FINDING AND REMOVING DUPLICATE ENTRIES IN A DATASET

A PROJECT REPORT

Submitted by

RISHIYANTH S (2303811724321090)

in partial fulfillment of requirements for the award of the course

AGB1252-FUNDAMENTALS OF DATA SCIENCE USING R

in

ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

K. RAMAKRISHNAN COLLEGE OF TECHNOLOGY

(An Autonomous Institution, affiliated to Anna University Chennai and Approved by AICTE, New Delhi)

SAMAYAPURAM – 621 112

JUNE- 2025

K. RAMAKRISHNAN COLLEGE OF TECHNOLOGY (AUTONOMOUS)

SAMAYAPURAM – 621 112

BONAFIDE CERTIFICATE

Certified that this project report on **"FINDING AND REMOVING DUPLICATE ENTRIES IN A DATASET"** is the bonafide work of **RISHIYANTH S(2303811724321090)** who carried out the project work during the academic year 2024 - 2025 under my supervision.

SIGNATURE

Dr.T. AVUDAIAPPAN, M.E.,Ph.D.,

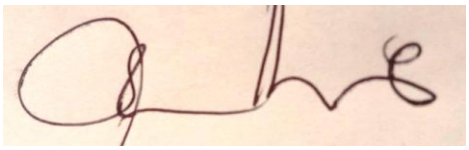
HEAD OF THE DEPARTMENT

ASSOCIATE PROFESSOR

Department of Artificial Intelligence

K.Ramakrishnan College of Technology (Autonomous)

Samayapuram–621112.



SIGNATURE

Mrs.S. MURUGAVALLI, M.E.,

SUPERVISOR

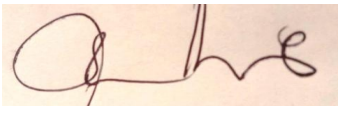
ASSISTANT PROFESSOR

Department of Artificial Intelligence

K.Ramakrishnan College of Technology (Autonomous)

Samayapuram–621112.

Submitted for the viva-voce examination held on ……02.06.2025…….



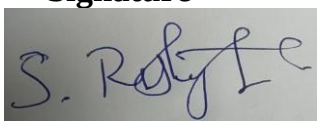
INTERNAL EXAMINER

EXTERNAL EXAMINER

DECLARATION

I declare that the project report on **"FINDING AND REMOVING DUPLICATE ENTRIES IN A DATASET"** is the result of original work done by us and best of our knowledge, similar work has not been submitted to **"ANNA UNIVERSITY CHENNAI"** for the requirement of Degree of **BACHELOR OF TECHNOLOGY**. This project report is submitted on the partial fulfilment of the requirement of the completion of the course **AGB1252-FUNDAMENTALS OF DATA SCIENCE USING R .**

Signature



RISHIYANTH S

Place: Samayapuram

Date:02.06.2025

ACKNOWLEDGEMENT

It is with great pride that I express our gratitude and in-debt to our institution "**K.Ramakrishnan College of Technology (Autonomous)**", for providing us with the opportunity to do this project.

I glad to credit honourable chairman **Dr. K. RAMAKRISHNAN**, **B.E.,** for having provided for the facilities during the course of our study in college.

I would like to express my sincere thanks to my beloved Executive Director **Dr. S. KUPPUSAMY, MBA, Ph.D.,** for forwarding to our project and offering adequate duration in completing our project.

I would like to thank **Dr. N. VASUDEVAN, M.Tech., Ph.D.,** Principal, who gave opportunity to frame the project the full satisfaction.

I whole heartily thanks to **Dr. T. AVUDAIAPPAN, M.E.,Ph.D.,** Head of the department, **ARTIFICIAL INTELLIGENCE** for providing his encourage pursuing this project.

I express my deep expression and sincere gratitude to my project supervisor **Ms.S.Murugavalli., M.E.,(Ph.D).,** Department of **ARTIFICIAL INTELLIGENCE,** for her incalculable suggestions, creativity, assistance and patience which motivated us to carry out this project.

I render our sincere thanks to Course Coordinator and other staff members for providing valuable information during the course.

I wish to express our special thanks to the officials and Lab Technicians of our departments who rendered their help during the period of the work progress.

INSTITUTE

Vision:

- To serve the society by offering top-notch technical education on par with global standards.

Mission:

- Be a center of excellence for technical education in emerging technologies by exceeding the needs of industry and society.
- Be an institute with world class research facilities.
- Be an institute nurturing talent and enhancing competency of students to transform them as all – round personalities respecting moral and ethical values.

DEPARTMENT

Vision:

- To excel in education, innovation, and research in Artificial Intelligence and Data Science to fulfil industrial demands and societal expectations.

Mission

- To educate future engineers with solid fundamentals, continually improving teaching methods using modern tools.
- To collaborate with industry and offer top-notch facilities in a conducive learning environment.
- To foster skilled engineers and ethical innovation in AI and Data Science for global recognition and impactful research.
- To tackle the societal challenge of producing capable professionals by instilling employability skills and human values.

PROGRAM EDUCATIONAL OBJECTIVES (PEO)

- **PEO1:** Compete on a global scale for a professional career in Artificial Intelligence and Data Science.
- **PEO2:** Provide industry-specific solutions for the society with effective communication and ethics.
- **PEO3** Enhance their professional skills through research and lifelong learning initiatives.

PROGRAM SPECIFIC OUTCOMES (PSOs)

- **PSO1:** Capable of finding the important factors in large datasets, simplify the data, and improve predictive model accuracy.
- **PSO2:** Capable of analyzing and providing a solution to a given real-world problem by designing an effective program.

PROGRAM OUTCOMES (POs)

Engineering students will be able to:

1. **Engineering knowledge:** Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.

2. **Problem analysis:** Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences

3. **Design/development of solutions:** Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations

4. **Conduct investigations of complex problems:** Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions

5. **Engineering tool usage:** Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations

6. **The engineer and The World:** Analyze and evaluate societal and environmental aspects while solving complex engineering problems for its impact on sustainability with reference to economy, health, safety, legal framework, culture and environment.

7. **Ethics:** Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.

8. **Individual and team work:** Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.
9. **Communication:** Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.
10. **Project management and finance:** Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.
11. **Life-long learning:** Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

ABSTRACT

Duplicate entries in datasets pose significant challenges to data quality, leading to inaccurate analysis and flawed decision-making. This project addresses the identification and removal of duplicate records to ensure data accuracy and consistency. By applying various data cleaning techniques and leveraging programming tools, the project explores methods to detect both exact and near-duplicate entries efficiently. The process improves dataset reliability, optimizes storage, and enhances the overall effectiveness of data-driven insights. Additionally, it emphasizes the importance of automating duplicate detection to handle large-scale datasets efficiently. The project ultimately contributes to building a robust data foundation, enabling more reliable and actionable outcomes across different applications.

ABSTRACT WITH POs AND PSOs MAPPING

CO 5 : BUILD DATABASES FOR SOLVING REAL-TIME PROBLEMS.

ABSTRACT	POs MAPPED	PSOs MAPPED
This project focuses on identifying and removing duplicate entries in a dataset to improve data quality and integrity. Duplicate data can lead to inaccurate analysis, increased storage costs, and unreliable results. The process involves using techniques such as exact matching, fuzzy matching, and machine learning algorithms to detect and eliminate redundant records. The goal is to ensure a clean, consistent, and reliable dataset suitable for further processing and analysis.	**PO1 -3** **PO2 -3** **PO3 -3** **PO4 -3** **PO5 -3** **PO6 -3** **PO7 -3** **PO8 -3** **PO9 -3** **PO10 -3** **PO11-3** **PO12 -3**	**PSO1 -3** **PSO2 -3** **PSO3 -3**

Note: 1- Low, 2-Medium, 3- High

TABLE OF CONTENTS

CHAPTER 1
INTRODUCTION

1.1 Objective

The objective of this study is to identify and eliminate duplicate entries within a dataset to enhance data quality, ensure accuracy, and improve the efficiency of data analysis processes. This involves implementing techniques to detect redundant records, applying appropriate data cleaning methods, and validating the integrity of the refined dataset.By removing duplicates, we reduce data redundancy and storage costs, while increasing the reliability of analytical insights.The process also contributes to better decision-making by providing consistent and trustworthy data.Automating duplicate detection can streamline data preprocessing workflows and support scalable data management practices.

1.2 Overview

This project focuses on identifying and removing duplicate entries in datasets to enhance data quality, ensure accuracy, and support reliable data analysis. Duplicate data can lead to misleading insights, inefficient processes, and flawed decision-making. The project involves the use of data cleaning techniques such as record comparison, key matching, and algorithmic deduplication to detect both exact and near-duplicate entries. Tools like Python (with libraries such as pandas) and SQL are used to implement and automate the process across different types of data sources. The goal is to streamline data preprocessing, maintain data integrity, and ensure that the final dataset is clean, consistent, and ready for effective analysis.

1.3 R Programming Concepts Used

In this project, several R programming concepts are utilized to identify and remove duplicate entries from a dataset. Data frames serve as the core structure for handling tabular data, with functions like read.csv() and data.frame() used for data import and manipulation. Key functions such as duplicated(), unique(), and anyDuplicated() help detect and manage duplicate rows. Indexing and subsetting techniques, such as df[!duplicated(df),], are employed to filter out redundant data. The **dplyr** package from the Tidyverse is extensively used, with functions like distinct(), filter(), select(), and group_by() enabling efficient and readable data cleaning workflows. Additional tools like summary(), str(), head(), and table() assist in exploring, analyzing, and validating the dataset before and after the cleaning process. Conditional statements like ifelse() are used for applying logic to identify and handle duplicates based on specific criteria, while apply-family functions (lapply(), sapply()) facilitate efficient iteration over data elements. Visualization tools such as ggplot2 may also be used to graphically identify patterns or anomalies related to duplicated data. Altogether, these R programming concepts work together to ensure the dataset is clean, consistent, and suitable for further analysis or modeling tasks.

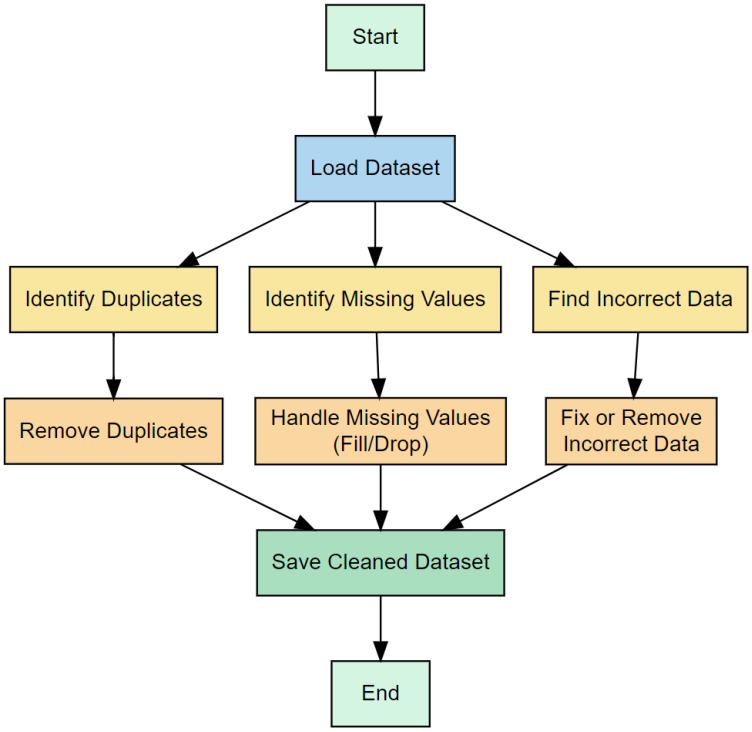
CHAPTER 2

PROJECT METHODOLOGY

2.1 Proposed Work

This project aims to systematically identify and remove duplicate entries in datasets to improve data quality and reliability. The work will start by gathering sample datasets, followed by an initial exploration to assess the extent and types of duplicates present. Various methods for detecting duplicates will be implemented, including exact matching on key attributes and approximate matching using similarity metrics to capture near-duplicates. Programming tools such as R will be used, employing functions like duplicated() and packages like **dplyr** to efficiently filter and clean the data. The process will include techniques to handle different scenarios, such as partial duplicates and inconsistent formatting. After duplicates are identified, the project will apply strategies to remove or merge records to maintain data integrity. Finally, the cleaned dataset will be validated through summary statistics and quality checks to ensure accuracy. The project will conclude by proposing automated workflows to make the duplicate detection and removal process scalable and repeatable for future datasets.

2.2 Block Diagram



CHAPTER 3

MODULE DESCRIPTION

3.1 Data Collection and Import Module

This module is responsible for acquiring the dataset from various sources such as CSV files, Excel spreadsheets, databases, or APIs. It includes tasks like loading data into the R environment using functions such as read.csv(), read_excel(), or database connection packages. Proper handling of different file formats and ensuring data is imported without corruption or loss is a key part of this module. It also includes initial data type assignment and setting up the dataset for further processing.

3.2 Data Exploration and Preprocessing Module

In this module, the imported dataset is thoroughly examined to understand its structure, content, and quality. Techniques such as summary(), str(), head(), and visualization tools like histograms or bar plots help identify missing values, outliers, inconsistencies, and potential duplicate records. Basic preprocessing steps include trimming whitespace, standardizing text cases (e.g., converting to lowercase), correcting common data entry errors, and handling missing or null values through imputation or removal. This module may also involve transforming data types (e.g., converting strings to factors or dates) to ensure consistency across columns. Additionally, normalization or standardization of data might be performed if duplicates are to be detected based on numerical thresholds. Exploratory data analysis (EDA) in this module provides insights into patterns and anomalies that could affect duplicate detection accuracy. Finally, this stage ensures the dataset is clean, consistent, and well-structured to improve the effectiveness of subsequent duplicate identification and removal steps.

3.3 Duplicate Detection Module

This module focuses on identifying duplicate entries within the dataset. It employs exact matching methods using functions like duplicated() to find identical rows or specific key columns that uniquely identify records. For more complex cases, approximate or fuzzy matching techniques are used, such as string similarity algorithms (e.g., Levenshtein distance) to detect near-duplicates caused by typographical errors or formatting differences. The module may also use grouping and filtering strategies from packages like **dplyr** to efficiently flag duplicates based on user-defined criteria.

3.4 Duplicate Removal and Validation Module

After duplicates are detected, this module handles their removal or consolidation. Strategies may include keeping the first occurrence, merging records with complementary information, or removing redundant entries altogether. The module ensures data integrity by performing consistency checks and validating the cleaned dataset against initial statistics. Validation methods might include comparing row counts, verifying unique key constraints, and conducting spot checks to ensure no critical data is lost during the cleaning process. Finally, this module prepares the dataset for downstream analysis or storage.

CHAPTER 4

CONCLUSION & FUTURE SCOPE

Conclusion

The presence of duplicate entries in datasets can significantly impact the accuracy and reliability of data analysis and decision-making. This project demonstrated effective methods for identifying and removing both exact and near-duplicate records using programming tools and data cleaning techniques. By applying these methods, data quality is enhanced, redundancy is minimized, and storage efficiency is improved. The cleaned and validated dataset provides a more trustworthy foundation for further analysis, reporting, and business intelligence. Moreover, automating the duplicate detection process can streamline future data management workflows, ensuring ongoing data integrity and consistency. Overall, this project highlights the critical role of data cleaning in maintaining high-quality datasets that support accurate and actionable insights.

Future Scope

- Future developments can incorporate advanced machine learning and AI-based techniques to enhance the detection of near-duplicates, especially in large, complex, and unstructured datasets such as text, images, and sensor data. These intelligent methods will improve accuracy, reduce manual intervention, and handle subtle variations that traditional methods might miss.
- The creation of automated, real-time duplicate detection systems integrated with big data platforms and cloud services will enable continuous monitoring and cleaning of datasets as they are collected or updated. This will help maintain data quality in dynamic environments, support scalable data management, and facilitate faster, more reliable analytics and decision-making.

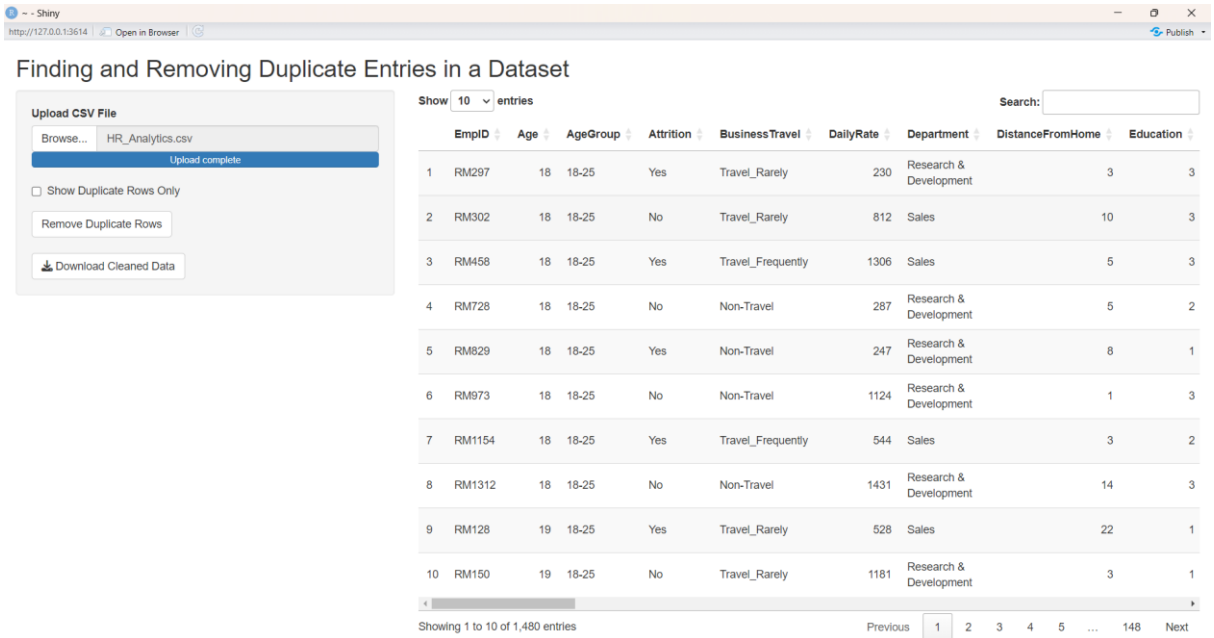
CHAPTER 5

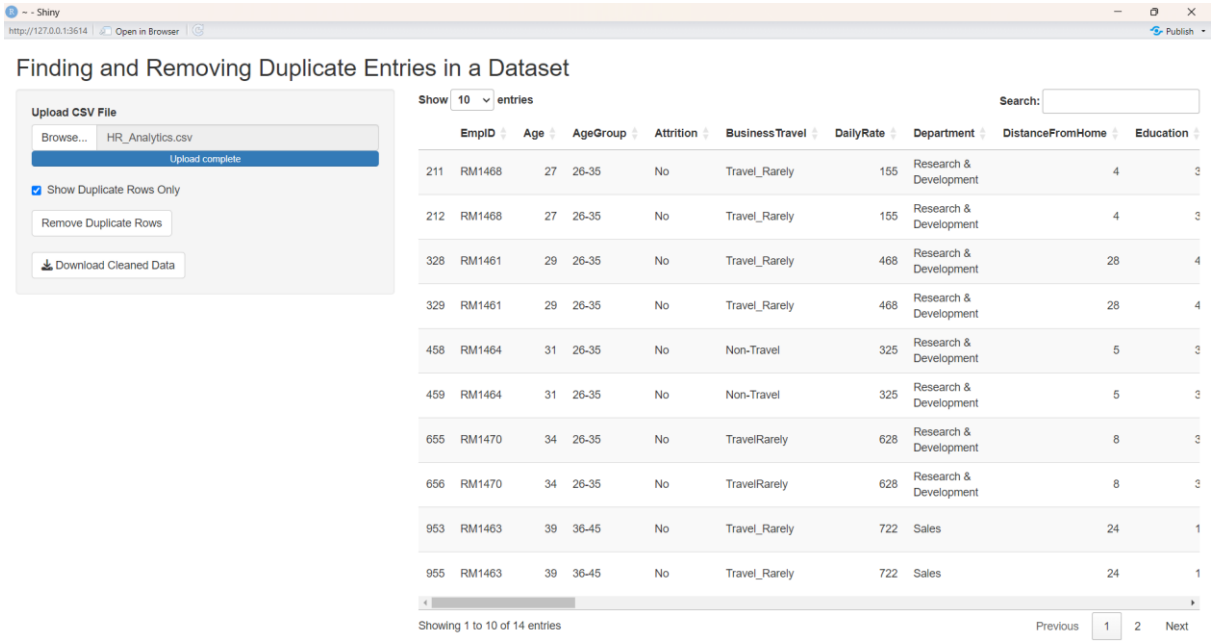
APPENDIX A – SOURCE CODE

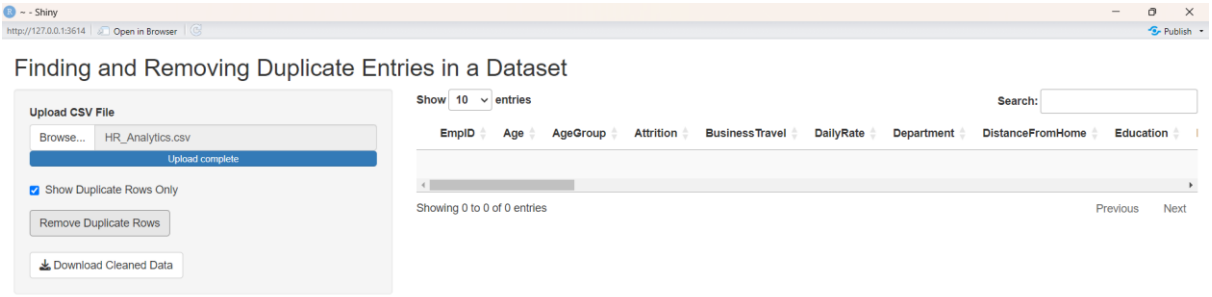
```
# Install packages if not already installed
# install.packages("shiny")
# install.packages("DT")

library(shiny)
library(DT)

ui <- fluidPage(
  titlePanel("Finding and Removing Duplicate Entries in a Dataset"),

  sidebarLayout(
    sidebarPanel(
      fileInput("file", "Upload CSV File", accept = c(".csv")),
      checkboxInput("showDup", "Show Duplicate Rows Only", value =
FALSE),
      actionButton("removeDup", "Remove Duplicate Rows"),
      br(),
      br(),
      downloadButton("downloadData", "Download Cleaned Data")
    ),

    mainPanel(
      DTOutput("table")
    )
  )
)
```

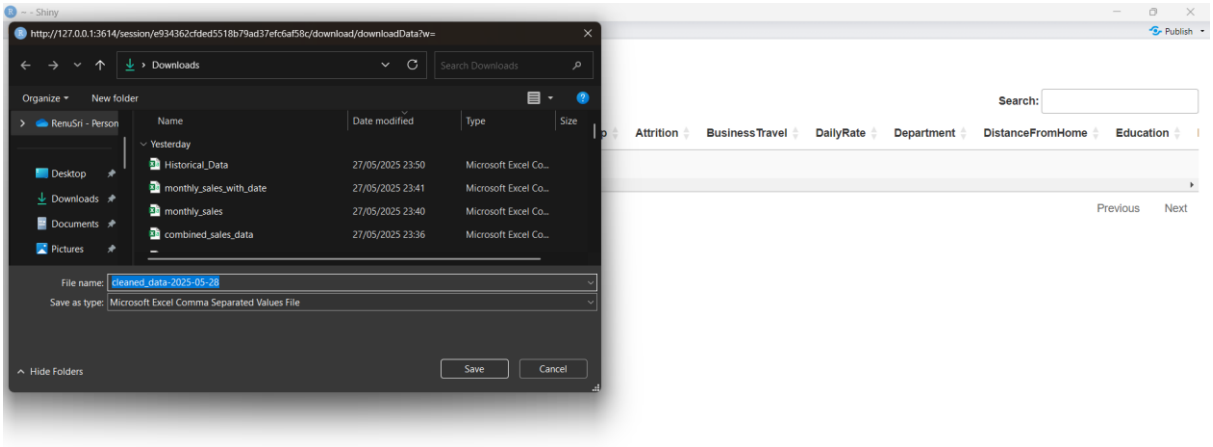
```
server <- function(input, output, session) {
  # Reactive value to store data
  data <- reactiveVal(NULL)

  # Load data from uploaded file
  observeEvent(input$file, {
    req(input$file)
    df <- read.csv(input$file$datapath, stringsAsFactors = FALSE)
    data(df)
  })

  # Reactive expression to identify duplicates
  duplicates <- reactive({
    df <- data()
    req(df)
    df[duplicated(df) | duplicated(df, fromLast = TRUE), ]
  })

  # Reactive expression for the data to display
  displayData <- reactive({
    req(data())
    if (input$showDup) {
      duplicates()
    } else {
      data()
    }
  })

  # Remove duplicates when button clicked
```

```
  observeEvent(input$removeDup, {
    df <- data()
    req(df)
    df <- df[!duplicated(df), ]
    data(df)
  })

  # Render data table
  output$table <- renderDT({
    req(displayData())
    datatable(displayData(), options = list(pageLength = 10, scrollX = TRUE))
  })

  # Download cleaned data
  output$downloadData <- downloadHandler(
    filename = function() {
      paste("cleaned_data-", Sys.Date(), ".csv", sep = "")
    },
    content = function(file) {
      write.csv(data(), file, row.names = FALSE)
    }
  )
}

shinyApp(ui, server)
```


~ - Shiny
http://127.0.0.1:3614
Open in Browser
Publish
Finding and Removing Duplicate Entries in a Dataset
Upload CSV File
Browse...
HR_Analytics.csv
Upload complete
Show Duplicate Rows Only
Remove Duplicate Rows
Download Cleaned Data
Show 10 entries
Search:
EmpID Age AgeGroup Attrition BusinessTravel DailyRate Department DistanceFromHome Education
1 RM297 18 18-25 Yes Travel_Rarely 230 Research & Development 3 3
2 RM302 18 18-25 No Travel_Rarely 812 Sales 10 3
3 RM458 18 18-25 Yes Travel_Frequently 1306 Sales 5 3
4 RM728 18 18-25 No Non-Travel 287 Research & Development 5 2
5 RM829 18 18-25 Yes Non-Travel 247 Research & Development 8 1
6 RM973 18 18-25 No Non-Travel 1124 Research & Development 1 3
7 RM1154 18 18-25 Yes Travel_Frequently 544 Sales 3 2
8 RM1312 18 18-25 No Non-Travel 1431 Research & Development 14 3
9 RM128 19 18-25 Yes Travel_Rarely 528 Sales 22 1
10 RM150 19 18-25 No Travel_Rarely 1181 Research & Development 3 1
Showing 1 to 10 of 1,480 entries
Previous 1 2 3 4 5 ... 148 Next


~ - Shiny
http://127.0.0.1:3614
Open in Browser
Publish
Finding and Removing Duplicate Entries in a Dataset
Upload CSV File
Browse...
HR_Analytics.csv
Upload complete
Show Duplicate Rows Only
Remove Duplicate Rows
Download Cleaned Data
Show 10 entries
Search:
EmpID Age AgeGroup Attrition BusinessTravel DailyRate Department DistanceFromHome Education
211 RM1468 27 26-35 No Travel_Rarely 155 Research & Development 4 3
212 RM1468 27 26-35 No Travel_Rarely 155 Research & Development 4 3
328 RM1461 29 26-35 No Travel_Rarely 468 Research & Development 28 4
329 RM1461 29 26-35 No Travel_Rarely 468 Research & Development 28 4
458 RM1464 31 26-35 No Non-Travel 325 Research & Development 5 3
459 RM1464 31 26-35 No Non-Travel 325 Research & Development 5 3
655 RM1470 34 26-35 No TravelRarely 628 Research & Development 8 3
656 RM1470 34 26-35 No TravelRarely 628 Research & Development 8 3
953 RM1463 39 36-45 No Travel_Rarely 722 Sales 24 1
955 RM1463 39 36-45 No Travel_Rarely 722 Sales 24 1
Showing 1 to 10 of 14 entries
Previous 1 2 Next


~ - Shiny
http://127.0.0.1:3614
Open in Browser
Publish
Finding and Removing Duplicate Entries in a Dataset
Upload CSV File
Browse...
HR_Analytics.csv
Upload complete
Show Duplicate Rows Only
Remove Duplicate Rows
Download Cleaned Data
Show 10 entries
Search:
EmpID
Age
AgeGroup
Attrition
BusinessTravel
DailyRate
Department
DistanceFromHome
Education
Showing 0 to 0 of 0 entries
Previous
Next


~ - Shiny
Publish
http://127.0.0.1:3614/session/e934362cfded5518b79ad37efc6af58c/download/downloadData?w=
Downloads
Search Downloads
Organize
New folder
RenuSri - Person
Desktop
Downloads
Documents
Pictures
Name
Date modified
Type
Size
Yesterday
Historical_Data
27/05/2025 23:50
Microsoft Excel Co...
monthly_sales_with_date
27/05/2025 23:41
Microsoft Excel Co...
monthly_sales
27/05/2025 23:40
Microsoft Excel Co...
combined_sales_data
27/05/2025 23:36
Microsoft Excel Co...
File name:
cleaned_data-2025-05-28
Save as type:
Microsoft Excel Comma Separated Values File
Hide Folders
Save
Cancel
Search:
Attrition
BusinessTravel
DailyRate
Department
DistanceFromHome
Education
Previous
Next

REFERENCES

1. **Grolemund, G., & Wickham, H. (2017). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly Media.**
   – A comprehensive guide to data cleaning and manipulation in R using tools like dplyr and tidyr.
2. R Core Team. (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org
   – Official documentation and resources for the R programming language.
3. Wickham, H., François, R., Henry, L., & Müller, K. (2023). *dplyr: A Grammar of Data Manipulation*. R package version 1.1.2. https://CRAN.R-project.org/package=dplyr
   – Reference for using dplyr, a core package for data filtering and duplicate removal.