DATA471 Assignment 5

kakauchad_300212228

2022-10-13

Assignment info

1. Read in data
2. Initial Data Analysis
3. Preprocess data
4. carry out tasks
5. Answer the questions

Read in the data

The data comprises 11 csv files providing weather and climate information for three locations in the Hong Kong region: Wong Chuk Hang (HKS), Tseung Kwan O (JKB), King's Park (KP). Each location has a file for each of: 'Rainfall (mm)' (RF), 'Wind Speed (km/h/)' (WSPD), 'Maximum Temperature (degrees Celsius)' (CLMMAXT), 'Minimum Temperature (degrees Celsius)' (CLMMINT), 'Mean Temperature (degrees Celsius)' (CLMTEMP).

File structure

The first two rows of the csv data files include merged columns, with file information and need to be removed in order to access the file data. Each file also includes file summary information in the file 3 rows that can be removed. The column names include pictograms and need to be converted to ASCII characters for easier referencing, because my keyboard doesn't have those keys.

The datafile contains several variables including individual date variables ('day', 'month', 'year'), with the 'year' variable reading in as a character vector, 'month' and 'day' read in as integer types. It will be more convenient to combine the 'day', 'month', and 'year' variables into a single 'Date' variable. The 'value' column contains the specific value of a metric on a given day and is captured as a character vector, so needs to be converted to a numeric.

I'll begin by defining a function to read in a file, remove the first two rows, capture the column headers and convert to ASCII for easier use. The function will also create a new 'Date' column combining the 'day', 'month', and 'year' columns, in date format. The function converts the 'value' column to a numeric.

```
# create a function to read in the climate files
climate_in <- function(file, path) {
```

```
  # the first two lines of data are merged so just take data from line 3
  raw <- read.csv(paste(path, file, sep = '/'), skip = 2, header = T, )

  # collect the file headers
  headers  <- read.csv(paste(path, file, sep = '/'), skip = 2, header = F, nrows = 1 )

  # convert headers to ASCII
  headers <- iconv(headers, "latin1", "ASCII", sub = "")
  # remove the straggling "/"
  headers <- substring(headers, 2)


  # and add the headers as columns
  colnames(raw) <- headers

  # convert Value column to float
  raw$Value <- as.numeric(raw$Value)

  # collect date values into a date column
  raw$Date <- paste(raw$Day, raw$Month, raw$Year, sep ="/")
  # convert Data column to date format
  raw$Date <- dmy(raw$Date)

  # Drop redundant columns... by keeping relevant columns
  raw <- raw[ ,c("Date", "Value", "data Completeness")]
  # and drop the last three rows
  raw <- raw[!is.na(raw$Date) , ]


  return(raw)
}
```

Having created the function, can collect some of the known file information to set up a loop to grab information from each file:

- location information
- metric type
- unit of measure

we can use this information to add columns for location, metric, and one for unit of measure.

```
# read in weather data from three regions in HK
# pre-load the location of files
path <- './DATA471_A5_datasets'
# collect names of data files
csv_files <- list.files(path)

# collect data_file information
# data files relate to locations:  collect the location names
locs <- c("Wong_Chuk_Hang", "Tseung_Kwan_O", "Kings_Park", "Kings_Park")
# and the short reference on the csv file
```

```
locs_ref <- c("HKS", "JKB", "_KP", "ALL")

# data files relate to a metric:  collect the metrics
measure <- c("Rainfall_mm", "Windspeed_kmh", "MaxTemp_C", "MinTemp_C", "MeanTemp_C")
# and the short reference on the csv file
measure_ref <- c("daily_RF", "daily_WS", "CLMMAXT_", "CLMMINT_", "CLMTEMP_")
```

```
# create for loop to grab data from each location
# add location, metric, and units as
full_data <- data.frame(matrix(ncol = 5, nrow = 0))
f_names <- c("Date", "Value", "data_Completeness", "Location", "Measure")
colnames(full_data) <- f_names

for (df in csv_files) {

  # collect some feature references
  n <- nchar(df)
  met <- substr(df, 1, 8)
  loc <- substr(df, n-6, n-4)

  # determine the measure:
  metric <- measure[which(measure_ref == met)]
  # determine the location
  station <- locs[which(locs_ref == loc)]

  # read data into a variable
  holding <- climate_in(df, path)
  holding$Location <- station
  holding$Measure <- metric

  full_data <- rbind(full_data, holding)
  rm(holding)
}
```

Print a summary of the data

```
# check the summary
summary(full_data)
```

```
##       Date                Value        data Completeness    Location
##  Min.   :1989-08-01   Min.   :  0.00   Length:125043      Length:125043
##  1st Qu.:1999-04-29   1st Qu.: 15.70   Class :character   Class :character
##  Median :2007-02-08   Median : 22.50   Mode  :character   Mode  :character
##  Mean   :2007-01-29   Mean   : 20.55
##  3rd Qu.:2014-11-20   3rd Qu.: 26.90
##  Max.   :2022-08-31   Max.   :324.00
##                       NA's   :1734
##    Measure
##  Length:125043
##  Class :character
##  Mode  :character
##
##
```

```
##
##
```

```r
# factorise and re-check the summary
full_data[, 3] <- as.factor(full_data[, 3])
full_data[, 4] <- as.factor(full_data[, 4])
full_data[, 5] <- as.factor(full_data[, 5])

summary(full_data)
```

```
##       Date                Value         data Completeness           Location
##  Min.   :1989-08-01   Min.   :  0.00    :   600          Kings_Park    :55095
##  1st Qu.:1999-04-29   1st Qu.: 15.70   #: 21703          Tseung_Kwan_O :33696
##  Median :2007-02-08   Median : 22.50   C:102740          Wong_Chuk_Hang:36252
##  Mean   :2007-01-29   Mean   : 20.55
##  3rd Qu.:2014-11-20   3rd Qu.: 26.90
##  Max.   :2022-08-31   Max.   :324.00
##                       NA's   :1734
##           Measure
##  MaxTemp_C    :34335
##  MeanTemp_C   :34335
##  MinTemp_C    :34335
##  Rainfall_mm  :11019
##  Windspeed_kmh:11019
##
##
```