Data Imputation in R: Handling Missing Values with readr Package

Name: Sesha Sai Chand Avula

Institution: IIIT Sricity

Course: IDA

Date: August 28, 2025

Overview:

This project implements a missing value imputation system in R using the readr package. It creates a CSV file with three integer columns, each with 12 records and two NA values. The NA values are imputed with the largest and second-largest values in each column, and the updated data is saved back to the CSV file. The project demonstrates proficiency in R programming, data handling, and CSV operations, using R 4.5.1 and readr for efficient file processing.

2 Task Description:

The objective is to generate a CSV file with three integer columns, each containing 12 records and two NA values. Each column's NA values are replaced with the largest and second-largest values from that column. For example, a column with values {10, 20, NA, 15, 25, 5, NA, 30, 12, 8, 22, 18} would have NA values replaced with 30 (largest) and 25 (second largest). The modified data is written back to the CSV, with checks to confirm no NA values remain.

3 Implementation Details:

3.1 Installing readr:

```
install.packages("readr")
library(readr)
cat("Loaded readr version:", packageVersion("readr"), "\n")
cat("Package initialization complete.\n")
```

Description: The readr package is installed from CRAN and loaded. The version is displayed to ensure proper setup and reproducibility.

3.2 Data Creation

```
set.seed(123)


col1 <- sample(1:100, 10, replace = TRUE)

col2 <- sample(50:200, 10, replace = TRUE)

col3 <- sample(10:80, 10, replace = TRUE)


col1[c(3, 7)] <- NA

col2[c(1, 9)] <- NA

col3[c(4, 8)] <- NA


sample_data <- data.frame(

  Column1 = col1,

  Column2 = col2,

  Column3 = col3

)


print(sample_data)


write_csv(sample_data, "sample_data.csv")


cat("\nCSV file 'sample_data.csv' created successfully!\n")

cat("File contains", nrow(sample_data), "records with", ncol(sample_data), "columns.\n")

cat("Each column has 2 NA values.\n")
```

Description: A reproducible dataset is created using set.seed(456). Three columns with 12 random integers each are generated in distinct ranges, with two NA values placed in different positions per column. The data is saved to datainput.csv.

3.3 Imputation Function

```
impute_na_with_largest <- function(column) {

  non_na_values <- column[!is.na(column)]


  sorted_values <- sort(non_na_values, decreasing = TRUE)
```

```
  largest <- sorted_values[1]
  second_largest <- sorted_values[2]

  na_positions <- which(is.na(column))

  if (length(na_positions) >= 1) {
    column[na_positions[1]] <- largest
  }
  if (length(na_positions) >= 2) {
    column[na_positions[2]] <- second_largest
  }

  return(column)
}
```

3.4 total code:

```
install.packages("readr")
library(readr)
set.seed(456)

col1 <- sample(1:100, 12, replace = TRUE)
col2 <- sample(50:200, 12, replace = TRUE)
col3 <- sample(10:80, 12, replace = TRUE)

col1[c(3, 7)] <- NA
col2[c(1, 9)] <- NA
col3[c(4, 8)] <- NA

sample_data <- data.frame(
  Column1 = col1,
  Column2 = col2,
  Column3 = col3
```

```
)

print(sample_data)

write_csv(sample_data, "datainput.csv")

cat("\nCSV file 'datainput.csv' created successfully!\n")
cat("File contains", nrow(sample_data), "records with", ncol(sample_data), "columns.\n")
cat("Each column has 2 NA values.\n")

impute_na_with_largest <- function(column) {
  non_na_values <- column[!is.na(column)]

  sorted_values <- sort(non_na_values, decreasing = TRUE)

  largest <- sorted_values[1]
  second_largest <- sorted_values[2]

  na_positions <- which(is.na(column))

  if (length(na_positions) >= 1) {
    column[na_positions[1]] <- largest
  }
  if (length(na_positions) >= 2) {
    column[na_positions[2]] <- second_largest
  }

  return(column)
}

data <- read_csv("datainput.csv", show_col_types = FALSE)

cat("Imputing NA values...\n")
```

```
data$Column1 <- impute_na_with_largest(data$Column1)
data$Column2 <- impute_na_with_largest(data$Column2)
data$Column3 <- impute_na_with_largest(data$Column3)

cat("Data after imputation:\n")
print(data)

write_csv(data, "datainput.csv")

remaining_na <- sum(is.na(data))
cat("\nVerification: Remaining NA values:", remaining_na, "\n")
if (remaining_na == 0) {
  cat("✓ All NA values have been successfully imputed!\n")
} else {
  cat("⚠ Some NA values remain. Please check the imputation logic.\n")
}
```

Input:

	Column1	Column2	Column3
1	35	NA	19
2	38	175	40
3	NA	191	74
4	27	62	NA
5	25	79	52
6	78	74	29
7	NA	162	39
8	73	89	NA
9	79	NA	47
10	90	165	24
11	83	171	13
12	43	77	25

Output after running code :

```
The downloaded binary packages are in
        /var/folders/xf/mjv2q8_x3tg828b8x3clxb3m0000gn/T//Rtmpga3ROm/downloaded_packages
   Column1 Column2 Column3
1       35      NA      19
2       38     175      40
3       NA     191      74
4       27      62      NA
5       25      79      52
6       78      74      29
7       NA     162      39
8       73      89      NA
9       79      NA      47
10      90     165      24
11      83     171      13
12      43      77      25

CSV file 'datainput.csv' created successfully!
File contains 12 records with 3 columns.
Each column has 2 NA values.
Imputing NA values...
Data after imputation:
# A tibble: 12 × 3
   Column1 Column2 Column3
     <dbl>   <dbl>   <dbl>
 1      35     191      19
 2      38     175      40
 3      90     191      74
 4      27      62      74
 5      25      79      52
 6      78      74      29
 7      83     162      39
 8      73      89      52
 9      79     175      47
10      90     165      24
11      83     171      13
12      43      77      25

Verification: Remaining NA values: 0
✓ All NA values have been successfully imputed!
```