consistent format, and then finding the relation between two attributes involves pretentious steps. Furthermore, we will come up with spectrum of data mining to show the rapid rise of data science and data mining as a professional field has lured in people from all backgrounds. Engineers, data scientist, data analyst, marketing and financial graduates, human resource everyone wants a piece of data science pie. Lastly, we will go through valuable version of techniques involved and data cleansing cycle. To add a cherry on this pie, a short glimpse of benefits of data cleansing is explained along with its future scope.

II. TECHNIQUES USED FOR DATA CLEANING

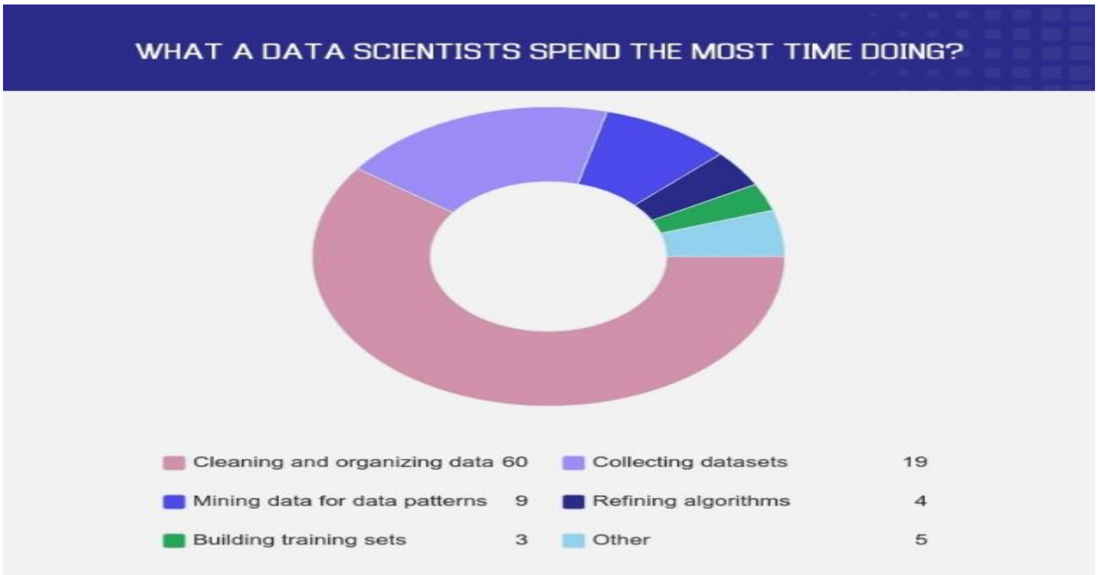


Fig. 1 TIME TAKEN BY EACH TASK (IN %)

Data analytics is all about data cleaning using various techniques. Before making it available to be explored for useful nuggets it is essential to get accustomed with the method of data cleaning. Every industry whether it be banking, telecom, healthcare, retail, hospitality, education, etc. dives in large ocean of data everyday. Over 2.5 quintillion bytes of data is generated and cleaned per day. That's why it is always analyzed and preferred to spend most of the time in cleaning and organizing data i.e. 60% by data scientists [2].

Data cleansing can be performed on different types of raw data. Identification of bunch of dirty data and then classifying them into various categories results in better outcome for further processes. So, in order to reduce the rate of false conclusions such mucky data's are divided into 5 pointers-

i. **Outlier:** If the data point differs significantly from other observations or if the data value lies outside of the frame of other values present in a particular dataset. Then, such data point/ values are called outliers. For e.g. consider names of people collaborating with a particular company for consecutive 10 years.

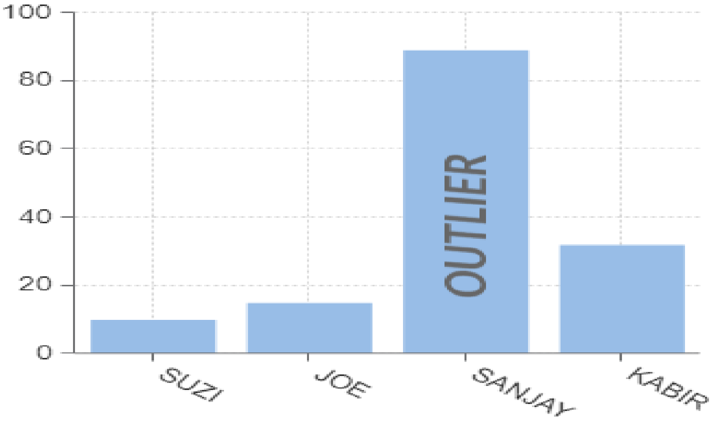
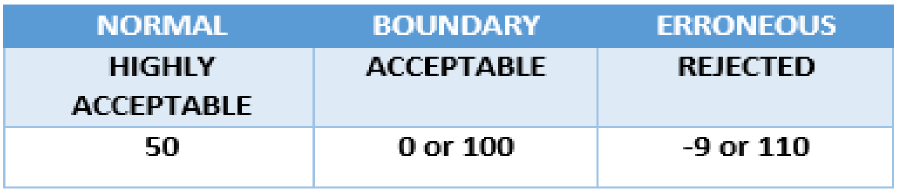


Fig. 2 DETECTION OF OUTLIER

ii. **Erroneous Data:** The data which contains error and gets rejected by the system are called erroneous data. These kind of data are neither processed nor accepted by any program. Let's consider one example- Set A contains marks of students between 0 and 100. In this case, any number above 100 or below 0 will be error as it can be seen in the given table below.

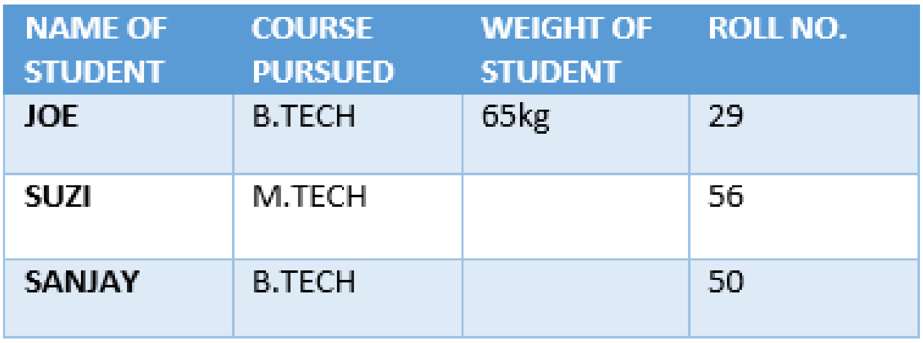


NORMAL	BOUNDARY	ERRONEOUS
HIGHLY ACCEPTABLE	ACCEPTABLE	REJECTED
50	0 or 100	-9 or 110

Fig. 3 ERRONEOUS DATA

iii. **Malicious Data:** Malicious Data is the data that forcefully causes the computer or any other system to perform undesirable actions without knowing the system's owner. Some of the examples of malicious data are- Viruses that infiltrate system or network, malware attacks, websites trying to gain your personal information, websites containing viruses that hack your system, phishing sites.

iv. **Missing Data:** The absence of data value in a particular record is called missing data. These missing data values can be obtained by various methods of data mining algorithms, regression analysis, clustering algorithms, and with the help of global constants. If the missing data value has no relation with other values or if the tuple is not important it can be completely ignored.

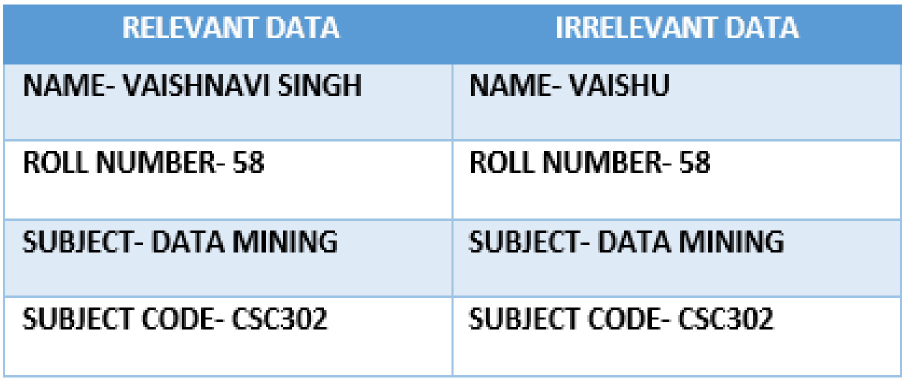


NAME OF STUDENT	COURSE PURSUED	WEIGHT OF STUDENT	ROLL NO.
JOE	B.TECH	65kg	29
SUZI	M.TECH		56
SANJAY	B.TECH		50

Fig. 4 HANDLING OF MISSING DATA

Here, the column named weight has no meaningful relation. So, it can be removed from table.

v. **Irrelevant Data:** Irrelevant Data is immaterial data which is not related to the subject. Such sort of data is inappropriate and considered as unimportant to be applied anywhere. Irrelevant data can be transformed into relevant one by just mere changing of its constraints. Say for example let's consider a table in which student details are required to be filled for a particular examination.



RELEVANT DATA	IRRELEVANT DATA
NAME- VAISHNAVI SINGH	NAME- VAISHU
ROLL NUMBER- 58	ROLL NUMBER- 58
SUBJECT- DATA MINING	SUBJECT- DATA MINING
SUBJECT CODE- CSC302	SUBJECT CODE- CSC302

Fig. 5 IRRELEVANT DATA

The table illustrates name of student in the second form to be irrelevant one because instead of original name, the student has provided her nickname (whereby this nickname is missing in student's record).

In all, there are 7 buckets for different types of data. These different types of data are dealt under various buckets mentioned here, with the use of individual techniques to make data ready for analysis.

A. PARSING

Once the data is collected the first step is to parse the data and then isolate it. Parsing the data involves breaking of data into simpler units for its better use. Let's consider a simple example of Parsing- Breaking of full name into three categories i.e. First Name, Middle Name, and Last Name.

B. CORRECTING

Correcting the data is the activity to replace ambiguous and wrong data with the correct one by applying data algorithms on data elements. The main motto behind correcting of data is to achieve consistency of data and recover the corrupted ones.

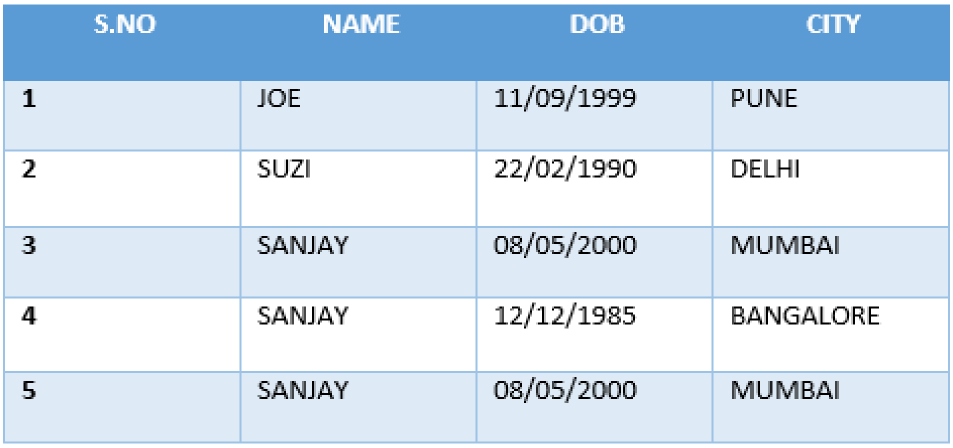
C. STANDARDIZING

This is one of the most used method for data organizing. Standardizing the data generally refers to bringing of data in a consistent format. In simpler words, one can say standardizing is done to bring all data in an acceptable format. These data values get useful for tracking. For example- conversion of units.

- Company-A deals with dollars whereas on the other hand, Company-B deals with rupees. So, while merging the data of both the companies, it is necessary to first bring the data of both the companies in an acceptable format. With the help of conversion of rupees into dollars or vice-versa.

D. MATCHING

Data matching is the task to find the repetition of a single entity to avoid duplication of data in a particular record. The removal of duplicated data results in more space and storage, it also avoids confusion. It is basically done to remove the risk of redundancy. Example- Consider the table given below.



S.NO	NAME	DOB	CITY
1	JOE	11/09/1999	PUNE
2	SUZI	22/02/1990	DELHI
3	SANJAY	08/05/2000	MUMBAI
4	SANJAY	12/12/1985	BANGALORE
5	SANJAY	08/05/2000	MUMBAI

Fig. 6 DATA MATCHING TABLE

i. **Match the data-** Sanjay was found 3 times. Amongst which 3rd and 5th row are one and the same.
ii. **Removal of duplicated data-** Deletion of either 3rd row or 5th row results in removal of duplicated data.

E. CONSOLIDATION

In this, the stored data is used to find relation between the two attributes. Combining two attributes into one by using data mining algorithms proves to be beneficial in finding data patterns and drawing insights.

F. HANDLING MISSING VALUE

The next bucket includes various means to handle the missing data. Missing value can be sorted out by following ways-

i. Ignoring of the whole row/ deleting the tuple, only if it's not important.
ii. Manually filling the data values when and only less number of missing values are required.
iii. Stating global constants (i.e. Null or NA), if less number of values are to be filled. The excess use of these constants disturbs the pattern.
iv. Use of mean and median to find out the missing values. Mean is used when there is normal distribution. Whereas, median is used when there is skewed distribution. There are two types of skewed distributions- left and right skewed [3].

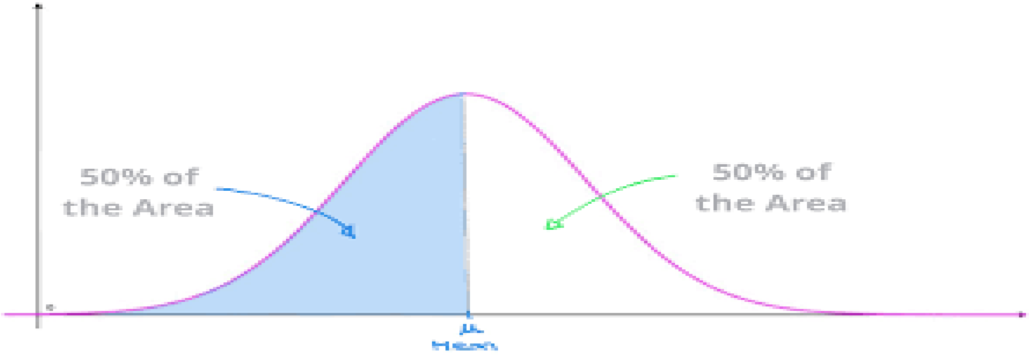


Fig. 7 NORMAL DISTRIBUTION CURVE

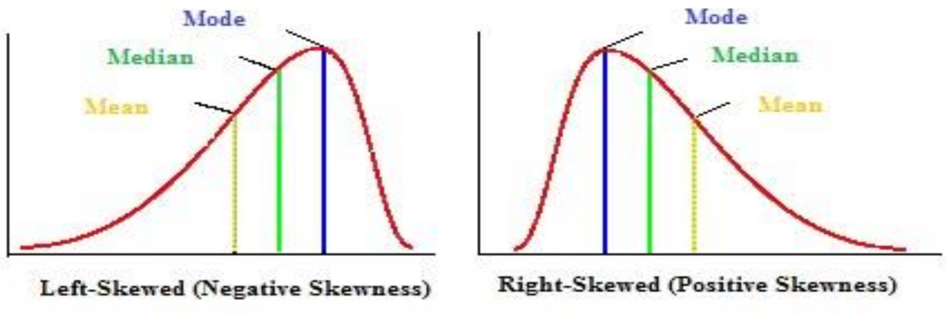


Fig. 8 LEFT AND RIGHT SKEWED DISTRIBUTION CURVE

v. With the help of data mining tools and algorithms such as clustering, decision tree, inference tools, etc.

G. NOISY DATA REMOVAL

Removal of noisy data by making use of binning technique not only helps in data smoothening but also helps in data cleansing. It consists of further three different techniques-

- **Equal Partitioning technique-** In this technique, original data value is divided into smaller intervals or bins. Suppose the given data is: 4,8,16,21,21,25,26,28,36. This dataset can be dived into smaller bins of equal numbers of data elements.
  Bin1= 4, 8, 16
  Bin2= 21, 21, 25
- **Bin mean-** Each value in a bin is replaced by the mean value of the bin. This approach is used to find average value of all the data elements present in dataset.
  Bin1= (4+8+16)/3 = 9 -> 9, 9, 9
  Bin2= (21+21+25)/3 = 22 -> 22, 22, 22
  Bin3= (26+28+36)/3 = 30 -> 30, 30, 30
- **Bin boundaries-** In smoothening by bin boundaries the values which are minimum and maximum are first identified. Then each value is replaced by closest bin boundary i.e. adapting the value of nearest boundary and then taking the minimum value in bin boundary for all the middle elements. Here, despite of middle values which are close to maximum element we prefer for minimum data value.
  Applying Bin Boundary on the same dataset:
  Bin1= 4, 8, 6 -> 4, 4, 16
  Bin2= 21, 21, 25 -> 21, 21, 25
  Bin3= 26, 28, 36 -> 26, 26, 36

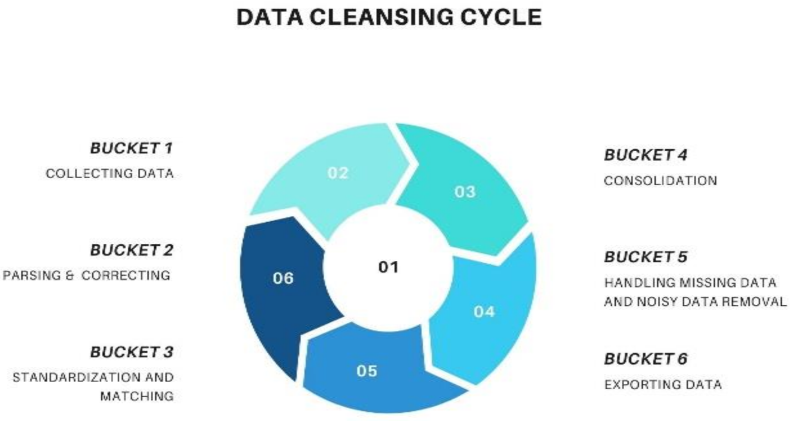


Fig. 9 DATA CLEANSING CYCLE

III. LITERATURE SURVEY

In this literature review, besides discussing what is data cleaning and how the data is cleaned we also came across the two similar kind of keywords i.e. data analyst and data scientist. Are the same or do they differ in their work? Data Analysts sift through data and seek to identify trends. Such as the business decisions which are made after drawing insights. They aim to create visual representations, graphs, and charts to showcase their data.

Whereas, data scientists deals with tasks which include hands-on-training in machine learning. They are pro in interpreting data and are expertise in the field of coding and mathematical modelling. Data scientists perform advanced programming and are capable of creating new processes for data modelling. They are highly skilled to work with algorithms, predictive models, and many more. These data scientists clean data by practicing programming languages (especially using R and Python) [4].

The broad spectrum of data science includes data mining, machine learning, artificial intelligence, cloud computing, and business intelligence to account for the degree of their importance, the complexity they show, and the business value that they grant. They grow through and analyze each and every field to provide a win-win situation for all sorts of businesses. The widening spectrum visualizes data for statistical modelling and provides neural network between them with the help of deep learning and machine learning algorithms [5].

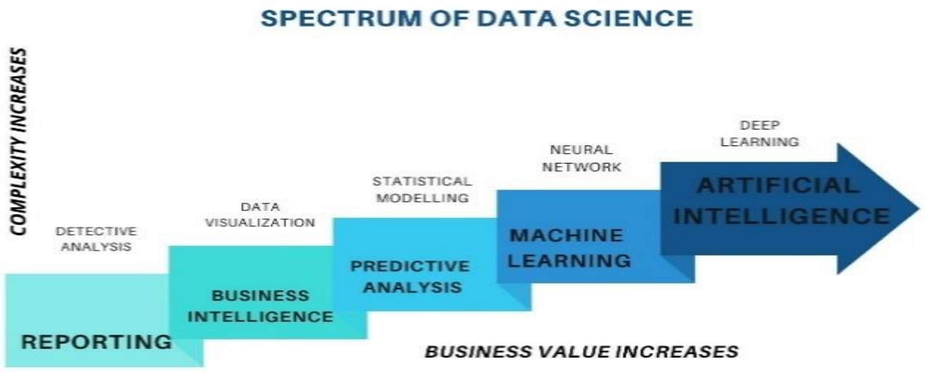


Fig. 10 SPECTRUM OF DATA SCIENCE

IV. BENEFITS OF DATA CLEANING

As cleaning is important to maintain a personal hygiene and good health especially when it comes to seeing after the covid-19 situations. Similarly, cleaning of data in data warehouse is also equally important to maintain a formatted data with a consistent representation. This will result in overall productivity with the ability to produce updated and correct information whenever needed.

- Whenever a Big Data is considered, the analysis of such big datasets are always expensive and difficult to handle. Therefore, to reduce the corrupted and inconsistent data always aid in easier management of data keeping it in a proper sequence. Such datasets are easy to access and avoids any confusion.
- The lesser the incomplete and incorrect data provided by the company or industry, the happier will be the clients. The same also ensures the less work-load on employees.
- Mapping of data and its functions to minimize the compliance risks and the threats related to data hacking and cyber/ malware attacks.
- Use of tools and techniques involved in data cleaning is an efficient way to allow improved decision making. Such data comes up with boost results and revenue.
- Removal of duplicate copies supports extra space, reduces waste and saves money.

V. CONCLUSION

Once the data is collected, data wrangling is performed to extract data. This data is cleaned through various methods and outliers are detected and deleted in this process. After scaling, it will analyze the data to form patterns and draw insights for its future use. Till here, we encountered with problems related to big data, data correcting, and analytical data and have also come across techniques used to solve such problems.

The scope of data cleaning is increasing day by day as it is becoming keystone for correcting vast amount of data errors. Explicitly, there are some methods which prove to be impractical on all sorts of data. But, what most important is to de-duplicate and link the data before preprocessing of it. Upgrading the quality of data in data warehouse seems to be necessity as concerned with the year 2020.

ACKNOWLEDGMENT

I am overwhelmed in all humbleness and gratefulness to acknowledge all those who have helped me to put these ideas, well above the level of simplicity and into something concrete. I would like to express my special thanks of gratitude to my Prof. Neha Vekhande for her continuous support as well as our HOD Dr. Biswajeet Champaty who gave me the golden opportunity to do research on topic “Data Cleaning” which also helped me in knowing about so many new techniques and algorithms. I would also thank my parents who have helped me a lot in guiding me from time to time.

REFERENCES

1. Arturas Mazeika Michael H.B’’ ohlen: Cleansing Databases of Misspelled Proper Nouns, Clean DB, Seoul, Korea, 2006.
2. Heiko Muller, Johann-Christoph Freytag. (2003). Problems, Methods, and Challenges in Comprehensive Data Cleansing, pp. 21.
3. 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.
4. Louardi BRADJI, Mahmoud BOUFAIDA. (2011). Open User Involvement in Data Cleaning for Data Warehouse Quality. International Journal of Digital Information and Wireless Communications (IJDIWC) 1(2), pp. 573.
5. Rohit Ananthakrishna1 Surajit Chaudhuri Venkatesh Ganti: Research Eliminating Fuzzy Duplicates in Data Warehouses. Proceedings of the 28th VLDB Conference, Hong Kong, China, 2002.