

INTELLICLEANSE

Capstone Project Phase - I

Vaishnavi R P

SRN : PES1PG23CA159

Dr. Lekha A

Associate Professor

Department of CA

Intellicleanse

Abstract



Data cleaning ensures data accuracy and readiness for analysis by addressing errors, inconsistencies, and inaccuracies. This online tool simplifies cleaning files like '.csv' with features like duplicate removal, missing value imputation, outlier detection, and data profiling. Its user-friendly interface helps scientists, analysts, and businesses transform raw data into actionable insights.

Intellicleanse

Introduction



- ▶ **Problem Scenario:** Inconsistent, redundant, and incomplete datasets hinder effective data analysis, leading to inaccurate insights and delayed decision-making. Manual cleaning processes are time-consuming and error-prone.
- ▶ **Proposed Solution:** Intellicleanse is a web application that automates data cleansing using machine learning algorithms. It identifies and resolves data issues like duplicates and missing values, improving data quality and decision-making efficiency.

Intellicleanse

Introduction



- ▶ **Purpose:** To enhance data quality by addressing issues such as missing values, outliers, and duplicates, enabling users to make accurate data-driven decisions.
- ▶ **Scope:** The application caters to data analysts, researchers, and enterprises by providing a comprehensive suite of data preprocessing tools, ensuring flexibility and scalability.

Intellicleanse

Related Work



- ▶ **Title:** Data Cleaning: Paving a Way for Accurate and Clean Data
- ▶ **Year:** 2024
- ▶ **Authors:** L. Ahuja, B. Singh, and R. Simonv
- ▶ **Key Findings:** This paper discusses techniques to address issues like missing values and duplicates, highlighting automated tools for efficient cleaning. It emphasizes the importance of clean data for better decision-making and predictive accuracy.

Intellicleanse

Related Work



- ▶ **Title:** CleanML: A Study for Evaluating the Impact of Data Cleaning on ML Classification Tasks
- ▶ **Year:** 2021
- ▶ **Authors:** Peng Li, Xi Rao, Jennifer Blase, Yue Zhang, Xu Chu, Ce Zhang
- ▶ **Key Findings:** CleanML systematically investigates the impact of data cleaning on machine learning (ML) classification tasks. It explores how cleaning various types of errors in real-world datasets influences ML performance using different algorithms and cleaning methods.

Intellicleanse

Related Work



- ▶ **Title:** Data Cleaning Methods
- ▶ **Year:** 2023
- ▶ **Authors:** Nilu Singh
- ▶ **Key Findings:** The paper highlights key techniques for improving data quality, including handling missing values, duplicates, and outliers. It emphasizes using tools like `isnull()`, `fillna()`, and `dropna()` for processing and underscores the importance of cleaning data to ensure accurate analysis.

Intellicleanse

Related Work



- ▶ **Title:** Data Cleaning Approaches in Data Mining
- ▶ **Year:** 2018
- ▶ **Authors:** P. V. S. V. Narayana and V. M. G. S. Pandit
- ▶ **Key Findings:** This study reviews key data cleaning techniques in data mining, focusing on handling noise, missing values, and inconsistencies. It provides insights into algorithms that ensure data quality for improved mining outcomes and discusses real-world applications highlighting the importance of robust cleaning methods.

Intellicleanse

Related Work



- ▶ **Title:** Data Cleaning for Data Analysis
- ▶ **Year:** 2019
- ▶ **Authors:** Huang Shan, E. Gubin
- ▶ **Key Findings:** This paper emphasizes the importance of data cleaning as a crucial preprocessing step in data analysis. It highlights common data issues such as missing values, outliers, and inconsistencies, and proposes methods to address these issues for improved data mining results.

Intellicleanse

Related Work



- ▶ **Title:** Data Cleaning: A Thorough Analysis and Survey on Unstructured Data
- ▶ **Year:** 2018
- ▶ **Authors:** Virender Kumar, Cherry Khosla
- ▶ **Key Findings:** This paper addresses the challenges of cleaning unstructured data, a less-explored domain compared to structured datasets. It surveys methodologies and tools tailored to handle diverse formats like text, multimedia, and semi-structured data.

Intellicleanse

Related Work



- ▶ **Title:** Possibilistic Data Cleaning
- ▶ **Year:** 2022
- ▶ **Authors:** Henning Koehler, Sebastian Link
- ▶ **Key Findings:** This paper explores a possibilistic approach to data cleaning, focusing on handling uncertainty and ambiguity in data. It introduces techniques based on fuzzy logic to clean datasets with incomplete or vague information.

Intellicleanse

Related Work



- ▶ **Title:** Time Series Data Cleaning: A Survey
- ▶ **Year:** 2019
- ▶ **Authors:** Xi Wang, Chen Wang
- ▶ **Key Findings:** This paper surveys techniques for cleaning time-series data, emphasizing methods to address unique challenges such as temporal dependencies, irregular intervals, and seasonality.

Intellicleanse

Related Work



- ▶ **Title:** Data Quality Measures and Data Cleansing for Research Information Systems
- ▶ **Year:** 2018
- ▶ **Authors:** Otmane Azeroual, Gunter Saake, Mohammad Abuosba
- ▶ **Key Findings:** The paper addresses the importance of data quality in research information systems (RIS) and explores strategies for data cleansing to ensure accurate and reliable data. It emphasizes error detection and corrective actions to improve decision-making and system trustworthiness. These measures aim to enhance the quality of research data across systems.

Intellicleanse

Related Work



- ▶ **Title:** A Review on Data Cleansing Methods for Big Data
- ▶ **Year:** 2019
- ▶ **Authors:** Fakhitah Ridzuan, Wan Mohd Nazmee Wan Zainon
- ▶ **Key Findings:** This paper reviews the data cleansing process and its challenges, specifically in the context of big data. It explores methods for identifying, detecting, and correcting data errors to enhance data quality, which is critical for accurate decision-making in organizations. The authors also discuss the limitations of traditional approaches and propose advanced techniques designed for the scale and complexity of big data.

► Hardware Requirements:

- ▶ **Processor:** Intel Core i5 (9th Gen) / AMD Ryzen 5 or higher.
- ▶ **RAM:** 8 GB (16 GB recommended).
- ▶ **Storage:** 256 GB SSD (minimum).
- ▶ **OS:** Windows 10/11, macOS Catalina+, or Linux (Ubuntu 20.04+).
- ▶ **Network:** High-speed internet.

► Software Requirements:

- ▶ **Frontend:** HTML, CSS, JavaScript
- ▶ **Backend:** Python, TensorFlow/Scikit-learn, NumPy, Pandas
- ▶ **Database:** MySQL
- ▶ **Server:** AWS

Intellicleanse

Functional Requirements



► Data Upload, Preview, and Profiling:

Users can upload .csv or .xlsx files and view an instant preview with a summary report showing column names, data types, missing values, duplicates, and key statistics.

► Redundancy and Consistency Cleaning:

Detect and remove duplicate rows or columns, and handle missing values with configurable options like mean, median, custom values, or flagging for review.

► Outlier Detection and Management with Visualizations:

Identify statistical outliers and visualize them using histograms, scatter plots, and box plots to allow intuitive management (e.g., capping, flagging, or removal).

Intellicleanse

Functional Requirements



► Data Standardization and Validation:

Normalize data formats (e.g., dates, numerical values, text case) and allow users to define and apply validation rules to ensure schema consistency and compliance.

► Automated Data Transformation Pipelines:

Create reusable workflows for repetitive cleaning tasks like data formatting, redundancy removal, or missing value imputation.

► Collaborative Cleaning:

Enable multiple users to work collaboratively on datasets with tracked changes, including the ability to save versions.

Intellicleanse

Functional Requirements



- ▶ **Export Options with Cleaning Logs:**

Download cleaned data in formats like .csv or .xlsx along with a detailed log of the changes made during the cleaning process.

- ▶ **Enhanced Cleaning Insights:**

Extend profiling functionality with statistical insights, such as data distributions, correlations, and warnings for inconsistent patterns, to guide the cleaning process.

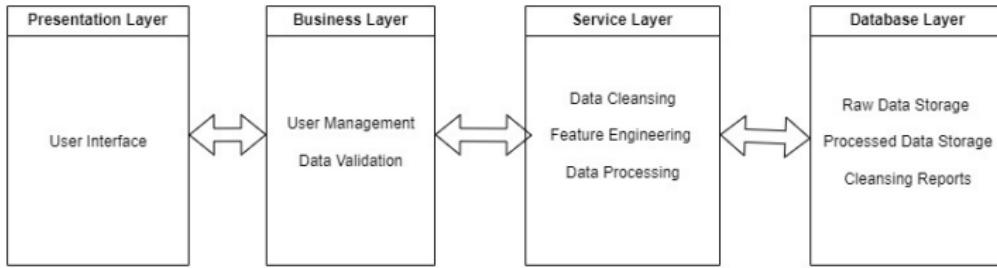
Intellicleanse

Non-Functional Requirements



- ▶ **Performance:**
The system should handle 100 concurrent users with fast page load times under typical internet conditions.
- ▶ **Security:**
User data must be encrypted, and all communications should occur over HTTPS.
- ▶ **Scalability:**
The system must scale horizontally to handle increasing user demand.
- ▶ **Availability:**
The system should have minimal downtime with failover mechanisms for continuous service.
- ▶ **Usability:**
The application should be responsive and user-friendly across all devices.

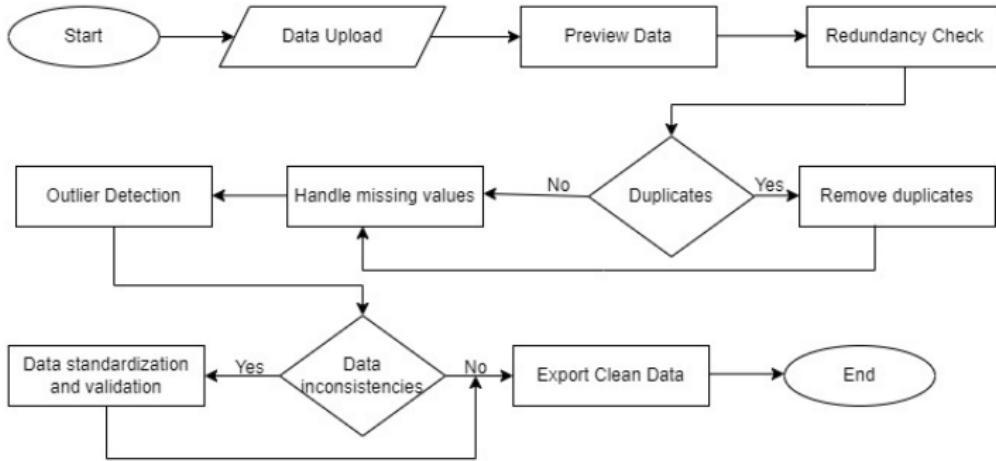
Intellicleanse Architecture Diagram



ARCHITECTURE

Intellicleanse

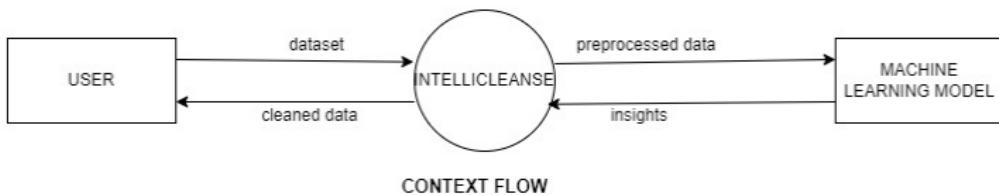
Process Flow Diagram



PROCESS FLOW

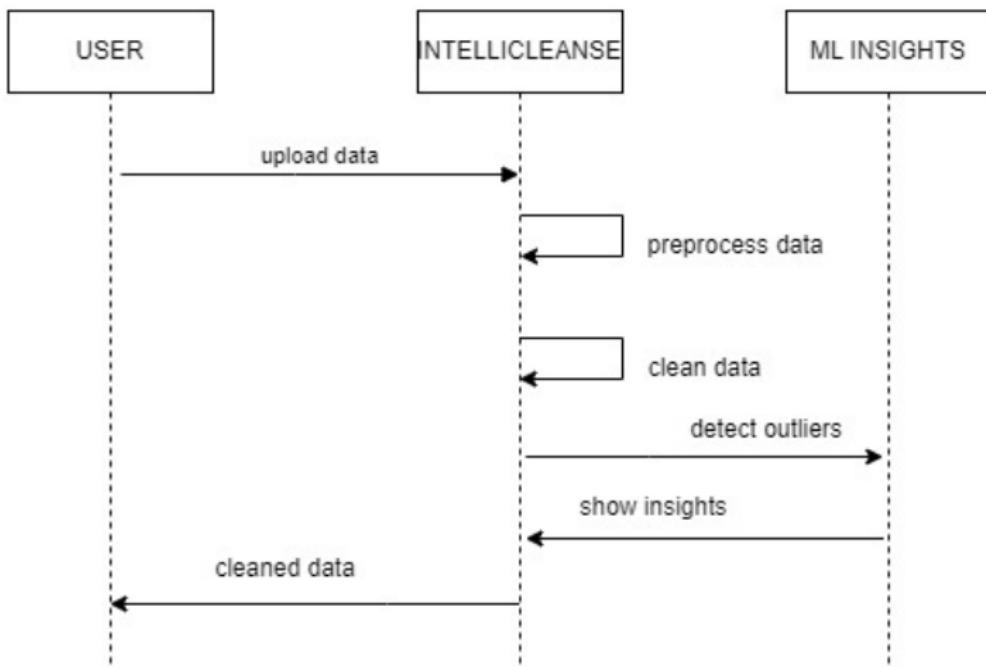
Intellicleanse

Context Flow Diagram



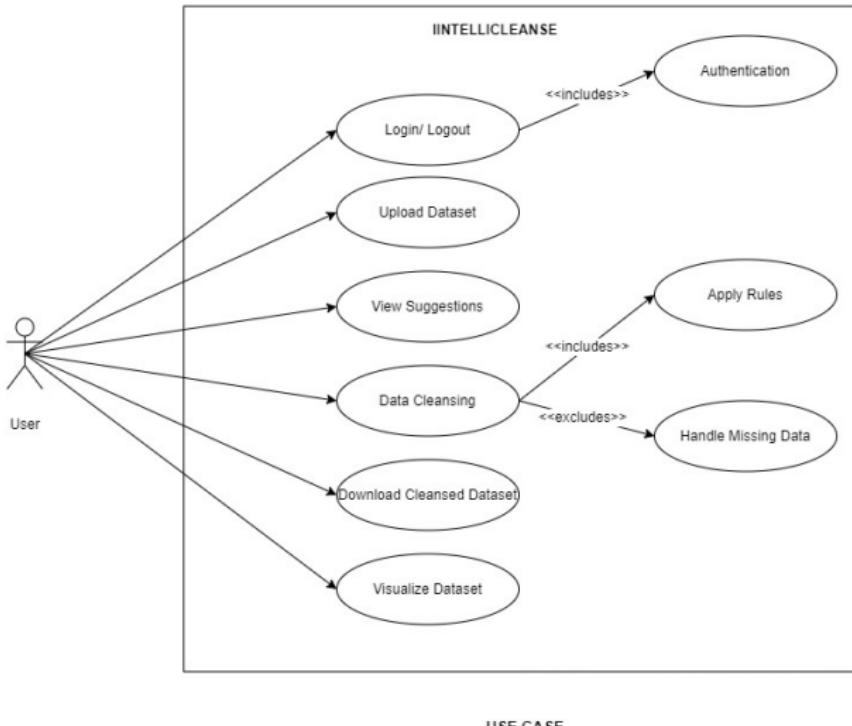
Intellicleanse

Sequence Diagram



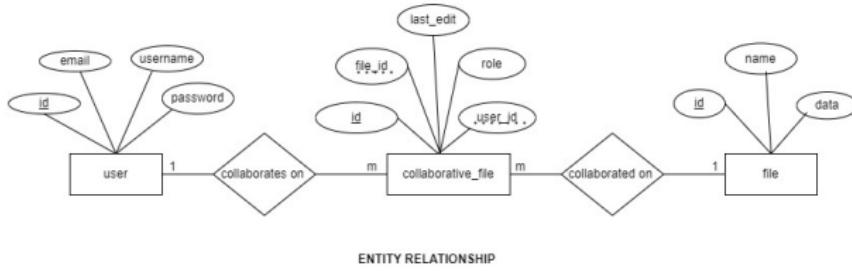
Intellicleanse

Use Case Diagram



Intellicleanse

Entity Relationship Diagram



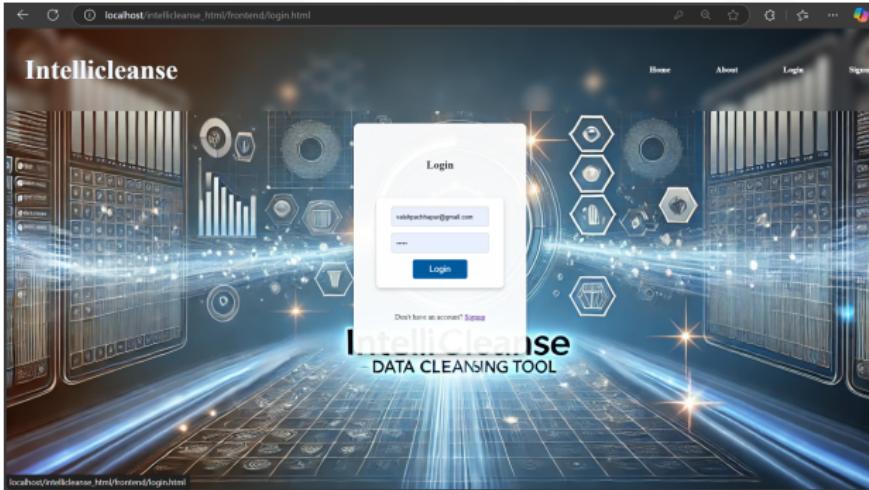
Intellicleanse Implementation



Intellicleanse Implementation



Intellicleanse Implementation



Intellicleanse Implementation



Intellicleanse Implementation



localhost:intellicleanse.html/frontend/upload.html

Upload Dataset

Choose a file Upload Download Report

Upload Status:
File "index_dataset_with_missing_values.csv" uploaded successfully and stored in database.

Dataset Preview:

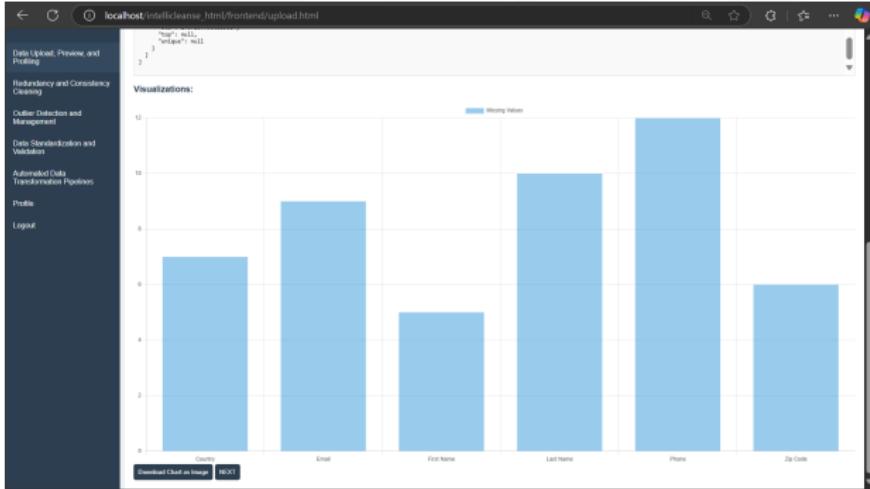
Country	Email	First Name	Last Name	Phone	Zip Code
India	riya.neety24@hotmail.com	Riya	Neety	9233064791	321995
India	aditya.gupta0@gmail.com	Aditya	Gupta	731024121	735307
India		Ayan	Gupta	8952007149	217990
India	riya.kumar16@hotmail.com	Riya	Kumar	880775538	481937
India	riya.patel0@hotmail.com	Ayan	Patel	711373427	218909
India	riya.verma60@gmail.com	Riya	Verma	9890216586	
India	riya.chopra40@yahoo.com	Riya	Chopra	8390278629	522738
	riya.patel64@gmail.com	Ishleen	Patel	7989100538	630954

Summary Report:

```
[{"values": [{"parent": "Riya", "child": "Neety", "count": 1}, {"parent": "Aditya", "child": "Gupta", "count": 1}, {"parent": "Ayan", "child": "Gupta", "count": 1}, {"parent": "Riya", "child": "Kumar", "count": 1}, {"parent": "Ayan", "child": "Patel", "count": 1}, {"parent": "Riya", "child": "Verma", "count": 1}, {"parent": "Riya", "child": "Chopra", "count": 1}, {"parent": "Ishleen", "child": "Patel", "count": 1}], "stats": [{"country": "India", "count": 8}, {"missing": "0", "count": 8}], {"registers": 8, "missing_registers": 0} ]
```

Visualizations:

Intellicleanse Implementation



Intellicleanse Implementation



localhost/intellicleanse.html/frontend/cleaning.html

Redundancy and Consistency Cleaning

Upload a New File:

Choose File: No file chosen
Get Latest File

Latest File Preview:

Country	Email	First Name	Last Name	Phone	Zip Code
India	kishna.agrawal20@rediffmail.com	Kishna	Agarwal	9135632040	569802
India	ananya.reddy32@yahoo.com	Ananya	Reddy	9444664230	180748
India	aditya.reddy76@rediffmail.com	Aditya	Reddy	9921504273	350216
India	aditya.mehra20@yahoo.com	Aditya	Mehra	9230087247	832157
null	kishna.chopra02@hotmail.com	Kishna	Chopra	8705649631	206180
India	null	Ayan	Joshi	9660107453	431421
India	ditya.reddy65@gmail.com	Ditya	Reddy	7374068877	800518
India	kishna.verma53@hotmail.com	Kishna	Verma	7945017319	784210
India	aditya.kumar33@gmail.com	Aditya	Kumar	9025766059	601008

Data Cleaning Options:

Handle Missing Values: Custom Value ▾
Nil
Close Date

Cleaning Report:

```
[{"Country": "India", "Email": "kishna.agrawal20@rediffmail.com", "First Name": "Kishna", "Last Name": "Agarwal", "Phone": "9135632040", "Zip Code": "569802"}, {"Country": "India", "Email": "ananya.reddy32@yahoo.com", "First Name": "Ananya", "Last Name": "Reddy", "Phone": "9444664230", "Zip Code": "180748"}, {"Country": "India", "Email": "aditya.reddy76@rediffmail.com", "First Name": "Aditya", "Last Name": "Reddy", "Phone": "9921504273", "Zip Code": "350216"}, {"Country": "India", "Email": "aditya.mehra20@yahoo.com", "First Name": "Aditya", "Last Name": "Mehra", "Phone": "9230087247", "Zip Code": "832157"}, {"Country": "null", "Email": "kishna.chopra02@hotmail.com", "First Name": "Kishna", "Last Name": "Chopra", "Phone": "8705649631", "Zip Code": "206180"}, {"Country": "India", "Email": null, "First Name": "Ayan", "Last Name": "Joshi", "Phone": "9660107453", "Zip Code": "431421"}, {"Country": "India", "Email": "ditya.reddy65@gmail.com", "First Name": "Ditya", "Last Name": "Reddy", "Phone": "7374068877", "Zip Code": "800518"}, {"Country": "India", "Email": "kishna.verma53@hotmail.com", "First Name": "Kishna", "Last Name": "Verma", "Phone": "7945017319", "Zip Code": "784210"}, {"Country": "India", "Email": "aditya.kumar33@gmail.com", "First Name": "Aditya", "Last Name": "Kumar", "Phone": "9025766059", "Zip Code": "601008"}]
```

Intellicleanse Implementation

localhost/intellicleanse.html/frontend/outlier.html

Data Upload, Preview and Profiling
Redundancy and Consistency Checking
Outlier Detection and Management
Data Standardization and Validation
Automated Data Transformation Pipelines
Profile
Layout

Outlier Detection and Management

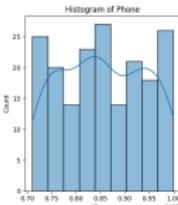
Upload a CSV or Excel file [Choose File](#) [outlier_data - 3 values.csv](#)

Outlier Detection Results

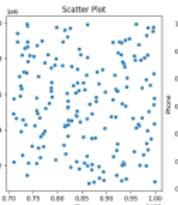
Outliers detected: ("Phone":0,"Zip Code":0)

Visualizations:

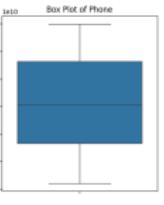
Histogram of Phone



Scatter Plot



Box Plot of Phone



Download Visualization

Intellicleanse Implementation



localhost/intellicleanse.html/frontend/datastandardization.html

Data Standardization

Standardize Data

Country	Email	First Name	Last Name	Phone	Zip Code
india	riya.reddy24@gmail.com	riya	weddy	923364791	321955
india	aditya.gupta12@gmail.com	aditya	gupta	731224121	725287
india		ayyaa	gupta	932857349	211963
india	riya.kumar15@gmail.com	riya	kumar	880787333	461937
india	ayyan.patel62@hotmail.com	ayyan	patel	7113714007	216699
india	riya.verma9@gmail.com	riya	verma	9890219596	
india	riya.chopra8@yahoo.com	riya	chopra	829977899	522738
india	ishan.patel14@gmail.com	ishan	patel	7098109230	638554
india	aditya.verma6@outlook.com	aditya	verma	9023675878	320637
india	diya.gupta10@gmail.com	diya	gupta	7106569812	575643

Insights:

Total Rows: 10
Total Columns: 6

Average Values Per Column:

- Country: NaN
- Email: NaN
- Phone: 8481580228.30
- Zip Code: NaN

Download Report

Intellicleanse

Application in Real World



- ▶ **Data Quality Enhancement:** Used by businesses to clean and preprocess large datasets, ensuring accuracy and consistency, particularly in industries like e-commerce, healthcare, and finance.
- ▶ **Data-driven Decision Making:** Helps organizations detect and eliminate data anomalies (outliers, redundancies) and standardize data for more reliable reporting and analysis.
- ▶ **Collaboration Efficiency:** Enables teams to collaboratively clean and refine data, improving the speed and quality of data analysis for actionable insights.

- ▶ D. J. Hand, H. Mannila, and P. Smyth, *Principles of Data Mining*, MIT Press, 2001.
- ▶ I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed., Morgan Kaufmann, 2016.
- ▶ M. J. Pazzani, *Data Mining: Machine Learning Perspectives*, AAAI Press, 2000.
- ▶ P. D. Turney, *Types of Data Preprocessing*, 1999.
- ▶ D. J. Salgado, A. L. A. Coelho, and M. R. Silva, *Data Cleansing: Problems and Techniques in Data Science*, Springer, 2020.
- ▶ W. H. Inmon, *Building the Data Warehouse*, 4th ed., Wiley, 2005.
- ▶ P. V. S. V. Narayana and V. M. G. S. Pandit, *Data Cleaning Approaches in Data Mining*, Springer, 2018.

- ▶ H. V. Jagadish, A. A. Imran, and M. A. U. Zaman, *Data Quality and Cleaning in Data Science*, Springer, 2020.
- ▶ A. M. MacQueen, *Efficient Methods for Data Preprocessing and Cleaning*, 2015.
- ▶ R. Agerri and A. R. M. A. Yousaf, *Data Cleaning Techniques for Data Mining: A Review*, International Journal of Computer Applications, 2015.
- ▶ M. M. M. A. Mahmud, *Data Quality and Cleaning Challenges in Machine Learning Models*, 2017.
- ▶ S. C. G. J. H. P. J. Geurts, *Data Quality and the Impact of Data Cleansing on Predictive Performance*, Journal of Machine Learning Research, 2019.
- ▶ B. L. Masnadi-Shirazi, *Outlier Detection and Data Preprocessing in Data Science*, Springer, 2014.
- ▶ J. L. Ambroise and R. N. Dubes, *Data Cleaning and Normalization: A Comprehensive Survey*, Journal of Data Science, 2002.



Thank You

Vaishnavi R P

SRN : PES1PG23CA159

Department of Computer Applications