Basic Data cleaning and export

Anuradha Vyas

2024-01-18

Contents

1. Load required packages

```
library(writexl) #to export excel sheet
library(dplyr)
library(openxlsx)
library(readxl) #to read excel
```

2. Create and Print Dataset

```
Original_Data <- data.frame(
  A = c(1, 2, 3, 4, 5),
  B = c("Red", "Blue", "Green", "Yellow", "Red"),
  C = c(10.5, 15.2, 8.7, 12.0, 9.3),
  D = c(TRUE, FALSE, TRUE, FALSE, TRUE),
  E = c("Monday", "Tuesday", "Wednesday", "Thursday", "Friday")
)

# Print the new dataset
print(Original_Data)
```

```
##   A      B    C     D         E
## 1 1    Red 10.5  TRUE    Monday
## 2 2   Blue 15.2 FALSE   Tuesday
## 3 3  Green  8.7  TRUE Wednesday
## 4 4 Yellow 12.0 FALSE  Thursday
## 5 5    Red  9.3  TRUE    Friday
```

3. Export Data to Excel

```
File_path <- "C:/Users/smile/Desktop/Original_data_1.xlsx"
write_xlsx(Original_Data, File_path)
```

4. Rename Columns

```
Original_Data_rename <- rename(Original_Data, id = A, color = B,
                        age = C, T_F = D,
                        Days = E)
Original_Data_rename
```

```
##   id  color  age   T_F      Days
## 1  1    Red 10.5  TRUE    Monday
## 2  2   Blue 15.2 FALSE   Tuesday
## 3  3  Green  8.7  TRUE Wednesday
## 4  4 Yellow 12.0 FALSE  Thursday
## 5  5    Red  9.3  TRUE    Friday
```

5.Count Data by Category

```
count(Original_Data_rename, id) #id is not converted to categorical
```

```
##   id n
## 1  1 1
## 2  2 1
## 3  3 1
## 4  4 1
## 5  5 1
```

```
count(Original_Data_rename, T_F)  # Count the occurrences of 'T_F'
```

```
##     T_F n
## 1 FALSE 2
## 2  TRUE 3
```

```
count(Original_Data_rename, T_F, color) # Count the occurrences of unique combinations of 'T_F' and 'co
```

```
##     T_F  color n
## 1 FALSE   Blue 1
## 2 FALSE Yellow 1
## 3  TRUE  Green 1
## 4  TRUE    Red 2
```

6. Simplify Color Column

```
Original_Data_rename <- mutate(Original_Data_rename, color.simplified = case_when(
  color == "Red" ~ "1",
  color == "Blue" ~ "2",
  color == "Green" ~ "3",
  color == "Yellow" ~ "4"
))

select(Original_Data_rename, c(color, color.simplified))
```

```
##    color color.simplified
## 1    Red               1
## 2   Blue               2
## 3  Green               3
## 4 Yellow               4
## 5    Red               1
```

7. Calculate Mean and SD by Group

```
Original_Data_rename
```

```
##   id  color  age   T_F      Days color.simplified
## 1  1    Red 10.5  TRUE    Monday               1
## 2  2   Blue 15.2 FALSE   Tuesday               2
```

```
## 3  3  Green  8.7  TRUE Wednesday                3
## 4  4 Yellow 12.0 FALSE  Thursday                4
## 5  5    Red  9.3  TRUE    Friday                1
```

```
Original_Data_rename |>
    mutate(midschool.school.complete = case_when(
      age< 11 ~ "No_midschool",
      age >11 ~ "Midschool")) |>
    group_by(midschool.school.complete) |>
    summarize(age.mean = mean(age), age.sd= sd(age))
```

```
## # A tibble: 2 x 3
##   midschool.school.complete age.mean age.sd
##   <chr>                        <dbl>  <dbl>
## 1 Midschool                     13.6  2.26
## 2 No_midschool                   9.5  0.917
```

```
#count(Original_Data_rename, midschool.school.complete)
```

8. Add New Data to Excel file

```
#create new data
New_data <- Original_Data |>select(A, B, C)
print(New_data)
```

```
##   A      B    C
## 1 1    Red 10.5
## 2 2   Blue 15.2
## 3 3  Green  8.7
## 4 4 Yellow 12.0
## 5 5    Red  9.3
```

```
# Load the existing workbook
WB <- loadWorkbook(File_path)

addWorksheet(WB, sheetName = "New_data")
writeData(WB, sheet = "New_data", x = New_data)
```

9. Add Additional Data to new sheet in same excel file

```
#create new data 2
Additional_data <- Original_Data |>select(A, D, E)
print(Additional_data)
```

```
##   A     D         E
## 1 1  TRUE    Monday
```

```
## 2 2 FALSE   Tuesday
## 3 3  TRUE Wednesday
## 4 4 FALSE  Thursday
## 5 5  TRUE    Friday
```

```
# Add the new_data to a new sheet in the existing workbook
addWorksheet(WB, sheetName = "Additional_data")
writeData(WB, sheet = "Additional_data", x = Additional_data)

# Save the modified workbook
saveWorkbook(WB, File_path, overwrite = TRUE)
```

10. Remove Columns and Combine Data

```
New_data_remove <- Original_Data |>select(-A, -B,-C)
print(New_data_remove)
```

```
##       D         E
## 1  TRUE    Monday
## 2 FALSE   Tuesday
## 3  TRUE Wednesday
## 4 FALSE  Thursday
## 5  TRUE    Friday
```

11. Edit Data and Save to new sheet in same excel file

```
# Add the new_data to a new sheet in the existing workbook

addWorksheet(WB, sheetName = "combined_data_NA")
writeData(WB, sheet = "combined_data_NA", x = cbind(Additional_data, New_data))

# Save the modified workbook
saveWorkbook(WB, File_path, overwrite = TRUE)
```

```
Edited_data <- Original_Data[-c(3,4),]
print(Edited_data)
```

```
##   A    B    C     D       E
## 1 1  Red 10.5  TRUE  Monday
## 2 2 Blue 15.2 FALSE Tuesday
## 5 5  Red  9.3  TRUE  Friday
```

```
addWorksheet(WB, sheetName = "Edited_data")
writeData(WB, sheet = "Edited_data", x = Edited_data)

# Save the modified workbook
saveWorkbook(WB, File_path, overwrite = TRUE)
```

12. Read Data from different Sheets in same excel file

```
#read data sheet 1
New_data_1 <- read_xlsx("C:/Users/smile/Desktop/Original_data_1.xlsx", sheet ="New_data")
print(New_data_1)
```

```
## # A tibble: 5 x 3
##       A B          C
##   <dbl> <chr>  <dbl>
## 1     1 Red     10.5
## 2     2 Blue    15.2
## 3     3 Green    8.7
## 4     4 Yellow  12
## 5     5 Red      9.3
```

```
#read data sheet 2
Additional_data_1 <- read_xlsx("C:/Users/smile/Desktop/Original_data_1.xlsx", sheet ="Additional_data")
print(Additional_data_1)
```

```
## # A tibble: 5 x 3
##       A D     E
##   <dbl> <lgl> <chr>
## 1     1 TRUE  Monday
## 2     2 FALSE Tuesday
## 3     3 TRUE  Wednesday
## 4     4 FALSE Thursday
## 5     5 TRUE  Friday
```

```
#read data sheet 1
combined_data_NA_1 <- read_xlsx("C:/Users/smile/Desktop/Original_data_1.xlsx", sheet ="combined_data_NA
```

```
## New names:
## * `A` -> `A...1`
## * `A` -> `A...4`
```

```
print(combined_data_NA_1)
```

```
## # A tibble: 5 x 6
##   A...1 D     E         A...4 B          C
##   <dbl> <lgl> <chr>     <dbl> <chr>  <dbl>
## 1     1 TRUE  Monday        1 Red     10.5
## 2     2 FALSE Tuesday       2 Blue    15.2
## 3     3 TRUE  Wednesday     3 Green    8.7
## 4     4 FALSE Thursday      4 Yellow  12
## 5     5 TRUE  Friday        5 Red      9.3
```