POORNIMA COLLEGE OF ENGINEERING JAIPUR

A Presentation On

ANALYSIS & RECOMMENDATIONS OF DATA CLEANING PROCESS FOR VARIOUS TYPES OF DATA

Group: 3CSC13 3rd SEM NSP

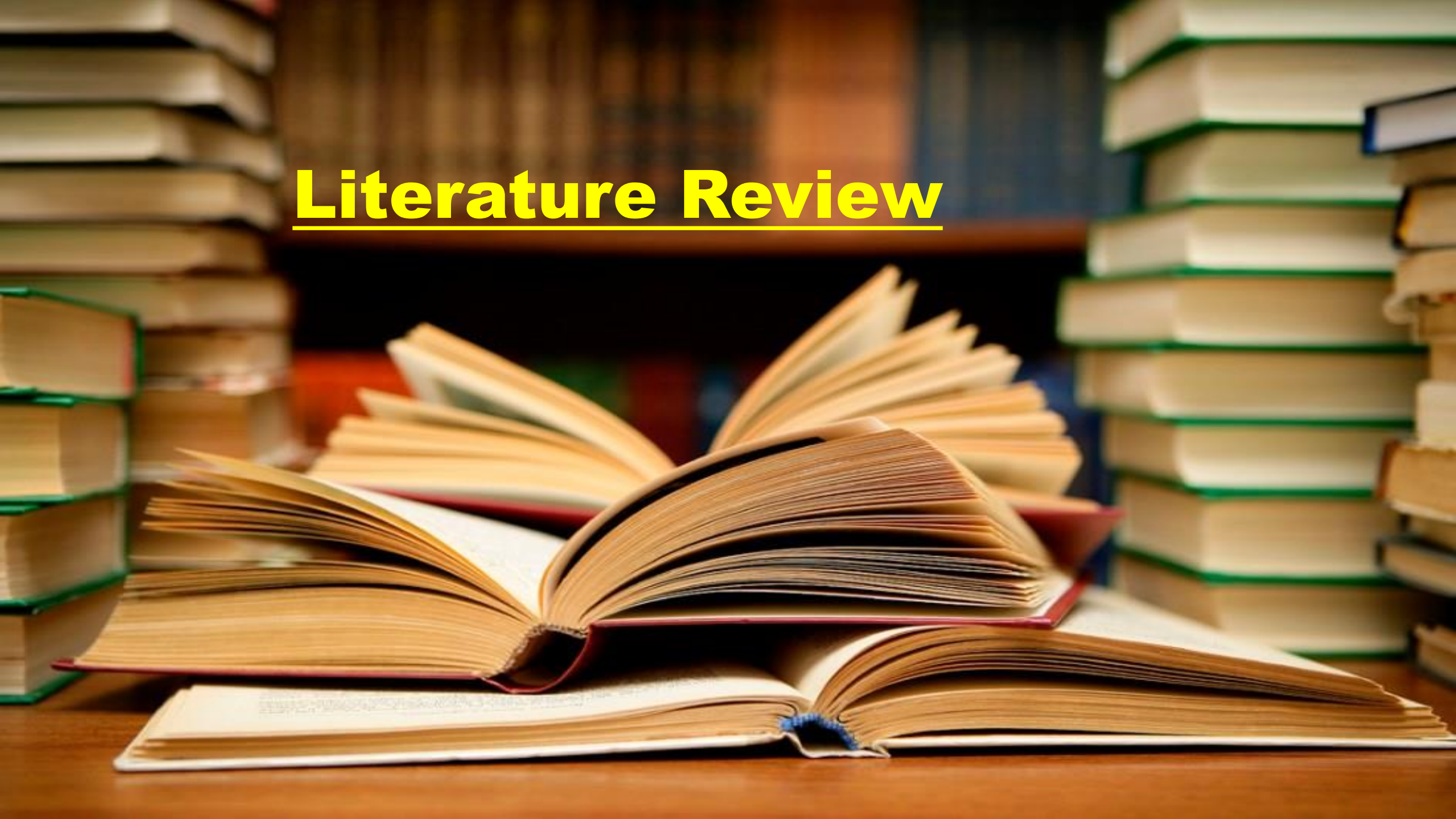
Submitted To :-
Mrs. SONAM GOUR Ma'am,
Mrs. Neelam Chaplot Ma'am,
NSP GUIDE,
Professor,
Department of CSE

Submitted By :-
Sejal Jain PCE20CS171, 20EPCCS171
SANSKAR SHARMA PCE20CS166, 20EPCCS166
SACHIN YADAV PCE20CS161, 20EPCCS161
RAHUL KHANDELWAL PCE20CS152, 20EPCCS152
YASH TRIPATHI PCE20CS200, 20EPCCS200

OUTLINES

- ABSTRACT
- SUMMARY
- APPROACH FOR DATA CLEANING
- DATA CLEANING ITERATIVE MODEL
- PROBLEM STATEMENT AND OBJECTIVE
- COMPARISION OF DATA CLEANING TOOLS
- RESEARCH DONE IN PAST
- DIFFERENT APPROACHES AND COMPARITIVE ANALYSIS
- FUTURE SCOPE
- REFERENCES

Literature Review



ABSTRACT

- The process of identifying and removing the errors
- Quality and consistency becomes significant
- Problem of data cleansing and the identification of potential errors for incorrect or inconsistent data
- False conclusion and misdirect investment

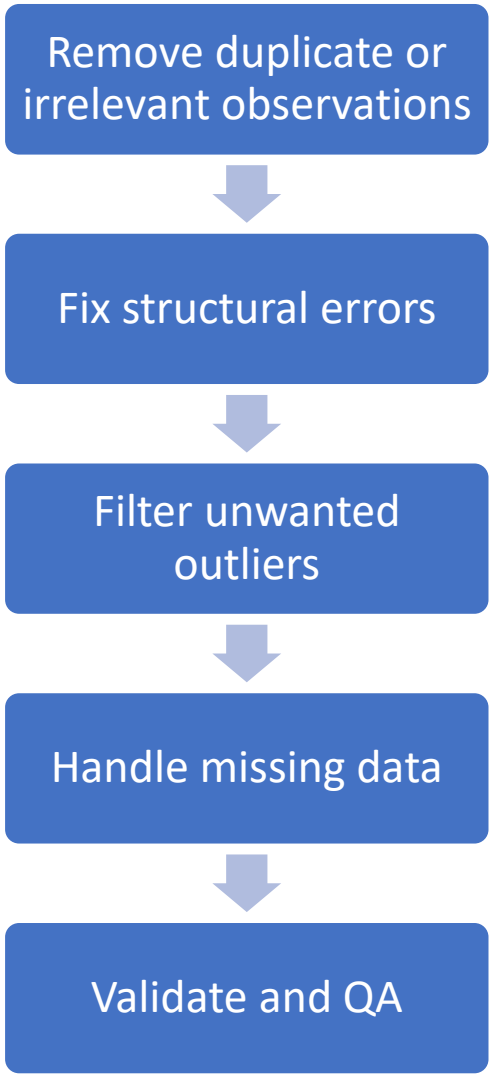
SUMMARY

- Various Data Cleaning Algorithms And Techniques
- Wide Variety Of Situations
- Data Cleaning Is Very Necessary Part
- Cleaning Methods And Approaches Depend Upon

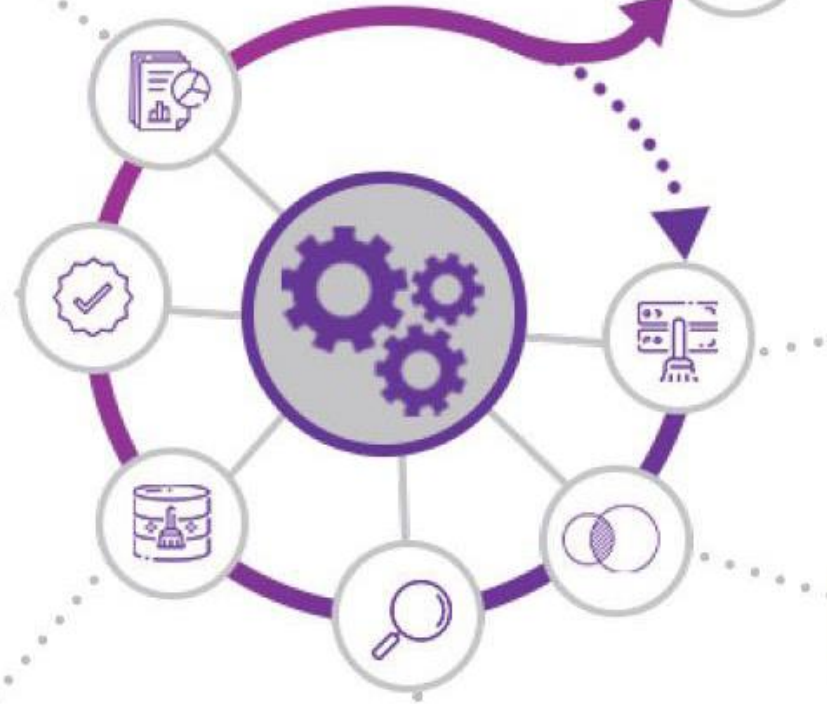
  The Type Of Data Comparison Of Data Cleaning Tools

  And Determines The Best Tool.

APPROACH FOR DATA CLEANING



DATA CLEANING ITERATIVE MODEL



6. REPORT

Report changes in the data-cleaning guidelines that you defined earlier

7. SCALE

Understand the intensity, type of work, and time, effort required to clean the complete data set

5. VERIFY

Verify against the rulebook - Verify using a naive model

1. GOALS

Validity, Accuracy, Completeness, Consistency, Uniformity

2. SAMPLE

Simple Random, Systematic, Stratified, Clustered

4. CLEAN

Format, Corrupt Data, Duplication & Irrelevance, Missing Data, Outliers

3. INSPECT

EDA can provide a quick, yet detailed idea of outliers, anomalies, and data inconsistencies

PROBLEM STATEMENT:

To analyse a data set containing uncleaned data and then apply various data cleaning methods on the data set according to the requirement.

OBJECTIVE:

- To apply the general understanding of data cleaning.
- To analyze a dataset.
- To use the data cleaning tools.
- To apply various data cleaning methods on a data set.

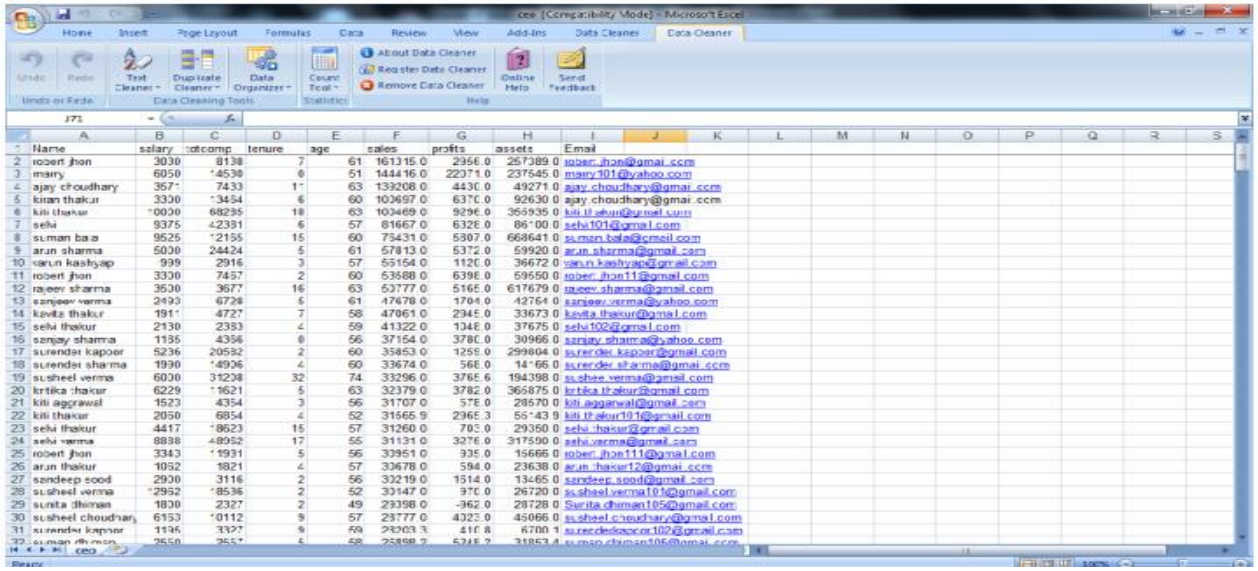
DESIGN OF SOLUTION

Color	Martin Scorsese	240	116866727	The Wolf of Wall StreetÂ	English	USA	100000000	2013	8.2
Color	Shane Black	195	408992272	Iron Man 3Â	English	USA	200000000	2013	7.2
color	Quentin Tarantino	187	54116191	The Hateful EightÂ	English	USA	44000000	2015	7.9
Color	Kenneth Lonergan	186	46495	MargaretÂ	English	usa	14000000	2011	6.5
Color	Peter Jackson	186	258355354	The Hobbit: The Desolation of SmaugÂ	English	USA	225000000	2013	7.9
	N/A	183	330249062	Batman v Superman: Dawn of JusticeÂ	English	USA	250000000	202	6.9
Color	Peter Jackson	-50	303001229	The Hobbit: An Unexpected JourneyÂ	English	USA	180000000	2012	7.9
Color	Edward Hall	180		RestlessÂ	English	UK		2012	7.2
Color	Joss Whedon	173	623279547	The AvengersÄ	English	USA	220000000	2012	8.1
Color	Joss Whedon	173	623279547	The AvengersÂ	English	USA	220000000	2012	8.1
	Tom Tykwer	172	27098580	Cloud AtlasA	English	Germany	102000000	2012	-7.5
Color	Null	158	102515793	The Girl with the Dragon TattooÂ	English	USA	90000000	2011	7.8
Color	Christopher Spencer	170	59696176	Son of GodÂ	English	USA	22000000	2014	5.6
Color	Peter Jackson	164	255108370	The Hobbit: The Battle of the Five ArmiesÂ	English	New Zealand	250000000	2014	7.5
Color	Tom Hooper	158	148775460	Les MisÃ©rablesÂ	English	USA	61000000	2012	7.6
Color	Tom Hooper	158	148775460	Les MisÃ©rablesÂ	English	USA	61000000	2012	7.6

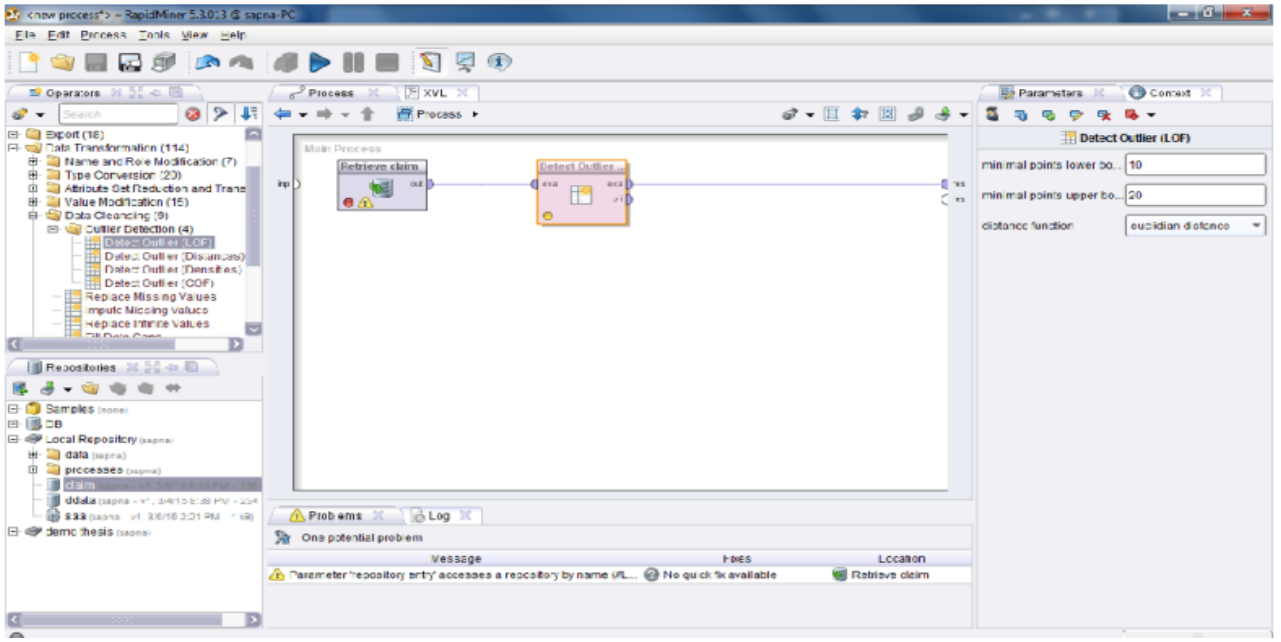
Example of Unclean Data Set

	Name	Height	Roll	Department	Address
0	A	5.2	55	CSE	polashi
1	B	5.7	99	EEE	banani
2	C	5.6	15	BME	farmgate
3	D	5.5	80	CSE	mirpur
4	E	5.3	1	ME	dhanmondi
5	F	5.8	12	ME	ishwardi
6	G	5.6	47	CE	khulna
7	H	5.5	104	CSE	uttara

Example of Cleaned Data Set



MS EXCEL WITH DATA CLEANER



RAPIDMINOR

WINPURE CLEAN AND MATCH

Comparison of Data Cleaning tools

Tools / Problems	MS Excel with data cleaner	RapidMinor	Winpure Clean & Match
Missing Values	No	Yes	Yes
Availability	Desktop	Desktop	Desktop
Duplication	Yes	Yes	Yes uses the matching
Illegal Values Elimination	No	No	Yes
Misspelling	No	No	No
Merge	No	Yes	Yes
File Format	Excel	CSV, Database, Excel, Access, binary, XML	Text files, Excel , commercial DBMS,
Ease of use	Moderate	Moderate	High

Research Done In Past

- Data Cleaning for Misspelled Proper Nouns (Border Detection Algorithm)
- Robust and Efficient Fuzzy Match for Online Data Cleaning(Fuzzy Match similarity Algorithm)
- Data Cleaning by Clustering and Association Methods (Data Mining Algorithms)

DIFFERENT APPROACHES AND COMPARATIVE ANALYSIS

	Border Detection Data Algorithm	Data Mining Algorithm-Attribute Correction Algorithm	Fuzzy Match Similarity Function Algorithm
Features	Simple, effective to compute clusters in the validated against reference to match the reference data then string data.	The given attributes are validated against reference to match the reference data to provide cleansing solution fuzzy match similarity (fms) that explicitly considers IDF token weights and input errors while comparing tuples.	data mining techniques in the area of attribute correction are: context-independent attribute correction implemented using clustering and robust results . If one techniques and context-dependent
Significance / performance	It produces good cleansing results for string data with large distances between centers of clusters and small distances within the clusters	Quality of fms is better than ed (edit distance) using two Datasets.	Algorithm shows better results for longer strings

- PROJECT STATUS
- PLANNING
- FUTURESCOPE


BIG
DATA
USE
ONE
SINCE
SOFTWARE
REQUIRING
UBIQUITOUS
RADIO-FREQUENCY
COMPLEXITY
SOLID
WIRELESS
SAN
TOLERABLE
NEEDED
QUALITIES
SIZE
PETABYTES
INTERNET
TECHNOLOGIES
PROCESSING
BURIED
LOGS
ANALYTICS
PARALLEL
BUSINESS
MOVING
ORGANIZATIONS
USING
TYPES
WORKING
GARTNER
MASSIVELY
BIOLOGICAL
RELATED
ALSO
PERFORMANCE
HUNDREDS
CREATED
DISK
ZETTABYTES
GENOMICS
COMPUTING
TOOLS
SET
SHARED
TIME
COMBAT
SIGNIFICANT
RELATIONAL
WITHIN
PROCESS
INCLUDE
SYSTEMS
DEFINITION
NETWORKS
BIOGEOCHEMICAL
RECONSIDER
SEARCH
INFORMATION
OPPORTUNITIES
CONNECTOMICS
RECORDS
DESKTOP
CURRENTLY
FC
TENS
CAPACITY
FORMS
PRESENTATIONS
WORLD'S
PRACTITIONERS
NOW
USED
DISTRIBUTED
CAPTURE
MAY
MANAGEMENT
MANAGE
GROW
SOCIAL
LARGE
STORE
DEFINING
COST
CONTINUES
TERABYTES
MPP
RESEARCH
STORAGE
EVERY
CASE
LARGER
CITATION
COMPLEX
DATABASES
INDEXING
DIFFICULTY
TARGET
ABILITY
SETS
EXAMPLES
APPLIED
ELAPSED
CURRENT
THOUGHT
SENSOR
ARCHIVES
AMOUNT
DESCRIBING

REFERENCES

[1] STEPS TO DO DATA CLEANING: https://www.tableau.com/learn/articles/what-is-data-cleaning

[2] DATA CLEANING: CURRENT APPROACHES AND ISSUES : Vaishali Chandrakant Wangikar and Ratnadeep R. Deshmukh

https://www.researchgate.net/publication/278301609_Data_Cleaning_Current_Approaches_and_Issues

[3] STUDY OF DATA CLEANING & COMPARISON OF DATA CLEANING TOOLS : Sapna Devi, Dr Arvind Kalia

https://www.ijcsmc.com/docs/papers/March2015/V4I3201599a30.pdf

[4] AN OVERVIEW STUDY ON DATA CLEANING, ITS TYPES AND ITS METHODS FOR DATA MINING : S.Lakshmi AND Dr S.V

https://acadpubl.eu/hub/2018-119-12/articles/6/1564.pdf

[5] ITERATIVE PROCESS FOR DATA CLEANING

https://innotescus.io/data-cleaning/complete-guide-iterative-process-for-data-cleaning/

[6] FUZZY MATCHING ALGORITHM

https://nanonets.com/blog/fuzzy-matching-fuzzy-logic/#:~:text=Fuzzy%20Matching%20(also%20called%20Approximate,Priceline%20in%20the%20graphic%20below

[7] BORDER DETECTION ALGORITHM

https://www.researchgate.net/publication/220803114_Cleansing_Databases_of_Misspelled_Proper_Nouns

[8] A review on data cleaning methods for Big Data : Fakhitah Ridzuan

https://www.researchgate.net/publication/338348131_A_Review_on_Data_Cleansing_Methods_for_Big_Data


THANK YOU