DM101_LabExam2

Odato,CJ

2025-02-21

1. Load these datasets and integrate them into one table. Save this as an Excel document using this syntax:

```
library(readxl)
SYNOPTIC_INFORMATION <- read_excel("C:/Users/Admin/Desktop/R/DM101_LabExam2/SYNOPTIC INFORMATION.xlsx")
METEOROLOGICAL_DATA_1 <- read_excel("C:/Users/Admin/Desktop/R/DM101_LabExam2/METEOROLOGICAL DATA_1.xlsx
HUMIDITY <- read_excel("C:/Users/Admin/Desktop/R/DM101_LabExam2/HUMIDITY.xlsx")
```

1. Load these datasets and integrate them into one table. Save this as an Excel document using this syntax:

```
library(openxlsx)
merged_table <- merge(HUMIDITY,METEOROLOGICAL_DATA_1, by = "SYNOPTIC STATION", all = TRUE)
merged_table <- merge(merged_table,SYNOPTIC_INFORMATION, by = "SYNOPTIC STATION", all = TRUE)
write.xlsx(merged_table, "C:/Users/Admin/Desktop/R/DM101_LabExam2/Combined Data.xlsx ")
```

2. Divide the new saved Excel dataset into three Excel datasets according to synoptic station.

You must do this because your group will conduct a separate data preprocessing for each synoptic station. Moreover, explain why we need to conduct a separate data preprocessing for each synoptic station. [5 points]

```
# Split data by Synoptic Station
station_A <- subset(merged_table,`SYNOPTIC STATION` == "A")
station_B <- subset(merged_table,`SYNOPTIC STATION` == "B")
station_C <- subset(merged_table,`SYNOPTIC STATION` == "C")

# Save each dataset separately
write.xlsx(station_A, "C:/Users/Admin/Desktop/R/DM101_LabExam2/Synoptic_Station_A.xlsx")
write.xlsx(station_B, "C:/Users/Admin/Desktop/R/DM101_LabExam2/Synoptic_Station_B.xlsx")
write.xlsx(station_C, "C:/Users/Admin/Desktop/R/DM101_LabExam2/Synoptic_Station_C.xlsx")
```

Each synoptic station collects weather data from different locations, leading to variations in temperature, humidity, and pressure. Separate preprocessing is needed to handle missing or inconsistent data, standardize formats, and prevent data contamination. This ensures accurate trend analysis, improves forecasting models, and enhances overall data reliability.

##Conduct data cleaning by estimating the missing values and detecting significant outliers foreach synoptic station data. Is it justifiable to completely remove the detected outliers [if thereare any]? Can your group suggest a better approach aside from simply removing these outliers[if there are any]? [15 points]

```r
library(zoo)
```

```
##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
library(openxlsx)

# Ensure numeric conversion for both TEMPERATURE and RAINFALL in Station C
station_A$TEMPERATURE <- as.numeric(as.character(station_A$TEMPERATURE))
station_B$TEMPERATURE <- as.numeric(as.character(station_B$TEMPERATURE))
station_C$TEMPERATURE <- as.numeric(as.character(station_C$TEMPERATURE))
station_C$RAINFALL <- as.numeric(as.character(station_C$RAINFALL))  # Convert RAINFALL

# Function to fill missing values with mean
fill_missing_values <- function(data, column_name) {
  if (!column_name %in% colnames(data)) {
    stop(paste("Column", column_name, "not found in dataset!"))
  }

  # Compute mean safely
  mean_value <- mean(data[[column_name]], na.rm = TRUE)

  # Replace missing values if mean is valid
  if (!is.na(mean_value)) {
    data[[column_name]][is.na(data[[column_name]])] <- mean_value
  }

  return(data)
}

# Apply function to Temperature column for each station
station_A <- fill_missing_values(station_A, "TEMPERATURE")
station_B <- fill_missing_values(station_B, "TEMPERATURE")
station_C <- fill_missing_values(station_C, "TEMPERATURE")
station_C <- fill_missing_values(station_C, "RAINFALL")  # Apply to RAINFALL

# Save cleaned datasets
write.xlsx(station_A, "C:/Users/Admin/Desktop/R/DM101_LabExam2/Synoptic_Station_A_Cleaned.xlsx")
write.xlsx(station_B, "C:/Users/Admin/Desktop/R/DM101_LabExam2/Synoptic_Station_B_Cleaned.xlsx")
write.xlsx(station_C, "C:/Users/Admin/Desktop/R/DM101_LabExam2/Synoptic_Station_C_Cleaned.xlsx")

#OUTLIERS
```